دانشگاه گیلان
۱۳۵۳_1974

چاپ اول

ژنتیک آماری صفات کمّی

پیوستگی ژنی، نقشه های ژنتیکی و QTL

تألیف:
رونلینگ وو چانگ-ژینگ ما جورج کاسلا
ترجمه:
دکتر نوید قوی حسین زاده
استاد دانشکده علوم کشاورزی دانشگاه گیلان
دکتر بابک ربیعی
استاد دانشکده علوم کشاورزی دانشگاه گیلان

اداره چاپ و انتشارات دانشگاه گیلان

ژنتیک آماری صفات کمّی
پیوستگی ژنی، نقشه‌های ژنتیکی و QTL

تألیف: رونلینگ وو چانگ-ژینگ ما جورج کاسلا
ترجمه: دکتر نوید قوی حسین زاده، دکتر بابک ربیعی

Statistical Genetics of Quantitative Traits

Linkage, Maps and QTL

By:
Rongling Wu Chang-Xing Ma George Casella
Translated by:
Navid Ghavi Hossein-Zadeh, Ph. D
Babak Rabiei, Ph. D

University of Guilan Press

هدف این کتاب فراهم آوردن ابزارهای مورد نیاز برای درک و مدل‌سازی تنوع ژنتیکی صفات کمّی بر اساس داده‌های ژنومی جمع‌آوری شده در تحقیقات مکان‌یابی برای متخصصان علم ژنتیک و مجهز کردن متخصصان علم آمار به وحدت نظر و ایده‌هایی در ارتباط با کشف رمز و رازهای ژنتیکی با استفاده از مهارت‌های محاسباتی آن‌ها است. همچنین این کتاب قصد دارد تا توجه محققان را به سوی انجام تحقیقات چند رشته‌ای جلب نموده و آنها را وارد نمونه‌ها و الگوهای جدیدی در علم ژنومیک نماید. در این کتاب، نظریه‌های آماری و محاسباتی که برای مکان‌یابی ژنتیکی به‌کار گرفته می‌شوند، مرحله به مرحله توسعه می‌یابند و به همراه مثال‌هایی برای نشان دادن کاربردهای ژنومیک آماری معرفی می‌شوند.

9786001532207
ISBN: 978-600-153-220-7

بسم الله الرحمن الرحیم

ژنتیک آماری صفات کمی

پیوستگی ژنی، نقشه‌های ژنتیکی و QTL

تألیف:

رونلینگ وو چانگ - ژینگ‌ما جورج کاسلا

ترجمه:

دکتر نوید قوی‌حسین‌زاده

استاد دانشکده علوم کشاورزی دانشگاه گیلان

دکتر بابک ربیعی

استاد دانشکده علوم کشاورزی دانشگاه گیلان

انتشارات دانشگاه گیلان

۱۳۹۹

دانشگاه گیلان
۱۳۵۳-1974

شابک: ۷ ـ۲۲۰ ـ ۱۵۳ ـ ۶۰۰ ـ ۹۷۸

سرشناسه	: وو، رونگلینگ Wu, Rongling
عنوان و نام پدیدآور	: ژنتیک آماری صفات کمی، پیوستگی ژنی، نقشه‌های ژنتیکی و QTL/ تالیف رونلینگ وو، چانگ - ژینگ‌ما، جورج کاسلا ؛ ترجمه نوید قوی‌حسین‌زاده، بابک ربیعی ؛ ویراستار علمی حبیب الله سمیع‌زاده‌لاهیجی.
مشخصات نشر	: رشت: دانشگاه گیلان، ۱۳۹۸.
مشخصات ظاهری	: ۴۳۹ص.: جدول، نمودار؛ ۲۱/۵×۱۴/۵ س‌م.
شابک	: 978-600-153-220-7
وضعیت فهرست نویسی	: فیپا
یادداشت	: عنوان اصلی: Statistical genetics of quantitative traits : linkage, maps and QTL.
یادداشت	: چاپ قبلی: جهاد دانشگاهی، واحد صنعتی اصفهان، مرکز انتشارات، ۱۳۹۷.
یادداشت	: کتابنامه:ص.[۴۲۰] - ۴۳۴.
یادداشت	: واژه‌نامه.
موضوع	: ژنتیک -- روش‌های آماری
موضوع	: Genetics -- Statistical methods
موضوع	: ژن نگاشت
موضوع	: Gene mapping
شناسه افزوده	: ما، چانگ-شینگ
شناسه افزوده	: Ma, Chang-Xing
شناسه افزوده	: کسلا، جورج، ۱۹۵۱ - م.
شناسه افزوده	: Casella, George
شناسه افزوده	: قوی حسین‌زاده، نوید، ۱۳۵۹ -۰۹۱۱۳۳۹۱۱۰۱
شناسه افزوده	: ربیعی، بابک، ۱۳۴۸ - ۰۹۱۱۱۳۷۷۹۴۴
شناسه افزوده	: سمیع‌زاده لاهیجی، حبیب‌الله، ۱۳۴۵ -، ویراستار
شناسه افزوده	: دانشگاه گیلان
رده بندی کنگره	: QH۴/۴۳۸
رده بندی دیویی	: ۵۷۶/۵۳
شماره کتابشناسی ملی	: ۶۰۳۶۰۵۵

اداره چاپ و انتشارات دانشگاه گیلان

نام کتاب	:	ژنتیک آماری صفات کمی، پیوستگی ژنی، نقشه های ژنتیکی و QTL
مؤلفان	:	رونلینگ وو، چانگ-ژینگ‌ما، جورج کاسلا
مترجمان	:	دکتر نوید قوی‌حسین‌زاده، دکتر بابک ربیعی
ویراستار علمی	:	دکتر حبیب‌اله سمیع‌زاده‌لاهیجی
ویراستار ادبی	:	هوشنگ سپهری
نوبت چاپ	:	اول، ۱۳۹۹
ناشر	:	انتشارات دانشگاه گیلان
شمارگان		۱۰۰۰ جلد

هر گونه چاپ و تکثیر صرفاً در اختیار انتشارات دانشگاه گیلان است.

دیباچه

بیشتر صفات موجود در طبیعت و حائز اهمیت در کشاورزی به‌صورت کمّی به ارث می‌رسند. مطالعهٔ این صفات به‌دلیل ماهیت پیچیده وراثت آن‌ها مشکل است. با این حال، پیشرفت‌های اخیر در فناوری‌های ژنومی، مفاهیم بنیادی و اساسی برای آشکارسازی راز و رمزهای موجود در تنوع ژنتیکی صفات کمّی را فراهم کرده است. فنّاوری‌های ژنومی، ویژگی‌های مولکولی نشانگرهای چندشکل را در سرتاسر ژنوم تعیین می‌کنند و سپس برای شناسایی و مکان‌یابی ژن‌ها یا جایگاه‌های ژنی صفات کمّی (QTL) بر اساس تجزیهٔ پیوستگی استفاده می‌شوند.

تجزیهٔ آماری ابزار مهمی برای تجزیهٔ داده‌های ژنومی است که در حال حاضر تا حد زیادی برای انواعی از گونه‌ها در دسترس قرار گرفته و توضیحات دقیقی در ارتباط با تنوع ژنتیکی صفات کمّی در بین گونه‌ها، جمعیت‌ها، خانواده‌ها و افراد ارائه داده است. «لندر و بوتستین»[1] (۱۹۸۹) مقاله روش شناختی با ارزشی در خصوص مکان‌یابی فاصله‌ای منتشر کردند که متخصصان علم ژنتیک را قادر می‌سازد تا تک QTLهایی که فنوتیپ یک صفت را کنترل می‌کنند، شناسایی و آثار آن‌ها را ارزیابی کنند. امروزه، مکان‌یابی فاصله‌ای ابزار آماری مهمی جهت مطالعهٔ ژنتیکی صفات کمّی در سطح مولکولی است و منجر به کشف هزاران QTL مسئول ایجاد تنوع در گیاهان، حیوانات و انسان‌ها شده است. در مطالعه‌ای که به‌تازگی در مجله *Science* به چاپ رسید، «لی»[2] و همکاران (۲۰۰۶) قادر شدند تا اساس مولکولی کاهش ریزش دانه غلات (فرایندی پایه‌ای و اولیه در انتخاب برای اهلی شدن برنج) را در QTL شناسایی شده با روش مکان‌یابی فاصله‌ای تشخیص دهند. در میان بسیاری از مثال‌های دیگر در خصوص موفقیت مکان‌یابی فاصله‌ای، همسانه‌سازی وضعیتی QTLهای مرتبط با اندازه و شکل میوه در گوجه فرنگی (فراری و همکاران[3]، ۲۰۰۰) و QTLهای کنترل کننده پنجه، بازتابش نور و معماری دانه در ذرت (دابلی و همکاران[4]، ۱۹۹۷؛ گالاوتی و همکاران[5]، ۲۰۰۴؛ وانگ و همکاران[6]، ۲۰۰۵) قابل ذکر هستند.

برای اینکه مکان‌یابی فاصله‌ای در کاربردهای عملی مختلف استفاده شود، این روش در طی ۱۵ سال گذشته تا حد زیادی تغییر یافته و اصلاح شده است. روش‌های آماری مناسبی برای مکان‌یابی

---

[1] - Lander and Botstein
[2] - Li *et al.*
[3] - Frary *et al.*
[4] - Doebly *et al.*
[5] - Gallavotti *et al.*
[6] - Wang *et al.*

QTL از راه تلاش‌های گروهی متخصصان ژنتیک آماری معرفی شده‌اند؛ اما این روش‌ها به‌طورکلی اهداف و کاربردهای مختلفی داشته و به‌صورت پراکنده در بسیاری از مقالات ارائه شده‌اند. وجود کتابی که پیشرفت‌های آماری مکان‌یابی ژنتیکی را توضیح دهد، برای بسیاری از محققان به‌ویژه افرادی که علاقه وافری به ایجاد ارتباط بین ژنتیک و آمار دارند، می‌تواند مفید بوده و آن‌ها را با این حوزه در حال گسترش و پیشرفت هر چه سریع‌تر آشنا نماید.

هدف این کتاب فراهم آوردن ابزارهای مورد نیاز برای درک و مدل‌سازی تنوع ژنتیکی صفات کمّی بر اساس داده‌های ژنومی جمع‌آوری شده در تحقیقات مکان‌یابی برای متخصصان علم ژنتیک و مجهز کردن متخصصان علم آمار به وحدت نظر و ایده‌هایی در ارتباط با کشف رمز و رازهای ژنتیکی با استفاده از مهارت‌های محاسباتی آن‌ها است. همچنین این کتاب قصد دارد تا توجه محققان را به سوی انجام تحقیقات چند رشته‌ای جلب کند و آن‌ها را وارد نمونه‌ها و الگوهای جدیدی در علم ژنومیک نماید. در این کتاب، نظریه‌های آماری و محاسباتی که برای مکان‌یابی ژنتیکی به‌کار گرفته می‌شوند، مرحله به مرحله توسعه می‌یابند و به همراه مثال‌هایی برای نشان دادن کاربردهای ژنومیک آماری معرفی می‌شوند.

این کتاب ۱۴ فصل دارد و به‌طورکلی به سه بخش تقسیم می‌شود. بخش اول، مشتمل بر فصول ۱ و ۲، مقدمه مناسبی در خصوص ژنتیک و آمار برای درک مفاهیم عمومی ژنتیک و روش‌های آماری جهت مکان‌یابی ژنتیکی فراهم می‌آورد. بخش دوم، مشتمل بر فصول ۳ تا ۷، سعی می‌کند تا پوشش جامع و کاملی در خصوص تجزیه پیوستگی ژن‌ها (لینکاژ) با استفاده از نشانگرهای مولکولی ارائه کند. در این بخش، مدل‌ها و روش‌های مورد استفاده برای تجزیه پیوستگی و تهیه نقشه به‌طور سیستماتیک برای طرح‌های مختلف مانند بک کراس/F_2 (فصل ۳)، تلاقی‌های آزاد[۱] (فصل ۴)، لاین‌های خویش‌آمیخته نوترکیب یا RILs[۲] (فصل ۵) و شجره‌های ساختاری[۳] یا با ساختار مشخص (فصل ۷) و برای انواع خاصی از نشانگرها مانند نشانگرهای فاقد طبقه‌بندی و بی‌شکل[۴] (فصل ۶) و نشانگرهای غالب (فصل ۷) معرفی شده‌اند. بخش سوم، مشتمل

---

[۱]- منظور از تلاقی‌های آزاد، جمعیت‌های حاصل از تلاقی آزاد ژنوتیپ‌های مختلف که در گیاهان دگرگشن بیشتر دیده می‌شود است (مترجمین).

[۲]- لاین‌های خویش‌آمیخته نوترکیب یکی از جمعیت‌های مهم برای مکان‌یابی QTL ها و تهیه نقشه‌های پیوستگی هستند و منظور از آنها جمعیت‌های F_2 به بعد است (مترجمین).

[۳]- شجره‌های ساختاری جمعیت‌هایی هستند که شجره‌نامه آنها کاملاً مشخص است و می‌توان شجره‌نامه افراد را تا چندین نسل قبل‌تر از نتاج جمعیت مورد مطالعه رسم نمود (مترجمین).

[۴]- منظور از نشانگرهای فاقد طبقه‌بندی و بی‌شکل، نشانگرهای دارای انحراف از تفرق مندلی است. به دلایل ژنتیکی یا زیستی ممکن است بعضی از نشانگرها از نسبت‌های تفرق مندلی (۱:۲:۱ برای نشانگرهای هم بارز و ۱:۱ برای

بر فصول ۸ تا ۱۴، مدل‌های آماری و الگوریتم‌های مکان‌یابی QTL را تحت پوشش قرار می‌دهد. عناوین این بخش شامل تجزیه ارتباطی ساده فنوتیپ-نشانگر (فصل ۸)، ساختار آماری مکان‌یابی فاصله‌ای (فصل ۹)، مکان‌یابی فاصله‌ای بر اساس تجزیه تابعیت یا رگرسیون (فصل ۱۰) و بر اساس روش حداکثر درستنمایی (فصل ۱۱)، تعیین فاصله اطمینان و آستانه (فصل ۱۲)، مکان‌یابی فاصله‌ای مرکب با استفاده از نشانگرهای چندگانه به عنوان متغیر همراه (فصل ۱۳) و مکان‌یابی فاصله‌ای برای مکان‌یابی جمعیت‌های غیرهمخون (فصل ۱۴) هستند. در بخش ضمایم، نظریه‌های آماری عمومی که به‌طور مستقیم در ارتباط با روش‌های مکان‌یابی ژنتیکی هستند و برنامه‌های زبان R برای تعدادی از مثال‌های استفاده شده در این کتاب، ارایه شده‌اند. صفحه وب (http://www.buffalo.edu/~cxma/book/) نیز برای این کتاب تهیه شده و فهرست کاملی از برنامه‌ها و الگوریتم‌های نوشته شده در برنامه MatLab یا R برای تمامی مثال‌ها ارائه شده است.

نوشتن کتاب در چنین حوزه‌ای که به سرعت در حال گسترش و تغییر است مشکل است، ولی اگر بخواهیم دقیق‌تر صحبت کنیم، دارای هیجان زیادی است. در تابستان سال ۱۹۹۷، «وو»[1] یکسری از سخنرانی‌ها را در خصوص روش‌های آماری شناسایی QTLها برای دانشجویان تحصیلات تکمیلی و دانشکده در دانشگاه جنگلداری Nanjing چین ایراد نمود. در ترم بهاره ۲۰۰۲، Wu دوره ژنتیک آماری را در سطح کارشناسی ارشد در دانشگاه فلوریدا تدریس نمود و سپس در بهار ۲۰۰۳ به‌عنوان همکار مدرس به کاسلا[2] و در بهار ۲۰۰۵ به Ma پیوست (کاسلا و Ma از نویسندگان کتاب حاضر هستند). این دوره در حال حاضر به وسیله Wu در دانشگاه فلوریدا و به وسیله Ma در دانشگاه ایالتی نیویورک در Buffalo به‌صورت منظم تدریس می‌شود. همهٔ ما تاکنون سخنرانی‌ها یا دوره‌های کوتاه مدت متعددی در ارتباط با ژنتیک آماری در مکان‌ها و زمان‌های مختلف ارائه نموده‌ایم. نظر به شور و اشتیاق دانشجویان و حضار در هر زمان و مکانی به شدت تحت‌تأثیر قرار گرفتیم تا در خصوص این حوزه جذاب آگاهی کسب کنیم. تمامی این عوامل ما را تشویق کرد تا کتابی بنویسیم که بتواند روش‌های پایه‌ای تحقیقات ژنتیک آماری را پوشش دهد. مفاهیم، مدل‌ها و الگوریتم‌های مرتبط با مکان‌یابی ژنتیکی به وسیله تعداد زیادی از نویسندگان در مجلات آماری و ژنتیکی مختلفی منتشر شده‌اند، ولی بخشی از مطالب موجود در این کتاب از برنامه‌های تحقیقاتی مشترک در طی ۵ سال گذشته به‌دست آمده‌اند. بدین وسیله از

نشانگرهای غالب) پیروی نکنند و نسبت‌های ژنوتیپی متفاوتی در جمعیت مورد مطالعه داشته باشند. اینگونه نشانگرها در این کتاب به نشانگرهای بی‌شکل و فاقد طبقه‌بندی معرفی شده‌اند (مترجمین).

[1] - Wu
[2] - Casella

آن گروه از نویسندگانی که نقش آن‌ها به‌دلیل محدودیت مطالب، در این کتاب ذکر نشده است، به‌طور ویژه عذرخواهی می‌کنیم.

طی نوشتن این کتاب بسیاری از همکاران و دوستان ما در دانشگاه فلوریدا و خارج از آن، کمک‌های با ارزشی را از جنبه‌های مختلف نمودند. Wu به گرمی از استاد راهنمای دوره فوق دکترای خود، دکتر Zhao-Bang Zeng در دانشگاه کارولینای شمالی برای راهنمایی‌های چشمگیرش جهت آموختن ژنتیک آماری تشکرمی‌کنم. دکتر Bruce Walsh در دانشگاه آریزونا، بازخوانی‌های متفکرانه و با ارزشی از متن کتاب در طی مراحل مختلف انجام داد. چندین داور ناشناس اظهار نظرات سازنده‌ای را ارائه دادند که به‌طور قابل‌توجهی شکل ارائه این کتاب را بهبود بخشیدند. دانشجویان یا افراد فوق دکترایی که در سخنرانی‌ها و کلاس‌های ما حاضر شدند یا با ما در مکان‌های مختلف در زمینه ژنتیک آماری کار می‌کنند، پیشنهادات خلاقانه زیادی ارائه داده‌اند که سبب بهبود شکل ارائه این کتاب شده است. از دانشجویان یا افراد فوق دکترای ذیل که در گذشته یا در حال حاضر در این گروه حضور داشته یا دارند تشکر ویژه می‌شود: Yuehua Cui، Wei Hou، Hongying Li، Min Lin، Tian Liu، Fei Long، Xiang-Yang Lou، Qing Lu، Damaris Santana، Zhaojie Wang، ZuohengWang، JiashengWu، SongWu، Jie Yang، John Yap، Li Zhang، Wei Zhao و Yun Zhu. داده‌های مورد استفاده برای مثال‌های ارائه شده در این کتاب با محبت فراوان به وسیله دکتر James Cheverud در دانشگاه واشینگتن (در رابطه با موش)، دکتر Junyi Gai در دانشگاه کشاورزی Nanjing (برای سویا)، Rory Todhunter در دانشگاه کورنل (داده‌های مربوط به سگ)، دکتر Stan Wullschleger و دکتر Tongming Yin در Oak Ridge National Laboratory (برای درخت صنوبر) و دکتر Jun Zhu در دانشگاه Zhejiang (در مورد برنج) فراهم شده است.

از گروه آمار و مؤسسه علوم کشاورزی و مواد غذایی دانشگاه فلوریدا برای نوشتن این کتاب و اجرای برنامه‌های تحقیقاتی خویش سپاسگزاریم. در پایان، مدیون و ممنون خانواده‌های محترممان به‌دلیل حمایت مداومشان از فعالیت‌های تحقیقاتی خود در طی سال‌های متمادی هستیم. این کتاب تا حدودی با گرانت NSF به شماره ۰۵۴۰۷۴۵ حمایت شده است.

Gainesville, FL
Rongling Wu
Buffalo, NY
Chang-Xing Ma
George Casella
December 2006

عنوان **صفحه**

فصل اول: ژنتیک پایه

۱-۱- مقدمه ۲
۱-۲- ژن‌ها و کروموزوم‌ها ۲
۱-۳- میوز ۳
۱-۴- قوانین مندل ۶
۱-۴-۱- قانون اول مندل ۶
۱-۴-۲- قانون دوم مندل ۷
۱-۵- پیوستگی ژنی و مکان‌یابی ۸
۱-۶- تداخل ۱۲
۱-۷- ژنتیک کمّی ۱۳
۱-۷-۱- ویژگی‌های ژن‌ها در جمعیت ۱۳
۱-۷-۲- یک مدل عمومی ژنتیک کمّی ۱۴
۱-۷-۳- مدل‌های ژنتیکی برای جمعیت‌های F_2 و تلاقی برگشتی ۱۵
۱-۷-۴- مدل اپیستازی ۱۷
۱-۷-۵- وراثت پذیری و برآورد آن ۱۹
۱-۷-۶- معماری ژنتیکی ۲۱
۱-۷-۷- برآورد تعداد ژن ۲۱
۱-۸- ژنتیک مولکولی ۲۳
۱-۹- SNP ۲۷
۱-۱۰- تمرین‌ها ۲۸
۱-۱۱- نکات ۲۹
۱-۱۱-۱- ایجاد مدل آثار ژنتیکی نامساوی با تابع گاما ۲۹
۱-۱۱-۲- ایجاد مدل آثار ژنتیکی نامساوی با سری‌های هندسی ۳۱

فصل دوم: آمار پایه

۲-۱- مقدمه ۳۴

۲-۱-۱- جمعیت‌ها و مدل‌ها ۳۴
۲-۱-۲- نمونه‌ها ۳۶
۲-۲- برآورد درستنمایی ۳۷
۲-۳- آزمون فرض ۴۲
۲-۳-۱- آزمون مربع کای پیرسون ۴۲
۲-۳-۲- آزمون‌های نسبت درستنمایی ۴۵
۲-۳-۳- روش مبتنی بر شبیه‌سازی ۵۰
۲-۳-۴- برآورد بیزین ۵۲
۲-۴- تمرین‌ها ۵۸

فصل سوم: تجزیهٔ پیوستگی و ایجاد نقشه ژنتیکی

۳-۱- مقدمه ۶۰
۳-۲- طرح آزمایش ۶۱
۳-۳- تفرق مندلی ۶۲
۳-۳-۱- آزمون الگوهای تفرق نشانگر ۶۳
۳-۴- الگوهای تفرق در یک خانواده تنی ۶۵
۳-۵- تجزیه دو نقطه‌ای ۶۸
۳-۵-۱- تلاقی برگشتی مضاعف ۶۹
۳-۵-۲- تلاقی‌های مضاعف بین افراد- F_2 ..... ۷۲
۳-۶- تجزیه سه نقطه‌ای ۷۸
۳-۷- درستنمایی چند مکانی و مرتب‌سازی مکان‌های ژنی ۸۰
۳-۸- برآورد با تعداد زیادی جایگاه ژنی ۸۴
۳-۹- درستنمایی‌های مختلط و احتمالات ترتیب ۸۶
۳-۱۰- توابع تهیه نقشه ۸۸
۳-۱۰-۱- فرمول ماتر ۸۸
۳-۱۰-۲- تابع نقشه مورگان ۸۹
۳-۱۰-۳- تابع نقشه هالدین ۹۰

۳-۱۰-۴- تابع نقشه کوسامبی ۹۲
۳-۱۱- تمرین‌ها ۹۶
۳-۱۲- نکات: الگوریتم‌ها و نرم‌افزارهای تهیه نقشه ۹۷
فصل چهارم: مدل کلی برای تجزیهٔ پیوستگی در تلاقی‌های کنترل شده
۴-۱- مقدمه.....۱۰۶
۴-۲- نشانگرهای حاوی اطلاعات کامل: یک مدل دیپلوتیپی ۱۰۷
۴-۲-۱- تجزیه دو نقطه‌ای ۱۰۷
۴-۲-۲- یک فرمول کلی‌تر.....۱۱۲
۴-۲-۳- تجزیه سه نقطه‌ای ۱۱۴
۴-۲-۴- ارائهٔ فرمول کلی‌تر ۱۱۹
۴-۳- نشانگرهای حاوی اطلاعات کامل: مدل ژنوتیپی ۱۲۲
۴-۳-۱- دیپلوتیپ‌های والدینی ۱۲۲
۴-۴- مدل‌سازی توأم برای پیوستگی، دیپلوتیپ والدینی و ترتیب ژنی ۱۲۷
۴-۵- نشانگرهای حاوی اطلاعات جزیی.....۱۳۳
۴-۵-۱- مدل‌سازی توأم پیوستگی و دیپلوتیپ والدینی ۱۳۳
۴-۵-۲- مدل‌سازی توأم پیوستگی، دیپلوتیپ والدینی و ترتیب ژنی ۱۳۵
۴-۶- تمرین‌ها ۱۳۷
۴-۶-۱- یک نشانگر غالب ۱۳۹
۴-۶-۲- یک نشانگر غالب و یک نشانگر هم بارز F_2.....۱۳۹
۴-۴- دو نشانگر غالب.....۱۳۹
۴-۷- نکات ۱۳۹
۴-۷-۱- تجزیه پیوستگی.....۱۴۰
۴-۷-۲- احتمال دیپلوتیپی ۱۴۲
۴-۷-۳- ترتیب ژنی ۱۴۳
۴-۷-۴- تجزیه M نقطه‌ای ۱۴۳

فصل پنجم: تجزیهٔ پیوستگی با استفاده از لاین‌های اینبرد نوترکیب

۵-۱- مقدمه ۱۴۸
۵-۲- RIL های حاصل از خویش‌آمیزی ۱۴۹
۵-۲-۱- تجزیه دو نقطه‌ای ۱۴۹
۵-۲-۲- تجزیه سه نقطه‌ای ۱۵۳
۵-۳- RIL های حاصل از آمیزش‌های برادر- خواهری ۱۶۱
۵-۳-۱- تجزیه دو نقطه‌ای ۱۶۱
۵-۳-۲- تجزیه سه نقطه‌ای ۱۶۲
۵-۴- کاهش اریبی ۱۶۳
۵-۴-۱- RIL های حاصل از خودگشنی ۱۶۳
۵-۴-۲- RIL های حاصل از آمیزش برادر- خواهری ۱۶۵
۵-۵- RIL های چند مسیری ۱۶۵
۵-۶- تمرین‌ها ۱۶۷
۵-۷- نکته ۱۶۸

فصل ششم: تجزیهٔ پیوستگی نشانگرهای اریب و با دسته‌بندی نادرست

۶-۱- مقدمه ۱۷۲
۶-۲- تنوع گامتی در زنده‌مانی ۱۷۲
۶-۲-۱- مدل تک ژنی ۱۷۲
۶-۲-۲- مدل دو ژنی ۱۷۷
۶-۲-۳- شبیه‌سازی ۱۸۲
۶-۳- تنوع در زنده‌مانی زیگوتی ۱۸۴
۶-۳-۱- مدل تک ژنی ۱۸۴
۶-۳-۲- مدل دو ژنی ۱۸۵
۶-۳-۳- شبیه‌سازی ۱۸۷
۶-۴- دسته‌بندی نادرست ۱۸۸
۶-۴-۱- مدل تک ژنی ۱۸۸

۶-۲-۴- مدل دو ژنی ۱۹۲
۶-۵- شبیه‌سازی ۱۹۷
۶-۶- تمرین‌ها ۲۰۰
۶-۱- مدل تک ژنی در F_2 ..... ۲۰۰
۶-۲- مدل دو ژنی در F_2 ..... ۲۰۱
فصل هفتم: ملاحظات خاص در تجزیهٔ پیوستگی
۷-۱- مقدمه ۲۰۴
۷-۲- تجزیه پیوستگی در شجره‌های پیچیده ۲۰۴
۷-۲-۱- یک خانواده هسته ۲۰۴
۷-۲-۲- برآورد چند نقطه‌ای اشتراک‌گذاری آلل‌های یکسان اجدادی ۲۰۹
۷-۲-۳- یک شجره پیچیده ۲۱۱
۷-۳- تجزیه اطلاعات نشانگرهای غالب ۲۲۵
۷-۳-۱- مقدمه ۲۲۵
۷-۳-۲- تجزیه تفرق ۲۲۵
۷-۳-۳- تجزیهٔ پیوستگی ۲۳۲
۷-۴- تمرین‌ها ۲۳۷
فصل هشتم: تجزیهٔ فنوتیپ‌های نشانگر
۸-۱- مقدمه ۲۴۲
۸-۲- مدل تابعیت QTL ۲۴۳
۸-۳- تجزیه در نشانگر ۲۴۶
۸-۳-۱- آزمون t دو نمونه‌ای ۲۴۶
۸-۳-۲- تجزیه واریانس ۲۴۹
۸-۳-۳- تجزیه ژنتیکی ۲۵۲
۸-۴- دور شدن از نشانگر ۲۵۵
۸-۴-۱- درستنمایی ۲۵۵
۸-۵- محاسبه توان آزمون ۲۶۱

۸-۶- تجزیه اثر متقابل نشانگر ۲۶۳
۸-۶-۱- تجزیه واریانس (ANOVA) ۲۶۳
۸-۶-۲- تجزیه ژنتیکی ۲۶۹
۸-۷- تجزیهٔ نشانگری کل ژنوم ۲۷۳
۸-۸- تمرین‌ها ۲۷۸

فصل نهم: ساختار مکان‌یابی QTL

۹-۱- مقدمه ۲۸۴
۹-۲- مدل مختلط ۲۸۵
۹-۲-۱- ارائه فرمول ۲۸۵
۹-۲-۲- ساختار، تنظیم و برآورد ۲۸۸
۹-۳- ساختار ژنتیکی جمعیت مدل مختلط ۲۸۹
۹-۳-۱- تلاقی برگشتی/F_2 ..... ۲۹۰
۹-۳-۲- دگرآمیزی یا دگرباروری ۲۹۰
۹-۳-۳- لاین‌های اینبرد نوترکیب (RIL) ۲۹۱
۹-۳-۴- جمعیت‌های طبیعی ۲۹۱
۹-۴- ساختار ژنتیک کمّی مدل مختلط ۲۹۱
۹-۴-۱- مدل افزایشی-غالبیت ۲۹۲
۹-۴-۲- مدل افزایشی-غالبیت-اپیستازی ۲۹۴
۹-۴-۳- مدل اپیستازی ضرب‌پذیر ۲۹۵
۹-۴-۴- مدل مکانیسمی ۲۹۷
۹-۵- تنظیم آزمایشی مدل مختلط ۲۹۸
۹-۶- برآورد در مدل مختلط ۲۹۹
۹-۷- الگوریتم‌های محاسباتی برای مدل مختلط ۳۰۲
۹-۷-۱- الگوریتم EM ۳۰۲
۹-۷-۲- EM مونت کارلو ۳۰۲
۹-۷-۳- EM اتفاقی ۳۰۳

۹-۷-۴- یک الگوریتم EM/آمیخته نیوتن-رافسون ۳۰۳
۹-۷-۵- برخی از جنبه‌های احتیاطی..... ۳۰۵
۹-۷-۶- روش‌های بیزی..... ۳۰۵
۹-۷-۷- برآورد تعداد اجزاء در یک مدل مختلط..... ۳۰۸
۹-۸- تمرین‌ها ۳۱۰

فصل دهم: مکان‌یابی فاصله‌ای با استفاده از تجزیه تابعیت

۱۰-۱- مقدمه..... ۳۱۲
۱۰-۲- مدل تابعیت خطی ۳۱۳
۱۰-۳- مکان‌یابی فاصله‌ای در جمعیت تلاقی برگشتی..... ۳۱۳
۱۰-۳-۱- احتمالات شرطی..... ۳۱۳
۱۰-۳-۲- مدل تابعیت شرطی..... ۳۱۵
۱۰-۳-۳- برآورد و آزمون ۳۱۷
۱۰-۴- مکان‌یابی فاصله‌ای QTL در یک جمعیت F_2 ..... ۳۲۱
۱۰-۵- نکات قابل توجه ۳۲۴
۱۰-۶- تمرین‌ها ۳۲۶

فصل یازدهم: مکان‌یابی فاصله‌ای بااستفاده از روش حداکثر درستنمایی

۱۱-۱- مقدمه..... ۳۳۰
۱۱-۲- مکان‌یابی فاصله‌ای QTL در تلاقی برگشتی ۳۳۱
۱۱-۲-۱- درستنمایی ۳۳۱
۱۱-۲-۲- به حداکثر رساندن درستنمایی ۳۳۳
۱۱-۳- آزمون فرض..... ۳۴۲
۱۱-۳-۱- ارائه مدل در صورت وجود نوترکیبی مضاعف ۳۴۹
۱۱-۳-۲- ارایه مدل در صورت وجود تداخل ۳۵۰
۱۱-۴- مکان‌یابی فاصله‌ای QTL در جمعیت F_2..... ۳۵۲
۱۱-۴-۱- عدم وجود نوترکیبی مضاعف..... ۳۵۲
۱۱-۴-۲- استقلال ۳۵۷

۱۱-۴-۳- تداخل ۳۶۰
۱۱-۴-۴- آزمون کردن فرض‌ها ۳۶۳
۱۱-۵- عوامل مؤثر بر شناسایی QTL..... ۳۶۴
۱۱-۶- روش‌های مکانیابی QTL..... ۳۶۵
۱۱-۶-۱- تعداد QTLها ۳۶۵
۱۱-۶-۲- موقعیت QTLهای انفرادی ۳۷۰
۱۱-۷- تمرین‌ها ۳۷۱

فصل دوازدهم: تجزیهٔ دقت و آستانهٔ بحرانی

۱۲-۱- مقدمه..... ۳۷۴
۱۲-۲- تعیین آستانهٔ بحرانی ۳۷۵
۱۲-۲-۱- پس‌زمینه ۳۷۵
۱۲-۲-۲- تقریب‌های تجزیه‌ای ۳۷۶
۱۲-۲-۳- مطالعات شبیه‌سازی ۳۸۰
۱۲-۲-۴- آزمون‌های تبدیل یا جایگشت..... ۳۸۲
۱۲-۲-۵- یک روش سریع ۳۸۳
۱۲-۲-۶- آماره امتیازی ۳۸۵
۱۲-۳- دقت برآورد پارامتر ۳۸۷
۱۲-۳-۱- ماتریس واریانس-کوواریانس مجانبی..... ۳۸۸
۱۲-۳-۲- مطالعات شبیه‌سازی ۳۹۰
۱۲-۴- فواصل اطمینان برای موقعیت QTL ۳۹۲
۱۲-۵- تمرین‌ها ۳۹۵

فصل سیزدهم: مکان‌یابی مرکب QTL

۱۳-۱- مقدمه..... ۳۹۸
۱۳-۲- مکان‌یابی فاصله‌ای مرکب در جمعیت تلاقی برگشتی..... ۳۹۸
۱۳-۲-۱- درستنمایی ۳۹۹
۱۳-۲-۲- به حداکثر رساندن درستنمایی ۴۰۰

۱۳-۲-۳- آزمون فرض ۴۰۲
۱۳-۳- مکان‌یابی فاصله‌ای مرکب در جمعیت F_2 ..... ۴۰۳
۱۳-۴- توجیه آماری مکان‌یابی فاصله‌ای ترکیبی ۴۰۵
۱۳-۴-۱- (کو) واریانس‌های شرطی نشانگر ۴۰۶
۱۳-۴-۲- واریانس شرطی QTL ۴۰۹
۱۳-۴-۳- انتخاب نشانگر ۴۱۲
۱۳-۵- مقایسه بین مکان‌یابی فاصله‌ای مرکب و مکان‌یابی فاصله‌ای ۴۱۳
۱۳-۶- مکان‌یابی فاصله‌ای چندگانه ۴۱۷
۱۳-۷- تمرین‌ها ۴۱۹

فصل چهاردهم: مکان‌یابی QTL در شجره‌های دگربارور

۱۴-۱- مقدمه ۴۲۲
۱۴-۲- مدل با آثار ثابت برای یک خانواده تنی ۴۲۳
۱۴-۲-۱- مقدمه ۴۲۳
۱۴-۲-۲- مدل مختلط برای دیپلوتیپ والدینی ۴۲۳
۱۴-۲-۳- مدل ژنتیک کمّی ۴۲۸
۱۴-۲-۴- تجزیه درستنمایی ۴۲۹
۱۴-۲-۵- برازش فنوتیپ‌های نشانگر ۴۳۳
۱۴-۲-۶- آزمون‌های فرض ۴۳۵
۱۴-۲-۷- اثر فازهای پیوستگی ۴۳۹
۱۴-۳- مدل مکان‌یابی با آثار تصادفی در یک شجره پیچیده ۴۴۱
۱۴-۳-۱- مقدمه ۴۴۱
۱۴-۳-۲- مدل آماری ۴۴۲
۱۴-۳-۳- IBD در یک QTL ۴۴۴
۱۴-۳-۴- درستنمایی ۴۴۶
۱۴-۳-۵- آزمون فرض ۴۴۸
۱۴-۴- تمرین‌ها ۴۵۲

ضمائم

ضمیمه A: نتایج و الگوریتم‌های آماری عمومی ۴۶۰
ضمیمه B: برنامه‌های R ۴۶۵
ضمیمه C: فهرست منابع ۴۷۷
واژه‌نامه ۴۹۳

فصل اول

ژنتیک پایه

۱-۱- مقدمه

در علم ژنتیک پیشرفت‌های زیادی صورت گرفته است و نیز اطلاعات زیادی در ارتباط با ساز و کارهای مولکولی دقیق انتقال ژنتیکی از والدین به فرزندان در دسترس قرار گرفته است. در این فصل، به‌طور خلاصه اصطلاحات و اصول ژنتیک مندلی، ژنتیک جمعیت، ژنتیک کمّی و جنبه‌های مولکولی ژنتیک را به منظور درک بهتر روش‌های پژوهشی که در این کتاب ارائه خواهند شد، مرور می‌کنیم. بیشتر مطالب و توصیف‌های ارائه شده برای ژنتیک کلاسیک مندلی از کتاب «بایلی»[1] (۱۹۶۱) اقتباس شده است. برای کسب اطلاعات بیشتر در خصوص ژنتیک مدرن (پیشرفته) به کتاب‌های ژنتیک عمومی‌تر که در پایان این فصل معرفی شده‌اند، مراجعه کنید.

۱-۲- ژن‌ها و کروموزوم‌ها

ژن‌ها واحدهای گسسته‌ای هستند که ویژگی‌های زیستی آن‌ها، انتقال اطلاعات وراثتی از والدین به فرزندان است. ژن‌ها به‌طور طبیعی، بدون تغییر از نسلی به نسل دیگر انتقال می‌یابند و معمولاً به‌صورت جفتی وجود دارند. اگر جفت ژن موجود مشابه باشد، فرد را هموزیگوت و اگر ژن‌ها نامشابه باشند فرد را هتروزیگوت می‌گویند. برای مثال، اگر دو ژن به صورت A و a داشته باشیم، دو نوع هموزیگوت یعنی AA و aa و یک نوع هتروزیگوت یعنی Aa وجود دارند. این حالات ژن‌ها را آلل می‌نامند. با یک جفت آلل، سه ژنوتیپ ممکن وجود خواهند داشت که به‌صورت AA، Aa و aa نشان داده می‌شوند. به‌طورکلی، تعداد ژن‌ها بسیار زیاد بوده است و در درون هسته سلول‌ها به‌صورت منظم و خطی روی اجسام ذره‌بینی به نام کروموزوم‌ها قرار دارند. کروموزوم‌ها به‌صورت جفت‌های مشابه یا همولوگ[2] وجود دارند که تعداد این جفت‌ها برای هر گونه ثابت است. برای مثال، مگس سرکه ۴ جفت، درخت کاج ۱۲ جفت، موش خانگی ۲۰ جفت و انسان ۲۳ جفت کروموزوم دارند. مجموع این جفت کروموزوم‌ها، ژنوم یک موجود زنده خاص را تشکیل می‌دهند. یکی از جفت کروموزوم‌های موجود در ژنوم، کروموزوم‌های جنسی[3] (که به‌طور معمول به صورت **X** و **Y** نشان داده می‌شوند) هستند که از نظر ژنتیکی تعیین کننده جنسیت هستند. جفت کروموزوم‌های دیگر، کروموزوم های غیرجنسی[4] هستند که بیان تظاهر بیشتر صفات دیگر را به عهده دارند.

---

[1] - Bailey
[2] - Homologous
[3] - Sex chromosomes
[4] - Autosomes

هر جفت ژن دارای مکان یا جایگاهی معین یا لوکوس[1] روی کروموزوم خاصّی است. از آنجایی که کروموزوم‌ها به صورت جفت هستند، جایگاه‌های ژنی و ژن‌های مستقر روی آن‌ها نیز به صورت جفت وجود دارند؛ بنابراین، اگرچه یک مکان ژنی ممکن است به وسیله هر ژنی از سری‌های چندآللی (بیش از دو آلل یا حالت چندآللی) که تعیین کننده صفت خاصی است، اشغال شود، اما هر جایگاه ژنی دارای ترتیب خطی ثابتی است. مهم‌ترین هدف یک پروژه نقشه‌یابی ژنوم، مکان‌یابی ژن‌های مؤثر بر بروز صفات روی کروموزوم‌های مربوطه است.

۱-۳- میوز[2]

زمانی که سلول‌های بدنی معمولی تقسیم و تکثیر می‌شوند، هستهٔ سلول متحمل یک فرایند تقسیم تحت‌عنوان میتوز[3] می‌گردد که منجر به ایجاد دو سلول دختری می‌شود که هر کدام دارای مجموعهٔ کاملی از جفت کروموزوم‌ها دقیقاً همانند والدینشان هستند. ولی در تولید سلول‌های تولیدمثلی یا گامت‌ها (تخمک و اسپرم) ساز و کار متفاوتی وجود دارد که میوز نامیده می شود. این فرآیند (میوز) تضمین می‌کند که فقط یک کروموزوم از هر جفت کروموزوم همولوگ به هر گامت می‌رسد. همچنین هر گامت فقط یک ژن از هر جفت ژن را دارا است. تعداد کروموزوم‌ها در گامت، عدد هاپلوئیدی نامیده می‌شود و در مقابل، ترکیب کامل تخم بارور شده یا زیگوت[4] قرار دارد که دیپلوئید نامیده می‌شود.

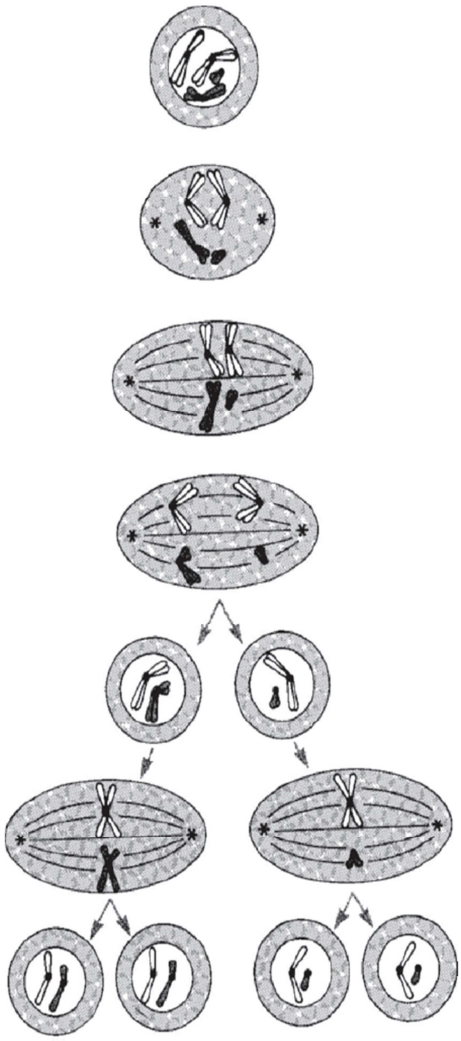
توصیف فرایند زیستی تقسیم میوز در شکل ۱-۱ ارائه شده است. کروموزوم‌ها تا پیش از زمانی‌که در شروع اولین تقسیم میوز قابل مشاهده باشند، تکثیر و دو برابر شده‌اند. هر جفت از این نسخه‌های تکثیر شده در محل سانترومر به یکدیگر می‌پیوندند. سانترومر ناحیه کوچکی است که در آن ناحیه، دو بازوی کروموزوم به یکدیگر متصل می‌شوند. جفت‌های دو برابر شده در سراسر آنافاز اول به‌صورت به هم پیوسته باقی می‌مانند. یکی از جفت کروموزوم‌های همولوگ (یک جفت تکثیر شده) به سمت یکی از قطب‌ها و همولوگ دیگر (جفت تکثیر شده دیگر) به سمت قطب دیگر حرکت می‌کند. محصول مستقیم و بی‌واسطه اولین تقسیم میوز دو سلول است که هر یک دارای یک مجموعه کروموزومی دیپلوئید هستند؛ اما هر جفت کروموزوم همولوگ در یکی از این سلول‌ها، حاوی یک جفت کروموزوم با منشأ مادری یا یک جفت کروموزوم با منشأ پدری است. جور شدن بین دو سلول به صورت تصادفی است و هر سلول تولید شده به‌طور معمول دارای تعدادی جفت

---

[1] - Locus
[2] - Meiosis
[3] - Mitosis
[4] - Zygote

کروموزوم با منشأ مادری و بقیه با منشأ پدری است. در تقسیم دوم میوز، تعداد کروموزوم‌ها نصف می‌شود و هر یک از دو محصول تولید شده از تقسیم اول میوز، سبب ایجاد سلول‌های دختری مشابه با نصف تعداد معمول کروموزوم‌ها می‌شوند.

اهمیت تقسیم کاهشی میوز از آن جهت است که می‌تواند سبب حفظ کروموزوم‌های دیپلوئید، در اثر اختلاط گامت نر (اسپرم) با گامت ماده (تخمک) پس از لقاح شود. دومین خصوصیت اصلی میوز این است که تبادل مواد ژنتیکی بین جفت کروموزوم‌های همولوگ رخ می‌دهد؛ بنابراین، مجموعه کروموزومی گامت هاپلوئید، حاوی ترکیبی از کروموزوم‌ها است، به‌طوری که مقداری از پدر و بخش دیگر از مادر به‌دست آمده است.



سلول پیش‌ساز گامت (مادر مگاسپور) در آغاز فرایند میوز، DNA تا این مرحله همانندسازی کرده و مضاعف شده است.

اولین تقسیم میوز: جفت شدن همولوگی

اولین تقسیم میوز: جفت کروموزوم‌های همانندسازی شده همولوگ در بخش استوای دوک در یک خط قرار گرفته‌اند، رشته های کروموزومی همانندسازی شده با یکدیگر باقی مانده‌اند، اعضای هریک به سمت قطب ها کشیده می‌شوند.

شکل‌گیری دو سلول دختری: هر سلول دختری حاوی دو کروموزوم است که قبلاً مضاعف شده‌اند.

تقسیم دوم میوز: DNA دو برابر نشده است، ولی سانترومرها و کروموزوم‌هایی که قبلاً مضاعف شده‌اند، در حال تفکیک شدن از یکدیگر هستند. هر سلول، دو سلول دختری مشابه تولید می‌کند که DNA و کروموزوم‌های آن‌ها به نصف کاهش می‌یابند.

شکل ۱-۱- شکل شماتیک تقسیم میوز در یک فرد نر فرضی که دارای یک جفت کروموزوم غیرجنسی یا پیکری (اتوزوم) مشابه (به رنگ سفید) و یک جفت غیرمشابه XY (به صورت سیاه) است.
اقتباس از کاوالی-فورزا و بودمر[1] (۱۹۷۱).

[1] - Cavalli-Sforza and Bodmer

۱-۴- قوانین مندل

۱-۴-۱- قانون اول مندل

ژن‌ها در تمامی سلول‌های یک موجود بالغ، به استثنای گامت‌ها، به‌صورت جفتی وجود دارند. گامت‌ها فقط دارای یک ژن از هر جفت معین هستند؛ بنابراین، اگر یک فرد بالغ دارای ژنوتیپ AA باشد، تمامی گامت‌های تولید شده از نوع A هستند. ولی اگر ژنوتیپ آن Aa باشد، دو نوع گامت ممکن A و a هستند و به‌طور معمول این گامت‌ها به تعداد مساوی تولید می‌شوند. زمانی‌که باروری رخ می‌دهد، اسپرمی که حامل یک ژن از والد نر است با تخمکی که حامل یک ژن از والد ماده است ترکیب شده و بنابراین یک جفت کامل تشکیل می‌شود. سپس تخم بارور شده، یا زیگوت، رشد می‌کند تا موجود زنده‌ای را تولید نماید که در هر سلول بدنی آن، یکی از ژن‌ها از یک والد و ژن دیگر از والد دوم به‌دست آمده است. فرد جدید سلول‌های تولیدمثلی خود را تولید می‌کند و از این رو این فرایند می‌تواند ادامه یابد.

توضیحات فوق قانون اول مندل را تشکیل می‌دهند که قانون تفرق نامیده می‌شود. این قانون بیان می‌کند که خصوصیات و ویژگی‌های هر فرد به وسیله جفت ژن‌هایی کنترل می‌شوند که در طی تشکیل سلول‌های تولیدمثلی، تفرق یافته یا جدا شده و بنابراین وارد گامت‌های مختلف می‌شوند. این جفت ژن‌ها در هنگام باروری و لقاح، بازیابی شده و منجر به تولید انواع مختلفی از نتاج در نسبت‌های معین و مشخصی می‌شوند. بدین صورت، تفرق سبب اختلاط ژن‌ها و توزیع مجدد آن‌ها در نسل بعد می‌شود. همچنین ممکن است خصوصیات و صفات، خود تفرق نشان دهند، ولی روش دقیق‌تری که این وضعیت رخ می‌دهد به ماهیت ژن‌های مربوطه و روابط غالب و مغلوبی آن‌ها بستگی دارد.

فرض کنید دو فرد با ژنوتیپ AA و aa را تلاقی دهیم. تمامی گامت‌های حاصل از فرد اول A و تمامی گامت‌های حاصل از فرد دوم a خواهد بود؛ بنابراین، تمامی زیگوت‌های F_1 از نوع هتروزیگوت Aa خواهند شد. حال دو فرد حاصل از نسل F_1 را باهم آمیزش می‌دهیم تا نسل جدید F_2 به‌وجود آید. هر فرد هتروزیگوت F_1 با ژنوتیپ Aa دو نوع گامت A و a با تعداد مساوی تولید می‌کند. چهار مسیر برای شکل‌گیری زیگوت در زمان باروری وجود دارد: یک ژن A از هر والد، یک ژن a از هر والد، A از والد نر و a از والد ماده یا A از والد ماده و a از والد نر؛ بنابراین، انتظار داریم سه نوع نتاج به‌صورت AA، Aa و aa با نسبت‌های ۱:۲:۱ در نسل F_2 تولید شوند. ولی اگر A غالب باشد، دو دسته اول از نظر فنوتیپی غیر قابل تشخیص بوده و دو کلاس فنوتیپی A و a با نسبت ۱:۳ خواهیم داشت.

اگر یکی از فرزندان هتروزیگوت F_1 با والد هموزیگوت خود تلاقی یابد، جمعیت تلاقی برگشتی تولید می‌شود. ژنوتیپ فرد در تلاقی برگشتی فقط به هتروزیگوت F_1 که دارای دو نوع گامت A و a

به تعداد مساوی است، بستگی دارد؛ بنابراین، نسبت تفرق ژنوتیپ‌ها در تلاقی برگشتی به‌صورت ۱:۱ است.

۱-۴-۲- قانون دوم مندل

قانون دوم مندل بیان می‌کند که وقتی دو یا چند جفت ژن به‌طور هم‌زمان تفرق می‌یابند، این ژن‌ها به‌طور مستقل از هم عمل می‌کنند. این قانون، قانون جور شدن مستقل نامیده می‌شود. در بسیاری از موارد، این قانون کافی است، ولی در معرض استثنائات بسیار مهم و خاصی قرار دارد. این استثنائات به‌دلیل پدیده پیوستگی (لینکاژ) که عنوان اصلی این کتاب است، ایجاد می‌شوند.

دو جفت ژن **A** و **B** را به‌ترتیب با آلل‌های A و a و B و b در نظر بگیرید. اگر دو فرد را آمیزش دهیم که یکی برای A و B و دیگری برای a و b هموزیگوت باشند (یعنی آمیزش AABB × aabb)، واضح است که تمامی فرزندان AaBb خواهند شد. این بدان علت است که والد اول گامت‌هایی تولید می‌کند که همگی AB و والد دوم نیز گامت‌هایی تولید می‌کند که همگی ab هستند. حال تلاقیAaBb AaBb × را در نظر بگیریم. اگر تفرق دو ژن به‌صورت مستقل باشد، هر یک از این افراد چهار نوع گامت یعنی AB، Ab، aB و ab را به تعداد مساوی تولید خواهند کرد. ترکیب این چهار نوع گامت از دو والد سبب ایجاد ۱۶ ترکیب یا ژنوتیپ می‌شود که همهٔ آن‌ها متفاوت نیستند. حالات ممکنه مختلف به ساده‌ترین صورت در شکل ۱-۲ نشان داده شده‌اند. همان‌طور که شکل ۱-۲ نشان می‌دهد، در حقیقت ۹ ژنوتیپ متمایز به‌صورت AABB(۱)، AABb(۲)، AAbb(۱)، AaBB(۲)، AaBb(۴)، Aabb(۲)، aaBB(۱)، aaBb(۲) و aabb(۱) وجود دارند که اعداد ارائه شده در پرانتز تعداد هر یک از ژنوتیپ‌ها را نشان می‌دهد. ولی اگر هر جفت ژن، رابطه غالب/مغلوبی نشان دهد، فقط چهار کلاس متفاوت فنوتیپی به صورت AB، Ab، aB و ab به‌ترتیب با نسبت ۹:۳:۳:۱ وجود خواهد داشت.

گامت‌ها

گامت‌ها	AB	Ab	aB	ab
AB	AABB (AB)	AABb (AB)	AaBB (AB)	AaBb (AB)
Ab	AABb (AB)	AAbb (Ab)	AaBb (AB)	Aabb (Ab)
aB	AaBB (AB)	AaBb (AB)	aaBB (aB)	aaBb (aB)
ab	AaBb (AB)	Aabb (Ab)	aaBb (aB)	aabb (ab)

شکل ۱-۲- تفرق ژنی در آمیزش AaBb AaBb × که دو جفت ژن در آن دخیل هستند. زمانی که هر یک از جفت ژن‌ها از خود غالبیت نشان دهند، فنوتیپ‌های حاصله در پرانتز نشان داده شده‌اند. درجه غالبیت با میزان تیره رنگ بودن سلول‌ها توصیف شده است.

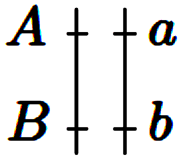
۱-۵- پیوستگی ژنی و مکان‌یابی

قانون دوم مندل برای ژن‌هایی به‌کار می‌رود که روی کروموزوم‌های متفاوت قرار دارند. ژن‌هایی که روی یک کروموزوم قرار دارند، تمایل دارند باهم باقی بمانند. ژن‌هایی که روی یک کروموزوم قرار دارند، ژن‌های هماهنگ[1] و آن‌هایی که روی کروموزوم‌های متفاوت قرار دارند، ناهماهنگ[2] نامیده می‌شوند. میزان باهم باقی ماندن ژن‌های هماهنگ به درجه نزدیکی آن‌ها بستگی دارد.

به منظور مشاهده آنچه که اساساً در پیوستگی دخالت دارد، شکل‌گیری گامت‌ها به وسیله هتروزیگوت AaBb را در نظر بگیرید. اگر جفت ژن‌های A، a و B، b روی یک کروموزوم قرار

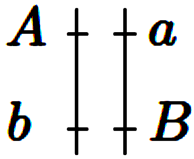
[1] - Syntenic
[2] - Nonsyntenic

داشته باشند، ما می‌توانیم به‌طور دقیق‌تری ترکیب جفت کروموزوم‌های همولوگ را تشخیص دهیم؛ بنابراین، یک کروموزوم ممکن است حاوی A و B و دیگری حاوی a و b باشد، یعنی:



(۱-۱)

که دو خط عمودی، دو کروموزوم همولوگ را نشان می‌دهند. در حالت دیگر، A و b ممکن است روی یک کروموزوم و a و B روی کروموزوم هومولوگ دیگر قرار داشته باشند، یعنی:



(۲-۱)

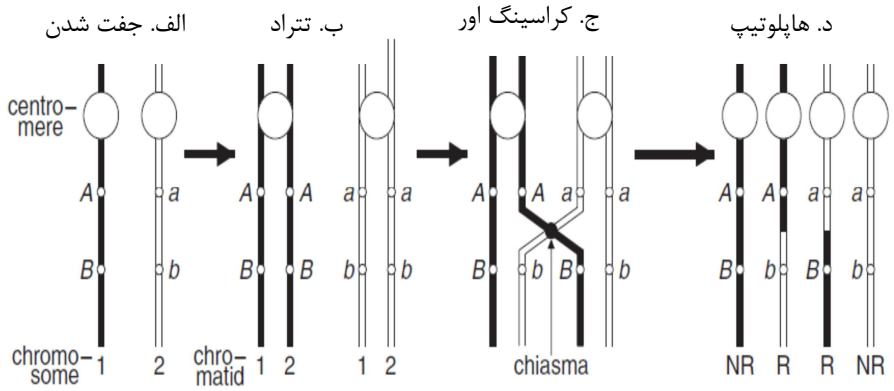
تعریف ۱-۱. برخی واژه‌گان بنیادی: برای آلل‌های A و B، ترتیبی که در شکل (۱-۱) نشان داده شده است، حالت اتصال[1] است و به‌صورت AB/ab نوشته می‌شود. به ترتیب نشان داده شده در شکل (۱-۲) حالت انفصال[2] گفته و به‌صورت Ab/aB نوشته می‌شود. به ترتیب نسبی غیر آلل‌ها (یعنی A در برابر B و b، a در برابر B و b) در جایگاه‌های ژنی مختلف در طول یک کروموزوم، فاز پیوستگی اطلاق می‌شود.

در مرحلهٔ اول میوز، دو کروموزوم ۱ و ۲ در ارتباط با ژن‌های مربوطه به‌صورت پهلو به پهلو کنار یکدیگر قرارمی‌گیرند. اگر ژنوتیپ مادری AB/ab باشد، می‌توانیم ترتیب قرارگیری خطی را به‌صورت شکل ۱-۳ الف نمایش دهیم. سپس هر یک از کروموزوم‌های جفت شده، همانندسازی می‌کنند تا دو رشته خواهری (کروماتید) را تشکیل دهند. این کروماتیدهای خواهری در ناحیه‌ای بنام سانترومر به یکدیگر متصل هستند. کروموزوم‌های همولوگ جفت‌هایی را تشکیل می‌دهند، به‌طوری که هر مجموعه حاصله شامل چهار کروماتید است که تحت‌عنوان تتراد شناخته می‌شوند (شکل ۱-۳ ب). در این مرحله، کروماتیدهای غیر خواهری به‌صورت نیمه تصادفی در نواحی که کیاسما نامیده می‌شوند به یکدیگر می‌چسبند. هر کیاسما نقطه‌ای است که در آن، بین دو کروماتید غیرخواهری می‌تواند کراسینگ اور رخ دهد (شکل ۱-۳ ج). کیاسماها به‌طور کاملاً تصادفی رخ

---

[1] - Coupling
[2] - Repulsion

نمی‌دهند، چون آن‌ها احتمالاً دورتر از سانترومر تشکیل می‌شوند و پیدا کردن دو کیاسما در نقاط نزدیک به یکدیگر معمولی نیست.



شکل ۱-۳- نمایش کراسینگ‌اور بین ژن‌های پیوسته **A** و **B**.

هر گامت، یک کروماتید از تتراد دریافت می‌کند تا بتواند یک هاپلوتیپ تشکیل دهد (شکل ۱-۳ د). چون ممکن است بیش از یک کراس‌اور در یک کروموزوم رخ دهد، از این‌رو برخی کروموزوم‌های هاپلوئید دارای قطعاتی از دو کروموزوم والدینی هستند. تعداد این قطعات باتوجه به تعداد کراس‌اورهایی که در فرایند تشکیل کروماتید و تبدیل آن به کروموزوم رخ می‌دهد، تعیین می‌شود. اگر کراس‌اور اتفاق نیافتد، در این صورت کروموزوم نسخه‌ای از کروموزوم کامل والدینی خواهد بود. اگر یک کراس‌اور بین دو ژن A و B رخ دهد، کروموزوم شامل دو قطعه خواهد شد که هر قطعه متعلق به یکی از والدین خواهد بود. در حالت قبل، گامت‌های حاصله بایستی AB یا ab باشند که مشابه کروموزوم‌های والدینی هستند. در حالت اخیر که یک نقطه برای تبادل وجود دارد، ترکیبات جدید Ab و aB را داریم که انواع نوترکیب نامیده می‌شوند. به‌طورکلی، اگر تعداد نقاط تبادل بین دو جایگاه ژنی زوج باشد، نتیجه نهایی به‌صورت AB یا ab قابل تشخیص خواهد شد؛ اما اگر تعداد نقاط تبادل فرد باشد، نتیجه به صورت Ab یا aB خواهد شد.

وجود پیوستگی بدان معناست که گامت‌های بیشتری به شکل AB و ab (گامت‌های نوع والدینی) و گامت‌های کمتری به صورت Ab و aB (گامت‌های نوترکیب) وجود خواهند داشت. فرض کنید سهم گامت‌های نوترکیب به صورت r (که آنرا نرخ نوترکیبی می‌نامیم) و سهم گامت‌های والدینی 1-r باشد. نرخ نوترکیبی می‌تواند بر اساس تعداد نوترکیب‌های مورد انتظار در نتاج در حال تفرق برآورد شود (فصل ۳ را ملاحظه کنید). به‌طورکلی، نباید انتظار داشته باشیم که

نرخ نوترکیبی بیشتر از نیم (۵۰ درصد) باشد اگرچه در شرایط غیر معمول ممکن است تمایلی برای کروموزوم‌های به ارث رسیده از یک والد یا دودمان خاصی وجود داشته باشد که به صورت غیر تصادفی با هم مرتبط شوند.

باتوجه به تعریف نرخ نوترکیبی، چنین استنباط می‌شود که در حالت خاص که $r = \frac{1}{2}$ است، می‌توان گفت که یا پیوستگی وجود ندارد و یا تفرق ژن‌ها به‌صورت مستقل رخ می‌دهد. در حقیقت، اگر دو جایگاه ژنی در روی یک کروموزوم فاصله زیادی داشته باشند، تعداد نقاط تبادل فرد و زوج مساوی خواهند بود (یعنی ۵۰ درصد از هر کدام)، به‌طوری‌که این حالت به سرعت از حالتی که در آن جایگاه‌ها روی کروموزوم‌های متفاوتی قرار دارند، قابل تشخیص نخواهد بود. در حالت دیگر، اگر دو جایگاه ژنی به یکدیگر نزدیک باشند، فراوانی نقاط تبادل کم و نرخ نوترکیبی مربوطه نیز کوچک خواهد بود. تا حدودی، می‌توانیم از این حالت اخیر به‌عنوان معیاری برای اندازه‌گیری فاصله بین دو جایگاه ژنی استفاده کنیم.

معیار بهتر، اندازه‌گیری آن از راه تعیین تراکم نقاط تبادل مشخص است.

تعریف ۱-۲- فاصله نقشه: فاصله نقشه در بین هر دو جایگاه ژنی برابر با متوسط تعداد نقاط تبادلی (کیاسما) است که در هر قطعه ایجاد می‌شود.

فاصله نقشه کمیتی است که به‌طور معمول افزایشی است. رابطه بسیار ساده‌ای بین نرخ نوترکیبی و فاصله نقشه برای یک جفت ژن در ساده‌ترین حالت یعنی عدم وجود تداخل[1]، وجود دارد. چنین رابطه‌ای تابع نقشه نامیده می‌شود و در بخش ۳-۱۰ مورد بحث قرار خواهد گرفت. زمانی‌که نرخ نوترکیبی بین ژن‌های موجود روی یک کروموزوم منفرد با انجام آزمایش‌های مناسب پیوستگی برآورد می‌شوند، تبدیل آن‌ها به فواصل نقشه و تهیه نقشه کروموزومی موضوع ساده‌ای است. از آنجایی که هیچ مدرکی مبنی بر یکنواختی فراوانی کراسینگ‌اور در طول کروموزوم‌ها وجود ندارد، از این‌رو نمی‌توان گفت که انطباق خیلی زیادی بین فاصله ژنتیکی و فاصله حقیقی بین ژن‌ها وجود دارد.

زمانی‌که تعداد زیادی ژن در نظر گرفته شوند، در خصوص ترتیب و نظم خطی آن‌ها روی هر کروموزوم مباحثی مطرح می‌شود. جایگاه‌های ژنی هر موجود زنده در گروه‌های پیوستگی[2] قرار می‌گیرند، به‌طوری که هر جایگاه ژنی در یک گروه پیوستگی به هر جایگاه ژنی در گروه دیگر پیوسته نیست. در هر حال، ژن‌های هر گروه پیوستگی می‌توانند به‌صورت خطی منظم شوند. برای

---

[1] - Interference
[2] - Linkage group

ژن‌هایی که به حد کافی به هم نزدیک هستند، نرخ نوترکیبی بین هر جفت ژن ممکن است در تجزیه‌های مقدماتی و اولیه، به‌عنوان معیار مستقیم فاصله بین ژن‌ها استفاده شود. برای آن که ژن‌ها به صورت افزایشی و مستقل باقی بمانند، به جای نرخ نوترکیبی باید متوسط تعداد کراس‌اورها را مدنظر قرار دهیم (که فقط فراوانی تعداد فرد کراس‌اورها را اندازه‌گیری می‌کند)؛ بنابراین، لازم است بدانیم چگونه نرخ نوترکیبی مشاهده شده بین تعداد زیادی از جفت ژن‌های موجود روی یک کروموزوم می‌توانند به‌صورت یکنواخت بر اساس یک نقشه کروموزومی برازش یابند. این حالت اساساً بر توسعه مدل‌های نظری و الگوریتم‌های آماری برای ایجاد نقشه‌های ژنتیکی پیوستگی متکی است که یکی از مباحث اصلی این کتاب است.

۱-۶- تداخل

در ساده‌ترین حالت، فرض بر این است که نقاط تبادل (کیاسما) به‌صورت تصادفی رخ می‌دهند، به‌طوری که الگوی کراسینگ‌اور در هر بخش از کروموزوم به‌صورت مستقل از الگوی مربوط به هر بخش دیگر است. ولی در عمل، وجود حالت غیرتصادفی معمول است که اچ. جی. مولر[1] (۱۹۱۶) آن‌را تداخل نامید. به‌طور معمول، وقتی این حالت مثبت است، وقوع یک نقطه تبادل از شکل گیری نقاط تبادل دیگر در نزدیک آن جلوگیری می‌کند. مدل‌های مختلفی برای توصیف پدیده تداخل وجود دارند و برخی از آن‌ها مستلزم وقوع نرخ نوترکیبی بیش از ۵۰٪ در کروموزوم‌هایی هستند که دارای طول زیادی هستند.

همان‌گونه که در بخش ابتدایی ذکر شد، هر کروموزوم در زمان مناسب تقسیم میوزی به‌صورت طولی به یک جفت کروموزوم دختری مشابه (کروماتید) شکسته می‌شود. در ابتدا دو کروماتید به وسیله سانترومر به همدیگر نگه داشته می‌شوند (شکل ۱-۳ ب). همانطور که در شکل ۱-۳ ج نیز نشان داده شده است، همیشه کراسینگ‌اور بین کروماتیدهای حاصل از کروموزوم‌های جفت همولوگ رخ می‌دهند؛ بنابراین در حقیقت پدیده کراسینگ‌اور شامل تمامی چهار کروماتید یا رشته مربوط به هر جفت از کروموزوم‌های همولوگ است. یک جفت کروموزوم همولوگ که به وسیله کراسینگ‌اور به صورت متحد و واحد در می‌آیند، اغلب بی‌والانت[2] نامیده می شوند.

ممکن است وقوع چندین نقطه تبادل یا کیاسما را تصور کنیم که هر یک از آن‌ها در این مرحله دارای ترتیب X مانند کروماتیدها هستند که در شکل ۱-۳ ج نشان داده شده‌اند. همچنین می‌توانیم بین دو نوع تداخل تمایز قائل شویم.

---

[1] - H. J. Muller
[2] - Bivalent

تعریف ۱-۳- انواع تداخل: یک نوع تداخل، تداخل کیاسمایی است که در آن، وقوع یک کیاسما روی شانس وقوع کیاسمای دیگر در نزدیکی آن، اثر می‌گذارد. نوع دیگر، تداخل کروماتیدی است که یک رابطه غیرتصادفی بین جفت رشته‌های دخیل در یک کیاسما و جفت رشته‌های دخیل در کیاسمای بعدی است.

تداخل کیاسمایی معمول و متداول است و در برخی از توزیع‌ها، واریانس این نوع تداخل تقریباً به میزان یک چهارم میانگین توزیع است. از سوی دیگر، شناسایی تداخل کروماتیدی بسیار مشکل‌تر است و شواهد و مدارک مبنی بر وجود آن محدود است. ثابت شده است که تداخل کیاسمایی به تنهایی قادر به ایجاد نرخ نوترکیبی بیش از ۵۰ درصد نیست (ماتر[1]، ۱۹۳۸).

۱-۷- ژنتیک کمّی

۱-۷-۱- ویژگی‌های ژن‌ها در جمعیت

تفرق مندلی منجر به ایجاد نسبت‌های تفرق ساده و قابل پیش‌بینی در نتاج یک نوع آمیزش می‌شود، ولی فقط برای جمعیت‌هایی به‌کار می‌رود که از دو والد با ژنوتیپ شناخته شده حاصل شده‌اند؛ اما آمیزش‌های متفاوتی می‌توانند به‌طور هم‌زمان در جمعیت‌های آزمایشی یا طبیعی انجام شوند که در آن‌ها نسبت ژنوتیپ‌های مختلف، میانگین وزنی نسبت‌های تفرق در تمامی انواع آمیزش‌های ممکنه و وزن‌ها برابر با فراوانی‌های نسبی انواع آمیزش‌های مختلف هستند. خصوصیات ژن‌های جمعیت می‌توانند با فراوانی‌های آللی، فراوانی‌های ژنوتیپی و قانون هاردی-واینبرگ توصیف شوند.

یک ژن با دو آلل A و a به‌ترتیب با فراوانی‌های p_1 و p_0 در یک جمعیت در نظر بگیرید. فرض کنید P_2، P_1 و P_0 به ترتیب فراوانی سه ژنوتیپ AA، Aa و aa باشند. زمانی که آمیزش‌ها تصادفی باشند، نسبت ژنوتیپ‌های مختلف از یک مدل ریاضی تبعیت می‌کند که به‌صورت مستقل به وسیله ریاضی‌دان انگلیسی هاردی[2] (۱۹۰۸) و پزشک آلمانی واینبرگ[3] (۱۹۰۸) ارائه شد. این مدل شناخته شده که امروزه قانون هاردی-واینبرگ نامیده می‌شود، بیان می‌کند که اگر آمیزش بین افراد در یک جمعیت به‌صورت تصادفی باشد، رابطه بین فراوانی‌های ژنوتیپی به صورت زیر خواهد بود:

$$P_1^2 = 4P_2P_0 \qquad (۳-۱)$$

---

[1] - Mather
[2] - Hardy
[3] - Weinberg

و هر یک از فراوانی‌های ژنوتیپی از نسلی به نسل دیگر بدون تغییر باقی می‌مانند. جمعیتی که از رابطه (۱-۳) تبعیت کند، در تعادل هاردی-واینبرگ خواهد بود که در آن، فراوانی‌های ژنوتیپی می‌توانند به ترتیب به‌صورت $P_2 = p_1^2$ ، $P_1 = 2p_1p_0$ و $P_0 = p_0^2$ بیان شوند. روش‌هایی وجود دارند که وجود یا عدم وجود تعادل هاردی-واینبرگ را در یک جمعیت آزمون می‌کنند (فالکونر و مک کی[1]، ۱۹۹۶ و لینچ و والش[2]، ۱۹۹۸).

۱-۷-۲- یک مدل عمومی ژنتیک کمّی

ژنی که در یک جمعیت در حال تفرق است ممکن است روی فنوتیپ یک صفت اثر گذارد. برای صفتی که دارای توارث پیچیده یا کمّی است، تعداد ژن‌هایی که آن‌را کنترل می‌کنند ممکن است زیاد بوده و روابط آن‌ها با محیط ممکن است پیچیده باشد. مطالعه اساس ژنتیکی یک صفت کمّی یکی از موضوعات ژنتیک کمّی است.

یک صفت کمّی با ارزش فنوتیپی P را در نظر بگیرید که به وسیله عوامل ژنتیکی (G)، محیطی (E) و اثر متقابل ژنوتیپ و محیط (G×E) کنترل می‌شود. رابطه بین این ارزش‌ها به‌صورت زیر بیان می‌شود:

$$P = G + E + G \times E \qquad (۱-۴)$$

با فرض اینکه کلیه عوامل موجود در معادله (۱-۴) مستقل از هم باشند، واریانس فنوتیپی صفت به مؤلفه‌های واریانس ژنتیکی، محیطی و اثر متقابل بین ژنوتیپ × محیط تفکیک می‌شود:

$$V_P = V_G + V_E + V_{G\times E} \qquad (۱-۵)$$

در آمار، واریانس به‌طورکلی با علامت V یا σ^2 نشان داده می‌شود. واریانس ژنتیکی، V_G یا σ_G^2، به‌واسطه آثار تمامی ژن‌هایی است که صفت را کنترل می‌کنند. یک ژن با ژنوتیپ‌های AA، Aa و aa در نظر بگیرید که ارزش‌ها و فراوانی‌های ژنوتیپی آن در یک جمعیت با تعادل هاردی- واینبرگ به صورت زیر بیان می‌شود:

ژنوتیپ	ارزش ژنوتیپی	فراوانی
AA	$\mu_2 = \mu + a$	$P_2 = p_1^2$
Aa	$\mu_1 = \mu + d$	$P_2 = 2p_1 p_0$
aa	$\mu_0 = \mu - a$	$P_0 = p_0^2$

[1] - Falconer and Mackay
[2] - Lynch and Walsh

سه ژنوتیپ مختلف با علامت j نشان داده می‌شوند (j=۲ برای AA، j=۱ برای Aa و j=۰ برای aa). ارزش‌های ژنوتیپی شامل میانگین کل صفت (μ)، اثر افزایشی (a) ژن به‌واسطه جایگزینی آلل A با a، یا اثر غالبیت (d) ژن به‌واسطه بر هم کنش بین آلل‌های A و a در ژن است. اگر غالبیت وجود نداشته باشد، d = ۰، اگر آلل A بر آلل a غالب باشد، d مثبت و اگر آلل a بر آلل A غالب باشد، d منفی خواهد شد. اگر d برابر $+a$ یا $-a$ باشد، غالبیت کامل و اگر d بزرگ تر از $+a$ یا کمتر از $-a$ باشد، فوق‌غالبیت وجود خواهد داشت. درجه غالبیت با نسبت $\frac{d}{a}$ بیان می‌شود.

میانگین جمعیت برای سه ژنوتیپ با فراوانی‌های مختلف به‌صورت زیر محاسبه می‌شود:

$$\overline{\mu} = \sum_{j=0}^{2} P_j \mu_j = (p_1 - p_0)a + 2p_1 p_0 d$$

و واریانس ژنتیکی این ژن به صورت زیر است:

$$\sigma_g^2 = \sum_{j=0}^{2} P_j (\mu_j - \overline{\mu})^2 = 2p_1 p_0 [a + (p_1 - p_0)d]^2 + 4p_1^2 p_0^2 d^2$$

$$= 2p_1 p_0 \alpha^2 + 4p_1^2 p_0^2 d^2 = \sigma_a^2 + \sigma_d^2$$

که در آن، $\alpha = a + (p_1 - p_0)d$ اثر متوسط جایگزینی آلل‌های a به جای A است (فالکونر و مک کی، ۱۹۹۶). قسمت اول عبارت واریانس ژنتیکی، جزء واریانس ژنتیکی افزایشی (σ_A^2) و بخش دوم، مؤلفه واریانس ژنتیکی غالبیت (σ_D^2) است. این دو بخش به آسانی می‌توانند بسط یابند تا شامل آثار تمامی کنترل‌کننده یک صفت شوند. اگر برهم کنش بین ژن‌ها نادیده گرفته شوند، واریانس‌هایی که به وسیله تمامی ژن‌ها توجیه می‌شوند، به‌صورت زیر بیان می‌شوند:

$$\sigma_A^2 = \sum \sigma_a^2 \quad \sigma_D^2 = \sum \sigma_d^2 \quad \sigma_G^2 = \sum \sigma_g^2 .$$

۱-۷-۳- مدل‌های ژنتیکی برای جمعیت‌های F_2 و تلاقی برگشتی

تفکیک واریانس ژنتیکی می‌تواند برای حالات ژنتیکی مختلف انجام شود. دو جمعیت والدینی P_1 و P_2 را در نظر بگیرید که به ترتیب با آلل‌های مطلوب $A_1, \ldots, A_m$ و آلل‌های نامطلوب $a_1, \ldots, a_m$ برای m جایگاه ژنی تثبیت شده‌اند. دو والد آمیزش داده می‌شوند تا F_1 را تولید کنند. F_1 یا با یکی از والدین آمیزش داده می‌شود تا یک تلاقی برگشتی تولید شود و یا با خودش تلاقی داده می‌شود تا یک F_2 ایجاد نماید.

فرض کنید a_k و d_k به ترتیب آثار افزایشی و غالبیت ژن k و r_{kl} نرخ نوترکیبی بین دو ژن k و l باشد. یک جفت ژن A_k و A_l را در نظر بگیرید که ارزش‌های ژنوتیپی (عبارت بالایی) و فراوانی‌های ژنوتیپی (عبارت پایینی) آن‌ها در جمعیت F_2 به‌صورت زیر است:

	A_lA_l	A_la_l	a_la_l
A_kA_k	$\mu+a_k+a_l$ $\frac{1}{4}(1-r_{kl})^2$	$\mu+a_k+d_l$ $\frac{1}{2}r_{kl}(1-r_{kl})$	$\mu+a_k-a_l$ $\frac{1}{4}r_{kl}^2$
A_ka_k	$\mu+d_k+a_l$ $\frac{1}{2}r_{kl}(1-r_{kl})$	$\mu+d_1+d_2$ $\frac{1}{2}[r_{kl}^2+(1-r_{kl})^2]$	$\mu+d_k-a_l$ $\frac{1}{2}r_{kl}(1-r_{kl})^2$
a_ka_k	$\mu-a_k+a_l$ $\frac{1}{4}r_{kl}^2$	$\mu-a_k+d_l$ $\frac{1}{2}r_{kl}(1-r_{kl})$	$\mu-a_k-a_l$ $\frac{1}{4}(1-r_{kl})^2$

(۱-۶)

ارزش‌های ژنوتیپی بالا از آثار افزایشی و غالبیت دو ژن تشکیل شده‌اند، زیرا برهم کنش ژن‌ها نادیده گرفته شده است. به‌دست آوردن فراوانی‌های ژنوتیپی در F_2 که بر حسب نرخ نوترکیبی بین دو ژن بیان می‌شوند، نیاز به آگاهی از تجزیه پیوستگی دارد که در بخش ۳-۵ توصیف می‌شود. از روی روابط (۱-۶) می‌توانیم واریانس ژنتیکی را به‌صورت زیر به‌دست آوریم:

$$\sigma_G^2=\frac{1}{2}\sum_{k=1}^{m}a_k^2+\frac{1}{4}\sum_{k=1}^{m}d_k^2 +\frac{1}{2}\sum_{k=1}^{m}\sum_{l=1,k\neq l}^{m}(1-2r_{kl})a_ka_l+\frac{1}{4}\sum_{k=1}^{m}\sum_{l=1,k\neq l}^{m}(1-2r_{kl})^2d_kd_l \qquad (۱-۷)$$

جملهٔ اول در سمت راست معادله (۱-۷) برای جمعیت F_2 برابر با واریانس افزایشی درون جایگاه‌های ژنی، جمله دوم واریانس غالبیت درون جایگاه‌های ژنی، جمله سوم کوواریانس بین آثار افزایشی جایگاه‌های ژنی متفاوت و جمله چهارم کوواریانس بین آثار غالبیت جایگاه‌های ژنی متفاوت است.

برای جمعیت تلاقی‌های برگشتی که در آن‌ها اثر غالبیت به‌دلیل ناکافی بودن درجه آزادی نمی‌تواند تعریف شود، می‌توانیم واریانس ژنتیکی مشابه ولی ساده‌تری را به‌صورت زیر استنتاج کنیم:

$$\sigma_G^2 = \frac{1}{4}\sum_{k=1}^{m}\sigma_k^2 + \frac{1}{4}\sum_{k\neq l}^{m}(1-2r_{kl})a_k a_l \qquad (۱-۸)$$

از روی معادله (۱-۸)، واریانس ژنتیکی در تلاقی برگشتی شامل واریانس ژنتیکی افزایشی و کوواریانس بین آثار افزایشی جایگاه‌های ژنی متفاوت است.

۱-۷-۴- مدل اپیستازی

ژن‌ها ممکن است صفات کمّی را با برهم کنش خود تحت‌تأثیر قرار دهند. اثری که به‌واسطه برهم کنش ژنی باشد به وسیله دبلیو. باتسون[1] (۱۹۰۲) تحت‌عنوان اپیستازی نام‌گذاری شده است. از لحاظ فیزیولوژیکی، اپیستازی وابستگی آثار ژنی یک جایگاه ژنی به آثار جایگاه‌های ژنی دیگر است. فیشر[2] (۱۹۱۸) برای اولین بار واریانس ژنتیکی را با استفاده از اصل حداقل مربعات به اجزای افزایشی، غالبیت و اپیستازی تفکیک نمود. بعدها کوکرهام[3] (۱۹۵۴) واریانس اپیستازی دو ژنی را به اجزای برهم کنش افزایشی× افزایشی، افزایشی× غالبیت، غالبیت× افزایشی و غالبیت× غالبیت تفکیک نمود. روش‌های زیادی برای تخصیص اپیستازی (به مدل واریانس ژنتیکی) وجود دارند، ولی ما اپیستازی را با استفاده از روش متر و جینکز[4] (۱۹۸۲) مدل‌سازی خواهیم کرد.

دو ژن یکی به‌صورت **A** با سه ژنوتیپ AA، Aa و aa و دومی به صورت **B** با سه ژنوتیپ BB، Bb و bb را در نظر بگیرید. این دو ژن ۹ ژنوتیپ دو ژنی را شکل می‌دهند که ارزش‌های ژنوتیپی آن‌ها به وسیله $\mu_{j_1 j_2}$ نشان داده شده و می‌توانند به مؤلفه‌های مختلف تقسیم شوند:

---

[1] - W. Bateson
[2] - Fisher
[3] - Cockerham
[4] - Mather and Jinks

$$
\begin{aligned}
&\mu_{j_1 j_2} = \mu && \text{میانگین کل}\\
&+(j_1-1)a_1 + (j_2-1)a_2 && \text{آثار افزایشی}\\
&+j_1(2-j_1)d_1 + j_2(2-j_2)d_2 && \text{آثار غالبیت}\\
&+(j_1-1)(j_2-1)i_{aa} && \text{اثر افزایشی× افزایشی}\\
&+(j_1-1)j_2(2-j_2)i_{ad} && \text{اثر افزایشی× غالبیت}\\
&+j_1(2-j_1)(j_2-1)i_{da} && \text{اثر غالبیت × افزایشی}\\
&+j_1(2-j_1)j_2(2-j_2)i_{dd} && \text{اثر غالبیت × غالبیت}
\end{aligned}
\qquad (۹-۱)
$$

در این روابط:

$$
j_1, j_2 = \begin{cases} 2 & \text{برای AA یا BB} \\ 1 & \text{برای Aa یا Bb} \\ 0 & \text{برای aa یا bb} \end{cases}
$$

خط دوم معادله (۱-۹) آثار افزایشی ژن‌های منفرد را نشان می‌دهد، خط سوم آثار غالبیت ژن‌های منفرد است و خطوط چهارم، پنجم، ششم و هفتم به ترتیب آثار اپیستازی افزایشی× افزایشی (i_{aa})، افزایشی× غالبیت (i_{ad})، غالبیت× افزایشی (i_{da}) و غالبیت× غالبیت (i_{dd}) بین دو ژن را نشان می‌دهند.

ارزش‌ها و فراوانی‌های ژنوتیپی مربوط به دو ژنی که به‌طور هم‌زمان با نرخ نوترکیبی r در جمعیت F_2 در حال تفرق هستند، در جدول ۱-۱ ارائه شده‌اند. توجه داشته باشید که فراوانی‌های ژنوتیپی بر حسب r محاسبه شده‌اند. بر اساس جدول ۱-۱، واریانس ژنتیکی صفت نیز می‌تواند به‌دست آید.

جدول ۱-۱- ارزش‌ها (بالا) و فراوانی‌های (پایین) ژنوتیپی ۹ ژنوتیپ ممکن در دو ژن **A** و **B**

	BB	Bb	Bb
AA	$\mu+a_1+a_2+i_{aa}$	$\mu+a_1+d_2+i_{ad}$	$\mu+a_1-a_2+i_{aa}$
	$\frac{1}{4}[(1-r)^2]$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r^2$
Aa	$\mu+d_1+a_2+i_{da}$	$\mu+d_1+d_2+i_{dd}$	$\mu+d_1-a_2-i_{da}$
	$\frac{1}{2}r(1-r)$	$\frac{1}{2}[r^2+(1-r)^2]$	$\frac{1}{4}r(1-r)$
aa	$\mu-a_1+a_2-i_{aa}$	$\mu-a_1+d_2-i_{ad}$	$\mu-a_1-a_2+i_{aa}$
	$\frac{1}{4}r^2$	$\frac{1}{4}r(1-r)$	$\frac{1}{4}(1-r)^2$

۱-۷-۵- وراثت پذیری و برآورد آن

بر اساس معادله (۱-۵)، کل واریانس فنوتیپی یک صفت کمّی می‌تواند به مؤلفه‌های واریانس ژنتیکی، محیطی و برهم کنش بین ژنوتیپ × محیط تفکیک شود. نسبت واریانس ژنتیکی به واریانس فنوتیپی تحت‌عنوان وراثت پذیری عمومی تعریف می‌شود، یعنی:

$$H^2=\frac{V_G}{V_G+V_E+V_{G\times E}} \qquad (۱۰-۱)$$

همان‌گونه که در بالا نشان داده شد، اثر یا واریانس ژنتیکی می‌تواند به آثار یا واریانس‌های افزایشی (A) و غیرافزایشی (NA) تفکیک شود؛ بنابراین، داریم:

$$P=G+E+G\times E$$
$$=A+NA+E+A\times E+NA\times E$$

و

$$V_P=V_G+V_E+V_{G\times E}$$
$$=V_A+V_{NA}+V_E+V_{A\times E}+V_{NA\times E}$$

این در شرایطی است که تمامی آثار مستقل از یکدیگر باشند.

اثر یا واریانس غیر افزایشی مجموع اثر یا واریانس غالبیت و اپیستازی است. از آنجایی‌که اثر افزایشی می‌تواند از والدین به فرزندان به ارث برسد، در حالی‌که آثار غیرافزایشی چنین خصوصیتی ندارند، ازاین‌رو نسبت واریانس افزایشی به کل واریانس فنوتیپی را تحت‌عنوان وراثت پذیری خصوصی تعریف می‌کنیم، یعنی:

$$h^2 = \frac{V_A}{V_A + V_{NA} + V_E + V_{A\times E} + V_{NA\times E}} \qquad (۱۱-۱)$$

وراثت‌پذیری خصوصی کمیتی است که نشان می‌دهد چه مقدار از ارزش فنوتیپی یک صفت کمّی بدون تغییر از نسلی به نسل دیگر منتقل می‌شود. دو فراسنجه وراثت‌پذیری یعنی روابط (۱-۱۰) و (۱-۱۱) به‌طور مرسوم برای توصیف درجه کنترل ژنتیکی کلّی یک صفت به کار می‌روند که شامل اثر تمامی ژن‌های کنترل کننده یک صفت است (لینچ و والش، ۱۹۹۸). این دو پارامتر در حال حاضر برای توصیف اثر یا میزان مشارکت هر یک از ژن‌های کنترل کننده صفت به‌کار می‌روند، به شرطی که این ژن‌ها بتوانند با استفاده از روش‌هایی مثل مکان‌یابی ژنتیکی که در فصل‌های ۸ الی ۱۴ توضیح داده شده‌اند، شناسایی شوند.

در عمل، واریانس‌های ژنتیکی می‌توانند بر اساس تئوری ژنتیک کمّی که به وسیله کوکرهام (۱۹۵۴، ۱۹۶۳) پایه‌گذاری شد، برآورد شوند. بر اساس این تئوری، مجموعه‌ای از والدین با یکدیگر تلاقی داده می‌شوند تا تلاقی‌های چندگانه‌ای را در یک طرح آمیزشی ایجاد نمایند. نتاج حاصل از طرح آمیزشی در قالب یک طرح آزمایشی خاص ارزیابی می‌شوند و سپس داده‌های فنوتیپی جمع آوری شده با روش‌های آماری نظیر تجزیه واریانس تجزیه می‌شوند تا واریانس‌های آزمایشی مختلف به‌دست آیند. بر پایه شباهت بین خویشاوندان، واریانس‌های آزمایشی محاسبه شده جهت برآورد واریانس‌های ژنتیکی افزایشی و غالبیت و از این‌رو وراثت‌پذیری‌های عمومی و خصوصی مورد استفاده قرار می‌گیرند.

در مقایسه با مدل‌های کوکرهام، متر و جینکز (۱۹۸۲) روشی متفاوت بر اساس تفاوت نسل‌ها برای برآورد آثار ژنتیکی یا مؤلفه‌های واریانس پیشنهاد دادند. فرض کنید که مواد مورد مطالعه شامل سه نسل والدین اینبرد P_1 و P_2، دوررگ‌های F_1 و F_2 های در حال تفرق باشند که در شرایط مشابهی رشد می‌نمایند. واریانس فنوتیپی صفت برای دو لاین (رگه) خالص (V_{P_1} و V_{P_2}) و فرزندان F_1 آن‌ها (V_{F_1}) کاملاً ناشی از عوامل محیطی است، درحالی‌که واریانس فنوتیپی همان صفت در F_2 (V_{F_2}) شامل مجموع واریانس ژنتیکی، محیطی و ژنوتیپ× محیط است؛ بنابراین، واریانس ژنتیکی صفت می‌تواند به‌صورت زیر برآورد شود:

$$V_G = V_{F_2} - V_{F_1} \qquad (۱۲-۱)$$

یا

$$V_G = V_{F_2} - \frac{1}{4}(V_{P_1} + V_{P_2} + 2V_{F_1}) \qquad (۱۳-۱)$$

با در نظر گرفتن تعداد نسل‌های بیشتر، برآورد مؤلفه‌های واریانس ژنتیکی (افزایشی و غالبیت) امکان‌پذیر خواهد شد (متر و جینکز، ۱۹۸۲).

۱-۷-۶- معماری ژنتیکی

بیشتر صفات کمّی به‌وسیله تعداد زیادی از ژن‌ها که با یکدیگر برهم‌کنش دارند و نیز مجموعهٔ منظمی از عوامل محیطی کنترل می‌شوند (فالکونر و مک کی، ۱۹۹۶). نظریه مرسوم پلی ژنیک برای صفات کمّی (متر، ۱۹۴۳) ژن‌های نسبتاً زیادی را در نظر می‌گیرد که هر یک دارای آثار مساوی و نسبتاً کوچک بوده و تا حد زیادی به‌صورت افزایشی عمل می‌کنند. طی سال‌ها مشاهده شده است که یک صفت کمّی ممکن است معماری ژنتیکی پیچیده‌ای را از خود نشان دهد (مک کی، ۱۹۹۶ و ۲۰۰۱) که به‌صورت زیر بیان می‌شود:

۱- ممکن است به وسیله تعداد نسبتاً زیادی جایگاه ژنی کنترل شود، برای مثال، تعداد ۵۰ ژن، مطابق با تحقیقی که شریمپتون و روبرتسون (۱۹۹۸a,b) انجام دادند.

۲- عمل ژن‌ها ممکن است به‌صورت افزایشی، غالبیت، اپیستازی با ژن‌های دیگر و برهم‌کنش با عوامل محیطی باشد.

۳- میزان اثر هر ژن می‌تواند تا حد قابل توجهی متغیر باشد.

۴- ژن‌های مشابه ممکن است از راه آثار پلیوتروپی روی صفات فنوتیپی مختلف تأثیرگذار باشند.

۵- ژن‌های اثرگذار بر روی صفت ممکن است به‌صورت تصادفی یا بر اساس الگوی معینی در سرتاسر ژنوم توزیع شده باشند.

باتوجه به استفاده گسترده از مکان‌یابی ژنتیکی برای تجزیه صفات کمّی و افزایش شواهد و مدارک، راهکار سوّمی مشاهده شده است که پیشنهاد می‌کند معمولاً تعداد کمی ژن بخش نسبتاً زیادی از تنوع صفات را توجیه می‌کنند. به همین دلیل، مدل پلی ژنیک سنتی ممکن است به‌وسیله مدل الیگوژنیک جدیدی که بر اساس آن، تعداد کمی از ژن‌های اصلی و بزرگ‌اثر هر یک با اثر زیاد به همراه تعداد زیادی از ژن‌های کوچک‌اثر هر یک با اثر کم تنوع ژنتیکی یک صفت کمّی را توجیه کنند (برای مرور بیشتر به منبع مک کی، ۱۹۹۶، مراجعه شود).

۱-۷-۷- برآورد تعداد ژن

تعداد واقعی ژن‌هایی که یک صفت کمّی را کنترل می‌کنند یکی از مهم‌ترین عناصر معماری ژنتیکی صفت است. تعداد ژن می‌تواند با روش‌های بیومتری برآورد شود، اگر چه این حالت به برقراری برخی از فرضیات مهم بستگی دارد (لند[1]، ۱۹۸۱؛ لینچ و والش، ۱۹۹۸). تعداد ژن‌هایی که با این روش برآورد می‌شود اساساً تعداد مؤثر ژن‌هایی را نشان می‌دهد که بخش عمده‌ای از تنوع

[1] - Lande

ژنتیکی صفت را به خود اختصاص داده است. کاربردی‌ترین روشی که برای برآورد تعداد ژن‌ها استفاده می‌شود، بر پایه میانگین و واریانس فنوتیپی دو لاین والدینی و آمیخته‌های آن‌ها یعنی F_1، F_2 و تلاقی‌های برگشتی است. روش بیومتری برای تعیین تعداد ژن‌های مؤثر اولین بار به وسیله کاستل[1] (۱۹۲۱) پیشنهاد شد.

فرض کنید دو لاین والدینی متفاوت شامل (P_1) هموزیگوت برای تمامی آلل‌های افزایش‌دهنده صفت و (P_2) هموزیگوت برای تمامی آلل‌های کاهش‌دهنده صفت وجود دارند. این دو لاین با یکدیگر تلاقی داده می‌شوند تا نسل‌های F_1 و F_2 تولید شوند. در مجموع تعداد m_e ژن مؤثر ناپیوسته هر یک دارای اثر مشابه (a) وجود دارند که کاملاً افزایشی است. میانگین فنوتیپی لاین P_1 و P_2 می‌تواند به ترتیب به‌صورت زیر نوشته شود:

$$\mu_{P_1} = \mu + \sum_{i=1}^{m_e} a = \mu + m_e a$$

$$\mu_{P_2} = \mu + \sum_{i=1}^{m_e} a = \mu - m_e a$$

که تفاوت آن‌ها عبارت است از:

$$\Delta = \mu_{P_1} - \mu_{P_2} = 2m_e a \qquad (۱-۱۴)$$

و در آن، میانگین کل (μ) حذف شده است. بر اساس معادله (۱-۷)، واریانس ژنتیکی F_2 مجدداً می‌تواند به‌صورت زیر نوشته شود:

$$V_G = \frac{1}{2}\sum_{i=1}^{m_e} a^2 = \frac{1}{2} m_e a^2 \qquad (۱-۱۵)$$

رابطهٔ بالا تحت فرضیاتی که در بخش بالا ذکر شد، می‌تواند نوشته شود. با ترکیب معادلات (۱-۱۴) و (۱-۱۵) برآوردگر Castle-Wright برای تعداد ژن‌ها حاصل خواهد شد که به‌صورت زیر است:

$$\widehat{m}_e = \frac{\Delta^2}{8V_G} \qquad (۱-۱۶)$$

که V_G به‌وسیله معادله (۱-۱۲) یا (۱-۱۳) برآورد می‌شود. واریانس نمونه‌گیری $\hat{m}_e$ نیز می‌تواند به صورت زیر تخمین زده شود:

---

[1] - Castle

$$Var(\widehat{m}_e) = \widehat{m}_e^2[\frac{4(V_{P_1}+V_{P_2})}{\Delta^2} + \frac{Var(V_G)}{V_G^2}] \qquad (۱۷-۱)$$

که در آن:

$$Var(V_G) = \frac{2V_{F_2}^2}{n_{F_2}+2} + \frac{2V_{F_1}^2}{n_{F_1}+2}$$

که n_{F_2} و n_{F_1}، تحت شرایطی که معادله (۱-۱۲) استفاده شود، به عنوان اندازه نمونه هستند.

پس از برآوردگر Castle-Wright، مطالعات دیگری به‌منظور برآورد بهتر تعداد ژن‌ها انجام شدند. لند (۱۹۸۱) برآوردگر Castle-Wright را برای استفاده در جمعیت‌های دگربارور تعمیم داد. زنگ و همکاران[1] (۱۹۹۰) و زنگ (۱۹۹۲) برخی از فرضیات مهم نظیر عدم وجود پیوستگی و یکسان بودن اثر افزایشی ژن‌ها که برای برآوردگر Castle-Wright باید صادق باشند را برطرف نمود. آثار اپیستاتیک بین ژن‌های مختلف در مطالعه وو[2] (۱۹۹۶) در نظر گرفته شد که سبب تعمیم برآورد تعداد ژن‌ها برای ارائه شد تصویر کامل‌تری از ساختار ژنتیکی شد. مدل "وو" اختصاصاً برآورد فراسنجه‌های ژنتیکی بیشتری را با در نظر گرفتن نسل‌های P_1، P_2، F_1، F_2 و تلاقی‌های برگشتی در یک آزمایش ممکن می‌سازد. در اظهار نظر کلی، استفاده از روش‌های بیومتری برای برآورد تعداد ژن‌ها، صرف‌نظر از اهمیت آن در درک کلیات تئوری ژنتیک کمّی، در عمل محدود شده است. روش عملی‌تر برای برآورد تعداد ژن بر پایه مکان‌یابی ژنتیکی با استفاده از نشانگرهای مولکولی است که در این حالت، ارتباط بین نشانگرها و تنوع فنوتیپی به وسیله مدل‌های آماری تجزیه و آزمون می‌شود (لندر و بوتستین[3]، ۱۹۸۹).

۱-۸- ژنتیک مولکولی

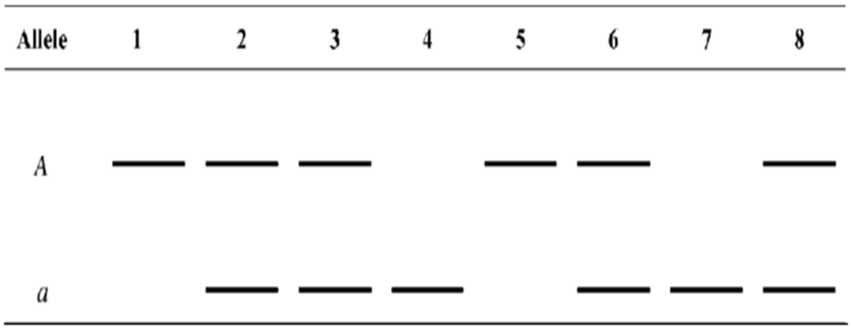
ژنتیک مولکولی که برای تجزیه پیوستگی به‌کار گرفته می‌شود، در ارتباط با فناوری نشانگرهای مولکولی است. نشانگرهای ژنتیکی مولکولی فنوتیپ‌هایی هستند که به آسانی مورد ارزیابی و سنجش قرار می‌گیرند و دارای انطباق مستقیم ۱:۱ با تنوع توالی DNA در منطقه خاصی از ژنوم هستند. در اصل، سنجش یک نشانگر ژنتیکی تحت‌تأثیر عوامل محیطی قرار نمی‌گیرد. نشانگرهای ژنتیکی چند شکلی‌های توالی DNA هستند که از توارث مندلی تبعیت می‌کنند. نشانگر ژنتیکی ایده‌آل برای تهیه نقشه‌های ژنومی، نشانگرهای هم‌بارز، چندآللی و با تنوع بسیار زیاد (یعنی تقریباً

---

[1] - Zeng *et al.*
[2] - Wu
[3] - Lander and Botstein

در هر خانواده‌ای تفرق می‌یابند) هستند؛ اما برخی از نشانگرهای غالب نیز در شرایط خاصی بسیار سودمند و توانمند هستند.

نشانگرهای مولکولی انواع بسیار متفاوتی دارند. چند شکلی‌های موجود در طول قطعات ناشی از هضم (RFLPs) از اولین نشانگرهای ژنتیکی بودند که به‌طور گسترده برای تهیه نقشه‌های ژنومی و مطالعه جمعیت‌ها مورد استفاده قرار گرفتند. نشانگرهای RFLP با استفاده از آنزیم‌های اندونوکلئاز محدود کننده برای ایجاد شکافت دقیق در قطعه DNA ژنومی حاوی توالی ژنی خاص به‌دست می‌آیند. اگر دو موجود زنده از نظر فاصله بین نقاط برش اندونوکلئازهای محدودکننده خاص متفاوت باشند، هر یک طول‌های متفاوتی از قطعات را در هنگامی‌که DNA آن‌ها به وسیله آنزیم محدودکننده هضم می‌شود، تولید خواهند کرد. سپس قطعات می‌توانند به وسیله ژل الکتروفورز جداسازی شوند. دیاگرام موجود در شکل ۱-۴ تفرق یک نشانگر RFLP را در میان ۸ موجود زنده مختلف نشان می‌دهد. افراد ۱ و ۵ برای آلل‌های RFLP بزرگ‌تر و افراد ۴ و ۷ برای آلل‌های RFLP کوچک‌تر هموزیگوت بوده و بقیه هتروزیگوت هستند. نشانگرهای RFLP به‌صورت هم‌بارز توارث می‌یابند و اجازه می‌دهند هر سه ژنوتیپ مربوط به یک جایگاه ژنی منفرد شناسایی شوند.



شکل ۱-۴- دیاگرام الگوی تفرق یک نشانگر RFLP که به وسیله ژل الکتروفورز شناسایی شده است.

تشخیص RFLP ها به هیبرید شدن کاوشگر[1] DNA نشانه گذاری شده با قطعات DNA تک رشته‌ای تقلیب شده[2] نیاز دارد. کاوشگر می‌تواند DNA کلون شده با توالی‌های شناخته شده یا ناشناخته باشد. بیشتر تنوع RFLP به‌واسطه تفاوت‌های درج/حذف است که بین مناطق شناسایی آنزیم محدودکننده واقع شده‌اند و سهم اندکی از تنوع به‌دلیل تنوع توالی درون مناطق

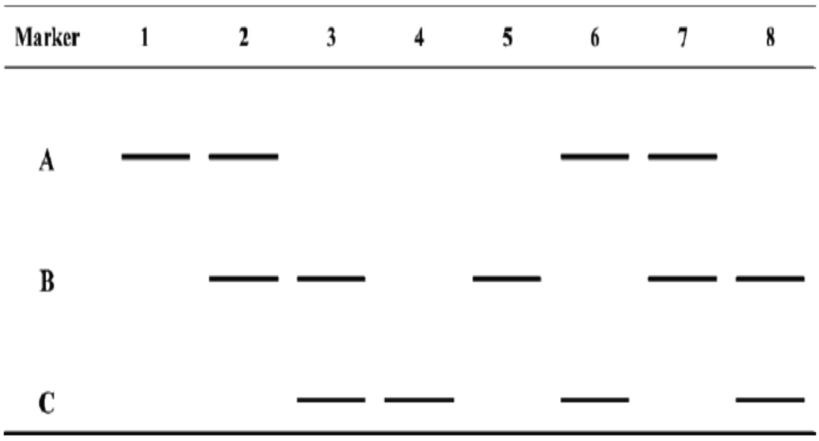
[1] - Probe
[2] - Denatured single-strand DNA fragment

محدودکننده است. RFLP ها به DNA ژنومی زیادی نیاز دارند و انجام آن‌ها در مقایسه با روش‌های مبتنی بر واکنش زنجیره‌ای پلی‌مراز دارای مشکلات بیشتری است.

واکنش زنجیره‌ای پلی‌مراز (PCR[1]) روش مفیدی برای به‌دست آوردن نشانگرهای ژنتیکی مبتنی بر تکثیر قطعات خاص DNA از روی مقادیر کمی از الگوهای DNA ژنومی فراهم می‌سازد. نشانگرهای تکثیر شده به‌وسیله PCR یا مبتنی بر قطعات DNA ژنومی ناشناخته که از نظر اندازه تفاوت دارند هستند (نشانگرهای با توارث هم‌بارز)، یا فقط در برخی افراد (نه در همه) تکثیر می‌شوند (نشانگرهای غالب) و یا به‌طور متمایزی به‌وسیله آنزیم‌های محدودکننده برش داده می‌شوند. همچنین PCR می‌تواند به آسانی تعداد زیادی نشانگر ژنتیکی بر اساس تکثیر DNA ژنومی الگو با استفاده از یک آغازگر کوتاه منفرد[2] تولید نماید (DNA چندشکل تکثیر شده به صورت تصادفی یا[3]RAPD). این نشانگرها توالی‌های DNA ناشناخته‌ای هستند که در جهت مخالف توالی آغازگر در کنار یکدیگر قرار گرفته‌اند. روش توارث نشانگرهای RAPD معمولاً غالب است، توالی موجود یا تکثیر شده و یا تکثیر نمی‌گردد و یک نسخه از دو نسخه به آسانی قابل تشخیص نیست. سیستم نشانگر ناشناخته دیگری که مبتنی بر PCR است، چند شکلی طولی قطعات تکثیر شده (AFLP) است. AFLP ها به وسیله برش DNA ژنومی با استفاده از یک جفت اندونوکلئاز محدودکننده و سپس اتصال سازگارسازها[4] به انتهای قطعات DNA تولید می‌شوند. مجموعه‌ای از قطعات به‌طور انتخابی از روی این قطعات اتصال یافته تکثیر می‌شوند. این قطعات DNA می‌توانند روی ژل‌های توالی‌یاب DNA حل شده و تعداد باندها (نشانگرهای) بیشتری در مقایسه با RAPD ها تولید می‌کنند که اغلب تعداد آن‌ها ۲۰ تا ۴۰ نشانگر (باند) به ازای هر خط ژل است. شکل ۱-۵ دیاگرام تفرق سه نشانگر غالب **A**، **B** و **C** (RAPD یا AFLP) را در ۸ فرد نشان می‌دهد. در نشانگر **A**، افراد ۱، ۲، ۶ و ۷ یک باند را نشان داده‌اند که نشان می‌دهد این افراد برای آلل غالب هموزیگوت بوده و یا هتروزیگوت هستند. برای این نشانگر، افراد ۳، ۴، ۵ و ۸ که هیچ باندی نداده‌اند، باید برای آلل مغلوب هموزیگوت باشند. استنباط مشابهی برای الگوهای در حال تفرق نشانگرهای **B** و **C** می‌تواند صورت گیرد.

---

[1] - Polymerase Chain Reaction
[2] - Single short primer
[3] - Random Amplified Polymorphic DNA
[4] - Adapters



شکل ۱-۵- دیاگرام الگوی تفرق سه نشانگر غالب که با ژل الکتروفورز شناسایی شده‌اند.

نشانگرهای ریزماهواره که همچنین توالی‌های تکراری ساده (SSRs) نیز نامیده می‌شوند، مبتنی بر تکثیر PCR ناحیه ژنومی حاوی یک توالی ساده تکرار شده (یک، دو یا سه نوکلئوتیدی) هستند. تعداد تکرار ریزماهواره‌ها تا حد زیادی می‌تواند متغیر باشد، به‌طوری که بیشتر افراد هتروزیگوت هستند و نسبت افرادی که دارای ژنوتیپ نشانگر مشابه هستند، کم است. ریزماهواره‌ها ممکن است تعداد زیادی آلل در یک جایگاه منفرد SSR (بیش از ۷۰ یا ۸۰) داشته باشند و به‌صورت نشانگرهای هم‌بارز به ارث می‌رسند. فرد ۱ برای بزرگ‌ترین آلل هموزیگوت، فرد ۲ برای بزرگ‌ترین و دومین بزرگ‌ترین آلل هتروزیگوت بوده و افراد دیگر می‌توانند به همین صورت مشاهده شوند.

در عمل، معمولاً دو روش برای شناسایی جایگاه‌های ریزماهواره مناسب جهت استفاده به‌عنوان نشانگرهای ژنتیکی مورد استفاده قرار می‌گیرند. برای برخی گونه‌ها نظیر انسان، موش، *Arabidopsis* و برنج که تاکنون تعداد زیادی از توالی‌های DNA در آن‌ها شناسایی شده‌اند، ریزماهواره‌ها ممکن است با جستجوی در بانک‌های اطلاعاتی توالی DNA برای توالی‌هایی که حاوی تکرارهای ساده هستند، شناسایی شوند؛ اما برای بیشتر گونه‌های گیاهی و جانوری، تلاش‌های بیشتری برای هیبرید شدن و توالی‌یابی مورد نیاز است تا ریزماهواره‌های مناسب برای استفاده به عنوان نشانگرهای ژنتیکی شناسایی شوند.

SNP -۹-۱

چندشکلی تک نوکلئوتیدی (SNP[1]) مکانی در ژنوم است که در آن، توالی‌های DNA تعدادی از افراد از نظر یک تک نوکلئوتید A، T، C یا G با یکدیگر متفاوت است. برای مثال، دو قطعه توالی یابی شده DNA دو فرد متفاوت، AAGCCTA و AAGCTTA، دارای تفاوت تک نوکلئوتیدی است. در این مورد، دو نوکلئوتیدی که سبب تفاوت دو فرد شده‌اند، C و T هستند که دو آلل نامیده می‌شوند. SNP ها به عنوان جدیدترین نشانگرها، بیشترین توجهات و تمرکز را در ژنتیک به خود اختصاص داده‌اند، چون فراوانی زیادی دارند و برای تعیین ژنوتیپ خودکار در مقیاس وسیع به خوبی مناسب هستند. مجموعه متراکمی از نشانگرهای SNP امکان مطالعه اساس ژنتیکی بیماری‌های پیچیده را با استفاده از روش‌های مطالعه جمعیت‌ها فراهم می‌آورند.

SNPها می‌توانند از راه روش‌های اختصاصی یا غیراختصاصی توالی DNA شناسایی شوند. شناسایی غیراختصاصی توالی مبتنی بر احاطه، شکافت یا تغییر حرکت در طی الکتروفورز یا کروماتوگرافی مایع هترودوپلکس‌های ناجور تشکیل شده بین مولکول‌های DNA آللی یا مولکول‌های DNA تک‌رشته ای است که فرض می‌شود تحت شرایط تقلیب نشده دارای شکل‌های فضایی نسبتاً متفاوتی هستند. اگر چه شناسایی غیر اختصاصی توالی چند شکلی‌ها نقطه اتکای اصلی در کشف چندشکلی/جهش است، اما روش قابل قبولی برای تعیین ژنوتیپ نیست، زیرا اگر تعیین ژنوتیپ انجام شده ژنوتیپ حقیقی باشد، در این صورت روش مناسبی نخواهد بود. شناسایی اختصاصی توالی جهت تشخیص آلل‌ها بر اساس چهار ساز و کار کلی صورت می‌گیرد: هیبریداسیون اختصاصی آلل، مشارکت نوکلئوتیدی اختصاصی آلل، اتصال الیگونوکلئوتیدی اختصاصی آلل و برش اختصاصی آلل.

کتاب‌های مرجع ژنتیک عمومی برای مطالعه بیشتر

[1] Anthony J. F. Griffiths, William M. Gelbart, Richard C. Lewontin and Jeffrey H. Miller (2002) *Modern Genetic Analysis: Integrating Genes and Genomes*. 2nd edition. W. H. Freeman, New York.
[2] Daniel L. Hartl and Elizabeth W. Jones (2001) *Genetics: Analysis of Genes and Genomes*. 6th edition. Jones and Bartlett Publishers, Sudbury, Massachusetts.
[3] Benjamin Lewin (2005) *Essential Genes*. Prentice-Hall, Engewood Cliff, New Jersey.

[1] - Single Nucleutide Polymorphism

۱-۱۰- تمرین‌ها

۱-۱- معادلات (۱-۷) و (۱-۸) واریانس ژنتیکی (σ_G^2) ناشی از m ژن را به‌ترتیب در یک جمعیت F_2 و تلاقی برگشتی توصیف می‌کنند. نشان دهید چگونه این دو معادله به‌دست آمده‌اند.

نکته: دو ژن به هم پیوسته A و B را با نرخ نوترکیبی r در تلاقی برگشتی در نظر بگیرید. آثار افزایشی این دو ژن به ترتیب به صورت a_1 و a_2 نشان داده شده‌اند. ارزش‌های ژنتیکی چهار ژنوتیپ تلاقی برگشتی در دو ژن به همراه فراوانی‌های آن‌ها بر حسب r به صورت زیر تعریف می‌شوند:

ژنوتیپ	کد	ارزش ژنتیکی μ_j	فراوانی P_j
AaBb	۱	$\mu + a_1 + a_2$	$\frac{1}{2}(1-r)$
Aabb	۲	$\mu + a_1$	$\frac{1}{2}r$
aaBb	۳	$\mu + a_2$	$\frac{1}{2}r$
aabb	۴	μ	$\frac{1}{2}(1-r)$

واریانس ژنتیکی به‌واسطه این دو ژن به صورت زیر محاسبه می‌شود:

$$\sigma_G^2 = \sum_{j=1}^{4} P_j \mu_j^2 - \left(\sum_{j=1}^{4} P_j \mu_j \right)^2$$

$$= \frac{1}{4}(a_1^2 + a_2^2) + \frac{1}{4}(1-2r)a_1 a_2$$

این فرمول می‌تواند برای m ژن بسط یابد و برای نتاج F_2 نیز به صورت مشابه به‌دست می‌آید.

۱-۲- معادله (۱-۱۶) برآوردگر Castle-Wright برای تعداد ژن است. فرضیاتی که برای به‌دست آوردن آن استفاده می‌شوند را بیان کنید.

۱-۳- برآوردگر Castle-Wright را به گونه‌ای تغییر دهید تا با شرایط زیر انطباق یابد:

الف. هر ژن دارای هر دو اثر افزایشی و غالبیت باشد.

ب. برخی ژن‌ها دارای اثر افزایش‌دهنده و برخی دیگر دارای اثر کاهش‌دهنده روی P_1 باشند و بالعکس برای P_2.

ج. ژن‌های متفاوت روی یک کروموزوم به هم پیوسته باشند.

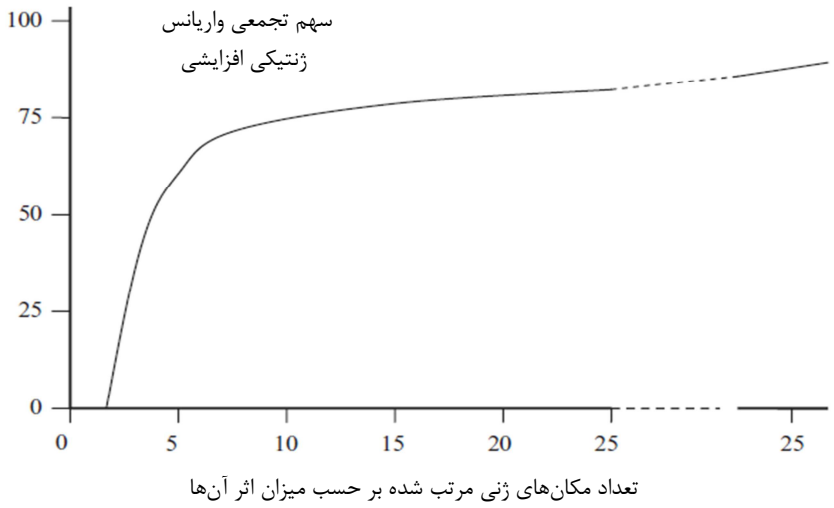
د. ژن‌های مختلف با یکدیگر اثر متقابل دارند تا آثار اپیستازی را نشان دهند.

ممکن است تعداد نسل‌های بیشتری را در مدل خود در نظر بگیرید که شما را قادر می‌سازد تا پارامترهای ژنتیکی بیشتری را برآورد کنید. زنگ و همکاران (۱۹۹۰)، زنگ (۱۹۹۲) و وو (۱۹۹۶) را برای بحث‌های کامل‌تر در خصوص موضوعات این بخش مطالعه کنید.

۱-۱۱- نکات

آنچه که در ادامه مطرح می‌شود، ارائه راهکاری برای ایجاد مدل با آثار ژنتیکی نابرابر برای جایگاه‌های ژنی متفاوت یک صفت کمّی است. همانگونه که در شکل ۱-۱۱-۱ نشان داده شده است، آثار ژنتیکی در بین ژن‌های مختلف و به ویژه در الگوی الیگوژنیک متفاوت هستند، یعنی تعداد اندکی از جایگاه‌های ژنی سهم زیادی از تنوع ژنتیکی را به خود اختصاص می‌دهند، ولی بسیاری از جایگاه‌های دیگر سهم ناچیزی از تنوع را توجیه می‌کنند (شریمپتون و روبرتسون، ۱۹۸۸a,b). این الگو اجازه می‌دهد تا بتوانیم آثار ژنتیکی جایگاه‌های ژنی را با تابع گاما یا سری‌های هندسی مدل‌سازی کنیم.

۱-۱۱-۱- ایجاد مدل آثار ژنتیکی نامساوی با تابع گاما



شکل ۱-۷- نمایش نظری گرافیکی رابطه بین تعداد جایگاه‌های ژنی تعیین کننده یک خصوصیت خاص و سهم تجمعی واریانس ژنتیکی افزایشی که به وسیله چنین جایگاه‌هایی در نظر گرفته می‌شود.

فرض کنید که n_e ژن مسئول بروز صفتی هستند که آثار ژنتیکی افزایشی و غالبیت آن‌ها متفاوت است. یک شجره سه نسله شامل P_1، P_2، F_1، F_2 و تلاقی‌های برگشتی (B_1 و B_2) در نظر

گرفته شده است. فرض کنید که اثر افزایشی (a) در یک جایگاه ژنی از توزیع گاما تبعیت کند (کیمورا[1]، ۱۹۷۹، هیل و راسباش[2]، ۱۹۸۶) که تابع چگالی احتمال آن به صورت زیر است:

$$f(a)=\frac{\alpha^{\beta}e^{-\alpha a}a^{\beta-1}}{\Gamma(\beta)},0\le a<\infty,0<\alpha,\beta<\infty$$

که در آن، α پارامتر مقیاسی توزیع گامای اثر افزایشی و β فراسنجه شکل مربوطه است. گشتاورها برای این توزیع عبارتند از:

برای a ,

$$\varepsilon(a)=\frac{\beta}{\alpha},\varepsilon(a^2)=\frac{\beta(1+\beta)}{\alpha^2},V(a)=\frac{\beta}{\alpha^2}$$

که ε نمایانگر امید ریاضی است. فراسنجه β می‌تواند برای اندازه‌گیری تساوی آثار افزایشی و غالبیت در ژن‌های مختلف مورد استفاده قرار گیرد. زمانی که $\beta\to\infty$، در حالت اثر ژنتیکی برابر، توزیع به همگرایی می‌رسد. بر اساس گشتاورهای فوق، عبارت زیر را می‌توانیم به‌دست آوریم:

$$\frac{\sum_{i=1}^{m_e}a_i^2}{m^*}=\overline{a^2}=\varepsilon(a^2)=\frac{1+\beta}{\beta}\left[\varepsilon(a)\right]^2$$

که $\bar{a}$ و $\overline{a^2}$ به ترتیب متوسط آثار افزایشی و مربع آثار افزایشی برای ژن‌های مربوطه است. با نادیده گرفتن اثر غالبیت، واریانس ژنتیکی F_2 به‌واسطه m^* ژن با اندازه نامساوی و ناپیوسته، مطابق با رابطه ۱-۷، می‌تواند به صورت ذیل بیان شود:

$$V_G=\frac{1}{2}\sum_{i=1}^{m^*}a_i^2=\frac{1}{2}m^*\varepsilon(a^2)=\frac{1}{2}m^*\left(\frac{1+\beta}{\beta}\right)(\bar{a})^2$$

و تفاوت بین دو لاین والدینی به صورت زیر نشان داده می‌شود:

$$\Delta=\mu_{P_1}-\mu_{P_2}=2m^*\bar{a}$$

از ترکیب دو معادله فوق می‌توان به برآوردی از m^* به صورت زیر دست یافت:

$$\hat{m}^*=\frac{\Delta^2}{8V_G}\left(\frac{1+\beta}{\beta}\right)=\hat{m}_e\left(\frac{1+\beta}{\beta}\right)$$

---

[1] - Kimura
[2] - Hill and Rasbash

که $\widehat{m}_e$ برآورد تعداد ژن در هنگامی است که تمامی ژن‌ها دارای اثر مساوی باشند (معادله (۱-۱۶)). پارامتر $\frac{(1+\beta)}{\beta}$ می‌تواند برای توصیف الگوی توزیع آثار ژنتیکی مورد استفاده قرار گیرد. اگر آثار منفرد (a_i) دارای توزیع نرمال باشند، سپس خواهیم داشت: $\frac{(1+\beta)}{\beta}=\frac{\pi}{2}=1.57$. در هر حال، توزیع‌هایی که دارای کشیدگی زیادی هستند، می‌توانند منجر به مقادیر $\frac{(1+\beta)}{\beta}$ بیش از $\frac{\pi}{2}$ شوند (مک کی و همکاران، ۱۹۹۲)؛ بنابراین، با فرض آثار ژنتیکی مشابه و تحت شرایطی که آثار ژنتیکی حقیقتاً در بین ژن‌ها مساوی نیستند، همیشه تعداد ژن کمتر از حد واقعی برآورد می‌شود.

۱-۱۱-۲- ایجاد مدل آثار ژنتیکی نامساوی با سری‌های هندسی

مدل الیگوژنیک بیان می‌کند که کنترل ژنتیکی یک صفت کمّی به وسیله تعداد کمی ژن با آثار بزرگ و تعداد زیادی ژن با اثر کوچک صورت می‌گیرد و اشاره ضمنی می‌نماید که توزیع آثار ژنتیکی افزایشی (a) ممکن است به وسیله سری‌های هندسی (لند و تامپسون، ۱۹۹۰) تخمین زده شود، به عبارت دیگر:

$$a, ar, ar^2, ar^3, ..., ar^m, r \neq 1$$

که r نسبتی است که بزرگی نسبی اثر افزایشی هر یک از m ژن را تعیین می‌کند. در این مورد، تفاوت بین دو والد و واریانس ژنتیکی F_2 می‌تواند به‌صورت زیر نوشته شود:

$$\Delta = 2a\left(\frac{1-r^m}{1-r}\right)$$

$$V_G = \frac{a^2}{2}\left(\frac{1-r^{2m}}{1-r^2}\right)$$

که منجر به رابطه زیر می‌شود:

$$\widehat{m}_e = \frac{A^2}{8V_G} = \frac{1-r^m}{1+r^m}\left(\frac{1+r}{1-r}\right) \qquad (۱-۱۹)$$

بنابراین خواهیم داشت:

$$r^m = \frac{1-\widehat{m}_e\left(\frac{1-r}{1+r}\right)}{1+\widehat{m}_e\left(\frac{1-r}{1+r}\right)}$$

و

$$\widehat{m} = \log_r\left[1-\widehat{m}_e\left(\frac{1-r}{1+r}\right)\right] - \log_r\left[1+\widehat{m}_e\left(\frac{1-r}{1+r}\right)\right] \qquad (۲۰-۱)$$

که با در نظر گرفتن $u = \frac{1+r^m}{1-r^m}$، می‌توانیم r را بر اساس معادله (۱-۱۹) به صورت زیر به‌دست آوریم:

$$r = \frac{\widehat{m}_e u - 1}{\widehat{m}_e u + 1}$$

برای مدل کنترل الیگوژنیک که در آن m بزرگ و r کوچک است، منطقی به نظر نمی‌رسد که فرض کنیم $u \simeq 1$ و این حالت منجر به عبارت زیر می‌شود:

$$r^* = \frac{\widehat{m}_e - 1}{\widehat{m}_e + 1}$$

بنابراین، تعداد ژن‌ها می‌تواند به‌طور تقریبی با جایگزینی r^* در معادله (۱-۲۰) برآورد شود.

فصل دوم

آمار پایه

۲-۱- مقدمه

حال که با اساس ژنتیک آشنا شدیم، به مقدمه‌ای در خصوص روش‌های آماری که در این کتاب از آن‌ها استفاده خواهد شد، اشاره می‌نماییم. بیشتر استنباط‌های آماری که انجام خواهیم داد بر اساس تجزیه درستنمایی خواهد بود و علاوه بر این‌که به ایجاد تابع درستنمایی مناسب برای مدل معینی می‌پردازیم، به ارائه روش‌هایی که برای محاسبه و بهینه‌سازی درستنمایی به‌کار می‌روند نیز خواهیم پرداخت. در این بخش برخی از مسایل پایه‌ای را بیان می‌کنیم و مسیر کاری خود را در جهت تجزیه درستنمایی یک مدل پیوستگی ژنتیکی قرار می‌دهیم.

۲-۱-۱- جمعیت‌ها و مدل‌ها

هدف از تجزیه آماری، به‌دست آوردن نتایجی در خصوص یک جمعیت از راه یک نمونه است. جمعیت به مجموعه‌ای از اجزا یا افراد گفته می‌شود که معمولاً نامحدود بوده و تمامی افراد قابل اندازه‌گیری نیستند، درحالی‌که یک نمونه، مجموعه کوچک‌تری از اجزایی است که از جمعیت بیرون کشیده شده و تمامی آن‌ها قابل اندازه‌گیری هستند. برای برقراری ارتباط نمونه با جمعیت و داشتن توانایی انجام استنباط، از مدل استفاده می‌کنیم.

مثال ۲-۱ **(ارتفاع گیاه گوجه فرنگی)**. فرض کنید داده‌های زیر (**y**) ارتفاع بوته مربوط به ۱۲ گیاه گوجه‌فرنگی از یک گونه خاص (بر حسب سانتی‌متر) باشد:

$y = (y_1, y_2, \ldots, y_{12}) = (79,82,85,87,100,101,102,103,124,125,126,127)$

ممکن است مدل ساده زیر را در نظر بگیریم. در این مدل، μ میانگین حقیقی ارتفاع بوته جمعیت و Y_i مشاهده (داده) مربوط به نمونه i ام است:

$$Y_i = \mu + \varepsilon_i \qquad (۱-۲)$$

که در آن، ε_i عبارت از خطای آزمایش است که به صورت ویژه دارای توزیع نرمال با میانگین صفر و واریانس σ^2 است.

برای مدل (۲-۱)، اغلب فرض می‌شود که ε_i از توزیع نرمال $N(0,\sigma^2)$ تبعیت کرده و تابع چگالی احتمال آن $N(\mu,\sigma^2)$ به‌صورت زیر ارایه می‌شود:

$$\phi(y \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

بنابراین، با فرض وجود توزیع نرمال، می‌توانیم مدل (۲-۱) را بـه صـورت $Y_i \sim \phi(y \mid \mu, \sigma^2)$ بنویسیم.

با پیچیده‌تر شدن فرایند توصیفی، مدل‌های مورد استفاده نیز پیچیده‌تـر خواهنـد شـد. در ایـن کتاب، انواع مختلفی از مدل‌ها را به همراه اشـکالی از آن‌هـا کـه اغلـب بـه وسـیله ماهیـت مسـئله مشخص می‌شوند، مورد آزمون قـرار خـواهیم داد. در ایـن بخـش برخـی از مثـال‌هـای مربـوط بـه مدل‌های دیگر را ارائه می‌دهیم:

الف. تابعیت خطی. برای توصیف رابطه خطی بین یک متغیر وابسته Y و یـک متغیـر مسـتقل (یـا متغیر همبسته) x می‌توانیم از این مدل استفاده کنیم:

$$Y_i = a + bx_i + \varepsilon_i$$

اگر تعداد زیادی متغیر همبسته مختلف داشته باشیم، می‌توانیم از مدل تابعیت چندگانـه‌ای بـه شکل $Y_i = a + \sum_j b_j x_{ij} + \varepsilon_i$ استفاده کنیم.

ب. برای توصیف رابطه بین یک متغیر همبسته و یک متغیر تصادفی برنولی (متغیـر کیفـی Y کـه فقط مقادیر ۰ و ۱ را می‌پذیرد، به‌طوری‌که $P(Y=1) = p$)، اغلب مدل تابعیت لجسـتیک مـورد استفاده قرار می‌گیرد که به شکل زیر تعریف می‌شود:

$$\text{logit}(p(x)) = a + bx$$

که logit به صورت $\text{logit}(p) = \left(\frac{p}{1-p}\right)$ تعریف شده و $P(Y=1 \mid x) = p(x)$ است.

ج. مدل‌های زیادی وجود دارند که برای توصیف فرایند رشد به‌صورت تابعی از زمان مـورد اسـتفاده قرار می‌گیرند. یکی از مدل‌های معروف در این خصوص، منحنی رشد لجستیک است که بـه‌صـورت زیر ارائه می‌شود:

$$g(t) = \frac{a}{1 + be^{-rt}}$$

که t اختصاصاً به‌عنوان معیار اندازه‌گیری زمان است. اگر رشد واقعی را مشاهده کنیم، می‌توانیم از این مدل به شکل $Y_i = \frac{a}{1 + be^{-rt_i}} + \varepsilon_i$ استفاده نماییم.

د. آخرین مدلی که در این بخش توصیف خواهد شد کاربرد بیشتری در زمینه مکان‌یابی QTL دارد (فصل ۹ را مشاهده کنید) و آن، مدل مختلطی است که به‌صورت زیر ارائه می‌شود:

$$Y_i \begin{cases} \mu_1 + \varepsilon_i & \text{با احتمال p} \\ \mu_2 + \varepsilon_i & \text{با احتمال 1-p} \end{cases}$$

یا معادل با آن:

$$Y_i \sim \begin{cases} N(\mu_1, \sigma^2) & \text{با احتمال p} \\ N(\mu_2, \sigma^2) & \text{با احتمال 1-p} \end{cases}$$

همچنین می‌توانیم مدل مختلط را به‌صورت زیر بنویسیم:

$$Y_i = \mu_1 I(X=1) + \mu_2 I(X=0) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

که $P(X=1) = p$ و I تابع موهومی است و در صورت صحت استدلال صورت گرفته برابر ۱ و در غیر این‌صورت برابر صفر است.

۲-۱-۲- نمونه‌ها

برای شروع تحقیق و پژوهش، اصولاً باید یک نمونه تصادفی که مجموعه کوچکی از اجزایی است که نمایانگر جمعیت هستند را از جمعیت مربوطه استخراج کنیم. یک نمونه تصادفی، اساساً نمونه‌ای n عضوی است و طوری از جمعیت استخراج می‌شود که تمامی مجموعه n عضوی شانس مساوی برای قرار گرفتن در نمونه استخراج شده را داشته باشند. در هر حال، در عمل امیدواریم که نمونه‌ای داشته باشیم که مستقل و دارای توزیع مشابه بوده یا به عبارت دیگر iid[1] باشد.

برای اینکه اختصاصی شویم، اگر جمعیتی با تابع چگالی احتمال $f(y|\theta)$ به‌طوری‌که θ دارای پارامترهای ناشناخته بوده و یک نمونه iid به صورت $Y_1, Y_2, ..., Y_n$ داشته باشیم، می‌خواهیم هر متغیر از رابطه $Y_i \sim f(y|\theta)$ (یعنی دارای توزیع مشابه) تبعیت کرده و متغیرها مستقل باشند. اگر یک نمونه iid را به صورت $y_1, y_2, ..., y_n$ استخراج کنیم، تابع چگالی احتمال نمونه به‌صورت زیر خواهد بود:

$$f(y_1, y_2, ..., y_n | \theta) = \prod_{i=1}^{n} f(y_i | \theta)$$

[1] - Independent and identically distributed

مثال ۲-۲ **(چگالی نمونه نرمال)**. فرض کنید که $Y_1, Y_2, ..., Y_n$ یک نمونه iid از جمعیت $N(\mu, \sigma^2)$ باشد. در این حالت چگالی نمونه عبارت است از:

$$\begin{aligned}\phi(y_1, y_2, ..., y_n \mid \mu, \sigma^2) &= \prod_{i=1}^{n} \phi(y_i \mid \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}\end{aligned} \tag{۲-۲}$$

چگالی نمونه، رابطه بین نمونه و مدل را تعریف می‌کند و بنابراین تنها وسیله‌ای است که برای برآورد پارامترهای ناشناخته جامعه در اختیار داریم. درحقیقت، روش‌های دیگری نیز وجود دارند، ولی هنگامی‌که همهٔ آن‌ها با برآوردهای مبتنی بر چگالی نمونه مقایسه می‌شوند، فاقد کارآیی هستند (کاسلا و برگر[1]، ۲۰۰۱ را مشاهده کنید). به‌علاوه، تمامی اطلاعاتی که نمونه درباره پارامترهای جامعه ارائه می‌دهد در تابع چگالی نمونه قرار دارند. نتیجه این امر آن است که باید روش‌های برآورد پارامتر خود را متناسب با چگالی نمونه انتخاب کنیم.

۲-۲- برآورد درستنمایی

معمول‌ترین روش برآورد که روش اختصاصی و بسیار مناسبی است، بر پایه چگالی نمونه استوار است. می‌توانیم تابع درستنمایی را به‌صورت زیر تعریف کنیم:

$$L(\theta \mid y_1, y_2, ..., y_n) = f(y_1, y_2, ..., y_n \mid \theta)$$

که صرفاً چگالی نمونه f را به‌عنوان تابعی از پارامترها در نظر گرفته و داده‌ها را ثابت نگه می‌دارد. روش برآورد حداکثر درستنمایی، برآورد θ را به‌گونه‌ای انجام می‌دهد که درستنمایی حداکثر شود.

مثال ۲-۳. **(تعیین برآوردگر حداکثر درستنمایی[2] (MLE))**. برای یافتن MLE های مثال ۲-۲، نیاز داریم تا مقادیر μ و σ^2 را که معادله (۲-۲) را به حداکثر می‌رسانند مشخص کنیم. چون توابع درستنمایی حاصل از نمونه‌های iid به‌صورت حال ضرب هستند (که به حداکثر رساندن

---

[1] - Casella and Berger
[2] - Maximum likelihood estimator

آن‌ها مطلوب به نظر نمی‌رسد)، اغلب روش ساده‌تر آن است که با لگاریتم درستنمایی (که به صورت log-likelihood شناخته می‌شود) کار کنیم. مقادیر حداکثر شده این تابع مشابه مقادیر مربوط به همان تابع اصلی است و انجام محاسبات نیز کاملاً ساده‌تر هستند.

باتوجه به معادله (۲-۲)، لگاریتم درستنمایی نرمال عبارت است از:

$$\log L(\mu,\sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

با مشتق‌گیری و برابر صفر قرار دادن عبارت فوق خواهیم داشت:

$$\frac{\partial}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) = 0$$

$$\frac{\partial}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i-\mu)^2 = 0$$

که MLE (برآورد حداکثر درستنمایی) میانگین برابر با $\hat{\mu} = \bar{y}$ و واریانس برابر با $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}$ است. برای داده‌های مثال ۲-۱، داریم: $\hat{\mu}$ = ۱۰۳/۴۱ و $\hat{\sigma}^2$ = ۳۰۳/۲۴.

منطق و اساس پشت پرده برآورد حداکثر درستنمایی به‌صورت زیر است. اگر نمونه $y_1, y_2, ..., y_n$ مدنظر باشد، سپس مقادیر y_i بایستی از روی مناطقی از f حاصل شوند که دارای احتمال بالایی هستند، یعنی نمونه بایستی تحت نما (مد) f قرار گیرد. در خصوص موقعیت این نما اطلاعی نداریم، چون مقدار θ را نمی‌دانیم. در هر حال، برای نمونه معین، می‌توانیم مقدار θ را به‌گونه‌ای بیابیم که $f(y_1, y_2, ..., y_n \mid \theta)$ و یا معادل آن، $L(\theta \mid y_1, y_2, ..., y_n)$ به حداکثر برسد. این فرض باعث می‌شود تا نمونه در بالاترین منطقه احتمال قرار گیرد و برآوردگر حاصله در این حالت، برآوردگر حداکثر درستنمایی (MLE) است.

مثال ۲-۴. **(مجدداً داده‌های گوجه فرنگی را ملاحظه نمایید).** به‌عنوان مثالی که تا حدودی واقعی‌تر است، به ارتفاع گوجه‌فرنگی‌ها در مثال ۲ باز می‌گردیم، ولی در حال حاضر می‌خواهیم بدانیم که شواهدی مبنی بر وجود کنترل ژنتیکی در ارتفاع وجود دارد. به‌طور اختصاصی، اگر ژن "ارتفاع" در نتاج F_2 بر اساس قانون اول مندل در حال تفرق باشد، انتظار داریم تا ژنوتیپ‌های AA:Aa:aa بر اساس نسبت ۱:۲:۱ تفرق یابند. فرض می‌کنیم که ژنی مرتبط با ارتفاع وجود دارد و بر اساس نسبت p:q:1-p-q برای ژنوتیپ‌های AA:Aa:aa در حال تفرق باشد. زمانی که

مقادیر p و q ناشناخته باشند و زمانی که در خصوص ژنوتیپ گیاهی که مشاهده می‌کنیم اطلاعی نداشته باشیم، مدل ۲-۱ یک مدل مختلط می‌شود:

$$Y_i = \begin{cases} \mu_{AA} + \varepsilon_i & \text{با احتمال p} \\ \mu_{Aa} + \varepsilon_i & \text{با احتمال q} \\ \mu_{aa} + \varepsilon_i & \text{با احتمال 1-p-q} \end{cases}$$

که در آن، ε_i دارای توزیع نرمال با میانگین صفر و واریانس σ^2 است، یعنی: $\varepsilon_i \sim N(0,\sigma^2)$.

اگر ϕ_{AA} دارای چگالی نرمال با میانگین μ_{AA} و واریانس σ^2 و ϕ_{Aa} و ϕ_{aa} را نیز به‌صورت مشابه تعریف کنیم، با نوشتن $y = (y_1, y_2, ..., y_n)$، توابع درستنمایی و لگاریتم درستنمایی را می‌توانیم به‌صورت زیر بیان کنیم:

$$L(\mu_{AA}, \mu_{Aa}, \mu_{aa}, \sigma^2 \mid y) = \prod_{i=1}^{n} \left[p\phi_{AA}(y_i) + q\phi_{Aa}(y_i) + (1-p-q)\phi_{aa}(y_i) \right]$$

$$\log L(\mu_{AA}, \mu_{Aa}, \mu_{aa}, \sigma^2 \mid y) = \sum_{i=1}^{n} \log \left[p\phi_{AA}(y_i) + q\phi_{Aa}(y_i) + (1-p-q)\phi_{aa}(y_i) \right]$$

اگر ژن بر اساس نسبت ۱:۲:۱ در حال تفرق باشد، در این صورت جمعیت F_2 مندلی در اختیار خواهیم داشت.

همچنین می‌توانیم برآوردهای حداکثر درستنمایی را از راه مشتق‌گیری به‌دست آوریم. اگر چه مشاهده خواهید کرد که هیچ‌گونه جواب واضح و مشخصی وجود ندارد، ولی معادلات منجر به ایجاد جواب‌های تکراری می‌شوند. این جواب قابلیت کاربرد کلی‌تری خواهد داشت، بنابراین برای متوجه شدن این قضیه، معادلات درستنمایی را برای مدل مختلط عمومی حل کرده و سپس به مثال باز خواهیم گشت.

با فرض دارا بودن مشاهدات iid به صورت $y_1, y_2, ..., y_n$ از مدل مختلط عمومی:

$$Y_i \sim \sum_{j=1}^{k} p_j f(y_i \mid \theta_j)$$

$$\sum_j p_j$$

لگاریتم درستنمایی عبارت است از:

$$LogL(\theta \mid y) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} p_j f(y_i \mid \theta_j) \right)$$

همچنین می‌توانیم بنویسیم: $\theta = (\theta_1, ..., \theta_k)$.

حال، فرض می‌کنیم که p و q شناخته شده هستند، ولی بعداً به این حالت باز خواهیم گشت. مشتق معادله بالا نسبت به $\theta_{j'}$ معادله زیر را ایجاد می‌کند:

$$\frac{\partial}{\partial\theta_{j'}}\log L(\theta\,|\,y)=\sum_{i=1}^{n}\frac{\sum_{j=1}^{k}p_j\frac{\partial}{\partial\theta_{j'}}f(y_i\,|\,\theta_j)}{\sum_{j=1}^{k}p_j f(y_i\,|\,\theta_j)}$$

توجه داشته باشید که اگر $\theta_{j'}$ هیچ مؤلفه‌ای برابر با θ_j نداشته باشد، مشتق صورت کسر برابر با صفر خواهد بود، مگر آنکه $j=j'$. این حالت به تفصیل در مثال ۲-۵ توضیح داده خواهد شد.

در حال حاضر، برای ساده نمودن می‌توان نوشت:

$$\frac{\partial}{\partial\theta_{j'}}f(y_i\,|\,\theta_j)=f(y_i\,|\,\theta_j)\frac{\partial}{\partial\theta_{j'}}\log(f(y_i\,|\,\theta_j))$$

$$P_j(y_i)=\frac{p_j f(y_i\,|\,\theta_j)}{\sum_{j=1}^{k}p_j f(y_i\,|\,\theta_j)}$$

به‌عبارت دیگر:

$$\frac{\partial}{\partial\theta_{j'}}\log L(\theta\,|\,y)=\sum_{i=1}^{n}\sum_{j=1}^{n}P_j(y_i)\frac{\partial}{\partial\theta_{j'}}\log(f(y_i\,|\,\theta_j)) \qquad (۳-۲)$$

همچنین می‌توانیم برآوردهای حداکثر درستنمایی را با الگوریتم تکراری زیر حل نماییم و مقادیر اولیه را به‌صورت $\theta_j^{(0)}$، $P_j^{(0)}(y_i)$ و j=1,…,k در نظر می‌گیریم. برای t=0,1,…:

۱- برای j=1,…,k، $P_j^{(t+1)}(y_i)=\frac{p_j f(y_i\,|\,\theta_j^{(t)})}{\sum_{j=1}^{k}p_j f(y_i\,|\,\theta_j^{(t)})}$

۲- برای $j'=1,\ldots,k$، در عبارت زیر برای $\theta_{j'}^{(t+1)}$ می‌توان معادله را حل نمود:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}P_j^{(t+1)}(y_i)\frac{\partial}{\partial\theta_{j'}}\log(f(y_i\,|\,\theta_j))=0$$

۳- مقدار t افزایش یافته و به ۱ باز می‌گردد. این مراحل تا رسیدن به همگرایی تکرار می‌شوند.

مثال ۲-۵. **(مختلط نرمال).** اگر $f(y_i\,|\,\theta_j)$ دارای توزیعی به شکل $N(\mu_j,\sigma^2)$ باشد، در هنگامی‌که نسبت به $\mu_{j'}$ مشتق گرفته می‌شود، صورت کسر در معادله (۲-۳) فقط یک جمله دارد، جمله‌ای که در آن $j=j'$ است؛ اما مشتق معادله فوق نسبت به σ^2 دارای مجموع کامل

است، چون پارامتر σ^2 برای تمامی چگالی‌های موجود در مدل مختلط، مشترک است؛ بنابراین داریم:

$$\sum_{i=1}^{n} P_j^{(t+1)}(y_i)\frac{\partial}{\partial\mu_j}\log(f(y_i\mid\mu_j,\sigma^2)) = \sum_{i=1}^{n} P_j^{(t+1)}(y_i)\frac{1}{2\sigma^2}(y_i-\mu_j)$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} P_j^{(t+1)}(y_i)\frac{\partial}{\partial\mu_j}\log(f(y_i\mid\mu_j,\sigma^2)) = \sum_{i=1}^{n}\sum_{j=1}^{k} P_j^{(t+1)}(y_i)$$

$$\times\left[-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}(y_i-\mu_j)^2\right]$$

با مساوی صفر قرار دادن این معادلات و حل کردن آن‌ها، می‌توانیم هر μ_j را به‌طور جداگانه در نظر گرفته و عبارت زیر را به‌دست آوریم:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} y_i P_j^{(t+1)}(y_i)}{\sum_{i=1}^{n} P_j^{(t+1)}(y_i)}$$

برای σ^2، پس از جایگزینی $\mu_j^{(t+1)}$ به جای μ_j داریم:

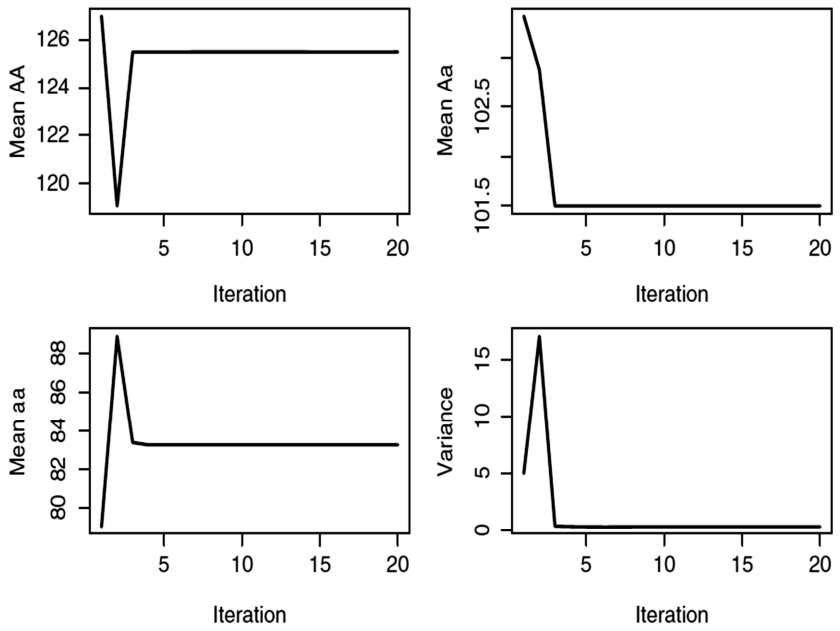
$$\sigma^{2(t+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} P_j^{(t+1)}(y_i)\left(y_i-\mu_j^{(t+1)}\right)^2}{n\sum_{j=1}^{k} P_j^{(t+1)}(y_i)}$$

مثال ۲-۶. **(مجدداً داده‌های گوجه فرنگی را در نظر بگیرید، داده‌ها تکمیل شده‌اند).** اگر ژنی بر اساس نسبت ۱:۲:۱ در حال تفرق باشد، در این صورت جمعیت F_2 مندلی خواهیم داشت. در اینجا پارامترهای توزیع نرمال تحت چنین سناریویی برآورد شده‌اند.

با استفاده از مراحلی که در بالا توضیح داده شدند، برآوردهای زیر به‌دست می‌آیند:

$$\hat{\mu}_{AA} = 125.5 \qquad \hat{\mu}_{Aa} = 101.5 \qquad \hat{\mu}_{aa} = 83.25 \qquad \hat{\sigma}^2 = 0.324$$

توالی تکرارها در شکل ۲-۱ نشان داده شده است که در آن، همگرایی سریع الگوریتم مشاهده می‌شود. یک برنامه R برای به‌دست آوردن مجدد شکل ۲-۱ در ضمیمه B-2 ارایه شده است.



شکل ۲-۱- گراف‌های تکرارها برای برآورد حداکثر درستنمایی (MLE) مربوط به مثال ۲-۶. شکل‌ها همگرایی برآوردهای مربوط به سه میانگین و واریانس را نشان می‌دهند.

۲-۳- آزمون فرض

یکی از فعالیت‌های عمده در تجزیه و تحلیل‌های آماری، آزمون فرض‌ها است. در این بخش، تعدادی از روش‌های کلاسیک و نوین را مرور می‌کنیم.

۲-۳-۱- آزمون مربع کای پیرسون

ابتدا حالتی را در نظر می‌گیریم که در آن، دو کلاس ژنوتیپی وجود دارد. فرض کنید که N فرد با احتمال p برای قرار گرفتن در اولین کلاس داریم. اگر افراد مستقل از هم باشند، احتمال مشاهده n_1 فرد در اولین کلاس (و $n_2=N-n_1$ فرد در دومین کلاس) با توزیع دوجمله‌ای زیر بیان می‌شود:

$$P(n_1,n_2)=\frac{N!}{n_1!n_2!}p^{n_1}(1-p)^{n_2}$$

میانگین توزیع برابر N_p و واریانس آن برابر $N_p(1-p)$ است. در نمونه‌های بزرگ، تقریب توزیع دوجمله‌ای به‌طور اختصاصی به وسیله توزیع نرمال با همان میانگین و واریانس صورت می‌گیرد؛ بنابراین:

$$Z = \frac{n_1 - N_p}{(N_p(1-p))^{\frac{1}{2}}} \qquad (۴-۲)$$

عبارت بالا تقریباً یک متغیر تصادفی نرمال استاندارد است و از این‌رو Z^2 تقریباً یک متغیر تصادفی مربع کای با یک درجه آزادی است. می‌توان نشان داد که:

$$Z^2 = x_1^2 = \frac{(n_1 - N_p)^2}{N_p} + \frac{(n_2 - N(1-p))^2}{N(1-p)} \qquad (۵-۲)$$

که فرمول معمول برای آماره مربع کای پیرسون است، یعنی، مجموع مربعات تفاوت‌ها بین مقادیر مشاهده شده (O) و مورد انتظار (E) تقسیم بر مقادیر مورد انتظار:

$$x^2 = \sum \frac{(O-E)^2}{E} \qquad (۶-۲)$$

توجه کنید که در آزمون نسبت تفرق ۱:۱، معادله (۲-۶) می‌تواند به صورت $\frac{(n_1 - n_2)^2}{N}$ نوشته شود. در حقیقت، آزمون تفرق می‌تواند به‌طور مستقیم با استفاده از معادله (۲-۴) بر اساس آزمون نرمال استاندارد صورت گیرد که نتایج مشابهی را همانند آزمون مربع کای پیرسون معادله (۲-۵) ایجاد می‌کند؛ اما آزمون مربع کای پیرسون از این لحاظ مزیت دارد که آماره آزمون (۲-۶) می‌تواند به مواردی تعمیم یابد که بیش از دو دسته ژنوتیپی در آن وجود دارد. در چنین مواردی، توزیع چند جمله‌ای برای ایجاد مدل مشاهدات استفاده می‌شود و تقریب کای مربع که در معادله (۲-۶) ارائه شده است به کار می‌رود.

به‌طورکلی، زمانی‌که m حالت مشاهده شده وجود داشته باشد، آماره کای مربع پیرسون از محاسبه m حالت مورد انتظار حاصل می‌شود که به‌طور مجانبی دارای توزیع کای مربع با m-k درجه آزادی است. در این حالت، m، تعداد حالات مشاهده شده (سلول‌ها) منهای یک و k، تعداد پارامترهایی است که برای محاسبه مقادیر مورد انتظار برآورد می‌شوند (بخش ۲-۳-۲ را مشاهده کنید).

برای تلاقی برگشتی که در آن دو کلاس ژنوتیپی وجود دارد، m=1 است. این حالت به این دلیل است که فقط یک کلاس می‌تواند به‌طور اختیاری پر شود، چون زمانی که یک عدد به اولین کلاس انتساب داده شود، عدد مربوط به کلاس دوم مشخص خواهد شد. برای F_2 که درمجموع دارای سه کلاس ژنوتیپی برای نشانگر هم‌بارز است، m=2 است.

اگر N خیلی بزرگ نباشد، توزیع دو جمله‌ای ممکن است به وسیله توزیع نرمال به خوبی تقریب زده نشود. در چنین مواردی، بهترین عملی که می‌توان انجام داد، محاسبه مقدار صحیحی از

p است (بخش ۲-۳-۳ را مشاهده کنید). همچنین می‌توانیم از آماره χ^2 تصحیح شده پیوسته استفاده کنیم:

$$x^2 = \sum \frac{(O-E)^2 - |O-E| + \frac{1}{4}}{E} \qquad (۷-۲)$$

که به‌دلیل سهولت استفاده و افزایش صحت آزمون به‌کار گرفته می‌شود.

***مثال ۲-۷*. (جمعیت دو رگه F_1).** مثالی از یین و همکاران[1] (۲۰۰۱) را مورد استفاده قرار می‌دهیم که نقشه پیوستگی ژنتیکی را با استفاده از نشانگرهای مولکولی در جمعیت دورگ‌های F_1 بین دو گونه مختلف درخت تبریزی، یعنی درخت تبریزی لرزان و درخت تبریزی سفید که دارای خصوصیات بینابینی بین دو گونه بودند، تشکیل دادند. جمعیتی که برای تهیه نقشه مورد استفاده قرار گرفته بود، شامل ۱۰۳ درخت دو رگه بود که هر یک با استفاده از نشانگرهای RAPD تعیین ژنوتیپ شدند. برخی از نشانگرها در یک والد هتروزیگوت و در والد دیگر صفر یا پوچ بودند، حال آنکه نشانگرهای دیگر الگوی معکوسی داشتند. زمانی که این نشانگرها همانند نشانگرهای تلاقی برگشتی دو طرفه تفرق یابند، این جمعیت F_1، یک جمعیت آزمون تلاقی برگشتی دوطرفه کاذب[2] نامیده می شود (گراتاپاگلیا و سدروف[3]، ۱۹۹۴). در جمعیت آزمون تلاقی برگشتی دوطرفه کاذب، می‌توان دو نقشه اختصاصی برای والدین ایجاد نمود. در این مثال، ۶ نشانگر انتخاب شدند که گروه پیوستگی ۱۶ را در نقشه پیوستگی ژنتیکی درخت تبریزی تشکیل دادند (یین و همکاران، ۲۰۰۱).

جدول ۲-۱- آماره آزمون پیرسون برای آزمون تفرق مندلی ۱:۱ با استفاده از نشانگرهای RAPD در یک جمعیت دو رگه درخت تبریزی با خصوصیات بینابینی دو گونه والد.

			χ^2 پیرسون	
نشانگر	n_1	n_2	χ^2	p-value
I18-1090	۴۴	۵۹	۲/۱۸۴	۰/۱۳۹
W2-1050	۴۵	۵۸	۱/۶۴۱	۰/۲۰۰
AK12-2700	۵۱	۵۲	۰/۰۱۰	۰/۹۲۰
N18-1605	۴۹	۵۴	۰/۲۴۳	۰/۶۲۲
Ak17-1200	۴۰	۶۳	۵/۱۳۵	۰/۰۲۳
I13-1080	۴۱	۶۲	۴/۲۸۲	۰/۰۳۸

[1] - Yin *et al.*
[2] - Two-way pseudo-test backcross
[3] - Grattapaglia and Sederoff

برای آزمون اینکه این ۶ نشانگر آمیزش آزمون از تفرق مندلی ۱:۱ در جمعیت نقشه‌یابی تبعیت می‌کنند، از سه روش مختلف استفاده شد. بر اساس آزمون مربع کای پیرسون (۲-۶)، با فرض اندازه نمونه زیاد، نشانگرهای I18-1090، W2-1050، AK12-2700 و N18-1605 از نسبت ۱:۱ تبعیت می‌کنند ($p>0.05$)، ولی نشانگرهای Ak17-1200 و I13-1080 ($p<0.05$) از این نسبت تبعیت نمی‌کنند (جدول ۲-۱). این نتایج پیشنهاد می کنند که دو نشانگر اخیر از تفرق مورد انتظار مندلی (۱:۱) انحراف نشان می‌دهند.

در این بخش، اندازه نمونه بزرگ بوده و آمار تصحیح شده پیوسته (۲-۷) پاسخ‌های کاملاً مشابه ارائه می‌دهد. همچنین می‌توان پی برد که وقتی چند آزمون انجام می‌دهیم، خطر رد شدن فرض صفر فقط به دلیل شانس وجود دارد؛ بنابراین، اگرچه دو نشانگر معنی‌دار هستند، بهتر است این نتیجه را فقط به‌عنوان مدرکی دال بر معنی‌دار بودن در نظر بگیریم و باید تحقیق و پژوهش بیشتری در این زمینه صورت گیرد.

۲-۳-۲- آزمون‌های نسبت درستنمایی

آزمون فرض می‌تواند با کارآیی و سودمندی بیشتری با استفاده از روش درستنمایی انجام شود. در ابتدا یک وضعیت کلی را در نظر می‌گیریم که در آن، مشاهدات iid به‌صورت $Y=(Y_1,\ldots,Y_n)$ از جمعیتی با تابع چگالی $f(y\mid\theta)$ استخراج شده‌اند که در آن، θ بردار پارامترها است. برای مثال، در شرایطی که اعداد نرمال را مدل‌سازی می‌کنیم $\theta=(\mu,\sigma^2)$، یا در شرایطی که احتمالات را مدل‌سازی می‌کنیم $\theta=(p_1,\ldots,p_k)$ و $\sum_j p_j=1$ است.

با فرض در اختیار داشتن نمونه‌ای به‌صورت $y=(y_1,\ldots,y_n)$، تابع درستنمایی برای این داده‌ها به‌صورت $L(\theta\mid y)=\prod_{i=1}^{n} f(y_i\mid\theta)$ است. برای آزمون فرض زیر:

$H_0:\theta=\theta_0$ در برابر $H_1:\theta\neq\theta_0$

نسبت حداکثر توابع درستنمایی را با هر دو فرض ارزیابی می‌کنیم:

$$\lambda(y)=\frac{\max_{\theta:\theta=\theta_0} L(\theta\mid y)}{\max_{\theta} L(\theta\mid y)}$$

مقدار عددی λ کمتر از ۱ است، چون مخرج کسر مقدار بیشینه را برای مجموعه‌های بزرگ‌تر محاسبه کرده و از این‌رو باید عدد بزرگ‌تری باشد. بعلاوه، مقادیر λ نزدیک ۱ از فرض H_0 حمایت می‌کنند، چون تابع درستنمایی محدود شده (تحت فرض صفر) به تابع درستنمایی محدود نشده (تحت فرض مقابل یا فرض یک) نزدیک است که در عوض باید به حقیقت نزدیک باشد. از سوی دیگر، مقادیر کم λ منجر به رد H_0 می‌شوند.

برای انجام یک آزمون حقیقی، می‌توانیم p-value را به صورت زیر محاسبه کنیم:

$$P(\lambda(Y) < \lambda(y) \mid H_0 \quad is \quad true) = p(y) \qquad (۸-۲)$$

و رد فرض H_0 برای مقادیر کم $p(y)$، نشان می‌دهد که $p(y) < 0.01$. به‌طورکلی، تشخیص توزیع حقیقی احتمال معادله (۲-۸) امر آسانی نیست، ولی نتیجه تقریبی که به‌دست می‌آید، اغلب مفید خواهد بود (ضمیمه A-1 را مشاهده کنید).

معمولاً λ به $-2\log\lambda$ تبدیل می‌شود و در این حالت فرض صفر وقتی رد می‌شود که $-2\log\lambda$ بزرگ باشد که یک «سناریوی معمول» است؛ اما نتیجه ثانویه و مهم‌تری وجود دارد که به‌صورت زیر توصیف می‌شود (برای کسب جزئیات فنی در خصوص این قضیه، ضمیمه A-1 را مشاهده کنید).

اگر $Y_1,\ldots,Y_n$ یک نمونه تصادفی با چگالی $f(x\mid\theta)$ باشد، در سطح تقریبی α آزمون فرض:

$$H_0:\theta\in\theta_0 \quad \text{در مقابل} \quad H_1:\theta\notin\theta_0$$

با در نظر گرفتن شرایط زیر سبب رد فرض H_0 می‌شود:

$$-2\log\lambda(x) > x^2_{\nu,\alpha}$$

که $x^2_{\nu,\alpha}$ مقدار کای مربع جدول در سطح احتمال α و درجه آزادی ν است. ν برابر با تفاوت بین تعداد پارامترهای اختصاص داده شده به H_0 و تعداد پارامترهای موجود در H_1 است؛ بنابراین، برای مثال اگر f دارای توزیع نرمال به صورت $N(\mu,\sigma^2)$ باشد، آزمون فرض $H_0:\mu=\mu_0,\sigma^2=\sigma_0^2$ در مقابل $H_0:\mu\neq\mu_0,\sigma^2\neq\sigma_0^2$ دارای دو درجه آزادی است، در حالی‌که در آزمون $H_0:\mu=\mu_0$ در مقابل $H_0:\mu\neq\mu_0$، σ^2 برای هر دو فرض، همان پارامتر آزاد است.

مثال ۲-۸. **(مثال اولین آزمون).** همانند اولین مثال ساده‌ای که ذکر شد، به مثال ۲-۲ بازگردید که در آن، iid های $Y_1,Y_2,\ldots,Y_n$ حاصل از جمعیت $N(\mu,\sigma^2)$ را در اختیار داشتیم. برای آزمون $H_0:\mu=\mu_0$ در مقابل $H_0:\mu\neq\mu_0$ می‌توانیم عبارت زیر را محاسبه کنیم:

$$\lambda(y) = \frac{\max_{\mu=\mu_0} L(\mu,\sigma^2\mid y)}{\max_{\mu,\sigma^2} L(\mu,\sigma^2\mid y)}$$

$$= \frac{\max_{\mu=\mu_0} L(\mu,\sigma^2\mid y)}{L(\hat{\mu},\hat{\sigma}^2\mid y)}$$

در این فرمول مشاهده می‌کنید که حداکثر مقدار مخرج کسر برای مقادیر پارامترها با جایگزینی MLEها برآوردهای حداکثر درستنمایی آن‌ها) به‌دست می‌آیند (مثال ۲-۳ را ببینید). جهت به حداکثر رساندن صورت کسر، $\mu = \mu_0$ قرار داده و σ^2 را به حداکثر می‌رسانیم و در نتیجه عبارت به‌دست می‌آید:

$$\max_{\mu=\mu_0} L(\mu, \sigma^2 \mid y) = L(\mu_0, \hat{\sigma}_0^2 \mid y)$$

که $\hat{\sigma}_0^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(y_i - \mu_0)^2$ است. به این ترتیب، آماره LR (آماره نسبت درستنمایی) برابر است با:

$$\lambda = \frac{L(\mu_0, \hat{\sigma}_0^2 \mid y)}{L(\hat{\mu}, \hat{\sigma}_0^2 \mid y)}$$

$$= \frac{\left(\frac{1}{2\pi\hat{\sigma}_0^2}\right)^{\frac{n}{2}} \exp\left\{\frac{1}{2\hat{\sigma}_0^2}\sum_{i=1}^{n}(y_i - \mu_0)^2\right\}}{\left(\frac{1}{2\pi\hat{\sigma}^2}\right)^{\frac{n}{2}} \exp\left\{\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \mu\ )^2\right\}}$$

$$= \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)$$

که در حقیقت معادل با آزمون t معمولی است.

در گام بعدی، به آزمون نسبت درستنمایی و مدل دیگری بر پایه توزیع چند جمله‌ای می‌پردازیم. با استفاده از این توزیع و روش مربوطه می‌توان یک آزمایش کلاسیک را تجزیه نمود.

مثال ۲-۹. **(داده‌های مورگان[1] (۱۹۰۹))**. در سال ۱۹۰۹، مورگان آزمایشی روی مگس سرکه انجام داد و دو لاین (رگه) اینبرد را با صفات ژنوتیپی زیر با یکدیگر آمیزش داد:

a: ارغوانی	A: قرمز	رنگ چشم
b: کوتاه وستیجیال[2]	B: نرمال	طول بال

[1] - Morgan
[2] - Vestigial

وی سپس ۲۸۳۹ تلاقی بـه صـورت AABB× aabb ایجـاد نمـود و چهـار ژنوتیـپ AaBb، Aabb، aaBb و aabb را مشاهده نمود (توجـه نماییـد کـه دو ژنوتیـپ میـانی از نـوع نوترکیـب هستند).

می‌توان مدلی بر اساس توزیع چند جمله‌ای برای آزمایش مورگان در نظر گرفت، یعنـی توزیـع گسسته‌ای که برای فراوانی‌های مدل مورد استفاده قرار می‌گیرد. فرض کنید که متغیر تصـادفی Y می‌تواند هر یک از مقادیر k یعنی اعداد صحیح ۱، ۲، ...، k، هر یک بـا احتمـال $p_1, p_2, ..., p_k$ را به خود اختصاص دهد. به‌طور دقیق‌تر می‌توان نوشت:

$$P(Y = j) = p_j,\ j = 1, ..., k$$

توجه شود که اگر k برابر با ۲ باشد، دارای توزیع دو جمله‌ای هستیم.

حال اگر یک نمونه تصادفی iid بـه صـورت $Y_1, ..., Y_n$ داشـته باشـیم و $p = (p_1, p_2, ..., p_k)$ باشد که $\sum_j p_j = 1$، سپس چگالی نمونه (تابع درستنمایی) عبارت است از:

$$L(p \mid y) = \prod_{i=1}^{n} f(y_i \mid p) = p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$$

که n_j تعداد $y_1, ..., y_n$ است که برابر با j است.

مثال ۲-۱۰. **(ادامه داده‌های مورگان (۱۹۰۹))**. برای آزمـایش مورگـان، چهـار دسـته (k=4) بـه صورت چهار ژنوتیپ AaBb، Aabb، aaBb و aabb داریم. تعداد n= 2839 مشاهده داریم کـه به صورت زیر ثبت شده‌اند:

AaBb	Aabb	aaBb	aabb
۱۳۳۹	۱۵۱	۱۵۴	۱۱۹۵

به‌طوری‌که $n_1 = 1339$ و غیره.

الگوهای معمول برای فرض‌های صفر در احتمالات سلولی (p_j)، به‌صورت احتمالات مساوی برای تمامی سلول‌ها نوشته می‌شود؛ به عبارت دیگر در این موارد، فـرض صـفر عبـارت اسـت از یکسـان بودن فراوانی تمامی خانه‌های جدول، یعنی:

$H_0 : p_1 = p_2 = ... = p_k$ در مقابل H_1: فرض صفر صحیح نیست.

البته، تحت فرض H_0، تمامی $p_j'^s$ ها (یعنی احتمال مورد انتظار هـر خانـه جـدول) برابـر $\frac{1}{k}$ است. به‌عنوان مثالی که دارای جذابیت بیشتری است و مثالی که برای آزمـایش مورگـان نیـز قابـل استفاده است، فرض کنید k=4 بوده و می‌خواهیم با استفاده از آزمون نسـبت درسـتنمایی، آزمـون فرض‌های زیر را انجام دهیم:

(۲-۹) $H_0 : p_1 = p_4, p_2 = p_3$ در مقابل H_1: فرض صفر فوق صحیح نیست.

همانند حالت قبل محاسبه را ادامه می‌دهیم:

$$\lambda(y) = \frac{\max_{p_1=p_4, p_2=p_3} L(p \mid y)}{\max_p L(p \mid y)} = \frac{\max_{p_1=p_4, p_2=p_3} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}}{\max_p p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}} \qquad (۲-۱۰)$$

برای به حداکثر رساندن صورت کسر، به یاد آورید که $\sum_j p_j = 1$ اشاره ضمنی دارد که $p_k = 1 - \sum_{j=1}^{k-1} p_j$ (بنابراین درحقیقت فقط k-1 پارامتر در مسئله کلی وجود دارد). اگر $p_1 = p_4 = \frac{p}{2}$، سپس $p_2 = p_3 = \frac{(1-p)}{2}$ که در آن p تنها پارامتر ناشناخته است. صورت کسر معادله (۲-۱۰) نیز به‌صورت زیر بیان می‌شود:

$$\max_p \left(\frac{p}{2}\right)^{n_1} \left(\frac{1-p}{2}\right)^{n_2} \left(\frac{1-p}{2}\right)^{n_3} \left(\frac{p}{2}\right)^{n_4} = \max_p \left(\frac{p}{2}\right)^{n_1+n_4} \left(\frac{1-p}{2}\right)^{n_2+n_3}$$

این حالت، درستنمایی دو جمله‌ای است و گرفتن لگاریتم و سپس مشتق از آن نشان خواهد داد که برآوردهای حداکثر درستنمایی مربوط به p تحت فرض H_0 عبارت است از: $\hat{p}_0 = \frac{(n_1+n_4)}{(n_1+n_2+n_3+n_4)} = \frac{(n_1+n_4)}{n}$. برای مخرج کسر، می‌توان نوشت:

$p_4 = 1 - p_1 - p_2 - p_3$ و با گرفتن لگاریتم از آن می‌توان عبارت زیر را به‌دست آورد:

$$L(p \mid y) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 \log(1 - p_1 - p_2 - p_3)$$

و مشتق‌گیری نسبت به p_1، p_2 و p_3 نشان می‌دهد که $\hat{p}_j = \frac{n_j}{n}, j = 1, \ldots, 4$. به این ترتیب، آمار آزمون درستنمایی به صورت زیر به‌دست می‌آید:

$$\lambda(y) = \frac{\hat{p}_0^{n_1+n_4} (1-\hat{p}_0)^{n_2+n_3}}{\hat{p}_1^{n_1} \hat{p}_2^{n_2} \hat{p}_3^{n_3} \hat{p}_4^{n_4}} = \frac{\left(\frac{n_1+n_4}{2}\right)^{n_1+n_4} \left(\frac{n_2+n_3}{2}\right)^{n_2+n_3}}{n_1^{n_1} n_2^{n_2} n_3^{n_3} n_4^{n_4}}$$

برای انجام آزمون فرض، می‌توانیم از این تقریب استفاده کنیم که $-2\log\lambda(y)$ دارای توزیع کای مربع است. ولی نخست باید درجه آزادی صحیح را به‌دست آوریم.

تحت فرض H_1، هیچ گونه محدودیتی برای پارامترها وجود ندارد، مگر آنکه جمع آن‌ها باید برابر با یک باشد، بنابراین سه پارامتر آزاد وجود دارد. تحت فرض H_0، دو محدودیت اضافی در نظر

می‌گیریم، بنابراین تحت فرض H_0 یک پارامتر آزاد وجود دارد و درجه آزادی کای مربع برابر با تفاوت تعداد پارامترهای آزاد در فرض‌های H_0 و H_1 یعنی ۲=۱-۳ است.

مثال ۲-۱۱. **(داده‌های مورگان (۱۹۰۹)، اولین نتیجه).** چهار ژنوتیپ غیرنوترکیب (AaBb و aabb) و نوترکیب (Aabb و aaBb) وجود دارد و فرض (۲-۹) مشخص می‌کند که ژنوتیپ‌های غیر نوترکیب و نوترکیب با یک پارامتر مشترک و معمول تفرق می‌یابند ولی نیازی نیست که این دو پارامتر مشترک و معمول با هم برابر باشند. فقط فرض جایگزین است که بیان می‌کند این حالت وجود ندارد.

مقدار مشاهده شده $\lambda(y)$ برابر $\lambda(y) = 0.0164$ با $-2\log\lambda(y) = 8.217$ است که با توزیع کای مربع با دو درجه آزادی قابل مقایسه است. با در نظر گرفتن حداکثر خطای نوع اول برابر با ۰/۰۵، $x^2_{2,0.05} = 5.99$ منجر به رد فرض صفر می‌شود.

۲-۳-۳- روش مبتنی بر شبیه‌سازی

تقریب کای مربع که در بخش ۲-۳-۲ مورد استفاده قرار گرفت روی تقریب مجانبی تأکید داشته که اعتبار آن به دارا بودن اندازه‌های سلولی بزرگ بستگی دارد[۱]. بعلاوه، تفاوت در اندازه سلول‌ها نیز می‌تواند روی بسندگی تقریب اثر گذارد.

اگر در خصوص بسندگی تقریب تردید و گمان وجود دارد و یا اگر تفسیر شواهد موجود در داده مشکل باشد، منطقی است که راهکار دیگری را برای ارزیابی شواهد بر علیه فرض H_0 بیازماییم (درحقیقت این حالت در بیشتر موارد ایده خوبی است).

در این بخش، تکنیک شبیه‌سازی را توصیف می‌کنیم که گاهی اوقات تحت‌عنوان خود راه اندازی پارامتری[۲] شناخته می‌شود (افرون و تیبشیرانی[۳]، ۱۹۹۳)؛ این روش، روش مبتنی بر شبیه‌سازی با در نظر گرفتن تمامی اهداف است. نخست این روش به‌طورکلی توصیف شده و سپس برای داده‌های مورگان به‌کار گرفته می‌شود.

---

[۱]- یعنی تعداد افراد در هر یک از حالات مشاهده شده و مورد انتظار بزرگ باشد (مترجمین)

[2] - Parametric bootstrap

[3] - Efron and Tibshirani

ارزیابی فرض مبتنی بر شبیه‌سازی

با فرض دارا بودن یک نمونه iid به صورت $Y=(Y_1,...,Y_n)$ و یک تابع چگالی به‌صورت $f(x|\theta)$ ، می‌توان فرض $H_0:\theta\in\theta_0$ را در مقابل $H_1:\theta\notin\theta_0$ به‌صورت زیر مورد ارزیابی قرار داد:

۱- برآورد θ با استفاده از $MLE\ \hat{\theta}$ و محاسبه آماره درستنمایی مشاهده شده به صورت $-2\log\lambda(y_t^*)$.

۲- ایجاد $t=1,...,M$ نمونه iid جدید به صورت $Y^*=(Y_1^*,...,Y_n^*)$ که در آن $Y^*\sim f(x|\hat{\theta})$ است. سپس، $-2\log\lambda(y_t^*)$ را محاسبه می‌کنیم.

۳- می‌توانیم p-value را برای آزمون به‌صورت زیر محاسبه کنیم:

$$\hat{p}(y)=\frac{1}{M}\sum_{t=1}^{M}I\left(\lambda(y_t^*)>\lambda(y)\right)$$

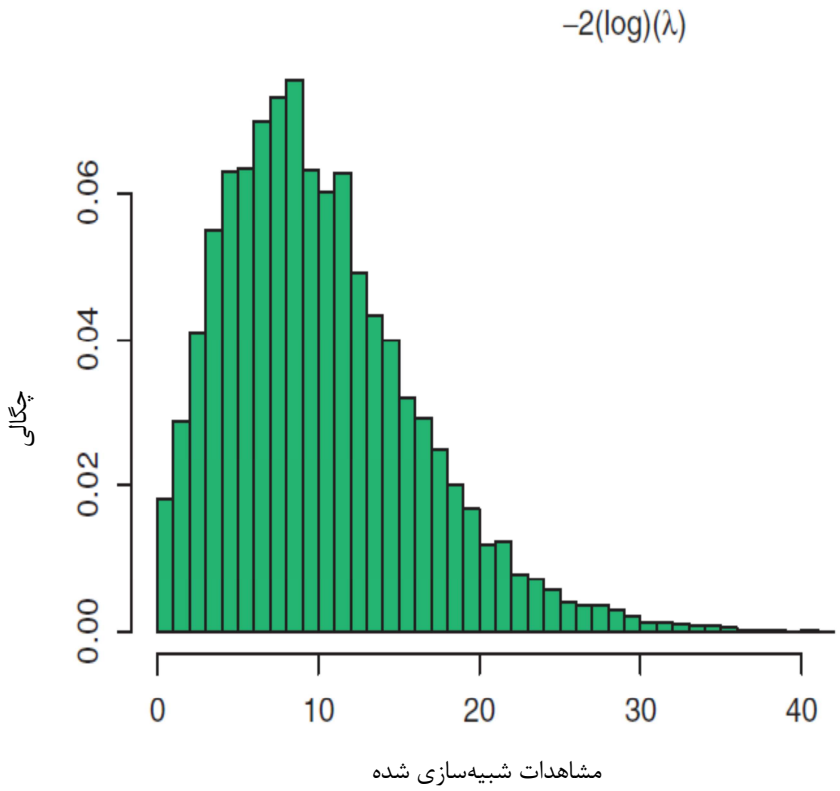
که I(.) تابع موهومی است که در شرایطی که استدلال انجام شده درست باشد، برابر با ۱ و در غیر این صورت برابر با صفر است. هیستوگرام $\lambda(y_t^*)$ نیز می‌تواند رسم شود.

مثال ۲-۱۲. **(داده‌های مورگان (۱۹۰۹)، نتیجه‌گیری دوم)**. برای ارزیابی فرض صفر بر اساس شبیه‌سازی فرض‌های (۲-۹)، متغیرهای تصادفی $Y_1^*,Y_2^*,...$ مربوط به توزیع چند جمله‌ای را با در نظر گرفتن n=2839 و بردار احتمال زیر شبیه سازی می‌کنیم:

$$p=\left(\frac{1339}{2839},\frac{151}{2839},\frac{154}{2839},\frac{1195}{2839}\right)=(0.471,0.054,0.053,0.421)$$

شکل ۲-۲ نتایج شبیه سازی ۱۰۰۰۰ مقدار از $-2\log\lambda(y_t^*)$ را نشان می‌دهد و این نتایج از نتایج حاصل از آزمون کای مربع مثال ۲-۱۱ کاملاً متفاوت است. مقدار مشاهده شده ۸/۲۱۷ = $-2\log\lambda(y_t^*)$ در اینجا درست در نقطه وسط توزیع قرار دارد و مقدار p-value برآورد شده ۰/۵۷۱۸ است و بدین معناست که ۵۷۱۸ عدد از ۱۰۰۰۰ متغیر تصادفی شبیه سازی شده بزرگ‌تر از مقدار مشاهده شده آمار آزمون هستند. این حالت سبب می‌شود تا آمار دقیقاً در میانه توزیع قرار گیرد و بنابراین منجر به پذیرش فرض صفر شود.

باید ذکر شود که نتیجهٔ کلی مشخص و واضح نیست، ولی شواهد موجود مسلماً به این نتیجه اشاره دارند که H_0 یک فرض قابل دفاع بوده و تقریب مجانبی توزیع کای مربع در این مورد، بهترین نیست.



شکل ۲-۲- هیستوگرام ۱۰۰۰۰ مقدار از آماره نسبت درستنمایی $-2\log\lambda$ برای داده‌های مورگان.

۲-۳-۴- برآورد بیزین

در کل این فصل، روش کلاسیکی را برای آمار توصیف کردیم که بر این اساس، شواهد خود را بر پایه ارزیابی خطا در آزمایش‌های تکرار شده قرار دادیم. یک روش جایگزین بنام روش بیزین وجود دارد که اساساً متفاوت از روش کلاسیک است. در این روش، تخمین بر اساس باور قبلی آزمایشگر و اینکه چگونه این باور به وسیله داده‌ها تغییر می‌کند، است.

قرار دادن این دو روش به صورت متضاد و مخالف هم سودمند و مناسب نیست، بلکه بهتر است هر مسئله‌ای را از نزدیک آزمون نموده و روشی که بهترین پاسخ را ارائه می‌دهد، انتخاب کنیم.

در روش کلاسیک، پارامتر θ ناشناخته ولی دارای مقدار ثابتی است. یک نمونه تصادفی به صورت $X_1,\ldots,X_n$ از جمعیتی که با θ مشخص شده است، استخراج می‌شود. بر اساس مقادیر مشاهده شده نمونه، مقدار θ به‌دست می‌آید. در روش بیزین، θ به‌عنوان کمیتی در نظر گرفته

می‌شود که تنوع آن می‌تواند به وسیله توزیع احتمال (که توزیع پیشین نامیده می‌شود) توصیف شود. این توزیع یک توزیع غیر عینی و ذهنی است که بر اساس باور آزمایشگر شکل می‌گیرد و قبل از آنکه داده‌ها مشاهده شوند، فرموله می‌شود. سپس نمونه‌ای از جمعیتی که به وسیله θ مشخص شده گرفته می‌شود و توزیع پیشین با استفاده از این اطلاعات از نمونه، نوسازی و تجدید می‌شود. توزیع پیشین تجدید شده، توزیع پسین نامیده می‌شود.

اگر توزیع پیشین را با $\pi(\theta)$ و توزیع نمونه‌گیری را با $f(x|\theta)$ نشان دهیم، سپس توزیع پسین، یعنی توزیع شرطی θ به شرط نمونه x با استفاده از قانون بیز به صورت زیر محاسبه می‌شود:

$$\pi(\theta|x)=\frac{f\left(x|\theta\right)\pi(\theta)}{\int f\left(x|\theta\right)\pi(\theta)d\theta}=\frac{f\left(x|\theta\right)\pi(\theta)}{m(x)} \qquad (۱۱-۲)$$

که $m(x)=\int f\left(x|\theta\right)\pi(\theta)d\theta$ توزیع حاشیه‌ای **x** است.

زمانی‌که $\pi(\theta|x)$ به‌دست می‌آید، دارای تمامی اطلاعات درباره پارامتر θ است. می‌توانیم این توزیع را رسم و شکل حاصله را مشاهده کنیم و شاید مکانی را که مدها قرار گرفته‌اند و اینکه توزیع متقارن است یا خیر، مشاهده کنیم. همچنین می‌توانیم به‌صورت خاصی میانگین و واریانس پسین را برای به‌دست آوردن برآوردگر نقطه‌ای و مقدار پراکنش و گسترش آن محاسبه کنیم. این اطلاعات با فرمول زیر ارائه می‌شوند:

$$E(\theta|x)=\int\theta\pi(\theta|x)d\theta$$

$$Var(\theta|x)=\int\left[\theta-E(\theta|x)\right]^2\pi(\theta|x)d\theta$$

حال برآورد بیزی را با استفاده از چند مثال شرح می‌دهیم.

مثال ۲-۱۳. **(برآورد بیزی در توزیع نرمال).** فرض می‌کنیم که $X\sim N(\theta,\sigma^2)$ و توزیع پیشین θ به صورت $X\sim N(\mu,\tau^2)$ باشد (در این بخش فرض می‌کنیم که σ^2 ، μ و τ^2 شناخته شده هستند)؛ بنابراین، توزیع پسین θ عبارت است از:

$$\pi(\theta \mid x) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(x-\mu)^2}$$

$$\propto \left[\frac{\sqrt{\sigma^2+\tau^2}}{\sqrt{2\pi\sigma^2\tau^2}} e^{-\frac{\sigma^2+\tau^2}{2\sigma^2\tau^2}\left(\theta-E(\theta|x)\right)^2} \right] \left[\frac{1}{\sqrt{2\pi\left(\sigma^2+\tau^2\right)}} e^{-\frac{1}{2\sigma^2\tau^2}(x-\mu)^2} \right]$$

که توزیع عبارت موجود در اولین کروشه به صورت پسین و دومین توزیع به‌صورت حاشیه‌ای است. محاسبه تا حدودی طول کشیده و کسل کننده بوده و به کامل شدن مربع نما بستگی دارد.

با بررسی، مشاهده می‌شود که توزیع پسین θ نیز نرمال است و میانگین و واریانس آن به‌صورت زیر هستند:

$$E(\theta \mid x) = \frac{\tau^2}{\tau^2+\sigma^2} x + \frac{\sigma^2}{\sigma^2+\tau^2} y$$
$$Var(\theta \mid x) = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2} \qquad (۱۲-۲)$$

برآوردگر بیزی ترکیب خطی از میانگین‌های پیشین و نمونه است. همچنین توجه کنید که چون τ^2، یعنی واریانس پیشین، تمایل دارد به سمت بی‌نهایت میل کند، برآوردگر بیزی به سمت میانگین نمونه میل می‌کند. می‌توانیم این حالت را بدین صورت تفسیر کنیم که زمانی که اطلاعات پیشین مبهم‌تر و نامشخص‌تر باشند، برآوردگر بیزی وزن بیشتری را به اطلاعات نمونه می‌دهد. از سوی دیگر، اگر اطلاعات پیشین خوب و مناسب باشند، به‌طوری که $\sigma^2 > \tau^2$، سپس وزن بیشتری به میانگین پیشین داده می‌شود.

مثال ۲-۱۴. **(برآورد بیزی در توزیع دو جمله‌ای).** فرض می‌کنیم $X_1,\ldots,X_n$ یک برنولی iid (p) باشد؛ بنابراین $Y=\sum X_i$ یک توزیع دو جمله‌ای به‌صورت $binomial(n,p)$ است؛ بنابراین، توزیع پیشین p بدین صورت در نظر گرفته می‌شود:

$$beta(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

و توزیع پسین p عبارت است از:

$$\pi(p \mid y) \propto \left[\binom{n}{y} p^{y}(1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]$$

$$\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

که مجدداً یک توزیع بتا با پارامترهای $y+\alpha$ و $n-y+\beta$ است. همچنین می‌توانیم p را با میانگین توزیع پسین به‌صورت زیر برآورد کنیم:

$$E(p \mid y) = \frac{y+\alpha}{\alpha+\beta+n}$$

در نظر داشته باشید که چگونه برآورد بیزی p شکل گرفته است. میانگین توزیع پیشین $\frac{\alpha}{\alpha+\beta}$ است که بهترین برآورد صورت گرفته از p بدون مشاهده داده است. با نادیده گرفتن اطلاعات پیشین، احتمالاً از $MLE(p) = \frac{y}{n}$ به عنوان برآورد خود از p استفاده می‌کنیم. برآورد بیزی p تمامی این اطلاعات را ترکیب می‌کند. حالتی که در آن، اطلاعات فوق ترکیب می‌شوند، تحت شرایطی که $E(p \mid y)$ را به‌صورت زیر می‌نویسیم واضح و مشخص است:

$$E(p \mid y) = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right)$$

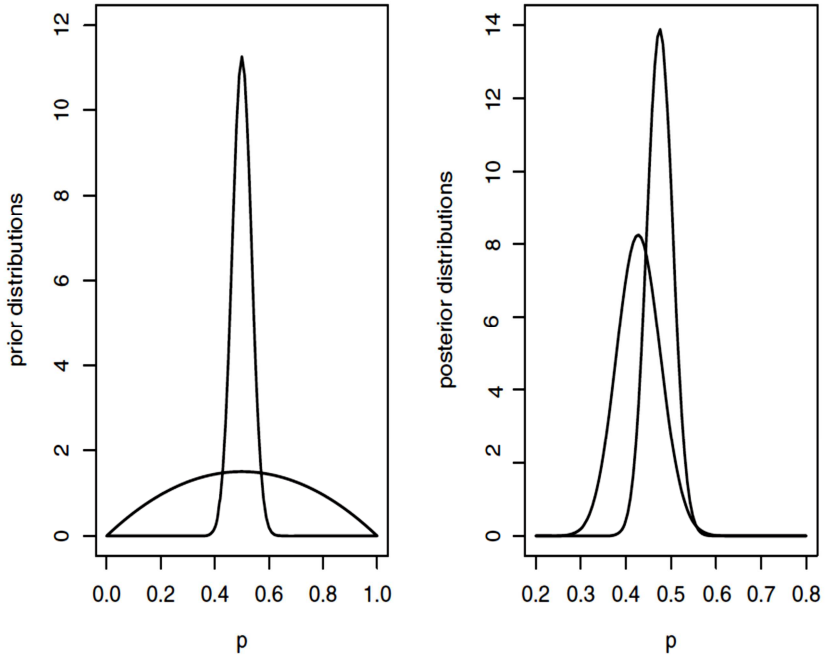
بنابراین، p_B به‌صورت ترکیب خطی از میانگین پیشین و میانگین نمونه با در نظر گرفتن وزن‌هایی است که به‌صورت α، β و n تخصیص می‌یابند.

مثال ۲-۱۵. **(برآورد بیزی در توزیع دو جمله‌ای).** داده‌های مثال ۲-۷ را مجدداً در نظر بگیریم و در اینجا p یعنی سهم واقعی را با استفاده از برآوردگر بیزی برآورد می‌کنیم. نشانگر I18-1090 را با مقادیر ۴۴ و ۵۹ در دو کلاس ژنوتیپی در نظر بگیرید. MLE سهم واقعی اولین ژنوتیپ به‌صورت $\frac{44}{(44+59)} = 0.427$ است.

فرض کنید که p را با استفاده از برآوردگر بیزی برآورد می‌کنیم. اگر آزمایشگر بر این باور باشد که ژن‌ها به‌طور مستقل از هم تفرق می‌یابند، در این صورت $p = \frac{1}{2}$ است. سپس می‌توانیم این باور را در یک توزیع بتا منعکس کنیم که در اطراف $\frac{1}{2}$ متقارن شده است. حال یک توزیع بتا را به

صورت beta (2,2) انتخاب می‌کنیم که وزن زیادتری را به توزیع پیشین نمی‌دهد. تحت این توزیع، برآوردگر بیزی p به صورت $\frac{(44+2)}{(44+59+2+2)}=0.430$ است که به MLE شباهت زیادی دارد؛ بنابراین، برای انتخاب این توزیع پیشین، اطلاعات داده بسیار قوی بوده و برآوردگر بیزی حقیقتاً مشابه MLE است.

اگر آزمایشگر بسیار مصمم باشد که مقدار p نزدیک $\frac{1}{2}$ است، این حالت می‌تواند با افزایش مقادیر پارامتر در توزیع پیشین انعکاس یابد. اگر $\alpha=\beta=100$ باشد، میانگین پیشین به صورت $\frac{1}{2}$ باقی می‌ماند، ولی واریانس پیشین کاهش می‌یابد. برآوردگر بیزی حاصله به صورت $\frac{(44+100)}{(44+59+200)}=0.475$ است و توزیع پسین در حال حاضر کاملاً متقارن است. این حالت در شکل ۲-۳ نشان داده شده است که تقارن توزیع‌های پیشین مشاهده می‌شود، ولی یکی از این توزیع ها تمرکز یافته‌تر است. توزیع‌های پسین حاصله نزدیک تابع درستنمایی هستند و زمانی که $\alpha=\beta=2$ باشد، چولگی ایجاد شده و زمانی که $\alpha=\beta=100$ باشد، به صورت متقارن و دورتر از تابع درستنمایی خواهند بود.



شکل ۲-۳- گراف‌های توزیع‌های پیشین (شکل سمت چپ) و توزیع‌های پسین (شکل سمت راست) مثال ۲-۱۵.

توزیع beta (2,2) پیشین بسیار مسطح و تخت بوده و منجر به ایجاد توزیع پسین دارای چولگی می‌گردد که از تابع درستنمایی قابل تشخیص نیست. توزیع beta (100,100) پیشین دارای نقطه رأسی بسیار بلندی است و منجر به ایجاد توزیع پسین متقارن و دارای نقطه رأسی بلند می‌شود.

۲-۴- تمرین‌ها

۲-۱- الف. همانگونه که در بخش ۲-۳-۱ توصیف شد، تقریب توزیع دوجمله ای به نرمال را شرح دهید. برای مقادیر N برابر با ۵، ۱۵ و ۵۰ و مقادیر p برابر با ۰/۲۵ و ۰/۵، هیستوگرام توزیع دوجمله‌ای و چگالی نرمال را رسم کنید. در هر مورد ناحیه دنباله ۰/۰۵ و ۰/۰۱ را محاسبه کنید.

ب. نشان دهید که اگر $p=\frac{1}{2}$ باشد، معادله (۲-۶) می‌تواند به صورت $\frac{(n_1-n_2)^2}{N}$ ساده شود.

۲-۲- برای وضعیت توصیف شده در مثال ۲-۸:

الف. فرمول مربوطه را برای $\hat{\sigma}_0^2$ مورد بررسی قرار دهید.

ب. نشان دهید که:

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\mu_0)^2}=\frac{1}{1+\frac{(\bar{y}-\mu_0)^2}{\hat{\sigma}^2}}$$

(برای برابری دوم، از عبارت $\sum_{i=1}^{n}(y_i-\mu_0)^2=\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{y}-\mu_0)^2$ استفاده کنید).

ج. بررسی نمایید که در آخرین شکل، تحت شرایطی که $\frac{|\bar{y}-\mu+0|}{\hat{\sigma}}$ مقداری بزرگ باشد، فرض H_0 رد خواهد شد. این آزمون را معادل آزمون t در نظر بگیرید.

۲-۳- الف. MLE های معادله (۲-۱۰) را به‌دست آورید.

فصل سوم

تجزیهٔ پیوستگی و ایجاد نقشه ژنتیکی

۳-۱- مقدمه

پیوستگی، تمایل ژن‌ها برای به ارث رسیدن با یکدیگر است، چون آن‌ها روی یک کروموزوم و در نزدیک یکدیگر قرار گرفته‌اند. تجزیهٔ پیوستگی نشانگرها اساس و پایه ایجاد نقشه پیوستگی ژنتیکی و در گام بعدی، تجزیه مولکولی صفات کمّی با استفاده از نقشه است. تجزیهٔ پیوستگی بر پایه تفرق هم‌زمان نشانگرهای نزدیک به هم و انتقال هم‌زمان آن‌ها به فرزندان نسل بعد است. پیش شرط تجزیه پیوستگی بین دو نشانگر ترتیب آللی شناخته شده آن‌ها (یعنی فازهای پیوستگی) روی کروموزوم‌های همولوگ است، به‌طوری‌که هاپلوتیپ‌های والدینی (یا غیر نوترکیب) در برابر هاپلوتیپ‌های غیر والدینی (یا نوترکیب) بتوانند به آسانی تشخیص داده شوند. در بسیاری از گونه‌های اهلی گیاهان و حیوانات، شجره‌هایی را که دارای فاز مشخص نقشه‌یابی هستند می‌توانند با استفاده از یک جمعیت در حال تفرق نظیر تلاقی برگشتی یا F_2 حاصل از دو رگه خویش‌آمیخته یکسان ایجاد شوند (تعاریف بخش ۱-۵ را به یاد آورید). تئوری‌های تجزیه پیوستگی در چنین شجره‌های در حال تفرق، به خوبی تهیه شده و بسط یافته‌اند.

پیوستگی نشانگرها می‌تواند بر حسب نرخ نوترکیبی یا فاصله ژنتیکی بین آن‌ها اندازه‌گیری شود. کار تجزیه پیوستگی، شناسایی موقعیت نسبی دو یا چند نشانگر روی یک کروموزوم است. تجزیه پیوستگی می‌تواند برای یک جفت نشانگر (تجزیه دو نقطه‌ای) یا سه نشانگر (تجزیه سه نقطه‌ای) به‌طور همزمان انجام شود. تجزیه‌های دو یا سه نقطه‌ای اطلاعات پایه‌ای را برای ایجاد نقشه پیوستگی ژنتیکی که کل ژنوم یا بخشی از آن را پوشش می دهد، فراهم می‌آورند. تابع نقشه که نرخ نوترکیبی را به فاصله ژنتیکی تبدیل می‌کند، می‌تواند با استفاده از تجزیه سه نقطه‌ای به‌دست آید. انواع مختلفی از تابع نقشه وجود دارند که بستگی به فرض وجود یا عدم وجود تداخل کراس‌اوری بین فواصل نشانگرهای مجاور دارند.

در این فصل، اصول اساسی تجزیهٔ پیوستگی را در یک شجره حاصل از دو لاین اینبرد (رگه خویش‌آمیخته) توضیح خواهیم داد. جزئیات رویه برای برآورد و آزمون پیوستگی بین نشانگرهای مختلف به صورت نظری به‌دست خواهد آمد و با مثال‌های واقعی نشن داده خواهد شد. روش پایه برای به‌دست آوردن توابع نقشه را نیز شرح خواهیم داد. تجزیه پیوستگی ژنتیکی انسان در کتاب‌های اوت[1] (۱۹۹۱) و لانژ[2] (۱۹۹۷) توضیح داده شده است. الگوریتم‌ها و نرم‌افزارهای تهیه نقشه در پایان این فصل معرفی می‌شوند.

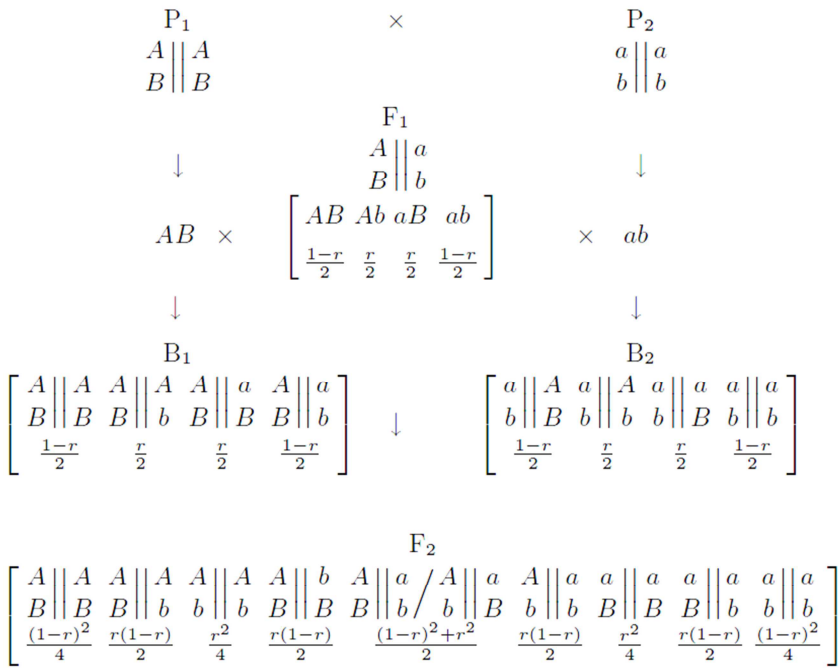
---

[1] - Ott
[2] - Lange

۳-۲- طرح آزمایش

تمایل آلل‌های ژن‌های مختلف روی یک کروموزوم مشابه برای رسیدن به هاپلوتیپ یکسان در طی میوز به درجه پیوستگی بین این ژن‌ها بستگی دارد؛ بنابراین، انجام تجزیهٔ پیوستگی به‌طور حساسی بر شجره‌های در حال تفرقی تکیه دارد که در آن‌ها، انواع گامت‌های نوترکیب و والدینی می‌توانند شمارش شوند. شجره یک جمعیت تلاقی برگشتی یا یک جمعیت F_2 حاصل از دو لاین اینبرد متفاوت، توانمندترین و کارآترین وسیله برای تجزیه پیوستگی است.

در عمل، دو لاین خویش‌آمیخته‌ای (اینبردی) که برای دو آلل جایگزین هر ژن همولوگ هستند، به‌عنوان والدهای P_1 و P_2 برای تولید نتاج F_1 تلاقی داده می‌شوند؛ بنابراین تمامی افراد F_1 برای تمامی ژن‌ها هتروزیگوت هستند. این F_1 های هتروزیگوت یا می‌توانند با هر یک از والدین خود تلاقی برگشتی داده شوند تا دو جمعیت تلاقی برگشتی (B_1 و B_2) ایجاد شوند یا افراد F_1 می‌توانند با یکدیگر تلاقی یابند تا نسل F_2 ایجاد شود. دیاگرامی که بتواند این روش تلاقی را نشان دهد در شکل ۳-۱ ارائه شده است.



شکل ۳-۱- طرح آزمایش مورد استفاده برای تجزیه پیوستگی نشانگرها.

دو نشانگر **A**، با آلل‌های A و a و **B**، با آلل‌های B و b را در نظر بگیرید. دو رگه اینبرد والدینی، P_1 و P_2، به ترتیب برای آلل‌های بزرگ و کوچک این دو ژن هموزیگوت هستند. والد P_1 در طی میوز، گامت یا هاپلوتیپ AB تولید می‌کند در حالی‌که والد P_2 گامت ab ایجاد می‌کند. این دو گامت با یکدیگر ترکیب می‌شوند تا هتروزیگوت F_1 با ژنوتیپ AaBb تولید شود. F_1 چهار نوع گامت مختلف تولید خواهد کرد که دو عدد از آن‌ها (AB و ab) از نوع غیرنوترکیب (والدینی) و دو نوع دیگر (Ab و aB) از نوع نوترکیب هستند. نرخ نوترکیبی بین دو ژن با r نمایش داده می‌شود؛ بنابراین، این دو گروه گامت دارای فراوانی‌های $\frac{(1-r)}{2}$ (غیر نوترکیب) و $\frac{r}{2}$ (نوترکیب) هستند. زمانی که F_1 با یکی از والدین خالص تلاقی برگشتی انجام می‌دهد، چهار ژنوتیپ تلاقی برگشتی تولید خواهد شد که این ژنوتیپ‌ها دارای فراوانی‌هایی برابر با فراوانی گامت‌های F_1 خواهند بود. تلاقی گامت‌های افراد F_1 با یکدیگر سبب ایجاد F_2 می‌شود که در آن، ۱۶ ترکیب گامتی به صورت ۹ ژنوتیپ با فراوانی‌های ترکیبی برای ژنوتیپ‌های مشابه حاصل می‌شوند.

۳-۳- تفرق مندلی

یکی از اولین گام‌ها در پروژه تهیه نقشه‌های ژنومی تعیین این مطلب است که آیا نشانگرهای منفرد از نسبت‌های تفرق مندلی در شجره‌های آزمایشی تبعیت می‌کنند یا خیر. فقط پس از تعیین ماهیت نسبت‌های هر یک از نشانگرها است که می‌توان با استفاده از روش‌های آماری مناسب تجزیه پیوستگی بعدی را انجام داد.

حالت کلی را در نظر می‌گیریم که در آن، آمیزش بین دو لاین اینبرد متفاوت سبب تولید k ژنوتیپ در یک نشانگر در نسبت‌های مورد انتظار $\lambda_1 : \ldots : \lambda_k$ می‌شود. فراوانی نسبی مورد انتظار کلاس ژنوتیپی iام به صورت $\phi_i = \frac{\lambda_i}{\left(\sum_{i=1}^{k} \lambda_i\right)}$ محاسبه می‌شود. اعدادی که درحقیقت در کلاس‌های m ام مشاهده می‌شوند، به ترتیب به‌صورت $n_1, \ldots, n_k$ هستند که $n = n_1 + \ldots + n_k$ هستند. قصد داریم نسبت تفرق مشاهده شده را با مقدار مورد انتظار مقایسه کنیم. برای یک نشانگر هم‌بارز، نسبت مورد انتظار در یک تلاقی برگشتی به صورت ۱:۱ و در F_2 به صورت ۱:۲:۱ است. برای یک نشانگر غالب، در تلاقی برگشتی که با یک مغلوب خالص صورت گرفته، نسبت مورد انتظار به صورت ۱:۱ و در F_2 به صورت ۳:۱ است.

روش‌های پایه برای آزمون الگوهای تفرق نشانگر، شامل آزمون دوجمله‌ای، آزمون نرمال استاندارد، آزمون کای مربع پیرسون و آزمون کای مربع نسبت درستنمایی است. دو آزمون نخست

تحت شرایطی مورد استفاده قرار می‌گیرند که دو کلاس ژنوتیپی در شجره وجود داشته باشد، در حالی‌که دو آزمون اخیر می‌توانند به شرایطی تعمیم یابند که در آن‌ها بیش از دو کلاس ژنوتیپی وجود دارد. در این بخش، شکل کلی‌تر آزمون‌های کای مربع نسبت درستنمایی و کای مربع پیرسون توصیف می‌شوند.

۳-۳-۱- آزمون الگوهای تفرق نشانگر

فرض الگوهای تفرق نشانگر می‌تواند به وسیله آزمون کای مربع پیرسون (۲-۶) یا آزمون نسبت درستنمایی بررسی شوند. در حالت اخیر، تابع درستنمایی، مشروط به اینکه تعداد متفاوتی از افراد در مجموعه‌ای از N نتاج مشاهده شده باشند، از روی توزیع چند جمله‌ای حاصل شده و به‌صورت زیر ارائه می‌شود:

$$L\left(p_1,\ldots,p_k\right)=\frac{n!}{n_1!\ldots n_k!}\prod_{i=1}^{k}p_i^{n_i} \qquad (۱-۳)$$

مقداری از p_i که سبب بیشینه شدن لگاریتم تابع درستنمایی (و بنابراین تابع درستنمایی) می‌شود برابر با $\hat{p}_i=\frac{n_i}{n}$ است، یعنی نسبت حقیقی مشاهده شده در نمونه است. مقادیر $\hat{p}_i$ برآوردهای حداکثر درستنمایی (MLE) p_i هستند. برای آزمون فرض صفر:

$$H_0: p_1=p_{10},\ldots,p_k=p_{k0}$$

وقتی که p_{i0} ها مقادیر مشخصی دارند، از آماره نسبت درستنمایی استفاده می‌کنیم:

$$\begin{aligned}-2\log\lambda &= 2(\ln L_1-\ln L_0)\\ &= 2\left[\sum_{i=1}^{k}n_i\ln\left(\frac{n_i}{n}\right)-\sum_{i=1}^{k}n_i\ln(p_{i0})\right]\end{aligned} \qquad (۲-۳)$$

در این رابطه L_0 تابع درستنمایی مربوط به مقادیر مفروضی است که جایگزین p_i ها شده‌اند و L_1 تابع درستنمایی مربوط به برآوردهای حداکثر درستنمایی است که جایگزین p_i ها شده‌اند.

مقادیر p (p-value) برابر با احتمالی است که یک متغیر تصادفی کای مربع با k-1 درجه آزادی از مقدار $-2\log\lambda$ بیشتر شود.

مثال ۳-۱. **(جمعیت DH).** دو رگه اینبرد، IR64 نیمه پا کوتاه و Azucena پا بلند برای تولید جمعیت F_1 با همدیگر تلاقی داده شدند. با دو برابر کردن کروموزوم‌های هاپلوئید گامت‌های حاصل از هتروزیگوت F_1، یک جمعیت دابل هاپلوئید (DH) شامل ۱۲۳ لاین ایجاد شد (هوآنگ و

همکاران[1]، ۱۹۹۷). چنین جمعیت DH معادل جمعیت تلاقی برگشتی است، چون تفرق نشانگر در آن از نسبت ۱:۱ تبعیت می‌کند. با ۱۲۳ رگه DH، هوآنگ و همکاران در مجموع ۱۷۵ نشانگر چندشکل (شامل ۱۴۶ RFLP، ۸ ایزوزایم، ۱۴ RAPD و ۱۲ ژن کلون شده) را تعیین ژنوتیپ کردند تا نقشه پیوستگی مربوطه را تهیه کنند که به خوبی ۱۲ کروموزوم برنج را پوشش دهد.

اگر n_1 و n_0 تعداد گیاهان برای دو ژنوتیپ مختلف در جمعیت DH باشند، آزمون χ^2 معادله (۲-۶) و آزمون نسبت درستنمایی معادله (۳-۲) را برای آزمون تفرق این نشانگرهای آمیزش آزمون (تست کراس) از نسبت مندلی مورد انتظار ۱:۱ استفاده می‌کنیم. جدول ۳-۱ نتایج ۶ نشانگر موجود روی کروموزوم ۱ را ارائه می‌دهد. نتایج حاصل از آزمون نسبت درستنمایی با نتایج حاصل از آزمون پیرسون مطابقت دارد. بر اساس مقادیر p محاسبه شده از روی توزیع χ^2 با یک درجه آزادی، مشخص شد که نشانگرهای RG472، RG246 و U10 بر اساس نسبت ۱:۱ تفرق می‌یابند و W1 از نسبت ۱:۱ انحراف نشان می‌دهد.

***مثال ۳-۲*. (جمعیت F_2).** چورود و همکاران (۱۹۹۶) ۷۵ نشانگر ریزماهواره را در جمعیتی متشکل از ۵۳۵ نتاج حاصل از تلاقی دو سویه بزرگ (LG/J) و کوچک (SM/J) موش تعیین ژنوتیپ کردند. برای انجام آزمون‌های تفرق، ۹ نشانگر مستقر روی کروموزوم شماره ۲ موش را به عنوان مثال انتخاب کردیم. n_2 ، n_1 و n_0 تعداد موش‌ها برای سه ژنوتیپ در هر نشانگر در جمعیت F_2 هستند. نتایج آزمون χ^2 و آزمون نسبت درستنمایی با یکدیگر هماهنگ بود که پیشنهاد می‌کند ۹ نشانگر از کروموزوم شماره ۲ موش در سطح معنی‌دار ۰/۰۱ بر اساس نسبت مندلی ۱:۲:۱ (جدول ۳-۱) تفرق می‌یابند. توجه داشته باشید که آماره آزمون محاسبه شده در F_2 دارای توزیع χ^2 با دو درجه آزادی است، چون سه ژنوتیپ موجود، سه دسته مستقل را نشان می‌دهند.

---

[1] - Huang *et al.*

جدول ۳-۱- آماره‌های آزمون نسبت درستنمایی و پیرسون برای آزمودن نسبت تفرق مندلی ۱:۱ در جمعیت دابل هاپلوئید برنج و نسبت ۱:۳:۱ در جمعیت F_2 موش.

نشانگر	n_2	n_1	n_0	پیرسون		درستنمایی	
				χ^2	p-value	$-2\log\lambda$	p-value
جمعیت DH							
RG472		۵۲	۵۸	۰/۳۲۷	۰/۵۹۲	۰/۳۲۷	۰/۵۹۲
RG246		۶۲	۵۴	۰/۵۵۲	۰/۴۰۸	۰/۵۵۲	۰/۴۰۷
K5		۶۶	۴۱	۵/۸۴۱	۰/۰۰۹	۵/۸۹۵	۰/۰۰۹
U10		۵۲	۳۸	۲/۱۷۸	۰/۰۹۱	۲/۱۸۷	۰/۰۹۰
RG532		۶۹	۴۴	۵/۵۳۱	۰/۰۱۱	۵/۵۷۷	۰/۰۱۰
W1		۸۳	۳۴	۲۰/۵۲۱	۰/۰۰۰	۲۱/۱۶۸	<۰/۰۰۱
جمعیت F_2							
D_2Mit382	۱۲۱	۲۳۶	۱۲۰	۰/۰۵۷	۰/۴۸۶	۰/۰۵۷	۰/۴۸۶
D_2Mit72	۱۲۱	۲۶۲	۱۱۵	۱/۵۰۲	۰/۲۳۶	۱/۵۱۱	۰/۲۳۵
D_2Mit205	۱۳۲	۲۵۵	۱۲۵	۰/۱۹۹	۰/۴۵۳	۰/۱۹۸	۰/۴۵۳
D_2Mit38	۱۳۹	۲۳۲	۱۴۸	۶/۱۴۱	۰/۰۲۳	۶/۱۲۲	۰/۰۲۳
D_2Mit93	۱۳۳	۲۴۴	۱۱۶	۱/۲۲۳	۰/۲۷۱	۱/۲۱۲	۰/۲۷۳
D_2Mit389	۱۲۸	۲۵۳	۱۳۶	۰/۴۸۲	۰/۳۹۳	۰/۴۷۷	۰/۳۹۴
D_2Mit17	۱۳۳	۲۲۶	۱۴۴	۵/۶۵۲	۰/۰۳۰	۵/۶۱۷	۰/۰۳۰
D_2Mit260	۱۴۳	۲۵۰	۱۲۲	۲/۱۵۰	۰/۱۷۱	۲/۱۰۳	۰/۱۷۵
D_2Mit25	۱۲۱	۲۳۶	۱۵۵	۷/۶۴۱	۰/۰۱۱	۷/۳۲۷	۰/۰۱۳

۳-۴- الگوهای تفرق در یک خانواده تنی

برای نشانگری که در یک خانواده تنی حاصل از دو رگه دگرگشن در حال تفرق است انواع بسیار متفاوتی از تفرق‌ها وجود خواهند داشت. حداکثر چهار آلل نشانگر، علاوه بر آلل خاموش، ممکن است در یک جایگاه ژنی منفرد تفرق یابد؛ بنابراین، تعداد آلل‌ها ممکن است در بین جایگاه‌های ژنی متغیر باشد. فرض می‌کنیم که هر یک از آلل‌های نشانگر که با a، b، c و d نشان داده می‌شوند، نسبت به یکدیگر دارای حالت هم‌بارز بوده، ولی نسبت به آلل خاموش که به‌صورت o نمایش داده می‌شود، غالب باشند. همچنین فرض می‌کنیم که تمامی نشانگرها دستخوش تفرق دقیق مندلی می‌شوند. بسته به اینکه چگونه آلل‌های مختلف در دو والدی که برای تلاقی مورد

استفاده قرار می‌گیرند ترکیب شده‌اند، درمجموع ۱۸ نوع تلاقی ممکنه برای جایگاه ژنی نشانگر وجود خواهد داشت (جدول ۳-۲). بر اساس الگوهای باندی نشانگرهای والدینی و فرزندی، این نوع تلاقی‌ها می‌توانند به ۷ گروه مختلف طبقه‌بندی شوند (همچنین مالیپارد و همکاران[1]، ۱۹۹۷ را مشاهده کنید):

الف. جایگاه‌های ژنی که در هر دو والد هتروزیگوت هستند و بر اساس نسبت ۱:۱:۱:۱ تفرق می‌یابند، یا شامل چهار آلل ab×cd هستند یا دارای سه آلل غیر خاموش ab×ac بوده، یا سه آلل غیر خاموش و یک آلل خاموش ab×co و یا دو آلل خاموش و دو آلل غیر خاموش به صورت ao×bo دارند.

ب. جایگاه‌های ژنی که در هر دو والد هتروزیگوت بوده و بر اساس نسبت ۱:۲:۱ تفرق می‌یابند شامل سه گروه هستند:

ب ۱. یک والد دو آلل غالب متفاوت و والد دیگر یک آلل غالب و یک آلل خاموش دارد، برای مثال ab×ao.

ب ۲. عکس حالت ب ۱.

ب ۳. هر دو والد دارای ژنوتیپ‌های مشابه با دو آلل هم‌بارز هستند: ab×ab

ج. جایگاه‌های ژنی که در هر دو والد هتروزیگوت هستند و بر اساس نسبت ۳:۱ تفرق می‌یابند، ao×ao.

د. جایگاه‌های ژنی که در ترکیب و شکل‌گیری آمیزش آزمون بین والدین وجود داشته و بر اساس نسبت ۱:۱ تفرق می‌یابند که شامل دو گروه هستند:

د ۱. یک والد به‌صورت هتروزیگوت و والد دیگر هموزیگوت است که عبارتند از: سه آللی ab×cc، دو آللی ab×cc، ab×oo و bo×aa و یک آللی (با سه آلل خاموش) ao×oo است.

د. عکس حالت د ۱.

تلاقی نشانگر نوع الف، هر چهار ژنوتیپ نشانگر را در نتاج تولید می‌کند و به‌صورت "کاملاً اطلاعات دهنده[2]" در نظر گرفته می‌شود. انواع دیگر تلاقی‌های نشانگری به‌عنوان "اطلاعات دهنده نسبی[3]" محسوب می‌شوند، چون چهار ژنوتیپ ممکن نتاج به‌واسطه فنوتیپ‌های غیر قابل شناسایی مخلوط شده و قابل تشخیص نیستند. توجه کنید که تلاقی نشانگری "بند د" می‌تواند در شرایطی که فقط والد هتروزیگوت حائز اهمیت باشد، به صورت کاملاً اطلاعات دهنده مشاهده شود. هر یک

---

[1] - Maliepard *et al.*
[2] - Fully informative
[3] - Partially informative

از انواع نشانگری ب ۳ و ج در هر دو والد دارای ژنوتیپ یکسانی هستند و بنابراین انواع تلاقی نشانگری متقارن نامیده می‌شوند. انواع دیگر نشانگرها، ژنوتیپ‌های نشانگری مختص والد هستند و انواع تلاقی نشانگری نامتقارن نامیده می‌شوند. انواع تلاقی نشانگری د ۱ و د ۲ تلاقی‌های برگشتی آزمون کاذب دو طرفه[1] نامیده می‌شوند (گراتاپاگلیا و سدروف[2]، ۱۹۹۴). به‌دلیل آن‌که نشانگرهای غالب به اندازه نشانگرهای هم‌بارز در چنین طرح‌هایی حاوی اطلاعات هستند، تلاقی‌های برگشتی آزمون کاذب به حد زیادی در تهیه نقشه پیوستگی خانواده‌های تنی گونه‌های دور جفت که در آن‌ها ایجاد رگه‌های اینبرد هموزیگوت مشکل یا حتی غیر ممکن است، مورد استفاده قرار می‌گیرند.

برای نشانگرهایی که حاوی اطلاعات نسبی هستند و انواع آمیزش آن‌ها بین دو والد غیر متقارن است، حالات متقابل (برای مثال ب ۱ در مقابل ب ۲ و د ۱ در مقابل د ۲) اطلاعات متفاوتی را برای تعیین خصوصیات فاز پیوستگی وقتی که نشانگرها جفت شده‌اند و بنابراین به صورت دو گروه متمایز حضور دارند، تأمین می‌کند (فصل ۴ را مشاهده کنید).

برای یک پروژه نقشه‌یابی، معمولاً هر دو والد و فرزند تعیین ژنوتیپ می‌شوند. بر اساس الگوی تفرق داده‌های حاصل از باندهای نشانگرها، می‌توانیم نوع تلاقی را که نشانگر معینی به آن تعلق دارد، همانند آنچه که در جدول ۳-۲ ارایه شده است، تعیین کنیم. برای مثال، اگر باندهای a، b، ac و bc را در نتاج و باندهای ab و c را در والدین مشاهده کنیم، می‌توان به آسانی استنباط کرد که این نشانگر به تلاقی نوع ۳ (الف) تعلق دارد. برای انواع تلاقی نشانگری نامتقارن (الف، ب ۱، ب ۲، د ۱ و د ۲)، اطلاعات نشانگری هر دو والد و فرزندانشان برای انجام استنباط در خصوص انواع تلاقی نشانگری مورد نیاز است؛ اما برای انواع تلاقی نشانگری متقارن (ب ۳ و ج)، فقط الگوی تفرق نشانگر در نتاج برای انجام استنباط در مورد آن‌ها مورد نیاز است. با فرض وجود یک نوع تلاقی نشانگری، می‌توان از راهکارهایی که در بخش قبلی معرفی شدند برای تعیین اینکه نشانگر از الگوی تفرق خاصی تبعیت می‌کند یا خیر، استفاده نمود (جدول ۳-۲). به عنوان مثال، برای نوع تلاقی نشانگری الف، چهار کلاس ممکن نشانگری بر اساس نسبت ۱:۱:۱:۱ تفرق می‌یابند، البته به غیر از دو کلاسی (که بر اساس نسبت ۱:۱ تفرق می‌یابند) که در مثال ارائه شده در جدول ۳-۱ اشاره شده‌اند.

---

[1] - Two-way pseudo-test backcrosses
[2] - Grattapaglia and Sederoff

جدول ۳-۲- ترکیبات ممکن حاصل از تلاقی ژنوتیپ نشانگر و الگوهای باندی مشاهده شده برای والدین و فرزندان آن‌ها.

نوع تلاقی			والد			فرزند		
			تلاقی	باند مشاهده شده	تفسیر	باندهای مشاهده شده	تفرق	تعداد فنوتیپ‌ها
A		۱	ab×cd	ab×cd	نامتقارن	ac, ad, bc, bd	۱:۱:۱:۱	۴
		۲	ab×ac	ab×ac	نامتقارن	a, ac, ba, bc	۱:۱:۱:۱	۴
		۳	ab×co	ab×c	نامتقارن	ac,a, bc, b	۱:۱:۱:۱	۴
		۴	ao×bo	a×b	نامتقارن	ab, a, b,o	۱:۱:۱:۱	۴
B	B_1	۵	ab×ao	ab×a	نامتقارن	ab, 2a, b	۱:۲:۱	۳
	B_2	۶	ao×ab	a×ab	نامتقارن	ab, 2a, b	۱:۲:۱	۳
	B_3	۷	ab×ab	ab×ab	متقارن	a, 2ab, b	۱:۲:۱	۳
C		۸	ao×ao	a×a	متقارن	3a, o	۳:۱	۲
D	D_1	۹	ab×cc	ab×c	نامتقارن	ac, bc	۱:۱	۲
		۱۰	ab×aa	ab×a	نامتقارن	a, ab	۱:۱	۲
		۱۱	ab×oo	ab×o	نامتقارن	a, b	۱:۱	۲
		۱۲	bo×aa	b×a	نامتقارن	ab, a	۱:۱	۲
		۱۳	ao×oo	a×o	نامتقارن	a, o	۱:۱	۲
	D_2	۱۴	cc×ab	c×ab	نامتقارن	ac, bc	۱:۱	۲
		۱۵	aa×ab	a×ab	نامتقارن	a, ab	۱:۱	۲
		۱۶	oo×ab	o×ab	نامتقارن	a, b	۱:۱	۲
		۱۷	aa×bo	a×b	نامتقارن	ab, a	۱:۱	۲
		۱۸	oo×ao	o×a	نامتقارن	a, o	۱:۱	۲

۳-۵- تجزیه دو نقطه‌ای

تجزیه دو نقطه‌ای راهکاری آماری برای برآورد و آزمون نرخ نوترکیبی بین دو نشانگر مختلف است. تجزیه دو نفطه‌ای اساسی را برای به‌دست آوردن تابع نقشه و تهیه نقشه‌های پیوستگی ژنتیکی فراهم می‌آورد. در این بخش، روش‌های آماری تجزیه پیوستگی را به صورت جداگانه برای جمعیت‌های F_2 و تلاقی برگشتی نشان خواهیم داد، چون این نوع جمعیت‌ها نیاز به استراتژی‌های تجزیه‌ای مختلف دارند.

۳-۵-۱- تلاقی برگشتی مضاعف

با استفاده از طرح توصیف شده در شکل ۳-۱، با تلاقی F_1ها با هر یک از دو والد، دو تلاقی برگشتی ایجاد می‌کنیم. اجازه دهید تا تلاقی برگشتی با والد P_2 را در نظر بگیریم. فراوانی‌های مورد انتظار و تعداد مشاهده شده چهار ژنوتیپ تولید شده در این تلاقی برگشتی می‌تواند به شکل جدول زیر نشان داده شود:

ژنوتیپ	AB/ab	Ab/ab	Ab/ab	aB/ab
فراوانی	$\frac{1}{2}(1-r)$	$\frac{1}{2}(1-r)$	$\frac{1}{2}r$	$\frac{1}{2}r$
تعداد مشاهده شده	n_1	n_2	n_3	n_4
	غیر نوترکیب		نوترکیب	
فراوانی	1-r		R	
تعداد مشاهده شده	$n_{NR}=n_1+n_2$		$n_R=n_3+n_4$	

زمانی که F_1 دارای فاز پیوستگی شناخته شده باشد (یعنی، آلل A با آلل B جفت شده باشد) تعداد ژنوتیپ‌های تلاقی برگشتی حاصل از گامت‌های غیرنوترکیب یا نوترکیب می‌تواند محاسبه شود که به ترتیب آن‌ها را به‌صورت n_{NR} و n_R نشان می‌دهیم. مجموع n_{NR} و n_R تعداد کل ژنوتیپ‌ها یعنی n است. نرخ نوترکیبی بین دو ژن می‌تواند از روی تعداد مشاهده شده چهار ژنوتیپ متفاوت تلاقی برگشتی برآورد شود. تابع درستنمایی r به شرط داده‌های نشانگری $n=(n_{NR},n_R)$ به‌صورت زیر است:

$$L(r\mid n)=\frac{n!}{n_{NR}!n_R!}(1-r)^{n_{NR}}r^{n_R} \qquad (۳-۳)$$

برآورد حداکثر درستنمایی (MLE) r می‌تواند با مشتق‌گیری از لگاریتم تابع درستنمایی معادله (۳-۳) نسبت به r و برابر صفر قرار دادن مشتق حاصل و حل کردن معادله لگاریتم درستنمایی حاصل شود. با انجام محاسبات، MLE مربوط به r به‌صورت زیر به‌دست می‌آید:

$$\hat{r}=\frac{n_R}{n}=\frac{n_3+n_4}{n_1+n_2+n_3+n_4} \qquad (۴-۳)$$

از روی قضیه A-3، واریانس $\hat{r}$ می‌تواند به وسیله فرمول زیر تقریب زده شود (تمرین ۳-۱):

$$\mathrm{var}(\hat{r})=\frac{1}{\frac{\partial^2}{\partial r^2}\log(r\mid n)\big|_{r=\hat{r}}}=\frac{\hat{r}(1-r)}{n} \qquad (۵-۳)$$

مشاهده می‌شود که دقت $\hat{r}$ به اندازه نمونه مورد استفاده و مقدار r بستگی دارد. دقت برآورد نرخ نوترکیبی نیز به وسیله دو عامل تحت تأثیر قرار می‌گیرد. اولاً، افزایش اندازه نمونه همیشه منجر به افزایش دقت برآورد r می‌شود. ثانیاً، نرخ نوترکیبی برای یک جفت نشانگر با پیوستگی قوی (دارای مقدار r کم) نسبت به آن‌هایی که پیوستگی ضعیفی (دارای مقدار r بالا) دارند، می‌تواند به شکل بهتری برآورد شود.

همیشه توجه داشته باشید که مدل احتمال متغیر تصادفی n_R دارای توزیع دوجمله‌ای با تعداد n آزمایش و احتمال موفقیت r به صورت $n_R \sim binomial-(n,r)$ است؛ بنابراین، $E(n_R)=nr$ و $Var(n_R)=nr(1-r)$ است که با پاسخ‌های درستنمایی مطابقت دارد. فاصله اطمینان تقریبی ۹۵ درصد برای r به صورت زیر تعریف می‌شود:

$$\hat{r}-2\sqrt{\frac{\hat{r}(1-\hat{r})}{n}} \le r \le \hat{r}+2\sqrt{\frac{\hat{r}(1-\hat{r})}{n}} \qquad (۶-۳)$$

MLE مربوط به r نیز می‌تواند برای تعیین درجه پیوستگی بین دو نشانگر استفاده شود. اگر شواهدی مبنی بر این وجود داشته باشد که دو نشانگر به‌طور کامل به هم پیوسته‌اند (یعنی $\hat{r}=0$)، در این صورت F_1 هتروزیگوت مضاعف فقط گامت‌های غیرنوترکیب تولید می‌کند. اگر شواهدی وجود داشته باشد که پیوستگی وجود ندارد (نوترکیبی آزاد)، بنابراین $\hat{r}=0.5$، در این صورت F_1 هاپلوتیپ‌های نوترکیب و غیر نوترکیب را به نسبت مساوی تولید می‌کند. به‌طورکلی، درجه پیوستگی بین دو نشانگر مورد نظر می‌تواند از لحاظ آماری با فرموله کردن دو فرض جایگزین زیر آزمون شود:

$$\begin{cases} H_0 : r=0.5 \\ H_1 : r<0.5 \end{cases} \qquad (۷-۳)$$

که H_0 به مدل کاهش یافته مربوط می‌شود که در آن، داده‌ها با توجه به محدودیت $r=0.5$ برازش شده و H_1 به مدل کاملی مربوط می‌شود که داده‌ها بدون وجود چنین محدودیتی برازش می‌یابند. آماره آزمون برای سنجش این دو فرض، نسبت لگاریتم درستنمایی زیر است:

$$\begin{aligned} LR &= -2\log\left[\frac{L(r=0.5\,|\,n)}{L(\hat{r}\,|\,n)}\right] \\ &= 2\left[n_R\log\left(\frac{n_R}{n}\right)+n_{NR}\log\left(\frac{n_{NR}}{n}\right)-n\log\left(\frac{1}{2}\right)\right] \end{aligned} \qquad (۸-۳)$$

که دارای توزیع مجانبی x^2 با یک درجه آزادی است.

آماره آزمون مشابهی که برای تشخیص پیوستگی مورد استفاده قرار می‌گیرد، مقدار LOD (اوت، ۱۹۹۱) است. مقدار LOD تبدیل آماره نسبت درستنمایی است که به صورت زیر ارائه می‌شود:

$$LOD = \log_{10}\left[\frac{L\left(\hat{r}\,|\,n\right)}{L\left(r=0.5\,|\,n\right)}\right] = 0.217LR$$

مقدار LOD=3 به عنوان نقطه آستانه بحرانی برای اعلام وجود پیوستگی است.

***مثال ۳-۳*. مثال ۳-۱ را مجدداً مشاهده کنید.** ۶ نشانگر ابتدایی را روی کروموزوم شماره ۱ برنج در نظر بگیرید (هوآنگ و همکاران، ۱۹۹۷). برای تشریح مسئله، فقط پیوستگی بین دو نشانگر نزدیک به هم را در نظر می‌گیریم. برای هر جفت نشانگری که در نظر گرفته می‌شود، تمامی ۱۲۳ رگه DH به صورت گروه‌های نوترکیب (n_R) و غیر نوترکیب (n_{NR}) مرتب می‌شوند که در قالب جدول ۳-۳ نشان داده شده‌اند.

با استفاده از معادلات (۳-۴)، (۳-۵) و (۳-۶)، نرخ نوترکیبی برای هر جفت نشانگر، خطای نمونه‌برداری و فواصل اطمینان آن‌ها در تلاقی‌های برگشتی برآورد می‌شود. آزمون‌های نسبت درستنمایی با محاسبه LR و مقدار LOD انجام می‌شوند. همچنین آماره آزمون x^2 پیرسون بر اساس رابطه زیر محاسبه می‌شود:

$$x^2 = \frac{\left(n_R - \frac{n}{2}\right)^2}{\frac{n}{2}} + \frac{\left(n_{NR} - \frac{n}{2}\right)^2}{\frac{n}{2}} = \frac{\left(n_R - n_{NR}\right)^2}{n}$$

که دارای توزیع x^2 با یک درجه آزادی است. اگرچه چهار ژنوتیپ برای دو نشانگر وجود دارد، فقط دو دسته، یعنی نوترکیب و غیر نوترکیب، به صورت مستقل از هم هستند که تأکید می‌نماید درجه آزادی برابر یک است.

سه آزمون، نتایج مشابهی را در ارتباط با نرخ‌های نوترکیبی به‌دست می‌دهند. این جفت نشانگرها به صورت معنی‌داری به یکدیگر پیوسته هستند، چون آزمون فرض نشان می‌دهد که نرخ‌های نوترکیبی به صورت معنی‌داری کمتر از ۰/۵ هستند[۱].

جدول ۳-۳- تجزیه پیوستگی نشانگرها در جمعیت DH برنج.

جفت نشانگر	n_R	n_{NR}	$\hat{r}$	SE $(\hat{r})$	فاصله اطمینان	LR	LOD	x^2	p-value
RG472-RG246	۹۰	۱۹	۰/۱۷	۰/۰۴	$(0.10, 0.25)$	۵۰/۲	۱۰/۹	۴۶/۲	<۰/۰۰۱
RG246-k5	۹۰	۱۵	۰/۱۴	۰/۰۳	$(0.07, 0.21)$	۵۹/۴	۱۲/۹	۵۳/۶	<۰/۰۰۱
K5-U10	۷۶	۴	۰/۰۵	۰/۰۲	$(0.00, 0.10)$	۷۹/۱	۱۷/۲	۶۴/۸	<۰/۰۰۱
U10-RG532	۸۱	۵	۰/۰۶	۰/۰۳	$(0.01, 0.11)$	۸۱/۱	۱۷/۶	۶۷/۲	<۰/۰۰۱
RG532-w1	۹۳	۱۷	۰/۱۵	۰/۰۳	$(0.09, 0.22)$	۵۷/۸	۱۲/۵	۵۲/۵	<۰/۰۰۱

۳-۵-۲- تلاقی‌های مضاعف بین افراد- F_2

هنگامی‌که F_1 های هتروزیگوت با یکدیگر آمیزش می‌کنند یک جمعیت F_2 در حال تفرق تولید خواهد شد که در آن، ۱۶ ترکیب از چهار گامت ماده و چهار گامت نر برای دو نشانگر به صورت ۹ ژنوتیپ قابل تشخیص از یکدیگر حاصل می‌شوند. تعداد مشاهده شده این ۹ ژنوتیپ می‌تواند به صورت منظم و در قالب ماتریس زیر بیان شود:

$$n = \begin{array}{c} \\ BB \\ Bb \\ bb \end{array} \begin{array}{c} \begin{array}{ccc} AA & Aa & aa \end{array} \\ \begin{bmatrix} n_{22} & n_{12} & n_{02} \\ n_{21} & n_{11} & n_{01} \\ n_{20} & n_{10} & n_{00} \end{bmatrix} \end{array} \qquad (۹-۳)$$

[۱]- نرخ نوترکیبی (r) بین دو ژن یا دو نشانگر پیوسته می‌تواند بین صفر و ۰/۵ باشد ($0 \le r < 0.5$). هر قدر نرخ نوترکیبی به سمت صفر باشد، نشان‌دهنده فاصله کم‌تر بین نشانگرها و پیوستگی قوی‌تر بین آنها و بر عکس، هرچه نرخ نوترکیبی به سمت ۰/۵ نزدیک شود، نشان‌دهنده فاصله بیشتر بین نشانگرها و پیوستگی ضعیف‌تر بین آنها است. اگر نرخ نوترکیبی برابر با ۰/۵ یا بزرگ‌تر از ۰/۵ باشد، نشان می‌دهد که دو نشانگر مستقل از هم هستند (مترجمین).

که اولین و دومین اندیس n به ترتیب نشان‌دهنده تعداد آلل‌های بزرگ در نشانگرهای **A** و **B** هستند. هتروزیگوت مضاعف حاصله AaBb ترکیبی از دو دیپلوتیپ، یکی به صورت $[AB][ab]$ که از ترکیب گامت‌های AB و ab مشتق شده و دیگری $[Ab][aB]$ است که از ترکیب گامت‌های Ab و aB حاصل شده است.

به استثنای هتروزیگوت مضاعف، فراوانی‌های ۸ ژنوتیپ دیگر می‌تواند از راه حاصل‌ضرب فراوانی‌های گامتی مربوطه محاسبه شود. توجه داشته باشید اگر دو گامتی که برای تشکیل یک ژنوتیپ با یکدیگر متحد می‌شوند، متفاوت باشند، فراوانی‌های ژنوتیپی باید از راه دو برابر کردن این حاصل‌ضرب‌ها به‌دست آیند. فراوانی هتروزیگوت مضاعف برابر مجموع فراوانی‌های دو دیپلوتیپ است، یعنی $\frac{1}{2}(1-r)^2$ برای دیپلوتیپ حاصل از ترکیب گامت‌های AB و ab و $\frac{1}{2}r^2$ برای دیپلوتیپ حاصل از ترکیب گامت‌های Ab و aB. فراوانی‌های ژنوتیپی برای نشانگرهای A و B می‌تواند به صورت منظم و به شکل زیر بیان شود:

$$F = \begin{array}{c} \\ BB \\ Bb \\ bb \end{array} \begin{array}{c} \begin{array}{ccc} AA & Aa & aa \end{array} \\ \begin{bmatrix} \frac{1}{4}(1-r)^2 & \frac{1}{2}r(1-r) & \frac{1}{4}r^2 \\ \frac{1}{2}r(1-r) & \frac{1}{2}\left[(1-r)^2+r^2\right] & \frac{1}{2}r(1-r) \\ \frac{1}{4}r^2 & \frac{1}{2}r(1-r) & \frac{1}{4}(1-r)^2 \end{bmatrix} \end{array} \qquad (۱۰-۳)$$

تعداد نوترکیب‌هایی که دخیل هستند را می‌توانیم برای هر ژنوتیپ شمارش کنیم. هر یک از ژنوتیپ‌های AABB و aabb، به دلیل اتحاد دو گامت غیر نوترکیب، هیچ گونه نوترکیبی ندارند. ژنوتیپ‌های AABb، AaBB، aaBb و Aabb دارای یک نوترکیب و ژنوتیپ‌های aaBB و AAbb دارای دو نوترکیب هستند. برای ژنوتیپ AaBb، دو دیپلوتیپ متفاوت حاوی تعداد متفاوتی از نوترکیب‌ها هستند که شامل صفر نوترکیب برای دیپلوتیپ حاصل از اتحاد گامت‌های AB و ab و ۲ نوترکیب برای دیپلوتیپ حاصل از اتحاد گامت‌های Ab و aB است. بر اساس فراوانی‌های دیپلوتیپی، تعداد مورد انتظار نوترکیب‌ها برای ژنوتیپ AaBb به‌صورت زیر محاسبه می‌شود:

$$\phi = \frac{0\times\frac{1}{2}(1-r)^2 + 2\times\frac{1}{2}r^2}{\frac{1}{2}(1-r)^2 + \frac{1}{2}r^2} = \frac{2r^2}{(1-r)^2 + r^2} \qquad (۱۱-۳)$$

تعداد مورد انتظار نوترکیب‌ها در ۹ ژنوتیپ نیز می‌تواند به صورت منظم و به شکل زیر بیان شود:

$$R = \begin{array}{c} \\ BB \\ Bb \\ bb \end{array} \begin{array}{c} \begin{array}{ccc} AA & Aa & aa \end{array} \\ \begin{bmatrix} 0 & 1 & 2 \\ 1 & \phi & 1 \\ 2 & 1 & 0 \end{bmatrix} \end{array} \qquad (۱۲-۳)$$

با استفاده از ماتریس‌های **n**و Fحاصل از ماتریس‌های (۳-۹) و (۳-۱۰)، تابع درستنمایی r برای داده‌های نشانگری به صورت زیر ارائه می‌شود:

(۱۳-۳)

$$L(r\,|\,n) = \frac{n!}{n_{22}!\ldots n_{00}!}\left[\frac{1}{4}(1-r)^2\right]^{n_{22}+n_{00}}\left[\frac{1}{2}r(1-r)\right]^{n_{12}+n_{21}+n_{01}+n_{10}}$$
$$\times\left[\frac{1}{4}r^2\right]^{n_{02}+n_{20}}\left[\frac{1}{2}(1-r)^2+\frac{1}{2}r^2\right]^{n_{11}}$$

MLE نرخ نوترکیبی می‌تواند با مشتق گیری از $\log L(r\,|\,n)$ نسبت به r و برابر صفر قرار دادن مشتق و حل کردن تابع درجه سه حاصله (تمرین ۳-۵) به‌دست آید.

روش دیگر برآورد r می‌تواند با استفاده از الگوریتم EM[۱] (لندر و گرین[۲]، ۱۹۸۷) انجام شود. اگر سلول‌های AaBb را به دو دیپلوتیپ به‌صورت $[AB][ab]$ و $[Ab][aB]$ بشکنیم، داده‌های

---

[۱]- الگوریتم EM روشی است برای برآورد حداکثر درستنمایی یک پارامتر که مخفف کلمات Expectation به معنی مورد انتظار و Maximization به معنی حداکثر کردن است. در این روش، ابتدا یک برآورد مقدماتی به عنوان مقدار مورد انتظار برای پارامتر مربوطه در نظر گرفته می‌شود و سپس با حداکثر کردن تابع درستنمایی، مقدار جدیدی برای آن به‌دست می‌آید. سپس، این دو مرحله یکی پس از دیگری چندین بار انجام می‌شوند تا در نهایت به یک همگرایی رسیده و مقدار نهایی با مقدار مورد انتظار تفاوتی نداشته باشد (مترجمین).

[2] - Lander and Green

گمشده دارای توزیع دو جمله‌ای و به‌صورت $z \sim binomial(n_{11}, \frac{\phi}{2})$ معرفی شده و از درستنمایی داده‌های کامل به‌صورت زیر به‌دست می‌آیند:

(۱۴-۳)

$$L(r \mid n, z) = \frac{n!}{n_{22}!\ldots n_{00}!}\left[\frac{1}{4}(1-r)^2\right]^{n_{22}+n_{00}}\left[\frac{1}{2}r(1-r)\right]^{n_{12}+n_{21}+n_{01}+n_{10}}$$
$$\times\left[\frac{1}{4}r^2\right]^{n_{02}+n_{20}}\left[\frac{1}{2}(1-r)^2\right]^{n_{11}-z}\left[\frac{1}{2}r^2\right]^{z}$$

الگوریتم EM نیز به‌صورت زیر پیش می‌رود. در لگاریتم درستنمایی مورد انتظار داده‌های کامل، z را با $\frac{n_{11}\phi}{2}$ جایگزین می‌کنیم. بر اساس تعداد نوترکیب‌های موجود در هر ژنوتیپ (ماتریس **R**) و مشاهدات ژنوتیپ‌های مختلف (ماتریس **n**)، توالی EM را که در MLE مربوط به r به همگرایی می‌رسد، به صورت زیر داریم:

$$\hat{r} = \frac{1}{2n}\left[n_{12} + n_{21} + n_{01} + n_{10} + 2(n_{02} + n_{20}) + \phi n_{11}\right] \qquad (۱۵-۳)$$

معادلات (۳-۱۱) و (۳-۱۵) نشان‌دهنده دو مرحله متوالی در الگوریتم EM هستند. در مرحله E، تعداد مورد انتظار نوترکیب‌ها، ϕ، با استفاده از معادله (۳-۱۱) محاسبه می‌شود. در مرحله M، ϕ برآورد شده برای نوسازی برآورد r با استفاده از معادله (۳-۱۵) بکار گرفته می‌شود. این روش تا جایی تکرار می‌شود که برآوردها در مقدار ثابتی به همگرایی برسند.

واریانس $\hat{r}$ می‌تواند به‌طور مستقیم از روی درستنمایی داده‌های مشاهده شده به صورت زیر تقریب زده شود:

$$\mathrm{var}(\hat{r}) \approx -\frac{1}{\frac{\partial^2}{\partial r^2}\log L(r \mid n)\big|_{r=\hat{r}}} = \frac{\hat{r}(1-\hat{r})(1-2\hat{r}+2\hat{r}^2)}{2(1-3\hat{r}+3\hat{r}^2)n} \qquad (۱۶-۳)$$

همانند تلاقی برگشتی، درجه پیوستگی بین دو نشانگر تحت بررسی می‌تواند با فرموله کردن دو فرض جایگزین یکدیگر آزمون شود. آماره آزمون LR یا مقدار LOD می‌تواند با استفاده از درستنمایی‌ها تحت مدل کامل $L(r \mid n)$ معادله (۳-۱۳) و مدل کاهش یافته $L(r = 0.5)$ محاسبه شده و به صورت زیر بیان شود:

$$L(r=0.5)=\frac{n!}{n_{22}!\ldots n_{00}!}\left[\frac{1}{16}\right]^{n_{00}+n_{02}+n_{20}+n_{22}}$$
$$\times\left[\frac{2}{16}\right]^{n_{12}+n_{21}+n_{01}+n_{10}}\left[\frac{4}{16}\right]^{n_{11}} \qquad (۱۷-۳)$$

مقدار LR دارای توزیع کای مربع با یک درجه آزادی است، چون فقط یک پارامتر r وجود دارد که مدل‌های کامل و کاهش یافته را از هم متمایز می‌سازد.

***مثال ۳-۴*. مثال ۳-۲ را دوباره مشاهده کنید.** می‌توانیم از داده‌های F_2 که به وسیله چورود و همکاران (۱۹۹۶) ایجاد شده بود، استفاده کنیم و پیوستگی بین ۷ نشانگر موجود روی کروموزوم شماره ۲ موش را برآورد کنیم. برای هر جفت نشانگر، تمامی موش‌های F_2 به صورت ۹ ژنوتیپ متفاوت مرتب شدند (ماتریس (۳-۹) را مشاهده کنید). بر اساس ماتریس (۳-۹)، ۹ ژنوتیپ مجدداً به ۴ گروه تقسیم می‌شوند که حاوی دو نوترکیب ($n_{2R}=n_{20}+n_{02}$)، یک نوترکیب ($n_{1R}=n_{21}+n_{12}+n_{10}+n_{01}$)، بدون نوترکیب ($n_{0R}=n_{22}+n_{00}$) و تعداد نامعینی از نوترکیب‌ها (n_{11}) هستند. مشاهدات هر یک از این چهار گروه در جدول ۳-۴ نشان داده شده‌اند.

جدول ۳-۴- تجزیه پیوستگی نشانگرها در جمعیت F_2 موش‌ها.

جفت نشانگر	n_{2R}	n_{1R}	n_{0R}	n_{11}	$\hat{r}$	SE ($\hat{r}$)	LR	LOD	χ^2	p-value
D2mit362-D2mit72	۹	۱۲۰	۱۵۱	۱۶۹	۰/۱۷	۰/۰۱۴	۲۵۲/۲	۵۴/۸	۲۵۳/۴	<۰/۰۰۱
D2mit72-D2mit205	۹	۹۷	۱۷۸	۱۹۶	۰/۱۳	۰/۰۱۲	۳۶۳/۲	۷۸/۹	۳۸۶/۸	<۰/۰۰۱
D2mit205-D2mit38	۰	۶۹	۲۲۶	۲۰۲	۰/۰۷	۰/۰۰۸	۵۹۹/۰	۱۳۰/۱	۶۵۶/۳	<۰/۰۰۱
D2mit38-D2mit93	۱	۲۶	۲۳۹	۲۱۸	۰/۰۳	۰/۰۰۶	۷۸۶/۵	۱۷۰/۸	۸۱۹/۴	<۰/۰۰۱
D2mit93-D2mit389	۳	۸۸	۱۹۵	۱۹۴	۰/۱۰	۰/۰۱۰	۴۵۱/۴	۹۸/۰	۴۷۲/۷	<۰/۰۰۱
D2mit389-D2mit17	۱	۱۰۸	۲۰۶	۱۷۳	۰/۱۲	۰/۰۱۱	۴۳۱/۷	۹۳/۷	۴۸۰/۹	<۰/۰۰۱
D2mit17-D2mit260	۸	۱۰۳	۲۰۰	۱۷۵	۰/۱۳	۰/۰۱۲	۳۸۹/۵	۸۴/۶	۴۵۵/۰	<۰/۰۰۱
D2mit260-D2mit25	۱۲	۱۵۱	۱۷۱	۱۶۰	۰/۲۰	۰/۰۱۵	۲۴۱/۲	۵۲/۴	۲۷۹/۴	<۰/۰۰۱

با استفاده از n_{2R}، n_{1R}، n_{0R} و n_{11} و معادلات (۳-۱۵) و (۳-۱۶)، MLE های مربوط به نرخ‌های نوترکیبی، به همراه خطاهای نمونه‌برداری آن‌ها، با استفاده از الگوریتم EM برآورد شدند. نرخ‌های نوترکیبی برآورد شده بر اساس نسبت درستنمایی (LR یا LOD) و روش آزمون x^2 آزمون شدند. همانند تلاقی برگشتی، فقط دو دسته از گامت‌هایی که این ۹ ژنوتیپ را شکل می‌دهند، نوترکیب و غیرنوترکیب، مستقل هستند، اگر چه ۹ دسته ژنوتیپی متفاوت در F_2 وجود دارند که نشان می‌دهند فقط یک درجه آزادی وجود دارد. نتایج سه روش برای آزمون‌های پیوستگی مطابقت داشت (جدول ۳-۴).

برای نشان دادن روش محاسبه در الگوریتم EM، مراحل تکراری را با استفاده از یک جفت نشانگر در جدول ۳-۴ به تفصیل بیان کردیم. از روی معادلات (۳-۱۱) و (۳-۱۵)، داریم:

تکرار	مرحله E	مرحله M
0	-	$r^{(0)} = 0.1$
1	$\phi^{(1)} = 0.01220$	$r^{(1)} = 0.158265$
2	$\phi^{(2)} = 0.03415$	$r^{(2)} = 0.166527$
3	$\phi^{(3)} = 0.03839$	$r^{(3)} = 0.168123$
4	$\phi^{(3)} = 0.03924$	$r^{(3)} = 0.168445$
5	$\phi^{(4)} = 0.03942$	$r^{(4)} = 0.168511$
6	$\phi^{(5)} = 0.03945$	$r^{(5)} = 0.168524$
7	$\phi^{(6)} = 0.03946$	$r^{(6)} = 0.168527$

با فرض مقدار ابتدایی $r = 0.1$، تعداد مورد انتظار نوترکیب‌ها را برای هتروزیگوت مضاعف، $\phi^{(1)}$، در مرحله اول با استفاده از معادله (۳-۱۱) محاسبه می‌کنیم. $\phi^{(1)}$ محاسبه شده برای محاسبه $r^{(1)}$ همان مرحله با استفاده از معادله (۳-۱۵) مورد استفاده قرار می‌گیرد. چنین تکرارهایی که حاوی مراحل E و M باشند تا جایی تکرار می‌شوند که تفاوت مقادیر برآورد شده r بین دو تکرار متوالی کمتر از مقدار عددی مربوطه، برای مثال 10^{-5} باشد. در این مثال، $r^{(5)}$ در تکرار ۵ به‌صورت ۰/۱۶۸۵۲۴ برآورد شده است که کمتر از $r^{(6)}$ برآورد شده در تکرار ۶ است که مقدار کمتر بودن آن برابر با $0.000003 < 10^5$ است؛ بنابراین، تکرار متوقف می‌شود و برآورد $r = 0.1685$ در تکرار ۶ به‌عنوان MLE مربوط به r استفاده می‌شود.

۳-۶- تجزیه سه نقطه‌ای

برای تجزیه پیوستگی بین دو نشانگر، فقط نیاز به برآورد یک پارامتر یعنی نرخ نوترکیبی است. این امر منعکس‌کننده فراوانی‌های نسبی گامت‌های نوترکیب و غیر نوترکیب مربوط به دو نشانگر است؛ اما زمانی‌که سه نشانگر به‌طور هم‌زمان در یک تجزیه سه نقطه‌ای مد نظر قرار می‌گیرند، به برآورد سه نرخ نوترکیبی برای سه جفت نشانگر ممکنه نیاز خواهیم داشت. در مقایسه با تجزیه دو نقطه‌ای، تجزیه سه نقطه‌ای دو مزیت دارد: ۱- ممکن است دقت برآورد نرخ‌های نوترکیبی را در زمانی که نشانگرها به‌طور کامل حاوی اطلاعات نیستند افزایش دهد (تامپسون، ۱۹۸۴؛ وو و همکاران، b ۲۰۰۰) و ۲- روشی را ارائه می‌دهد که بر اساس آن می‌توان ترتیب بهینه نشانگرهای مختلف را تعیین نمود. در این بخش، راهکار کلی تجزیه سه نقطه‌ای بحث می‌شود.

سه نشانگر به‌صورت **A**، **B** و **C** بدون در نظر گرفتن ترتیب خاص برای F_1 های هتروزیگوت سه گانه در نظر بگیرید که از روی آن، یک تلاقی برگشتی سه گانه یا F_2 تولید می‌شود. اجازه دهید تا در ابتدا یک تلاقی برگشتی به صورت $ABC/abc \times abc/abc$ در نظر بگیریم. مجموع ۸ گروه از ژنوتیپ‌های نشانگری در فرزندان تلاقی برگشتی می‌تواند بر اساس تعداد نوترکیب‌ها بین جفت نشانگرهای **A** و **B** و بین جفت نشانگرهای **B** و **C** به چهار گروه دسته‌بندی شوند. این چهار گروه به صورت ژنوتیپ‌های AbC/abc و aBc/abc (یک نوترکیب از هر جفت)، Abc/abc و aBc/abc (یک نوترکیب فقط از جفت دوم) و ABC/abc (هیچ نوترکیبی برای هر جفت) هستند. فرض کنیم که n_{ij} تعداد ژنوتیپ های حاوی i نوترکیب بین نشانگرهای A و B و j نوترکیب بین نشانگرهای B و C است و g_{ij} مقادیر نرخ نوترکیبی توأم هستند. n_{ij} و g_{ij} می‌توانند به صورت زیر بیان شوند:

جفت A و B	جفت B و C		کل
	نوترکیب	غیر نوترکیب	
نوترکیب	n_{11}	n_{10}	$n_{1.}$
غیر نوترکیب	n_{01}	n_{00}	$n_{0.}$
کل	$n_{.1}$	$n_{.0}$	n
نوترکیب	g_{11}	g_{10}	r_{AB}
غیر نوترکیب	g_{01}	g_{00}	$1-r_{AB}$
کل	r_{BC}	$1-r_{BC}$	

نرخ نوترکیبی بین نشانگرهای A و B، یعنی r_{AB}، فراوانی ژنوتیپ‌های نوترکیب این دو نشانگر را بدون توجه به اینکه بین نشانگرهای B و C نوترکیبی وجود دارد یا خیر، نشان می‌دهد. به‌طور

مشابه، نرخ نوترکیبی بین نشانگرهای B و C، یعنی r_{BC}، فراوانی ژنوتیپ‌های نوترکیب این دو نشانگر را بدون وقوع نوترکیبی بین اولین جفت نشانگرها یا بین دومین جفت آن‌ها نشان می‌دهد. بر اساس این روابط، نرخ‌های نوترکیبی بر حسب سطوح حاشیه‌ای (مقادیر کم یا زیاد) g_{ij} بیان می‌شوند.

تابع درستنمایی نیز برابر است با:

$$L(n \mid g_{11}, g_{10}, g_{01}, g_{00}) = \frac{n!}{n_{11}!n_{10}!n_{01}!n_{00}!} g_{11}^{n_{11}} g_{10}^{n_{10}} g_{01}^{n_{01}} g_{00}^{n_{00}}$$

و باید واضح باشد که MLE های احتمالات نوترکیبی توأم g_{ij} با فرمول $\hat{g}_{ij} = \frac{n_{ij}}{n}$ و MLEهای مربوط به r_{AB}، r_{BC} و r_{AC} می‌توانند به‌طور همزمان به صورت زیر برآورد شوند:

$$\begin{aligned} \hat{r}_{AB} &= \hat{g}_{10} + \hat{g}_{11} = \frac{n_{10} + n_{11}}{n} \\ \hat{r}_{BC} &= \hat{g}_{01} + \hat{g}_{11} = \frac{n_{01} + n_{11}}{n} \\ \hat{r}_{BC} &= \hat{g}_{01} + \hat{g}_{10} = \frac{n_{01} + n_{10}}{n} \end{aligned} \qquad (۱۸-۳)$$

با حل کردن معادلات g_{11}، g_{10}، g_{01} و g_{00}، نتیجه به صورت زیر حاصل می‌شود:

$$\begin{aligned} g_{11} &= \frac{1}{2}(r_{AB} + r_{BC} - r_{AC}) \\ g_{10} &= \frac{1}{2}(r_{AB} - r_{BC} + r_{AC}) \\ g_{01} &= \frac{1}{2}(-r_{AB} + r_{BC} + r_{AC}) \\ g_{00} &= 1 - g_{11} - g_{10} - g_{01} = 1 - \frac{1}{2}(r_{AB} + r_{BC} + r_{AC}) \end{aligned} \qquad (۱۹-۳)$$

و برای اطمینان از وجود احتمالات نوترکیبی توأم $g_{ij} \geq 0$، محدودیت‌های نامعادلات زیر باید برقرار باشند:

$$\begin{aligned} r_{AC} &\leq r_{AB} + r_{BC} \\ r_{BC} &\leq r_{AB} + r_{AC} \\ r_{AB} &\leq r_{BC} + r_{AC} \end{aligned}$$

این بدان معناست که نرخ نوترکیبی بین نشانگرها در دو نقطه انتهایی[1] مساوی یا کمتر از مجموع نرخ‌های نوترکیبی بین هر دو نشانگر نزدیک به یکدیگر است.

قضیه ۳-۵. برای سه نشانگر مرتب شده، نرخ نوترکیبی بین دو نشانگر که هر کدام در یک انتها قرار گرفته‌اند، همیشه بزرگ‌تر یا مساوی نرخ نوترکیبی بین دو نشانگر نزدیک به هم است؛ بنابراین، برای ترتیب معین نشانگری **A-B-C**، $r_{AC} \geq r_{AB}$ و $r_{AC} \geq r_{BC}$.

برهان: ترتیب نشانگری **A-B-C** را در نظر بگیرید. بر اساس تعریف نرخ نوترکیبی کوچک‌تر و یا مساوی ۰/۵، تعداد غیرنوترکیب‌های بین نشانگرهای A و B همیشه بزرگ‌تر یا مساوی تعداد نوترکیب‌ها مشروط بر وقوع همان پیشامد بین نشانگرهای B و C است، یعنی، $g_{01} \geq g_{11}$ و $g_{00} \geq g_{10}$. افزودن g_{10} به هر دو سمت اولین معادله سبب ایجاد رابطه زیر می‌شود:

$$g_{01} + g_{10} > g_{11} + g_{10}$$

که بر اساس معادله (۳-۱۸) منجر به ایجاد رابطه $r_{AC} \geq r_{AB}$ می‌شود. به‌طور مشابه، می‌توانیم ثابت کنیم که $r_{AC} \geq r_{BC}$.

تجزیه سه نقطه‌ای در جمعیت F_2 پیچیده‌تر است. استراتژی که در بخش فوق برای طرح تلاقی برگشتی فرض شد درحقیقت بر اساس تفرق ژنوتیپ‌های گامت‌های F_1 است چون والد برگشتی[2] هیچ مشارکتی در تفرق ژنوتیپ‌های تلاقی برگشتی ندارد؛ اما در نتاج F_2، هر دو والد در تفرق ژنوتیپی نقش دارند و بنابراین تجزیه سه نقطه‌ای در F_2 باید بر اساس تفرق ژنوتیپ زیگوت صورت گیرد. در فصل ۴، از یک مدل آماری کلی که بحث جمعیت F_2 را برای برآورد نرخ‌های نوترکیبی در تجزیه‌های سه نقطه‌ای پوشش می‌دهد، استفاده می‌کنیم.

۳-۷- درستنمایی چند مکانی[3] و مرتب‌سازی مکان‌های ژنی

برای مجموعه معینی از داده‌ها که حاوی چند نشانگر است، ترتیب نشانگرها بر اساس دلایل قبلی که ذکر شدند شناخته شده نیست. ترتیب بهینه نشانگرها که برای تجزیه پیوستگی حائز اهمیت است، می‌تواند با مقایسه درستنمایی‌های چند مکانی برای تمامی انواع ترتیب‌های ممکنه تعیین شود. مجدداً تلاقی برگشتی سه گانه $ABC/abc \times abc/abc$ را بدون وجود اطلاعات در

---

[1] - Endpoint
[2] - Recurrent parent
[3] - Multilocus likelihood

خصوص ترتیب نشانگرها در نظر می‌گیریم. اهمیت ندارد که چگونه ترتیب این نشانگرها مشخص می‌شوند، ولی این تلاقی برگشتی دربردارنده ۸ ژنوتیپ است که بر حسب تعداد نوترکیب‌های بین جفت نشانگرها به چهار گروه دسته‌بندی می‌شوند (بخش ۳-۴). r_{AB}، r_{AC} و r_{BC} به ترتیب نرخ نوترکیبی بین جفت نشانگرهای **A** و **B**، جفت نشانگرهای **A** و **C** و جفت نشانگرهای **B** و **C** هستند. این چهار گروه از ژنوتیپ‌های تلاقی برگشتی به همراه تعداد مشاهده شده و فراوانی‌های مورد انتظارشان برای هر یک از این سه ترتیب نشانگری در جدول ۳-۵ نشان داده شده است. توجه کنید که به‌دست آوردن فراوانی مورد انتظار سه گامت نشانگری بر اساس این فرض است که وقوع نوترکیبی بین فواصل نشانگری مختلف مستقل است. برای مثال، با در نظر گرفتن اولین گروه گامت‌ها، تحت این فرض می‌توان نوشت:

احتمال (عدم وجود نوترکیبی در ABC یا abc)

= احتمال (عدم وجود نوترکیبی در AB یا ab)× احتمال (عدم وجود نوترکیبی در BC یا bc)

$= (1-r_{AB})(1-r_{BC})$

جدول ۳-۵- فراوانی‌های مورد انتظار چهار گروه از گامت‌ها در تلاقی برگشتی با در نظر گرفتن سه ترتیب ژنی ممکن.

نوع گامت	مشاهده	فراوانی مورد انتظار بر اساس ترتیب		
		A-B-C	**A-C-B**	**B-A-C**
ABC یا abc	n_{00}	$(1-r_{AB})(1-r_{BC})$	$(1-r_{AC})(1-r_{BC})$	$(1-r_{AB})(1-r_{AC})$
ABc یا abC	n_{01}	$(1-r_{AB})r_{BC}$	$r_{AC}r_{BC}$	$(1-r_{AB})r_{AC}$
Abc یا aBC	n_{10}	$r_{AB}(1-r_{BC})$	$r_{AC}(1-r_{BC})$	$r_{AB}r_{AC}$
AbC یا aBc	n_{11}	$r_{AB}r_{BC}$	$(1-r_{AC})r_{BC}$	$r_{AB}(1-r_{AC})$

MLE های سه نرخ نوترکیبی برای هر ترتیب را می‌توان با بیشینه نمودن تابع درستنمایی تحت آن ترتیب به‌دست آورد. توجه شود که در طرح تلاقی برگشتی، برآوردهای حداکثر درستنمایی به ترتیب نشانگر بستگی ندارند و به‌صورت زیر بیان می‌شوند:

$$\hat{r}_{AB} = \frac{n_{10} + n_{11}}{n} = \hat{g}_{10} + \hat{g}_{11}$$

$$\hat{r}_{BC} = \frac{n_{01} + n_{11}}{n} = \hat{g}_{01} + \hat{g}_{11}$$

$$\hat{r}_{AC} = \frac{n_{01} + n_{10}}{n} = \hat{g}_{01} + \hat{g}_{10}$$

همان‌گونه که در جدول ۳-۵ مشاهده می‌شود، مقدار درستنمایی محاسبه شده مشاهدات بسته به ترتیب نشانگرها متفاوت خواهد بود. برای مثال، برای ترتیب خاص A-B-C، داریم:

$$\begin{aligned} L_{ABC} &\propto (1-r_{AB})^{n_{00}+n_{10}} (1-r_{BC})^{n_{00}+n_{10}} (r_{AB})^{n_{10}+n_{11}} (r_{BC})^{n_{01}+n_{11}} \\ &= \left[(1-r_{AB})^{n_{00}+n_{10}} (r_{AB})^{n_{10}+n_{11}}\right]\left[(1-r_{BC})^{n_{00}+n_{10}} (r_{BC})^{n_{01}+n_{11}}\right] \end{aligned} \qquad (۲۰-۳)$$

درحقیقت درستنمایی حاصل‌ضرب دو جمله‌ای‌ها بوده و می‌تواند برای MLE های $\hat{r}_{AB}$ و $\hat{r}_{BC}$ حل شود. واریانس‌های نمونه‌گیری برای برآورد نرخ نوترکیبی (r_{AB} و r_{BC}) نیز به صورت زیر هستند:

$$\begin{aligned} Var(\hat{r}_{AB}) &\approx \frac{\hat{r}_{AB}\left(1-\hat{r}_{AB}\right)}{n} \\ Var(\hat{r}_{BC}) &\approx \frac{\hat{r}_{BC}\left(1-\hat{r}_{BC}\right)}{n} \end{aligned} \qquad (۲۱-۳)$$

به‌طور مشابه، مقادیر درستنمایی می‌توانند برای دو ترتیب نشانگری دیگر (جدول ۳-۵ را مشاهده کنید) که به صورت L_{ACB} و L_{CAB} نشان داده می‌شوند، محاسبه گردند (تمرین ۳-۶).

ترتیب نشانگرها که به مقدار بیشینه درستنمایی بستگی دارد، می‌تواند با ترتیب بهینه‌ای که به وسیله داده‌ها تأیید می‌شود، در نظر گرفته شود؛ بنابراین، با مقایسه سه مقدار درستنمایی، L_{ABC}، L_{ACB} و L_{CAB}، می‌توانیم ترتیب نشانگرها را با بیشترین احتمال ممکن تعیین کنیم. روش مشابهی نیز می‌تواند برای تعیین ترتیب نشانگری بهینه در نتاج F_2 مورد استفاده قرار گیرد، ولی بر خلاف تلاقی برگشتی، MLE های مربوط به نرخ‌های نوترکیبی در بین ترتیب‌های حاصله متفاوت خواهند بود. این مبحث در فصل ۴ مورد توجه قرار خواهد گرفت.

محاسبه درستنمایی (۳-۲۰) بر اساس فرض وجود کراس‌اورهای مستقل در فواصل نشانگری متفاوت است. همان‌گونه که به وسیله اسپید و همکاران[1] (۱۹۹۲) نشان داده شده است، نحوهٔ

[1] - Speed *et al.*

مرتب کردن جایگاه‌های ژنی که بتواند درستنمایی را با فرض عدم وجود تـداخل و حتـی در زمـانی که واقعاً تداخل وجود دارد، بیشینه کند حقیقتاً برآورد سازگاری از ترتیب صـحیح جایگـاه‌هـا ارائـه می‌دهد. این خصوصیت استفاده از روش فوق را برای ترتیب‌های ژنی با این شرط کـه انـدازه نمونـه مورد استفاده به حد کافی بزرگ است، بسط می‌دهد.

مثال ۳-۶. **(تجزیه سه نقطه‌ای).** مجدداً مثال ۳-۱ را برای تهیه نقشه جمعیت بـرنج بـا n=۱۲۳ گیاه را در نظر بگیرید. سه نشـانگر، RG472، RG246 و k5 را روی کرومـوزوم شـماره ۱ بـرنج در نظر بگیرید. هتروزیگوت F_1 (AaBbCc) حاصل از ژنوتیپ‌های AABBCC و aabbcc هشت نـوع مختلف گامت هاپلوئید تولید می‌کند. از این‌رو دابل هاپلوئیدها (DH) برای هر نوع گامت بـه‌صـورت زیر مشاهده می‌شوند (n=۱۰۰ پس از حذف مشاهدات گمشده در سه نشانگر):

DH	*ABC / ABC*	*ABc / ABc*	*AbC / AbC*	*Abc / Abc*	*aBC / aBC*	*aBc / aBc*	*abC / abC*	*abc / abc*
تعداد مشاهدات	۳۸	۲	۲	۵	۱۱	۱	۱۰	۳۱

این ۸ نوع گامت بر اساس توزیع نوترکیب‌ها بین نشانگرها به‌صورت چهار گـروه (جـدول ۳-۵ را مشاهده کنید) مرتب می‌شوند. مشاهدات هر گـروه بـه صـورت ۶۹=۳۱+۳۸= n_{00}، ۱۲=۱۰+۲= n_{01}، ۱۶=۱۱+۵= n_{10} و ۳=۱+۲= n_{11} هسـتند. بـا اسـتفاده از معادلـه (۳-۱۸)، MLE هـای سـه نـرخ نوترکیبی بین هر جفت از این سه نشانگر بدین صورت برآورد می‌شوند:

$$\hat{r}_{AB} = \frac{16+3}{100} = 0.19$$

$$\hat{r}_{BC} = \frac{12+3}{100} = 0.15$$

$$\hat{r}_{AC} = \frac{16+12}{100} = 0.28$$

واریانس‌هـای نمونـه‌گیـری $Var(\hat{r}_{AB})$، $Var(\hat{r}_{AC})$ و $Var(\hat{r}_{BC})$ بـه ترتیـب بـه صـورت $\frac{0.19(1-0.19)}{100} = 0.0015$، $\frac{0.15(1-0.15)}{100} = 0.0013$ و $\frac{0.28(1-0.28)}{100} = 0.0020$ تقریب زده می‌شوند.

با گرفتن لگاریتم مقادیر درستنمایی برای هر یک از ترتیب‌های نشانگری و با محاسبه $\hat{r}_{AB}$، $\hat{r}_{AC}$ و $\hat{r}_{BC}$ به وسیله معادله (۳-۲۰) داریم:

$$\log L_{ABC}$$
$$=\left[\left(69+12\right)\log(1-0.19)+(16+3)\log(0.19)\right]$$
$$+\left[(69+16)\log(1-0.15)+(12+3)\log(0.15)\right]=-90.89$$
$$\log L_{ACB}$$
$$=\left[(69+3)\log(1-0.28)+(12+16)\log(0.28)\right]$$
$$+\left[(69+16)\log(1-0.15)+(12+3)\log(0.15)\right]=-101.57$$
$$\log L_{BAC}$$
$$=\left[\left(69+12\right)\log(1-0.19)+(16+3)\log(0.19)\right]$$
$$+\left[(69+3)\log(1-0.28)+(12+16)\log(0.28)\right]=-107.92$$

چون $\log L_{ABC}$ در بین سه مقدار مذکور بزرگ‌ترین مقدار را دارد، نتیجه‌گیری می‌کنیم که این سه نشانگر دارای ترتیب **A-B-C** هستند. اگرچه، ترتیب نشانگری بهینه تعیین شده است، ولی مبحث مهم و حل نشده دیگر آن است که چقدر احتمال دارد که این سه نشانگر نسبت به دو نوع ترتیب دیگر، **A-C-B** و **B-A-C**، دارای ترتیب **A-B-C** باشند. این مبحث در بخش ۳-۹ بیان خواهد شد.

۳-۸- برآورد با تعداد زیادی جایگاه ژنی

در اصل، مسائل مربوط به تعیین ترتیب جایگاه‌های ژنی و برآورد فاصله بین آن‌ها می‌تواند به‌طور هم‌زمان با مقایسه درستنمایی‌هایی که برای تمامی فواصل بین جایگاه‌ها و برای تمامی ترتیب‌های ممکنه جایگاه‌ها بیشینه می‌شوند، صورت گیرند. در هر حال، تعداد ترتیب های ممکنه و همچنین زمان و حافظه رایانه مورد نیاز برای محاسبه درستنمایی چند جایگاهی، به سرعت با افزایش تعداد جایگاه‌ها افزایش می‌یابد. حتی پس از مقایسه دو نوع ترتیب جایگاهی که معکوس همدیگر هستند، هنوز $\frac{m!}{2}$ ترتیب ممکن برای m جایگاه وجود دارد. ارزیابی درستنمایی برای تمامی ترتیب‌های ممکن به‌سرعت باافزایش تعداد جایگاه‌های ژنی عملی نمی‌شود($\frac{2!}{2}=1,\ldots,\frac{10!}{2}=1,814,400$)؛

بنابراین ضروری است که پیش از پرداختن به تجزیه تفصیلی درستنمایی این ترتیب‌های ژنی، تعداد کمی از ترتیب‌های تقریبی را ایجاد نماییم.

دو روش اصلی برای ایجاد ترتیب‌های تقریبی وجود دارند. یک روش در نظر گرفتن تعداد کمی نشانگر است که ترتیب آن‌ها می‌تواند با تجزیه درستنمایی مشخص شود و سپس تعیین مکان نشانگرهای باقیمانده است که در یک زمان، در یکی از فواصل بین نشانگرهای موجود در نقشه قرار می‌گیرند. در هر مرحله، اثری که روی درستنمایی موقعیت مکانی نشانگرهای اضافی در هر فاصله مورد نظر وارد می‌شود، مورد ارزیابی قرار می‌گیرد و آن موقعیتی که بیشترین درستنمایی را ایجاد کند، انتخاب می‌شود. برای مثال، اگر فردی با دو نشانگر کار خود را شروع کند، اولین نشانگر اضافی در یکی از سه فاصله مربوطه و نشانگر بعدی در یکی از چهار فاصله مربوطه و ... قرار می‌گیرند، به‌طوری که کل روش فقط به ارزیابی $3+4+...+m=\frac{(m-1)(m+3)}{2}$ درستنمایی نیاز خواهد داشت.

دومین روش برای ایجاد ترتیب‌های تقریبی، تجزیه تمامی جفت جایگاه‌ها با استفاده از تجزیه پیوستگی دو نقطه‌ای و سپس قرار دادن $\frac{m(m-1)}{2}$ برآورد نرخ نوترکیبی (یا بیشینه مقدار LOD) در معرض برخی از روش‌های مرتب‌سازی است. مسئله تبدیل یک ماتریس مربع عدم تشابهات (یا فاصله) بین افراد در مجموعه‌ای از نقاط مختصات که موقعیت‌های نسبی افراد را مشخص می‌کنند، در تجزیه اکتشافی چند متغیری[1] به خوبی شناخته شده است. مجموعه کلی روش‌هایی که برای حل کردن این مشکل بکار می‌روند، تحت‌عنوان مقیاس‌بندی چند بعدی شناخته می‌شوند. هر یک از روش‌های مقیاس‌بندی چند بعدی می‌توانند برای ماتریس m در m برآوردهای نرخ نوترکیبی که بتواند نمایش دو بعدی از m نقطه را ایجاد نماید، بکار گرفته می‌شود. به استثنای قطعات کروموزومی بسیار کوتاه، m نقطه به جای آنکه روی یک خط راست قرار گیرند، روی یک منحنی نعل اسبی شکل قرار می‌گیرند، چون نرخ نوترکیبی بین دو نشانگر که از یکدیگر بسیار دور هستند، حداکثر $\frac{1}{2}$ است، ولی مجموع نرخ‌های نوترکیبی بین جایگاه‌های ژنی نزدیک به هم که با هر ترتیبی مرتب شده باشند، احتمالاً بیش از $\frac{1}{2}$ می‌شود. فرض می‌شود که این راهکار بسیار متنوع است. زمانی که ترتیب‌های تقریبی ایجاد شدند، این ترتیب‌ها می‌توانند در معرض تجزیه تفصیلی درستنمایی چند نقطه‌ای قرار گیرند.

---

[1] - Multivariate exploratory analysis

۳-۹- درستنمایی‌های مختلط و احتمالات ترتیب

پرسش بدیهی و مسلمی که مطرح می‌شود، مربوط به برآورد احتمالات ترتیب ژن‌ها است. در این بخش، درستنمایی کلی را به‌دست می‌آوریم و سپس با استفاده از آن، می‌توانیم به‌طور توأم نرخ‌های نوترکیبی و احتمالات ترتیب را در تجزیه سه نقطه‌ای برآورد کنیم.

درستنمایی‌ها برای هر یک از این سه نوع ترتیب بدین صورت بیان می‌شوند:

$$L_{ABC} \propto (1-r_{AB})^{n_{00}+n_{10}}(1-r_{BC})^{n_{00}+n_{10}}(r_{AB})^{n_{10}+n_{11}}(r_{BC})^{n_{01}+n_{11}}$$

$$L_{ACB} \propto (1-r_{AC})^{n_{00}+n_{11}}(1-r_{BC})^{n_{00}+n_{10}}(r_{AC})^{n_{01}+n_{10}}(r_{BC})^{n_{01}+n_{11}}$$

$$L_{BAC} \propto (1-r_{AB})^{n_{00}+n_{10}}(1-r_{AC})^{n_{00}+n_{11}}(r_{AB})^{n_{10}+n_{11}}(r_{AC})^{n_{01}+n_{10}}$$

اما هیچ یک از این توابع درستنمایی مربوط به مدل کامل و برای داده‌ها نیست. برای آن مدل، مشاهده y را از روی یکی از سه ترتیب به‌دست می‌آوریم و ترتیب مورد نظر با احتمال معینی ارائه می‌شود. این احتمال، احتمال ناشناخته‌ای است که نشان می‌دهد آن ترتیب صحیح است؛ بنابراین، اگر این احتمالات را با p نشان دهیم، با در نظر گرفتن اندیس مناسب، تابع درستنمایی برای مدل کامل به‌صورت زیر ارایه می‌شود که یک مدل مختلط است:

$$L_F = p_{ABC}L_{ABC} + p_{ACB}L_{ACB} + p_{BAC}L_{BAC} \qquad (۲۲-۳)$$

می‌دانیم که به ترتیب ژنی با احتمال معینی دست خواهیم یافت و اگر از ترتیب مورد نظر آگاهی داشته باشیم، بخش صحیح درستنمایی مورد استفاده را خواهیم شناخت. این امر پیشنهاد می‌کند که می‌توانیم داده‌های گمشده $z=(z_1,z_2,z_3)$ را برای ارائه ترتیب ژنی وارد نماییم. برای مثال، تعیین کرده‌ایم که z می‌تواند دقیقاً یک ۱ و دو صفر باشد که ۱ نشان‌دهنده ترتیب ژنی است. درستنمایی توأم داده‌های کامل (n,z) بدین صورت است:

$$L_C = \left[p_{ABC}L_{ABC}\right]^{z_1}\left[p_{ACB}L_{ACB}\right]^{z_2}\left[p_{BAC}L_{BAC}\right]^{z_3} \qquad (۲۳-۳)$$

و چگالی حاصل از داده‌های گمشده به صورت زیر بیان می‌شود:

$$k(z,n) = \frac{\left[p_{ABC}L_{ABC}\right]^{z_1}\left[p_{ACB}L_{ACB}\right]^{z_2}\left[p_{BAC}L_{BAC}\right]^{z_3}}{p_{ABC}L_{ABC} + p_{ACB}L_{ACB} + p_{BAC}L_{BAC}}$$

که باید آن‌را به‌عنوان توزیع چند جمله‌ای با یک مشاهده شناسایی کنیم تا آن‌را در یکی از سه سلول (خانه جدول) قرار دهیم که احتمالات سلول نیز بدین صورت ارائه می‌شوند:

$$E(z_1) = P_{ABC} = \frac{p_{ABC}L_{ABC}}{p_{ABC}L_{ABC} + p_{ACB}L_{ACB} + p_{BAC}L_{BAC}} \qquad (۲۴-۳)$$

و به‌طور مشابه P_{ACB} و $P_{BAC} = 1 - p_{ABC} - p_{ACB}$ را تعریف می‌کنیم.

با استفاده از معادلات (۳-۲۳) و (۳-۲۴)، یک الگوریتم EM را به‌صورت زیر ایجاد می‌کنیم. لگاریتم درستنمایی مورد انتظار داده‌های کامل عبارت است از:

$$E(L_C) = P_{ABC}\log(p_{ABC}L_{ABC}) + P_{ACB}\log(p_{ACB}L_{ACB}) + P_{BAC}\log(p_{BAC}L_{BAC}) \quad (۳-۲۵)$$

مشتق‌گیری از $E(L_C)$ نسبت به نرخ‌های نوترکیبی سبب ایجاد MLE های مربوط به r می‌شوند که با استفاده از آن‌ها، می‌توانیم P_{ABC} را به کمک رابطه زیر برآورد کنیم:

$$\hat{P}_{ABC} = \frac{p_{ABC}L_{ABC}}{p_{ABC}L_{ABC} + p_{ACB}L_{ACB} + p_{BAC}L_{BAC}}$$

و به‌طور مشابه برای P_{ACB} و $P_{BAC} = 1 - p_{ABC} - p_{ACB}$ (به شباهت موجود با مدل مختلط بخش ۲-۲ توجه نمایید).

مثال ۳-۷. مثال ۳-۶ برای تجزیه سه نقطه‌ای نشانگرهای RG472، RG246 و K5 در جمعیت DH برنج را مجدداً در نظر بگیرید. بر اساس مشاهدات هر نوع گامت، داریم $n_{00} = 69$، $n_{01} = 12$، $n_{10} = 16$ و $n_{11} = 3$ که MLE های مربوط به نرخ نوترکیبی را به صورت زیر به‌دست می‌دهد:

$$\hat{r}_{AB} = 0.19 \qquad \hat{r}_{BC} = 0.15 \qquad \hat{r}_{AC} = 0.28$$

برای برآورد P، مقادیر p باید تخصیص یابند. زمانی که اطلاعاتی در اختیار نداریم، نمی‌توانیم کار زیادی انجام دهیم، ولی $p_{ABC} = p_{ACB} = p_{BAC} = 0.333$ است. از این‌رو برآورد P_{ABC} به صورت زیر خواهد بود:

$$\hat{P}_{ABC} = \frac{\frac{1}{3}L_{ABC}}{\frac{1}{3}L_{ABC} + \frac{1}{3}L_{ACB} + \frac{1}{3}L_{BAC}}$$

که $\hat{P}_{ACB}$ و $\hat{P}_{BAC}$ نزدیک به صفر هستند. این نتایج پیشنهاد می‌دهند که احتمال بسیار زیادی (۰/۹۹۹۹۷۷) وجود دارد تا داده‌ها از ترتیب نشانگری **A-B-C** پیروی کنند.

۳-۱۰- توابع تهیه نقشه

تابع نقشه یک تابع ریاضی است که نرخ نوترکیبی (r) بین دو جایگاه ژنی را به فاصله ژنتیکی (d) تبدیل می‌کند. نرخ نوترکیبی یک مقیاس فاصله افزایشی نیست. سه نشانگر را به صورت A، B و C در نظر بگیرید. اگر نرخ نوترکیبی بین نشانگرهای A و B و بین نشانگرهای B و C به‌طور مفروض برابر $r = 0.3$ باشد، سپس نرخ نوترکیبی بین نشانگرهای A و C نمی‌تواند 2r باشد، چون مقدار آن بیش از ۵۰ درصد خواهد بود[1]؛ بنابراین نیاز داریم تا نرخ نوترکیبی r را به فاصله نقشه افزایشی d تبدیل کنیم.

تعریف ۳-۸. [فاصله نقشه] . فاصله نقشه بین دو جایگاه ژنی به صورت تعداد مورد انتظار کراس‌اورهای رخ داده بین آن‌ها روی یک کروماتید منفرد در طی میوز تعریف می‌شود.

دو غیر آلل که هر یک مربوط به یک جایگاه ژنی هستند، تحت شرایطی که هیچ کراس‌اوری رخ نداده باشد یا حتی تعداد زوجی از کراس‌اورها بین دو جایگاه ژنی رخ داده باشد، می‌توانند از روی کروموزوم‌های والدینی مشابه حاصل شوند و اگر تعداد کراس‌اورهای بین دو جایگاه فرد باشند، از کروموزوم‌های والدینی مختلف به وجود می‌آیند؛ بنابراین، می‌توانیم با استفاده از تعداد پیشامدهای نوترکیب، یک مدل نظری را برای بیان نرخ نوترکیبی بین دو جایگاه ژنی بر حسب فاصله یا طول نقشه آن‌ها فرموله نماییم.

۳-۱۰-۱- فرمول ماتر[2]

ماتر (۱۹۳۸) فرمولی را به‌دست آورد که نرخ نوترکیبی بین دو جایگاه ژنی A و B را به تعداد تصادفی کیاسماهای (یعنی کراس‌اورها) X رخ داده روی فاصله $[A, B]$ مجموعه کروماتیدی مرتبط ساخت. بر اساس این استنتاج، نرخ نوترکیبی بین دو جایگاه ژنی یعنی r، نصف احتمال کیاسمایی است که در تمامی چهار رشته تتراد بین جایگاه‌ها رخ می‌دهد. از لحاظ ریاضی، این حالت می‌تواند به صورت زیر بیان شود:

$$r = \frac{1}{2} prob(X > 0) = \frac{1}{2}\left[1 - prob(X = 0)\right] \qquad (۳-۲۶)$$

---

[1]- نرخ نوترکیبی یا r حداکثر می‌تواند ۵۰ درصد یا ۰/۵ باشد (مترجمین)

[2] - Mather

که $prob(X=0)$ احتمال نبود کیاسما بین دو جایگاه ژنی است. فاصله نقشه ژنتیکی d که A و B را از هم جدا می‌کند، به‌طور کلی به‌صورت $\frac{1}{2}E(X)$ یعنی تعداد مورد انتظار کیاسماها در $[A,B]$ برای تتراد تعریف می‌شود، چون هر کراس‌اور شامل دو کروماتید است.

در ابتدا فرمول ماتر (۳-۲۶) می‌تواند با توجه به این نکته که اگر و فقط اگر تعداد کراس‌اورها در گامت بین جایگاه‌ها فرد باشد، اثبات شود. احتمال اینکه گامت پس از وقوع x کیاسما بین A و B نوترکیب باشد، با r_x نشان داده می‌شود. واضح است که $r_0=0$ است. برای $x>0$، وقوع مجدد کیاسما به‌صورت زیر نشان داده می‌شود:

$$r_x=\frac{1}{2}r_{x-1}+\frac{1}{2}(1-r_{x-1})=\frac{1}{2}$$

یک گامت پس از x کراس‌اور نوترکیب است، به‌شرطی که پس از x-۱ کراس‌اور نوترکیب بوده و در کراس‌اور x مشارکت نداشته باشد یا اگر پس از وقوع x-۱ کراس‌اور غیرنوترکیب بوده و در کراس‌اور x مشارکت داشته است و این احتمالات به‌نسبت باهم برابر هستند؛ بنابراین، برای تمامی مقادیر $x>0$، $r_x=\frac{1}{2}$ خواهد بود. این حالت پیشنهاد می‌کند وقتی که حداقل یک کیاسما بین دو جایگاه ژنی وجود دارد، احتمال به‌دست آوردن یک رشته نوترکیب برابر $\frac{1}{2}$ است. این استدلال فرمول ماتر را ثابت می‌کند.

۳-۱۰-۲- تابع نقشه مورگان[1]

تابع نقشه مورگان ساده‌ترین تابع نقشه‌ای است که فرض می‌کند ۱) حداکثر یک کراس‌اور در فاصله بین دو جایگاه ژنی به وقوع می‌پیوندد و ۲) احتمال وقوع کراس‌اور در یک فاصله متناسب با طول نقشه در آن فاصله است (مورگان، ۱۹۲۸). تحت این فرضیات، احتمال وقوع کیاسما در فاصله واحدهای نقشه d برابر با تعداد مورد انتظار کراس‌اورها به ازای گامت در این فاصله بوده و بنابراین تا 2d است (تعریف فوق در مورد d را مشاهده کنید) که رابطه زیر را ارائه می‌دهد:

$$r=\frac{1}{2}[1-prob(X=0)]=\frac{1}{2}[1-(1-2d)]=d$$

[1] - Morgan

این تابع فقط زمانی برقرار است که $0 \le d \le \frac{1}{2}$ باشد، چون برای $d > \frac{1}{2}$، نرخ‌های نوترکیبی بیش از $\frac{1}{2}$ حاصل خواهد شد؛ بنابراین ممکن است به‌عنوان تقریبی برای فواصل کوتاه استفاده شود، ولی برای قطعات بلند کروموزوم قابلیت بکارگیری ندارد.

۳-۱۰-۳- تابع نقشه هالدین[1]

تابع نقشه هالدین، دومین تابع نقشه ساده، فرض می‌کند که کراس‌اورها به‌صورت تصادفی و مستقل از یکدیگر رخ می‌دهند (هالدین، ۱۹۱۹). با این فرض، وقوع کراس‌اورهای بین دو جایگاه ژنی موجود روی یک کروموزوم می‌تواند از رویه پویسون[2] پیروی کند (یعنی، وقوع کراس‌اورها در هر نقطه‌ای بین جایگاه‌های ژنی دارای احتمال برابر هستند)، به‌طوری‌که تعداد کراس‌اورهای بین جایگاه‌های ژنی می‌تواند با توزیع پویسون مدل‌سازی شود. زمانی که فاصله نقشه d به عنوان متوسط تعداد کراس‌اورها به ازای هر کروماتید در یک فاصله معین تعریف شود، به‌طورکلی متوسط تعداد کراس‌اورها برای تتراد برابر 2d است. فرض وجود رویه پویسون اشاره ضمنی دارد که احتمال عدم وقوع کیاسما در فاصله مربوطه یعنی $\Pr ob(X=0)$ برابر با e^{-2d} است. با استفاده از فرمول ماتر، می‌توان تابع نقشه هالدین را به‌صورت زیر ارائه داد:

$$r = \frac{1}{2}\left[1 - \Pr ob\left(X=0\right)\right] = \frac{1}{2}\left(1 - e^{-2d}\right) \qquad (۲۷-۳)$$

که عکس آن به صورت زیر است:

$$d = -\frac{1}{2}\ln(1-2r) \qquad (۲۸-۳)$$

تابع نقشه هالدین می‌تواند به روش دیگری نیز به‌دست آید. به‌دلیل آنکه فاصله نقشه (d) بین دو جایگاه ژنی به صورت متوسط تعداد کراس‌اورها تعریف می‌شود، احتمال تعداد کراس‌اورها می‌تواند با استفاده از توزیع پویسون بر حسب d به صورت زیر بیان شود:

کراس‌اور	۰	۱	۲	۳	...	x	...
احتمال	e^{-d}	$\frac{d}{1!}e^{-d}$	$\frac{d^2}{2!}e^{-d}$	$\frac{d^3}{3!}e^{-d}$	...	$\frac{d^x}{x!}e^{-d}$	...

[1] - Haldane
[2] - Poisson

به‌دلیل آنکه مقدار نرخ نوترکیبی r برای فاصله ژنتیکی d برابر با مجموع احتمالات تمامی تعداد کراس‌اورهای فرد است، خواهیم داشت:

$$r = e^{-d}\left(\frac{d}{1!}+\frac{d^3}{3!}+\frac{d^5}{5!}+\ldots\right)$$
$$=\frac{1}{2}\left(1-e^{-2d}\right)$$

جمع‌پذیری تابع هالدین می‌تواند با این فرض که سه جایگاه ژنی با ترتیب **A-B-C** قرار دارند، ایجاد شود. یک گامت در ارتباط با A و C نوترکیب است، اگر و فقط اگر در ارتباط با A و B نوترکیب بوده ولی در ارتباط با B و C نوترکیب نباشد، یا اگر در ارتباط با B و C نوترکیب بوده ولی در ارتباط با A و B نوترکیب نباشد؛ بنابراین، با فرض وجود استقلال، سه نرخ نوترکیبی ممکن در بین این سه جایگاه ژنی دارای روابط زیر هستند:

$$r_{AC} = r_{AB}(1-r_{BC}) + r_{BC}(1-r_{AB}) = r_{AB} + r_{BC} - 2r_{AB}r_{BC} \qquad (۳-۲۹)$$

یا:

$$1-2r_{AC} = (1-2r_{AB})(1-2r_{BC})$$

با فرض $r_{AB} = \frac{1}{2}(1-e^{-2d_{AB}})$ و $r_{BC} = \frac{1}{2}(1-e^{-2d_{BC}})$ که d ها فواصل نقشه‌ای بین جایگاه‌های ژنی مربوطه هستند، داریم:

$$r_{AC}$$
$$= r_{AB} + r_{BC} - 2r_{AB}r_{BC}$$
$$=\frac{1}{2}\left(1-e^{-2d_{AB}}\right)+\frac{1}{2}\left(1-e^{-2d_{BC}}\right)-2.\frac{1}{2}\left(1-e^{d_{AB}}\right)\frac{1}{2}\left(1-e^{d_{BC}}\right)$$
$$=\frac{1}{2}\left(1-e^{-2d_{AB}}+1-e^{-2d_{BC}}-1+e^{-2d_{AB}}+e^{-2d_{BC}}-e^{-2d_{AB}}e^{-2d_{BC}}\right)$$
$$=\frac{1}{2}\left[1-e^{-2(d_{AB}+d_{BC})}\right]$$
$$=\frac{1}{2}\left(1-e^{-2d_{AC}}\right)$$

که منجر به $d_{AC} = d_{AB} + d_{BC}$ می‌شود.

در عمل، تابع نقشه هالدین ممکن است در فواصل کوچک دقیق نباشد. مشاهدات تجربی نشان داده‌اند که احتمال داشتن دو کراس‌اور در مجاورت یکدیگر، اغلب کمتر از احتمال پیش‌بینی شده در تابع نقشه هالدین است. در هر حال، هالدین نیز در مقاله ۱۹۱۹ خود روش مشتق‌گیری را که به

ایجاد توابع نقشه مختلف تعمیم داد، معرفی نمود. یکی از کاربردهای این معادله تعمیم یافته، مشتق تابع نقشه کوسامبی[1] (۱۹۴۴) بود که در عمل هم ساده و هم قابل توجیه است.

۳-۱۰-۴- تابع نقشه کوسامبی

در فواصل بسیار کوتاه، تداخل کامل است، به‌طوری که فرض می‌شود که جایگاه B بین جایگاه‌های A و C قرار دارد، نوترکیبی بین A و B بر عدم وجود نوترکیبی بین B و C تأکید می‌کند و بر عکس و بنابراین یا r_{AB} و یا r_{BC} برابر صفر است؛ بنابراین نرخ‌های نوترکیبی در فواصل کوتاه تقریباً افزایشی بوده و باعث تأیید رابطه زیر می‌شوند:

$$r_{AC} = r_{AB} + r_{BC} \qquad (۳۰-۳)$$

حال آنکه در فواصل بلند، معادله (۳-۲۹) دقیق‌تر است. زمانی که نشانگرها در فواصل متوسط قرار گرفته باشند، رابطه بین نرخ‌های نوترکیبی به‌صورت زیر بیان می‌شود:

$$r_{AC} = r_{AB} + r_{BC} - r_{AB} r_{BC} \qquad (۳۱-۳)$$

در مجموع، مدل کلی توصیف کننده این روابط، می‌تواند به‌صورت زیر نوشته شود:

$$r_{AC} = r_{AB}(1 - r_{BC}) + r_{BC}(1 - r_{AB}) = r_{AB} + r_{BC} - 2crr_{AB} r_{BC}$$

که بر اساس معادله (۳-۱۹):

$$c = \frac{\frac{1}{2}\left(r_{AB} + r_{BC} - r_{AC}\right)}{r_{AB} r_{BC}} = \frac{g_{11}}{r_{AB} r_{BC}} \qquad (۳۲-۳)$$

معادله (۳-۳۲) میزان انحراف نوترکیبی‌های مشاهده شده در فواصل مختلف در بین نشانگرها را از نوترکیبی‌هایی که فرض می‌شود به‌صورت مستقل رخ می‌دهند، اندازه‌گیری می‌کند (مولر[2]، ۱۹۱۶). این انحراف (c) از استقلال به‌صورت ضریب تطابق[3] تعریف می‌شود، در حالی‌که I =1-c تداخل[4] نامیده می‌شود. عدم وجود تداخل (I = ۰) و مقادیر مثبت (I>۰) و منفی (I<۰) تداخل به ترتیب به c=۱، c<۱ و c>۱ مربوط می‌شوند (اوت، ۱۹۹۱). ضرایب c و I به عنوان خصوصیت فواصل $[A,B]$ و $[B,C]$ هستند.

---

[1] - Kosambi
[2] - Muller
[3] - Coefficient of coincidence
[4] - Interference

سپس، قصد داریم تا تابع $r = f(d)$ را پیدا کنیم که می‌تواند رابطه نرخ‌های نوترکیبی را، به همان صورتی که در معادلات (۳-۲۹) تا (۳-۳۱) توصیف شد، در فواصل ژنتیکی مختلف نشان دهد. فرض کنید که تابع f از رابطه زیر پیروی می‌کند:

$$f(d+h) = f(d) + f(h) - 2cf(d)f(h)$$

با یادآوری معادله مشتقی هالدین (۱۹۱۹)، خواهیم داشت:

$$\frac{f(d+h)-f(d)}{h} = \frac{f(h)}{h} - 2cf(d)\frac{f(h)}{h} \qquad (۳۳-۳)$$

اگر در فواصل کوتاه به $r = f(d) = d$ نیاز داشته باشیم، زمانی که h به سمت صفر میل می‌کند، $\frac{f(h)}{h}$ به سمت یک میل می‌کند و مشتق تابع بر اساس معادله (۳-۳۳) به‌صورت زیر بیان می‌شود:

$$f'(d) = 1 - 2c_0 f(d) = 1 - 2c_0 r$$

که c_0 تطابق حاشیه‌ای نامیده می‌شود و وجه تمایز آن از c به این دلیل است که در زمانی که یکی از دو فاصله به صفر می‌رسند، به‌عنوان حد مطرح می‌شود. زمانی که c_0 مقدار ثابت غیر صفر باشد، این معادله مشتقی دارای جواب زیر است:

$$r = \int \frac{1}{1-2c_0 r} dr = -\frac{1}{2c_0}\ln(1-2c_0 r)$$

که اگر $c_0 = 1$ باشد، این تابع همان تابع نقشه هالدین (۳-۲۷) است؛ اما وجود تداخل پیشنهاد می‌کند که مقادیر کوچکتر c_0 برای مقادیر کوچک‌تر r مناسب‌تر هستند. این حالت منجر به ایجاد تعداد نامحدودی از توابع نقشه می‌شود، ولی کوسامبی (۱۹۴۴) اشاره کرد که اگر c_0 تابع مناسبی از r باشد، در این صورت یک تابع نقشه منفرد می‌تواند به‌دست آید. ساده‌ترین تابع r که در فاصله $0 < r < 1$ افزایش می‌یابد و مقدار آن در r=0 برابر صفر و در $r = \frac{1}{2}$ برابر یک می‌شود، به‌صورت $c_0 = 2r$ است. از این‌رو معادله مشتقی به‌صورت زیر بیان می‌شود:

$$f'(d) = 1 - 2c_0 r = 1 - 4r^4$$

سپس با انتگرال‌گیری می‌توان تابع را به‌صورت زیر به‌دست آورد:

(۳۴-۳)

$$d = \frac{1}{4}\ln\frac{1+2r}{1-2r}$$

و معکوس آن به صورت زیر است:

$$r = \frac{1}{2}\frac{e^{2d} - e^{-2d}}{2e^{2d} + e^{-2d}} \qquad (۳۵-۳)$$

که این تابع به عنوان تابع نقشه کوسامبی شناخته می‌شود که به‌طور گسترده در تهیه نقشه‌های پیوستگی مورد استفاده قرار می‌گیرد. از روی معادله (۳-۳۳)، فرمول افزایشی کوسامبی را برای نرخ‌های نوترکیبی جایگاه‌های ژنی A-B-C به صورت زیر داریم:

$$\begin{aligned} r_{AC} &= \frac{1}{2}\tanh \quad 2d_{AC} \\ &= \frac{1}{2}\tanh(2d_{AB} + 2d_{BC}) \\ &= \frac{\frac{1}{2}\tanh 2d_{AB} + \frac{1}{2}\tanh 2d_{BC}}{1 + \tanh 2d_{AB}\tanh 2d_{BC}} \\ &= \frac{r_{AB} + r_{BC}}{1 + 4r_{AB}r_{BC}} \end{aligned} \qquad (۳۶-۳)$$

این رابطه، مشابه نقش افزایش سرعت به نظریه اختصاصی نسبیت است.

بسیاری از تغییرات این توابع نقشه کلاسیک برای وضعیت‌های مختلف و اغلب با در نظر گرفتن تداخل انجام شده است. تابع کارتر و فالکونر[1] (۱۹۵۱) تداخل نسبتاً قوی را در نظر می‌گیرد. سطوح متغیر تداخل به وسیله فلسنشتاین[2] (۱۹۷۹) مدل‌سازی شد. تمامی این توابع می‌توانند از روی معادله مشتقی هالدین (۱۹۱۹) با فرض تطابق‌های جانبی (c) متفاوت به‌دست آیند (لیبرمن و کارلین[3]، ۱۹۸۴).

مثال ۳-۸- مجدداً مثال ۳-۶ را برای تجزیه سه نقطه‌ای در برنج مشاهده کنید. نرخ‌های نوترکیبی برای سه نشانگر، RG472 (A)، RG246 (B) و K5 (C) در مثال ۳-۶ برآورد شدند. بهترین ترتیب سه نشانگر در مثال ۳-۷ به صورت A-B-C تعیین شد. در این بخش، از توابع نقشه هالدین و کوسامبی برای برآورد فواصل ژنتیکی این سه نشانگر به صورت زیر استفاده می‌کنیم:

---

[1] - Carter and Falconer
[2] - Felsenstein
[3] - Liberman and Karlin

جفت نشانگر	هالدین		کوسامبی	
	r	d (cM)	r	d (cM)
A-B	۰/۱۹۰	۲۳/۹۰۲	۰/۱۹۰	۲۰/۰۰۳
B-C	۰/۱۵۰	۱۷/۸۳۴	۰/۱۵۰	۱۵/۴۷۶
A-C	۰/۲۸۰	۴۱/۰۴۹	۰/۲۸۰	۳۱/۶۴۲
(A-C)	(۰/۲۸۳)	(۴۱/۷۳۶)	(۰/۳۰۵)	(۳۵/۴۷۹)

نرخ نوترکیبی بین نشانگرهای A و C در دو انتها می‌تواند به‌طور مستقیم بر اساس روش توصیف شده در تجزیه سه نقطه‌ای برآورد شود. این نرخ نوترکیبی به‌صورت غیرمستقیم نیز می‌تواند با استفاده از معادلات (۳-۲۹) و (۳-۳۶) بر اساس برآورد نرخ‌های نوترکیبی دیگر یعنی r_{AB} و r_{BC} برآورد شود. سپس فاصله ژنتیکی بین نشانگرهای A و C را با استفاده از برآوردهای حاصل از روش غیر مستقیم (که در پرانتز نشان داده شده است) محاسبه می‌کنیم. برآوردهای فاصله ژنتیکی بین A و C و بین روش‌های مستقیم و غیر مستقیم برای تابع نقشه هالدین مطابقت دارند، حال آنکه برای تابع کوسامبی متفاوت هستند. این امر به دلیل فرض‌های متفاوتی است که برای به‌دست آوردن این دو تابع نقشه مورد استفاده قرار می‌گیرند. تابع نقشه هالدین فرض می‌کند که تداخل بین دو فاصله نشانگری نزدیک به هم وجود ندارد، در حالی‌که وجود این فرض برای تابع نقشه کوسامبی ضرورت ندارد.

در پایان این بخش، توابع نقشه هالدین و کوسامبی را به‌صورت زیر خلاصه می‌کنیم:

تابع	r(d)	d(r)
هالدین	$r=\frac{1}{2}(1-e^{-2d})$	$d=-\frac{1}{2}\ln(1-2r)$
کوسامبی	$r=\frac{1}{2}\frac{e^{2d}-e^{-2d}}{2e^{2d}+e^{-2d}}$	$d=\frac{1}{4}\ln\frac{1+2r}{1-2r}$

مقایسه بین این دو تابع با رسم نمودار فاصله ژنتیکی در برابر نرخ نوترکیبی و بالعکس صورت گرفته است (شکل ۳-۲). برای دو نشانگر که تا حد زیادی به همدیگر پیوسته هستند (یعنی مقدار r کوچک است)، این دو تابع، فواصل ژنتیکی مشابهی ایجاد می‌نمایند؛ اما واگرایی موجود در فاصله ژنتیکی برآورد شده به وسیله دو تابع با افزایش مقادیر r افزایش می‌یابد. زمانی که r نزدیک به ۰/۵ باشد، دو تابع تمایل به همگرایی دارند.

۳-۱۱- تمرین‌ها

۳-۱- بررسی کنید که MLE مربوط به r ارائه شده در معادله (۳-۴) از راه مشتق‌گیری درستنمایی (۳-۳) حاصل می‌شود.

۳-۲- در ارتباط با تجزیه دو نقطه‌ای، یک جمعیت F_2 را در نظر بگیرید که در آن دو نشانگر تعیین ژنوتیپ شده‌اند. دو نشانگر فوق، ۹ ژنوتیپ قابل شناسایی هر یک با مشاهدات زیر ایجاد می‌کنند:

$$n = \begin{array}{c} \\ BB \\ Bb \\ bb \end{array} \begin{array}{c} \begin{array}{ccc} AA & Aa & aa \end{array} \\ \begin{bmatrix} 107 & 20 & 3 \\ 24 & 175 & 18 \\ 5 & 41 & 93 \end{bmatrix} \end{array}$$

الف. با استفاده از الگوریتم EM، نرخ نوترکیبی بین این دو نشانگر را برآورد کنید.

ب. آزمون کنید که آیا این دو نشانگر با استفاده از آزمون نسبت درستنمایی و آزمون x^2 به‌طور معنی‌داری به هم پیوسته هستند یا خیر.

ج. خطای نمونه‌برداری نرخ نوترکیبی برآورد شده و فاصله اطمینان نرخ نوترکیبی را برآورد کنید.

۳-۳- در ارتباط با ترتیب نشانگرها، سه نشانگر مختلف برای جمعیت تلاقی برگشتی در ارتباط با والد aabbcc تعیین ژنوتیپ شده‌اند. ژنوتیپ‌های سه نشانگر به‌صورت زیر مشاهده شده‌اند:

DH	ABC/abc	ABc/abc	AbC/abc	Abc/abc	aBC/abc	aBc/abc	abC/abc	abc/abc
تعداد مشاهدات	٦٠	١	٢	١	١٤	٠	٩	١٥

الف. نرخ‌های نوترکیبی بین هر جفت از نشانگرها را برآورد کنید.

ب. سطح معنی‌دار این نرخ‌های نوترکیبی برآورد شده را آزمون کنید.

ج. بهترین ترتیب این سه نشانگر را تعیین نمایید.

د. فواصل ژنتیکی بین هر جفت نشانگر را با استفاده از توابع نقشه هالدین و کوسامبی محاسبه کنید.

ه. چگونه می‌توانید نرخ نوترکیبی و فاصله ژنتیکی بین نشانگرها را در دو انتها برآورد نمایید؟ نتایج حاصل از روش‌های مختلف را با یکدیگر مقایسه نمایید.

۳-۴- در ارتباط با تجزیه دو نقطه‌ای، در جمعیت F_2 دوم، ۹ ژنوتیپ برای دو نشانگر به‌صورت زیر مشاهده شده است:

$$n = \begin{array}{c} \\ BB \\ Bb \\ bb \end{array} \begin{array}{c} \begin{array}{ccc} AA & Aa & aa \end{array} \\ \begin{bmatrix} 0 & 8 & 113 \\ 18 & 202 & 27 \\ 2 & 202 & 27 \end{bmatrix} \end{array}$$

با استفاده از الگوریتم EM، نرخ نوترکیبی را برآورد کنید. اگر برآورد شما بزرگ‌تر از ۰/۵ باشد که تعریف این پارامتر را نقض می‌کند، چگونه می‌توان آن‌را توضیح داد؟

۳-۵- MLE حاصل از معادله (۳-۱۳) را با الف) حل کردن معادله درجه سوم و ب) با استفاده از الگوریتم EM به‌دست آورید.

۳-۶- در ارتباط با درستنمایی معادله (۳-۲۰):

الف. نشان دهید که این رابطه می‌تواند به صورت حاصل‌ضرب درستنمایی‌های دو جمله‌ای نوشته شود و MLE های مربوط به $\hat{r}_{AB}$ و $\hat{r}_{BC}$ را بیابید. نشان دهید که معادله (۳-۲۱) تقریبی از واریانس‌های مربوطه است.

ب. MLE های دو ترتیب دیگر را بیابید (جدول ۳-۵ را مشاهده کنید). کدامیک MLE های کلی و عمومی هستند؟

۳-۱۲- نکات: الگوریتم‌ها و نرم‌افزارهای تهیه نقشه

نقشه‌های پیوستگی ژنتیکی به‌عنوان ابزار قدرتمندی برای مطالعه ساختار و سازماندهی ژنوم گونه‌ها و شناسایی و تشخیص جایگاه‌های ژنی کنترل کننده صفات کمّی استفاده می‌شوند. تهیه نقشه‌های پیوستگی شامل دو مرحله است: ۱- گروه‌بندی نشانگرها در گروه‌های پیوستگی و ۲- تعیین ترتیب نشانگرها در هر گروه پیوستگی. نشانگرها می‌توانند بر اساس روابط پیوستگی در گروه‌های پیوستگی قرار گیرند. درجه پیوستگی بین هر دو نشانگر که بر حسب مقدار نوترکیبی اندازه‌گیری می‌شود، می‌تواند با روش تجزیه دو و سه نقطه‌ای برآورد شود. نشانگرها می‌توانند بر اساس ماتریسی که شامل تمامی نوترکیبی‌های بین جفت نشانگرها، به وسیله تجزیه خوشه‌ای[1] گروه‌بندی

[1] - Cluster analysis

شوند. معیار خاصی برای مشخص کردن نقطه برش[1] استفاده می‌شود تا تعداد گروه‌های پیوستگی مشخص شود.

پس از تخصیص گروه‌های پیوستگی، ترتیب نشانگرهای درون هر گروه با به حداقل رساندن تفاوت‌های موجود بین نرخ‌های نوترکیبی حاصل از داده‌های جفتی و نرخ‌های محاسبه شده در نقشه به‌دست می‌آیند؛ اما زمانی‌که تعداد نشانگرها افزایش می‌یابد، تعداد ترتیب‌های ممکنه نشانگری نیز به سرعت افزایش می‌یابند که نشان‌دهنده ایجاد چالش قابل توجه در محاسبات است. در حقیقت، تعیین ترتیب نشانگرها می‌تواند به‌صورت یک مورد خاص در مسئله بازرگانان در حال مسافرت (TSP) در نظر گرفته شود (ویلسون[2]، ۱۹۸۸؛ السون و بوهنکه[3]، ۱۹۹۰؛ فالک[4]، ۱۹۹۲؛ لیو[5]، ۱۹۹۸؛ گاسپین و شیکس[6]، ۱۹۹۷؛ مستر و همکاران[7]، ۲۰۰۳) که به‌عنوان یک مسئله کامل‌چندجمله‌ای مجهول کلاسیک در ریاضیات و علوم رایانه محسوب می‌شود. مسئله تعیین ترتیب نشانگرها می‌تواند با انجام جستجوهای گسترده مورد بررسی و مدیریت قرار گیرد. هنگامی‌که تعداد نشانگرها زیاد است، یعنی بیش از ۳۰ نشانگر، انجام جستجوی جامع در مورد تمامی ترتیب‌های ممکن، بی‌نهایت وقت‌گیر است. به همین دلیل، الگوریتمی که برای به‌دست آوردن جواب تقریبی بهینه به‌کار گرفته می‌شود، یک روش عملی برای تجزیه پیوستگی در مقیاس وسیع است.

چندین الگوریتم تقریبی برای تهیه نقشه به وسیله «تان و فو»[8] (۲۰۰۶) معرفی شده‌اند که شامل تعیین ترتیب (بوتو و چاکراوارتی[9]، ۱۹۸۷)، اتصال شبیه‌سازی شده[10] (تامپسون، ۱۹۸۴؛ ویکز و لانژ[11]، ۱۹۸۷)، باز و بست (لاتروپ و همکاران[12]، ۱۹۸۵)، الگوریتم لندر و گرین[13] (۱۹۸۷) و درستنمایی گام به گام (لاتروپ و همکاران، ۱۹۸۴) هستند. بسته‌های نرم‌افزاری که از این الگوریتم‌ها استفاده می‌کنند شامل LINKAGE (لاتروپ و همکاران، ۱۹۸۴)،

---

[1] - Cutoff
[2] - Wilson
[3] - Olson and Boehnke
[4] - Falk
[5] - Liu
[6] - Gaspin and Schiex
[7] - Mester *et al.*
[8] - Tan and Fu
[9] - Buetow and Chakravarti
[10] - Simulated annealing
[11] - Weeks and Lange
[12] - Lathrop *et al.*
[13] - Lander and Green

MAPMAKER/EXP (لندر و همکاران، ۱۹۸۷)، LINKAGE MAP (اپیگ و ایکر[1]، ۱۹۸۳)، JoinMap (استام[2]، ۱۹۹۳)، LINKAGE-1 (سوئیتر و همکاران[3]، ۱۹۸۳) و GMendel (اچت و همکاران[4]، ۱۹۹۲) هستند. به‌طورکلی، عملکرد تمامی این الگوریتم‌ها و نرم افزارها ممکن است به وسیله خطاهای آزمایشی، اندازه نمونه و تداخل نوترکیبی و کراس‌اورهای مضاعف تحت‌تأثیر قرار گیرد.

حائز اهمیت است که دو روش اخیر برای تهیه نقشه را ذکر کنیم. اولین راهکار، الگوریتم تکاملی و ژنتیکی (GEA) مستر و همکاران (۲۰۰۳) است که به‌طور سازگاری با پیروی از فرآیندهای تکاملی یک جمعیت نظیر جهش، نوترکیبی و انتخاب در جستجوی یک راه حل بهینه است. دومین راهکار به وسیله تان و فو (۲۰۰۶) پیشنهاد شد که از اصل مکان‌یابی مجاور (الیس[5]، ۱۹۹۷) برای تهیه نقشه پیوستگی ژنتیکی با استفاده از یک نقشه کوچک و افزودن نشانگرها به آن در یک زمان استفاده کردند. در مقایسه با الگوریتم‌های دیگر، الگوریتم ترتیبی تان و فو مزایای مهمی از نظر سرعت محاسباتی و دقت تشخیص ترتیب صحیح نشانگرها دارد.

اصول آماری خاصی در پس‌زمینه هر یک از الگوریتم‌های تهیه نقشه و بسته‌های نرم‌افزاری ذکر شده وجود دارند. برای کمک به فهم خواننده، برخی توضیحات مربوط به اصول آماری پس‌زمینه الگوریتم لندر و گرین را که به‌طور گسترده‌ای برای تهیه نقشه موجودات زنده آزمایشی مورد استفاده قرار می‌گیرد را ارائه می‌کنیم. خوانندگان نیز می‌توانند به شم[6] (۱۹۹۸) مراجعه کنند که کاربرد الگوریتم لندر و گرین را در مقایسه با الگوریتم الستون و استوارت[7] (۱۹۷۱) مرور نموده است.

الگوریتم لندر-گرین بر اساس فرمول پنهانی مارکوف در خصوص الگوی توارث چندین نشانگر مرتب شده است. اطلاعات پیوستگی یک هاپلوتیپ می‌تواند به‌صورت رشته‌ای از f آلل حاصل از هاپلوتیپ‌های والد پدری و m آلل حاصل از هاپلوتیپ‌های والد مادری خلاصه شود. در یک نشانگر منفرد، اطلاعات پیوستگی هر یک از اعضای غیر پایه‌گذار[8] جمعیت می‌تواند فقط با دو متغیر دو تایی، یکی برای آلل هاپلوتیپ پدری و دیگری برای آلل هاپلوتیپ مادری خلاصه شود. یک شجره با N فرد غیر پایه‌گذار دارای 2N گامت حاوی اطلاعات بالقوه است، به‌طوری که الگوی توارث در یک

---

[1] - Eppig and Eicher
[2] - Stam
[3] - Suiter *et al.*
[4] - Echt *et al.*
[5] - Ellis
[6] - Sham
[7] - Elston and Stewart
[8] - Nonfounding

نشانگر منفرد می‌تواند به وسیله برداری از 2N جزء یا عنصر توصیف شود. هر عنصر از این بردار توارثی، v، وضعیت والدینی آلل را در نشانگر موجود در یک گامت توصیف می‌کند: اگر آلل مربوط به والد پدری باشد، به‌طور قراردادی به صورت صفر و اگر آلل مادری باشد به صورت یک تعریف می‌شود؛ بنابراین 2^{2n} بردار توارثی ممکن که هر یک توصیف کننده الگوی متفاوتی از انتقال آلل در جایگاه ژنی هستند، وجود دارد. برای هر بردار توارثی ممکن در هر جایگاه ژنی، احتمال داده‌های فنوتیپی مرتبط با جایگاه ژنی می‌تواند به صورت مجموع حاصل‌ضرب‌های احتمالات شرطی فنوتیپ‌ها (X) مشروط بر ژنوتیپ (g) و احتمالات شرطی ژنوتیپ (g) مشروط به بردار توارث (v) بدین صورت تجزیه شود:

$$P(X \mid v) = \sum_{G} P(X \mid g) P(g \mid v)$$

این رابطه، مجموع تمامی ترکیبات ژنوتیپ‌های ممکن (G) را در جایگاه ژنی شامل می‌شود؛ بنابراین، این احتمال تابعی از میزان نفوذ جایگاه ژنی و پارامترهای جمعیت است. 2^{2n} عدد از چنین احتمالاتی که یکی برای هر یک از بردارهای توارث ممکنه است، می‌تواند به‌صورت عناصر قطری یک ماتریس مربع $2^{2n} \times 2^{2n}$ مرتب شود که این ماتریس به‌صورت Q نشان داده می‌شود. زمانی که تمامی 2^{2n} بردار توارثی ممکن پیش از در نظر گرفتن داده‌های فنوتیپی دارای احتمال حضور مساوی باشند، درستنمایی داده‌ها در هر جایگاه ژنی متناسب با مجموع این 2^{2n} احتمال است. این حالت می‌تواند به صورت ماتریسی به شکل زیر نشان داده شود:

$$P(X) \propto 1^T Q 1$$

که X نشان‌دهنده داده‌های فنوتیپی در جایگاه ژنی و ۱ یک بردار ستونی است که تمامی 2^{2n} عنصر آن یک هستند. زمانی که k نشانگر مرتب شده به‌طور توأم در یک تجزیه چند نقطه‌ای در نظر گرفته شوند، درستنمایی توأم می‌تواند به صورت زیر فاکتور گرفته شود:

$$P(X_1, X_2, ..., X_k) = P(X_1) P(X_1 \mid X_2) P(X_3 \mid X_1, X_2) ... P(X_k \mid X_1, X_2, ..., X_{k-1})$$

که $X_1, X_2, ..., X_k$ داده‌های فنوتیپی در k نشانگر مرتب شده است. ضرورت مشروط کردن به تمامی نشانگرهای قبلی می‌تواند با آگاهی از اینکه در هر جایگاه ژنی، احتمال داده‌های فنوتیپی تابعی از بردار توارث خود داده‌ها بوده و به‌طور شرطی مستقل از بردارهای توارثی تمامی ماتریس‌های قبلی به شرط بردار توارث جایگاه ژنی قبلی است، مورد اجتناب قرار گیرد؛ به عبارت دیگر، k بردار توارث از k نشانگر مرتب شده، جزیی از زنجیره پنهان شده مارکوف هستند. احتمال شرطی یک بردار توارثی v_{i+1} در نشانگرهای i+۱ مشروط به بردار توارثی v_i در نشانگرهای i، برابر $r_i^j (1-r_i)^{2N-j}$ است که r_i نرخ نوترکیبی بین نشانگرهای i و i+۱ است و j تعداد تغییرات در عناصر بردار توارثی از v_i تا v_{i+1} است که هر کدام با یک احتمال شرطی مشخص مرتبط هستند،

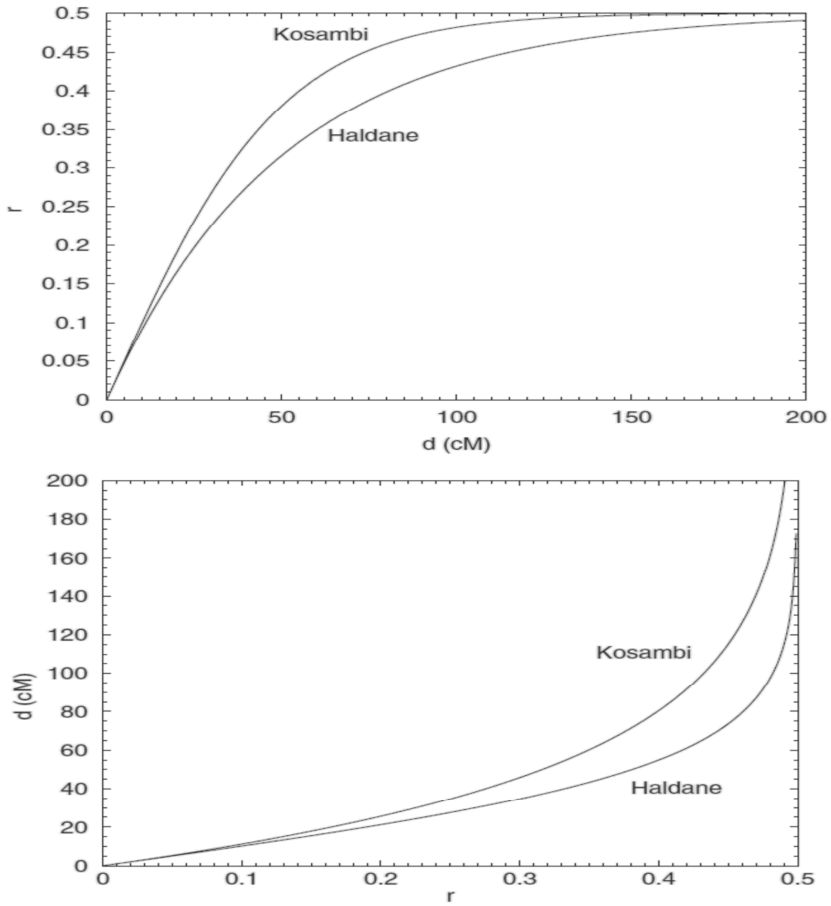
به‌طوری که $\left[2^{2n}\right]^2$ احتمال شرطی وجود دارد که می‌تواند در یک ماتریس مربع انتقالی به نام $\mathbf{T_i}$ مرتب شوند. درستنمایی توأم داده‌های فنوتیپی چند مکانی به صورت زیر ارایه می‌شود:

$$P(X_1, X_2, ..., X_k) \propto 1^T Q_1 T_1 Q_2 T_2 ... T_{k-1} Q_k 1$$

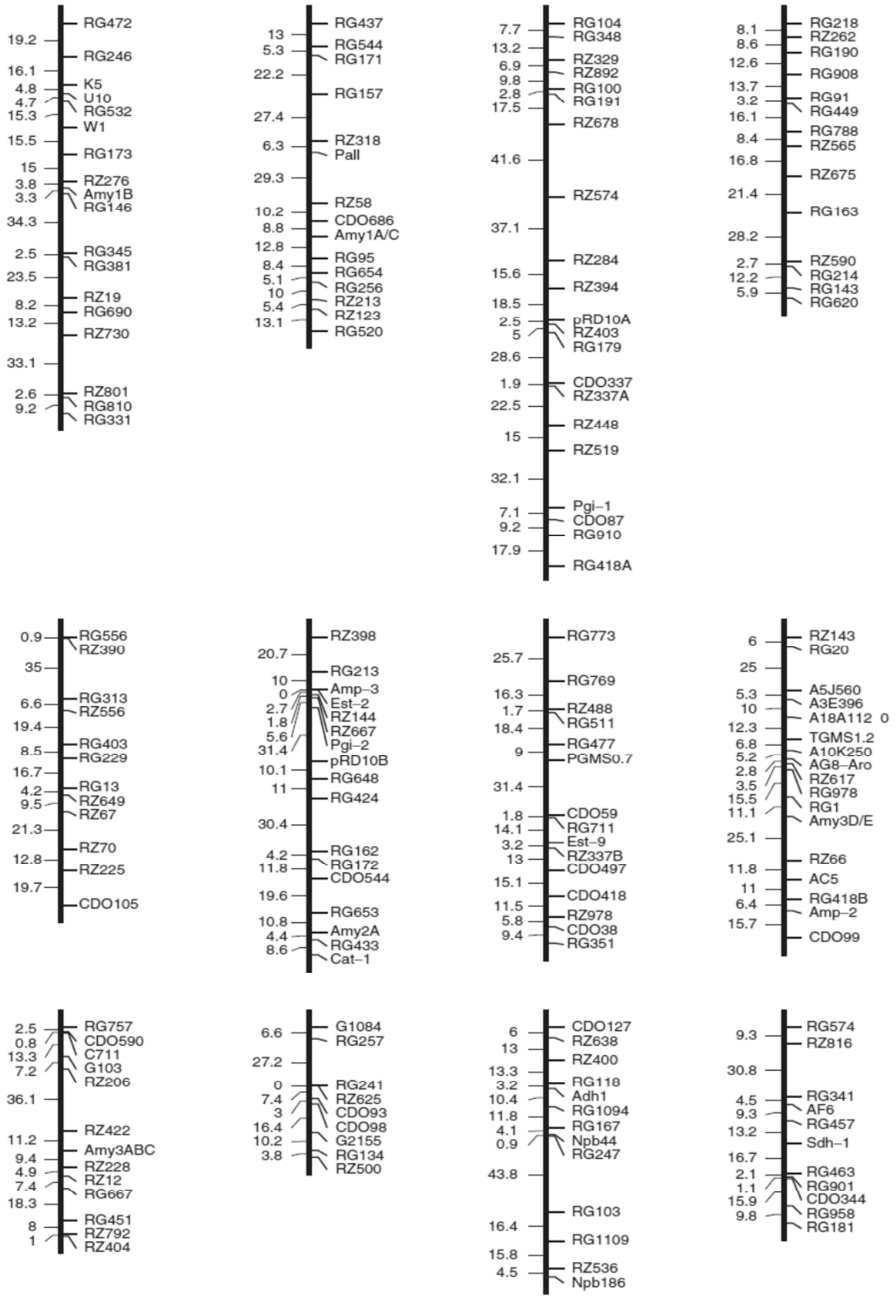
این رابطه شکل پایه‌ای و اصولی الگوریتم لندر-گرین است که مراحل پاک سازی مختلفی به آن افزوده شده تا تعداد عملیات‌های ریاضی ضروری را کاهش دهد.

مثال ۳-۹. **(نقشه پیوستگی ژنتیکی)**مثال ۳-۱ را مجدداً در نظر بگیرید. ۱۲۳ گیاه دابل هاپلوئید (DH) حاصل از دو لاین اینبرد، IR64 نیمه پاکوتاه و Azucena پا بلند، برای مجموعاً ۱۷۵ نشانگر چندشکل تعیین ژنوتیپ شدند (هوانگ و همکاران، ۱۹۹۷). تجزیه دو نقطه‌ای برای برآورد نرخ‌های نوترکیبی جفتی این ۱۷۵ نشانگر نیز انجام شد. بر اساس تجزیه خوشه‌ای ماتریس ۱۷۵×۱۷۵ برای توابع نوترکیبی، این نشانگرها در قالب ۱۲ گروه مختلف مرتب شدند که هر یک نشان‌دهنده یک کروموزوم برنج است (شکل ۳-۳).

الگوریتم لاندر-گرین برای تعیین بهترین ترتیب نشانگرهای خوشه‌بندی شده در هر گروه مورد استفاده قرار گرفت (شکل ۳-۳). فواصل ژنتیکی بین تمامی نشانگرهای مجاور برآورد شد و فواصل ژنتیکی مربوطه با استفاده از یک تابع نقشه محاسبه شدند. در این مثال، تابع نقشه کوسامبی مورد استفاده قرار گرفت. در شکل ۳-۳ فواصل ژنتیکی و اسامی نشانگرها در طرف چپ و راست کروموزوم‌ها ارایه شده‌اند.



شکل ۳-۲- فاصله ژنتیکی (d) به عنوان تابعی از نرخ نوترکیبی (r) و بالعکس که با توابع هالدین و کوسامبی برآورد شده‌اند.



شکل ۳-۳- نقشه پیوستگی ژنتیکی حاصل از ۱۳۵ نشانگر RFLP و ۴۰ نشانگر ایزوزیم و RAPD برای ۱۲۳ گیاه DH که از والدین Azucena پا بلند و IR64 پا کوتاه به‌دست آمده است.

فصل چهارم

مدل کلی برای تجزیهٔ پیوستگی در تلاقی‌های کنترل شده

۴-۱- مقدمه

روش‌های آماری تجزیه پیوستگی در یک جمعیت تلاقی برگشتی یا F_2 حاصل از دو لاین اینبرد در فصل قبل توصیف شدند. مزیت تجزیهٔ پیوستگی با استفاده از تلاقی لاین‌های اینبرد آن است که فاز پیوستگی والدینی بین جایگاه‌های ژنتیکی متفاوت شناخته شده است و بنابراین الگوی تفرق نشانگرها می‌تواند تعیین و پیوستگی بر حسب نرخ نوترکیبی اندازه‌گیری و آزمون شود. در هر حال، تجزیه‌های مبتنی بر لاین‌های اینبرد برای گونه‌های دگربارور که در آن‌ها تولید لاین‌های خالص هموزیگوت از راه انجام آمیزش‌های همخونی متوالی امکان پذیر نیست، مناسب نیست.

جوامع دگربارور دو خصوصیت مهم دارند که تجزیهٔ پیوستگی در آن‌ها را از نظر کیفی متفاوت از تجزیه پیوستگی در تلاقی‌های لاین‌های اینبرد می‌نماید. اولین خصوصیت آن است که به‌طور کلی تعداد آلل‌ها و روش توارث نشانگرها از یک جایگاه ژنی به جایگاه دیگر متفاوت است. برخی نشانگرها ممکن است تعداد آلل‌های بیشتری نسبت به نشانگرهای دیگر داشته باشند، برخی از نشانگرها هم‌بارز هستند، در حالی‌که نشانگرهای دیگر غالب بوده و برخی از نشانگرها در یک والد هتروزیگوت، ولی در والد دیگر تثبیت شده‌اند، حال آنکه مسئله در مورد نشانگرهای دیگر به صورت عکس است. خصوصیت دوم، عدم قطعیت در خصوص فازهای پیوستگی بین جایگاه‌های ژنی متفاوت است. چون برآورد پیوستگی بین نشانگرهای مختلف بر اطلاعاتی تکیه دارد که در خصوص فازهای پیوستگی موجود است، تعیین فاز پیوستگی صحیح قبل از برآورد پیوستگی ضرورت می‌یابد.

استراتژی مرسوم برای برآورد فازهای پیوستگی در نظر گرفتن تمامی فازهای پیوستگی ممکن برای تمامی جفت نشانگرهای معین و انتخاب محتمل‌ترین فاز پیوستگی بر اساس حداقل نرخ نوترکیبی و حداکثر مقدار درستنمایی است (ریتر و همکاران[1]، ۱۹۹۰؛ ریتر و سالامینی[2]، ۱۹۹۶؛ مالیپارد و همکاران[3]، ۱۹۹۷؛ ریدوت و همکاران[4]، ۱۹۹۸)؛ اما این استراتژی همیشه از نظر آماری مؤثر نیست، چون برآورد حداقل نرخ نوترکیبی ممکن است از روی فاز پیوستگی نادرست به‌دست آید. وو و همکاران (b ۲۰۰۲) یک الگوریتم کلی برای برآورد همزمان پیوستگی و فازهای پیوستگی والدینی در انواع نشانگرهای مولکولی پیوسته در یک خانواده تنی حاصل از دو والد غیراینبرد ارائه کردند. لو و همکاران (۲۰۰۴) یک مدل یکنواخت را برای تشخیص پیوستگی، فازهای پیوستگی

---

[1] - Ritter *et al.*
[2] - Ritter and Salamini
[3] - Maliepaard *et al.*
[4] - Ridout *et al.*

والدینی و ترتیب ژنی برای انواع نشانگرها پیشنهاد دادند. در این فصل، یک چارچوب کلی برای برآورد هم‌زمان پیوستگی و فازهای پیوستگی ارائه می‌شود که می‌تواند به‌عنوان تعمیم تجزیه پیوستگی در تلاقی لاین‌های اینبرد باشد. بیشتر مطالبی که ارائه می‌شود از مقاله لو و همکاران (۲۰۰۴) اقتباس شده است.

۴-۲- نشانگرهای حاوی اطلاعات کامل: یک مدل دیپلوتیپی

در یک خانواده تنی مشتق از دو والد P و Q مربوط به یک گونه دگربارور، بیش از چهار آلل نشانگری، علاوه بر آلل خاموش، ممکن است در یک جایگاه ژنی منفرد در حال تفرق باشند. فرض کنیم هر یک از آلل‌های نشانگر که با علایم a، b، c و d نشان داده می‌شوند، نسبت به همدیگر هم‌بارز و نسبت به آلل خاموش که با o نمایش داده می شود، غالب باشند. همچنین فرض کنیم که آلل‌های متفاوت در دو والدی که برای تلاقی استفاده می‌شوند، ترکیب شده و مجموعاً ۱۸ نوع تلاقی ممکن برای جایگاه نشانگر وجود دارد (جدول ۳-۲). در بخش ۳-۴، نشان دادیم که چگونه این نوع تلاقی‌ها بر اساس الگوهای باندی نشانگرهای والدینی و نتاج به ۷ گروه طبقه‌بندی می‌شوند. تجزیه تفرق اجازه می‌دهد تا نوع احتمالی تلاقی حاصل از داده‌های خام را تعیین کنیم.

در این بخش، یک مدل کلی برای نشانگرهای حاوی اطلاعات کامل که در آن‌ها، چهار نوع ژنوتیپ قابل تشخیص از نظر فنوتیپی در خانواده‌های تنی وجود دارد، ارائه خواهیم داد. مدل‌های تجزیه پیوستگی برای نشانگرهای حاوی اطلاعات ناقص که در آن‌ها برخی از ژنوتیپ‌های مختلف از نظر فنوتیپی مشابه هستند، در بخش‌های بعدی مورد بحث قرار خواهد گرفت.

۴-۲-۱- تجزیه دو نقطه‌ای

تعریف فاز پیوستگی را به یاد آورید (تعریف ۱-۱). فاز پیوستگی آرایش آلل‌ها را در یک جفت ژن هتروزیگوت روی کروموزوم‌های همولوگ در یک والد منفرد توصیف می‌کند. در صورتی که بدانیم هر یک از کروموزوم‌های همولوگ، چه آلل‌هایی از هر یک از نشانگرها را در یک والد حمل می‌کنند، در این صورت می‌توان فاز پیوستگی بین هر دو نشانگر پیوسته را تعیین نمود؛ بنابراین، مسئله تعیین فاز پیوستگی به عنوان یک مسئله علامت‌گذاری کروموزوم‌های والدینی با استفاده از آلل‌های نشانگرهای مورد نظر مطرح می‌شود. ارتباط آلل‌های نشانگر و کروموزوم‌های همولوگ می‌تواند با محاسبه احتمالات ژنوتیپ‌های نشانگر مشروط به وضعیت نشانگرهای پیوسته در خانواده تنی استحکام یابد.

دو نشانگر حاوی اطلاعات کامل را به صورت **A** و **B** در یک خانواده تنی در نظر بگیرید. برای اولین نشانگر A، کروموزوم‌های والدینی می‌توانند به‌صورت اختیاری با آلل‌هایشان علامت‌گذاری

شوند. فرض کنید کروموزوم‌های والدینی نشانگر A به صورت $A_1 \mid \mid A_2$ (یا برای ساده‌سازی به صورت $1 \mid \mid 2$) برای والد P و $A_3 \mid \mid A_4$ (یا برای ساده‌سازی $3 \mid \mid 4$) برای والد Q باشند که $\mid \mid$ به ترتیب نشان‌دهنده دو کروموزوم همولوگ در سمت راست و چپ است. تلاقی والدین P و Q منجر به ایجاد چهار ژنوتیپ مختلف در این نشانگر در نتاج حاصل به صورت A_1A_3، A_1A_4، A_2A_3 و A_2A_4 یا 13، 14، 23 و 24 می‌شود. فاز پیوستگی بین آلل‌های نشانگر A و B می‌تواند با انتساب آلل‌های جایگزین نشانگر B به کروموزوم همولوگ متفاوت مشروط به علامت تعریف شده نشانگر A تعیین شود. برای هر والد، دو فاز پیوستگی ممکن وجود دارد؛ بنابراین، زمانی که دو والد با یکدیگر آمیزش می‌کنند، چهار ترکیب فازی امکان‌پذیر است که یکی از آن‌ها می‌تواند به‌صورت شماتیک به شکل زیر بیان شود:

$$\overset{\mathrm{P}\times\mathrm{Q}}{\Phi = \begin{matrix} A & 1 \\ B & 1 \end{matrix}\left\|\begin{matrix} 2 & 3 \\ 2 & 3 \end{matrix}\right\|\begin{matrix} 4 \\ 4 \end{matrix}} \quad \text{یا} \quad \overset{\mathrm{P}\times\mathrm{Q}}{[11][22]\times[33][44]} \qquad (۱-۴)$$

برای یک والد خاص، ترکیب کروموزوم‌های تعیین فاز شده، دیپلوتیپ والدینی نامیده می‌شود که همانند آنچه که در رابطه (۴-۱) نمایش داده شد، به صورت $[..][..]$ نمایش داده می‌شود. r به عنوان نرخ نوترکیبی بین دو نشانگر است. با فرض اینکه دیپلوتیپ‌های دو والد شناخته شده هستند، همانند آنچه که در بخش بالا نشان داده شد، الگوی تفرق همزمان دو نشانگر می‌تواند به شکل ماتریسی به‌صورت زیر بیان شود:

$$H = \begin{array}{c} \\ ۱۳ \\ ۱۴ \\ ۲۳ \\ ۲۴ \end{array} \begin{array}{c} \begin{array}{cccc} ۱۳ & ۱۴ & ۲۳ & ۲۴ \end{array} \\ \begin{bmatrix} \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{r^2}{4} \\ \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} & \frac{r^2}{4} & \frac{r(1-r)}{4} \\ \frac{r(1-r)}{4} & \frac{r^2}{4} & \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} \\ \frac{r^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} \end{bmatrix} \end{array} \qquad (۲-۴)$$

که هر یک از سلول‌ها نشان‌دهنده ژنوتیپ دو نشانگر در افراد تنی یعنی نتاج هستند. ستون‌ها نیز به نشانگر A و ردیف‌ها به نشانگر B تعلق دارند.

تعداد مورد انتظار پیشامدهای نوترکیب (یعنی تعداد r) که بین دو نشانگر می‌تواند رخ دهد نیز به شکل ماتریسی زیر بیان می‌شود:

$$D = \begin{array}{c} \\ ۱۳ \\ ۱۴ \\ ۲۳ \\ ۲۴ \end{array} \begin{array}{c} \begin{array}{cccc} ۱۳ & ۱۴ & ۲۳ & ۲۴ \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{bmatrix} \end{array} \qquad (۳-۴)$$

جایی که:

$$n = \left(n_{j1j2}\right)_{4\times4} = \begin{array}{c} \\ ۱۳ \\ ۱۴ \\ ۲۳ \\ ۲۴ \end{array} \begin{array}{c} \begin{array}{cccc} ۱۳ & ۱۴ & ۲۳ & ۲۴ \end{array} \\ \begin{bmatrix} n_{11} & n_{12} & n_{13} & n_{14} \\ n_{21} & n_{22} & n_{23} & n_{24} \\ n_{31} & n_{32} & n_{33} & n_{34} \\ n_{41} & n_{42} & n_{43} & n_{44} \end{bmatrix} \end{array} \qquad (۴-۴)$$

ماتریس مشاهدات نتاج است که در آن $j_1, j_2 = 1$ برای ۱۳، ۲ برای ۱۴، ۳ برای ۲۳ یا ۴ برای ۳۴ است که به ترتیب نشان دهنده فنوتیپ‌های نشانگر در A و B است. توجه کنید که یک نشانگر دارای فنوتیپ خاصی است که به وسیله ژنوتیپ خود تعیین می‌شود. اگر $n_{j_1 j_2}$ از توزیع چندجمله‌ای تبعیت کند، تابع درستنمایی ژنوتیپ‌های نشانگر تحت ترکیبات دیپلوتیپی والدینی که در قالب رابطه (۴-۱) نشان داده شده است، به صورت زیر بیان می‌شود:

$$L(r \mid n) \propto r^{(2n_2+n_3+n_4)}(1-r)^{(2n_1+n_3+n_4)} \qquad (۵-۴)$$

که در آن:

$$\begin{aligned} n_1 &= n_{11} + n_{22} + n_{33} + n_{44} \\ n_2 &= n_{14} + n_{23} + n_{32} + n_{41} \\ n_3 &= n_{12} + n_{21} + n_{34} + n_{43} \\ n_4 &= n_{13} + n_{31} + n_{24} + n_{42} \end{aligned} \qquad (۶-۴)$$

برآوردهای حداکثر درستنمایی (MLE) نرخ نوترکیبی (r) با واریانس‌های نمونه‌ای بزرگ مربوطه به‌صورت زیر هستند:

$$\hat{r} = \frac{2n_2 + n_3 + n_4}{2n}$$

$$Var(\hat{r}) = \frac{\hat{r}(1-\hat{r})}{n}$$

فرض وجود پیوستگی ممکن است به صورت زیر فرموله شود:

$$H_0 : r = 0.5 \text{ در مقابل } H_1 < 0.5 \qquad (۷-۴)$$

که H_0 برابر با r = ۰/۵ به معنی عدم وجود پیوستگی معنی‌دار است.

آماره آزمون برای سنجش فرض (۴-۷) به‌صورت نسبت لگاریتم نسبت درستنمایی (LR) مدل کامل بر مدل کاهش یافته محاسبه می‌شود:

$$LR = -2\log\left[\frac{L(\hat{r}\,|\,n)}{L(r = 0.5\,|\,n)}\right]$$

آماره آزمون LR می‌تواند به صورتی مشاهده شود که از لحاظ مجانبی دارای توزیع کای مربع با یک درجه آزادی است.

مثال ۴-۱. یک خانواده تنی حاصل از دو والد غیر اینبرد به ترتیب با دیپلوتیپ‌های شناخته شده $[11][22]$ و $[33][44]$ را در نظر بگیرید. دو نشانگر حاوی اطلاعات کامل A و B برای هر فرد تنی تعیین ژنوتیپ شدند. تعداد مشاهدات مربوط به هر یک از ۱۶ ژنوتیپ دو نشانگری به‌صورت زیر بودند:

$$n = \begin{array}{c} \\ ۱۳ \\ ۱۴ \\ ۲۳ \\ ۲۴ \end{array} \begin{array}{c} \begin{array}{cccc} ۱۳ & ۱۴ & ۲۳ & ۲۴ \end{array} \\ \begin{bmatrix} n_{11}=24 & n_{12}=8 & n_{13}=7 & n_{14}=1 \\ n_{21}=11 & n_{22}=36 & n_{23}=4 & n_{24}=8 \\ n_{31}=7 & n_{32}=0 & n_{33}=35 & n_{34}=2 \\ n_{41}=2 & n_{42}=6 & n_{43}=12 & n_{44}=37 \end{bmatrix} \end{array}$$

بر اساس فراوانی مورد انتظار هر سلول بر حسب نرخ نوترکیبی (r) بین دو نشانگر، می‌توانیم تابع درستنمایی را به‌صورتی که با معادله (۴-۵) توصیف شد، فرموله نماییم. محاسبه برآوردگر حداکثر درستنمایی r به‌صورت زیر دشوار نخواهد بود:

$$\begin{aligned} \hat{r} &= \frac{2n_2+n_3+n_4}{2n} \\ &= \frac{2\times 7+33+28}{200\times 2} = 0.1875 \end{aligned} \qquad (۴-۸)$$

که در آن، n_2 = ۱+۴+۰+۲ = ۷، n_3 = ۸+۱۱+۲+۱۲ = ۳۳ و n_4 = ۷+۷+۸+۶ = ۲۸ با معادله (۴-۶) تعریف شده‌اند.

آماره آزمون LR برابر است با:

$$LR = -2\left[\log\left(L\left(\hat{r}\,|\,n\right)\right) - \log\left(L(r=0.5\,|\,n)\right)\right] = -2(-193.0+277.3) = 168.6$$

که بزرگ‌تر از $x^2_{0.05}(1) = 3.84$ است؛ بنابراین، فرض صفر بالا رد می‌شود و به این مفهوم است که پیوستگی بین دو نشانگر معنی‌دار است.

از روی رابطه بالا می‌توان مشاهده کرد که برآورد r برای دو نشانگر حاوی اطلاعات کامل می‌تواند فقط در طی یک مرحله به‌دست آید و اصلاً نیازی به الگوریتم تکراری EM ندارد. استفاده از الگوریتم EM در بخش زیر، با استفاده از معادلات (۴-۱۲) و (۴-۱۳)، به‌منظور به‌دست آوردن چارچوب کلی برای برآورد پیوستگی بین انواع نشانگرها، شامل نشانگرهای با اطلاعات کامل توصیف می‌شود.

۴-۲-۲- یک فرمول کلی‌تر

G_{j_1} و G_{j_2} $(j_1, j_2 = 1,2,3,4)$ را به ترتیب به‌عنوان چهار ژنوتیپ نتاج برای نشانگرهای A و B در ترتیب داده شده در ماتریس‌های H و D بالا در نظر بگیرید. با فرض اینکه n نتاج موجود در خانواده تنی از هم مستقل باشند، مجدداً می‌توانیم درستنمایی داده‌های نشانگر n را که با معادله (۴-۵) ارائه شد، تحت ترکیب دیپلوتیپی والدینی که در رابطه (۴-۱) نشان داده شده است را به‌صورت زیر بنویسیم:

$$L(r \mid n) = \prod_{i=1}^{n} L_i(r \mid n) = \prod_{i=1}^{n} \sum_{j_1=1}^{4} \sum_{j_2=1}^{4} x_{ij_1} P\left(G_{j_1} G_{j_2}\right) x_{ij_2} \qquad (۹-۴)$$

در این رابطه، x_{ij_k} متغیر شاخص و موهومی توصیف‌کننده ژنوتیپ j_kام نشانگر M^k برای نتاج i است که اگر ژنوتیپ مشاهده شده نشانگر با G_{j_k} هماهنگ و یکجور باشد، مقدار عددی آن برابر با ۱ و در غیر این‌صورت برابر با ۰ است، $P\left(G_{j_1} G_{j_2}\right)$ نیز احتمال توأم j_1امین ژنوتیپ نشانگر A و j_2امین ژنوتیپ نشانگر B (همانند ماتریس (۴-۲)) است. معادله (۴-۹) می‌تواند به شکل ماتریسی زیر نوشته شود:

$$L(r \mid n) = \prod_{i=1}^{n} m_{ij_1}^T H m_{ij_2} \qquad (۱۰-۴)$$

که m_{ij_k} بردار چهار بعدی متغیر موهومی x_{ij_k} برای نشانگر $\mathbf{M^k}$ است.

تابع درستنمایی فنوتیپ‌های مشاهده شده نشانگر (**n**) که در معادله (۴-۱۰) ارائه شده است، بر اساس ماتریس نرخ نوترکیبی (**H**) حاصل از ژنوتیپ‌های قابل تشخیص برای نشانگرهای حاوی اطلاعات کامل ایجاد می‌شود. می‌توانیم یک ماتریس طرح (I) تعریف کنیم که ژنوتیپ‌های نشانگر (**H**) را به ماتریس فنوتیپی نشانگر (**P**) مرتبط سازد؛ بنابراین، داریم:

$$L(r \mid n) = \prod_{i=1}^{n} m_{ij_1}^T \left(I_{b_1}^T H I_{b_2}\right) m_{ij_2} = \prod_{i=1}^{N} m_{ij_1}^T P m_{ij_2} \qquad (۱۱-۴)$$

که در این رابطه، b_k تعداد ژنوتیپ‌های (فنوتیپ ها) قابل تشخیص در نتاج برای نشانگر $\mathbf{M^k}$ است که در مورد یک نشانگر حاوی اطلاعات کامل برابر ۴ است، $P = I_{b_1}^T H I_{b_2}$ ماتریس

$(b_1 \times b_2)$ احتمالات فنوتیپی توأم دو نشانگر و I_{b_k} ماتریس طرح $(4 \times b_k)$ است که برای تخصیص الگوی تفرق یک نوع نشانگر با در نظر گرفتن ترکیب والدینی معین طرح ریزی شده است. برای نشانگرهای حاوی اطلاعات کامل داریم:

$$I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

البته این ماتریس تحت شرایطی است که ترکیب دیپلوتیپی والدینی (۴-۱) در نظر گرفته شده باشد. همان‌گونه که در بخش پیشین مشاهده شد، ماتریس‌های طرح مشابه نیز می‌توانند برای انواع تلاقی نشانگری دیگر (جدول ۳-۲) و ترکیبات دیپلوتیپی دیگر طراحی شوند.

وو و همکاران (b ۲۰۰۲) یک روش کلی برای برآورد نرخ نوترکیبی بین هر نوع نشانگر از راه حداکثر نمودن لگاریتم تابع درستنمایی معادله (۴-۱۱) ارایه دادند. این روش به همراه الگوریتم EM با استفاده از رویه زیر ارائه می‌شود.

مرحله E: در مرحله τ، با استفاده از ماتریس H و بر اساس برآورد فعلی $r^{(\tau)}$، تعداد مورد انتظار پیشامدهای نوترکیبی بین نشانگرهای A و B برای نتاج i با در نظر گرفتن ترکیب دیپلوتیپی والدینی محاسبه می‌شود:

$$D_{ij_1j_2}^{(\tau+1)} = \frac{m_{ij_1}^T \left[I_{b_1}^T \left(H \circ D \right) I_{b_2} \right] m_{ij_2}}{m_{ij_1}^T P m_{ij_2}} \qquad (۱۲-۴)$$

که $\circ$ نشان‌دهنده حاصل‌ضرب اجزای نظیر به نظیر دو ماتریس است.

مرحله M: $r^{(\tau+1)}$ را با در نظر گرفتن ترکیب مورد نظر دیپلوتیپ والدینی با استفاده از معادله زیر محاسبه کنید:

$$r^{(\tau+1)} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j_1=1}^{b_1} \sum_{j_2}^{b_2} D_{ij_1j_2}^{(\tau+1)} \qquad (۱۳-۴)$$

این مراحل بین معادلات (۴-۱۲) و (۴-۱۳) تا جایی که r در یک مقدار ثابت به همگرایی برسد، تکرار می‌شوند. این مقدار ثابت و پایدار نشان‌دهنده برآورد حداکثر درستنمایی (MLE) نرخ نوترکیبی بین نشانگر A و B با در نظر گرفتن ترکیب مورد نظر دیپلوتیپ والدینی است.

برای هر جفت نشانگر، ترکیبات دیپلوتیپ والدینی چندگانه‌ای باتوجه به هر یک از نرخ‌های نوترکیبی برآورد شده در نظر گرفته شده و برای هر یک از آن‌ها مقدار درستنمایی مربوطه محاسبه می‌شود. محتمل‌ترین ترکیب دیپلوتیپی می‌تواند بر اساس حداقل نرخ نوترکیبی و حداکثر مقدار درستنمایی به‌دست آید.

۴-۲-۳- تجزیه سه نقطه‌ای

الگوریتم‌های آماری برای برآورد نرخ نوترکیبی بر پایه تجزیه دو نقطه‌ای، به‌ویژه برای نشانگرهای حاوی اطلاعات ناقص (نسبی) ممکن است توانمند نباشند. ریدوت و همکاران (۱۹۹۸) مثالی را شرح دادند که در آن، تجزیه سه نقطه‌ای نسبت به تجزیه دونقطه‌ای می‌تواند روابط پیوستگی بیشتری را بین سه جایگاه ژنی شناسایی کند.

سه نشانگر با ترتیب **A-B-C** در نظر بگیرید. نسبت به نشانگر A، نشانگر B دو حالت ممکن برای انتساب آلل‌های آن به دو کروموزوم همولوگ دارد. به‌طور مشابه، دو آرایش آللی برای نشانگر C وقتی که نشانگر B ثابت است، وجود دارد؛ بنابراین، برای یک والد، ۴ = ۲×۲ حالت دیپلوتیپی ممکن وجود دارد. فرض کنیم که دیپلوتیپ‌های دو والد شناخته شده بوده و به‌صورت زیر باشند:

$$\Phi = \begin{array}{c} \\ A1 \\ B1 \\ C1 \end{array} \begin{array}{c} \mathrm{P} \times \mathrm{Q} \\ \left\| \begin{matrix} 2 & 3 \\ 2 & 3 \\ 2 & 3 \end{matrix} \right\| \end{array} \begin{array}{c} \\ 4 \\ 4 \\ 4 \end{array} \quad \text{یا} \quad \begin{array}{c} \mathrm{P} \times \mathrm{Q} \\ [111][222]\times[333][444] \end{array} \qquad (۱۴-۴)$$

r_{AB}، r_{BC} و r_{AC} به ترتیب نرخ‌های نوترکیبی بین نشانگرهای A و B، بین نشانگرهای B و C و بین نشانگرهای A و C هستند. این نرخ‌های نوترکیبی، احتمالات مربوط به حالتی هستند که یک کراس‌اور بین نشانگرهای A و B و بین نشانگرهای B و C رخ دهد. اگر یک کراس‌اور در هر فاصله رخ دهد، این پیشامد با G_{11} و اگر هیچ کراس‌اوری در هر فاصله رخ ندهد، این پیشامد با G_{00} نشان داده می‌شود، حال آنکه این پیشامد در زمانی که یک کراس‌اور فقط در فاصله اول یا فقط در فاصله دوم رخ دهد، به ترتیب با G_{10} و G_{01} نشان داده می شود. احتمالات این پیشامدها به ترتیب با g_{00}، g_{01}، g_{10} و g_{11} نشان داده می‌شوند که مجموع آن‌ها برابر با ۱ است. بر طبق تعریف نرخ نوترکیبی، به‌صورت احتمال وقوع کراس‌اور بین یک جفت ژن، واضح است که:

$$\begin{aligned} r_{AB} &= g_{10} + g_{11} \\ r_{BC} &= g_{01} + g_{11} \\ r_{AC} &= g_{01} + g_{10} \end{aligned} \qquad (۱۵-۴)$$

و

$$
\begin{aligned}
g_{11} &= \frac{1}{2}\left(r_{AB} + r_{BC} - r_{AC}\right) \\
g_{10} &= \frac{1}{2}\left(r_{AB} + r_{AC} - r_{BC}\right) \\
g_{01} &= \frac{1}{2}\left(r_{BC} + r_{AC} - r_{AB}\right) \\
g_{00} &= 1 - \frac{1}{2}\left(r_{AB} + r_{AC} + r_{BC}\right)
\end{aligned}
\qquad (۱۶-۴)
$$

تلاقی بین دو والد که به صورت شماتیک در رابطه (۴-۱۴) نشان داده شده است، مجموعاً سبب ایجاد ۶۴ = ۴×۴×۴ ژنوتیپ برای سه نشانگر حاوی اطلاعات کامل می‌شود. فراوانی‌های هر ژنوتیپ می‌تواند بر حسب g_{00}، g_{01}، g_{10} و g_{11} (جدول ۴-۱ را مشاهده کنید) و بنابراین r_{AB}، r_{BC} و r_{AC} به وسیله معادله ۴-۱۶ بیان شود. برای هر ژنوتیپ، تعداد کراس‌اورهایی که بین هر جفت نشانگر مجاور هم رخ می‌دهند، در قالب جدول ۴-۲ نشان داده شده‌اند که از روی آن، اشکال نزدیک به هم می‌توانند برای برآورد g ها به‌دست آیند. در مثال بعد، روشی برای برآورد g ها و بنابراین نرخ‌های نوترکیبی با استفاده از معادله (۴-۱۵) ارائه می‌شود.

جدول ۴-۱- ماتریس احتمال توأم (**H**) بین سه نشانگر بر حسب تعداد کراس‌اورهای رخ داده بین A و B و همچنین بین B و C تحت ترکیب خاص دیپلوتیپ والدینی که با رابطه (۴-۱۴) نشان داده شد.

نشانگر **A**	نشانگر **B**	نشانگر **C**			
		13	14	23	24
13	13	g_{00}^2	$g_{00}g_{01}$	$g_{01}g_{00}$	g_{01}^2
13	14	$g_{00}g_{11}$	$g_{00}g_{10}$	$g_{01}g_{11}$	$g_{01}g_{10}$
13	23	$g_{11}g_{00}$	$g_{11}g_{01}$	$g_{10}g_{00}$	$g_{10}g_{01}$
13	24	g_{11}^2	$g_{11}g_{10}$	$g_{10}g_{11}$	g_{10}^2
14	13	$g_{00}g_{10}$	$g_{00}g_{11}$	$g_{01}g_{10}$	$g_{01}g_{11}$
14	14	$g_{00}g_{01}$	g_{00}^2	g_{01}^2	$g_{01}g_{00}$
14	23	$g_{11}g_{10}$	g_{11}^2	g_{10}^2	$g_{10}g_{11}$
14	24	$g_{11}g_{01}$	$g_{11}g_{00}$	$g_{10}g_{01}$	$g_{10}g_{00}$
23	13	$g_{10}g_{00}$	$g_{10}g_{01}$	$g_{11}g_{00}$	$g_{11}g_{01}$
23	14	$g_{10}g_{11}$	g_{10}^2	g_{11}^2	$g_{11}g_{10}$
23	23	$g_{01}g_{00}$	g_{01}^2	g_{00}^2	$g_{00}g_{01}$
23	24	$g_{01}g_{11}$	$g_{01}g_{10}$	$g_{00}g_{11}$	$g_{00}g_{10}$
24	13	g_{10}^2	$g_{10}g_{11}$	$g_{11}g_{10}$	g_{11}^2
24	14	$g_{10}g_{01}$	$g_{10}g_{00}$	$g_{11}g_{01}$	$g_{11}g_{00}$
24	23	$g_{01}g_{10}$	$g_{01}g_{11}$	$g_{00}g_{10}$	$g_{00}g_{11}$
24	24	g_{01}^2	$g_{01}g_{00}$	$g_{00}g_{01}$	g_{00}^2

مثال ۴-۲. سه نشانگر حاوی اطلاعات کامل به‌صورت A، B و C برای خانواده تنی حاصل از دو والد دگربارور به ترتیب با دیپلوتیپ‌های شناخته شده $[222][111]$ و $[444][333]$ تعیین ژنوتیپ شده‌اند. مشاهدات مربوط به ۶۴ ژنوتیپ سه نشانگر در جدول ۴-۳ ارایه شده است.

بر اساس فراوانی مورد انتظار هر سلول بر حسب احتمالات کراس‌اورهای بین سه نشانگر (g ها) که در جدول ۴-۱ نشان داده شده‌اند، می‌توانیم یک تابع درستنمایی که در ادامه با معادله (۴-۲۱)

ارائه می‌شود را فرموله کنیم. همچنین می‌توانیم برآوردگر حداکثر درستنمایی مربوط به g ها را به‌صورت زیر به‌دست آوریم:

$$\hat{g}_{00} = \frac{2\left(n_{111}+n_{222}+n_{333}+n_{444}\right)+n_1}{2n} = \frac{2\times 225+147}{2\times 400} = 0.7462 \qquad (۱۷-۴)$$

که در آن:

$$n_1 = n_{121}+n_{131}+n_{211}+n_{221}+n_{311}+n_{331}+n_{112}+n_{122}+n_{212}+n_{242}+n_{422}+ n_{422}+n_{113}+n_{133}+n_{313}+n_{343}+n_{433}+n_{443}+n_{224}+n_{244}+n_{334}+n_{344}+n_{424}+n_{434}$$

جدول ۴-۲- ماتریس‌های (G_{00}، G_{01}، G_{10} و G_{11}) مربوط به پیشامدهای کراس‌اور مختص هر فاصله در بین سه نشانگر A، B و C تحت ترکیب خاص دیپلوتیپ والدینی ارایه شده در رابطه (۴-۱۴).

نشانگر **A**	نشانگر **B**	نشانگر **C**															
		G_{00}				G_{01}				G_{10}				G_{11}			
		13				14				23				24			
13	13	2	1	1	0	0	1	1	2	0	0	0	0	0	0	0	0
13	14	1	1	0	0	0	0	1	1	0	1	0	1	1	0	1	0
13	23	1	0	1	1	0	1	0	1	0	0	1	1	1	1	0	0
13	24	0	0	0	0	0	0	0	0	0	1	1	2	2	1	1	0
14	13	1	1	0	0	0	0	1	1	1	0	1	0	0	1	0	1
14	14	1	2	0	1	1	0	2	1	0	0	0	0	0	0	0	0
14	23	0	0	0	0	0	0	0	0	1	0	2	1	1	2	0	1
14	24	0	1	0	1	1	0	1	0	0	0	1	1	1	1	0	0
23	13	1	0	1	0	0	1	0	1	1	1	0	0	0	0	1	1
23	14	0	0	0	0	0	0	0	0	1	2	0	1	1	0	2	1
23	23	1	0	2	1	1	2	0	1	0	0	0	0	0	0	0	0
23	24	0	0	1	1	1	1	0	0	0	1	0	1	1	0	1	0
24	13	0	0	0	0	0	0	0	0	2	1	1	0	0	1	1	2
24	14	0	1	0	1	1	0	1	0	1	1	0	0	0	0	1	1
24	23	0	0	1	1	1	1	0	0	1	0	1	0	0	1	0	1
24	24	0	1	1	2	2	1	1	0	0	0	0	0	0	0	0	0

$$\hat{g}_{01} = \frac{2(n_{441} + n_{332} + n_{223} + n_{114}) + n_2}{2n} = \frac{2\times5+68}{2\times400} = 0.0975 \quad (۱۸-۴)$$

که در آن:

$$n_2 = n_{221} + n_{241} + n_{331} + n_{341} + n_{421} + n_{431} + n_{112} + n_{132} + n_{312} + n_{342} + n_{432} + n_{442} + n_{113} + n_{123} + n_{213} + n_{243} + n_{423} + n_{443} + n_{124} + n_{134} + n_{214} + n_{224} + n_{314} + n_{334}$$

$$\hat{g}_{10} = \frac{2(n_{441} + n_{332} + n_{223} + n_{114}) + n_3}{2n} = \frac{2\times5+101}{2\times400} = 0.1388 \quad (۱۹-۴)$$

که:

$$n_3 = n_{211} + n_{231} + n_{311} + n_{321} + n_{421} + n_{431} + n_{122} + n_{142} + n_{312} + n_{342} + n_{412} + n_{442} + n_{133} + n_{143} + n_{213} + n_{243} + n_{413} + n_{433} + n_{124} + n_{134} + n_{234} + n_{244} + n_{324} + n_{344}$$

$$\hat{g}_{11} = \frac{2(n_{441} + n_{332} + n_{223} + n_{114}) + n_4}{2n} = \frac{2+14}{2\times400} = 0.0175 \quad (۲۰-۴)$$

که n_4 برابر است با:

$$n_4 = n_{121} + n_{131} + n_{231} + n_{241} + n_{321} + n_{341} + n_{132} + n_{142} + n_{212} + n_{242} + n_{412} + n_{432} + n_{123} + n_{143} + n_{313} + n_{343} + n_{413} + n_{423} + n_{214} + n_{234} + n_{314} + n_{324} + n_{424} + n_{434}$$

از این‌رو نرخ‌های نوترکیبی به صورت زیر برآورد می‌شوند:

$$\hat{r}_{AB} = 0.1563 \qquad \hat{r}_{BC} = 0.1150 \qquad \hat{r}_{AC} = 0.2363$$

از آنجایی که برآوردهای مربوط به g برای سه نشانگر حاوی اطلاعات کامل فقط با استفاده از یک مرحله به‌دست می‌آیند، نیازی به الگوریتم EM تکراری نیست. با معادلات (۴-۲۲) و (۴-۲۳) زیر، یک چارچوب EM کلی برای برآورد پیوستگی بین هر نوع از نشانگرها، از جمله نشانگرهای حاوی اطلاعات جزیی به‌دست می‌آوریم.

جدول ۴-۳- مشاهدات مربوط به ژنوتیپ‌های سه نشانگری در یک خانواده تنی حاصل از رابطه (۴-۱۴).

نشانگر **A**	نشانگر **B**	نشانگر **C**			
		۱۳	۱۴	۲۳	۲۴
۱۳	۱۳	n_{111} = ۴۹	n_{112} = ۳	n_{113} = ۷	n_{114} = ۱
۱۳	۱۴	n_{121} = ۲	n_{122} = ۶	n_{123} = ۰	n_{124} = ۱
۱۳	۲۳	n_{131} = ۳	n_{132} = ۱	n_{133} = ۱۵	n_{134} = ۱
۱۳	۲۴	n_{141} = ۰	n_{142} = ۱	n_{143} = ۰	n_{144} = ۳
۱۴	۱۳	n_{211} = ۱۷	n_{212} = ۲	n_{213} = ۳	n_{214} = ۰
۱۴	۱۴	n_{221} = ۶	n_{222} = ۶۴	n_{223} = ۴	n_{224} = ۹
۱۴	۲۳	n_{231} = ۰	n_{232} = ۰	n_{233} = ۰	n_{234} = ۰
۱۴	۲۴	n_{241} = ۰	n_{242} = ۰	n_{243} = ۲	n_{244} = ۱۱
۲۳	۱۳	n_{311} = ۱۳	n_{312} = ۱	n_{313} = ۰	n_{314} = ۱
۲۳	۱۴	n_{321} = ۱	n_{322} = ۱	n_{323} = ۰	n_{324} = ۰
۲۳	۲۳	n_{331} = ۱۱	n_{332} = ۰	n_{333} = ۵۳	n_{334} = ۵
۲۳	۲۴	n_{341} = ۰	n_{342} = ۲	n_{343} = ۱	n_{344} = ۱۱
۲۴	۱۳	n_{411} = ۱	n_{412} = ۰	n_{413} = ۱	n_{414} = ۰
۲۴	۱۴	n_{421} = ۲	n_{422} = ۴	n_{423} = ۰	n_{424} = ۱
۲۴	۲۳	n_{431} = ۱	n_{432} = ۰	n_{433} = ۸	n_{434} = ۲
۲۴	۲۴	n_{441} = ۰	n_{442} = ۶	n_{443} = ۶	n_{444} = ۵۹

۴-۲-۴- ارائهٔ فرمول کلی‌تر

همانند تجزیه دو نقطه‌ای، درستنمایی احتمال کراس‌اور، $g=(g_{00},g_{01},g_{10},g_{11})$، برای داده‌های سه نشانگر تحت ترکیب دیپلوتیپ والدینی ارائه شده در معادله (۴-۱۴) به‌صورت زیر بیان می‌شود:

$$L(g\mid n)=\prod_{i=1}^{n}\left(m_{ij_1}^{T}\otimes m_{ij_2}^{T}\right)\left(I_{p_1}^{T}\otimes I_{p_2}^{T}\right)\left(HI_{p_3}\right)m_{ij_3} \qquad (۴-۲۱)$$

که در آن، j_1, j_2, j_3 = ۱ برای رشته ۱۳، ۲ برای رشته ۱۴، ۳ برای رشته ۲۳ و ۴ برای رشته ۲۴ به ترتیب برای ژنوتیپ‌های نشانگرهای A، B و C هستند، $\otimes$ نشان‌دهنده حاصل‌ضرب

کرونکر[۱] است. بردارها و ماتریس‌های این رابطه همانند آنچه هستند که در تجزیه دو نقطه ای تعریف شدند. ماتریس H یعنی ماتریس احتمال توأم ژنوتیپ‌های سه نشانگری در جدول ۴-۱ ارائه شده است که به صورت $H=\left(H_{j_1j_2j_3}\right)_{16\times4}$ است، در حالی که جدول ۴-۲ ماتریس‌های تعداد کراس‌اورها یعنی (G_{00}، G_{01}، G_{10} و G_{11}) را برای دو فاصله A-B و B-C تحت ترکیب دیپلوتیپ والدینی ارائه شده در رابطه (۴-۱۴) نشان می‌دهد.

برای به‌دست آوردن برآوردهای حداکثر درستنمایی (MLE) احتمالات کراس‌اور (و بنابراین نرخ های نوترکیبی) بین سه نشانگر از الگوریتم EM استفاده می‌شود. معادلات کلی تکرار $(\tau+1)$ ام مرحله EM به‌صورت زیر فرموله می‌شوند:

مرحله E: تعداد مورد انتظار کراس‌اورهای مربوط به هر فاصله مرتبط با $G_{00}=\left(G^{00}_{j_1j_2j_3}\right)_{16\times4}$، $G_{01}=\left(G^{01}_{j_1j_2j_3}\right)_{16\times4}$، $G_{10}=\left(G^{10}_{j_1j_2j_3}\right)_{16\times4}$ و $G_{11}=\left(G^{11}_{j_1j_2j_3}\right)_{16\times4}$ را برای نتاج i تحت ترکیب دیپلوتیپ والدینی مورد نظر را به ترتیب به صورت زیر محاسبه کنید:

$$
\begin{aligned}
G^{00(\tau+1)}_{ij_1j_2j_3} &= \frac{\left[m^T_{ij_1}\otimes m^T_{ij_2}\right]\left[I^T_{b_1}\otimes I^T_{b_2}\right]\left[\left(G^{(\tau)}_{00}\circ H^{(\tau)}I_{b_3}\right)\right]m_{ij_3}}{\left[m^T_{ij_1}\otimes m^T_{ij_2}\right]\left[I^T_{b_1}\otimes I^T_{b_2}\right]\left(H^{(\tau)}I_{b_3}\right)m_{ij_3}}\\
G^{01(\tau+1)}_{ij_1j_2j_3} &= \frac{\left[m^T_{ij_1}\otimes m^T_{ij_2}\right]\left[I^T_{b_1}\otimes I^T_{b_2}\right]\left[\left(G^{(\tau)}_{01}\circ H^{(\tau)}I_{b_3}\right)\right]m_{ij_3}}{\left[m^T_{ij_1}\otimes m^T_{ij_2}\right]\left[I^T_{b_1}\otimes I^T_{b_2}\right]\left(H^{(\tau)}I_{b_3}\right)m_{ij_3}}\\
G^{10(\tau+1)}_{ij_1j_2j_3} &= \frac{\left[m^T_{ij_1}\otimes m^T_{ij_2}\right]\left[I^T_{b_1}\otimes I^T_{b_2}\right]\left[\left(G^{(\tau)}_{10}\circ H^{(\tau)}I_{b_3}\right)\right]m_{ij_3}}{\left[m^T_{ij_1}\otimes m^T_{ij_2}\right]\left[I^T_{b_1}\otimes I^T_{b_2}\right]\left(H^{(\tau)}I_{b_3}\right)m_{ij_3}}\\
G^{11(\tau+1)}_{ij_1j_2j_3} &= \frac{\left[m^T_{ij_1}\otimes m^T_{ij_2}\right]\left[I^T_{b_1}\otimes I^T_{b_2}\right]\left[\left(G^{(\tau)}_{11}\circ H^{(\tau)}I_{b_3}\right)\right]m_{ij_3}}{\left[m^T_{ij_1}\otimes m^T_{ij_2}\right]\left[I^T_{b_1}\otimes I^T_{b_2}\right]\left(H^{(\tau)}I_{b_3}\right)m_{ij_3}}
\end{aligned}
\qquad (۴-۲۲)
$$

---

[۱]- حاصل‌ضرب کرونکر (Kronecker product) نوعی حاصل‌ضرب در جبر ماتریس است که در آن، هر جزء ماتریس اول در تمامی اجزای ماتریس دوم ضرب می‌شود (مترجمین).

مرحله M: احتمالات کراس‌اورهای مربوط به هر فاصله یعنی g_{00}، g_{01}، g_{10} و g_{11} را با استفاده از روابط زیر محاسبه کنید:

$$
\begin{aligned}
g_{00}^{(\tau+1)} &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j_1=1}^{b_1}\sum_{j_2=1}^{b_2}\sum_{j_3=1}^{b_3}\varsigma_{ij_1j_2j_3}^{00(\tau+1)} \\
g_{01}^{(\tau+1)} &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j_1=1}^{b_1}\sum_{j_2=1}^{b_2}\sum_{j_3=1}^{b_3}\varsigma_{ij_1j_2j_3}^{01(\tau+1)} \\
g_{10}^{(\tau+1)} &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j_1=1}^{b_1}\sum_{j_2=1}^{b_2}\sum_{j_3=1}^{b_3}\varsigma_{ij_1j_2j_3}^{10(\tau+1)} \\
g_{11}^{(\tau+1)} &= \frac{1}{2n}\sum_{i=1}^{n}\sum_{j_1=1}^{b_1}\sum_{j_2=1}^{b_2}\sum_{j_3=1}^{b_3}\varsigma_{ij_1j_2j_3}^{11(\tau+1)}
\end{aligned}
\qquad (۲۳-۴)
$$

مراحل E و M بر اساس معادلات (۴-۲۲) و (۴-۲۳) تکرار می‌شوند تا وقتی که این احتمالات در مقادیر ثابت و پایداری به همگرایی برسند. MLE های مربوط به g ها می‌توانند تبدیل شوند تا MLE های نرخ‌های نوترکیبی یعنی r_{AB}، r_{BC} و r_{AC} به‌دست آیند، چون MLE ها با تبدیل پارامتری، ثابت هستند. از آنجایی که تمامی نرخ‌های نوترکیبی ممکن در بین سه نشانگر برآورد می‌شوند، از این‌رو تجزیه سه نقطه‌ای اطلاعات مهمی در خصوص ترتیب نشانگرها فراهم می‌آورد. برای مثال، اگر r_{AB} یا r_{BC} با فرض اینکه ترتیب نشانگرها A-B-C باشد، بزرگتر از r_{AC} برآورد شوند، در این صورت ترتیب نشانگری بالا بر اساس قضیه ۳- ۱ احتمالاً اشتباه است. وقتی که هر سه نشانگر مجاور هم در یک گروه پیوستگی در معرض تجزیه سه نقطه‌ای قرار گیرند و ترتیب نشانگرها تأیید شود، می‌توانیم دو برآورد متفاوت از نرخ نوترکیبی برای یک فاصله نشانگری مشابه به‌دست آوریم (به استثنای فواصل مربوط به دو انتهای یک گروه پیوستگی). بهترین راه برای ترکیب این برآوردها به‌دست آوردن یک میانگین وزنی است که در آن، وزن‌ها معکوس واریانس بین دو برآورد جداگانه هستند (ریدوت و همکاران، ۱۹۹۸).

محتمل‌ترین ترکیب دیپلوتیپ والدینی در بین سه نشانگر با انتخاب بیشینه مقادیر درستنمایی تحت تمامی ترکیبات دیپلوتیپی ممکن تعیین می‌شود. فقط با وجود محتمل‌ترین ترکیب دیپلوتیپی، MLE های نرخ‌های نوترکیبی بهینه می‌شوند.

۴-۳- نشانگرهای حاوی اطلاعات کامل: مدل ژنوتیپی

در بخش ۴-۲، روش کلی تجزیه پیوستگی نشانگرهای حاوی اطلاعات کامل را در خانواده‌های تنی حاصل از دو والد تعیین فاز شده ارائه دادیم. زمانی که دو والد آمیزش کننده دارای دیپلوتیپ‌های شناخته شده باشند، الگوهای تفرق همزمان نشانگر در نتاج آن‌ها می‌تواند پیش‌بینی شود. در عمل، دیپلوتیپ‌های والدینی نمی‌توانند به‌طور مستقیم مشاهده شوند، بلکه فقط ژنوتیپ‌ها می‌توانند مشاهده شوند، به‌طوری که به ایجاد مدل‌های آماری برای برآورد پیوستگی بر اساس ژنوتیپ‌های مشاهده شده نیاز داریم. یک راهکار برای تجزیه پیوستگی بر اساس داده‌های ژنوتیپی، محاسبه مقادیر درستنمایی تحت تمامی دیپلوتیپ‌های والدینی ممکن و برآورد نرخ نوترکیبی تحت محتمل‌ترین ترکیب دیپلوتیپی است. در این بخش، یک مدل توأم (ترکیبی) که احتمالات دیپلوتیپی والدینی را با یکدیگر ترکیب می‌کند، معرفی می‌شود.

۴-۳-۱- دیپلوتیپ‌های والدینی

دو نشانگر با ترتیب A-B را در نظر بگیرید. هر والد، P یا Q، دو دیپلوتیپ متفاوت بین دو نشانگر دارد؛ بنابراین، وقتی دو والد آمیزش می‌کنند، چهار ترکیب دیپلوتیپی (Φ) وجود دارند که به‌صورت زیر بیان می‌شوند:

$$\begin{array}{c} \mathrm{P} \times \mathrm{Q} \\ \left.\begin{array}{l} \Phi_{11} = \begin{array}{cc} A & 1 \\ B & 1 \end{array} \left\| \begin{array}{c} 2 \\ 2 \end{array} \times \begin{array}{c} 3 \\ 3 \end{array} \right\| \begin{array}{c} 4 \\ 4 \end{array} \\ \Phi_{12} = \begin{array}{cc} A & 1 \\ B & 1 \end{array} \left\| \begin{array}{c} 2 \\ 2 \end{array} \times \begin{array}{c} 3 \\ 4 \end{array} \right\| \begin{array}{c} 4 \\ 3 \end{array} \\ \Phi_{21} = \begin{array}{cc} A & 1 \\ B & 2 \end{array} \left\| \begin{array}{c} 2 \\ 1 \end{array} \times \begin{array}{c} 3 \\ 3 \end{array} \right\| \begin{array}{c} 4 \\ 4 \end{array} \\ \Phi_{22} = \begin{array}{cc} A & 1 \\ B & 2 \end{array} \left\| \begin{array}{c} 2 \\ 1 \end{array} \times \begin{array}{c} 3 \\ 4 \end{array} \right\| \begin{array}{c} 4 \\ 3 \end{array} \end{array}\right\} \end{array} \qquad (۲۴-۴)$$

که اولین (ξ) و دومین (ζ) اندیس Φ به ترتیب نشان‌دهنده دو دیپلوتیپ ممکن والد P و Q هستند. هر یک از این ترکیبات دیپلوتیپی الگوی متفاوتی از تفرق همزمان نشانگری را ایجاد می‌کنند که به‌صورت ماتریسی و به همراه نرخ نوترکیبی r بدین صورت بیان می‌شود:

$$H_{11}=\begin{array}{c} \text{A/B} \\ 13 \\ 14 \\ 23 \\ 24 \end{array}\begin{array}{c} \begin{array}{cccc} 13 & 14 & 23 & 24 \end{array} \\ \begin{bmatrix} \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{r^2}{4} \\ \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} & \frac{r^2}{4} & \frac{r(1-r)}{4} \\ \frac{r(1-r)}{4} & \frac{r^2}{4} & \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} \\ \frac{r^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} \end{bmatrix} \end{array}$$

$$H_{12}=\begin{array}{c} 13 \\ 14 \\ 23 \\ 24 \end{array}\begin{bmatrix} \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} & \frac{r^2}{4} & \frac{r(1-r)}{4} \\ \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{r^2}{4} \\ \frac{r^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} \\ \frac{r(1-r)}{4} & \frac{r^2}{4} & \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} \end{bmatrix} \qquad (۴-۲۵)$$

$$H_{21}=\begin{array}{c} \text{A/B} \\ 13 \\ 14 \\ 23 \\ 24 \end{array}\begin{array}{c} \begin{array}{cccc} 13 & 14 & 23 & 24 \end{array} \\ \begin{bmatrix} \frac{r(1-r)}{4} & \frac{r^2}{4} & \frac{r^2}{4} & \frac{r(1-r)}{4} \\ \frac{r^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{r^2}{4} \\ \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} \\ \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} & \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} \end{bmatrix} \end{array}$$

$$H_{22} = \begin{array}{c} 13 \\ 14 \\ 23 \\ 24 \end{array} \begin{bmatrix} \frac{r^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} \\ \frac{r(1-r)}{4} & \frac{r^2}{4} & \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} \\ \frac{r(1-r)}{4} & \frac{(1-r)^2}{4} & \frac{r^2}{4} & \frac{r(1-r)}{4} \\ \frac{(1-r)^2}{4} & \frac{r(1-r)}{4} & \frac{r(1-r)}{4} & \frac{r^2}{4} \end{bmatrix}$$

تعداد مورد انتظار پیشامدهای نوترکیب (یعنی تعداد r) که تحت ترکیبات دیپلوتیپی مختلف بین دو نشانگر رخ می‌دهند نیز می‌تواند به شکل ماتریسی زیر بیان شود:

$$D_{11} = \begin{array}{c|cccc} A/B & 13 & 14 & 23 & 24 \\ \hline 13 & 0 & 1 & 1 & 2 \\ 14 & 1 & 0 & 2 & 1 \\ 23 & 1 & 2 & 0 & 1 \\ 24 & 2 & 1 & 1 & 0 \end{array}$$

$$D_{12} = \begin{array}{c} 13 \\ 14 \\ 23 \\ 24 \end{array} \begin{bmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \\ 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix}$$

$$D_{21} = \begin{array}{c} 13 \\ 14 \\ 23 \\ 24 \end{array} \begin{bmatrix} 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \end{bmatrix}$$

$$D_{22} = \begin{array}{c} 13 \\ 14 \\ 23 \\ 24 \end{array} \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}$$

دو نشانگر برای خانواده تنی با مشاهدات $n = \left(n_{j_1 j_2}\right)_{4\times4}$ که با ماتریس (۴-۴) توصیف شـدند، تعیین ژنوتیپ می‌شوند. سپس درستنمایی‌های چهار ترکیب دیپلوتیپی به‌صورت زیر بیان می‌شوند:

$$\begin{aligned} L_{11}(r \mid n) &\propto r^{(2n_2+n_3+n_4)}(1-r)^{(2n_1+n_3+n_4)} \\ L_{12}(r \mid n) &\propto r^{(2n_4+n_1+n_2)}(1-r)^{(2n_3+n_1+n_2)} \\ L_{21}(r \mid n) &\propto r^{(2n_3+n_1+n_2)}(1-r)^{(2n_4+n_1+n_2)} \\ L_{22}(r \mid n) &\propto r^{(2n_1+n_3+n_4)}(1-r)^{(2n_2+n_3+n_4)} \end{aligned} \qquad (۲۶-۴)$$

که n_1، n_2، n_3 و n_4 در معادله (۵-۴) تعریف شدند.

می‌توان مشاهده کرد که MLE مربوط به r تحت ترکیب دیپلوتیپی Φ_{11} برابر بـا یـک منهـای MLE مربوط به r تحت Φ_{22} است و روابط مشابهی بین Φ_{12} و Φ_{21} وجود دارد. اگر چه مقادیر

درستنمایی دو طرفی یکسان و مشابهی بین Φ_{11} و Φ_{22} و نیز بین Φ_{12} و Φ_{21} وجود دارند، اما هنوز می‌توان MLE مناسب مربوط به r را در هر یک از این جفت‌ها انتخاب نمود، چون یکی از یک جفت MLE های مربوط به r از ۰/۵ بزرگ‌تر است. راهکارهای مرسوم برای برآورد پیوستگی و دیپلوتیپ‌های والدینی، برآورد نرخ‌های نوترکیبی و مقادیر درستنمایی تحت هر یک از چهار ترکیب و انتخاب یک تخمین صحیح از r با بالاترین مقدار درستنمایی هستند.

مثال ۴-۳. یک خانواده تنی حاصل از دو والد دگربارور را در نظر بگیرید. ژنوتیپ‌های دو نشانگر برای این دو والد به صورت 12/12 و 34/34 هستند، ولی اطلاعی در خصوص دیپلوتیپ‌های آن‌ها نداریم. دو نشانگر حاوی اطلاعات کامل A و B برای هر فرد تنی تعیین ژنوتیپ شدند که مشاهدات مربوط به مجموعاً ۱۶ ژنوتیپ دو نشانگری به‌صورت زیر است:

$$n = \begin{array}{c} \\ 13 \\ 14 \\ 23 \\ 24 \end{array} \begin{array}{c} \begin{array}{cccc} 13 & 14 & 23 & 24 \end{array} \\ \begin{bmatrix} n_{11}=3 & n_{12}=18 & n_{13}=2 & n_{14}=4 \\ n_{21}=21 & n_{22}=5 & n_{23}=4 & n_{24}=0 \\ n_{31}=1 & n_{32}=2 & n_{33}=6 & n_{34}=13 \\ n_{41}=4 & n_{42}=1 & n_{43}=13 & n_{44}=3 \end{bmatrix} \end{array}$$

بنابراین داریم:

$$n_2 = n_{14}+n_{23}+n_{32}+n_{41}=14 \quad n_1 = n_{11}+n_{22}+n_{33}+n_{44}=17$$
$$n_4 = n_{13}+n_{31}+n_{24}+n_{42}=4 \quad n_3 = n_{12}+n_{21}+n_{34}+n_{43}=65$$

برای MLE های مربوط به r، داریم:

$\hat{r}=0.4850$ تحت Φ_{11}، $L_{11}=-138.5394$

$\hat{r}=0.1950$ تحت Φ_{12}، $L_{12}=-98.6785$

$\hat{r}=0.8050$ تحت Φ_{21}، $L_{21}=-98.6785$

$\hat{r}=0.5150$ تحت Φ_{22}، $L_{22}=-138.5394$

دیپلوتیپ‌های صحیح والدینی، Φ_{12} هستند یعنی $[12][12]$ برای والد P و $[34][34]$ برای والد Q.

همانند معادله (۴-۱۱)، می‌توانیم درستنمایی را به شکل ماتریسی با در نظر گرفتن ترکیب دیپلوتیپ والدینی $\Phi_{\xi\xi}$ بنویسیم، یعنی:

$$L_{\xi\zeta}(r \mid n) = \prod_{i=1}^{n} m_{ij_1}^{T} \left(I_{p_1}^{T} H_{\xi\zeta} I_{p_2} \right) m_{ij_2}$$
$$= \prod_{i=1}^{n} m_{ij_1}^{T} P_{\xi\zeta} m_{ij_2} \qquad (۲۷-۴)$$

الگوریتم EM برای برآورد احتمالات وقوع کراس‌اور و نرخ‌های نوترکیبی بر اساس معادلات (۴-۲۲) و (۴-۲۳) توسعه و بسط داده شده‌اند.

۴-۴- مدل‌سازی توأم برای پیوستگی، دیپلوتیپ والدینی و ترتیب ژنی

الگوی تفرق هر دو نشانگر حاوی اطلاعات کامل در یک خانواده تنی به وسیله دیپلوتیپ‌های دو والد دگربارور تعیین می شود؛ بنابراین، پیوستگی بین این دو نشانگر حاصل از تفرق هم‌زمان آن‌ها، وقتی که ترکیب صحیح دیپلوتیپ والدینی تعیین می‌شود، می‌تواند برآورد گردد. مدل‌های نشان داده شده در بخش ۴-۳ اجازه برآورد هم‌زمان پیوستگی و دیپلوتیپ‌های والدینی را فراهم می‌کنند؛ اما زمانی که تعداد نشانگرهای مورد مطالعه سه نشانگر یا بیشتر باشند، مدل توأم پیوستگی و هاپلوتیپ والدینی کفایت نمی‌کند، چون ترتیب‌های ژنی مختلف نیز ممکن است استنباط و برآورد پیوستگی را تحت‌تأثیر قرار دهند. در این بخش، مدلی که پیوستگی، دیپلوتیپ والدینی و ترتیب ژنی را ترکیب می‌کند، معرفی خواهد شد.

سه نشانگر را در یک گروه پیوستگی که دارای سه ترتیب ممکن **A-B-C**، **A-C-B** و **B-A-C** هستند در نظر بگیرید. o_1، o_2 و o_3 احتمالات مربوط به وقوع این ترتیب‌ها در ژنوم والدینی هستند. بدون کنار گذاشتن کلیات، برای یک ترتیب معین، آرایش آلل‌های اولین نشانگر بین دو کروموزوم همولوگ می‌تواند برای یک والد تثبیت شود؛ بنابراین، تغییر ترتیب آلل‌ها در دو نشانگر دیگر منجر به ایجاد ۴ = ۲×۲ دیپلوتیپ والدینی خواهد شد. ژنوتیپ سه نشانگری والد P (12/12/12) ممکن است چهار دیپلوتیپ ممکن به صورت [222][111]، [221][112]، [212][121] و [211][122] داشته باشد. در ارتباط با ترتیب ثابت آللی [..2][..1] برای اولین نشانگر، احتمالات ترتیب آللی [.2.][.1.] و [.1.][.2.] برای دومین نشانگر به ترتیب به صورت p_1 و $1-p_1$ و احتمالات ترتیب آللی [..2][..1] و [..1][..2] برای سومین نشانگر به ترتیب به صورت p_2 و $1-p_2$ است. فرض کنید که ترتیب آلل‌ها بین دومین و سومین نشانگر مستقل باشند. احتمال این چهار دیپلوتیپ سه نشانگری می‌تواند به ترتیب به صورت p_1p_2، $p_1(1-p_2)$، $p_2(1-p_1)$ و $(1-p_1)(1-p_2)$ باشد. چهار دیپلوتیپ والد Q نیز می‌تواند ایجاد شوند که احتمالات آن‌ها به ترتیب به صورت q_1q_2، $q_1(1-q_2)$، $q_2(1-q_1)$ و

$(1-q_1)(1-q_2)$ تعریف می‌شوند؛ بنابراین، زمانی که والدین P و Q آمیزش می‌کنند، ۱۶ = ۴×۴ ترکیب دیپلوتیپی ممکن (که احتمالات آن‌ها برابر با حاصل‌ضرب مقادیر احتمال دیپلوتیپی مربوطه است) وجود دارند.

r_{AB} نرخ نوترکیبی بین نشانگرهای A و B و r_{BC} و r_{AC} نیز به‌طور مشابه تعریف می‌شوند. همچنین، G_{00} نشان‌دهنده عدم وجود کراس‌اور بین نشانگرهای A و B و بین نشانگرهای B و C، G_{01} فقط یک کراس‌اور در دومین فاصله، G_{10} فقط یک کراس‌اور در اولین فاصله و G_{11} یک کراس‌اور هم در فاصله اول و هم در فاصله دوم هستند. همان‌گونه که قبلاً نشان داده شد، احتمالات این پیشامدها که به ترتیب با g_{00}، g_{01}، g_{10} و g_{11} نشان داده می‌شوند، برای تعریف سه نرخ نوترکیبی مورد استفاده قرار می‌گیرند.

جدول ۴-۱ یک ماتریس (16×4) را برای فراوانی‌های ژنوتیپی سه نشانگر تحت ترکیب دیپلوتیپی خاص $[111][222]\times[333][444]$ و ترتیب ژنی **A-B-C** توصیف می‌کند. برای هر ترتیبی (O_k)، ۱۶ ترکیب دیپلوتیپی و بنابراین ۱۶ عدد از چنین ماتریسی وجود دارند که با $H_{x_1^k x_2^k y_1^k y_2^k}$ نشان داده می‌شود که در آن x_1^k برای $[11.][22.]$ برابر با یک یا برای $[12.][21.]$ برابر با دو است که دو ترتیب آللی جایگزین نشانگر B را در والد P نشان می‌دهد، x_1^k برای $[1.1][2.2]$ برابر با یک یا برای $[1.2][2.1]$ برابر با دو است و نشان‌دهنده دو ترتیب آللی جایگزین نشانگر C در والد P است، y_1^k, y_2^k برابر با ۱ یا دو بوده و معنای مشابهی برای والد Q دارد. بر اساس ریدوت و همکاران (۱۹۹۸) و وو و همکاران (۲۰۰۲ b)، عناصر $H_{x_1^k x_2^k y_1^k y_2^k}$ بر حسب g_{00}، g_{01}، g_{10} و g_{11} (جدول ۴-۱) بیان می‌شوند. به‌طور مشابه، ۱۶ ماتریس (16×4) برای تعداد مورد انتظار کراس‌اورهای ایجاد شده برای G_{00}، G_{01}، G_{10} و G_{11} و برای یک ترتیب نشانگری معین وجود دارند که به ترتیب به صورت $G^{00}_{x_1^k x_2^k y_1^k y_2^k}$، $G^{01}_{x_1^k x_2^k y_1^k y_2^k}$، $G^{10}_{x_1^k x_2^k y_1^k y_2^k}$ و $G^{11}_{x_1^k x_2^k y_1^k y_2^k}$ تعریف می‌شوند (جدول ۴-۲ را مشاهده کنید).

فراوانی‌های ژنوتیپی توأم سه نشانگر می‌تواند به‌صورت مخلوطی از ۱۶ ترکیب دیپلوتیپی و سه ترتیب نشانگری که با احتمالات وقوع خود وزن داده شده‌اند، به‌صورت زیر بیان شوند:

$$H=\sum_{k=1}^{3} o_k \sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2} p_1^{2-x_1^k}\left(1-p_1\right)^{x_1^k-1} p_2^{2-x_2^k}\left(1-p_2\right)^{x_2^k-1} q_1^{2-y_1^k}\left(1-q_1\right)^{y_1^k-1} q_2^{2-y_2^k}\left(1-q_2\right)^{y_2^k-1} H_{x_1^k x_2^k y_1^k y_2^k} \qquad (۲۸-۴)$$

به‌طور مشابه، تعداد مورد انتظار پیشامدهای نوترکیبی که در درون ژنوتیپ نتاج، مخلوطی از دیپلوتیپ‌های مختلف و ترتیب‌های متفاوت نشانگرها است و به‌صورت زیر بیان می‌شوند:

$$G_{00}=\sum_{k=1}^{3}o_k\sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2}p_1^{2-x_1^k}\left(1-p_1\right)^{x_1^k-1}p_2^{2-x_2^k}\left(1-p_2\right)^{x_2^k-1}$$
$$q_1^{2-y_1^k}\left(1-q_1\right)^{y_1^k-1}q_2^{2-y_2^k}\left(1-q_2\right)^{y_2^k-1}G^{00}_{x_1^k x_2^k y_1^k y_2^k}$$
$$G_{01}=\sum_{k=1}^{3}o_k\sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2}p_1^{2-x_1^k}\left(1-p_1\right)^{x_1^k-1}p_2^{2-x_2^k}\left(1-p_2\right)^{x_2^k-1}$$
$$q_1^{2-y_1^k}\left(1-q_1\right)^{y_1^k-1}q_2^{2-y_2^k}\left(1-q_2\right)^{y_2^k-1}G^{01}_{x_1^k x_2^k y_1^k y_2^k} \qquad (۲۹-۴)$$
$$G_{10}=\sum_{k=1}^{3}o_k\sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2}p_1^{2-x_1^k}\left(1-p_1\right)^{x_1^k-1}p_2^{2-x_2^k}\left(1-p_2\right)^{x_2^k-1}$$
$$q_1^{2-y_1^k}\left(1-q_1\right)^{y_1^k-1}q_2^{2-y_2^k}\left(1-q_2\right)^{y_2^k-1}G^{10}_{x_1^k x_2^k y_1^k y_2^k}$$
$$G_{11}=\sum_{k=1}^{3}o_k\sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2}p_1^{2-x_1^k}\left(1-p_1\right)^{x_1^k-1}p_2^{2-x_2^k}\left(1-p_2\right)^{x_2^k-1}$$
$$q_1^{2-y_1^k}\left(1-q_1\right)^{y_1^k-1}q_2^{2-y_2^k}\left(1-q_2\right)^{y_2^k-1}G^{11}_{x_1^k x_2^k y_1^k y_2^k}$$

همچنین ماتریس‌ها به صورت زیر تعریف می‌شوند:

$$P_1=\sum_{k=1}^{3}o_k\sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2}p_1\,p_2^{2-x_2^k}\left(1-p_2\right)^{x_2^k-1}$$
$$q_1^{2-y_1^k}\left(1-q_1\right)^{y_1^k-1}q_2^{2-y_2^k}\left(1-q_2\right)^{y_2^k-1}H_{x_1^k x_2^k y_1^k y_2^k}$$
$$P_2=\sum_{k=1}^{3}o_k\sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2}p_1^{2-x_2^k}\left(1-p_1\right)^{x_2^k-1}p_2 \qquad (۳۰-۴)$$
$$q_1^{2-y_1^k}\left(1-q_1\right)^{y_1^k-1}q_2^{2-y_2^k}\left(1-q_2\right)^{y_2^k-1}H_{x_1^k x_2^k y_1^k y_2^k}$$
$$G_{10}=\sum_{k=1}^{3}o_k\sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2}p_1^{2-x_1^k}\left(1-p_1\right)^{x_1^k-1}p_2^{2-x_2^k}\left(1-p_2\right)^{x_2^k-1}$$

$$Q_1=\sum_{k=1}^{3}o_k\sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2}p_1^{2-x_1^k}\left(1-p_1\right)^{x_1^k-1}p_2^{2-x_2^k}\left(1-p_2\right)^{x_2^k-1}$$
$$q_1q_2^{2-y_2^k}\left(1-q_2\right)^{y_2^k-1}H_{x_1^kx_2^ky_1^ky_2^k}$$
$$Q_2=\sum_{k=1}^{3}o_k\sum_{x_1^k=1}^{2}\sum_{x_2^k=1}^{2}\sum_{y_1^k=1}^{2}\sum_{y_2^k=1}^{2}p_1^{2-x_1^k}\left(1-p_1\right)^{x_1^k-1}p_2^{2-x_2^k}\left(1-p_2\right)^{x_2^k-1} \qquad (۳۱-۴)$$
$$q_1^{2-y_1^k}\left(1-q_1\right)^{y_1^k-1}q_2H_{x_1^kx_2^ky_1^ky_2^k}$$

احتمالات وقوع سه ترتیب نشانگری، مخلوطی از تمامی ترکیبات دیپلوتیپی هستند که به شکل ماتریسی به‌صورت زیر بیان می‌شوند:

$$O_1=o_1\sum_{x_1^1=1}^{2}\sum_{x_2^1=1}^{2}\sum_{y_1^1=1}^{2}\sum_{y_2^1=1}^{2}p_1^{2-x_1^1}\left(1-p_1\right)^{x_1^1-1}p_2^{2-x_2^1}\left(1-p_2\right)^{x_2^1-1}$$
$$q_1^{2-y_1^1}\left(1-q_1\right)^{y_1^1-1}q_2^{2-y_2^1}\left(1-q_2\right)^{y_2^1-1}H_{x_1^1x_2^1y_1^1y_2^1}$$
$$O_2=o_2\sum_{x_1^2=1}^{2}\sum_{x_2^2=1}^{2}\sum_{y_1^2=1}^{2}\sum_{y_2^2=1}^{2}p_1^{2-x_1^2}\left(1-p_1\right)^{x_1^2-1}p_2^{2-x_2^2}\left(1-p_2\right)^{x_2^2-1} \qquad (۳۲-۴)$$
$$q_1^{2-y_1^2}\left(1-q_1\right)^{y_1^2-1}q_2^{2-y_2^2}\left(1-q_2\right)^{y_2^2-1}H_{x_1^2x_2^2y_1^2y_2^2}$$
$$O_3=o_3\sum_{x_1^3=1}^{2}\sum_{x_2^3=1}^{2}\sum_{y_1^3=1}^{2}\sum_{y_2^3=1}^{2}p_1^{2-x_1^3}\left(1-p_1\right)^{x_1^3-1}p_2^{2-x_2^3}\left(1-p_2\right)^{x_2^3-1}$$
$$q_1^{2-y_1^3}\left(1-q_1\right)^{y_1^3-1}q_2^{2-y_2^3}\left(1-q_2\right)^{y_2^3-1}H_{x_1^3x_2^3y_1^3y_2^3}$$

از الگوریتم EM برای برآورد MLE های نرخ نوترکیبی در بین سه نشانگر استفاده می‌شود. معادلات کلی تکرار $(\tau+1)$ ام مرحله EM به‌صورت زیر فرموله می‌شوند:

مرحله E: همانند مرحله τ، تعداد مورد انتظار پیشامدهای نوترکیبی را برای فرد i مرتبط با $G_{00}=\left(G_{j_1j_2j_3}^{00}\right)_{16\times4}$، $G_{01}=\left(G_{j_1j_2j_3}^{01}\right)_{16\times4}$، $G_{10}=\left(G_{j_1j_2j_3}^{10}\right)_{16\times4}$ و $G_{11}=\left(G_{j_1j_2j_3}^{11}\right)_{16\times4}$ برای $\left(j_1j_2j_3\right)$ امین ژنوتیپ نتاج (که j_1، j_2 و j_3 به ترتیب نشان دهنده ژنوتیپ‌های نتاج برای سه نشانگر منفرد هستند) با استفاده از فرمول‌هایی که شکل‌های مشابه معادلات (۴-۲۲) دارند و H و G های آن در اینجا به ترتیب با معادلات (۴-۲۸) و (۴-۲۹) تعریف می‌شوند، محاسبه

نمایید. توجه کنید که ماتریس‌های H و G معادلات (۴-۲۸) و (۴-۲۹) نیز به وسیله p_1، p_2، q_1، q_2 و k ($k=1,2,3$) تعیین می‌شوند.

مرحله M: با استفاده از معادلاتی که شکل‌های مشابه معادلات (۴-۲۳) دارند، $g_{00}^{(\tau+1)}$، $g_{01}^{(\tau+1)}$، $g_{10}^{(\tau+1)}$ و $g_{11}^{(\tau+1)}$ را محاسبه کنید. ولی در اینجا نیز به تجدید مقادیر $p_1^{(\tau+1)}$، $p_2^{(\tau+1)}$، $q_1^{(\tau+1)}$، $q_2^{(\tau+1)}$ **و** $o_k^{(\tau+1)}$ با استفاده از روابط زیر نیاز داریم:

$$p_1^{(\tau+1)} = \frac{1}{n}\sum_{j_1=1}^{4}\sum_{j_2=1}^{4}\sum_{j_3=1}^{4}\frac{p_{1\,j_1j_2j_3}^{(\tau)}}{h_{j_1j_2j_3}^{(\tau)}}n_{j_1j_2j_3}$$

$$p_2^{(\tau+1)} = \frac{1}{n}\sum_{j_1=1}^{4}\sum_{j_2=1}^{4}\sum_{j_3=1}^{4}\frac{p_{2\,j_1j_2j_3}^{(\tau)}}{h_{j_1j_2j_3}^{(\tau)}}n_{j_1j_2j_3}$$

$$q_1^{(\tau+1)} = \frac{1}{n}\sum_{j_1=1}^{4}\sum_{j_2=1}^{4}\sum_{j_3=1}^{4}\frac{q_{1\,j_1j_2j_3}^{(\tau)}}{h_{j_1j_2j_3}^{(\tau)}}n_{j_1j_2j_3}$$

$$q_2^{(\tau+1)} = \frac{1}{n}\sum_{j_1=1}^{4}\sum_{j_2=1}^{4}\sum_{j_3=1}^{4}\frac{q_{2\,j_1j_2j_3}^{(\tau)}}{h_{j_1j_2j_3}^{(\tau)}}n_{j_1j_2j_3}$$

$$o_k^{(\tau+1)} = \frac{1}{n}\sum_{j_1=1}^{4}\sum_{j_2=1}^{4}\sum_{j_3=1}^{4}\frac{o_{kj_1j_2j_3}^{(\tau)}}{h_{j_1j_2j_3}^{(\tau)}}n_{j_1j_2j_3}$$

در این روابط، $n_{j_1j_2j_3}$ نشان‌دهنده تعداد نتاج دارای ژنوتیپ خاص برای سه نشانگر و $h_{j_1j_2j_3}$، $p_{1j_1j_2j_3}$، $p_{2j_1j_2j_3}$، $q_{1j_1j_2j_3}$ و $q_{2j_1j_2j_3}$ به ترتیب $(j_1j_2j_3)$ امین عنصر ماتریس های **H**، $\mathbf{P_1}$، $\mathbf{P_2}$، $\mathbf{Q_1}$ و $\mathbf{Q_2}$ هستند.

مراحل E و M تا جایی تکرار می‌شوند که g_{00}، g_{01}، g_{10} و g_{11} در مقادیر ثابتی که دقت قابل قبولی دارند، همگرا شوند. با استفاده از MLE های مربوط به g ها، MLE های نرخ نوترکیبی r_{AB}، r_{AC} و r_{BC} بر اساس خصوصیت ثابت MLE ها به‌دست می‌آید.

مثال ۴-۴. **(مدل‌سازی توأم پیوستگی، دیپلوتیپ والدینی و ترتیب ژنی)**. یین و همکاران (۲۰۰۴) نقشه پیوستگی با تراکم بالا تهیه شده با نشانگرهای ریزماهواره و AFLP را در جمعیت تلاقی برگشتی $(T\times D)\times D$، بین دو گونه درخت تبریزی *Populus trichocarpa* (T) و *P. deltoids* (D) گزارش نمودند. بر خلاف لاین‌های اینبرد معمولی، والدین درخت تبریزی مورد

استفاده برای تلاقی‌ها هتروزیگوت هستند؛ بنابراین، این تلاقی برگشتی همانند خانواده‌های تنی، دارای یک والد به صورت والد F_1 و والد دیگر به‌صورت والد D است. سه نشانگر یعنی CA/CGA-580RD (A)، AM11-1060 (B) و A7-690 (C) برای آمیزش آزمون انتخاب شدند که در F_1 به‌صورت هتروزیگوت (۱۲) ولی در والد D (۲۲) هموزیگوت هستند. ترتیب دقیق این سه نشانگر و دیپلوتیپ آن‌ها برای والد F_1 مشخص نیست. مدلی که در بخش ۴-۴ توصیف شد را برای برآورد همزمان پیوستگی، دیپلوتیپ والدینی و ترتیب ژنی مورد استفاده قرار خواهیم داد. مشاهدات این سه نشانگر به صورت زیر ارائه می‌شود:

نشانگر **A**	نشانگر **B**	نشانگر **C**	
		12	22
۱۲	۱۲	۳۳	۱
۱۲	۲۲	۴	۷
۲۲	۱۲	۳	۴
۲۲	۲۲	۰	۲۱

p_1، $1-p_1$ و p_2 به ترتیب احتمالات چهار دیپلوتیپ ممکن والد D به صورت $[111][222]$، $[121][212]$، $[111][222]$ و $[112][221]$ و o_1، o_2 و o_3 به ترتیب احتمالات سه ترتیب ژنی ممکن برای سه نشانگر **A**، **B** و **C** هستند. با استفاده از g_{00}، g_{01}، g_{10} و g_{11} برای نشان دادن احتمالات تعداد کراس‌اورهای مختص هر فاصله، یک ماتریس **H** (4×2) بر پایه معادله (۴-۲۸)، دو ماتریس **P** (4×2) بر اساس معادله (۴-۳۰) و سه ماتریس **O** (4×2) بر اساس معادله (۴-۳۲) به‌دست می‌آوریم. درضمن، چهار ماتریس **G** (4×2) بر اساس معادله (۴-۲۹) حاصل می‌گردد.

با استفاده از الگوریتم EM که در بالا توصیف شد، MLE های مربوط به پارامترهای ناشناخته به صورت $\hat{g}_{00}=0.5621$، $\hat{g}_{01}=0.1924$، $\hat{g}_{10}=0.0925$، $\hat{g}_{11}=0.1507$ و $\hat{p}_1=0.9$، $\hat{p}_2=0.8$، $\hat{o}_1=0$، $\hat{o}_2=0.9989$ و $\hat{o}_3=0.0011$ محاسبه می‌شوند. دیپلوتیپ والد D برای سه نشانگر بر اساس ترتیب **A-C-B** بدین صورت ارائه می‌شود:

CA/CGA-580RD (A)	1	2
A7-690 (C)	1	2
AM11-1060 (B)	2	1

مقادیر برآورد شده g تا حد زیادی برای برآورد MLE های مربوط به نرخ‌های نوترکیبی یعنی $\hat{r}_{AB} = 0.35$، $\hat{r}_{BC} = 0.28$ و $\hat{r}_{AC} = 0.24$ استفاده می‌شوند.

۴-۵- نشانگرهای حاوی اطلاعات جزیی

در بخش ۴-۲-۱، یک ماتریس $\mathbf{H}$ (4×4) را برای فراوانی‌های ژنوتیپی توأم بین دو نشانگر حاوی اطلاعات کامل تعریف نمودیم. برای یک نشانگر حاوی اطلاعات کامل، نسبت ۱:۱ بین ژنوتیپ و فنوتیپ وجود دارد. تعداد زیادی از نشانگرهایی که از لحاظ عملی مفید هستند، همانطور که در جدول ۳-۱ نشان داده شده‌اند، اطلاعات دهنده جزیی هستند که برای آن‌ها، چنین انطباق ۱:۱ وجود ندارد. مدل‌های کلی که در بخش بالا توصیف شدند، می‌توانند تغییر یابند تا پیوستگی و دیپلوتیپ‌های والدینی برای نشانگرهای اطلاعات دهنده نسبی نیز برآورد شوند.

۴-۵-۱- مدل‌سازی توأم پیوستگی و دیپلوتیپ والدینی

بر خلاف نشانگرهای حاوی اطلاعات کامل، چهار روش ممکن شکل‌گیری ژنوتیپ برای تلاقی نشانگری تیپ ۱۲×۱۲ سبب ایجاد سه فنوتیپ خواهند شد، چون شکل‌های ژنوتیپی ۱۲ و ۲۱ از لحاظ فنوتیپی مشابه هستند. به این دلیل، ماتریس $\mathbf{H}$ (4×4) برای فراوانی‌های ژنوتیپی دو نشانگر به‌صورت یک ماتریس با ابعاد کوچک‌تر برای فراوانی‌های فنوتیپی نشانگر کاهش خواهد یافت. وو و همکاران (b ۲۰۰۲) ماتریس‌های تلاقی ویژه‌ای (I) را که فراوانی‌های ژنوتیپی را به فراوانی‌های فنوتیپی برای انواع مختلف نشانگر مرتبط می‌سازد، طرح ریزی نمودند (ماتریس تلاقی (I) در اینجا به‌صورت ماتریسی تعریف می‌شود که ژنوتیپ نشانگر را به فنوتیپ نشانگر برای یک نوع از نشانگر اطلاعات دهنده نسبی مرتبط می‌سازد). در این بخش، از عبارت $H' = I_{b_1}^T H I_{b_2}$ برای یک ماتریس $(b_1 \times b_2)$ از فراوانی‌های فنوتیپی دو نشانگر اطلاعات دهنده نسبی استفاده می‌شود که در آن b_1 و b_2 به ترتیب تعداد فنوتیپ‌های قابل تشخیص برای نشانگرهای A و B هستند. مطابق با آن، داریم:

$$(DH)' = I_{b_1}^T (D \circ H) I_{b_2}$$
$$P' = I_{b_1}^T P I_{b_2}$$
$$Q' = I_{b_1}^T Q I_{b_2}$$

باید یادآور شویم که علامت $\circ$ به معنای ضرب جزء به جزء دو ماتریس است.

سپس الگوریتم EM می‌تواند برای برآورد نرخ نوترکیبی بین هر دو نشانگر اطلاعات دهنده نسبی بسط داده شود.

مرحله E: در مرحله τ، بر اساس ماتریس $(DH)'$ که از برآورد فعلی $r^{(\tau)}$ به‌دست آمده است، تعداد مورد انتظار پیشامدهای نوترکیبی بین دو نشانگر را برای ژنوتیپ نتاج مورد نظر محاسبه کنید:

$$D_{j_1 j_2}^{(\tau+1)} = \frac{(dh)'^{(\tau)}_{j_1 j_2}}{h'^{(\tau)}_{j_1 j_2}} n_{j_1 j_2} \qquad (۳۳-۴)$$

که $(dh)'_{j_1 j_2}$ و $h'_{j_1 j_2}$، به ترتیب $(j_1 j_2)$ امین عنصر ماتریس‌های $(DH)'$ و H' هستند.

مرحله M: $r^{(\tau+1)}$، $p^{(\tau+1)}$ و $q^{(\tau+1)}$ را با استفاده از معادلات زیر محاسبه کنید:

(۳۴-۴)

$$r^{(\tau+1)} = \frac{1}{2n} \sum_{j_1=1}^{b_1} \sum_{j_2=1}^{b_2} D_{j_1 j_2}^{(\tau+1)}$$

$$p^{(\tau+1)} = \frac{1}{n} \sum_{j_1=1}^{b_1} \sum_{j_2=1}^{b_2} \frac{p'^{(\tau)}_{j_1 j_2}}{h'^{(\tau)}_{j_1 j_2}} n_{j_1 j_2}$$

$$q^{(\tau+1)} = \frac{1}{n} \sum_{j_1=1}^{b_1} \sum_{j_2=1}^{b_2} \frac{q'^{(\tau)}_{j_1 j_2}}{h'^{(\tau)}_{j_1 j_2}} n_{j_1 j_2}$$

که $p'_{j_1 j_2}$ و $q'_{j_1 j_2}$، به ترتیب $(j_1 j_2)$ امین عنصر ماتریس‌های P' و Q' هستند.

در هر مرحله، ماتریس‌های $(DH)'$ و H' به وسیله r، p و q که جدیداً برآورد شده‌اند، تجدید و نوسازی می‌شوند. مراحل E و M بین معادلات (۴-۳۳) و (۴-۳۴) تا جایی تکرار می‌شوند که برآوردها در مقدار ثابت و پایداری به همگرایی برسند.

مثال ۴-۵. **(مدل‌سازی توأم پیوستگی و دیپلوتیپ والدینی)**. می‌توانیم از مثال مربوط به سه نشانگر غالب استفاده کنیم تا مدل واحد و یکنواختی را برای برآورد هم‌زمان پیوستگی و دیپلوتیپ والدینی ارائه نماییم. یک تلاقی بین دو هتروزیگوت سه‌گانه با ژنوتیپ ao/bo/co برای نشانگرهای

A، B و C انجام شد. به این دلیل که این سه نشانگر غالب هستند، تلاقی سبب تولید ۸ (نسبت به ۲۷) ژنوتیپ قابل تشخیص با مشاهدات زیر می‌شود:

نشانگر **A**	نشانگر **B**	نشانگر **C**	
		c-	oo
a-	b-	۲۸	۴
a-	oo	۱۲	۳
oo	b-	۱	۸
oo	oo	۲	۲

در ابتدا می‌توانیم از تجزیه دو نقطه‌ای برای برآورد نرخ‌های نوترکیبی و دیپلوتیپ‌های والدینی بین تمامی جفت‌های ممکنه سه نشانگر استفاده کنیم. برای یک نوع تلاقی نشانگری غالب $a|o|\times a|o|$ و $a|o|\times o|a|$، ماتریس تلاقی به ترتیب به‌صورت زیر تعریف می‌شود:

$$I_2=\begin{bmatrix}1&1&1&0\\0&0&0&1\end{bmatrix} \quad \text{یا} \quad \begin{bmatrix}1&1&0&0\\0&0&1&1\end{bmatrix}$$

با استفاده از ماتریس‌های تلاقی فوق که برای کاهش ابعاد ماتریس H بکار گرفته می‌شوند، از الگوریتم EM برای برآورد نرخ نوترکیبی بین ماتریس‌های A و B به صورت $r_{AB}=0.376$ استفاده می‌کنیم که دیپلوتیپ‌های والدینی برآورد شده آن به صورت $[ao][ob]\times[ab][oo]$ یا $[ab][oo]\times[ao][ob]$ هستند. به‌طور مشابه، دو نرخ نوترکیبی دیگر و دیپلوتیپ‌های والدینی مربوطه به صورت $r_{BC}=0.386$، $[bo][oc]\times[bc][oo]$ یا $[bc][oo]\times[bo][oc]$ و $r_{AC}=0.184$، $[ac][oo]\times[ac][oo]$ برآورد می‌شوند. از روی تجزیه دو نقطه‌ای، یکی از دو والد دارای آلل‌های غالب از نشانگرهای A و B است که با آلل‌های غالب نشانگر C جفت نمی‌شوند.

۴-۵-۲- مدل‌سازی توأم پیوستگی، دیپلوتیپ والدینی و ترتیب ژنی

سه نشانگر اطلاعات دهنده نسبی را با تعداد فنوتیپ‌های قابل تشخیصی که به ترتیب به صورت b_1، b_2 و b_3 نشان داده می‌شوند، در نظر بگیرید. $H'=\left(I_{b_1}^T\otimes I_{b_2}^T\right)HI_{b_3}$ به صورت یک ماتریس $\left(b_1b_2\times b_3\right)$ برای فراوانی‌های ژنوتیپی سه نشانگر اطلاعات دهنده نسبی تعریف می‌شود. به‌طور مشابه، $\left(HD_{00}\right)'=\left(I_{b_1}^T\otimes I_{b_2}^T\right)\left(D_{00}\circ H\right)I_{b_3}$، $\left(HD_{01}\right)'=\left(I_{b_1}^T\otimes I_{b_2}^T\right)\left(D_{01}\circ H\right)I_{b_3}$،

$(HD_{10})' = \left(I_{b_1}^T \otimes I_{b_2}^T\right)(D_{10} \circ H)\, I_{b_3}$، $(HD_{11})' = \left(I_{b_1}^T \otimes I_{b_2}^T\right)(D_{11} \circ H)\, I_{b_3}$،
$P_1' = \left(I_{b_1}^T \otimes I_{b_2}^T\right) P_1 I_{b_3}$، $P_2' = \left(I_{b_1}^T \otimes I_{b_2}^T\right) P_2 I_{b_3}$، $Q_1' = \left(I_{b_1}^T \otimes I_{b_2}^T\right) Q_1 I_{b_3}$،
$Q_2' = \left(I_{b_1}^T \otimes I_{b_2}^T\right) Q_2 I_{b_3}$، $O_1' = \left(I_{b_1}^T \otimes I_{b_2}^T\right) O_1 I_{b_3}$، $O_2' = \left(I_{b_1}^T \otimes I_{b_2}^T\right) O_2 I_{b_3}$ و
$O_3' = \left(I_{b_1}^T \otimes I_{b_2}^T\right) O_3 I_{b_3}$ تعریف می‌شوند.

با استفاده از روشی که در بخش ۴-۳ تشریح شده از الگوریتم EM برای برآورد MLE های مربوط به نرخ‌های نوترکیبی در بین سه نشانگر اطلاعات‌دهنده نسبی و همچنین ترکیبات دیپلوتیپی والدینی آن‌ها و ترتیب‌های ژنی استفاده می‌شود.

مثال ۴-۶. مثال ۴-۵ را مجدداً مشاهده کنید. از تجزیه سه نقطه‌ای استفاده می‌کنیم تا دیپلوتیپ‌های والدینی و ترتیب‌های ژنی را برای برآورد پیوستگی بین این سه نشانگر غالب برآورد کنیم. ترتیب ژنی برآورد شده به‌صورت **B-A-C** است. دیپلوتیپ‌های والدینی برآورد شده نیز به‌صورت زیر هستند:

$$\begin{array}{c} Bb \\ Aa \\ Cc \end{array}\left\|\begin{array}{cc} o & b \\ o\times & o \\ o & o \end{array}\right\|\begin{array}{c} o \\ a \\ c \end{array} \qquad \begin{array}{c} Bb \\ Ao \\ Co \end{array}\left\|\begin{array}{cc} o & b \\ a\times & a \\ c & c \end{array}\right\|\begin{array}{c} o \\ o \\ o \end{array}$$

MLE های نرخ های نوترکیبی به صورت $\hat{r}_{AB} = 0.309$، $\hat{r}_{AC} = 0.184$ و $\hat{r}_{BC} = 0.379$ هستند.

۴-۶- تمرین‌ها

۴-۱- نشانگرهای حاوی اطلاعات کامل

فرض کنید یک خانواده تنی حاصل از دو والد دگربارور با دیپلوتیپ‌های $[22][11]$ و $[44][33]$ ، به ترتیب برای دو نشانگر حاوی اطلاعات کامل A و B وجود دارد. مشاهدات مربوط به هر یک از ۱۶ ژنوتیپ دو نشانگری به‌صورت زیر است:

	13	14	23	24
13	$n_{11}=70$	$n_{12}=13$	$n_{13}=9$	$n_{14}=2$
14	$n_{21}=9$	$n_{22}=82$	$n_{23}=2$	$n_{24}=13$
23	$n_{31}=14$	$n_{32}=4$	$n_{33}=75$	$n_{34}=15$
24	$n_{41}=4$	$n_{42}=13$	$n_{43}=9$	$n_{44}=66$

$n_{j_1 j_2}$ را به‌عنوان تعداد نتاج در یک سلول در نظر بگیرید.

الف. ماتریس H را برای فراوانی‌های ژنوتیپی بر حسب نرخ نوترکیبی r بین دو نشانگر بنویسید.

ب. MLE مربوط به r را برآورد کرده و مشتق آن را بگیرید.

ج. آزمون کنید که پیوستگی معنی‌دار وجود دارد یا خیر.

۴-۲- یک نشانگر هم بارز F_2

یکی از دو نشانگر (که B نامیده می‌شود) به سیستم آللی مشابه بین دو والد بالا افزوده می‌شود. چنین نشانگرهایی، نشانگرهای هم بارز F_2 نامیده می‌شوند. دیپلوتیپ‌های دو والد به‌صورت $[22][11]$ و $[42][31]$ هستند. در این تلاقی، شکل‌های ژنوتیپی با فنوتیپ مشابه، با یکدیگر تلفیق خواهند شد. مشاهدات به صورت زیر هستند:

	11	12	22
13	$n_{11}=70$	$n_{12}+n_{13}=22$	$n_{14}=2$
14	$n_{21}=9$	$n_{22}+n_{23}=84$	$n_{24}=13$
23	$n_{31}=14$	$n_{32}+n_{33}=79$	$n_{34}=15$
24	$n_{41}=4$	$n_{42}+n_{43}=22$	$n_{44}=66$

الف. ماتریس H را برای فراوانی‌های ژنوتیپی بر حسب نرخ نوترکیبی r بین دو نشانگر بنویسید.

ب. نشان دهید که چگونه الگوریتم EM می‌تواند برای برآورد r استفاده شود.

ج. r را با استفاده از برنامه‌ای برای الگوریتم EM برآورد کنید.

د. آزمون کنید که این دو نشانگر به‌صورت معنی‌دار به یکدیگر پیوسته هستند.

تذکرات: در این وضعیت، ماتریس های H و D به ترتیب به صورت زیر تجزیه می شوند:

$$H_{11} = \begin{array}{c} \\ 13 \\ 14 \\ 23 \\ 24 \end{array} \begin{array}{c} \begin{array}{ccc} 11 & 12 & 22 \end{array} \\ \begin{bmatrix} \frac{(1-r)^2}{4} & \frac{r(1-r)}{4}+\frac{r(1-r)}{4} & \frac{r^2}{4} \\ \frac{r(1-r)}{4} & \frac{(1-r)^2}{4}+\frac{r^2}{4} & \frac{r(1-r)}{4} \\ \frac{r(1-r)}{4} & \frac{r^2}{4}+\frac{(1-r)^2}{4} & \frac{r(1-r)}{4} \\ \frac{r^2}{4} & \frac{r(1-r)}{4}+\frac{r(1-r)}{4} & \frac{(1-r)^2}{4} \end{bmatrix} \end{array}$$

$$D = \begin{array}{c} \\ 13 \\ 14 \\ 23 \\ 24 \end{array} \begin{array}{c} \begin{array}{ccc} 11 & 12 & 22 \end{array} \\ \begin{vmatrix} 0 & 1 & 2 \\ 1 & 2\phi & 1 \\ 1 & 2\phi & 1 \\ 2 & 1 & 0 \end{vmatrix} \end{array}$$

که $\phi = \frac{r^2}{(1-r)^2 + r^2}$ است. از این بخش، باید MLE مربوط به r را تعیین کنید و مراحل الگوریتم EM را نشان دهید.

۴-۳- دو نشانگر هم بارز F_2

مسئله ۴-۲ را برای وقتی که هر دو نشانگر، نشانگرهای هم بارز F_2 باشند، انجام دهید.

۴-۶-۱- یک نشانگر غالب

اگر یکی از دو نشانگر بالا (که B نامیده می‌شود) غالب باشد، می‌توان مشاهدات را به‌صورت زیر تجزیه نمود:

	10	00
13	$n_{11}+n_{12}+n_{13}=92$	$n_{14}=2$
14	$n_{21}+n_{22}+n_{23}=93$	$n_{24}=13$
23	$n_{31}+n_{32}+n_{33}=93$	$n_{34}=15$
24	$n_{41}+n_{42}+n_{43}=26$	$n_{44}=66$

مسائل مشابه تمرین ۴-۲ را انجام دهید.

۴-۶-۲- یک نشانگر غالب و یک نشانگر هم بارز F_2

اگر یکی از دو نشانگر (B) غالب و دیگری (A) هم بارز F_2 باشد، می‌توان مشاهدات را به صورت زیر تجزیه نمود:

	10	00
11	$n_{11}+n_{12}+n_{13}=92$	$n_{14}=2$
12	$n_{21}+n_{22}+n_{23}+n_{31}+n_{32}+n_{33}=186$	$n_{24}+n_{34}=28$
22	$n_{41}+n_{42}+n_{43}=26$	$n_{44}=66$

مسائل مشابه تمرین ۴-۲ را انجام دهید.

۴-۴- دو نشانگر غالب

اگر هر دو نشانگر غالب باشند، داریم:

	10	00
10	$n_{11}+n_{12}+n_{13}+n_{21}+n_{22}+n_{23}+n_{31}+n_{32}+n_{33}=278$	$n_{14}+n_{24}+n_{34}=30$
00	$n_{41}+n_{42}+n_{43}=26$	$n_{44}=66$

همانند تمرین ۴-۲ عمل کنید.

۴-۷- نکات

در این فصل چارچوب کلی تجزیه پیوستگی را برای هر نوع نشانگر معرفی نمودیم. برای گونه‌های دگرگشن، دیپلوتیپ‌های بین نشانگرهای مختلف برای والدینی که برای تلاقی استفاده می‌شوند،

ناشناخته هستند. همچنین مدل‌های توأمی را که پیوستگی، دیپلوتیپ والدینی و ترتیب ژنی را ترکیب می‌کنند، توصیف نمودیم. لو و همکاران (۲۰۰۴) مطالعات شبیه‌سازی گسترده‌ای را برای تشریح مزایای مدل‌سازی توأم انجام دادند. در این بخش، سعی می‌کنیم تا نتایج اصلی مطالعات لو و همکاران (۲۰۰۴) را خلاصه نماییم.

۴-۷-۱- تجزیه پیوستگی

فرض کنید ۵ نشانگر با ترتیب شناخته شده M_1-M_2-M_3-M_4-M_5 روی یک کروموزوم قرار دارند. این ۵ نشانگر به‌طور متفاوتی با ترتیب ۱:۱:۱:۱، ۱:۲:۱، ۳:۱، ۱:۱ و ۱:۱:۱:۱ در حال تفرق هستند. دو والد، با دیپلوتیپ‌های داده شده برای ۵ نشانگر (جدول ۴-۴)، تلاقی داده می‌شوند تا خانواده تنی در حال تفرق را ایجاد کنند. این خانواده تنی با درجات متفاوت پیوستگی شبیه‌سازی می‌شود (r = ۰/۰۵ در مقابل r = ۰/۲۰).

همان‌گونه که انتظار می‌رود، نشانگرهایی که اطلاعات بیشتری ارایه می‌دهند یا نشانگرهایی که پیوستگی محکم‌تری دارند، دقت بیشتری در برآورد پیوستگی نسبت به نشانگرهایی دارند که اطلاعات کمتر یا پیوستگی ضعیف‌تری دارند (جدول ۴-۴). مدل‌های توأم برای تجزیه دو نقطه‌ای می‌توانند برآورد عالی از دیپلوتیپ والدینی فراهم کنند. برای مثال، MLE احتمال دیپلوتیپ والدینی (p یا q) نزدیک به ۱ یا ۰ است (جدول ۴-۴) که نشان می‌دهد همیشه می‌توانیم دیپلوتیپ‌های والدینی را به‌طور دقیقی برآورد کنیم. ولی برای دو نشانگر متقارن (یعنی نشانگرهای M_2 و M_3 در این مثال)، دو مجموعه از MLE ها، $\hat{p}=1$ ، $\hat{q}=0$ و $\hat{p}=0$ ، $\hat{q}=1$، آماره آزمون نسبت درستنمایی مشابهی را ارایه می‌دهند؛ بنابراین، تجزیه دو نقطه‌ای نمی‌تواند دیپلوتیپ‌های والدینی را برای نشانگرهای متقارن حتی زمانی که دو والد دیپلوتیپ‌های متفاوتی دارند، ارائه دهد.

دقت برآورد پیوستگی هنگامی افزایش می‌یابد که تجزیه سه نقطه‌ای انجام شود (جدول ۴-۵)، ولی این حالت بستگی به انواع متفاوت نشانگرها و درجات متفاوت پیوستگی دارد. مزیت تجزیه سه نقطه‌ای در برابر تجزیه دو نقطه‌ای برای نشانگرهای اطلاعات دهنده نسبی نسبت به نشانگرهای حاوی اطلاعات کامل و برای نشانگرهای دارای پیوستگی کمتر نسبت به نشانگرهای دارای پیوستگی محکم‌تر بارها گزارش شده است. برای مثال، خطای نمونه برداری MLE نرخ نوترکیبی (با فرض r = ۰/۲۰) بین نشانگرهای M_2 و M_3 حاصل از تجزیه دو نقطه‌ای ۰/۰۸۴۸ است، در حالیکه این مقدار در تجزیه سه نقطه‌ای وقتی که نشانگر حاوی اطلاعات کامل M_1 را ترکیب می‌کنیم به ۰/۰۷۵۸ کاهش می‌یابد و وقتی که نشانگر اطلاعات‌دهنده نسبی M_4 را ترکیب می‌کنیم به ۰/۰۹۳۹ افزایش می‌یابد. تجزیه سه نقطه‌ای می‌تواند به وضوح دیپلوتیپ‌های والدین مختلف را

تا مادامی‌که یکی از سه نشانگر نامتقارن باشند، تعیین نماید. در مثال مذکور، با استفاده از نشانگرهای نامتقارن M_1 یا M_4، دیپلوتیپ‌های دو والد برای دو نشانگر متقارن (M_2 و M_3) می‌تواند تعیین شود. مدل تجزیه سه نقطه‌ای می‌تواند محتمل‌ترین ترتیب ژنی را تعیین کند. در تجزیه‌های سه نقطه‌ای که نشانگرهای M_1 و M_3، نشانگرهای M_2 و M_4 و نشانگرهای M_3 و M_5 را ترکیب می‌کنند، MLE های احتمالات ترتیب ژنی تقریباً برابر با ۱ هستند که نشان می‌دهد ترتیب ژنی برآورد شده با ترتیب مفروض سازگار است.

جدول ۴-۴- برآورد تجزیه دو نقطه‌ای نرخ نوترکیبی ($\hat{r} \pm SD$) و احتمال دیپلوتیپ والدینی برای والدین P ($\hat{p}$) و Q ($\hat{q}$) برای ۵ نشانگر در یک خانواده تنی با n = ۱۰۰.

نشانگر	دیپلوتیپ والدینی P[a] × Q[a]	r = ۰/۰۵ $\hat{r}$	$\hat{p}$	$\hat{q}$	r = ۰/۲۰ $\hat{r}$	$\hat{p}$	$\hat{q}$
M_1	| | 1 2 × | | 3 4						
		0.0530±0.0183			0.2097±0.0328		
M_2	| | 1 2 × | | 1 2		0.9960	0.9972		0.9882	0.9878
		0.0464±0.0303			0.2103±0.0848		
M_3	| | 1 0 × | | 0 1		1 (0^b)	0 (1^b)		1 (0^b)	0 (1^b)
		0.0463±0.0371			0.1952±0.0777		
M_4	| | 1 2 × | | 2 2		1	1.0^c		1	1.0^c
		0.0503±0.0231			0.2002±0.0414		
M_5	| | 1 2 × | | 3 4		1	1.0^c		1	1.0^c

MLE مربوط به r بین دو نشانگر برای مقایسه ارایه شده است، حال آنکه MLE های p و q فقط برای نشانگر دوم ارایه شده‌اند. [a] نشان‌دهنده دیپلوتیپ والدینی هر والد برای ۵ نشانگر مفروض است و خطوط عمودی نشان دهنده دو کروموزوم همولوگ هستند. [b] مقادیر درون پرانتز، جواب ممکنه دوم را نشان می‌دهند. برای هر دو نشانگر متقارن (M_2 و M_3)، $\hat{p}=1$ ، $\hat{q}=0$ و $\hat{p}=0$ ، $\hat{q}=1$، آماره آزمون نسبت درستنمایی مشابهی را ارایه می‌دهد؛ بنابراین، زمانی که دو والد دیپلوتیپ‌های متفاوتی برای نشانگرهای متقارن داشته باشند، دیپلوتیپ‌های والدینی آن‌ها نمی‌توانند به‌طور صحیحی از روی تجزیه دو نقطه‌ای تعیین شوند. [c] دیپلوتیپ والدینی والد Q نمی‌تواند در این دو حالت برآورد شود، چون نشانگر ۴ در این والد هموزیگوت است.

۴-۷-۲- احتمال دیپلوتیپی

دو نشانگر غالب به هم پیوسته برای نشان دادن مزیت مدل‌سازی توأم پیوستگی و دیپلوتیپ والدینی شبیه‌سازی شدند. در تجزیه دو نقطه‌ای، دو ترکیب دیپلوتیپ والدینی متفاوت در نظر گرفته می‌شوند:

(۱): $[aa][oo]\times[aa][oo]$ (سیس×سیس)

(۲): $[ao][oa]\times[ao][oa]$ (ترانس×ترانس)

MLE پیوستگی تحت ترکیب (۲) که در آن دو آلل غالب در فاز ترانس قرار دارند، به دقت MLE پیوستگی تحت ترکیب (۱) که در آن دو غیر آلل غالب در فاز سیس قرار دارند، نیست. برای مجموعه داده‌های معین با فاز پیوستگی ناشناخته، روش مرسوم برای برآورد نرخ نوترکیبی، محاسبه مقادیر درستنمایی تحت تمامی ترکیبات ممکن فاز پیوستگی (یعنی سیس×سیس، سیس×ترانس، ترانس×سیس و ترانس×ترانس) است. ترکیبات سیس×سیس و ترانس×ترانس دارای مقدار درستنمایی یکسانی هستند، به‌طوری که MLE یک ترکیب برابر با ۱ منهای MLE ترکیب دوم است (جدول ۴-۶). رابطه مشابهی برای سیس×ترانس و ترانس×سیس صادق است. محتمل‌ترین ترکیب فازی، ترکیبی است که دارای بزرگ‌ترین درستنمایی و یک MLE منطقی برای نرخ نوترکیبی ($r \le 0.5$) باشد.

برای مجموعه داده‌های شبیه‌سازی شده از $[aa][oo]\times[aa][oo]$، به آسانی می‌توان وضعیت سیس×سیس را به‌عنوان بهترین برآورد ترکیب فازی انتخاب نمود، چون دارای مقادیر بزرگ‌تر درستنمایی و مقادیر کوچک‌تر $\hat{r}$ است. تلفیق مدل پیوستگی دیپلوتیپ‌های والدینی، می‌تواند دقت برآورد پیوستگی قابل مقایسه‌ای را برای داده‌های $[aa][oo]\times[aa][oo]$ فراهم کند و دقیقاً دیپلوتیپ‌های والدینی را تعیین نماید (MLE های p و q و جدول ۴-۶ را مشاهده نمایید). مدل واحد و یکنواخت[1] دارای مزیت زیادی در برابر مدل مرسوم برای داده‌های $[ao][oa]\times[ao][oa]$ است. برای این مجموعه داده، تحت تمامی چهار ترکیب دیپلوتیپی ممکن (جدول ۴-۶) درستنمایی مشابهی به‌دست می‌آید. در این مورد، سیس×ترانس یا ترانس×سیس انتخاب می‌شود، چون این دو ترکیب فازی با مقادیر کمتر r مرتبط هستند. ولی این برآورد از r (۰/۰۳۹۳) دارای اریب است، چون بسیار کمتر از مقدار مفروض ۰/۲۰ است. مدل یکنواخت و واحد، دقت برآورد پیوستگی مشابهی را برای داده‌های مشتق از $[ao][oa]\times[ao][oa]$ همانند وقتی که تجزیه بر اساس ترکیب صحیح دیپلوتیپی انجام

[1] - Unified model

می‌شود، ارایه می‌دهد (جدول ۴-۶). همچنین مدل یکنواخت و واحد می‌تواند دقیقاً دیپلوتیپ‌های والدینی را تعیین کند ($\hat{p} = \hat{q} = 0$).

۴-۷-۳- ترتیب ژنی

مزیت به‌کارگیری تجزیه پیوستگی با ترتیب‌های ژنی را در تجزیه سه نقطه‌ای امتحان می‌کنیم. فرض می‌شود که سه نشانگر غالب دارای دو ترکیب دیپلوتیپی مختلف به شکل زیر هستند:

(۱): $[aaa][ooo]\times[aaa][ooo]$

(۲): $[aao][ooa]\times[aao][ooa]$

روش مرسوم، محاسبه مقادیر درستنمایی تحت سه ترتیب ژنی ممکن و انتخاب یک درستنمایی حداکثر برای برآورد پیوستگی است. تحت ترکیب (۱)، محتمل‌ترین ترتیب ژنی می‌تواند به خوبی تعیین شده و بنابراین، نرخ‌های نوترکیبی بین سه نشانگر به خوبی برآورد شوند، چون مقدار درستنمایی ترتیب صحیح همیشه بزرگ‌تر از مقادیر درستنمایی ترتیب‌های غیر صحیح (جدول ۴-۷) است؛ اما تحت ترکیب (۲)، برآوردهای پیوستگی همیشه دقیق نیستند، چون با فراوانی ۲۰ درصد، ترتیب‌های ژنی به درستی تعیین نمی‌شوند. برآوردهای r تا حد زیادی از مقادیر حقیقی آن‌ها بر اساس ترتیب ژنی غلط انحراف خواهند داشت (جدول ۴-۷). مدل یکنواخت ترکیب کننده ترتیب ژنی می‌تواند برآورد بهتری از پیوستگی نسبت به راهکار مرسوم، به ویژه بین نشانگرهایی که دارای آلل‌های غالب در فاز ترانس هستند، ارایه کند؛ بنابراین، محتمل‌ترین ترتیب ژنی می‌تواند به‌طور همزمان با برآورد پیوستگی از روی مدل تعیین شود.

۴-۷-۴- تجزیه M نقطه‌ای

با در نظر گرفتن وابستگی پیشامدهای نوترکیبی در بین فواصل نشانگری مختلف، می‌توان تجزیه سه نقطه‌ای را بسط داد و تجزیه پیوستگی یک تعداد اختیاری نشانگرها را انجام داد. فرض کنید m نشانگر مرتب شده روی یک گروه پیوستگی وجود داشته باشد. احتمالات توأم ژنوتیپی m نشانگر یک ماتریس $\left(4^{m-1}\times 4\right)$ تشکیل می‌دهد. تعداد $2^{m-1}\times 2^{m-1}$ ماتریس احتمالی این‌چنینی مشروط بر ترکیب دیپلوتیپی والدینی خاص وجود دارند. برآوردهای منطقی نرخ‌های نوترکیبی به خصوصیات محتمل‌ترین ترکیب فازی بر اساس مقادیر درستنمایی محاسبه شده تکیه دارند.

احتمالات ژنوتیپی توأم m نشانگر می‌تواند به صورت تابعی از احتمال وقوع یا عدم وقوع کراس‌اور در بین دو نشانگر مجاور هم بیان شود، یعنی $g_{l_1 l_2 \ldots l_{m-1}}$ که $l_1, l_2, \ldots, l_{m-1}$ متغیرهای شاخص و موهومی هستند که به ترتیب نشان دهنده وقوع کراس‌اور بین نشانگرهای M_1 و M_2،

نشانگرهای M_2 و M_3، ... و نشانگرهای M_{m-1} و M_m هستند. اگر یک کراس‌اور بین دو نشانگر مجاور وجود داشته باشد، این شاخص به‌صورت ۱ و در غیر این‌صورت به صورت ۰ نشان داده می‌شود. چون هر شاخص می‌تواند به‌صورت یک یا صفر باشد، از این‌رو مجموعاً تعداد 2^{m-1} تابع g وجود خواهد داشت.

احتمال این‌که یک کراس‌اور به‌طور اختصاصی در فاصله $g_{l_1 l_2 \ldots l_{m-1}}$ رخ دهد، می‌تواند با استفاده از الگوریتم EM برآورد شود. در مرحله E، تعداد مورد انتظار کراس‌اورهای مختص هر فاصله محاسبه می‌شوند (معادله (۴-۲۲) را برای تجزیه سه نقطه‌ای مشاهده کنید). در مرحله M، یک معادله ساده برای برآورد احتمال $g_{l_1 l_2 \ldots l_{m-1}}$ مورد استفاده قرار می‌گیرد. MLE های مربوط به $g_{l_1 l_2 \ldots l_{m-1}}$ برای برآورد تعداد $\frac{m(m-1)}{2}$ نرخ نوترکیبی بین تمامی جفت نشانگرهای ممکنه مورد استفاده قرار می‌گیرد. با مقایسه مقادیر و بزرگی این نرخ‌های نوترکیبی، می‌توانیم اطلاعات کافی در خصوص ترتیب نشانگرها به‌دست آوریم.

جدول ۴-۵- برآورد تجزیه سه نقطه‌ای نرخ نوترکیبی ($\hat{r} \pm SD$) و احتمالات دیپلوتیپ والدینی برای والد P ($\hat{p}$) و Q ($\hat{q}$) برای پنج نشانگر در یک خانواده تنی با n = ۱۰۰

نشانگر	دیپلوتیپ والدینی P × Q	$\hat{r}$ حالت ۱	$\hat{r}$ حالت ۲	$\hat{p}$	$\hat{q}$	$\hat{r}$ حالت ۱	$\hat{r}$ حالت ۲	$\hat{p}$	$\hat{q}$
نرخ نوترکیبی = 0.05									
	| | | |								
M_1	1 2 3 4								
	| | | |	0.0511±0.0175							
M_2	1 2 1 2		0.1008±0.0298	0.9978	0.9986				
	| | | |	0.0578±0.0269				0.0557±0.0312			
M_3	1 0 × 0 1			0.9977	0		0.0988±0.0277	1	0
	| | | |	0.0512±0.0307				0.0476±0.0280			
M_4	1 2 2 2		0.0932±0.0301	1	1/0			1	1/0
	| | | |	0.0514±0.0229							
M_5	1 2 3 4			1	1				
	| | | |								
نرخ نوترکیبی = 0.20									
	| | | |								
M_1	1 2 3 4								
	| | | |	0.2026±0.0348							
M_2	1 2 1 2		0.3282±0.0482	0.9918	0.9916				
	| | | |	0.2240±0.0758				0.2408±0.0939			
M_3	1 0 × 0 1			0.9944	0		0.3241±0.0488	1	0
	| | | |	0.1927±0.0613				0.1824±0.0614			
M_4	1 2 2 2		0.3161±0.0502	1	1/0			1	1/0
	| | | |	0.2017±0.0393							
M_5	1 2 3 4			1	1				
	| | | |								

حالت ۱ نرخ نوترکیبی بین دو نشانگر مجاور هم را نشان می‌دهد، حال آنکه حالت ۲ نرخ نوترکیبی بین دو نشانگر جدا شده به وسیله نشانگر سوم را نشان می‌دهد. برای توضیحات دیگر جدول ۴-۴ را مشاهده کنید.

جدول ۴-۶- مقایسه برآورد پیوستگی و دیپلوتیپ والدینی بین دو نشانگر غالب در یک خانواده تنی با n=۱۰۰ با استفاده از مدل مرسوم و یکنواخت.

	مدل مرسوم				مدل یکنواخت
	سپس×سپس	ترانس×سپس	سپس×ترانس	ترانس×ترانس	
داده شبیه‌سازی شده از سپس×سپس					
ترکیب دیپلوتیپی صحیح	درست	نادرست	نادرست	نادرست	
لگاریتم درستنمایی[a]	-46.2	-92.3	-92.3	-46.2	
$\hat{r}$ تحت هر ترکیب دیپلوتیپی	0.1981±0.0446	0.5000±0.0000	0.5000±0.0000	0.8018±0.0446	
ترکیب دیپلوتیپی برآورد شده	انتخاب شده				
$\hat{r}$ تحت ترکیب دیپلوتیپی صحیح	0.1981±0.0446				0.1982±0.0446
احتمال دیپلوتیپی برای والد P ($\hat{p}$)					1.0000±0.0000
احتمال دیپلوتیپی برای والد Q ($\hat{q}$)					1.0000±0.0000
داده شبیه‌سازی شده از ترانس×ترانس					
ترکیب دیپلوتیپی صحیح	نادرست	نادرست	نادرست	درست	
لگاریتم درستنمایی[a]	-89.6	-89.6	-89.6	-89.6	
$\hat{r}$ تحت هر ترکیب دیپلوتیپی	0.8573±0.1253	0.0393±0.0419	0.0393±0.0419	0.1426±0.1253	
ترکیب دیپلوتیپی برآورد شده		انتخاب شده	انتخاب شده		
$\hat{r}$ تحت ترکیب دیپلوتیپی صحیح				0.1426±0.1253	0.1426±0.1253
احتمال دیپلوتیپی برای والد P ($\hat{p}$)					0.0000±0.0000
احتمال دیپلوتیپی برای والد Q ($\hat{q}$)					0.0000±0.0000

[a] مقادیر لگاریتم درستنمایی ارایه شده در این بخش، مقادیر حاصل از شبیه‌سازی تصادفی برای هر ترکیب دیپلوتیپی به وسیله مدل مرسوم هستند.

فصل پنجم

تجزیهٔ پیوستگی با استفاده از لاین‌های اینبرد نوترکیب

۵-۱- مقدمه

لاین‌های اینبرد نوترکیب (RILs[1]) توانایی اثبات شده و قابل قبولی برای مکان‌یابی QTL ها دارند. این لاین‌ها یا از راه خویش‌آمیزی‌های مکرر و یا با آمیزش‌های برادر-خواهری مکرر نتاج F_1 حاصل از تلاقی بین دو لاین اینبرد می‌توانند به‌وجود آیند. به‌دلیل خویش‌آمیزی مداوم برای تعداد نسل‌های نسبتاً کافی (برای مثال ۷ تا ۱۰ نسل)، RIL ها تمایل دارند تا برای تمامی ژن‌ها هموزیگوت و خالص باشند. با چنین ژنوتیپ‌های تثبیت شده‌ای (با نادیده گرفتن جهش)، RIL ها همیشه می‌توانند تکثیر یابند و اجازه تکرار ژنوتیپ‌های مشابه را در مقیاس زمانی و مکانی می‌دهند و از این راه می‌توان بسیاری از مسائل ژنتیکی و زیستی اساسی را مورد ارزیابی قرار داد؛ بنابراین، RIL ها کراس‌اورهایی را که در هر میوز و در هر نسل رخ می‌دهند، جمع می‌کنند و بنابراین سهم زیگوت‌های نوترکیب در RIL ها (یعنی احتمال این‌که دو جایگاه ژنی به هم پیوسته دارای آلل‌های والدینی متفاوت باشند) بیش از آن چیزی است که در F_2 وجود دارد. در نتیجه، RIL ها که تا حد زیادی به‌واسطه تلاش‌های گسترده متخصصان علم ژنتیک و اصلاح نژاد در دسترس قرار گرفته‌اند (تریگیل و همکاران[2]، ۲۰۰۲؛ کنسرسیوم صفات پیچیده[3]، ۲۰۰۴) ابزار قدرتمندی را برای نقشه یابی دقیق QTLها فراهم نموده‌اند.

علاقه‌مندی به تجزیه ژنتیکی RIL ها می‌تواند به مطالعات ابتدایی یانینگز[4] (۱۹۱۷) و رابینز[5] (۱۹۱۸) باز گردد. در تحقیق هالدین و وادینگتون (۱۹۳۱) جزئیات اساسی تجزیه پیوستگی در RIL های تولید شده از خویش‌آمیزی یا آمیزش‌های برادر-خواهری را تشریح کردند. امروزه، تجزیه داده‌های RIL سبب افزایش توجهات نسبت به تحقیقات عملی‌تر و نظری‌تر شده است (برومن، ۲۰۰۵؛ تاچر و همکاران، ۲۰۰۵؛ مارتین و هوسپیتال، ۲۰۰۶). در این فصل، برخی از نظریه‌های اساسی تجزیه پیوستگی در RIL های تولید شده از راه خویش‌آمیزی و آمیزش‌های برادر- خواهری توصیف خواهد شد. روش‌ها و الگوریتم‌های آماری برای تجزیه RIL نیز مورد بررسی قرار خواهد گرفت.

---

[1] - Recombinant Inbred Lines
[2] - Threadgill *et al.*
[3] - Complex Trait Consortium
[4] - Jennings
[5] - Robbins

۵-۲- RIL های حاصل از خویش‌آمیزی

۵-۲-۱- تجزیه دو نقطه‌ای

یک جفت نشانگر به صورت A (با دو آلل A و a) و B (با دو آلل B و b) با نرخ نوترکیبی r در نظر بگیرید. دو لاین اینبرد AABB و aabb را تلاقی دادیم تا هتروزیگوت F_1 ایجاد شود. F_1ها خودگشن شدند تا F_2های در حال تفرق را ایجاد کنند و هر یک از ژنوتیپ‌های F_2 مجدداً خودگشن شدند تا F_3 را تولید کنند. این فرآیند خودگشنی برای تعداد زیادی نسل تکرار می‌شود. در نسل t، سهم کل ۱۰ نوع زیگوت (دیپلوتیپ‌ها، یعنی، ژنوتیپ‌های دارای فاز مشخص) بر حسب ۵ حالت به‌صورت زیر بیان می‌شوند:

سهم	دیپلوتیپ	وضعیت
C_t	AABB, aabb	۱
D_t	AAbb, aaBB	۲
E_t	AABb, AaBB, Aabb, aaBb	۳
F_t	$[AB][ab]$	۴
G_t	$[Ab][aB]$	۵

از کروشه‌های دوگانه برای نشان دادن دیپلوتیپ‌های مختلف زیگوت مضاعف AaBb استفاده نمودیم. ژنوتیپ‌های فوق خودگشن می‌شوند تا نسل t+1 را ایجاد کنند که دارای سهم دیپلوتیپی جدید زیر هستند:

$$\begin{aligned}
C_{t+1} &= C_t + \frac{1}{2}E_t + \frac{1}{4}(1-r)^2 F_t + \frac{1}{4}r^2 G_t \\
D_{t+1} &= D_t + \frac{1}{2}E_t + \frac{1}{4}r^2 F_t + \frac{1}{4}(1-r)^2 G_t \\
E_{t+1} &= \frac{1}{4}r(1-r)\ (F_t + G_t) \\
F_{t+1} &= \frac{1}{2}(1-r)^2 F_t + \frac{1}{2}r^2 G_t \\
G_{t+1} &= \frac{1}{2}r^2 F_t + \frac{1}{2}(1-r)^2 G_t
\end{aligned} \tag{۱-۵}$$

چون هر ژنوتیپ از دو گامت یا هاپلوتیپ تشکیل یافته است، ازاین‌رو $C_{t+1} + D_{t+1} + E_{t+1} + F_{t+1} + G_{t+1} = 2$ است، به‌طوری که $C_1 = D_1 = E_1 = G_1 = 0$ و

$F_1 = 2$ است. مشتق معادلات (۵-۱) می‌تواند به‌صورت زیر زمانی که خویش آمیزی انجام می‌شود، هموزیگوت‌ها فقط خودشان را تولید می‌کنند، به‌طوری که C_t و D_t فقط در C_{t+1} و D_{t+1} سهم دارند. ولی خودگشنی هتروزیگوت‌ها سبب ایجاد تفرق خواهد شد. برای مثال، وقتی AABb خودگشن شود، $\frac{1}{4}AABB:\frac{1}{2}AABb:\frac{1}{4}AAbb$ را ایجاد می‌نماید؛ بنابراین، سهم E_t به E_{t+1} برابر $\frac{1}{2}E_t$ است. ولی سهم آن برای C_{t+1} و D_{t+1} دو برابر است، چون در C_{t+1} یا D_{t+1} به میزان دو برابر سهم E_t تعداد کلاس بیشتری وجود دارد. هتروزیگوت مضاعف چهار هاپلوتیپ AB، Ab، aB و ab را با فراوانی‌های $\frac{1}{2}(1-r)$، $\frac{1}{2}r$، $\frac{1}{2}r$ و $\frac{1}{2}(1-r)$ برای دیپلوتیپ $[AB][ab]$ یا با فراوانی‌های $\frac{1}{2}r$، $\frac{1}{2}(1-r)$، $\frac{1}{2}(1-r)$ و $\frac{1}{2}r$ برای دیپلوتیپ $[Ab][aB]$ تولید می‌کند؛ بنابراین، ضرایب سهم F_t یا G_t به C_{t+1} و D_{t+1} باید برابر با حاصل‌ضرب فراوانی‌های هاپلوتیپی مربوطه باشند.

زمانی که t به سمت بی‌نهایت میل کند، خواهیم داشت: $E_\infty = F_\infty = G_\infty = 0$ و D_∞ سهم نهایی زیگوت‌های کراس‌اوری است. حال $C_t - D_t = c_t$ و $F_t - G_t = d_t$ را در نظر می‌گیریم؛ بنابراین، با تفریق معادلات C_{t+1} و D_{t+1} و همچنین F_{t+1} و G_{t+1} داریم:

$$\begin{aligned} c_{t+1} &= c_t + \frac{1}{4}(1-2r)d_t \\ d_{t+1} &= \frac{1}{2}(1-2r)d_t \end{aligned} \qquad (۵-۲)$$

اگر پارامتر λ را در نظر بگیرید که برای تمامی مقادیر t داریم: $c_{t+1} + \lambda d_{t+1} \equiv c_t + \lambda d_t$. از روی معادله (۵-۲) نیز داریم:

$$c_t + \lambda d_t = c_t + \frac{1}{4}(1-2r)d_t + \frac{1}{2}\lambda(1-2r)d_t$$

که منجر به عبارت زیر می‌شود:

$$\lambda = \frac{1-2r}{2(1+2r)}$$

بر اساس تعریف و طبیعت RIL ها، داریم: $C_\infty + D_\infty = 1$ و $C_\infty - D_\infty = c_\infty$ که پیشنهاد می‌کند $D_\infty = \frac{1}{2}(1-c_\infty)$. توجه کنید که:

$$c_\infty = c_\infty + \lambda d_\infty = c_1 + \lambda d_1 = \frac{2r}{1+2r}$$

زمانی که $d_\infty = 0$ و $c_1 = 0$، $d_1 = 2$ است؛ بنابراین، ساده و مشخص است که عبارت زیر را خواهیم داشت:

$$D_\infty = \frac{1}{2}\left(1 - \frac{1-2r}{1+2r}\right) = \frac{2r}{1+2r} \qquad (۵-۳)$$

سهم D_∞ فقط فراوانی زیگوت‌های نوترکیب برای دو نشانگر A و B در RIL ها است که با R تعریف می‌شود. بر اساس رابطه بین دو پارامتر R و r در معادله (۵-۳)، نرخ نوترکیبی را از روی نسبت هموزیگوت‌های نوترکیب به صورت زیر محاسبه می‌کنیم:

$$r = \frac{R}{2(1-R)} \qquad (۵-۴)$$

بنابراین، با برآورد R با استفاده از یک مجموعه داده عملی، معادله (۵-۴) می‌تواند موجب برآورد r شود.

مثال ۵-۱. ژانگ و همکاران (۲۰۰۴) نقشه پیوستگی مولکولی سویا را با استفاده از ۱۸۴ RIL مشتق شده از ۷ نسل خودگشنی F_1 بین واریته‌های Kefeng No. 1 و Nannong 1138-2 گزارش کردند. نقشه، ۳۵۹۶ سانتی مورگان از ژنوم سویا را با ۴۵۲ نشانگر که روی ۲۱ گروه پیوستگی قرار داشتند، پوشش داد. دو جفت نشانگر، sat_300 (A) در برابر sat_384 (B) و sat_384 (C) در برابر sat_265 (D) برای تجزیه پیوستگی با استفاده از مشاهدات ارایه شده در جدول ۵-۱ انتخاب شدند.

جفت نشانگرهای بالا را در نظر بگیرید. درستنمایی مشاهدات نشانگر برای نسبت هموزیگوت‌های نوترکیب (R) می‌تواند به صورت زیر ایجاد شود:

$$L(R) = \frac{n!}{n_{22}!\,n_{20}!\,n_{02}!\,n_{00}!} R^{n_{20}+n_{02}} (1-R)^{n_{22}+n_{00}}$$

بیشینه نمودن لگاریتم درستنمایی منجر به برآورد حداکثر درستنمایی (MLE) مربوط به R به صورت زیر می‌شود:

$$\hat{R} = \frac{n_{20}+n_{02}}{n} \qquad (۵-۵)$$

معادله (۵-۵) برای برآورد سهم زیگوت‌های کراس‌اور به صورت $\widehat{R}=\frac{7+7}{175}=0.080$ برای اولین (حالت سمت چپ) جفت و $\widehat{R}=\frac{0+6}{175}=0.034$ برای دومین (حالت سمت راست) جفت مورد استفاده قرار می‌گیرد. سپس، معادله (۵-۳) برای برآورد نرخ‌های نوترکیبی به صورت $\hat{r}=\frac{0.080}{2(1-0.080)}=0.043$ برای اولین جفت نشانگر و $\hat{r}=\frac{0.034}{2(1-0.034)}=0.018$ برای دومین جفت نشانگر استفاده می‌شود.

رابینز[1] (۱۹۱۸) سهم انواع زیگوت‌های مختلف را در هر نسل از روی معادله (۵-۱) به‌دست آورد که به صورت زیر بیان می‌شوند:

$$C_n=\frac{1-\left(\frac{1}{2}-r\right)^t}{1+2r}+\frac{1}{2}\left(\frac{1}{2}-r+r^2\right)^{t-1}-\left(\frac{1}{2}\right)^{t-1}$$

$$D_n=\frac{2r+\left(\frac{1}{2}-r\right)^t}{1+2r}+\frac{1}{2}\left(\frac{1}{2}-r+r^2\right)^{t-1}-\left(\frac{1}{2}\right)^{t-1}$$

با استفاده از این دو معادله و نرخ‌های نوترکیبی برآورد شده، سهم هموزیگوت‌های غیرنوترکیب و نوترکیب در نسل‌های ۲ تا ۱۰ (جدول ۵-۲) برآورد شدند. الگوهای پویای این نسبت‌ها برای پیش بینی ترکیبات ژنتیکی در طول نسل‌ها مفید خواهد بود.

سهم هموزیگوت‌های مضاعف تا حد زیادی در نسل‌های اولیه افزایش می‌یابند. درجه و دوره این افزایش بستگی به میزان نرخ نوترکیبی دارد. در مقایسه با پیوستگی ضعیف، پیوستگی قوی با درجات بیشتر افزایش و تعداد کمتر نسل‌هایی که در طی آن، سهم هموزیگوت‌ها به مقدار ثابت و پایدار می‌رسند، مرتبط است.

جدول ۵-۱- مشاهدات هموزیگوت‌های غیرنوترکیب و نوترکیب در دو جفت نشانگر در جمعیت RIL حاصل از خودگشنی سویا.

نشانگر sat_300	نشانگر sat_384	
	BB	bb
AA	n_{22} = ۸۷	n_{20} = ۷
aa	n_{02} = ۷	n_{00} = ۷۴

[1] - Robbins

نشانگر sat_384	نشانگر sat_265	
	DD	dd
CC	n_{22} = ۹۹	n_{20} = ۰
cc	n_{02} = ۶	n_{00} = ۷۰

۵-۲-۲- تجزیه سه نقطه‌ای

سهم هموزیگوت‌های نوترکیب

مجدداً سه نشانگر را با ترتیب A-B-C در نظر می‌گیریم. در یک جمعیت RIL، ۸ هاپلوتیپ، ABC، abc، ABc، abC، aBC، Abc، AbC، aBc یا ژنوتیپ‌های AABBCC، aabbcc، AABBcc، aabbCC، aaBBCC، AAbbcc، AAbbCC و aaBBcc وجود دارند. فراوانی هاپلوتیپ‌ها به چهار گروه به صورت $P(ABC)=P(abc)$، $P(ABc)=P(abC)$، $P(aBC)=P(Abc)$، $P(Abc)=P(aBC)$، $P(AbC)=P(aBc)$ دسته‌بندی می‌شوند که به ترتیب نشان می‌دهند که یک هاپلوتیپ دارای ۰ کراس‌اور در فواصل نشانگری A-B و B-C، فقط یک کراس‌اور در فاصله دوم، فقط یک کراس‌اور در فاصله اول و دو کراس‌اور هر یک در فواصل مختلف است. فرض کنید $g_{(j_1j_2)}$ ($\sum_{j_1,j_2} g_{(j_1j_2)}=1$) احتمال آن باشد که یک هاپلوتیپ دارای j_1(صفر یا ۱) کراس‌اور در اولین فاصله نشانگری و j_2(صفر یا ۱) کراس‌اور در دومین فاصله نشانگری است. از $n_{(j_1j_2)}$ ($\sum_{j_1,j_2} n_{(j_1j_2)}=n$) برای نشان دادن مشاهدات ژنوتیپی مربوطه استفاده می‌کنیم.

تابع درستنمایی گروه‌های ژنوتیپی نشانگر می‌تواند به‌صورت زیر فرموله شود:

$$L\left(n_{(j_1j_2)}\right)=\frac{n!}{n_{(00)}!n_{(01)}!n_{(10)}!n_{(11)}!}g_{(00)}^{n_{(00)}}g_{(01)}^{n_{(01)}}g_{(10)}^{n_{(10)}}g_{(11)}^{n_{(11)}} \qquad (۶-۵)$$

که MLE مربوط به $g_{(j_1j_2)}$ به صورت زیر به‌دست می‌آید:

$$\hat{g}_{(j_1j_2)}=\frac{n_{(j_1j_2)}}{n} \qquad (۷-۵)$$

که در آن:

$$\begin{aligned} n_{(00)}&=n_{ABC}+n_{abc}\\ n_{(01)}&=n_{ABc}+n_{abC}\\ n_{(10)}&=n_{Abc}+n_{aBC}\\ n_{(11)}&=n_{AbC}+n_{aBc} \end{aligned} \qquad (۸-۵)$$

ما از نشانه‌گذاری مخصوص آللی برای نشان دادن مشاهدات نشانگرها استفاده می‌کنیم. جدول ۵-۳ ساختار داده‌های مشاهدات ژنوتیپی را برای سه نشانگر به هم پیوسته A، B و C نشان می‌دهد.

به‌صورت اختصاصی‌تر، احتمالات وقوع کراس‌اور می‌تواند بر حسب سهم زیگوت‌های نتاج بیان شود، یعنی:

$$\hat{g}_{(00)} = \frac{n_{ABC} + n_{abc}}{n} = \hat{P}(ABC) + \hat{P}(abc)$$
$$\hat{g}_{(01)} = \frac{n_{ABc} + n_{abC}}{n} = \hat{P}(ABc) + \hat{P}(abC)$$
$$\hat{g}_{(10)} = \frac{n_{Abc} + n_{aBC}}{n} = \hat{P}(Abc) + \hat{P}(aBC)$$
$$\hat{g}_{(11)} = \frac{n_{AbC} + n_{aBc}}{n} = \hat{P}(AbC) + \hat{P}(aBc)$$

اگر R_{AB}، R_{BC} و R_{AC} به ترتیب به عنوان سهم هموزیگوت‌های نوترکیب بین نشانگرهای A و B، بین B و C و بین A و C باشد، بر اساس تعریف نرخ نوترکیبی، داریم:

$$\begin{aligned} R_{AB} &= g_{(10)} + g_{(11)} \\ R_{BC} &= g_{(01)} + g_{(11)} \\ R_{AC} &= g_{(01)} + g_{(10)} \end{aligned} \qquad (۹-۵)$$

بنابراین دستیابی به معادلات زیر به سهولت انجام می‌گیرد:

$$\begin{aligned} g_{(00)} &= 2P(ABC) = 2P(abc) = \frac{1}{2}\left(2 - R_{AB} - R_{BC} - R_{AC}\right) \\ g_{(01)} &= 2P(ABc) = 2P(abC) = \frac{1}{2}\left(-R_{AB} + R_{BC} + R_{AC}\right) \\ g_{(10)} &= 2P(Abc) = 2P(aBC) = \frac{1}{2}\left(R_{AB} - R_{BC} + R_{AC}\right) \\ g_{(11)} &= 2P(AbC) = 2P(aBc) = \frac{1}{2}\left(R_{AB} + R_{BC} - R_{AC}\right) \end{aligned} \qquad (۱۰-۵)$$

و

$$\begin{aligned} R_{AB} &= P(AbC) + P(Abc) + P(aBC) + P(aBc) = 2P(AbC) + 2P(Abc) \\ R_{BC} &= P(AbC) + P(abC) + P(ABc) + P(aBc) = 2P(AbC) + 2P(ABc) \\ R_{AC} &= P(ABc) + P(Abc) + P(aBC) + P(abC) = 2P(ABc) + 2P(Abc) \end{aligned}$$

بنابراین، سهم هموزیگوت‌های نوترکیب می‌تواند از روی فراوانی‌های هاپلوتیپی برآورد شود.

آزمون ترتیب ژنی در RIL

از تجزیه سه نقطه‌ای می‌توان برای یافتن محتمل‌ترین ترتیب نشانگرها بر اساس سهم هموزیگوت‌های نوترکیب مربوطه استفاده نمود. سه نشانگر مرتب نشده A، B و C را در نظر بگیرید که مشاهدات آن‌ها در جدول ۵-۳ ارایه شده است. همانگونه که در بخش فوق نشان داده شد، ۸ ژنوتیپ سه نقطه‌ای در چهار گروه مرتب شده‌اند و اندازه نمونه برای هر گروه با $n_{j_1 j_2}$ نمایش داده می‌شود. با فرض سه ترتیب نشانگری ممکن و مختلف، فراوانی مورد انتظار گامت‌ها برای هر ترتیب در جدول ۵-۴ ارایه شده است.

بر اساس جدول ۵-۴، سه درستنمایی برای هر نوع ترتیب ویژه را بدین صورت ایجاد می‌کنیم:

$$\begin{aligned} L_{ABC} &\propto R_{AB}^{n_{10}+n_{11}} R_{BC}^{n_{01}+n_{11}} \left(1-R_{AB}\right)^{n_{00}+n_{01}} \left(1-R_{BC}\right)^{n_{00}+n_{10}} \\ L_{ACB} &\propto R_{AC}^{n_{01}+n_{10}} R_{BC}^{n_{01}+n_{11}} \left(1-R_{AC}\right)^{n_{00}+n_{11}} \left(1-R_{BC}\right)^{n_{01}+n_{11}} \\ L_{BAC} &\propto R_{AB}^{n_{10}+n_{11}} R_{AC}^{n_{01}+n_{10}} \left(1-R_{AB}\right)^{n_{00}+n_{01}} \left(1-R_{AC}\right)^{n_{00}+n_{11}} \end{aligned} \qquad (۱۱-۵)$$

هر درستنمایی منفرد سبب تولید MLE نسبت هموزیگوت‌های نوترکیب برای هر ترتیب می‌شود که عبارت است از:

$$\begin{aligned} \widehat{R}_{AB} &= \frac{n_{(10)}+n_{(11)}}{n} \\ \widehat{R}_{BC} &= \frac{n_{(01)}+n_{(11)}}{n} \\ \widehat{R}_{AC} &= \frac{n_{(01)}+n_{(10)}}{n} \end{aligned} \qquad (۱۲-۵)$$

این برآوردها در درستنمایی رابطه (۵-۱۱) وارد می‌شوند و مقادیر درستنمایی برای ترتیب‌های مختلف محاسبه می‌شوند. محتمل‌ترین ترتیب نشانگری دارای بیشترین مقدار درستنمایی است.

جدول ۵-۲- سهم هتروزیگوت‌های غیرنوترکیب و نوترکیب در نسل‌های مختلف برای جمعیت RIL حاصل از خودگشنی.

سهم یا نسبت	نسل								
	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
sat_300-sat_384									
C_t	۰/۴۵۷	۰/۶۷۷	۰/۷۷۸	۰/۸۲۴	۰/۸۴۵	۰/۸۵۴	۰/۸۵۹	۰/۸۶۱	۰/۸۶۲
D_t	۰/۰۰۱	۰/۰۱۲	۰/۰۱۸	۰/۰۲۰	۰/۰۲۱	۰/۰۲۲	۰/۰۲۲	۰/۰۲۲	۰/۰۲۲
sat_384-sat_265									
C_t	۰/۴۸۲	۰/۷۱۹	۰/۸۳۴	۰/۸۸۹	۰/۹۱۶	۰/۹۲۸	۰/۹۳۵	۰/۹۳۸	۰/۹۳۹
D_t	۰/۰۰۰	۰/۰۰۵	۰/۰۰۷	۰/۰۰۸	۰/۰۰۹	۰/۰۰۹	۰/۰۰۹	۰/۰۰۹	۰/۰۰۹

جدول ۵-۳- مشاهدات هموزیگوت‌های غیرنوترکیب و نوترکیب برای سه نشانگر A، B و C با ترتیب ناشناخته در جمعیت RIL حاصل از خودگشنی.

نشانگر A	نشانگر B	نشانگر C	
		CC	cc
AA	BB	n_{ABC}	n_{ABc}
AA	bb	n_{AbC}	n_{Abc}
aa	BB	n_{aBC}	n_{aBc}
aa	bb	n_{abC}	n_{abc}

جدول ۵-۴- فراوانی‌های مورد انتظار گامت‌ها برای سه ترتیب نشانگری ممکن مختلف که بر حسب نسبت زیگوت‌های نوترکیب بیان شده‌اند.

نوع گامت	مشاهده	فراوانی مورد انتظار در هر ترتیب		
		A-B-C	A-C-B	B-A-C
ABC یا abc	$n_{(00)}$	$(1-R_{AB})(1-R_{BC})$	$(1-R_{AC})(1-R_{BC})$	$(1-R_{AB})(1-R_{AC})$
ABc یا abC	$n_{(01)}$	$(1-R_{AB})R_{BC}$	$R_{AC}R_{BC}$	$(1-R_{AB})R_{AC}$
Abc یا aBC	$n_{(10)}$	$R_{AB}(1-R_{BC})$	$R_{AC}(1-R_{BC})$	$R_{AB}R_{AC}$
AbC یا aBc	$n_{(11)}$	$R_{AB}R_{BC}$	$(1-R_{AC})R_{BC}$	$R_{AB}(1-R_{AC})$

فواصل ژنتیکی و تابع نقشه‌یابی

اگر r_{AB}، r_{BC} و r_{AC} به ترتیب نرخ‌های نوترکیبی بین نشانگرهای A و B، بین B و C و بین A و C باشند، با این فرض که هیچ‌گونه تداخلی در پیشامدهای کراس‌اوری طی میوز وجود ندارد، رابطه بین سه نرخ نوترکیبی به صورت زیر تعیین می‌شود:

$$r_{AC} = r_{AB} + r_{BC} - 2r_{AB}r_{BC} \qquad (۱۳-۵)$$

یا

$$(1-2r_{AC}) = (1-2r_{AB})(1-2r_{BC})$$

بر اساس معادله (۵-۴)، رابطه زیر را بین سه R خواهیم داشت:

$$\left(\frac{1-2R_{AC}}{1-R_{AC}}\right) = \left(\frac{1-2R_{AB}}{1-R_{AB}}\right)\left(\frac{1-2R_{BC}}{1-R_{BC}}\right) \qquad (۱۴-۵)$$

که می‌تواند بدین صورت تغییر یابد:

$$(1-2R_{AC}) = \frac{(1-2R_{AB})(1-2R_{BC})}{1-2R_{AB}R_{BC}} \qquad (۱۵-۵)$$

یا

$$(1-R_{AC}) = \frac{(1-R_{AB})(1-R_{BC})}{1-2R_{AB}R_{BC}}$$

به‌علاوه داریم:

$$R_{AC} = \frac{R_{AB} + R_{BC} - 3R_{AB}R_{BC}}{1-2R_{AB}R_{BC}} \qquad (۱۶-۵)$$

در حالت عدم وجود تداخل در هر میوز، فاصله ژنتیکی (d) می‌تواند بر حسب نرخ نوترکیبی (r) از راه تابع نقشه هالدین $d=-\frac{1}{2}\ln\left(1-2r\right)$ بیان شود. بر اساس این تابع، فاصله ژنتیکی با استفاده از نسبت هموزیگوت‌های نوترکیب به‌صورت زیر محاسبه می‌شود:

$$d=-\frac{1}{2}\ln\left[\frac{1-2R}{1-R}\right] \qquad (۱۷-۵)$$

عدم استقلال نوترکیبی‌ها

زمانی که هیچ‌گونه تداخلی بین فواصل نشانگری مختلف وجود نداشته باشد، نرخ‌های نوترکیبی از رابطه (۵-۱۳) تبعیت می‌کنند. واضح است که این رابطه برای نسبت هموزیگوت‌های نوترکیب صادق نیست، یعنی:

$$R_{AC}\neq R_{AB}+R_{BC}-2R_{AB}R_{BC}$$

این بدان معناست که حتی اگر هیچ‌گونه تداخلی در هر میوز وجود نداشته باشد، هنوز نوترکیبی در فواصل مختلف در RIL ها مستقل نیست. درجه عدم استقلال در داده‌های RIL که در آن، هیچ گونه تداخلی در میوز وجود ندارد، به‌صورت زیر اندازه‌گیری می‌شود:

$$\rho=\frac{R_{AB}+R_{BC}-R_{AC}}{2R_{AB}R_{BC}}$$

$$=\frac{2g_{(11)}}{2R_{AB}R_{BC}}$$

در مجموع، از چهار نسبت زیر به ترتیب برای اندازه‌گیری درجه عدم استقلال زیگوت‌های نوترکیب دوگانه ($\rho_{(00)}$)، زیگوت‌های غیرنوترکیب دوگانه ($\rho_{(11)}$) و دو زیگوت نوترکیب منفرد ($\rho_{(01)}$ و $\rho_{(10)}$) استفاده نمودیم:

$$\rho_{(11)} = \frac{g_{(11)}}{R_{AB}R_{BC}} \geq 1$$
$$\rho_{(10)} = \frac{g_{(10)}}{R_{AB}\left(1-R_{BC}\right)} \leq 1$$
$$\rho_{(01)} = \frac{g_{(01)}}{\left(1-R_{AB}\right)R_{BC}} \leq 1 \qquad (۱۸-۵)$$
$$\rho_{(00)} = \frac{g_{(00)}}{\left(1-R_{AB}\right)\left(1-R_{BC}\right)} \geq 1$$

آزمون تداخل میوزی

در معادلات بالا، فرض کردیم که هیچ تداخلی در میوز وجود ندارد، ولی این فرض ممکن است برای مجموعه داده‌های RIL صحیح نباشد. مارتین و هاسپیتال (۲۰۰۶) ضریب تطابق مولر (۱۹۱۶) را در میوزهای منفرد برای آزمودن تداخل حقیقی در هر میوز برای داده‌های RIL با استفاده از معادله زیر تعمیم دادند:

$$I = 1 - \frac{\hat{g}_{(11)}}{\breve{g}_{(11)}}$$
$$= 1 - \frac{\left(1-2R_{AB}R_{BC}\right)\left(R_{AB}+R_{BC}-R_{AC}\right)}{3-2R_{AB}-2R_{BC}} \qquad (۱۹-۵)$$

در این رابطه، $\hat{g}_{(11)}$ فراوانی مشاهده شده زیگوت‌های دوگانه نوترکیب و $\breve{g}_{(11)}$ فراوانی زیگوت‌های دوگانه نوترکیب در شرایطی است که تداخل وجود ندارد. بدون وجود تداخل، R_{AC} زیگوت‌های نوترکیب دوگانه را با معادله (۵-۱۶) نشان دادیم که همان $\breve{g}_{(11)}$ است. اگر I = ۰ باشد، تداخل وجود ندارد.

مثال ۵-۲. ژانگ و همکاران (۲۰۰۴) نقشه پیوستگی مولکولی سویا را با استفاده از ۱۸۴ RIL حاصل از ۷ نسل خودگشنی F_1 بین واریته‌های Kefeng No. 1 و Nannong 1138-2 گزارش نمودند. نقشه ۳۵۹۶ سانتی مورگان از ژنوم سویا را پوشش داد که در آن ۴۵۲ نشانگر در ۲۱ گروه پیوستگی قرار داشتند. مجموعه سه نشانگر sat_300 (A)، sat_384 (B) و sat_265 (C)، با مشاهدات جدول ۵-۵ برای تجزیه پیوستگی انتخاب می‌شوند:

جدول ۵-۵- مشاهدات هموزیگوت‌های غیرنوترکیب و نوترکیب برای سه ترتیب ناشناخته نشانگری A، B و C در جمعیت RIL حاصل از خودگشنی در سویا.

نشانگر A	نشانگر B	نشانگر C	
		CC	cc
AA	BB	n_{ABC} = ۸۷	n_{ABc} = ۰
AA	bb	n_{AbC} = ۰	n_{Abc} = ۷
aa	BB	n_{aBC} = ۷	n_{aBc} = ۰
aa	bb	n_{abC} = ۶	n_{abc} = ۶۱

بـر اسـاس مشـاهدات ژنـوتیپی، $n_{(00)} = n_{ABC} + n_{abc} = 148$، $n_{(01)} = n_{ABc} + n_{abc} = 6$، $n_{(10)} = n_{Abc} + n_{aBC} = 14$ و $n_{(11)} = n_{AbC} + n_{aBc} = 0$ را با استفاده از معادله (۵-۸) محاسبه کردیم که بـرای محاسـبه سـه نسـبت هموزیگـوت‌هـای نوترکیـب بـه‌صـورت $\widehat{R}_{AB} = 0.0833$، $\widehat{R}_{BC} = 0.0357$ و $\widehat{R}_{AC} = 0.1190$ با استفاده از معادله (۵-۱۲) به‌کـار مـی‌رونـد. بـا در نظـر گرفتن توابع درستنمایی (رابطه ۵-۱۱)، مقـادیر درسـتنمایی بـرای سـه ترتیـب ممکـن نشـانگری به‌صورت زیر خواهد بود:

برای ترتیب A-B-C:

$$L_1 \propto (148+6)\log(1-0.08)+(148+14)\log(1-0.04)+(14+0)\log(0.08)$$
$$+(6+0)\log(0.04) = -74.07$$

برای ترتیب A-C-B:

$$L_2 \propto (148+0)\log(1-0.12)+(148+14)\log(1-0.04)+(6+14)\log(0.12)$$
$$+(6+0)\log(0.04) = -87.21$$

برای ترتیب B-A-C:

$$L_3 \propto (148+6)\log(1-0.08)+(148+0)\log(1-0.12)+(14+0)\log(0.08)$$
$$+(6+14)\log(0.12) = -109.51$$

بر طبق مقادیر درستنمایی بالا، A-B-C به عنوان ترتیب بهینه برای این نشانگرها تعیین مـی-شود.

با استفاده از $\widehat{g}_{(j_1 j_2)}$ که از روی معادله (۵-۷) با ترتیب نشانگری A-B-C محاسبه شـد، سـهم هموزیگوت‌های نوترکیب را با استفاده از معادله (۵-۹) و نرخ‌های نوترکیبی و فواصل نقشـه‌ای را بـا استفاده از معادلات (۵-۴) و (۵-۱۷) به‌صورت زیر محاسبه می‌نماییم:

	A-B	B-C	A-C
R_{AB}	۰/۰۸۳۳	۰/۰۳۵۷	۰/۱۱۹۰ *۰/۱۱۰۸
r_{AB}	۰/۰۴۵۴	۰/۰۱۸۵	۰/۰۶۷۵ *۰/۰۶۲۳
d_{AB}	۴/۷۶	۱/۸۹	۷/۲۶ *۶/۶۵

مقادیر نشان داده شده با ستاره از روی رابطه بین سه R با استفاده از معادلـه (۵-۱۶) محاسـبه شدند. بر اساس مقادیر R برآورد شده، آزمون عدم استقلال نوترکیبی بـا معادلـه (۵-۱۸) و تـداخل میوزی با معادله (۵-۱۹) به صورت زیر انجام شد:

	مقدار یا ارزش
$\rho_{(11)}$	۰
$\rho_{(10)}$	۱/۰۳۷۰
$\rho_{(01)}$	۱/۰۹۰۹
$\rho_{(00)}$	۰/۹۲۶۷
I	۱/۰۰۰۰

مشاهده می‌شود که فواصل نشانگری نزدیک و مجاور، درجاتی از عدم اسـتقلال و تـداخل را در بین نشانگرهای sat_300، sat_384 و sat_256 در RILs سویا نشان می‌دهند.

۵-۳- RIL های حاصل از آمیزش‌های برادر- خواهری

۵-۳-۱- تجزیه دو نقطه‌ای

گونه‌های زیادی نظیر موش وجود دارند که خودگشنی برای آن‌ها امکان ندارد. بـرای ایـن گونـه‌هـا، RIL ها می‌توانند با آمیزش برادر- خواهری مکرر نتـاج F_1 حاصـل از تلاقـی بـین دو لایـن اینبـرد AABB... و aabb... تولید شوند. مشابه با RIL های حاصل از خودگشنی، هالدین و وادینگتون[1] (۱۹۳۱) ارتباط بین نرخ نوترکیبی و نسبت هموزیگـوت‌هـای نوترکیـب را در RIL هـای حاصـل از آمیزش برادر- خواهری به‌دست آوردند. با در نظر گرفتن دو نشانگر روی کرومـوزوم‌هـای اتـوزومی، آمیزش برادر- خواهری در F_1 منجر به ایجاد ۱۰ دیپلوتیپ مختلف در F_2 می‌شود کـه بـرای آن‌هـا،

[1] - Waddington

مجموعاً ۵۵ = $\frac{1}{2}(10\times11)$ نوع آمیزش برادر- خواهری برای ایجاد نسل F_3 وجود دارد. هالدین و وادینگتون (۱۹۳۱) نشان دادند که زمانی آمیزش برادر- خواهری برای نسل‌های زیادی صورت می‌گیرد، ارتباط بین نسبت هموزیگوت‌های نوترکیب (R) و نرخ نوترکیبی (r) از شکل‌های زیر تبعیت می‌کند:

$$R=\frac{4r}{1+6r} \qquad (۲۰-۵)$$

$$r=\frac{R}{2(2-3R)} \qquad (۲۱-۵)$$

در خصوص تجزیه برادر- خواهری، ژن‌های وابسته به جنس موجود روی کروموزوم X به‌طور متفاوتی از ژن‌های اتوزومی به ارث می‌رسند. آمیزش اصلی ژن‌های وابسته به جنس به صورت AABB.×ab… است. هالدین و وادینگتون (۱۹۳۱) نیز روابط مشابهی را برای دو ژن وابسته به جنس به‌صورت زیر بیان کردند:

$$R=\frac{8r}{3(1+4r)} \qquad (۲۲-۵)$$

$$r=\frac{3R}{4(2-3R)} \qquad (۲۳-۵)$$

۵-۳-۲- تجزیه سه نقطه‌ای

روش تجزیه سه نقطه‌ای در RIL های حاصل از خودگشنی می‌تواند برای RIL های حاصل از آمیزش برادر- خواهری تغییر یابد. سه نشانگر به صورت A-B-C در نظر بگیرید. با فرض عدم وجود تداخل میوزی، سه نسبت از هموزیگوت‌های نوترکیب برای نشانگرهای A، B و C در مورد آمیزش برادر- خواهری از رابطه زیر تبعیت می‌کنند:

$$\left(\frac{2-4R_{AC}}{2-3R_{AC}}\right)=\left(\frac{2-4R_{AB}}{2-3R_{AB}}\right)=\left(\frac{2-4R_{BC}}{2-3R_{BC}}\right) \qquad (۲۴-۵)$$

$$\left(1-2R_{AC}\right)=\frac{\left(1-2R_{AB}\right)\left(1-2R_{BC}\right)}{1-3R_{AB}R_{BC}} \qquad (۲۵-۵)$$

یا

$$R_{AC}=\frac{2R_{AB}+2R_{BC}-7R_{AB}R_{BC}}{2\left(1-3R_{AB}R_{BC}\right)} \qquad (۲۶-۵)$$

فاصله ژنتیکی از نسبت هموزیگوت‌های نوترکیب با تابع نقشه هالدین به‌صورت زیر محاسبه می‌شود:

$$d = -\frac{1}{2}\ln\left[\frac{2-4R}{2-3R}\right] \qquad (۲۷-۵)$$

۵-۴- کاهش اریبی

۵-۴-۱- RIL های حاصل از خودگشنی

مارتین و هاسپیتال (۲۰۰۶) مشخص کردند که برآورد نرخ نوترکیبی (r) از نسبت هموزیگوت‌های نوترکیب برآورد شده (R) با استفاده از معادله (۵-۴) به صورت زیر نوشته می‌شود:

$$\tilde{r} = \frac{\hat{R}}{2\left(1-\hat{R}\right)} \qquad (۲۸-۵)$$

این رابطه دارای اریبی است، چون این دو پارامتر (R و r) به‌طور خطی با یکدیگر مرتبط نیستند. برآورد نااریب $\hat{R}$ می‌تواند با معادله (۵-۵) به‌دست آید. در نتیجه، $\hat{R}$ می‌تواند از مجموع مقدار حقیقی و باقیمانده تصادفی با میانگین صفر (یعنی $\hat{R} = R + e\left(0,\sigma^2\right)$) به‌دست آید که $e\left(0,\sigma^2\right)$ نشان‌دهنده متغیر تصادفی با میانگین صفر و واریانس σ^2 است.

با بازنویسی برآوردگر r از راه معادله (۵-۲۸) و سپس انجام بسط سری‌های تیلور در $\in_R \equiv \frac{\left(\hat{R}-R\right)}{\left(1-R\right)}$ داریم:

$$\tilde{r} = \frac{\hat{R}}{2\left(1-\hat{R}\right)}$$

$$= -\frac{1}{2} + \frac{1}{2\left(1-\hat{R}\right)}$$

$$= \frac{R + \in_R + \in_R^2 + \in_R^3 + \ldots}{2\left(1-R\right)}$$

چون $\hat{R}$ نااریب است، ارزش مورد انتظار $\tilde{r}$ می‌تواند به صورت زیر به‌دست آید:

$$E\left[\tilde{r}\right]=r+\frac{E\left[\left(\hat{R}-R\right)^2\right]}{2\left(1-R\right)^3}+\ldots$$

که عبارات با ترتیب بزرگ‌تر با گشتاورهای درجه بالاتر $\left(\hat{R}-R\right)$ مرتبط هستند. چون امید ریاضی $E\left[\left(\hat{R}-R\right)^2\right]$ با $\frac{\hat{R}\left(1-\hat{R}\right)}{n}$ برآورد می‌شود، برآوردگر r که بیشترین اریبی آن برطرف شده است، می‌تواند به‌صورت زیر برآورد شود:

$$\begin{aligned}\hat{r}&=\frac{\hat{R}}{2\left(1-\hat{R}\right)}-\frac{\hat{R}\left(1-\hat{R}\right)}{2n\left(1-\hat{R}\right)^3}\\&=\tilde{r}\left[1-\frac{1}{n-m}\right]\\&=\frac{m\left(n-m-1\right)}{2\left(n-m\right)^2}\end{aligned}\qquad(۲۹-۵)$$

که m تعداد زیگوت‌های نوترکیب در RIL ها است. برآوردگر r هنوز دارای اریبی است (اگرچه اریب باقیمانده در حال حاضر فقط در ترتیب $\frac{1}{n^2}$ است)، مگر در شرایطی که تمامی تصحیحات رتبه‌ای در نظر گرفته شوند.

مثال ۵-۳. مثال ۵-۱ را مجدداً مشاهده کنید. نرخ‌های نوترکیبی برای دو جفت نشانگر برای مقادیر تصحیح نشده ($\tilde{r}$) و تصحیح شده ($\hat{r}$) به ترتیب با استفاده از معادلات (۵-۲۸) و (۵-۲۹) برآورد شدند و به‌صورت زیر ارائه می‌شوند:

جفت نشانگر	$\tilde{r}$	$\hat{r}$
sat_300-sat_384	۰/۰۴۳۵	۰/۰۴۳۲
sat_384-sat_265	۰/۰۱۷۸	۰/۰۱۷۶

معادله تصحیح نشده، نرخ نوترکیبی را برای دو جفت نشانگر در این مثال نسبتاً بیشتر از حد برآورد می‌نماید؛ بنابراین، در عمل، معادله تصحیح شده توصیه می‌شود، به‌ویژه برای وقتی که اندازه نمونه کوچک بوده و پیوستگی ضعیف باشد.

۵-۴-۲- RIL های حاصل از آمیزش برادر- خواهری

مارتین و هاسپیتال (۲۰۰۶) نیز روشی را برای کاهش اریبی برآوردگر r حاصل از R در RIL های مشتق از آمیزش برادر- خواهری به‌دست آوردند. معادله تصحیح شده برای از بین بردن اریبی اصلی در برآورد نرخ نوترکیبی به‌صورت زیر ارائه می‌شود:

$$\begin{aligned}\hat{r} &= \frac{\hat{R}}{4-6\hat{R}} - \frac{24\hat{R}\left(1-\hat{R}\right)}{N\left(4-6\hat{R}\right)^3} \\ &= \tilde{r}\left[1-\frac{6(n-m)}{(2n-m)^2}\right] \\ &= \frac{m}{4n-m}\left[1-\frac{6(n-m)}{(2n-3m)^2}\right]\end{aligned} \qquad (۵-۳۰)$$

معادلات (۵-۲۹) و (۵-۳۰) زمانی برای کاهش اریبی به‌دست می‌آیند که نرخ نوترکیبی با استفاده از نسبت هموزیگوت‌های نوترکیب برآورد شود. در بخش نکات ۵-۷، یک روش کلی برای کاهش اریبی ارایه خواهد شد که خوانندگان علاقه‌مند می‌توانند هر نوع فرمول تصحیح شده مفیدی را از روی آن استنباط نمایند.

۵-۵- RIL های چند مسیری

سویه‌های اینبرد نوترکیب از خویش‌آمیزی مکرر نتاج F_1 حاصل از تلاقی بین دو لاین همخون ۱ و ۲ به‌دست می‌آیند. RIL های حاصل از دو لاین اینبرد، RIL های دو مسیری[1] (۲×۱) نامیده می‌شوند. باتوجه به پیچیدگی ژنوم، RIL ها از روی تلاقی‌های چندگانه بین تعداد زیادی از لاین‌های اینبرد تولید می‌شوند. به‌تازگی RIL های چهار مسیری، (۲×۱)× (۴×۳) و RIL های هشت مسیری $[(1\times2)\times(3\times4)]\times[(5\times6)\times(7\times8)]$، در گونه‌های مدل نظیر موش (تریجیل و همکاران[2]، ۲۰۰۲؛ کنسرسیوم صفات پیچیده، ۲۰۰۴) تولید شده‌اند. شکل ۵-۱ روشی را برای تولید RIL های هشت مسیری نشان می‌دهد که در آن، نتاج از روی ۸ لاین اینبرد والدینی تولید می‌شوند که به‌طور مکرر خودگشن شده یا آمیزش های برادر- خواهری نموده‌اند. این روش لاین‌های اینبرد جدیدی تولید خواهد کرد که ژنوم آن ترکیبی از ۸ سویه والدینی است؛ بنابراین، RIL های دارای مسیرهای بیشتر می‌توانند دارای توانایی بیشتری برای مطالعه الگوی تنوع ژنتیکی

---

[1] - Two-way RIL
[2] - Threadgill et al.

در جوامع طبیعی نسبت به منابع رایج مکان‌یابی باشند و حقیقتاً می‌توانند برای تشریح و موشکافی هر گونه بیماری پیچیده انسانی با استفاده از مدل‌های موش بکار گرفته شوند.

بر اساس پژوهش هالدین و وادینگتون (۱۹۳۱)، برومن (۲۰۰۵) یک روش کلی را برای به‌دست آوردن فرمول‌های مربوط به بیان نسبت زیگوت‌های نوترکیب بر حسب نرخ نوترکیبی پیشنهاد داد. این روش می‌تواند برای انواع مختلف RIL ها و همچنین برای ژن‌های به هم پیوسته موجود روی کروموزوم X استفاده شود (جدول ۵-۶).

۵-۶- تمرین‌ها

۵-۱- به‌عنوان دو طرح مکان‌یابی مختلف، شباهت‌ها و تفاوت‌های موجود بین طرح‌های RIL و تلاقی برگشتی را مورد بحث قرار دهید. طرح RIL دارای چه مزیت‌هایی نسبت به طرح تلاقی برگشتی است؟

۵-۲- نشان دهید که فاصله ژنتیکی محاسبه شده از روی R (یعنی $d^* = -\frac{1}{2}\ln(1-2R)$)، افزایشی نیست، یعنی، برای سه نشانگر A-B-C، داریم $d^*_{AC} \neq d^*_{AB} + d^*_{BC}$.

۵-۳- برای طرح تلاقی برگشتی، برآوردهای نرخ نوترکیبی حاصل از تجزیه‌های دو و سه نقطه‌ای، باید در شرایطی که هیچ داده گمشده‌ای وجود ندارد، مشابه باشند. نشان دهید که چگونه برآوردهای نرخ نوترکیبی حاصل از تجزیه‌های دو و سه نقطه‌ای برای طرح RIL متفاوت هستند.

۵-۴- اگر برآورد مشابهی برای نسبت هموزیگوت‌های نوترکیب بین دو نشانگر معین برای RILهای حاصل از خودگشنی و RIL های حاصل از آمیزش برادر- خواهری به‌دست آید، نشان دهید چگونه نرخ‌های نوترکیبی مربوطه برای این دو نوع مختلف RIL متفاوت هستند.

۵-۷- نکته

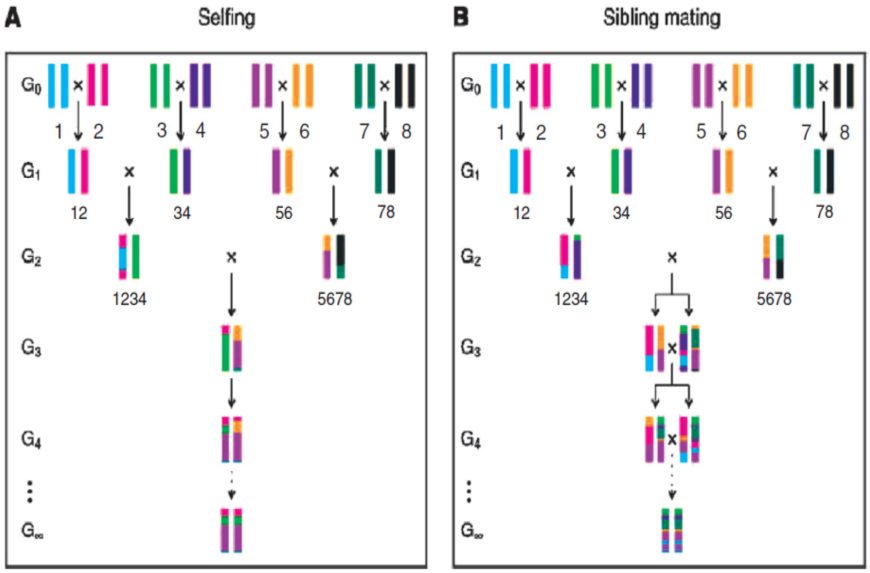
یک روش کلی برای به‌دست آوردن معادله کاهش اریبی ارائه دادیم. یک رابطه اختیاری بین R و r به صورت $r = f(R)$ در نظر بگیرید. برآوردگر نرخ نوترکیبی از راه $\tilde{r} = f(\hat{R})$ دارای اریبی است، چون f غیر خطی است. مقدار اریبی می‌تواند از روی بسط سری تیلور به‌صورت زیر محاسبه شود:

$$E\left[f(\hat{R})\right] = f(R) + f'(R)E\left[\hat{R}-R\right] + \frac{1}{2}f''(R)E\left[\left(\hat{R}-R\right)^2\right] + \ldots$$

چون $\hat{R}$ نااریب است، امید ریاضی $E\left[\hat{R}-R\right]$ به صفر می‌رسد و اریب عمده در برآوردگر r از واریانس $\hat{R}$ ناشی می‌شود که تقریباً برابر $\frac{\hat{R}\left(1-\hat{R}\right)}{n}$ است. با جایگزینی $f(R)$ با $f(\hat{R})$، برآوردگر صحیح عبارت است از:

$$\hat{r} = f(\hat{R}) - \frac{f''(\hat{R})\hat{R}\left(1-\hat{R}\right)}{2n}$$

که در آن، $f''(\hat{R})$ مشتق دوم f ارزیابی شده در نقطه $\hat{R}$ است. توجه کنید که از عبارات با رتبه بالاتر چشم‌پوشی کرده‌ایم. اگرچه این حالت معمولاً به‌عنوان یک مشکل محسوب نمی‌شود، اما رسماً نیاز به تصحیح دارد.



شکل ۵-۱- روشی برای تولید RIL های ۸ مسیری از راه خودگشنی (A) یا آمیزش برادر- خواهری (B).
اقتباس از برومن (۲۰۰۵).

جدول ۵-۶- خلاصه‌ای از روابط بین نرخ نوترکیبی و نسبت هموزیگوت‌های نوترکیب در انواع مختلف RIL.

	خودگشنی	آمیزش برادر- خواهری	
		اتوزوم	کروموزوم X
۲ مسیری	$\frac{2r}{1+2r}$	$\frac{4r}{1+6r}$	$\frac{8r}{3(1+4r)}$
۴ مسیری	$\frac{3r}{1+2r}$	$\frac{6r}{1+6r}$	$\frac{4r}{1+4r}$
۸ مسیری	$\frac{r(4-r)}{1+2r}$	$\frac{7r}{1+6r}$	$\frac{14r}{3(1+4r)}$
۱۶ مسیری	$\frac{r(5-3r+r^2)}{1+2r}$	$\frac{r(8-r)}{1+6r}$	$\frac{r(16-r)}{3(1+4r)}$
2^t مسیری	$1-\frac{(1-r)^{t-1}}{1+2r}$	$1-\frac{(1-r)^{t-2}}{1+6r}$	-

جدول ۴-۷- مقایسه برآورد پیوستگی و ترتیب ژنی در میان سه نشانگر غالب در یک خانواده تنی با n =۱۰۰ با استفاده از مدل مرسوم و یکنواخت.

	M_1-M_2-M_3	M_1-M_3-M_2	M_2-M_1-M_3	مدل یکنواخت
داده شبیه‌سازی شده از $[aaa][ooo]\times[aaa][ooo]$				
ترتیب ژنی صحیح	درست	نادرست	نادرست	
بهترین ترتیب ژنی برآورد شده (٪[a])	100	0	0	
$\hat{r}_{12}$	0.2047±0.0422			0.2048±0.0422
$\hat{r}_{23}$	0.1980±0.0436			0.1985±0.0434
$\hat{r}_{13}$	0.3245±0.0619			0.3235±0.0618
$\Pr ob(M_1-M_2-M_3)(\hat{o}_1)$				0.9860±0.0105
$\Pr ob(M_1-M_3-M_2)(\hat{o}_2)$				0.0060±0.0071
$\Pr ob(M_2-M_1-M_3)(\hat{o}_3)$				0.0080±0.0079
داده شبیه‌سازی شده از $[aao][ooa]\times[aao][ooa]$				
ترتیب ژنی صحیح	درست	نادرست	نادرست	
بهترین ترتیب ژنی برآورد شده (٪[a])	80	11	9	
$\hat{r}_{12}$	0.1991±0.0456	0.8165±0.1003	0.9284±0.0724	0.2104±0.0447
$\hat{r}_{23}$	0.1697±0.0907	0.8220±0.0338	0.1636±0.0608	0.2073±0.0754
$\hat{r}_{13}$	0.3218±0.0755	0.2703±0.0586	0.7821±0.0459	0.2944±0.0929
$\Pr ob(M_1-M_2-M_3)(\hat{o}_1)$				0.9952±0.0058
$\Pr ob(M_1-M_3-M_2)(\hat{o}_2)$				0.0045±0.0058
$\Pr ob(M_2-M_1-M_3)(\hat{o}_3)$				0.0003±0.0015

[a] درصدهای ۲۰۰ شبیه‌سازی که دارای بیشترین درستنمایی برای ترتیب ژنی معین برآورد شده از روش مرسوم هستند. در این مثال که برای آزمون مزیت بکارگیری ترتیب‌های ژنی استفاده شد، فرض شده است که فازهای پیوستگی شناخته شده هستند.

فصل ششم

تجزیهٔ پیوستگی نشانگرهای اریب[1] و با دسته‌بندی نادرست

۱- منظور از نشانگرهای اریب، نشانگرهای دارای انحراف از تفرق مندلی هستند (مترجمین).

۶-۱- مقدمه

تاکنون روش‌های آماری تجزیهٔ پیوستگی نشانگرهای مختلف در یک تلاقی کنترل شده توصیف شدند. ولی این روش‌ها فقط می‌توانند برای نشانگرهایی مناسب باشند که تفرق آن‌ها از نسبت مندلی (۱:۱ برای تلاقی برگشتی یا ۱:۲:۱ برای F_2) تبعیت می‌کند. در یک آزمایش مولکولی عملی، بسیاری از نشانگرها ممکن است به‌واسطه وجود برخی از دلایل ژنتیکی یا زیستی (برای مثال تفاوت در قابلیت بقاء که به صورت توانایی فرد برای زنده ماندن تعریف می‌شود) از نسبت‌های مندلی انحراف یابند. اگر تفرق نشانگر به وسیله آثار بقاء و زنده مانی دچار انحراف شود، می‌توانیم چنین آثاری را با استفاده از روش‌های تغییر یافته در تجزیه پیوستگی وارد کنیم. بایلی[1] (۱۹۶۱) چندین روش را برای تجزیه نشانگرهای اریب با جزئیات بیشتر مورد بحث قرار داد. در این فصل، روش‌های برآورد نرخ نوترکیبی را بین نشانگرهای با قابلیت بقاء متفاوت توضیح خواهیم داد.

انواع متفاوت زنده‌مانی که سبب انحراف از تفرق مندلی می‌شوند، ممکن است به شیوه‌های مختلفی ایجاد شوند. اولاً گامت‌هایی که حامل ژن‌های مختلف هستند، ممکن است قابلیت بقاء نابرابر داشته باشند، به‌طوری که احتمال آمیزش یک نوع نسبت به نوع دیگر بیشتر باشد؛ بنابراین، یک فرد با ژنوتیپ Aa دو نوع گامت مختلف یعنی A و a تولید می‌کنند، ولی ممکن است به‌صورت مساوی در دسترس قرار نگیرند. ثانیاً ممکن است قابلیت بقاء انواع مختلف زیگوت متفاوت باشد. در تلاقی برگشتی Aa×aa، انتظار می‌رود تعداد افراد Aa و aa مساوی باشند، ولی این افراد ممکن است دارای بقاء و زنده‌مانی نابرابر باشند. برای تلاقی برگشتی، تفاوت زنده‌مانی گامتی و زیگوتی اثر یکسانی بر تجزیه پیوستگی دارد. ولی برای F_2، مدل‌های مختلفی برای تشخیص آثار تفاوت در زنده‌مانی گامتی و زیگوتی بر تجزیه پیوستگی مورد نیاز است.

۶-۲- تنوع گامتی در زنده‌مانی

۶-۲-۱- مدل تک ژنی

فرض کنید یکی از دو نشانگر A و B مورد استفاده برای تجزیه پیوستگی، به وسیله قابلیت بقاء متفاوت تحت‌تأثیر قرار گیرد. همچنین، r نرخ نوترکیبی بین دو نشانگر و u قابلیت بقاء آلل A نسبت به آلل جایگزین آن یعنی a برای نشانگر A باشد.

---

[1] - Bailey

تلاقی برگشتی AaBb×aabb

برای تلاقی برگشتی مضاعف AaBb×aabb، تعداد افراد $n=(n_1,n_2,n_3,n_4)$ و فراوانی‌های مورد انتظار هر یک از ۴ ژنوتیپ را به‌صورت زیر مشاهده می‌نماییم:

گامت	فراوانی	
	مورد انتظار	مشاهده شده
AB	$\frac{u}{1+u}(1-r)$	n_1
Ab	$\frac{u}{1+u}r$	n_2
aB	$\frac{1}{1+u}r$	n_3
ab	$\frac{1}{1+u}(1-r)$	n_4
کل	۱	n

لگاریتم درستنمایی داده‌ها نیز به صورت زیر است:

$$\log L(u,r\mid n)=\log\left[\frac{n!}{n_1!n_2!n_3!n_4!}\left(\frac{u}{1+u}(1-r)\right)^{n_1}\left(\frac{u}{1+u}r\right)^{n_2}\left(\frac{1}{1+u}r\right)^{n_3}\left(\frac{1}{1+u}(1-r)\right)^{n_4}\right]$$

$$=(n_1+n_2)\log u+(n_2+n_3)\log r+(n_1+n_4)\log(1-r)-n\log(1+u)+\log\left[\frac{n!}{n_1!n_2!n_3!n_4!}\right]$$

که امتیازات آن به‌صورت زیر هستند:

$$\begin{cases} S_r=\frac{\partial}{\partial r}\log L(u,r\mid n)=\frac{n_2+n_3}{r}-\frac{n_1+n_4}{1-r} \\ S_u=\frac{\partial}{\partial u}\log L(u,r\mid n)=\frac{n_1+n_2}{u}-\frac{n}{1-u} \end{cases}$$

و ماتریس اطلاعات آن نیز عبارت است از:

$$I=-E\begin{bmatrix} \frac{\partial^2}{\partial r^2}\log L(u,r\mid n) & \frac{\partial^2}{\partial r\,\partial u}\log L(u,r\mid n) \\ \frac{\partial^2}{\partial u\,\partial r}\log L(u,r\mid n) & \frac{\partial^2}{\partial u^2}\log L(u,r\mid n) \end{bmatrix}$$

$$=\begin{bmatrix} \frac{n}{r(1-r)} & 0 \\ 0 & \frac{n}{u(1+u)^2} \end{bmatrix}$$

بنابراین MLE های نرخ نوترکیبی و ضریب زنده‌مانی به همراه واریانس‌های نمونه‌ای بزرگ آن‌ها عبارتند از:

$$\hat{r}=\frac{n_2+n_3}{n}$$

$$\hat{u}=\frac{n_1+n_2}{n_3+n_4}$$

و:

$$\mathrm{var}\left(\hat{r}\right)=\frac{r\left(1-r\right)^2}{n}$$

$$\mathrm{var}\left(\hat{u}\right)=\frac{u\left(1+u\right)^2}{n}$$

می‌توان مشاهده کرد که MLE نرخ نوترکیبی برای طرح تلاقی برگشتی بدون توجه به وجود اختلال در زنده‌مانی برای یکی از دو نشانگر تغییر نیافته است. آزمون پیوستگی بین نشانگرهای A و B می‌تواند با استفاده از معادله (۳-۸) انجام شود. تفاوت بین n_1+n_2 و n_3+n_4 میزان انحراف از نسبت تفرق واحد را برای A:a (یعنی نسبت ۱:۱) اندازه‌گیری می‌کند. فرض وجود اختلال در زنده‌مانی می‌تواند به‌صورت زیر فرموله شود:

$H_0: u=1$ در برابر $H_1 \neq 1$ (۱-۶)

که H_0 به مدل کاهش یافته (یعنی عدم وجود اختلال در زنده‌مانی) تعلق دارد.

آماره آزمون برای سنجش فرض‌های (۶-۱) به‌صورت نسبت لگاریتم درستنمایی (LR) فرض تحت مدل کامل (یعنی H_1) به مدل کاهش یافته (یعنی H_0) محاسبه می‌شود:

$$LR = -2\log\left[\frac{L\left(u=1,\hat{r}\right)}{L\left(\hat{u},\hat{r}\right)}\right]$$

آماره آزمون LR آماره‌ای است که از لحاظ مجانبی دارای توزیع کای مربع با یک درجه آزادی است.

مثال ۶-۱. تفاوت در زنده‌مانی ژنوتیپ‌ها در یک نمونه تلاقی برگشتی دوگانه به‌صورت زیر بوده است:

گامت	AB	Ab	aB	ab
تعداد مشاهده شده	n_1 = ۵۰	n_2 = ۲	n_3 = ۱	n_4 = ۳۲

بنابراین، MLE های نرخ نوترکیبی و ضریب زنده‌مانی به همراه واریانس‌های نمونه‌ای بزرگ آن‌ها عبارتند از:

$$\hat{r} = \frac{2+1}{85} = 0.0353$$
$$\hat{u} = \frac{50+2}{1+32} = 1.576$$

و:

$$\mathrm{var}\left(\hat{r}\right) = \frac{0.0353(1-0.0353)}{85} = 0.00040$$
$$\mathrm{var}\left(\hat{u}\right) = \frac{1.576(1+1.576)^2}{85} = 0.1230$$

لگاریتم درستنمایی برای عدم وجود اختلال در زنده‌مانی به صورت زیر است:

$$LL_0 = 50\log\left[\frac{1-0.0353}{2}\right] + 2\log\left[\frac{0.0353}{2}\right]$$
$$+1\log\left[\frac{0.0353}{2}\right] + 32\log\left[\frac{1-0.0353}{2}\right] = -71.8961$$

به‌طور مشابه، لگاریتم درستنمایی برای اختلال در زنده‌مانی در جایگاه ژنی A عبارت است از:

$$LL_1 = 50\log\left[\frac{1.576}{1+1.576}\times(1-0.0353)\right] + 2\log\left[\frac{1.576}{1+1.576}\times 0.0353\right]$$
$$+1\log\left[\frac{1}{1+1.576}\times 0.0353\right] + 32\log\left[\frac{1}{1+1.576}\times(1-0.0353)\right] = -69.755$$

آماره LR برای اختلال در زنده‌مانی در جایگاه ژنی A در برابر عدم وجود اختلال برابر است با:

$$LR=-2\log\left[\frac{L\left(u=1,\hat{r}\right)}{L\left(\hat{u},\hat{r}\right)}\right]=-2\left[-71.8961-(-69.7545)\right]=4.283$$

مقایسه مقدار LR با جدول کای مربع $x^2_{0.05}(1)$ = ۳/۸۴۱، مشخص می‌کند که اختلال در زنده‌مانی در جایگاه ژنی A معنی‌دار است.

آمیزش مضاعف AaBb×AaBb

برای تلاقی مضاعف AaBb×AaBb، فراوانی‌های ۹ ژنوتیپ F_2 برای نشانگرهای A و B در قالب ماتریسی می‌تواند به شکل ماتریسی زیر بیان شود:

(۲-۶)

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array}\begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \begin{bmatrix} \frac{u^2}{(1+u)^2}(1-r)^2 & \frac{2u}{(1+u)^2}r(1-r) & \frac{u}{(1+u)^2}r^2 \\ \frac{2u}{(1+u)^2}r(1-r) & \frac{2u}{(1+u)^2}\left[(1-r)^2+r^2\right] & \frac{2u}{(1+u)^2}r(1-r) \\ \frac{1}{(1+u)^2}r^2 & \frac{2}{(1+u)^2}r(1-r) & \frac{1}{(1+u)^2}(1-r)^2 \end{bmatrix} \end{array}$$

که $P=\left\{p_{ij}, i=2,1,0;\ j=2,1,0\right\}$ است.

ماتریس داده‌ها نیز به صورت زیر ارایه می‌شود:

$$n=\begin{array}{c} \\ AA \\ Aa \\ aa \end{array}\begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \begin{bmatrix} n_{22} & n_{21} & n_{20} \\ n_{12} & n_{11} & n_{10} \\ n_{02} & n_{01} & n_{00} \end{bmatrix} \end{array}$$

تابع درستنمایی داده‌ها برابر است با:

$$L(u,r\mid n)=\frac{n!}{n_{22}!n_{21}!\ldots n_{00}!}\prod_{i=0}^{2}\prod_{j=0}^{2}p_{ij}^{n_{ij}}$$

بر اساس ماتریس (۶-۲)، محاسبه MLE ضریب زنده‌مانی به صورت زیر دشوار نیست:

$$\hat{u}=\frac{2\left(n_{22}+n_{21}+n_{20}\right)+n_{12}+n_{11}+n_{10}}{2\left(n-n_{22}-n_{21}-n_{20}\right)-\left(n_{12}+n_{11}+n_{10}\right)}$$

MLE نرخ نوترکیبی در F_2 نیز می‌تواند از روی ماتریس (۶-۲) به‌دست آید. ولی الگوریتم EM ضروری و مورد نیاز است، چون نمی‌توان هیچ شکل ثابتی برای $\hat{r}$ به‌دست آورد. با مشاهده دقیق ماتریس (۶-۲)، متوجه می‌شویم که تعداد مورد انتظار نوترکیب‌ها در اجزای ماتریس مربوطه حقیقتاً هنگامی‌که یک نشانگر اختلال در زنده‌مانی را نشان می‌دهد، تغییر نمی‌کند. این امر به‌دلیل آن است که ضرایب هر عنصر ماتریس دربردارنده هیچ‌گونه اطلاعاتی در خصوص نرخ نوترکیبی نیست؛ بنابراین، می‌توانیم $\hat{r}$ را به‌طور مستقیم با استفاده از فرمول‌های مرحله M برآورد کنیم:

$$\hat{r}=\frac{1}{2n}\left[n_{12}+n_{21}+n_{10}+n_{01}+2\left(n_{02}+n_{20}\right)+\phi n_{11}\right]$$

که در آن:

$$\phi=\frac{2r^2}{(1-r)^2+r^2}$$

برای محاسبه تعداد مورد انتظار نوترکیب‌ها برای هتروزیگوت مضاعف در مرحله E استفاده می‌شود (فصل ۳ را مشاهده کنید). در پایان، تجزیه پیوستگی در F_2 نیز همانند تلاقی برگشتی، به وسیله آثار زنده‌مانی یک نشانگر تحت‌تأثیر قرار نمی‌گیرد.

واریانس نمونه گیری MLE های نرخ نوترکیبی و ضریب زنده‌مانی در F_2 می‌تواند به‌صورت زیر به‌دست آید:

$$\operatorname{var}\left(\hat{r}\right)=\frac{r\left(1-r\right)\left(1-2r+2r^2\right)\left(1+u\right)^2}{2n\left(\left(2r^2-2r+1\right)u^2+\left(8r^2-8r+2\right)u+2r^2-2r+1\right)}$$

$$\operatorname{var}\left(\hat{u}\right)=\frac{u\left(1+u\right)^2}{2n}$$

به‌طور مشابه، می‌توانیم فرض وجود اختلال در زنده‌مانی را آزمون کنیم.

۶-۲-۲- مدل دو ژنی

اکنون به حالت کلی که در آن هر دو ژن به وسیله آثار زنده‌مانی دچار اختلال شده‌اند، باز می‌گردیم. فرض کنید که علاوه بر پارامتر u که در مبحث بالا برای افزایش نسبی قابلیت بقاء یا

زنده‌مانی A در برابر a تعریف شد، v نیز زنده‌مانی B نسبت به b باشد. همچنین فرض می‌کنیم که دو اثر زنده‌مانی به‌طور مستقل از هم عمل می‌کنند.

تلاقی برگشتی AaBb×aabb

مشاهدات و مقادیر مورد انتظار یک تلاقی برگشتی مضاعف به‌صورت زیر است:

گامت	فراوانی	
	مورد انتظار	مشاهده شده
AB	$\frac{uv}{d}(1-r)$	n_1
Ab	$\frac{u}{d}r$	n_2
aB	$\frac{v}{d}r$	n_3
Ab	$\frac{1}{d}(1-r)$	n_4
کل	۱	n

که در آن:

$$d = uv(1-r) + ur + vr + (1-r)$$

درستنمایی داده‌ها نیز به صورت زیر است:

$$L(u,r\,|\,n) = \frac{n!}{n_1!n_2!n_3!n_4!}\left[\frac{uv}{d}(1-r)\right]^{n_1}\left[\frac{u}{d}r\right]^{n_2}\left[\frac{v}{d}r\right]^{n_3}\left[\frac{1}{d}(1-r)\right]^{n_4}$$

MLE های سه پارامتر می‌توانند به صورت معمول به‌دست آیند:

$$\hat{r} = \frac{\sqrt{n_2 n_3}}{\sqrt{n_1 n_4} + \sqrt{n_2 n_3}}$$

$$\hat{u} = \sqrt{\frac{n_1 n_2}{n_3 n_4}}$$

$$\hat{v} = \sqrt{\frac{n_1 n_3}{n_2 n_4}}$$

واریانس‌های نمونه‌گیری این پارامترها نیز بدین صورت حاصل می‌شوند:

$$\mathrm{var}\left(\hat{r}\right)=\frac{r^2\left(1-r\right)^2}{h}$$

$$\mathrm{var}\left(\hat{u}\right)=\frac{u^2}{h}$$

$$\mathrm{var}\left(\hat{v}\right)=\frac{v^2}{h}$$

که در آن:

$$h=\frac{4n_1n_2n_3n_4}{n_1n_2n_3+n_1n_2n_4+n_1n_3n_4+n_2n_3n_4}$$

می‌توانیم آماره آزمون نسبت لگاریتم درستنمایی را برای آزمون وجود پیوستگی و آثار زنده‌مانی دو نشانگر فرموله کنیم. با تجزیه بالا، آزمون و تشخیص پیوستگی بین دو نشانگر اریب با استفاده از فرمولی که هیچ کدام از دو نشانگر و یا فقط یکی از آن‌ها دارای انحراف از نسبت تفرق مندلی هستند، نمی‌تواند انجام شود.

مثال ۶-۲. زنده‌مانی متفاوت دو نشانگر مثال ۶-۱ را در نظر بگیرید.

بنابراین، MLE های نرخ نوترکیبی و ضریب زنده‌مانی به همراه واریانس‌های نمونه‌ای بزرگ آن‌ها عبارتند از:

$$\hat{r}=\frac{\sqrt{2\times1}}{\sqrt{50\times32}+\sqrt{2\times1}}=0.0341$$

$$\hat{u}=\sqrt{\frac{50\times2}{1\times32}}=1.7678$$

$$\hat{v}=\sqrt{\frac{50\times1}{2\times32}}=0.8839$$

و

$$h=\frac{4n_1n_2n_3n_4}{n_1n_2n_3+n_1n_2n_4+n_1n_3n_4+n_2n_3n_4}=2.5786$$

$$\mathrm{var}\left(\hat{r}\right)=\frac{0.0341^2\left(1-0.0341\right)^2}{2.5786}=0.000422$$

$$\mathrm{var}\left(\hat{u}\right)=\frac{1.7678^2}{2.5786}=1.2119$$

$$\mathrm{var}\left(\hat{v}\right)=\frac{0.8839^2}{2.5786}=0.3030$$

لگاریتم درستنمایی اختلال در زنده‌مانی در جایگاه‌های ژنی A و B بدین صورت است:

$$LL_2=50\log\left[\left(1-0.0341\right)1.7678\times\frac{0.8839}{2.5655}\right]+2\log\left[0.0341\times\frac{1.7678}{2.5655}\right]$$

$$+1\log\left[0.0341\times\frac{0.8839}{2.5655}\right]+32\log\left[\frac{1-0.0341}{2.5655}\right]=-69.734$$

که $\hat{d}=\hat{u}\hat{v}\left(1-\hat{r}\right)+\hat{u}\hat{r}+\hat{v}\hat{r}+\left(1-\hat{r}\right)=2.5655$ است.

آماره LR برای وجود اختلال در زنده‌مانی در جایگاه‌های ژنی A و B در برابر عدم وجود اختلال عبارت است از:

$$LR=-2\log\left[\frac{L\left(u=1,v=1,\hat{r}\right)}{L\left(\hat{u},\hat{v},\hat{r}\right)}\right]=-2\left[-71.8961-(-69.7344)\right]=4.3234$$

مقایسه مقدار LR با مقدار جدول کای مربع یعنی $x^2_{0.05}(2)$ = ۵/۹۹۱، نشان می‌دهد که اختلال در زنده‌مانی معنی‌دار نیست.

تلاقی مضاعف F_2 های AaBb×AaBb

زمانی که هر دو نشانگر دارای انحراف از تفرق مندلی باشند، فراوانی‌های ۹ ژنوتیپ را در جمعیت F_2 بدین صورت داریم:

(۶-۳)

$$\begin{array}{c} AA \\ Aa \\ aa \end{array}\begin{bmatrix} \frac{u^2v^2}{d^2}(1-r)^2 & \frac{2u^2v}{d^2}r(1-r) & \frac{u^2}{d^2}r^2 \\ \frac{2uv^2}{d^2}r(1-r) & \frac{2uv}{d^2}\left[(1-r)^2+r^2\right] & \frac{2u}{d^2}r(1-r) \\ \frac{v^2}{d^2}r^2 & \frac{2v}{d^2}r(1-r) & \frac{1}{d^2}(1-r)^2 \end{bmatrix}$$

$$= \begin{bmatrix} (1-r)^2 & 2r(1-r) & r^2 \\ 2r(1-r) & 2\left[(1-r)^2+r^2\right] & 2r(1-r) \\ r^2 & 2r(1-r) & (1-r)^2 \end{bmatrix} \circ \begin{bmatrix} \frac{u^2v^2}{d^2} & \frac{u^2v}{d^2} & \frac{u^2}{d^2} \\ \frac{u\,v^2}{d^2} & \frac{uv}{d^2} & \frac{u}{d^2} \\ \frac{v^2}{d^2} & \frac{v}{d^2} & \frac{1}{d^2} \end{bmatrix} \qquad (۶-۴)$$

$$= F_r \circ F_{uvr}$$

که $\circ$ برای نشان دادن حاصل‌ضرب جزء به جزء عناصر دو ماتریس استفاده می‌شود، F_r فقط با r و F_{uvr} با u، v و r مرتبط است.

الگوریتم EM می‌تواند برای محاسبه MLE مربوط به r بر اساس ماتریس (۶-۳) استفاده شود، ولی به‌دست آوردن MLE مشکل است، چون ضرایب مربوط به هر جزء از این ماتریس دارای r است؛ اما با تفکیک ماتریس (۶-۳) به دو ماتریس جزئی‌تر در ماتریس (۶-۴)، می‌توانیم نتیجه‌گیری را ساده نماییم. MLE مربوط به r بر اساس اولین جزء ماتریس که به عنوان F_r داده شده است (همانگونه که می‌توان در بخش ۳-۵-۲ مشاهده نمود)، شامل مرحله E برای محاسبه تعداد مورد انتظار پیشامدهای نوترکیبی برای هتروزیگوت مضاعف F_2 یعنی AaBb با استفاده از $\phi = \frac{2r^2}{(1-r)^2+r^2}$ است. MLE مربوط به r بر اساس جزء دوم ماتریس یعنی F_{uvr}، می‌تواند از راه بیشینه نمودن درستنمایی ایجاد شده با ماتریس (۶-۳) و ماتریس مشاهدات n به‌دست آید. درستنمایی به صورت رابطه زیر ارایه می‌شود:

$$\log L = 1^T\left(n \circ \log F\right)1$$
$$= 1^T\left[n \circ \log\left(F \circ F_{uvr}\right)\right]1$$
$$= 1^T\left(n \circ \log F\right)1 + 1^T\left(n \circ F_{uvr}\right)1$$

که در آن $1 = (1,1,1)$ است.

با ترکیب این دو مرحله به مرحله M برای برآورد r می‌رسیم که عبارت است از:

$$\hat{r} = \frac{1}{2n}\left[n_{12} + n_{21} + n_{10} + n_{01} + 2\left(n_{02} + n_{20}\right) + \phi n_{11}\right] + \frac{(u-1)(v-1)(1-r)r}{d} \qquad (۵-۶)$$

که ضرایب زنده‌مانی به‌صورت زیر برآورد می‌شوند:

$$\hat{u} = \frac{\left[2\left(n_{22} + n_{21} + n_{20}\right) + n_{12} + n_{11} + n_{10}\right]\left[vr + (1-r)\right]}{\left[2\left(n - n_{22} - n_{21} - n_{20}\right) - \left(n_{12} + n_{11} + n_{10}\right)\right]\left[v(1-r) + r\right]}$$

$$\hat{v} = \frac{\left[2\left(n_{22} + n_{12} + n_{02}\right) + n_{12} + n_{11} + n_{01}\right]\left[ur + (1-r)\right]}{\left[2\left(n - n_{22} - n_{12} - n_{02}\right) - \left(n_{21} + n_{11} + n_{01}\right)\right]\left[u(1-r) + r\right]}$$

زمانی که هر دو نشانگر در F_2 دارای انحراف از تفرق مندلی باشند، برآورد پیوستگی تحت‌تأثیر قرار خواهد گرفت. وجود پیوستگی و آثار زنده‌مانی می‌تواند با استفاده از آزمون نسبت درستنمایی آزمون شود.

ماتریس اطلاعات Fisher برای r، u و v می‌تواند به صورت زیر به‌دست آید:

$$2n\begin{bmatrix} \frac{u\left(1-2r+2r^2\right)\left(uv+(v+1)^2\right)-2uv+v+2rv(r-1)}{r\left(2r^2-2r+1\right)(1-r)d^2} & \frac{v^2-1}{d^2} & \frac{u^2-1}{d^2} \\ \frac{v^2-1}{d^2} & \frac{(vr+1-r)(rv-v-r)}{ud^2} & \frac{2r-1}{d^2} \\ \frac{u^2-1}{d^2} & \frac{2r-1}{d^2} & \frac{(ur+1-r)(uv-u-r)}{ud^2} \end{bmatrix}$$

واریانس‌های نمونه‌گیری برای MLE های مربوط به نرخ نوترکیبی و ضرایب زنده‌مانی در F_2 نیز می‌توانند از روی معکوس ماتریس اطلاعات Fisher به‌دست آیند.

۶-۲-۳- شبیه‌سازی

تاکنون معادلات برآورد نرخ نوترکیبی بین نشانگرهای اریب توصیف شدند، ولی مشخص نیست که چگونه زنده‌مانی گامتی روی برآورد نرخ نوترکیبی اثر می‌گذارد. این امر می‌تواند با مطالعات شبیه‌سازی مورد آزمایش قرار گیرد. یک مجموعه داده‌های ژنوتیپی نشانگری در یک جمعیت تلاقی برگشتی با اندازه‌های نمونه مختلف (n = ۸۰ یا n = ۲۰۰) برای دو نشانگر با قابلیت زنده‌مانی

متفاوت که پیوستگی قوی (r = ۰/۰۵) یا پیوستگی ضعیف (r = ۰/۳۵) نشان می‌دهند شبیه‌سازی شد. داده‌های شبیه‌سازی شده با استفاده از مدل زنده‌مانی (۱) که در بخش ۶-۲-۲ پیشنهاد شد و همچنین مدل پیوستگی (۲) که برای نشانگرهای طبیعی استفاده شدند، تجزیه شدند.

جدول ۶-۱- MLE های نرخ نوترکیبی بین دو نشانگر اریب با درجات متفاوت اختلال در زنده‌مانی گامتی در نتاج تلاقی برگشتی. خطای نمونه‌برداری (SE) برآوردها با توجه به ۱۰۰ تکرار شبیه‌سازی ارایه شده‌اند.

	مدل ۱			مدل ۲
(u,v,r)	$\hat{r}\pm SE$	$\hat{u}\pm SE$	$\hat{v}\pm SE$	$\hat{r}_0\pm SE$
n = ۸۰				
(1.0,1.0,0.05)	0.053±0.021	1.099±0.438	1.067±0.409	0.056±0.022
(1.0,1.0,0.35)	0.359±0.054	1.064±0.226	1.014±0.237	0.359±0.052
(1.0,1.0,0.05)	0.053±0.021	1.099±0.438	1.067±0.409	0.056±0.022
(1.0,1.0,0.35)	0.359±0.054	1.064±0.226	1.014±0.237	0.359±0.052
(2.0,2.0,0.05)	0.062±0.023	2.410±1.112	2.119±0.989	0.053±0.020
(2.0,2.0,0.35)	0.362±0.056	2.156±0.596	2.045±0.544	0.313±0.047
n = ۲۰۰				
(1.0,1.0,0.05)	0.047±0.017	1.123±0.453	1.054±0.435	0.050±0.016
(1.0,1.0,0.35)	0.354±0.033	0.996±0.144	1.017±0.163	0.355±0.033
(1.0,1.0,0.05)	0.053±0.021	1.099±0.438	1.067±0.409	0.056±0.022
(1.0,1.0,0.35)	0.359±0.054	1.064±0.226	1.014±0.237	0.359±0.052
(2.0,2.0,0.05)	0.049±0.018	2.115±0.857	2.230±0.945	0.042±0.014
(2.0,2.0,0.35)	0.349±0.042	2.043±0.365	1.989±0.329	0.302±0.034

همان‌گونه که انتظار می‌رود، نرخ نوترکیبی می‌تواند به‌طور منطقی به وسیله دو مدل، وقتی که حداقل یک نشانگر اریب نیست، برآورد شود (u = ۱ و یا v = ۱)(جدول ۶-۱). ولی زمانی که هر دو نشانگر به وسیله زنده‌مانی گامتی دچار اریب شوند، برآورد مدل دوم تا حد زیادی اریب خواهد شد، به ویژه زمانی که دو نشانگر شدیداً به هم پیوسته هستند، حال آنکه مدل اول همیشه می‌تواند یک برآورد منطقی از r ارایه نماید. برآورد مدل دوم به‌ویژه برای اندازه نمونه کوچک، نسبت به مدل اول دقیق‌تر است (دارای مقادیر خطای نمونه‌برداری کوچک‌تری است) که اساساً به‌دلیل آن است که در مدل قبلی، پارامترهای کمتری (یک پارامتر) نسبت به مدل اخیر (سه پارامتر) برآورد می‌شود. ولی این مزیت در دقت سبب اریب بیشتری می‌شود.

ضرایب زنده‌مانی گامتی (u و v) می‌توانند به‌طور ضعیفی با یک اندازه نمونه کوچک برآورد شوند، اگرچه این حالت روی برآورد نرخ نوترکیبی اثری ندارد (جدول ۶-۱). برای افزایش صحت و

دقت برآورد u و v، به یک اندازه نمونه بزرگ نیاز است. یافته‌های مشابهی برای مطالعات شبیه‌سازی جمعیت F_2 به‌دست آمده است.

۶-۳- تنوع در زنده‌مانی زیگوتی

۶-۳-۱- مدل تک ژنی

تأثیر تنوع در زنده‌مانی زیگوتی می‌تواند در سطح ژنوتیپ نیز رخ دهد. برای تلاقی برگشتی، مدل‌هایی که برای مطالعهٔ اثر اختلال در زنده‌مانی گامتی به‌دست آمده‌اند، می‌توانند به‌طور مستقیم برای آزمون اختلال در زنده‌مانی زیگوتی مورد استفاده قرار گیرند. در این بخش بحث خود را در خصوص ایجاد مدل برای تنوع زندهمانی زیگوتی در نتاج F_2 متمرکز خواهیم کرد.

یک طرح F_2 را در نظر بگیرید که در آن، سه ژنوتیپ در هر نشانگر وجود دارند. اگر فقط یک نشانگر (که A نامیده می‌شود) آثار زنده‌مانی زیگوتی را نشان دهد، می‌توانیم فرض کنیم که زنده‌مانی AA نسبت به Aa به‌صورت u_1 و زنده‌مانی Aa نسبت به aa به صورت u_2 است؛ بنابراین، سهم کلی سه ژنوتیپ باید برای AA به‌صورت $\frac{u_1u_2}{U}$، برای Aa به صورت $\frac{2u_2}{U}$ و برای aa به صورت $\frac{1}{U}$ باشد که $U = u_1u_2 + 2u_2 + 1$ است. فراوانی‌های مورد انتظار ۹ ژنوتیپ F_2 تابعی از ضرایب زنده‌مانی و نرخ نوترکیبی بوده و به صورت زیر مرتب می‌شوند:

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array}\begin{array}{c} \begin{array}{ccc} \text{BB} & \text{Bb} & \text{bb} \end{array} \\ \begin{bmatrix} \frac{u_1u_2}{U}(1-r)^2 & \frac{2u_1u_2}{U}r(1-r) & \frac{u_1u_2}{U}r^2 \\ \frac{2u_2}{U}r(1-r) & \frac{2u_2}{U}\left[(1-r)^2+r^2\right] & \frac{2u_2}{U}r(1-r) \\ \frac{1}{U}r^2 & \frac{2}{U}r(1-r) & \frac{1}{U}(1-r)^2 \end{bmatrix} \end{array} \qquad (۶-۶)$$

MLE مربوط به r برای یک F_2 طبیعی با استفاده از الگوریتم EM به‌دست می آید، چون ضرایب درون هر یک از اجزای ماتریس (۶-۶) به نرخ نوترکیبی بستگی ندارد. می‌توان مشاهده نمود که برآورد r تحت‌تأثیر زنده مانی متفاوت زیگوتی یک نشانگر قرار نمی‌گیرد. پارامترهای u_1 و u_2 به‌صورت زیر به‌دست می‌آیند:

$$\hat{u}_1 = \frac{2\left(n_{22}+n_{21}+n_{20}\right)}{n_{12}+n_{11}+n_{10}}$$

$$\hat{u}_2 = \frac{n_{12}+n_{11}+n_{10}}{2\left(n_{02}+n_{01}+n_{00}\right)}$$

همچنین ماتریس اطلاعات Fisher عبارت است از:

$$I = n\begin{bmatrix} \frac{2\left(2r^2-2r+1\right)\left(u_1u_2+1\right)+4u_2(1-r)^2}{Ur(1-r)\left(2r^2-2r+1\right)} & 0 & 0 \\ 0 & \frac{u_2\left(2u_2+1\right)}{u_1U^2} & -\frac{1}{U^2} \\ 0 & -\frac{1}{U^2} & \frac{u_1+2}{u_2U^2} \end{bmatrix}$$

واریانس‌های نمونه‌برداری MLE های نرخ نوترکیبی و ضرایب زنده‌مانی زیگوتی در F_2 نیز می‌توانند به‌صورت زیر به‌دست آیند:

$$\operatorname{var}\left(\hat{r}\right) = \frac{Ur\left(1-r\right)\left(2r^2-2r+1\right)}{2n\left[\left(2r^2-2r+1\right)\left(u_1u_2+1\right)+2u_2\left(1-2r\right)^2\right]}$$

$$\text{۶}\ \operatorname{var}\left(\hat{u}_1\right) = \frac{U\left(u_1+2\right)u_1}{2nu_2}$$

$$\operatorname{var}\left(\hat{u}_2\right) = \frac{U\left(2u_2+2\right)u_2}{2n}$$

۶-۳-۲- مدل دو ژنی

اکنون وضعیتی را در نظر می‌گیریم که در آن هر دو نشانگر دارای آثار زنده‌مانی هستند. آثار زنده‌مانی نشانگر دوم (B) می‌تواند به‌طور مشابه با استفاده از v_1 و v_2 نشان داده شود. اگر دو ژن در معرض اختلال در زنده‌مانی قرار گیرند، عناصر ماتریس (۶-۶) باید تغییر یابند تا آثار نشانگر دوم در نظر گرفته شود:

$$
F = \begin{array}{c} \\ AA \\ Aa \\ aa \end{array}
\begin{array}{c}
\begin{array}{ccc} BB & Bb & bb \end{array} \\
\begin{bmatrix}
\frac{u_1u_2v_1v_2}{d}(1-r)^2 & \frac{2u_1u_2v_2}{d}r(1-r) & \frac{u_1u_2}{d}r^2 \\
\frac{2u_2v_1v_2}{d}r(1-r) & \frac{2u_2v_2}{d}\left[(1-r)^2+r^2\right] & \frac{2u_2}{d}r(1-r) \\
\frac{v_1v_2}{d}r^2 & \frac{2v_2}{d}r(1-r) & \frac{1}{d}(1-r)^2
\end{bmatrix}
\end{array}
\tag{۶-۷}
$$

$$
= \begin{bmatrix}
(1-r)^2 & 2r(1-r) & r^2 \\
2r(1-r) & 2\left[(1-r)^2+r^2\right] & 2r(1-r) \\
r^2 & 2r(1-r) & (1-r)^2
\end{bmatrix}
\circ
\begin{bmatrix}
\frac{u_1u_2v_1v_2}{d} & \frac{u_1u_2v_2}{d} & \frac{u_1u_2}{d} \\
\frac{u_2v_1v_2}{d} & \frac{u_2v_2}{d} & \frac{u_2}{d} \\
\frac{v_1v_2}{d} & \frac{v_2}{d} & \frac{1}{d}
\end{bmatrix}
$$

$$
= F_r \circ F_{u_1u_2v_1v_2r}
$$

که در آن:

$$
d = u_1u_2v_1v_2(1-r)^2 + 2u_2v_1v_2r(1-r) + v_1v_2r^2 + 2u_1u_2v_2r(1-r)
$$
$$
+2u_2v_2\left[(1-r)^2+r^2\right] + 2v_2r(1-r) + u_1u_2r^2 + 2u_2r(1-r) + (1-r)^2
$$

هر دو ماتریس مؤلفه در رابطه (۶-۷) دارای نرخ نوترکیبی هستند. الگوریتم EM می تواند برای برآورد r در اولین ماتریس مؤلفه یعنی F_r، فرموله شود. توأم با اثر ماتریس مؤلفه دوم یعنی $F_{u_1u_2v_1v_2r}$، برای برآورد r، MLE کلی r را می‌توانیم بدین صورت بیان کنیم:

$$
\hat{r} = \frac{1}{2n}\left(n_{12}+n_{21}+n_{10}+n_{01}+2\left(n_{02}+n_{20}\right)+\phi n_{11}\right) - \frac{r(1-r)}{2d}\frac{\partial d}{\partial r} \tag{۶-۸}
$$

که در آن:

$$
\frac{\partial d}{\partial r} = 2\left[u_1u_2v_1v_2(r-1) + u_2v_1v_2(1-2r) + v_1v_2r + u_1u_2v_2(1-2r) + u_2v_2(4r-2) + v_2(1-2r) + u_1u_2r + u_2(1-2r) - 1 + r\right]
$$

چهار ضریب زنده‌مانی u_1، u_2، v_1 و v_2 می‌توانند به‌طور همزمان بدین صورت برآورد شوند:

$$\hat{u}_1 = \frac{\left(n_{22}+n_{21}+n_{20}\right)d}{n\left[u_2v_1v_2\left(1-r\right)^2+u_2r^2+2u_2v_2r\left(1-r\right)\right]}$$

$$\hat{u}_2 = \frac{\left(n_{22}+n_{21}+n_{20}+n_{12}+n_{11}+n_{10}\right)d}{n\left[\left(u_1v_1v_2+2v_2\right)\left(1-r\right)^2+\left(u_1+2v_2\right)r^2+2\left(v_1v_2+u_1v_2+1\right)r\left(1-r\right)\right]}$$

$$\hat{v}_1 = \frac{\left(n_{22}+n_{21}+n_{02}\right)d}{n\left[u_1u_1v_2\left(1-r\right)^2+v_2r^2+2u_2v_2r\left(1-r\right)\right]}$$

$$\hat{v}_2 = \frac{\left(n_{22}+n_{12}+n_{02}+n_{21}+n_{11}+n_{01}\right)d}{n\left[\left(u_1u_2v_1+2u_2\right)\left(1-r\right)^2+\left(v_1+2u_2\right)r^2+2\left(u_1u_2+u_2v_1+1\right)r\left(1-r\right)\right]}$$

۶-۳-۳- شبیه‌سازی

اثر زنده‌مانی زیگوتی بر برآورد نرخ نوترکیبی از راه مطالعات شبیه‌سازی بررسی می‌شود. دو نشانگر که درجات متفاوتی از زنده‌مانی زیگوتی را در جمعیت F_2 با اندازه‌های نمونه n = ۸۰ و n = ۲۰۰ نشان می‌دادند، شبیه‌سازی کردیم. این دو نشانگر می‌توانند دارای پیوستگی قوی (r = ۰/۰۵) یا ضعیف (r = ۰/۳۵) باشند. داده‌های شبیه‌سازی شده با استفاده از مدل زنده‌مانی (۱) که در بخش ۶-۳ پیشنهاد شد و همچنین مدل پیوستگی (۲) که برای نشانگرهای معمولی استفاده می‌شود، تجزیه شدند.

زمانی‌که فقط یک نشانگر اریب باشد، دقت برآورد تحت‌تأثیر اختلال زنده‌مانی قرار نمی‌گیرد (جدول ۶-۲)؛ اما اگر هر دو نشانگر اریب باشند، از نظر دقت برآورد، مدل ۱ بهتر از مدل ۲ عمل می‌کند. این حالت به‌ویژه برای دو نشانگری که پیوستگی ضعیفی دارند و در آن‌ها، برآورد r ممکن است اریب زیادی به سمت پایین داشته باشد، صحیح است. مهم‌تر اینکه، برآورد حاصل از مدل ۲ در شرایطی که اندازه نمونه بیش از دو برابر شود (از ۸۰ تا ۲۰۰)، می‌تواند تا حدودی بهبود یابد. باید توجه نمود که مدل ۱ ممکن است صحت و دقت برآورد کمتری برای اندازه نمونه کوچک داشته باشد، ولی این حالت می تواند در هنگامی‌که اندازه نمونه افزایش می‌یابد تا حد زیادی بهبود یابد (جدول ۶-۲).

مدلی که در بخش ۶-۳ توصیف شد، اجازه برآورد ضرایب اختلال در زنده‌مانی گامتی (u_1، u_2 و v_1 و v_2) را برای دو نشانگر نیز فراهم می‌کند. برآوردهای این دو ضریب تحت تاثیر اندازه نمونه و درجه پیوستگی قرار می‌گیرند. همانگونه که انتظار می‌رود، هرچه اندازه نمونه بزرگ‌تر باشد،

برآوردها بهتر می‌شوند. ولی وقتی که دو نشانگر پیوستگی قوی دارند، این دو ضریب می‌توانند بهتر از زمانی برآورد شوند که پیوستگی بین دو نشانگر ضعیف باشد.

۶-۴- دسته‌بندی نادرست

۶-۴-۱- مدل تک ژنی

در عمل، ممکن است به‌واسطه خطاهای انسانی، یک ژنوتیپ به عنوان ژنوتیپ دیگری دسته‌بندی شود. برای مثال، در تلاقی برگشتی مضاعف AaBb×aabb، ممکن است نسبت λ آلل A به‌طور نادرست به صورت a دسته‌بندی شده باشد (صرف نظر از این که این دو آلل با B یا b مرتبط هستند یا خیر). فراوانی‌های مورد انتظار مناسب و تعداد مشاهده شده ژنوتیپ‌ها یا گامت‌ها در یک جمعیت تلاقی برگشتی به‌صورت زیر نشان داده می‌شوند:

گامت	فراوانی	
	مورد انتظار	مشاهده شده
AB	$\frac{1}{2}(1-r)(1-\lambda)$	n_1
Ab	$\frac{1}{2}r(1-\lambda)$	n_2
aB	$\frac{1}{2}\left[r+\lambda(1-r)\right]$	n_3
ab	$\frac{1}{2}\left[(1-r)+\lambda r\right]$	n_4
کل	۱	n

جدول ۶-۲- MLE های نرخ نوترکیبی بین دو نشانگر اریب با درجات مختلف اختلال در زنده‌مانی زیگوتی در نتاج تلاقی برگشتی. خطای نمونه‌برداری (SE) برآوردها با استفاده از ۱۰۰ تکرار شبیه‌سازی ارایه شده‌اند.

	مدل ۱					مدل ۲
(u_1,u_2,v_1,v_2,r)	$\hat{r}\pm SE$	$\hat{u}_1\pm SE$	$\hat{u}_2\pm SE$	$\hat{v}_1\pm SE$	$\hat{v}_2\pm SE$	$\hat{r}_0\pm SE$
n = 80						
(1.0,1.0,1.0,1.0,0.05)	0.045±0.017	1.190±0.699	1.332±1.370	1.101±0.613	1.090±0.772	0.050±0.017
(1.0,1.0,1.0,1.0,0.35)	0.359±0.050	1.076±0.298	1.027±0.299	0.965±0.273	1.040±0.316	0.360±0.049
(1.0,1.0,2.0,2.0,0.05)	0.044±0.020	1.267±1.181	1.548±2.243	2.654±2.346	2.534±2.469	0.049±0.016
(1.0,1.0,2.0,2.0,0.35)	0.349±0.052	1.051±0.291	1.018±0.346	2.092±0.575	2.457±1.257	0.352±0.048
(2.0,2.0,2.0,2.0,0.05)	0.044±0.021	2.598±2.435	4.982±8.378	2.438±2.313	2.456±3.352	0.039±0.015
(2.0,2.0,2.0,2.0,0.35)	0.347±0.056	2.157±0.584	4.545±20.383	2.038±0.523	2.346±1.427	0.282±0.036
n = 200						
(1.0,1.0,1.0,1.0,0.05)	0.047±0.011	1.067±0.435	1.108±0.472	1.069±0.409	1.049±0.404	0.049±0.010
(1.0,1.0,1.0,1.0,0.35)	0.352±0.032	1.000±0.186	1.008±0.171	1.022±0.210	1.024±0.182	0.353±0.032
(1.0,1.0,2.0,2.0,0.05)	0.048±0.012	1.004±0.292	1.037±0.402	2.232±1.001	2.192±0.940	0.050±0.011
(1.0,1.0,2.0,2.0,0.35)	0.351±0.034	1.009±0.167	1.057±0.237	2.011±0.314	2.051±0.432	0.350±0.030
(2.0,2.0,2.0,2.0,0.05)	0.051±0.014	2.117±0.637	2.258±1.093	2.058±0.660	2.151±1.213	0.042±0.011
(2.0,2.0,2.0,2.0,0.35)	0.351±0.033	2.052±0.362	2.101±0.656	2.030±0.318	2.038±0.534	0.285±0.024

مقادیر مورد انتظار به این دلیل به‌دست آمده‌اند که نسبتی از هر یک از دو کلاس گامتی اول در بردارنده A (برابر با λ)، به کلاس گامتی متناظر دربردارنده a، ولی با دسته‌بندی یکسان B یا b انتقال یافته است.

برآورد حداکثر درستنمایی دو پارامتر r و λ سر راست و آسان است؛ بنابراین داریم:

$$\hat{r}=\frac{n_2\left(n_1+n_3\right)}{n_2\left(n_1+n_3\right)+n_1\left(n_2+n_4\right)}$$

$$\hat{\lambda}=\frac{n_3n_4-n_1n_2}{\left(n_1+n_3\right)\left(n_2+n_4\right)}$$

واریانس‌های نمونه‌برداری این برآوردگرها عبارتند از:

$$\mathrm{var}\left(\hat{r}\right)=\frac{2r\left(1-r\right)}{n}\left[\frac{1}{1-\lambda}-2r\left(1-r\right)\right]$$

$$\mathrm{var}\left(\hat{\lambda}\right)=\frac{2\left(1-r\right)^{2}}{n}\left[\frac{\lambda}{1-\lambda}+2r\left(1-r\right)\right]$$

اگرچه A به‌طور نادرست به صورت a دسته بندی شده است، ولی ممکن است دسته بندی نادرست معکوس یعنی $a \to A$ نیز داشته باشیم. فرمول‌های مربوط به این الگوی جایگزین نیز می توانند با تغییر مناسب علایم مشاهده‌ای n_1، n_2، n_3 و n_4 به‌دست آیند.

زمانی که دسته‌بندی نادرست در شکل ژنوتیپ‌های F_2 رخ می‌دهند، فرمول‌های مشابهی می‌توانند برای توصیف آثار آن‌ها بر برآوردهای نرخ نوترکیبی ارائه شوند. بدین منظور، فرض کنیم که چنین دسته‌بندی نادرستی از مشخص نبودن آلل‌های انفرادی ایجاد شود؛ بنابراین، فرمول جمعیت تلاقی برگشتی بالا می‌تواند بسط داده شود تا دسته‌بندی نادرست در F_2 مدل‌سازی شود. فراوانی‌های مورد انتظار ژنوتیپ‌های F_2 پس از دسته بندی نادرست ژن A به صورت زیر بیان می‌شوند:

(۶-۹)

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array}\begin{array}{c} \begin{array}{ccc} \mathrm{BB} & \mathrm{Bb} & \mathrm{bb} \end{array} \\ \begin{bmatrix} \frac{1}{4}(1-\lambda)^2(1-r)^2 & \frac{1}{4}(1-\lambda)^2 r(1-r) & \frac{1}{4}(1-\lambda)^2 r^2 \\ \frac{1}{2}(1-\lambda)(1-r)\left(r+\lambda(1-r)\right) & \frac{1}{2}(1-\lambda)\left[r^2+(1-r)^2+2\lambda r(1-r)\right] & \frac{1}{2}(1-\lambda)r(1-r+\lambda r) \\ \frac{1}{4}\left[r+\lambda(1-r)\right]^2 & \frac{1}{2}\left[\lambda r^2+\lambda(1-r)^2+\left(1+\lambda^2\right)r(1-r)\right] & \frac{1}{4}\left[(1-r)+\lambda r\right]^2 \end{bmatrix} \end{array}$$

از الگوریتم EM برای برآورد نرخ نوترکیبی (r) استفاده می‌کنیم. یافتن تعداد مورد انتظار نوترکیب‌های درون هر جزء ماتریس (۶-۹) دشوار نیست به‌صورت زیر بیان می‌شود:

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array}\begin{array}{c} \begin{array}{ccc} \mathrm{BB} & \mathrm{Bb} & \mathrm{bb} \end{array} \\ \begin{bmatrix} 0 & 1 & 2 \\ \phi_1 & 2\phi_2 & \phi_3 \\ 2\phi_4 & \phi_5 & \phi_6 \end{bmatrix} \end{array}$$

که در آن:

$$\phi_1 = \frac{r(1-r)}{r(1-r)+\lambda(1-r)^2}$$
$$\phi_2 = \frac{r^2+\lambda r(1-r)}{r^2+(1-r)^2+2\lambda r(1-r)} \qquad (۱۰-۶)$$
$$\phi_3 = 1+\frac{\lambda r^2}{r(1-r)+\lambda r^2}$$
$$\phi_4 = \frac{r}{r+\lambda(1-r)}$$
$$\phi_5 = \frac{2\lambda r^2+(1+\lambda^2)r(1-r)}{\lambda r^2+\lambda(1-r)^2+(1+\lambda^2)r(1-r)}$$
$$\phi_6 = \frac{2\lambda r}{(1-r)+\lambda r}$$

همچنین نرخ نوترکیبی می‌تواند با استفاده از معادله زیر برآورد شود:

$$\hat{r} = \frac{1}{2n}\left(n_{21}+2n_{20}+\phi_1 n_{12}+2\phi_2 n_{11}+\phi_3 n_{10}+2\phi_4 n_{02}+\phi_5 n_{01}+\phi_6 n_{00}\right) \qquad (۱۱-۶)$$

در مرحله E، تعداد مورد انتظار نوترکیب‌ها برای هر ژنوتیپ با استفاده از معادله (۶-۱۰) محاسبه می‌شود. این تعداد مورد انتظار برای تجدید برآورد r با معادله (۶-۱۱) در مرحله M استفاده می‌شود. این دو مرحله تا زمانی که r در مقدار ثابت و پایداری همگرا شود، تکرار می‌شوند.

می‌توان مشاهده کرد که برآورد r تحت تأثیر درجه دسته‌بندی نادرست (λ) قرار گرفته است، ولی به وسیله برآورد λ تحت تأثیر قرار نمی‌گیرد. MLE مربوط به λ در F_2 با حل معادله چندجمله‌ای درجه سوم زیر به‌دست می‌آید:

$$\frac{n_1}{1-\lambda} = \frac{n_2}{\lambda+\frac{r}{1-r}} + \frac{n_3}{\lambda+\frac{1-r}{r}} + \frac{n_{11}}{\lambda+\frac{r^2+(1-r)^2}{2r(1-r)}}$$

که در آن:

$$n_1 = 2(n_{22}+n_{21}+n_{20})+n_{12}+n_{11}+n_{10}$$
$$n_2 = n_{12}+2n_{02}+n_{01}$$
$$n_3 = n_{10}+2n_{00}+n_{01}$$

واریانس‌های نمونه‌برداری $\hat{r}$ و $\hat{\lambda}$ می‌توانند به‌صورت زیر به‌دست آیند:

$$\mathrm{var}\left(\hat{r}\right)=\frac{\left\{r^2\left[12\left(1-\hat{\lambda}\right)^2\hat{r}\left(\hat{r}-2\right)+13\hat{\lambda}^2-34\hat{\lambda}+21\right]-\left(\hat{\lambda}^2-10\hat{\lambda}+9\right)r+2\right\}\hat{r}\left(1-\hat{r}\right)}{n\left(1-\hat{\lambda}\right)\left[2-\hat{\lambda}-6\left(1-\hat{\lambda}\right)\hat{r}\left(1-\hat{r}\right)\right]}$$

$$\mathrm{var}\left(\hat{\lambda}\right)=\frac{\hat{\lambda}\left(1-\hat{\lambda}\right)+2\left(1-\hat{\lambda}\right)^2\hat{r}\left(1-\hat{r}\right)}{n}$$

۶-۲-۴- مدل دو ژنی

زمانی که دو ژن به‌صورت نادرست دسته‌بندی شده باشند، باید یک نسبت اضافی برای آلل B که به‌طور نادرست به شکل b دسته‌بندی شده است را در نظر بگیریم. اجازه دهید λ_1 و λ_2 را به ترتیب به‌عنوان نسبت‌های دسته‌بندی نادرست نشانگرهای A و B در نظر بگیریم. با فرض اینکه این دو نسبت مستقل هستند، تعداد مورد انتظار هر یک از ۴ ژنوتیپ تلاقی برگشتی به همراه تعداد مشاهده شده آن‌ها را به‌صورت زیر خواهیم داشت:

گامت	تعداد مورد انتظار	تعداد مشاهده شده
AB	$\frac{1}{2}(1-\lambda_1)(1-\lambda_2)(1-r)$	n_1
Ab	$\frac{1}{2}(1-\lambda_1)\left[r+\lambda_2(1-r)\right]$	n_2
aB	$\frac{1}{2}(1-\lambda_2)\left[r+\lambda_1(1-r)\right]$	n_3
ab	$\frac{1}{2}\left[r(\lambda_1+\lambda_2)+(1-r)(1+\lambda_1\lambda_2)\right]$	n_4
کل	۱	n

MLE نرخ نوترکیبی و نسبت‌های دسته‌بندی نادرست می‌تواند به صورت زیر به‌دست آید:

$$\hat{r} = 1 - \frac{n_1 n}{2(n_1+n_2)(n_1+n_3)}$$

$$\hat{\lambda}_1 = \frac{(n_3+n_4)-(n_1+n_2)}{n}$$

$$\hat{\lambda}_2 = \frac{(n_2+n_4)-(n_1+n_3)}{n}$$

واریانس‌های نمونه‌برداری MLE های مربوط به r، λ_1 و λ_2 عبارتند از:

$$\mathrm{var}(\hat{r}) = \frac{n n_1 \left[n(n_1^2 + n_2 n_3) - n_1(n_1+n_2)(n_1+n_3) \right]}{4(n_1+n_2)^3 (n_1+n_3)^3}$$

$$\mathrm{var}(\hat{\lambda}_1) = \frac{1-\lambda_1^2}{n}$$

$$\mathrm{var}(\hat{\lambda}_2) = \frac{1-\lambda_2^2}{n}$$

هنگامی‌که هر دو نشانگر در F_2 به صورت نادرست دسته‌بندی شده باشند، برآورد نرخ نوترکیبی مشکل است. با این فرض که دسته‌بندی نادرست به دلیل مشخص نبودن آلل‌های انفرادی ایجاد می‌شود، فراوانی‌های مورد انتظار ژنوتیپ‌های F_2 را پس از آنکه هر دو ژن A و B به صورت نادرست دسته‌بندی می‌شوند، به‌صورت زیر بیان می‌کنیم:

ژن‌ها		فراوانی مورد انتظار	مشاهدات
AA	BB	$\frac{(1-\lambda_1)^2(1-\lambda_2)^2(1-r)^2}{4}$	n_{22}
AA	Bb	$\frac{(1-\lambda_1)^2(1-\lambda_2)(1-r)\left(r+\lambda_2(1-r)\right)}{2}$	n_{21}
AA	Bb	$\frac{(1-\lambda_1)^2\left(r+\lambda_2(1-r)\right)^2}{4}$	n_{20}
Aa	BB	$\frac{(1-\lambda_1)(1-\lambda_2)^2(1-r)\left(r+\lambda_1(1-r)\right)}{2}$	n_{12}
Aa	Bb	$(1-\lambda_1)(1-\lambda_2)\left[2r(1-r)(\lambda_1+\lambda_2)+r^2+(1-r)^2(2\lambda_1\lambda_2+1)\right]$	n_{11}
Aa	bb	$\frac{(1-\lambda_1)\left(r+\lambda_2(1-r)\right)\left((1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r\right)}{2}$	n_{10}
aa	BB	$\frac{(1-\lambda_2)^2\left(r+\lambda_1(1-r)\right)^2}{4}$	n_{02}
aa	Bb	$\frac{\left((1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r\right)\left(r+\lambda_1(1-r)\right)(1-\lambda_2)}{2}$	n_{01}
aa	bb	$\frac{\left((1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r\right)^2}{4}$	n_{00}

تعداد مورد انتظار نوترکیب‌ها در هر یک از اجزای جدول بالا به‌صورت زیر بیان می‌شود:

که در آن:

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array}\begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \begin{bmatrix} 0 & \phi_1 & 2\phi_2 \\ \phi_3 & 2\phi_4 & \phi_5 \\ 2\phi_6 & \phi_7 & \phi_8 \end{bmatrix} \end{array}$$

$$\phi_1 = \frac{r(1-r)}{r(1-r)+\lambda_2(1-r)^2}$$

$$\phi_2 = \frac{r}{r+\lambda_2(1-r)}$$

$$\phi_3 = \frac{r(1-r)}{r(1-r)+\lambda_1(1-r)^2}$$

$$\phi_4 = \frac{r(1-r)(\lambda_1+\lambda_2)+r^2}{2r(1-r)(\lambda_1+\lambda_2)+r^2+(1-r)^2(2\lambda_1\lambda_2+1)}$$

$$\phi_5 = \frac{2r^2(\lambda_1+\lambda_2)+(\lambda_2^2+2\lambda_1\lambda_2+1)r(1-r)}{\left[(1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r\right](r+\lambda_2(1-r))}$$

$$\phi_6 = \frac{r}{r+\lambda_1(1-r)}$$

$$\phi_7 = \frac{2r^2(\lambda_1+\lambda_2)+(\lambda_1^2+2\lambda_1\lambda_2+1)r(1-r)}{\left[(1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r\right](r+\lambda_1(1-r))}$$

$$\phi_8 = \frac{2(\lambda_1+\lambda_2)r}{(1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r} \qquad (۱۲-۶)$$

نرخ نوترکیبی می‌تواند با استفاده از معادله زیر برآورد شود:

$$\hat{r} = \frac{1}{2n}\left(\phi_1 n_{21} + 2\phi_2 n_{20} + \phi_3 n_{12} + 2\phi_4 n_{11} + \phi_5 n_{10} + 2\phi_6 n_{02} + \phi_7 n_{01} + \phi_8 n_{00}\right) \qquad (۱۳-۶)$$

الگوریتم EM برای برآورد r فرموله می‌شود. در مرحله E، تعداد مورد انتظار نوترکیب‌ها برای هر ژنوتیپ با استفاده از معادلات (۶-۱۲) محاسبه می‌شود. این تعداد مورد انتظار برای تجدید برآورد r با استفاده از معادله (۶-۱۳) در مرحله M بکار گرفته می‌شود. این دو مرحله تا جایی که r در مقدار ثابت و پایداری همگرا شود، تکرار می‌شوند.

مجدداً مشاهده می‌شود که برای دو نشانگر که به صورت نادرست دسته‌بندی شده‌اند، برآورد r تحت تأثیر برآورد λ قرار نمی‌گیرد، اگرچه قبلاً دیدیم که برآورد r تحت تأثیر درجه دسته‌بندی نادرست λ قرار می‌گیرد. MLE های λ_1 و λ_2 در F_2 با حل معادلات چند جمله‌ای درجه ۳ به‌دست می‌آیند:

$$\frac{n_1}{1-\lambda_1}=\frac{n_2}{\lambda_1+\frac{r}{1-r}}+\frac{n_3}{\lambda_1+\frac{1-r+\lambda_2 r}{(1-r)\lambda_2+r}}+\frac{n_{11}}{\lambda_1+\frac{r^2+(1-r)^2+2\lambda_2 r(1-r)}{2r(1-r)+2\lambda_2(1-r)^2}}$$

$$\frac{m_1}{1-\lambda_2}=\frac{m_2}{\lambda_2+\frac{r}{1-r}}+\frac{m_3}{\lambda_2+\frac{1-r+\lambda_1 r}{(1-r)\lambda_1+r}}+\frac{n_{11}}{\lambda_2+\frac{r^2+(1-r)^2+2\lambda_1 r(1-r)}{2r(1-r)+2\lambda_1(1-r)^2}}$$

که در آن:

$$n_1=2(n_{22}+n_{21}+n_{20})+n_{12}+n_{11}+n_{10}$$
$$n_2=n_{12}+2n_{02}+n_{01}$$
$$n_3=n_{10}+2n_{00}+n_{01}$$
$$m_1=2(n_{22}+n_{12}+n_{02})+n_{21}+n_{11}+n_{01}$$
$$m_2=n_{21}+2n_{20}+n_{10}$$

واریانس‌های نمونه‌گیری MLE های مربوط به r، λ_1 و λ_2 عبارتند از:

$$\mathrm{var}(\hat{r})=\frac{\left\{(1-r)\begin{bmatrix}24(1+\lambda_1^2\lambda_2^2+\lambda_2^2-2\lambda_2^2\lambda_1-2\lambda_1-2\lambda_2+\lambda_1^2+4\lambda_1\lambda_2-2\lambda_1^2\lambda_2)r^4\\+(100\lambda_2-56\lambda_2^2+100\lambda_1-56\lambda_1^2-240\lambda_1\lambda_2-44-84\lambda_1^2\lambda_2^2+140\lambda_1^2\lambda_2+140\lambda_2^2\lambda_1)r^3\\+(-144\lambda_2^2\lambda_1+42\lambda_1^2+42\lambda_2^2+110\lambda_1^2\lambda_2^2-144\lambda_1^2\lambda_2+34-76\lambda_2+212\lambda_1\lambda_2-76\lambda_1)r^2\\+(59\lambda_2^2\lambda_1-12+59\lambda_1^2\lambda_2-64\lambda_1^2\lambda_2^2+27\lambda_1-11\lambda_1^2-82\lambda_1\lambda_2-11\lambda_2^2+27\lambda_2)r\\+(2+\lambda_2^2-7\lambda_2^2\lambda_1-7\lambda_1^2\lambda_2-3\lambda_1+14\lambda_1^2\lambda_2^2-3\lambda_2+14\lambda_1\lambda_2+\lambda_1^2)\end{bmatrix}\right\}}{\left\{2n(1-\lambda_2)(1-\lambda_1)\left[6(\lambda_1\lambda_2+1-\lambda_1-\lambda_2)r^2+2(3\lambda_1-5\lambda_1\lambda_2-3+3\lambda_2)r+4\lambda_1\lambda_2-\lambda_1-\lambda_2+2\right]\right\}}$$

$$\mathrm{var}(\hat{\lambda}_1)=\frac{1-\lambda_1^2}{2n}$$

$$\mathrm{var}(\hat{\lambda}_2)=\frac{1-\lambda_2^2}{2n}$$

۶-۵- شبیه‌سازی

اثر دسته‌بندی نادرست نشانگر بر برآورد نرخ نوترکیبی از راه شبیه‌سازی مورد بررسی قرار می‌گیرد. دو نشانگر برای جمعیت‌های تلاقی برگشتی و F_2 با درجات مختلف دسته‌بندی نادرست نشانگر (λ_1 و λ_2) شبیه‌سازی شدند. اندازه نمونه‌ها ۸۰ و ۲۰۰ در نظر گرفته شد و درجات پیوستگی (۰/۳۵ و r = ۰/۰۵) بودند. داده‌های شبیه‌سازی شده با دو مدل (۱) که دسته‌بندی نادرست نشانگر را وارد می‌کند و مدل (۲) که آنرا در نظر نمی‌گیرد، تجزیه شدند.

زمانی که $\lambda_2 = \lambda_1 = ۰$ باشد (یعنی دسته بندی نادرست نشانگر وجود نداشته باشد)، مدل‌های ۱ و ۲ به‌طور مشابه نتایج خوبی ایجاد می‌کنند (جداول ۶-۳ و ۶-۴). زمانی که یکی از نشانگرها تا حدودی دارای دسته بندی نادرست باشد، مدل ۱ به سرعت نامناسب عمل کرده و برآورد آن به ویژه برای حالت پیوستگی قوی بسیار نااریب می‌شود. برای مثال، زمانی که $\lambda_2 = ۰/۱$ باشد، مدل ۱ مقدار حقیقی r = ۰/۰۵ را بیش از ۰/۰۹ برآورد می‌کند. این مقدار وقتی که اندازه نمونه افزایش می‌یابد، بهبود نمی‌یابد. مدل ۱ می‌تواند به‌طورکلی برآوردهای خوبی از r را حتی زمانی که هر دو نشانگر به صورت نادرست دسته‌بندی می‌شوند، فراهم کند. صحت برآورد r با مدل ۱ می‌تواند با افزایش اندازه نمونه تا حد زیادی افزایش یابد. چون مدل ۲ تعداد پارامترهای کمتری نسبت به مدل ۱ دارد که باید برآورد شوند، مدل ۱ صحت برآورد بهتری نسبت به مدل ۲ نشان می‌دهد. ولی این مزیت مدل ۱ باوجود اریب بالای برآورد آن‌ها، مفید و مطلوب نیست. صحت برآورد مدل ۱ با افزایش اندازه نمونه افزایش می‌یابد.

تلاقی برگشتی و F_2 روندهای مشابهی را از نظر برآورد پارامتر نشان می‌دهند. ولی مشاهده شده است که از نظر صحت و دقت برآورد، F_2 (جدول ۶-۴) نسبت به تلاقی برگشتی (جدول ۶-۳) بهتر است. این نتیجه ممکن است به‌دلیل آن باشد که F_2 دارای مقدار اطلاعات بیشتری نسبت به تلاقی برگشتی است.

جدول ۶-۳- MLE های نرخ نوترکیبی بین دو نشانگر دارای دسته‌بندی نادرست در نتاج تلاقی برگشتی. خطای نمونه‌برداری (SE) برآوردها که از ۱۰۰ تکرار شبیه‌سازی هستند به‌دست آمده‌اند.

	مدل ۱			مدل ۲
(λ_1,λ_2,r)	$\hat{r}\pm SE$	$\hat{\lambda}_1\pm SE$	$\hat{\lambda}_2\pm SE$	$\hat{r}_0\pm SE$
n = 80				
(0.0,0.0,0.05)	0.060±0.107	−0.020±0.119	−0.019±0.116	0.054±0.026
(0.0,0.0,0.35)	0.349±0.060	0.014±0.098	0.003±0.121	0.354±0.052
(0.0,0.1,0.05)	0.040±0.095	0.004±0.109	0.109±0.106	0.099±0.032
(0.0,0.1,0.35)	0.351±0.059	−0.015±0.102	0.107±0.108	0.366±0.051
(0.2,0.3,0.05)	0.033±0.124	0.198±0.107	0.304±0.096	0.215±0.047
(0.2,0.3,0.35)	0.339±0.090	0.208±0.106	0.311±0.103	0.382±0.048
n = 200				
(0.0,0.0,0.05)	0.043±0.069	0.002±0.077	0.002±0.075	0.050±0.015
(0.0,0.0,0.35)	0.353±0.040	0.001±0.070	0.001±0.071	0.354±0.034
(0.0,0.1,0.05)	0.041±0.069	0.005±0.073	0.098±0.072	0.093±0.021
(0.0,0.1,0.35)	0.350±0.040	−0.006±0.062	0.090±0.067	0.363±0.034
(0.2,0.3,0.05)	0.043±0.073	0.206±0.064	0.310±0.068	0.220±0.028
(0.2,0.3,0.35)	0.347±0.062	0.201±0.069	0.302±0.069	0.385±0.035

جدول ۶-۴- MLE های نرخ نوترکیبی بین دو نشانگر دارای دسته‌بندی نادرست در نتاج F_2. خطای نمونه‌برداری (SE) برآوردها که از ۱۰۰ تکرار شبیه‌سازی هستند به‌دست آمده‌اند.

	مدل ۱			مدل ۲
(λ_1,λ_2,r)	$\hat{r}\pm SE$	$\hat{\lambda}_1\pm SE$	$\hat{\lambda}_2\pm SE$	$\hat{r}_0\pm SE$
n = 80				
(0.0,0.0,0.05)	0.050±0.049	−0.045±0.183	−0.033±0.169	0.048±0.018
(0.0,0.0,0.35)	0.352±0.049	−0.000±0.082	−0.005±0.073	0.354±0.044
(0.0,0.1,0.05)	0.053±0.052	−0.005±0.064	0.100±0.065	0.098±0.024
(0.0,0.1,0.35)	0.369±0.060	−0.014±0.077	0.111±0.078	0.385±0.053
(0.2,0.3,0.05)	0.073±0.063	0.196±0.079	0.296±0.071	0.225±0.034
(0.2,0.3,0.35)	0.341±0.085	0.193±0.075	0.296±0.084	0.372±0.047
n = 200				
(0.0,0.0,0.05)	0.050±0.040	−0.002±0.045	−0.003±0.046	0.049±0.012
(0.0,0.0,0.35)	0.350±0.030	−0.005±0.048	−0.005±0.046	0.349±0.028
(0.0,0.1,0.05)	0.049±0.036	0.001±0.042	0.102±0.041	0.097±0.015
(0.0,0.1,0.35)	0.356±0.034	0.001±0.051	0.096±0.045	0.371±0.030
(0.2,0.3,0.05)	0.053±0.045	0.201±0.045	0.292±0.048	0.215±0.021
(0.2,0.3,0.35)	0.350±0.056	0.200±0.051	0.293±0.045	0.373±0.030

۶-۶- تمرین‌ها

۶-۱- مدل تک ژنی در F_2

در یک آزمایش نقشه یابی، فردی می‌خواهد پیوستگی بین دو نشانگر A و B را در F_2 برآورد کند. یکی از نشانگرها (A) به‌واسطه اختلال در زنده‌مانی گامتی دارای اریبی است (از نسبت‌های ژنوتیپی مورد انتظار مندلی انحراف دارد)، ولی نشانگر دیگر (B) طبیعی است. ۹ ژنوتیپ در نتاج F_2 به‌صورت زیر مشاهده شده‌اند:

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \begin{bmatrix} n_{22}=111 & n_{21}=32 & n_{20}=1 \\ n_{12}=56 & n_{11}=123 & n_{10}=13 \\ n_{02}=7 & n_{01}=32 & n_{00}=25 \end{bmatrix} \end{array}$$

برای برآورد نرخ نوترکیبی (r) بین دو نشانگر بر اساس ماتریس فوق، الگوریتم EM را در حالات زیر به‌دست آورید:

الف. هر دو نشانگر A و B هم بارز هستند. نشان دهید چرا برآورد r به اختلال در زنده‌مانی نشانگر A بستگی ندارد.

ب. نشانگر اریب A غالب است.

ج. نشانگر طبیعی B غالب است.

د. هر دو نشانگر A و B غالب هستند.

ه. نشان دهید چرا برآورد r هنگامی‌که حداقل یکی از نشانگرها غالب باشد، به وسیله اختلال در زنده‌مانی یک نشانگر تحت تأثیر قرار می‌گیرد.

۶-۲- مدل دو ژنی در F_2

برای دو نشانگر A و B که به‌واسطه زنده‌مانی گامتی دارای اریب هستند، مشاهدات ۹ ژنوتیپ برای F_2 به صورت زیر به‌دست آمده‌اند:

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array}\begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \begin{bmatrix} n_{22}=265 & n_{21}=33 & n_{20}=3 \\ n_{12}=48 & n_{11}=39 & n_{10}=2 \\ n_{02}=2 & n_{01}=6 & n_{00}=2 \end{bmatrix} \end{array}$$

الگوریتم EM را برای برآورد نرخ نوترکیبی (r) بین دو نشانگر بر اساس ماتریس (۶-۲) بالا در حالات زیر محاسبه کنید:

الف. نشانگرهای A و B هم بارز هستند.

ب. نشانگر اریب A غالب است.

ج. نشانگر طبیعی B غالب است.

د. هر دو نشانگر A و B غالب هستند.

ه. نشان دهید چرا هنگامی‌که حداقل یکی از نشانگرها غالب است، برآورد r به وسیله اختلال در زنده‌مانی یک نشانگر تحت‌تأثیر قرار می‌گیرد.

۶-۳- اگر یک یا دو نشانگر به‌واسطه زنده‌مانی زیگوتی دارای اریبی باشند، به ترتیب موارد مشابه با آنچه که در تمرین‌های ۶-۱ و ۶-۲ مورد نیاز است را انجام دهید.

۶-۴- اگر ماتریس‌های تمرین‌های ۶-۱ و ۶-۲ حاوی نتایجی مبنی بر دسته‌بندی نادرست یکی از نشانگرها (برای مثال نشانگر A) باشند، موارد زیر را انجام دهید:

الف. ضریب دسته‌بندی نادرست نشانگر را محاسبه کرده و معنی‌داری آن را آزمون کنید.

ب. نرخ نوترکیبی را محاسبه کرده و معنی‌داری آن‌را آزمون کنید.

ج. بر اساس نتایج مطالعات شبیه‌سازی، صحت و دقت ممکن برآوردهای r خود را تجزیه کنید.

۶-۵- اگر هر دو نشانگر به صورت نادرست دسته‌بندی شده باشند، موارد مشابه با آنچه که در تمرین ۶-۳ ارائه شده را انجام دهید.

فصل هفتم

ملاحظات خاص در تجزیهٔ پیوستگی

۷-۱- مقدمه

در فصل‌های پیشین، تجزیهٔ پیوستگی را در یک تلاقی کنترل شده حاصل از لاین‌های اینبرد یا دگربارور بحث کردیم. ما یک چارچوب کلی را برای انجام تجزیه پیوستگی با هر نوع نشانگر، یعنی نشانگرهای حاوی اطلاعات کامل یا نسبی، شرح دادیم و دیدیم که امتیازدهی به صورت کیفی بر اساس حضور یا عدم حضور یک باند DNA برای یک آلل صورت گرفت. برای یک تلاقی دگربارور، روشی برای برآورد همزمان پیوستگی، دیپلوتیپ والدینی و ترتیب ژنی ارایه شد؛ اما در عمل شرایط متعددی وجود دارند که در آن‌ها، روش‌هایی را که تا این بخش برای تجزیه پیوستگی توصیف شده‌اند، نمی‌توانند به درستی و به‌طور مناسب مورد استفاده قرار گیرند. در این فصل، دو مبحث مهم را در نظر خواهیم گرفت که معمولاً تا حد زیادی در تجزیه پیوستگی مطرح هستند. اولین مبحث، در ارتباط با تجزیه پیوستگی در شجره‌های پیچیده است، در حالی که بحث دوم، تجزیه پیوستگی بر اساس اطلاعات کمّی نشانگرها است.

۷-۲- تجزیه پیوستگی در شجره‌های پیچیده

پیش‌شرط استفاده از یک تلاقی منفرد و ساده در تجزیه پیوستگی، داشتن اندازه نمونه تا حد کافی بزرگ است. اگرچه این امر می‌تواند به سادگی در گیاهان به‌دست آید، اما تولید تعداد بسیار زیاد نتاج از یک تلاقی منفرد برای بیشتر حیوانات و انسان‌ها مشکل یا حتی غیرممکن است. برای این گونه موجودات زنده، روش انجام تجزیه پیوستگی، ترکیب تمامی افراد ممکن (شامل والدین) حاصل از چندین خانواده خویشاوند یا غیرخویشاوند است. در این بخش، راهکاری برای تجزیه پیوستگی در شجره‌های چند نسلی دارای ساختار معرفی خواهد شد که با چندین والد (یا افراد پایه‌گذار[1]) شروع می‌شوند. به دلیل آنکه دیپلوتیپ‌های هر فرد پایه‌گذار ناشناخته‌اند، ازاین‌رو این راهکار باید به همراه برآورد احتمالات دیپلوتیپی والدینی بکار گرفته شود.

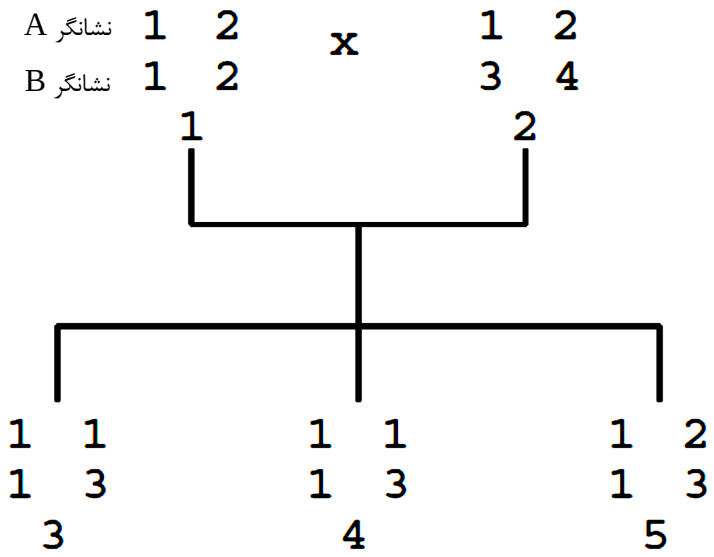
۷-۲-۱- یک خانواده هسته

برای درک اساس تجزیه پیوستگی در یک شجره پیچیده، مطالعه خود را با یک خانواده هسته منفرد شروع می‌کنیم. فرض کنید که از دو والد، سه فرزند تولید شود. همه این افراد که با شماره‌های ۱ تا ۵ نشان داده شده‌اند، برای دو نشانگر A و B تعیین ژنوتیپ می‌شوند (شکل ۷-۱).

---

[1]- افراد یا والدین پایه‌گذار، افرادی هستند که در رأس شجره قرار دارند و اولین نسل در شجره مورد نظر هستند، به طوری که والدین این افراد در شجره مشخص نیست (مترجمین).

نشانگر A دارای دو آلل مختلف است، حال آنکه نشانگر B چهار آلل دارد. توجه کنید که اگرچه ژنوتیپ‌های هر فرد می‌تواند از روی تجزیه ژل مشخص شود، اما فاز پیوستگی بین دو نشانگر نمی‌تواند اندازه‌گیری شود. در این مثال، فازهای پیوستگی هر دو والد و فرزند شماره ۵ شناخته شده نیستند، اما چون فازهای پیوستگی مختلف برای این افراد روی استنباط و برآورد نرخ نوترکیبی (r) بین دو نشانگر اثر می‌گذارند، ازاین‌رو این افراد باید به‌طور صحیح ارزیابی شوند.



شکل ۷-۱- یک خانواده هسته با سه فرزند

هر یک از سه فردی که دارای فاز ناشناخته هستند، دارای دو فاز ممکن هستند که عبارتند از:

برای والد ۱ $\Lambda_1 = \begin{matrix}1 & 2\\ 2 & 1\end{matrix}$ یا $\Lambda_1 = \begin{matrix}1 & 2\\ 1 & 2\end{matrix}$

برای والد ۲ $\overline{\Lambda}_2 = \begin{matrix}1 & 2\\ 4 & 3\end{matrix}$ یا $\Lambda_2 = \begin{matrix}1 & 2\\ 3 & 4\end{matrix}$

برای فرزند ۵ $\Lambda_5 = \begin{matrix}1 & 2\\ 1 & 3\end{matrix}$ یا $\overline{\Lambda}_5 = \begin{matrix}1 & 2\\ 3 & 1\end{matrix}$

که ۸ ترکیب دیپلوتیپی ممکن را شکل می‌دهند. در تجزیه پیوستگی مرسوم، فرض می‌شود که این ترکیبات دارای احتمال یکسان $\frac{1}{8}$ هستند. جدول ۷-۱ تعداد نوترکیب‌ها (R) و

غیر نوترکیب‌های (NR) ممکن را برای هر یک از این نتاج و درستنمایی حاصل را تحت ترکیب دیپلوتیپی مختلف نشان می‌دهد.

بر اساس درستنمایی هر ترکیب دیپلوتیپی، یک درستنمایی کلی برای این خانواده به صورت زیر ایجاد می‌کنیم:

$$\begin{aligned} L(r) &= \frac{1}{8}\left(L_1 + \ldots + L_8\right) \\ &= \frac{1}{8}\left[2r\left(1-r\right)^5 + r^2\left(1-r\right)^4 + r^4\left(1-r\right)^2 + r^5\left(1-r\right)\right] \\ &= \frac{1}{4}r\left(1-r\right)\left[1-3r\left(1-r\right)\right] \end{aligned} \qquad (۱-۷)$$

با بیشینه نمودن درستنمایی (۷-۱)، MLE مربوط به نرخ نوترکیبی (r) به صورت $\hat{r}$ = ۰/۲۱ به‌دست می‌آید. یک راهکار تقریبی برای برآورد r با استفاده از درستنمایی پیچیده (۷-۱)، محاسبه درستنمایی برای رشته‌ای از مقادیر ثابت r به صورت $r = \left(0.00, 0.01, 0.05, 0.10, 0.15, \ldots\right)$ است. آن مقدار r که منجر به بیشینه شدن درستنمایی شود، به‌عنوان MLE نرخ نوترکیبی در نظر گرفته می‌شود (شکل ۷-۲). از لحاظ مجانبی، منحنی درستنمایی به شکل منحنی گوس (منحنی توزیع نرمال) نزدیک شده و به آن می‌رسد (آت، ۱۹۹۱).

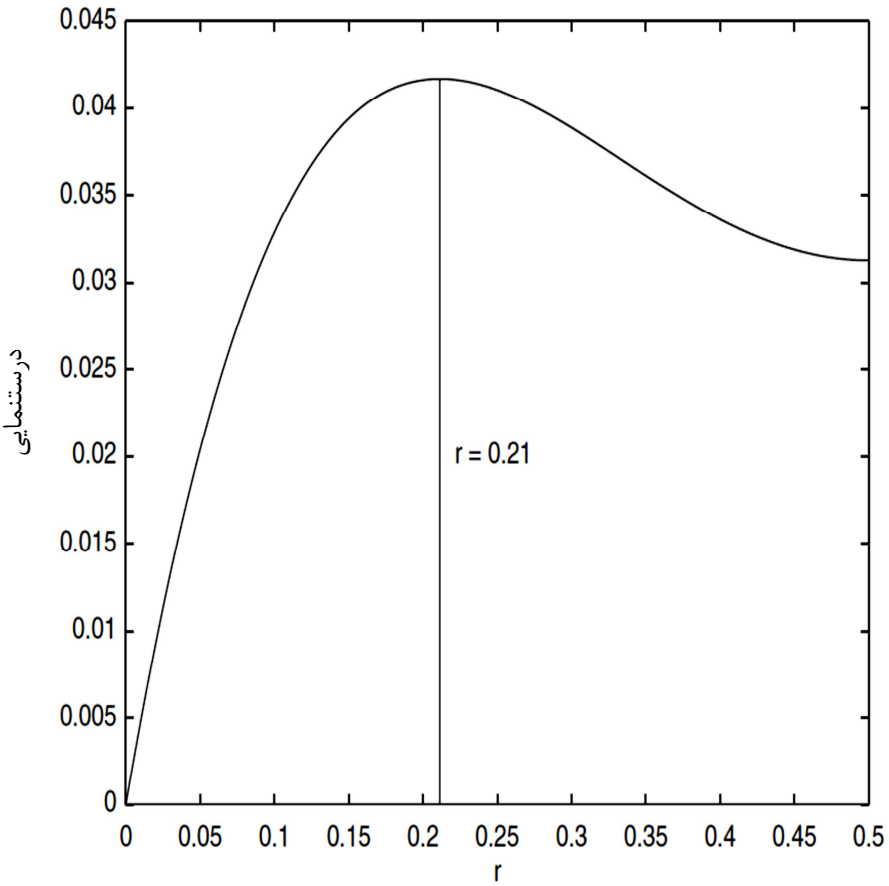
روشی که در بخش بالا برای تجزیه پیوستگی تشریح شد، احتمال مساوی را برای تمامی ۸ ترکیب دیپلوتیپی ممکن فرض می‌کند. ولی ممکن است این فرض صحیح نباشد. محتمل‌ترین ترکیب دیپلوتیپی می‌تواند از روی داده‌ها، با محاسبه MLE مربوط به r با استفاده از درستنمایی ارایه شده در جدول ۷-۲ تعیین شود. نتایج هر احتمال دیپلوتیپی در بخش زیر نشان داده شده است.

می‌توان مشاهده کرد که ترکیب دیپلوتیپی بهینه با بزرگترین درستنمایی، به صورت $\Lambda_1 - \Lambda_2 - \Lambda_5$ یا $\Lambda_1 - \Lambda_2 - \overline{\Lambda}_5$ است که احتمال آن به صورت زیر محاسبه می‌شود:

$$\hat{p}_2 \text{ یا } \hat{p}_1 = \frac{\hat{L}_1 or \hat{L}_2}{\hat{L}_1 + \ldots + \hat{L}_8} = 0.2787$$

جدول ۷-۱- تعداد نوترکیب‌ها (R) و غیرنوترکیب‌ها (NR) برای سه نتاج حاصل از یک خانواده با ترکیبات دیپلوتیپی مختلف برای والدین ۱ و ۲ و فرزند شماره ۵ که در شکل ۷-۱ نشان داده شده‌اند. همچنین درستنمایی‌ها برای هر ترکیب دیپلوتیپی ارایه شده‌اند.

ترکیب دیپلوتیپی				فرزند			
احتمال	والد ۱	والد ۲	فرزند ۵	۳	۴	۵	درستنمایی
$p_1=\frac{1}{8}$	Λ_1	Λ_2	Λ_5	NR,NR	NR,NR	NR, R	$L_1=r(1-r)^5$
$p_2=\frac{1}{8}$	Λ_1	Λ_2	$\overline{\Lambda}_5$	NR,NR	NR,NR	NR,R	$L_2=r(1-r)^5$
$p_3=\frac{1}{8}$	Λ_1	$\overline{\Lambda}_2$	Λ_5	NR,NR	NR, R	NR,R	$L_3=r^2(1-r)^4$
$p_4=\frac{1}{8}$	Λ_1	$\overline{\Lambda}_2$	$\overline{\Lambda}_5$	NR,R	NR,R	R,R	$L_4=r^4(1-r)^2$
$p_5=\frac{1}{8}$	$\overline{\Lambda}_1$	Λ_2	Λ_5	R,NR	R,NR	R,R	$L_5=r^4(1-r)^2$
$p_6=\frac{1}{8}$	$\overline{\Lambda}_1$	Λ_2	$\overline{\Lambda}_5$	R,NR	R,NR	NR,NR	$L_6=r^2(1-r)^4$
$p_7=\frac{1}{8}$	$\overline{\Lambda}_1$	$\overline{\Lambda}_2$	Λ_5	R,R	R,R	R,NR	$L_7=r^5(1-r)$
$p_8=\frac{1}{8}$	$\overline{\Lambda}_1$	$\overline{\Lambda}_2$	$\overline{\Lambda}_5$	R,R	R,R	NR,R	$L_8=r^5(1-r)$



شکل ۷-۲- پروفیل درستنمایی محاسبه شده برای یک خانواده هسته در شکل ۷-۳.

جدول ۷-۲- MLE نرخ نوترکیبی و ترکیب شده با مقادیر درستنمایی تحت ترکیبات مختلف دیپلوتیپی.

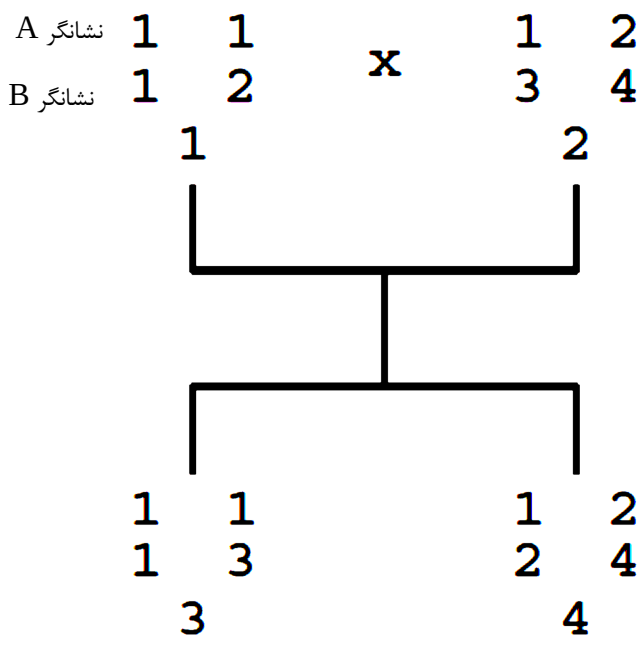
ترکیب دیپلوتیپی				$\hat{r}$	لگاریتم درستنمایی
احتمال	والد ۱	والد ۲	فرزند ۵		
$\hat{p}_1 = 0.2787$	Λ_1	Λ_2	Λ_5	$\frac{1}{6}$	$\log \hat{L}_1 = -2.7034$
$\hat{p}_2 = 0.2787$	Λ_1	Λ_2	$\overline{\Lambda}_5$	$\frac{1}{6}$	$\log \hat{L}_2 = -2.7034$
$\hat{p}_3 = 0.0913$	Λ_1	$\overline{\Lambda}_2$	Λ_5	$\frac{1}{3}$	$\log \hat{L}_3 = -3.8191$
$\hat{p}_4 = 0.0650$	Λ_1	$\overline{\Lambda}_2$	$\overline{\Lambda}_5$	$\frac{1}{2}$	$\log \hat{L}_4 = -4.1589$
$\hat{p}_5 = 0.0650$	$\overline{\Lambda}_1$	Λ_2	Λ_5	$\frac{1}{2}$	$\log \hat{L}_5 = -4.1589$
$\hat{p}_6 = 0.0913$	$\overline{\Lambda}_1$	Λ_2	$\overline{\Lambda}_5$	$\frac{1}{3}$	$\log \hat{L}_6 = -3.8191$
$\hat{p}_7 = 0.0650$	$\overline{\Lambda}_1$	$\overline{\Lambda}_2$	Λ_5	$\frac{1}{2}$	$\log \hat{L}_7 = -4.1589$
$\hat{p}_8 = 0.0650$	$\overline{\Lambda}_1$	$\overline{\Lambda}_2$	$\overline{\Lambda}_5$	$\frac{1}{2}$	$\log \hat{L}_8 = -4.1589$

یک روش دیگر برای تعیین ترکیب دیپلوتیپی بهینه، ایجاد یک درستنمایی مخلوط از تمامی ترکیبات دیپلوتیپی وزن داده شده با احتمال فرضی آن‌ها است. می‌توان الگوریتمی را بکار گرفت تا برآورد همزمان r و p ها را انجام دهد. ولی این روش نیاز به اندازه نمونه بزرگ‌تر نسبت به روش تشریح شده در بالا دارد، چون تعداد پارامترهای بیشتری برآورد می‌شوند.

۷-۲-۲- برآورد چند نقطه‌ای اشتراک‌گذاری آلل‌های یکسان اجدادی

دو آلل در یک جایگاه ژنی منفرد، یکسان اجدادی (IBD) نامیده می‌شوند، اگر کپی‌های مشابه یک آلل روی کروموزوم مادری و پدری باشند، به‌طوری که هر دو کپی با تکثیر DNA از توالی اجدادی مشابه بدون دخالت هر گونه جهشی حاصل می‌شوند. مفهوم IBD استفاده گسترده‌ای برای درک ژنتیک جمعیت داشته و همچنین استفاده متعددی در تجزیه پیوستگی و تهیه نقشه‌های ژنتیکی دارد. برآورد احتمالات IBD هنگامی می‌تواند از روی تجزیه پیوستگی به‌دست آید که ژنوتیپ‌های چندین نشانگر به هم پیوسته در دسترس باشند. برای خانواده‌ای که در شکل ۷-۳ نشان داده شده

است، میانگین IBD به اشتراک گذاشته شده برای برادر- خواهران در شرایطی که دو نشانگر ناپیوسته باشند، برای نشانگر A برابر ۰/۲۵ و برای نشانگر B برابر صفر است.



شکل ۷-۳- ژنوتیپ‌های دو نشانگر مجاور در یک خانواده هسته.

برآورد IBD برای نشانگر حاوی اطلاعات کمتر (A) می‌تواند از راه درستنمایی توأم دو نشانگر تحت شرایطی به‌دست آید که دو نشانگر به هم پیوسته باشند. r را به عنوان نرخ نوترکیبی بین دو نشانگر در نظر بگیرید. برای نشانگر A، تعیین والد منشأ دو آلل مشابه ۱ که به وسیله والد ۱ حمل می‌شود، ممکن نیست؛ بنابراین، نمی‌توان فاز پیوستگی والد ۱ را تعیین نمود. به‌طور اختیاری یک آلل از والد پدری (که به صورت 1^p نشان داده شده است) و دیگری از والد مادری (که به صورت 1^m نشان داده شده است) را در نظر می‌گیریم. دو فاز ممکن والد ۱ به‌صورت زیر بیان می‌شوند:

$$\Lambda_1 = \begin{matrix} 1^p \\ 1 \end{matrix}\left|\begin{matrix} 1^m \\ 2 \end{matrix}\right| \quad \text{یا} \quad \overline{\Lambda}_1 = \begin{matrix} 1^p \\ 2 \end{matrix}\left|\begin{matrix} 1^m \\ 1 \end{matrix}\right|$$

به‌طور مشابه والد ۲ نیز دارای دو فاز ممکن به صورت زیر است:

$$\Lambda_2 = \begin{matrix} 1 \\ 3 \end{matrix}\left|\begin{matrix} 2 \\ 4 \end{matrix}\right| \quad \text{یا} \quad \overline{\Lambda}_2 = \begin{matrix} 1 \\ 4 \end{matrix}\left|\begin{matrix} 2 \\ 3 \end{matrix}\right|$$

فرزند شماره ۳ دارای دو آلل مشابه برای نشانگر A است، ولی یکی و فقط یکی از آن‌ها، یا 1^m یا 1^p، از والد ۱ منشأ گرفته‌اند؛ بنابراین، فاز پیوستگی ممکن آن بایستی بدین صورت باشد:

$$\Lambda_3^p = \begin{matrix} 1^p \\ 1 \end{matrix}\left|\begin{matrix} 1 \\ 3 \end{matrix}\right| \text{ یا } \Lambda_3^m = \begin{matrix} 1^m \\ 1 \end{matrix}\left|\begin{matrix} 1 \\ 3 \end{matrix}\right|$$

برای فرزند شماره ۳ این امکان وجود ندارد که ترکیب زیر را داشته باشد:

$$\begin{matrix} 1^p \\ 3 \end{matrix}\left|\begin{matrix} 1 \\ 1 \end{matrix}\right| \text{ یا } \begin{matrix} 1^m \\ 3 \end{matrix}\left|\begin{matrix} 1 \\ 1 \end{matrix}\right|$$

چون گامت‌هایی که این آرایش ژنوتیپی را شکل دهند، وجود ندارند. به دلیل مشابه، فرزند شماره ۴ فقط دو فاز ممکن به صورت زیر دارد:

$$\Lambda_4^p = \begin{matrix} 1^p \\ 2 \end{matrix}\left|\begin{matrix} 2 \\ 4 \end{matrix}\right| \text{ یا } \Lambda_4^m = \begin{matrix} 1^m \\ 2 \end{matrix}\left|\begin{matrix} 2 \\ 4 \end{matrix}\right|$$

با ترکیب چهار فرد، ۱۶ ترکیب دیپلوتیپی خواهیم داشت که درستنمایی‌ها و احتمالات IBD برای هر نشانگر در جدول ۷-۳ ارائه شده است. با فرض وجود احتمال مشابه برای ۱۶ ترکیب دیپلوتیپی، درستنمایی توأم را برای به‌دست آوردن MLE مربوط به r ایجاد می‌کنیم. با داشتن $\hat{r}$، مشارکت IBD برای هر نشانگر می‌تواند محاسبه شود که برای نشانگر A برابر است با:

$$\frac{0.5\left[4r(1-r)^3+4r^3(1-r)\right]}{2(1-r)^4+4r(1-r)^3+4r^2(1-r)^2+4r^3(1-r)+2r^4}=\frac{r(1-r)^3+r^3(1-r)}{1-2r(1-r)}$$

و برای نشانگر B برابر با صفر است.

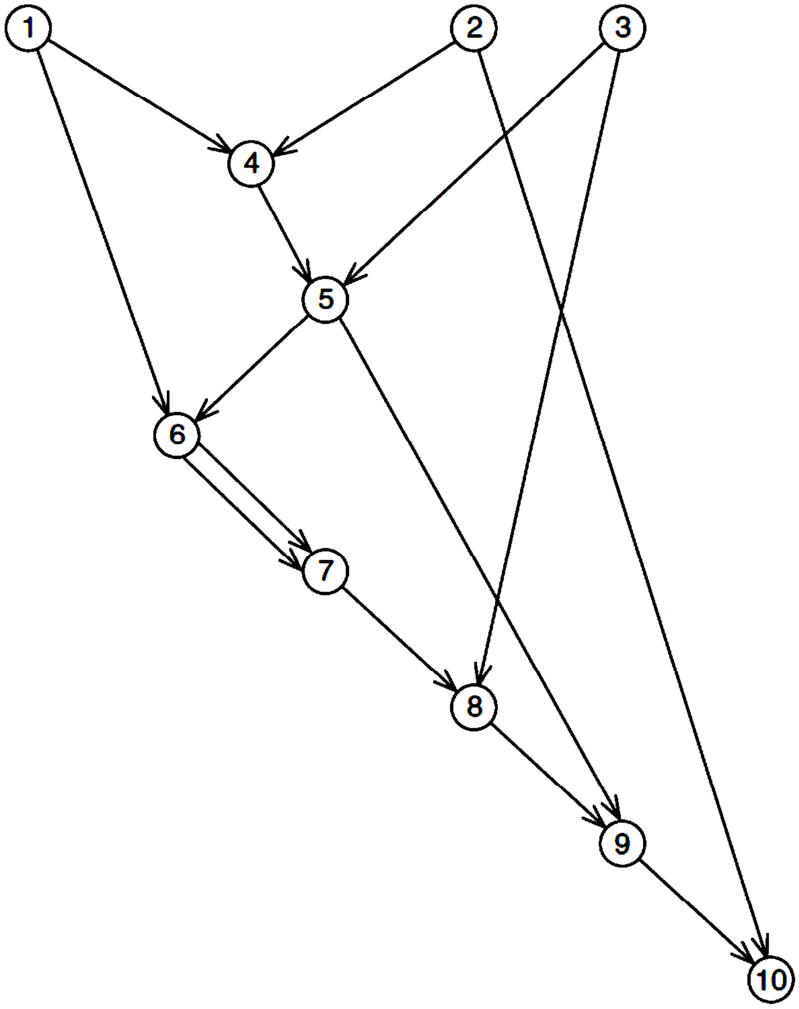
زمانی که تعداد زیادی نشانگر برای تجزیه IBD در نظر گرفته می‌شوند، درستنمایی می‌تواند بر حسب مدل مارکوف[1] فاکتوربندی شود، به‌طوری که IBD یک نشانگر و احتمال انتقال از یک نشانگر به نشانگر دیگر را مورد توجه قرار داد.

۷-۲-۳- یک شجره پیچیده

چون اعضای خانواده‌های مختلف می‌توانند از راه اجداد مشترک خود با یکدیگر مرتبط شوند، از این‌رو شجره‌ای که در نظر می‌گیریم یک شبکه پیچیده را تشکیل می‌دهد. شکل ۷-۴ نمودار چنین شجره پیچیده‌ای را نشان می‌دهد که می‌تواند در گونه‌های مختلف زیادی مورد استفاده قرار گیرد. فرض کنید مجموعاً ۱۰ فرد وجود دارند که در دو نشانگر چند آللی A و B و یک نشانگر دو آللی

[1] - Markov

C تعیین ژنوتیپ شده‌اند. آلل‌ها در هر یک از این نشانگرها با اعداد انگلیسی نشان داده شده‌اند. افراد ۱، ۲ و ۳ افراد پایه‌گذار هستند که والدین آن‌ها قابل شناسایی نیستند، حال آنکه افراد ۵ تا ۱۰ افراد غیر پایه‌گذار هستند که از روی افراد دارای والدین مشخص مشتق شده‌اند. تلاقی افراد ۱ و ۲، فرد ۴ را به وجود می‌آورد که با فرد ۳ تلاقی کرده تا فرد ۵ را ایجاد نماید. فرد ۱ نیز با فرد ۵ تلاقی داده شده است تا فرد ۶ را به وجود آورد و این فرد خودگشن شده است تا فرد ۷ را ایجاد کند. فرد ۸ نیز از تلاقی افراد ۳ و ۷ ایجاد می‌شود. تلاقی افراد ۸ و ۵ سبب ایجاد فرد ۹ می‌شود که با فرد ۲ تلاقی داده شده است تا فرد ۱۰ را به وجود آورد. این روابط شجره‌ای بعداً در جدول ۷-۴ توصیف شده است. ستون ۱ نشان‌دهنده کل افرادی است که مورد مطالعه قرار گرفته‌اند، ستون ۲ و ۳ منشأ والدینی آن‌ها (توجه کنید که اطلاعات والدینی افراد پایه‌گذار موجود نیست) و ستون‌های ۴ تا ۶ ژنوتیپ‌های سه نشانگر A، B و C را به ترتیب برای هر فرد نشان می‌دهند. در این مثال، فاز پیوستگی هر فرد برای این نشانگرها ناشناخته است.



شکل ۷-۴- ساختار یک شجره فرضی.

ایجاد رابطه درستنمایی

پیش‌شرط تجزیه پیوستگی، وجود فاز پیوستگی شناخته شده بین دو نشانگر مورد نظر برای افراد پایه‌گذاری است که نتاج در حال تفرق را ایجاد می‌کنند. در مثال ارایه شده در جدول ۷-۴ که احتمالاً در عمل یک حالت واقعی است، داده‌های نشانگر برای افراد پایه گذار غیراینبرد، فقط ژنوتیپ‌ها را نشان می‌دهند (که به صورت ς نشان داده شده‌اند)، ولی هیچ‌گونه اطلاعاتی در خصوص شکل فاز پیوستگی یا دیپلوتیپ ارایه نمی‌کنند؛ اما همیشه می‌توان یک دیپلوتیپ ممکن را

برای یک فرد پایه‌گذار فرض نمود. برای دو نشانگر نخست، هر یک از سه فرد پایه‌گذار دارای دو دیپلوتیپ جایگزین (که به صورت Λ یا $\overline{\Lambda}$ نشان داده شده‌اند) هستند که ۸ ترکیب مختلف (جدول ۷-۵) را شکل می‌دهند.

جدول ۷-۳- درستنمایی‌ها و احتمالات IBD برای هر نشانگر تحت ۱۶ ترکیب دیپلوتیپی ممکن برای شجره نشان داده شده در شکل ۷-۳.

ترکیب فاز					فرزند				
احتمال	۱	۲	۳	٤	۳	٤	درستنمایی $\propto$	IBD (A)	IBD (B)
$p_1=\frac{1}{16}$	Λ_1	Λ_2	Λ_3^p	Λ_4^p	NR, NR	R, NR	$r(1-r)^3$	۰/۵	۰
$p_2=\frac{1}{16}$	Λ_1	Λ_2	Λ_3^p	Λ_4^m	NR, NR	NR, NR	$(1-r)^4$	۰	۰
$p_3=\frac{1}{16}$	Λ_1	Λ_2	Λ_3^m	Λ_4^p	R, NR	R, NR	$r^2(1-r)^2$	۰	۰
$p_4=\frac{1}{16}$	Λ_1	Λ_2	Λ_3^m	Λ_4^m	R, NR	NR, NR	$r(1-r)^3$	۰/۵	۰
$p_5=\frac{1}{16}$	Λ_1	$\overline{\Lambda}_2$	Λ_3^p	Λ_4^p	NR, R	R, R	$r^3(1-r)$	۰/۵	۰
$p_6=\frac{1}{16}$	Λ_1	$\overline{\Lambda}_2$	Λ_3^p	Λ_4^m	NR, R	NR, R	$r^2(1-r)^2$	۰	۰
$p_7=\frac{1}{16}$	Λ_1	$\overline{\Lambda}_2$	Λ_3^m	Λ_4^p	R, R	R, R	r^4	۰	۰
$p_8=\frac{1}{16}$	Λ_1	$\overline{\Lambda}_2$	Λ_3^m	Λ_4^m	R, R	NR, R	$r^3(1-r)$	۰/۵	۰
$p_9=\frac{1}{16}$	$\overline{\Lambda}_1$	Λ_2	Λ_3^p	Λ_4^p	R, NR	NR, NR	$r(1-r)^3$	۰/۵	۰
$p_{10}=\frac{1}{16}$	$\overline{\Lambda}_1$	Λ_2	Λ_3^p	Λ_4^m	R, NR	R, NR	$r^2(1-r)^2$	۰	۰
$p_{11}=\frac{1}{16}$	$\overline{\Lambda}_1$	Λ_2	Λ_3^m	Λ_4^p	NR, NR	NR, NR	$(1-r)^4$	۰	۰
$p_{12}=\frac{1}{16}$	$\overline{\Lambda}_1$	Λ_2	Λ_3^m	Λ_4^m	NR, NR	R, NR	$r(1-r)^3$	۰/۵	۰
$p_{13}=\frac{1}{16}$	$\overline{\Lambda}_1$	$\overline{\Lambda}_2$	Λ_3^p	Λ_4^p	R, R	NR, R	$r^3(1-r)$	۰/۵	۰
$p_{14}=\frac{1}{16}$	$\overline{\Lambda}_1$	$\overline{\Lambda}_2$	Λ_3^p	Λ_4^m	R, R	R, R	r^4	۰	۰
$p_{15}=\frac{1}{16}$	$\overline{\Lambda}_1$	$\overline{\Lambda}_2$	Λ_3^m	Λ_4^p	NR, R	NR, R	$r^2(1-r)^2$	۰	۰
$p_{16}=\frac{1}{16}$	$\overline{\Lambda}_1$	$\overline{\Lambda}_2$	Λ_3^m	Λ_4^m	NR, R	R, R	$r^3(1-r)$	۰/۵	۰

اولین ترکیب دیپلوتیپی را در نظر بگیرید. افراد پایه‌گذار ۱ و ۲ آمیزش می‌کنند تا فرد ۴ را ایجاد کنند که بر اساس مشاهدات ژل، ژنوتیپ ۱۳ را در نشانگر A و ژنوتیپ ۱۲ را در نشانگر B دارد. ممکن است آلل‌های دو نشانگر در فرد ۴ در دو فاز پیوستگی بدین صورت مرتب شوند:

$$\Lambda_4 = \begin{matrix}1\\2\end{matrix}\left|\begin{matrix}3\\1\end{matrix}\right| \qquad \text{یا} \qquad \overline{\Lambda}_4 = \begin{matrix}1\\1\end{matrix}\left|\begin{matrix}3\\2\end{matrix}\right| \tag{۲-۷}$$

بر اساس فازهای پیوستگی در نظر گرفته شده برای افراد پایه‌گذار ۱ و ۲، ترتیب دوم به‌دلیل نبود تولید گامت ۳۲ به وسیله افراد پایه‌گذار هنوز وجود ندارد. تنها ترتیب ممکن همان ترتیب اول است که تشکیل آن به‌واسطه ترکیب گامت نوترکیب ۱۲ از فرد پایه‌گذار ۱ و گامت نوترکیب ۳۱ از فرد پایه‌گذار ۲ امکان‌پذیر است. اگر r نرخ نوترکیبی بین این دو نشانگر باشد، عبارت $P\left(\Lambda(t)\,|\,\Lambda(t-1)\right)$ احتمال شرطی دیپلوتیپ یک فرد (نسل t) مشروط به دیپلوتیپ‌های والدینیش (نسل t- ۱) و عبارت $P\left(\Lambda(t-1)\,|\,\varsigma(t-1)\right)$ احتمال شرطی دیپلوتیپ یک والد (نسل t- ۱) مشروط به ژنوتیپ مشاهده شده آن است؛ بنابراین، امکان تولید ژنوتیپ مشاهده شده برای فرد ۴ باید بدین صورت باشد:

$$\begin{aligned} P\left(\varsigma_4\,|\,\Lambda_1,\Lambda_2\right) &= P\left(\Lambda_4\,|\,\Lambda_1,\Lambda_2\right) + P\left(\overline{\Lambda}_4\,|\,\Lambda_1,\Lambda_2\right) \\ &= P\left(12\,|\,\Lambda_1\right)P\left(31\,|\,\Lambda_2\right) \\ &= \frac{1}{4}r^2 \end{aligned} \tag{۳-۷}$$

که در آن، $P\left(\overline{\Lambda}_4\,|\,\Lambda_1,\Lambda_2\right) = ۰$ است. برای فرد ۵ حاصل از فرد پایه گذار ۳ و هیبرید ۴، تنها دیپلوتیپ ممکن به‌صورت زیر است:

$$\Lambda_5 = \begin{matrix}1\\2\end{matrix}\left|\begin{matrix}1\\2\end{matrix}\right|$$

بنابراین، احتمال ژنوتیپی فرد ۵ به‌صورت زیر بیان می‌شود:

(۴-۷)

$$\begin{aligned} P\left(\varsigma_5\,|\,\Lambda_3,\varsigma_4\right) &= P\left(\varsigma_5\,|\,\Lambda_3,\Lambda_4\right)P\left(\Lambda_4\,|\,\varsigma_4\right) \\ &= P\left(\Lambda_5\,|\,\Lambda_3,\Lambda_4\right)P\left(\Lambda_4\,|\,\varsigma_4\right) + P\left(\overline{\Lambda}_5\,|\,\Lambda_3,\Lambda_4\right)P\left(\overline{\Lambda}_4\,|\,\varsigma_4\right) \\ &= \frac{1}{2}r\left(1-r\right) \end{aligned}$$

که در آن، $P\left(\Lambda_4 \mid \varsigma_4\right) = ۱$ و $P\left(\overline{\Lambda}_5 \mid \Lambda_3, \Lambda_4\right) = ۰$ است. برای فرد ۶ حاصل از افراد ۱ و ۵، آلل‌های ۱ و ۲ را در نشانگر A و آلل‌های ۱ و ۲ را در نشانگر B مشاهده نمودیم. اگرچه این فـرد ممکن است ترتیب‌های آللی به شکل زیر داشته باشد:

$$\Lambda_6 = \begin{matrix} 2 \\ 2 \end{matrix}\left|\begin{matrix} 1 \\ 1 \end{matrix}\right| \qquad \text{یا} \qquad \overline{\Lambda}_6 = \begin{matrix} 2 \\ 1 \end{matrix}\left|\begin{matrix} 1 \\ 2 \end{matrix}\right|$$

فقط مورد اخیر مشروط به ژنوتیپ فرد ۵ امکان‌پذیر اسـت؛ بنـابراین، احتمـال ژنـوتیپی فـرد ۶ عبارت است از:

$$P\left(\varsigma_6 \mid \Lambda_1, \varsigma_5\right) = P\left(\overline{\Lambda}_6 \mid \Lambda_1, \Lambda_5\right) = \frac{1}{2} r \qquad (۵-۷)$$

جدول ۷-۴- ژنوتیپ‌های مشاهده شده برای سه نشانگر مفروض در یک شجره با ساختار مشخص که در شکل ۷-۴ نشان داده شده است.

ID فرد	والد ۱	والد ۲	ژنوتیپ نشانگر		
			A	B	C
۱	-	-	۱۲	۱۲	۱۲
۲	-	-	۳۴	۱۳	۱۲
۳	-	-	۱۳	۱۲	۱۲
۴	۱	۲	۱۳	۱۲	۱۲
۵	۳	۴	۱۱	۲۲	۱۲
۶	۱	۵	۱۲	۱۲	۱۲
۷	۶	۶	۱۲	۱۲	۱۲
۸	۳	۷	۱۳	۱۲	۲۲
۹	۵	۸	۱۳	۱۲	۱۲
۱۰	۲	۹	۱۴	۲۳	۱۱

خودگشنی فرد ۶ سبب تولید فرد هتروزیگوت مضاعف ۷ با دو دیپلوتیپ ممکن به صـورت زیـر می‌شود:

$$\Lambda_7 = \begin{matrix} 1 \\ 1 \end{matrix}\left|\begin{matrix} 2 \\ 2 \end{matrix}\right| \qquad \text{یا} \qquad \overline{\Lambda}_7 = \begin{matrix} 1 \\ 2 \end{matrix}\left|\begin{matrix} 2 \\ 1 \end{matrix}\right|$$

که کل احتمال ژنوتیپی باید بدین صورت باشد:

$$\begin{aligned}
&P(\varsigma_7 \mid \varsigma_6) = P(\Lambda_7 \mid \varsigma_6) + P(\overline{\Lambda}_7 \mid \varsigma_6) \\
&= P(\Lambda_7 \mid \Lambda_6) P(\Lambda_6 \mid \varsigma_6) + P(\Lambda_7 \mid \overline{\Lambda}_6) P(\overline{\Lambda}_6 \mid \varsigma_6) \\
&+ P(\overline{\Lambda}_7 \mid \Lambda_6) P(\Lambda_6 \mid \varsigma_6) + P(\overline{\Lambda}_7 \mid \overline{\Lambda}_6) P(\overline{\Lambda}_6 \mid \varsigma_6) \\
&= P(\Lambda_7 \mid \overline{\Lambda}_6) P(\overline{\Lambda}_6 \mid \varsigma_6) + P(\overline{\Lambda}_7 \mid \overline{\Lambda}_6) P(\overline{\Lambda}_6 \mid \varsigma_6) \\
&= \frac{1}{2} r^2 + \frac{1}{2}(1-r)^2
\end{aligned} \tag{۶-۷}$$

که در این روابط، $P(\Lambda_6 \mid \varsigma_6) = ۰$ و $P(\overline{\Lambda}_6 \mid \varsigma_6) = ۱$. فرد ۷ با فرد پایه گذار ۳ آمیزش می‌کند تا فرد ۸ را با آلل‌های مشاهده شده ۱ و ۳ در نشانگر A و آلل‌های ۱ و ۲ در نشانگر B تولید نماید. ممکن است فرد ۸ یکی از دو دیپلوتیپ زیر را داشته باشد:

$$\Lambda_8 = \begin{matrix} 1 & 3 \\ 2 & 1 \end{matrix} \qquad \text{یا} \qquad \overline{\Lambda}_8 = \begin{matrix} 1 & 3 \\ 1 & 2 \end{matrix}$$

کل احتمال ژنوتیپی فرد ۸ مشروط به ژنوتیپ والدینش بدین صورت بیان می‌شود:

$$\begin{aligned}
&P(\varsigma_8 \mid \Lambda_3, \varsigma_7) = P(\Lambda_8 \mid \Lambda_3, \varsigma_7) + P(\overline{\Lambda}_8 \mid \Lambda_3, \varsigma_7) \\
&= P(\Lambda_8 \mid \Lambda_3, \varsigma_7) P(\Lambda_7 \mid \varsigma_7) + P(\Lambda_8 \mid \Lambda_3, \overline{\Lambda}_7) P(\overline{\Lambda}_7 \mid \varsigma_7) \\
&+ P(\overline{\Lambda}_8 \mid \Lambda_3, \varsigma_7) P(\Lambda_7 \mid \varsigma_7) + P(\overline{\Lambda}_8 \mid \Lambda_3, \overline{\Lambda}_7) P(\overline{\Lambda}_7 \mid \varsigma_7) \\
&= \frac{r\left[r^3 + (1-r)^3 + (1-r)^2\right]}{4\left[r^2 + (1-r)^2\right]}
\end{aligned} \tag{۷-۷}$$

که در آن، $P(\Lambda_8 \mid \Lambda_3, \varsigma_7) = \frac{1}{4} r^2$ است.

به ترتیب با هاپلوتیپ‌های ۱۲ و ۳۱ از فرد Λ_8 حاصل از Λ_7 و Λ_3 که احتمال هر یک برابر با $\frac{1}{2} r$ است (توجه کنید که Λ_7 فقط از $\overline{\Lambda}_6$ به‌دست می‌آید، به‌طوری که فراوانی هاپلوتیپی ۱۲ برابر با $\frac{1}{2} r$ است):

$$P\left(\Lambda_8 \mid \Lambda_3, \overline{\Lambda}_7\right) = \frac{1}{4} r\left(1-r\right)$$

$$P\left(\overline{\Lambda}_8 \mid \Lambda_3, \Lambda_7\right) = \frac{1}{4}\left(1-r\right)^2$$

$$P\left(\overline{\Lambda}_8 \mid \Lambda_3, \overline{\Lambda}_7\right) = \frac{1}{4} r\left(1-r\right)$$

$$P\left(\Lambda_7 \mid \varsigma_7\right) = \frac{r^2}{\left(1-r\right)^2 + r^2}$$

$$P\left(\overline{\Lambda}_7 \mid \varsigma_7\right) = \frac{\left(1-r\right)^2}{\left(1-r\right)^2 + r^2}$$

جدول ۷-۵- هشت ترکیب دیپلوتیپی برای افراد پایه‌گذار.

ترکیب دیپلوتیپی	فرد پایه‌گذار ۱	فرد پایه‌گذار ۲	فرد پایه‌گذار ۳	احتمال
۱ = $\Lambda_1\Lambda_2\Lambda_3$	$\begin{matrix}1 & 2\\ 1 & 2\end{matrix}$	$\begin{matrix}3 & 4\\ 3 & 1\end{matrix}$	$\begin{matrix}1 & 3\\ 1 & 2\end{matrix}$	$p_1 p_2 p_3$
۲ = $\Lambda_1\Lambda_2\overline{\Lambda}_3$	$\begin{matrix}1 & 2\\ 1 & 2\end{matrix}$	$\begin{matrix}3 & 4\\ 3 & 1\end{matrix}$	$\begin{matrix}1 & 3\\ 2 & 1\end{matrix}$	$p_1 p_2 \left(1-p_3\right)$
۳ = $\Lambda_1\overline{\Lambda}_2\Lambda_3$	$\begin{matrix}1 & 2\\ 1 & 2\end{matrix}$	$\begin{matrix}3 & 4\\ 1 & 3\end{matrix}$	$\begin{matrix}1 & 3\\ 1 & 2\end{matrix}$	$p_1 \left(1-p_2\right) p_3$
۴ = $\Lambda_1\overline{\Lambda}_2\overline{\Lambda}_3$	$\begin{matrix}1 & 2\\ 1 & 2\end{matrix}$	$\begin{matrix}3 & 4\\ 1 & 3\end{matrix}$	$\begin{matrix}1 & 3\\ 2 & 1\end{matrix}$	$p_1 \left(1-p_2\right)\left(1-p_3\right)$
۵ = $\overline{\Lambda}_1\Lambda_2\Lambda_3$	$\begin{matrix}1 & 2\\ 2 & 1\end{matrix}$	$\begin{matrix}3 & 4\\ 3 & 1\end{matrix}$	$\begin{matrix}1 & 3\\ 1 & 2\end{matrix}$	$\left(1-p_1\right) p_2 p_3$
۶ = $\overline{\Lambda}_1\Lambda_2\overline{\Lambda}_3$	$\begin{matrix}1 & 2\\ 2 & 1\end{matrix}$	$\begin{matrix}3 & 4\\ 3 & 1\end{matrix}$	$\begin{matrix}1 & 3\\ 2 & 1\end{matrix}$	$\left(1-p_1\right) p_2 \left(1-p_3\right)$
۷ = $\overline{\Lambda}_1\overline{\Lambda}_2\Lambda_3$	$\begin{matrix}1 & 2\\ 2 & 1\end{matrix}$	$\begin{matrix}3 & 4\\ 1 & 3\end{matrix}$	$\begin{matrix}1 & 3\\ 1 & 2\end{matrix}$	$\left(1-p_1\right)\left(1-p_2\right) p_3$
۸ = $\overline{\Lambda}_1\overline{\Lambda}_2\overline{\Lambda}_3$	$\begin{matrix}1 & 2\\ 2 & 1\end{matrix}$	$\begin{matrix}3 & 4\\ 1 & 3\end{matrix}$	$\begin{matrix}1 & 3\\ 2 & 1\end{matrix}$	$\left(1-p_1\right)\left(1-p_2\right)\left(1-p_3\right)$

توجه: دو فاز پیوستگی ممکن برای هر فرد پایه‌گذار با Λ یا $\overline{\Lambda}$ نشان داده شده است که در آن اندیس‌ها شماره افراد مختلف هستند.

با معلوم بودن ژنوتیپ فرد ۹ و همچنین دیپلوتیپ‌های افراد ۸ و ۵، دیپلوتیپ فرد ۹ فقط بایستی به صورت زیر باشد:

$$\Lambda_9 = \begin{matrix} 1 \\ 2 \end{matrix} \left| \begin{matrix} 3 \\ 1 \end{matrix} \right|$$

که احتمال آن بدین صورت است:

(۸-۷)

$$\begin{aligned}
&P(\zeta_9 \mid \zeta_5, \zeta_8) = P(\Lambda_9 \mid \zeta_5, \zeta_8) \\
&= P(\Lambda_9 \mid \Lambda_5, \Lambda_8) P(\Lambda_5, \Lambda_8 \mid \zeta_5, \zeta_8) + P\left(\Lambda_9 \mid \overline{\Lambda}_5, \Lambda_8\right) P\left(\overline{\Lambda}_5, \Lambda_8 \mid \zeta_5, \zeta_8\right) \\
&+ P\left(\Lambda_9 \mid \Lambda_5, \overline{\Lambda}_8\right) P\left(\Lambda_5, \overline{\Lambda}_8 \mid \zeta_5, \zeta_8\right) + P\left(\Lambda_9 \mid \overline{\Lambda}_5, \overline{\Lambda}_8\right) P\left(\overline{\Lambda}_5, \overline{\Lambda}_8 \mid \zeta_5, \zeta_8\right) \\
&= P(\Lambda_9 \mid \Lambda_5, \Lambda_8) P(\Lambda_5 \mid \zeta_5) P(\Lambda_8 \mid \zeta_8) + P\left(\Lambda_9 \mid \overline{\Lambda}_5, \Lambda_8\right) P\left(\overline{\Lambda}_5 \mid \zeta_5\right) P(\Lambda_8 \mid \zeta_8) \\
&+ P\left(\Lambda_9 \mid \Lambda_5, \overline{\Lambda}_8\right) P(\Lambda_5 \mid \zeta_5) P\left(\overline{\Lambda}_5 \mid \zeta_5\right) + P\left(\Lambda_9 \mid \overline{\Lambda}_5, \overline{\Lambda}_8\right) P\left(\overline{\Lambda}_5 \mid \zeta_5\right) P\left(\overline{\Lambda}_8 \mid \zeta_8\right) \\
&= P(\Lambda_9 \mid \Lambda_5, \Lambda_8) P(\Lambda_5 \mid \zeta_5) P(\Lambda_8 \mid \zeta_8) + P\left(\Lambda_9 \mid \Lambda_5, \overline{\Lambda}_8\right) P(\Lambda_5 \mid \zeta_5) P\left(\overline{\Lambda}_8 \mid \zeta_8\right) \\
&= \frac{r^3(1-r) + r(1-r)^3 + (1-r)^4}{2\left[r^3 + (1-r)^3 + (1-r)^2\right]}
\end{aligned}$$

که در آن:

$$P(\Lambda_5 \mid \zeta_5) = 1$$

$$P(\overline{\Lambda}_5 \mid \zeta_5) = 0$$

$$P(\Lambda_9 \mid \Lambda_5, \Lambda_8) = \frac{1}{2}(1-r)$$

$$P(\Lambda_9 \mid \Lambda_5, \overline{\Lambda}_8) = \frac{1}{2} r$$

$$P(\overline{\Lambda}_8 \mid \zeta_8) = \frac{P(\Lambda_8 \mid \Lambda_3, \zeta_7)}{P(\zeta_8 \mid \Lambda_3, \zeta_7)}$$

$$= \frac{P(\Lambda_8 \mid \Lambda_3, \Lambda_7) P(\Lambda_7 \mid \zeta_7) + P(\Lambda_8 \mid \Lambda_3, \overline{\Lambda}_7) P(\overline{\Lambda}_7 \mid \zeta_7)}{P(\zeta_8 \mid \Lambda_3, \zeta_7)}$$

$$= \frac{r^3 + (1-r)^3}{r^3 + (1-r)^3 + (1-r)^2}$$

$$P(\Lambda_8 \mid \zeta_8) = 1 - P(\overline{\Lambda}_8 \mid \zeta_8)$$

برای فرد آخر، فقط یک فاز ممکن وجود دارد:

$$\Lambda_{10} = \begin{matrix} 4 \\ 3 \end{matrix} \left| \begin{matrix} 1 \\ 2 \end{matrix} \right|$$

و احتمال ژنوتیپی آن بدین صورت است:

$$P(\zeta_{10} \mid \Lambda_2, \zeta_9) = P(\zeta_{10} \mid \Lambda_2, \Lambda_9) = \frac{1}{4} r(1-r) \qquad (۹-۷)$$

با تشکیل احتمال هر یک از ۷ فرد غیر پایه‌گذار تحت ترکیب دیپلوتیپی $\Lambda_1\Lambda_2\Lambda_3$، می‌توانیم درستنمایی نرخ نوترکیبی ناشناخته را به شرط نشانگرها (M) با رابطهٔ زیر نشان دهیم:

(۱۰-۷)

$$L_{\Lambda_1\Lambda_2\Lambda_3}(r \mid M) = P(\zeta_4 \mid \Lambda_1, \Lambda_2) P(\zeta_5 \mid \Lambda_3, \zeta_4) P(\zeta_6 \mid \Lambda_1, \zeta_5) P(\zeta_7 \mid \zeta_6)$$
$$P(\zeta_8 \mid \Lambda_3, \zeta_7) P(\zeta_9 \mid \zeta_5, \zeta_8) P(\zeta_{10} \mid \Lambda_2, \zeta_9)$$
$$= \frac{1}{2^{10}} r^6 (1-r)^2 \left[r^3(1-r) + r(1-r)^3 + (1-r)^4 \right]$$

برای به‌دست آوردن MLE نرخ نوترکیبی، می‌توان از راهکار شبکه‌ای با در نظر گرفتن یک دامنه برای r استفاده نمود. بر اساس معادله (۷-۱۰)، سعی می‌کنیم یک پروفیل درستنمایی از

r = ۰ تا r = ۰/۵ را ترسیم کنیم (شکل ۷-۵). ولی نتیجه‌گیری می‌شود که برآورد بهینه r تحت اولین ترکیب دیپلوتیپی در بین سه فرد پایه‌گذار (جدول ۷-۵) برابر ۰/۷۱۵ است. روشن است که این برآورد برای داده‌های جدول ۷-۴ بهترین برآورد نیست، چون خارج از محدوده نرخ نوترکیبی است. همچنین این حالت نشان می‌دهد که نیاز به جستجوی برآورد r تحت ۷ ترکیب دیپلوتیپی دیگر داریم. باید به این نکته توجه نمود که چون فقط دو نشانگر نخست برای این تجزیه در نظر گرفته شده‌اند، نتایج به‌طور کامل برای جفت ترکیبات دیپلوتیپی ۱ و ۲، ۳ و ۴، ۵ و ۶ و ۷ و ۸ مشابه خواهند بود. به‌دلیل دیپلوتیپ‌های متفاوت نشانگر سوم که در نظر گرفته نشدند، هر جفت از این ترکیبات، متفاوت هستند. MLE نهایی r تحت ترکیب دیپلوتیپی ۳ یا ۴ برابر با ۰/۲۸۵ برآورد شد که به بزرگ‌ترین درستنمایی تعلق دارد.

الگوریتم EM

بر اساس درستنمایی (۷-۱۰)، می‌توانیم شکل دقیق‌تری برای الگوریتم EM به منظور برآورد نرخ نوترکیبی به‌دست آوریم. در این درستنمایی، عبارت آخر دارای ترکیبی از نوترکیب‌ها و غیر نوترکیب‌ها است که در آن، تعداد مورد انتظار نوترکیب‌ها باید برابر با مجموع دو احتمال زیر باشد:

$$\phi_1 = \frac{r^3(1-r)}{r^3(1-r)+r(1-r)^3+(1-r)^4} \qquad (۷-۱۱)$$

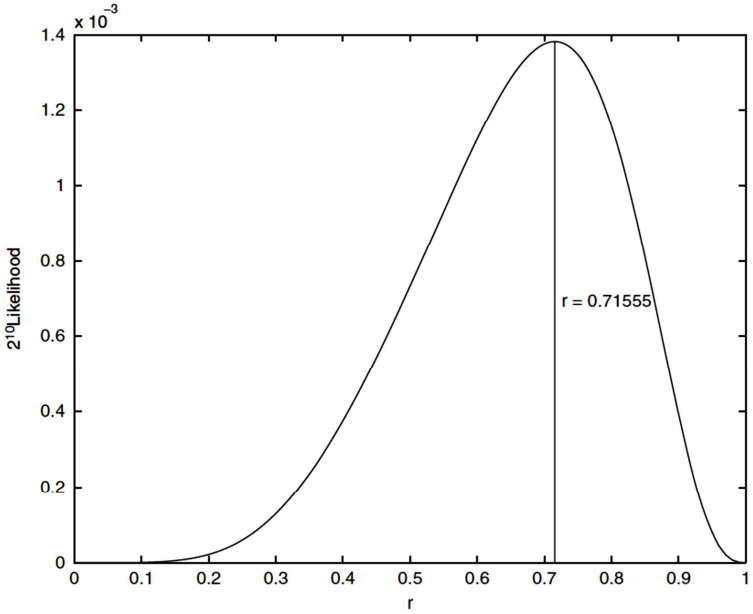
و:

$$\phi_2 = \frac{r(1-r)^2}{r^3(1-r)+r(1-r)^3+(1-r)^4} \qquad (۷-۱۲)$$

الگوریتم EM برای به‌دست آوردن MLE مربوط به r به‌کار گرفته می‌شود. در مرحله E، تعداد مورد انتظار پیشامدهای نوترکیبی از راه معادلات (۷-۱۱) و (۷-۱۲) محاسبه می‌شوند. در مرحله M، r با استفاده از رابطه زیر برآورد می‌شود:

$$\hat{r} = \frac{1}{12}\left(6+3\phi_1+\phi_2\right) \qquad (۷-۱۳)$$

معادلات (۷-۱۱) تا (۷-۱۳) تا جایی تکرار می‌شوند که برآورد ثابت و پایداری از r به‌دست آید. اگر مقدار اولیه r برابر با ۰/۵ در نظر گرفته شود، r پس از ده تکرار در مقدار ۰/۷۱۵ همگرا می‌شود که با برآورد حاصل از راهکار شبکه‌ای منطبق است. MLE بهینه r تحت ترکیب دیپلوتیپی ۳ (جدول ۷-۵) برابر با ۰/۲۸۵ است.



شکل ۷-۵- پروفیل درستنمایی محاسبه شده برای خانواده پیچیده نشان داده شده در شکل ۷-۴.

تجزیه همزمان نرخ نوترکیبی و احتمال دیپلوتیپی

درستنمایی توأم می‌تواند برای برآورد هم‌زمان نرخ نوترکیبی و دیپلوتیپ هر فرد پایه‌گذار فرموله شود. داده‌های نشانگری داده شده در جدول ۷-۴ را در نظر بگیرید. اگرچه دیپلوتیپ افراد پایه گذار ۱ تا ۳ ناشناخته است، ولی می‌توانیم احتمالی را برای هر یک از آن‌ها فرض کنیم که حامل دیپلوتیپ خاصی باشند. فرض کنیم p_1 احتمال آن باشد که فرد پایه‌گذار ۱ دارای دیپلوتیپ Λ_1 است و بنابراین احتمال دیپلوتیپ Λ_2 برابر $1-p_1$ است. به‌طور مشابه، p_2 و p_3 به ترتیب به‌عنوان احتمالات دیپلوتیپی افراد پایه‌گذار ۲ و ۳ تعریف می‌شوند؛ بنابراین، ترکیب دیپلوتیپی $\Lambda_1\Lambda_2\Lambda_3$ دارای احتمال $p_1p_2p_3$ است. احتمال دیپلوتیپی ترکیبات دیگر نیز می‌توانند تعریف شوند (جدول ۷-۵).

فرد ۴، فرزند افراد پایه‌گذار ۱ و ۲، دارای ژنوتیپ دو نشانگری ۱۳/۱۲ برای نشانگرهای A و B است. فرد ۴ دو دیپلوتیپ Λ_4 و $\overline{\Lambda}_4$ (۷-۲) دارد، حال آنکه افراد پایه‌گذار ۱ و ۲ ممکن است دیپلوتیپ‌هایی داشته باشند که در جدول ۷-۵ تعریف شده‌اند. احتمال ژنوتیپی فرد ۴ مشروط به والدینش یعنی افراد پایه‌گذار ۱ و ۲، بدین صورت بیان می‌شود:

$$
\begin{aligned}
&P(\zeta_4 \mid \zeta_1, \zeta_2) \\
&= P(\Lambda_4 \mid \zeta_1, \zeta_2) + P(\overline{\Lambda}_4 \mid \zeta_1, \zeta_2) \\
&= p_1 p_2 \left[P(\Lambda_4 \mid \Lambda_1, \Lambda_2) + P(\overline{\Lambda}_4 \mid \Lambda_1, \Lambda_2) \right] \\
&+ p_1 (1-p_2) \left[P(\Lambda_4 \mid \Lambda_1, \overline{\Lambda}_2) + P(\overline{\Lambda}_4 \mid \Lambda_1, \overline{\Lambda}_2) \right] \\
&+ (1-p_1) p_2 \left[P(\Lambda_4 \mid \overline{\Lambda}_1, \Lambda_2) + P(\overline{\Lambda}_4 \mid \overline{\Lambda}_1, \Lambda_2) \right] \\
&+ (1-p_1)(1-p_2) \left[P(\Lambda_4 \mid \overline{\Lambda}_1, \overline{\Lambda}_2) + P(\overline{\Lambda}_4 \mid \overline{\Lambda}_1, \overline{\Lambda}_2) \right]
\end{aligned}
\tag{۱۴-۷}
$$

که احتمالات دو دیپلوتیپ Λ_4 و $\overline{\Lambda}_4$ فرد ۴ که از روی هر یک از چهار ترکیب دیپلوتیپی ممکن افراد پایه‌گذار ۱ و ۲ به‌دست می‌آیند، در قالب جدول زیر بیان می‌شود:

دیپلوتیپ والدینی			فرزند ٤	
۱	۲	احتمال	Λ_4	$\overline{\Lambda}_4$
Λ_1	Λ_2	$p_1 p_2$	$P(\Lambda_4 \mid \Lambda_1, \Lambda_2) = ۰$	$P(\overline{\Lambda}_4 \mid \Lambda_1, \Lambda_2) = \frac{1}{4} r^2$
Λ_1	$\overline{\Lambda}_2$	$p_1 (1-p_2)$	$P(\Lambda_4 \mid \Lambda_1, \overline{\Lambda}_2) = ۰$	$P(\overline{\Lambda}_4 \mid \Lambda_1, \overline{\Lambda}_2) = \frac{1}{4} r (1-r)$
$\overline{\Lambda}_1$	Λ_2	$(1-p_1) p_2$	$P(\Lambda_4 \mid \overline{\Lambda}_1, \Lambda_2) = ۰$	$P(\overline{\Lambda}_4 \mid \overline{\Lambda}_1, \Lambda_2) = \frac{1}{4} r (1-r)$
$\overline{\Lambda}_1$	$\overline{\Lambda}_2$	$(1-p_1)(1-p_2)$	$P(\Lambda_4 \mid \overline{\Lambda}_1, \overline{\Lambda}_2) = ۰$	$P(\overline{\Lambda}_4 \mid \overline{\Lambda}_1, \overline{\Lambda}_2) = \frac{1}{4} (1-r)^2$

بر خلاف معادله (۷-۳)، احتمال ژنوتیپی فرد ۴ مشروط به ژنوتیپ‌های والدینش، وقتی که تمامی ترکیبات دیپلوتیپی در نظر گرفته می‌شوند، به‌صورت زیر به‌دست می‌آید:

$$
P(\zeta_4 \mid \zeta_1, \zeta_2) = \frac{1}{4} \left[p_1 p_2 r^2 + (p_1 + p_2 - 2 p_1 p_2) r (1-r) + (1-p_1)(1-p_2)(1-r)^2 \right]
$$

با استفاده از راهکار مشابه، احتمالات ژنوتیپی فرزند دیگر نیز به‌دست می‌آید. درنهایت، درستنمایی توأم به صورت زیر می‌شود:

$$
\begin{aligned}
L(r \mid M) = {} & P(\zeta_4 \mid \zeta_1, \zeta_2) P(\zeta_5 \mid \zeta_3, \zeta_4) P(\zeta_6 \mid \zeta_1, \zeta_5) P(\zeta_7 \mid \zeta_6) \\
& P(\zeta_8 \mid \zeta_3, \zeta_7) P(\zeta_9 \mid \zeta_5, \zeta_8) P(\zeta_{10} \mid \zeta_2, \zeta_9)
\end{aligned}
\tag{۱۵-۷}
$$

با بیشینه نمودن درستنمایی (۷-۱۵)، نرخ نوترکیبی r و احتمالات دیپلوتیپی سه فرد پایه‌گذار (p_1 ، p_2 و p_3) می‌توانند برآورد شوند.

تجزیه سه نقطه‌ای

ایده تجزیه پیوستگی پیشنهاد شده برای یک خانواده هسته پیچیده می‌تواند به تجزیه سه نقطه‌ای بسط یابد. جدول ۷-۴ را به یاد آورید که در آن، سه نشانگر A، B و C تعیین ژنوتیپ شده‌اند. فرض کنید g_{00}، g_{01}، g_{10} و g_{11} به ترتیب احتمال وجود هیچ نوترکیب برای هر دو فاصله نشانگری A-B و B-C، فقط یک نوترکیب برای فاصله B-C، فقط یک نوترکیب برای فاصله A-C و دو نوترکیب هر یک برای یکی از فواصل متفاوت باشند. ژنوتیپ سه نشانگری فرد ۴ به‌صورت ۱۳/۱۲/۱۲ است که ممکن است چهار دیپلوتیپ متفاوت داشته باشد که عبارت است از:

$$\Lambda_4^1 = \begin{matrix} 1 \\ 1 \\ 1 \end{matrix}\left|\begin{matrix} 3 \\ 2 \\ 2 \end{matrix}\right| \qquad \Lambda_4^2 = \begin{matrix} 1 \\ 1 \\ 2 \end{matrix}\left|\begin{matrix} 3 \\ 2 \\ 1 \end{matrix}\right| \qquad \Lambda_4^3 = \begin{matrix} 1 \\ 2 \\ 1 \end{matrix}\left|\begin{matrix} 3 \\ 1 \\ 2 \end{matrix}\right| \qquad \Lambda_4^4 = \begin{matrix} 1 \\ 2 \\ 2 \end{matrix}\left|\begin{matrix} 3 \\ 1 \\ 1 \end{matrix}\right|$$

چون فرد ۴ از افراد پایه‌گذار ۱ و ۲ مشتق شده است، دیپلوتیپ بالا باید به وسیله دیپلوتیپ‌های دو فرد پایه‌گذار تعیین شود. هر یک از افراد پایه‌گذار دارای ۴ دیپلوتیپ هستند که مجموعاً ۱۶ ترکیب دیپلوتیپی را تشکیل می‌دهند. فرض کنید که یکی از ترکیبات دیپلوتیپ والدینی بدین صورت باشد:

$$\Lambda_1 = \begin{matrix} 1 \\ 1 \\ 1 \end{matrix}\left|\begin{matrix} 2 \\ 2 \\ 2 \end{matrix}\right| \text{(برای فرد پایه‌گذار ۱)}$$

$$\Lambda_2 = \begin{matrix} 3 \\ 1 \\ 1 \end{matrix}\left|\begin{matrix} 4 \\ 3 \\ 2 \end{matrix}\right| \text{(برای فرد پایه‌گذار ۲)}$$

می‌توان مشاهده کرد که دیپلوتیپ‌های والدینی فوق نمی‌توانند دیپلوتیپ‌های Λ_4^1 و Λ_4^2 را تولید کنند، ولی امکان دارد که Λ_4^3 و Λ_4^4 را به ترتیب با احتمال‌های $g_{11}g_{01}$ و $g_{10}g_{00}$ تولید نمایند. این حالت می‌تواند با احتمال شرطی زیر بیان شود:

$$P\left(\zeta_4 \mid \Lambda_1, \Lambda_2\right) = P\left(\Lambda_4^3 \mid \Lambda_1, \Lambda_2\right) + P\left(\Lambda_4^4 \mid \Lambda_1, \Lambda_2\right)$$
$$= g_{11}g_{01} + g_{10}g_{00}$$

روابط مشابهی برای نتاج ۵ تا ۱۰ نیز می‌تواند به‌دست آید. از تمامی این روابط برای ایجاد درستنمایی استفاده می‌شود که از روی آن، پارامترهای g_{00}، g_{01}، g_{10} و g_{11} می‌توانند برآورد شوند. سپس، نرخ‌های نوترکیبی بین سه نشانگر یعنی r_{AB}، r_{BC} و r_{AC} بر اساس ترکیب دیپلوتیپ فرد پایه‌گذار مورد نظر به‌صورت زیر برآورد می‌شوند:

$$\hat{r}_{AB} = \hat{g}_{10} + \hat{g}_{11}$$
$$\hat{r}_{BC} = \hat{g}_{01} + \hat{g}_{11}$$
$$\hat{r}_{AC} = \hat{g}_{01} + \hat{g}_{10}$$

همانند تجزیه دو نقطه‌ای، احتمالات دیپلوتیپی سه فرد پایه‌گذار می‌توانند در تابع درستنمایی وارد شده و به‌طور همزمان با نرخ‌های نوترکیبی برآورد شوند. مزایای تجزیه سه نقطه‌ای در هنگام مقایسه با تجزیه دو نقطه‌ای، برآورد دقیق نرخ‌های نوترکیبی و توانایی بیشتر تشخیص پیوستگی معنی‌دار است. اشکال آن نیز از این نظر است که به منابع محاسباتی بیشتری نیاز دارد.

۷-۳- تجزیه اطلاعات نشانگرهای غالب

۷-۳-۱- مقدمه

در بسیاری از شرایط، نشانگرهای هم‌بارز به‌دلیل محتوای اطلاعاتی بیشتری که ارایه می‌دهند، نسبت به نشانگرهای غالب ترجیح داده می‌شوند. یک نشانگر مولکولی هم‌بارز اجازه تمایز و تشخیص ژنوتیپ‌های هموزیگوت را از ژنوتیپ‌های هتروزیگوت روی ژل الکتروفورز می‌دهد. در مقابل، در نشانگرهای غالب، افراد هموزیگوت غالب و هتروزیگوت بر اساس حضور یا عدم حضور باندها روی ژل قابل شناسایی نیستند؛ اما نشانگرهای غالب، نظیر AFLP، به‌دلیل هزینه پایین و خصوصیات ساده مولکولی آن‌ها بسیار سودمند و مفید هستند. به این دلیل، امکان استخراج اطلاعات هم‌بارز از روی نشانگرهای غالبی که دارای خصوصیات ساده هستند، شایستگی انجام تحقیق و پژوهش را دارد.

۷-۳-۲- تجزیه تفرق

مدل‌های مختلط

فرض کنید یک سیستم نشانگر غالب مبتنی بر PCR برای امتیازدهی ژنوتیپ‌های یک جمعیت در جایگاه ژنی یک نشانگر استفاده می‌شود. به جای امتیازدهی حضور یا عدم حضور باند، از تکنیک بازتابش[1] برای اندازه‌گیری شدت درجه تکثیر یک قطعه از باند روی ژل الکتروفورز استفاده می‌کنیم؛ بنابراین، انتظار می‌رود که شدت باندها برای افراد هموزیگوت غالب بیش از افراد هتروزیگوت باشد، چون برای افراد هتروزیگوت، مقدار فرآورده‌های PCR فقط به همان مقدار هموزیگوت‌ها است؛ اما چون تمایز بین شدت باند در میان ژنوتیپ‌ها نامشخص است، شدت باندهایی که اندازه‌گیری می‌شوند، در بین افراد حاصل از ژنوتیپ‌های مشابه متغیر است. این حالت

[1] - Fluorescence

اساساً به‌واسطه تنوع تصادفی است که طی سنجش نشانگر رخ می‌دهد. یک مدل مختلط برای انتساب هر فرد به یکی از کلاس‌های ژنوتیپی و برآورد الگوهای تفرق ژنوتیپ‌ها در نتاج ایجاد شده است (پیفو و کخ[1]، ۲۰۰۰؛ یانسن و همکاران[2]، ۲۰۰۱).

خانواده‌ای را در نظر بگیرید که در آن، یک نشانگر غالب در حال تفرق است. شدت باند (y_{ij}) برای فرد i ام در کلاس ژنوتیپی j ام ($j=1,\ldots,k$) می‌تواند با استفاده از مدل خطی زیر بیان شود:

$$y_{ij}=\mu_j+e_{ij} \qquad (۱۶-۷)$$

که μ_j میانگین مورد انتظار شدت باند برای ژنوتیپ j و e_{ij} خطای باقیمانده است که فرض می‌شود دارای توزیع نرمال با میانگین صفر و واریانس σ^2 است.

برای در نظر گرفتن مجدد تنوع ژنوتیپی، فرض می‌کنیم که هر مشاهده y_i از مدل مختلط زیر پیروی می‌کند:

$$y_i \sim \sum_{j=1}^{k}\gamma_j\left(y_i \mid \mu_j,\sigma^2\right)$$

که $\gamma_1,\ldots,\gamma_k$ نسبت‌های اختلاط ناشناخته (پیشین) هستند به گونه‌ای که $0\le\gamma_i\le 1$، $\sum\gamma_j=1$ و $f\left(.\mid\mu_j,\sigma^2\right)$ چگالی نرمال با میانگین μ_j و واریانس σ^2 است.

از روی نمونه $y=\left(y_1,\ldots,y_n\right)$، تابع درستنمایی بدین صورت ارایه می‌شود:

$$L\left(\Omega \mid y_{ij}\right)=\prod_{i=1}^{n}\sum_{j=1}^{k}\gamma_j f_j\left(y_i \mid \mu_j,\sigma^2\right) \qquad (۱۷-۷)$$

که در آن، $\Omega=\left(\mu_j,\sigma^2,\gamma_j\right)^T$ بوده و لگاریتم درستنمایی عبارت است از:

$$\log L\left(\Omega\right)=\sum_{i=1}^{n}\log\left[\sum_{j=1}^{k}\gamma_j f_j\left(y_i \mid \mu_j,\sigma^2\right)\right] \qquad (۱۸-۷)$$

---

[1] - Piepho and Koch
[2] - Jansen *et al.*

با مشتق‌گیری تابع بالا نسبت به یک Ω ناشناخته (Ω_l)، داریم:

$$\frac{\partial}{\partial \Omega_l}\log L(\Omega)=\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\gamma_j \frac{\partial}{\partial \Omega_l} f_j\left(y_i \mid \mu_j,\sigma^2\right)}{\sum_{j=1}^{k}\gamma_j f_j\left(y_i \mid \mu_j,\sigma^2\right)}$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\gamma_j \frac{\partial}{\partial \Omega_l} f_j\left(y_i \mid \mu_j,\sigma^2\right)}{\sum_{j=1}^{k}\gamma_j f_j\left(y_i \mid \mu_j,\sigma^2\right)}\frac{\partial}{\partial \Omega_l}\log f_j\left(y_i \mid \mu_j,\sigma^2\right) \qquad (۱۹-۷)$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{k}\Gamma_{ij}\frac{\partial}{\partial \Omega_l}\log f_j\left(y_i \mid \mu_j,\sigma^2\right)$$

و داریم:

$$\Gamma_{ij}=\frac{\gamma_j f_j\left(y_i \mid \mu_j,\sigma^2\right)}{\sum_{j=1}^{k}\gamma_j f_j\left(y_i \mid \mu_j,\sigma^2\right)} \qquad (۲۰-۷)$$

که می‌تواند به‌عنوان احتمال پسین ژنوتیپ نشانگر j برای فرزند i در نظر گرفته شود. سپس، معادلات (۷-۱۹) و (۷-۲۰) با مجموعه پارامترهای بسط یافته $\{\Omega,\Gamma\}$ تکرار می‌شوند که $\Gamma=\left\{\Gamma_{ij},\, j=1,\ldots,k; i=1,\ldots,n\right\}$ است. مشروط به Γ، برای مقادیر صفر $\frac{\partial}{\partial \Omega_l}\log L(\Omega)$ حل معادله را انجام می‌دهیم تا به برآوردهای Ω دست یابیم:

$$\gamma_j=\frac{1}{n}\sum_{i=1}^{n}\Gamma_{ij}$$

$$\mu_j=\frac{\sum_{i=1}^{n}\Gamma_{ij}y_j}{\sum_{i=1}^{n}\Gamma_{ij}}$$

$$\sigma^2=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\Gamma_{ij}\left(y_j-\mu_j\right)^2$$

سپس برآوردها برای نوسازی و تجدید مقادیر Γ استفاده می‌شوند و این فرآیند تا جایی تکرار می‌شود که به همگرایی برسد. مقادیر در هنگام همگرایی همان MLE ها هستند. احتمال پسین ژنوتیپ نشانگر یک فرد، مشروط به ارزش فنوتیپی آن یعنی y_i، می‌تواند با جایگزینی پارامترها با

MLE های مربوطه ارزیابی شود. بر اساس احتمال‌های پسین برآوردهای مربوطه، افراد ممکن است به یکی از ژنوتیپ های k منتسب شوند.

برای خانواده در حال تفرق F_2، نسبت‌های اختلاط بر اساس دلایل قبلی شناخته شده‌اند، یعنی برای AA، $\gamma_1 = \frac{1}{4}$، برای Aa، $\gamma_2 = \frac{1}{2}$ و برای aa، $\gamma_3 = \frac{1}{4}$ است. در این حالت، مدل کامل برای برآورد تمامی شش پارامتر آزاد $\Omega_F = \left(\mu_j, \sigma^2, \gamma_j\right)^T$ به یک مدل کاهش یافته تبدیل می‌شود که در آن، فقط چهار پارامتر آزاد $\Omega_R = \left(\mu_j, \sigma^2\right)^T$ برآورد می‌شوند؛ بنابراین، برای آزمودن انحراف معنی‌دار از نسبت تفرق ۱:۲:۱، آماره آزمون نسبت درستنمایی به صورت زیر برآورد می‌شود:

$$LR = -2\left[\log L\left(\Omega_R\right) - \log L\left(\Omega_F\right)\right]$$

که از لحاظ مجانبی دارای توزیع کای مربع با سه درجه آزادی است (تفاوت تعداد پارامترهای آزاد بین Ω_R و Ω_F). اگر آزمون معنی‌دار نشود، نتیجه می‌گیریم که نشانگر غالب از نسبت ۱:۲:۱ تبعیت می‌کند. انحراف از نسبت ۱:۲:۱ در نسل F_2 حاصل از دو لاین اینبرد، ممکن است به‌واسطه عوامل زیادی نظیر وجود بیش از یک جایگاه ژنی، تفاوت در ماندگاری و بقاء افراد و غیره باشد.

رابطهٔ خطی بین تعداد کپی یک آلل غالب و مقادیر کمّی آن‌ها نیز می‌تواند آزمون شود. مدل کاهش یافته برای این آزمون به مقادیر میانگین (μ_j) سه ژنوتیپ نشانگر به صورت ۰ (aa): ۱ (Aa): ۲ (AA) مرتبط است، در حالی‌که مدل کامل برای برآورد این مقادیر میانگین، فرموله شده است.

نرخ تخصیص[1] صحیح

استفاده از احتمال پسین برآورد شده برای تعیین کلاس ژنوتیپی یک فرد عاری از خطا نیست. در این بخش، توضیح می‌دهیم که چگونه نرخ تخصیص صحیح (CAR)، یعنی احتمال این که یک فرد انتخاب شده به‌صورت تصادفی به‌طور صحیح دسته‌بندی شود، محاسبه می‌شود (پیفو و کخ[2]، ۲۰۰۰). فرض کنید حدود دسته‌بندی y_L و y_R هستند، به‌طوری که افراد به صورت زیر دسته‌بندی می‌شوند:

---

[1] - Allocation rate
[2] - Piepho and Koch

شرایط برای مقدار y	دسته‌بندی
$y < y_L$	ژنوتیپ = AA
$y_L < y < y_R$	ژنوتیپ = Aa
$y_R < y$	ژنوتیپ = aa

حد دسته‌بندی y_L نقطه تقاطع بین $\gamma_1 f_1(y_i \mid \mu_1, \sigma)$ و $\gamma_2 f_2(y_i \mid \mu_2, \sigma)$ برای $\mu_1 < y_L < \mu_2$ است، یعنی نقطه‌ای که در آن، برای $\mu_1 < y_L < \mu_2$، $\gamma_2 f_2(y_i \mid \mu_2, \sigma) = \gamma_1 f_1(y_i \mid \mu_1, \sigma)$ است. حد دسته‌بندی y_R نقطه تقاطع بین $\gamma_2 f_2(y_i \mid \mu_2, \sigma_2)$ و $\gamma_3 f_3(y_i \mid \mu_3, \sigma)$ برای $\mu_2 < y_R < \mu_3$ است. چون واریانس باقیمانده به‌صورت همگن فرض می‌شود:

$$y_L = \frac{\sigma^2 \log\left(\frac{\gamma_1}{\gamma_2}\right)}{\mu_2 - \mu_1} + \frac{\mu_2 + \mu_1}{2}$$

$$y_R = \frac{\sigma^2 \log\left(\frac{\gamma_2}{\gamma_3}\right)}{\mu_3 - \mu_2} + \frac{\mu_3 + \mu_2}{2}$$

اگر $F(.)$ نشان‌دهنده تابع توزیع تجمعی نرمال استاندارد و CAR_j نشان‌دهنده نرخ تخصیص صحیح برای j امین مؤلفه باشد، خواهیم داشت:

$$CAR_1 = F\left(\frac{\mu_L - \mu_1}{\sigma}\right)$$

$$CAR_2 = F\left(\frac{y_R - y_2}{\sigma}\right) - F\left(\frac{y_L - y_2}{\sigma}\right)$$

$$CAR_3 = 1 - F\left(\frac{y_R - y_3}{\sigma}\right)$$

نرخ دسته‌بندی صحیح کل نیز عبارت است از:

$$CAR = \sum_{j=1}^{k} \gamma_j CAR_j$$

CAR را از راه برآوردهای نمونه‌ای برای پارامترها برآورد نمودیم. این روش، ارزیابی خام و اولیه‌ای برای CAR حقیقی فراهم می‌کند.

تبدیل داده‌ها

مدل مختلطی که در بخش بالا برای دسته‌بندی ژنوتیپ‌های نشانگر در یک نشانگر غالب استنتاج شد، بر اساس فرض نرمال بودن داده‌ها انجام شد. در عمل، این فرض ممکن است همیشه صادق نباشد و داده‌های شدت باند ممکن است از راه تبدیل داده‌ها به حالت نرمال نزدیک‌تر شوند.

تبدیل‌های لگاریتمی و ریشه دوم، انواع متداول تبدیل داده‌ها هستند که در بیشتر موارد عملکرد آن‌ها مشابه است. اگر داده‌ها دارای چولگی باشند، روش‌های تبدیل داده به گونه‌ای عمل می‌کنند که داده‌ها را به حالت متقارن نزدیک نمایند و واریانس‌های همگن‌تری ایجاد کنند. به‌طور خاص، تبدیل لگاریتمی زمانی مناسب است که میانگین نسبتی از انحراف معیار بوده و تبدیل ریشه دوم زمانی مناسب است که میانگین نسبتی از واریانس باشد.

گوتیرز و همکاران (۱۹۹۵) کاربرد مدل‌های مختلط نرمال را برای دسته کلی‌تری از تبدیل داده‌ها یعنی تبدیل‌های نمایی باکس و کاکس[1] (۱۹۶۴) بحث نمودند. این دسته از تبدیل‌ها بسیار انعطاف پذیر بوده و در حالت خاص، شامل تبدیل لگاریتمی هستند که به وسیله تابع زیر نشان داده می‌شوند:

$$f(y,\lambda)=y^{\lambda}=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda} & \text{اگر } \lambda\neq 0\\ \log(y) & \text{اگر } \lambda=0\end{cases}$$

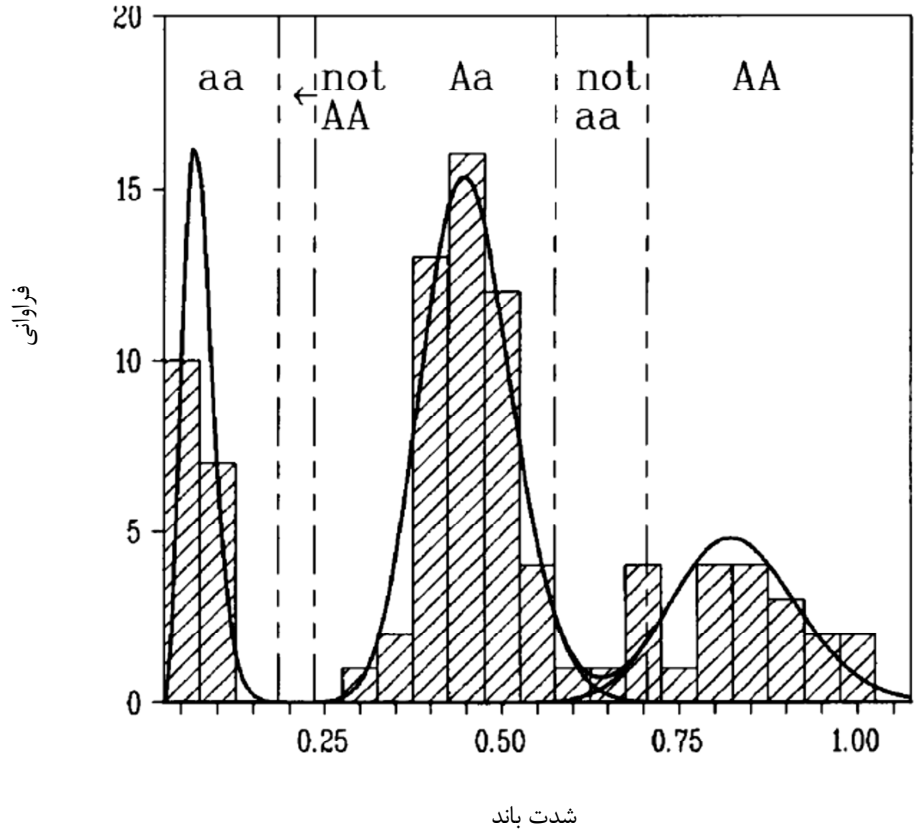
توجه کنید که ۱= λ معادل عدم انجام تبدیل داده‌ها است. ساده‌ترین راه برای به‌دست آوردن MLE مربوط به λ به وسیله جستجوی شبکه‌ای است (گوتیرز و همکاران، ۱۹۹۵). روش‌های ممکن دیگر برای تبدیل داده‌ها شامل گرفتن ریشه دوم است که می‌تواند ناهمگنی واریانس باقیمانده را به‌واسطه اثر مقیاس کاهش دهد (یانسن و همکاران، ۲۰۰۱). زمانی که یک روش تبدیل داده مناسب انتخاب شود، ممکن است مدل مختلط نرمال برای تبدیل داده‌ها برازش یابد.

***مثال ۷-۱*. (نتاج F_2 در گوجه فرنگی).** یانسن و همکاران (۲۰۰۱) مثالی از روش تجزیه هم‌بارز یک نشانگر غالب AFLP را در ۸۸ گیاه گوجه فرنگی حاصل از تفرق نتاج F_2 گزارش کردند. هیستوگرام شدت باند این نتاج در شکل ۷-۶ نشان داده شده است که در آن، یک توزیع سه نمایی[2] می‌تواند مشاهده شود. به نظر می‌رسد که شدت باند دارای ارتباط خطی با تعداد کپی است

[1] - Box and Cox
[2] - Trimodal distribution

و واریانس به‌طور خطی با افزایش میانگین افزایش می‌یابد. تبدیل ریشه دوم برای رفع ناهمگنی واریانس ژنوتیپ‌های متفاوت استفاده می‌شود.

مدل مختلط نرمال برای برازش داده‌های تبدیل شده به روش ریشه دوم استفاده می‌شود. داده‌ها می‌توانند به خوبی با سه توزیع مختلط، همان‌گونه که در هیستوگرام‌ها رسم شده است، برازش یابند (شکل ۷-۶). مقداری هم‌پوشانی بین توزیع‌های Aa و AA وجود دارد. افرادی که شدت باند آن‌ها در مناطق هم‌پوشاننده قرار دارد، برای دسته‌بندی توصیه نمی‌شوند. CAR برای دسته‌بندی صحیح ژنوتیپ نشانگر بیش از ۹۸ درصد است.



شکل ۷-۶- توزیع‌های مختلط شدت باند در ۸۷ گیاه F_2 گوجه فرنگی. توزیع‌های سه ژنوتیپ aa، Aa و AA شناسایی شده‌اند، ولی مقداری هم‌پوشانی بین توزیع‌های aa و Aa (نه AA) و بین توزیع‌های Aa و AA (نه aa) وجود دارد. اقتباس از یانسن و همکاران (۲۰۰۱).

۷-۳-۳- تجزیهٔ پیوستگی

نشانگرهای غالب مثل AFLP ها و RAPD ها به‌طور معمول بر اساس حضور یا عدم حضور باند در ژل الکتروفورز تجزیه می‌شوند. این نوع تجزیه، اجازه تمایز بین هموزیگوت‌های غالب و هتروزیگوت‌ها را از یکدیگر نمی‌دهند. چنین تمایزی می‌تواند در شرایطی صورت گیرد که شدت باندها به‌صورت کمّی اندازه‌گیری شود. در بخش ۷-۳-۲، یک مدل مختلط آماری از راه الگوریتم EM بکار گرفته شد (پیفو و کخ، ۲۰۰۰) که برای شناسایی و تشخیص هموزیگوت‌های غالب و هتروزیگوت‌ها برای نشانگرهای منفرد بر اساس شدت کمّی باندها توضیح داده شد. این مدل می‌تواند از نشانگرهای غالب، اطلاعات هم‌بارز بیشتری برای تجزیه پیوستگی وقتی که شدت باندها کمّی می‌شود، استخراج نماید.

فرض کنید n گیاه F_2 حاصل از یک هتروزیگوت مضاعف F_1 (AaBb) مشتق از دو لاین اینبرد وجود دارد. دو نشانگر A و B در مجموع ۹ ژنوتیپ را در جمعیت F_2 به صورت AABB، AABb، AAbb، AaBB، AaBb، Aabb، aaBB، aaBb و aabb تولید می‌کنند که فراوانی‌های مربوطه به صورت تابعی از نرخ نوترکیبی (r) بین دو نشانگر است که می‌تواند به شکل ماتریسی نشان داده شود. برای نشانگرهای غالب، AA و Aa و همچنین BB و Bb نمی‌توانند تشخیص داده شوند، چون هر دوی آن‌ها یک باند روی ژل ایجاد می‌کنند؛ اما اگر شدت باندهای این دو نشانگر اندازه‌گیری شوند، می‌توانیم انتظار داشته باشیم که AA یا BB دارای ارزش‌های بیشتری نسبت به Aa یا Bb باشند. در هر حال، وجود چنین تفاوتی ممکن است به‌واسطه خطاهای تصادفی که می‌توانند فرض شوند از توزیع نرمال تبعیت می‌کنند، پوشیده شوند و نامعلوم بمانند.

مقدار کمّی شدت باند در j امین ژنوتیپ نشانگر A برای i امین گیاه F_2 شامل ۲ مؤلفه است که یکی ارزش میانگین μ_{1j} و دیگری واریانس $\sigma^2_{1j_1}$ است. توزیع حاشیه‌ای $y_i = \left(y_{1i} \quad y_{2i}\right)$ برای دو نشانگر A و B به‌صورت مخلوطی از ۹ توزیع نرمال دو متغیره است:

$$f\left(y_i \mid \Omega\right) = \sum_{j_0=1}^{2}\sum_{j_2=0}^{2} \gamma_{j_1 j_2} f\left(y_{1i} \mid \mu_{1j_1}, \sigma^2_{1j_1}\right) f\left(y_{2i} \mid \mu_{2j_2}, \sigma^2_{2j_2}\right) \qquad (۲۱-۷)$$

که $\Omega = \left(\gamma_{j_1 j_2}, \mu_{1j_1}, \sigma^2_{1j_1}, \mu_{2j_2}, \sigma^2_{2j_2}\right)$ پارامترهای ناشناخته‌ای هستند که برآورد می‌شوند، $j_1, j_2 = 0$ برای aa یا bb، ۱ برای Aa یا Bb و ۲ برای AA یا BB و $\gamma_{j_1 j_2}$ فراوانی ژنوتیپی j_1 امین ژنوتیپ در نشانگر A و j_2 امین ژنوتیپ در نشانگر B است. با فرض اینکه واریانس در سطوح مختلف یکنواخت است، مدل مربوطه می‌تواند ساده شود، برای مثال:

$$\sigma^2_{1j_1} = \sigma^2_1, \sigma^2_{2j_2} = \sigma^2_2 \qquad (۲۲-۷)$$

یا:

$$\sigma_{1j_1}^2 = \sigma_{2j_2}^2 = \sigma^2 \qquad (۲۳-۷)$$

در حالی‌که مدل (۷-۲۳) از نظر محاسباتی ساده‌تر است، مدل (۷-۲۲) ممکن است به واقعیت نزدیک‌تر باشد، چون نوع و کمیت فرآورده اندازه‌گیری شده در یک موقعیت باندی در بین نشانگرها متغیر است.

اگر $z_{j_1j_2i}$ به عنوان یک متغیر تصادفی با ۱ = $z_{j_1j_2i}$ برای ژنوتیپ مشاهده شده نشانگر و ۰ = $z_{j_1j_2i}$ برای سایر موارد در نظر گرفته شود، $z_i = \left(z_{00i},...,z_{22i}\right)$ از توزیع چند متغیره‌ای با عدد ثابت ۱ و احتمالات سلولی $\gamma_{j_1j_2}$ تبعیت می‌کند. در حالیکه y_i مشاهده می‌شود، z_i مشاهده نمی‌شود (گمشده است). هم y_i و هم z_i داده‌های کامل را تشکیل می‌دهند؛ بنابراین، الگوریتم EM می‌تواند برای فراهم کردن برآوردی از Ω مورد استفاده قرار گیرد (مک لاکلان و کریشنان[1]، ۱۹۹۷). درستنمایی داده‌های کامل عبارت است از:

$$\log L(\Omega) = \sum_{i=1}^{n}\sum_{j_1=0}^{2}\sum_{j_2=0}^{2} z_{j_1j_2i}\log\left[\gamma_{j_1j_2} f\left(y_{1i}\mid\mu_{1j_1},\sigma_{1j_1}^2\right) f\left(y_{2i}\mid\mu_{2j_2},\sigma_{2j_2}^2\right)\right]$$

امید ریاضی مورد انتظار لگاریتم درستنمایی داده‌های کامل مشروط به داده‌های مشاهده شده y_i، با استفاده از برآوردهای فعلی پارامتر $\Omega^{(t)}$ ممکن است بدین صورت بیان شود:

$$Q\left(\Omega;\Omega^{(t)}\right) = E_{\Omega^{(t)}}\left[\log L\left(\Omega^{(t)}\right)\mid y_i\right] \qquad (۲۴-۷)$$

چون $Q\left(\Omega;\Omega^{(t)}\right)$ در $z_{j_1j_2i}$ خطی است، $Q\left(\Omega;\Omega^{(t)}\right)$ از راه جایگزینی $z_{j_1j_2i}$ با امید ریاضی مربوطه مشروط به y_i، در لگاریتم درستنمایی داده‌های کامل که در $Q\left(\Omega;\Omega^{(t)}\right)$ ارزیابی شده‌اند، محاسبه می‌شود، یعنی $z_{j_1j_2i}$ با عبارت زیر جایگزین می‌شود:

$$\Gamma_{j_1j_2i}^{(t)} = \frac{\gamma_{j_1j_2} f\left(y_{1i}\mid\mu_{1j_1}^{(t)},\sigma_{1j_1}^{2(t)}\right) f\left(y_{2i}\mid\mu_{2j_2}^{(t)},\sigma_{2j_2}^{2(t)}\right)}{\sum_{j_1=0}^{2}\sum_{j_2=0}^{2}\gamma_{j_1j_2} f\left(y_{1i}\mid\mu_{1j_1}^{(t)},\sigma_{1j_1}^{2(t)}\right) f\left(y_{2i}\mid\mu_{2j_2}^{(t)},\sigma_{2j_2}^{2(t)}\right)} \qquad (۲۵-۷)$$

در مرحله E، $\Gamma_{j_1j_2i}^{(t)}$ با استفاده از برآوردهای فعلی پارامتر، نوسازی و تجدید می‌شود و در مرحله M، $Q\left(\Omega;\Omega^{(t)}\right)$ به‌طور مرتبط با پارامترها بیشینه می‌شود. معادلات برآوردکننده میانگین‌ها و واریانس‌ها جواب‌های مشخصی دارند، در حالیکه معادله مربوط به r یک معادله

[1] - McLachnan and Krishnan

چندجمله‌ای درجه سه است که ممکن است به‌صورت عددی و یا با فرمول‌های مشخص حل شود. معادلات برآورد میانگین‌ها و واریانس‌ها در مرحله M به صورت زیر هستند:

$$\mu_{1j_1}^{(t)} = \frac{\sum_{i=1}^{n}\sum_{j_2=0}^{2}\Gamma_{j_1j_2i}^{(t)}y_{1i}}{\sum_{i=1}^{n}\sum_{j_2=0}^{2}\Gamma_{j_1j_2i}^{(t)}}$$

$$\mu_{2j_2}^{(t)} = \frac{\sum_{i=1}^{n}\sum_{j_1=0}^{2}\Gamma_{j_1j_2i}^{(t)}y_{2i}}{\sum_{i=1}^{n}\sum_{j_1=0}^{2}\Gamma_{j_1j_2i}^{(t)}} \qquad (۲۶-۷)$$

$$\sigma_1^{2(t)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j_1=0}^{2}\sum_{j_2=0}^{2}\Gamma_{j_1j_2i}^{(t)}\left(y_{1i}-\mu_{1j_1}^{(t)}\right)^2$$

$$\sigma_2^{2(t)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j_1=0}^{2}\sum_{j_2=0}^{2}\Gamma_{j_1j_2i}^{(t)}\left(y_{2i}-\mu_{2j_2}^{(t)}\right)^2$$

$r^{(t)}$ به وسیله ریشه غیرپیچیده زیر نوسازی و تجدید می‌شود:

$$\frac{\partial Q\left(\Omega;\Omega^{(t)}\right)}{\partial r} = \sum_{i=1}^{n}\sum_{j_1=0}^{2}\sum_{j_2=0}^{2}\Gamma_{j_1j_2i}^{(t)}S_{j_1j_2} = 0$$

که در آن:

$$S_{j_1j_2} = \frac{1}{\Gamma_{j_1j_2}}\frac{\partial\Gamma_{j_1j_2}}{\partial r}$$

که $Q\left(\Omega;\Omega^{(t)}\right)$ را به حداکثر می‌رساند. اگر جواب این معادله بزرگ‌تر از ۰/۵ باشد، $r^{(t+1)}$ = ۰/۵ و اگر جواب کوچک‌تر از صفر باشد، $r^{(t+1)}$ = ۰ خواهد بود.

پیفو (۲۰۰۱) مطالعه شبیه‌سازی برای مقایسه برآوردگر نرخ نوترکیبی از راه در نظر گرفتن نشانگرهای غالب به‌صورت هم‌بارز کمّی و از راه حضور یا عدم حضور باند انجام داد. مشخص شد که روش کمّی برآوردهای دقیق‌تری از r نسبت به روش کیفی، برای مقادیر متفاوت r، میانگین‌ها و واریانس‌های متفاوت و اندازه‌های نمونه متفاوت نشان می‌دهد.

روش EM که در بخش بالا توصیف شد، می‌تواند برای برآورد هم‌زمان هم‌بارزی نشانگر و پیوستگی بین سه نشانگر غالب A، B و C که شدت باندهای آن‌ها تعیین شده است، استفاده شود. در نظر گرفتن همزمان این سه نشانگر ممکن است سبب افزایش دقت تجزیه پیوستگی شده و اطلاعاتی در خصوص ترتیب نشانگرها فراهم آورد. با فرض این که سه نشانگر دارای ترتیب A-B-C باشند، فراوانی‌های ۲۷ ژنوتیپ نشانگری مختلف AABBCC، AABBCc، ...، aabbcc

در جمعیت F_2 می‌تواند به صورت تابعی از مقادیر نوترکیبی r_{AB}، r_{AC} و r_{BC} بین سه نشانگر بیان شود، یعنی:

(۲۷-۷)

$$H_{ABC}=\begin{array}{c} \\ BBCC \\ BBCc \\ BBcc \\ BbCC \\ BbCc \\ Bbcc \\ bbCC \\ bbCc \\ bbcc \end{array}\begin{array}{c} \begin{array}{ccc} AA & Aa & aa \end{array} \\ \begin{bmatrix} \frac{1}{4}(1-r_{AB})^2(1-r_{BC})^2 & \frac{1}{2}r_{AB}(1-r_{AB})r_{BC}(1-r_{BC}) & \frac{1}{4}r_{AB}^2r_{BC}^2 \\ \frac{1}{2}r_{AB}(1-r_{AB})^2r_{BC}(1-r_{BC}) & \frac{1}{2}r_{AB}(1-r_{AB})\phi_{BC} & \frac{1}{4}r_{AB}^2r_{BC}(1-r_{BC}) \\ \frac{1}{4}(1-r_{AB})^2r_{BC}^2 & \frac{1}{2}r_{AB}(1-r_{AB})r_{BC}(1-r_{BC}) & \frac{1}{4}r_{AB}^2(1-r_{BC})^2 \\ \frac{1}{2}r_{AB}(1-r_{AB})(1-r_{BC})^2 & \frac{1}{2}\phi_{AB}r_{BC}(1-r_{BC}) & \frac{1}{4}r_{AB}(1-r_{AB})r_{BC}^2 \\ r_{AB}(1-r_{AB})r_{BC}(1-r_{BC}) & \frac{1}{2}\phi_{AB}\phi_{BC} & r_{AB}(1-r_{AB})r_{BC}(1-r_{BC}) \\ \frac{1}{2}r_{AB}(1-r_{AB})r_{BC}^2 & \frac{1}{2}\phi_{AB}r_{BC}(1-r_{BC}) & \frac{1}{2}(r_{AB})(1-r_{AB})(1-r_{BC})^2 \\ \frac{1}{4}r_{AB}^2(1-r_{BC})^2 & \frac{1}{2}r_{AB}(1-r_{AB})r_{BC}(1-r_{BC}) & \frac{1}{4}(1-r_{AB})^2r_{BC}^2 \\ \frac{1}{2}r_{AB}^2r_{BC}(1-r_{BC}) & \frac{1}{2}r_{AB}(1-r_{AB})\phi_{BC} & \frac{1}{2}(1-r_{AB})^2r_{BC}(1-r_{BC}) \\ \frac{1}{4}r_{AB}^2r_{BC}^2 & \frac{1}{2}r_{AB}(1-r_{AB})r_{BC}(1-r_{BC}) & \frac{1}{4}(1-r_{AB})^2(1-r_{BC})^2 \end{bmatrix} \end{array}$$

که $\phi_{AB}=1-2r_{AB}+2r_{AB}^2$ و $\phi_{BC}=1-2r_{BC}+2r_{BC}^2$ است. دو ماتریس مشابه از فراوانی‌های ژنوتیپی H_{ACB} (بر حسب r_{AC} و r_{BC}) و H_{BAC} (بر حسب r_{AB} و r_{AC}) نیز می‌توانند برای ترتیب‌های نشانگری A-C-B و B-A-C به‌دست آیند.

زمانی که ترتیب نشانگری صحیح ناشناخته باشد، الگوریتمی باید فرموله شود تا محتمل‌ترین ترتیب نشانگری برای تجزیه پیوستگی برآورد شود. اگر q_1 و q_2 احتمالات ترتیب‌های A-B-C و A-C-B باشند، احتمال ترتیب B-A-C برابر ($1-q_1-q_2$) می‌شود. در مدل مختلط نرمال ۲۷، فراوانی هر مؤلفه به‌صورت میانگین وزنی فراوانی یک ژنوتیپ تحت سه نشانگر مختلف است؛ به عبارت دیگر، ماتریس فراوانی‌های ۲۷ ژنوتیپ در سه نشانگر می‌تواند به‌صورت زیر نوشته شود:

$$H=q_1H_{ABC}+q_2H_{ACB}+(1-q_1-q_2)H_{BAC}$$

عناصر این ماتریس نشان‌دهنده فراوانی‌های ژنوتیپی استقرایی $\pi_{j_1 j_2 j_3}$ هستند که j_1، j_2 و j_3 ژنوتیپ‌های هر یک از سه نشانگر هستند.

با استفاده از اصول مشابه با تجزیه دو نقطه‌ای، می‌توان توزیع حاشیه‌ای $y_i = (y_{1i}, y_{2i}, y_{3i})$ را بر حسب اختلاطی از ۲۷ توزیع نرمال سه متغیره فرموله نمود. الگوریتم EM برای برآورد نرخ‌های نوترکیبی r_{AB}، r_{AC} و r_{BC}، احتمالات ترتیب نشانگر، q_1 و q_2 و سه میانگین و واریانس باقیمانده در هر نشانگر بسط یافته است. برآوردهای مربوط به بیشترین احتمالات ترتیب نشانگری به عنوان برآوردهای بهینه در نظر گرفته می‌شوند.

۷-۴- تمرین‌ها

۷-۱- بر اساس اطلاعات ژنوتیپی ارائه شده در جدول ۷-۴، معادله (۷-۵) را اثبات کنید، یعنی:

$$P\left(\varsigma_6 \mid \Lambda_1, \varsigma_5\right) = P\left(\overline{\Lambda}_6 \mid \Lambda_1, \Lambda_5\right) = \frac{1}{2} r.$$

۷-۲- جدول ۷-۴ را به یاد آورید. روشی را برای تجزیه پیوستگی بین نشانگرهای A و B توصیف کردیم. روش مشابهی را نشان دهید که بتواند تجزیهٔ پیوستگی را بین نشانگرهای B و C انجام دهد.

الف. درستنمایی مربوطه را نوشته و نرخ نوترکیبی r_{AC} را تحت ترکیب دیپلوتیپی بهینه فرد پایه‌گذار برآورد کنید.

ب. درستنمایی توأمی را بنویسید که ترکیب دیپلوتیپی فرد پایه‌گذار و نرخ‌های نوترکیبی را با یکدیگر ترکیب کند.

۷-۳- مورتون[1] (۱۹۵۶) شجره پیچیده‌ای را برای تجزیه پیوستگی elliptocytosis[2] خانوادگی گزارش نمود (شکل ۷-۷). این حالت بی‌نهایت نادر بوده و فراوانی جمعیتی آلل بیماری بسیار پایین است. تحقیقات روی این شجره و شجره‌های دیگر نشان داد که توارث این بیماری با توارث غالب اتوزومی با شدت نفوذ ژنی کامل در توافق است. فراوانی پایین بیماری- آلل و توارث غالب آن اجازه امتیازدهی به هر فرد بیماری را به‌صورت Dd و هر فرد سالم را به‌صورت dd می‌دهد و هر آلل D به صورت مشابه اجدادی (IDB) در نظر گرفته می‌شود. نشان دهید که تابع درستنمایی این خانواده می‌تواند به صورت زیر بیان شود:

---

[1] - Morton

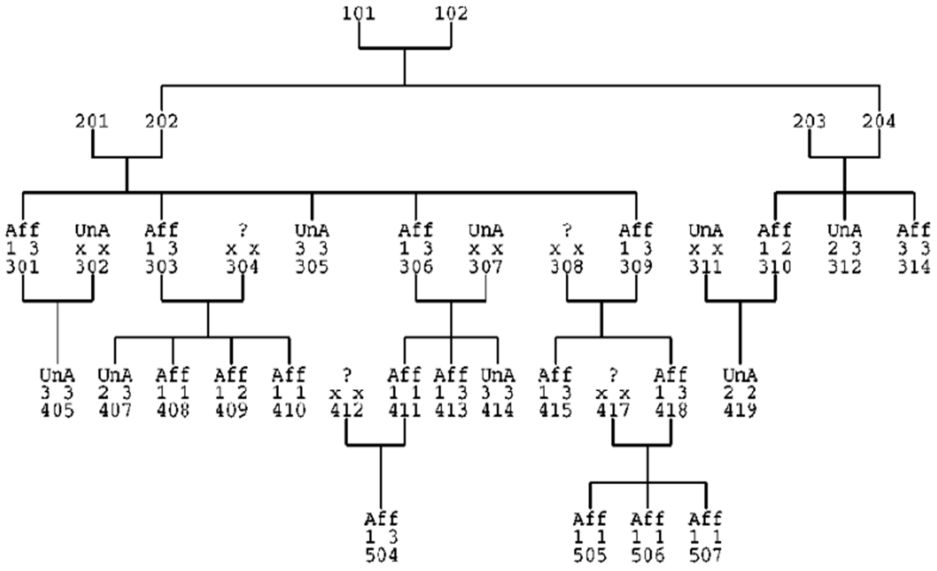
[2] - Elliptocytosis یک عارضه ژنتیکی است که در آن، گلبول‌های قرمز خون بدشکل می‌شوند (مترجمین).

$$
\begin{aligned}
L \propto\ & 810r(1-r)^{19}+324r(1-r)^{18}+180r(1-r)^{17}+72r(1-r)^{16}+90r^{3}(1-r)^{17}\\
&+72r^{3}(1-r)^{16}+40r^{3}(1-r)^{15}+24r^{3}(1-r)^{14}+90r^{4}(1-r)^{15}+20r^{4}(1-r)^{13}\\
&+90r^{5}(1-r)^{15}+432r^{5}(1-r)^{14}+20r^{5}(1-r)^{13}+104r^{5}(1-r)^{12}\\
&+1800r^{6}(1-r)^{14}+558r^{6}(1-r)^{13}+440r^{6}(1-r)^{12}+176r^{6}(1-r)^{11}\\
&+90r^{7}(1-r)^{13}+324r^{7}(1-r)^{12}+120r^{7}(1-r)^{10}+360r^{8}(1-r)^{12}\\
&+378r^{8}(1-r)^{11}+80r^{8}(1-r)^{10}+76r^{8}(1-r)^{9}+4r^{8}(1-r)^{4}+180r^{9}(1-r)^{11}\\
&+522r^{9}(1-r)^{10}+80r^{9}(1-r)^{9}+100r^{9}(1-r)^{8}+10r^{9}(1-r)^{3}+180r^{10}(1-r)^{10}\\
&+846r^{10}(1-r)^{9}+40r^{10}(1-r)^{8}+216r^{10}(1-r)^{7}+18r^{10}(1-r)^{4}+4r^{10}(1-r)^{2}\\
&+117r^{11}(1-r)^{9}+378r^{11}(1-r)^{8}+260r^{11}(1-r)^{7}+72r^{11}(1-r)^{6}\\
&+45r^{11}(1-r)^{3}+180r^{12}(1-r)^{8}+396r^{12}(1-r)^{7}+40r^{12}(1-r)^{5}+18r^{12}(1-r)^{2}\\
&+270r^{13}(1-r)^{7}+234r^{13}(1-r)^{6}+40r^{13}(1-r)^{5}+52r^{13}(1-r)^{4}+180r^{14}(1-r)^{6}\\
&+180r^{14}(1-r)^{5}+80r^{14}(1-r)^{4}+16r^{14}(1-r)^{3}+90r^{15}(1-r)^{5}+162r^{15}(1-r)^{4}\\
&+20r^{15}(1-r)^{3}+180r^{16}(1-r)^{4}+72r^{16}(1-r)^{3}+90r^{17}(1-r)^{3}
\end{aligned}
$$

نقطه اوج پروفیل درستنمایی برای مقادیر r در نرخ نوترکیبی ۰/۰۵ برابر با ۳/۳۱ برآورد شد.

۷-۴- دو روش برای تجزیه پیوستگی نشانگرهای غالب معرفی شد. اولین روش بر اساس مشاهدات کیفی حضور یا عدم حضور باند بود، در حالی‌که دومین روش از اندازه‌گیری کمّی باندها بر اساس تکنیک‌های بازتابش استفاده نمود. حائز اهمیت است که چگونگی عمل این دو روش مختلف در شرایط پیچیده واقعی که می‌تواند از راه مطالعات شبیه‌سازی به شکل زیر مورد تحقیق قرار گیرد، مقایسه شود:

مقادیر کمّی شدت باند را با فرض این که دو نشانگر به هم پیوسته در جمعیت F_2 بر اساس نسبت ۱:۲:۱ تفرق می‌یابند بر اساس توزیع نرمال شبیه‌سازی می‌کنیم. روش آماری که در این فصل توصیف می‌شود، برای برآورد نرخ نوترکیبی بین دو نشانگر مفروض استفاده می‌شود. سپس نشانگرهایی که به صورت کمّی شبیه‌سازی شده‌اند، به صورت نشانگرهای (۱/۰)- غالب در نظر گرفته می‌شوند. الگوریتم EM که در فصل ۴ معرفی شد، برای برآورد نرخ نوترکیبی نشانگرهای غالب به کار گرفته می‌شود. نتایج برآوردهای حاصل از این تیمارهای کمّی و کیفی را مقایسه کنید.



شکل ۷-۷- یک شجره پیچیده در انسان. اقتباس از مورتون (۱۹۵۶).

فصل هشتم

تجزیهٔ فنوتیپ‌های نشانگر

۸-۱- مقدمه

در فصول قبل، مسائل آماری متعدد مرتبط با تجزیه پیوستگی نشانگرهای مولکولی و تهیه نقشه‌های ژنتیکی پیوستگی توصیف شدند. یکی از مهم‌ترین اهداف تجزیه این نشانگرها شناسایی ترتیب آن‌ها روی کروموزوم‌ها بود که به کمک آن‌ها، فرد بتواند مکان‌های ژنی صفات کمّی (QTL ها) عملکردی تعیین‌کننده تنوع فنوتیپی پیچیده را در مناطق خاص ژنوم مکان‌یابی کند. شناسایی QTL ها در سطح ژنوم، موقعیت و آثار آن‌ها، دارای اهمیت اساسی برای ژنتیک کشاورزی، تکامل و زیست- پزشکی است.

روش‌های متعددی برای مکان‌یابی QTL ها ارائه شده‌اند (هوشله و همکاران[1]، ۱۹۹۷؛ لینچ و والش، ۱۹۹۸). این روش‌ها می‌توانند در قالب آزمون‌های t و تجزیه واریانس، تجزیه حداقل مربعات (LS)، تجزیه حداکثر درستنمایی (ML) و تجزیه بیزین دسته‌بندی شوند. این روش‌ها از نظر نیازهای محاسباتی، کارآیی آن‌ها از نظر استخراج اطلاعات، قابلیت انعطاف‌پذیری آن‌ها در ارتباط با بکارگیری ساختارهای متفاوت داده‌ها و قابلیت مکان‌یابی QTL های چندگانه متفاوت هستند. روش ساده LS از نظر سرعت محاسباتی توانمند و کارا است، ولی نمی‌تواند تمامی اطلاعات حاصل از داده‌ها را استخراج کند و به طرح‌های آمیزشی خاصی محدود می‌شود. تکنیک مکان‌یابی فاصله‌ای بر اساس حداکثر درستنمایی لندر و بوتستین (۱۹۸۹) یکی از پرکاربردترین روش‌های تجزیه QTL در تلاقی‌های کنترل شده یا شجره‌های دارای ساختار مشخص است.

روش مکان‌یابی فاصله‌ای به مکان‌یابی فاصله‌ای مرکب[2] (زنگ[3]، ۱۹۹۴) و مکان‌یابی فاصله‌ای چندگانه[4] (کائو و همکاران[5]، ۱۹۹۹) توسعه یافته است.

در این فصل، دربارهٔ آزمون t، تجزیه واریانس و تجزیه تابعیت (رگرسیون) نشانگرهای چندگانه و انجام آزمون‌های آماری فقط بر اساس اطلاعات نشانگرهای منفرد DNA بحث خواهیم کرد. برای تجزیه تک نشانگری، هیچ نقشه ژنتیکی مورد نیاز نیست و محاسبات بر اساس میانگین‌ها و واریانس‌های درون هر یک از کلاس‌های ژنوتیپی انجام می‌شود. تجزیه نشانگر می‌تواند بسط یابد تا شامل تمامی نشانگرهای ژنوم شود (ژو، ۲۰۰۳). اگرچه تجزیه تک نشانگری همانگونه که در بخش‌های بعدی این فصل نشان داده شده است، اثر QTL و موقعیت آن‌را ترکیب می‌کند، اما نتایج حاصل از آن استفاده از مکان‌یابی فاصله‌ای پیشرفته‌تر را برای شناسایی QTL های واقع در یک

---

[1] - Hoeschele *et al.*
[2] - Composite interval mapping
[3] - Zeng
[4] - Multiple interval mapping
[5] - Kao *et al.*

فاصله ژنومی احاطه شده به وسیله دو نشانگر به هم پیوسته تسهیل می‌کند (روتمن[1] و چورود، ۱۹۹۷). مکان‌یابی فاصله‌ای به روش حداکثر درستنمایی (ML) و بسط آن به مکان‌یابی فاصله‌ای مرکب در فصول ۱۱ و ۱۳ ارائه خواهند شد.

۸-۲- مدل تابعیت QTL

قبل از معرفی یک روش پیشرفته آماری برای مکان‌یابی QTL، ابتدا یک مثال فرضی از یک طرح تلاقی برگشتی موش‌ها که در جدول ۸-۱ نشان داده شده است را در نظر بگیرید. این مثال شامل ۱۰ موش تعیین فنوتیپ شده از نظر وزن بدن، y و یک QTL با دو ژنوتیپ شناخته شده، Qq (که با ۱ نشان داده شده است) و qq (که با ۰ نشان داده شده است) است. به نظر می‌رسد موش‌هایی که ژنوتیپ Qq مربوط به QTL را حمل می‌کنند، سنگین‌تر از موش‌هایی باشند که ژنوتیپ qq را حمل می‌کنند، اگرچه موش‌هایی که ژنوتیپ مشابهی را حمل می‌کنند، دقیقاً وزن بدن مشابهی ندارند. برای آزمون اینکه چنین حالتی حقیقتاً وجود دارد و نیز برآورد اثر QTL روی وزن بدن، یک مدل تابعیت ساده را به‌صورت زیر فرموله می‌کنیم:

$$y_i = \mu + z_i a + e_i \qquad (۱-۸)$$

که y_i ارزش فنوتیپی موش i ام، μ میانگین کل، z_i متغیر شاخص و موهومی که ژنوتیپ QTL موش i ام را مشخص می‌کند و به صورت زیر تعریف می‌شود:

$$z_i = \begin{cases} 1 & \text{اگر ژنوتیپ QTL به صورت Qq باشد} \\ 0 & \text{اگر ژنوتیپ QTL به صورت qq باشد} \end{cases}$$

a اثر افزایشی QTL و e_i خطای تصادفی است که به‌طور خاص فرض می‌شود که دارای توزیع نرمال به صورت $N\left(0,\sigma^2\right)$ است.

مدل خطی (۸-۱) می‌تواند بسط یابد تا آثار ژنتیکی QTL را در جمعیت F_2 با سه ژنوتیپ QTL به صورت QQ (که با ۲ نشان داده می‌شود)، Qq (که با ۱ نشان داده می‌شود) و qq (که با ۰ نشان داده می‌شود) برآورد نماید. مدل مربوط به F_2 بدین صورت نوشته می‌شود:

$$y_i = \mu + z_{1i} a + z_{2i} d + e_i \qquad (۲-۸)$$

که دارای پارامتر اضافی d است که همان اثر غالبیت QTL است و متغیرهای شاخص z_{1i} و z_{2i} که به صورت زیر تعریف می‌شوند:

---

[1] - Routman

$$z_{1i} = \begin{cases} 1 & \text{اگر ژنوتیپ QTL به صورت QQ باشد} \\ -1 & \text{اگر ژنوتیپ QTL به صورت qq باشد} \end{cases}$$

و:

$$z_{2i} = \begin{cases} 0 & \text{اگر ژنوتیپ QTL به صورت QQ یا qq باشد} \\ 1 & \text{اگر ژنوتیپ QTL به صورت Qq باشد} \end{cases}$$

روش‌های حداقل مربعات استاندارد (LS) می‌توانند برای برآورد عرض از مبدأ ناشناخته مدل یعنی μ و ضرایب تابعیت یعنی a و یا d، استفاده شوند؛ بنابراین، با آزمون مستقیم a و یا d می‌توان تعیین کرد که این QTL اثر معنی‌داری روی وزن بدن می‌گذارد یا خیر.

مثال ۸-۱. یک جمعیت کوچک تلاقی برگشتی با ۱۰ موش را در نظر بگیرید. هر موش برای دو نشانگر تعیین ژنوتیپ شده و همچنین وزن بدن نیز اندازه‌گیری شده است. داده‌های نشانگر و داده‌های فنوتیپی در جدول ۸-۱ ارئه شده‌اند. با این جدول، روشی برای برآورد و آزمون اثر ژنتیکی QTL یعنی a، روی وزن بدن موش‌ها ارایه می‌نماییم؛ بنابراین داریم:

$$y = \begin{pmatrix} 30 \\ 32 \\ 28 \\ 29 \\ 29 \\ 22 \\ 20 \\ 21 \\ 20 \\ 21 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \qquad e = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \end{pmatrix}$$

و $b = (\mu, a)^T$ است. ازاین‌رو، رابطه $y = Xb + e$ را خواهیم داشت. برآوردهای حداقل مربعات (LS) پارامترها به صورت زیر هستند:

$$\hat{b} = \left(X^T X\right)^{-1}\left(X^T y\right) = \left(20.8, 8.8\right)^T$$

و:

$$\hat{\sigma}^2 = \frac{1}{n}\left(y - Xb\right)^T\left(y - Xb\right) = 1.2$$

جدول ۸-۱- ساختار داده‌ها برای دو نشانگر تعیین ژنوتیپ شده و وزن بدن موش در یک طرح تلاقی برگشتی.

نمونه	نشانگر		وزن بدن
	A	B	
۱	۱	۱	۳۰
۲	۱	۱	۳۲
۳	۱	۱	۲۸
۴	۱	۱	۲۹
۵	۱	۱	۲۹
۶	۰	۰	۲۲
۷	۰	۱	۲۰
۸	۰	۰	۲۱
۹	۰	۰	۲۰
۱۰	۰	۰	۲۱

با این برآوردها، مجموع مربعات کل (SST) را محاسبه می‌کنیم:

$$\sum_{i=1}^{10}\left(y_i - \mu\right)^2 = 205.6$$

و مجموع مربعات باقیمانده (SSE) برابر است با:

$$\sum_{i=1}^{10}\left(y_i - \mu - z_i a\right)^2 = 12.0$$

اکنون، معنی‌داری اثر ژنتیکی QTL (a) با محاسبه مقدار F آزمون می‌شود:

$$F = \frac{\dfrac{\left(SST - SSE\right)}{(2-1)}}{\dfrac{SSE}{(10-2)}} = 129.07$$

که در مقایسه با مقدار بحرانی جدول F یعنی $F_{0.05,(2,8)}$ = ۴/۴۶، نتیجه می‌گیریم که این QTL اثر معنی‌داری بر وزن بدن در جمعیت تلاقی برگشتی موش‌ها دارد.

۸-۳- تجزیه در نشانگر

مدل آماری QTL فرض می‌کند که ژنوتیپ‌های QTL می‌توانند در یک جمعیت مکان‌یابی مشاهده شوند. در عمل چنین چیزی امکان ندارد. آنچه که می‌توانیم انجام دهیم آن است که از نشانگرهای قابل مشاهده برای پیش‌بینی چنین QTL های غیر قابل مشاهده‌ای از راه پیوستگی بین نشانگرها و QTL ها استفاده کنیم؛ بنابراین، با انجام تجزیه ارتباط بین نشانگرها و فنوتیپ‌ها، هنوز می‌توانیم اثر QTL مفروض را بر تنوع فنوتیپی نتیجه‌گیری کنیم.

استفاده از تک نشانگر برای شناسایی QTL محدود است، چون نمی‌تواند تعیین کند که QTL در کدام سمت نشانگر، چپ یا راست، واقع شده است. در هر حال، تجزیه‌های تک نشانگری، اگرچه نمی‌توانند موقعیت QTL را برآورد کنند، اما برای آزمون ابتدایی وجود QTL مفید هستند. در بخش زیر، دو روش برای آزمون تجزیهٔ پیوستگی بر اساس آمارهای آزمون t و F معرفی می‌کنیم.

۸-۳-۱- آزمون t دو نمونه‌ای

داده‌های تلاقی برگشتی موش‌ها برای دو نشانگر مولکولی به هم پیوسته A و B تعیین ژنوتیپ شده و در جدول ۸-۱ ارایه شده‌اند. دو ژنوتیپ در هر نشانگر به صورت ۱ و ۰ تعریف می‌شوند. پیوستگی بین این دو نشانگر می‌تواند از روی انطباق ژنوتیپ‌های آن‌ها در بین نمونه‌ها، به استثنای موش‌های ۵ و ۶، مشاهده شود. نرخ نوترکیبی بین دو نشانگر $r = \frac{2}{10} = 0.2$ است. این دو نشانگر به‌طور جداگانه تجزیه خواهند شد.

به نشانگر A نگاه کنید، به نظر می‌رسد که دو گروه ژنوتیپ‌های نشانگر از نظر وزن بدن تفاوت دارند. طبیعتاً پرسشی دربارهٔ اینکه این تفاوت در وزن بدن بین ژنوتیپ‌های ۱ و ۰ در نشانگر A از نظر آماری معنی‌دار است یا خیر مطرح می‌شود. این فرضیه می‌تواند با آزمون t دو نمونه‌ای سنجیده شود. اگر μ_1 و μ_0 به ترتیب میانگین حقیقی صفت در دو گروه متفاوت از موش‌ها با ژنوتیپ ۱ و ۰ و m_1 و m_0 به ترتیب میانگین‌های نمونه‌ای مربوط به آن‌ها باشند، فرض‌های آزمون می‌توانند به‌صورت زیر نوشته شوند:

$$H_0 : \mu_1 = \mu_0$$
$$H_1 : \mu_1 \neq \mu_0$$

آماره آزمون t مورد استفاده برای آزمون معنی‌داری تفاوت بین دو میانگین عبارت است از:

$$t = \frac{m_1 - m_0}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}} \qquad (۳-۸)$$

که s^2 تلفیق واریانس دو نمونه یا همان میانگین وزنی واریانس دو نمونه بوده و برابر است با:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}$$

که در آن، n_1، n_0، s_1^2 و s_0^2 به ترتیب اندازه و واریانس مربوط به نمونه‌های حاصل از دو گروه متفاوت نشانگر هستند. فرض صفر (H_0) در شرایطی رد می‌شود که آماره آزمون t محاسبه شده بیشتر از یا مساوی با مقدار بحرانی حاصل از توزیع t باشد. اگر نقطه بحرانی بالاتر α را با $t_{(\alpha,\nu)}$ نشان دهیم، وقتی که $t > t_{(0.025,\nu)}$ باشد، فرض صفر رد خواهد شد که $t_{(0.025,\nu)}$ مقدار بحرانی t در آزمون دوطرفه برای سطح معنی‌دار ۰/۰۵ با درجه آزادی $\nu = n_1 + n_0 - 2$ است.

مثال ۸-۲. در مثالی که با $n_1 = n_0 = ۵$ در جدول ۸-۱ ارایه شده است، $m_1 = ۲۹/۶$، $m_0 = ۲۰/۸$، $s_1 = ۰/۸۳۶۷$، $s_0 = ۱/۵۱۶۶$ و $s = ۱/۵۰$ را برای نشانگر A محاسبه نمودیم. همچنین $t = \frac{29.6 - 20.8}{\sqrt{1.5^2\left(\frac{1}{5} + \frac{1}{5}\right)}} = ۱۱/۳۶۰۸$ را محاسبه می‌کنیم. در مقایسه با مقدار بحرانی $t_{(0.025,5+5-2)} = ۲/۳۰۶۰$، نتیجه می‌گیریم که نشانگر A به‌طور معنی‌داری با وزن بدن مرتبط است.

در این مثال، همچنین دریافتیم که ژنوتیپ ۱ برای هر یک از دو نشانگر تمایلی را برای وزن بیشتر نسبت به ژنوتیپ ۲ از خود نشان می‌دهد. چون نتیجه حاصل از آزمون t فقط وجود تفاوت را نشان داد، نمی‌توانیم رسماً اظهار نظری دربارهٔجهت تفاوت داشته باشیم (اگر آزمون یک‌طرفه انجام می‌شد، می‌توانستیم در خصوص جهت تفاوت اظهار نظر کنیم، ولی به‌طور معمول آزمون دو دنباله انجام می‌شود که در این صورت نمی‌توان جهت تفاوت را مشخص نمود). به هر حال بیشتر افراد با این نتیجه غیر رسمی که ژنوتیپ ۱ در هر نشانگر نسبت به ژنوتیپ ۲ منجر به وزن بیشتر می‌شود، مشکلی ندارند.

مثال ۸-۳. **(ارتفاع گیاه گوجه فرنگی).** مجدداً به داده‌های مثال ۲-۱ نگاه می‌کنیم، اما اکنون داده‌های نشانگر نیز ارائه می‌شود. فرض کنید داده‌های زیر (y) را در خصوص ارتفاع (بر حسب

سانتی‌متر) ۱۲ گیاه گوجه فرنگی از یک گونه خاص که به دو دسته نشانگری گروه‌بندی شده‌اند، در اختیار داریم:

$$M_1M_1 \quad : \quad y = (79,82,100,102,124)$$
$$M_1M_2 \quad : \quad y = (85,87,101,103,125,126,127)$$

میانگین‌ها و واریانس‌های مشاهده شده کلاس‌های نشانگری به‌صورت زیر هستند:

$$\hat{y}_{M_1M_1} = 97.4 \quad \bar{y}_{M_1M_2} = 101.71$$
$$s_{M_1M_1} = 18.10 \quad s_{M_1M_2} = 18.33$$

برای آزمون H_0، طبیعی به نظر می‌رسد که از آزمون t دو نمونه‌ای استفاده کنیم که یکی از اولین روش‌ها است. آماره t برای آزمون H_0 به صورت زیر است:

$$t = \frac{y_{M_1M_1} - y_{M_1M_2}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

که s^2 میانگین وزنی واریانس دو گروه و n_1 و n_2 اندازه نمونه‌ها در هر گروه هستند.

یک مشکل در خصوص استفاده از این آماره برای آزمون H_0 آن است که به‌دلیل وجود توزیع مخلوط، این آماره دارای توزیع صفر t استیودنت معمولی نیست. از این‌رو، تنظیم این آماره، یعنی محاسبه p-value، استدلال معتبری را ارائه نخواهد داد. یک روش نزدیک به این حالت، استفاده از آزمون تبدیل[1] و به‌دست آوردن توزیع صفر آماره آزمون بر اساس تبدیل‌ها است.

در این وضعیت، انجام آزمون تبدیل کاملاً ساده است. داده‌ها به‌صورت تصادفی به یکی از دو گروه منتسب شده و آماره t حاصله محاسبه می‌شود. قضیه اثبات شده آن است که تحت فرض H_0 هیچ‌گونه تفاوتی بین دو گروه نشانگر وجود ندارد و از این‌رو مشاهده مربوط به هر یک از دو گروه تقریباً می‌تواند مشابه باشد. این انتساب تصادفی به دفعات زیاد تکرار می‌شود و آماره t حاصله تشکیل هیستوگرامی را می‌دهد که به‌عنوان توزیع صفر عمل می‌کند.

مثال ۸-۴. **(ادامه مثال ارتفاع گیاه گوجه فرنگی)**. برای داده‌های مثال ۸-۳، آماره t داده‌های مشاهده شده ۰/۹۶۵- است و هیستوگرام حاصل از ۵۰۰۰ آماره t تبدیل شده در شکل ۸-۱ ارائه شده است. همچنین برنامه‌ای که هیستوگرام را تولید کرده است در ضمیمه ب.۲ ارائه شده است.

---

[1] - Permutation test

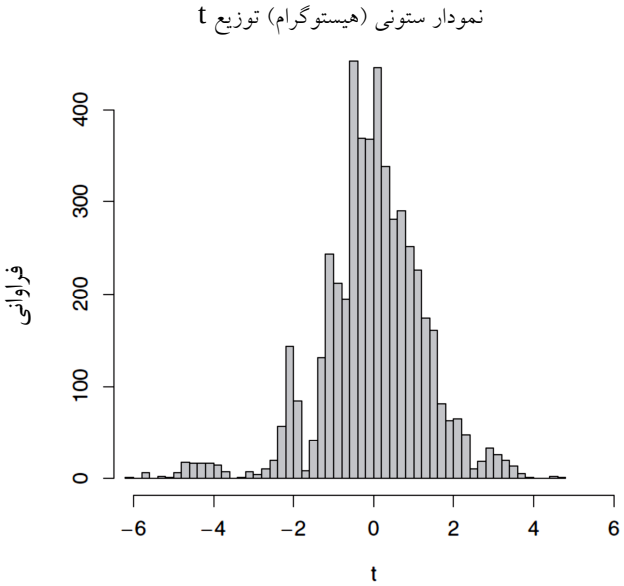
نقاط برش ۵ درصد دنباله بالا و پایین توزیع تبدیل به ترتیب ۱/۹۷ و ۲/۱۱- هستند. بر اساس این توزیع صفر، نتیجه می‌گیریم که هیچ QTL در نزدیکی این نشانگر وجود ندارد.

۸-۳-۲- تجزیه واریانس

برای جمعیت F_2، سه گروه متفاوت از ژنوتیپ‌های نشانگر وجود دارند که می‌توانند به ترتیب به صورت ۲، ۱ و ۰ در هر نشانگر تعریف شوند (جدول ۸-۲ را مشاهده کنید). برای آزمون تفاوت کلی بین سه ژنوتیپ، می‌توان از تجزیه واریانس معمولی (ANOVA) استفاده نمود. میانگین مربعات ناشی از تفاوت بین سه ژنوتیپ نشانگر بازتابی از درجه ارتباط نشانگر با QTL مفروض برای یک صفت خاص است، درحالیکه میانگین مربعات ناشی از تفاوت درون ژنوتیپ‌ها بازتابی از واریانس باقیمانده است. نسبت این دو میانگین مربعات، یعنی مقدار F، آماره آزمونی است که برای آزمون معنی‌دار بودن تفاوت بین سه ژنوتیپ نشانگر استفاده می‌شود.

مقدار F محاسبه شده با مقدار بحرانی حاصل از توزیع F یعنی $F_{0.05,(2,n-3)}$، مقایسه می‌شود. واریانس ژنتیکی حاصل از یک نشانگر معنی‌دار می‌تواند با برابر قرار دادن میانگین مربعات مورد انتظار یا همان امید ریاضی میانگین مربعات (جدول ۸-۳) با میانگین‌های مربعات (MS) و حل کردن معادلات حاصله برآورد شود:

$$\sigma_g^2 = \frac{MS_1 - MS_2}{k} \qquad (۸-۴)$$



شکل ۸-۱- توزیع صفر آماره t مثال ۸-۴ بر اساس تبدیل‌ها.

جدول ۸-۲- ساختار داده‌ها برای دو نشانگر تعیین ژنوتیپ شده و وزن بدن موش در طرح F_2.

نمونه	نشانگر		وزن بدن
	A	B	
۱	۲	۲	۳۰
۲	۲	۱	۳۲
۳	۲	۰	۲۸
۴	۱	۲	۲۹
۵	۱	۱	۲۹
۶	۱	۰	۲۲
۷	۱	۰	۲۰
۸	۰	۲	۲۱
۹	۰	۱	۲۰
۱۰	۰	۰	۲۱

سهمی از واریانس فنوتیپی یک صفت کمّی که به وسیله نشانگر توضیح داده می‌شود، یعنی وراثت‌پذیری عمومی، به صورت زیر برآورد می‌شود:

$$R^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \qquad (۵-۸)$$

این سهم تا حد زیادی به عنوان یک پارامتر جهت ارزیابی میزان مشارکت نشانگر A در تنوع ژنتیکی استفاده می‌شود.

جدول ۸-۳- خلاصه‌ای از تجزیه واریانس برای بررسی تفاوت بین سه گروه ژنوتیپی در جمعیت F_2.

منابع تغییر	درجه آزادی	میانگین مربعات	مقدار F	امید ریاضی میانگین مربعات
بین ژنوتیپ‌های نشانگر	۲	MS_1	$\frac{MS_1}{MS_2}$	$\sigma_e^2 + k\sigma_q^2$
درون ژنوتیپ‌های نشانگر	n-۳	MS_2		σ_e^2

توجه: $k=\dfrac{3}{\left(\dfrac{1}{n_2}+\dfrac{1}{n_1}+\dfrac{1}{n_0}\right)}$ میانگین هارمونیک و n_2، n_1 و n_0 اندازه‌های نمونه سه ژنوتیپ نشانگر مختلف در جمعیت F_2 هستند.

تفاوت کلی بین سه ژنوتیپ نشانگر در جمعیت F_2 ممکن است به‌واسطه اثر افزایشی، غالبیت یا هر دو باشد. معنی‌داری این دو اثر نیز می‌تواند با استفاده از آزمون t بررسی شود. برای آزمون اثر افزایشی نشانگر، آماره آزمون به صورت زیر بیان می‌شود:

$$t_1=\frac{m_2-m_0}{\sqrt{s^2\left(\dfrac{1}{n_2}+\dfrac{1}{n_0}\right)}} \qquad (۶-۸)$$

که در آن:

$$s^2=\frac{(n_2-1)s_2^2+(n_0-1)s_0^2}{n_2+n_0-2}$$

و برای آزمون اثر غالبیت نشانگر، آماره آزمون را به صورت زیر بیان می‌کنیم:

$$t_2=\frac{m_1-\dfrac{1}{2}(m_2+m_0)}{\sqrt{s^2\left(\dfrac{1}{4n_2}+\dfrac{1}{n_1}+\dfrac{1}{4n_0}\right)}} \qquad (۷-۸)$$

که در آن:

$$s^2=\frac{(n_2-1)s_2^2+(n_1-1)s_1^2+(n_0-1)s_0^2}{n_2+n_1+n_0-3}$$

s_2^2، s_1^2 و s_0^2 به ترتیب واریانس‌های نمونه‌ای در سه گروه مختلف نشانگر در F_2 هستند.

مثال ۸-۵. در جدول ۸-۲ مثالی برای جمعیت F_2 با ۱۰ موش که هر یک از نظر وزن بدن اندازه‌گیری شده و برای نشانگرهای هم‌بارز A و B تعیین ژنوتیپ شده‌اند، ارایه شده است. میانگین مربعات تفاوت‌های بین و درون ژنوتیپی را برای نشانگر A به صورت MS_1 = ۶۵/۴۷ و MS_2 = ۱۰/۶۷ محاسبه می‌کنیم که از روی آن مقدار F برابر با ۶/۱۴ محاسبه می‌شود. با مقایسه F محاسبه شده با مقدار بحرانی یعنی $F_{0.05,(2,10-3=7)}$ = ۴/۷۳۷۴، نتیجه‌گیری می‌شود که این

نشانگر به‌طور معنی‌داری با وزن بدن در ارتباط است. واریانس ژنتیکی ناشی از این نشانگر با استفاده از معادله (۸-۴) برابر با ۵/۰۲۳۳ = σ_g^2 محاسبه می‌شود و سپس وراثت‌پذیری برابر با ۰/۳۲۰۲ برآورد می‌شود.

برای نشانگر A، سه گروه ژنوتیپی به ترتیب شامل ۳= n_2، ۴=n_1 و ۳=n_0 فرد و سه میانگین نمونه‌برداری به ترتیب برابر با ۳۰ = m_2، ۲۵ = m_1 و ۲۰/۶۷ = m_0 و سه واریانس نمونه‌گیری به ترتیب برابر با ۴ = s_2^2، ۲۲= s_1^2 و ۰/۳۳۳۳ = s_0^2 محاسبه شدند. آماره آزمون t نیز برای آثار افزایشی و غالبیت به ترتیب با معادلات (۸-۶) و (۸-۷) برابر با ۷/۷۶۵۸ = t_1 و ۰/۱۲۲۰- = t_2 محاسبه می‌شوند. با مقایسه مقادیر محاسبه شده t با مقدار بحرانی t یعنی ۲/۱۳۱۸ = $t_{(0.025,\nu=3+3-2=4)}$ برای آزمون اثر افزایشی و ۱/۸۹۴۶ = $t_{(0.025,\nu=3+4+3-3=7)}$ برای اثر غالبیت، نتیجه می‌گیریم که نشانگر A اثر افزایشی معنی‌داری را نشان می‌دهد، ولی اثر غالبیت بر وزن بدن معنی‌دار نیست.

روش محاسباتی مشابهی برای نشانگر B انجام می‌شود. این نشانگر دارای مقدار F برابر با ۰/۸۵ است که نشان می‌دهد هیچ ارتباط معنی‌داری با وزن بدن در موش‌ها ندارد. مقادیر t برای آزمون آثار افزایشی و غالبیت به ترتیب برابر با ۱/۲۲۶۴ = t_1 و ۰/۵۶۳۵ = t_2 محاسبه می‌شوند. مشاهده می‌شود که هر دو اثر افزایشی و غالبیت غیر معنی‌دار هستند.

۸-۳-۳- تجزیه ژنتیکی

چرا می‌توانیم حضور یک QTL فرضی را برای یک صفت کمّی از راه انجام یک آزمون t یا تجزیه واریانس (ANOVA) بر اساس میانگین کلاس‌های نشانگر نتیجه‌گیری کنیم؟ یک QTL فرضی پیوسته با یک نشانگر را با نرخ نوترکیبی r در نظر بگیرید. ارزش‌های ژنوتیپی مورد انتظار شرطی مرتبط با هر ژنوتیپ نشانگر از روی احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ نشانگر و از روی ارزش‌های ژنوتیپی ژنوتیپ‌های مختلف QTL محاسبه می‌شوند. مشروط به ژنوتیپ‌های شناخته شده نشانگر، یعنی Aa(۱) و aa(۰)، می‌توانیم احتمالات شرطی دو ژنوتیپ QTL، یعنی Qq(۱) و qq(۰) را برای جمعیت تلاقی برگشتی به‌صورت زیر به‌دست آوریم:

ارزش ژنوتیپی نشانگر	ارزش ژنوتیپی QTL	
	μ_1	μ_0
m_1	۱-r	r
m_0	r	۱-r

ارزش‌های ژنتیکی این دو ژنوتیپ QTL در جمعیت تلاقی برگشتی، به ترتیب می‌تواند بدین صورت نشان داده شود:

$$\mu_0 = \mu - \frac{1}{2}a \qquad \text{و} \qquad \mu_1 = \mu + \frac{1}{2}a$$

برای هر ژنوتیپ نشانگر، دو ژنوتیپ متفاوت QTL مخلوط شده و وزن آن‌ها احتمالات شرطی مربوط به آن‌ها است؛ بنابراین، ارزش‌های ژنوتیپی شرطی مورد انتظار مرتبط با ژنوتیپ‌های متفاوت نشانگر می‌توانند بدین صورت محاسبه شوند:

$$m_1 = (1-r)\mu_1 + r\mu_0 = (1-r)\left(\mu + \frac{1}{2}a\right) + r\left(\mu - \frac{1}{2}a\right) = \mu + \frac{1}{2}(1-2r)a$$

$$m_0 = r\mu_1 + (1-r)\mu_0 = r\left(\mu + \frac{1}{2}a\right) + (1-r)\left(\mu - \frac{1}{2}a\right) = \mu - \frac{1}{2}(1-2r)a$$

بنابراین، تفاوت میانگین دو نشانگر برابر است با:

$$\mu_1 - \mu_0 = (1-2r)a \qquad (۸-۸)$$

اگر a تفاوت معنی‌داری از صفر نداشته باشد، آماره آزمون t بر اساس معادله (۸-۳) کوچک‌تر از مقدار بحرانی آماره خواهد بود. در این حالت، آزمون t می‌تواند اطلاعاتی در خصوص معنی‌داری اثر QTL فراهم کند، همچنین، آزمون t غیر معنی‌دار ممکن است به‌واسطه عدم وجود پیوستگی بین نشانگر و QTL (r = ۰/۵) بر اساس معادله (۸-۵) ایجاد شده باشد؛ بنابراین، آزمون t فقط یک آزمون مرکب برای اثر QTL و پیوستگی QTL- نشانگر بوده و اثر آن‌ها را مخلوط می‌کند.

برای جمعیت F_2، میانگین سه کلاس نشانگر به‌صورت مشابه می‌تواند با محاسبه احتمالات شرطی زیر به‌دست آیند:

ارزش ژنوتیپی نشانگر	ارزش ژنوتیپی QTL		
	μ_2	μ_1	μ_0
m_2	$(1-r)^2$	$2r(1-r)$	r^2
m_1	$r(1-r)$	$(1-r)^2 + r^2$	$r(1-r)$
m_0	r^2	$2r(1-r)$	$(1-r)^2$

و ارزش‌های ژنوتیپی QTL مورد نظر به‌دست آید. میانگین سه کلاس نشانگری به صورت زیر نوشته می‌شوند:

$$m_2 = (1-r)^2 \mu_2 + 2r(1-r)\mu_1 + r^2\mu_0$$
$$= \mu + (1-2r)a + 2r(1-r)d$$
$$m_1 = r(1-r)\mu_2 + \left[(1-r)^2 + r^2\right]\mu_1 + r(1-r)\mu_0$$
$$= \mu + (1-2r+2r^2)d$$
$$m_0 = r^2\mu_2 + 2r(1-r)\mu_1 + (1-r)^2\mu_0$$
$$= \mu - (1-2r)a + 2r(1-r)d$$

آزمون‌های اثر افزایشی (a) مربوط به معادله (۶-۸) و اثر غالبیت (d) مربوط به معادله (۸-۷) به ترتیب معادل آزمون این حالت است که پارامترهای مرکب:

$$\frac{1}{2}(m_2 - m_0) = (1-2r)a = a - 2ra$$

و:

$$m_1 - \frac{1}{2}(m_2 - m_0) = (1-2r)^2 d = d - 4r(1-r)d$$

برابر با صفر هستند یا خیر.

به کمک تجزیه فوق، اگرچه آزمون t و ANOVA می‌تواند برای آزمون معنی‌داری تفاوت‌های بین کلاس‌های نشانگر استفاده شوند، اما این آزمون‌ها نمی‌توانند میانگین‌های ژنوتیپی QTL و نرخ نوترکیبی بین یک نشانگر منفرد و یک QTL را تفکیک کنند. اگر تفاوت بین کلاس‌های نشانگر معنی‌دار باشد، همان‌گونه که در مورد دو نشانگر در وزن بدن موش نشان داده شد، هنوز نمی‌دانیم که آیا این تفاوت به‌واسطه پیوستگی قوی (r کوچک) بین نشانگر و یک QTL کوچک اثر است یا پیوستگی ضعیف (r بزرگ) بین نشانگر و یک QTL بزرگ اثر است. در حقیقت، با مقایسه ساده میانگین بین کلاس‌های نشانگر، آثار افزایشی و غالبیت QTL ها به ترتیب به میزان 2r و $4r(1-r)$ کمتر از حد واقعی برآورد می‌شوند. همچنین، تحت شرایطی که دو یا چند QTL روی کروموزوم مشابهی قرار داشته باشند، آزمون t و ANOVA نمی‌توانند آثار QTL های انفرادی را روی فنوتیپ جدا کنند. دو پارامتر اختلاط یافته، یعنی میانگین‌های ژنتیکی QTL و نرخ نوترکیبی، می‌توانند با استفاده از روش‌هایی که در بخش زیر توضیح داده می‌شوند، جداسازی شوند.

۸-۴- دور شدن از نشانگر

تاکنون، تجزیه تک نشانگری را شرح دادیم که در آن فرض نشد که QTL در جایگاه نشانگر قرار دارد. پیرو تحقیق دورج و همکاران (۱۹۹۷)، روشی بر اساس طرح تلاقی برگشتی شرح داده می‌شود که در آن، دو ژنوتیپ در هر نشانگر یا QTL وجود دارد.

۸-۴-۱- درستنمایی

فرض کنید ژنوتیپ مشاهده شده نشانگر به‌صورت M_1M_1 یا M_1M_2 و مطابق با این ژنوتیپ مشاهده شده، ژنوتیپ QTL به‌صورت Q_1Q_1 یا Q_1Q_2 با احتمالات زیر باشد:

	M_1M_1	M_1M_2
Q_1Q_1	$\frac{1-r}{2}$	$\frac{r}{2}$
Q_1Q_2	$\frac{r}{2}$	$\frac{1-r}{2}$

که r نرخ نوترکیبی بین نشانگر و QTL است. فرض کنید که یک صفت فنوتیپی (y) از توزیع نرمال تبعیت می‌کند. با برقراری ارتباط بین فنوتیپ‌های صفت با میانگین‌های فنوتیپی ژنوتیپ‌های نشانگر، به‌صورت مدل مختلط زیر است:

$$\begin{aligned} &\text{مشاهده شده} \quad M_1M_1 \rightarrow y \sim (1-r)N(\mu_1,\sigma^2)+rN(\mu_2,\sigma^2) \\ &\text{مشاهده شده} \quad M_1M_2 \rightarrow y \sim rN(\mu_1,\sigma^2)+(1-r)N(\mu_2,\sigma^2) \end{aligned} \tag{۸-۹}$$

که μ_1 و μ_2 به‌ترتیب میانگین‌های فنوتیپی (یا ارزش‌های ژنوتیپی) صفت برای ژنوتیپ‌های QTL یعنی Q_1Q_1 و Q_1Q_2 هستند.

در مدل (۸-۹)، میانگین و واریانس توزیع‌ها عبارتند از (تمرین ۸-۳):

$$\begin{aligned} &\mu_{M_1M_1}=(1-r)\mu_1+r\mu_2 \\ &\mu_{M_1M_2}=r\mu_1+(1-r)\mu_2 \\ &\sigma^2_{M_1M_1}=\sigma^2_{M_1M_2}=\sigma^2+r(1-r)(\mu_1-\mu_2)^2 \end{aligned} \tag{۸-۱۰}$$

توجه نمایید که اگر هیچ پیوستگی بین نشانگرها و QTL وجود نداشته باشد، یعنی $r=\frac{1}{2}$ باشد، در نتیجه $\mu_{M_1M_1}$ و $\mu_{M_1M_2}$ مساوی خواهند بود؛ بنابراین، فرض عدم وجود پیوستگی عبارت است از:

$$H_0 : r = \frac{1}{2} \qquad \text{یا} \qquad H_0 : \mu_{M_1M_1} - \mu_{M_1M_2} = (1-2r)(\mu_1 - \mu_2) = 0$$

و توجه به این نکته حائز اهمیت است که تحت فرض H_0، نمی‌توان بیان کرد که آیا $r = \frac{1}{2}$ است یا $\mu_1 = \mu_2$ است؛ اما در هر دو حالت فوق، هیچ‌گونه تنوع فنوتیپی عملی قابل تشخیصی وجود ندارد.

اگر فرض کنیم که $y_1 ..... y_{n_1}$ متعلق به گروه نشانگری M_1M_1 و $y_{n_1+1}, ..., y_n$ مربوط به گروه نشانگری M_1M_2 هستند، سپس تابع درستنمایی بر اساس مدل (۸-۹) عبارت است از:

$$\begin{aligned} L(\mu_1, \mu_2, \sigma^2, r \mid y) = & \prod_{i=1}^{n_1} (1-r)\, f(y_i \mid \mu_1, \sigma^2) + r f(y_i \mid \mu_2, \sigma^2) \\ & \times \prod_{i=n_1+1}^{n} r f(y_i \mid \mu_1, \sigma^2) + (1-r)\, f(y_i \mid \mu_2, \sigma^2) \end{aligned} \qquad (۸-۱۱)$$

برای آزمون فرض صفر عدم وجود پیوستگی یعنی H_0: عدم وجود QTL، می‌توانیم از آماره نسبت درستنمایی زیر استفاده کنیم:

$$\lambda = \frac{\max_{\mu_1 = \mu_{12}, \sigma^2, r} L(\mu_1, \mu_2, \sigma^2, r \mid y)}{\max_{\mu_1, \mu_2, \sigma^2, r} L(\mu_1, \mu_{12}, \sigma^2, r \mid y)} \qquad (۸-۱۲)$$

اگر آماره آزمون λ خیلی کوچک باشد، فرض صفر رد خواهد شد. در حالت دیگر، می‌توانیم λ را به $-2\log\lambda$ تبدیل کنیم که اغلب دارای توزیع کای مربع تقریبی است و در صورت بزرگ بودن آن، فرض صفر رد خواهد شد؛ اما مدل مختلط، فرض توزیع کای مربع را غیرمعتبر می‌سازد و آنچه که به‌طور معمول انجام می‌شود یک آزمون تبدیل روی مقادیر $-2\log\lambda$ یا معادل آن یعنی امتیاز LOD است.

آزمون نسبت درستنمایی

برای محاسبه آماره آزمون λ باید هم صورت و هم مخرج کسر (۸-۱۲) را به حداکثر برسانیم. صورت کسر آسان است، چون با فرض صفر $\mu_1 = \mu_{12} = \mu$، درستنمایی (۸-۱۱) به‌صورت زیر خواهد شد:

$$L(\mu_1, \mu_{12}, \sigma^2, r \mid y) = L(\mu, \sigma^2, r \mid y) = \prod_{i=1}^{n} f(y_i \mid \mu, \sigma^2) \qquad (۸-۱۳)$$

کـه در آن، برآوردهـای حـداکثر درسـتنمایی μ و σ^2 بـه صـورت $MLEs\hat{\mu}=\bar{\mu}$ و $\hat{\sigma}^2=\left(\frac{1}{n}\right)\sum_i\left(y_i-\bar{y}\right)^2$ خواهد بود (تمرین ۸-۵).

برای به حداکثر رساندن مخرج معادله (۸-۱۲)، نیـاز بـه حـداکثر کـردن درسـتنمایی (۸-۱۱) است. برای انجام ایـن امـر، ابتـدا از درسـتنمایی (۸-۱۱) لگـاریتم مـی‌گیـریم و سـپس از لگـاریتم درستنمایی مشتق گرفته و آن‌را برابر با صفر قرار می‌دهـیم و معادلـه را حـل مـی‌کنـیم. در اینجـا، برخی از جزئیات نشان داده شده است و بقیه جزئیات در تمـرین ۸-۶ نشـان داده خواهـد شـد. بـا مشتق‌گیری نسبت به μ_1 خواهیم داشت:

(۸-۱۴)

$$\begin{aligned}\frac{\partial}{\partial\mu_1}\log L\left(\mu_1,\mu_2,\sigma^2,r\,|\,y\right)&=\frac{\partial}{\partial\mu_1}\sum_{i=1}^{n_1}\log\left((1-r)\,f\left(y_i\,|\,\mu_1,\sigma^2\right)+rf\left(y_i\,|\,\mu_2,\sigma^2\right)\right)\\&+\frac{\partial}{\partial\mu_1}\sum_{i=n_1+1}^{n}\log\left(rf\left(y_i\,|\,\mu_1,\sigma^2\right)+(1-r)\,f\left(y_i\,|\,\mu_2,\sigma^2\right)\right)\\&=\sum_{i=1}^{n_1}\frac{(1-r)\frac{\partial}{\partial\mu_1}f\left(y_i\,|\,\mu_1,\sigma^2\right)}{(1-r)\,f\left(y_i\,|\,\mu_1,\sigma^2\right)+rf\left(y_i\,|\,\mu_2,\sigma^2\right)}\\&+\sum_{i=1}^{n_1}\frac{r\frac{\partial}{\partial\mu_1}f\left(y_i\,|\,\mu_1,\sigma^2\right)}{rf\left(y_i\,|\,\mu_1,\sigma^2\right)+(1-r)\,f\left(y_i\,|\,\mu_2,\sigma^2\right)}\end{aligned}$$

حال با استفاده از تمرین ۸-۴ مشتق می‌گیریم و به شکل زیر تعریف می‌کنیم:

$$\begin{aligned}P_1\left(y_i\right)&=\frac{(1-r)\,f\left(y_i\,|\,\mu_1,\sigma^2\right)}{(1-r)\,f\left(y_i\,|\,\mu_1,\sigma^2\right)+rf\left(y_i\,|\,\mu_2,\sigma^2\right)}\\P_2\left(y_i\right)&=\frac{rf\left(y_i\,|\,\mu_1,\sigma^2\right)}{rf\left(y_i\,|\,\mu_1,\sigma^2\right)+(1-r)\,f\left(y_i\,|\,\mu_2,\sigma^2\right)}\end{aligned}\qquad(۸-۱۵)$$

تا عبارت زیر را به‌دست آوریم:

(۸-۱۶)

$$\frac{\partial}{\partial\mu_1}\log L\left(\mu_1,\mu_2,\sigma^2,r\,|\,y\right)=\sum_{i=1}^{n_1}P_1\left(y_i\right)\left(y_i-\mu_1\right)+\sum_{i=n_1+1}^{n}P_2\left(y_i\right)\left(y_i-\mu_1\right)$$

این معادله را برابر با صفر قرار می‌دهیم و μ_1 را به‌دست می‌آوریم:

(۱۷-۸)

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n_1} P_1(y_i)(y_i) + \sum_{i=n_1+1}^{n} P_2(y_i)(y_i)}{\sum_{i=1}^{n_1} P_1(y_i) + \sum_{i=n_1+1}^{n} P_2(y_i)}$$

به‌طور مشابه می‌توانیم برای μ_2 و σ^2 نیز معادله را حل کنیم و روابط زیر را به‌دست آوریم:

(۱۸-۸)

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n_1}\left[1-P_1(y_i)\right](y_i) + \sum_{i=n_1+1}^{n}\left[1-P_2(y_i)\right](y_i)}{\sum_{i=1}^{n_1}\left[1-P_1(y_i)\right] + \sum_{i=n_1+1}^{n}\left[1-P_2(y_i)\right]}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n_1}\left[P_1(y_i)(y_i-\hat{\mu}_1)^2 + \left(1-P_1(y_i)\right)(y_i-\hat{\mu}_2)^2\right]\right.$$

$$\left. + \sum_{i=n_1+1}^{n}\left[P_2(y_i)(y_i-\hat{\mu}_1)^2 + \left(1-P_2(y_i)\right)(y_i-\hat{\mu}_2)^2\right]\right)$$

البته معادلات (۸-۱۷) و (۸-۱۸) درستنمایی را برای تمامی پارامترهای آن حل نمی‌کنند، چون $p_1(y_i)$ و $p_2(y_i)$ به پارامترها و همچنین به r بستگی دارند. برای حل این مشکل، بخشی از داده‌های طرح را برای محاسبه پارامترها یعنی μ_1، μ_2 و σ^2 استفاده می‌کنیم و پس از برآورد μ_1، μ_2 و σ^2 از مقادیر به‌دست آمده برای برآورد مجدد r و سپس p_1 و p_2 استفاده می‌کنیم. این عمل را تا جایی تکرار می‌کنیم که مقادیر به یک همگرایی برسند.

مراحل فوق به‌طور اختصاصی عبارتند از:

۱- r را ثابت در نظر بگیرید.

۲- از معادلات (۸-۱۷) و (۸-۱۸) برای برآورد μ_1، μ_2 و σ^2 استفاده کنید.

۳- با استفاده از برآوردهای $\hat{\mu}_1$، $\hat{\mu}_2$ و $\hat{\sigma}^2$، عبارت زیر را به حداکثر برسانید:

$$\log L(\mu_1,\mu_2,\sigma^2,r\,|\,y) = \sum_{i=1}^{n_1}\log\left((1-r)\,f\left(y_i\,|\,\hat{\mu}_1,\hat{\sigma}^2\right) + rf\left(y_i\,|\,\hat{\mu}_2,\hat{\sigma}^2\right)\right)$$

$$+ \sum_{i=n_1+1}^{n}\log\left(rf\left(y_i\,|\,\hat{\mu}_1,\hat{\sigma}^2\right) + (1-r)\,f\left(y_i\,|\,\hat{\mu}_2,\hat{\sigma}^2\right)\right)$$

۴- تا رسیدن به همگرایی، مراحل ۲ و ۳ را تکرار کنید.

برای به حداکثر رساندن مرحله ۳، تمرین ۸-۷ را مشاهده کنید. اکنون در قالب مثال زیر به شرح مسئله می‌پردازیم:

مثال ۸-۶. (**MLE ها برای ارتفاع بوته گوجه فرنگی**). فرض کنید که داده‌های زیر (y) را در خصوص ارتفاع بوته (بر حسب سانتی‌متر) ۱۲ گیاه گوجه فرنگی از یک گونه خاص که به دو کلاس نشانگری گروه‌بندی شده‌اند، در اختیار داریم:

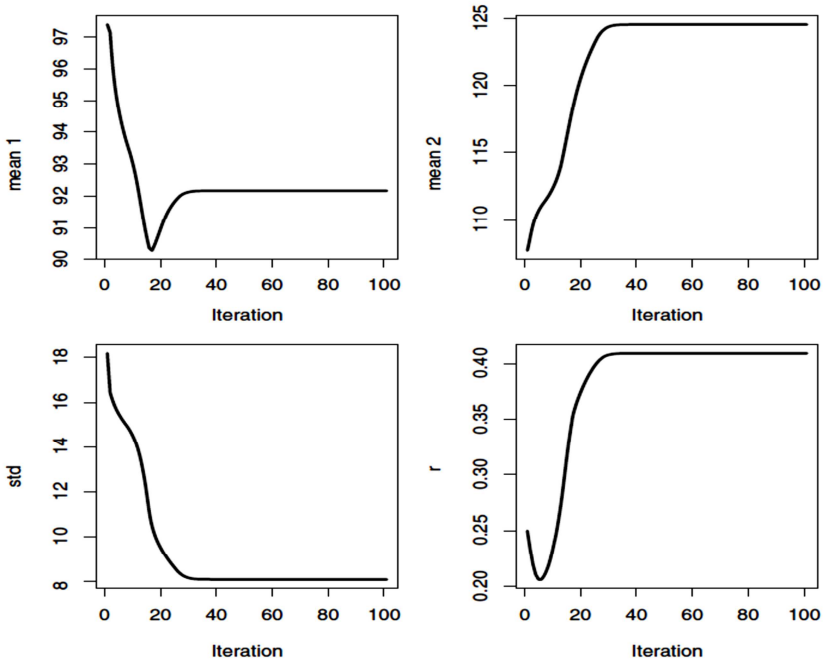
$$M_1M_1 : y = (79, 82, 100, 102, 124)$$
$$M_1M_2 : y = (85, 87, 101, 103, 125, 126, 127)$$

میانگین‌ها و واریانس‌های مشاهده شده کلاس‌های نشانگر به‌صورت $\bar{y}_{M_1M_1} = 97.4$، $\bar{y}_{M_1M_2} = 101.71$، $s_{M_1M_1} = 18.10$ و $s_{M_1M_2} = 18.33$ هستند. سپس MLE های مربوط به μ_1، μ_2، σ^2 و r را محاسبه می‌کنیم:

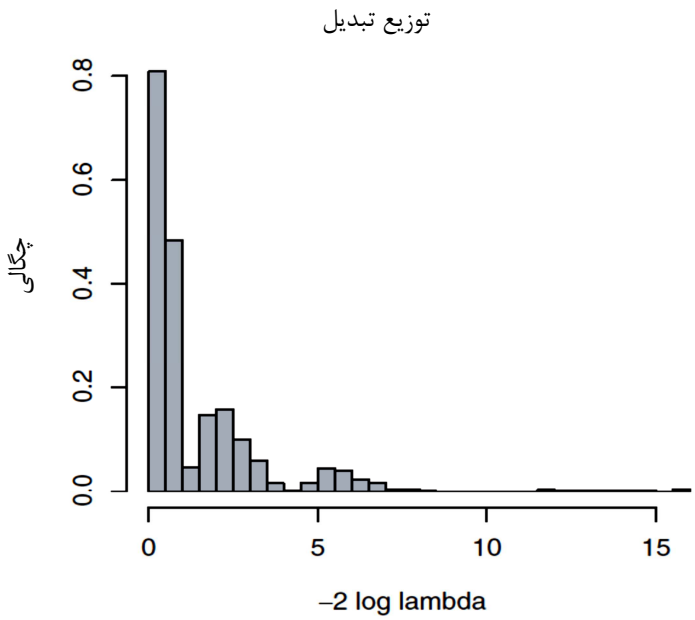
$$\hat{\mu}_1 = 92.169 \quad \hat{\mu}_2 = 124.561 \quad \hat{\sigma} = 8.08 \;\; \hat{r} = 0.409$$

شکل ۸-۲ را برای همگرایی برآوردها و ضمیمه B.2 را برای برنامه R مربوطه مشاهده کنید. آماره آزمون برابر با ۳/۵۳۹ = $-2\log\lambda$ محاسبه می‌شود.

برای ارزیابی معنی‌داری آزمون فرض، از آزمون تبدیل استفاده می‌کنیم. این آزمون برای ۵۰۰۰ نمونه اجرا و برای هر یک $-2\log\lambda$ محاسبه می‌شود. توزیع مربوطه در شکل ۸-۳ نشان داده شده است. از روی ۵۰۰۰ تبدیل، نقطه برش ۰/۹۵ برابر با ۵/۶۴۵ به‌دست می‌آید، بنابراین آماره معنی‌دار نبوده و پیوستگی نداریم.



شکل ۸-۲- MLE ها برای داده‌های گوجه فرنگی مثال ۸-۶.



شکل ۸-۳- توزیع تبدیل آماره نسبت درستنمایی $-2\log\lambda$ با فرض صفر بر پایه ۵۰۰۰ تبدیل مثال ۸-۶.

۸-۵- محاسبه توان آزمون

به‌طور عملی حائز اهمیت است که حداقل اندازه نمونه مورد نیاز برای دستیابی به سطح معنی‌دار از پیش تعیین شده محاسبه شود. محاسبه توان آزمون برای تشخیص QTL های معنی‌دار به نوع جمعیت مکان‌یابی بستگی دارد. برای مثال، تلاقی برگشتی می‌تواند فقط اثر افزایشی یک QTL را برآورد کند، درحالی‌که F_2 می‌تواند آثار افزایشی و غالبیت را برآورد نماید؛ بنابراین، باید راهکارهای متفاوتی برای محاسبه توان آزمون در این دو طرح مورد استفاده قرار گیرد. در این بخش، رویه کلی محاسبه توان تجزیه QTL در F_2 معرفی خواهد شد که از روی آن، خوانندگان علاقه‌مند می‌توانند طرح‌های پیچیده‌تر را مدنظر قرار دهند.

فرض کنید یک جمعیت F_2 وجود دارد که در آن نشانگرهای منفرد تعیین ژنوتیپ شده‌اند تا ارتباط آن‌ها با QTL کنترل‌کننده صفت کمّی شناسایی شود. سه ژنوتیپ نشانگر بر اساس نسبت $1(AA):2(Aa):1(aa)$ تفرق می‌یابند که نشان می‌دهد اندازه نمونه n به‌صورت $\frac{n}{4}$، $\frac{n}{2}$ و $\frac{n}{4}$ به سه ژنوتیپ اختصاص یافته است. بر اساس تجزیه‌های بخش ۸-۳-۳، تفاوت بین دو ژنوتیپ هموزیگوت نشانگر می‌تواند بازتابی از اثر افزایشی یک QTL مفروض باشد، اگرچه اثر QTL با مکان QTL اختلاط یافته است. مجدداً معادله (۸-۶) را برای آزمون اثر افزایشی QTL بدین صورت می‌نویسیم:

$$t_1 = \frac{m_2 - m_0}{\sqrt{s^2\left(\frac{4}{n}+\frac{4}{n}\right)}} = \frac{(1-2r)2a}{\sqrt{\frac{8s^2}{n}}}$$

زمانی که n بزرگ است، آماره آزمون محاسبه شده یعنی $\hat{t}_1$، از یک توزیع تقریباً نرمال به صورت $\hat{t}_1 \sim N(t_1,1)$ تبعیت می‌کند؛ بنابراین، توان تشخیص تفاوت $(m_2 - m_0)$ برای یک آزمون دو دنباله بدین صورت بیان می‌شود:

$$1-\beta = \Pr ob\left(\hat{t}_1 > z_{\frac{\alpha}{2}}\right) = 1-\Phi\left(z_\alpha - t\right) \qquad (۱۹-۸)$$

که z_α مقدار بحرانی آزمون با اطمینان $(1-\alpha)$ و با فرض صفر $t_1 = ۰$ است و $\Phi\left(z_\alpha - t\right)$ تابع توزیع تجمعی نرمال استاندارد است.

برای خطای نوع اول (α) و نوع دوم (β) معین در آزمون t، اندازه نمونه n برای تشخیص اثر افزایشی QTL با رابطه زیر تعیین می‌شود:

$$n = 8\left[\frac{z_{\frac{\alpha}{2}} + z_{\beta}}{\frac{(1-2r)2a}{s}}\right]^2 \qquad (۲۰-۸)$$

اگر جستجوی QTL در سرتاسر ژنوم صورت گیرد، خطای نوع اول (α) برای هر آزمون اساساً باید کمتر از مقدار افزایش احتمال مثبت کاذب[1] در آزمون کل ژنوم باشد.

به‌طور مشابه، آماره آزمون اثر غالبیت بر اساس رابطه (۸-۷) به صورت زیر است:

$$t_2 = \frac{\frac{m_1 - (m_2 + m_0)}{2}}{\sqrt{s^2\left(\frac{2}{n} + \frac{1}{n} + \frac{1}{n}\right)}}$$

از روی این رابطه، اندازه نمونه مورد نیاز برای شناسایی اثر غالبیت به صورت زیر تعیین می‌شود:

(۲۱-۸)

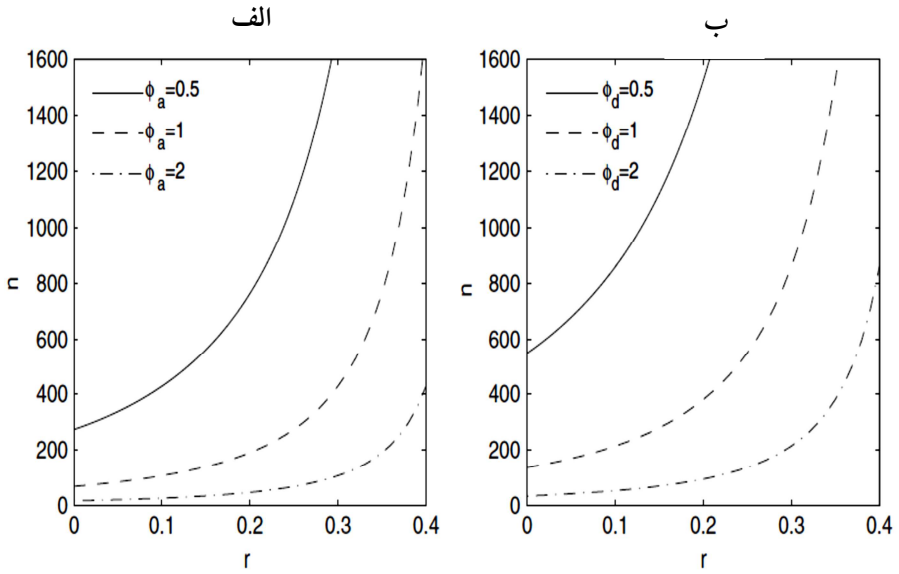
$$n = 4\left[\frac{z_{\frac{\alpha}{2}} + z_{\beta}}{\frac{(1-2r)d}{s}}\right]^2$$

از روی معادلات (۸-۲۰) و (۸-۲۱) می‌توان مشاهده کرد که اندازه نمونه برای شناسایی یک QTL به عوامل زیادی بستگی دارد که عبارتند از: ۱- میزان بزرگی اثر (افزایشی یا غالبیت) یک QTL (هر چه این اثر بزرگ‌تر باشد اندازه نمونه مورد نیاز کوچک‌تر است)، ۲- درجه پیوستگی بین نشانگر مورد نظر و QTL مفروض (هرچه پیوستگی قوی‌تر باشد، اندازه نمونه مورد نیاز کوچک‌تر است) و ۳- واریانس باقیمانده درون یک کلاس ژنوتیپی QTL (هرچه واریانس باقیمانده کوچک‌تر باشد اندازه نمونه مورد نیاز کوچک‌تر است).

مثال ۸-۷. یک متخصص ژنتیک موش قصد دارد تا یک مطالعه مکان‌یابی مولکولی برای شناسایی QTL های مؤثر بر وزن بدن موش انجام دهد. اگرچه، اثر دقیق QTL ناشناخته است، اما $\Phi_a = \frac{a}{s}$ یا $\Phi_d = \frac{d}{s}$ برای تعریف آثار ژنتیکی نسبت به انحراف استاندارد باقیمانده استفاده می‌شود. فرض کنید سطح معنی‌داری برای شناسایی QTL برابر با ۰/۰۱ = α باشد. انتظار می‌رود

[1] - False-positive

توان ۰/۹۰ = $(1-\beta)$ برای آزمایش QTL به‌دست آید. اندازه‌های نمونه مورد نیاز برای دامنه‌ای از نرخ‌های نوترکیبی به ترتیب در سطوح مختلف آثار افزایشی و غالبیت QTL به ترتیب با استفاده از معادلات (۸-۲۰) و (۸-۲۱) محاسبه می‌شوند (شکل‌های ۸-۴ الف و ۸-۴ ب).



شکل ۸-۴- اندازه‌های نمونه مورد نیاز در نرخ‌های نوترکیبی مختلف، الف. اثر افزایشی، ب. اثر غالبیت.

فقط در شرایطی که QTL دقیقاً در جایگاه نشانگر قرار داشته باشد، یک اندازه نمونه کوچک برای QTL بزرگ اثر با $(\Phi_a, \Phi_d = 2)$ مورد نیاز است. اندازه نمونه مورد نیاز به‌صورت نمایی با نرخ نوترکیبی بین نشانگر و QTL افزایش می‌یابد که میزان افزایش به‌صورت چشمگیری برای QTL های کوچک اثر یعنی $(\Phi_a, \Phi_d = 0.5)$ بزرگ‌تر از QTL های بزرگ اثر یعنی $(\Phi_a, \Phi_d = 2)$ است. تحت شرایطی که سطح پیوستگی بین نشانگر و QTL مشابه باشد، اندازه نمونه مورد نیاز برای شناسایی اثر غالبیت، بزرگ‌تر از اندازه نمونه مورد نیاز برای شناسایی اثر افزایشی است.

۸-۶- تجزیه اثر متقابل نشانگر

۸-۶-۱- تجزیه واریانس (ANOVA)

اثر متقابل ژنتیکی بین QTL های مختلف نقش مهمی در کنترل و بیان صفت ایفا می‌کند. تجزیه واریانس دو طرفه بر پایه نشانگرهای مختلف می‌تواند برای برآورد و آزمون اثر اصلی نشانگرها و آثار

متقابل بین آن‌ها انجام گیرد (روتمن و چورود، ۱۹۹۷). یک جمعیت مکان‌یابی را در نظر بگیرید که در آن، l ژنوتیپ در هر نشانگر وجود دارد. روش استاندارد تجزیه واریانس دوطرفه برای تجزیه همزمان دو نشانگر مورد استفاده قرار می‌گیرد که نتایج آن در جدول ۸-۴ خلاصه شده‌اند.

بر پایه ساختار امید ریاضی میانگین مربعات ناشی از آثار اصلی و متقابل (با فرض این‌که همه آن‌ها تصادفی هستند)، می‌توانیم آماره آزمون F را برای هر اثر به ترتیب در مقایسه با مقادیر بحرانی $F_{0.05,(l-1,(l-1)^2)}$ و $F_{0.05,((l-1)^2,n-l^2-1)}$ محاسبه کنیم. واریانس‌های ژنتیکی برای اثر اصلی نشانگرهای A و B و اثر متقابل بین آن‌ها بدین صورت محاسبه می‌شود:

$$\sigma_A^2 = \frac{MS_1 - MS_3}{k_1}$$

$$\sigma_B^2 = \frac{MS_2 - MS_3}{k_2}$$

$$\sigma_{AB}^2 = \frac{MS_3 - MS_4}{k_3}$$

که میانگین‌های هارمونیک به صورت متفاوتی برای تلاقی برگشتی و F_2 محاسبه می‌شوند.

جدول ۸-۴- خلاصه تجزیه واریانس دوطرفه به‌هدف تشخیص آثاراصلی و متقابل بین نشانگرهای مختلف در یک جمعیت مکان‌یابی.

منبع تغییرات	درجه آزادی	میانگین مربعات	آماره F	امید ریاضی میانگین مربعات
اثر اصلی ناشی از نشانگر A	$l-1$	MS_1	$\frac{MS_1}{MS_3}$	$\sigma_e^2 + k_3\sigma_{AB}^2 + k_1\sigma_A^2$
اثر اصلی ناشی از نشانگر B	$l-1$	MS_2	$\frac{MS_2}{MS_3}$	$\sigma_e^2 + k_3\sigma_{AB}^2 + k_2\sigma_B^2$
اثر متقابل	$(l-1)^2$	MS_3	$\frac{MS_3}{MS_4}$	$\sigma_e^2 + k_3\sigma_{AB}^2$
خطای باقیمانده	$n-l^2-1$	MS_4		σ_e^2

برای جمعیت تلاقی برگشتی داریم:

$$k_1 = \frac{2}{\frac{1}{n_1} + \frac{1}{n_0}}$$

که n_1 و n_0 مشاهدات دو ژنوتیپ برای نشانگر A هستند:

$$k_2 = \frac{2}{\frac{1}{n_1'} + \frac{1}{n_0'}}$$

که n_1' و n_0' مشاهدات دو ژنوتیپ برای نشانگر B هستند و:

$$k_3 = \frac{4}{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}$$

که در آن، n_{11}، n_{10}، n_{01} و n_{00} مشاهدات چهار ژنوتیپ برای نشانگرهای A و B هستند. برای جمعیت F_2 نیز داریم:

$$k_1 = \frac{3}{\frac{1}{n_2} + \frac{1}{n_1} + \frac{1}{n_0}}$$

که n_2، n_1 و n_0 مشاهدات سه ژنوتیپ برای نشانگر A هستند.

$$k_2 = \frac{3}{\frac{1}{n_2'} + \frac{1}{n_1'} + \frac{1}{n_0'}}$$

که n_2'، n_1' و n_0' نیز مشاهدات سه ژنوتیپ برای نشانگر B هستند؛ و نیز:

$$k_3 = \frac{9}{\frac{1}{n_{22}} + \frac{1}{n_{21}} + \frac{1}{n_{20}} + \frac{1}{n_{12}} + \frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{02}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}$$

که در آن، $n_{22},\ldots,n_{00}$ نیز مشاهدات ۹ ژنوتیپ برای نشانگرهای A و B هستند.

نسبت هر یک از این واریانس‌های ژنتیکی بر واریانس فنوتیپی، یعنی وراثت‌پذیری خاص نشانگری می‌تواند با رابطه زیر محاسبه شود:

$$R_A^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_e^2}$$

$$R_B^2 = \frac{\sigma_B^2}{\sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_e^2} \qquad (۲۲-۸)$$

$$R_{AB}^2 = \frac{\sigma_{AB}^2}{\sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_e^2}$$

برای F_2، اثر اصلی برای هر نشانگر، A و B، می‌تواند به مؤلفه‌های اثر افزایشی (a) و غالبیت (d) تقسیم شود، درحالی‌که اثر متقابل بین نشانگرهای مختلف در F_2 می‌تواند به اثر افزایشی× اثر افزایشی (I_{aa})، اثر افزایشی × اثر غالبیت (I_{ad})، اثر غالبیت× اثر افزایشی (I_{da}) و اثر غالبیت× اثر غالبیت (I_{dd}) تفکیک شود. تمامی این آثار می‌توانند با یک آماره آزمون t آزمون شوند که به همراه مقادیر بحرانی مربوطه به ترتیب به‌صورت زیر هستند:

$$t_{Aa}=\frac{\left(m_{22}+m_{20}\right)-\left(m_{02}+m_{00}\right)}{\sqrt{s^2\left(\frac{1}{n_{22}}+\frac{1}{n_{20}}+\frac{1}{n_{02}}+\frac{1}{n_{00}}\right)}},t_{(0.025,\nu=n_{22}+n_{20}+n_{02}+n_{00}-4)} \qquad (۲۳-۸)$$

که در آن:

(۲۴-۸)

$$s^2=\frac{\left(n_{22}-1\right)s_{22}^2+\left(n_{20}-1\right)s_{20}^2+\left(n_{02}-1\right)s_{02}^2+\left(n_{00}-1\right)s_{00}^2}{n_{22}+n_{20}+n_{02}+n_{00}-4}$$

$$t_{Ad}=\frac{\left(m_{12}+m_{10}\right)-\frac{1}{2}\left(m_{22}+m_{20}+m_{02}+m_{00}\right)}{\sqrt{s^2\left(\frac{1}{4n_{22}}+\frac{1}{4n_{20}}+\frac{1}{n_{12}}+\frac{1}{n_{10}}+\frac{1}{4n_{02}}+\frac{1}{4n_{00}}\right)}},t_{(0.025,\nu=n_{22}+n_{20}+n_{12}+n_{10}+n_{02}+n_{00}-6)}$$

که s^2 برابر است با:

$$s^2=\frac{\left(n_{22}-1\right)s_{22}^2+\left(n_{20}-1\right)s_{20}^2+\left(n_{12}-1\right)s_{12}^2}{n_{22}+n_{20}+n_{12}+n_{10}+n_{02}+n_{00}-6}$$

$$t_{Ba}=\frac{\left(m_{22}+m_{02}\right)-\left(m_{20}+m_{00}\right)}{\sqrt{s^2\left(\frac{1}{n_{22}}+\frac{1}{n_{20}}+\frac{1}{n_{02}}+\frac{1}{n_{00}}\right)}},t_{(0.025,\nu=n_{22}+n_{20}+n_{02}+n_{00}-4)} \qquad (۲۵-۸)$$

که در آن:

$$s^2=\frac{\left(n_{22}-1\right)s_{22}^2+\left(n_{20}-1\right)s_{20}^2+\left(n_{02}-1\right)s_{02}^2+\left(n_{00}-1\right)s_{00}^2}{n_{22}+n_{20}+n_{02}+n_{00}-4}$$

(۸-۲۶)

$$t_{Bd}=\frac{\left(m_{21}+m_{01}\right)-\frac{1}{2}\left(m_{22}+m_{20}+m_{02}+m_{00}\right)}{\sqrt{s^2\left(\frac{1}{4n_{22}}+\frac{1}{4n_{20}}+\frac{1}{n_{21}}+\frac{1}{n_{01}}+\frac{1}{4n_{02}}+\frac{1}{4n_{00}}\right)}},t_{\left(0.025,\nu=n_{22}+n_{20}+n_{12}+n_{10}+n_{02}+n_{00}-6\right)}$$

و داریم:

$$s^2=\frac{\left(n_{22}-1\right)s_{22}^2+\left(n_{21}-1\right)s_{21}^2+\left(n_{20}-1\right)s_{20}^2}{n_{22}+n_{21}+n_{20}+n_{02}+n_{01}+n_{00}-6}+\frac{\left(n_{02}-1\right)s_{02}^2+\left(n_{01}-1\right)s_{01}^2+\left(n_{00}-1\right)s_{00}^2}{n_{22}+n_{21}+n_{20}+n_{02}+n_{01}+n_{00}-6}$$

$$t_{aa}=\frac{\left(m_{22}+m_{00}\right)-\left(m_{20}+m_{02}\right)}{\sqrt{s^2\left(\frac{1}{n_{22}}+\frac{1}{n_{20}}+\frac{1}{n_{02}}+\frac{1}{n_{00}}\right)}},t_{\left(0.025,\nu=n_{22}+n_{20}+n_{02}+n_{00}-4\right)} \qquad (۸-۲۷)$$

و s^2 برابر است با:

$$s^2=\frac{\left(n_{22}-1\right)s_{22}^2+\left(n_{20}-1\right)s_{20}^2+\left(n_{02}-1\right)s_{02}^2+\left(n_{00}-1\right)s_{00}^2}{n_{22}+n_{20}+n_{02}+n_{00}-4}$$

(۸-۲۸)

$$t_{ad}=\frac{\left[m_{21}-\frac{1}{2}\left(m_{22}+m_{20}\right)\right]-\left[m_{01}-\frac{1}{2}\left(m_{02}+m_{00}\right)\right]}{\sqrt{s^2\left(\frac{1}{4n_{22}}+\frac{1}{n_{21}}+\frac{1}{4n_{20}}+\frac{1}{4n_{02}}+\frac{1}{n_{01}}+\frac{1}{4n_{00}}\right)}},t_{\left(0.025,\nu=n_{22}+n_{21}+n_{20}+n_{02}+n_{01}+n_{00}-6\right)}$$

که در آن:

$$s^2=\frac{\left(n_{22}-1\right)s_{22}^2+\left(n_{21}-1\right)s_{21}^2+\left(n_{20}-1\right)s_{20}^2}{n_{22}+n_{21}+n_{20}+n_{02}+n_{01}+n_{00}-6}+\frac{\left(n_{02}-1\right)s_{02}^2+\left(n_{01}-1\right)s_{01}^2+\left(n_{00}-1\right)s_{00}^2}{n_{22}+n_{21}+n_{20}+n_{02}+n_{01}+n_{00}-6}$$

(۸-۲۹)

$$t_{da}=\frac{\left[m_{12}-\frac{1}{2}\left(m_{22}+m_{02}\right)\right]-\left[m_{10}-\frac{1}{2}\left(m_{20}+m_{00}\right)\right]}{\sqrt{s^2\left(\frac{1}{4n_{22}}+\frac{1}{4n_{20}}+\frac{1}{n_{12}}+\frac{1}{n_{10}}+\frac{1}{4n_{02}}+\frac{1}{4n_{00}}\right)}},t_{\left(0.025,\nu=n_{22}+n_{20}+n_{12}+n_{10}+n_{02}+n_{00}-6\right)}$$

که در آن:

$$s^2=\frac{(n_{22}-1)s_{22}^2+(n_{20}-1)s_{20}^2+(n_{12}-1)s_{12}^2}{n_{22}+n_{20}+n_{12}+n_{10}+n_{02}+n_{00}-6}+\frac{(n_{10}-1)s_{10}^2+(n_{02}-1)s_{02}^2+(n_{00}-1)s_{00}^2}{n_{22}+n_{20}+n_{12}+n_{10}+n_{02}+n_{00}-6}$$

(۳۰-۸)

و

$$t_{dt}=\frac{\left[m_{11}-\frac{1}{2}(m_{21}+m_{01})\right]-\frac{1}{2}\left\{\left[m_{12}-\frac{1}{2}(m_{22}+m_{02})\right]+\left[m_{10}-\frac{1}{2}(m_{20}+m_{00})\right]\right\}}{\sqrt{s^2\left(\frac{1}{4n_{22}}+\frac{1}{n_{12}}+\frac{1}{4n_{02}}+\frac{1}{4n_{21}}+\frac{1}{n_{11}}+\frac{1}{4n_{01}}+\frac{1}{4n_{20}}+\frac{1}{n_{10}}+\frac{1}{4n_{00}}\right)}},t_{(0.025,\nu=n_{22}+n_{12}+n_{02}+n_{21}+n_{11}+n_{01}+n_{20}+n_{10}+n_{00}-9)}$$

s^2 برابر است با:

$$s^2=\frac{(n_{22}-1)s_{22}^2+(n_{12}-1)s_{12}^2+(n_{02}-1)s_{02}^2+(n_{21}-1)s_{21}^2+(n_{11}-1)s_{11}^2}{n_{22}+n_{12}+n_{02}+n_{21}+n_{11}+n_{01}+n_{20}+n_{10}+n_{00}-9}$$
$$+\frac{(n_{01}-1)s_{01}^2+(n_{20}-1)s_{20}^2+(n_{10}-1)s_{10}^2+(n_{00}-1)s_{00}^2}{n_{22}+n_{12}+n_{02}+n_{21}+n_{11}+n_{01}+n_{20}+n_{10}+n_{00}-9}$$

$s_{22}^2,\ldots,s_{00}^2$ واریانس‌های نمونه‌ای برای ۹ ژنوتیپ در نشانگرهای A و B هستند.

مثال ۸-۸. مثال ۳-۲ را مجدداً مشاهده کنید. یک جمعیت F_2 حاصل از تلاقی بین نژادهای اینبرد (خالص) بزرگ (LG/J) و کوچک (SM/J) موش برای توصیف آثار متقابل بین نشانگرها بر وزن بدن بالغ (چورود و همکاران، ۱۹۹۶) استفاده شده است. وزن ۵۳۵ موش F_2 به صورت هفتگی از هفته اول تا سن ۱۰ هفتگی اندازه‌گیری شد. وزن ۱۰ هفتگی که وزن بلوغ نامیده می‌شود، تجزیه خواهد شد. ۷۶ ریزماهواره چندشکل در سرتاسر ۱۹ کروموزوم اتوزوم موش امتیازبندی شدند. به عنوان مثال، تجزیه را روی یک نشانگر (D1Mit7) موجود روی کروموزوم ۱ و نشانگر دوم (D2Mit17) موجود روی کروموزوم ۲ متمرکز خواهیم کرد. برآوردهای مربوط به میانگین مربعات، مقادیر F، اجزای واریانس ژنتیکی و سهم هر اثر در توجیه واریانس فنوتیپی کل در جدول ۸-۵ ارایه شده است.

جدول ۸-۵- نتایج تجزیه واریانس دو نشانگری در یک جمعیت F_2 موش.

منبع تغییر	درجه آزادی	میانگین مربعات	مقدار F	Pr>F	سهم هر اثر
اثر اصلی ناشی از نشانگر D1Mit7	۲	۱۱۵/۲۰	۷/۶۱	<۰/۰۰۰۱	۰/۰۲۶
اثر اصلی ناشی از نشانگر D2Mit17	۲	۸۷/۴۰	۵/۷۷	۰/۰۰۶۵	۰/۰۱۹
اثر متقابل	۴	۱۲/۲۵	۰/۸۱	۰/۵۱۹۵	
خطای باقیمانده	۸۳۸	۱۵/۱۴			

هر دو نشانگر D1Mit7 و D2Mit17 برای آثار اصلی مربوطه روی وزن بدن بالغ موش‌ها که به ترتیب ۲/۶ و ۱/۹ درصد از واریانس فنوتیپی را توجیه می‌کنند معنی‌دار هستند. اثر متقابل بین دو نشانگر معنی‌دار نشد. بر اساس معادلات (۸-۲۳) و (۸-۲۴)، آماره t برای آثار افزایشی و غالبیت مربوط به نشانگر D1Mit7 به ترتیب t_{Aa} = −۱/۷۸۶۶ (p = ۰/۰۷۵۶) و t_{Ad} = ۱/۷۲۳۲ (p = ۰/۰۸۵۷) محاسبه شد. این دو اثر در مقایسه با مقادیر بحرانی مربوطه یعنی $t_{(0.025,\nu=186)}$ = ۱/۹۷۲۸ و $t_{(0.025,\nu=378)}$ = ۱/۹۶۶۳ معنی‌دار نیستند. آماره‌های آزمون t برای آثار افزایشی و غالبیت نشانگر D2Mit17 نیز از راه معادلات (۸-۲۵) و (۸-۲۶) به ترتیب برابر با t_{Ba} = −۳/۱۶۹۳ (p = ۰/۰۰۱۸) و t_{Bd} = −۰/۲۳۴۹ (p = ۰/۸۱۴۴) محاسبه شدند که نشان می‌دهد اثر افزایشی با توجه به مقدار بحرانی $t_{(0.025,\nu=186)}$ = ۱/۹۷۲۸ معنی‌دار و اثر غالبیت با توجه به مقدار بحرانی $t_{(0.025,\nu=378)}$ = ۱/۹۶۶۳ معنی‌دار نیست.

۸-۶-۲- تجزیه ژنتیکی

اثر اصلی نشانگر فقط بازتابی از اثر افزایشی نشانگر برای یک جمعیت تلاقی برگشتی است، ولی برای یک جمعیت F_2 بیانگر هر دو اثر افزایشی و غالبیت نشانگر است. اثر متقابل بین دو نشانگر مختلف در تلاقی برگشتی نیز در حقیقت نشان‌دهنده اثر ژنتیکی افزایشی× افزایشی در سطح نشانگر است که در F_2 می‌تواند به چهار مؤلفه افزایشی× افزایشی، افزایشی× غالبیت، غالبیت× افزایشی و غالبیت× غالبیت در سطح نشانگر تقسیم شود. در عمل، برآورد هر یک از این مؤلفه‌های اثر متقابل حائز اهمیت فراوان است (چورود و روتمن، ۱۹۹۵).

همان‌گونه که برای حالت تک نشانگری نشان داده شد، تجزیه نشانگر با موقعیت و اثر QTL اختلاط می‌یابد. این حالت، برای تجزیه دو نشانگری نیز صحیح است، ولی توضیح این‌که چگونه موقعیت‌های QTL و عوامل ژنتیکی مختلف اختلاط می‌یابند، جالب توجه است.

دو نشانگر غیر پیوسته A (با آلل‌های A و a) و B (با آلل‌های B و b) را در نظر بگیرید. فرض کنید که هر نشانگر با یک QTL مختلف، یعنی P (با آلل‌های P و p) یا Q (با آلل‌های Q و q) به ترتیب با نرخ‌های نوترکیبی r_1 و r_2 پیوسته باشند. فرض می‌شود $\omega_{j_1 j_2|k_1 k_2}$ احتمال شرطی دو ژنوتیپ QTL به صورت $j_1 j_2$ مشروط به دو ژنوتیپ نشانگری به صورت $k_1 k_2$ باشند. اگر این دو نشانگر روی کروموزوم‌های متفاوتی قرار گرفته باشند، احتمالات شرطی $\omega_{j_1 j_2|k_1 k_2}$ می‌توانند به صورت زیر بیان شوند:

ژنوتیپ نشانگر	QTL P	
	Qq(1)	qq(0)
Bb(1)	$1-r_2$	r_2
bb(0)	r_2	$1-r_2$

$\otimes$

نشانگر M	QTL P	
	Pp(0)	Pp(0)
Aa(1)	$1-r_1$	r_1
Aa(0)	r_1	$1-r_1$

برای تلاقی برگشتی و:

ژنوتیپ نشانگر	QTL P		
	QQ(2)	Qq(1)	qq(0)
BB(2)	$(1-r_2)^2$	$2r_2(1-r_2)$	r_2^2
Bb(1)	$r_2(1-r_2)$	$(1-r_2)^2+r_2^2$	$r_2(1-r_2)$
bb(0)	r_2^2	$2r_2(1-r_2)$	$(1-r_2)^2$

$\otimes$

نشانگر M	QTL P		
	PP(2)	Pp(1)	Pp(0)
AA(2)	$(1-r_1)^2$	$2r_1(1-r_1)$	r_1^2
Aa(1)	$r_1(1-r_1)$	$(1-r_1)^2+r_1^2$	$r_1(1-r_1)$
Aa(0)	r_1^2	$2r_1(1-r_1)$	$(1-r_1)^2$

فرض کنید $\mu_{j_1 j_2}$ (j_1, j_2 = ۰ و ۱ برای تلاقی برگشتی و ۲ و ۱ و ۰ برای F_2) و $m_{k_1 k_2}$ ($k_1 k_2$ = ۰ و ۱ برای تلاقی برگشتی و ۲ و ۱ و ۰ برای F_2) به ترتیب ارزش‌های ژنوتیپی دو QTL و دو نشانگر باشند. ارزش‌های ژنوتیپی QTL می‌توانند به مؤلفه‌های متفاوتی تقسیم شوند که عبارتند از: میانگین کل (μ)، آثار افزایشی QTL های P (a_1) و Q (a_2) و اثر افزایشی× اثر افزایشی (i_{aa}) برای تلاقی برگشتی یا میانگین کل (μ)، آثار افزایشی (a_1) و غالبیت (d_1) در QTL P، آثار افزایشی (a_2) و غالبیت (d_2) در QTL Q و همچنین اثر افزایشی× اثر افزایشی (i_{aa})،

اثر افزایشی × اثر غالبیت (i_{ad})، اثر غالبیت× اثر افزایشی (i_{da}) و اثر غالبیت× اثر غالبیت (i_{dd}) برای F_2 (معادلات (۱-۹) و ماتریس (۹-۹) را مشاهده کنید).

همچنین از یک فرمول عمومی برای بیان ارزش‌های ژنوتیپی نشانگر بر حسب ارزش‌های ژنوتیپی QTL استفاده می‌کنیم:

$$m_{k_1k_2} = \sum_{j_1=0}^{2}\sum_{j_2=0}^{2} \omega_{j_1j_2|k_1k_2}\mu_{j_1j_2} \qquad (۳۱-۸)$$

به‌صورت اختصاصی‌تر، بر اساس معادله (۸-۳۱)، ارزش‌های ژنوتیپی نشانگر را بر حسب آثار انفرادی QTL برای تلاقی برگشتی به صورت زیر به‌دست می‌آوریم:

$$m_{11} = \mu + (1-r_1)a_1 + (1-r_2)a_2 + (1-r_1)(1-r_2)i_{aa}$$
$$m_{10} = \mu + (1-r_1)a_1 + r_2a_2 + (1-r_1)r_2i_{aa}$$
$$m_{01} = \mu + r_1a_1 + (1-r_2)a_2 + r_1(1-r_2)i_{aa}$$
$$m_{00} = \mu + r_1a_1 + r_2a_2 + r_1r_2i_{aa}$$

آثار اصلی نشانگرهای A و B به ترتیب بدین صورت بیان می‌شوند:

$$M_A = \frac{1}{2}(m_{11}+m_{10}) - \frac{1}{2}(m_{01}+m_{00}) = (1-2r_1)a_1 + \frac{1}{2}(1-r_1)i_{aa} \qquad (۳۲-۸)$$

$$M_B = \frac{1}{2}(m_{11}+m_{01}) - \frac{1}{2}(m_{10}+m_{00}) = (1-2r_2)a_2 + \frac{1}{2}(1-r_2)i_{aa} \qquad (۳۳-۸)$$

حال آنکه اثر متقابل بین دو نشانگر بدین صورت بیان می‌شود:

$$I_{AB} = (m_{11}+m_{00}) - (m_{10}+m_{01}) = (1-2r_1)(1-2r_2)i_{aa} \qquad (۳۴-۸)$$

جالب است بدانیم که برآورد اثر اصلی یک نشانگر در تجزیه دو نشانگری نه فقط با موقعیت یک QTL پیوسته با نشانگر اختلاط می‌یابد، بلکه تحت تأثیر اثر اپیستازی بین این QTL و هر QTL دیگر مستقر در یک موقعیت متفاوت نیز قرار می‌گیرد. برآورد اثر متقابل بر اساس اطلاعات نشانگر به وسیله اثر اپیستازی افزایشی× افزایشی بین دو QTL فرضی و نرخ نوترکیبی بین آن‌ها با نشانگرها اختلاط می‌یابد.

آثار اصلی و متقابل دو QTL مختلف در F_2 که با معادله (۱-۹) نشان داده شده و به شکل جدول در ماتریس (۹-۹) ارایه شده است، اجازه می‌دهند مشخصات آثار افزایشی و غالبیت و تمامی چهار اثر ممکن اپیستازی به‌دست آیند. با استفاده از معادله (۸-۳۱)، ارزش‌های ژنوتیپی نشانگر بر حسب آثار QTL برای F_2 ها به‌صورت زیر محاسبه می‌شود:

$$m_{22} = \mu + (1-2r_1)a_1 + 2r_1(1-r_1)d_1 + (1-2r_2)a_2 + 2r_2(1-r_2)d_2$$
$$+(1-2r_1)(1-2r_2)i_{aa} + 2r_2(1-r_2)(1-2r_1)i_{ad}$$
$$+2r_1(1-2r_2)(1-r_1)i_{da} + 4r_1(1-r_1)r_2(1-r_2)i_{dd}$$
$$m_{21} = \mu + (1-2r_1)a_1 + 2r_1(1-r_1)d_1 + (1-2r_2+2r_2^2)d_2$$
$$+(1-2r_2+2r_2^2)(1-2r_1)i_{ad} + 2r_1(1-r_1)(1-2r_2+2r_2^2)i_{dd}$$
$$m_{20} = \mu + (1-2r_1)a_1 + 2r_1(1-r_1)d_1 - (1-2r_2)a_2 + 2r_2(1-r_2)d_2$$
$$-(1-2r_1)(1-2r_2)i_{aa} + 2r_2(1-r_2)(1-2r_1)i_{ad}$$
$$-2r_1(1-2r_2)(1-r_1)i_{da} + 4r_1(1-r_1)r_2(1-r_2)i_{dd}$$
$$m_{12} = \mu + (1-2r_1+2r_1^2)d_1 + (1-2r_2)a_2 + 2r_2(1-r_2)d_2$$
$$(1-2r_1+2r_1^2)(1-2r_2)i_{da} + 2(1-2r_1+2r_1^2)r_2(1-r_2)i_{dd}$$
$$m_{11} = \mu + (1-2r_1+2r_1^2)d_1 + (1-2r_2+2r_2^2)d_2$$
$$+(1-2r_1+2r_1^2)(1-2r_2+2r_2^2)i_{dd}$$
$$m_{10} = \mu + (1-2r_1+2r_1^2)d_1 - (1-2r_2)a_2 + 2r_2(1-r_2)d_2$$
$$-(1-2r_1+2r_1^2)(1-2r_2)i_{da} + 2(1-2r_1+2r_1^2)r_2(1-r_2)i_{dd}$$
$$m_{02} = \mu - (1-2r_1)a_1 + 2r_1(1-r_1)d_1 + (1-2r_2)a_2 + 2r_2(1-r_2)d_2$$
$$-(1-2r_1)(1-2r_2)i_{aa} - 2r_2(1-r_2)(1-2r_1)i_{ad}$$
$$+2r_1(1-2r_2)(1-r_1)i_{da} + 4r_1(1-r_1)r_2(1-r_2)i_{dd}$$
$$m_{01} = \mu - (1-2r_1)a_1 + 2r_1(1-r_1)d_1 + (1-2r_2+2r_2^2)d_2$$
$$-(1-2r_2+2r_2^2)(1-2r_1)i_{ad} + 2r_1(1-r_1)(1-2r_2+2r_2^2)i_{dd}$$
$$m_{00} = \mu - (1-2r_1)a_1 + 2r_1(1-r_1)d_1 - (1-2r_2)a_2 + 2r_2(1-r_2)d_2$$
$$+(1-2r_1)(1-2r_2)i_{aa} - 2r_2(1-r_2)(1-2r_1)i_{ad}$$
$$-2r_1(1-2r_2)(1-r_1)i_{da} + 4r_1(1-r_1)r_2(1-r_2)i_{dd}$$

آثار افزایشی و غالبیت نشانگرهای A و B نیز به ترتیب بدین صورت به‌دست می‌آیند:

$$A_A=\frac{1}{4}(m_{22}+m_{20})-\frac{1}{4}(m_{02}+m_{00})=(1-2r_1)a_1+2(1-2r_1)r_2(1-r_2)i_{ad}$$

$$D_A=\frac{1}{2}(m_{12}+m_{10})-\frac{1}{4}(m_{22}+m_{20}+m_{02}+m_{00})=(1-2r_1)^2d_1+2(1-2r_1)^2r_2(1-r_2)i_{dd}$$

$$A_B=\frac{1}{4}(m_{22}+m_{02})-\frac{1}{4}(m_{20}+m_{00})=(1-2r_2)a_2+2r_1(1-r_1)(1-2r_2)i_{da}$$

$$D_B=\frac{1}{2}(m_{21}+m_{01})-\frac{1}{4}(m_{22}+m_{20}+m_{02}+m_{00})=(1-2r_1)d_2+2r_1(1-r_1)(1-2r_2)^2i_{dd}$$

آثار اپیستازی افزایشی× افزایشی (I_{aa})، افزایشی × غالبیت (I_{ad})، غالبیت× افزایشی (I_{da}) و غالبیت× غالبیت (I_{dd}) در سطح نشانگر می‌تواند بدین صورت بیان شود:

$$I_{aa}=\frac{1}{4}(m_{22}+m_{00})-\frac{1}{4}(m_{20}+m_{02})=(1-2r_2)(1-2r_1)i_{aa}$$

$$I_{ad}=\frac{1}{2}\left[m_{21}-\frac{1}{2}(m_{22}+m_{20})\right]-\frac{1}{2}\left[m_{01}-\frac{1}{2}(m_{02}+m_{00})\right]=(1-2r_2)^2(1-2r_1)i_{ad}$$

$$I_{da}=\frac{1}{2}\left[m_{12}-\frac{1}{2}(m_{22}+m_{02})\right]-\frac{1}{2}\left[m_{10}-\frac{1}{2}(m_{20}+m_{00})\right]=(1-2r_1)^2(1-2r_2)i_{da}$$

$$\begin{aligned}I_{dd}&=\left[m_{11}-\frac{1}{2}(m_{21}+m_{01})\right]-\frac{1}{2}\left\{\left[m_{12}-\frac{1}{2}(m_{22}+m_{02})\right]+\left[m_{10}-\frac{1}{2}(m_{20}+m_{00})\right]\right\}\\&=\left[m_{11}-\frac{1}{2}(m_{12}+m_{10})\right]-\frac{1}{2}\left\{\left[m_{21}-\frac{1}{2}(m_{22}+m_{20})\right]+\left[m_{01}-\frac{1}{2}(m_{02}+m_{00})\right]\right\}\\&=(1-2r_1)^2(1-2r_2)^2i_{aa}\end{aligned}$$

با تجزیه بالا، می‌توانیم مشاهده کنیم که چه مقدار از برآوردهای مربوط به هر یک از آثار نشانگر تحت‌تأثیر موقعیت QTL و آثار متقابل QTL-QTL قرار گرفته‌اند.

۸-۷- تجزیهٔ نشانگری کل ژنوم

هدف از تهیه نقشه ژنتیکی، شناسایی QTL های مؤثر بر یک صفت کمّی، تعداد آن‌ها، توزیع ژنومی، آثار ژنتیکی و حساسیت به سیگنال‌های محیطی مختلف است. تجزیه‌های جداگانه نشانگرهای منفرد برای به‌دست آوردن دقیق تمامی اطلاعات در خصوص QTL کافی نیست، چون شبکه اثر متقابل آن‌ها نمی‌تواند به صورت سیستماتیک شناسایی شود. ژو (۲۰۰۳) روشی را برای

برآورد همزمان آثار نشانگر در کل ژنوم پیشنهاد داد. در این روش فرض می‌شود که چگالی نشانگر نسبتاً بالا است و بنابراین آثار نشانگر تقریباً آثار QTL مرتبط با نشانگرها را نشان می‌دهد.

یک تلاقی برگشتی را در نظر بگیرید که در آن یک صفت کمّی برای فرد i اندازه‌گیری شده و با y_i نشان داده شده است. مدل آماری برای y_i در تمامی m نشانگر، هر یک با دو ژنوتیپ ۱ و ۰ و در کل ژنوم بدین صورت نوشته می‌شود:

$$y_i = b_0 + \sum_{k=1}^{m} x_{ik} b_k + e_i \qquad (۳۵-۸)$$

که در آن، b_0 میانگین کل، x_{ik} یک متغیر ساختگی (مصنوعی) نشان‌دهنده ژنوتیپ k امین نشانگر برای فرد i، b_k اثر k امین نشانگر و e_i خطای باقیمانده است که از توزیع نرمال $N\left(0,\sigma^2\right)$ تبعیت می‌کند. متغیر ساختگی به‌صورت ۰ و ۱ = x_{ik} تعریف می‌شود که به ژنوتیپ نشانگر بستگی دارد. معادله (۸-۳۵) یک مدل تابعیت چندگانه است که ضریب تابعیت جزیی b_k اثر نشانگر k ام مرتبط با صفت است.

چگونگی برآورد ضرایب تابعیت جزیی در معادله (۸-۳۵) از نظر آماری یک مسئله چالش‌زا است، چون تعداد نشانگرها در مدل خطی ممکن است بزرگ‌تر از تعداد افراد بوده و آثار تعداد زیاد نشانگر درحقیقت نزدیک به صفر است. ژو (۲۰۰۳) یک روش بیزین را برای برآورد پارامتر به‌کار گرفت. در مفهوم بیزین، هر جزئی از جمله پارامترها به‌عنوان یک متغیر تصادفی با توزیع خاص در نظر گرفته می‌شوند. برای مثال هر b_k از یک توزیع نرمال با میانگین صفر و واریانس σ_k^2 نمونه‌برداری می‌شود. تمامی متغیرها به‌صورت قابل مشاهده و غیر قابل مشاهده دسته‌بندی و مرتب می‌شوند. متغیرهای قابل مشاهده همان ارزش‌های فنوتیپی یعنی $y=\{y_i\}$ و داده‌های نشانگر هستند، حال آنکه متغیرهای غیر قابل مشاهده شامل $b=\{b_k\}$ و v= توزیع پیشین، درستنمایی و توزیع پسین هستند. چارچوب بیزین از سه جزء توزیع پیشین، درستنمایی و توزیع پسین تشکیل یافته است. توزیع پیشین، توزیع متغیرهای غیر قابل مشاهده است. درستنمایی، متغیرهای توزیع قابل مشاهده است که به‌صورت تابعی از متغیرهای غیر قابل مشاهده بیان می‌شود. توزیع پسین که نیاز به استنباط کردن از روی تجزیه بیزی دارد، توزیع شرطی پارامترها به شرط داده‌های مشاهده شده است.

روش زنجیره مارکوف مونت کارلو (MCMC) تجزیه بیزین را روشی انعطاف‌پذیر نموده است. تجزیه بیزین با استفاده از MCMC به مشخص و واضح بودن توزیع پسین نیاز ندارد، بلکه نمونه‌ای از متغیرهای غیرقابل مشاهده را از روی توزیع پسین توأم استخراج می‌کند. از روی نمونه پسین توأم، برآوردهای مطلوب بیزی، نظیر میانگین‌های پسین و واریانس‌های پسین می‌تواند به‌دست آید.

توصیف بعدی تجزیه بیزی بر اساس ایده ژو (۲۰۰۳) خواهد بود که توزیع‌های پیشین زیر را انتخاب نمود:

$$P(b_0) \propto 1,$$
$$P(\sigma_e^2) \propto \frac{1}{\sigma_e^2},$$
$$P(b_k) = N(0, \sigma_k^2),$$
$$P(\sigma_k^2) \propto \frac{1}{\sigma_k^2}$$

توزیع پیشین توأم متغیرهای غیر قابل مشاهده یعنی $P(b,v)$ حاصل‌ضرب مقادیر پیشین پارامترهای انفرادی است؛ بنابراین، درستنمایی عبارت است از:

(۳۶-۸)

$$P(y \mid b,v) = \prod_{i=1}^{n} P(y_i \mid b, \sigma_e^2) \propto \left[\sigma_e^2\right]^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{k=1}^{m} x_{ik} b_k\right)^2\right]$$

همچنین توزیع پسین توأم به شکل زیر است:

$$P(b,v \mid y) \propto P(y \mid b,v)\, P(b,v) \qquad (۳۷-۸)$$

برای تجزیه بیزین به کمک MCMC، متغیرهای غیرقابل مشاهده از روی توزیع پسین توأم بالا نمونه‌برداری می‌شوند. نمونه‌برداری بر اساس مراحل زیر انجام می‌شود:

۱- مقادیر ابتدایی تمامی متغیرهای غیرقابل مشاهده که به‌صورت زیر نمایش داده می‌شوند را تعیین کنید:

$$\left(b_0^{(0)}, b_1^{(0)}, \ldots, b_k^{(0)}, \sigma_e^{2(0)}, \sigma_1^{2(0)}, \ldots, \sigma_k^{2(0)}\right)$$

تمامی پارامترهای b با مقدار صفر و پارامترهای مقیاسی v با یک عدد مثبت شروع می‌شوند.

۲- میانگین جمعیت b_0 را نوسازی و تجدید نمایید. توزیع پسین شرطی b_0 به صورت نرمال با میانگین

$$\bar{b}_0 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{m} x_{ik} b_k^{(0)}\right)$$

و واریانس

$$\overline{s}_0^2 = \sigma_e^{2(0)}$$

است که از روی آن می‌توانیم یک b_0 جدید که به صورت $b_0^{(1)}$ نمایش داده می‌شود را نمونه‌برداری کنیم که در کلیه مراحل بعدی نمونه‌برداری جایگزین $b_0^{(0)}$ می‌شود.

۳- ضرایب تابعیت جزیی b_k را برای $k = 1,\ldots,m$ نوسازی کنید. توزیع پسین شرطی برای b_k به صورت نرمال با میانگین

$$\overline{b}_k = \left[\sum_{i=1}^{n} x_{ik}^2 + \frac{\sigma_e^{2(0)}}{\sigma_k^{2(0)}}\right]^{-1} \sum_{i=1}^{n} x_{ik}\left(y_i - b_0^{(0)} - \sum_{l\neq k}^{m} x_{il} b_l^{(0)}\right)$$

و واریانس

$$\overline{s}_k^2 = \left[\sum_{i=1}^{n} x_{ik}^2 + \frac{\sigma_e^{2(0)}}{\sigma_k^{2(0)}}\right]^{-1} \sigma_e^{2(0)}$$

است که برای نمونه‌برداری b_k استفاده می‌شود تا $b_0^{(1)}$ به‌دست آید که جایگزین $b_0^{(0)}$ در تمامی مراحل بعدی نمونه‌گیری می‌شود.

۴- واریانس باقیمانده $\sigma^{2(0)}$ را نوسازی کنید. واریانس باقیمانده از روی معکوس توزیع کای مربع مقیاسی نمونه‌برداری می‌شود:

$$\sigma_e^{2(1)} = \frac{1}{x_n^2}\sum_{i=1}^{n} x_{ik}\left(y_i - b_0^{(0)} - \sum_{l\neq k}^{m} x_{il} b_l^{(0)}\right)$$

که X_n^2 عدد تصادفی نمونه‌برداری شده از توزیع کای مربع با n درجه آزادی است. واریانس $\sigma_e^{2(0)}$ فوراً با $\sigma_e^{2(1)}$ جایگزین می‌شود.

۵- واریانس‌های ضرایب تابعیت جزیی σ_k^2 برای $k = 1,\ldots,m$ نوسازی می‌شود. آن‌ها می‌توانند از روی عکس توزیع کای مربع مقیاسی نمونه‌برداری شوند:

$$\sigma_k^{2(1)} = \frac{b_k^{2(0)}}{X_1^2}$$

که X_1^2 عدد تصادفی نمونه‌برداری شده از توزیع کای مربع با یک درجه آزادی است.

۶- مراحل ۲ تا ۵ را تکرار کنید. در این نقطه، یک رفت و برگشت از MCMC تکمیل شده و نمونه‌برداری از دور بعدی می‌تواند ادامه یابد. زمانی که زنجیره در توزیع ثابت به همگرایی برسد، پارامترهای نمونه‌برداری شده از توزیع پسین توأم تبعیت می‌کنند.

مثال ۸-۹. ژو (۲۰۰۳) از روش بیزین برای شناسایی آثار نشانگر در کل ژنوم گیاه جو استفاده کرد (تینکر و همکاران، ۱۹۹۶). جمعیت مکان‌یابی از n = ۱۴۵ لاین دابل هاپلوئید (DH) تشکیل شده بود که بسط مدل تلاقی برگشتی می‌تواند مستقیماً در این جمعیت مورد استفاده قرار گیرد. مجموع m = ۱۲۷ نشانگر مولکولی، ۷ گروه پیوستگی را تشکیل دادند که ۱۵۰۰ سانتی‌مورگان از ژنوم جو را پوشش می‌داد. ماده آزمایشی در محیط‌های مختلف کشت شده بود. چندین صفت زراعی برای هر گیاه اندازه‌گیری و بر اساس میانگین بین محیط‌ها تجزیه شدند، ولی فقط نتایج وزن دانه در این بخش نشان داده می‌شوند.

شکل ۸-۵-الف آثار نشانگر را در طول ۷ گروه پیوستگی به هم چسبیده نشان می‌دهد. ۴ منطقه کاندیدا ارتباط معنی‌داری را بین نشانگرها و وزن دانه نشان دادند که با نتایج بدست آمده در تجزیه تک نشانگری انطباق دارد (شکل ۸-۵-ب). تفاوت معنی‌دار بین دو روش از آن جهت است که آثار نشانگری که با تجزیه چندنشانگری تشخیص داده شده‌اند به صورت بارزتری تفاوت موجود را نسبت به آثار نشانگری تشخیص داده شده با تجزیه تک نشانگری نشان می‌دهند. این نتیجه ممکن است نشان دهد که تجزیه چند نشانگری، روابط معنی‌دار کاذب کمتری را در مقایسه با تجزیه تک نشانگری نشان می‌دهد.

ژو (۲۰۰۳) آزمون جالبی را برای بررسی توزیع آثار نشانگر در سرتاسر ژنوم نیز انجام داد. آثار نشانگری شناسایی شده با تجزیه چند نشانگری و تک نشانگری می‌توانند به وسیله توزیع گاما با مقیاس برآورد شده (۰/۰۵۷۹ در مقابل ۰/۱۱۳۴) و پارامترهای شکل (۰/۲۲۳۳ در مقابل ۱/۱۳۹۶) برازش شوند که تا حد زیادی بین دو روش متفاوت هستند؛ بنابراین، تجزیه چند نشانگری یک توزیع L شکل ایجاد کرد (شکل ۸-۶-الف)، درحالیکه تجزیه تک نشانگری توزیع زنگوله‌ای شکل (شکل ۸-۶-ب) ایجاد می‌کند. تجزیه‌های چندگانه به‌کار گرفته شده در مفهوم بیزین ابزار قدرتمندی برای مطالعه کل ساختار ژنتیکی یک صفت کمّی فراهم می‌آورند.

۸-۸- تمرین‌ها

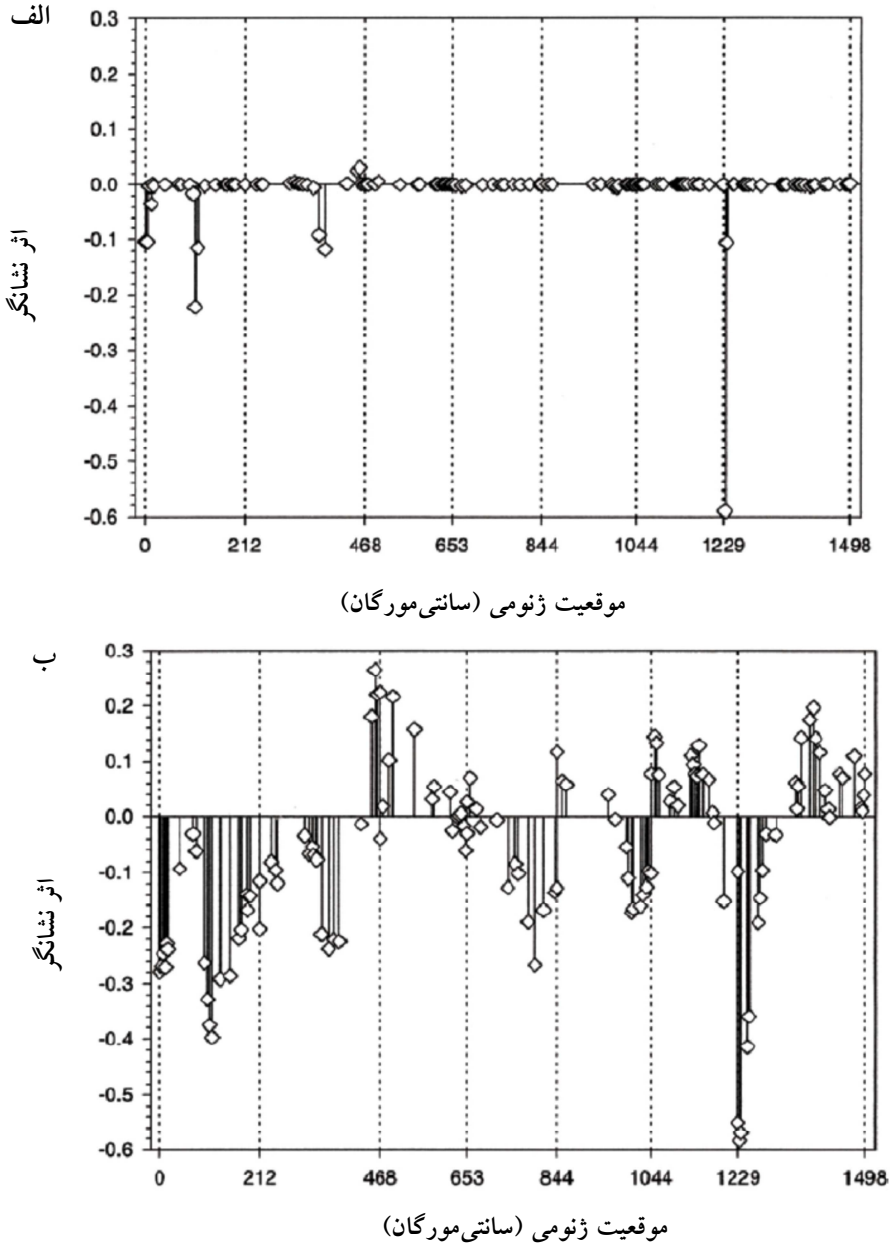
۸-۱- باتوجه به داده‌های جدول ۸-۱، برای نشانگر B آزمون کنید که آیا تفاوت وزن بدن بین ژنوتیپ‌های ۱ و ۰ از نظر آماری معنی‌دار است یا خیر؟

۸-۲- باتوجه به مثال ۸-۵، جدول تجزیه واریانس را برای نشانگر B تشکیل داده و آزمون‌های t بعدی را انجام دهید. در مورد این نشانگر چه نتیجه‌ای می‌توانید بگیرید؟

۸-۳- برای مدل تلاقی برگشتی (۸-۹)، توجه کنید که اگر ژنوتیپ M_1M_1 مشاهده شود، چگالی y بدین صورت ارایه می‌شود:

$$\frac{(1-r)}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{(y-\mu_1)^2}{2\sigma^2}\right\}+\frac{r}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{(y-\mu_{12})^2}{2\sigma^2}\right\}$$

از این مطلب استفاده کنید تا میانگین‌ها و واریانس‌های معادله (۸-۱۰) را مورد بررسی قرار دهید.



شکل ۸-۵- آثار نشانگری وزن دانه در جو در بین نشانگرهای مختلف موجود روی ژنوم.
الف- تجزیه بیزین چند نشانگری؛ ب- تجزیه تابعیت تک نشانگری. هفت گروه پیوستگی با خطوط عمودی نقطه‌چین جدا شده‌اند. اقتباس از ژو (۲۰۰۳).

۸-۴- اگر f چگالی توزیع نرمال یعنی $f(y)=\frac{(1)}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{(y-\mu)^2}{2\sigma^2}\right\}$ باشد، نشان دهید که:

الف. $\frac{\partial}{\partial\mu}f=(y-\mu)f$

ب. $\frac{\partial}{\partial\sigma^2}f=\left(\frac{(y-\mu)^2-\sigma^2}{2\sigma^{\frac{3}{2}}}\right)f$

۸-۵- با توجه به رابطه (۸-۱۳)، عبارات درستنمایی و MLE ها را مورد بررسی قرار دهید.

۸-۶- معادلات (۸-۱۷) و (۸-۱۸) را مورد بررسی قرار دهید.

۸-۷- برای لگاریتم درستنمایی در بخش (۳) از تکرار قسمت ۸-۴-۱ نشان دهید که:

$$\frac{\partial}{\partial r}\log L\left(\hat{\mu}_1,\hat{\mu}_{12},\hat{\sigma}^2,r\,|\,y\right)=\sum_{i=1}^{n_1}\left(\frac{1-P_1(y_i)}{r}-\frac{P_1(y_i)}{1-r}\right)+\sum_{i=n_1+1}^{n}\left(\frac{P_2(y_i)}{r}-\frac{1-P_2(y_i)}{1-r}\right)$$

و این رابطه را برابر با صفر قرار دهید تا عبارت زیر به‌دست آید:

$$r=\frac{1}{n}\left(\sum_{i=1}^{n_1}1-P_1(y_i)+\sum_{i=n_1+1}^{n}P_2(y_i)\right)$$

۸-۸- در قسمت ۸-۶-۲، فرض کردیم که دو نشانگر از یکدیگر مستقل هستند. ولی اگر نشانگرها در منطقه ژنومی مشابهی به هم پیوسته باشند، موارد زیر را به‌طور جداگانه برای تلاقی برگشتی و F_2 انجام دهید:

الف. از احتمالات شرطی ژنوتیپ‌های QTL به‌شرط ژنوتیپ‌های نشانگر مشتق بگیرید.

ب. ارزش‌های ژنوتیپی نشانگر را بر حسب ارزش‌های ژنوتیپی QTL بیان کنید.

ج. نشان دهید چگونه آثار متقابل در سطح نشانگر، زمانی که از آن‌ها برای نشان دادن آثار متقابل QTL استفاده می‌شود، کمتر از حد معمول برآورد می‌شوند.

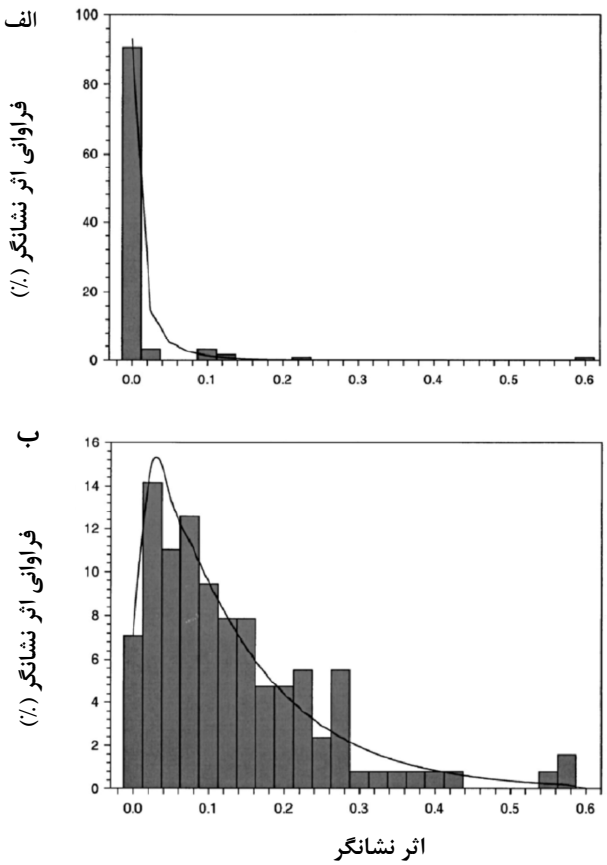
۸-۹- دوبلی و همکاران (۱۹۹۵) دو نشانگر را به عنوان QTL کاندیدا برای متوسط طول میانگره‌های رویشی در پنجه اصلی اولیه یک جمعیت F_2 ۱۸۳ بوته‌ای حاصل از لاین‌های اینبرد، Teosinte-M1l× Teosinte-M3l، تعیین ژنوتیپ کردند. مشاهدات و ارزش‌های صفت برای هر ژنوتیپ در این دو QTL به صورت زیر ارایه شده‌اند (کائو و ژنگ، ۲۰۰۲).

P	Q			کل
	۲	۱	۰	
		مشاهدات		
۲	۸	۲۲	۳	۳۳
۱	۲۰	۴۲	۲۴	۸۶
۰	۱۱	۲۱	۱۰	۴۲
کل	۳۹	۸۵	۳۷	۱۶۱
		ارزش‌های فنوتیپی		میانگین
۲	۱۰۱/۶۰	۶۶/۵۰	۶۱/۱۱	۷۴/۵۲
۱	۸۳/۶۲	۴۷/۵۵	۴۰/۹۴	۵۴/۰۹
۰	۴۷/۸۰	۵۴/۵۷	۱۷/۹۸	۴۴/۰۸
کل	۷۷/۲۱	۵۴/۱۹	۳۶/۳۷	۵۵/۶۷

الف. آزمون کنید که آیا این دو QTL کاندیدا به هم پیوسته هستند یا خیر؟
ب. آثار افزایشی هر QTL کاندیدا را برآورد کنید.
ج. آثار غالبیت هر QTL کاندیدا را برآورد کنید.
د. آثار ژنتیکی اپیستازی بین دو QTL را برآورد کنید.
ه. اگر این دو QTL کاندیدا به صورت نشانگر مشاهده شوند، بحث کنید که چگونه برآورد آثار ژنتیک افزایشی، غالبیت و اپیستازی با موقعیت QTL ها و آثار اپیستازی اختلاط می‌یابند.

۸-۱۰- از دست رفتن داده‌های نشانگر در تجزیه نشانگری کل ژنوم

از دست رفتن برخی از داده‌های ژنوتیپی برای برخی از نشانگرها معمول است. ژنوتیپ نشانگرهای از دست رفته می‌توانند به‌صورت تصادفی با احتمال خاصی به‌طور توأم از روی نزدیکترین نشانگرها و فنوتیپ‌های مجاور که از دست نرفته‌اند، استنباط شوند (ژو، ۲۰۰۳). احتمال حاصل از نشانگرها به عنوان احتمال پیشین است. پس از انطباق آثار نشانگر از راه فنوتیپ، احتمال مربوطه به احتمال پسین تبدیل می‌شود که برای تولید ژنوتیپ نشانگر از دست رفته استفاده می‌شود. الگوریتمی را برای نمونه‌گیری از داده‌های از دست رفته در قالب MCMC (به‌صورتی که در بخش ۸-۷ نشان داده شد)، ارایه کنید.



شکل ۸-۶- تقریب گاما از آثار نشانگری برای وزن دانه در جو. الف- تجزیه بیزین چند نشانگری. ب- تجزیه تابعیت تک نشانگری. اقتباس از ژو (۲۰۰۳).

فصل نهم

ساختار مکان‌یابی QTL

۹-۱- مقدمه

در فصل قبل، راهکارهای آماری ارزیابی ارتباط بین نشانگرها و فنوتیپ‌ها توضیح داده شدند. چنین تجزیهٔ ارتباطی می‌تواند شواهدی را برای کنترل ژنتیکی تنوع صفت فراهم کند، ولی خیلی دقیق نیست، چون آثار ژنتیکی مرتبط با ژنوتیپ‌های نشانگر به وسیله موقعیت QTL کارکردی و اثر واقعی آن اختلاط می‌یابد. البته اگر نشانگرها بسیار متراکم باشند و در موقعیت QTL ها قرار داشته باشند، یک تجزیه ارتباط ساده نشانگر-فنوتیپ ممکن است مفید باشد. تهیه چنین نقشه‌های بسیار متراکم و اشباع شده برای بیشتر گونه‌ها امکان‌پذیر نیست. روش‌های تجزیه قدرتمندی برای جداسازی آثار یک QTL از موقعیت آن مورد نیاز است.

بر خلاف نشانگرهای مولکولی، موقعیت ژنومی QTL ها ناشناخته است و باید بر اساس تجزیه ارتباط بین فنوتیپ‌ها و نشانگرها به آن‌ها پی برد. نقش روش‌های آماری، شناسایی، مکان‌یابی و برآورد QTLهای کارکردی با استفاده از نشانگرهای خنثی[1] و با جایگاه ژنوم شناخته شده است. در ابتدای این فصل، به‌طور سیستماتیک راهکارهای مکان‌یابی QTL و روش‌های آماری موجود برای برآورد موقعیت‌ها و آثار QTL ها را بر اساس نقشه‌های پیوستگی ژنتیکی معرفی خواهیم کرد. تمامی این مباحث هسته مرکزی این کتاب را تشکیل می‌دهند.

یکی از مهم‌ترین اصول آماری مکان‌یابی QTL، بر اساس مدل مختلط انجام می‌شود (مک لاچلان و پیل، ۲۰۰۰) که در آن فرض می‌شود هر مشاهده به یکی از g گروه ژنوتیپی QTL غیر قابل مشاهده تعلق داشته و هر گروه به‌طور مناسبی با یک توزیع چگالی احتمال مربوط به بعضی از توزیع‌های پارامتری مدل‌سازی می‌شود. این مدل چارچوبی را فراهم می‌آورد که به وسیله آن مشاهدات می‌توانند در قالب گروه‌های ژنوتیپی مشخص گروه‌بندی شوند. روش‌های مکان‌یابی QTL بر پایه مدل مختلط شامل سه مرحله هستند:

۱- مدل مختلط کامل را برای ارتباط ممکن بین فنوتیپ-ژنوتیپ مدل‌سازی کنید.

۲- انواع آثار QTL، یعنی تصادفی یا ثابت را مشخص کنید.

۳- پارامترهای ناشناخته QTL را با استفاده از روش‌های آماری برآورد کنید.

در این فصل، مباحث ژنتیکی و آماری اصولی مرتبط با خصوصیات مدل‌های مختلط مکان‌یابی QTL معرفی خواهد شد. توضیح این مباحث کمک خواهد کرد تا نتایج را بهتر مورد استفاده قرار دهیم و مدل‌های آماری جدید را برای مسائل جدید بسط دهیم.

---

[1]- منظور از نشانگرهای خنثی این است که نقشی در تنوع و تغییرات صفت کمّی مورد نظر ندارند (مترجمین).

۹-۲- مدل مختلط

۹-۲-۱- ارائه فرمول

اساس ژنتیکی روش‌های آماری مکان‌یابی آن است که QTL شناسایی شده به دو یا چند ژنوتیپ مختلف در نتاج جمعیت مکان‌یابی در حال تفرق باشد؛ بنابراین، مکان‌یابی QTL می‌تواند با استفاده از هر جمعیت مکان‌یابی انجام شود. مدل‌های آماری برای مطالعه آثار QTL می‌توانند به‌صورت دقیق‌تری با این فرض که QTL مفروض با استفاده از معادله (۸-۱) برای تلاقی برگشتی یا معادله (۸-۲) برای F_2 شناخته می‌شود، اختصاصی باشند. جدول ۹-۱ مثال‌هایی را برای ساختار داده‌های یک طرح تلاقی برگشتی ارایه می‌دهد که با مدل تابعیت QTL مطابقت دارد.

در عمل، ژنوتیپ‌های QTL برای یک صفت قابل مشاهده نیستند، اگرچه آن‌ها حقیقتاً تنوع ژنتیکی صفت را تعیین می‌کنند. برای یک جمعیت مکان‌یابی، الگوی تفرق QTL می‌تواند پیش‌بینی شود، برای مثال، نسبت ۱:۱ برای تلاقی برگشتی (BC) و نسبت ۱:۲:۱ برای F_2. به‌طور تصادفی یک موش، i را از جدول ۹-۱ انتخاب می‌کنیم، احتمال این‌که این موش حامل ژنوتیپ QTL به صورت Qq(1) یا qq(0) باشد، به ترتیب برابر با $\frac{1}{2}$ و $\frac{1}{2}$ است. به‌طور مشابه، چنین احتمالاتی به‌صورت $\frac{1}{4}$ برای QQ(2)، $\frac{1}{2}$ برای Qq(1) و $\frac{1}{4}$ برای qq(0) در طرح F_2 خواهد بود (جدول ۹-۲)؛ بنابراین می‌توان فرض کرد که هر موش به یکی و فقط یکی از این ژنوتیپ‌های QTL با احتمالی که در بالا برای طرح‌های مختلف در نظر گرفته شد تعلق دارد.

جدول ۹-۱- وزن بدن موش (y) و یک QTL مفروض در طرح تلاقی برگشتی.

نمونه	ژنوتیپ QTL (z)	وزن بدن (y)
۱	Qq(1)	۳۰
۲	Qq(1)	۳۲
۳	Qq(1)	۲۸
۴	Qq(1)	۲۹
۵	Qq(1)	۲۹
۶	qq(0)	۲۲
۷	qq(0)	۲۰
۸	qq(0)	۲۱
۹	qq(0)	۲۰
۱۰	qq(0)	۲۱

برای موش i ام، تابع درستنمایی فنوتیپ آن یعنی y_i می‌تواند در قالب مدل مختلط به‌صورت زیر فرموله شود:

(۹-۱)

$$y_i \sim p\left(\phi_j,\eta \mid y_i\right)=\begin{cases}\frac{1}{2} f_1\left(y_i;\phi_1,\eta\right)+\frac{1}{2} f_0\left(y_i;\phi_0,\eta\right) & \text{برای BC}\\ \frac{1}{4} f_2\left(y_i;\phi_2,\eta\right)+\frac{1}{2} f_1\left(y_i;\phi_1,\eta\right)+\frac{1}{4} f_0\left(y_i;\phi_0,\eta\right) & \text{برای } F_2\end{cases}$$

که در آن، $f_j\left(y_i;\phi_j,\eta\right)$ یک نوع تابع چگالی احتمال با پارامتر اختصاصی ϕ_j (مختص هر جزء یعنی ژنوتیپ) و پارامتر مشترک η (مشترک برای تمامی اجزاء) است. نسبت اختلاط هر جزء در مدل مختلط (۹-۱) همان فراوانی‌های ژنوتیپی در جمعیت مکان‌یابی کامل هستند.

جدول ۹-۲- وزن بدن موش که به وسیله یک QTL فرضی در حال تفرق با احتمالات پیشین مختلف در طرح تلاقی برگشتی یا F_2 تحت تأثیر قرار گرفته است.

نمونه	وزن بدن	تلاقی برگشتی		F_2		
		Qq(1)	qq(0)	QQ (2)	Qq(1)	qq(0)
۱	۳۰	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
۲	۳۲	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
۳	۲۸	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
۴	۲۹	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
۵	۲۹	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
۶	۲۲	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
۷	۲۰	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
۸	۲۱	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
۹	۲۰	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
۱۰	۲۱	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

مدل (۹-۱) می‌تواند برای تشخیص وجود یک QTL در حال تفرق در جمعیت مکان‌یابی مورد استفاده قرار گیرد، ولی برای تعیین موقعیت کروموزومی QTL قابل استفاده نیست. موقعیت QTL می‌تواند با استنباط ژنوتیپ‌های QTL از روی الگوی تفرق نشانگرهایی که به QTL پیوسته هستند، برآورد شود. داده‌های خام برای مکان‌یابی QTL شامل اطلاعات مرتب شده نشانگر و فنوتیپ‌های اندازه‌گیری شده افراد مختلف است (جداول ۸-۱ و ۸-۲). با اضافه کردن اطلاعات نشانگر (M)، تابع درستنمایی موش i ام برای جمعیت تلاقی برگشتی، همان‌گونه که در معادله (۹-۱) نشان داده شد، می‌تواند مجدداً به‌صورت زیر نوشته شود:

$$y_i \sim p\left(\phi_j, \eta \mid y_i, M\right) = \omega_{1|i} f_1\left(y_i; \phi_1, \eta\right) + \omega_{0|i} f_0\left(y_i; \phi_0, \eta\right) \qquad (۹-۲)$$

که $\omega_{j|i}$ احتمال شرطی ژنوتیپ j ام QTL مشروط به ژنوتیپ‌های نشانگر برای موش i ام است.

به‌طور مشابه، درستنمایی یک موش برای F_2 می‌تواند به‌صورت زیر نوشته شود:

(۳-۹)

$$y_i \sim p\left(\phi_j,\eta \mid y_i, M\right) = \omega_{2|i} f_2\left(y_i;\phi_2,\eta\right) + \omega_{1|i} f_1\left(y_i;\phi_1,\eta\right) + \omega_{0|i} f_0\left(y_i;\phi_0,\eta\right)$$

ماتریس‌های احتمال شرطی برای تلاقی برگشتی و F_2 در فصل ۱۰ ارائه شده‌اند.

۹-۲-۲- ساختار، تنظیم و برآورد

توزیع‌های مختلط محدودی می‌توانند برای مدل‌سازی انواع گسترده‌ای از پدیده‌های تصادفی استفاده شوند (مک لاچان و پیل، ۲۰۰۰ را مشاهده کنید) و راهکارهای آماری صحیحی برای شناسایی ژنوتیپ‌های ناشناخته QTL فراهم کنند. در یک روش مبتنی بر مدل مختلط نرمال به‌منظور جداسازی ژنوتیپ‌های QTL برای یک جمعیت مکان‌یابی عمومی، فرض می‌شود که هر مشاهده که شامل فنوتیپ (y_i) و اطلاعات نشانگر (M) است، از مخلوط تعداد مشخصی (J) از چگالی‌های احتمالی تشکیل شده است:

$$p\left(\Omega \mid y_i, M\right) = \omega_{1|i} f_1\left(y_i;\mu_1,\sigma^2\right) + \ldots + \omega_{J|i} f_J\left(y_i;\mu_J,\sigma^2\right) \qquad (۴-۹)$$

که در آن، $\Omega = \left(\omega_{1|i},\ldots,\omega_{J|i},\mu_1,\ldots,\mu_J,\sigma^2\right)$ بردار پارامترهای ناشناخته و $\omega_{j|i}$ها نسبت‌های اختلاط هستند که مجموع آن‌ها برابر با یک است و احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ‌های نشانگر هستند و $f_j\left(y_i;\mu_j,\sigma^2\right)$ به‌طورکلی چگالی توزیع نرمال برای ژنوتیپ j ام QTL با میانگین μ_j و واریانس σ^2 بوده و به‌صورت زیر بیان می‌شود:

$$f_j\left(y_i;\mu_j,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{\left(y_i-\mu_j\right)^2}{2\sigma^2}\right] \qquad (۵-۹)$$

فرض می‌شود که تمامی توزیع‌ها متعلق به یک توزیع پارامتری مشابه هستند.

با داشتن اطلاعات پایه در خصوص یک مدل مختلط، ایجاد و توسعه یک روش آماری برای مکان‌یابی QTL مستلزم انجام سه گام اصلی است:

۱- ساختارهای مدل مختلط شامل نسبت‌های اختلاط (که به صورت فراوانی ژنوتیپ‌های QTL در نظر گرفته می‌شوند) و توابع چگالی (که بر حسب آثار QTL و واریانس باقیمانده بیان می‌شوند) را به‌دست آورید.

۲- وضعیت مکان‌یابی QTL را از روی یک طرح آزمایشی با تخصیص آثار QTL و سایر آثار مدل به‌صورت ثابت یا تصادفی بهینه‌سازی نمایید.

۳- یک روش آماری را برای برآورد پارامترهای ناشناخته تعریف شده در مدل مختلط ایجاد کنید.

اولین مرحله نیاز به فرموله کردن مدل‌های ژنتیکی کامل بر اساس طرح‌های آمیزشی مختلف، انواع مختلف نشانگرها، ساختارهای جمعیتی مختلف و آثار ژنتیکی مختلف دارد. مرحله دوم با مدل‌سازی و تجزیه یک آزمایش مکان‌یابی بر طبق استراتژی‌های نمونه گیری خاص مرتبط است و در آن نتایج حاصل از مکان‌یابی QTL توضیح داده شده و مورد استفاده قرار می‌گیرند. مرحله سوم به الگوریتم‌های آماری و محاسباتی به کمک تکنیک-های رایانه‌ای وابسته است. این سه مرحله در چارچوب مدل مختلط با یکدیگر ترکیب می‌شوند تا به‌صورت کارآمد و مفید پارامترهای ناشناخته مورد نظر به‌دست آیند.

در عمل، بسیاری از روش‌های آماری مکان‌یابی به منظور به‌دست آوردن مدل‌های ژنتیکی که به‌طور بهتری ساختار ژنتیکی یک صفت پیچیده را در یک جمعیت خاص آشکار سازند، روی ساختار مدل مختلط تمرکز می‌کنند. بسیاری از روش‌ها تلاش می‌کنند تا وضعیت مدل مختلط را بر اساس طرح‌های آزمایشی مختلف مورد آزمایش قرار دهند و بدین صورت استفاده بهتری از نتایج مکان‌یابی داشته باشند. روش‌های دیگر ممکن است با به‌کارگیری یک الگوریتم محاسباتی کارآتر روی برآورد مدل مختلط در یک شجره پیچیده تمرکز کنند. این سه جنبه مکان‌یابی QTL را از منظر ژنتیکی، آزمایشی و آماری توضیح خواهیم داد.

۹-۳- ساختار ژنتیکی جمعیت مدل مختلط

همان‌گونه که در بالا توضیح داده شد، یک مدل مختلط نرمال معین حاوی دو ساختار مختلف یعنی نسبت‌های اختلاط $\omega=\left(\omega_{1|i},...,\omega_{J|i}\right)$ و توزیع‌های نرمال تعیین شده به وسیله پارامترهای مدل یعنی $\Omega_m=\left(\mu_1,...,\mu_J,\sigma^2\right)$ است. نسبت‌های اختلاط در حقیقت فراوانی ژنوتیپ‌های QTL در QTL مورد نظر در یک جمعیت هستند. برای یک جمعیت مکان‌یابی کامل حاصل از دو لاین اینبرد، فراوانی ژنوتیپ‌های QTL می‌توانند بر اساس قانون اول مندل پیش‌بینی شوند. برای مثال، فراوانی ژنوتیپ‌های QTL برای تلاقی برگشتی $\frac{1}{2}$ و $\frac{1}{2}$ یا برای F_2 به صورت $\frac{1}{4}$، $\frac{1}{2}$ و $\frac{1}{4}$ هستند؛ اما وقتی که اطلاعات نشانگر مرتبط با QTL مفروض پیوسته با آن باشد، فراوانی ژنوتیپ‌های QTL با شرط ژنوتیپ خاص نشانگر (یعنی احتمالات شرطی) از قانون مندل

تبعیت نمی‌کنند، بلکه بستگی به نرخ نوترکیبی یا عدم تعادل پیوستگی بین نشانگر و QTL دارند. احتمالات شرطی ژنوتیپ‌های QTL بر ژنوتیپ‌های نشانگر همانند پلی هستند که اطلاعات نشانگر شناخته شده را با اطلاعات QTL ناشناخته مرتبط می‌کنند. در بخش زیر، روشی برای به‌دست آوردن احتمالات شرطی در جمعیت‌های مکان‌یابی مختلف ارائه خواهد شد.

۹-۳-۱- تلاقی برگشتی/F_2

دو لاین اینبرد مختلف آمیزش می‌کنند تا هتروزیگوت F_1 را ایجاد کنند. زمانی که F_1 با والدین اصلی تلاقی برگشتی می‌کند، یا خودگشن می‌شود و یا با برادر-خواهران آمیزش می‌کند، ژن‌های مختلف به‌طور همزمان تفرق می‌یابند و منجر به ایجاد انواع غیروالدینی و نوترکیب می‌شوند که نسبت‌های آن‌ها بستگی به درجه پیوستگی بین ژن‌های مختلف دارد. از این‌رو با مشاهده تعداد انواع نوترکیب، می‌توانیم پیوستگی را برآورد کنیم؛ اما نوترکیب‌های بین یک نشانگر و QTL قابل مشاهده نیستند، چون جایگاه QTL ناشناخته است و نیاز داریم تا احتمالات شرطی ژنوتیپ‌های QTL را مشروط به ژنوتیپ‌های شناخته شده نشانگر بر حسب نرخ‌های نوترکیبی آن‌ها به‌دست آوریم. مزیت تلاقی برگشتی یا F_2 به عنوان یک جمعیت مکان‌یابی به فاز پیوستگی مشخص بین تمامی ژن ها مربوط می‌شود که دقیقاً می‌تواند از روی منشأ والدینی آلل‌ها تعیین شود. در بخش‌های ۱۰-۳ و ۱۰-۴، روشی برای به‌دست آوردن این احتمالات شرطی در تلاقی‌های برگشتی (جدول ۱۰-۳) و جمعیت‌های F_2 (جدول ۱۰-۶) ارایه می‌شود.

۹-۳-۲- دگرآمیزی یا دگرباروری[1]

گونه‌های زیادی نظیر درختان جنگلی، به‌دلیل خصوصیات زیستی پیچیده خود نمی‌توانند لاین‌های اینبرد تولید کنند، ولی می‌توانند آمیزش کنند تا یک جمعیت در حال تفرق تولید کنند. این نوع آمیزش، دگرباروری یا دگرآمیزی نامیده می‌شود. این امکان وجود دارد تا از نتاج حاصل از دگرباروری به عنوان یک جمعیت دگرآمیخته[2] استفاده شود، چون در طی میوز کراسینگ‌اور رخ می‌دهد. در حقیقت، لاین‌های دگرآمیخته[3] می‌توانند به‌عنوان والدین در برخی از جایگاه‌های ژنی هموزیگوت باشند، درحالی‌که در تعداد زیادی از جایگاه‌ها هتروزیگوت هستند و بنابراین تلاقی‌های کنترل شده آن‌ها می‌تواند برای برخی از جایگاه‌ها شبیه تلاقی برگشتی و برای جایگاه‌های دیگر شبیه F_2 بوده یا انواع جدیدی از تلاقی را نمایش دهد. ژنوتیپ‌های QTL در تلاقی‌های دگرآمیزی

---

[1] - Outbred crosses
[2] - Outbred population
[3] - Outbred lines

می‌توانند با احتمالات شرطی خود مشروط بر ژنوتیپ‌های نشانگر که به‌صورت تابعی از نرخ نوترکیبی بیان می‌شود تفسیر شوند. روشی برای تعیین فاز پیوستگی صحیح قبل از برآورد پیوستگی ضروری است. توضیح جزئیات مکان‌یابی QTL در جمعیت‌های دگربارور در فصل آخر ارائه شده است.

۹-۳-۳- لاین‌های اینبرد نوترکیب (RIL)

لاین‌های اینبرد نوترکیب (RILs) ابزار قدرتمندی برای مکان‌یابی ژنتیکی هستند. این لاین‌ها می‌توانند با خودگشنی مکرر و یا با آمیزش‌های مکرر برادر-خواهری نتاج F_1 حاصل از تلاقی بین دو لاین اینبرد ایجاد شوند. RIL ها می‌توانند به صورت یک جمعیت مکان‌یابی دائمی برای استفاده‌های متعدد عمل کنند، چون برای دو آلل مربوطه در تمامی مکان‌های ژنی پایدار و هموزیگوت هستند. برخی لاین‌ها مشابه انواع والدینی (غیرنوترکیب) هستند، حال آنکه بقیه لاین‌ها، نوترکیب هستند. احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ‌های نشانگر بر حسب نسبت زیگوت‌های نوترکیب به‌دست می‌آیند (جدول ۱۰-۷ را مشاهده کنید).

۹-۳-۴- جمعیت‌های طبیعی

برای برخی از گونه‌ها که امکان انجام تلاقی وجود ندارد، تفرق همزمان بین نشانگر و QTL می‌تواند با عدم تعادل پیوستگی (LD[1]) مشخص شود. LD ارتباطات غیرتصادفی بین جایگاه‌های ژنی مختلف را در یک جمعیت بیان می‌کند و می‌تواند برای تجزیه یک جمعیت با ساختار نامشخص استفاده شود. تجزیه پیوستگی به‌طور گسترده‌ای برای مکان‌یابی ژنتیکی از راه تشخیص درجه LD بین نشانگر و QTL استفاده شده است. بر خلاف تلاقی‌های کنترل شده، احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ‌های نشانگر بر حسب مقادیر LD بیان می‌شوند (لو و همکاران، ۲۰۰۳).

۹-۴- ساختار ژنتیک کمّی مدل مختلط

به‌طور کلی فرض می‌شود که صفت کمّی مورد نظر (y) در یک QTL فرضی دارای توزیع نرمال با میانگین μ_j و واریانس σ^2 است. مشارکت ژنتیکی QTL در فنوتیپ صفت در میانگین یا واریانس توزیع نرمال انعکاس می‌یابد. برای ژنوتیپ j ام یک QTL معین، می‌توانیم میانگین مورد انتظار آن را (μ_j) به اجزای مختلف شامل میانگین کلی (μ)، اثر افزایشی (a)، اثر غالبیت (d) و آثار

[1] - Linkage disequilibrium

اپیستازی افزایشی× افزایشی (i_{aa})، افزایشی× غالبیت (i_{ad})، غالبیت × افزایشی (i_{da}) و غالبیت × غالبیت (i_{dd}) تفکیک کنیم. اگر فرض کنیم که این آثار ثابت هستند، در این صورت می‌توانند به‌طور مستقیم در مدل مختلط برآورد شوند. همان‌طور که می‌دانیم واریانس آثار ثابت صفر است و از این‌رو واریانس فقط شامل واریانس باقیمانده (σ^2) در درون یک ژنوتیپ QTL خاص است. مرحله اصلی یک روش مکان‌یابی مبتنی بر مدل ثابت، تخصیص مؤلفه‌های ژنتیکی μ_j بر اساس مسائل ژنتیکی مختلف یا اهداف مکان‌یابی است.

۹-۴-۱- مدل افزایشی-غالبیت

در بسیاری از موارد، منطقی است که فرض کنیم هیچ گونه اثر متقابل غیر آللی (اپیستازی) بین QTL های مختلف وجود ندارد. یک QTL را با سه ژنوتیپ در نظر بگیرید که ارزش‌های ژنوتیپی و واریانس‌های باقیمانده آن به‌صورت زیر تعریف می‌شوند:

ژنوتیپ	QQ (2)	Qq (1)	qq (0)
میانگین کل	μ	μ	μ
اثر	a	d	-a
ارزش ژنوتیپی	$\mu_2 = \mu + a$	$\mu_1 = \mu + d$	$\mu_0 = \mu - a$
باقیمانده	σ^2	σ^2	σ^2

(۹-۶)

در بیشتر موارد، واریانس باقیمانده در بین سه گروه ژنوتیپی QTL برای سهولت محاسبات، یکسان فرض می‌شود. زمانی که واریانس‌های باقیمانده مختص ژنوتیپ‌ها (*heteroscedasticity*) باشند، تکنیک‌های آماری خاصی در دسترس هستند که باید مورد استفاده قرار گیرند.

زمانی که دو یا چند QTL برازش شود، تعداد ژنوتیپ‌های QTL بیشتری در مدل مختلط خواهیم داشت. برای یک جمعیت F_2، ارزش‌های ژنوتیپی، $\mu_{j_1 j_2}\ (j_1, j_2 = 0,1,2)$، نه ژنوتیپ QTL برای دو QTL مفروض، P و Q، تحت مدل افزایشی-غالبیت می‌تواند به‌صورت زیر تعریف شود:

(۹-۷)

$$\mu_{j_1 j_2} = \begin{array}{c} \\ P \\ 2 \\ 1 \\ 0 \end{array} \begin{array}{c} Q \\ \begin{array}{ccc} 2 & 1 & 0 \end{array} \\ \begin{bmatrix} \mu_{22}=\mu+a_1+a_2 & \mu_{21}=\mu+a_1+d_2 & \mu_{20}=\mu+a_1-a_2 \\ \mu_{12}=\mu+d_1+a_2 & \mu_{11}=\mu+d_1+d_2 & \mu_{10}=\mu+d_1-a_2 \\ \mu_{02}=\mu-a_1+a_2 & \mu_{01}=\mu-a_1+d_2 & \mu_{00}=\mu-a_1-a_2 \end{bmatrix} \end{array}$$

که a_1 و a_2 آثار افزایشی و d_1 و d_2 به ترتیب آثار غالبیت این دو QTL هستند.

مثال ۹-۱. در مکان‌یابی وزن بدن برای یکی از نتاج F_2 موش، ارزش‌های ژنوتیپی توأم دو QTL، P و Q، برآورد شدند. سپس ارزش‌های ژنوتیپی در QTL های منفرد به صورت میانگین‌های حاشیه‌ای محاسبه شدند. برآورد میانگین‌های ژنوتیپی حاشیه‌ای یا توأم در جدول ۹-۳ ارائه شده‌اند. آثار افزایشی و غالبیت QTL منفرد می‌توانند بر اساس میانگین‌های حاشیه‌ای با استفاده از معادله (۹-۶) یا ارزش‌های ژنوتیپی توأم با استفاده از معادله (۹-۷) برآورد شوند.

جدول ۹-۳- ارزش‌های ژنتیکی توأم و میانگین‌های حاشیه‌ای دو QTL برای وزن بدن در نتاج F_2.

P	Q			حاشیه‌ای
	۲	۱	۰	
۲	$\mu_{22}=$ ۳۰	$\mu_{21}=$ ۳۲	$\mu_{20}=$ ۲۸	$\mu_{2.}=$ ۳۰
۱	$\mu_{12}=$ ۲۹	$\mu_{11}=$ ۲۹	$\mu_{10}=$ ۲۱	$\mu_{1.}=$ ۲۵
۰	$\mu_{02}=$ ۲۱	$\mu_{01}=$ ۲۰	$\mu_{00}=$ ۲۱	$\mu_{0.}=$ ۲۰/۶۷
حاشیه‌ای	$\mu_{.2}=$ ۲۶/۶۷	$\mu_{.1}=$ ۲۷	$\mu_{.0}=$ ۲۲/۷۵	

فرض کنید m بردار میانگین ژنوتیپی، a بردار اثر و D ماتریس طرح باشد که m و a را به یکدیگر مرتبط می‌سازد. رابطه $m=Da$ منجر به ایجاد معادله زیر می‌شود:

$$a=D^{-1}m \qquad (۹-۸)$$

برای میانگین‌های حاشیه‌ای، داریم: $m=(\mu_2,\mu_1,\mu_0)^T$ و $a=(\mu,a,d)^T$ و:

$$D=\begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 1\\ 1 & -1 & 0\end{bmatrix}$$

برای ارزش‌های ژنوتیپی توأم داریم:

$m=\left(\mu_{22},\mu_{21},\mu_{20},\mu_{12},\mu_{11},\mu_{10},\mu_{02},\mu_{01},\mu_{00}\right)^T$ ، $a=\left(\mu,a_1,d_1,a_2,d_2\right)^T$ و:

$$D=\begin{bmatrix}1 & 1 & 0 & 1 & 0\\ 1 & 1 & 0 & 0 & 1\\ 1 & 1 & 0 & -1 & 0\\ 1 & 0 & 1 & 1 & 0\\ 1 & 0 & 1 & 0 & 1\\ 1 & 0 & 1 & -1 & 0\\ 1 & -1 & 0 & 1 & 0\\ 1 & -1 & 0 & 0 & 1\\ 1 & -1 & 0 & -1 & 0\end{bmatrix}$$

با استفاده از داده‌های جدول ۹-۳، آثار ژنتیکی به‌صورت ۲۴/۷۱ و ۲۵/۳۳۵ = μ، ۱/۹۶ و ۴/۶۶۵ = a و ۲/۲۹ و ۰/۳۳۵- = d به ترتیب برای Q و P از روی میانگین‌های حاشیه‌ای و به‌صورت ۲۴/۵۶ = μ، ۴/۶۷ = a_1، ۰/۳۵۷ = d_1، ۱/۹۹ = a_2 و ۲/۳۲ = d_2 از روی ارزش‌های ژنوتیپی توأم برآورد می‌شوند.

۹-۴-۲- مدل افزایشی-غالبیت-اپیستازی

اگر دو یا چند QTL با اثر متقابل خود بر یک صفت کمّی مؤثر باشند، آثار اپیستازی آن‌ها باید در ارزش‌های ژنوتیپی QTL موجود در مدل مختلط مدل‌سازی شود. از روش ماتر و جینکز (۱۹۸۲) برای توصیف اپیستازی استفاده می‌کنیم که با استفاده از آن، ارزش‌های ژنوتیپی همانند زمانی که دو QTL در نظر گرفته می‌شوند، بیان می‌گردند:

$$\mu_{j_1 j_2} = \begin{array}{c} \\ P \\ 2 \\ 1 \\ 0 \end{array} \begin{array}{c} Q \\ \begin{array}{ccc} 2 & 1 & 0 \end{array} \\ \begin{bmatrix} \mu+a_1+a_2+i_{aa} & \mu+a_1+d_2+i_{ad} & \mu+a_1-a_2-i_{aa} \\ \mu+d_1+a_2+i_{da} & \mu+d_1+d_2+i_{dd} & \mu+d_1-a_2-i_{da} \\ \mu-a_1+a_2-i_{aa} & \mu-a_1+d_2-i_{ad} & \mu-a_1-a_2+i_{aa} \end{bmatrix} \end{array}$$

(۹-۹)

که i_{aa}، i_{ad}، i_{da} و i_{dd} به ترتیب آثار اپیستازی افزایشی× افزایشی، افزایشی× غالبیت، غالبیت× افزایشی و غالبیت× غالبیت هستند.

مثال ۹-۲. مجدداً مثال ۹-۱ را مشاهده کنید. بر اساس جدول ۹-۳، می‌توانیم چهار نوع اثر اپیستازی را نیز برآورد کنیم. در اینجا، $a=\left(\mu,a_1,d_1,a_2,d_2,i_{aa},i_{ad},i_{da},a_{dd}\right)^T$ بوده و داریم:

$$D=\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

با استفاده از معادله (۹-۸)، آثار ژنتیکی را به‌صورت $\mu =$ ۲۵، $a_1 =$ ۴، $d_1 =$ ۰، $a_2 =$ ۰/۵، $d_2 =$ ۱، $i_{aa} =$ ۰/۵، $i_{ad} =$ ۲، $i_{da} =$ ۳/۵ و $i_{dd} =$ ۳ برآورد می‌کنیم.

۹-۴-۳- مدل اپیستازی ضرب‌پذیر

اساس فیزیولوژیک اپیستازی با مدل‌سازی رابطه بین ژن‌ها و فرآورده‌های آن‌ها در بسیاری از آزمایشات گیاهی و حیوانی مورد مطالعه قرار گرفته است. مینویل (۱۹۸۷) از دیدگاه نظری نشان

داد که اثر متقابل ضرب‌پذیر بین یک جفت ژن ممکن است شکل مهمی از اپیستازی برای کنترل صفات پیچیده باشد، همان‌گونه که به وسیله آروناچلام (۱۹۷۷) نیز توصیف شده بود. مدل اپیستازی ضرب‌پذیر فرض می‌کند که ژنوتیپ‌ها در یک جفت ژن دارای ارزش‌های ژنوتیپی برابر با حاصل‌ضرب ارزش‌های ژنوتیپی در دو جایگاه مختلف هستند (شنل و کوکرهام، ۱۹۹۲؛ لی و وو، ۱۹۹۶؛ اتو و فلدمن، ۱۹۹۷). برای مثال، اگر ارزش‌های ژنوتیپی برای ژنوتیپ دوم P به صورت μ_2 و برای ژنوتیپ دوم Q به صورت μ_2' باشند، سپس ارزش ژنوتیپ ۲۲ در QTL توأم به صورت $\mu_{22} = \mu_2\mu_2'$ است. تحت مدل اپیستازی ضرب‌پذیر، ارزش ژنوتیپی دو QTL می تواند با معادله زیر مدل‌سازی شود:

$$\begin{array}{cc} & Q \\ P & \begin{array}{ccc} 2 & 1 & 0 \end{array} \\ \mu_{j_1 j_2} = \begin{array}{c} 2 \\ 1 \\ 0 \end{array} & \begin{bmatrix} \mu_{22} = \mu_2\mu_2' & \mu_{21} = \mu_2\mu_1' & \mu_{20} = \mu_2\mu_0' \\ \mu_{12} = \mu_1\mu_2' & \mu_{11} = \mu_1\mu_1' & \mu_{10} = \mu_1\mu_0' \\ \mu_{02} = \mu_0\mu_2' & \mu_{01} = \mu_0\mu_1' & \mu_{00} = \mu_0\mu_0' \end{bmatrix} \end{array} \qquad (۱۰-۹)$$

گرچه آثار متقابل ضرب‌پذیر بین یک جفت QTL در نظر گرفته شده است، اما هنوز دو حالت خاص، یعنی عمل کاملاً ضرب‌پذیر (هر دو بین و درون جایگاه‌های ژنی) و فعالیت افزایشی خالص (بدون عمل غالبیت)، نیز می‌توانند با اعمال محدودیت‌هایی در نظر گرفته شوند.

مثال ۹-۳. مثال ۹-۱ را مجدداً مشاهده کنید. در اینجا سعی می‌کنیم تا پارامترهای اثر ژنتیکی را با استفاده از داده‌های موجود در جدول ۹-۳ تحت مدل اپیستازی ضرب‌پذیر برآورد کنیم. می‌توانیم از روش حداقل مربعات غیرخطی برای برآورد سه ارزش ژنوتیپی برای هر QTL بر اساس ماتریس (۹-۱۰) استفاده کنیم. این برآوردها به‌صورت $\mu_2 =$ ۵/۶۹۵۴، $\mu_1 =$ ۵/۰۳۰۸، $\mu_0 =$ ۳/۹۰۶۵، $\mu_2' =$ ۵/۴۶۲۴، $\mu_1' =$ ۵/۵۶۴۹ و $\mu_0' =$ ۴/۷۵۵ هستند.

شنل و کوکرهام (۱۹۹۲) یک مدل تجزیه‌ای برای برآورد آثار افزایشی و غالبیت و انواع آثار اپیستازی ناشی از دو QTL با اثر متقابل ضرب‌پذیر را پیشنهاد نمودند. برآوردگرهای این آثار بدین صورت بیان می‌شوند:

$$\begin{aligned} &\mu = uu', a_1 = \alpha u', d_1 = \delta u', a_2 = u\alpha', d_2 = u\delta', \\ &i_{aa} = \alpha\alpha', i_{ad} = \alpha\delta', i_{da} = \delta\alpha', i_{dd} = \delta\delta' \end{aligned} \qquad (۱۱-۹)$$

کـــــــــــــه در آن، $u=\left(\frac{1}{4}\mu_2+\frac{1}{2}\mu_1+\frac{1}{4}\mu_0\right)$، $u'=\left(\frac{1}{4}\mu'_2+\frac{1}{2}\mu'_1+\frac{1}{4}\mu'_0\right)$،

$\alpha=\frac{1}{2}(\mu_2-\mu_1)+\frac{1}{2}(\mu_1-\mu_0)$، $\alpha'=\frac{1}{2}(\mu'_2-\mu'_1)+\frac{1}{2}(\mu'_1-\mu'_0)$،

$\delta=\mu_2-2\mu_1+\mu_0$ و $\delta'=\mu'_2-2\mu'_1+\mu'_0$ است.

از ایـــن‌رو آثـــار ژنتیکـــی بـــه صـــورت $\mu=$ ۲۶/۲۳۴۹، $a_1=$ ۴/۷۷۳۶، $d_1=$ -۲/۴۵۳۳۱، $a_2=$ ۱/۷۳۸۵۶، $d_2=$ -۴/۴۸۵۰۱، $i_{aa}=$ ۰/۳۱۶۳۴، $i_{ad}=$ -۰/۸۱۶۰۷۴، $i_{da}=$ -۰/۱۶۲۵۷۸ و $i_{dd}=$ ۰/۴۱۹۴۰۷ برآورد می‌شوند.

۹-۴-۴- مدل مکانیسمی

ارزش‌های ژنوتیپی QTL های موجود در مدل مختلط نیز می‌توانند بر اساس اصول مکانیسمی به‌وجود آورنده آن‌ها مدل‌سازی شوند. برای مثال، ارزش ژنوتیپی ژنوتیپ QQ مربوط به QTL کنترل‌کننده یک صفت کمّی می‌تواند به صورت تابعی از متغیرهای مستقل دیگر، زمان یا عامل محیطی بیان شود:

$$y=\begin{cases} f(x) & \text{برای قوانین مربوط به اندازه‌گیری رشد موجودات} \\ g(t) & \text{برای مدل‌های رشدی} \\ h(z) & \text{برای استانداردهای واکنش} \end{cases} \qquad (۹-۱۲)$$

که y صفت زیستی مورد مطالعه، x اندازه بدن، t سن و z یک متغیر محیطی نظیر درجه حرارت، مواد غذایی یا شدت نور است. شکل‌های توابع ریاضی $f(x)$، $g(t)$ و $h(z)$ که می‌توانند خطی یا غیرخطی باشند، به‌طورکلی متفاوت بوده و به سئوالات خاص مورد نظر بستگی دارند. به‌طورکلی، ایجاد توابع ریاضی مناسب بر اساس نیکویی برازش داده‌های مشاهده شده هستند (نیکلاس، ۱۹۹۴). در حقیقت، این توابع ریاضی از دیدگاه بهینگی به‌دست می‌آیند. برای مثال، وست و همکاران (۱۹۹۹، ۱۹۹۷) یک سیستم شبکه‌ای فراکتالی را برای جذب و توزیع داخلی متابولیت‌ها پیشنهاد دادند تا قوانین "سنجش توان یک چهارم[1]" را که در جهان زنده به‌صورت گسترده وجود دارند توضیح دهند. بعلاوه، وست و همکاران (۲۰۰۱، ۱۹۹۹، ۱۹۹۷) توضیح دادند که چرا رشد یک موجود زنده از یک منحنی سیگموئیدی بر اساس اصول پایه‌ای تخصیص انرژی متابولیکی بین نگهداری بافت موجود و تولید زیست-توده جدید تبعیت می‌کند.

[1] - Quarter-power scaling

تاکنون، دو ساختار کلیدی در مدل مختلط مورد بحث قرار گرفت: نسبت‌های اختلاط و مدل‌های آثار ژنتیکی. همان‌گونه که می‌توان مشاهده نمود، احتمالات شرطی با فراوانی‌های آللی و پیوستگی یا ارتباط بین نشانگرها و QTL (موضوع ژنتیک مندلی یا جمعیت) وابسته هستند، حال آنکه توزیع نرمال با آثار کمّی QTL روی صفات مرتبط است (موضوع ژنتیک کمّی). بسته به ماهیت صفتی که مدنظر است، پیچیدگی‌های این دو نوع ساختار متفاوت خواهد بود. برخی از روش‌ها نظیر مکان‌یابی بر اساس عدم تعادل پیوستگی یا مکان‌یابی پلی‌پلوئیدی، بر استخراج احتمالات شرطی تمرکز و تأکید دارند. در حالی که روش‌های دیگر روی مدل‌های آثار ژنتیکی که می‌توانند برای تقریب پدیده‌های زیستی مورد استفاده قرار گیرند، تأکید دارند. این دو ساختار، اساس و اصالت یک روش آماری جدید برای مکان‌یابی QTL را تعیین می‌کنند.

۹-۵- تنظیم آزمایشی مدل مختلط

در مدل مختلط (۹-۴)، فرض می‌شود که اجزای مختلف نرمال با تعداد معلوم یا نامعلومی از ژنوتیپ‌های QTL مشخص می‌شوند. آثار ژنتیکی QTL های فرضی روی فنوتیپ که در توزیع‌های نرمال گنجانده می‌شوند، می‌توانند مستقیماً به وسیله روش مدل ثابت برآورد شوند. روش مدل ثابت در شرایطی مفید است که آثار ژنتیکی مورد نظر بتوانند به آسانی تخصیص یابند، همانند آنچه که در حالت تلاقی‌های کنترل شده حاصل از لاین‌های اینبرد خالص (لندر و بوتستین، ۱۹۸۹؛ هالی و نوت، ۱۹۹۲؛ زنگ، ۱۹۹۴) یا لاین‌های دگربارور (لین و همکاران، ۲۰۰۳) وجود دارد.

نتاج حاصل از یک تلاقی کنترل شده به منظور جمع‌آوری داده‌های فنوتیپی صفات کمّی مورد نظر باید در قالب یک طرح آزمایشی مشخص مورد مطالعه قرار گیرند. مدل مختلط اجازه برآورد آثار آزمایشی ناشی از هر نوع متغیر همبسته نظیر جنس یا عوامل محیطی ناپیوسته و آثار متقابل بین QTL و تیمارها را می‌دهد. دو طرح آزمایشی برای برآورد اثر متقابل QTL × محیط مورد استفاده قرار می‌گیرند. در طرح ۱، مجموعه یکسانی از ژنوتیپ‌ها که اطلاعات مربوط به نشانگرها در آن‌ها ثبت شده است، در محیط‌های مختلف ارزیابی می‌شوند. چنین طرحی برای گونه‌هایی نظیر کلون ها و RILs که می‌توانند به‌طور ژنتیکی افراد مشابهی تولید نمایند، امکان‌پذیر است. در طرح ۲، فقط مجموعه‌های متفاوتی از افراد حاصل از یک جمعیت مکان‌یابی در محیط‌های مختلف پرورش می‌یابند، ولی برای این مجموعه‌ها، سیستم‌های نشانگری مشابهی تعیین ژنوتیپ می‌شوند. این طرح برای بسیاری از گونه‌ها که در آن‌ها ژنوتیپ‌ها نمی‌توانند تکرار شوند، استفاده می‌شود.

ژنوتیپ‌های مشابه ممکن است به‌طور متفاوتی در دامنه‌ای از محیط‌ها عمل نمایند. تنوع ژنتیکی که سبب چنین انعطاف‌پذیری فنوتیپی[1] می‌شود، موجود زنده‌ای را ایجاد می‌کند که ظرفیت تحمل و مقاومت در برابر تنوع محیطی را دارد (شلیچتینگ، ۱۹۸۶). این نقش تحت تأثیر حساسیت آللی و تنظیم ژنی قرار دارد (ویا و همکاران، ۱۹۹۵). مفهوم حساسیت آللی بیان می‌کند که انعطاف‌پذیری از آثار متغیر جایگاه‌های ژنی ایجاد می‌شود و مستقیماً در تنوع صفات انعطاف‌پذیر[2] نقش دارند. فرضیه تنظیم ژنی بیان می‌کند که جایگاه‌های ژنی خاص بدون تغییر میانگین‌های درون محیط‌های مورد نظر، سبب تغییرات صفت بین محیط می‌شوند. این فرضیه‌ها متقابلاً منحصر به فرد نیستند، ولی تفاوت آن‌ها در اثر محیط بر بیان ژن‌های اثرگذار روی صفت قرار دارد: یا به‌طور مستقیم برای حساسیت آللی و یا به‌طور غیرمستقیم برای ژن‌های تنظیم کننده (شلیچتینگ و پیگلیوکسی، ۱۹۹۸).

این دو فرضیه عمومی می‌توانند در قالب مدل مختلط مکان‌یابی QTL آزمون شوند (وو، ۱۹۹۸؛ لیپز و مک کای، ۲۰۰۰). حساسیت آللی بیان می‌کند که آلل‌ها آثار متغیری روی فنوتیپ در محیط‌های مختلف دارند که این امر دلالت دارد بر اینکه QTL اثرگذار بر حساسیت یک صفت به تغییرات محیطی باید در نواحی ژنومی یکسانی مکان‌یابی شود تا تنوع ژنتیکی صفت در یک محیط را توجیه نماید. تظاهر متغیر آلل‌ها در این مناطق QTL در میان محیط‌ها، کوواریانس بین عملکرد صفت و محیط را توجیه خواهد کرد. تنظیم ژنی بیان می‌کند که ژن‌های تنظیم کننده خاص با روشن کردن یا تنظیم ژن‌های ساختاری که مستقیماً بر صفت اثر می‌گذارند، پاسخ می‌دهند؛ بنابراین، بر اساس این فرضیه، مناطق QTL که تنوع ژنی در حساسیت فنوتیپی را توضیح می‌دهند، از مناطقی که در تنوع درون یک محیط معین نقش دارند، متمایز خواهند بود.

۹-۶- برآورد در مدل مختلط

اساس آماری مکان‌یابی QTL ها در مدل مختلط (۹-۴) در یک جمعیت در حال تفرق است. برآورد پارامترهای ناشناخته موجود در مدل مختلط گام مهم بعدی برای یک پروژه مکان‌یابی است. روش‌های آماری زیادی وجود دارند که می‌توانند برآورد پارامترهای ناشناخته مربوط به مکان‌یابی QTL ها را تحت چارچوب مدل مختلط برآورد نمایند. در این بخش، روش حداکثر درستنمایی را با جزئیات آن توضیح می‌دهیم و بحث مربوط به روش‌های دیگر در بخش ۹-۷ ارائه می‌شود.

با استفاده از روش حداکثر درستنمایی برای برآورد پارامترها در مدل‌های مختلط، برآوردهای نقطه‌ای ($\hat{\omega}_{1|i}, \hat{\mu}_i, ..., \hat{\omega}_{J|i}, \hat{\mu}_J, \hat{\sigma}^2$) پارامترها با حداکثر کردن درستنمایی به‌دست می‌آیند:

---

[1] - Phenotypic plasticity
[2] - Plastic traits

$$\Omega = \left(\omega_{1|i}, \mu_1, ..., \omega_{J|i}, \mu_J, \sigma^2\right)$$

پارامترهای ناشناخته یعنی Ω شامل نسبت‌های اختلاط یعنی $\omega = \left(\omega_{1|i}, ..., \omega_{J|i}\right)$ که مخصوص جایگاه‌های QTL است، پارامترهای مدل یعنی $\Omega_m = \left(\mu_1, ..., \mu_J, \sigma^2\right)$ که مخصوص آثار QTL است و نیز واریانس باقیمانده خواهد بود.

یک مجموعه داده مشاهده شده برای یک پروژه مکان‌یابی (برای مثال جداول ۸-۱ و ۸-۲ را مشاهده کنید)، شامل فنوتیپ‌های صفت کمّی (y) و ژنوتیپ نشانگر (M) برای تمامی افراد تعیین ژنوتیپ شده است. یک مجموعه داده کامل شامل این مشاهدات به همراه ژنوتیپ‌های QTL (g) و موقعیت‌های QTL (d) است که داده‌های از دست رفته یا گمشده هستند.

درستنمایی پارامترهای ناشناخته مدل θ و نسبت‌های اختلاط ناشناخته یعنی ω (و بنابراین موقعیت‌های d ناشناخته QTL یعنی d) برای فرد i ام می‌تواند به‌صورت زیر بیان شود:

$$p\left(d, \theta \mid y_i, M_i\right) = \sum_{j=1}^{J} \omega_{j|i} \left(g_i \mid d, M_i\right) f_j \left(y_i \mid g_i, \theta\right) \qquad (۹-۱۳)$$

که برای تمامی ژنوتیپ‌های ممکنه فرد i ام جمع می‌شود. به‌طور واضح وابستگی وزن‌ها به g_i، d و M_i نشان داده شده است.

فرض کنید n فرزند مستقل در یک جمعیت مکان‌یابی وجود دارند. تابع درستنمایی مدل مختلط (۹-۱۳) برای بردار پارامترها یعنی Ω می‌تواند با محاسبه حاصل‌ضرب‌های لگاریتم چگالی‌های اختلاط در هر y_i به صورت زیر تشکیل شود:

$$L\left(d, \theta \mid y, M\right) = \prod_{i=1}^{n} p\left(d, \theta \mid y_i, M_i\right)$$

تقریباً همیشه ساده‌تر است که با لگاریتم تابع درستنمایی کار کنیم:

$$\log L\left(d, \theta \mid y, M\right) = \log \prod_{i=1}^{n} p\left(d, \theta \mid y_i, M_i\right)$$

$$= \sum_{i=1}^{n} \log p\left(d, \theta \mid y_i, M_i\right) \qquad (۹-۱۴)$$

$$= \sum_{i=1}^{n} \log \sum_{j=1}^{J} \left[\omega_{j|i} \left(g_i \mid d, M_i\right) f_j \left(y_i \mid g_i, \Omega_m\right)\right]$$

برآورد حداکثر درستنمایی Ω از راه ریشه مناسب معادله لگاریتم درستنمایی به‌دست می‌آید:

$$\frac{\partial}{\partial \Omega} \log L\left(d, \theta \mid y_i, M_i\right) = \sum_{i=1}^{n} \log \frac{\partial}{\partial \Omega} p\left(d, \theta \mid y_i, M_i\right) = 0 \qquad (۹-۱۵)$$

برای ساده کردن عبارت بالا، می‌توانیم بنویسیم: $\omega_{j|i}(g_i)=\omega_{j|i}(g_i\,|\,d,M_i)$ و $f_j(y_i)=f_j(y_i\,|\,g_i,\Omega_m)$ و داریم:

$$\log\frac{\partial}{\partial\Omega}p(d,\theta\,|\,y_i,M_i)=\frac{1}{p(d,\theta\,|\,y_i,M_i)}\frac{\partial}{\partial\Omega}\sum_{j=1}^{J}\omega_{j|i}(g_i)\,f_j(y_i)$$

$$=\sum_{j=1}^{J}\frac{\omega_{j|i}(g_i)\,f_j(y_i)}{p(d,\theta\,|\,y_i,M_i)}\frac{\partial}{\partial\Omega}\log\left[\omega_{j|i}(g_i)\,f_j(y_i)\right] \qquad (۱۶-۹)$$

$$=\sum_{j=1}^{J}\prod\nolimits_{j|i}\frac{\partial}{\partial\omega}\log\omega_{j|i}(g_i)+\sum_{j=1}^{J}\prod\nolimits_{j|i}\frac{\partial}{\partial\theta}\log f_j(y_i)$$

که در آن، $\prod_{j|i}$ به صورت زیر تعریف می‌شود:

$$\prod\nolimits_{j|i}=\frac{\omega_{j|i}(g_i)\,f_j(y_i)}{p(d,\theta\,|\,y_i,M_i)} \qquad (۱۷-۹)$$

که می‌تواند به صورت احتمال پسین ژنوتیپ j ام QTL برای فرد i ام در نظر گرفته شود. ازاین‌رو می‌توانیم الگوریتم EM با مجموعه پارامترهای بسط یافته یعنی $\{\Omega,\Pi\}$ را بکار ببریم که در آن: $\Pi=\left\{\prod_{j|i};\ j=1,\ldots,J,\ i=1,\ldots,n\right\}$ است.

برای حل معادلات (۹-۱۵)، معادلات (۹-۱۶) و (۹-۱۷) تکرار می‌شوند. با فرض اینکه $\Pi_{j|i}$ شناخته شده باشد، معادله (۹-۱۶) را برابر با صفر قرار داده و معادله را حل می‌کنیم. با استفاده از مقادیر به‌دست آمده برای ω و θ، مجدداً $\Pi_{j|i}$ را می‌نویسیم و این مراحل را تکرار می‌کنیم تا جایی که به همگرایی برسیم. این الگو می‌تواند به‌عنوان یک مثال کاربردی از الگوریتم EM در نظر گرفته شود (دمپستر و همکاران، ۱۹۷۷؛ منگ و روبین، ۱۹۹۳؛ همچنین بخش ۹-۷-۱ را مشاهده کنید).

زمانی که مدل مختلط برازش می‌شود، خوشه‌بندی احتمالی داده‌ها در قالب g خوشه می‌تواند بر حسب احتمالات پسین برازش شده برای هر یک اجزای داده‌ها یعنی $\prod_{j|i}\left(y_i\,|\,\hat{\Omega}\right)$ به‌دست آید که $\hat{\Omega}$ نشان‌دهنده برآورد حداکثر درستنمایی از Ω است. گروه‌بندی کامل داده‌ها در قالب J خوشه از راه انتساب هر یک از مشاهدات به جزئی که بالاترین احتمال پسین برآورد شده را دارا است، صورت می‌گیرد.

۹-۷- الگوریتم‌های محاسباتی برای مدل مختلط

در این بخش، برخی از موارد مربوط به حل معادلات درستنمایی بخش ۹-۶ را مرور می‌کنیم. همچنین برخی از جنبه‌های دیگر مربوط به برازش مدل‌های مختلط را مورد بحث قرار می‌دهیم.

۹-۷-۱- الگوریتم EM

هر تکرار شامل دو مرحله است. اولاً، در مرحله E، لگاریتم درستنمایی داده‌های کامل روی توزیع شرطی متغیرهای موهومی با شرط داده‌های مشاهده شده با استفاده از برآورد فعلی بردار پارامترها میانگین‌گیری می‌شود. چون لگاریتم درستنمایی داده‌های کامل در این متغیرهای موهومی به‌صورت خطی است (به‌عنوان یک نتیجه از فرض نرمال بودن توزیع داده‌ها)، مرحله E الگوریتم EM به سادگی با جایگزینی آن‌ها با مقادیر فعلی مورد انتظار شرطی آن‌ها، یعنی احتمالات پسین اعضای مؤلفه‌ای ارائه شده در معادله (۹-۱۷)، انجام می‌شود. سپس، در مرحله M، معادله (۹-۱۵) (معادلات درستنمایی) را با شرط Π برای مقادیر برابر با صفر حل می‌کنیم تا پارامترهای Ω را برآورد کنیم. معادله درستنمایی می‌تواند به دو عبارت موجود در معادله (۹-۱۶) تجزیه شود. عبارت اول به پیوستگی ژنتیکی بین مکان‌های ژنی (همانگونه که در یک تلاقی کنترل شده در نظر گرفته شد) و یا به ارتباط ژنتیکی بین مکان‌های ژنی (همانگونه که در یک جمعیت طبیعی دیدیم)، باز می‌گردد و عبارت دوم به رابطه بین فنوتیپ-ژنوتیپ مربوط می‌شود. سپس این برآوردها برای به‌دست آوردن مقدار جدید Π در مرحله E استفاده می‌شوند و این فرآیند (مراحل E و M) تا رسیدن به همگرایی تکرار می‌شود. مقادیر به‌دست آمده در هنگام همگرایی همان برآوردهای حداکثر درستنمایی پارامترها هستند.

الگوریتم EM دارای همگرایی قابل اعتمادی است که در آن، بدون توجه به نقطه آغازین، درستنمایی (۹-۱۵) پس از هر تکرار EM افزایش می‌یابد و این همگرایی یک حداکثر محلی (یا یک نقطه ایستگاهی) است.

۹-۷-۲- EM مونت کارلو

در هر دور از الگوریتم EM، معادله درستنمایی می‌تواند با استفاده از تعداد N جستجوی مونت کارلو برآورد شود:

$$\frac{\partial}{\partial\Omega}\log L\left(d,\theta\,|\,y,M\right)\triangleq\frac{1}{N}\sum_{k=1}^{N}\frac{\partial}{\partial\Omega}\sum_{i=1}^{n}\log\pi_j^{(k)}\left(g_i\right)+\frac{1}{N}\sum_{k=1}^{N}\frac{\partial}{\partial\Omega}\sum_{i=1}^{n}\log f_j^{(k)}\left(y_i\right)$$

که در k نمونه مونت کارلو، QTL (ژنوتیپ‌های j) با شرط y و M و برآورد فعلی پارامتر ایجاد می‌شوند. برای نمونه‌های ژنوتیپی نسبتاً بزرگ، EM مونت کارلو، خصوصیات EM حقیقی را حفظ

خواهد کرد. نمونه‌های مونت کارلو می‌توانند برای برآورد نسبت درستنمایی در مرحلـه نهـایی EM نیز استفاده شوند.

۹-۷-۳- EM اتفاقی

در هر دور از الگوریتم EM، معادله درستنمایی می‌تواند با استفاده از یک بررسی منفرد مونت کارلو برآورد شود:

$$\frac{\partial}{\partial\Omega}\log L\left(d,\theta\mid y,M\right)\hat{=}\frac{\partial}{\partial\Omega}\log\sum_{i=1}^{n}\omega_{j|i}^{(k)}\left(g_i\right)+\frac{\partial}{\partial\Omega}\log\sum_{i=1}^{n}f_j^{(k)}\left(y_i\right)$$

که در k امین دور از EM یک ژنوتیپ منفرد QTL (ژنوتیپ j) بـا شـرط y و m و برآوردهـای فعلی پارامترها با استفاده از توزیع $\Pi^{(k)}$ ایجاد می‌شـود. ایـن عبـارت مـی‌توانـد بـه‌عنـوان معادلـه درستنمایی استاندارد در نظر گرفته شود. توزیع پسـین بـرآورد پارامترهـا در طـی تعـداد زیـادی از دورهای EM به‌دست می‌آیند و پس از یک دوره سوختن[1] مناسب به‌طور تقریبی در برآورد حداکثر درستنمایی متمرکز می‌شوند و میانگین توزیع می‌تواند به عنوان بـرآورد ML در نظـر گرفتـه شـود (سلیوکس و دیبولت، ۱۹۸۵). این توزیع پسین می‌تواند برای پـارامتر مـورد نظـر نیـز رسـم شـود. همچنین، یک مرحله کوتاه مقدماتی از EM اتفاقی می‌تواند برای بـه‌دسـت آوردن مقـادیر آغـازین برای روش EM مونت کارلو اجرا شود.

۹-۷-۴- یک الگوریتم EM/آمیخته نیوتن-رافسون[2]

برای برخی از پارامترهای QTL که با استفاده از الگوریتم EM برآورد می‌شوند، معـادلات حـداکثر درستنمایی حاصل به دلیل ساختارهای غیرخطی پیچیده‌ای که دارند، ممکن است به آسـانی قابـل حل نباشند. این پارامترها می‌توانند به کمک روش‌های عددی دیگـر بـرآورد شـوند. مسـتقیم‌تـرین روش، ارزیابی ساده درستنمایی برای دامنه‌ای از مقادیر پارامترهـای مـورد نظـر و انتخـاب مقـداری است که بیشترین درستنمایی را ایجاد نماید. این رویه جستجوی شـبکه‌ای[3] مـی‌توانـد بـه صـورت کاملاً پیچیده انجام شود، به این صورت که پارامتر می‌تواند به دهـک‌هـایی تقسـیم شـده و سـپس دهکی که حاوی برآورد حداکثر درستنمایی است، می‌تواند به ده قسمت تقسیم شود. چون بسیاری از دورهای جستجو که می‌توانند به صورت ارقام معنی‌دار باشند، در جواب‌ها مورد نیاز هستند. آنچه که سودمند است آن است که نمودار درستنمایی را ترسیم کنـیم تـا نشـان دهـیم کـه چگونـه بـه

---

[1] - Burning period
[2] - Newton-Raphson Hybrid
[3] - Grid search

تغییرات مقادیر پارامتر پاسخ نشان داده و چگونه در برابر یک حداکثر محلی که با یک حداکثر عمومی مغشوش می‌شود، محافظت می‌گردد.

یک روش جایگزین با استفاده از تکرارهای نیوتن-رافسون ارایه شده است که ترکیب آن با الگوریتم EM یک روش آمیخته را برای برآورد پارامترهای مورد نظر در معادلات غیرخطی لگاریتم درستنمایی تشکیل می‌دهد. به‌طور مختصر، برخی از مقادیر ابتدایی برای برآورد انتخاب می‌شوند و سپس این مقادیر با استفاده از تابع امتیاز تغییر می‌یابند. مقدار تغییر یافته به نوبت تغییر می‌یابد و این فرایند تا زمانی ادامه می‌یابد که تکرارهای متوالی به میزان کمتر از برخی از مقادیر ویژه و اختصاصی با یکدیگر تفاوت داشته باشند (برای توصیف مناسب این روش ویر، ۱۹۹۶ را مشاهده کنید).

برای پارامتر φ، MLE مورد نیاز به صورت $\hat{\varphi}$ و مقدار ابتدایی یا فرضی به صورت φ' نشان داده می‌شود. امتیاز S_φ یعنی مشتق لگاریتم درستنمایی نسبت به φ در نقطه $\hat{\varphi}$ برابر با صفر است و آن امتیاز می‌تواند با قضیه تیلور[1] بدین صورت بسط داده شود:

$$S_{\hat{\varphi}} = 0 = S_{\varphi'} + \left(\hat{\varphi} - \varphi'\right)\left[\frac{\partial S_\varphi}{\partial \varphi}\right]_{\varphi=\varphi'}$$

که عبارات با درجه بالاتر در $\left(\hat{\varphi} - \varphi'\right)$ نادیده گرفته می‌شوند. مرتب کردن مجدد این عبارت مقدار تقریبی φ'' را برای $\hat{\varphi}$ ایجاد می‌کند:

$$\varphi'' = \varphi' - S_{\varphi'} / \left[\frac{\partial S_\varphi}{\partial \varphi}\right]_{\varphi=\varphi'} = \varphi' + S_{\varphi'} / I\left(\varphi'\right)$$

که در آن، $I\left(\varphi\right) = \left[\frac{\partial S_\varphi}{\partial \varphi}\right]_{\varphi=\varphi'}$ ماتریس اطلاعات است.

مقدار ابتدایی φ' با افزودن امتیاز تقسیم شده به وسیله اطلاعات که هر دو در مقادیر ابتدایی ارزیابی شده‌اند، تغییر داده می‌شود. سپس مقدار جدید به عنوان مقدار ابتدایی جهت اعمال تغییر بیشتر مورد استفاده قرار می‌گیرد:

$$\varphi''' = \varphi'' + S\left(\varphi''\right) / I\left(\varphi''\right)$$

و تکرار مراحل تا رسیدن به همگرایی ادامه می‌یابد.

چون اطلاعات $I\left(\varphi\right)$ به‌طور معمول یک ماتریس است، روش تکرار نیاز به معکوس ماتریس دارد:

$$\varphi'' = \varphi' + I^{-1}\left(\varphi'\right) S\left(\varphi'\right)$$

---

[1] - Taylor's theorem

به‌طور واضح، این روش زمانی قابل انجام نیست که اطلاعات برابر صفر بوده یا ماتریس اطلاعات، یک ماتریس منفرد یا ویژه[۱] باشد.

۹-۷-۵- برخی از جنبه‌های احتیاطی

گاهی اوقات ممکن است در هنگام استفاده از روش حداکثر درستنمایی برای مدل‌های مختلط QTL مشکلاتی وجود داشته باشد. اولاً، برای برخی از انتخاب‌های خانواده‌های پارامتری، درستنمایی می‌تواند بدون حد و مرز و بی پایان باشد. ثانیاً، در شرایط پیچیده، تابع درستنمایی می‌تواند تعداد زیادی نقاط حداکثری محلی داشته باشد که هر یک از آن‌ها ممکن است برآوردهای مختلفی (و احتمالاً معقول) برای مقادیر مورد نظر ارائه دهند. در این موارد، انتخاب یکی از این برآوردهای نقطه‌ای پارامترها در مقابل سایر برآوردها مشکل است.

ثالثاً، مهم‌تر از سایر نکات، زمانی که یک روش مکان‌یابی QTL با خصوصیات ژنتیکی ژن‌های جمعیت تلفیق می‌شود، با افزایش فضای ابعادی برای پارامترهای ناشناخته مواجه خواهیم شد. اگرچه فقط وجود فضای پارامتری با ابعاد بزرگ لزوماً مضر نیست، اما برای جستجوی برآوردگر ML باید مراقبت بیشتری صورت گیرد. پیچیدگی اضافی (نه فقط ML) از آن جهت است که عدم قطعیت در خصوص تعداد واقعی QTL های صفات کمّی منجر به مشکلات اضافی در زمینه برازش و انتخاب مدل می‌شود.

۹-۷-۶- روش‌های بیزی

استفاده از روش بیزی می‌تواند از بروز برخی از مشکلات مرتبط با روش حداکثر درستنمایی جلوگیری نماید. در روش حداکثر درستنمایی، پارامترهای ناشناخته به عنوان متغیرهای ناشناخته (متغیرهای غیر قابل مشاهده) شناخته می‌شوند و تابع درستنمایی در این متغیرها به حداکثر می‌رسد. در روش بیزی، هر پارامتر غیر قابل مشاهده دارای یک توزیع پیشین است و سپس توزیع پسین هر متغیر غیر قابل مشاهده به شرط داده‌ها (متغیرهای قابل مشاهده) استنباط می‌شود. خلاصه آماره‌های توزیع پسین (برای مثال، میانگین، مد یا میانه) می‌تواند به عنوان برآوردهای بیزی متغیرهای غیر قابل مشاهده در نظر گرفته شوند (کارلین و لوئیس، ۱۹۹۸). برآورد فاصله‌ای نیز می‌تواند با آزمون توزیع پسین به‌دست آید.

---

[۱]- ماتریس‌های ویژه یا منفرد ماتریس‌هایی هستند که دترمینان برابر با صفر دارند و در نتیجه این‌گونه ماتریس‌ها معکوس ندارند (مترجمین).

با استفاده از نمادسازی مشابه روش حداکثر درستنمایی، متغیرهای مشاهده شده با y (داده‌های فنوتیپی) و M (داده‌های نشانگر) و متغیرهای غیرقابل مشاهده با θ (پارامترهای مدل)، d (جایگاه QTL ها)، g (ژنوتیپ‌های QTL) و $\Omega_m = \left(\mu_1,...,\mu_J,\sigma^2\right)$ یعنی آثار QTL و واریانس باقیمانده نشان داده می‌شوند. توزیع پسین متغیرهای غیر قابل مشاهده با رابطه زیر نشان داده می‌شود:

(۱۸-۹)

$$\pi\left(d,\Omega_m,g \mid y,M\right) = \frac{\pi\left(y,M,d,\Omega_m,g\right)}{\pi\left(y,M\right)}$$

$$= \frac{p\left(y,M \mid g,\Omega_m\right)\pi\left(g \mid d\right)\pi\left(d,\Omega_m\right)}{\pi\left(y\right)} \propto \pi\left(y,M \mid g,\Omega_m\right)\pi\left(g \mid d\right)\pi\left(d,\Omega_m\right)$$

که در آن، $\pi(.)$ عبارت کلی برای چگالی احتمال، $p\left(y,M \mid g,\Omega_m\right)$ چگالی احتمال داده‌ها با شرط ژنوتیپ‌های QTL، $\pi\left(g \mid d\right)$ چگالی احتمال ژنوتیپ‌های QTL تمامی مشاهدات مشروط به موقعیت‌های آن‌ها یعنی d و $\pi\left(d,\Omega_m\right)$ توزیع احتمال پیشین متغیرهای غیر قابل مشاهده است. چون مخرج کسر تابعی از پارامترها نیست، بنابراین می‌تواند نادیده گرفته شود.

استقلال پیشین پارامترها به‌صورت زیر در نظر گرفته می شود:

$$\pi\left(d,\Omega_m\right) = \pi\left(d\right)\pi\left(\mu\right)\pi\left(\sigma^2\right)\prod_{k=1}^{\dot{n}}\left[\pi\left(a_k\right)\pi\left(d_k\right)\right]$$

توجه کنید که k نشان‌دهنده k امین QTL و $\dot{n}$ نشان‌دهنده تعداد QTL ها نسبت به تعداد ژنوتیپ‌های QTL است. میزان تأثیر انتخاب نوع توزیع پیشین در فضای پارامتری بر استنباط نهایی به‌عنوان معیاری از قدرت و توانمندی محسوب شده و برای هر کاربردی نیازمند کنترل است. توزیع پیشین می‌تواند بر اساس مطالعات یا اطلاعات مرتبط از منابع انتخاب شود. به‌طورکلی، زمانی که هیچ اطلاعاتی در ارتباط با موقعیت QTL ها در دسترس نباشد، انتخاب طبیعی برای مقدار اولیه d دارای توزیع یکنواخت است. تخصیص مقادیر ابتدایی توأم برای μ، a_k، d_k و σ^2 سبب ساده شدن شکل توزیع می‌شود، در حالی که افزایش پراکنش‌ها سبب کاهش اثر مقدار اولیه می‌شود.

در روش بیزی، پارامترهای ژنتیکی بر اساس توزیع پسین حاشیه‌ای استنباط می‌شوند که می‌تواند از روی توزیع پسین توأم (۱۸-۹) با تلفیق تمامی عوامل ناشناخته دیگر به‌دست آید. می‌توانیم بردار $\Omega = \left(d,g,\Omega_m\right)$ را به $\Omega = \left[\Omega_l \Omega_{-l}\right]$ تقسیم کنیم که Ω_l عنصر واحدی از

متغیرهای غیر قابل مشاهده و Ω_{-l} باقیمانده متغیرهای غیر قابل مشاهده است که فاقد Ω_l است. توزیع پسین حاشیه ای Ω_l با عبارت زیر بیان می‌شود:

(۹-۱۹)

$$\Pi(\Omega_l \mid y) = \int \pi(\Omega_l, \Omega_{-l} \mid y)\, d\Omega_{-l} \propto \int \pi(y \mid \Omega_l, \Omega_{-l}) \pi(\Omega_l, \Omega_{-l})\, d\Omega_{-l}$$

میانگین این توزیع پسین حاشیه‌ای برآوردگر بیزی کاندید برای Ω_l است. این توزیع حاشیه‌ای به ندرت دارای یک شکل واضح است و غالباً از انتگرال‌گیری عددی اجتناب می‌شود، چون ابعاد Ω_{-l} ممکن است زیاد باشد. ولی توزیع (۹-۱۹) می‌تواند با ایجاد آن به‌وسیله الگوریتم زنجیره مارکوف مونت کارلو (نظیر نمونه بردار گیبس) در فضای پارامتری تقریب زده شود.

زنجیره مارکوف یک توالی تصادفی از حالات مختلف است:

$$\left\{\left(d^0, g^0, \Omega_m^0\right), \left(d^1, g^1, \Omega_m^1\right), \ldots, d^N, g^N, \Omega_m^N\right\}$$

که از یک نقطه اختیاری $\left(d^0, g^0, \Omega_m^0\right)$ آغاز شده و دارای چگالی پسین مثبت بوده و از قوانین ساده‌ای تبعیت می‌کند که سه ناشناخته d، g و Ω_m را تغییر می‌دهند. هر مرحله در این زنجیره، چرخه‌ای از سه مرحله کوچک‌تر است که در اولین مرحله، اطلاعات d و سپس g و در نهایت Ω_m تجدید و نوسازی می‌شوند. مرحله‌ای که در آن d نوسازی می‌شود، معمولاً یک مرحله متروپولیس-هاستینگز[1] است (هاستینگز، ۱۹۷۰)، در حالی که مراحل نوسازی اطلاعات g و Ω_m همان مراحل نمونه‌گیری گیبس هستند (گمان و گمان، ۱۹۸۴). برای بررسی کامل الگوریتم‌های زنجیره مارکوف مونت کارلو (MCMC)، روبرت و کاسلا (۲۰۰۴) را مشاهده کنید.

الگوریتم MCMC با این واقعیت توجیه می‌شود که اگر h هر تابعی از ناشناخته‌ها باشد که در ارتباط با توزیع تعادلی π به صورت مربع و قابل انتگرال‌گیری باشد، سپس:

$$\overline{h}_N = \frac{1}{N}\sum_{t=1}^{N} h\left(d^t, g^t, \Omega_m^t\right) \to E_\pi\left[h\left(d, g, \Omega_m\right) \mid y\right] \qquad (۹-۲۰)$$

به‌طور حتم وقتی که $N \to \infty$

که در آن، $\left(d^t, g^t, \Omega_m^t\right)$ نمونه‌هایی از زنجیره مارکوف هستند. میانگین‌های تجربی نمونه‌های MCMC حاصل از معادله (۹-۲۰) می‌توانند برای به‌دست آوردن برآوردگر بیزی پارامترهای ناشناخته $\left(d, g, \Omega_m\right)$ مورد استفاده قرار گیرند. چگالی‌های پسین حاشیه‌ای پارامترها نیز از روی مقادیر نمونه حاصل می‌شوند.

---

[1] - Metropolis-Hastings

۹-۷-۷- برآورد تعداد اجزاء در یک مدل مختلط

در بسیاری از موارد عملی، تعداد ژنوتیپ‌های QTL (و بنابراین تعداد QTL ها) در یک مدل مختلط ناشناخته است. برآورد تعداد QTL های یک صفت کمّی مسئله مهمی در تحقیقات ژنتیک کمّی است. از نظر آماری، سنجش تعداد QTL ها معادل آزمون تعداد مؤلفه‌ها در یک توزیع مختلط است. این مسئله دشوار هنوز به خوبی در منابع آماری حل نشده است (مقاله مروری لو و همکاران، ۲۰۰۱ را مشاهده کنید).

از نقطه نظر تجزیه درستنمایی، تعیین تعداد مؤلفه‌ها در یک مدل مختلط می‌تواند بر اساس آزمون فرض صفر: یک نمونه تصادفی متعلق به J_0 مؤلفه مختلط است، در برابر فرض مقابل: نمونه به J_1 مؤلفه مختلط تعلق دارد، بررسی شود که $J_1 > J_0$ است. استفاده از آزمون نسبت درستنمایی که تحت فرض صفر به‌طور مجانبی دارای توزیع کای مربع است گمراه‌کننده است. در هر حال، نظریه کلاسیک کای مربع مجانبی در خصوص آزمون نسبت درستنمایی در مبحث مدل‌های مختلط جایگاهی ندارد. لو و همکاران (۲۰۰۱) پیشنهاد کردند که استفاده از روش آزمون نسبت درستنمایی بر اساس معیار اطلاعات کولباک-لیبلر[1] و با قضیه پیشنهاد شده به وسیله وونگ (۱۹۸۹) صورت گیرد. این نویسندگان نشان دادند که تحت شرایط معمول، توزیع محدودکننده آماره نسبت درستنمایی، در زمانی که مدل‌های مورد مقایسه به‌صورت آشیانه‌ای بوده یا در حال هم‌پوشانی هستند، به صورت مجموع وزنی متغیرهای تصادفی x_1^2 بوده و در هنگامی‌که مدل‌های مورد مقایسه به صورت غیر آشیانه‌ای باشند، توزیع نرمال است.

روش دیگر برای تشخیص تعداد QTL ها بر اساس معیارهای انتخاب مدل بیزی است. این روش از نمونه‌های حاصل از زنجیره‌های مارکوف برای محاسبه عامل بیزی استفاده می‌کند (جفریز، ۱۹۶۱؛ کاس و رافتری، ۱۹۹۵).

برای مقایسه دو مدل M_1 و M_2 فرض می‌شود که دو تعداد متفاوت از QTL ها در مبحث مدل مختلط برازش می‌یابند. نسبت احتمال پسین M_1 بر M_2 برای داده‌های y می‌تواند با استفاده از قضیه بیز به‌صورت زیر بیان شود:

$$\frac{\Pr(M_1 \mid y)}{\Pr(M_2 \mid y)} = \frac{\Pr(M_1)\Pr(y \mid M_1)}{\Pr(M_2)\Pr(y \mid M_2)}$$

که اولین عامل نسبت احتمال پیشین است و:

$$B = \frac{\Pr(M_1 \mid y)}{\Pr(M_2 \mid y)}$$

---

[1] - Kullback-Leibler

نسبت احتمالات حاشیه‌ای y مشروط به دو مدل است و عامل بیزی[1] نامیده می‌شود. نیوتن و رافتری (۱۹۹۴) و کاس و رافتری (۱۹۹۵) روش‌هایی را برای برآورد احتمال حاشیه‌ای داده‌ها تحت هر دو مدل فوق بحث نمودند.

گرین (۱۹۹۵) و ریچاردسون و گرین (۱۹۹۷) روش MCMC برگشت‌پذیر پرشی[2] را برای شبیه‌سازی توزیع پسین تعداد مؤلفه‌ها پیشنهاد دادند. این روش برای برآورد تعداد QTL در یک آزمایش مکان‌یابی به‌کار گرفته شده است (استفانز و فیش، ۱۹۹۸). روش‌های پرشی برگشت‌پذیر ایجاد یک زنجیره مارکوف ارگودیک[3] را با توزیع پسین توأم پارامترها و مدل به عنوان توزیع ثابت برای آن اجازه می‌دهند. حرکات بین مدل‌ها با فرض نمودن دوره‌ای یک حرکت به یک مدل مختلف و رد نمودن آن با احتمال مناسب، جهت تضمین اینکه زنجیره‌ای با توزیع بدون تغییر مورد نیاز است، به‌دست می‌آید. به‌طور ایده آل، این حرکات مفروض بگونه‌ای طرح‌ریزی می‌شوند تا احتمال زیاد پذیرش را دارا باشند به‌طوری که الگوریتم، مدل‌های مختلف را به حد کفایت جستجو می‌کند، اگرچه دستیابی به این حالت همیشه از نظر عملی امکان‌پذیر نیست.

---

[1] - Bayes factor
[2] - Reversible jump MCMC
[3] - Ergodic Markov chain

۹-۸- تمرین‌ها

۹-۱- سه وظیفه عمده مکان‌یابی QTL بر اساس مدل مختلط چیست؟

۹-۲- در ارجاع به مثال ۹-۱، از معادله (۹-۸) و داده‌های موجود در جدول ۹-۳ برای تأیید برآوردهای اثر ارایه شده در مثال استفاده کنید. از هر دو ماتریس D استفاده کرده و محاسبات را برای میانگین‌های حاشیه‌ای و میانگین‌های توأم انجام دهید.

۹-۳- مشابه تمرین ۹-۲، مقادیر برآوردهای اثر ارایه شده در مثال ۹-۲ را تأیید کنید.

۹-۴- چگونه می‌توانید تعداد QTL ها را برای یک صفت پیچیده در مفهوم مدل مختلط تعیین کنید؟

فصل دهم

مکان‌یابی فاصله‌ای با استفاده از تجزیه تابعیت

۱۰-۱- مقدمه

تجزیهٔ ژنتیکی صفات کمّی شامل دو مرحله اصلی است: ۱- شناسایی جایگاه ژنومی QTL های مؤثر بر صفت کمّی با استفاده از نقشه پیوستگی ژنتیکی ایجاد شده از نشانگرهای ژنتیکی و ۲- برآورد آثار ژنتیکی QTL ها روی فنوتیپ. اگر ژنوتیپ‌های یک QTL فرضی برای تمامی افراد شناخته شده بود، جایگاه ژنومی آن می‌توانست به آسانی با استفاده از تجزیه پیوستگی نشانگر تعیین شود. علاوه بر آن، آثار ژنتیکی QTL دقیقاً برآورد شده و به وسیله آزمون‌های ساده t یا ANOVA آزمون شوند؛ اما امکان مشاهده مستقیم ژنوتیپ‌های QTL وجود ندارد و در مقابل، آن‌ها را باید از روی اطلاعات فنوتیپی و نشانگر مشاهده شده استنباط نمود. همانگونه که در فصل ۸ مشاهده شد، تجزیهٔ یک نشانگر نمی‌تواند به‌صورت واضحی آثار ژنتیکی یک QTL را از نرخ نوترکیبی بین نشانگرها و QTL جدا نماید.

برای تعیین اثر ژنتیکی و موقعیت QTL، تجزیه آماری پیشرفته‌تری باید انجام شود. ایده متمرکز برای برآورد انفرادی اثر و موقعیت QTL، فرموله کردن یک مدل آماری برای داده‌های فنوتیپی و نشانگر مشاهده شده بر حسب QTL مورد نظر است که بین دو نشانگر مجاور هم قرار دارد. این روش که مکان‌یابی فاصله‌ای نامیده می‌شود، می‌تواند بر مشکل اختلاط نرخ نوترکیبی نشانگر-QTL و آثار QTL از راه احتمالات شرطی ژنوتیپ‌های QTL ناشناخته مشروط به ژنوتیپ‌های مشاهده شده نشانگر غلبه نماید. QTL و دو نشانگری که آن را احاطه کرده‌اند، اجازه کاربرد تجزیه سه نقطه‌ای را برای به‌دست آوردن احتمالات شرطی در یک جمعیت در حال تفرق ویژه نظیر تلاقی برگشتی، RIL یا F_2 می‌دهند.

مدل‌های مکان‌یابی فاصله‌ای می‌توانند بر اساس مفهوم نظریه‌های تابعیت یا حداکثر درستنمایی (ML) فرموله شوند. از راه روش‌های تجزیه‌ای و عددی، کائو (۲۰۰۰) تفاوت‌های بین مکان‌یابی فاصله‌ای مبتنی بر تجزیه تابعیت یا ML را بر حسب میانگین مربعات خطای برآورد پارامترها و توان شناسایی QTL مقایسه کرد. این تفاوت‌ها زمانی افزایش می‌یابند که موارد زیر صادق باشند: ۱- سهم زیادی از واریانس صفت به وسیله QTL توجیه می‌شود، ۲- QTL در ناحیه میانی فاصله نشانگری قرار دارد، ۳- فاصله نشانگری بزرگ است، ۴- بین QTL های مختلف آثار متقابل اپیستازی وجود دارد، ۵- QTL ها دارای آثار متفاوتی هستند و ۶- QTL های مختلف پیوسته هستند. در عمل، روش تابعیت از نظر محاسباتی کارآتر از روش ML است، به ویژه زمانی که تعداد QTL های در نظر گرفته شده زیاد باشند.

در این فصل، اصول روش‌های مکان‌یابی فاصله‌ای مبتنی بر تجزیه تابعیت، استنباط‌های آماری آن‌ها در خصوص برآورد پارامترها و آزمون فرض و رویه کاربرد عملی آن‌ها معرفی خواهند شد.

مثال‌هایی برای نشان دادن مزیت مکان‌یابی QTL با تجزیه تابعیت ذکر خواهد شد. در فصل ۱۱، روش‌های مکان‌یابی فاصله‌ای QTL بر اساس حداکثر درستنمایی تشریح خواهد شد.

۱۰-۲- مدل تابعیت خطی

مدل خطی QTL مشروط به ژنوتیپ‌های یک فاصله نشانگری برای جمعیت تلاقی برگشتی می‌تواند به‌صورت زیر بیان شود:

$$y_i = \mu + x_{j|i}a + e_i \qquad (۱-۱۰)$$

که a اثر واقعی QTL و $x_{j|i}$ متغیر شاخص یا موهومی است که به صورت احتمال شرطی ژنوتیپ j ام QTL مشروط به ژنوتیپ نشانگر در فرزند i ام تعریف می‌شود. بر اساس معادله (۱-۱۰)، وقتی که متغیر شاخص یا موهومی مشخص شود، اثر QTL (یعنی a) می‌تواند برآورد شود.

مدل تابعیت شرطی برای جمعیت F_2 باید به‌صورت زیر فرموله شود:

$$y_i = \mu + x_{1j|i}a + x_{2j|i}d + e_i \qquad (۲-۱۰)$$

که $x_{1j|i}$ و $x_{2j|i}$ متغیرهای موهومی یا شاخصی هستند که احتمالات شرطی آثار ژنتیکی افزایشی و غالبیت QTL را مشروط به ژنوتیپ فاصله نشانگری در فرزند i ام تعیین می‌کنند.

روش تجزیه تابعیت برای مکان‌یابی QTL ها رگرسیون یا تابعیت ارزش‌های فنوتیپی یک صفت کمّی بر ارزش‌های ژنوتیپی مورد انتظار شرطی و برآورد پارامترهای ناشناخته با استفاده از روش حداقل مربعات کلاسیک است. ارزش‌های ژنوتیپی مورد انتظار شرطی مرتبط با هر ژنوتیپ نشانگر از روی احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ نشانگر و از روی ارزش‌های ژنوتیپی ژنوتیپ‌های مختلف QTL محاسبه می‌شوند.

۱۰-۳- مکان‌یابی فاصله‌ای در جمعیت تلاقی برگشتی

۱۰-۳-۱- احتمالات شرطی

در حالی‌که جدول ۹-۱ جدول QTL و جدول ۸-۱ جدول نشانگر برای یک جمعیت تلاقی برگشتی با ۱۰ موش است، جدول ۱۰-۱ اطلاعات این دو جدول را تلفیق کرده است. فرض کنید یک QTL فرضی وجود دارد که به وسیله دو نشانگر به هم پیوسته M و N احاطه شده است. برای یک ژنوتیپ نشانگر قابل مشاهده، دو امکان برای حمل یک ژنوتیپ QTL یعنی $Qq\,(1)$ یا $qq\,(0)$ وجود دارد. همچنین جدول ۱۰-۱ ژنوتیپ‌های نشانگر-QTL-نشانگر را نیز نشان

می‌دهد. فرض کنید $\omega_{1|i}$ و $\omega_{0|i}$ احتمالات شرطی دو ژنوتیپ QTL مشروط بـه یـک ژنوتیـپ دو نشانگری برای موش i ام باشند.

مقادیر $\omega_{1|i}$ و $\omega_{0|i}$ به نرخ‌های نوترکیبی بـین دو نشـانگر (r)، بـین نشـانگر M و QTL (r_1) و بین QTL و نشانگر N (r_2) بستگی دارند. از راه تلاقی برگشتی هتروزیگوت سه‌گانه F_1 با یک والـد، هشت ژنوتیپ سه نقطه‌ای یا سه‌مکانی (نشانگر-QTL-نشانگر) ایجاد خواهد شد: ۱۱۱، ۱۰۱، ۱۱۰، ۱۰۰، ۰۱۱، ۰۰۱، ۰۱۰ و ۰۰۰. همانگونه که با تجزیه سه نقطه‌ای در فصل چهارم نشـان داده شـد، این فراوانی‌های ژنوتیپی با فرض عدم وجود کراس‌اور مضاعف به‌دست مـی‌آینـد و در جـدول ۱۰-۲ بیان شده‌اند؛ بنابراین، احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ‌های نشانگر در فاصله مربوطه در جمعیت تلاقی برگشتی می‌تواند بر اساس قضیه بیز به‌دست آینـد کـه در جـدول ۱۰-۳ بیان شده‌اند.

جدول ۱۰-۱- ژنوتیپ‌های QTL، ژنوتیپ‌های فاصله نشانگری، ژنوتیپ‌های توأم نشانگر-QTL و وزن بدن موش در یک طرح تلاقی برگشتی.

موش	QTL (z_i)	فاصله M-N (x_i^*)	ژنوتیپ M-QTL-N	احتمال (QTL به شرط فاصله نشانگر) $\left(x_{j\|i}\right)$	وزن بدن (y_i)
1	1	11	111 101	$\omega_{1\|11}$	30
2	1	11	111 101	$\omega_{1\|11}$	32
3	1	11	111 101	$\omega_{1\|11}$	28
4	1	11	111 101	$\omega_{1\|11}$	29
5	۱	10	110 100	$\omega_{1\|10}$	29
6	0	01	001 011	$\omega_{0\|01}$	22
7	0	00	000 010	$\omega_{0\|00}$	20
8	0	00	000 010	$\omega_{0\|00}$	21
9	0	00	000 010	$\omega_{0\|00}$	20
10	0	00	000 010	$\omega_{0\|00}$	21

توجه: ستون اول ژنوتیپ‌های QTL، ستون دوم ژنوتیپ‌های فاصله مربوط به دو نشانگر M و N است که QTL را احاطه کرده‌اند و ستون سوم، ژنوتیپ‌های سه نقطه‌ای (نشانگر-QTL-نشانگر) است.

اگر دو نشانگر پیوستگی زیادی داشته باشند، می‌توان فرض زیر را برای ساده کردن روابط در نظر گرفت:

$$r_1 + r_2 \approx r \qquad (۳-۱۰)$$

که از این واقعیت تبعیت می‌کند که $r = r_1 + r_2 - 2r_1r_2$ و فرض بر این است که $r_1r_2 \approx 0$ و بنابراین، جدول ۱۰-۳ می‌تواند به وسیله جدول ۱۰-۴ تقریب زده شود. برای سه جایگاه ژنی با ترتیب M-QTL-N، از g_{11} (بخش ۳-۶ را مشاهده کنید) برای نشان دادن فراوانی نرخ‌های نوترکیبی دوگانه (یک نوترکیبی در هر یک از دو فاصله M-QTL و QTL-N) استفاده می‌شود. همان‌گونه که با معادله (۳-۱۹) نشان داده شد، $g_{11} = \frac{1}{2}\left(r_1 + r_2 - r\right)$ است؛ بنابراین، اگر هیچ نوترکیبی دوگانه‌ای وجود نداشته باشد، معادله (۱۰-۳) را خواهیم داشت؛ به عبارت دیگر، برای یک نقشه بسیار متراکم و اشباع شده، می‌توانیم فرض کنیم که هیچ نوترکیبی دوگانه‌ای بین فواصل نزدیک به هم رخ نمی‌دهد.

۱۰-۳-۲- مدل تابعیت شرطی

با جایگزینی احتمالات شرطی جدول ۱۰-۳ یا جدول ۱۰-۴ در جدول ۱۰-۱ می‌توانیم متغیر وابسته $x_{j|i}$ را برای مدل تابعیت شرطی (۱۰-۱) ایجاد کنیم. به‌طورکلی، برای تلاقی برگشتی با اندازه n، معادله (۱۰-۱) را به شکل ماتریسی به‌صورت زیر می‌نویسیم:

$$y = Xb + e \qquad (۴-۱۰)$$

که در آن، $y = \left(y_i\right)_{n\times 1}$، $b = \left(\mu, a\right)^T$، $e = \left(e_i\right)_{n\times 1}$ بوده و داریم:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix}_{n\times 2}$$

و زمانی که ماتریس (۱۰-۳) در نظر گرفته شود، داریم:

$$X_1 = \left(1 \frac{(1-r_1)(1-r_2)}{1-r}\right)_{n_1 \times 1}$$

$$X_2 = \left(1 \frac{(1-r_1) r_2}{r}\right)_{n_2 \times 2}$$

$$X_3 = \left(1 \frac{r_1 (1-r_2)}{r}\right)_{n_3 \times 1}$$

$$X_4 = \left(1 \frac{r_1 r_2}{1-r}\right)_{n_4 \times 1}$$

که در آن، n_1، n_2، n_3 و n_4 به ترتیب اندازه نمونه‌ها برای چهار ژنوتیپ مختلف نشانگر یعنی ۱۱، ۱۰، ۰۱ و ۰۰ هستند. در مثال خاصی که در جدول ۱۰-۱ اشاره شده است، این مشاهدات به ترتیب برابر با ۴، ۱، ۱ و ۴ هستند.

جدول ۱۰-۲- فراوانی‌های ژنوتیپی توأم نشانگر-QTL-نشانگر در جمعیت تلاقی برگشتی.

فاصله نشانگر		ژنوتیپ QTL	
ژنوتیپ	فراوانی	1	0
11	$\frac{1}{2}(1-r)$	$\frac{1}{2}(1-r_1)(1-r_2)$	$\frac{1}{2} r_1 r_2$
10	$\frac{1}{2} r$	$\frac{1}{2}(1-r_1) r_2$	$\frac{1}{2} r_1 (1-r_2)$
01	$\frac{1}{2} r$	$\frac{1}{2} r_1 (1-r_2)$	$\frac{1}{2}(1-r_1) r_2$
00	$\frac{1}{2}(1-r)$	$\frac{1}{2} r_1 r_2$	$\frac{1}{2}(1-r_1)(1-r_2)$

جدول ۱۰-۳- احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ فاصله نشانگری در جمعیت تلاقی برگشتی.

ژنوتیپ فاصله نشانگری	ژنوتیپ QTL	
	1	0
11	$\frac{(1-r_1)(1-r_2)}{1-r}$	$\frac{r_1 r_2}{1-r}$
10	$\frac{(1-r_1)r_2}{r}$	$\frac{r_1(1-r_2)}{r}$
01	$\frac{r_1(1-r_2)}{r}$	$\frac{(1-r_1)r_2}{r}$
00	$\frac{r_1 r_2}{1-r}$	$\frac{(1-r_1)(1-r_2)}{1-r}$

جدول ۱۰-۴- احتمالات شرطی تقریبی ژنوتیپ‌های QTL مشروط به ژنوتیپ فاصله نشانگری در جمعیت تلاقی برگشتی، با فرض عدم وجود نوترکیبی دوگانه.

ژنوتیپ فاصله نشانگری	ژنوتیپ QTL	
	1	0
11	۱	۰
10	$1-\theta$	θ
01	θ	$1-\theta$
00	۰	۱

توجه: نسبت $\theta = \frac{r_1}{r}$ و $1-\theta = \frac{r_2}{r}$ است.

۱۰-۳-۳- برآورد و آزمون

اگر r_1 یا r_2 شناخته شده باشند، امکان جایگزینی این مقادیر در معادله (۱۰-۳) وجود داشته و از این‌رو آن را به‌عنوان یک تابعیت خطی ساده با μ به عنوان عرض از مبدأ y و a به‌عنوان شیب خط حل می‌کنیم. در حقیقت، حتی اگر r_1 یا r_2 ناشناخته باشند، می‌توانیم ماتریس طرح $\mathbf{X}$ را با فرض نمودن موقعیت یک QTL در چندین موقعیت (برای مثال، هر ۱ یا دو سانتی‌مورگان) بین دو نشانگر محاسبه کنیم.

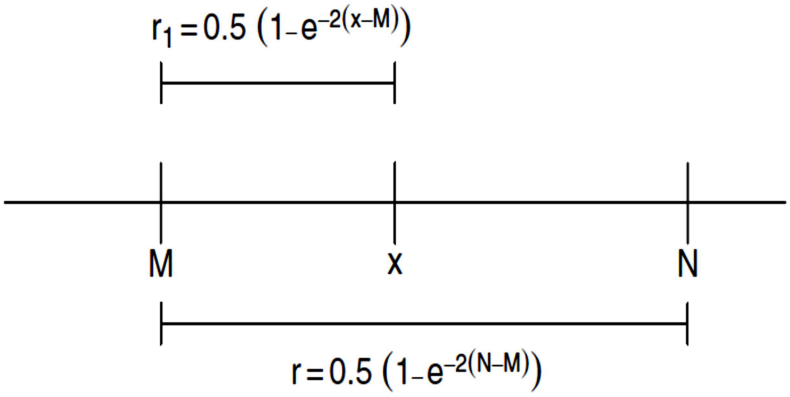
توجه کنید یک QTL در هر ۱ یا ۲ سانتی‌مورگان از نشانگر M تا N جستجو شود، نیاز داریم تا از یک تابع نقشه برای تبدیل فاصله نقشه به نرخ نوترکیبی استفاده کنیم. با فرض نقطه x داریم:

$$r_1(x)=\frac{1}{2}\left(1-e^{-2d(x)}\right),\quad r=\frac{1}{2}\left(1-e^{-2d}\right)$$

و:

$$\theta=\frac{r_1(x)}{r}$$

که تابع نقشه هالدین در نظر گرفته شده و $d(x)$ و d به ترتیب فواصل نقشه‌ای (بر حسب مورگان) بین QTL و نشانگر سمت چپ یعنی M و بین دو نشانگر هستند (شکل ۱۰-۱ را مشاهده کنید).



شکل ۱۰-۱- نمایش مکان‌یابی فاصله‌ای QTL بر اساس دو نشانگر مجاور M و N.

از این‌رو، از تابعیت خطی برای برازش μ و a برای هر موقعیت فرضی QTL استفاده می‌شود. این حالت برآوردهای حداقل مربعات بردار **b** و واریانس باقیمانده را به‌صورت زیر فراهم می‌آورد:

$$\hat{b}=\left(X'X\right)^{-1}\left(X'y\right) \tag{۱۰-۵}$$

$$\hat{\sigma}^2=\frac{1}{n}\left(y-Xb\right)^T\left(y-Xb\right) \tag{۱۰-۶}$$

همچنین، مجموع مربعات و میانگین مربعات تابعیت (رگرسیون) و باقیمانده و نیز آماره F برای واریانس تابعیت محاسبه شده و بنابراین آزمون a (شیب خط) انجام می‌شود. موقعیتی که بهترین مدل برازش یافته را ارایه می‌دهد (یعنی، کوچک‌ترین میانگین مربعات خطا را دارد)، محتمل‌ترین موقعیت QTL و بهترین برآورد اثر آن را ارائه می‌دهد.

به‌طور اختصاصی‌تر، آزمون فرض برای وجود یک QTL در یک موقعیت معین در درون فاصله نشانگری (یعنی، معنی‌داری $\hat{a}$) می‌تواند بدین صورت فرموله شود:

$H_0 : a = 0$ در مقابل $H_1 : a \neq 0$

مدل تحت فرض H_1 یک مدل کامل با QTL است که به صورت زیر بیان می‌شود:

$$\hat{y}_i = \hat{\mu} + x_{j|i}\hat{a}$$

حال آن که مدل تحت فرض H_0 یک مدل کاهش یافته بدون QTL است و بدین صورت بیان می‌شود:

$$\tilde{y}_i = \tilde{\mu}$$

مجموع مربعات کل (SST) مجموع $\left(y_i - \tilde{\mu}\right)^2$ است، در حالی که مجموع مربعات خطا (SSE) مجموع $\left(y_i - \hat{\mu} - x_{j|i}\hat{a}\right)^2$ است. برای آزمون H_0، می‌توانیم آماره آزمون F را به صورت زیر محاسبه کنیم:

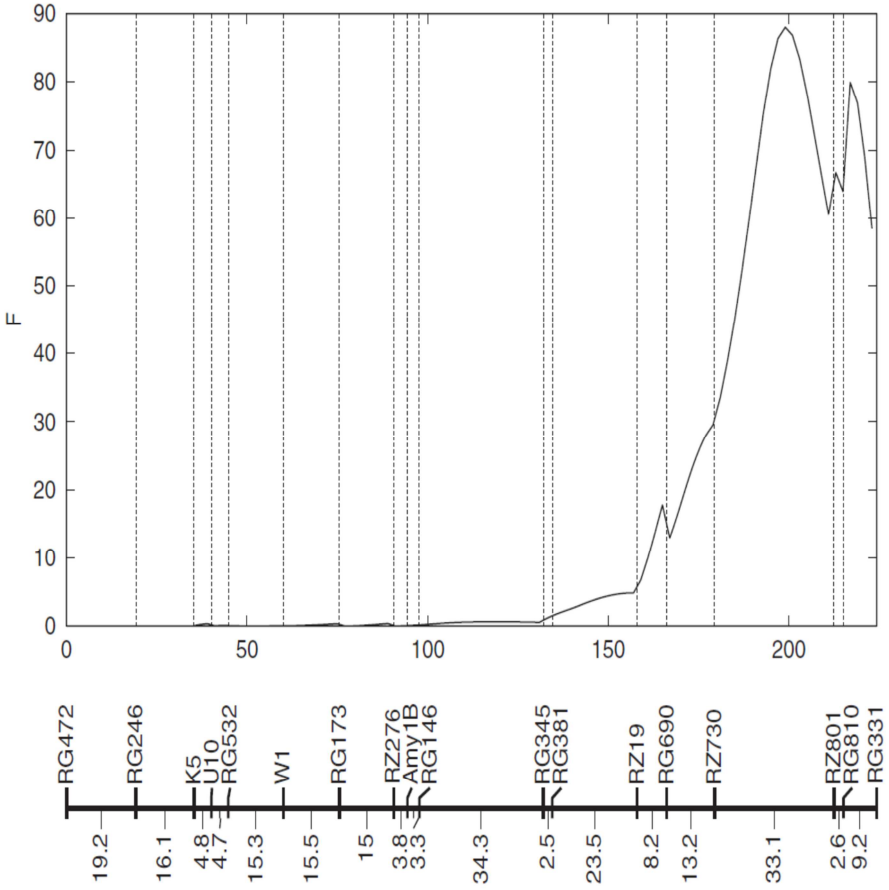
$$F = \frac{\left(SST - SSE\right)/\left(2-1\right)}{SSE/\left(n-2\right)} \qquad (۷-۱۰)$$

با مقایسه مقدار محاسبه شده F با مقدار F جدول یعنی $F_{0.05,(1,n-2)}$، وجود QTL می‌تواند آزمون شود.

مثال ۱۰-۱. مثال ۳-۱ را مجدداً مشاهده کنید. یک جمعیت DH که معادل با جمعیت تلاقی برگشتی است، با استفاده از دو لاین اینبرد نیمه پاکوتاه IR64 و پابلند Azucena ایجاد شد (هوانگ و همکاران، ۱۹۹۷). این جمعیت DH شامل ۱۲۳ لاین بود که هر یک برای ۱۳۵ نشانگر RFLP و ۴۰ نشانگر آیزوزایم و RAPD تعیین ژنوتیپ شده و برای صفات فنوتیپی مختلف تعیین فنوتیپ شدند. در این مثال، فنوتیپ انتخاب شده برای مکان‌یابی QTL ارتفاع بوته است که در ۱۰ هفته پس از انتقال نشاء برنج به مزرعه اندازه‌گیری شد. با مرتب کردن داده‌های نشانگر و داده‌های فنوتیپی به شکلی که در جدول ۸-۱ نشان داده شده است، می‌خواهیم از روش‌های تابعیت (رگرسیون) مبتنی بر حداقل مربعات (LS) برای مکان‌یابی یک QTL برای ارتفاع بوته با استفاده از نقشه پیوستگی ژنتیکی تهیه شده از این نشانگرهای تعیین ژنوتیپ شده استفاده کنیم (شکل ۳-۳).

شکل ۱۰-۲ پروفیل مقادیر F که با معادله (۱۰-۷) محاسبه شده‌اند را در موقعیت‌های مختلف در طول گروه پیوستگی کروموزوم ۱ که با ۱۸ نشانگر ایجاد شده است، نشان می‌دهد. یک نقطه اوج مشخص در ۱۹۹ سانتی‌مورگانی فاصله نشانگری [RZ730-RZ801] وجود دارد که مقدار F آن ۸۸/۰۲ بوده و خیلی بیشتر از مقدار F جدول یعنی $F_{0.05,(1,85)}$ = ۳/۹۵ است (۸۷ بوته برنج به دلیل داده‌های گمشده برای این دو نشانگر کنار گذاشته شدند). این حالت حضور یک QTL معنی‌دار را در اطراف موقعیت فوق پیشنهاد می‌کند.

برآوردهای LS پارامترهای مدل یعنی μ، a و σ^2 در نقطه اوج پروفیل مقدار F به‌دست می‌آیند. این مقادیر برابر با $\hat{\mu} = 86.9$، $\hat{a} = 42.9$ و $\hat{\sigma}^2 = 286.6$ هستند. واریانس ژنتیکی ارتفاع بوته به‌واسطه این QTL شناسایی شده با استفاده از معادله (۸-۴) برابر با $\sigma_g^2 = 271.8$ محاسبه می‌شود. این حالت اجازه محاسبه سهم واریانس فنوتیپی توجیه شده به وسیله این QTL ($R^2 = 0.50$) را با استفاده از معادله (۸-۵) می‌دهد.



شکل ۱۰-۲- پروفیل مقدار F در طول کروموزوم ۱ برای آزمون QTL که ارتفاع بوته را در سن ۱۰ هفتگی یک جمعیت DH برنج کنترل می‌کند (هوانگ و همکاران، ۱۹۹۷). اسامی نشانگرها و فواصل آن‌ها در زیر پروفیل ارایه شده‌اند.

۱۰-۴- مکان‌یابی فاصله‌ای QTL در یک جمعیت F_2

از استراتژی مشابهی می‌توان برای مکان‌یابی یک QTL در حال تفرق در یک جمعیت F_2 نیز بر اساس معادله (۱۰-۲) استفاده نمود. همانگونه که برای جمعیت تلاقی برگشتی نشان داده شد، به احتمالات شرطی ژنوتیپ‌های QTL مشروط به فواصل نشانگری نیاز داریم. جدول ۱۰-۲ ماتریس فراوانی ژنوتیپ‌های سه نقطه‌ای (نشانگر-QTL-نشانگر) را برای تلاقی برگشتی ارایه می‌دهد که معادل ماتریس ژنوتیپ‌های گامتی سه نقطه‌ای تولید شده به وسیله هتروزیگوت مضاعف F_1 است. از آنجایی که F_2 از ترکیب گامت‌های F_1 به‌دست می‌آید، ماتریس فراوانی برای ژنوتیپ‌های سه

نقطه‌ای در F_2 به‌صورت حاصل‌ضرب کرونکر دو ماتریس معادله (۱۰-۲) است که هر یک مربوط به یک والد مختلف با ژنوتیپ‌های مشابه است که از هم جدا شده و به‌صورت یک ماتریس (9×3) تبدیل شده است. ژنوتیپ‌های مشابه ممکن است به وسیله ترکیبات گامتی مختلف تولید شوند. برای مثال، ژنوتیپ نشانگر-QTL-نشانگر $12/12/12$ در F_2 می‌تواند با چهار ترکیب مختلف ایجاد شود:

$$12/12/12=\begin{cases}|111|\times|222|\\|121|\times|212|\\|112|\times|221|\\|211|\times|122|\end{cases}$$

بنابراین، احتمال توأم این ژنوتیپ هتروزیگوت سه گانه برابر با مجموع احتمالات این چهار ترکیب است. ماتریس تجزیه شده (9×3) در جدول ۱۰-۵ ارایه شده است.

جدول ۱۰-۵- فراوانی‌های توأم ژنوتیپی نشانگر-QTL-نشانگر در جمعیت F_2.

فاصله نشانگری		ژنوتیپ QTL		
ژنوتیپ	فراوانی	2	1	0
22	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}(1-r_1)^2(1-r_2)^2$	$\frac{1}{2}r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{4}r_1^2r_2^2$
21	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r_2(1-r_1)^2(1-r_2)$	$\frac{1}{2}r_1(1-r_1)(1-2r_2+2r_2^2)$	$\frac{1}{4}r_1^2r_2(1-r_2)$
20	$\frac{1}{4}r^2$	$\frac{1}{4}(1-r_1)^2r_2^2$	$\frac{1}{2}r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{4}r_1^2(1-r_2)^2$
12	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r_1(1-r_1)(1-r_2)^2$	$\frac{1}{2}r_2(1-r_2)(1-2r_1+2r_1^2)$	$\frac{1}{4}r_1(1-r_1)r_2^2$
11	$\frac{1-2r+2r^2}{2}$	$\frac{1}{4}r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{2}(1-2r_1+2r_1^2)(1-2r_2+2r_2^2)$	$\frac{1}{4}r_1r_2(1-r_1)(1-r_2)$
10	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r_1(1-r_1)r_2^2$	$\frac{1}{2}r_2(1-r_2)(1-2r_1+2r_1^2)$	$\frac{1}{4}r_1(1-r_1)(1-r_2)^2$
02	$\frac{1}{4}r^2$	$\frac{1}{4}r_1^2(1-r_2)^2$	$\frac{1}{2}r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{4}(1-r_1)^2r_2^2$
01	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r_1^2r_2(1-r_2)$	$\frac{1}{2}r_1(1-r_1)(1-2r_2+2r_2^2)$	$\frac{1}{4}r_2(1-r_1)^2(1-r_2)$
00	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}r_1^2r_2^2$	$\frac{1}{2}r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{4}(1-r_1)^2(1-r_2)^2$

از روی جدول ۱۰-۵، ضرایب مرتبط با اثر افزایشی (a) و اثر غالبیت (d) مربوط به QTL در F_2 (جدول ۱۰-۶) که متغیرهای توصیفی هستند، می‌توانند بر حسب موقعیت ژنومی QTL توصیف شده به وسیله نرخ نوترکیبی بین QTL و نشانگرها به‌دست آیند. چون موقعیت QTL ناشناخته است، استراتژی جستجوی شبکه‌ای که در آن، یک QTL در هر موقعیتی در بین دو نشانگر در نظر گرفته می‌شود، برای تابعیت (رگرسیون) ارزش صفت بر متغیر توصیفی استفاده می‌شود تا μ، a و d و آماره F برای آزمون معنی‌داری تابعیت محاسبه شود و بیشترین مقدار F برای تابعیت به‌عنوان محتمل‌ترین موقعیت QTL شناسایی شود.

مثال ۱۰-۲. مثال ۳-۲ را مجدداً مشاهده کنید. چورود و همکاران (۱۹۹۶) یک نقشه پیوستگی را با استفاده از ۷۵ نشانگر ریزماهواره در یک جمعیت ۵۳۵ فردی F_2 حاصل از دو سویه بزرگ (LG/J) و کوچک (SM/J) تهیه کردند. نتاج F_2 از نظر توده بدن در فواصل ۱۰ هفته‌ای که از سن ۷ روزگی شروع شده بود، اندازه‌گیری شدند. داده‌های خام وزن بدن برای آثار هر متغیر همبسته (کوواریت) به واسطهٔ مادر، تعداد نوزاد در هنگام تولد و شکم زایش تصحیح شدند، ولی برای اثر جنسیت تصحیح نشدند. همچنین، از کروموزوم ۱ متشکل از ۹ نشانگر استفاده شد تا QTL مؤثر بر وزن بدن در سن ۱۰ هفتگی با روش‌های تابعیت مبتنی بر LS مکان‌یابی شود.

شکل ۱۰-۳ پروفیل مقادیر F که با معادله (۱۰-۷) محاسبه شده‌اند را در موقعیت‌های مختلف گروه پیوستگی کروموزوم ۱ که با ۱۸ نشانگر تهیه شده است نشان می‌دهد. نقطه اوج مشخصی در ۴۶ سانتی‌مورگانی فاصله نشانگری [D2MIT389-D2MIT17] وجود دارد که مقدار F آن ۱۶/۴۳۴ است که بسیار بیشتر از مقدار بحرانی ۳/۸۶۲ = $F_{0.05,(1,451)}$ است. این حالت وجود یک QTL معنی‌دار را در اطراف موقعیت فوق پیشنهاد می‌دهد.

برآوردهای LS پارامترهای مدل یعنی μ، a، d و σ^2 در نقطه اوج پروفیل مقدار F به‌دست می‌آیند. این برآوردها عبارت از $\hat{\mu} = 23.0$، $\hat{a} = 1.6$، $\hat{d} = 1.4$ و $\hat{\sigma}^2 = 21.7$ هستند. واریانس ژنتیکی ارتفاع بوته به‌واسطه این QTL شناسایی شده با استفاده از معادله (۸-۴) برابر با $\sigma_g^2 = 1.5$ محاسبه می‌شود. این حالت اجازه محاسبه سهم واریانس فنوتیپی توجیه شده به وسیله این QTL ($R^2 = 0.06$) را با استفاده از معادله (۸-۵) می‌دهد.

در مقایسه با جمعیت تلاقی برگشتی، جمعیت F_2 دارای اطلاعات بیشتری است، چون اجازه انجام آزمون‌های معنی‌داری آثار افزایشی و غالبیت QTL شناسایی شده را می‌دهد. مدل کامل با اثر افزایشی بدین صورت نوشته می‌شود:

$$\hat{y}_i = \hat{\mu} + x_{1j|i}\hat{a} + x_{2j|i}\hat{d}$$

حال آن که مدل کاهش یافته بدون اثر افزایشی به‌صورت زیر نوشته می‌شود:

$$\tilde{y}_i = \tilde{\mu} + x_{2j|i}\tilde{d}$$

با استفاده از این دو مدل مقابل هم می‌توانیم مجموع مربعات کل (SST) و مجموع مربعات باقیمانده (SSE) و درنهایت آماره آزمون F را محاسبه کنیم (معادله (۱۰-۷) را مشاهده کنید). مدل کاهش یافته بدون اثر غالبیت نیز به‌صورت زیر فرموله می‌شود:

$$\tilde{y}_i = \tilde{\mu} + x_{1j|i}\tilde{a}$$

که از روی آن می‌توان آماره‌های آزمون را به صورت مشابه محاسبه نمود.

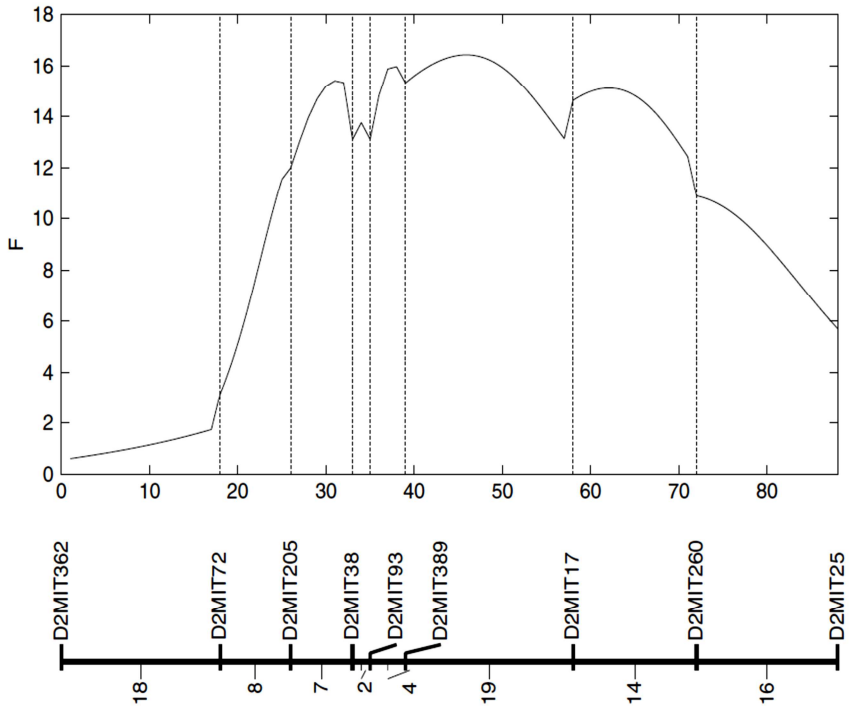
۱۰-۵- نکات قابل توجه

روش تابعیت ساده مکان‌یابی QTL در مقایسه با روش حداکثر درستنمایی مبتنی بر مدل مختلط مورد تحقیق و پژوهش قرار گرفته و در فصل بعد توضیح داده می‌شود. هیچ تفاوت معنی‌داری بین دو روش بر حسب خطاهای برآورد پارامترها و توان آماری تشخیص داده نشده است. اگرچه پروفیل‌های آماره آزمون، تفاوت اندکی را بین این دو روش نشان دادند، اما تفاوت فقط در سطوح جزیی تشخیص داده شد (هالی و نوت، ۱۹۹۲). مزیت اصلی روش تابعیت، سادگی بسیار زیاد محاسبات آن و سهولت اجرای آن با استفاده از بسیاری از برنامه‌های آماری رایانه‌ای در دسترس است.

ژو (۱۹۹۵) متوجه شد که برآورد واریانس باقیمانده ایجاد شده با روش تابعیت با بخشی از واریانس QTL تداخل می‌یابد. وی بعداً یک روش جایگزین که از آن به‌عنوان حداقل مربعات وزنی مکرر (IRLS[1]) یاد می‌شود را پیشنهاد داد تا نقص مربوط به اختلاط پارامترها در روش تابعیت تصحیح شود و کارآیی این روش افزایش یابد (ژو، ۱۹۹۸). روش IRLS خصوصیات سادگی و سرعت روش تابعیت معمولی را دارد. مشابه روش تابعیت موجود، این روش می‌تواند در مکان‌یابی QTL به همراه آزمون‌های تبدیل یا جایگشت و ایجاد فواصل اطمینان با روش خود راه‌اندازی[2] مفید باشد.

---

[1] - Iteratively reweighted least squares
[2] - Bootstrapping



شکل ۱۰-۳- پروفیل مقدار F در طول کروموزوم ۱ برای آزمون یک QTL که وزن بدن را در سن ۱۰ هفتگی در یک جمعیت F_2 موش کنترل می‌کند (چورود و همکاران، ۱۹۹۶). اسامی نشانگرها و فواصل آن‌ها در زیر پروفیل ارایه شده‌اند.

۱۰-۶- تمرین‌ها

۱۰-۱- پس از مطالعه فصل‌های ۸ و ۱۰، مزایای مکان‌یابی فاصله‌ای را نسبت به تجزیه تک نشانگری مقایسه کنید. ممکن است از یک مثال برای مکان‌یابی برنج یا موش، همان‌گونه که در این کتاب برای بیان این مزایا نشان داده شد، استفاده کنید.

۱۰-۲- از قضیه بیز برای به‌دست آوردن جدول ۱۰-۳ استفاده کنید.

۱۰-۳- داده‌های گمشده در مکان‌یابی QTL

مشکلات داده‌های گمشده در مکان‌یابی ژنتیکی معمول هستند. یک راه چاره برای داده‌های گمشده که به‌طور معمول استفاده می‌شود، آن است که افرادی که دارای فنوتیپ‌ها یا نشانگرهای گمشده هستند را به سادگی کنار بگذاریم. ولی این روش نباید مؤثرترین راهکار باشد. اگر داده‌ها به‌صورت کاملاً تصادفی گمشده باشند، می‌توان داده‌های گمشده را با مقادیر مورد انتظار آن‌ها مشروط به داده‌های موجود نشانگرها یا افراد دیگر با استفاده از الگوریتم لاندر و گرین (۱۹۸۷) جایگزین نمود. منابع ژنتیکی و آماری مرتبط برای رفع مشکلات مربوط به داده‌های گمشده و مشخص کردن این که چگونه این روش‌ها می‌توانند برای مکان‌یابی QTL استفاده شوند را مطالعه کنید.

۱۰-۴- مکان‌یابی فاصله‌ای با استفاده از لاین‌های اینبرد نوترکیب (RILs). RIL ها می‌توانند از راه خودگشنی مکرر یا آمیزش برادر-خواهری نتاج F_1 حاصل از تلاقی بین دو لاین اینبرد به‌دست آیند. RIL ها برای تمامی ژن‌ها تثبیت شده‌اند، ژنوتیپ‌های هموزیگوت ۲ و ۰ هستند و می‌توانند به‌عنوان یک جمعیت دائمی مکان‌یابی برای اهداف متعدد مورد استفاده قرار گیرند.

دو نشانگر مجاور M و N را در نظر بگیرید که یک QTL فرضی را در یک جمعیت RIL احاطه کرده‌اند. فرض کنید R_1 و R_2 به ترتیب نسبت زیگوت‌های نوترکیب بین نشانگر M و QTL و بین QTL و نشانگر N باشند. مارتین و هاسپیتال (۲۰۰۶) احتمالات شرطی ژنوتیپ‌های QTL را مشروط به ژنوتیپ‌های نشانگر بر حسب R_1 و R_2 در جمعیت خودگشنی RIL به‌دست آوردند که در جدول ۱۰-۷ ارائه شده‌اند.

فرض کنید r_1 و r_2 به ترتیب نرخ‌های نوترکیبی بین نشانگر M و QTL و بین QTL و نشانگر N باشند. رابطه بین R و r برای دو جایگاه ژنی در جمعیت خودگشنی RIL به وسیله هالدین و وادینگتون (۱۹۳۱) به‌صورت زیر بیان شده است:

$$R_1 = \frac{2r_1}{1+2r_1}, \quad r_1 = \frac{R_1}{2\left(1-R_1\right)}$$
$$R_2 = \frac{2r_2}{1+2r_2}, \quad r_2 = \frac{R_2}{2\left(1-R_2\right)} \qquad (۸-۱۰)$$

الف. با استفاده از f_2 و f_0 که به ترتیب نشان‌دهنده چگالی‌های دو ژنوتیپ هموزیگوت QTL یعنی ۲ و ۰ هستند، مدل مختلط مبتنی بر درستنمایی را بنویسید.

ب. احتمالات پسین ژنوتیپ‌های QTL را بر اساس احتمالات شرطی پیشین در جدول ۱۰-۷ تعریف کنید.

ج. بر اساس روش‌های تشریح شده برای جمعیت‌های تلاقی برگشتی و F_2، الگوریتم محاسباتی EM را برای برآورد ارزش‌های ژنوتیپی (μ_2 و μ_0) و واریانس باقیمانده (σ^2) نشان دهید.

د. به‌دست آوردن شکل دقیقی برای برآورد نسبت زیگوت‌های نوترکیب (R_1 و R_2) مشکل است. از راهکار شبکه‌ای برای برآورد این سهم‌ها استفاده کنید. توجه کنید که $0 \le R_1$ و $R_2 \le 0.5$.

ه. پس از برآورد R_1 و R_2، از معادله (۱۰-۸) برای برآورد MLE های نرخ‌های نوترکیبی و بنابراین موقعیت QTL فرضی به وسیله تابع نقشه استفاده کنید.

و. آزمون کنید که آیا QTL با رابطه فرض صفر زیر معنی‌دار است یا خیر:

$$H_0 : \mu_2 = \mu_0$$

آستانه بحرانی برای اعلام حضور یک QTL می‌تواند از روی آزمون‌های تبدیل یا جایگشت تعیین شود.

جدول ۱۰-۷- احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ‌های نشانگر در جمعیت خودگشنی RIL.

فاصله نشانگری		ژنوتیپ QTL	
ژنوتیپ	اندازه	2	0
22	n_{22}	$1-\frac{R_1R_2(3-2R_1-2R_2)}{2(1-R_1)(1-R_2)}$	$\frac{R_1R_2(3-2R_1-2R_2)}{2(1-R_1)(1-R_2)}$
20	n_{20}	$-\frac{2R_1-R_1R_2(3+2R_1-2R_2)}{2R_2+R_1(2-6R_2)}$	$\frac{2R_1-R_1R_2(3+2R_1-2R_2)}{2R_2+R_1(2-6R_2)}$
02	n_{02}	$\frac{2R_1-R_1R_2(3+2R_1-2R_2)}{2R_2+R_1(2-6R_2)}$	$-\frac{2R_1-R_1R_2(3+2R_1-2R_2)}{2R_2+R_1(2-6R_2)}$
00	n_{00}	$\frac{R_1R_2(3-2R_1-2R_2)}{2(1-R_1)(1-R_2)}$	$1-\frac{R_1R_2(3-2R_1-2R_2)}{2(1-R_1)(1-R_2)}$

فصل یازدهم

مکان‌یابی فاصله‌ای با استفاده از روش حداکثر درستنمایی

۱۱-۱- مقدمه

همان‌گونه که در فصل ۱۰ بحث شد، مکان‌یابی فاصله‌ای برای جداسازی آثار QTL و پیوستگی QTL-نشانگر توانمند است. مکان‌یابی فاصله‌ای بر اساس روش تابعیت از نظر محاسباتی سریع‌تر از روش حداکثر درستنمایی (ML) است و نتایج این دو روش در برخی موارد خاص قابل مقایسه است. کائو (۲۰۰۰) شرایطی که تحت آن، روش‌های مبتنی بر تابعیت و ML نتایج متفاوتی ایجاد می‌کنند را از نظر تجزیه و تحلیلی از راه شبیه‌سازی مورد بحث قرار داد. مقایسات بیشتر به وسیله مایر (۲۰۰۵) انجام شد و این دو روش را از نظر توان آزمون، دقت برآورد اثر و جایگاه QTL و برآورد واریانس باقیمانده وقتی که چندین QTL پیوسته در کنترل ژنتیکی یک صفت کمّی دخیل هستند، مقایسه نمود.

در هر حال، مشخص شده است که روش ML دارای چند خصوصیت آماری قابل توجه نظیر سازگاری و کارآیی مجانبی است و بنابراین توانایی زیادی برای برآورد دقیق پارامترهای QTL دارد. علاوه بر آن، روش ML دارای قابلیت تفسیر بهتری نسبت به مدل تابعیت از نظر مدل ژنتیکی است که این امر قابلیت آن را برای مسائل مکان‌یابی ژنتیکی عملی نشان می‌دهد.

در مواردی که چندین QTL پیوسته در یک منطقه ژنومی مشابه قرار دارند، مکان‌یابی‌فاصله‌ای مبتنی بر ML عموماً از مکان‌یابی فاصله‌ای مبتنی بر تابعیت از نظر توان تشخیص QTL و دقت برآورد پارامترها توانایی بیشتری دارد (مایر، ۲۰۰۵). این برتری با بیشتر شدن فواصل بین نشانگرها و بزرگ‌تر شدن اندازه جمعیت افزایش می‌یابد. همچنین، اگر QTL های پیوسته در وضعیت ناجفت (ترانس) باشند، تفاوت‌های بین دو روش زیاد خواهد بود. همان‌طورکه به‌وسیله کائو و همکاران (۱۹۹۹) پیشنهاد شد، روش ML به‌عنوان یک روش توانمند برای مکان‌یابی هم‌زمان چند QTL در نظر گرفته می‌شود.

در این فصل، اصول پایه‌ای روش مکان‌یابی فاصله‌ای مبتنی بر ML، استنباط‌های آماری آن در خصوص برآورد پارامترها و آزمون فرض‌ها و روش کاربرد عملی آن توصیف خواهد شد. مثال‌هایی نیز برای مکان‌یابی فاصله‌ای مبتنی بر ML ارائه می‌شود. تمامی منابع موجود در خصوص مکان‌یابی فاصله‌ای فرض می‌کنند که هیچ نوترکیبی مضاعف یا تداخل میوزی در فواصل مختلف بین نشانگرها و QTL ها وجود ندارد. مکان‌یابی فاصله‌ای QTL که این دو فرض را در طرح تلاقی برگشتی و F_2 در نظر نمی‌گیرد، توصیف خواهند شد.

۱۱-۲- مکان‌یابی فاصله‌ای QTL در تلاقی برگشتی

مدل مختلط موضوع اصلی مکان‌یابی فاصله‌ای مبتنی بر حداکثر درستنمایی است که به وسیله لندر و بوتستین (۱۹۸۹) پیشنهاد شد. در فصل ۹، چارچوب کلی ساختار یک مدل مختلط مکان‌یابی را از نقطه نظر جمعیتی و ژنتیک کمّی شرح دادیم و سپس الگوریتم‌های آماری مورد استفاده در برآورد پارامترهای QTL به‌ویژه در مدل مختلط را مرور کردیم. تحت این چارچوب، مکان‌یابی فاصله‌ای QTL ها در تلاقی یک لاین اینبرد می‌تواند به‌عنوان یک حالت خاص در نظر گرفته شود که در آن، مسائل آماری مربوط به آزمون‌های معنی‌داری و برآورد پارامترها به خوبی مورد بررسی قرار گرفته‌اند. بحث خود را روی دو طرح ژنتیکی ساده یعنی تلاقی برگشتی و F_2 یا لاین‌های اینبرد نوترکیب (RILs) حاصل از دو لاین اینبرد متضاد متمرکز خواهیم کرد.

یک QTL ممکن به همراه نشانگر مجاور سمت چپ M و نشانگر مجاور سمت راست N، با نرخ‌های نوترکیبی r_1 بین M و QTL، r_2 بین QTL و N و r بین M و N را در نظر بگیرید. فرض می‌کنیم که با یک جمعیت تلاقی برگشتی کار می‌کنیم و دو آلل ممکنه هر نشانگر را با ۱ و ۰ نشان می‌دهیم. جدول ۱۰-۲ ژنوتیپ‌های ممکنه نشانگر و احتمالات آن‌ها را ارائه نموده است. مجدداً یادآوری می‌کنیم که ما، ژنوتیپ QTL را مشاهده نمی‌کنیم، بلکه با استفاده از اطلاعات نشانگر استنباط شده است.

۱۱-۲-۱- درستنمایی

برای ایجاد تابع درستنمایی، فرض می‌کنیم که f_1 و f_0 مطابق با دو ژنوتیپ QTL یعنی یک و صفر، توابع چگالی احتمال مربوطه باشند. ازاین‌رو چگالی‌های مشاهدات در کلاس‌های نشانگر، مخلوطی از دو ژنوتیپ ممکنه QTL هستند که به‌صورت زیر ارائه می‌شوند:

	کلاس نشانگر		چگالی		
	ژنوتیپ	مشاهده	۱		۰
۱	۱۱	n_1	$1 f_1(y)$	+	$0 f_0(y)$
۲	۱۰	n_2	$(1-\theta) f_1(y)$	+	$\theta f_0(y)$
۳	۰۱	n_3	$\theta f_1(y)$	+	$(1-\theta) f_0(y)$
۴	۰۰	n_4	$0 f_1(y)$	+	$1 f_0(y)$
کل		n	$\frac{1}{2} f_1(y)$	+	$\frac{1}{2} f_0(y)$

که در آن‌ها هیچ نوترکیبی مضاعفی بین دو فاصله نشانگر-QTL در نظر گرفته نشده است. چون داده‌های فنوتیپی برای هر کلاس نشانگر مشاهده می‌شوند، از این‌رو یک تابع درستنمایی داریم که به چهار کلاس تفکیک می‌شود:

$$\begin{aligned} L &= \prod_{i=1}^{n}\left[\frac{1}{2} f_1(y_i)+\frac{1}{2} f_0(y_i)\right] \\ &= \prod_{i=1}^{n_1} f_1(y_i) \\ &\times\prod_{i=1}^{n_2}\left[(1-\theta) f_1(y_i)+\theta f_0(y_i)\right] \\ &\times\prod_{i=1}^{n_3}\left[\theta f_1(y_i)+(1-\theta) f_0(y_i)\right] \\ &\times\prod_{i=1}^{n_4} f_0(y_i) \end{aligned} \qquad (۱-۱۱)$$

فرض صفر برای عدم وجود QTL عبارت است از:

$$H_0 : f_1(y_i) = f_0(y_i) = f(y_i) \quad \text{یعنی:} \quad \mu_1 = \mu_0 = \mu$$

و برای آزمون این فرض در موقعیت x که در آن $\theta = \frac{r_1(x)}{r}$ است، آماره نسبت درستنمایی را مطابق با شکل ۱۰-۱ رسم می‌کنیم:

$$\lambda(x) = \frac{\max_{H_0} L\left(f(y_i)\right)}{\max_{H_1} L\left(f_1(y_i), f_0(y_i), \theta\right)} \qquad (۲-۱۱)$$

برای مقادیر x بین نشانگرهای M و N، معمولاً در فواصل ۲ سانتی‌مورگان از هم مقدار معادله (۱۱-۲) یا $-2\log\lambda$ را محاسبه کرده و از این‌رو نمودار درستنمایی رسم می‌شود. سپس، حداکثر آماره آزمون را در طول فاصله M-N پیدا می‌کنیم.

محاسبه سطح معنی‌دار $\max_x \lambda(x)$ می‌تواند یک مشکل باشد. اغلب اگرچه آماره آزمون نسبت درستنمایی به‌صورت توزیع x^2 تقریب زده می‌شود، ولی حداکثر آماره آزمون نسبت درستنمایی چنین نیست. چندین راه برای ادامه کار وجود دارد:

الف. مجانب‌ها برای یک نقشه متراکم: لندر و بوتستین (۱۹۸۹) از یک استدلال احتمالی جالب برای نشان دادن این موضوع استفاده کردند که امتیاز LOD نامیده می‌شود و تبدیل λ به صورت زیر است:

$$LOD(x) = -2\log_{10}\lambda(x)$$

وقتی که نشانگرها متراکم باشند (یعنی فاصله بین نشانگرها به سمت صفر میل می‌کند)، مقدار LOD به سمت کای مربع با دو درجه آزادی (x_2^2) میل می‌کند. در حالت خاص داریم:

$$P\left(\max_x LOD(x) > T \mid H_0\right) \to (C + 2Gt)\, x_2^2(t)$$

که در آن، $t = 2\log_{10} T$، C تعداد کروموزوم‌ها و G طول نقشه ژنتیکی است.

ب. تقریب‌هایی نظیر تقریب‌های پیفو (۲۰۰۱) و دیویس (۱۹۸۷، ۱۹۷۷).

ج. آزمون‌های تبدیل (جایگشت): اگر منابع محاسباتی کافی وجود داشته باشد، آزمون‌های تبدیل یا جایگشت ارائه شده به وسیله چرچیل و دورج (۱۹۹۴) می‌توانند مورد استفاده قرار گیرند. در این حالت، N نمونه جایگشتی تولید می‌شود و برای هر یک از آن‌ها $\max_x \lambda(x)$ محاسبه می‌شود.

در این بخش، استفاده از آزمون تبدیل به منظور مقایسه آماره برای داده‌های مشاهده شده با این توزیع مرجع شرح داده می‌شود. تمامی این تقریب‌ها بعداً در فصل ۱۲ بحث می‌شوند.

۱۱-۲-۲- به حداکثر رساندن درستنمایی

اکنون فرض می‌کنیم که توزیع صفات فنوتیپی می‌تواند با استفاده از یک توزیع نرمال توصیف شود. به‌طور اختصاصی داریم:

$$f_0(y_i) = f\left(y_i \mid \mu_0, \sigma^2\right) = N\left(\mu_0, \sigma^2\right) \text{ و } f_1(y_i) = f\left(y_i \mid \mu_1, \sigma^2\right) = N\left(\mu_1, \sigma^2\right)$$

و درستنمایی به‌صورت زیر بیان می‌شود:

$$\begin{aligned} L\left(\mu_1, \mu_2, \sigma^2, \theta \mid y\right) &= \prod_{i=1}^{n_1} f\left(y_i \mid \mu_1, \sigma^2\right) \\ &\times \prod_{i=1}^{n_2}\left[(1-\theta)\, f\left(y_i \mid \mu_1, \sigma^2\right) + \theta f\left(y_i \mid \mu_0, \sigma^2\right)\right] \\ &\times \prod_{i=1}^{n_3}\left[\theta f\left(y_i \mid \mu_1, \sigma^2\right) + (1-\theta)\, f\left(y_i \mid \mu_0, \sigma^2\right)\right] \\ &\times \prod_{i=1}^{n_4} f\left(y_i \mid \mu_0, \sigma^2\right) \end{aligned} \qquad (۱۱-۳)$$

و از این‌رو داریم:

$$\lambda(\theta)=\frac{\max_{\mu_1=\mu_0=\mu,\sigma^2} L\left(\mu,\sigma^2,\theta\mid y\right)}{\max_{\mu_1,\mu_0,\sigma^2} L\left(\mu_1,\mu_0,\sigma^2,\theta\mid y\right)} \qquad (۴-۱۱)$$

برای به حداکثر رساندن صورت کسر، اگر $\mu_1=\mu_0=\mu$ باشد، از این‌رو درستنمایی برابر است با:

$$L\left(\mu,\sigma^2,\theta\mid y\right)=L\left(\mu,\sigma^2\mid y\right)=\prod_{i=1}^{n} f\left(y_i\mid\mu,\sigma^2\right) \qquad (۵-۱۱)$$

که به θ بستگی ندارد و MLE ها بدین صورت هستند: $\hat{\mu}=\overline{y}$ و $\hat{\sigma}^2=\left(\frac{1}{n}\right)\sum_i^n\left(y_i-\overline{y}\right)^2$. برای به حداکثر رساندن درستنمایی کامل (۱۱-۳)، همانند بخش ۲-۲ عمل کرده و تعریف می‌کنیم:

$$P_{1|i}=\frac{(1-\theta)\,f\left(y_i\mid\mu_1,\sigma^2\right)}{(1-\theta)\,f\left(y_i\mid\mu_1,\sigma^2\right)+\theta f\left(y_i\mid\mu_0,\sigma^2\right)}$$
$$P_{0|i}=\frac{\theta f\left(y_i\mid\mu_0,\sigma^2\right)}{(1-\theta)\,f\left(y_i\mid\mu_1,\sigma^2\right)+\theta f\left(y_i\mid\mu_0,\sigma^2\right)} \qquad (۶-۱۱)$$

که $P_{1|i}+P_{0|i}=1$ است.

با مشتق گرفتن از $\log L$ (نسبت به پارامترهای مجهول معادله درستنمایی یعنی μ_0، μ_1، σ^2 و θ) و برابر صفر قرار دادن مشتق‌ها و حل کردن معادلات لگاریتم درستنمایی داریم:

(۷-۱۱)

$$\frac{\partial}{\partial \mu_1}\log L = 0 \Rightarrow \hat{\mu}_1 = \frac{\sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_2+n_3} P_{1|i} y_i}{n_1 + \sum_{i=1}^{n_2+n_3} P_{1|i}}$$

$$\frac{\partial}{\partial \mu_0}\log L = 0 \Rightarrow \hat{\mu}_0 = \frac{\sum_{i=1}^{n_2+n_3} P_{0|i} y_i + \sum_{i=1}^{n_4} y_i}{\sum_{i=1}^{n_2+n_3} P_{0|i} + n_4}$$

$$\frac{\partial}{\partial \sigma^2}\log L = 0 \Rightarrow \hat{\sigma}^2 =$$

$$\frac{1}{n}\left(\sum_{i=1}^{n_1}\left(y_i - \hat{\mu}_1\right)^2 + \sum_{i=1}^{n_2+n_3}\left[P_{1|i}\left(y_i - \hat{\mu}_1\right)^2 + P_{0|i}\left(y_i - \hat{\mu}_0\right)^2\right] + \sum_{i=1}^{n_4}\left(y_i - \hat{\mu}_0\right)^2\right)$$

$$\frac{\partial}{\partial \theta}\log L = 0 \Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{n_2} P_{0|i} + \sum_{i=1}^{n_3} P_{1|i}}{n_2 + n_3}$$

معادله (۱۱-۷) سیستمی از معادلات درستنمایی است که در آن، جواب‌ها برای پارامترهای ناشناخته از راه تجزیه به‌دست نمی‌آید، چون هر برآوردی به برآورد پارامترهای دیگر بستگی دارد. غیر مشابه با مسائل عددی دیگر، هر برآورد نیز تابعی از احتمال پسین برای هر فرد به همراه ژنوتیپ مورد انتظار QTL است؛ بنابراین، این تکرار درحقیقت حالت خاصی از الگوریتم EM است که به وسیله دمپستر و همکاران (۱۹۷۷) و منگ و روبین (۱۹۹۳) پیشنهاد شده و همگرایی تکرارها را تضمین می‌کند (تمرین ۱۱-۴ را مشاهده کنید).

الگوریتم EM برای برآورد پارامترها مورد استفاده قرار می‌گیرد. در مرحله E، احتمالات پسین برای نتاج i ام تلاقی برگشتی که ژنوتیپ ۱ یا ۰ مربوط به QTL را حمل می‌کنند، با استفاده از معادله (۱۱-۶) محاسبه می‌شوند. در مرحله M، احتمالات پسین محاسبه شده برای برآورد پارامترها از راه معادله (۱۱-۷) استفاده می‌شوند. معادلات (۱۱-۶) و (۱۱-۷) تا جایی تکرار می‌شوند که برآوردها به مقدار ثابت و پایداری برسند. برآوردهای پایدار که درستنمایی را حداکثر خواهند کرد، به‌عنوان برآوردهای حداکثر درستنمایی (MLE) در نظر گرفته می شوند.

مثال ۱۱-۱. **(روش‌های مکان‌یابی فاصله‌ای).** جدول ۱۰-۱ را مجدداً مشاهده کنید که در آن، یک نمونه کوچک با ۱۰ موش در نظر گرفته شده که برای دو نشانگر M و N تعیین ژنوتیپ شده و برای وزن بدن تعیین فنوتیپ شده‌اند. داده‌های فنوتیپی و نشانگری مشاهده شده به همراه

احتمالات پسین ژنوتیپ QTL (۱ یا ۰) مشروط به ژنوتیپ‌های فاصله نشانگری بدین صورت بیان می‌شوند:

موش	فاصله M-N	وزن بدن (y_i)	ژنوتیپ QTL	
			1	0
1	11	30	1	0
2	11	32	1	0
3	11	28	1	0
4	11	29	1	0
5	10	29	$1-\theta$	θ
6	01	22	θ	$1-\theta$
7	00	20	0	1
8	00	21	0	1
9	00	20	0	1
10	00	21	0	1

برای آن که از الگوریتم EM برای برآورد موقعیت QTL (θ)، ارزش‌های ژنوتیپی QTL (μ_1 و μ_0) و واریانس باقیمانده (σ^2) در مکان‌یابی فاصله‌ای استفاده کنیم، نیاز داریم تا مقادیر اولیه این پارامترها را در اختیار داشته باشیم. معمولاً این کار می‌تواند از روی نمونه‌برداری برآورد شود که در این صورت برابر با $\mu = 25.2$ و $s^2 = 22.84$ هستند. چون μ_1 و μ_0 متفاوت هستند، از دو مقدار با تفاوت کم، به عنوان مقادیر ابتدایی آن‌ها استفاده می‌کنیم، یعنی: $\mu_1^{(0)} = 25.2$ و $\mu_0^{(0)} = 23.2$. ضمناً فرض می‌کنیم که $\sigma^{2(0)} = 22.84$ و $\theta^{(0)} = 0.5$ است. مراحل تکرار الگوریتم EM به‌صورت زیر انجام می‌شوند.

مرحله ۱- احتمالات پسین ($P_{1|i}$ و $P_{0|i}$) را برای این که موش منفرد (i) ژنوتیپ ۱ یا ۰ را حمل می‌کند، محاسبه کنید. در این مثال، این احتمالات فقط برای موش‌های ۵ و ۶ با استفاده از معادله (۱۱-۶) محاسبه می‌شوند یعنی:

$$P_{1|5}^{(1)} = \frac{\left(1-\theta^{(0)}\right)\exp\left[-\frac{\left(y_5-\mu_1^{(0)}\right)^2}{2\sigma^{2(0)}}\right]}{\left(1-\theta^{(0)}\right)\exp\left[-\frac{\left(y_5-\mu_1^{(0)}\right)^2}{2\sigma^{2(0)}}\right]+\theta^{(0)}\exp\left[-\frac{\left(y_5-\mu_0^{(0)}\right)^2}{2\sigma^{2(0)}}\right]}$$

$$= \frac{(1-0.5)\exp\left[-\frac{(29-25.2)^2}{2\times 22.84}\right]}{(1-0.5)\exp\left[-\frac{(29-25.2)^2}{2\times 22.84}\right]+0.5\exp\left[-\frac{(29-23.2)^2}{2\times 22.84}\right]} = 0.6035$$

$$P_{0|5}^{(1)} = \frac{\theta^{(0)}\exp\left[-\frac{\left(y_5-\mu_0^{(0)}\right)^2}{2\sigma^{2(0)}}\right]}{\left(1-\theta^{(0)}\right)\exp\left[-\frac{\left(y_5-\mu_1^{(0)}\right)^2}{2\sigma^{2(0)}}\right]+\theta^{(0)}\exp\left[-\frac{\left(y_5-\mu_0^{(0)}\right)^2}{2\sigma^{2(0)}}\right]}$$

$$= \frac{0.5\exp\left[-\frac{(29-23.2)^2}{2\times 22.84}\right]}{(1-0.5)\exp\left[-\frac{(29-25.2)^2}{2\times 22.84}\right]+0.5\exp\left[-\frac{(29-23.2)^2}{2\times 22.84}\right]} = 0.3965$$

و:

$$P_{1|6}^{(1)}=\frac{\left(1-\theta^{(0)}\right)\exp\left[-\frac{\left(y_6-\mu_1^{(0)}\right)^2}{2\sigma^{2(0)}}\right]}{\left(1-\theta^{(0)}\right)\exp\left[-\frac{\left(y_6-\mu_1^{(0)}\right)^2}{2\sigma^{2(0)}}\right]+\theta^{(0)}\exp\left[-\frac{\left(y_6-\mu_0^{(0)}\right)^2}{2\sigma^{2(0)}}\right]}$$

$$=\frac{(1-0.5)\exp\left[-\frac{(22-25.2)^2}{2\times 22.84}\right]}{(1-0.5)\exp\left[-\frac{(22-25.2)^2}{2\times 22.84}\right]+0.5\exp\left[-\frac{(22-23.2)^2}{2\times 22.84}\right]}=0.4520$$

$$P_{0|6}^{(1)}=\frac{\theta^{(0)}\exp\left[-\frac{\left(y_6-\mu_0^{(0)}\right)^2}{2\sigma^{2(0)}}\right]}{\left(1-\theta^{(0)}\right)\exp\left[-\frac{\left(y_6-\mu_1^{(0)}\right)^2}{2\sigma^{2(0)}}\right]+\theta^{(0)}\exp\left[-\frac{\left(y_6-\mu_0^{(0)}\right)^2}{2\sigma^{2(0)}}\right]}$$

$$=\frac{0.5\exp\left[-\frac{(22-23.2)^2}{2\times 22.84}\right]}{(1-0.5)\exp\left[-\frac{(22-25.2)^2}{2\times 22.84}\right]+0.5\exp\left[-\frac{(22-23.2)^2}{2\times 22.84}\right]}=0.5480$$

سپس، احتمالات پسین برای هر موش به صورت جدول زیر محاسبه می‌شوند:

موش	فاصله M-N	وزن بدن (y_i)	احتمال پسین	
			$P_{1\|i}^{(1)}$	$P_{0\|i}^{(1)}$
۱	۱۱	۳۰	۱	۰
۲	۱۱	۳۲	۱	۰
۳	۱۱	۲۸	۱	۰
۴	۱۱	۲۹	۱	۰
۵	۱۰	۲۹	۰/۶۰۳۵	۰/۳۹۶۵
۶	۰۱	۲۲	۰/۴۵۲۰	۰/۵۴۸۰
۷	۰۰	۲۰	۰	۱
۸	۰۰	۲۱	۰	۱
۹	۰۰	۲۰	۰	۱
۱۰	۰۰	۲۱	۰	۱

مرحله ۲- ارزش‌های ژنوتیپی QTL، واریانس باقیمانده و موقعیت QTL را با استفاده از معادله لگاریتم درستنمایی (۱۱-۷) برآورد کنید، یعنی:

$$\hat{\mu}_1^{(1)} = \frac{30+32+28+29+29\times 0.6986+22\times 0.4049}{1+1+1+1+0.6986+0.4049} = 28.97$$

$$\hat{\mu}_0^{(1)} = \frac{29\times 0.3014+22\times 0.5951+20+21+20+21}{0.3014+0.5951+1+1+1+1} = 21.35$$

$$\hat{\sigma}^{2(1)} = \frac{1}{10}\Big[(30-29.03)^2+(32-29.03)^2+(28-29.03)^2+(29-29.03)^2$$
$$+(29-29.03)^2\times 0.6986+(22-29.03)^2\times 0.4049$$
$$+(29-21.21)^2\times 0.3014+(22-21.21)^2\times 0.5951$$
$$+(20-21.21)^2+(21-21.21)^2+(20-21.21)^2+(21-21.21)^2\Big]$$
$$= 6.05$$

$$\hat{\theta}^{(1)} = \frac{0.3965+0.4520}{1+1} = 0.4242$$

بر اساس معادله (۱۱-۳)، از پارامترهای برآورد شده برای محاسبه مقدار لگاریتم درستنمایی به صورت زیر استفاده می‌کنیم:

$$\log L^{(1)} = \log\left(\frac{10}{\sqrt{2\pi}\times 2.46}\right)$$
$$+\left[-\frac{(30-28.97)^2}{2\times 6.05}\right]+\left[-\frac{(32-28.97)^2}{2\times 6.05}\right]+\left[-\frac{(28-28.97)^2}{2\times 6.05}\right]+\left[-\frac{(29-28.97)^2}{2\times 6.05}\right]$$
$$+\log\left\{(1-0.4242)\exp\left[-\frac{(29-28.97)^2}{2\times 6.05}\right]+0.4242\exp\left[-\frac{(29-28.97)^2}{2\times 6.05}\right]\right\}$$
$$+\log\left\{0.4242\exp\left[-\frac{(22-21.35)^2}{2\times 6.05}\right]+(1-0.4242)\exp\left[-\frac{(22-21.35)^2}{2\times 6.05}\right]\right\}$$
$$+\left[-\frac{(20-21.35)^2}{2\times 6.05}\right]+\left[-\frac{(21-21.35)^2}{2\times 6.05}\right]+\left[-\frac{(21-21.35)^2}{2\times 6.05}\right]+\left[-\frac{(21-21.35)^2}{2\times 6.05}\right]$$
$$=-47.74$$

مرحله ۳- مراحل ۱ و ۲ را تا جایی که برآوردهای تمامی پارامترها به مقادیر ثابت و پایداری برسند، تکرار کنید. در این مثال، متوجه شدیم که برآورد پارامترها در مرحله x همگرا می‌شوند. MLE های پارامترها به‌صورت $\hat{\mu}_1 = 29.6$، $\hat{\mu}_0 = 20.8$، $\hat{\sigma}^2 = 1.2$، $\hat{\theta} \approx 0$ و $\log L = -6.53$ به‌دست می‌آیند.

مثال ۱۱-۲. **(مکان‌یابی QTL در درختان تبریزی)**. برای تشریح مکان‌یابی فاصله‌ای QTL، داده‌های درخت تبریزی را در نظر می‌گیریم (یین و همکاران، ۲۰۰۲). جدول ۱۱-۱ بخشی از داده‌ها را نشان می‌دهد که اطلاعات نشانگر را برای چهار نشانگر CA/CGA-580RD، A7-690، AM11-1060 و AT2-850 ارایه داده است. فواصل نقشه‌ای تجمعی در نشانگرها به ترتیب ۱۵/۶، ۳۰/۷ و ۴۳/۹ سانتی‌مورگان هستند. صفتی که اندازه‌گیری شده است، ارتفاع (بر حسب متر) هر یک از ۵۷ درخت است.

با استفاده از تکرارهای بالا برای برآوردهای حداکثر درستنمایی پارامترها (MLE)، برنامه R را که در سایت اینترنتی این کتاب در دسترس است، به‌کار می‌گیریم. نتایج در جدول ۱۱-۲ ارائه شده است. برآورد نرخ نوترکیبی $\hat{\theta}$ موقعیت QTL ممکن را در فاصله مربوطه مشخص می‌کند.

برای تبدیل نرخ نوترکیبی به فاصله نقشه می‌توانیم از تابع نقشه هالدین (۱۰-۵) استفاده کنیم. بـه یاد آورید که $\theta = \frac{r_1(x)}{r}$ است و از این‌رو، برای فاصله M-N داریم:

$$\theta = \frac{r_1(x)}{r} = \frac{\frac{1}{2}\left(1-e^{-2d(x)}\right)}{\frac{1}{2}\left(1-e^{-2d}\right)}$$

جدول ۱۱-۱- بخشی از داده‌های نشانگر و ارتفاع درخت تبریزی.

مشاهدات	CA/CGA-580RD	A7-690	AM11-1060	AT2-850	ارتفاع (متر)
۱	۲	۲	۲	۲	۱۰/۸
۲	۲	۱	۱	۱	۱۱/۵
۳	۱	۱	۱	۱	۱۳/۸
۴	۱	۱	۲	۲	۱۳/۹
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
۵۴	۲	۲	۲	۲	۲۱/۰
۵۵	۲	۲	۱	۲	۲۲/۰
۵۶	۲	۱	۱	۱	۲۲/۱
۵۷	۲	۲	۲	۲	۲۶/۰

جدول ۱۱-۲- برآوردهای پارامترها و نسبت‌های درستنمایی برای داده‌های درخت تبریزی.

فاصله	$\hat{\mu}_1$	$\hat{\mu}_0$	$\hat{\sigma}$	$\hat{\theta}$	$-2\log\lambda$
۱	۱۸/۰۰۶	۱۶/۳۵۶	۲/۶۳۲	۰/۱۷۶	۳/۲۵۷
۲	۱۷/۸۲۶	۱۶/۸۹۱	۲/۷۰۸	۰/۱۰۱	۱/۴۷۶
۳	۱۸/۰۲۰	۱۶/۷۴۸	۲/۶۷۳	۱/۰۰۰	۳/۱۳۶

که نشان می‌دهد:

$$d(x) = -\frac{1}{2}\log(1-2r\theta)$$

$$r = \frac{1}{2}\left(1-e^{-2d}\right)$$

با استفاده از این روابط برای اولین فاصله، داریم:

$$r = \frac{1}{2}\left(1-e^{-2\times 0.156}\right) = 0.134$$

$$d_{xM} = -\frac{1}{2}\log(1-2\times 0.134\times 0.176) = 0.024$$

جدول ۱۱-۳ موقعیت QTL های برآورد شده را برای تمامی فواصل ارائه می‌دهد.

جدول ۱۱-۳- موقعیت‌های QTL برآورد شده برای داده‌های درخت تبریزی.

فاصله	$\hat{\theta}$	$d(x)$
$(0, 0.156)$	۰/۱۷۶	۲/۴ سانتی مورگان
$(0.156, 0.307)$	۰/۱۰۱	۱۶/۹ سانتی مورگان
$(0.307, 0.439)$	۱/۰۰۰	۴۳/۹ سانتی مورگان

۱۱-۳- آزمون فرض

پس از این که پارامترها برآورد شدند، باید معنی‌داری این برآوردها را ارزیابی کنیم. در آزمون فرض صفر:

H_0: هیچ QTL در تمامی فواصل نشانگری وجود ندارد.

وقتی می‌توانیم H_0 را رد کنیم که نسبت‌های درستنمایی معنی‌دار باشند. اگر LRT_i آماره نسبت درستنمایی در طول فاصله l باشد، سپس در صورتی H_0 رد می‌شود که:

$$\max_l LRT_t > t$$

که در آن، t یک نقطه برش تقریبی است. آماره آزمون نسبت لگاریتم درستنمایی (LR) تحت فرض H_0 ($\mu_1 = \mu_0 = \mu$) و H_1 ($\mu_1 \neq \mu_0$) با رابطه زیر محاسبه می‌شود:

$$LR = -2\ln\left[\frac{L\left(\tilde{\mu}, \tilde{\sigma}^2\right)}{L\left(\hat{\mu}_1, \hat{\mu}_0, \hat{\sigma}^2\right)}\right] \qquad (۸-۱۱)$$

که در آن، $\hat{\mu}_1$، $\hat{\mu}_0$ و $\hat{\sigma}^2$ همان MLE های پارامترهای مربوطه تحت فرض H_1 یعنی یک QTL وجود دارد (مدل کامل) و $\tilde{\mu}$ و $\tilde{\sigma}^2$ همان MLE پارامترهای مربوطه تحت فرض H_0 یعنی QTL وجود ندارد (مدل کاهش یافته) هستند.

لندر و بوتستین (۱۹۸۹) از امتیاز LOD به عنوان یک آماره آزمون به صورت زیر استفاده کردند:

$$LOD = -\log_{10}\left[\frac{L\left(\tilde{\mu},\tilde{\sigma}^2\right)}{L\left(\hat{\mu}_1,\hat{\mu}_0,\hat{\sigma}^2\right)}\right] \qquad (۹-۱۱)$$

که در حقیقت معادل LR است. ارتباط LOD و LR به صورت زیر است:

$$LOD = \frac{1}{2}\left(\log_{10} e\right) LR = 0.217 LR$$

تحت فرض H_0، امتیاز LOD یا آماره LR به‌صورت مجانبی دارای توزیعی همانند آماره مرکزی χ^2 است و درجات آزادی آن‌ها برابر با تعداد پارامترهای تثبیت شده در فرض صفر است. ولی این درجات در فرض مقابل متغیر هستند، چون نرخ نوترکیبی و اثر برآورد شده QTL همبسته هستند. برای مثال، اگر نرخ نوترکیبی بین یک نشانگر و QTL در مقدار ۰/۵ ثابت باشد، سپس مقدار اثر QTL فرضی برای تجزیه تک نشانگری، جزئی و بی‌اهمیت است؛ بنابراین، مشخص نیست که چند درجه آزادی باید برای چنین تجزیه نشانگر-QTL پیوسته به آن استفاده شود. اگرچه این مشکل تا حدودی به وسیله مکان‌یابی فاصله‌ای کاهش می‌یابد، ولی هنوز مشخص نیست که آیا هم موقعیت QTL و هم اثر آن در فرض صفر تثبیت شده‌اند یا خیر.

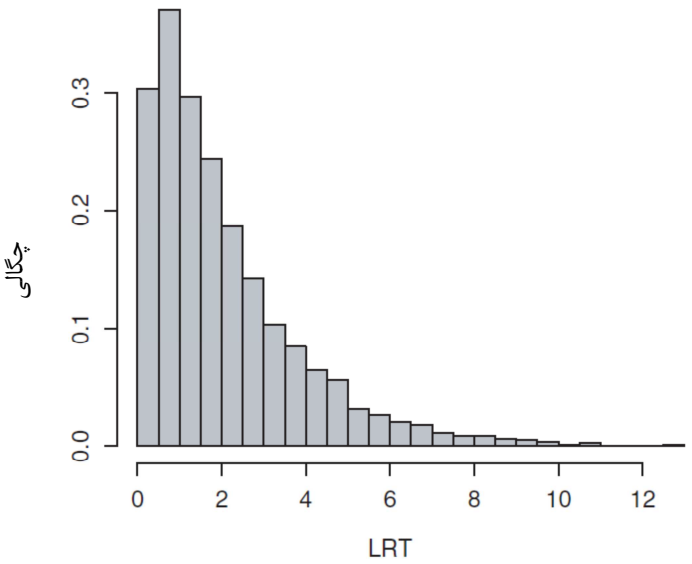
از آنجایی که موقعیت QTL (که با نسبت θ اندازه‌گیری می‌شود) فقط تحت فرض مقابل مشخص است، از این‌رو در آزمون معنی‌داری QTL به‌عنوان پارامتر مزاحم در نظر گرفته می‌شود. تحت فرض صفر، پارامتر مزاحم (موقعیت QTL) تعریف نشده است، به‌طوری که تکنیک‌های استاندارد برای به‌دست آوردن توزیع صفر آماره آزمون غیر مشروط به پارامتر مزاحم کاربردی ندارند. به این دلایل، یک پروفیل آماره آزمون در طول فاصله مجاز برای پارامتر مزاحم ایجاد می‌شود و حداکثر آن پروفیل برای انجام فقط یک آزمون انتخاب می‌شود. این حالت که جستجوی شبکه‌ای نامیده می‌شود، فرض می‌کند که QTL در هر موقعیتی با دو نشانگر احاطه می‌شود و هیچ نیازی به در نظر گرفتن معادله لگاریتم درستنمایی برای موقعیت QTL در تکرارها نیست، ولی در عوض به θ اجازه می‌دهد تا به‌صورت معینی از نظر رتبه تغییر کند. چون فرض می‌شود که QTL در جایی از فاصله ژنومی قرار دارد که با دو نشانگر مجاور احاطه شده است، از این‌رو θ مقداری بین ۰ تا ۱ را به خود اختصاص خواهد داد. در هر موقعیت فرضی، امتیاز LOD یا مقدار LR محاسبه می‌شود که اجازه جستجوی سیستماتیک یک QTL معنی‌دار را در سرتاسر یک گروه

پیوستگی معین یا حتی کل ژنوم می‌دهد. اگر در ابتدا QTL در نشانگر سمت راست در نظر گرفته شود و سپس در هر ۱ تا ۲ سانتی‌مورگان از این نشانگر حرکت نماییم، می‌توانیم امتیاز LOD یا مقدار LR را در برابر موقعیت نقشه‌ای QTL رسم نماییم. اگر آماره آزمون در یک ناحیه بیشتر از آستانه بحرانی از پیش تعیین شده باشد، QTL معنی‌دار در ناحیه حداکثری (اوج) پروفیل LOD یا LR مشخص می‌شود. مقدار θ مربوط به حداکثر آماره آزمون نسبت درستنمایی در طول یک گروه پیوستگی، برآورد بهینه‌ای از موقعیت QTL است.

چون به‌دست آوردن توزیع‌های آزمون به وسیله تجزیه مشکل است، آستانه بحرانی می‌تواند به‌صورت تجربی از روی آزمون‌های تبدیل یا جایگشت تعیین شود (چرچیل و دورج، ۱۹۹۴). برای داده‌های درخت تبریزی مربوط به مثال ۱۱-۲، ارزش‌های فنوتیپی را تبدیل می‌کنیم، تمامی ارتباطات را گسسته و از یک توزیع صفر تقلید می‌کنیم. برای هر جایگشت و تبدیلی که صورت می‌گیرد، $\max_l LRT_l$ را محاسبه کرده، هیستوگرام را رسم می‌کنیم و نقطه برش را به‌دست می‌آوریم. شکل ۱۱-۱ هیستوگرام مربوط به ۵۰۰۰ بار جایگشت داده‌ها و نتیجه حداکثر آماره LR است. با استفاده از این مطلب به‌عنوان یک مرجع، با برش ۵ درصدی از دنباله بالا، عدد ۵/۷۵۷ برای آماره نسبت درستنمایی به‌دست می‌آید که نشان می‌دهد نتوانسته‌ایم یک QTL معنی‌دار بیابیم[۱].

---

[۱]- اگر با یک خط افقی از ناحیه ۵ درصد هیستوگرام شکل ۱۱-۱ را برش دهیم، این خط هیستوگرام را در چند نقطه قطع می‌کند که بیشترین مقدار آن (که طبق شکل عدد ۵/۷۵۷ است) به عنوان آستانه معنی‌دار QTL خواهد بود، یعنی در تجزیه QTL هر ناحیه‌ای از ژنوم که مقدار آماره نسبت درستنمایی ($LRT = -2\log\lambda$) بیشتر از این مقدار آستانه داشته باشد، به عنوان QTL معنی‌دار شناسایی خواهد شد. در این مثال، مطابق با جدول ۱۱-۲ هیچ ناحیه‌ای دارای این مقدار LRT نبود (مترجمین).



شکل ۱۱-۱- هیستوگرام توزیع صفر $\max -2\log\lambda$ برای داده‌های مثال ۱۱-۲.
توزیع بر اساس ۵۰۰۰ بار جایگشت از داده‌ها به‌دست آمده است. نقطه برش ۵ درصد برای دنباله بالا، عدد ۵/۷۵۷ را به‌عنوان آستانه معنی‌داری LRT ارائه می‌دهد.

همچنین، می‌توان از توزیع جایگشت برای محاسبه مقدار p (p-value) استفاده نمود. حداکثر مشاهده شده (در این مثال) به صورت ۳/۲۵۷ بود و از ۵۰۰۰ جایگشت، عدد ۱۰۱۶ برای مقادیر بزرگ‌تر آماره آزمون به‌دست آمد که مقدار p آن برابر با $\frac{1016}{5000}=0.203$ است.

مثال ۱۱-۳. مثال ۱۱-۳ را مجدداً مشاهده کنید. دو لاین اینبرد، IR64 نیمه پاکوتاه و Azucena پابلند، با هم تلاقی داده شدند تا هتروزیگوت F_1 ایجاد شود. کروموزوم‌های هاپلوئید گرده‌های (گامت‌های) F_1 دو برابر شدند تا ۱۲۳ گیاه دابل هاپلوئید (DH) را تولید کنند. این گیاهان DH که معادل نتاج تلاقی برگشتی هستند، برای ۱۳۵ نشانگر RFLP و نشانگر ۴۰ ایزوزیم و RAPD تعیین ژنوتیپ شده و از روی آن‌ها یک نقشه پیوستگی که کل ژنوم ۱۲ کروموزومی برنج را پوشش می‌داد، ایجاد شد (یان و همکاران، ۱۹۹۸؛ شکل ۳-۳). هر یک از لاین‌های DH برای ارتفاع بوته در هر یک از ۱۰ هفته متوالی اندازه‌گیری شد.

برای بررسی وجود یک QTL در طول نقشه پیوستگی از روش حداکثر درستنمایی استفاده می‌شود. از کروموزوم ۱ به‌عنوان یک مثال برای توصیف این روش استفاده می‌شود. در جدول ۱۱-۴

فواصل تجمعی بین جفت نشانگرها برای ۱۸ نشانگر موجود روی کروموزوم ۱ برنج، به همراه اندازه‌های نمونه، میانگین‌ها و واریانس‌های ارتفاع گیاه اندازه‌گیری شده در سن ۱۰ هفتگی برای هر فاصله نشانگری ارائه شده است. از لحاظ نظری، این برآوردهای نمونه‌ای باید در بین فواصل نشانگری مختلف مشابه باشند، اما برآوردهای فوق به دلیل داده‌های گمشده که در مکان‌یابی QTL معمول هستند، متفاوت هستند.

با استفاده از برآورد این آماره‌های نمونه به‌عنوان مقادیر ابتدایی پارامترهای ناشناخته، اطلاعات موجود در جدول ۱۱-۴ برای جستجوی یک QTL در طول کروموزوم با فرض وجود QTL در هر ۲ سانتی‌مورگانی در هر فاصله نشانگر مورد استفاده قرار می‌گیرد. جدول ۱۱-۵ نتایج جستجوی QTL را در هر فاصله ۴ سانتی‌مورگانی ارایه می‌دهد که شامل برآوردهای دو ارزش ژنوتیپی QTL، واریانس باقیمانده، مقادیر لگاریتم درستنمایی تحت فرض‌های صفر و مقابل و LR در موقعیت‌های QTL مفروض است. با ترسیم مقادیر LR در برابر طول گروه پیوستگی، می‌توانیم مشاهده کنیم که چگونه آماره آزمون در موقعیت‌های QTL فرضی تغییر می‌کند. این احتمال وجود دارد که نقطه اوج پروفیل LR در ۲۱۷ سانتی‌مورگانی از اولین نشانگر در فاصله نشانگری RG810-RG331 مرتبط با MLE موقعیت QTL باشد.

برای آزمون اینکه QTL در نقطه اوج LR از نظر آماری معنی‌دار است، می‌توانیم به‌صورت تجربی آستانه بحرانی را بر اساس آزمون‌های جایگشت تعیین کنیم (فصل ۱۲ را برای جزئیات مشاهده کنید). با تغییر دادن داده‌های فنوتیپی، ۱۰۰۰ مقدار حداکثر LR برآورد می‌شود که صدک ۹۹ ام آن به‌عنوان نقطه برش (۱۲/۶۴) برای آزمون وجود یک QTL در طول کروموزوم در سطح معنی‌داری $\alpha = 0.01$ استفاده می‌شود.

یک نتیجه جانبی محاسبه آماره آزمون، برآورد سهم واریانس فنوتیپی توجیه شده به وسیله QTL (R^2) است که می‌تواند به‌صورت زیر محاسبه شود:

$$R^2 = \frac{\hat{\sigma}_T^2 - \hat{\sigma}_R^2}{\hat{\sigma}_R^2} \qquad (۱۰-۱۱)$$

که در آن، $\hat{\sigma}_T^2$ برآوردی از واریانس فنوتیپی کل (یعنی $\tilde{\sigma}^2$ تحت فرض صفر) و $\hat{\sigma}_R^2$ برآوردی از واریانس باقیمانده است (یعنی $\hat{\sigma}^2$ تحت فرض جایگزین).

اثر افزایشی QTL شناسایی شده روی کروموزوم ۱ برنج برابر با $\hat{a} = \hat{\mu}_1 - \hat{\mu}_0 = 30.97$ برآورد می‌شود که نسبت به انحراف معیار ارتفاع گیاه در جمعیت DH، ۱۳۴/۹ درصد است. با استفاده از معادله (۱۱-۱۰)، سهم واریانس فنوتیپی توجیه شده به وسیله QTL برابر با ۴۴/۵ درصد محاسبه می‌شود.

جدول ۱۱-۴- فواصل نقشه اندازه‌گیری شده بر حسب سانتی‌مورگان (cM) برای ۱۸ نشانگر موجود روی کروموزوم ۱ به همراه اندازه‌های نمونه، میانگین‌ها و واریانس‌های هر یک از ۱۷ فاصله نشانگری در یک جمعیت DH برنج.

نشانگر	فاصله تجمعی	فاصله جفتی	اندازه نمونه	میانگین	واریانس
RG472	0	19.2	96	109.63	539.48
RG246	19.2	16.1	100	109.85	524.21
K5	35.3	4.8	96	109.27	509.49
U10	40.1	4.7	78	110.49	506.71
RG532	44.8	15.3	100	109.85	537.80
W1	60.1	15.5	102	109.57	511.86
RG173	75.6	15	97	109.68	509.24
RZ276	90.6	3.8	104	110.20	529.99
Amy1B	94.4	3.3	104	109.85	536.55
RG146	97.7	34.3	99	109.44	543.70
RG345	132	2.5	105	109.71	528.72
RG381	134.5	23.5	97	109.37	520.59
RZ19	158	8.2	91	109.96	510.37
RG690	166.2	13.2	104	109.69	531.14
RZ730	179.4	33.1	89	110.17	538.40
RZ801	212.5	2.6	96	109.90	521.97
RG810	215.1	9.2	105	109.93	532.13
RG331	224.3				

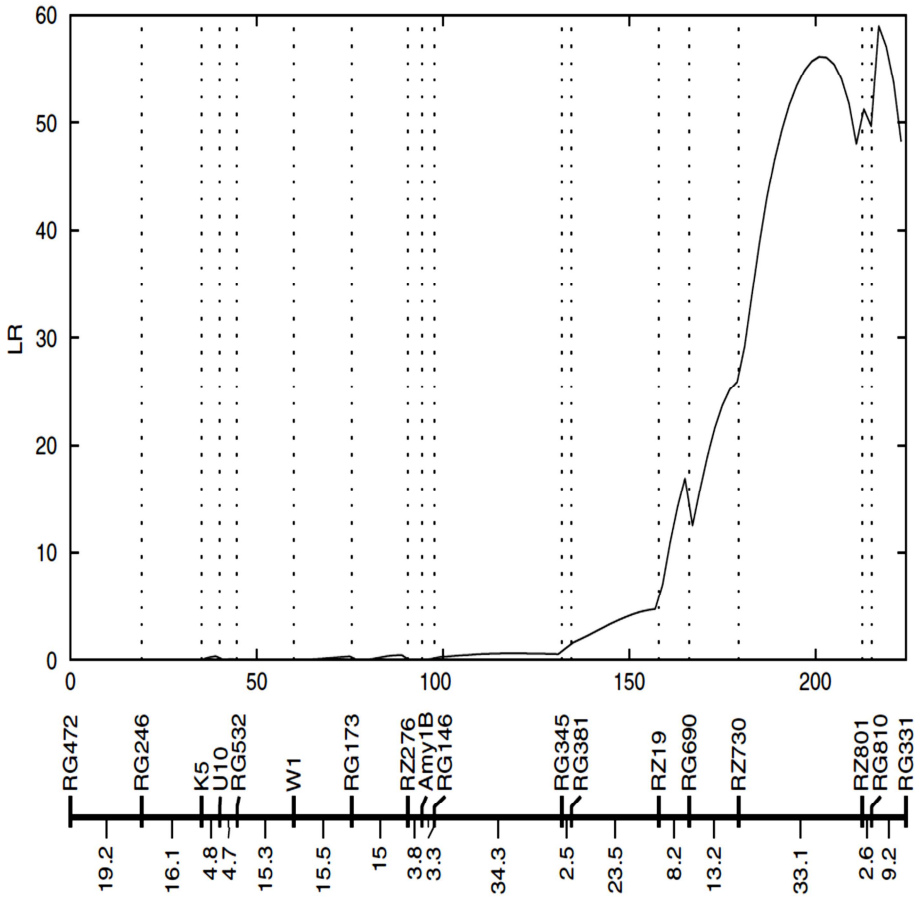
نکته: چون نشانگرهای مختلف از دست رفته‌اند، برآورد میانگین‌ها و واریانس‌های نمونه در بین فواصل نشانگری متفاوت هستند.

جدول ۱۱-۵- فرآیند برآورد مکان‌یابی فاصله‌ای QTL برای کروموزوم ۱ در یک جمعیت DH برنج.

n	LR	$LogL_0$	σ^2	μ_0	μ_1	موقعیت	فاصله
1	1	109.835	109.798	541.929	-433.818	0	95
1	5	109.789	109.85	541.928	-433.818	0	95
1	9	109.736	109.907	541.922	-433.818	0.001	95
1	13	109.685	109.959	541.91	-433.817	0.002	95
1	17	109.646	109.996	541.899	-433.816	0.004	95
2	21	109.761	108.94	497.122	-425.197	0.029	94
2	25	109.665	109.051	497.197	-425.204	0.014	94
2	29	109.506	109.193	497.266	-425.209	0.003	94
2	33	109.324	109.325	497.29	-425.211	0	94
3	37	108.182	110.686	477.832	-324.262	0.219	72
4	41	109.577	110.522	513.786	-345.029	0.03	76
5	45	108.813	109.88	515.15	-445.035	0.051	98
5	49	109.108	109.652	515.355	-445.055	0.011	98

ادامه جدول ۱۱-۵- فرآیند برآورد مکان‌یابی فاصله‌ای QTL برای کروموزوم ۱ در یک جمعیت DH برنج.

n	LR	$LogL_0$	σ^2	μ_0	μ_1	موقعیت	فاصله
5	53	109.562	109.383	515.418	-445.06	0.001	98
5	57	110.078	109.149	515.237	-445.045	0.031	98
6	61	108.572	109.336	489.133	-424.434	0.019	94
6	65	107.921	109.519	488.784	-424.408	0.073	94
6	69	107.171	109.688	488.169	-424.361	0.166	94
6	73	106.532	109.779	487.584	-424.303	0.281	94
7	77	109.198	109.789	509.19	-439.93	0.007	97
7	81	107.208	110.15	508.087	-439.882	0.102	97
7	85	104.375	110.519	504.806	-439.756	0.355	97
7	89	104.6	110.332	505.947	-439.705	0.458	97
8	93	111.159	109.952	534.307	-469.621	0.031	103
9	97	107.169	109.662	547.736	-448.043	0.116	98
10	101	105.543	109.882	542.842	-447.655	0.319	98
10	105	104.908	110.177	541.152	-447.601	0.427	98
10	109	104.561	110.451	539.543	-447.552	0.524	98
10	113	104.592	110.641	538.702	-447.519	0.59	98
10	117	104.935	110.721	538.838	-447.506	0.617	98
10	121	105.442	110.71	539.635	-447.508	0.613	98
10	125	105.982	110.641	540.656	-447.519	0.59	98
10	129	106.481	110.545	541.62	-447.534	0.561	98
11	133	105.749	110.98	515.684	-436.028	1.094	96
12	137	104.165	111.73	501.76	-394.034	1.939	87
12	141	103.212	112.374	495.53	-393.686	2.634	87
12	145	102.494	112.917	489.624	-393.33	3.347	87
12	149	102.098	113.277	485.604	-393.014	3.979	87
12	153	102.046	113.419	484.362	-392.777	4.453	87
12	157	102.324	113.349	485.997	-392.638	4.731	87
13	161	99.411	116.583	440.496	-407.358	10.928	91
13	165	98.17	117.625	419.984	-404.359	16.926	91
14	169	97.924	118.561	433.685	-393.583	15.806	88
14	173	96.276	119.96	400.347	-390.697	21.578	88
14	177	96.079	120.437	391.781	-388.902	25.168	88
15	181	94.988	121.898	364.697	-382.823	29.192	87
15	185	91.994	125.186	269.907	-377.98	38.879	87
15	189	91.895	126.999	235.728	-374.168	46.504	87
15	193	92.113	128.047	220.819	-371.591	51.657	87
15	197	92.313	128.583	215.013	-370.005	54.829	87
15	201	92.475	128.694	216.105	-369.339	56.16	87
15	205	92.641	128.359	225.014	-369.701	55.436	87
15	209	92.967	127.307	248.828	-371.535	51.769	87
16	213	94.335	123.975	302.895	-410.949	51.269	96
17	217	93.769	124.741	292.775	-449.026	58.997	105
17	221	93.759	124.256	300.457	-451.654	53.741	105



شکل ۱۱-۲- پروفیل مقدار LR در طول کروموزوم ۱ برای آزمون QTL کنترل کننده ارتفاع گیاه در سن ۱۰ هفتگی در یک جمعیت DH برنج (هوانگ و همکاران، ۱۹۹۷).

برنامه‌های دیگری برای تهیه نقشه‌های QTL در دسترس هستند. برای مثال، بسته نرم‌افزاری rQTL در نرم‌افزار آماری R وجود دارد که رایگان بوده و روی اینترنت نیز در دسترس است. همچنین، برنامه‌ای به نام QTL Cartographer وجود دارد که قابلیت دانلود آن بدون هزینه است.

۱۱-۳-۱- ارائه مدل در صورت وجود نوترکیبی مضاعف

درستنمایی (۱۱-۱) با فرض عدم وجود نوترکیبی مضاعف یعنی $r = r_1 + r_2$ ایجاد شد. اگر نوترکیبی مضاعف رخ دهد، ولی هیچ تداخلی فرض نشود، رابطه $r = r_1 + r_2 - 2r_1r_2$ را خواهیم

داشت. در این حالت، احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ‌های نشانگر در جدول ۱۰-۳ بیان شده است. الگوریتم EM می‌تواند بـرای بـرآورد r_1 ، r_2، ارزش‌هـای ژنـوتیپی QTL و واریانس باقیمانده استفاده شود. فرض می‌شود که $\omega_{j|i}$ احتمال شرطی کلی (پیشین) برای فرزنـد i ام با ژنوتیپ j ام QTL (۱ یـا ۰) اسـت. در مرحلـه E، احتمـال پسـین مربوطـه بـرای فرزنـد i ام به‌صورت زیر تعریف و محاسبه می‌شود:

$$P_{j|i} = \frac{\omega_{j|i} f_j(y_i)}{\omega_{1|i} f_1(y_i) + \omega_{0|i} f_0(y_i)} \qquad (۱۱-۱۱)$$

در مرحله M، پارامترها با رابطه زیر برآورد می‌شوند:

$$\begin{aligned}
\hat{\mu}_j &= \frac{\sum_{i=1}^{n} P_{j|i} y_i}{\sum_{j=1}^{n} P_{j|i}} \\
\hat{\sigma}^2 &= \frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=0}^{1}\left(y_i - \hat{\mu}_j\right)^2 P_{j|i}\right] \\
\hat{r}_1 &= \frac{1}{n}\left[\sum_{i=1}^{n_1+n_2} P_{0|i} + \sum_{i=1}^{n_3+n_4} P_{1|i}\right] \\
\hat{r}_2 &= \frac{1}{n}\left[\sum_{i=1}^{n_1+n_3} P_{0|i} + \sum_{i=1}^{n_2+n_4} P_{1|i}\right]
\end{aligned} \qquad (۱۲-۱۱)$$

مراحل E و M تا جایی تکرار می‌شوند که پارامترها همگرا شوند. برآوردهـا در حالـت همگرایـی همان MLE های پارامترها هستند.

۱۱-۳-۲- ارایه مدل در صورت وجود تداخل

ترتیب نشانگر-QTL را به صورت M-QTL-N در نظر بگیرید. فرض کنید که g_{11}، g_{10}، g_{01} و g_{00} به ترتیب احتمال دو کراس‌اور یکی بـین M و QTL و دیگـری بـین QTL و N، فقـط یـک کراس‌اور بین M و QTL، فقط یک کـراس‌اور بـین QTL و N و عـدم وجـود کـراس‌اور بـین سـه جایگاه ژنی باشند.

بر اساس تجزیه سه نقطه‌ای در جدول ۴-۲، فراوانی‌های توأم ژنوتیپی نشانگر-QTL-نشانگر بـر حسب g_{11}، g_{10}، g_{01} و g_{00} به‌دست می‌آیند که در جدول ۱۱-۹ ارایه شده‌اند. به‌جای استفاده از جدول ۱۱-۷، در اینجا از احتمالات شرطی که عموماً به صورت $\omega_{j|i}$ بیان شده و از روی جـدول ۱۱-۹ به‌دست آمده است، استفاده می‌کنیم تا درستنمایی را مشابه با معادله (۱۱-۷) ایجـاد کنـیم.

احتمالات توأم ژنوتیپ‌های نشانگر-QTL بر حسب این چهار g (جدول ۱۱-۶) بیان می‌شوند کـه از روی آن، احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ‌های فاصله نشانگری ($\omega_{j|i}$) مـی-توانند به‌دست آیند.

جدول ۱۱-۶- فراوانی‌های ژنوتیپی توأم سه نقطه‌ای در جمعیت تلاقی برگشتی.

فاصله نشانگر		ژنوتیپ QTL	
ژنوتیپ	اندازه نمونه	1	0
11	n_1	g_{00}	g_{11}
10	n_2	g_{01}	g_{10}
01	n_3	g_{10}	g_{01}
00	n_4	g_{11}	g_{00}

برای بـرآورد احتمـالات g، ارزش‌هـای ژنـوتیپی QTL و واریـانس باقیمانـده از الگـوریتم EM استفاده می‌شود. در مرحله E، احتمال پسین را به شکلی مشـابه معادلـه (۱۱-۱۱) تعریـف کـرده و محاسبه می‌کنیم. در مرحله M، ارزش‌های ژنوتیپی QTL و واریانس باقیمانـده را بـا معادلـه (۱۱-۱۲) و احتمالات g را با معادله زیر برآورد می‌نماییم:

$$\hat{g}_{00} = \frac{1}{n_1 + n_4}\left[\sum_{i=1}^{n_1} P_{1|i} + \sum_{i=1}^{n_4} P_{0|i}\right]$$

$$\hat{g}_{01} = \frac{1}{n_2 + n_3}\left[\sum_{i=1}^{n_2} P_{1|i} + \sum_{i=1}^{n_3} P_{0|i}\right]$$

$$\hat{g}_{10} = \frac{1}{n_2 + n_3}\left[\sum_{i=1}^{n_3} P_{1|i} + \sum_{i=1}^{n_2} P_{0|i}\right]$$

$$\hat{g}_{11} = \frac{1}{n_1 + n_4}\left[\sum_{i=1}^{n_4} P_{1|i} + \sum_{i=1}^{n_1} P_{0|i}\right]$$

پس از محاسبه MLE های مربوط به g، نرخ‌های نوترکیبی به صورت زیر برآورد می‌شوند:

$$\begin{aligned} \hat{r}_1 &= \hat{g}_{10} + \hat{g}_{11} \\ \hat{r}_2 &= \hat{g}_{01} + \hat{g}_{11} \\ \hat{r} &= \hat{g}_{01} + \hat{g}_{10} \end{aligned} \qquad (۱۳-۱۱)$$

ضریب تداخل میوزی درجه‌ای را اندازه‌گیری می‌کند که از راه آن، نوترکیبی‌ها در دو فاصله مجاور هم به وسیله یکدیگر تحت‌تأثیر قرار می‌گیرند[1] و با رابطهٔ زیر محاسبه می‌شود:

$$\hat{I} = 1 - \frac{\hat{g}_{11}}{\hat{r}_1 \hat{r}_2} \qquad (۱۴-۱۱)$$

آزمون معنی‌داری تداخل می‌تواند با محاسبه نسبت دو لگاریتم درستنمایی تحت فرض صفر ($I = 0$) و فرض مقابل ($I \neq 0$) که به ترتیب با توجه به بخش‌های ۱۱-۳-۱ و ۱۱-۳-۲ برآورد شده‌اند، فرموله شود. این نسبت دارای توزیع مجانبی x^2 با یک درجه آزادی است.

۱۱-۴- مکان‌یابی فاصله‌ای QTL در جمعیت F_2

همانند جمعیت تلاقی برگشتی، سه حالت مختلف را برای F_2 در نظر خواهیم گرفت: ۱- هیچ نوترکیبی مضاعفی در فاصله نشانگری وجود ندارد، ۲- نوترکیبی‌ها در فواصل مختلف مستقل هستند و ۳- تداخل میوزی بین فواصل مختلف رخ می‌دهد. در هر وضعیت، الگوریتم EM را برای برآورد پارامترها تهیه خواهیم کرد.

۱۱-۴-۱- عدم وجود نوترکیبی مضاعف

در طرح F_2، هر نشانگر می‌تواند سه ارزش داشته باشد و زمانی که یک جفت نشانگر برای تشکیل یک فاصله در نظر گرفته شود، این حالت منجر به ایجاد ۹ کلاس نشانگری می‌شود که احتمالات شرطی آن‌ها در جدول ۱۰-۵ ارائه شده است. احتمال شرطی ژنوتیپ QTL مشروط به ژنوتیپ نشانگر به‌صورت نسبت فراوانی ژنوتیپی توأم نشانگر-QTL به فراوانی ژنوتیپی نشانگر محاسبه می‌شود. نسبت های r_1، r_2 و r به ترتیب نرخ‌های نوترکیبی بین نشانگر M (با دو آلل M و m) و QTL (با دو آلل Q و q)، بین QTL و نشانگر N (با دو آلل N و n) و بین دو نشانگر مجاور همدیگر هستند.

فرض کنید هیچ نوترکیبی مضاعفی وجود نداشته باشد. در این صورت، $r = r_1 + r_2$ یا $r_1 r_2 \approx 0$ خواهد بود. از این‌رو، با تعریف $\theta = \frac{r_1}{r}$ و $\eta = \frac{r^2}{\left[(1-r)^2 + r^2\right]}$، احتمالات شرطی ژنوتیپ‌های QTL مشروط به کلاس‌های مختلف نشانگر به‌دست می‌آید (جدول ۱۱-۷).

---

[1]- به عبارت دیگر، ضریب تداخل میوزی میزان تأثیر دو نوترکیبی ایجاد شده در دو فاصله مجاور هم بر یکدیگر است (مترجمین).

درستنمایی

مشابه بسط روابط در بخش ۱۱-۲، فرض کنید چگالی‌های f_2 ، f_1 و f_0 را مطابق با ژنوتیپ‌های QTL داریم. بر اساس جدول ۱۱-۷ تابع درستنمایی به صورت زیر است:

(۱۵-۱۱)

$$
\begin{aligned}
L = & \prod_{i=1}^{n_{22}} \left[1 f_2(y_i) + 0 f_1(y_i) + 0 f_0(y_i) \right] \\
& \times \prod_{i=1}^{n_{21}} \left[(1-\theta) f_2(y_i) + \theta f_1(y_i) + 0 f_0(y_i) \right] \\
& \times \prod_{i=1}^{n_{20}} \left[(1-\theta)^2 f_2(y_i) + 2\theta(1-\theta) f_1(y_i) + \theta^2 f_0(y_i) \right] \\
& \times \prod_{i=1}^{n_{12}} \left[\theta f_2(y_i) + (1-\theta) f_1(y_i) + 0 f_0(y_i) \right] \\
& \times \prod_{i=1}^{n_{11}} \left[\eta\theta(1-\theta) f_2(y_i) + \left(1 - 2\eta\theta(1-\theta)\right) f_1(y_i) + \eta\theta(1-\theta) f_0(y_i) \right] \\
& \times \prod_{i=1}^{n_{10}} \left[0 f_2(y_i) + (1-\theta) f_1(y_i) + \theta f_0(y_i) \right] \\
& \times \prod_{i=1}^{n_{02}} \left[\theta^2 f_2(y_i) + 2\theta(1-\theta) f_1(y_i) + (1-\theta)^2 f_0(y_i) \right] \\
& \times \prod_{i=1}^{n_{01}} \left[0 f_2(y_i) + \theta f_1(y_i) + (1-\theta) f_0(y_i) \right] \\
& \times \prod_{i=1}^{n_{00}} \left[0 f_2(y_i) + 0 f_1(y_i) + 1 f_0(y_i) \right]
\end{aligned}
$$

برای آزمون معنی‌داری QTL، فرض صفر یعنی عدم وجود QTL در نشانگر بدین صورت فرموله می‌شود:

$$H_0 : f_2(y_i) = f_1(y_i) = f_0(y_i) \equiv f(y_i)$$

و برای آزمون این فرض در موقعیت x که در آن $\theta = \frac{r_1(x)}{r}$ است، آماره نسبت درستنمایی مشابه با آنچه که در بخش ۱۱-۲ انجام شد، به شکل زیر است:

(۱۶-۱۱)

$$\lambda(x)=\frac{\max_{H_0} L\left(f\left(y_i\right)\right)}{\max_{H_1} L\left(f_2\left(y_i\right), f_1\left(y_i\right), f_0\left(y_i\right), \theta\right)}$$

جدول ۱۱-۷- احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ‌های نشانگر در جمعیت F_2.

فاصله نشانگری		ژنوتیپ QTL		
ژنوتیپ	اندازه نمونه	QQ (2)	Qq (1)	qq (0)
MMNN (22)	n_{22}	1	0	0
MMNn (21)	n_{21}	$1-\theta$	θ	0
MMnn (20)	n_{20}	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2
MmNN (12)	n_{12}	θ	$1-\theta$	0
MmNn (11)	n_{11}	$\eta\theta(1-\theta)$	$1-2\eta\theta(1-\theta)$	$\eta\theta(1-\theta)$
Mmnn (10)	n_{10}	0	$1-\theta$	0
mmNN (02)	n_{02}	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$
mmNn (01)	n_{01}	0	θ	$1-\theta$
mmnn (00)	n_{00}	0	0	1
کل	n	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

توجه کنید که $\theta=\frac{r_1}{r}$ و $\eta=\frac{r^2}{\left[(1-r)^2+r^2\right]}$ است.

حداکثر رساندن درستنمایی

اکنون به حداکثر رساندن درستنمایی (۱۱-۱۵) باز می‌گردیم، ولی در ابتدا دو فرض را در نظر می‌گیریم:

۱- فرض می‌کنیم که چگالی‌های $f_j\left(y_i\right)\left(j=2,1,0\right)$ چگالی‌های نرمال به شکل زیر هستند:

$$f_j\left(y_i\right)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\frac{\left(y_i-\mu_j\right)^2}{2\sigma^2}\right]$$

که دارای میانگین‌های $\mu_2 = \mu + a$ ، $\mu_1 = \mu + d$ و $\mu_0 = \mu - a$ و واریانس σ^2 هستند.

۲- از عبارت $\omega_{j|i}$ برای نشان دادن وزن‌های ژنوتیپ‌های QTL مربوط به فرد i ام استفاده می‌کنیم. برای مثال، اگر فرد i دارای ژنوتیپ نشانگر ۱۲ باشد، سپس:

$$\omega_{2|i} = \theta$$
$$\omega_{1|i} = 1 - \theta$$
$$\omega_{0|i} = 0$$

با استفاده از این روابط، می‌توانیم لگاریتم درستنمایی (۱۱-۱۵) را مشروط به مشاهدات y و فاصله نشانگری M-N به‌صورت زیر بنویسیم:

$$\log L\left(\Omega \mid y, M-N\right) = \sum_{i=1}^{n} \log \sum_{j=0}^{2} \left[\omega_{j|i} f_j\left(y_i\right)\right] \qquad (۱۱-۱۷)$$

که در آن:

$$\Omega = \left(\mu_2, \mu_1, \mu_0, \sigma^2, \theta\right) = \left(\Omega_q, \theta\right)$$

با مشتق گرفتن از لگاریتم درستنمایی نسبت به هر یک از پارامترهای ناشناخته موجود در بردار بالا داریم:

(۱۱-۱۸)

$$\frac{\partial}{\partial \Omega} \log L\left(\Omega \mid y, M-N\right) = \sum_{i=1}^{n} \left(\frac{\omega_{j|i} f_j\left(y_i\right)}{\sum_{j'=0}^{2} \left[\omega_{j'|i} f_{j'}\left(y_i\right)\right]} \right) \left(\frac{\partial}{\partial \Omega_q} \log f_j\left(y_i\right) + \frac{1}{\omega_{j|i}} \frac{\partial}{\partial \theta} \omega_{j|i} \right)$$
$$= \sum_{i=1}^{n} P_{j|i} \left(\frac{\partial}{\partial \Omega_q} \log f_j\left(y_i\right) + \frac{1}{\omega_{j|i}} \frac{\partial}{\partial \theta} \omega_{j|i} \right)$$

که در آن:

$$P_{j|i} = \frac{\omega_{j|i} f_j\left(y_i\right)}{\sum_{j'=0}^{2} \left[\omega_{j'|i} f_{j'}\left(y_i\right)\right]} \qquad (۱۱-۱۹)$$

که می‌تواند به‌صورت احتمالات پسین ژنوتیپ‌های QTL برای فرد i ام مشروط به ژنوتیپ نشانگر در نظر گرفته شود. توجه شود که:

$$\sum_{i=1}^{2} P_{j|i} = 1$$

در احتمالات پسین، $\omega_{j|i}$ ها همان احتمالات پیشین برای فرد i ام هستند که ژنوتیپ j ام QTL را حمل می‌کنند، حال آن که $f_j\left(y_i\right)$ ها درستنمایی‌های فردی هستند. احتمال پسین این که یک فرد ژنوتیپ QTL خاصی را حمل می‌کند، می‌تواند برای استنباط هویت فرد مورد استفاده قرار گیرد.

با گرفتن مشتق جزیی از معادله (۱۱-۸) نسبت به هر پارامتر ناشناخته و برابر صفر قرار دادن آن، MLE های ارزش ژنوتیپی μ_j برای ژنوتیپ j ام QTL به‌صورت زیر به‌دست می‌آید:

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} P_{j|i} y_i}{\sum_{i=1}^{n} P_{j|i}}, \quad j = 2,1,0 \qquad (۱۱-۲۰)$$

و MLE مربوط به σ^2 در جمعیت مکان‌یابی برابر است با:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{2}\left[P_{j|i}\left(y_i - \hat{\mu}_j\right)^2\right] \qquad (۱۱-۲۱)$$

آثار افزایشی (a) و غالبیت (d) مربوط به QTL نیز می‌توانند از روی MLE های مربوط به μ_jها به‌صورت زیر برآورد شوند:

$$\hat{a} = \frac{1}{2}\left(\hat{\mu}_2 - \hat{\mu}_0\right)$$

$$\hat{d} = \hat{\mu}_1 - \frac{1}{2}\left(\hat{\mu}_2 + \hat{\mu}_0\right)$$

به‌طور مشابه، MLE مربوط به موقعیت QTL (که با نسبت θ اندازه‌گیری می‌شود) می‌تواند با روش جستجو کردن که در آن، یک QTL فرضی در فواصل ۲ سانتی‌مورگانی در هر فاصله نشانگری فرض می‌شود، به‌دست آید. آماره نسبت درستنمایی ($\lambda\left(x\right)$) که با معادله (۱۱-۱۶) محاسبه می‌شود، در مقابل طول گروه پیوستگی (یعنی x) رسم می‌شود. نقطه اوج منحنی $\lambda\left(x\right)$ به برآورد بهینه موقعیت QTL تعلق دارد.

مثال ۱۱-۴. **(داده‌های F_2 موش).** چورود و همکاران (۱۹۹۶) یک نقشه پیوستگی را با استفاده از ۷۵ نشانگر ریزماهواره در جمعیتی متشکل از ۵۳۵ نتاج F_2 متعلق به دو سویه بزرگ (LG/J) و کوچک (SM/J) تهیه نمودند. نتاج F_2 برای توده بدنی در فواصل ۱۰ هفته‌ای که از سن ۷ روزگی شروع شد، اندازه‌گیری شدند. وزن‌های خام برای آثار متغیرهای همبسته مربوط به مادر، تعداد همزادان در تولد و شکم زایش، تصحیح شدند، ولی این تصحیح برای اثر جنسیت انجام نشد.

کروموزوم ۱ شامل ۹ نشانگر به عنوان مثالی برای مکان‌یابی QTL مؤثر بر وزن بدن در ۱۰ هفتگی (جدول ۱۱-۸) با استفاده از روش مکان‌یابی فاصله‌ای مبتنی بر ML استفاده شد.

نقشه حاصل از روش درستنمایی در شکل ۱۱-۳ نشان داده شده است. حداکثر درستنمایی در فاصله نشانگری [D2MIT93-D2MIT389] در ۳۷ سانتی‌مورگانی از اولین نشانگر سمت چپ کروموزوم قرار دارد که پیشنهاد می‌کند ممکن است یک QTL در این مکان وجود داشته باشد. LR در نقطه اوج ۳۳/۸۱ است که بیشتر از آستانه بحرانی ۱۴/۵۸ محاسبه شده از روی آزمون‌های جایگشت است و نشان می‌دهد که QTL شناسایی شده معنی‌دار است.

۱۱-۴-۲- استقلال

در بخش ۱۱-۴-۱، احتمالات شرطی با فرض عدم وجود نوترکیبی مضاعف در فواصل مختلف به‌دست آمدند. این فرض ممکن است زمانی صحیح باشد که چگالی نقشه بالا باشد (نقشه اشباع باشد) که در این مورد، حاصل‌ضرب نرخ‌های نوترکیبی بین نشانگر M و QTL (r_1) و بین QTL و نشانگر N (r_2) نزدیک به صفر می‌شود.

جدول ۱۱-۸- بخشی از داده‌های موش حاصل از چورود و همکاران (۱۹۹۶). شش نشانگر ابتدایی، D1Mit3، D1Mit20، D1Mit7، D1Mit11، D1Mit14 و D1Mit17 روی کروموزوم ۱ به ترتیب در ۰، ۶/۳، ۴۱/۶، ۵۲/۵، ۷۷/۶ و ۱۱۹/۲ سانتی‌مورگانی نشان داده شده‌اند. صفت فنوتیپی مربوطه وزن بدن در سن ۱۰ هفتگی است.

افزایش وزن	D1Mit17	D1Mit14	D1Mit11	D1Mit7	D1Mit20	D1Mit3	شماره
۲۹/۸۰۴	۳	۲	۳	۲	۳	۱	۴۴
۲۰/۶۲۲	۲	۲	۲	۲	۲	۳	۴۵
۲۵/۹۵۹	۲	۲	۲	۱	۲	۲	۴۶
۲۷/۴۸۴	۲	۱	۱	۱	۳	۱	۴۷
۲۴/۳۳۸	۱	۲	۲	۳	۲	۳	۵۱
۳۴/۹۸۰	۳	۲	۲	۲	۲	۲	۵۴
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
۳۰/۱۹۹	۲	۲	۲	۳	۲	۲	۶۳۴
۳۰/۳۰۳	۲	۱	۱	۱	۱	۱	۶۳۶
۳۲/۶۷۰	۲	۲	۲	۲	۲	۲	۶۳۷
۳۶/۸۴۹	۲	۱	۲	۲	۲	۲	۶۳۸
۲۵/۲۹۱	۲	۲	۲	۳	۲	۳	۶۴۰
۳۷/۸۲۴	۱	۳	۲	۳	۲	۲	۶۴۱
۱۶/۴۴۱	۲	۳	۲	۳	۱	۱	۶۴۲

در اینجا، مدلی را شرح می‌دهیم که فرض عدم وجود نوترکیبی مضاعف را نادیده می‌گیرد، ولی وقایع نوترکیبی را مستقل فرض می‌کند.

در جدول ۱۰-۵ فراوانی‌های توأم فاصله نشانگری (M-N) و ژنوتیپ‌های QTL در افراد F_2 تحت فرض استقلال نوترکیبی، ولی با در نظر گرفتن نوترکیبی مضاعف نشان داده شده است. همانند آنچه که قبلاً دیدیم، فرض کنید $\omega_{j|i}$ احتمال شرطی ژنوتیپ j ام QTL ($j = 2,1,0$) مشروط به ژنوتیپ فاصله نشانگری در فرزند i ام باشد. مشابه درستنمایی (۱۱-۱۵) برای عدم وجود نوترکیبی مضاعف، درستنمایی مربوط به احتمالات شرطی حاصل از جدول ۱۰-۵ نیز می‌تواند تهیه شود. با به حداکثر رساندن درستنمایی، الگوریتم EM می‌تواند برای فراهم کردن MLE های مربوط به پارامترها بکار گرفته شود.

در مرحله E، احتمال پسین ژنوتیپ j ام QTL برای فرزند i ام مشروط به ژنوتیپ فاصله نشانگری و ارزش فنوتیپی آن می‌تواند با استفاده از معادله (۱۱-۱۹) محاسبه شود. در مرحله M، ارزش‌های ژنوتیپی QTL و واریانس باقیمانده از روی احتمالات پسین با استفاده از معادلات (۱۱-۲۰) و (۱۱-۲۱) برآورد می‌شوند، در حالی که نرخ‌های نوترکیبی نشانگر-QTL با رابطه زیر برآورد می‌شوند:

$$\hat{r}_1 = \frac{1}{n}\left[\sum_{i=1}^{n_{22}+n_{21}+n_{20}}\left(P_{1|i}+2P_{0|i}\right)\right.$$
$$\left.+\sum_{i=1}^{n_{12}+n_{11}+n_{10}}\left(P_{2|i}+2\phi_1 P_{1|i}+P_{0|i}\right)+\sum_{i=1}^{n_{02}+n_{01}+n_{00}}\left(2P_{2|i}+P_{1|i}\right)\right]$$

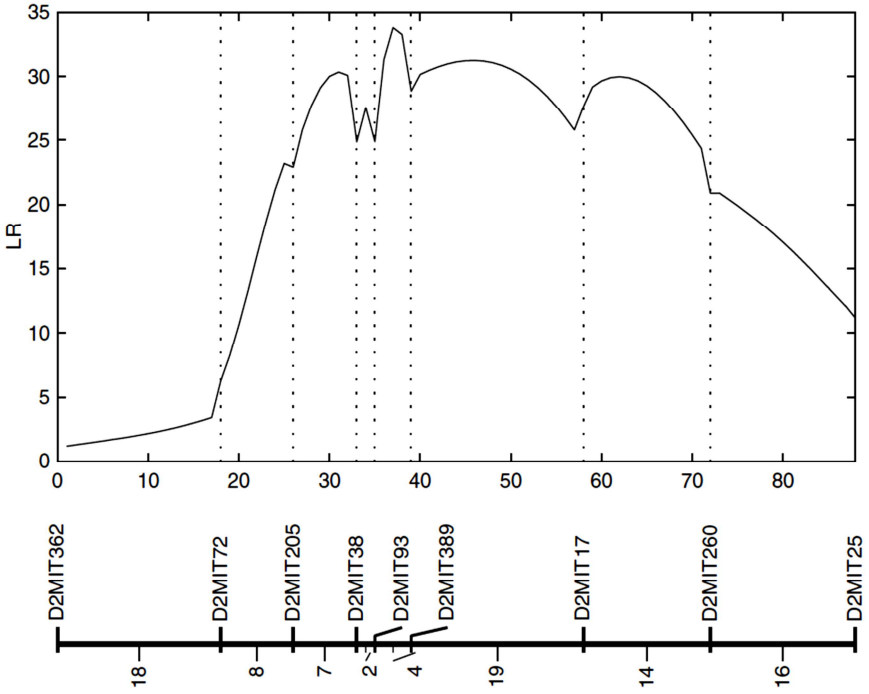
$$\hat{r}_2 = \frac{1}{n}\left[\sum_{i=1}^{n_{22}+n_{12}+n_{02}}\left(P_{1|i}+2P_{0|i}\right)\right.$$
$$\left.+\sum_{i=1}^{n_{21}+n_{11}+n_{01}}\left(P_{2|i}+2\phi_2 P_{1|i}+P_{0|i}\right)+\sum_{i=1}^{n_{20}+n_{10}+n_{00}}\left(2P_{2|i}+P_{1|i}\right)\right]$$

که در آن:

$$\phi_1 = \frac{r_1^2}{\left(1-r_1\right)^2+r_1^2}$$

$$\phi_2 = \frac{r_2^2}{\left(1-r_2\right)^2+r_2^2}$$

مراحل E و M تا رسیدن به همگرایی تکرار می شوند. برآوردها در هنگام همگرایی به عنوان MLE های پارامترها در نظر گرفته می‌شود.



شکل ۱۱-۳- پروفیل LR در طول کروموزوم ۱ برای آزمون QTL که وزن بدن را در سن ۱۰ هفتگی در جمعیت F_2 موش مثال ۱۱-۴ کنترل می‌کند (چورود و همکاران، ۱۹۹۶).

۱۱-۴-۳- تداخل

این امکان وجود دارد که نوترکیبی‌ها در فواصل مختلف با یکدیگر تداخل داشته باشند، به‌طوری که ضریب تداخل باید در مدل مکان‌یابی در نظر گرفته شود. همانند آنچه که در تلاقی برگشتی مشاهده شد، g_{11}، g_{10}، g_{01} و g_{00} به عنوان احتمالات وقوع نوترکیبی در دو فاصله با ترتیب M-QTL-N تعریف می‌شوند. بر اساس تجزیه سه نقطه‌ای در جدول ۴-۲، فراوانی‌های ژنوتیپی توأم نشانگر-QTL-نشانگر را بر حسب g_{11}، g_{10}، g_{01} و g_{00} به‌دست می‌آوریم که در جدول (۱۱-۹) ارایه شده‌اند. به جای استفاده از جدول ۱۱-۷، در اینجا از احتمالات شرطی حاصل از جدول (۱۱-۹) که عموماً با $\omega_{j|i}$ نشان داده می‌شود برای ایجاد درستنمایی همانند معادله (۱۱-۷) استفاده می‌کنیم.

جدول ۱۱-۹- فراوانی‌های ژنوتیپی توأم سه نقطه‌ای در جمعیت F_2.

فاصله نشانگری		ژنوتیپ QTL		
ژنوتیپ	اندازه نمونه	QQ(2)	Qq(1)	Qq(0)
MMNN(22)	n_{22}	g_{00}^2	$2g_{00}g_{11}$	g_{11}^2
MMNn(21)	n_{21}	$2g_{00}g_{01}$	$2(g_{00}g_{10}+g_{01}g_{11})$	$2g_{11}g_{10}$
MMnn(20)	n_{20}	g_{01}^2	$2g_{01}g_{10}$	g_{10}^2
MmNN(12)	n_{12}	$2g_{00}g_{10}$	$2(g_{00}g_{11}+g_{01}g_{10})$	$2g_{01}g_{11}$
MmNn(11)	n_{11}	$2(g_{00}g_{01}+g_{00}g_{10})$	$2(g_{00}^2+g_{01}^2+g_{10}^2+g_{11}^2)$	$2(g_{00}g_{01}+g_{00}g_{10})$
Mmnn(10)	n_{10}	$2g_{01}g_{11}$	$2(g_{00}g_{11}+g_{01}g_{10})$	$2g_{00}g_{10}$
mmNN(02)	n_{02}	g_{10}^2	$2g_{10}g_{01}$	g_{01}^2
mmNn(01)	n_{01}	$2g_{11}g_{10}$	$2(g_{00}g_{10}+g_{01}g_{11})$	$2g_{00}g_{01}$
Mmnn(00)	n_{00}	g_{11}^2	$2g_{00}g_{11}$	g_{00}^2

با تعریف و محاسبه احتمالات پسین بر اساس معادله (۱۱-۱۹)، MLE های دقیق برای t امـین احتمالات وقوع کراس‌اورها به‌دست می‌آیند که به صورت زیر هستند:

$$\hat{g}_{00}=\frac{1}{2n}\left[\sum_{i=1}^{n_{22}}\left(2P_{2|i}+P_{1|i}\right)+\sum_{i=1}^{n_{21}}\left(2P_{2|i}+\phi_1P_{1|i}\right)+\sum_{i=1}^{n_{12}}\left(2P_{2|i}+\phi_2P_{1|i}\right)\right.$$
$$+\sum_{i=1}^{n_{11}}\left(\phi_3P_{2|i}+2\phi_4P_{1|i}+\phi_3P_{0|i}\right)$$
$$\left.+\sum_{i=1}^{n_{10}}\left(\phi_2P_{1|i}+P_{0|i}\right)+\sum_{i=1}^{n_{01}}\left(\phi_2P_{1|i}+P_{0|i}\right)+\sum_{i=1}^{n_{00}}\left(P_{1|i}+2P_{0|i}\right)\right] \qquad (۱۱-۲۲)$$

(۲۳-۱۱)

$$\hat{g}_{01}=\frac{1}{2n}\left[\sum_{i=1}^{n_{21}}\left(P_{2|i}+\left(1-\phi_1\right)P_{1|i}\right)+\sum_{i=1}^{n_{20}}\left(2P_{2|i}+P_{1|i}\right)+\sum_{i=1}^{n_{12}}\left(\left(1-\phi_2\right)P_{1|i}+P_{0|i}\right)\right.$$
$$+\sum_{i=1}^{n_{11}}\left(\phi_3P_{2|i}+2\phi_4'P_{1|i}+\phi_3P_{0|i}\right)$$
$$\left.+\sum_{i=1}^{n_{10}}\left(P_{2|i}+\left(1-\phi_2\right)P_{1|i}\right)+\sum_{i=1}^{n_{02}}\left(P_{1|i}+2P_{0|i}\right)+\sum_{i=1}^{n_{01}}\left(\left(1-\phi_1\right)P_{1|i}+P_{0|i}\right)\right]$$

(۲۴-۱۱)

$$\hat{g}_{10}=\frac{1}{2n}\left[\sum_{i=1}^{n_{21}}\left(\phi_1P_{1|i}+P_{0|i}\right)+\sum_{i=1}^{n_{20}}\left(P_{1|i}+2P_{0|i}\right)+\sum_{i=1}^{n_{12}}\left(P_{2|i}+\left(1-\phi_2\right)P_{1|i}\right)\right.$$
$$+\sum_{i=1}^{n_{11}}\left(\left(1-\phi_3\right)P_{2|i}+2\phi_4''P_{1|i}+\left(1-\phi_3\right)P_{0|i}\right)$$

(۲۵-۱۱)

$$\left.+\sum_{i=1}^{n_{10}}\left(\left(1-\phi_2\right)P_{1|i}+P_{0|i}\right)+\sum_{i=1}^{n_{02}}\left(2P_{2|i}+P_{1|i}\right)+\sum_{i=1}^{n_{01}}\left(P_{2|i}+\phi_1P_{1|i}\right)\right]$$

$$\hat{g}_{11}=\frac{1}{2n}\left[\sum_{i=1}^{n_{22}}\left(P_{1|i}+2P_{0|i}\right)+\sum_{i=1}^{n_{21}}\left(\phi_1P_{1|i}+P_{0|i}\right)+\sum_{i=1}^{n_{12}}\left(\phi_1P_{1|i}+P_{0|i}\right)\right.$$
$$+\sum_{i=1}^{n_{11}}\left(\left(1-\phi_3\right)P_{2|i}+2\left(1-\phi_4-\phi_4'-\phi_4''\right)P_{1|i}+\left(1-\phi_3\right)P_{0|i}\right)$$
$$\left.+\sum_{i=1}^{n_{10}}\left(P_{2|i}+\phi_2P_{1|i}\right)+\sum_{i=1}^{n_{01}}\left(P_{0|i}+\left(1-\phi_1\right)P_{1|i}\right)+\sum_{i=1}^{n_{00}}\left(2P_{2|i}+P_{1|i}\right)\right]$$

که در این روابط داریم:

$$\phi_1 = \frac{g_{00}g_{10}}{g_{00}g_{10} + g_{01}g_{11}}$$

$$\phi_2 = \frac{g_{00}g_{11}}{g_{00}g_{11} + g_{01}g_{10}}$$

$$\phi_3 = \frac{g_{00}g_{01}}{g_{00}g_{01} + g_{10}g_{11}}$$

$$\phi_4 = \frac{g_{00}^2}{g_{00}^2 + g_{01}^2 + g_{10}^2 + g_{11}^2}$$

$$\phi_4' = \frac{g_{01}^2}{g_{00}^2 + g_{01}^2 + g_{10}^2 + g_{11}^2}$$

$$\phi_4'' = \frac{g_{10}^2}{g_{00}^2 + g_{01}^2 + g_{10}^2 + g_{11}^2}$$

MLE های مربوط به g ها می‌توانند برای برآورد r_1 و r_2 با استفاده از معادله (۱۱-۱۳) استفاده شوند و سپس درجه تداخل با معادله (۱۱-۱۴) برآورد شود. معنی‌دار بودن تداخل می‌تواند با محاسبه نسبت درستنمایی تحت فرض صفر (I = ۰) و فرض مقابل (I ≠ ۰) آزمون شود.

۱۱-۴-۴- آزمون کردن فرض‌ها

وجود یک QTL در حال تفرق پیوسته با نشانگرهای شناخته شده، می‌تواند به وسیله آزمون نسبت درستنمایی آزمون شود. در یک جمعیت مکان‌یابی، فرض‌های مربوط به آزمون QTL فرضی در یک جمعیت F_2 می‌تواند به صورت زیر فرموله شود:

$$H_0: \ \mu_2 = \mu_1 = \mu_0 \equiv \mu \qquad (۱۱-۲۶)$$

H_1 : حداقل یکی از تساوی‌ها صادق نباشد

این فرض صفر معادل $a = d = 0$ در جمعیت F_2 است. همانند تلاقی برگشتی، مقدار LR برای فرض (۱۱-۲۶) می‌تواند محاسبه شده و با آستانه بحرانی حاصل از آزمون‌های جایگشت مقایسه شود. ولی روش آزمون‌های معنی‌داری در F_2 باید جستجو و مشخص شود:

۱- وجود یک QTL معنی‌دار

۲- اثر افزایشی (a) مربوط به QTL

۳- اثر غالبیت (d) QTL

فرض صفر برای این سه آزمون به ترتیب با اعمال محدودیت‌های زیر فرموله می‌شود:

۱- $H_0:\ \mu_2 = \mu_1 = \mu_0$

۲- $H_0:\ \frac{1}{2}(\mu_2 - \mu_0) = 0$

۳- $H_0:\ \mu_1 = \frac{1}{2}(\mu_2 + \mu_0)$

مطابق با هر یک از این آزمون‌ها، آماره آزمون همانند آنچه که مشاهده شد، با استفاده از معادلات (۱۱-۸) و (۱۱-۹) محاسبه می‌شود.

مثال ۱۱-۵. مثال ۱۱-۴ را مجدداً مشاهده کنید که در آن یک QTL معنی‌دار روی کروموزوم شماره ۱ (شکل ۱۱-۳) برای وزن بدن ۱۰ هفتگی در جمعیت F_2 موش شناسایی شد. جمعیت F_2 اجازه انجام آزمون آثار افزایشی (a) و غالبیت (d) QTL را می‌دهد. MLE های مربوط به a و d به ترتیب ۱/۵۹ و ۱/۷۱ هستند. پارامترها تحت فرض صفر $a=0$ و $d=0$ با به حداکثر رساندن درستنمایی‌ها برآورد می‌شوند که به ترتیب به صورت $L\left(\tilde{d},\tilde{\sigma}^2\right)$ و $L\left(\tilde{a},\tilde{\sigma}^2\right)$ نشان داده می‌شوند.

نسبت‌های لگاریتم درستنمایی بین فرض‌های صفر و مقابل برای آزمون‌های معنی‌داری a و d به ترتیب به صورت ۱۲/۵۹ و ۱۰/۵۵ محاسبه می‌شوند. چون هیچ پارامتری در فرض صفر برای آزمون‌های جداگانه a و d قابل شناسایی نیست، تصور می‌شود که نسبت‌های درستنمایی هر یک از این دو آزمون می‌تواند به صورت مجانبی از یک توزیع χ^2 با یک درجه آزادی تبعیت کند؛ بنابراین، برای اندازه نمونه بزرگ، آستانه بحرانی می‌تواند از روی جدول توزیع χ^2 به‌دست آید؛ اما اگر اندازه نمونه استفاده شده به حد کافی بزرگ نباشد، آستانه می‌تواند به صورت تجربی به وسیله مطالعات شبیه‌سازی با داده‌هایی که تحت فرض صفر شبیه‌سازی شده‌اند، تعیین شود. در این مثال با اندازه نمونه مناسب، آستانه از روی جدول به‌دست می‌آید که پیشنهاد می‌کند هم اثر افزایشی و هم اثر غالبیت QTL شناسایی شده معنی‌دار هستند.

۱۱-۵- عوامل مؤثر بر شناسایی QTL

توان آماری شناسایی QTL های در حال تفرق به عوامل زیادی بستگی دارد. این عوامل شامل اندازه نمونه‌های اندازه‌گیری شده برای نشانگرهای مولکولی و صفات کمّی، حد مجاز خطای نوع اول، سهم QTL در حال تفرق در واریانس‌های فنوتیپی مشاهده شده، فواصل نوترکیبی بین QTL و نشانگرهای ژنتیکی، طرح آزمایشی خاص مورد استفاده و روش تجزیه آماری هستند. آثار این

عوامل روی توان شناسایی QTL و برآورد پارامترها می‌تواند با تجزیه یا از راه شبیه‌سازی رایانه‌ای مورد بررسی قرار گیرند. نتایج حاصل از مطالعات شبیه‌سازی به‌طور گسترده‌ای با نتایجی که از تجزیه نمونه‌های بزرگ به‌دست آمده‌اند، مطابقت دارند. در بیشتر موارد مکان‌یابی QTL، روش‌های مبتنی بر شبیه‌سازی مونت کارلو به‌طور گسترده‌ای استفاده می‌شوند، چون استفاده از روش‌های تجزیه‌ای اغلب مشکل است.

تصورات غلط زیادی در ارتباط با توان تشخیص QTL و طرح آزمایشی بهینه وجود دارند. در بیشتر موارد، توانایی تشخیص یک QTL در حال تفرق که با احتمال بیشتری در یک جمعیت در حال تفرق است، به تعیین ژنوتیپ حداقل ۵۰۰ فرد و اغلب بیشتر از این نیاز دارد. اندازه تعداد زیادی از آزمایشات بسیار کوچک‌تر از آن است که بتواند بزرگی آثار مربوطه را که به‌طور منطقی انتظار می‌رود، پیدا کند. حتی در مواردی که هزینه‌های تعیین فنوتیپ نسبت به هزینه‌های تعیین ژنوتیپ خیلی زیاد است، طرح‌های آزمایشی که در آن‌ها فواصل بین نشانگرها خیلی زیاد است، مناسب هستند و کاهش فواصل نشانگری تا حد زیر ۲۰ سانتی‌مورگان حقیقتاً هیچ اثری برای بیشتر طرح‌های آزمایشی ندارند. توان آزمون به ازای یک ژنوتیپ منفرد می‌تواند تا حد زیادی با تکرار نتاج، تعیین ژنوتیپ انتخابی، ادغام نمونه‌ها و نمونه‌گیری پی در پی یا تکراری افزایش یابد و اثر این تکنیک‌ها تجمعی است. به استثنای تکرار نتاج، تکنیک‌های دیگر مختص صفات بوده و بنابراین، برای آزمایش‌هایی که فقط تعداد اندکی صفت را در نظر می‌گیرند، مناسب‌ترین تکنیک‌ها هستند.

۱۱-۶- روش‌های مکان‌یابی QTL

مشکل مکان‌یابی QTL ها برای صفات فنوتیپی مشاهده شده می‌تواند به دو مرحله تقسیم شود: ۱- شناسایی ابتدایی حضور QTL های در حال تفرق و شمارش آن‌ها در جمعیت مکان‌یابی و ۲ - تعیین موقعیت ژنومی یا مکان‌یابی دقیق QTL ها. مرحله اول بستگی به آزمون و تعیین مدل برای تعداد مؤلفه‌های موجود در مدل مختلط، بدون نیاز به اطلاعات نشانگری دارد، حال آن که مرحله دوم روی اطلاعات نشانگری تمرکز می‌کند تا موقعیت‌های QTL ها را بر اساس تفرق همزمان نشانگر-QTL مشخص نماید.

۱۱-۶-۱- تعداد QTL ها

حضور QTL های در حال تفرق در یک جمعیت مکان‌یابی پیش شرط مکان‌یابی QTL است. بسیاری از مطالعات مولکولی که به‌صورت ضعیف طراحی می‌شوند، در شناسایی QTL های معنی‌دار موفق نمی‌شوند زیرا از جمعیت مکان‌یابی اشتباه که در آن هیچ QTL وجود ندارد،

استفاده می‌کنند. همچنین، اطلاعات قبلی در خصوص QTL ها کمک می‌کند تا نتایج حاصل از مکان‌یابی مولکولی تأیید شود. الگوی تفرق QTL می‌تواند به خوبی با استفاده از تلاقی‌های لاین‌های اینبرد بر اساس توارث مندلی پیش‌بینی شود. در این فصل، روش حداکثر درستنمایی بر اساس مدل مختلط برای برآورد آثار ژنتیکی QTL ها استفاده شد. در اینجا این روش برای شناسایی هر تعداد از QTL ها در جمعیت مکان‌یابی بسط داده می‌شود.

یک جمعیت مکان‌یابی با اندازه n در نظر بگیرید که در آن، تعداد k عدد QTL در حال تفرق برای صفت فنوتیپی y وجود دارند که J ژنوتیپ را تشکیل می‌دهند. در هر QTL، به ترتیب دو ژنوتیپ QTL که برای تلاقی‌های برگشتی با ۱ و ۰ نشان داده می‌شوند و سه ژنوتیپ QTL که برای F_2 به صورت ۲، ۱ و ۰ نشان داده می‌شوند، وجود دارند؛ بنابراین، J برای تلاقی برگشتی 2^k یا برای F_2 به صورت 3^k است. با در نظر گرفتن تلاقی برگشتی، یک مدل مختلط که تفرق تمامی k عدد QTL فرضی را مشخص نماید، می‌تواند بدین صورت نوشته شود:

(۲۷-۱۱)

$$L(\Omega \mid y)=\prod_{i=1}^{n}\left[\left(\frac{1}{2}\right)^{k} f_{11....1}(y_i)+\left(\frac{1}{2}\right)^{k} f_{11....0}(y_i)+...+\left(\frac{1}{2}\right)^{k} f_{00....0}(y_i)\right]$$

که $\Omega=\left(\mu_{11...1},\mu_{11...0},...,\mu_{00...0},\sigma^2\right)$ و $\left(\frac{1}{2}\right)^k$ احتمال پیشین ژنوتیپ QTL در جمعیت مکان‌یابی است. MLE های Ω می‌توانند با عبارت زیر برآورد شوند:

$$\hat{\mu}_{11...1}=\frac{\sum_{i=1}^{n} P_{11...1|i}\, y_i}{\sum_{i=1}^{n} P_{11...1|i}}$$

$$\hat{\mu}_{11...0}=\frac{\sum_{i=1}^{n} P_{11...0|i}\, y_i}{\sum_{i=1}^{n} P_{11...0|i}} \qquad (۲۸-۱۱)$$

$$\hat{\mu}_{00...0}=\frac{\sum_{i=1}^{n} P_{00...0|i}\, y_i}{\sum_{i=1}^{n} P_{00...0|i}}$$

$$\hat{\sigma}^2=\frac{1}{n}\left[\left[(y_i-\mu_{11...1})^2+...+(y_i-\mu_{00...0})^2\right]\right]$$

که در آن:

$$P_{11...1|i} = \frac{\left(\frac{1}{2}\right)^k f_{11....1}(y_i)}{\left(\frac{1}{2}\right)^k f_{11....1}(y_i) + \left(\frac{1}{2}\right)^k f_{11....0}(y_i) + ... + \left(\frac{1}{2}\right)^k f_{00....0}(y_i)}$$

$$P_{11...0|i} = \frac{\left(\frac{1}{2}\right)^k f_{11....0}(y_i)}{\left(\frac{1}{2}\right)^k f_{11....1}(y_i) + \left(\frac{1}{2}\right)^k f_{11....0}(y_i) + ... + \left(\frac{1}{2}\right)^k f_{00....0}(y_i)}$$

.

.

.

$$P_{00...0|i} = \frac{\left(\frac{1}{2}\right)^k f_{00....0}(y_i)}{\left(\frac{1}{2}\right)^k f_{11....1}(y_i) + \left(\frac{1}{2}\right)^k f_{11....0}(y_i) + ... + \left(\frac{1}{2}\right)^k f_{00....0}(y_i)} \quad (۲۹-۱۱)$$

احتمالات پسین یک ژنوتیپ QTL خاص برای فرد i ام مشروط به ارزش فنوتیپی آن هستند. برآوردهای حداکثر درستنمایی نیز به صورت زیر محاسبه می‌شوند:

۱- مقادیر اولیه $\Omega^{(0)} = \left(\mu_{11...1}^{(0)}, \mu_{11...0}^{(0)}, ..., \mu_{00...0}^{(0)}, \sigma^{2(0)}\right)$ را برای پارامترهای ناشناخته مشخص کنید.

۲- احتمالات پسین $P_{11...1|i}^{(0)}, P_{11...0|i}^{(0)}, ..., P_{00...0|i}^{(0)}$ را با استفاده از معادله (۱۱-۲۹) محاسبه کنید.

۳- برآورد جدیدی از پارامترهای ناشناخته $\Omega^{(1)} = \left(\mu_{11...1}^{(0)}, \mu_{11...0}^{(0)}, ..., \mu_{00...0}^{(0)}, \sigma^{2(1)}\right)$ را با استفاده از معادله (۱۱-۲۸) به‌دست آورید.

۴- مراحل ۲ و ۳ را تا زمانی که برآورد Ω در مقدار ثابتی همگرا شود، تکرار کنید.

مقادیر ثابت به عنوان MLE های پارامترهای ناشناخته در نظر گرفته می‌شوند. این MLE ها می‌توانند برای محاسبه مقادیر تابع درستنمایی (۱۱-۲۷) جایگزین شوند.

تجزیه‌های مدل مختلط (۱۱-۲۷) بالا برای برآورد تعداد QTL های مؤثر بر صفت انجام می‌شود. در آمار، برآورد تعداد QTL معادل تعیین تعداد مؤلفه‌های مختلط است که می‌توانند به بهترین شکل داده‌های فنوتیپی را توضیح دهند. k_1 و k_2 دو عدد جایگزین QTL هستند که به ترتیب تعداد ژنوتیپ‌های QTL را به صورت 2^{k_1} و 2^{k_2} برای تلاقی برگشتی نشان می‌دهند. لو و همکاران (۲۰۰۱) یک مدل آماری را برای تشخیص تعداد مؤلفه‌های مختلط بر اساس آماره نسبت درستنمایی محاسبه شده از معیار اطلاعاتی کولباخ-لیبلر[1] پیشنهاد دادند. تحت قضیه‌ای که به وسیله وونگ (۱۹۸۹) پیشنهاد شد، نسبت درستنمایی بین این فرض صفر که جمعیت مکان‌یابی حاوی k_1 عدد QTL بوده و فرض مقابل که جمعیت مکان‌یابی حاوی k_2 عدد QTL باشد، به صورت مجانبی دارای توزیعی به صورت مجموع وزنی متغیرهای مستقل تصادفی χ^2 با یک درجه آزادی تحت قاعده کلی است. با محاسبه نسبت درستنمایی تحت دو فرض جایگزین، مدل لو و همکاران (۲۰۰۱) می‌تواند برای تعیین تعداد QTL هایی که در جمعیت مکان‌یابی در حال تفرق هستند، استفاده شود.

به‌طور جایگزین، برآورد تعداد QTL های در حال تفرق برای یک صفت کمّی در جمعیت می‌تواند بر اساس معیار انتخاب مدل، AIC (آکائیک، ۱۹۹۷) یا BIC (شوآرز، ۱۹۷۸) باشد. AIC یا BIC، بسته به درستنمایی با توجه به تعداد پارامترهایی که برآورد می‌شوند و اندازه نمونه مورد استفاده، محاسبه می‌شوند. مدلی که کوچکترین مقدار AIC یا BIC را ارایه دهد، در نظر گرفته می‌شود تا به بهترین شکل، داده‌های مکان‌یابی تجزیه شده را توضیح دهد.

رویه مشابهی می‌تواند برای جمعیت F_2 فرموله شود که برای آن، تابع درستنمایی با فرض k عدد QTL به صورت زیر بیان می‌شود:

$$L(y)=\prod_{i=1}^{n}\left[\left(\frac{1}{3}\right)^{k} f_{22....2}(y_i)+\left(\frac{1}{3}\right)^{k} f_{22....0}(y_i)+...+\left(\frac{1}{3}\right)^{k} f_{00....0}(y_i)\right]$$

که احتمال پیشین ژنوتیپ QTL به صورت 3^k است. محاسبه نسبت درستنمایی در دو حالت مختلف که هر یک برای تعداد مختلفی از QTL ها هستند، یعنی k_1 و k_2، برای تعیین تعداد بهینه QTL های واقعی موجود در جمعیت مکان‌یابی F_2 استفاده می‌شود.

مثال ۱-۶. مثال ۳-۱ را مجدداً مشاهده کنید. دو لاین اینبرد، IR64 نیمه پاکوتاه و Azucena پا بلند، برای تولید هتروزیگوت F_1 تلاقی داده شدند. کروموزوم‌های هاپلوئید برای دانه‌های گرده F_1 دو برابر شدند تا ۱۲۳ گیاه دابل هاپلوئید (DH) تولید کنند. این گیاهان DH،

[1] - Kullback-Leibler

معادل با نتاج تلاقی برگشتی، برای ۱۳۵ نشانگر RFLP و ۴۰ نشانگر ایزوزیم و RAPD تعیین ژنوتیپ شدند که از روی آن‌ها یک نقشه پیوستگی پوشش‌دهنده کل ژنوم ۱۲ کروموزومی برنج تهیه شد (یان و همکاران، ۱۹۹۸؛ شکل ۳-۳). هر یک از این لاین‌های DH برای ارتفاع گیاه پس از آن که گیاهان به مدت ۱۰ هفته در مزرعه رشد کردند، اندازه‌گیری شدند.

تعداد QTL های موجود در جمعیت DH برای ارتفاع گیاه در سن ۱۰ هفتگی برآورد می‌شود. با فرض این که $k = 0,1,2,...,8$ QTL برای صفت وجود دارد، نسبت‌های لگاریتم درستنمایی (LR) تحت دو فرض صفر و مقابل، یعنی تعداد QTL ها برابر با k و k-۱ محاسبه می‌شوند. بر اساس معیارهای اطلاعات AIC و BIC، پیشنهاد می‌کنیم که احتمالاً یک QTL بزرگ اثر وجود دارد که ارتفاع گیاهان را در سن ۱۰ هفتگی در برنج کنترل می‌کند (جدول ۱۱-۱۰). مکان‌یابی ژنتیکی بیشتر به همراه نقشه‌های پیوستگی تهیه شده با نشانگرهای مولکولی (شکل ۳-۳) می‌توانند برای تعیین موقعیت این QTL استفاده شوند.

جدول ۱۱-۱۰- معیارهای اطلاعات AIC و BIC با فرض تعداد مختلف QTL ها در یک جمعیت DH برنج.

تعداد QTL	مقدار درستنمایی	LR	AIC	BIC
۰	-۴۸۳/۰۹۹۹		۹۶۸/۱۹۹۸۹	۹۷۰/۸۶۳۳۳
		۲/۸۵۴۱		
۱	-۴۸۰/۲۴۵۸		۹۶۴/۴۹۱۶۸	۹۶۹/۸۱۸۵۶
		۰/۲۲۹۶		
۲	-۴۸۰/۰۱۶۳		۹۶۸/۰۳۲۵	۹۷۸/۶۸۶۲۶
		۴/۳۸۱۰		
۳	-۴۷۵/۶۳۵۲		۹۶۷/۲۷۰۴۴	۹۸۸/۵۷۷۹۶
		۴/۸۰۸۵		
۴	-۴۷۰/۸۲۶۷		۹۷۳/۶۵۳۳۶	۱۰۱۶/۲۶۸۴
		۳/۶۲۴۸		
۵	-۴۶۷/۲۰۱۹		۹۹۸/۴۰۳۸	۱۰۸۳/۶۳۳۸
		۳۴/۲۵۴۰		
۶	-۴۳۲/۹۴۷۹		۹۹۳/۸۹۵۸	۱۱۶۴/۳۵۵۹
		۰/۷۳۴۳		
۷	-۴۳۲/۲۱۳۶		۱۱۲۰/۴۲۷۲	۱۴۶۱/۳۴۷۵
		۹/۰۶۲۹		
۸	-۴۲۳/۱۵۰۷		۱۳۵۸/۳۰۱۴	۲۰۴۰/۱۴۱۸

۱۱-۶-۲- موقعیت QTL های انفرادی

با آگاهی از تعداد QTL های در حال تفرق در یک جمعیت مکان‌یابی، نقشه پیوستگی ایجاد شده از روی نشانگرهای مولکولی برای تعیین مکان این QTL ها استفاده می‌شود. به‌طورکلی، روش جستجوی شبکه‌ای از موقعیت‌های ژنومی که فضای بین یک نشانگر تا نشانگر بعدی را به‌طور یکنواخت مدنظر قرار می‌دهد، برای جستجوی حضور تمامی QTL های ممکنه در سرتاسر ژنوم استفاده می‌شود. مکان‌یابی فاصله‌ای که یک QTL را در یک زمان در نظر می‌گیرد، استفاده از تابع درستنمایی مبتنی بر مدل مختلط را به صورت زیر امکان‌پذیر می‌سازد:

$$L(\Omega \mid y) = \prod_{i=1}^{n} \left[\omega_{1|i} f_1(y_i) + \omega_{0|i} f_0(y_i) \right] \qquad (۳۰-۱۱)$$

برای تلاقی برگشتی بوده و برای جمعیت F_2 نیز داریم:

$$L(\Omega \mid y) = \prod_{i=1}^{n} \left[\omega_{2|i} f_2(y_i) + \omega_{1|i} f_1(y_i) + \omega_{0|i} f_0(y_i) \right] \qquad (۳۱-۱۱)$$

که $\omega_{j|i}$ ها احتمالات شرطی ژنوتیپ‌های QTL مشروط به ژنوتیپ نشانگر برای فرد i ام هستند؛ بنابراین، با برآورد نسبت لگاریتم درستنمایی حداکثری در امتداد ژنوم، می‌توانیم موقعیت QTL را تعیین کنیم. تعداد QTL های شناسایی شده از روی نقشه پیوستگی ژنتیکی از تعداد QTL های شناسایی شده با تجزیه خالص فنوتیپی (بدون نشانگر) کوچک‌تر است، البته در شرایطی که نقشه پیوستگی به خوبی تمامی ژنوم را پوشش ندهد.

از لحاظ نظری، مدل‌های QTL چندگانه می‌توانند در جهت تشخیص آثار QTL نسبت به مدل تک QTL کارآتر باشند، چون اثر متقابل QTL-QTL برای QTL چندگانه در نظر گرفته می‌شوند. ولی این وضعیت ممکن است در عمل همیشه درست نباشد. تجزیه و مدل‌ریزی همزمان چند QTL احتمالاً منجر به سنگین شدن محاسبات می‌شود.

۱۱-۷- تمرین‌ها

۱۱-۱- مزایا و معایب تجزیه تابعیت و راهکارهای حداکثر درستنمایی را در مکان‌یابی QTL مقایسه کنید.

۱۱-۲- از قضیه بیز برای به‌دست آوردن جدول ۱۰-۳ استفاده کنید.

۱۱-۳- الف- تأیید کنید که معادله (۱۱-۳) درستنمایی کامل است.

ب- تأیید کنید که معادله (۱۱-۵) درستنمایی تحت H_0 بوده و مستقل از θ است.

۱۱-۴- برای معادلات درستنمایی بخش ۱۱-۴-۱، نشان دهید که:

$$E\left(\omega_{j|i} \mid y,\Omega\right) = P_{j|i}$$

$$E\left(\omega_{j|i}^2 \mid y,\Omega\right) = P_{j|i}$$

$$E\left(\omega_{j1|i}\omega_{j2|i} \mid y,\Omega\right) = 0$$

که امید ریاضی توزیع $\omega_{j|i}$ مشروط به داده‌های y و پارامترهای دیگر است. این حالت مرحله E الگوریتم EM است که در آن، ژنوتیپ‌های مورد انتظار شرطی QTL با استفاده از معادله (۱۱-۱۹) تعیین می‌شوند و در مرحله M با استفاده از معادلات (۱۱-۲۰) و (۱۱-۲۱) هستند، پارامترهای ناشناخته با استفاده از مقدار پسین محاسبه شده از مرحله E برآورد می‌شوند. این فرایند تکراری بین مراحل E و M تا زمانی تکرار می‌شود که برآوردها همگرا شوند.

۱۱-۵- مکان‌یابی فاصله‌ای

زمانی که مدل مکان‌یابی فاصله‌ای تهیه شد، سه حالت مختلف برای تشکیل گامت‌ها در طی میوز در نظر گرفته شدند: ۱- هیچ نوترکیبی مضاعفی وجود نداشت، یعنی $r = r_1 + r_2$، ۲- نوترکیبی‌ها مستقل بودند، یعنی $r = r_1 + r_2 - 2r_1r_2$ و ۳- تداخل بین فواصل مختلف وجود داشت، یعنی $r = r_1 + r_2 - 2\left(1-I\right)r_1r_2$ (بخش ۱۱-۳-۲ را برای تعاریف r_1، r_2، r و I مشاهده کنید). حالت ۳ کلی‌ترین حالت بوده و می‌تواند زمانی که I =۱ باشد، به حالت ۱ و زمانی که I = ۰ باشد، به حالت ۲ کاهش می‌یابد. تداخل I، دارای محدودیت‌هایی است (ات، ۱۹۹۱):

$$1-\min\left(\frac{1}{r_1},\frac{1}{r_2}\right) \le I \le 1-\max\left[0,\frac{r_1+r_2-\frac{1}{2}}{2r_1r_2}\right]$$

الف. مطالعات شبیه‌سازی را در ارتباط با توان و دقت برآورد پارامترها، برای مقایسه این حالات با در نظر گرفتن مقادیر مختلف I انجام دهید.

ب. برای افرادی که در زمینه مکان‌یابی QTL کار می‌کنند، یک دستورالعمل راهنما تهیه کنید تا مدل تجزیه‌ای مناسبی برای داده‌های خود انتخاب کنند.

۱۱-۶- مکان‌یابی اپیستازی

اپیستازی به‌عنوان وابستگی بیان یک ژن به بیان ژن‌های دیگر تعریف می‌شود. نشان دهید چگونه توابع درستنمایی (۱۱-۱) و (۱۱-۱۵) می‌توانند برای مکان‌یابی دو QTL اپیستاتیک در جمعیت مکان‌یابی بسط یابند.

فصل دوازدهم

تجزیهٔ دقت و آستانهٔ بحرانی

۱۲-۱- مقدمه

در فصل قبل، اصول اساسی مکان‌یابی فاصله‌ای QTL ها را بـه روش حـداکثر درسـتنمایی تشـریح کردیم، ولی دو پرسش اساسی باید در ارتباط با تجزیه QTL صفات پیچیده مورد توجه قرار گیـرد. اولاً، آستانه بحرانی آماره آزمون که می‌تواند برای تشخیص معنی‌داری آماری QTL اسـتفاده شـود، چیست؟، ثانیاً پس از آن که QTL معنی‌دار تعیین شد، آیـا برآوردهـای درسـتنمایی و پارامترهـای QTL به حد کافی دقیق هستند تا استنباط علمی در خصوص موقعیت و اثر QTL صـورت گیـرد؟ از نظر آماری، این دو پرسش جنبه‌های متفاوتی از مکان‌یابی QTL را نشان می‌دهند، ولی ما آن‌هـا را در اینجا در یک فصل مجزا به‌عنوان دو موضوع مهم مرتبط به هم بـرای مکـان‌یـابی QTL ارائـه می‌دهیم.

تمامی روش‌های آماری مکان‌یابی QTL بر تعیین آستانه‌های معنـی‌داری (یـا مقـادیر بحرانـی) مناسب تأکید دارند که با استفاده از آن‌ها، آماره آزمون برآورد شده از یک مجموعه داده‌های خـاص مورد مقایسه قرار می‌گیرد تا مشخص شود که آیا QTL معنی‌دار وجود دارد یا خیر؟ اغلـب تعیـین آستانه‌های بحرانی مشکل است، چون این امر تا حد زیادی بر توزیع آماره آزمون (امتیـاز LOD یـا نسبت‌های درستنمایی) با این فرض صفر که QTL وجـود نـدارد، تکیـه دارد. خصوصـیات تـوزیعی آماره آزمون به وسیله عوامل زیادی نظیر اندازه نمونه‌ها، توزیع صفات کمّی و غیره تأثیر مـی‌پـذیرد. به‌دلیل اهمیت این مسئله، منابع آماری زیادی در خصوص تعیین آسـتانه‌هـای بحرانـی وجـود دارد (لندر و بوتستین، ۱۹۸۹؛ ربای و همکاران، ۱۹۹۴؛ دورج و ربـای، ۱۹۹۶؛ پیفـو، ۲۰۰۱؛ چرچیـل و دورج، ۱۹۹۴؛ دوپوئیس و سیگموند، ۱۹۹۹؛ زو و همکاران، ۲۰۰۴). در اولین بخـش از ایـن فصـل، چند روش برای تعیین آستانه‌های بحرانی و مزایا و معایب آن‌ها معرفی خواهد شد.

برای یک QTL معنی‌دار شناسایی شده که آماره آزمون آن بیشـتر از آسـتانه بحرانـی از پـیش تعیین شده است، ارزیابی صحت و دقت برآورد آثار ژنتیکی QTL حائز اهمیت است. ارزیـابی دقـت یک روش آماری برای مکان‌یابی QTL مـی‌توانـد از راه مطالعـات شـبیه‌سـازی انجـام شـود. بـرای مجموعه‌ای از پارامترها که هر یک دارای مقدار معینی هستند، داده‌های مصنوعی بـر پایـه سـاختار مدل شبیه‌سازی می‌شوند و در هر یک از این تکرارها، پارامترها برآورد می‌شوند. سـپس میـانگین و میانگین مربعات خطا با استفاده از جفت‌هـای حاصـل از تمـامی تکرارهـای شـبیه‌سـازی محاسـبه مـی‌شـوند؛ امـا بـرای داده‌هـای واقعـی، دقـت بـرآورد پارامترهـا مـی‌توانـد بـا محاسـبه واریـانس-کوواریـانس‌هـای مجـانبی بـرآورد پـارامترهـا آزمـون شـود. نظریـه‌هـایی بـرای بـرآورد واریـانس-کوواریانس‌های مجانبی برآوردهای مربوط به یک مدل مختلط ایجاد شده‌اند (لوئیس، ۱۹۸۲؛ منـگ و روبین، ۱۹۹۱). کائو و زنگ (۱۹۹۷) و چن (۲۰۰۵) این نظریه‌هـا را بـرای ارزیـابی دقـت بـرآورد

پارامترهای QTL بسط دادند. در بخش دوم این فصل، مسائل کلیدی تکنیکی برای برآورد واریانس-کوواریانس‌های مجانبی پارامترهای مربوط به موقعیت و اثر QTL توضیح داده خواهند شد. خوانندگان علاقه‌مند به مقالات اصلی ارجاع شده در این بخش مراجعه نمایند.

۱۲-۲- تعیین آستانهٔ بحرانی

۱۲-۲-۱- پس‌زمینه

در یک آزمایش علمی، همیشه لازم است تا احتمال دستیابی به نتایج غلط را بدانیم. برای یک مطالعه مکان‌یابی QTL، این نتایج عبارتند از: ۱- یک QTL در حال تفرق وجود دارد، در حالی که در عالم واقعیت وجود ندارد و ۲- QTL شناسایی نمی‌شود، در حالی که در عالم واقعیت وجود دارد. خطای نوع اول منجر به ایجاد حالت مثبت کاذب (نوع یک) می‌شود، حال آن که خطای نوع دوم منجر به ایجاد حالت منفی کاذب (نوع دو) می‌شود. احتمال مثبت کاذب (یعنی سطح معنی‌دار) می‌تواند با انتخاب آستانه معنی‌دار مناسب کنترل شود. نرخ منفی کاذب با توجه به وضعیت آزمایش و بزرگی آثار QTL تعیین می‌شود (یانسن، ۱۹۹۴؛ ون ایجن، ۱۹۹۹).

باوجود اهمیت تعیین آستانه‌های بحرانی برای تعیین QTL های معنی‌دار، تشخیص آستانه‌های بحرانی یکی از مشکل‌ترین مسائل در زمینه مکان‌یابی QTL است. دو مسئله وجود دارد که تعیین آستانه بحرانی را مشکل می‌سازد (لندر و شورک، ۱۹۹۴؛ چرچیل و دورج، ۱۹۹۴). اولاً، توزیع آماره آزمون (LOD یا LR) تحت فرض صفر نمی‌تواند به خوبی تعیین شود، چون شرایط وجود توزیع χ^2 مجانبی برای آماره آزمون LR نقض می‌شود. این حالت از این حقیقت ناشی می‌شود که با فرض صفر عدم وجود QTL، موقعیت QTL قابل شناسایی نیست و به‌عنوان یک پارامتر مزاحم محسوب می‌شود. همچنین، قابلیت اعتماد تقریب‌های مجانبی می‌تواند با عوامل دیگری نظیر اندازه محدود نمونه و توزیع صفت مورد مطالعه تحت‌تأثیر قرار گیرد. ثانیاً، زمانی که آزمون در کل ژنوم صورت می‌گیرد، آنچه که به‌طور معمول انجام می‌شود، ایجاد مشکل چند آزمونی است، چون آزمون‌هایی که در امتداد طول گروه پیوستگی انجام می‌شوند، به‌دلیل ماهیت پیوستگی به‌صورت متقابل مستقل نیستند. توصیف ساختار ارتباط بین چنین آزمون‌هایی مشکل است؛ بنابراین، برای کنترل نرخ خطای نوع اول ژنومی، مقادیر آستانه بحرانی آماره آزمون باید تصحیح شوند.

در حال حاضر، سه روش مختلف برای محاسبه مقدار آستانه بحرانی در سرتاسر ژنوم وجود دارد: ۱- روش‌های تجزیه‌ای، ۲- مطالعات شبیه‌سازی و ۳- آزمون‌های جایگشتی. پیفو (۲۰۰۱) یک روش تقریبی سریع را برای محاسبه آستانه بحرانی جهت ارزیابی معنی‌داری کل ژنومی پیشنهاد کرد. اخیراً، زو و همکاران (۲۰۰۴) یک آمار امتیازی را با هدف افزایش کارآیی محاسباتی

تعیین آستانه بحرانی به‌دست آوردند. در این بخش، هر یک از این روش‌ها را معرفی خواهیم کرد. مقایسه بین برخی از این روش‌ها نیز انجام خواهد شد.

۱۲-۲-۲- تقریب‌های تجزیه‌ای

روش تجزیه‌ای بستگی به توزیع آماره آزمون مربوطه دارد. به‌طور معمول، پروفیل آماره آزمون LR برای بخش‌هایی از موقعیت‌های ممکن QTL در یک گروه پیوستگی یا کل ژنوم تهیه شده و حداکثر LR (MLR) به عنوان یک آماره آزمون عمومی استفاده می‌شود. در یک موقعیت فرضی QTL، آماره آزمون LR با در نظر گرفتن فرض صفر دارای توزیع مجانبی χ^2 با درجه آزادی برابر با تعداد آثار QTL مربوطه است؛ اما با فرض صفر عدم وجود QTL، همان‌گونه که در بخش بالا ذکر شد، موقعیت QTL شناسایی نمی‌شود و بنابراین، آماره آزمون LR به صورت مجانبی از توزیع χ^2 تبعیت نمی‌کند. بر اساس نتایج دیویس (۱۹۷۷ و ۱۹۸۷)، چندین محقق فرمول‌های تقریبی برای تعیین آستانه‌های بحرانی در طرح‌های خاصی که در آن‌ها آستانه‌های دقیق در دسترس نیستند، به‌دست آوردند (ربای و همکاران، ۱۹۹۴؛ دورج و ربای، ۱۹۹۶؛ پیفو، ۲۰۰۱).

زمانی که یک QTL در موقعیت خاصی آزمون می‌شود، سطح معنی‌دار عبارت از احتمالی است که فرض صفر صحیح رد خواهد شد. اگر LR(x) آماره آزمون نسبت درستنمایی در موقعیت x باشد، سطح معنی‌دار جزئی این آزمون به‌صورت زیر خواهد بود:

$$\text{Prob}(\text{LR}(x) > T_\alpha^P \mid \text{عدم وجود QTL در موقعیت } x) = \alpha \qquad (۱-۱۲)$$

که T_α^P آستانه بحرانی برای آزمون معنی‌داری QTL در سطح α در نقطه خاص است. احتمالی که با معادله (۱۲-۱) تعیین می‌شود، نرخ خطای مقایسه‌ای نامیده می‌شود. نرخ خطای مقایسه‌ای می‌تواند با تعیین توزیع آماره آزمون در یک نقطه خاص ژنوم کنترل شود. این آزمون نقطه‌ای خاص در حالت عملی تحت شرایطی امکان‌پذیر است که علاقه‌مندیم آزمون کنیم که آیا یک QTL خاص شناسایی شده در یک جمعیت، در جمعیت‌های مختلف نیز وجود دارد یا خیر؟ زمانی که آزمون در یک نقطه منفرد ژنوم انجام می‌شود، آماره آزمون عموماً از یک توزیع χ^2 با یک درجه آزادی تبعیت می‌کند.

در مکان‌یابی QTL، نیازمند انجام آزمون‌های زیادی در هر فاصله نشانگری در طول ژنوم هستیم. اگر QTL در یک فاصله آزمون شده وجود نداشته باشد، نیاز داریم تا احتمال شناسایی اشتباه هر QTL را در آن فاصله کنترل کنیم. احتمال یافتن حداقل یک QTL در یک فاصله وقتی که در حقیقت QTL وجود ندارد، نرخ خطای آزمایشی نامیده می‌شود. برای طرح تلاقی برگشتی، MLR برای یک فاصله نشانگری کامل دارای توزیعی بین x_1^2 و x_2^2 است:

$$x_{1,\alpha}^2 < T_\alpha^I < x_{2,\alpha}^2 \qquad (۲-۱۲)$$

که T_α^I آستانه بحرانی آزمون معنی‌داری QTL با سطح α در فاصله مربوطه است. توزیع χ^2 با دو درجه آزادی از این حقیقت ناشی می‌شود که دو پارامتر در تلاقی برگشتی، یعنی موقعیت QTL و اثر افزایشی QTL تحت فرض صفر تداخل دارند. با استدلال می‌توان نشان داد که توزیع MLR برای فاصله کوچک‌تر نسبت به فاصله بزرگ‌تر به x_1^2 نزدیک‌تر است.

در یک مطالعه ژنومی آزمایشی، به حضور QTL در کل ژنوم علاقه‌مندتر هستیم؛ بنابراین، تعیین توزیع و آستانه مناسب برای MLR در نقطه‌ای از طول ژنوم کامل که دارای مقدار حداکثر است، حائز اهمیت است. تحت فرض صفر (یعنی هیچ QTL در کل ژنوم با طول کلی L وجود ندارد)، شانس (α) MLR بیشتر از T_α هر نقطه‌ای از ژنوم عبارت است از:

Prob (MLR $0 \le x \le L > T_\alpha^P \mid \alpha =$ عدم وجود QTL در ژنوم)

چند روش تجزیه‌ای تقریبی برای محاسبه آستانه T_α برای هر سطح معنی‌دار α پیشنهاد شده است (لندر و بوتستین، ۱۹۸۹؛ دوپوئیس و سیگموند، ۱۹۹۹؛ ربای و همکاران، ۱۹۹۴). این روش‌ها اساساً برای دنباله‌های نقشه‌های ژنتیکی بسیار متراکم و بسیار پراکنده مناسب هستند. اینکه چگونه T_α را به‌صورت جداگانه برای این دو حالت مقابل هم تعیین کنیم، اغلب بر اساس بحث لندر و بوتستین (۱۹۸۹) بحث خواهیم کرد. تقریب‌های بیشتری برای نقشه‌های متراکم (کوچک‌تر از ۱ سانتی‌مورگان) و نقشه‌های پراکنده به وسیله دوپوئیس و سیگموند (۱۹۹۹) ارائه شده‌اند.

آستانه بحرانی کل ژنوم برای یک نقشه پراکنده

در مکان‌یابی QTL، تعیین سطح معنی‌دار کل آزمایش معمولاً مطلوب است، مگر آن که فرد بخواهد مشخص کند که QTL های شناسایی شده قبلی در آزمایش جدید وجود دارند. به همین دلیل، معنی‌داری کل ژنوم- احتمال به‌دست آوردن آماره آزمونی بیشتر از آستانه بحرانی در نقطه‌ای روی کل ژنوم به وسیله شانس- باید مورد استفاده قرار گیرد. آستانه بحرانی کل ژنوم نه فقط به تعداد و طول کروموزوم‌ها بستگی دارد، بلکه به تعداد نشانگرها (یعنی چگالی) روی کروموزوم‌ها نیز بستگی دارد. در یک سطح معنی‌دار یکسان ژنومی، زمانی که فقط تعداد اندکی نشانگر به ازای هر کروموزوم آزمون می‌شود (یعنی حالت نقشه پراکنده)، آستانه پایین‌تری نسبت به زمانی که تعداد زیادی نشانگر به ازای هر کروموزوم (یعنی نقشه متراکم) آزمون می‌شود، مورد نیاز است (لندر و بوتستین، ۱۹۸۹؛ ون اویجن، ۱۹۹۹).

برای یک نقشه پراکنده که در آن، نشانگرها با فاصله زیاد روی ژنوم پراکنده شده‌اند، می‌توانیم با اطمینان فرض کنیم که احتمالات عدم وجود QTL در درون فواصل نشانگری مختلف تقریباً مستقل هستند. α و p را به ترتیب برای سطوح معنی‌دار کل ژنوم و کل فاصله برای تعیین QTL معنی‌دار در نظر بگیرید. اگر m فاصله نشانگری در نظر گرفته شود، خواهیم داشت:

$$1-\alpha = \Pr ob$$ (عدم وجود QTL در ژنوم)

$$= \prod_{i=1}^{m} \Pr ob$$ (عدم وجود QTL در فاصله i)

$$= \prod_{i=1}^{m} (1-p)$$

$$= (1-p)^m$$

$$\approx 1-mp$$

یا:

$$p \approx \frac{\alpha}{m} \qquad (۳-۱۲)$$

سطح معنی‌دار کل فاصله‌ای مورد نیاز برای اعلام وجود ژنوم گسترده یک QTL در ژنومی با m فاصله نشانگری با استفاده از معادله (۱۲-۳) محافظه‌کارانه است. این حالت می‌تواند با استفاده از نامعادله بونفرونی ساده که بدین شکل بیان می‌شود، اثبات شود:

(۱۲-۴)

$$\Pr ob\left(MLR_{0\le x\le L} > T_\alpha\right) \le \prod_{i=1}^{m} \Pr ob\left(MLR_{0\le x\le l} > T_\alpha^{I}\right)$$

البته با این فرض که هر فاصله نشانگری طولی برابر با l دارد. این حالت معادل $\alpha \le \prod_{i=1}^{m} p = mp$ یا رابطه زیر است:

$$p \ge \frac{\alpha}{m}$$

بر اساس معادله (۱۲-۳)، آستانه تقریبی برای آزمون معنی‌داری یک QTL در کل ژنوم برای تلاقی برگشتی به‌صورت زیر است:

$$x^2_{1,\frac{\alpha}{m}} < T_\alpha < x^2_{2,\frac{\alpha}{m}} \qquad (۱۲-۵)$$

مثال ۱۲-۱. هوآنگ و همکاران (۱۹۹۷) یک جمعیت دابل هاپلوئید (DH) ۱۲۳ لاینی را با دو لاین اینبرد، IR64 نیمه پاکوتاه و Azucena پا بلند تهیه کردند. این جمعیت DH با دو برابر کردن کروموزوم‌های هاپلوئید گامت‌های حاصل از هتروزیگوت F_1 تهیه شد و بنابراین معادل جمعیت تلاقی برگشتی بود، چون تفرق نشانگری آن از نسبت ۱:۱ تبعیت می‌کند. نقشه پیوستگی (شکل ۳-۳) با مجموع ۱۷۵ نشانگر چند شکل (شامل ۱۴۶ RFLP، ۸ ایزوزیم، ۱۴ RAPD و ۱۲ ژن کلون شده) که نشان‌دهنده پوشش خوبی از ۱۲ کروموزوم برنج بود، تهیه شد. نقشه تهیه شده، دارای طولی برابر با ۲۰۰۵ سانتی‌مورگان با ۱۶۳ فاصله نشانگری است که متوسط فاصله ۱۱/۵ سانتی‌مورگان داشته و دارای ۶ شکاف (فاصله) بیش از ۳۵ سانتی‌مورگان است.

سه صفت دانه (طول دانه، عرض دانه و نسبت طول به عرض) برای هر یک از لاین‌های DH اندازه‌گیری شد. مکان‌یابی فاصله‌ای برای شناسایی QTL هایی که این صفات را کنترل می‌کنند، مورد استفاده قرار گرفت. هوآنگ و همکاران (۱۹۹۷) ۱۲ مورد از چنین QTL هایی که روی ۵ کروموزوم مختلف قرار داشتند را گزارش نمودند. جدول ۱۲-۱، QTL هایی را برای صفات دانه، موقعیت آن‌ها، مقادیر LR، سطوح معنی‌دار کل فاصله (p) بر اساس $x^2_{1,p}$ و $x^2_{2,p}$ (معادله (۱۲-۲) را مشاهده کنید) و سطوح معنی‌دار تصحیح شده کل ژنوم (α) نشان می‌دهد.

آستانه بحرانی کل ژنوم برای نقشه متراکم

فرض کنید که نقشه بی‌نهایت متراکمی وجود دارد که در آن، نشانگرها در هر نقطه‌ای از ژنوم مستقر هستند. برای چنین نقشه متراکمی، وقوع آماره‌های آزمون کاذب بزرگ در فواصل مجاور هم مستقل نخواهند بود. برای یک نقشه بی‌نهایت متراکم و نامحدود (یعنی تعداد فواصل یا m به سمت صفر میل کند)، سطح معنی‌دار جزئی مورد نیاز برای آزمون هر فاصله تمایل دارد تا محدوده غیر صفر مستقل از m داشته باشد. اگر اندازه نمونه مورد استفاده بزرگ باشد، لندر و بوتستین (۱۹۸۹) بحث نمودند که تغییر امتیاز LOD از مربع فرآیند انتشار (OU[1]) در نقشه بی‌نهایت متراکم تبعیت می‌کند.

بر اساس قضیه حد مرکزی، آماره آزمون نسبت لگاریتم درستنمایی (LR) در هر موقعیت (x) به‌طور مجانبی متناسب با مربع متغیر نرمال تصادفی $z^2(x)$ است:

$$LR(x) \sim z^2(x)$$

[1] - Orenstein-Uhlenbeck

که $z(x) \sim N(0,1)$ است. برای دو موقعیت x_1 و x_2، $LR(x_1)$ و $LR(x_2)$ همبسته هستند، به‌طوری که داریم:

$$Corr\left[z(x_1), z(x_2)\right] \sim 1-2r = e^{-2d}$$

که r و d به ترتیب نرخ نوترکیبی و فاصله ژنتیکی (اندازه‌گیری شده بر حسب مورگان) بین دو موقعیت ژنومی هستند و ارتباط بین r و d با تابع نقشه هالدین تعیین می‌شود. می‌تواند مشاهده شود که $z(x)$ فرآیند ثابت نرمال با تابع کوواریانس e^{-2d} است که می‌تواند با فرآیند انتشار OU توصیف شود.

بعلاوه، لندر و بوتستین (۱۹۸۹) مقادیر بحرانی مناسب کل ژنوم را برای تلاقی برگشتی بر اساس نظریه مجانبی با فرآیند OU به‌دست آوردند. با در نظر گرفتن یک ژنوم با C کروموزوم و طول ژنتیکی کل برابر با L (اندازه‌گیری شده بر حسب مورگان)، احتمال این که امتیاز LOD بیشتر از سطح بالایی T در حالتی باشد که در آن هیچ QTL وجود ندارد، بدین صورت تقریب زده می‌شود:

$$(C+2Lt)\,x^2(t) \qquad (۶-۱۲)$$

که $T=(2\ln 10)^{-1}t$ و $x^2(t)$ تابع توزیع تجمعی معکوس توزیع χ^2 با یک درجه آزادی هستند.

برای کنترل احتمال مثبت کاذب در سطح معنی‌دار کل ژنوم (α)، آستانه مناسب LOD بدین صورت تقریب زده می‌شود:

$$T_\alpha = (2\ln 10)^{-1} t_\alpha \qquad (۷-۱۲)$$

که t_α معادله $\alpha = (C+2Lt_\alpha)\,x^2(t_\alpha)$ را حل می‌نماید.

لندر و بوتستین (۱۹۸۹) پیشنهاد کردند که آستانه امتیاز LOD معمولاً باید بین ۲ تا ۳ باشد تا ۵ درصد خطای مثبت کاذب کلی را برای تشخیص QTL تضمین نماید. نظریه مبتنی بر انتشار OU مربوط به لندر و بوتستین برای جمعیت F_2 (دوپویس و سیگموند، ۱۹۹۹) استفاده شده بود که در آن، امتیاز LOD از توزیع χ^2 با دو درجه آزادی تبعیت می‌کرد که این امر به‌دلیل برازش مؤلفه‌های افزایشی و غالبیت و طرح‌های تلاقی عمومی‌تر است (ژو و همکاران، ۲۰۰۴).

۱۲-۲-۳- مطالعات شبیه‌سازی

برای تعداد زیادی از داده‌های عملی، نمی‌توانیم دربارهٔ توزیع آماره‌های آزمون اطلاعی داشته باشیم. بهترین روش در این مورد، برآورد مستقیم نرخ مثبت کاذب به وسیله شبیه‌سازی است. بر اساس

وضعیت ژنتیکی خاص، می‌توان به‌طور تصادفی هم داده‌های مطابق با قوانین توارث مندلی و هم داده‌های فنوتیپی که از یک توزیع نرمال با در نظر گرفتن فرض صفر (عدم وجود QTL) تبعیت می‌کنند را شبیه‌سازی نمود. داده‌های شبیه‌سازی شده با روش آماری مورد نظر تجزیه شده و مقدار آماره آزمون برای هر تکرار از راه شبیه‌سازی محاسبه می‌شود. توزیع مقادیر LR در مقابل تعداد تکرارهای شبیه‌سازی (که ۱۰۰۰ تکرار است) می‌تواند با یک توزیع χ^2 تقریب زده شود. صدک ۹۵ و ۹۹ توزیع حداکثری به عنوان مقادیر بحرانی تجربی استفاده می‌شوند تا حضور QTL برای یک صفت کمّی در سطوح معنی‌دار ۰/۰۵ و ۰/۰۱ بیان شود.

جدول ۱۲-۱- موقعیت کروموزومی و مقادیر LR مربوط به QTL های صفات دانه اندازه‌گیری شده در جمعیت DH. اقتباس از هوآنگ و همکاران (۱۹۹۷).

صفت	فاصله پیک	کروموزوم	MLR	مقدار p		مقدار α	
				$x^2_{1,p}$	$x^2_{2,p}$	$x^2_{1,\alpha}$	$x^2_{2,\alpha}$
طول	RZ730-RZ801	۱	۲۸/۶	۸/۹۰ e^{-8}	۶/۱۶ e^{-7}	۱/۴۵ e^{-5}	۱/۰۰ e^{-4}
	RZ519-RZ448	۳	۱۹/۲	۱/۱۸ e^{-5}	۶/۷۷ e^{-8}	۱/۹۲ e^{-3}	۱/۱۰ e^{-2}
	RZ337A-CDO337	۳	۲۸/۶	۸/۹۰ e^{-8}	۶/۱۶ e^{-7}	۱/۴۵ e^{-5}	۱/۰۰ e^{-4}
	G2155-RG134	۱۰	۲۶/۲	۳/۰۸ e^{-7}	۲/۰۴ e^{-6}	۵/۰۲ e^{-5}	۳/۳۳ e^{-4}
عرض	RG810-RG331	۱	۱۸/۲	۱/۹۹ e^{-5}	۱/۱۲ e^{-4}	۳/۲۴ e^{-3}	۱/۸۲ e^{-2}
	RZ318-RZ58	۲	۱۴/۰	۱/۸۳ e^{-4}	۹/۱۲ e^{-4}	۲/۹۸ e^{-2}	۱/۴۹ e^{-1}
	CDO87-Pgi-1	۳	۱۴/۴	۱/۴۸ e^{-4}	۷/۴۶ e^{-4}	۲/۴۱ e^{-2}	۱/۲۲ e^{-1}
	RG134-RZ500	۱۰	۱۴/۱	۱/۸۳ e^{-4}	۸/۶۷ e^{-4}	۲/۸۲ e^{-2}	۱/۴۱ e^{-1}
	RZ536-G186	۱۱	۱۵/۷	۷/۴۲ e^{-5}	۳/۹۰ e^{-4}	۱/۲۱ e^{-2}	۶/۳۵ e^{-2}
طول به عرض	RG157-RZ318	۲	۱۹/۶	۹/۵۵ e^{-6}	۵/۵۴ e^{-5}	۱/۵۶ e^{-3}	۹/۰۴ e^{-3}
	RZ519-Pgi-1	۳	۱۸/۵	۱/۷۰ e^{-5}	۹/۶۱ e^{-5}	۲/۸۷ e^{-3}	۱/۵۷ e^{-2}
	RG179-CDO337	۳	۲۳/۲	۱/۴۶۰۰ e^{-6}	۹/۱۶۶۱ e^{-6}	۲/۳۸ e^{-4}	۱/۴۹ e^{-3}

۱۲-۲-۴- آزمون‌های تبدیل یا جایگشت

چرچیل و دورج (۱۹۹۴) یک روش عددی مبتنی بر داده‌ها را پیشنهاد کردند که بر اساس مفهوم آزمون تبدیل یا جایگشت بود و مقادیر بحرانی تجربی را برای مکان‌یابی QTL و برای مجموعه داده‌های معینی برآورد می‌نماید. با تغییر تصادفی روابط بین داده‌های فنوتیپی و نشانگری در بین افراد، آزمون‌های جایگشت مجموعه داده‌های جدیدی را ایجاد می‌کنند که در آن، ارتباط اصلی نشانگر-QTL از بین رفته است. مزایای بزرگ این راهکار سادگی مفهومی آن، ماهیت عاری از توزیع و کاربرد کلی آن در ساختارهای جمعیتی مختلف است، اگرچه بار محاسباتی آن سنگین است. مراحل این روش به‌صورت زیر است:

۱- ژنوتیپ‌های نشانگر منفرد را به‌طور تصادفی با فنوتیپ صفت منفرد جفت نمائید تا نمونه جایگشت داده‌ها ایجاد شود (این حالت، فرض صفر مبنی بر عدم وجود ارتباط بین QTL و فنوتیپ را شبیه‌سازی می‌کند).

۲- مکان‌یابی فاصله‌ای را برای نمونه جایگشت یافته انجام دهید.

۳- مراحل ۱ و ۲ را چندین بار تکرار کنید تا توزیع تجربی آماره آزمون را در فرض صفر برای تعیین مقدار بحرانی مناسب جهت آماره آزمون در تجزیه داده‌های اصلی به‌دست آورید.

به‌طور اختصاصی‌تر، $X=\{X_1, X_2, ..., X_n\}$ را برای نشان دادن یک نمونه تصادفی با اندازه n در جمعیت استفاده کنید. در مکان‌یابی QTL، هر نقطه مشاهده‌ای X_i از روی مقدار یا ارزش صفت y_i و ژنوتیپ نشانگر $M_i=\{M_{i1}, M_{i2}, ..., M_{im}\}$ برای m نشانگر ایجاد شده است، یعنی:

$$X_i=\{y_i, M_i\}$$

یک نمونه جایگشتی با اندازه n که به‌صورت $X'=\{X'_1, X'_2, ..., X'_n\}$ نشان داده می‌شود، نمونه‌ای از نمونه اصلی بدون جایگزینی است، به‌گونه‌ای که ارتباطات تصادفی بین y_i و M_i وجود دارد، یعنی:

$$X'_k=\{y_{i_k}, M_{i_k}\}$$

که در آن، y_{i_k} بدون جایگزینی از روی $(M_1, ..., M_n)$ انتخاب می‌شود. از اینرو، در نمونه‌های جایگشت یافته، X'، y' و M' هیچ گونه ارتباط ذاتی ندارند و بنابراین، فرض صفر عدم وجود QTL شبیه‌سازی می‌شود.

برای N نمونه جایگشت شده، LR'_p آماره آزمون حداکثر درستنمایی برای موقعیت ژنومی خاص یا مقدار حداکثری آماره آزمون برای فاصله نشانگری یا برای تمامی ژنوم در p امین نمونه جایگشت یافته است. آستانه $\alpha\times100$ درصدی آماره آزمون با در نظر گرفتن فرض صفر می‌تواند به صورت تجربی بدین شکل برآورد شود:

$\hat{T}_\alpha = \alpha \times 100$ درصد از $\{LR'_1, LR'_2, ..., LR'_N\}$

مقدار N برای تجزیه داده‌های عملی چقدر است؟ چرچیل و دورج (۱۹۹۴) برای $\alpha = 0.05$ پیشنهاد می‌کنند که N باید حداقل ۱۰۰۰ باشد و برای $\alpha = 0.01$، N باید حداقل ۵۰۰۰ باشد.

۱۲-۲-۵- یک روش سریع

برای غلبه بر معایب روش‌های تجربی که به وسیله محاسبات زیاد آن‌ها ایجاد می‌شود، پیفو (۲۰۰۱) روش سریعی را برای محاسبه مقادیر آستانه تقریبی که نرخ خطای نوع اول کل ژنوم را برای تشخیص QTL بر اساس نتایج دیویس (۱۹۷۷) کنترل می‌کنند، پیشنهاد نمود. مشروط به ترکیبی از توزیع‌های نرمال با واریانس ثابت و پارامترهای مکانی وابسته به آثار QTL، آماره آزمون LR در موقعیت‌های متفاوت QTL (θ) محاسبه می‌شوند. با فرض این که مقادیر LR تابع پیوسته‌ای از θ باشند که به صورت $L(\theta)$ بیان می‌شود و مشروط به موقعیت QTL، $L(\theta)$ از یک توزیع χ^2 با k درجه آزادی تبعیت می‌کند که k تعداد آثار ژنتیکی برای QTL مفروض است. حد بالاتر نرخ خطای نوع اول کل کروموزوم (α) با رابطه زیر (پیفو، ۲۰۰۱) برآورد می‌شود:

$$\alpha = \Pr\left(x_k^2 > T\right) + \frac{VT^{\frac{1}{2}(k-1)} e^{-\frac{1}{2}T} 2^{-\left(\frac{1}{2}\right)^k}}{\Gamma\left(\frac{1}{2}k\right)} \qquad (۱۲-۸)$$

که $\Pr\left(x_k^2 > T\right)$ تابع توزیع تجمعی χ^2 با k درجه آزادی، T مقدار آستانه بحرانی برای آماره آزمون LR، $\Gamma(.)$ تابع گاما است و

$$V = \int_0^l \left| \frac{\partial \sqrt{L(\theta)}}{\partial \theta} \right| d\theta$$

$$= \left|\sqrt{L(0)} - \sqrt{L(\theta_1)}\right| + \left|\sqrt{L(\theta_1)} - \sqrt{L(\theta_2)}\right| + ... + \left|\sqrt{L(\theta_s)} - \sqrt{L(\theta_l)}\right|$$

که $\theta_1, ..., \theta_s$ نقاط برگشت متوالی (نقاط تغییر یا خمش) $\sqrt{L(\theta)}$ هستند (یعنی مقادیر θ که اولین مشتق $\frac{\partial \sqrt{L(\theta)}}{\partial \theta}$ علامت را تغییر می‌دهد). این تغییر علامت در نقطه حداقل و حداکثر $\sqrt{L(\theta)}$ رخ می‌دهد.

برای مقدار معینی از α، T ممکن است از روی معادله (۱۲-۸) با روش‌های عددی مشخص شود. مشکلی که در عمل وجود دارد، یافتن نقاط برگشت $\theta_1, ..., \theta_s$ است. در بیشتر موارد، این

حالت باید به صورت عددی صورت گیرد. معمولاً جستجوی شبکه‌ای برای همگی مقادیر θ انجام می‌شود و بنابراین، نقاط برگشت می‌توانند فقط برای صحتی که با اندازه مرحله خاصی از شبکه ارایه می‌شوند، تعیین شوند؛ بنابراین، استفاده از شبکه‌های نسبتاً ظریف (برای مثال بین ۱ و ۲ سانتی‌مورگانی) پیشنهاد می‌شود. تجزیه با در نظر گرفتن این فرض که هر نقطه روی شبکه دارای یک نقطه برگشت است، ساده‌سازی می‌شود.

با استفاده از نامعادله بونفرونی، پیفو (۲۰۰۱) نشان داد که نرخ خطای نوع اول کل ژنوم با رابطه زیر محاسبه می‌شود:

(۹-۱۲)

$$\gamma = C\Pr\left(x_k^2 > T\right) + \frac{\left(\sum_{c=1}^{C} V_c\right) T^{\frac{1}{2}(k-1)} e^{-\frac{1}{2}T} 2^{-\left(\frac{1}{2}\right)^k}}{\Gamma\left(\frac{1}{2}k\right)}$$

که C تعداد کروموزوم‌ها و V_c مقدار V برای c امین کروموزوم است. به‌جای انتخاب α_c مشابه برای هر کروموزوم، پیشنهاد می‌شود که مقدار معمول بحرانی T برای تمامی کروموزوم‌ها استفاده شود، در حالی که α_c ممکن است روی هر کروموزوم، متفاوت باشد.

مثال ۱۲-۲. **(تعیین تجربی آستانه‌های بحرانی)**. مثال‌های ۱۲-۴ و ۱۱-۵ به ترتیب نتایج مکان‌یابی QTL را در جمعیت DH مشابه تلاقی برگشتی در برنج و جمعیت F_2 موش گزارش می‌کنند. در این بخش، آستانه‌های بحرانی برای این دو مثال با استفاده از روش‌های تجزیه‌ای محاسبه می‌شوند، روش سریع پیفو (۲۰۰۱)، مطالعات شبیه‌سازی و آزمون‌های جایگشت به صورت زیر هستند:

روش	DH برنج		F_2 موش	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
روش سریع	۱۲/۹۶	۱۶/۱۸	۱۶/۴۵	۱۹/۸۶
مطالعات شبیه‌سازی	۹/۱۴	۱۲/۶۴	۱۰/۶۰	۱۴/۵۸
آزمون‌های جایگشت	۸/۲۰	۱۲/۷۹	۹/۷۱	۱۳/۴۹

می‌توان مشاهده کرد که سه روش، نتایج بسیار هماهنگ و منطبقی را برای آستانه‌های بحرانی ارایه داده‌اند. روش سریع نسبت به دو روش دیگر محافظه‌کارانه‌تر است، به ویژه زمانی که نشانگرها نسبتاً متراکم باشند (پیفو، ۲۰۰۱).

۱۲-۲-۶- آماره امتیازی

به تازگی، ژو و همکاران (۲۰۰۴) یک آماره امتیازی برای آزمون معنی‌داری مکان‌یابی QTL ژنوم گسترده برای تلاقی‌های آزمایشی پیشنهاد دادند. روش پیشنهادی بر اساس روش تکرار نمونه‌گیری از نظر محاسباتی نسبت به روش جایگشتی (به نسبت 10^2 و بالاتر) دارای توان کمتری بوده و برای طرح‌های آمیزشی پیچیده و مدل‌های ژنتیکی پیچیده که نمی‌توانند با روش‌های جایگشت و نظری بکار گرفته شوند، کاربرد دارد.

ایده این راهکار از روی این بحث به‌دست آمد که آماره آزمون نسبت درستنمایی برای آزمون حضور، معادل آماره آزمون امتیازی در نمونه‌های بزرگ است (کاکس و هینکلی، ۱۹۷۴). آماره امتیاز می‌تواند به وسیله مجموع بردارهای تصادفی مستقل تقریب زده شود و در نتیجه، توزیع آن می‌تواند به سادگی برای نمونه‌های بزرگ به‌دست آید. در مکان‌یابی فاصله‌ای، بردار پارامترهای آثار QTL با β و بردار پارامترهای مزاحم (شامل میانگین کل و واریانس باقیمانده) با η نشان داده می‌شوند. این دو نوع پارامتر به صورت $\Phi=(\beta,\eta)$ بیان می‌شوند. لگاریتم درستنمایی Φ در موقعیت معین x برای فرد i ام می‌تواند بدین صورت بیان شود:

$$L_i(\Phi;x)=\log\left[\sum_{j=1}^{J}\omega_{j|i}(x)\,f_j(y_i;\Phi)\right] \qquad (۱۰-۱۲)$$

البته با این فرض که تعداد J ژنوتیپ QTL برای جمعیت مکان‌یابی وجود دارد. $LR(x)$ نیز مقدار LR در موقعیت x و در سرتاسر ژنوم یا ناحیه خاصی از آن است. چون x می‌تواند هر نقطه‌ای از ژنوم باشد، توزیع $LR(x)$ می‌تواند به‌عنوان یک فرایند تصادفی در نظر گرفته شود.

توابع امتیاز فرد i ام برای β و η به ترتیب به‌صورت زیر محاسبه می‌شوند:

$$S_{\beta i}(\Phi;x)=\frac{\partial}{\partial\beta}L_i(\Phi;x)$$

$$S_{\eta i}(\Phi;x)=\frac{\partial}{\partial\eta}L_i(\Phi;x)$$

تحت فرض صفر $H_0:\beta=0$، پارامترهای مزاحم را با $\tilde{\eta}$ نشان می‌دهیم. $S_i(x)$ نیز تابعی از فرد i ام برای β است که در $H_0:\beta=0$ و $\eta=\tilde{\eta}$ ارزیابی می‌شود. بر اساس بسط سری

تیلور و قانون اعداد بزرگ، بسط $S_i(x)$ را به‌دست آورده (ژو و همکاران، ۲۰۰۴) و بدین صورت برآورد می‌کنیم:

$$\hat{S}_i(x) = S_{\beta i}(0,\tilde{\eta};x) - \left[\frac{\partial^2}{\partial\beta\partial\eta}L_i(0,\tilde{\eta};x)\right]\left[\frac{\partial^2}{\partial\eta^2}L_i(0,\tilde{\eta};x)\right]^{-1} S_{\eta i}(0,\tilde{\eta};x)$$

آماره آزمون امتیاز برای $H_0:\beta=0$ در مقابل $H_1:\beta\neq 0$ در موقعیت x، شکل زیر را ایجاد می‌کند:

$$W(x) = \hat{S}^T(x)\hat{V}^{-1}(x)\hat{S}(x) \qquad (۱۱-۱۲)$$

که در آن:

$$\hat{S}(x) = \sum_{i=1}^{n}\hat{S}_i(x)$$

$$\hat{V}(x) = n\hat{\Psi}(x,x)$$

که در آن، $\hat{\Psi}(x,x)$ حالت خاصی از رابطه زیر است:

$$\hat{\Psi}(x_1,x_2) = n^{-1}\sum_{i=1}^{n}\hat{S}_i(x_1)\hat{S}_i^T(x_2)$$

که به‌عنوان کوواریانس بین $n^{-\frac{1}{2}}S(x_1)$ و $n^{-\frac{1}{2}}S(x_2)$ در دو موقعیت معین x_1 و x_2 است. همان‌گونه که به وسیله کاکس و هینکلی (۱۹۷۴) نشان داده شد، $W(x)$ به صورت مجانبی معادل $LR(x)$ است. معنی‌داری ژنوم گسترده QTL می‌تواند با ارزیابی توزیع $\max_x W(x)$ از راه روش تکرار نمونه‌گیری آزمون شود. این روش به وسیله ژو و همکاران (۲۰۰۴) به صورت زیر توصیف شده است:

۱- متغیر مستقل تصادفی نرمال استاندارد را به صورت $G_i\ (i=1,\ldots,n)$ از روی $N(0,1)$ نمونه‌گیری کنید.

۲- عبارت زیر را محاسبه کنید:

$$S^*(d) = \sum_{i=1}^{n}\hat{S}_i(x)G_i$$

$$W^*(d) = S^{*T}(d)\hat{V}^{-1}(d)S^*(d)$$

$$M^* = \max_x W^*(x)$$

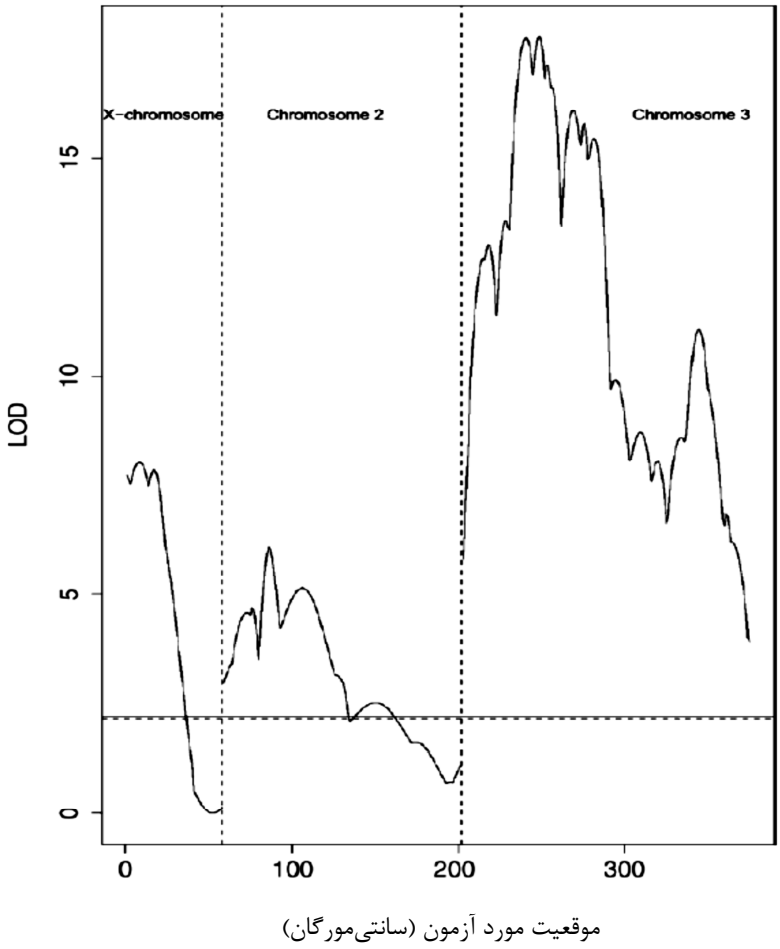
۳- مراحل ۱ و ۲ را R بار تکرار کنید (R عدد بزرگی است).

۴- برای نرخ خطای نوع اول کل ژنوم α، صدک $100(1-\alpha)$ ام مقادیر R مربوط به M^* را محاسبه کنید. اگر مقدار مشاهده شده LR بیشتر از این آستانه باشد، سپس فرض صفر را رد کنید.

مثال ۱۲-۳. ژو و همکاران (۲۰۰۴) از داده‌های منتشر شده مگس سرکه (ژنگ و همکاران، ۲۰۰۰) برای نشان دادن میزان مفید بودن آماره آزمون امتیاز استفاده کردند. یک نقشه پیوستگی شامل ۴۲ نشانگر برای کروموزوم‌های X، ۲ و ۳ در یک جمعیت تلاقی برگشتی بین *D. simulans* و *D. mauritiana* که شامل ۲۹۹ مگس بود، تهیه شد. مکان‌یابی فاصله‌ای برای جستجوی QTL که روی صفت کمّی شکل مگس مؤثر بود، استفاده شد. شکل ۱۲-۱ پروفیل امتیاز LOD محاسبه شده در هر ۱ سانتی‌مورگان را در طول سه کروموزوم نشان می‌دهد. آستانه بحرانی کل ژنوم به وسیله آزمون‌های جایگشت و روش آماره امتیاز تعیین شد که هر دو مورد مقادیر مشابه ۱۰/۰۸ و ۹/۹۶ را در سطح معنی‌دار ۵ درصد بر اساس ۱۰۰۰۰ جایگشت و تکرار نمونه‌گیری ارایه کردند (شکل ۱۲-۱). ولی این دو روش تا حد زیادی از نظر زمان محاسباتی با یکدیگر متفاوت هستند که زمان مورد نیاز برای آماره امتیاز ۱۳ ثانیه بود، در حالی که آزمون‌های جایگشت از ۶۰۰۰ ثانیه برای انجام محاسبات استفاده کردند.

۱۲-۳- دقت برآورد پارامتر

برآوردهای دقیق پارامترهای QTL برای مکان‌یابی ژنتیکی ضروری هستند. دقت برآورد پارامترها می‌تواند با برآورد واریانس‌های نمونه‌گیری برآوردهای حاصل از ماتریس واریانس-کوواریانس مجانبی تخمین زده شود. یک روش جایگزین، شبیه‌سازی چندین مجموعه داده‌های چندگانه (مرکب) از راه تقلید ساختار داده‌های مثال مورد نظر و برآورد میانگین‌ها و واریانس‌های نمونه‌گیری تکرارهای شبیه‌سازی شده است.



شکل ۱۲-۱- پروفیل امتیاز LOD در طول کروموزوم‌های X، ۲ و ۳. خطوط افقی آستانه‌های بحرانی ۹۵ درصد نمونه‌گیری مجدد (خطوط کامل و پر) و جایگشت (خطوط نقطه چین) هستند. اقتباس از زو و همکاران (۲۰۰۴).

۱۲-۳-۱- ماتریس واریانس-کوواریانس مجانبی

پس از به‌دست آوردن برآوردهای نقطه‌ای از راه الگوریتم EM، نیاز داریم تا ماتریس واریانس-کوواریانس مجانبی را به‌دست آورده و خطاهای نمونه‌برداری برآوردها را ارزیابی کنیم. در این بخش، روشی را برای به‌دست آوردن ماتریس واریانس-کوواریانس مجانبی برای موقعیت‌ها و آثار QTL ارایه می‌کنیم. تکنیک‌هایی که برای انجام این امر وجود دارند، مستلزم محاسبه ماتریس اطلاعات ناقص داده‌ها است که همان منفی مشتق درجه دوم لگاریتم درستنمایی داده‌های ناقص

است. بحث کامل‌تر در خصوص این روش به وسیله کائو و ژنگ (۱۹۹۷) ارایه شده است. همان‌گونه که در مطالب بالا دیدیم، الگوریتم EM به صورت خودکار، ماتریس واریانس-کوواریانس برآوردها را محاسبه نمی‌کند و بنابراین پیشنهاد می‌دهد که در نظر گرفتن چند مرحله اضافی برای انجام این عمل ضروری است. لوئیس (۱۹۸۲) و منگ و رابین (۱۹۹۱) یک روش کلی برای برآورد ماتریس واریانس-کوواریانس یک مدل مختلط ارایه نمودند. ایده اساسی این روش آن است که اطلاعات ناقص داده‌ها (مشاهده شده) می‌تواند با استخراج اطلاعات داده‌های گمشده از روی اطلاعات کامل داده‌ها به‌دست آید.

معادلات (۹-۲) و (۹-۳) مدل‌های مختلطی هستند که در آن‌ها، فنوتیپ (y_i) و اطلاعات نشانگر (M) داده‌های مشاهده شده ناقص هستند که به صورت Y_{obs} نشان داده شدند، در حالی که اطلاعات ژنوتیپ QTL موجود در احتمالات شرطی ($\omega_{j|i}$)، داده‌های غیر قابل مشاهده یا گمشده هستند که به صورت Y_{mis} نشان داده شدند. ترکیب داده‌های ناقص و گمشده، داده‌های کامل هستند که به صورت Y_{com} نشان داده می‌شوند.

تابع توزیع احتمال توأم داده‌های کامل می‌تواند به صورت زیر بیان شود:

$$f\left(Y_{com}\,|\,\Omega\right)=f\left(Y_{obs},Y_{mis}\,|\,\Omega\right)=f\left(Y_{obs}\,|\,\Omega\right)f\left(Y_{mis}\,|\,Y_{obs},\Omega\right) \qquad (۱۲-۱۲)$$

که در آن، $f\left(Y_{obs}\,|\,\Omega\right)$ چگالی داده‌های مشاهده بوده که همان درستنمایی مختلط است و $f\left(Y_{mis}\,|\,Y_{obs},\Omega\right)$ چگالی داده‌های گمشده مشروط به داده‌های مشاهده شده است. لگاریتم درستنمایی مربوط به $f\left(Y_{com}\,|\,\Omega\right)$ عبارت است از:

$$\log L\left(\Omega\,|\,Y_{com}\right)=\log L\left(\Omega\,|\,Y_{obs}\right)+\log f\left(Y_{mis}\,|\,Y_{obs},\Omega\right)$$

یعنی:

$$\log L\left(\Omega\,|\,Y_{obs}\right)=\log L\left(\Omega\,|\,Y_{com}\right)-\log f\left(Y_{mis}\,|\,Y_{obs},\Omega\right) \qquad (۱۳-۱۲)$$

با گرفتن مشتق‌های دوم و امیدهای ریاضی $f\left(Y_{mis}\,|\,Y_{obs},\Omega\right)$ و ارزیابی $\Omega=\hat{\Omega}$ برای معادله (۱۰-۱۸)، لوئیس (۱۹۸۲) اطلاعات مشاهده شده را به شکل زیر ارایه نمود:

$$I_{obs}\left(\hat{\Omega}\,|\,Y_{obs}\right)=I_{com}-I_{mis} \qquad (۱۴-۱۲)$$

که در آن، اطلاعات گمشده به صورت زیر است:

$$I_{mis}=E\left[-\frac{\partial^2\log f\left(Y_{mis}\,|\,Y_{obs},\Omega\right)}{\partial\Omega^2}\,|\,Y_{obs},\Omega\right]_{\Omega=\hat{\Omega}}$$

$$=E\left[S\left(Y_{com},\Omega\right)S^T\left(Y_{com},\Omega\right)|\,Y_{obs},\Omega\right]_{\Omega=\hat{\Omega}}$$

و اطلاعات کامل عبارت است از:

$$I_{com} = E\left[-\frac{\partial^2 \log f\left(Y_{com} \mid \Omega\right)}{\partial \Omega^2} \mid Y_{obs}, \Omega\right]_{\Omega=\hat{\Omega}}$$

که در آن، s بردار گرادیان لگاریتم درستنمایی است.

بر اساس معادله (۱۰-۱۹)، ماتریس اطلاعات مشاهده شده برای حالت استقلال، اما نه لزوماً دارای توزیع یکسان، می‌تواند بدین صورت بیان شود:

$$\begin{aligned} & I_{obs}\left(\hat{\Omega} \mid Y_{obs}\right) \\ & = \sum_{i=1}^{n} E\left[-\frac{\partial^2 \log f_i\left(Y_{com} \mid \Omega\right)}{\partial \Omega^2} \mid Y_{obs}, \Omega\right]_{\Omega=\hat{\Omega}} \\ & - \sum_{i=1}^{n} E\left[S_i\left(Y_{(com,i)}, \Omega\right) S_i^T\left(Y_{(com,i)}, \Omega\right) \mid Y_{(obs,i)}, \Omega\right]_{\Omega=\hat{\Omega}} \\ & - \sum_{i_1 \neq i_2} E\left[S_{i1}\left(Y_{(com,i1)}, \Omega\right) \mid Y_{(obs,i1)}, \Omega\right]_{\Omega=\hat{\Omega}} E\left[S_{i2}^T\left(Y_{(com,i2)}, \Omega\right) \mid Y_{(obs,i2)}, \Omega\right]_{\Omega=\hat{\Omega}} \end{aligned}$$

کائو و زنگ (۱۹۹۷) عبارات هر جزء موجود در ماتریس‌های اطلاعات کامل و گمشده را ارایه کردند.

۱۲-۳-۲- مطالعات شبیه‌سازی

از روی یک مجموعه داده واقعی، از مکان‌یابی فاصله‌ای برای برآورد موقعیت و آثار QTL و واریانس باقیمانده مربوط به آن استفاده شده است. دقت برآورد پارامترها می‌تواند از راه مطالعات شبیه‌سازی ارزیابی شود. در این بخش، یک طرح شبیه‌سازی پایه که برای دستیابی به این هدف استفاده شده است، تشریح می‌شود.

یک جمعیت تلاقی برگشتی را در نظر بگیرید که در آن، m نشانگر برای تهیه نقشه پیوستگی تعیین ژنوتیپ شده و یک صفت کمّی برای هر یک از n فرد تعیین فنوتیپ شده‌اند. مکان‌یابی فاصله‌ای به‌طور موفقیت‌آمیز، یک QTL (Q) را در موقعیت $\left(\hat{\theta}\right)$ بین یک جفت نشانگر مجاور M_l و M_{l+1} با نرخ نوترکیبی r_l ($l = 1, \ldots, m-1$) روی کروموزوم شناسایی کرده است. اثر افزایشی QTL به صورت $\hat{a}$ و دو پارامتر مزاحم به صورت $\hat{\mu}$ برای میانگین کل و $\hat{\sigma}^2$ برای واریانس باقیمانده برآورد شدند. بر اساس این برآوردها، نیازمند شبیه‌سازی یک نقشه پیوستگی مشابه با m نشانگر، ارزش های فنوتیپی توجیه شده به وسیله Q و خطاهای باقیمانده برای جمعیت تلاقی

برگشتی با اندازه n هستیم. مشاهدات ژنوتیپی نشانگر برای n فرد در m نشانگر شناخته شده از یک توزیع چند جمله‌ای نمونه‌گیری می‌شوند:

$$\left[(1-r_1)(1-r_2)\ldots(1-r_{m-1})\right]^{N_1}\left[(1-r_1)(1-r_2)\ldots(1-r_{m-1})\right]^{N_2}\ldots\left[r_1r_2\ldots r_{m-1}\right]^{N_{2m}}$$

که در آن، $n = N_1 + N_2 + \ldots + N_{2m}$ است.

مدل آماری مورد استفاده برای شبیه‌سازی ارزش فنوتیپی فرد i ام به شکل زیر است:

$$y_i = \hat{\mu} + x_i\hat{a} + e_i \qquad (۱۵-۱۲)$$

که در آن، x_i متغیر شاخص و موهومی است و به صورت ۱ یا ۰ بسته به ژنوتیپ QTL تعریف می‌شود و e_i خطای باقیمانده برای فرد i ام است که از توزیع $N\left(0,\hat{\sigma}^2\right)$ تبعیت می‌کند. برای هر فرد i ام که ژنوتیپ نشانگر مربوط به آن شناخته شده است، احتمال x_i برابر با ۱ یا ۰ بوده و با احتمال شرطی ژنوتیپ QTL مشروط به ژنوتیپ نشانگر برای نشانگرهای M_l و M_{l+1} تعیین می‌شود (جدول ۱۰-۳ یا جدول ۱۰-۴).

داده‌های شبیه‌سازی شده‌ای که به‌طور کامل از داده‌های واقعی تقلید و تبعیت می‌کنند، با مکان‌یابی فاصله‌ای تجزیه می‌شوند که در آن، مجموعه مشابهی از پارامترها (μ, a, σ^2, θ) می‌توانند برآورد شوند. این روش به دفعات زیاد (برای مثال ۱۰۰۰ بار) تکرار می‌شود که اجازه برآورد میانگین و خطاهای نمونه‌برداری برآوردها را می‌دهد. بر اساس میانگین‌ها و خطاهای نمونه‌برداری، می‌توان صحت و دقت هر یک از پارامترهای برآورد شده را تعیین نمود.

استفاده دیگر از مطالعات شبیه‌سازی، آزمون توان مکان‌یابی فاصله‌ای برای تجزیه یک مجموعه داده‌های حقیقی معین است. برای هر تکرار شبیه‌سازی، آستانه بحرانی برای اعلام وجود یک QTL می تواند با استفاده از روش‌هایی که در مطالب بالا اشاره شدند، تعیین شود. این آستانه با مقدار LR محاسبه شده در سطح معنی‌دار معینی مقایسه می‌شود. سهم تعداد تکرارهای شبیه‌سازی شده که در آن‌ها، QTL شناسایی شده در تعداد کلی تکرارهای شبیه‌سازی معنی‌دار شده است، می‌تواند به صورت تجربی به‌عنوان توان تشخیص QTL در آن مجموعه داده معین استفاده شود.

مثال ۱۲-۴. مثال ۳-۱ را مجدداً مشاهده کنید. دو لاین اینبرد، IR64 نیمه پاکوتاه و Azucena پا بلند با یکدیگر تلاقی داده شدند تا جمعیت دابل هاپلویید را ایجاد کنند. نقشه پیوستگی جمعیت تهیه و ارتفاع بوته‌ها در سن ۱۰ هفتگی اندازه‌گیری شدند. در مثال ۱۲-۴، یک QTL معنی دار بین نشانگرهای RG810 و RG331 روی کروموزوم ۱ شناسایی شد (شکل ۱۱-۲).

MLE های مربوط به پارامترهای QTL و برآورد خطاهـای نمونـه‌بـرداری (SE) MLE هـا بـه وسیله ماتریس واریانس-کوواریانس مجانبی برای QTL شناسایی شده در نقطه اوج پروفیل LR در جدول ۱۲-۲ ارایه شده است. مقادیر برآورد شده، برای شبیه‌سازی داده‌ها به وسیله تقلیـد سـاختار داده‌های مثال واقعی استفاده شدند. داده‌های شبیه‌سازی شده برای پارامترهای ارایه شده در جدول ۱۲-۲ برآورد می‌شوند و روش مشابهی بـرای ۱۰۰۰ بـار تکـرار مـی‌شـود. میـانگین‌هـا و خطاهـای نمونه‌برداری تمامی تکرارهای شبیه‌سازی شده برای هر پارامتر، به همراه توان تشخیص این QTL، در جدول ۱۲-۲ ارایه شده‌اند. مشخص شد که پارامترهای ژنتیکی برای این جمعیت DH می‌توانند با دقت خوبی برآورد شوند و توان تشخیص QTL های معنی‌دار در این جمعیت خیلی زیـاد اسـت. روش‌های مجانبی و شبیه‌سازی، دقت برآورد پایداری بـرای پارامترهـای موقعیـت و اثـر QTL و دو پارامتر مزاحم (میانگین کل، μ و واریانس باقیمانده، σ^2) فراهم می‌کنند.

جدول ۱۲-۲- تجزیه دقت مکان‌یابی QTL در جمعیت DH برنج به وسیله روش‌های مجانبی و شبیه‌سازی.

پارامتر	مجانبی		شبیه‌سازی		
	MLE	SE	میانگین	SE	توان
μ	۱۰۹/۳۲	۱/۷	۱۰۹/۳۴	۱/۸	
a	۳۰/۷۹	۳/۴	۳۱/۰۳	۳/۶	۰/۹۹
σ^2	۲۹۵/۴۶	۲۹/۵	۲۷۸/۵۵	۴۵/۲	
موقعیت	۲۱۵/۹۸	۱/۶	۲۱۶/۰۴	۱/۰	

۱۲-۴- فواصل اطمینان برای موقعیت QTL

در بخش قبل سعی نمودیم تا ماتریس واریـانس-کوواریـانس مجـانبی را بـرای تمـامی پارامترهـای QTL از جمله موقعیت QTL (θ) برآورد نماییم. برای اجتناب از عدم توانایی تشخیص بین اثـر و موقعیت QTL تحت فرض صفر، اغلب موقعیت QTL را به عنوان یک عامل ثابت در نظر می‌گیریم. اگر موقعیت QTL ثابت باشد، مشتق مجانبی ذکر شده فقط برای آثار QTL و واریـانس باقیمانـده که در قالب Ω مرتب شده‌اند، مورد نیاز است. در این حالت، نیاز داریم تا خطـای نمونـه‌بـرداری و فاصله اطمینان را برای موقعیت QTL بر اساس روش‌های دیگری برآورد کنیم.

لندر و بوتستین (۱۹۸۹) روش کـاهش امتیـاز LOD[1] را بـرای بـرآورد "فواصـل حمـایتی" یـا فواصل اطمینان (CI) موقعیت QTL بر اساس آزمون نسبت درستنمایی پیشنهاد دادند. با اسـتفاده از این روش، امتیاز LOD در دو طرف موقعیت QTL شناسایی شده به میزان یک یا دو واحـد کـم

[1] - Dropoff

می‌شود. سپس عرض کلی متعلق به ۱ یا ۲ واحد کاهش در مقدار LOD به عنوان CI مشخص می‌شود که به صورت مجانبی باید به ترتیب تقریباً معادل فاصله اطمینان ۹۶/۶ و ۹۹/۸ درصد CI - باشد (مانجین و همکاران، ۱۹۹۴)؛ اما آن‌ها متوجه شدند که روش کاهش امتیاز LOD که به وسیله لندر و بوتستین (۱۹۸۹) ارایه شد، به‌طور پایداری CI را به ویژه برای آثار کوچک QTL کمتر از حد برآورد می‌کند. آن‌ها یک آماره آزمون پیچیده به‌دست آوردند که به‌طور صحیحی CI را برای آثار کوچک QTL برآورد می‌کرد، ولی توزیع این آماره آزمون باید به‌صورت تجربی محاسبه شود (مانجین و همکاران، ۱۹۹۴).

فاصله اطمینان می‌تواند با شبیه‌سازی تعداد زیادی از نمونه‌ها نیز برآورد شود. اگر ۱۰۰۰ نمونه تهیه شود، حدود اطمینان ۹۵ درصد با تعیین ۲۵ عدد از پایین‌ترین و ۲۵ عدد از بالاترین برآوردها برای هر یک از پارامترها به‌دست می‌آید. داروا‌سی و همکاران (۱۹۹۳) واریانس خطای برآورد پارامترهای QTL را بر اساس ماتریس اطلاعات فیشر و با تکرار شبیه‌سازی برای طرح تلاقی برگشتی با استفاده از دسته‌ای از نشانگرها برآورد کردند. ۹۵ درصد CI برای هر پارامتر به صورت 2 SE ± برآورد شد. داروا‌سی و همکاران (۱۹۹۳) به‌طور مستقیم نیز فاصله اطمینان ۹۵ درصد را برای هر پارامتر با تکرار شبیه‌سازی برآورد نمودند. تمامی روش‌ها برای برآورد واریانس‌های آثار QTL خیلی دقیق بودند. برآوردهای حاصل از مشتق دوم، به میزان اندکی SE را برای میانگین‌های QTL مربوط به برآوردهای تجربی، به‌ویژه برای فضای بزرگ بین نشانگرها، بیشتر از حد برآورد کردند.

برای موقعیت نقشه QTL، به‌طورکلی برآوردهای فواصل اطمینان ۹۵ درصد تجربی و برآوردهای حاصل از چهار برابر SE تجربی مشابه بودند؛ اما برآوردهای مبتنی بر ماتریس مشتق دوم تمایل داشتند CI را برای فواصل نشانگر کوچک کمتر از حد و برای فواصل نشانگر بزرگ بیشتر از حد برآورد کنند. تفاوت‌ها در برخی موارد بیش از دو برابر بودند. به وضوح، برای این پارامتر، خصوصیات مجانبی ماتریس مشتق دوم حفظ نمی‌شود.

همچنین روش‌های تجربی برآورد CI شامل روش‌های پارامتری و غیر پارامتری خود راه انداز یا بوت استراپ[1] و جک نایف[2] هستند (افرون و تیبشیرانی، ۱۹۹۳). در روش پارامتری بوت استراپ، برآورد پارامترها ابتدا با هر یک از روش‌های مورد نظر انجام می‌شود. در مرحله دوم، تعداد زیادی از نمونه‌های خود راه انداز از روی توزیع‌های نظری فرضی به‌دست می‌آیند که فرض می‌کنند برآوردهای اصلی پارامترها همان مقادیر پارامترها هستند. سپس برآوردهای پارامترها برای هر نمونه

---

[1] - Bootstrap
[2] - jackknife

به‌دست می‌آیند. سپس CI برای هر پارامتر از روی توزیع‌های تجربی برآورد پارامترهای حاصل از نمونه‌های تولید شده حاصل می‌شود.

در روش بوت استراپ ناپارامتری (ویشر و همکاران، ۱۹۹۶)، تعداد زیادی از نمونه‌های خود راه‌انداز با نمونه‌گیری از تکرارهای حاصل از داده‌های اصلی ایجاد می‌شوند؛ بنابراین، در یک نمونه خاص، برخی از رکوردهای حقیقی بیش از یک بار ظاهر خواهند شد، در حالی که مشاهدات دیگر از دست رفته‌اند. اگر داده‌های حقیقی شامل حداقل چند صد نقطه باشند، بیرون کشیدن تعداد نامحدود نمونه‌های مختلف در این روش امکان‌پذیر می‌شود. سپس برآورد پارامترها برای هر نمونه انجام می‌شود و همانند روش بوت استراپ پارامتری، توزیع این برآوردها برای به‌دست آوردن حدود تجربی CI استفاده می‌شود. این روش اکیداً پارامتری نیست، چون فرضیات توزیع هنوز برای به‌دست آوردن برآورد پارامترها برای هر نمونه بکار گرفته می‌شوند. این روش در ارتباط با تناقض‌هایی که در فرضیات به‌دست آوردن برآورد پارامترها استفاده می‌شوند قدرتمندتر است.

۱۲-۵- تمرین‌ها

۱۲-۱- تعداد زیادی از داده‌ها برای مکان‌یابی QTL در موجودات زنده مختلف وجود دارد. سعی کنید روی یک نمونه واقعی کار کرده و تمرین کنید که چگونه آستانه بحرانی می‌تواند با استفاده از روش‌های مختلف تعیین شود.

۱۲-۲- با تثبیت یک QTL فرضی در یک موقعیت کروموزومی معین، کائو و زنگ (۱۹۹۷) فرمول‌های کلی برای برآورد MLE های آثار QTL و پارامترهای مزاحم و ماتریس واریانس-کوواریانس مجانبی MLE ها به‌دست آوردند. چن (۲۰۰۵) برآورد همزمان ماتریس واریانس-کوواریانس مجانبی را برای موقعیت و اثر QTL و پارامترهای مزاحم با استفاده از الگوریتم EM انجام داد. تفاوت بین این دو مطالعه در روش برآورد موقعیت QTL است. کائو و زنگ از روش شبکه‌ای استفاده کردند، حال آن که چن شکل دقیق برآورد موقعیت را استفاده کرد. روش چن برآورد مستقیم واریانس نمونه‌برداری موقعیت QTL بود.

اما در استنتاج‌های مربوط به تلاقی برگشتی، چن (۲۰۰۵) از احتمالات ساده، ولی تقریب زده شده و شرطی استفاده کرد (جدول ۱۰-۴). همان گونه که قبلاً مشاهده شد، این شکل تقریبی فقط زمانی عمل می‌کند که فواصل به‌طور منطقی کوچک باشند که این حالت در بسیاری از مطالعات مکان‌یابی عملی مشکل است. اگر روش چن می‌توانست احتمالات شرطی (جدول ۱۰-۳) را نیز مورد استفاده قرار دهد، در این صورت می‌توانست یک روش بسیار خوب برای مکان‌یابی ارائه کند. لطفاً سعی کنید به سئوالات زیر پاسخ دهید:

الف. نشان دهید چگونه MLE مربوط به موقعیت QTL و خطای نمونه‌برداری آن می‌تواند با در نظر گرفتن احتمالات شرطی جدول ۱۰-۳ برآورد شوند.

ب. ماتریس واریانس-کوواریانس مجانبی مربوط به اثر و موقعیت QTL و پارامترهای مزاحم را برای F_2 به‌دست آورید.

ج. چگونه احتمالات شرطی حقیقی F_2 می‌توانند برای برآورد ماتریس واریانس-کوواریانس مجانبی در نظر گرفته شوند.

فصل سیزدهم

مکان‌یابی مرکب QTL

۱۳-۱- مقدمه

روش فاصله‌ای لندر و بوتستین برای مکان‌یابی ژنومی QTL ها با جستجوی موقعیت QTL در سرتاسر ژنوم دارای مزیت است. ولی هنگامی‌که چند QTL روی یک گروه پیوستگی مشابهی وجود دارند این روش می‌تواند منجر به برآوردهای اریب از موقعیت‌ها و آثار QTL شود زیرا در این روش فقط یک فاصله نشانگری در یک زمان استفاده می‌شود و هیچ مکانیسمی برای کاهش اثر QTL های دیگر در خارج از این فاصله ندارد. به این دلیل، اگر یک QTL واقعی در نزدیکی فاصله نشانگری که QTL ندارد مستقر باشد، مکان‌یابی فاصله‌ای ممکن است هنوز یک QTL "خیالی" را به‌واسطه پیوستگی بین QTL واقعی و فاصله‌ای که آزمون می‌شود، تشخیص دهد (مارتینز و کورنو، ۱۹۹۲). اگرچه جستجوی همزمان چندین QTL در فواصل مختلف می‌تواند بر این مشکل غلبه کند، اما این امر مشکلات جدیدی را در زمینه برآورد پارامترها و شناسایی مدل فراهم می‌آورد.

در ارجاع به شکل ۱۰-۱، اگر هدف مکان‌یابی، برآورد اثر QTL در فاصله (M,N) باشد، در این صورت آثار اختلاط نشانگرهای (یا QTL های) مستقر در خارج از این فاصله می‌توانند وجود داشته باشند. یک روش معمول برای حذف آثار زمینه ژنتیکی، تلاش برای از بین بردن این اختلاط با استفاده از متغیرهای همبسته (کوواریت‌ها) یا کوفاکتورها است. این روش، مکان‌یابی فاصله‌ای مرکب نامیده می‌شود که آماره‌های آزمون را با ترکیب مکان‌یابی فاصله‌ای در دو نشانگر مجاور و تجزیه تابعیت چندگانه روی نشانگرهای دیگر ایجاد می‌کند. روش مکان‌یابی فاصله‌ای مرکب که به‌طور مستقل به وسیله زنگ (۱۹۹۳، ۱۹۹۴) و یانسن و استام (۱۹۹۴) پیشنهاد شد، برخی از استنباط‌های مربوط به QTL های خارج از فاصله مورد آزمون را از بین می‌برد. زنگ (۱۹۹۳) به‌طور خاص مزایا و معایب مکان‌یابی فاصله‌ای مرکب را نشان داد.

در این فصل، اصول مکان‌یابی فاصله‌ای مرکب و الگوریتم آن را برای برآورد پارامترهای موقعیت و اثر QTL معرفی می‌کنیم. مزایای مکان‌یابی فاصله‌ای مرکب را در برابر مکان‌یابی فاصله‌ای از راه مطالعات شبیه‌سازی و مثال‌های واقعی مقایسه خواهیم نمود. حوزه‌هایی که در آن‌ها مکان‌یابی فاصله‌ای مرکب می‌تواند خصوصیت کامل ساختار ژنتیکی صفات کمّی را بهبود بخشد نیز ذکر می‌شوند.

۱۳-۲- مکان‌یابی فاصله‌ای مرکب در جمعیت تلاقی برگشتی

تشریح مکان‌یابی فاصله‌ای مرکب را با استفاده از طرح تلاقی برگشتی و والدین هموزیگوتی که حامل آلل‌های نامطلوب هستند، شروع می‌کنیم. فرض کنید که این جمعیت تلاقی برگشتی دارای اندازه n است که در آن تعدادی از نشانگرها $(M_1,...,M_m)$ برای تهیه نقشه پیوستگی تعیین

ژنوتیپ شده‌اند. در یک QTL فرضی که با یک جفت نشانگر خاص M_t و M_{t+1} احاطه شده است، دو کلاس ژنوتیپی، هتروزیگوت Qq(1) و هموزیگوت qq(0) در تلاقی برگشتی وجود دارند.

یک مدل خطی برای فاصله تحت آزمون که با دو نشانگر مجاور احاطه شده و ترکیب آن با نشانگرهای دیگر می‌تواند بدین صورت بیان شود:

$$y_i = \mu + \left(x_{1|i}^* - x_{0|i}^*\right)a + \sum_{k=1}^{m-2} b_k x_{ki} + e_i \qquad (۱-۱۳)$$

که در آن، y_i صفت فنوتیپی i امین نتاج در تلاقی برگشتی، μ میانگین کل، a اثر ژنتیکی افزایشی QTL، $x_{1|i}^*$ و $x_{0|i}^*$ متغیرهای شاخص و موهومی هستند که ژنوتیپ QTL (Qq یا qq) فرزند i ام را در موقعیت مربوطه نشان می‌دهند و بر اساس اطلاعات فاصله $M_t - M_{t+1}$ تعریف می‌شوند، x_{ki} ژنوتیپ فرزند i ام برای نشانگر k ام واقع در خارج فاصله است که به صورت ۱ برای نشانگر هتروزیگوت و ۰ برای نشانگر هموزیگوت است، $b_1,\ldots,b_{m-2}$ ضرایب تابعیت جزئی (آثار افزایشی نشانگر) هستند که برآورد می‌شوند و e_i خطای باقیمانده است که آثار اعمال شده به وسیله QTL های دیگر و عوامل غیر ژنتیکی را شامل می‌شود و $e_i \sim N\left(0,\sigma^2\right)$ است.

این روش در اصل مکان‌یابی فاصله‌ای است، ولی آثار خارج از فاصله را نیز کنترل می‌کند؛ بنابراین، اساس این روش با موارد ذکر شده در بخش ۱۱-۲ بسیار شباهت دارد، البته به استثنای اینکه، ضرورت برآورد آثار اضافی موجب پیچیده‌تر شدن این روش شده است.

۱۳-۲-۱- درستنمایی

در ادامه بحث ارائه شده در بخش ۱۱-۲ و مشابه معادله (۱۱-۳)، تابع درستنمایی را به‌صورت زیر خواهیم داشت:

$$L\left(a,b,\sigma^2,\omega_{1|i},\omega_{0|i} \mid y\right) = \prod_{i=1}^{n}\left[\omega_{1|i} f_1\left(y_i \mid a,b,\sigma^2\right) + \omega_{0|i} f_0\left(y_i \mid b,\sigma^2\right)\right] \qquad (۲-۱۳)$$

یا:

$$L\left(a,b,\sigma^2,\theta \mid y\right)=\prod_{i=1}^{n_1} f_1\left(y_i \mid a,b,\sigma^2\right)$$
$$\times\prod_{i=1}^{n_2}\left[(1-\theta)\, f_1\left(y_i \mid a,b,\sigma^2\right)+\theta\, f_0\left(y_i \mid b,\sigma^2\right)\right]$$
$$\times\prod_{i=1}^{n_3}\left[\theta\, f_1\left(y_i \mid a,b,\sigma^2\right)+(1-\theta)\, f_0\left(y_i \mid b,\sigma^2\right)\right] \qquad (۳-۱۳)$$
$$\times\prod_{i=1}^{n_4} f_0\left(y_i \mid b,\sigma^2\right)$$

مدل مختلط مبتنی بر درستنمایی در معادله (۱۳-۲)، به ترتیب با شکل کلی نسبت‌های اختلاط یعنی $\omega_{1|i}$ و $\omega_{0|i}$ و احتمالات شرطی ژنوتیپ‌های QTL یعنی Qq و qq مشروط به ژنوتیپ فاصله نشانگری فرزند i ام اختصاص یافته است. احتمالات شرطی در تابع درستنمایی معادله (۱۳-۳) شکل ساده‌ای است که فرض می‌کند هیچ نوترکیبی مضاعفی وجود ندارد (جدول ۱۰-۴).

تابع چگالی احتمال نرمال ارزش صفت در هر کلاس ژنوتیپ QTL بدین صورت تعریف می‌شود:

$$f_1\left(y_i \mid a,b,\sigma^2\right)=N\left(a+X_i b^T,\sigma^2\right)$$
$$f_0\left(y_i \mid b,\sigma^2\right)=N\left(X_i b^T,\sigma^2\right)$$

زیرا:

$Qq\left(x_i=1\right)$ برای: $y_i=\mu+a+\sum_{k=1}^{m-2} b_k x_{ki}+e_i=a+X_i b^T+e_i$

$qq\left(x_i=0\right)$ برای: $y_i=\mu+\sum_{k=1}^{m-2} b_k x_{ki}+e_i=X_i b^T+e_i$

که در آن‌ها، $b=\left(\mu,b_1,\ldots,b_{m-2}\right)$ و $X_i=\left(1,x_{1i},x_{2i},\ldots,x_{(m-2)i}\right)$ است.

۱۳-۲-۲- به حداکثر رساندن درستنمایی

پارامترهای موجود در درستنمایی (۱۳-۲) با به‌کارگیری الگوریتم EM برآورد می‌شوند. این امر می‌تواند با استفاده از روشی مشابه با روشی که در بخش ۱۱-۲-۲ توصیف شد، انجام شود. در مرحله E، این احتمال که فرزند i ام ژنوتیپ معینی از QTL را مشروط به ژنوتیپ‌های نشانگر و ارزش فنوتیپی آن به‌صورت زیر حمل می‌کند، تعریف می‌کنیم:

$$P_{1|i}=\frac{\omega_{1|i}\,f_1\left(y_i\mid a,b,\sigma^2\right)}{\omega_{1|i}\,f_1\left(y_i\mid a,b,\sigma^2\right)+\omega_{0|i}\,f_0\left(y_i\mid b,\sigma^2\right)}$$
$$P_{0|i}=\frac{\omega_{0|i}\,f_0\left(y_i\mid b,\sigma^2\right)}{\omega_{1|i}\,f_1\left(y_i\mid a,b,\sigma^2\right)+\omega_{0|i}\,f_0\left(y_i\mid b,\sigma^2\right)} \qquad (۴-۱۳)$$

در مرحله M، پارامترها بر حسب احتمالات پسین با مشتق‌گیری از لگاریتم درستنمایی (۱۳-۲) نسبت به هر پارامتر و برابر صفر قرار دادن مشتق‌ها و حل کردن معادلات لگاریتم درستنمایی برآورد می‌شوند. این روش به‌صورت زیر توضیح داده می‌شود:

برای a داریم:

$$\frac{\partial}{\partial a}\ln L\left(a,b,\sigma^2,\theta\mid y\right)\propto\sum_{i=1}^{n}P_{1|i}\left(\frac{y_i-a-X_ib^T}{\sigma^2}\right)=0$$

که منجر به عبارت زیر می‌شود:

$$\hat{a}=\frac{\sum_{i=1}^{n}\left(y_i-X_ib^T\right)P_{1|i}}{\sum_{i=1}^{n}P_{1|i}}$$
$$=\frac{1}{c}\left(y-Xb^T\right)^T P_1 \qquad (۵-۱۳)$$

در این رابطه، $y=\{y_i\}_{n\times1}$، $X=\{X_i\}_{n\times1}$، $P_1=\{P_{1|i}\}_{n\times1}$ و $c=\sum_{i=1}^{n}P_{1|i}$.

برای b و σ^2، داریم:

$$\frac{\partial}{\partial b}\ln L\left(a,b,\sigma^2,\theta\mid y\right)\propto\sum_{i=1}^{n}\left[P_{1|i}X_i\left(y_i-a-X_ib^T\right)+P_{0|i}X_i\left(y_i-X_ib^T\right)\right]\frac{1}{\sigma^2}=0$$

و:

$$\frac{\partial}{\partial\sigma^2}\ln L\left(a,b,\sigma^2,\theta\mid y\right)\propto\sum_{i=1}^{n}\left[P_{1|i}X_i\left(y_i-a-X_ib^T\right)^2+P_{0|i}X_i\left(y_i-X_ib^T\right)^2\right]\frac{1}{2\sigma^4}-\frac{n}{2\sigma^2}=0$$

در نتیجه داریم:

$$\hat{b}=\left(X^TX\right)^{-1}X^T\left(y-P\hat{b}\right) \qquad (۶-۱۳)$$

$$\hat{\sigma}^2=\left(y-X\hat{b}^T\right)^T\left(y-Xb^T\right)-\hat{a}^2c \qquad (۷-۱۳)$$

برای برآورد موقعیت θ مربوط به QTL، دو روش می‌تواند استفاده شود که یک روش بر اساس شکل بسته آن نظیر معادله (۱۱-۷)، حاصل از درستنمایی (۱۳-۳) و روش دوم بر اساس

جستجوی شبکه‌ای تعدادی از موقعیت‌های ثابت در فاصله نشانگری مربوطه است. مشروط به مقادیر ابتدایی پارامترهای ناشناخته $\left(a,b,\sigma^2\right)$، مراحل E و M به ترتیب با استفاده از معادلات (۱۳-۴) و (۱۳-۵) تا (۱۳-۷) محاسبه می‌شوند و تا جایی که برآوردها ثابت و پایدار شوند، تکرار می‌شوند.

۱۳-۲-۳- آزمون فرض

معنی‌داری اثر ژنتیکی QTL شناسایی شده به وسیله مکان‌یابی فاصله‌ای مرکب می‌تواند با فرموله نمودن فرض‌های زیر آزمون شود:

$H_0: a=0$ در مقابل $H_1: a\neq 0$.

مقدار درستنمایی تحت فرض مقابل با قرار دادن MLE های پارامترها در درستنمایی (۱۳-۲) محاسبه می‌شود. درستنمایی با در نظر گرفتن فرض صفر به‌صورت در نظر گرفته می‌شود:

$$L\left(b,\sigma^2 \mid y\right)=\prod_{i=1}^{n} f\left(y_i \mid b,\sigma^2\right)$$

که از روی آن، MLE ها به صورت زیر به‌دست می‌آیند:

$$\tilde{b}=\left(X^T X\right)^{-1} X^T Y$$

$$\sigma^2=\frac{1}{n}\left(y-X\tilde{b}^T\right)^T\left(y-X\tilde{b}^T\right)$$

از این‌رو، آماره آزمون درستنمایی با استفاده از معادله زیر محاسبه می‌شود:

$$LR=-2\ln\left[\frac{L\left(a=0,\tilde{b},\tilde{\sigma}^2 \mid y\right)}{L\left(\hat{a},\hat{b},\hat{\sigma}^2 \mid y,M\right)}\right] \qquad (۱۳-۸)$$

توجه شود که محاسبه درستنمایی با در نظر گرفتن فرض مقابل (H_1) مستلزم داشتن اطلاعات نشانگر (M) است، در حالی که با در نظر گرفتن فرض صفر به اطلاعات نشانگر نیازی نیست. آستانه بحرانی برای اعلام وجود QTL با استفاده از روش‌های پیشنهاد شده در فصل ۱۲ تعیین می‌شود. در عمل، روش‌های تجربی بر اساس آزمون‌های تبدیل یا جایگشت، اگرچه از نظر محاسباتی مشکل و پیچیده هستند، اما به‌طور وسیعی استفاده می‌شوند، زیرا بر توزیع آماره‌های آزمون تأکیدی ندارند.

۱۳-۳- مکان‌یابی فاصله‌ای مرکب در جمعیت F_2

روش مشابهی برای مکان‌یابی فاصله‌ای مرکب می‌تواند برای F_2 ها که در آن سه ژنوتیپ QTL به ترتیب به صورت QQ(2)، Qq(1) و qq(0) نشان داده می‌شوند، ایجاد شده و بسط یابد. مدل خطی ارزش فنوتیپی فرزند i ام برای فاصله آزمون شده $M_t - M_{t+1}$، مشابه معادله (۱۳-۱) که به‌طور روشن و واضح آثار افزایشی و غالبیت را مدل‌ریزی می‌کند، می‌تواند به‌صورت زیر نوشته شود:

$$y_i = \mu + \left(x_{2|i}^* - x_{0|i}^*\right)a + x_{1|i}^* d + \sum_{k=1}^{m-2} x_{ki} b_k + \sum_{k=1}^{m-2} z_{ki} h_k + e_i \qquad (۹-۱۳)$$

که $x_{2|i}^*$، $x_{1|i}^*$ و $x_{0|i}^*$ متغیرهای شاخص و موهومی برای ژنوتیپ‌های QTL در نتاج i ام هستند و با ژنوتیپ فاصله نشانگری مربوطه مشخص می‌شوند، a و d آثار افزایشی و غالبیت QTL، x_{ki} و z_{ki} نیز متغیرهای موهومی برای ژنوتیپ‌های نشانگر M_k هستند و داریم:

$$\left.\begin{array}{ll} \text{برای } M_kM_k & \frac{1}{2} \\ \text{برای } M_km_k & 0 \\ \text{برای } m_km_k & -\frac{1}{2} \end{array}\right\} = x_{ki}$$

و:

$$\left.\begin{array}{ll} \text{برای } M_km_k & 1 \\ \text{برای دیگران} & 0 \end{array}\right\} = z_{ki}$$

و b_k و h_k آثار افزایشی و غالبیت مربوط به نشانگر M_k هستند.

تابع درستنمایی داده‌های فنوتیپی مشاهده شده (y) و داده‌های نشانگر (M) در F_2 به‌صورت زیر است:

$$L\left(\Omega \mid y, M\right) = \prod_{i=1}^{n} \left[\omega_{2|i} f_2\left(y_i\right) + \omega_{1|i} f_1\left(y_i\right) + \omega_{0|i} f_0\left(y_i\right)\right] \qquad (۱۰-۱۳)$$

که $\omega_{j|i}$ ها ($j = 2,1,0$) وزن‌های F_2 برای هر ژنوتیپ QTL مشروط به ژنوتیپ نشانگر در نتاج i ام بوده و به وسیله موقعیت QTL در فاصله نشانگری تعیین می‌شوند (جدول ۱۱-۷ را مشاهده کنید) و توزیع‌های نرمال به‌صورت زیر ارائه می‌شوند:

$$f_2(y_i)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\frac{\left(y_i-X_ib^T+a\right)^2}{2\sigma^2}\right]$$

$$f_1(y_i)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\frac{\left(y_i-X_ib^T+d\right)^2}{2\sigma^2}\right]$$

$$f_0(y_i)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\frac{\left(y_i-X_ib^T-a\right)^2}{2\sigma^2}\right]$$

و داریم:

$$X_ib^T=\mu+\sum_{k=1}^{m-2}x_{ki}b_k+\sum_{k=1}^{m-2}z_{ki}h_k$$

$$X_i=\left(1,x_{1i},\ldots,x_{(m-2)i},z_{1i},\ldots,z_{(m-2)i}\right)$$

$$b=\left(\mu,b_1,\ldots,b_{m-2},h_1,\ldots,h_{m-2}\right)$$

تابع درستنمایی (۱۳-۱۰) حاوی یک بردار ناشناخته $\Omega=\left(a,d,b,\sigma^2,\omega_{j|i}\right)$ است. همانند جمعیت تلاقی برگشتی، الگوریتم EM می‌تواند استفاده شود تا MLE های پارامترهایی (Ω) که درستنمایی (۱۳-۱۰) را به حداکثر می‌رسانند، محاسبه شوند.

در مرحله E، احتمال پسین QTL j ام برای نتاج i ام به‌صورت زیر تعریف می‌شود:

$$P_{j|i}=\frac{\omega_{j|i}f_j(y_i)}{\sum_{j'=0}^{2}\left[\omega_{j'|i}f_{j'}(y_i)\right]} \qquad (۱۳-۱۱)$$

که $P_{2|i}+P_{1|i}+P_{0|i}=1$ است. از این‌رو، در مرحله M، MLE های پارامترهای ناشناخته به‌صورت زیر به‌دست می‌آیند:

$$\hat{a} = \frac{\left(y - X\hat{b}^T\right)^T P_2}{2 1^T P_2}$$

$$\hat{d} = \frac{\left(y - X\hat{b}^T\right)^T P_1}{1^T P_1} - \hat{a} \qquad (۱۲-۱۳)$$

$$\hat{b} = \left(X^T X\right)^{-1} X^T \left[y - \left(2P_2 + P_1\right)\hat{a} - P_1\hat{d}\right]$$

$$\hat{\sigma}^2 = \frac{1}{n}\left[\left(Y - X\hat{b}^T\right)^T \left(y - X\hat{b}^T\right) - 4\left(1^T P_2\right)\hat{a}^2 - \left(1^T P_2\right)\left(a+d\right)^2\right]$$

که در آن، $P_2 = \left(P_{2|1}, ..., P_{2|n}\right)^T$ و $P_1 = \left(P_{1|1}, ..., P_{1|n}\right)^T$ هستند.

معادلات (۱۳-۱۱) و (۱۳-۱۲) یک روش تکراری برای حل کردن MLE های پارامترهای ناشناخته ارائه می‌دهند.

برای جمعیت F_2، آزمون های فرض شامل سه مرحله هستند. اولاً، حضور QTL می‌تواند بر اساس فرض‌های زیر آزمون شود:

$H_0 : a = d = 0$ در مقابل H_1: حداقل یکی از آن‌ها برابر صفر نیست (۱۳-۱۳)

ثانیاً، معنی‌داری اثر افزایشی QTL (a) می‌تواند به‌صورت زیر آزمون شود:

$H_0 : a = 0$ در مقابل $H_1 : a \neq 0$ (۱۳-۱۴)

ثالثاً، اثر غالبیت QTL (d) با استفاده از رابطه زیر آزمون می‌شود:

$H_0 : d = 0$ در مقابل $H_1 : d \neq 0$ (۱۳-۱۵)

در هر حالت، نسبت‌های لگاریتم درستنمایی با در نظر گرفتن فرض‌های صفر و مقابل محاسبه می‌شوند. آستانه بحرانی برای اولین فرض (۱۳-۱۳) می‌تواند به‌صورت تجربی از روی آزمون‌های جایگشت تعیین شود. آستانه بحرانی برای فرض‌های دوم (۱۳-۱۴) و سوم (۱۳-۱۵) می‌تواند به‌صورت تجربی به وسیله مطالعات شبیه‌سازی تعیین شوند. در حقیقت، فرض‌های (۱۳-۱۴) و (۱۳-۱۵) دارای پارامترهای غیر قابل شناسایی در فرض صفر آن‌ها هستند و بنابراین، نسبت‌های لگاریتم درستنمایی هر یک از آن‌ها می‌تواند از توزیع کای مربع با یک درجه آزادی تبعیت کند.

۱۳-۴- توجیه آماری مکان‌یابی فاصله‌ای ترکیبی

زنگ (۱۹۹۳) نشان داد که تحت فرض عدم وجود کراس‌اور و عدم اپیستازی، ضریب تابعیت جزیی حقیقی (پارامتر واقعی) یک صفت روی نشانگر فقط به QTL هایی بستگی دارد که روی فاصله احاطه شده با دو نشانگر مجاور بدون وابستگی به هر QTL دیگر قرار گرفته‌اند.

در حالت ایده‌آل، فرض می‌شود که آماره آزمون تهیه شده از یک فاصله نشانگر، یک QTL را حمل می‌کند که مستقل از آثار هر QTL ممکن دیگر در منطقه کروموزومی خارج از این فاصله است. این ارتباط در مدل مکان‌یابی فاصله‌ای مرکب مبتنی بر تابعیت، به‌عنوان یک خصوصیت از پارامترها محسوب می‌شود. بیشتر مباحث مربوط به اصول آماری مکان‌یابی فاصله‌ای مرکب از مطالعات زنگ (۱۹۹۳) ارایه می‌شود.

۱۳-۴-۱- (کو) واریانس‌های شرطی نشانگر

فرض کنید که دو نشانگر M_j، با دو آلل M_j و m_j و M_k، با دو آلل M_k و m_k، برای یک جمعیت تلاقی برگشتی تعیین ژنوتیپ شده و نرخ نوترکیبی بین این دو نشانگر r_{jk} است. مقادیر ۱ و ۰ را به ترتیب برای هتروزیگوت‌ها و هموزیگوت‌های هر نشانگر در جمعیت تلاقی برگشتی نشان می‌دهیم؛ بنابراین، فراوانی‌ها و ارزش‌های دو ژنوتیپ جمعیت تلاقی برگشتی به صورت زیر بیان می‌شوند:

	$M_jm_jM_km_k$	$M_jm_jm_km_k$	$m_jm_jM_km_k$	$m_jm_jm_km_k$
فراوانی	$\frac{1}{2}(1-r_{jk})$	$\frac{1}{2}r_{jk}$	$\frac{1}{2}r_{jk}$	$\frac{1}{2}(1-r_{jk})$
ارزش در M_1	1	1	0	0
ارزش در M_2	1	0	1	0

واریانس هر نشانگر می‌تواند به صورت زیر محاسبه شود:

(۱۶-۱۳)

$$\sigma_j^2 = \sigma_k^2 = \frac{1}{4}$$

و کوواریانس بین دو نشانگر برابر است با:

(۱۷-۱۳)

$$\sigma_{jk} = \frac{1}{4}(1-2r_{jk})$$

که در آن، $1-2r_{ij}$ همبستگی بین دو نشانگر است. بر اساس معادلات (۱۳-۱۶) و (۱۳-۱۷)، واریانس شرطی نشانگر M_k مشروط به نشانگر M_j به صورت زیر است:

$$\sigma_{k|j}^2 = \sigma_k^2 - \frac{\sigma_{kj}^2}{\sigma_j^2}$$
$$= \frac{1}{4} - \frac{\left[\frac{1}{4}\left(1-2r_{jk}\right)\right]^2}{4} \qquad (۱۸-۱۳)$$
$$= r_{jk}\left(1-r_{jk}\right)$$

با در نظر گرفتن سه نشانگر M_j، M_k و M_l، به‌طور مشابه کوواریانس بین نشانگرهای M_j و M_k مشروط به نشانگر M_l به صورت زیر به‌دست می‌آید:

$$\sigma_{jk|l} = \sigma_{jk} - \frac{\sigma_{jl}\sigma_{kl}}{\sigma_l^2}$$
$$= \frac{1}{4}\left[\left(1-2r_{jk}\right)-\left(1-2r_{jl}\right)\left(1-2r_{kl}\right)\right] \qquad (۱۹-۱۳)$$
$$= \begin{cases} 0 & \text{برای ترتیب } M_jM_lM_k \text{ یا } M_kM_lM_j \\ r_{kl}\left(1-r_{kl}\right)\left(1-2r_{jk}\right) & \text{برای ترتیب } M_jM_kM_l \text{ یا } M_lM_kM_j \\ r_{jl}\left(1-r_{jl}\right)\left(1-2r_{jk}\right) & \text{برای ترتیب } M_lM_jM_k \text{ یا } M_kM_jM_l \end{cases}$$

زیرا بدون استنباط، برای ترتیب $M_iM_lM_k$ یا $M_kM_lM_i$، $\left(1-2r_{ik}\right)=\left(1-2r_{il}\right)\left(1-2r_{kl}\right)$ است. معادله (۱۳-۱۹) بیان می‌کند که مشروط به نشانگر حدواسط، کوواریانس بین دو نشانگر مجاور، چنانچه انتظار می‌رود برابر با صفر است.

با استفاده از معادلات (۱۳-۱۸) و (۱۳-۱۹)، واریانس نشانگر M_j مشروط به نشانگرهای M_k و M_l به‌صورت زیر است:

$$\sigma^2_{j|kl} = \sigma^2_{j|k} - \frac{\sigma^2_{jl|k}}{\sigma^2_{l|k}}$$

$$= \sigma^2_{j|k} - \frac{\sigma^2_{jk|l}}{\sigma^2_{k|l}} \qquad (۲۰-۱۳)$$

$$= \begin{cases} \sigma^2_{j|k} & \text{برای ترتیب } M_jM_kM_l \text{ یا } M_lM_kM_j \\ \sigma^2_{k|j} & \text{برای ترتیب } M_jM_lM_k \text{ یا } M_kM_lM_j \\ \dfrac{r_{jk}\left(1-r_{jk}\right)r_{jl}\left(1-r_{jl}\right)}{r_{kl}\left(1-r_{kk}\right)} & \text{برای ترتیب } M_lM_jM_k \text{ یا } M_kM_jM_l \end{cases}$$

بر طبق معادله (۱۳-۲۰)، واریانس شرطی یک نشانگر مشروط به تمامی نشانگرهای دیگر، فقط به نشانگرهایی بستگی دارد که دقیقاً بعد از نشانگر مورد نظر قرار گرفته باشند. به‌طورکلی، این حالت می‌تواند بدین صورت بیان شود. واریانس نشانگر M_j مشروط به تمامی نشانگرهای دیگر برابر است با:

$$\sigma^2_{j|s_{-j}} = \sigma^2_{j|(j-1)(j+1)}$$

که در آن، s_{-j} مجموعه‌ای شامل تمامی نشانگرها به استثنای نشانگر M_j است.

به‌طور مشابه، با در نظر گرفتن چهار نشانگر به صورت $M_j < M_k$ و $M_l < M_m$، کوواریانس بین نشانگرهای M_j و M_k مشروط به M_l و M_m به صورت زیر به‌دست می‌آید:

$$\sigma^2_{jk|lm} = \sigma_{jk|l} - \frac{\sigma_{jm|l}\sigma_{km|l}}{\sigma^2_{m|l}}$$

$$= \sigma_{jk|m} - \frac{\sigma_{jl|m}\sigma_{kl|m}}{\sigma^2_{l|m}} \qquad (۲۱-۱۳)$$

$$= \begin{cases} 0 & \text{برای ترتیب } M_jM_mM_kM_l \text{ یا } M_jM_lM_kM_m \\ \sigma^2_{jk|l} & \text{برای ترتیب } M_jM_kM_lM_m \\ \sigma^2_{jk|m} & \text{برای ترتیب } M_lM_mM_jM_k \\ \dfrac{r_{lk}(1-r_{lk})r_{km}(1-r_{km})(1-2r_{jk})}{r_{lm}(1-r_{lm})} & \text{برای ترتیب } M_lM_jM_kM_m \end{cases}$$

معادله (۱۳-۲۱) حاوی دو نتیجه مهم است. اولاً، وقتی که روی یک نشانگر حد واسط شرط می‌گذاریم، کوواریانس بین هر دو نشانگر صفر خواهد بود، ثانیاً، کوواریانس بین دو نشانگر مجاور مشروط به تمامی نشانگرهای دیگر، فقط به دو نشانگری بستگی دارد که به هر یک از نشانگرهای مجاور نزدیک‌تر است؛ به عبارت دیگر، مشروط کردن به دو نشانگر مجاور، کوواریانس بین دو نشانگر درون فاصله را مستقل از تمامی نشانگرهای خارج فاصله نشانگری خواهد ساخت. از لحاظ ریاضی، این حالت می‌تواند به‌صورت زیر بیان شود:

$$\sigma^2_{jk|s_{-jk}} = \sigma^2_{jk|(j-1)(k+1)}$$

که در آن، s_{-jk} مجموعه‌ای شامل تمامی نشانگرها به استثنای نشانگرهای M_j و M_k است.

۱۳-۴-۲- واریانس شرطی QTL

در این بخش، واریانس شرطی نشانگر را بسط می‌دهیم تا واریانس شرطی QTL مورد نظر را به‌دست آوریم. نشانگر M_j و QTL(Q_u) را با دو آلل Q_u و q_u در نظر می‌گیریم که ژنوتیپ‌ها، فراوانی‌ها و ارزش‌های آن‌ها در یک جمعیت تلاقی برگشتی به‌صورت زیر هستند:

	$M_jm_jQ_uq_u$	$M_jm_jq_uq_u$	$m_jm_jQ_uq_u$	$m_jm_jq_uq_u$
فراوانی	$\frac{1}{2}(1-r_{ju})$	$\frac{1}{2}r_{ju}$	$\frac{1}{2}r_{ju}$	$\frac{1}{2}(1-r_{ju})$
ارزش برای M_j	1	1	0	0
ارزش برای Q_u	a	0	a	0

که در آن، r_{ju} نرخ نوترکیبی بین نشانگر و QTL و a_u اثر افزایشی Q_u است. به سادگی می‌توان نشان داد که: $\sigma_u^2 = \frac{1}{4}a^2$ و $\sigma_{ju} = \frac{1}{4}(1-2r_{ju})\sigma_u^2$.

با در نظر گرفتن صفت کمّی y که با U عدد QTL بدون اثر متقابل اپیستازی کنترل می‌شود، کوواریانس بین صفت y و نشانگر M_j مشروط به نشانگر M_k به‌صورت زیر است:

$$\sigma_{yj|k} = \sigma_{yj} - \frac{\sigma_{yk}\sigma_{jk}}{\sigma_k^2}$$

$$= \frac{1}{4}\sum_{u=1}^{U}\left[(1-2r_{uj}) - (1-2r_{uk})(1-2r_{jk})\right]a_u \qquad (۲۲-۱۳)$$

$$= \begin{cases} r_{jk}(1-r_{jk})\sum_{u\le j}(1-2r_{uj})a_u + \sum_{j<u<k} r_{uk}(1-r_{uk})(1-2r_{ju})a_u & \text{برای ترتیب } M_jM_k \\ r_{jk}(1-r_{jk})\sum_{u\ge j}(1-2r_{uj})a_u + \sum_{j<u<k} r_{uk}(1-r_{uk})(1-2r_{ju})a_u & \text{برای ترتیب } M_kM_j \end{cases}$$

که نشان می‌دهد کوواریانس بین یک نشانگر و صفت مشروط به نشانگر دوم، شامل آن QTL هایی که در آن سوی نشانگر شرطی قرار دارند، نیست. برای روشن شدن بیشتر، اجازه دهید دو نشانگر شرطی را در نظر بگیریم. به آسانی می‌توان عبارت زیر را به‌دست آورد:

(۱۳-۲۳)

$$\sigma^2_{yj|kl} = \sigma_{yj|k} - \frac{\sigma_{yl|k}\sigma_{jl|k}}{\sigma^2_{l|k}}$$

$$= \sigma_{yj|l} - \frac{\sigma_{yk|l}\sigma_{jk|l}}{\sigma^2_{l|m}}$$

$$= \begin{cases} \sigma_{yj|k} & \text{برای ترتیب } M_jM_kM_l \\ \sigma_{yj|l} & \text{برای ترتیب } M_jM_lM_k \\ \dfrac{r_{jk}\left(1-r_{jk}\right)}{r_{kl}\left(1-r_{kl}\right)} \sum\limits_{l<u\le j} r_{lu}\left(1-t_{lu}\right)\left(1-2r_{uj}\right)a_u + \\ \dfrac{r_{jk}\left(1-r_{jk}\right)}{r_{kl}\left(1-r_{kl}\right)} \sum\limits_{j<u\le k} r_{lu}\left(1-t_{lu}\right)\left(1-2r_{uj}\right)a_u & \text{برای ترتیب } M_lM_jM_k \end{cases}$$

عبارت کلی معادله (۱۳-۲۳) به صورت $\sigma_{yj|s_{-j}} = \sigma_{yj|(j-1)(j+1)}$ است. بر اساس فرمول‌های بالا، ضریب تابعیت جزئی ارزش صفت روی نشانگر M_j مشروط به باقیمانده نشانگرها به‌صورت زیر به‌دست می‌آید:

$$b_{yj|s_{-j}} = \frac{\sigma_{yj|s_{-j}}}{\sigma^2_{j|s_{-j}}}$$

$$= \frac{\sigma_{yj|(j-1)(j+1)}}{\sigma^2_{j|(j-1)(j+1)}}$$

$$= \sum_{j-1<u\le j} \frac{r_{(j-1)u}\left(1-r_{(j-1)u}\right)\left(1-2r_{uj}\right)}{r_{(j-1)j}\left(1-r_{(j-1)j}\right)} a_u \qquad (۱۳-۲۴)$$

$$+ \sum_{j<u<j+1} \frac{r_{u(j-1)}\left(1-r_{u(j-1)}\right)\left(1-2r_{ju}\right)}{r_{j(j-1)}\left(1-r_{j(j-1)}\right)}$$

که دارای دو عبارت جمع برای تمامی QTL های مستقر بین نشانگرهای M_{j-1} و M_j و برای تمامی QTL های مستقر بین نشانگرهای M_j و M_{j+1} است. این مطلب نشان می‌دهد که ضریب تابعیت جزئی فقط به QTL هایی بستگی دارد که بین نشانگرهای M_{j-1} و M_{j+1} قرار دارند، اگرچه این حالت برای وقتی که QTL ها آثار متقابل اپیستاتیک قوی از خود نشان

می‌دهند، کاربرد ندارد. در هر حالتی که اپیستازی وجود ندارد، اگر مکان‌یابی فاصله‌ای با یک جفت نشانگر مجاور با تجزیه تابعیت چندگانه نشانگرهای دیگر ترکیب شود، تمامی QTL های ممکن در خارج از فاصله مورد آزمون در ضرایب تابعیت چندگانه وجود خواهند داشت. به این دلیل، ترکیب مکان‌یابی فاصله‌ای و تجزیه تابعیت می‌تواند آماره‌های آزمون را برای تعیین موقعیت QTL در درون یک فاصله نشانگری مورد نظر محدود کند و آن را از آثار QTL های ممکن دیگر موجود در خارج از فاصله آزمون شده مستقل سازد. این امر یک اصل آماری برای مکان‌یابی فاصله‌ای مرکب ایجاد کرده است.

با نتیجه‌گیری‌های بعدی، زنگ (۱۹۹۳) نشان داد که واریانس نمونه‌گیری ضریب تابعیت جزیی می‌تواند کاهش یابد و بنابراین توان آماری مکان‌یابی QTL وقتی که این امر مشروط به نشانگرهای غیرپیوسته باشد، می‌تواند افزایش یابد. بر اساس تجزیه واریانس نمونه‌برداری ضریب تابعیت جزیی، ترکیب مکان‌یابی فاصله‌ای و تجزیه تابعیت چندگانه می‌تواند اثر چند QTL به هم پیوسته را بر آزمون فرض و بنابراین دقت آزمون و برآورد QTL وقتی که نشانگرهای مشروط به یکدیگر پیوسته هستند کاهش دهد، اگرچه این امر احتمالاً توان آماری را کاهش می‌دهد. نسبت بین توان و دقت مکان‌یابی QTL می‌تواند برای مسائل عملی متوازن شود. اگر فردی بخواهد چند QTL را در یک ناحیه ژنومی جدا نماید، باید نشانگرهای به هم پیوسته را برای تجزیه تابعیت انتخاب نماید. از سوی دیگر، اگر فردی امیدوار باشد که توان مکان‌یابی QTL را افزایش دهد، باید تجزیه تابعیت را با استفاده از نشانگرهای غیر پیوسته انجام دهد.

زنگ (۱۹۹۳) شکل همبستگی نمونه‌برداری بین دو ضریب تابعیت جزئی را روی نشانگرهای مختلف به‌دست آورد. وی نشان داد که چنین همبستگی به‌طورکلی صفر است، مگر آنکه دو نشانگر مجاور هم باشند. این خصوصیت به محاسبه همبستگی آماره‌های آزمون بین دو موقعیت مورد نظر در دو فاصله برای یک آزمون فاصله‌ای و تعیین آستانه معنی‌دار آماره آزمون با در نظر گرفتن فرض صفر برای انجام یک آزمون کلی که کل ژنوم را پوشش می‌دهد، کمک می‌کند.

زنگ (۱۹۹۳) تصور کرد که جنبه‌های نظری مکان‌یابی فاصله‌ای مرکب حاصل از تلاقی برگشتی برای جمعیت F_2 نیز کاربرد دارند. برای F_2 بدون وجود غالبیت، اصول نظریه از نظر کیفی معادل تلاقی برگشتی است. ولی در صورت وجود غالبیت در F_2، نیاز به معرفی یک پارامتر اضافی در تجزیه تابعیت جهت منعکس کردن انحراف غالبیت است.

۱۳-۴-۳- انتخاب نشانگر

مکان‌یابی فاصله‌ای مرکب، تلفیق مکان‌یابی فاصله‌ای و تجزیه تابعیت چندگانه است. انتخاب نشانگرها به‌عنوان کوفاکتورها بحث مهمی است. مشروط کردن به نشانگرهای به هم پیوسته، به

جداسازی QTL های به هم پیوسته چندگانه روی یک منطقه کمک می کند، اما توان آماری آزمون را کاهش می‌دهد. اگرچه مشروط کردن به نشانگرهای ناپیوسته می‌تواند سبب افزایش توان تجزیه از راه کاهش واریانس باقیمانده شود، اما در خصوص جداسازی QTL های پیوسته نمی‌توان کار زیادی انجام داد. این خصوصیات مکان‌یابی فاصله‌ای مرکب، انتخاب نشانگر را مشکل می‌سازد.
متأسفانه، هیچ‌گونه انحراف نظری برای انتخاب نشانگرها به‌عنوان کوفاکتورها وجود ندارد. زنگ (۱۹۹۴) تجزیه تابعیت گام به گام پیشرفتی و پسرفتی را برای افزودن یا کاستن نشانگرها پیشنهاد نمود. همانگونه که در مطالب بالا بحث شد، انتخاب نشانگر باید به صورت تجربی بر پایه هدف مطالعه انجام شود. مکان‌یابی با وضوح بالای QTL های پیوسته به نفع وارد نمودن نشانگرهای پیوسته بیشتر است، حال آنکه توان آماری می‌تواند از راه مشروط کردن به نشانگرهای ناپیوسته بیشتر افزایش یابد.

۱۳-۵- مقایسه بین مکان‌یابی فاصله‌ای مرکب و مکان‌یابی فاصله‌ای

زنگ (۱۹۹۴) مطالعات شبیه‌سازی گسترده‌ای را برای مقایسه مزایا و معایب مکان‌یابی فاصله‌ای مرکب و مکان‌یابی فاصله‌ای انجام داد. وی چهار کروموزوم با طول مساوی شبیه‌سازی کرد. ده QTL با آثار متفاوت برای یک صفت کمّی شبیه‌سازی شدند. مکان‌یابی فاصله‌ای مرکب مشروط به نشانگرهای پیوسته می‌تواند QTL های انفرادی که روی کروموزوم مشابهی قرار دارند را تشخیص دهد. مکان‌یابی فاصله‌ای مرکب مشروط به تمامی نشانگرها توان زیادتری را نشان می‌دهد، ولی این روش نمی‌تواند به خوبی QTL های پیوسته را روی برخی از کروموزوم‌ها جدا نماید. مکان‌یابی فاصله‌ای مرسوم توان کمتری دارد و علاوه بر آن، در برآورد موقعیت‌های QTL دارای اریبی است.

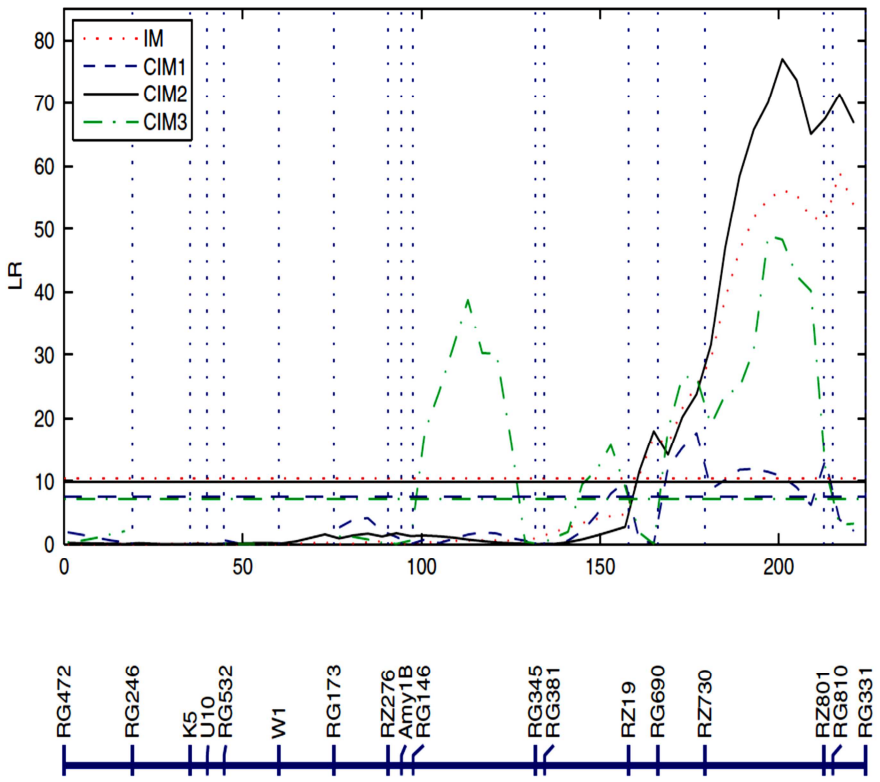
برای نشان دادن تفاوت‌های بین این دو روش متفاوت (یعنی مکان‌یابی فاصله‌ای و مکان‌یابی فاصله‌ای مرکب) از دو مثال استفاده می‌کنیم. در هر دو مثال، گروه‌های پیوستگی متعددی وجود دارند که هر یک متعلق به یک کروموزوم متفاوت هستند. سه گزینه برای انتخاب نشانگر وجود دارد: ۱- نشانگرهای پیوسته روی یک کروموزوم مشابه، ۲- نشانگرهای ناپیوسته روی کروموزوم‌های مختلف و ۳- تمامی نشانگرها روی کروموزوم‌های مشابه و مختلف. نتایج حاصل از این سه حالت با نتایج مکان‌یابی فاصله‌ای مقایسه می‌شوند.

مثال ۱۳-۱. مثال ۳-۱ را مجدداً مشاهده کنید. دو لاین اینبرد، IR64 نیمه پاکوتاه و Azucena پا بلند، با یکدیگر تلاقی داده شدند تا هتروزیگوت F_1 را ایجاد کنند. کروموزوم‌های هاپلوئید گرده‌های (گامت‌های) F_1 دو برابر شدند تا ۱۲۳ گیاه دابل هاپلوئید (DH) تولید شوند. این گیاهان DH که معادل نتاج تلاقی برگشتی هستند، برای ۱۳۵ نشانگر RFLP و ۴۰ نشانگر ایزوزیم و RAPD

تعیین ژنوتیپ شدند که از روی آن‌ها یک نقشه پیوستگی پوشش‌دهنده کل ژنوم و متشکل از ۱۲ کروموزوم، تهیه شد (یان و همکاران، ۱۹۹۸؛ شکل ۳-۳). هر یک از لاین‌های DH برای ارتفاع بوته در هر یک از ۱۰ هفته متوالی اندازه‌گیری شدند.

به‌عنوان مثال، اولین کروموزوم برای انجام مکان‌یابی فاصله‌ای مرکب (CIM) برای وزن گیاه در هفته دهم استفاده شد. در ابتدا، از مکان‌یابی فاصله‌ای (IM) برای جستجوی حضور QTL در سرتاسر کروموزوم ۱ برنج استفاده شد. آماره آزمون لگاریتم درستنمایی (LR) با تثبیت موقعیت یک QTL فرضی در هر ۲ سانتی‌مورگان در هر فاصله نشانگری محاسبه شد. پس از جستجوی QTL از اولین تا آخرین فاصله، پروفیل صاف و یکنواخت LR با استفاده از تعدادی از مقادیر ناپیوسته LR رسم شد (شکل ۱۳-۱). دو نقطه اوج LR به وسیله IM مشاهده شد که نقطه اوج اول پهن‌تر بوده و بین نشانگرهای RZ730 و RZ801 قرار داشت و نقطه اوج دیگر که باریک‌تر بود، بین نشانگرهای RG810 و RG331 قرار داشت. مقادیر LR در دو نقطه اوج (۵۵<) تا حد زیادی بیشتر از آستانه بحرانی در سطح معنی‌دار ۵ درصد (۱۰/۳۹) بودند که از بین بردن روابط اصلی فنوتیپ-نشانگر به وسیله ۱۰۰ آزمون جایگشت تعیین شد. مشخص شد که این دو QTL شناسایی شده به وسیله IM برای ارتفاع بوته، بسیار معنی‌دار هستند.

برای آزمون اینکه آیا هر یک از این دو نقطه اوج LR که به وسیله IM شناسایی شدند، متعلق به QTL های متفاوت هستند یا اینکه درمجموع یک QTL را روی کروموزوم ۱ نشان می‌دهند، مکان‌یابی فاصله‌ای مرکب مشروط به نشانگرهای پیوسته روی یک کروموزوم یکسان (CIM1) انجام شد. مطابق با خصوصیت CIM که به وسیله زنگ (۱۹۹۳) گزارش شد، CIM1 توان بیشتری برای جدا کردن چند QTL پیوسته دارد. همان‌گونه که در شکل ۱۳-۱ نشان داده شده است، CIM1 سه نقطه اوج LR مشخص را در فواصل نشانگری RG381-RZ19 و RG690-RZ730 و روی نشانگر RZ801 تشخیص داده است که از ۵ درصد آستانه بحرانی (۷/۴۷) بزرگ‌تر هستند. به نظر می‌رسد که CIM1 یک نقطه اوج بزرگ و تخت شناسایی شده به وسیله IM را به سه نقطه اوج کوچک‌تر و باریک‌تر شکسته است که پیشنهاد می‌کند CIM1 راه حل بهتری نسبت به IM برای مکان‌یابی QTL های به هم پیوسته فراهم می‌کند؛ اما چون CIM1 واریانس‌های نمونه‌گیری را افزایش داده است، از این‌رو برای دقت برآورد پارامترهای موقعیت و اثر QTL باید توجه و احتیاط بیشتری نمود.



شکل ۱۳-۱- نمودارهای مربوط به آماره‌های نسبت درستنمایی (LR ها) بر حسب موقعیت نقشه برای QTL فرضی در یک جمعیت DH برنج. IM، مکان‌یابی فاصله‌ای؛ CIM، مکان‌یابی فاصله‌ای مرکب؛ CIM1، نشانگرهای پیوسته روی یک کروموزوم؛ CIM2، نشانگرهای ناپیوسته روی کروموزوم‌های متفاوت؛ و CIM3، تمامی نشانگرها روی کروموزوم‌های مشابه و متفاوت. همان گونه که از روی آزمون‌های جایگشتی تعیین می‌شود آستانه‌های بحرانی برای IM، CIM1، CIM2 و CIM3 به ترتیب ۱۰/۳۹، ۷/۴۷، ۹/۷۹ و ۷/۱۳ هستند.

مکان‌یابی فاصله‌ای مرکب مشروط به نشانگرهای ناپیوسته (CIM2) می‌تواند توان مکان‌یابی QTL را با کنترل مقداری از واریانس باقیمانده افزایش دهد (زنگ، ۱۹۹۳). تعجب آور نیست که مشاهده کنیم CIM2 در این مثال، دو نقطه اوج بالاتر (>۷۰) در موقعیت‌های مشابه با IM شناسایی کرد که منجر به افزایش سطوح معنی‌دار می‌شود (در مقایسه با آستانه پنج درصد یعنی ۹/۷۹)؛ اما همان‌گونه که انتظار می‌رود، CIM2 برای جداسازی این دو نقطه اوج مزیتی نسبت به IM ندارد.

مکان‌یابی فاصله‌ای مرکب مشروط به هم نشانگرهای پیوسته و هم ناپیوسته (CIM3)، نتایج متفاوتی از نتایج IM، CIM1 و CIM2 نشان می‌دهد. این روش چهار نقطه اوج در فواصل

نشـانگری RG146-RG345، RG381-RZ19، RG690-RZ730 و RZ730-RZ801 تشـخیص می‌دهد که مقادیر LR آن‌ها بیش از آستانه ۵ درصـد (۷/۱۳) اسـت (شـکل ۱۳-۱). شـاید CIM3 مزایای CIM1 و CIM2 را حفظ کند و بنابراین منجر به جداسازی بهتـر تعـداد بیشـتری از QTL های پیوسته و نیز توان بهتر شناسایی QTL می‌شود.

به‌طور خلاصه، CIM3 مشروط بـه نشـانگرهای ناپیوسـته، اگرچـه توانـایی تشـخیص QTL را افزایش می‌دهد، نتایجی مشابه با نتایج IM به‌دست می‌آید. CIM1 مشروط به نشانگرهای پیوسـته، دارای وضوح بهتری برای چند QTL پیوسته است و توان آن نیز با وارد کردن نشانگرهای ناپیوسته در تجزیه تابعیت جزیی می‌تواند بهبود یابد. بـرای تعیـین قابلیـت اعتمـاد نتـایج تشـخیص QTL، مطالعات شبیه‌سازی می‌تواند با تقلید مثال بالا، بر اسـاس توزیـع ژنـومی QTL هـا، روش تـوارث و اندازه‌های اثر آن‌ها انجام شود.

مثال ۱۳-۲. مثال ۳-۲ را مجدداً مشاهده کنید. چورود و همکاران (۱۹۹۶) یـک نقشـه پیوسـتگی شامل ۷۵ نشانگر ریزماهواره را در یک جمعیت متشکل از ۵۳۵ نتاج F_2 حاصـل از دو سـویه بـزرگ (LG/J) و کوچک (SM/J) موش تهیه کردند. نتاج F_2 از نظر توده بدنی در فواصـل ۱۰ هفتگـی (از سن ۷ روزگی) اندازه‌گیری شدند. وزن‌های خام برای آثار هر یک از متغیرهای همبسته از نظر مادر، تعداد همزادان در زمان تولد و شکم زایش تصحیح شدند، ولـی تصـحیحی از نظـر عامـل جنسـیت انجام نشد. از کروموزوم ۲ شامل ۹ نشانگر برای مکـان‌یـابی QTL مـؤثر بـر وزن بـدن در سـن ۱۰ هفتگی با استفاده از مکان‌یابی فاصله‌ای (IM) و انواع متفاوت مکان‌یـابی فاصـله‌ای مرکـب (CIM) استفاده شد. در هر حالت، آزمون‌های جایگشت برای تعیین تجربـی آسـتانه‌هـای بحرانـی اسـتفاده شدند.

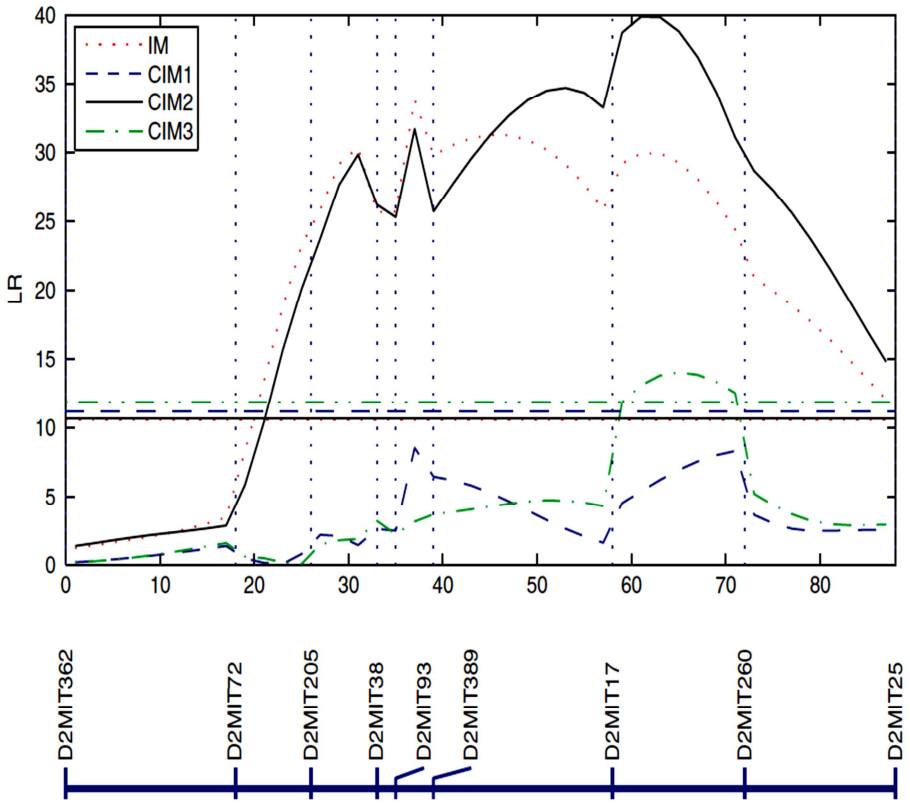
مکان‌یابی فاصله‌ای (IM) سبب ایجاد یک پروفیل تخت LR می‌شـود کـه بیشـتر کرومـوزوم را پوشش می‌دهد (شکل ۱۳-۲) و پیشنهاد می‌کند که بیش از یک QTL مـؤثر بـر وزن بـدن مـوش ممکن است روی این کروموزوم وجود داشته باشد. مکان‌یابی فاصله‌ای مرکب مشروط به نشانگرهای ناپیوسته (CIM2) پروفیل LR مشابهی ایجاد می‌کند، اگرچـه نقطـه اوج تشـخیص داده شـده بـه وسیله CIM2 بیش از نقطه اوج ایجاد شده به وسیله IM است. مکان‌یابی فاصله‌ای مرکب مشـروط به نشانگرهای پیوسته (CIM1)، علی‌رغم وجود تعدادی از نقاط اوج کوچک و متمـایز، هـیچ QTL را شناسایی نمی‌کند که پیشنهاد می‌کند CIM1 تمایل دارد تـوان آزمـون را در ایـن مثـال خـاص کاهش دهد. مکان‌یابی فاصله‌ای مرکب مشروط به نشانگرهای پیوسته و ناپیوسته (CIM3) پروفیـل LR متفاوتی را ایجاد می‌کند که نقطه اوج منفرد آن در فاصله نشانگری DZMIT17-DZMIT26

است و هیچ شواهدی مبنی بر وجود دو QTL وجود ندارد. برای آزمون قابلیت اعتماد هر یک از طرح‌های مکان‌یابی بالا، مطالعات شبیه‌سازی با تقلید این مثال مفید خواهند بود.

۱۳-۶- مکان‌یابی فاصله‌ای چندگانه

هم مکان‌یابی فاصله‌ای و هم مکان‌یابی فاصله‌ای مرکب، فقط یک QTL را در هر بار جستجوی QTL مدل‌سازی می‌کنند. اگرچه مکان‌یابی فاصله‌ای مرکب که ایده مکان‌یابی فاصله‌ای و تجزیه تابعیت نشانگر را ترکیب می‌کند، می‌تواند بر مشکل پیوستگی چند QTL غلبه کند، اما توانایی کمتری برای توصیف جزئیات ساختار ژنتیکی یک صفت کمّی دارد. چون یک صفت پیچیده ممکن است با چند QTL کنترل شود، از این‌رو داشتن یک روش مکان‌یابی که بتواند چندین QTL را به‌طور هم‌زمان مدل‌سازی و تمامی QTL های مسئول ایجاد تنوع صفت کمّی را شناسایی و تعیین موقعیت نماید، ضروری است. چنین روشی به وسیله زنگ و همکارانش پیشنهاد و به‌عنوان مکان‌یابی فاصله‌ای چندگانه نام‌گذاری شد (کائو و همکاران، ۱۹۹۹؛ زنگ و همکاران، ۲۰۰۰). کائو و زنگ (۱۹۹۷) فرمول‌های کلی مربوط به برآوردهای حداکثر درستنمایی موقعیت‌ها و آثار QTL های چندگانه را به‌دست آوردند.

مکان‌یابی فاصله‌ای چندگانه، چند QTL را به شیوه‌ای مدل‌سازی می‌کند که QTL ها بتوانند به‌طور مستقیم با استفاده همزمان از چند فاصله نشانگری کنترل شوند. ثابت شده است که این روش برای برآورد موقعیت‌ها و آثار QTL ها نسبت به مکان‌یابی فاصله‌ای مرسوم و مکان‌یابی فاصله‌ای مرکب دارای قدرت و دقت بیشتری است. بعلاوه، با جستجو و مکان‌یابی تمامی QTL های ممکن به‌طور همزمان در چند فاصله نشانگری، اجازه برآورد کامل ساختار ژنتیکی یک صفت کمّی از نظر تعداد QTL های مربوطه، آثار ژنتیکی آن‌ها، آثار پلیوتروپی و شبکه اپیستازی بین QTL های مختلف با استفاده از مکان‌یابی فاصله‌ای چندگانه فراهم است. حوزه تحقیقاتی که برای مکان‌یابی فاصله‌ای چندگانه باز است، شامل روش انتخاب مدل QTL های چندگانه، موقعیت‌های ژنومی آن‌ها و آثاری است که در مجموع بهترین برازش را برای داده‌های مشاهده شده فراهم می‌کنند.



شکل ۱۳-۲- نمودارهای مربوط به آماره‌های نسبت درستنمایی (LR ها) بر حسب موقعیت نقشه‌ای QTL فرضی. IM، مکان‌یابی فاصله‌ای؛ CIM، مکان‌یابی فاصله‌ای مرکب؛ CIM1، نشانگرهای پیوسته روی کروموزوم مشابه؛ CIM2، نشانگرهای ناپیوسته روی کروموزوم‌های متفاوت؛ CIM3، تمامی نشانگرها روی کروموزوم‌های مشابه و متفاوت.

۱۳-۷- تمرین‌ها

۱۳-۱- با توجه به درستنمایی‌های مکان‌یابی فاصله‌ای مرکب برای تلاقی برگشتی (۱۳-۲) و F_2 (۱۳-۱۰)، MLE های مربوط به پارامترهای مرحله M را همان‌طور که در معادلات (۱۳-۵) و (۱۳-۷) (برای تلاقی برگشتی) و (۱۳-۱۲) (برای F_2) نشان داده شد، به‌دست آورید.

۱۳-۲- زنگ (۱۹۹۳ و ۱۹۹۴) خصوصیات آماری مکان‌یابی فاصله‌ای مرکب را در یک طرح تلاقی برگشتی ساده که مجدداً در زیر تکرار شده است ارایه نمود:

الف. برای QTL های افزایشی (با نادیده گرفتن اپیستازی)، ضریب تابعیت جزئی مورد انتظار صفت به یک نشانگر فقط به آن QTL هایی بستگی دارد که در فاصله‌ای قرار گرفته‌اند که با دو نشانگر مجاور در بر گرفته شده‌اند و به وسیله آثار QTL های موجود در فواصل دیگر تحت تأثیر قرار نمی‌گیرند.

ب. مشروط کردن به نشانگرهای ناپیوسته در تجزیه تابعیت چندگانه، واریانس نمونه‌برداری آماره آزمون را از راه کنترل بخشی از تنوع ژنتیکی باقیمانده کاهش خواهد داد و بنابراین توان مکان‌یابی QTL افزایش خواهد یافت.

ج. مشروط کردن به نشانگرهای پیوسته در تجزیه تابعیت چندگانه شانس تداخل QTL های پیوسته چندگانه ممکن را در آزمون فرض و برآورد پارامترها کاهش خواهد داد، ولی واریانس نمونه‌برداری نیز افزایش خواهد یافت.

د. به‌طورکلی در تجزیه تابعیت چندگانه، دو ضریب تابعیت جزئی ارزش صفت بر دو نشانگر، غیر همبسته هستند، مگر آنکه دو نشانگر در مجاورت هم قرار داشته باشند.

نشان دهید که این خصوصیات برای طرح F_2 با سه ژنوتیپ مختلف در هر جایگاه ژنی کاربرد دارد.

تذکر: می‌توانید ۲، ۱ و ۰ را به ترتیب برای نشان دادن ژنوتیپ هموزیگوت آلل‌های یک والد، ژنوتیپ هتروزیگوت و ژنوتیپ هموزیگوت آلل‌های والد دیگر استفاده کنید.

۱۳-۳- مطالعات شبیه‌سازی را برای نشان دادن اینکه چه تعداد و چه نوعی از نشانگرها به عنوان کوفاکتور در مکان‌یابی فاصله‌ای مرکب دخالت دارند، انجام دهید.

۱۳-۴- مقاله کائو و زنگ (۱۹۹۷) را مطالعه کنید تا مشاهده کنید فرمول‌های کلی برای برآورد واریانس‌های نمونه‌گیری مجانبی MLE های مربوط به موقعیت‌ها و آثار QTL در مکان‌یابی فاصله‌ای چندگانه چگونه به‌دست می‌آیند. همچنین، نشان دهید که چگونه ماتریس کوواریانس مجانبی برای MLE ها می‌تواند با استفاده از روش مبتنی بر مدل مختلط لوئیس (۱۹۹۲) به‌دست آید.

فصل چهاردهم

مکان‌یابی QTL در شجره‌های دگربارور

۱۴-۱- مقدمه

روش‌های مکان‌یابی QTL اساساً برای تلاقی‌های آزمایشی، نظیر تلاقی برگشتی، هاپلوئید مضاعف، RIL ها یا F_2، ایجاد شده‌اند که از تلاقی دو لاین اینبرد به‌دست می‌آیند. به‌دلیل هموزیگوت بودن لاین‌های اینبرد، در چنین تلاقی‌هایی تفرق همزمان مندلی تمامی نشانگرها با دو آلل متفاوت می‌تواند مستقیماً مشاهده شود. در عمل، گروهی از گونه‌ها مانند درختان جنگلی یا حیوانات بزرگ که گونه‌های دگربارور[1] نامیده می‌شوند، نیز وجود دارند که تولید لاین‌های اینبرد در آن‌ها به‌واسطه فاصله نسل طولانی و هتروزیگوتی زیاد غیرممکن یا مشکل است، اگرچه دورگه‌های آزمایشی آن‌ها می‌توانند به صورت تجاری برای بهبود ژنتیکی مورد استفاده قرار گیرند.

بسیاری از جوامع دورگ تجاری در دسترس برای گونه‌های دگربارور می‌توانند مستقیماً به عنوان مواد آزمایشی مطالعات مکان‌یابی عمل کنند. ولی مدل‌های مکان‌یابی آماری که برای این جمعیت‌ها استفاده می‌شوند، باید متغیر باشند که به خصوصیات زیستی گونه‌ها بستگی دارد. برخی گونه‌ها، نظیر درختان جنگلی، می‌توانند خانواده دورگ بزرگی تولید کنند (گراتاپاگلیا و سدروف، ۱۹۹۴؛ برادشاو و استتلر، ۱۹۹۵)، به‌طوری که یک تلاقی منفرد برای به‌دست آوردن توان مورد نیاز کفایت می‌کند. از سوی دیگر، برای برخی گونه‌ها، نظیر سگ‌ها، تولید خانواده بزرگ و بنابراین خانواده‌های چندگانه که هر یک دارای اندازه کوچکی هستند، غیرممکن است و وارد نمودن برخی از والدین خویشاوند و غیرخویشاوند در این گونه‌ها مورد نیاز خواهد بود (بلیس و همکاران، ۲۰۰۲؛ تودانتر و همکاران، ۲۰۰۵). تجزیه پیوستگی برای این دو نوع مختلف از گونه‌های دگربارور به ترتیب در فصول ۴ و ۷ توضیح داده شدند. در این فصل، مدل‌های آماری را ارایه خواهیم کرد که در ارتباط با مکان‌یابی QTL برای گونه‌های دگربارور هستند. بر اساس خصوصیات زیستی گونه‌های دگربارور، دو نوع مختلف از مدل‌های مکان‌یابی توضیح داده خواهند شد که اولین مدل در مورد یک خانواده تنی دارای اندازه بزرگ و دومین نوع در خصوص تعداد زیادی از خانواده‌های خویشاوند و هر یک با اندازه کوچک خواهند بود.

برای یک لاین دگربارور معین، برخی نشانگرها ممکن است هتروزیگوت باشند، حال آنکه نشانگرهای دیگر ممکن است در سرتاسر ژنوم هموزیگوت باشند. تمامی نشانگرها ممکن است یا ممکن نیست دارای سیستم آللی یکسان در بین هر دو لاین دگربارور مورد استفاده در تلاقی باشند. همچنین، برای یک جفت ژن هتروزیگوت، آرایش آللی آن‌ها در امتداد دو کروموزوم همولوگ (یعنی فاز پیوستگی) نمی‌تواند از الگوی تفرق ژنوتیپ‌ها در تلاقی مشاهده شود. متأسفانه، وجود

---

[1] - Outcrossing species

تعداد ثابتی از آلل‌ها در نشانگرهای مختلف و فازهای پیوستگی شناخته شده آن‌ها، پیش شرط روش‌های مکان‌یابی آماری تشریح شده برای تلاقی برگشتی یا F_2 در فصول پیشین هستند.

۱۴-۲- مدل با آثار ثابت برای یک خانواده تنی

۱۴-۲-۱- مقدمه

چند مدل آماری مخصوص برای مکان‌یابی QTL در خانواده تنی پیشنهاد شده است (شفر-پرگل و همکاران، ۱۹۹۶؛ جانسون و همکاران، ۱۹۹۹، سونگ و همکاران، ۱۹۹۹). در برخی مطالعات، الگوریتم‌های آماری پیچیده‌تر نظیر روش بیزی متکی بر روش مونت کارلو زنجیره مارکوف برای در نظر گرفتن پیچیدگی‌های مربوط به مکان‌یابی خانواده تنی پیشنهاد شده‌اند (هوئشل و همکاران، ۱۹۹۷؛ سیلانپا و آرجاس، ۱۹۹۹). لین و همکاران (۲۰۰۳) یک مدل کلی برای مکان‌یابی خانواده‌های تنی از راه وارد کردن تعداد و آرایش آللی در مفهوم مدل مختلط ارایه نمودند. در این بخش، این مدل کلی با یک مثال تشریح شده و نشان داده خواهد شد.

۱۴-۲-۲- مدل مختلط برای دیپلوتیپ والدینی

برای یک خانواده تنی حاصل از دو والد دگربارور تا ۴ آلل نشانگر، علاوه بر آلل خاموش، می‌تواند در یک جایگاه منفرد وجود داشته باشد. همچنین، تعداد آلل‌ها ممکن است در تمام جایگاه‌های ژنی متغیر باشد. هر یک از آلل‌های نشانگر نسبت به آلل خاموش غالب هستند. در بخش ۳-۴، انواع ممکن تلاقی‌ها برای نشانگرهای در حال تفرق در جدول ۳-۲ ارایه شد.

دو لاین والدینی دگربارور به صورت P_1 و P_2 در نظر بگیرید که به ترتیب دارای یک مجموعه حاوی دو کروموزوم همولوگ به‌صورت ۱۲ و ۳۴ هستند. تلاقی بین این دو لاین، ۱۲×۳۴، منجر به ایجاد چهار جفت کروموزوم ممکن والدینی می‌شود که شامل ۱۳، ۱۴، ۲۳ و ۲۴ هستند. از اعداد عربی برجسته (ضخیم) برای نشان دادن کروموزوم‌های والدینی استفاده نمودیم. اگرچه تعداد زیادی از انواع مختلف نشانگرها در خانواده تنی حاصل از دو لاین والدینی دگربارور ممکن است وجود داشته باشند، اما به‌طورکلی تمامی نشانگرهای مشاهده شده به‌صورت ۱ و ۲ برای والد P_1 و ۳ و ۴ برای والد P_2 بیان می‌شوند و مهم نیست که از کدام نوع به‌دست آمده‌اند.

فرض کنید که یک QTL بین دو نشانگر مستقر است. چهار آلل QTL که به وسیله ۱ و ۲ برای والد P_1 و ۳ و ۴ برای والد P_2 نشان داده می‌شوند، برای ایجاد زیگوت‌های ۱۳، ۱۴، ۲۳ و ۲۴ که از نسبت ۱:۱:۱:۱ در خانواده تبعیت می‌کنند، در حال تفرق خواهند بود. نرخ‌های نوترکیبی بین دو نشانگر، بین نشانگر M_1 و QTL و بین QTL و نشانگر M_2 که به ترتیب با r_1، r و r_2 نشان داده می‌شوند، به‌صورت زیر بیان می‌شوند:

$$r=\begin{cases} r_1+r_2 & \text{برای عدم وجود نوترکیبی} \\ r_1+r_2-2r_1r_2 & \text{برای نوترکیبی مستقل} \\ r_1+r_2-2cr_1r_2 & \text{برای تداخل نوترکیبی} \end{cases}$$

که در آن، c ضریب تطابق بین نوترکیبی‌ها در دو فاصله مختلف است. تفاوت پیوستگی مختص والدین نادیده گرفته می‌شود. آلل‌های این دو نشانگر و QTL بین دو کروموزوم همولوگ در هر یک از چهار فاز پیوستگی برای هر والد مرتب می‌شوند.

ولی فازهای پیوستگی آللی دو نشانگر برای هر دو والد می‌تواند از راه تجزیه‌های پیوستگی نشانگرها با استفاده از راهبرد پیشنهاد شده در فصل ۴ مشخص شود؛ بنابراین، با در نظر گرفتن فاز پیوستگی نشانگری ثابت، ۴= ۲×۲ ترکیب والدینی (Φ ها) فاز پیوستگی (یا دیپلوتیپ) QTL نسبت به دو نشانگر به همراه ترتیب چهار ژنوتیپ نشانگر در نتاج به‌صورت شماتیک به شکل زیر هستند:

$$\begin{cases} \Phi_{11}=\begin{matrix}1\\1\\1\end{matrix}\left|\begin{matrix}2\\2\\2\end{matrix}\right|\times\begin{matrix}3\\3\\3\end{matrix}\left|\begin{matrix}4\\4\\4\end{matrix}\right| \rightarrow (13,14,23,24) \\ \Phi_{12}=\begin{matrix}1\\1\\1\end{matrix}\left|\begin{matrix}2\\2\\2\end{matrix}\right|\times\begin{matrix}3\\4\\3\end{matrix}\left|\begin{matrix}4\\3\\4\end{matrix}\right| \rightarrow (14,13,24,23) \\ \Phi_{21}=\begin{matrix}1\\2\\1\end{matrix}\left|\begin{matrix}2\\1\\2\end{matrix}\right|\times\begin{matrix}3\\3\\3\end{matrix}\left|\begin{matrix}4\\4\\4\end{matrix}\right| \rightarrow (23,24,13,14) \\ \Phi_{22}=\begin{matrix}1\\2\\1\end{matrix}\left|\begin{matrix}2\\1\\2\end{matrix}\right|\times\begin{matrix}3\\4\\3\end{matrix}\left|\begin{matrix}4\\3\\4\end{matrix}\right| \rightarrow (24,23,14,13) \end{cases} \qquad (۱-۱۴)$$

که در آن، اولین و دومین اندیس Φ به ترتیب نشان‌دهنده دو فاز ممکن والدین P_1 و P_2 و خطوط عمودی برای هر ترکیب دیپلوتیپی نشان‌دهنده دو کروموزوم والدینی 12 و 34 هر یک برای یک والد هستند. هر والد، بدون توجه به این که دارای کدام ترکیب دیپلوتیپی ممکن است، هشت گامت هاپلوئیدی (هاپلوتیپ) سه مکانی تولید خواهد کرد که فراوانی‌های هاپلوتیپی به نرخ‌های

نوترکیبی این سه جایگاه ژنی بستگی دارند. فراوانی‌های هشت هاپلوتیپ بـه صـورت p_{111} ، p_{121} ، p_{112} ، p_{122} ، p_{211} ، p_{221} ، p_{212} و p_{222} برای والـد P_1 بـه همـراه فراوانـی‌هـای هـاپلوتیپی دو نشانگر به صورت p_{11} ، p_{12} ، p_{21} و p_{22} در جدول ۱۴-۱ نشـان داده شـده اسـت. فراوانـی‌هـای هاپلوتیپی مربوط به والد P_2 نیز می‌توانند ارایه شوند.

فراوانی‌های هاپلوتیپی نشانگر-QTL می‌توانند بر حسب نرخ‌های نـوترکیبی بـین هـر نشـانگر و QTL (r_1 و r_2)، وقتی که نوترکیبی مضاعفی وجود ندارد (حالت ۱) یا نوترکیبی مستقل فرض شده است (حالت ۲)، بیان شوند؛ اما اگر تداخلی بین فواصل مختلف در طی میوز (حالت ۳) وجود داشته باشد، فراوانی‌های هاپلوتیپی باید به صورت متفاوت بیـان شـوند. فـرض کنیـد g_{00} ، g_{01} ، g_{10} و g_{11} به ترتیب احتمال عدم وقوع نوترکیبی در فواصل M_1-QTL و QTL-M_2، یک نوترکیبی فقط در فاصله M_1-QTL، یک نوترکیبی فقط در فاصله QTL-M_2 و یک نوترکیبی در هـر یـک از ایـن فواصل باشند. فراوانی‌های هاپلوتیپی در سه حالت بدین صورت بیان می‌شوند:

فراوانی هاپلوتیپ	حالت ۱	حالت ۲	حالت ۳
p_{111}	1	$\frac{1}{2}(1-r_1)(1-r_2)$	$\frac{1}{2}g_{00}$
p_{121}	0	$\frac{1}{2}\ r_1r_2$	$\frac{1}{2}g_{11}$
p_{112}	$\frac{1}{2}r_2$	$\frac{1}{2}\ (1-r_1)r_2$	$\frac{1}{2}g_{01}$
p_{122}	$\frac{1}{2}r_1$	$\frac{1}{2}r_1(1-r_2)$	$\frac{1}{2}g_{10}$
p_{211}	$\frac{1}{2}r_1$	$\frac{1}{2}r_1(1-r_2)$	$\frac{1}{2}g_{10}$
p_{221}	$\frac{1}{2}r_2$	$\frac{1}{2}\ (1-r_1)r_2$	$\frac{1}{2}g_{01}$
p_{212}	0	$\frac{1}{2}\ r_1r_2$	$\frac{1}{2}g_{11}$
p_{222}	۱	$\frac{1}{2}(1-r_1)(1-r_2)$	$\frac{1}{2}g_{00}$

جدول ۱۴-۱- فراوانی هاپلوتیپ‌های تشکیل شده به وسیله دو نشانگر و یک QTL موجود بین دو نشانگر در والد P_1.

نشانگرهای مجاور هم		QTL	
هاپلوتیپ	فراوانی	1	2
11	$p_{11}=\frac{1}{2}(1-r)$	p_{111}	p_{121}
12	$p_{12}=\frac{1}{2}r$	p_{112}	p_{122}
21	$p_{21}=\frac{1}{2}r$	p_{211}	p_{221}
22	$p_{22}=\frac{1}{2}(1-r)$	p_{212}	p_{222}

برای فراوانی هاپلوتیپ‌های دو نشانگر، اولین و دومین اندیس را به ترتیب برای آلل‌های نشانگرهای M_1 و M_2 در نظر بگیرید. برای فراوانی هاپلوتیپ‌های سه مکانی، دومین اندیس متعلق به آلل‌های QTL است، حال آنکه اولین و دومین اندیس به ترتیب به آلل‌های دو نشانگر تعلق دارند. می‌توان نشان داد که $p_{11}=p_{111}+p_{121}$ ، $p_{12}=p_{112}+p_{122}$ ، $p_{21}=p_{211}+p_{221}$ و $p_{22}=p_{212}+p_{222}$.

هشت هاپلوتیپ حاصل از والد P_1 به‌طور تصادفی با هشت هاپلوتیپ حاصل از والد P_2 ترکیب می‌شوند تا در مجموع ۶۴ ژنوتیپ زیگوت ایجاد کنند. فراوانی‌های این ژنوتیپ‌ها در خانواده تنی بر حسب فراوانی‌های هاپلوتیپی (که به‌صورت g ها بیان می‌شوند) محاسبه می‌شوند که در جدول ۱۴-۱ ارائه شده‌اند. به این امر توجه شده است که دو والد برای تلاقی دارای چهار ترکیب ممکن از دیپلوتیپ‌های والدینی هستند (همان گونه که با معادله (۱۴-۱) نیز نشان داده شده است)، ولی یک و فقط یک ترکیب صحیح است. تحت این ترکیبات دیپلوتیپی مختلف، دو والد، مجموعه مشابهی از ۶۴ ژنوتیپ زیگوت را تشکیل می‌دهند، ولی فراوانی‌های شکل‌گیری ژنوتیپ‌ها برای هر فاز اختصاصی هستند. فراوانی‌های ژنوتیپی ارائه شده در جدول ۱۴-۱ به ترکیبات دیپلوتیپی والدینی مختلف (Φ_{11}، Φ_{12}، Φ_{21} و Φ_{22}) تعلق دارند.

جدول ۱۴-۲- فراوانی‌های ژنوتیپی برای دو نشانگر و QTL احاطه شده با نشانگرها در یک خانواده تنی حاصل از دو والد دگربارور با ترکیبات دیپلوتیپی مختلف.

نشانگرهای مجاور هم			فاز	ژنوتیپ QTL			
			Φ_{11}	۱۳	۱۴	۲۳	۲۴
			Φ_{12}	۱۴	۱۳	۲۴	۲۳
			Φ_{21}	۲۳	۲۴	۱۳	۱۴
شماره	ژنوتیپ	فراوانی	Φ_{22}	۲۴	۲۱	۱۴	۱۳
				$G_{l_1l_2}$			
۱	۱۳/۱۳	p_{11}^2		p_{111}^2	$p_{111}p_{121}$	$p_{121}p_{111}$	p_{121}^2
۲	۱۳/۱۴	$p_{11}p_{12}$		$p_{111}p_{112}$	$p_{111}p_{122}$	$p_{121}p_{112}$	$p_{121}p_{122}$
۳	۱۳/۲۳	$p_{12}p_{11}$		$p_{111}p_{112}$	$p_{112}p_{121}$	$p_{122}p_{111}$	$p_{121}p_{122}$
۴	۱۳/۲۴	p_{12}^2		p_{112}^2	$p_{112}p_{122}$	$p_{122}p_{112}$	p_{122}^2
۵	۱۴/۱۳	$p_{11}p_{21}$		$p_{111}p_{211}$	$p_{111}p_{221}$	$p_{121}p_{211}$	$p_{121}p_{221}$
۶	۱۴/۱۴	$p_{11}p_{22}$		$p_{111}p_{212}$	$p_{111}p_{222}$	$p_{121}p_{212}$	$p_{121}p_{222}$
۷	۱۴/۲۳	$p_{12}p_{21}$		$p_{112}p_{211}$	$p_{112}p_{221}$	$p_{122}p_{211}$	$p_{122}p_{221}$
۸	۱۴/۲۴	$p_{12}p_{22}$		$p_{112}p_{212}$	$p_{112}p_{122}$	$p_{122}p_{212}$	$p_{122}p_{222}$
۹	۲۳/۱۳	$p_{21}p_{11}$		$p_{211}p_{111}$	$p_{211}p_{121}$	$p_{221}p_{111}$	$p_{221}p_{121}$
۱۰	۲۳/۱۴	$p_{21}p_{12}$		$p_{211}p_{112}$	$p_{211}p_{122}$	$p_{221}p_{112}$	$p_{221}p_{122}$
۱۱	۲۳/۲۳	$p_{22}p_{11}$		$p_{212}p_{111}$	$p_{212}p_{121}$	$p_{222}p_{111}$	$p_{222}p_{121}$
۱۲	۲۳/۲۴	$p_{22}p_{12}$		$p_{212}p_{112}$	$p_{212}p_{122}$	$p_{222}p_{112}$	$p_{222}p_{122}$
۱۳	۲۴/۱۳	p_{21}^2		p_{211}^2	$p_{211}p_{221}$	$p_{221}p_{211}$	p_{221}^2
۱۴	۲۴/۱۴	$p_{21}p_{22}$		$p_{211}p_{212}$	$p_{211}p_{222}$	$p_{221}p_{212}$	$p_{221}p_{222}$
۱۵	۲۴/۲۳	$p_{21}p_{22}$		$p_{211}p_{212}$	$p_{212}p_{221}$	$p_{222}p_{211}$	$p_{221}p_{222}$
۱۶	۲۴/۲۴	p_{22}^2		p_{212}^2	$p_{212}p_{222}$	$p_{222}p_{212}$	p_{222}^2

نکته: فراوانی‌های ژنوتیپی تحت ترکیب دیپلوتیپی والدینی $\Phi_{l_1l_2}\ (l_1,l_2=1,2)$ در یک ماتریس $G_{l_1l_2}$ با ابعاد (۴×۱۶) مرتب شده‌اند. با تصحیح ترتیب ژنوتیپ‌های QTL در خانواده تنی تحت ترکیبات دیپلوتیپی والدینی Φ_{11}، Φ_{12}، Φ_{21} و Φ_{22}، ماتریس‌های G_{11}، G_{12}، G_{21} و G_{22} می‌توانند ایجاد شوند.

p و q را به ترتیب به عنوان احتمالاتی در نظر بگیرید که به وسیله آن‌ها، اولین دیپلوتیپ در شمای (۱۴-۱) برای والدین P_1 و P_2 رخ می‌دهد؛ بنابراین، احتمالات مربوط به چهار ترکیب دیپلوتیپی والدینی به ترتیب به صورت $\phi_{11} = pq$، $\phi_{12} = p(1-q)$، $\phi_{21} = (1-p)q$ و $\phi_{22} = (1-p)(1-q)$ برای Φ_{11}، Φ_{12}، Φ_{21} و Φ_{22} خواهند بود. G_{11}، G_{12}، G_{21} و G_{22} نیز ماتریس‌های فراوانی ژنوتیپی تحت ترکیبات دیپلوتیپی مربوطه هستند (جدول ۱۴-۲)؛ بنابراین، فراوانی‌های ژنوتیپی مشاهده شده G در خانواده تنی باید مخلوطی از فراوانی‌های ژنوتیپی باشند که به وسیله احتمالات ترکیب دیپلوتیپی وزن داده شده‌اند. این فراوانی‌ها عبارتند از:

$$G = \phi_{11}G_{11} + \phi_{12}G_{12} + \phi_{21}G_{21} + \phi_{22}G_{22} \qquad (۱۴-۲)$$

که شامل یک ماتریس (۴×۱۶) با ۱۶ ردیف برای ژنوتیپ‌های دو نشانگر و چهار ستون برای ژنوتیپ‌های QTL در خانواده تنی است. بر اساس قضیه بیز، احتمالات شرطی ژنوتیپ QTL یعنی uv (با آلل‌های u= ۱ و ۲ که از والد P_1 به ارث می‌رسد و آلل‌های v= ۳ و ۴ که از والد P_2 به ارث می‌رسند)، به شرط ژنوتیپ‌های نشانگر با تقسیم G بر فراوانی‌های ژنوتیپ‌های نشانگر که در جدول ۱۴-۲ ارایه شده‌اند، برآورد می‌شوند. احتمال شرطی یک ژنوتیپ QTL خاص مشروط به ژنوتیپ نشانگر فرد i ام بدین صورت بیان می‌شود:

$$\omega_{uv|i} = \phi_{11}\omega^{11}_{uv|i} + \phi_{12}\omega^{12}_{uv|i} + \phi_{21}\omega^{21}_{uv|i} + \phi_{22}\omega^{22}_{uv|i} \qquad (۱۴-۳)$$

که سمت راست معادله حاوی احتمالات شرطی مختص دیپلوتیپ‌های والدینی است.

۱۴-۲-۳- مدل ژنتیک کمّی

در یک خانواده تنی، یک QTL چهار ژنوتیپ تولید می‌کند. μ_{uv} ارزش ژنوتیپ QTL است که آلل u را از والد P_1 و آلل v را از والد P_2 به ارث می‌برد. بر اساس نظریه ژنتیک کمّی، این ارزش ژنوتیپی می‌تواند به‌صورت زیر به آثار افزایشی و غالبیت تقسیم شود:

$$\mu_{uv} = \mu + \alpha_u + \beta_v + \gamma_{uv}$$

که در آن، μ میانگین کل، α_u و β_v به ترتیب آثار آللی (افزایشی) آلل‌های u و v و γ_{uv} اثر متقابل (غالبیت) در QTL است. با در نظر گرفتن تمامی آلل‌های ممکن و ترکیبات آللی بین دو والد، در مجموع چهار اثر افزایشی (α_1 و α_2 حاصل از والد P_1 و β_3 و β_4 حاصل از والد P_2) و چهار اثر غالبیت (γ_{13}، γ_{14}، γ_{23} و γ_{34}) وجود دارد. ولی این آثار افزایشی و غالبیت مستقل نبوده و بنابراین قابل برآورد نیستند. پس از تعیین پارامترها، دو اثر افزایشی مستقل، $\alpha = \alpha_1 = -\alpha_2$ و $\beta = \beta_3 = -\beta_4$ و اثر غالبیت $\gamma = \gamma_{13} = -\gamma_{14} = -\gamma_{23} = \gamma_{24}$ برآورد می‌شود.

اگر $u=(\mu_{uv})_{4\times1}$ و $a=(\mu,\alpha,\beta,\gamma)^T$ باشند که با یک ماتریس طرح D به یکدیگر مرتبط می‌شوند، داریم:

$$u=Da$$

که در آن:

$$D=\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 1 & -1 & -1\\ 1 & -1 & 1 & -1\\ 1 & -1 & -1 & 1\end{bmatrix}$$

MLE مربوط به a می‌تواند از روی MLE مربوط به u بدین صورت به‌دست آید:

$$\hat{a}=D^{-1}\hat{u}$$

۱۴-۲-۴- تجزیه درستنمایی

فرض کنید یک خانواده تنی با اندازه n حاصل از دو لاین دگربارور وجود دارد. یک QTL برای یک صفت کمّی در نظر بگیرید که به وسیله دو نشانگر احاطه شده است. مدل خطی ارزش صفت برای فرد i ام (y_i) که تحت‌تأثیر این QTL قرار می‌گیرد، عبارت است از:

$$y_i=\sum_{u=1}^{2}\sum_{v=3}^{4}\xi_{iuv}\mu_{uv}+e_i \qquad (۱۴-۴)$$

که ξ_{iuv} متغیر شاخص و موهومی برای ژنوتیپ‌های QTL است، به‌طوری که اگر ژنوتیپ فرد i ام به صورت uv در نظر گرفته شود، ارزش آن ۱ و در غیر این صورت، ارزش آن صفر است، e_i نیز خطای باقیمانده است که دارای توزیع نرمال با میانگین صفر و واریانس σ^2 است. احتمال اینکه فرد i ام ژنوتیپ uv را برای QTL حمل کند، می‌تواند از روی ژنوتیپ نشانگر آن استنباط و استخراج شود که این احتمال به صورت $\omega_{uv|i}$ بیان می‌شود.

لگاریتم درستنمایی مقادیر صفت (y) و اطلاعات نشانگری (M) به‌صورت زیر ارائه می‌شوند:

$$\log L(\Omega\,|\,y,M)=\sum_{i=1}^{n}\log\left[\sum_{u=1}^{2}\sum_{v=3}^{4}\omega_{uv|i}\,f_{uv}(y_i)\right] \qquad (۱۴-۵)$$

که در این رابطه، Ω بردار پارامترهای ناشناخته QTL است که شامل موقعیت QTL (r_1 و r_2)، احتمالات فازهای والدینی، ارزش‌های ژنوتیپی QTL (μ_{uv}) و واریانس باقیمانده (σ^2) است. دو پارامتر اول که با Ω_p نشان داده شده‌اند، شامل نسبت‌های اختلاط مدل بالا هستند، حال آنکه دو پارامتر دوم که با Ω_q نشان داده شده‌اند، پارامترهای ژنتیک کمّی هستند.

به‌طور معمول، MLE های بردار ناشناخته با مشتق‌گیری از تابع لگاریتم درستنمایی نسبت به هر یک از Ω_τ ناشناخته و برابر صفر قرار دادن مشتق‌ها و حل معادلات لگاریتم درستنمایی به‌دست می‌آیند. این رویه به صورت زیر نشان داده می‌شود:

$$\frac{\partial}{\partial\Omega_\tau}\log L\left(\Omega\mid y,M\right)$$

$$=\sum_{i=1}^{n}\sum_{u=1}^{2}\sum_{v=3}^{4}\frac{f_{uv}\left(y_i\right)\frac{\partial}{\partial\Omega_p}\omega_{uv|i}+\omega_{uv|i}\frac{\partial}{\partial\Omega_q}f_{uv}\left(y_i\right)}{\sum_{u'=1}^{2}\sum_{v'=3}^{4}\omega_{u'v'|i}f_{u'v'}\left(y_i\right)}$$

$$=\sum_{i=1}^{n}\sum_{u=1}^{2}\sum_{v=3}^{4}\left[\frac{\omega_{uv|i}f_{uv}\left(y_i\right)\frac{1}{\omega_{uv|i}}\frac{\partial}{\partial\Omega_p}\omega_{uv|i}}{\sum_{u'=1}^{2}\sum_{v'=3}^{4}\omega_{u'v'|i}f_{u'v'}\left(y_i\right)}+\frac{\omega_{uv|i}f_{uv}\left(y_i\right)\frac{\partial}{\partial\Omega_q}\log f_{uv}\left(y_i\right)}{\sum_{u'=1}^{2}\sum_{v'=3}^{4}\omega_{u'v'|i}f_{u'v'}\left(y_i\right)}\right]$$

$$=\sum_{i=1}^{n}\sum_{u=1}^{2}\sum_{v=3}^{4}\prod_{uv|i}\left[\frac{1}{\omega_{uv|i}}\frac{\partial}{\partial\Omega_p}\omega_{u'v'|i}+\frac{\partial}{\partial\Omega_q}\log f_{uv}\left(y_i\right)\right]$$

در رابطه بالا، $\prod_{uv|i}$ به صورت زیر تعریف می‌شود:

(۶-۱۴)

$$\prod_{uv|i}=\frac{\omega_{uv|i}f_{uv}\left(y_i\right)}{\sum_{u'=1}^{2}\sum_{v'=3}^{4}\omega_{u'v'|i}f_{u'v'}\left(y_i\right)}$$

که می‌تواند به عنوان احتمال پسین اینکه فرد i ام دارای ژنوتیپ uv برای QTL است، باشد. سپس الگوریتم EM با پارامترهای بسط یافته $\{\Omega,\Pi\}$ بکار گرفته می‌شود که در آن $\Pi=\{\Pi_{uv|i}\}$ است. مشروط به Π، معادله $\frac{\partial}{\partial\Omega_l}\log L\left(\Omega\mid y,M\right)$ را برابر با صفر قرار داده و حل می‌کنیم تا برآوردهای Ω را به‌دست آوریم (مرحله M). این برآوردها برای نوسازی و تجدید مقادیر Π استفاده می‌شوند (مرحله E) و این فرآیند تا زمان رسیدن به همگرایی تکرار می‌شود. مقادیر در هنگام همگرایی به عنوان برآوردهای حداکثر درستنمایی (MLE ها) هستند.

برآوردهای مربوط به ارزش‌ها و واریانس ژنوتیپی در مرحله M به صورت زیر به‌دست می‌آیند:

$$\hat{\mu}_{uv} = \frac{\sum_{i=1}^{n} \Pi_{uv|i} y_i}{\sum_{i=1}^{n} \Pi_{uv|i}}$$
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{u=1}^{2} \sum_{v=3}^{4} \Pi_{uv|i} \left(y_i - \hat{\mu}_{uv} \right)^2 \qquad (۷-۱۴)$$

محاسبه احتمالات دیپلوتیپ والدینی مشکل خواهد بود، چون این احتمالات در سلسله مراتب پایین‌تر نسبت‌های اختلاط، همان‌طور که با معادله (۱۴-۳) نیز نشان داده شده است، وارد شده‌اند.

با جایگزینی معادله (۱۴-۳) در لگاریتم درستنمایی (۱۴-۵) داریم:

$$\log L(\Omega \mid y, M) = \sum_{i=1}^{n} \log\left[\omega_{13} f_{13}(y_i) + \omega_{14} f_{14}(y_i) + \omega_{23} f_{23}(y_i) + \omega_{24} f_{24}(y_i) \right]$$
$$= \sum_{i=1}^{n} \log\Big\{ \left[pq\omega_{13|i}^{11} + p(1-q)\omega_{13|i}^{12} + (1-p)q\omega_{13|i}^{21} + (1-p)(1-q)\omega_{13|i}^{22} \right] f_{13}(y_i)$$
$$+ \left[pq\omega_{14|i}^{11} + p(1-q)\omega_{14|i}^{12} + (1-p)q\omega_{14|i}^{21} + (1-p)(1-q)\omega_{14|i}^{22} \right] f_{14}(y_i)$$
$$+ \left[pq\omega_{23|i}^{11} + p(1-q)\omega_{23|i}^{12} + (1-p)q\omega_{23|i}^{21} + (1-p)(1-q)\omega_{23|i}^{22} \right] f_{23}(y_i)$$
$$+ \left[pq\omega_{24|i}^{11} + p(1-q)\omega_{24|i}^{12} + (1-p)q\omega_{24|i}^{21} + (1-p)(1-q)\omega_{24|i}^{22} \right] f_{24}(y_i) \Big\}$$

که از روی آن، شکل‌های دقیق برآوردهای احتمالات دیپلوتیپ والدینی به‌صورت زیر به‌دست می‌آید:

(۸-۱۴)

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} \left(\psi_1 \prod_{13|i} + \psi_2 \prod_{14|i} + \psi_3 \prod_{23|i} + \psi_4 \prod_{24|i} \right)$$
$$\hat{q} = \frac{1}{n} \sum_{i=1}^{n} \left(\psi'_1 \prod_{13|i} + \psi'_2 \prod_{14|i} + \psi'_3 \prod_{23|i} + \psi'_4 \prod_{24|i} \right)$$

در این روابط، داریم:

$$\psi_1 = \frac{pq\omega_{13|i}^{11} + p(1-q)\,\omega_{13|i}^{12}}{\omega_{13|i}}$$

$$\psi_2 = \frac{pq\omega_{14|i}^{11} + p(1-q)\,\omega_{14|i}^{12}}{\omega_{14|i}}$$

$$\psi_3 = \frac{pq\omega_{23|i}^{11} + p(1-q)\,\omega_{23|i}^{12}}{\omega_{23|i}}$$

$$\psi_4 = \frac{pq\omega_{24|i}^{11} + p(1-q)\,\omega_{24|i}^{12}}{\omega_{24|i}}$$

$$\psi_1' = \frac{pq\omega_{13|i}^{11} + (1-p)\,q\omega_{13|i}^{12}}{\omega_{13|i}}$$

$$\psi_2' = \frac{pq\omega_{14|i}^{11} + (1-p)\,q\omega_{14|i}^{12}}{\omega_{14|i}}$$

$$\psi_3' = \frac{pq\omega_{23|i}^{11} + (1-p)\,q\omega_{23|i}^{12}}{\omega_{23|i}}$$

$$\psi_4' = \frac{pq\omega_{24|i}^{11} + (1-p)\,q\omega_{24|i}^{12}}{\omega_{24|i}}$$

دو روش برای برآورد نرخ نوترکیبی بین نشانگرها و QTL (r_1 یا r_2) وجود دارد که موقعیت QTL را وقتی که نوترکیبی مضاعف وجود ندارد (حالت ۱) یا نوترکیبی مستقل در نظر گرفته می‌شود (حالت ۲) نشان می‌دهد. اولین روش بر اساس جستجوی کل ژنوم است که موقعیت QTL را با ثابت در نظر گرفتن r_1 (و بنابراین r_2) برآورد می‌کند. با استفاده از روش جستجوی شبکه‌ای، می‌توانیم MLE مربوط به موقعیت QTL را از روی نقطه اوج پروفیل آماره آزمون نسبت لگاریتم درستنمایی در طول یک کروموزوم به‌دست آوریم. روش دوم بر اساس برآوردهای دقیق فراوانی‌های هاپلوتیپی یعنی p ها (و بنابراین نرخ‌های نوترکیبی) است. اگر بین فواصل مختلف تداخل در نظر گرفته شود (حالت ۳)، فقط برآورد دقیق مورد استفاده قرار می‌گیرد. ترکیب جدول ۱۴-۲ و معادله (۱۴-۳) ترکیبی از احتمالات شرطی را بر حسب p ایجاد خواهند کرد. شکل‌های دقیق MLE های مربوط به p می‌توانند با محدودیت $p_{121} = p_{212} = 0$ برای حالت ۱، محدودیت

$p_{121}+p_{212}=\left(p_{121}+p_{212}+p_{122}+p_{211}\right)\left(p_{121}+p_{212}+p_{112}+p_{221}\right)$ برای حالت ۲ و بدون ایجاد محدودیت برای حالت ۳ به‌دست آیند. در هر حالت، از MLE های مربوط به p ها برای به‌دست آوردن MLE های نرخ‌های نوترکیبی به وسیله جدول زیر استفاده می‌شود:

MLE	حالت ۱	حالت ۲	حالت ۳
$\hat{r}_1$	$\hat{p}_{122}+\hat{p}_{211}$	$\hat{p}_{121}+\hat{p}_{212}+\hat{p}_{122}+\hat{p}_{211}$	$\hat{p}_{122}+\hat{p}_{211}+\hat{p}_{121}+\hat{p}_{212}$
$\hat{r}_2$	$\hat{p}_{112}+\hat{p}_{221}$	$\hat{p}_{121}+\hat{p}_{212}+\hat{p}_{112}+\hat{p}_{221}$	$\hat{p}_{112}+\hat{p}_{221}+\hat{p}_{121}+\hat{p}_{212}$

توجه کنید که MLE های نرخ‌های نوترکیبی ممکن است در بین سه حالت متفاوت باشند و درجه این تفاوت‌ها بستگی به مقدار سازگاری و ثبات این حالات برای یک مجموعه داده مورد نظر دارد.

۱۴-۲-۵- برازش فنوتیپ‌های نشانگر

ما یک چارچوب کلی برای مکان‌یابی QTL در یک خانواده تنی بر اساس ژنوتیپ‌های زیگوت نشانگر تهیه کرده‌ایم؛ اما در عمل فقط فنوتیپ‌های زیگوت‌های نشانگر می‌توانند مشاهده شوند. تعداد فنوتیپ‌های زیگوت برای یک نشانگر، ۴، ۳ و ۲ فنوتیپ بوده و بستگی به انواع تلاقی نشانگرها دارد (شکل ۳-۳ را مشاهده کنید). می‌توانیم ماتریس‌های توافق مختلفی (I) را طراحی کنیم تا ژنوتیپ‌های زیگوت را به فنوتیپ‌های زیگوت برای انواع مختلف نشانگرهای ارایه شده در جدول ۳-۲ مرتبط سازیم؛ بنابراین، ماتریس فراوانی توأم دو نشانگر و یک QTL برای انواع خاص نشانگرها می‌تواند با استفاده از ماتریس‌های توافق مربوطه به‌دست آید که عبارت است از:

$$\dot{G}=\left(I_{M_1}\otimes I_{M_2}\right)G$$

در این رابطه، $\otimes$ ضرب کرونکر و I_{M_1} و I_{M_2} به ترتیب ماتریس‌های توافق برای نشانگرهای M_1 و M_2 هستند. بردار فراوانی ژنوتیپ‌های دو نشانگر نیز می‌تواند به شیوه مشابهی به‌دست آید.

الگو و ساختار یک ماتریس توافق (I) که ژنوتیپ‌های زیگوت را به فنوتیپ‌های زیگوت مرتبط می‌سازد، بستگی به نوع تلاقی این نشانگر دارد. اگر a، b، c و d آلل‌های یک نشانگر بین دو والد دگربارور $P_1\times P_2$ باشند که همگی نسبت به آلل صفر (o) غالب هستند، داریم:

$$I=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

و برای نوع تلاقی نشانگری $ab \times cd$، داریم:

$$I = \begin{cases} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \text{برای } ab \times ab \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} & \text{برای } ab \times ao \end{cases}$$

$$I = \begin{cases} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \text{برای } ao \times ab \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \text{برای } ab \times ao \end{cases}$$

$$I = \begin{cases} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \text{برای } ab \times ao \\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \text{برای } ab \times ao \end{cases}$$

$$I=\begin{cases}\begin{bmatrix}1&1&1&0\\0&0&0&1\end{bmatrix} & \text{برای } ab\times ao\\ \begin{bmatrix}1&1&0&1\\0&0&1&0\end{bmatrix} & \text{برای } ab\times ao\\ \begin{bmatrix}1&0&1&1\\0&1&0&0\end{bmatrix} & \text{برای } ab\times ao\\ \begin{bmatrix}1&0&0&0\\0&1&1&1\end{bmatrix} & \text{برای } ab\times ao\end{cases}$$

$$I=\begin{bmatrix}1&1&0&0\\0&0&1&1\end{bmatrix}$$

$$I=\begin{bmatrix}1&0&1&0\\0&1&0&1\end{bmatrix}$$

با در اختیار داشتن فراوانی‌های توأم نشانگر-QTL و فراوانی‌های نشانگر برای نوع معینی از نشانگر، احتمال شرطی ($\omega_{uv|i}$) یک ژنوتیپ QTL مشروط به ژنوتیپ نشانگر برای فرد i ام می‌تواند محاسبه شود. این حالت به‌عنوان یک اصل برای مکان‌یابی QTL در گونه‌های دگربارور استفاده می‌شود.

۱۴-۲-۶- آزمون‌های فرض

وجود یک QTL با اثر معنی‌دار در داخل یک فاصله نشانگری، می‌تواند با محاسبه آماره آزمون نسبت درستنمایی (LR) تحت فرض صفر (H_0: عدم وجود QTL) و فرض مقابل (H_1: وجود QTL) به صورت زیر بیان شود:

$$LR=-2\left[\log L_0\left(\mu_{uv}=\tilde{\mu},\tilde{\sigma}^2,\tilde{p},\tilde{q}\right)-\log L_1\left(\hat{\Omega}\right)\right]$$

چون موقعیت QTL تحت فرض صفر قابل شناسایی نیست، LR تحت فرض صفر ممکن است به‌صورت مجانبی دارای توزیع x^2 با چهار درجه آزادی نباشد. چرچیل و دورج (۱۹۹۴) روش آزمون تبدیل یا جایگشت را برای تعیین آستانه بحرانی به منظور اعلام حضور یک QTL در نرخ خطای نوع اول مشخصی پیشنهاد کردند.

فرض‌های آزمون در ارتباط با آزمون معنی‌داری آثار ژنتیکی افزایشی ($H_0 : \alpha = \beta = 0$) و غالبیت ($H_0 : \gamma = 0$) می‌توانند در نظر گرفته شوند. آماره آزمون نسبت لگاریتم درستنمایی (LR) برای هر آزمون محاسبه می‌شود که در آن، آستانه‌های بحرانی می‌توانند یا با مطالعات شبیه‌سازی و یا از روی توزیع x^2 به ترتیب با دو یا یک درجه آزادی تعیین شوند، البته با این شرط که اندازه نمونه مورد استفاده به قدر کافی بزرگ باشد.

در یک خانواده تنی حاصل از دو والد دگربارور، ممکن است یک QTL فرضی بر اساس نسبت ۱:۱:۱:۱ تفرق نیابد. در عمل، الگوی تفرق QTL معنی‌دار باید آزمون شود، چون این امر برای طراحی راهبرد اصلاحی کارا و سودمند حائز اهمیت است. این که QTL دو آللی بوده و بر اساس نسبت ۱:۲:۱ یا ۱:۱ تفرق می‌یابد، می‌تواند آزمون شود. برای مثال، این فرض که QTL معنی‌دار با نوع تفرق ab × ab مطابقت دارد، می‌تواند با فرض‌های زیر آزمون شود:

$$H_0 : \alpha = \beta$$
$$H_1 : \alpha \neq \beta$$

به‌طور مشابه، فرض لازم برای آزمون مطابقت تفرق QTL برای نوع $ab \times bb$ می‌تواند به صورت زیر فرموله شود:

$H_0 : \alpha = 0$ یا $\beta = 0$

H_1 : α و β برابر صفر نیستند

به‌طور معمول، مقادیر بحرانی برای تشخیص معنی‌داری این آثار بر اساس مطالعات شبیه‌سازی تعیین می‌شوند.

همچنین می‌توانیم آزمون فرض وقوع نوترکیبی‌های مضاعف یا تداخل نوترکیبی را به ترتیب با فرض‌های صفر زیر بیان کنیم:

$$H_0 : p_{121} = p_{212} = 0 \qquad (۹-۱۴)$$

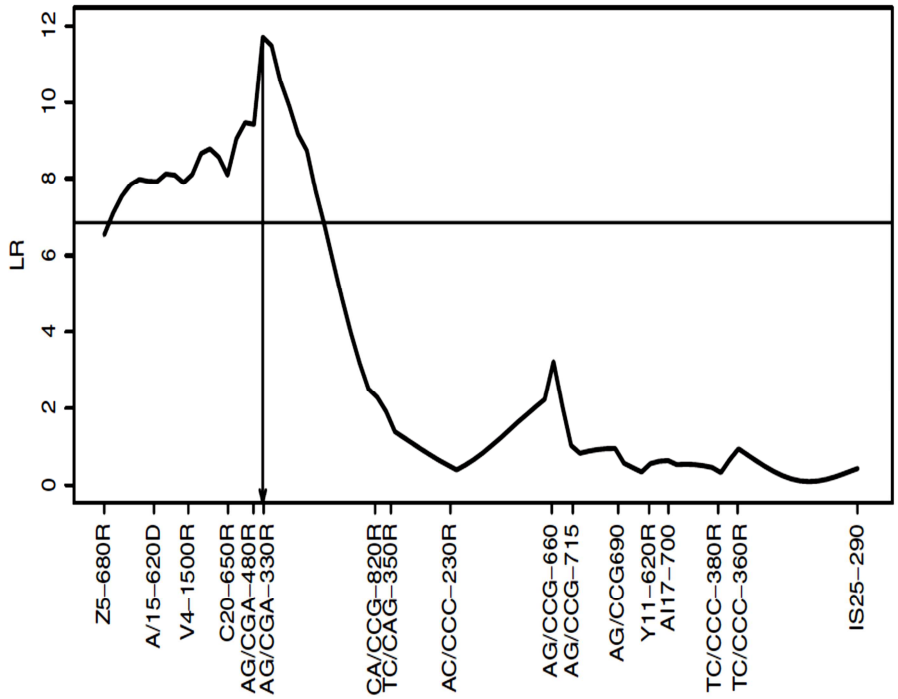
و:

$$H_0 : c = \frac{p_{121} + p_{212}}{\left(p_{121} + p_{212} + p_{122} + p_{211}\right)\left(p_{121} + p_{212} + p_{112} + p_{221}\right)} = 1 \qquad (۱۰-۱۴)$$

مقادیر LR تحت هر یک از دو فرض صفر (۱۴-۹) و (۱۴-۱۰) و فرض‌های مقابل آن‌ها محاسبه می‌شوند که هر یک از آن‌ها می‌توانند به صورت مجانبی دارای توزیع x^2 با یک درجه آزادی باشند.

مثال ۱۴-۱. اولین مثال‌ها در خصوص استفاده از یک تلاقی کنترل شده برای مکان‌یابی QTL درختان دگربارور، مطالعات مربوط به درخت صنوبر (برادشاو و استتلر، ۱۹۹۵) و اوکالیپتوس (گراتاپاگلیا و همکاران، ۱۹۹۵) و کاج loblolly (گروور و همکاران، ۱۹۹۴) هستند. در اینجا، از مثال مربوط به تلاقی بین دو گونه تبریزی، *Populus deltoids* و *P. euramericana*، استفاده می‌کنیم. نقشه پیوستگی ژنتیکی با استفاده از راهبردی که آمیزش آزمون کاذب نامیده می‌شود (گراتاپاگلیا و سدروف، ۱۹۹۴)، بر اساس ۹۰ ژنوتیپ که به‌طور تصادفی از خانواده هیبرید بین گونه‌ای F_1 انتخاب شده بودند، با استفاده از DNA های چندشکل تکثیر یافته تصادفی (RAPD)، چندشکلی‌های طولی قطعات تکثیر شده (AFLP) و توالی‌های تکراری ساده (ISSR) (یین و همکاران، ۲۰۰۲) تهیه شد. این نقشه شامل ۱۹ عدد از بزرگ‌ترین گروه‌های پیوستگی برای هر یک از نقشه‌های والدینی بود که به‌طور خام ۱۹ جفت کروموزوم را نشان می‌داد. ۹۰ ژنوتیپ دورگ که برای ایجاد نقشه استفاده شدند، با استفاده از نمونه‌های چوب جمع‌آوری شده از ساقه‌های ۱۱ ساله، در یک آزمایش مزرعه‌ای از نظر چگالی چوب اندازه‌گیری شدند. اندازه‌گیری هر ژنوتیپ ۴ تا ۶ بار تکرار شد تا خطای اندازه‌گیری کاهش یابد. میانگین این ژنوتیپ‌ها محاسبه و برای مکان‌یابی QTL استفاده شد.

QTL معنی‌دار برای چگالی چوب روی گروه پیوستگی D17 تشخیص داده شد که به وسیله یین و همکاران (۲۰۰۲) گزارش شد. در این مثال، برآورد تجربی مقدار بحرانی از روی ۱۰۰۰ آزمون جایگشت به‌دست آمد. مشخص شد که مقدار بحرانی برای اعلام حضور QTL روی کل گروه پیوستگی مورد نظر در سطح معنی‌داری ۵ درصد برابر ۶/۹ است. پروفیل LR برای مدل کامل در مقابل مدل کاهش یافته در طول گروه پیوستگی D17 دارای نقطه اوج تندی بین فاصله نشانگری باریک AG/CGA-480-AG/CGA-330 است (شکل ۱۴-۱). مقدار LR در این نقطه اوج ۱۱/۷ است که به خوبی از آستانه بحرانی تجربی در سطح معنی‌دار ۰/۰۵ بیشتر است.



شکل ۱۴-۱- پروفیل آماره آزمون نسبت لگاریتم درستنمایی (LR) برای شناسایی QTL در طول گروه پیوستگی D17 در مطالعه یین و همکاران (۲۰۰۲) با استفاده از تجزیه فاز ترکیبی. آستانه بحرانی بر اساس آزمون‌های جایگشت (چرچیل و دورج، ۱۹۹۴) روی خط افقی نشان داده شده است. اسامی نشانگرها در طول گروه پیوستگی در قسمت زیرین پروفیل ارایه شده‌اند. اقتباس از لین و همکاران (۲۰۰۳).

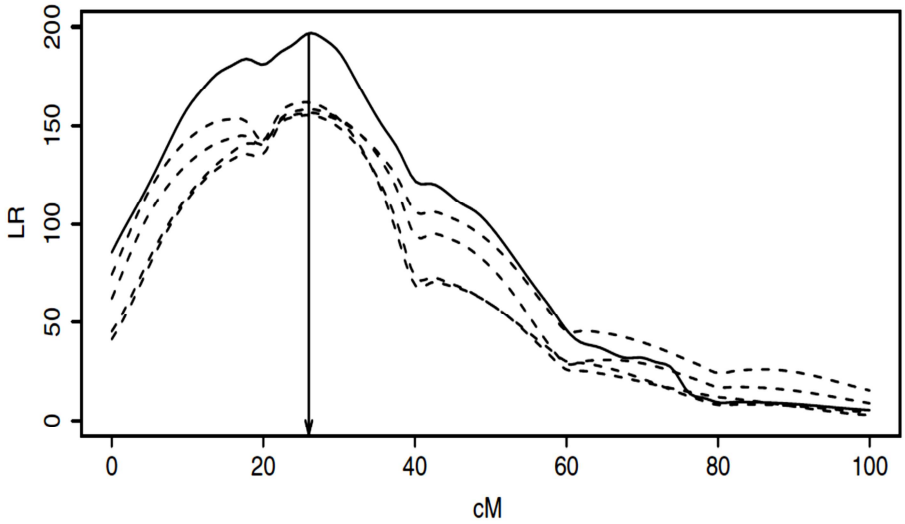
اثر افزایشی QTL معنی‌دار شناسایی شده ۰/۰۳۳ یا معادل ۷ درصد نسبت به میانگین کل است. این QTL در حدود ۳۰ درصد از واریانس فنوتیپی را برای چگالی چوب در تبریزی‌های دورگ توجیه می‌کند. MLE احتمال فاز یعنی p برابر با ۰/۸۲ است که پیشنهاد می‌کند به احتمال زیاد فاز پیوستگی Φ_{11} وجود دارد. این امر نشان می‌دهد که آلل مثبت این QTL که چگالی چوب را افزایش می‌دهد، در احتمال ۰/۸۲ با آلل‌های غالب دو نشانگر AG/CGA-480 و AG/CGA-330 که در مجاورت QTL واقع شده‌اند، در فاز جفت یا ترانس قرار دارد.

مواد مشابهی با استفاده از روش مکان‌یابی فاصله‌ای مرسوم که فاز پیوستگی ممکن QTL-نشانگر را به‌طور همزمان در نظر می‌گیرد، تجزیه شد. این روش که بر اساس فاز پیوستگی تفکیک شده است می‌تواند QTL معنی‌دار را برای چگالی چوب شناسایی کند، ولی نمی‌تواند فاز پیوستگی

صحیح را تعیین نماید، چون حداکثر مقادیر LR بین دو فاز پیوستگی ممکن، مشابه هستند. روش ما اطلاعات مهمی در خصوص ترتیب‌های غیر آللی روی کروموزوم‌های همولوگ فراهم می‌کند.

۱۴-۲-۷- اثر فازهای پیوستگی

یک مطالعه شبیه‌سازی به منظور آزمون اثر خصوصیات نادرست فاز پیوستگی روی شناسایی و برآورد پارامترهای QTL انجام شد. یک خانواده تنی برای شش نشانگر حاوی اطلاعات کامل که با فواصل یکسان (۲۰ سانتی‌مورگان) از هم قرار داشتند و پنج فاصله را تشکیل می‌دادند، شبیه‌سازی شد. فرض شد که یک QTL در ۲۶ سانتی‌مورگانی از اولین نشانگر (در داخل دومین فاصله) قرار دارد. ارزش‌های فنوتیپی برای این خانواده تنی با در نظر گرفتن یک مجموعه مخصوص از پارامترهای اثر QTL ناشناخته تحت ترکیب فاز والدینی Φ_{11} برای دو والد شبیه‌سازی می‌شوند. اندازه نمونه شبیه‌سازی شده ۴۰۰ و وراثت‌پذیری صفت برابر با ۰/۴ است.



شکل ۱۴-۲- پروفیل آماره‌های آزمون نسبت لگاریتم درستنمایی (LR) حاصل از یک تکرار تصادفی شبیه‌سازی برای تشخیص QTL در طول یک گروه پیوستگی تحت یک فاز مختلط (منحنی پیوسته) و چهار فاز جداگانه (منحنی‌های نقطه چین). وراثت‌پذیری صفت ۰/۴= H^2 و اندازه نمونه ۴۰۰ فرد است. باید توجه کرد که مجموعه مشابهی از داده‌های شبیه‌سازی شده با شرط $\alpha = 0.5$، $\beta = 0.5$ و $\gamma = 0.5$، برای این ۵ تجزیه متفاوت (یک فاز مختلط و چهار فاز جداگانه) استفاده می‌شود. اقتباس از لین و همکاران (۲۰۰۳).

داده‌های شبیه‌سازی شده تحت ترکیب فاز پیوستگی Φ_{11} با استفاده از مدل‌های مبتنی بر این فاز و سه فاز متفاوت دیگر یعنی Φ_{12}، Φ_{21} و Φ_{22} تجزیه شدند. چون فازهای پیوستگی مختلف فقط ترتیب جفت شدن کروموزوم والدینی را تغییر می‌دهند، مقادیر حداکثر LR حاصل از فاز پیوستگی صحیح Φ_{11} و سه فاز پیوستگی غیر صحیح Φ_{12}، Φ_{21} و Φ_{22} یکسان خواهند شد (شکل ۱۴-۲ را مشاهده کنید) که بیان می‌کند تجزیه‌های مبتنی بر فازهای جداگانه، هیچ توانی برای انتخاب محتمل‌ترین فازهای پیوستگی ندارند. همچنین، همان طور که به وسیله منحنی‌های پهن و خمیده نشان داده شده است، حداکثر مقدار LR از مدل فاز پیوستگی منفرد نمی‌تواند برای تعیین موقعیت دقیق QTL استفاده شود. شکل ۱۴-۲ مقادیر LR را در طول گروه پیوستگی در زمانی نشان می‌دهد که تمامی ترکیبات فاز پیوستگی به‌طور همزمان بر اساس داده‌های شبیه‌سازی شده مشابه در نظر گرفته می‌شوند. نقطه اوج بیشتر منحنی برای تجزیه فاز مختلط نشان می‌دهد که مدل مختلط برای تشخیص QTL معنی‌دار نسبت به تجزیه‌های جداگانه فازها بهتر است. زمانی که فاز پیوستگی مورد استفاده صحیح نباشد، علامت MLE های آثار افزایشی و غالبیت QTL بر عکس خواهد شد.

مثال ۱۴-۲. والشلگر و همکاران (۲۰۰۵) مکان‌یابی QTL را برای اجزای زیست توده (بیوماس) در یک جمعیت تلاقی برگشتی، $(T\times D)\times D$، حاصل از دو گونه صنوبر، *Populus trichocarpa Torr. & Gray* (T) و *P. deltoids Bartr.* (D) انجام دادند. جمعیت مکان‌یابی یک جمعیت تلاقی برگشتی با ۱۷۱ درخت بود که برای نشانگرهای ریزماهواره (SSR) و AFLP تعیین ژنوتیپ شده بودند. به دلیل ماهیت هتروزیگوت درختان صنوبر، چند نوع نشانگر مورد ارزیابی و مشاهده قرار گرفتند. از بین آن‌ها، نشانگرهای آمیزش آزمونی که در والد F_1 در حال تفرق بوده ولی در والد D در حال تفرق نبودند، شامل ۹۲ نشانگر SSR و ۵۵۶ نشانگر AFLP بودند. از این‌رو، نقشه پیوستگی ژنتیکی بر اساس والد F_1 از روی این نشانگرها با استفاده از استراتژی آمیزش آزمون کاذب (یین و همکاران، ۲۰۰۴) تهیه شد. این نقشه شامل ۱۹ گروه پیوستگی بود که معادل با تعداد کروموزوم‌های درخت صنوبر است.

اگر قصد داشته باشیم که یک QTL بین تلاقی[1] ($Qq\times Qq$) را با استفاده از یک جفت نشانگر آمیزش آزمون ($TD\times DD$) مکان‌یابی کنیم، والد F_1 ممکن است دو دیپلوتیپ به صورت

---

[1]- منظور از QTL بین تلاقی یا Intercross QTL، QTL هایی است که در یک جمعیت حاصل از تلاقی دو ژنوتیپ هتروزیگوت یکسان (مثلاً جمعیت F_2 حاصل از تلاقی بین ژنوتیپ‌های هتروزیگوت F_1 یا حاصل از خودگشنی ژنوتیپ‌های F_1) وجود دارد (مترجمین).

$[TQT][DqD]$ و $[TqT][DDD]$ داشته باشد. این دو دیپلوتیپ به ترتیب با احتمال p و p-۱ در مدل مکان‌یابی QTL وارد می‌شوند که منجر به شناسایی تعداد اندکی QTL معنی‌دار مؤثر بر صفات اجزای زیست توده در درختان صنوبر می‌شوند (جدول ۱۴-۲).

یک QTL بین تلاقی برای زیست توده ساقه در جایگاه نشانگر P_204_C2 در گروه پیوستگی ۶ مکان‌یابی شد، حال آنکه یک QTL بین تلاقی برای زیست توده ساقه، درصد زیست توده برگ، زیست توده روزمینی و کل زیست توده در مجاورت نشانگر T4_10 در گروه پیوستگی ۱۳ مشاهده شد.

۱۴-۳- مدل مکان‌یابی با آثار تصادفی در یک شجره پیچیده

۱۴-۳-۱- مقدمه

در مدل مختلط (۹-۴)، فرض می‌کنیم که اجزای نرمال مختلف به وسیله تعدادی از ژنوتیپ‌های شناخته شده یا ناشناخته QTL مشخص می‌شوند. آثار ژنتیکی QTL های فرضی روی فنوتیپ که از توزیع نرمال تبعیت می‌کنند، می‌توانند مستقیماً با استفاده از روش مدل آثار ثابت برآورد شوند. این روش تحت شرایطی مفید است که آثار ژنتیکی اساسی (همانند تلاقی‌های کنترل شده با اندازه بزرگ حاصل از لاین‌های خودبارور یا دگربارور) بتوانند به آسانی مشخص شوند (لندر و بوتستین، ۱۹۸۹؛ هالی و نوت، ۱۹۹۲؛ زنگ، ۱۹۹۴؛ لین و همکاران، ۲۰۰۳)؛ اما در بسیاری از جوامع دگربارور، اندازه یک خانواده منفرد می‌تواند محدود باشد، به‌طوری که تجزیه مرکب چند خانواده حاصل از والدین غیرخویشاوند یا خویشاوند صرفاً یک انتخاب محسوب می‌شود. چون این والدین هتروزیگوت هستند و در آن‌ها تعداد آلل‌های موجود در یک جایگاه ژنی و فاز پیوستگی جایگاه‌های ژنی مختلف ناشناخته است، تخصیص و تعیین آثار ژنتیکی مشکل است. یک روش قدرتمند مبتنی بر روش مدل آثار تصادفی (ژو و آچلی، ۱۹۹۵) که به دیپلوتیپ‌های والدینی و تعداد آلل‌های QTL وابسته نیست، می‌تواند مورد استفاده قرار گیرد.

تحت مدل ثابت، تمامی آثار موجود در میانگین مورد انتظار توزیع نرمال، ثابت بوده و می‌توانند به‌طور مستقیم از روی مدل مختلط برآورد شوند. از آنجایی که آثار ثابت دارای واریانس نیستند، از این‌رو واریانس فنوتیپی فقط شامل واریانس باقیمانده درون ژنوتیپ QTL است؛ اما در مدل آثار تصادفی، تمامی آثار موجود در میانگین مورد انتظار نمی‌توانند برآورد شوند، چون امیدهای ریاضی آن‌ها صفر است. ولی واریانس‌های آن‌ها می‌تواند با تفکیک واریانس یا کوواریانس فنوتیپی کل به اجزای مربوطه برآورد شوند؛ بنابراین، در جایی که روش مدل ثابت، اثر جایگزینی آللی (یا اثر آللی) را برآورد می‌کند، روش مکان‌یابی مدل تصادفی، واریانس QTL در حال تفرق را برآورد می‌کند.

جدول ۱۴-۳- شناسایی QTL های بین تلاقی و منشأ والدینی آلل‌های مطلوب QTL برای اجزای زیست‌توده در سال دوم در مزرعه برای یک خانواده هیبرید صنوبر.

صفت	گروه پیوستگی	موقعیت	فاز	$\hat{p}$	$\hat{a}$	$\hat{d}$	LR
زیست توده ساقه	۶	P_204_C2	DQT	۰/۷۱	۰/۱۵	-۰/۰۲	۲۹/۵۱
زیست توده شاخه	۱۳	T4_10	DQT	۰/۸۲	۰/۹۲	۰/۲۳	۳۵/۰۳
درصد برگ	۱۳	T4_10	DQT	۰/۸۱	۰/۷۹	۰/۱۸	۳۱/۰۳
زیست توده روزمینی	۱۳	T4_10	DQT	۰/۸۲	۰/۸۳	۰/۱۹	۳۲/۸۴
زیست توده کل	۱۳	T4_10	DQT	۰/۸۱	۰/۸۰	۰/۱۹	۳۱/۴۷

اساس نظری روش مدل تصادفی، شباهت فنوتیپی (یا کوواریانس) بین افرادی است که از نظر ژنتیکی خویشاوند هستند. واریانس فنوتیپی یک فرد حاصل از خانواده تنی دگربارور می‌تواند به واریانس ژنتیکی ناشی از QTL فرضی، واریانس ناشی از اثر خاص خانواده و واریانس مربوط به اثر باقیمانده تفکیک شود. از آنجایی که آثار باقیمانده بین برادر-خواهران می‌تواند مستقل در نظر گرفته شود، از این‌رو کوواریانس فنوتیپی بین دو فرد مختلف در درون یک خانواده باز هم می‌تواند به کوواریانس ژنتیکی ناشی از QTL و کوواریانس ناشی از اثر خاص خانواده تفکیک شود.

۱۴-۳-۲- مدل آماری

چند خانواده خویشاوند هر یک با تعداد متفاوتی برادر-خواهر را در نظر بگیرید. صفت کمّی (y) برای هر برادر-خواهر در درون هر خانواده اندازه‌گیری می‌شود. ارزش فنوتیپی برادر-خواهر j ام ($j = 1,\ldots,m_i$) در درون خانواده i ام ($i = 1,\ldots,n$)، به‌عنوان تابعی خطی از K عدد QTL و متغیرهای همبسته (کوواریت) ثابت دیگر بیان می‌شوند:

$$y_{ij} = \mu + \sum_{k=1}^{K} \alpha_k + \sum_{l=1}^{L} \beta_l X_l + e_{ij} \qquad (۱۱-۱۴)$$

در این رابطه، μ میانگین کل، α_k اثر k امین QTL، X_l ها برخی از متغیرهای همبسته (کوواریت) نظیر جنسیت یا سن، β_l اثر l امین متغیر همبسته که فرض می‌شود با خطاهای ژنتیکی و محیطی همبستگی ندارد و e_{ij} خطای محیطی تصادفی است. اندازه کل نمونه نیز به صورت $N = \sum_{i=1}^{n} m_i$ است.

فرض کنید یک QTL مؤثر بر صفت که دارای اثر افزایشی است، وجود دارد. چنین مدل QTL منفرد می‌تواند به‌صورت زیر بیان شود:

$$y_{ij} = \mu + a_{ij} + d_{ij} + g_{ij} + \sum_{l=1}^{L} \beta_l X_l + e_{ij} \qquad (۱۲-۱۴)$$

که در آن، $a_{ij} \sim N\left(0, \sigma_a^2\right)$ اثر ژنتیکی افزایشی، $d_{ij} \sim N\left(0, \sigma_d^2\right)$ اثر ژنتیکی غالبیت و $g_{ij} \sim N\left(0, \sigma_g^2\right)$ اثر افزایشی چند ژنی است که آثار ژن‌های ناپیوسته یا آثار خانوادگی دیگر از جمله عوامل محیطی مشترک فامیلی (به غیر از QTL مورد نظر) را نشان می‌دهد و $e_{ij} \sim N\left(0, \sigma_e^2\right)$ خطای محیطی است.

با فرض اینکه a_{ij}، d_{ij}، g_{ij} و e_{ij} متغیرهای تصادفی غیرهمبسته هر یک با امید ریاضی صفر هستند، از این‌رو واریانس کل برای یک مشاهده منفرد (y_{ij}) برابر است با:

$$\mathrm{var}\left(y_{ij}\right) = \sigma_a^2 + \sigma_d^2 + \sigma_g^2 + \sigma_e^2$$

کوواریانس بین دو برادر-خواهر j و j' در خانواده i نیز برابر است با:

$$\mathrm{cov}\left(y_{ij}, y_{ij'}\right) = \pi_{ia}\sigma_a^2 + \pi_{id}\sigma_d^2 + \phi_{ig}\sigma_g^2$$

که در آن، π_{ia} نسبت آلل‌های یکسان اجدادی (IBD) مشترک بین افراد فامیل j و j'، π_{id} متغیر دو وضعیتی است که نشان می‌دهد آیا j و j' آلل‌های یکسان اجدادی (IBD) را دریافت کرده‌اند یا خیر و ϕ_{ig} نسبت مورد انتظار IBD مشترک است که مطابق انتظار برای جفت‌های تنی معادل ۰/۵ است؛ بنابراین، ماتریس واریانس-کوواریانس کل برای y به‌صورت زیر است:

$$\Sigma = \Pi_a\sigma_a^2 + \Pi_d\sigma_d^2 + \Phi_g\sigma_g^2 + I\sigma_e^2 \qquad (۱۳-۱۴)$$

که در آن، Π_a ماتریس نسبت آلل‌های IBD مشترک نشانگر، Π_d ماتریس متغیر دو وضعیتی π_{id} و Φ_g ماتریس نسبت مورد انتظار آلل‌های IBD مشترک و I ماتریس واحد است.

ماتریس کوواریانس برای دو برادر-خواهر j و j' مربوط به یک خانواده مشابه را می‌توان به‌صورت زیر ارائه نمود:

$$\Sigma_i = Var\begin{bmatrix} y_{ij} \\ y_{ij'} \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix} = \sigma^2 R_i \qquad (۱۴-۱۴)$$

که σ^2 واریانس فنوتیپی کل بوده و داریم:

$$\rho_i = \pi_{ia}\left(\frac{\sigma_a^2}{\sigma^2}\right) + \pi_{id}\left(\frac{\sigma_d^2}{\sigma^2}\right) + \phi_{ig}\left(\frac{\sigma_g^2}{\sigma^2}\right) \qquad (۱۵-۱۴)$$

$$= \pi_{ia}h_a^2 + \pi_{id}h_d^2 + \phi_{ig}h_g^2$$

که در آن، σ_a^2، σ_d^2 و σ_g^2، همان طور که تعریف شدند، به ترتیب واریانس‌های افزایشی و غالبیت QTL و واریانس ژنتیکی چند ژنی بوده و h_a^2، h_d^2 و h_g^2 به ترتیب وراثت‌پذیری‌های افزایشی و غالبیت برای QTL فرضی و سهم واریانس فنوتیپی توجیه شده به وسیله اثر چند ژنی هستند. متغیرهای IBD دارای توزیع‌های گسسته هستند:

$$\pi_{ia} = \begin{cases} 0 & \text{اگر هیچ آللی IBD نیست} \\ 0.5 & \text{اگر یک آلل IBD است} \\ 1 & \text{اگر هر دو آلل IBD است} \end{cases}$$

و:

$$\pi_{id} = \begin{cases} 1 & \text{اگر هر دو آلل IBD است} \\ 0 & \text{در غیر این صورت} \end{cases}$$

معادله (۱۴-۱۵) برای مرتبط کردن افراد متعلق به خانواده‌های متفاوت ولی خویشاوند i و i' می‌تواند بسط داده شود که در این صورت ϕ_{ig} باید به صورت $\phi_{ii'g}$ نشان داده شود. تعیین $\phi_{ii'g}$ باید بر اساس نظریه ژنتیک کمّی انجام شود. برای مثال، $\phi_{ii'g} = \frac{1}{4}$ برای دو برادر-خواهر ناتنی، $\frac{1}{8}$ برای دو عموزاده و غیره (فالکونر و مک کی، ۱۹۹۶).

۱۴-۳-۳- IBD در یک QTL

چون ژنوتیپ‌های یک QTL نمی‌توانند مستقیماً مشاهده شوند، از این‌رو نیاز داریم تا از اطلاعات نشانگرهای مشاهده شده برای استنباط ژنوتیپ‌های پیوسته به نشانگرها استفاده کنیم. با استفاده از احتمال توأم دو جایگاه ژنی نشان داده شده است که IBD مورد انتظار یک جایگاه می‌تواند به صورت تابع خطی از IBD جایگاه دیگر بیان شود (هاسمن و الستون، ۱۹۷۲)، یک مدل خطی برای بیان IBD یک QTL فرضی (π_{ia}) بر حسب IBD دو نشانگر مجاور M_1 و M_2 ارایه شده است (فالکر و کاردون، ۱۹۹۴). فرض کنید r_1، r_2 و r به ترتیب نرخ‌های نوترکیبی بین نشانگر

M_1 و QTL، QTL و نشانگر M_2 و دو نشانگر باشند. فالکر و کاردون (۱۹۹۴) رابطه زیر را به‌دست آوردند:

$$\pi_{ia} = a + b_1\pi_1 + b_2\pi_2 \quad (۱۶-۱۴)$$

که در آن، π_1 و π_2 به ترتیب IBD های دو نشانگر هستند،

$$b_1 = \frac{(1-2r_1)^2 - (1-2r_2)^2 (1-2r)^2}{1-(1-2r)^4}$$

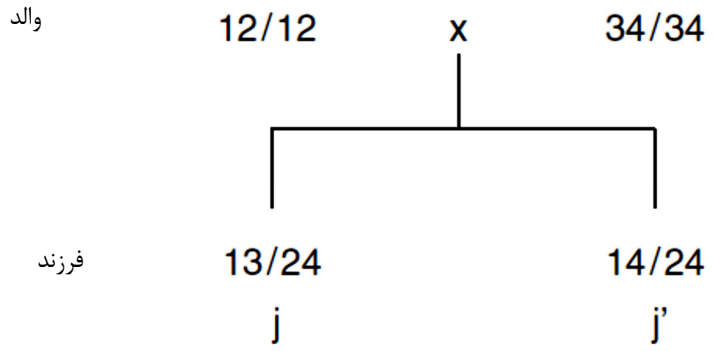
$$b_2 = \frac{(1-2r_2)^2 - (1-2r_1)^2 (1-2r)^2}{1-(1-2r)^4}$$

$$a = \frac{(1-b_1-b_2)}{2}$$

سپس IBD برای QTL در نظر گرفته شده در معادله (۱۴-۱۵) به وسیله معادله (۱۴-۱۶) جایگزین می‌شود.

مثال ۱۴-۳. دو والد دگربارور به ترتیب با ژنوتیپ‌های ۱۲/۱۲ و ۳۴/۳۴ در دو نشانگر را در نظر بگیرید. این دو والد برای تولید دو فرزند j و j' با ژنوتیپ‌های دو نشانگری ۱۳/۲۴ و ۱۴/۲۴ تلاقی داده می‌شوند (شکل ۱۴-۳ را مشاهده کنید). ضرایب IBD آلل‌های مشترک بین برادر-خواهر j و j' برابر با $\pi_1 = 0.5$ برای نشانگر A و $\pi_2 = 1$ برای نشانگر B هستند. فرض کنید یک QTL با نرخ‌های نوترکیبی $r_1 = 0.05$ و $r_2 = 0.10$ بین این دو نشانگر وجود دارد. از این‌رو، نرخ نوترکیبی بین دو نشانگر با این فرض که هیچ تداخلی برای کراس‌اور در میوز وجود نداشته باشد، برابر با ۰/۱۴ است. بر اساس معادله (۱۴-۱۶)، IBD مربوط به QTL به‌صورت زیر محاسبه می‌شود:

$$\pi_{ia} = \left[1 - \frac{(1-2\times0.05)^2 - (1-2\times0.10)^2(1-2\times0.14)^2}{1-(1-2\times0.14)^4} - \frac{(1-2\times0.10)^2 - (1-2\times0.05)^2(1-2\times0.14)^2}{1-(1-2\times0.14)^4}\right] + \frac{(1-2\times0.05)^2 - (1-2\times0.10)^2(1-2\times0.14)^2}{1-(1-2\times0.14)^4}\times0.5 + \frac{(1-2\times0.10)^2 - (1-2\times0.05)^2(1-2\times0.14)^2}{1-(1-2\times0.14)^4}\times1 = 0.6730$$



شکل ۱۴-۳- دیاگرام مربوط به تفرق نشانگر بین دو والد دگربارور.

۱۴-۳-۴- درستنمایی

خانواده‌های غیر خویشاوند

درستنمایی مشاهدات برای یک شجره می‌تواند بر حسب ساختار آن تهیه شود. دو نوع شجره در نظر می‌گیریم که در یکی، خانواده‌ها نسبت به یکدیگر خویشاوند نبوده و در دیگری، خانواده‌ها با مقادیر خویشاوندی مختلف با یکدیگر خویشاوند هستند. یکی از n خانواده غیر خویشاوند (i) که دارای m_i برادر-خواهر است را در نظر بگیرید. همان طور که در معادله (۱۴-۴) تعریف شد، R_i در این مورد یک ماتریس $m_i \times m_i$ است که روی قطرهای آن ۱ و در خارج قطر ρ_i قرار دارد. از این‌رو، تحت فرض نرمال، تابع چگالی احتمال به‌صورت زیر است:

$$f_i\left(y_i\right)=\frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}\left|R_i\right|}\exp\left[-\frac{1}{2\sigma^2}\left(Z_i^T R_i^{-1} Z_i\right)\right] \qquad (۱۷-۱۴)$$

کـه در آن، $Z_i = y_i - 1\mu - \sum_{l=1}^{L}\beta_l X_l$ ، $y_i = \left(y_1,...,y_{m_i}\right)^T$ یـک بـردار $\left(m_i \times 1\right)$ از فنوتیپ‌ها و 1 یک بردار $\left(m_1 \times 1\right)$ است که تمامی عناصر آن ۱ اسـت. بـرای n خـانواده مسـتقل، لگاریتم درستنمایی کلی نیز به شکل زیر است:

$$L\left(\Omega \mid y\right)=\log\left[f_1\left(y_i\right)\right]+...+\log\left[f_n\left(y_n\right)\right] \qquad (۱۸-۱۴)$$

پارامترهای ناشناخته $\Omega=\left(\mu,\beta_l,\sigma^2,h_a^2,h_d^2,h_g^2,r_1,r_2\right)$ موجود در درستنمایی (۱۴-۱۸) می‌توانند با استفاده از روش حداکثر درستنمایی برآورد شوند. به‌طور معمول، برآورد موقعیت QTL می‌تواند با در نظر گرفتن مقادیر ثابت و مشخص r_1 یا r_2 و تغییر آن در سرتاسر فاصـله نشـانگری بر اساس روش شبکه‌ای انجام شود. با مشتق‌گیری از تابع لگاریتم درستنمایی نسبت بـه μ ، β_l و σ^2 ، MLE ها به‌صورت زیر به‌دست می‌آیند:

$$\hat{\mu}=\left[\sum_{i=1}^{n}1^T R_i^{-1}1\right]^{-1}\left[\sum_{i=1}^{n}1^T R_i^{-1}y_i\right]$$

$$\hat{\beta}_l=\left(1^T R_i^{-1}1\right)^{-1}\left(1^T R^{-1}y\right)$$

$$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{n}\hat{Z}^T R_i^{-1}\hat{Z}_i$$

که در آن، $\hat{Z}_i = y_i - 1\hat{\mu} - \sum_{l=1}^{L}\hat{\beta}_l X_l$ و $N=\sum_{i=1}^{n}m_i$ هستند. با وارد کـردن $\hat{\mu}$ ، $\hat{\beta}_l$ و $\hat{\sigma}^2$ در تابع لگاریتم درستنمایی، MLE های مربوط به h_a^2 ، h_d^2 و h_g^2 با اسـتفاده از الگـوریتم سیمپلکس (نلدر و مید، ۱۹۶۵) برآورد می‌شوند.

خانواده‌های خویشاوند

اگر تمامی خانواده‌ها در شجره با یکدیگر خویشاوند باشند، تابع درستنمایی داده‌هـای فنـوتیپی (y) تحت فرض نرمال چندمتغیره به‌صورت زیر است:

$$L\left(\Omega \mid y\right)=\left(2\pi\right)^{\frac{N}{2}}\left|\Sigma\right|^{\frac{1}{2}}\exp\left[-\frac{1}{2}\left(Z^T\Sigma^{-1}Z\right)\right] \qquad (۱۹-۱۴)$$

در این رابطه، N تعداد کل نتاج حاصل از تمامی خانواده‌ها است، $Z=\{Z_i\}_{i=1}^{n}$، $y=\{y_{im_i}\}_{i=1}^{n}$ یک بردار $(N\times 1)$ از تمامی فنوتیپ‌ها است، 1 یک بردار $(N\times 1)$ است که تمامی عناصر آن ۱ هستند و:

$$\Sigma=\sigma^2\left(\Pi_a h_a^2+\Pi_d h_d^2+\Pi_g h_g^2+I\right)=\sigma^2 R \qquad (۲۰-۱۴)$$

که در آن، Π_a، Π_d و Π_g به ترتیب ماتریس‌های $(N\times N)$ برای ضرایب آلل‌های IBD در QTL مشترک بین هر جفت از افراد، متغیرهای دو وضعیتی که نشان می‌دهند چگونه آلل‌های IBD بین هر جفت از افراد مشترک هستند و ضرایب آلل‌های IBD به‌طورکلی بین هر جفت از افراد موجود در شجره تقسیم می‌شوند. با مشتق‌گیری از لگاریتم درستنمایی نسبت به μ، β_l و σ_e^2، MLE های مربوطه به‌صورت زیر به‌دست می‌آیند:

$$\hat{\mu}=\left(1^T H^{-1}1\right)^{-1}\left(1^T H^{-1}y\right)$$

$$\hat{\beta}_l=\left(1^T H^{-1}1\right)^{-1}\left(1^T H^{-1}y\right)$$

$$\hat{\sigma}^2=\frac{1}{N}\left(y-1\mu\right)^T H^{-1}\left(y-1\mu\right)$$

با وارد کردن $\hat{\mu}$، $\hat{\beta}_l$ و $\hat{\sigma}^2$ در تابع لگاریتم درستنمایی، MLE های مربوط به h_a^2، h_d^2 و h_g^2 با استفاده از الگوریتم سیمپلکس برآورد می‌شوند.

۱۴-۳-۵- آزمون فرض

پس از این که پارامترها برآورد شدند، فرض وجود QTL ها می‌تواند آزمون شود. این امر می‌تواند به‌صورت زیر انجام شود:

$$H_0: h_a^2=h_d^2=0$$
$$H_1: \text{حداقل یکی از وراثت‌پذیری های فوق برابر با صفر نیست} \qquad (۲۱-۱۴)$$

درستنمایی‌ها با در نظر گرفتن فرض‌های صفر ($L_0\left(\tilde{\Omega}\mid y\right)$) و مقابل ($L_1\left(\tilde{\Omega}\mid y\right)$) محاسبه می‌شوند که با استفاده از آن‌ها نسبت لگاریتم درستنمایی نیز محاسبه می‌شود:

$$LR=-2\left[\ln L_0\left(\tilde{\Omega}\mid y\right)-\ln L_1\left(\tilde{\Omega}\mid y\right)\right] \qquad (۲۲-۱۴)$$

که در این رابطه، $\tilde{\Omega}$ و $\hat{\Omega}$ به ترتیب MLE های پارامترها با در نظر گرفتن فرض‌های H_0 و H_1 هستند که H_0 به وسیله ژنوتیپ‌های نشانگر تحت‌تأثیر قرار نمی‌گیرد. مقدار بحرانی برای تأیید وجود یک QTL می‌تواند به‌صورت تجربی با آزمون‌های جایگشت تعیین شود. آزمون‌های مشابهی نیز می‌تواند برای واریانس‌های افزایشی یا غالبیت توجیه شده به وسیله QTLانجام شود. واریانس چند ژنی جهت توصیف نتایج حائز اهمیت است که معنی‌داری آن می‌تواند با تهیه نسبت لگاریتم درستنمایی مربوطه آزمون شود.

مثال ۱۴-۴. **(پروژه ژنوم سگ).** یک شجره سگ برای مکان‌یابی QTL مسئول دیسپلازی استخوان هیپ سگ[1] (CHD) با استفاده از نشانگرهای مولکولی تهیه شد. ۷ سگ پایه‌گذار Greyhounds و ۶ سگ پایه‌گذار از بازماندگان Labrador با یکدیگر تلاقی داده شدند و سپس تلاقی برگشتی F_1ها با Greyhounds و بازماندگان Labrador و نیز تلاقی بین F_1ها انجام شد. یک سری از تلاقی‌های بعدی در بین نتاج در نسل‌های مختلف منجر به تولید یک ساختار شجره‌ای پیچیده (شکل ۱۴-۴) شد که دامنه‌های فنوتیپی صفات کمّی مرتبط با CHD و شانس شناسایی QTL های در حال تفرق را به حداکثر رساند (بلیس و همکاران، ۲۰۰۲؛ تودهانتر و همکاران، ۲۰۰۵). در مجموع ۱۴۸ سگ حاصل از این جمعیت دگربارور برای ۲۴۰ نشانگر ریزماهواره مستقر روی ۳۸ جفت کروموزوم اتوزومی و یک جفت کروموزوم جنسی تعیین ژنوتیپ شدند (برین و همکاران، ۲۰۰۱). نقشه پیوستگی ژنوم سگ تهیه شده از این نشانگرها، پوشش خوبی از هر یک از کروموزوم‌ها را نشان می‌دهد. فواصل بین نشانگرهای مجاور در این نقشه پیوستگی بر حسب سانتی‌مورگان برآورد شد (برین و همکاران، ۲۰۰۱). سن سگ‌ها در هنگام تشخیص عارضه استخوانی شدن قسمت فوقانی ران (OSS) که یکی از معیارهای مهم برای ارزیابی CHD است، برای هر یک از سگ‌های مطالعه شده در سمت چپ و راست آن اندازه گیری شد.

از مدل مندلی با آثار تصادفی (۱۴-۱۲) برای مکان‌یابی OSS بر اساس نقشه پیوستگی تهیه شده با ۲۴۰ نشانگر استفاده خواهد شد. برای ساده کردن، اثر غالبیت QTL و آثار متغیرهای همبسته (کوواریت) نادیده گرفته شده و در نتیجه، مدل مربوطه به‌صورت زیر خواهد بود:

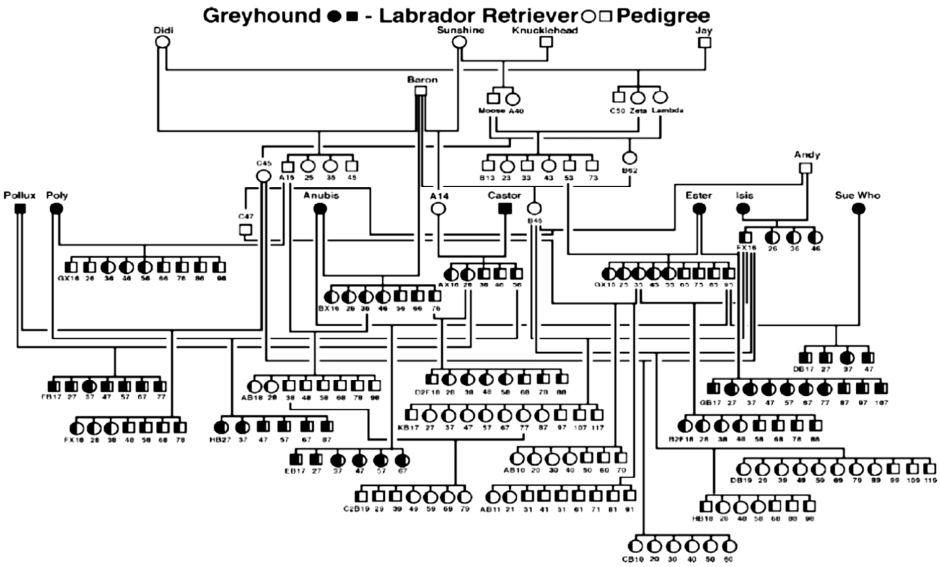
$$y_{ij} = \mu + a_{ij} + g_{ij} + e_{ij}$$

که اجازه برآورد میانگین کل (μ)، واریانس ژنتیکی توجیه شده به وسیله QTL (σ_a^2)، واریانس چند ژنی (σ_g^2) و واریانس باقیمانده (σ^2) را می‌دهد.

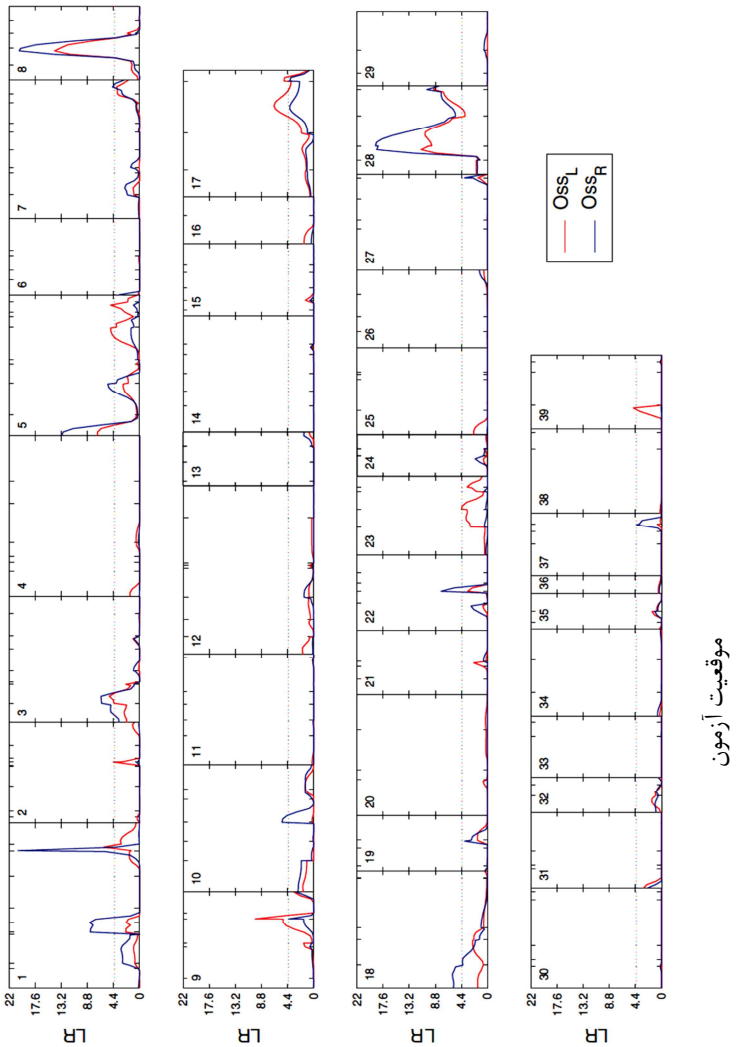
---

[1] - Canine Hip Displasia

همان‌طور که به وسیله نقاط اوج پروفیل نسبت لگاریتم درستنمایی ژنومی که نشان‌دهنده MLE های موقعیت‌های QTL هستند، نشان داده شده است (شکل ۱۴-۵)، ۱۳ عدد QTL برای OSS در سطح معنی‌دار کل ژنومی ۰/۱ درصد شناسایی شد. مقادیر برآوردی مختلفی در ارتباط با σ_a^2 وجود دارند که پیشنهاد می‌دهد QTL ها به‌طور متفاوتی در واریانس ژنتیکی نقش دارند. از QTL های شناسایی شده، ۶ مورد که روی نشانگرهای CFA1، CFA3، CFA5، CFA8، CFA9 و CFA28 شناسایی شدند، دارای اثر عمومی هستند که هم در سمت راست و هم در سمت چپ استخوان هیپ روی OSS اثر می‌گذارند (شکل ۱۴-۵). در بیشتر موارد، برآوردهای ثابت و پایدار σ_a^2 در سمت چپ و راست اشاره ضمنی دارد که این QTL ها نقش‌های یکسان مهمی در حساسیت CHD برای دو سمت استخوان هیپ ایفا می‌کنند. QTL های دیگر اختصاصی هستند، به‌طوری که دو QTL شناسایی شده روی نشانگرهای CFA17 و CFAX مسئول سمت چپ و پنج QTL شناسایی شده روی نشانگرهای CFA7، CFA10، CFA18، CFA22 و CFA37 مسئول سمت راست هستند. همان‌گونه که نشان داده شده است، این QTL های اختصاصی هیچ سهمی در CHD استخوان هیپ مخالف ندارند.



شکل ۱۴-۴- دیاگرام یک شجره دگربارور در سگ. مربع‌ها و دایره‌ها به ترتیب نشان‌دهنده نرها و ماده‌ها هستند. بخش پر شده و پر نشده هر علامت به ترتیب نشان‌دهنده سهم آلل‌های Greyhound و Labrador Retriever در آن سگ هستند.



شکل ۱۴-۵- پروفیل نسبت‌های لگاریتم درستنمایی (LR) بین مدل‌های کامل (یک QTL وجود دارد) و مدل‌های کاهش یافته (QTL وجود ندارد) برآورد شده از مدل فاصله‌ای برای OSS اندازه‌گیری شده در سمت چپ (OSS_L، قرمز) و راست (OSS_R، آبی) استخوان هیپ سگ در طول کل ژنوم از کروموزوم ۱ تا ۳۹ با استفاده از نقشه پیوستگی تهیه شده از نشانگرهای ریزماهواره. خط افقی نشان‌دهنده آستانه بحرانی در p=۰/۰۰۱ است که از روی آزمون‌های جایگشت تعیین شده است. اقتباس از لیو و همکاران (۲۰۰۶).

۱۴-۴- تمرین‌ها

۱۴-۱- مدل مکان‌یابی در خانواده‌های تنی فرض می‌کند که فازهای نشانگر قبل از تجزیه QTL شناخته شده هستند؛ به عبارت دیگر، مکان‌یابی QTL برای داده‌های حاصل از یک خانواده تنی مستلزم انجام دو مرحله است: ۱- تعیین فازهای نشانگر با استفاده از روش‌های تشریح شده در فصل ۴ و ۲- مکان‌یابی QTL با استفاده از فازهای شناخته شده نشانگر. از لحاظ نظری، مدل QTL خانواده تنی می‌تواند عدم قطعیت فازهای نشانگر را در مدل مختلط (۱۴-۴) مشارکت دهد.

الف. روشی را برای مدل‌سازی توأم فازهای نشانگر و فازهای نشانگر-QTL نشان دهید.

ب. مدل‌های مختلط و جداگانه را با هم مقایسه کنید. آیا مدل های مختلط و توأم نسبت به مدل‌های جداگانه دارای مزیت بیشتری هستند؟

ج. یک مثال واقعی برای مکان‌یابی تنی پیدا کنید. از مدل‌های آماری ارایه شده در این فصل برای تجزیه مثال و استنباط‌های آماری حاصل از تجزیه خود استفاده کنید.

۱۴-۲- مکان‌یابی QTL های با آثار اپیستازی در تلاقی‌های دگربارور

روش‌های مکان‌یابی تلاقی‌های دگربارور می‌توانند به آسانی با در نظر گرفتن ژنوتیپ‌های QTL بیشتری در مدل مختلط، به منظور مکان‌یابی QTL های با آثار اپیستازی بسط یابند. برای مکان‌یابی آثار اپیستازی در تلاقی‌های دگربارور، مدل‌سازی مشابهی می‌تواند انجام شود. برای جایگاه‌های ژنی حاوی اطلاعات کامل، چهار آلل متفاوت بین دو والد دگربارور در یک جایگاه ژنی وجود دارد. دو QTL را در نظر بگیرید که هر یک دارای چهار ژنوتیپ به صورت ۱۳، ۱۴، ۲۳ و ۲۴ در جمعیت نتاج دگربارور با اندازه n هستند. اگر $\mu_{u_1v_1/u_2v_2}$ ارزش ژنوتیپی ژنوتیپ u_1v_1/u_2v_2 مربوط به QTL و $u_1,u_2=1,2$ و $v_1,v_2=3,4$ باشند و $u=\left(\mu_{u_1v_1/u_2v_2}\right)$ بردار میانگین مربوطه باشد، پارامترهای آثار ژنتیکی برای دو QTL دارای اثر متقابل به صورت $a=\left(\mu,\alpha_1,\beta_1,\gamma_1,\alpha_2,\beta_2,\gamma_2,I_{\alpha\alpha},I_{\alpha\beta},I_{\beta\alpha},I_{\beta\beta},J_{\alpha\gamma},J_{\beta\gamma},K_{\gamma\alpha},K_{\gamma\beta},L_{\gamma\gamma}\right)^T$ مرتب می‌شوند که در آن:

الف. μ میانگین کل است.

ب. α_1 اثر افزایشی ناشی از جایگزینی آلل ۱ با ۲ در اولین QTL است.

پ. β_1 اثر افزایشی ناشی از جایگزینی آلل ۳ با ۴ در اولین QTL است.

ت. γ_1 اثر غالبیت ناشی از اثر متقابل بین آلل‌های حاصل از والدین متفاوت است.

ث. α_2 اثر افزایشی حاصل از جایگزینی آلل ۱ با ۲ در دومین QTL است.

ج. β_2 اثر افزایشی ناشی از جایگزینی آلل ۳ با ۴ در دومین QTL است.

چ. γ_2 اثر غالبیت ناشی از اثر متقابل بین آلل‌های حاصل از والدین متفاوت است.

ح. $I_{\alpha\alpha}$ اثر اپیستازی افزایشی× افزایشی ناشی از اثر متقابل مربوط به جایگزینی آلل ۱ با ۲ در اولین و دومین QTL است.

خ. $I_{\alpha\beta}$ اثر اپیستازی افزایشی× افزایشی ناشی از اثر متقابل مربوط به جایگزینی آلل ۱ با ۲ در اولین QTL و جایگزینی آلل‌های ۳ با ۴ در دومین QTL است.

د. $I_{\beta\alpha}$ اثر اپیستازی افزایشی× افزایشی ناشی از اثر متقابل مربوط به جایگزینی آلل ۳ با ۴ در اولین QTL و جایگزینی آلل‌های ۱ با ۲ در دومین QTL است.

ذ. $I_{\alpha\beta}$ اثر اپیستازی افزایشی× افزایشی ناشی از اثر متقابل مربوط به جایگزینی آلل ۳ با ۴ در اولین و دومین QTL است.

ر. $J_{\alpha\beta}$ اثر اپیستازی افزایشی× غالبیت ناشی از اثر متقابل مربوط به جایگزینی آلل ۱ با ۲ در اولین QTL و اثر غالبیت در دومین QTL است.

ز. $J_{\alpha\beta}$ اثر اپیستازی افزایشی× غالبیت ناشی از اثر متقابل مربوط به جایگزینی آلل ۳ با ۴ در اولین QTL و اثر غالبیت در دومین QTL است.

ژ. $K_{\alpha\beta}$ اثر اپیستازی غالبیت× افزایشی ناشی از اثر متقابل مربوط به اثر غالبیت در اولین QTL و جایگزینی آلل ۱ با ۲ در دومین QTL است.

ط. $K_{\alpha\beta}$ اثر اپیستازی غالبیت× افزایشی ناشی از اثر متقابل مربوط به اثر غالبیت در اولین QTL و جایگزینی آلل ۳ با ۴ در دومین QTL است.

ظ. $K_{\alpha\beta}$ اثر اپیستازی غالبیت×غالبیت ناشی از اثر متقابل بین آثار غالبیت در اولین و دومین QTL است.

رابطه بین بردار ارزش ژنوتیپی و بردار اثر ژنتیکی به صورت زیر است:

$$u = Da$$

که در آن، ماتریس طرح (D) به صورت زیر است:

$$D=\begin{bmatrix}
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&1&1&-1&-1&1&-1&1&-1&-1&-1&1&-1&-1\\
1&1&1&1&-1&1&-1&-1&1&-1&1&-1&-1&-1&1&-1\\
1&1&1&1&-1&-1&1&-1&-1&-1&-1&1&1&-1&-1&1\\
1&1&-1&-1&1&1&1&1&1&-1&-1&1&-1&-1&-1&-1\\
1&1&-1&-1&1&-1&-1&1&-1&-1&1&-1&1&-1&1&1\\
1&1&-1&-1&-1&1&-1&-1&1&1&-1&-1&1&1&-1&1\\
1&1&-1&-1&-1&-1&1&-1&-1&1&1&1&-1&1&1&-1\\
1&-1&1&-1&1&1&1&-1&-1&1&1&-1&1&-1&-1&-1\\
1&-1&1&-1&1&-1&-1&-1&1&1&-1&1&-1&-1&1&1\\
1&-1&1&-1&-1&1&-1&1&-1&-1&1&1&-1&1&-1&1\\
1&-1&1&1&-1&-1&1&1&1&-1&-1&-1&1&1&1&-1\\
1&-1&-1&1&1&1&1&-1&-1&-1&-1&-1&-1&1&1&1\\
1&-1&-1&1&1&-1&-1&-1&1&-1&1&1&1&1&-1&-1\\
1&-1&-1&1&-1&1&-1&1&-1&1&-1&1&1&-1&1&-1\\
1&-1&-1&1&-1&-1&1&1&1&1&1&-1&-1&-1&-1&1
\end{bmatrix}$$

با برآورد بردار ارزش ژنوتیپی، بردار اثر ژنتیکی می‌تواند به آسانی از رابطه $\hat{a}=D^T\hat{u}$ برآورد شود. درستنمایی برای دو QTL دارای اثر متقابل دگربارور می‌تواند به صورت زیر فرموله شود:

$$L(y)=\prod_{i=1}^{n}\sum_{u_1=1}^{2}\sum_{v_1=3}^{4}\sum_{u_2=1}^{2}\sum_{v_2=3}^{4}\omega_{u_1v_1/u_2v_2|i}\,f_{u_1v_1/u_2v_2}(y_i)$$

که در آن، $\omega_{u_1v_1/u_2v_2|i}$ احتمال شرطی ژنوتیپ QTL مشروط به ژنوتیپ نشانگر در فرد i و $f_{u_1v_1/u_2v_2}(y_i)$ تابع چگالی توزیع نرمال با میانگین $\mu_{u_1v_1/u_2v_2}$ و واریانس σ^2 است. احتمالات شرطی نیز می‌توانند بر حسب نرخ‌های نوترکیبی بین نشانگرها و QTL به‌دست آیند.

یک الگوریتم EM استاندارد می‌تواند برای برآورد ارزش‌های ژنوتیپی و واریانس باقیمانده تهیه شود. با تعریف:

$$\Pi_{u_1v_1/u_2v_2|i}=\frac{\omega_{u_1v_1/u_2v_2|i}\,f_{u_1v_1/u_2v_2}(y_i)}{\sum_{u_1'=1}^{2}\sum_{v_1'=3}^{4}\sum_{u_2'=1}^{2}\sum_{v_2'=3}^{4}\omega_{u_1'v_1'/u_2'v_2'|i}\,f_{u_1'v_1'/u_2'v_2'}(y_i)}$$

در مرحله E، MLE های پارامترهای ناشناخته در مرحله M به‌صورت زیر به‌دست می‌آیند:

$$\hat{\mu}_{u_1v_1/u_2v_2} = \frac{\sum_{u_1=1}^{2}\sum_{v_1=3}^{4}\sum_{u_2=1}^{2}\sum_{v_2=3}^{4}\Pi_{u_1v_1/u_2v_2|i}y_i}{\sum_{u_1=1}^{2}\sum_{v_1=3}^{4}\sum_{u_2=1}^{2}\sum_{v_2=3}^{4}\Pi_{u_1v_1/u_2v_2|i}}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{u_1=1}^{2}\sum_{v_1=3}^{4}\sum_{u_2=1}^{2}\sum_{v_1=3}^{4}\left(y_i - \hat{\mu}_{u_1v_1/u_2v_2}\right)^2\Pi_{u_1v_1/u_2v_2|i}$$

مراحل E و M تا جایی تکرار می‌شوند که برآوردها به مقدار ثابت و پایدار برسند. پس از آن که پارامترها برآورد شدند، چند آزمون فرض می‌تواند انجام شود.

الف. نشان دهید که چگونه حضور QTL آزمون شده و آستانه بحرانی برای اعلام حضور QTL تعیین می‌شود.

ب. آزمون‌های فرض برای آثار ژنتیکی مختلف شامل آثار افزایشی، غالبیت و اپیستازی، می‌توانند با استفاده از فرض‌های صفر مربوطه فرموله شوند:

$$H_0 : \alpha_1 = 0$$
$$H_0 : \beta_1 = 0$$
$$H_0 : \gamma_1 = 0$$
$$H_0 : \alpha_2 = 0$$
$$H_0 : \beta_2 = 0$$
$$H_0 : \gamma_2 = 0$$
$$H_0 : I_{\alpha\alpha} = 0$$
$$H_0 : I_{\alpha\beta} = 0$$
$$H_0 : I_{\beta\beta} = 0$$
$$H_0 : J_{\alpha\gamma} = 0$$
$$H_0 : J_{\beta\gamma} = 0$$
$$H_0 : K_{\gamma\alpha} = 0$$
$$H_0 : K_{\gamma\beta} = 0$$
$$H_0 : L_{\gamma\gamma} = 0$$

تحت فرض صفر $H_0 : \alpha_1 = 0$، ارزش‌های ژنوتیپی باید به‌صورت زیر محدود شوند:

$$\mu_{13/13} + \mu_{13/14} + \mu_{13/23} + \mu_{13/24}$$
$$+\mu_{14/13} + \mu_{14/14} + \mu_{14/23} + \mu_{14/24}$$
$$= \mu_{23/13} + \mu_{23/14} + \mu_{23/23} + \mu_{23/24}$$
$$+\mu_{24/13} + \mu_{24/14} + \mu_{24/23} + \mu_{24/24}$$

همان طور که در مطالب بالا توضیح داده شد، این محدودیت برای الگوریتم EM اعمال می‌شود که منجر به MLE های ارزش‌های ژنوتیپی با محدود نمودن α_1 برابر صفر خواهد شد. محدودیت‌ها و الگوریتم‌هایی را برای برآورد پارامترها در هر یک از فرض‌های صفر دیگر ایجاد کنید.

ج. از الگوریتم‌هایی که به‌عنوان مثال عملی برای مکان‌یابی QTL در جوامع دگربارور ایجاد نمودید، استفاده کنید.

۱۴-۳- همان‌طور که در معادله (۱۴-۱۶) نشان داده شد، نشان دهید که چگونه ارتباط بین IBD های نشانگر و QTL به‌دست می‌آیند.

۱۴-۴- برای مکان‌یابی QTL ها در یک شجره پیچیده، به ماتریس های IBD بین هر جفت از سگ‌ها نیاز داریم. از شکل ۱۴-۴ به عنوان مثال استفاده کنید:

الف. ماتریس (148×148) را برای نسبت آلل‌های IBD در یک QTL بنویسید.

ب. ماتریس (148×148) را برای متغیرهای دو وضعیتی به نحوی بنویسید که نشان دهد آیا هر جفت از سگ‌ها در QTL مربوطه به صورت IBD هستند یا خیر.

ج. ماتریس (148×148) را برای ضرایب IBD بین هر جفت از سگ‌ها برای تمامی ژن‌ها بنویسید.

۱۴-۵- مکان‌یابی QTL های با آثار اپیستازی در یک شجره پیچیده

اگر QTL های با آثار اپیستازی برای شجره شکل ۱۴-۴ مکان‌یابی شوند، باید واریانس‌های اپیستازی یعنی افزایشی× افزایشی ($\sigma^2_{i_{aa}}$)، افزایشی× غالبیت ($\sigma^2_{i_{ad}}$)، غالبیت × افزایشی ($\sigma^2_{i_{da}}$) و غالبیت × غالبیت ($\sigma^2_{i_{dd}}$) را در مدل با آثار تصادفی وارد نماییم. در این حالت، واریانس کل برای مشاهده منفرد (y_{ij}) بدین صورت بیان می‌شود:

$$\mathrm{var}\left(y_{ij}\right) = \sigma^2_{a_1} + \sigma^2_{d_1} + \sigma^2_{a_2} + \sigma^2_{d_2} + \sigma^2_{i_{aa}} + \sigma^2_{i_{ad}} + \sigma^2_{i_{da}} + \sigma^2_{i_{dd}} + \sigma^2_g + \sigma^2_e$$

کوواریانس بین دو برادر-خواهر j و j' از خانواده i برابر است با:

$$\operatorname{cov}\left(y_{ij}, y_{ij'}\right) = \pi_{ia_1}\sigma^2_{a_1} + \pi_{id_1}\sigma^2_{d_1} + \pi_{ia_2}\sigma^2_{a_2} + \pi_{id_2}\sigma^2_{d_2}$$
$$+ \pi_{ia_1}\pi_{ia_2}\sigma^2_{i_{aa}} + \pi_{ia_1}\pi_{id_2}\sigma^2_{i_{ad}} + \pi_{id_1}\pi_{id_2}\sigma^2_{i_{dd}} + \phi_{ig}\sigma^2_g$$

که اندیس درجه دوم π برای شناسایی یک QTL است.

الف. رابطه درستنمایی را برای دو QTL دارای اثر متقابل بنویسید.

ب. الگوریتم‌های مربوط به برآورد واریانس‌های افزایشی ($\sigma^2_{a_1}$ و $\sigma^2_{a_2}$) و غالبیت ($\sigma^2_{d_1}$ و $\sigma^2_{d_2}$) دو QTL مختلف و واریانس‌های اپیستازی آن‌ها را به‌دست آورید.

ضمائم

ضمیمه A

نتایج و الگوریتم‌های آماری عمومی

A.1- مجانب‌های درستنمایی

مباحث زیر از کتاب استنباط آماری[1]، ویرایش دوم، تألیف کاسلا و برگر (۲۰۰۱) اقتباس شده‌اند. این مباحث تا حدودی پیشرفته بوده و برای افرادی که می‌خواهند تصویر کامل‌تری از مجانب‌های درستنمایی داشته باشند، ارائه شده است.

پس از آن که داده‌های X=x مشاهده شدند، تابع درستنمایی یعنی $L(\theta|x)$، تابع کاملاً تعریف شده‌ای از متغیر θ بوده و آماره LR عبارت است از:

$$\lambda(x)=\frac{\sup_{\theta_0} L(\theta|x)}{\sup_{\theta} L(\theta|x)}$$

حتی اگر دو مجموعه منظم و کامل $L(\theta|x)$، برای تمامی مقادیر θ_0 و θ، نتوانند از نظر تجزیه‌ای به‌دست آیند، معمولاً می‌توانند به‌صورت عددی محاسبه شوند؛ بنابراین، آماره آزمون $\lambda(x)$ می‌تواند برای داده‌های مشاهده شده به‌دست آید، حتی اگر هیچ فرمول مناسبی برای تعریف کردن و مشخص نمودن $\lambda(x)$ وجود نداشته باشد.

برای تعریف یک سطح آزمون یعنی α، عدد ثابت c باید انتخاب شود، طوری که:

$$\sup_{\theta\in\theta_0} P_\theta\left(\lambda(x)\le c\right)\le\alpha \qquad \text{(A.1)}$$

قضیه A.1 (توزیع مجانبی LRT- H_0 ساده). برای آزمون $H_0:\theta=\theta_0$ در مقابل $H_1:\theta\neq\theta_0$، فرض کنید $X_1,\ldots,X_n$ به صورت $iid\ \ f(x|\theta)$ باشد، $\hat{\theta}$ نیز MLE مربوط به θ بوده و $f(x|\theta)$ نیز از شرایط معمول معین تبعیت می‌کند. سپس، تحت فرض H_0، وقتی که $n\to\infty$ میل کند، $-2\log\lambda(x)$ به سمت یک توزیع کای مربع با یک درجه آزادی میل خواهد کرد، یعنی:

$$-2\log\lambda(x)\to x_1^2$$

که x_1^2 متغیر تصادفی x^2 با یک درجه آزادی است.

---

[1]- Statistical Inference

قضیه A.1 می‌تواند به حالتی بسط یابد که در آن، فرض صفر در ارتباط با بردار پارامترها است. تعمیم زیر که آن‌را بدون اثبات نیز بیان می‌کنیم، اجازه می‌دهد که اطمینان حاصل کنیم معادله (A.1) حداقل برای نمونه‌های بزرگ صادق است.

قضیه A.2. فرض کنید $X_1,...,X_n$ نمونه‌ای تصادفی از یک pdf یا pmf مربوط به $f(x\mid\theta)$ باشد. تحت شرایط معمول معین، اگر $\theta\in\theta_0$ باشد، سپس توزیع آماره $-2\log\lambda(x)$ هنگامی‌که اندازه نمونه $n\to\infty$ میل می‌کند، یک توزیع کای مربع است. درجه آزادی توزیع محدود شده، تفاوت بین تعداد پارامترهای آزاد اختصاص یافته به $\theta\in\theta_0$ و تعداد پارامترهای آزاد اختصاص یافته به $\theta\in\theta$ است.

رد $H_0:\theta\in\theta_0$ برای مقادیر کوچک $\lambda(x)$ معادل رد مقادیر بزرگ $-2\log\lambda(x)$ است؛ بنابراین:

H_0 زمانی رد می‌شود که اگر و فقط اگر $-2\log\lambda(x)\geq x^2_{v,\alpha}$ باشد.

که در آن، v درجه آزادی بیان شده در قضیه A.2 است. اگر $\theta\in\theta_0$ و اندازه نمونه بزرگ باشد، احتمال خطای نوع اول به‌طور تقریبی α خواهد بود. در این روش، معادله (A.1) به‌طور تقریبی برای اندازه‌های نمونه بزرگ برقرار بوده و یک آزمون اندازه مجانبی (α) تعریف می‌شود. توجه کنید که این قضیه درحقیقت فقط به این موضوع دلالت دارد که:

برای هر

$$\lim_{n\to\infty}P_\theta(reject\quad H_0)=\alpha \qquad\qquad \theta\in\theta_0$$

و به این که $\sup_{\theta\in\theta_0}P_\theta(reject\quad H_0)$ در نقطه α همگرا می‌شود، دلالت نمی‌کند. این حالت معمولاً برای آزمون‌های اندازه مجانبی (α) رخ می‌دهد.

قضیه A.3. اگر $X_1,...,X_n$ یک نمونه تصادفی از یک pdf یا pmf مربوط به $f(x\mid\theta)$، $\hat{\theta}$ به عنوان MLE مربوط به θ و $h(\theta)$ تابع پیوسته‌ای از θ باشد، تحت شرایط معمول معین، داریم:

$$\sqrt{n}\left[h(\hat{\theta})-h(\theta)\right]\to n\left[0,v(\theta)\right]$$

در این رابطه، $Var\left(h\left(\hat{\theta}\right)\right)=v\left(\theta\right)$ می‌تواند به صورت زیر تقریب زده شود:

$$Var\left(h\left(\hat{\theta}\right)\right)\approx\frac{\left[h'(\theta)\right]^2|_{\theta=\hat{\theta}}}{-\frac{\partial^2}{\partial\theta^2}\log L\left(\theta\,|\,X\right)|_{\theta=\hat{\theta}}}$$

که در آن، مخرج کسر برابر با $\hat{I}_n\left(\hat{\theta}\right)$ یعنی تعداد اطلاعات مشاهده شده است.

شرایط منظم. قضایای A.1، A.2 و A.3 به شرایط معمول معین ارجاع شده‌اند. این شرایط از نظر ریاضی شرایط تکنیکی، نسبتاً خسته کننده و در بیشتر مسائل منطقی معمولاً مناسب و رضایت‌بخش هستند. ولی آن‌ها به عنوان یک عوامل زیان‌آور ضروری هستند.

این شرایط اساساً وابسته به قابلیت مشتق‌گیری تابع چگالی احتمال و توانایی تبادل و تغییر مشتق و انتگرال هستند. برای ارائه اصول کلی و روشن شدن جزئیات بیشتر، کاسلا و برگر (۲۰۰۱)، استوآرت، اورد و آرنولد (۱۹۹۹، فصل ۱۸)، فرگوسن (۱۹۹۶، بخش ۴) یا لمن و کاسلا (۱۹۹۸، بخش ۶-۳) را مشاهده کنید.

A.2- شکل کلی الگوریتم EM

اگرچه جزئیات طرز عمل عملکردهای الگوریتم EM را توضیح خواهیم داد، اما از ذکر برخی از جزئیات که دارای ماهیت ریاضی هستند، خودداری خواهیم کرد. برای مشاهده جزئیات بالا، به کاسلا و برگر (۲۰۰۱، فصل ۷) مراجعه نمایید.

به‌طور معمول، روابط آماری را با داده‌های y و تابع درستنمایی $L\left(\theta\,|\,y\right)$ آغاز می‌کنیم. سپس، داده‌های گمشده[1] با z معرفی می‌شوند (که گاهی اوقات، داده‌های حجیم شده یا تکمیلی[2] نامیده می‌شوند). آنگاه به $\left(y,z\right)$ به عنوان داده‌های کامل و $L\left(\theta\,|\,y,z\right)$ به عنوان درستنمایی داده‌های کامل مراجعه می‌کنیم (در مسائلی که با آن‌ها مواجه هستیم، چگونگی انتخاب $L\left(\theta\,|\,y,z\right)$ معمولاً مشخص است، چون این مطلب که چه نوعی از داده‌های گمشده محاسبات را آسان‌تر می‌سازند، مشخص و معلوم است). به‌طور مشابه، گاهی اوقات $L\left(\theta\,|\,y\right)$ درستنمایی داده‌های ناقص یا درستنمایی داده‌های مشاهده شده نامیده می‌شود. از روی این دو نوع درستنمایی، الگوریتم کامل EM ادامه می‌یابد.

بخش اصلی و ضروری الگوریتم، تعریف تابع چگالی احتمال داده‌های گمشده (k) است که به‌صورت زیر ارائه می‌شود:

---

[1] - Missing data
[2] - Augmented data

$$k(z\mid\theta,y)=\frac{L(\theta\mid y,z)}{L(\theta\mid y)} \tag{A.2}$$

که مجدداً به‌صورت زیر مرتب و برحسب لگاریتم نوشته می‌شود:

$$\log L(\theta\mid y)=\log L(\theta\mid y,z)-\log k(z\mid\theta,y) \tag{A.3}$$

الگوریتم EM یک الگوریتم تکراری است. در هر مرحله ($t=1,2,\ldots$) یک برآورد از θ به‌دست می‌آید. در این صورت، زمانی که $t\to\infty$ میل می‌کند، $\hat{\theta}^{(t)}\to\hat{\theta}$ میل کرده و در نتیجه $\hat{\theta}^{(t)}$ یک MLE واقعی از θ خواهد بود. برای به‌دست آوردن این توالی، با مقدار مورد انتظار معادله (A.3)، به شیوه زیر عمل می‌کنیم. در تکرار t ام، $\hat{\theta}^{(t)}$ را داریم و برای محاسبه $\hat{\theta}^{(t+1)}$ که اکنون آنرا با θ نشان می‌دهیم، نیز آماده‌ایم. برای انجام این کار، ابتدا مقدار مورد انتظار هر دو سمت معادله (A.3) (مرحله E) را نسبت به $k\left(z\mid\hat{\theta}^{(t+1)},y\right)$ نوشته و رابطه زیر را به‌دست می‌آوریم:

$$E_{\hat{\theta}^{(t)}}\left[\log L(\theta\mid y)\right]=E_{\hat{\theta}^{(t)}}\left[\log L(\theta\mid y,z)\right]-E_{\hat{\theta}^{(t)}}\left[\log k(z\mid\theta,y)\right] \tag{A.4}$$

یا:

$$\log L(\theta\mid y)=E_{\hat{\theta}^{(t)}}\left[\log L(\theta\mid y,z)\right]-E_{\hat{\theta}^{(t)}}\left[\log k(z\mid\theta,y)\right]$$

زمانی که امید ریاضی دارای توزیع z و $\log L(\theta\mid y)$ فاقد z باشد، نیازی به به‌دست آوردن امیدهای ریاضی نداریم.

حال در اینجا ظرافت و زیبایی الگوریتم EM مشخص می‌شود. $\hat{\theta}^{(t+1)}$ را طوری انتخاب می‌کنیم تا برابر با همان مقدار θ بوده و عبارت $E_{\hat{\theta}^{(t)}}\left[\log L(\theta\mid y,z)\right]$ یعنی درستنمایی مورد انتظار داده‌های کامل را به حداکثر برساند. به وسیله نظریه الگوریتم EM، این مقدار θ به‌طور خودکار، عبارت انتهایی معادله (A.4) یعنی $E_{\hat{\theta}^{(t)}}\left[\log k(z\mid\theta,y)\right]$ را کاهش می‌دهد؛ بنابراین، از رابطه زیر تبعیت می‌کند:

$$\log L\left(\hat{\theta}^{(t)}\mid y\right)\le\log L\left(\hat{\theta}^{(t+1)}\mid y\right)$$

و نباید رابطه $E_{\hat{\theta}^{(t)}}\left[\log k(z\mid\theta,y)\right]$ را محاسبه کنیم.

بنابراین، توالی EM یعنی $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, ..., \hat{\theta}^{(t)} \hat{\theta}^{(t+1)}$ منجر به افزایش مقدار تـابع درسـتنمایی یعنی $L(\theta | y)$ شده و تحت برخی از شرایط منعطف‌تر، به سمت MLE مربـوط بـه θ همگـرا خواهد شد.

ضمیمه B

برنامه‌های R

B.1- فصل دوم

(۱) برنامه R برای مثال ۲-۴

```
y<-c(79,82,85,87,100,101,102,103,124,125,126,127)
n<-length(y);nit<-20;
MAA<-array(max(y),dim=c(nit,1));
MAa<-array(mean(y),dim=c(nit,1));
Maa<-array(min(y),dim=c(nit,1));
S<-array(sd(y),dim=c(nit,1));S2<-array(sy,dim=c(nit,1));
S2[1]<-5;
for(i in 2:nit){
temp1<-.25*dnorm(y,mean=MAA[i-1],sd=S[i-1]);
temp2<-.5*dnorm(y,mean=MAa[i-1],sd=S[i-1]);
temp3<-.25*dnorm(y,mean=Maa[i-1],sd=S[i-1]);
PAA<-temp1/(temp1+temp2+temp3);
PAa<-temp2/(temp1+temp2+temp3);
Paa<-1-PAA-PAa;
MAA[i]<-sum(y*PAA)/sum(PAA);
MAa[i]<-sum(y*PAa)/sum(PAa);
Maa[i]<-sum(y*Paa)/sum(Paa);
#S[i]<-sqrt((1/nit)*(sum((y-MAA[i])^2)
+sum((y-MAa[i])^2)+sum((y-Maa[i])^2)))
S2[i]<-(sum(PAA*(y-MAA[i])^2)+sum(PAa*(y-MAa[i])^2)
+sum(Paa*(y-Maa[i])^2))/(n*sum(PAA+PAa+Paa))
S[i]<-sqrt(S2[i])
}
par(mfrow=c(2,2))
```

```
par(mar=c(4,4,1,1))
plot(MAA,type="l",lwd=2,main="",
ylab="Mean AA",xlab="Iteration")
plot(MAa,type="l",lwd=2,main="",
ylab="Mean Aa",xlab="Iteration")
plot(Maa,type="l",lwd=2,main="",
ylab="Mean aa",xlab="Iteration")
plot(S2,type="l",lwd=2,main="",
ylab="Variance",xlab="Iteration")
MAA[nit];MAa[nit];Maa[nit];S2[nit]
```

برای مثال ۲-R۱۲(۲) برنامه

```
y<-c(1339,154,151,1195);
m<-length(y);n<-sum(y);nsim<-10000;
#-----Log likelihood function---------
lambda<-function(p1,p2,p3,x1,x2,x3)
(x1*log(p1)+x2*log(p2)+x3*log(p3)
+(n-x1-x2-x3)*log(1-p1-p2-p3));
p1hat<-(y[1]+y[4])/(2*n);p2hat<-(y[2]+y[3])/(2*n);
Lnull<-lambda(p1hat,p2hat,p2hat,y[1],y[2],y[3])
Lmax<-lambda(y[1]/n,y[2]/n,y[3]/n,y[1],y[2],y[3]);
Tobs<- -2*(Lnull-Lmax);
Tsim<-array(0,dim=c(nsim,1));
for(i in 1:nsim)
{
x<-rmultinom(1,n,c(y[1]/n,y[2]/n,y[3]/n,y[4]/n))
Tsim[i]<- -2*(lambda((x[1]+x[4])/(2*n),
(x[2]+x[3])/(2*n), (x[2]+x[3])/(2*n),
```

```
x[1],x[2],x[3])-lambda(x[1]/n,
x[2]/n,x[3]/n,x[1],x[2],x[3]))
}
hist(Tsim,main=expression(-2(log)(lambda)),
xlab="Simulated Observations",xlim=c(0,40),
freq=F,col="green",breaks=50)
mean(Tsim>Tobs)
#-----------------------------------------------------
#The function rmultinom returns a random multinomial vector
#n=number of variables desired
#size=sum of cells
#prob=vector of cell probabilities
rmultinom <- function(n,size,prob){
K <- length(prob) # #{classes}
matrix(tabulate(sample(K, n*size,
repl = TRUE, prob)+K *0:(n-1),
nbins=n*K),
nrow=n, ncol=K, byrow=TRUE)}
```

B.2- فصل هشتم

(۱) برنامه R برای مثال ۸-۶

```
#This gets MLEs for the Tomato data
y1<-c(79,82,100,102,124);y2<-c(85,87,101,103,125,126,127)
n1<-length(y1);n2<-length(y2);n<-n1+n2
#-----Initialize Estimates-------
r<-.25;m1<-mean(y1);m2<-mean(y2);s<-sqrt(var(c(y1,y2)))
m1plot<-m1;m2plot<-m2;splot<-s;rplot<-r
#-----Start Iteration-----------------------------
```

```
nit<-100
for(i in 1:nit)
{
w1<-P1(y1,m1,m2,s,r);w2<-P2(y2,m1,m2,s,r)
m1<-(sum(w1*y1)+sum(w2*y2))/(sum(w1)+sum(w2))
m2<-(sum((1-w1)*y1)+sum((1-w2)*y2))/(sum((1-w1))+sum((1-w2)))
s<-sqrt((sum(w1*(y1-m1)^2+(1-w1)*(y1-m2)^2)
+sum(w2*(y2-m1)^2+(1-w2)*(y2-m2)^2))/n)
r<-(sum(1-w1)+sum(w2))/n
r<-min(r,.5)
m1plot<-c(m1plot,m1);m2plot<-c(m2plot,m2)
splot<-c(splot,s);rplot<-c(rplot,r)
}
#----------------Plot---------------------------
par(mfrow=c(2,2))
par(mar=c(4,4,1,1))
plot(m1plot,type="l",lwd=2,main="",
ylab="mean 1",xlab="Iteration")
plot(m2plot,type="l",lwd=2,main="",
ylab="mean 2",xlab="Iteration")
plot(splot,type="l",lwd=2,main="",
ylab="std",xlab="Iteration")
plot(rplot,type="l",lwd=2,main="",
ylab="r",xlab="Iteration")
#----------Define Functions for P1 and P2
P1<-function(y,t1,t2,w,r)
{
temp1<-(1-r)*dnorm(y,mean=t1,sd=w)
temp2<-r*dnorm(y,mean=t2,sd=w)
```

```
return(temp1/(temp1+temp2))
}
P2<-function(y,t1,t2,w,r)
{
temp1<-r*dnorm(y,mean=t1,sd=w)
temp2<-(1-r)*dnorm(y,mean=t2,sd=w)
return(temp1/(temp1+temp2))
}
```

(۲) برنامه R برای مثال ۸-۶

```
#This gets permutation distribution for the Tomato data
y1<-c(79,82,100,102,124);y2<-c(85,87,101,103,125,126,127)
n1<-length(y1);n2<-length(y2);n<-n1+n2
y<-c(y1,y2)
#-----Initialize Estimates-------
r<-.25;m1<-mean(y1);m2<-mean(y2);s<-sd(y)
Lplot<- logL(y1,y2,m1,m2,s,r)
#-----Start Iteration-----------------------------
nperm<-5000 #number of permutation statistics
nit<-50 #number of iterations to find MLEs
for(j in 1:nperm)
{
yp<-sample(y);y1<-yp[1:n1];y2<-yp[(n1+1):n]
for(i in 1:nit)
{
w1<-P1(y1,m1,m2,s,r);w2<-P2(y2,m1,m2,s,r)
m1<-(sum(w1*y1)+sum(w2*y2))/(sum(w1)+sum(w2))
m2<-(sum((1-w1)*y1)+sum((1-w2)*y2))/(sum((1-w1))+sum((1-w2)))
```

```
s<-sqrt((sum(w1*(y1-m1)^2+(1-w1)*(y1-m2)^2)
+sum(w2*(y2-m1)^2+(1-w2)*(y2-m2)^2))/n)
r<-(sum(1-w1)+sum(w2))/n
r<-min(r,.5)
}
L<- logL(y1,y2,m1,m2,s,r);Lplot<-c(Lplot,L)
}
#----------------Calculate Statistics-----------
m<-mean(y);s<-sd(y)
LH0<-logL(y1,y2,m,m,s,.5)
Lplot<- -2*(LH0-Lplot)
sort(Lplot)[.95*nperm]
sort(Lplot)[.05*nperm]
#----------------Plot---------------------------
hist(Lplot,main="Permutation Distribution",
freq=F,xlab="-2 log lambda")
#----------Define Functions for P1 and P2
P1<-function(y,t1,t2,w,r)
{
temp1<-(1-r)*dnorm(y,mean=t1,sd=w)
temp2<-r*dnorm(y,mean=t2,sd=w)
return(temp1/(temp1+temp2))
}
P2<-function(y,t1,t2,w,r)
{
temp1<-r*dnorm(y,mean=t1,sd=w)
temp2<-(1-r)*dnorm(y,mean=t2,sd=w)
return(temp1/(temp1+temp2))
}
```

```
#-------------Define Functions for log likelihood----------
logL<-function(y1,y2,t1,t2,w,r)
{
S1<-sum(log((1-r)*dnorm(y1,t1,s)+r*dnorm(y1,t2,s)))
S2<-sum(log(r*dnorm(y2,t1,s)+(1-r)*dnorm(y2,t2,s)))
return(S1+S2)
}
```

(۳) برنامه R برای مثال ۱۱-۲

```
#This gets MLEs for the Poplar data
data<-read.table("PoplarMarker3.txt",sep = ",",header=F)
M<-data[,1:4];nm<-dim(M);
Y<-data[,5];n<-length(Y)
mY<-mean(Y);sY<-sd(Y)
LL0<-sum(dnorm(Y,mY,sY,log=T))
#----------Create Classes-------------------
#----------Class1=11, Class2=10,Class3=01,Class4=00
Cl<-array(1,c(nm[1], (nm[2]-1)));nc<-dim(Cl)
for(i in 1:nc[1])
{
for(j in 1:nc[2])
{
c1<-1*(M[i,j]==2)*(M[i, (j+1)]==2);
c2<-2*(M[i,j]==2)*(M[i, (j+1)]==1);
c3<-3*(M[i,j]==1)*(M[i, (j+1)]==2);
c4<-4*(M[i,j]==1)*(M[i, (j+1)]==1);
Cl[i,j]<-c1+c2+c3+c4
}
```

```
}
ML<-array(0,c(nc[2],4));LL<-array(0,c(nc[2],1))
#---------Define the Y classes for each Interval-----------
for(j in 1:nc[2])
{
y1<-(Cl[,j]==1)*Y;y1<-y1[y1!=0]
y2<-(Cl[,j]==2)*Y;y2<-y2[y2!=0]
y3<-(Cl[,j]==3)*Y;y3<-y3[y3!=0]
y4<-(Cl[,j]==4)*Y;y4<-y4[y4!=0]
#---------Get MLEs--------------------------------
#-----Initialize Estimates-------
T<-c(mean(y1),mean(y4),sY,.25)
#-----Start Iteration-----------------------------
nit<-1000
for(i in 1:nit){T<-c(MLE(y1,y2,y3,y4,T[1],T[2],T[3],T[4]))}
ML[j,]<-T
LL[j]<-2*(logL(y1,y2,y3,y4,T[1],T[2],T[3],T[4])-LL0)
}
print(cbind(ML,LL))
#-------------------------------------------------------
#----------Define Functions for P1 and P2
P1<-function(y,t1,t2,w,v)
{
temp1<-(1-v)*dnorm(y,mean=t1,sd=w)
temp2<-v*dnorm(y,mean=t2,sd=w)
return(temp1/(temp1+temp2))
}
P2<-function(y,t1,t2,w,v)
{
```

```
temp1<-v*dnorm(y,mean=t1,sd=w)
temp2<-(1-v)*dnorm(y,mean=t2,sd=w)
return(temp1/(temp1+temp2))
}
#-------------Define log likelihood---------------
logL<-function(y1,y2,y3,y4,t1,t2,w,v)
{
S1<-sum(dnorm(y1,t1,w,log=T))
S2<-sum((1-v)*dnorm(y2,t1,w,log=T)+v*dnorm(y2,t2,w,log=T))
S3<-sum(v*dnorm(y3,t1,w,log=T)+(1-v)*dnorm(y3,t2,w,log=T))
S4<-sum(dnorm(y4,t2,w,log=T))
return(S1+S2+S3+S4)
}
#--------------Define MLE Solver------------------------
MLE<-function(y1,y2,y3,y4,t1,t2,w,v)
{
n1<-length(y1);n2<-n1+length(y2)
n3<-n2+length(y3);n<-n3+length(y4)
w1<-P1(y2,t1,t2,w,v);w2<-P2(y3,t1,t2,w,v)
t1<-(sum(y1)+sum(w1*y2)+sum(w2*y3))/(n1+sum(w1)+sum(w2))
t2<-(sum((1-w1)*y2)+sum((1-w2)*y3)+sum(y4))/(sum((1-w1))
+sum((1-w2))+n-n3)
s1<-sum((y1-t1)^2)
s2<-sum(w1*(y2-t1)^2+(1-w1)*(y2-t2)^2)
s3<-sum(w2*(y3-t1)^2+(1-w2)*(y3-t2)^2)
s4<-sum((y4-t2)^2)
w<-sqrt((s1+s2+s3+s4)/n)
v<-(sum(1-w1)+sum(w2))/(n3-n1)
return(c(t1,t2,w,v))
}
```

(۴) برنامه R برای مثال ۱۱-۲

```
#This gets permutation distribution of MLEs
#for the Poplar data
data<-read.table("PoplarMarker1.txt",
sep=",",header=F)
M<-data[,1:4];nm<-dim(M);
Y<-data[,5];n<-length(Y)
mY<-mean(Y);sY<-sd(Y)
LL0<-sum(dnorm(Y,mY,sY,log=T))
#----------Create Classes-------------------
#----------Class1=11, Class2=10,Class3=01,Class4=00
Cl<-array(1,c(nm[1], (nm[2]-1)));nc<-dim(Cl)
for(i in 1:nc[1])
{
for(j in 1:nc[2])
{
c1<-1*(M[i,j]==2)*(M[i, (j+1)]==2);
c2<-2*(M[i,j]==2)*(M[i, (j+1)]==1)
c3<-3*(M[i,j]==1)*(M[i, (j+1)]==2);
c4<-4*(M[i,j]==1)*(M[i, (j+1)]==1)
Cl[i,j]<-c1+c2+c3+c4
}
}
#---------Start Permutation Loop-----------------------
#---------Initialize Estimates----------------------------
nperm<-250;nit<-100;LRT<-1;
for(k in 1:nperm)
{
```

```
yp<-sample(Y);LL<-array(0,c(nc[2],1))
#-----Get maximum of likelihood ratio-------------
for(j in 1:nc[2])
{
y1<-(Cl[,j]==1)*yp;y1<-y1[y1!=0]
y2<-(Cl[,j]==2)*yp;y2<-y2[y2!=0]
y3<-(Cl[,j]==3)*yp;y3<-y3[y3!=0]
y4<-(Cl[,j]==4)*yp;y4<-y4[y4!=0]
T<-c(mean(y1),mean(y4),sY,.25)
for(i in 1:nit){T<-c(MLE(y1,y2,y3,y4,T[1],T[2],T[3],T[4]))}
LL[j]<-2*(logL(y1,y2,y3,y4,T[1],T[2],T[3],T[4])-LL0)
}
LRT<-c(LRT, max(LL))
}
hist(LRT,freq=F)
#-------------------------------------------------------
#-------------------------------------------------------
#----------Define Functions for P1 and P2
P1<-function(y,t1,t2,w,v)
{
temp1<-(1-v)*dnorm(y,mean=t1,sd=w)
temp2<-v*dnorm(y,mean=t2,sd=w)
return(temp1/(temp1+temp2))
}
P2<-function(y,t1,t2,w,v)
{
temp1<-v*dnorm(y,mean=t1,sd=w)
temp2<-(1-v)*dnorm(y,mean=t2,sd=w)
return(temp1/(temp1+temp2))
```

```
}
#-------------Define log likelihood---------------
logL<-function(y1,y2,y3,y4,t1,t2,w,v)
{
S1<-sum(dnorm(y1,t1,w,log=T))
S2<-sum((1-v)*dnorm(y2,t1,w,log=T)+v*dnorm(y2,t2,w,log=T))
S3<-sum(v*dnorm(y3,t1,w,log=T)+(1-v)*dnorm(y3,t2,w,log=T))
S4<-sum(dnorm(y4,t2,w,log=T))
return(S1+S2+S3+S4)
}
#--------------Define MLE Solver------------------------
MLE<-function(y1,y2,y3,y4,t1,t2,w,v)
{
n1<-length(y1);n2<-n1+length(y2)
n3<-n2+length(y3);n<-n3+length(y4)
w1<-P1(y2,t1,t2,w,v);w2<-P2(y3,t1,t2,w,v)
t1<-(sum(y1)+sum(w1*y2)+sum(w2*y3))/(n1+sum(w1)+sum(w2))
t2<-(sum((1-w1)*y2)+sum((1-w2)*y3)+sum(y4))/(sum((1-w1))
+sum((1-w2))+n-n3)
s1<-sum((y1-t1)^2)
s2<-sum(w1*(y2-t1)^2+(1-w1)*(y2-t2)^2)
s3<-sum(w2*(y3-t1)^2+(1-w2)*(y3-t2)^2)
s4<-sum((y4-t2)^2)
w<-sqrt((s1+s2+s3+s4)/n)
v<-(sum(1-w1)+sum(w2))/(n3-n1)
return(c(t1,t2,w,v))}
```

ضمیمه C

فهرست منابع

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**, 716-723.

Arunachlum, V. (1977). Heterosis for characters governed by two genes. *Journal of Genetics* **63**, 15-24.

Bailey, N. T. J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Clarendon Press, Oxford.

Bateson, W. (1902). *Mendel's Principles of Heredity: A Defence*. Cambridge University Press, London. Bliss, S., R. J. Todhunter, R. Quaas, G. Casella, R. L. Wu, G. Lust, A. J. Williams, S. Hamilton, N. L. Dykes, A. Yeager, R. O. Gilbert, N. I. Burton-Wurster, and G. M. Acland (2002). Quantitative genetics of traits associated with canine hip dysplasia in a canine pedigree constructed by mating dysplastic Labrador Retrievers with unaffected Greyhounds. *American Journal of Veterinary Research* **63**, 1029-1035.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformation (with discussion). *Journal of the Royal Statistical Society Series B* **26**, 211-252.

Bradshaw, H. D. Jr. and R. F. Stettler (1995). Molecular genetics of growth and development in *Populus*. IV. mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. *Genetics* **139**, 963-973.

Breen, M., S. Jouquand, C. Renier, C. S. Mellersh, C. Hitte, N. G. Holmes, A. Cheron, N. Suter, F. Vignaux, A. E. Bristow, C. Priat, E. McCann, C. Andre, S. Boundy, P. Gitsham, R. Thomas, W. L. Bridge, H. F. Spriggs, E. J. Ryder, A. Curson, J. Sampson, E. A. Ostrander, M. M. Binns and F. Galibert (2001). Chromosome specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes. *Mammalian Genome* **11**, 1784-1795.

Broman, K. W. (2005). The genomes of recombinant inbred lines. *Genetics* **169**, 1133-1146.

Buetow, K. N. and A. Chakravarti (1987). Multipoint gene mapping using seriation. *American Journal of Human Genetics* **41**, 189-201.

Carlin, B. P. and T. A. Louis (1998). *Bayes and Emperical Bayes Methods for Data Analysis*. Chapman Hall, New York.

Carter, T. C. and D. S. Falconer (1951). Stocks for detecting linkage in the mouse, and the theory of their design. *Journal of Genetics* **50**, 307-323.

Casella, G. and R. L. Berger (2001). *Statistical Inference*, Second Edition. Brooks- Cole.

Castle, W. E. (1921). An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. *Science* **54**, 223.

Cavalli-Sforza, L. L. and W. F. Bodmer (1971). *The Genetics of Human Populations*. W. H. Freeman and Company, San Francisco.

Celeux, G. and J. Diebolt (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* **2**, 73-82.

Chen, Z. H. (2005). The full EM algorithm for the MLEs of QTL effects and positions and their estimated variances in multiple-interval mapping. *Biometrics* **61**, 474-480.

Cheverud, J. M. and E. J. Routman(1995). Epistasis and its contribution to genetic variance components. *Genetics* **139**, 1455-1461.

Cheverud, J. M., E. J. Routman, F. A. M. Duarte, B. van Swinderen, K. Cothran and C. Perel (1996). Quantitative trait loci for murine growth. *Genetics* **142**, 1305-1319.

Churchill, G. A. and R. W. Doerge(1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-971.

Cockerham, C. C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**, 859-882.

Cockerham, C. C. 1963. Estimation of genetic variances. In: *Statistical Genetics and Plant Breeding*, edited by W. D. Hanson and H. F. Robinson. National Academy of Sciences-National Research Council, Washington, D. C. pp. 53-99.

Complex Trait Consortium (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics* **36**, 1133-1137.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall, London.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander (2001). High-resolution haplotype structure in the human genome. *Nature Genetics* **29**, 229-232.

Darvasi, A., A. Weinreb, V. Minke, J. I. Weller, and M. Soller (1993). Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134**, 943-951.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247-254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33-43.

Dawson, E., G. R. Abecasis, S. Bumpstead, Y. Chen, S. Hunt, D. M. Beare, J. Pabial, T. Dibling, E. Tinsley, S. Kirby, D. Carter, M. Papaspyridonos, S. Livingstone, R. Ganske, E. Lohmussaar, J. Zernant, N. Tonisson, M. Remm, R. Magi, T. Puurand, J. Vilo, A. Kurg, K. Rice, P. Deloukas, R. Mott, A. Metspalu, D. R. Bentley, L. R. Cardon and I. A. Dunham (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544-548.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39**, 1-38.

Doebley, J., A. Stec, and C. Gustus (1995). *Teosinte branched1* and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* **141**, 333-346.

Doebley, J., A. Stec, and L. Hubbard (1997). The evolution of apical dominance in maize. *Nature* **386**, 485-488.

Doerge, R. W. and A. Rebai (1996). Significance thresholds for QTL interval mapping tests. *Heredity* **76**, 459-464.

Dupuis, J. and D. Siegmund (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**, 373-386.

Echt, C., S. Knapp and B.-H. Liu (1992). Genome mapping with non-inbred crosses using GMendel 2.0. *Maize Genet Cooperative Newsletter* **66**, 27-29.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Ellis, T. H. N. (1997). Neighbor mapping as method for ordering genetic markers. Genetical Research **69**, 35-43.

Elston, R. C. and J. Stewart (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523-542.

Eppig, J. and E. M. Eicher (1983). The mouse linkage map. *Journal of Heredity* **74**, 213-231.

F. Chen, R. M. Biyashev, D. A. Kudrna, and B. J. Steffenson (1996). Regions of the genome that affect agronomic performance in two-row barley. *Crop Science* **36**, 1053-1062.

Falconer, D. S. and T. F. C. Mackay (1996). *Introduction to Quantitative Genetics*, Fourth Edition. Longman, New York.

Falk, C. T. (1992). Preliminary ordering of multiple linked loci using pairwise linkage data. *Genetic Epidemiology* **9**, 367-375.

Felsenstein, J. (1979). Alternative methods of phylogenetic inference and their interrelationship. *Systematic Zoology 28*, 49-62.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, London.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399-433.

Frary, A., T. C. Nesbitt, A. Frary, S. Grandillo, E. van der Knaap, B. Cong, J. P. Liu, J. Meller, R. Elber, K. B. Alpert, and S. D. Tanksley (2000). *fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**, 85-88.

Fulker, D. W. and L. R. Cardon (1994). A sib-pair approach to interval mapping of quantitative trait loci. *American Journal of Human Genetics* **54**, 1092-1103.

Gallavotti, A., Q. Zhao, J. Kyozuka, R. B. Meeley, M. K. Ritter, J. F. Doebley, M. E. Pe, and R. J. Schmidt (2004). The role of barren stalk1 in the architecture of maize. *Nature* **432**, 630-635.

Gaspin, C. and T. Schiex (1997). Genetic algorithms for genetic mapping. In *Proceedings of the Third European Conference on Evolutionary Computing*, pp. 145-156. Nimes, France.

German, S., and D. German (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741.

Grattapaglia, D. and R. R. Sederoff (1994). Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* **137**, 1121-1137.

Grattapaglia, D., F. L. G. Bertolucci, R. Penchel and R. R. Sederoff (1996). Genetic mapping of quantitative trait loci controlling growth and wood quality traits in *Eucalyptus grandis* using a maternal half-sib family and RAPD markers. *Genetics* **144**, 1205-1214.

Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Griffiths, A. J. F., W. M. Gelbart, R. C. Lewontin and J. H. Miller (2002). *Modern Genetic Analysis: Integrating Genes and Genomes*, Second Edition. W. H. Freeman, New York.

Groover, A. T., M. Devey, T. Fiddler, J. Lee, R. Megraw, T. Mitchel-Olds, B. Sherman, S. Vujcic, C. Williams and D. Neale (1994). Identification of quantitative trait loci influencing wood specific gravity in an outbred pedigree of loblolly pine. *Genetics* **138**, 1293-1300.

Gutierrez, R. G., R. J. Carroll, N.Wang, G.-H. Lee, and B. H. Taylor (1995). Analysis of tomato root initiation using a normal mixture distribution. *Biometrics* **51**, 1461-1468.

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics* **8**, 299-309.

Haldane, J. B. S. and C. H. Waddington (1931). Inbreeding and linkage. *Genetics* **16**, 357-374.

Haley, C. S. and S. A. Knott (1992). A simple method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315-324.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science* **28**, 49-50.

Hartl, D. L. and E. W. Jones (2001) *Genetics: Analysis of Genes and Genomes*, Sixth Edition. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Haseman, J. K. and R. C. Elston (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavioral Genetics* **2**, 3-19.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Hill, W. G. and J. Rasbash, (1986). Models of long term artificial selection in finite population. *Genetical Research* **48**, 4150.

Hoeschele, I., P. Ulimari, F. E. Grignola, Q. Zhang and K. M. Gage (1997). Advances in statistical methods to map quantitative trait loci in outbreed populations. *Genetics 147*, 1445-1457.

Huang, N., A. Parco, T. Mew, G. Magpantay, S. McCouch, E. Guiderdoni, J. Xu, P. Subudhi, E. R. Angeles and G. S. Khush (1997). RFLP mapping of isozymes,

Jansen, J., A. G. de Jong, and J. W. van Ooijen (2001). Constructing dense genetic linkage maps. *Theoretical and Applied Genetics* **102**, 1113-1122.

Jansen, R. C. (1994). Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics* **138**, 871-881.

Jansen, R. C. and P. Stam. (1994). High resolution mapping of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447-1455.

Jansen, R. C., H. Geerlings, A. J. Van Oeveren and R. C. Van Schaik (2001). A comment on codominant scoring of AFLP markers. *Genetics* **158**, 925-926.

Jeffreys, H. (1961). *Theory of Probability*, Third Edition. Oxford University Press, New York.

Jennings, H. S. (1917). The numerical results of diverse systems of breeding with respect to two characters, etc. *Genetics* **2**, 97-154.

Johnson, D. L., R. C. Jansen and J. A. M. Van Arendonk (1999). Mapping quantitative trait loci in a selectively genotyped outbred population using a mixture model approach. *Genetical Research* **73**, 75-83.

Kao, C.-H. (2000). On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics* **156**, 855-865.

Kao, C.-H. and Z.-B. Zeng (1997). General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 653-665.

Kao, C.-H. and Z.-B. Zeng (2002). Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* **160**, 1243-1261.

Kao, C.-H., Z.-B. Zeng, and R. D. Teasdale (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203-1216.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

Kimura, M. (1979). The neutral theory of molecular evolution. *Scientific American* **241**, 98-126.

Kosambi, D. D. (1944). The estimation of map distance from recombination values. *Annals of Eugenics* **12**, 172-175.

Lande, R. (1981). The minimum number of genes contributing to quantitative variation between and within populations. *Genetics* **99**, 541-553.

Lande, R. and R. Thompson (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743-756.

Lander, E. S. and D. Botstein (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-199.

Lander, E. S. and N. J. Schork (1994). Genetic dissection of complex traits. *Science* **265**, 2037-2048.

Lander, E. S. and P. Green (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences USA* **84**, 2363-2367.

Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. Daly, S. Lincoln and L. Newburg (1987). MapMaker: an interactive computer package for constructing genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174-181.

Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*. Springer Verlag, New York.

Lathrop, G. M., J. M. Lalouel, C. Julier and J. Ott (1984). Strategies for multilocus linkage analysis in human. *Proceedings of the National Academy of Sciences USA* **81**, 3443-3446.

Lathrop, G. M., J. M. Lalouel, C. Julier and J. Ott (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics* **37**, 482-498.

Lehmann, E. and G. Casella (1998). *Theory of Point Estimation*. Springer-Verlag, New York.

Leips, J. and T. F. C. Mackay (2000). Quantitative trait loci for lifespan in *Drosophila melanogaster*: interactions with genetic background and larval density. *Genetics* **155**, 1773-1788.

Lewin, B. (2005) *Essential Genes*. Prentice-Hall, Engewood Cliff, New Jersey.

Li, B. and R. L. Wu. (1996). Genetic causes of heterosis in juvenile aspen: a quantitative comparison across intra-and interspecific hybrids. *Theoretical and Applied Genetics* **93**, 380-391.

Li, C. B., A. L. Zhou, T. Sang (2006). Rice domestication by reducing shattering. *Science* **311**, 1936-1939.

Liberman, U. and S. Karlin (1984). Theoretical models of genetic map functions. *Theoretical Population Biology* **25**, 331-346.

Lin, M., X.-Y. Lou, M. Chang and R. L. Wu (2003). A general statistical framework for mapping quantitative trait loci in non-model systems: Issue for characterizing linkage phases. *Genetics* **165**, 901-913.

Liu, B.-H. (1998). *Statistical Genomics: Linkage, Mapping, and QTL Analysis*. CRC Press, New York.

Liu, T., R. J. Todhunter, S. Wu, W. Hou, R. Mateescu, Z. W. Zhang, N. I. Burton-Wurster, G. M. Acland, G. Lust and R. L. Wu (2007) A random model for mapping imprinted quantitative trait loci in a structured pedigree: An implication for mapping canine hip dysplasia. *Genomics* (in press).

Lo, Y. T., N. R. Mendell, and D. B. Rubin (2001). Testing the number of components in a normal mixture, *Biometrika* **88**, 767-778.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B* **44**, 226-233.

Lu, Q., Y. H. Cui and R. L. Wu (2004). A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full-sib family. *BMC Genetics* **5**, 20.

Lynch, M. and B.Walsh (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Mackay, T. F. C. (1996). The nature of quantitative genetic variation revisited: lessons from *Drosophila* bristles. *BioEssays* **18**, 113-121.

Mackay, T. F. C. (2001). Quantitative trait loci in *Drosophila*. *Nature Reviews Genetics* **2**, 11-20.

Mackay, T. F. C., R. F. Lyman, and M. S. Jackson (1992). Effects of P element inserts on quantitative traits in *Drosophila melanogaster*. *Genetics* **130**, 315-332.

Maliepaard, C., J. Jansen and J. W. van Ooijen (1997). Linkage analysis in a fullsib family of an outbreeding plant species: overview and consequences for applications. *Genetical Research* **70**, 237-250.

Mangin, B., B. Goffinet, and A. Rebai (1994). Constructing confidence intervals for QTL location. *Genetics* **138**, 1301-1308.

Martin, O. C. and F. Hospital (2006). Two and three-locus tests for linkage analysis using recombinant inbred lines. *Genetics* **173**, 451-459.

Martinez, O. and R. N. Curnow (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480-488.

Mather, K. (1938). Crossing-over. *Biological Review* **13**, 252-292.

Mather, K. (1943). Polygenic inheritance and natural selection. *Biological Review* **18**, 32-65.

Mather, K. and J. L. Jinks (1982). *Biometrical Genetics*, Third Edition. Chapman and Hall, London.

Mayer, M. (2005). A comparison of regression interval mapping and multiple interval mapping for linked QTL. *Heredity* **94**, 599-605.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. Wiley, New York. Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood via the ECM algorithm: A general framework. *Biometrika* **80**, 267-278.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. Wiley, New York.

Meng, X.-L. and D. B. Rubin (1991). Using EM to obtain asymptotic variancecovariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86**, 899-909.

Mester, D., Y. Ronin, D. Minkov, E. Nevo and A. Korol (2003). Constructing largescale genetic maps using an evolutionary strategy algorithm. *Genetics* **165**, 2269-2282.

Minvielle, F. (1987). Dominance is not necessary for heterosis: a two-locus model. *Genetical Research* **49**, 245-247.

Morgan, T. H. (1909). What are "factors" in Mendelian explanations? *American Breeders Association Reports* **5**, 365-368.

Morgan, T. H. (1928). *The Theory of Genes*. Yale University Press, New Haven, Connecticut.

Morton, N. E. (1956). The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. *American Journal of Human Genetics* **8**, 80-96.

Muller, H. J. (1916). The mechanism of crossing over. *American Naturalist* **50**, 193-207.

Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. *Computer Journal* **7**, 308-313.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society Series B* **56**, 3-48.

Niklas, K. L. (1994). *Plant Allometry: The Scaling of Form and Process*. University of Chicago Press, Chicago, Illinois.

Olson, J. M., and M. Boehnke (1990). Monte Carlo comparison of preliminary methods of ordering multiple genetic loci. *American Journal of Human Genetics* **47**, 470-482.

Ott, J. (1991). *Analysis of Human Genetic Linkage.* 2nd edition. John Hopkins University Press, Baltimore, Maryland.

Otto, S. P. and M. W. Feldman (1997). Deleterious mutations, variable epistatic interactions, and the evolution of recombination. *Theoretical Population Biology* **51**, 134-147.

Piepho, H. P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics* **157**, 425-432.

Piepho, H. P. and G. Koch (2000). Codominant analysis of banding data from a dominant marker system by normal mixtures. *Genetics* **155**, 1459-1468.

RAPD and QTLs for grain shape, brown planthopper resistance in a doubled haploid rice population. *Molecular Breeding* **3**, 105-113.

Rebai, A., B. Goffinet, and B. Mangin (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics* **138**, 235-240

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B* **59**, 731-792.

Ridout, M. S., S. Tong, C. J. Vowden, and K. R. Tobutt (1998). Three-point linkage analysis in crosses of allogamous plant species. *Genetical Research* **72**, 111-121.

Ritter, E. and F. Salamini (1996). The calculation of recombination frequencies in crosses of allogamous plant species with applications to linkage mapping. *Genetical Research* **67**, 55-65.

Ritter, E., C. Gebhardt, and F. Salamini (1990). Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. *Genetics* **125**, 645-654.

Robbins, R. B. (1918). Some applications of mathematics to breeding problems. III. *Genetics* **3**, 375-379.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods*, Second Edition. Springer Verlag, New York.

Routman, E. J. and J. M. Cheverud (1997). Gene effects on a quantitative trait: two locus epistatic effects measured at microsatellite markers and at estimated QTL. *Evolution* **51**, 1654-1662.

Scheiner, S. M. (1993). Genetics and evolution of phenotypic plasticity. *Annual Review of Ecology and Systematics* **24**, 35-68.

Schlichting, C. D. (1986). The evolution of phenotypic plasticity in plants. *Annual Reviews of Ecology and Systematics* **17**, 667693.

Schlichting, C. D. and M. Pigliucci (1998). *Phenotypic Evolution: A Reaction Norm Perspective*. Sinauer Associates, Sunderland, Massachusetts.

Schnell, F. W. and C. C. Cockerham (1992). Multiplicative vs. arbitrary gene action in heterosis. *Genetics* **131**, 461-469.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

Schäfer-Pregl, R., F. Salamini, and C. Gebhardt (1996). Models for mapping quantitative trait loci (QTL) in progeny of non-inbred parents and their behaviour in the presence of distorted segregation ratios. *Genetical Research* **67**, 43-54.

Sham, P. (1998). *Statistics in Human Genetics*. Wiley, New York.

Shrimpton, A. E., and A. Robertson (1988a). The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*. I: allocation of third chromosome sterno-pleural bristle effects to chromosome sections. *Genetics* **118**, 437-443.

Shrimpton, A. E., and A. Robertson (1988b). The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*. II: Distribution of third chromosome bristle effects within chromosome sections. *Genetics* **118**, 445-459.

Sillanpa, M. J. and E. Arjas (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151**, 1605-1619.

Song, J. Z., M. Soller, and A. Genizt (1999). The full-sib intercross line (FSIL): a QTL mapping design for outcrossing species. *Genetical Research* **73**, 61-73.

Speed, T. P., M. S. McPeek, and S. N. Evans (1992). Robustness of the noninterference model for ordering genetic markers. *Proceedings of the National Academy of Sciences USA* **89**, 3103-3106.

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *Plant Journal* **3**, 739-744.

Stephens, D. A. and R. D. Fisch (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* **54**, 1334-1347.

Stuart, A., J. K. Ord, and S. Arnold (1999). *Kendall's Advanced Theory of Statistics* Volume 2A. Oxford University Press, New York.

Suiter, K. A., J. F. Wendel and J. Case (1983). SLINKAGE-1: a PASCAL computer program for the detection and analysis of genetic linkage. *Journal of Heredity* **74**, 203-204.

Tan, Y.-D. and Y.-X. Fu (2006). A novel method for estimating linkage map. *Genetics* **173**, 2383-2390.

Teuscher, F., V. Guiard, P. E. Rudolph, and G. A. Brockmann (2005). The map expansion obtained with recombinant inbred strains and intermated recombinant inbred populations for finite generation designs. *Genetics* **170**, 875-879.

The International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**, 789-794.

Thompson, E. A. (1984). Information gain in joint linkage analysis. *IMA Journal of Mathematics Applied in Medicine and Biology* **1**, 31-49.

Threadgill, D. W., K. W. Hunter, and R. W. Williams (2002). Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mammalian Genome* **13**, 175-178.

Tinker, N. A., D. E. Mather, B. G. Rossnagel, K. J. Kasha, A. Kleinhofs, P. M. Hayes, D. E. Falk, T. Ferguson, L. P. Shugar, W. G. Legge, R. B. Irvine, T. M. Choo, K. G. Briggs, S. E. Ullrich, J. D. Franckowiak, T. K. Blake, R. J. Graf, S. M. Dofing, M. A. Saghai Maroof, G. J. Scoles, D. Hoffman, L. S. Dahleen, A. Kilian,

Todhunter, R. J., R. Mateescu, G. Lust, N. I. Burton-Wurster, N. L. Dykes, S. P. Bliss, A. J. Williams, M. Vernier-Singer, E. Corey, C. Harjes, R. L. Quaas, Z. Zhang, R. O. Gilbert, D. Volkman, G. Casella, R. L. Wu, and G. M. Acland (2005). Quantitative trait loci for hip dysplasia in a crossbreed canine pedigree. *Mammalian Genome* **16**, 720-730.

van Ooijen, J. W. (1999). LOD significance thresholds for QTL analysis in experimental populations of diploid species. *Heredity* **83**, 613-624.

Vaughn, T. T., L. S. Pletscher, A. Peripato, K. King-Ellison, E. Adams, C. Erikson, and J. M. Cheverud (1999). Mapping quantitative trait loci for murine growth- a closer look at genetic architecture. *Genetical Research* **74**, 313-322.

Via, S., R. Gomulkiewicz, G. de Jong, S. E. Scheiner, C. D. Schlichting, and P. van Tienderen (1995). Adaptive phenotypic plasticity: consensus and controversy. *Trends in Ecology and Evolution* **10**, 212-217.

Visscher, P. M., C. S. Haley, and R. Thompson (1996). Marker-assisted introgression in backcross breeding programs. *Genetics* **144**, 1923-1932.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307-333.

Wall, J. D. and J. K. Pritchard (2003). Haplotype blocks and the structure of linkage disequilibrium in the human genome. *Nature Reviews Genetics* **4**, 587-597.

Wang, H., T. Nussbaum-Wagler, B. L. Li, Q. Zhao, Y. Vigouroux, M. Faller, K. Bomblies, L. Lukens, and J. F. Doebley (2005). The origin of the naked grains of maize. *Nature* **436**, 714-719.

Weeks, D. and K. Lange (1987). Preliminary ranking procedures for multilocus ordering. *Genomics* **1**, 236-242.

Weinberg, W. (1908). Uber den Nachweis der Vererbung beim Menschen. *Jh. Ver. vaterl Naturk. Wurttemb.* **64**, 369-382.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates. Sunderland, Massachusetts.

West, G. B., J. H. Brown and B. J. Enquist (1997). A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122-126.

West, G. B., J. H. Brown and B. J. Enquist (1999). The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science* **284**, 1677-1679.

Wilson, S. (1988). A major simplification in the preliminary ordering of linked loci. *Genetic Epidemiology* **5**, 75-80.

Wu, R. L. (1996). Quantitative genetic dissection of complex traits in a QTL-mapping pedigree. *Theoretical and Applied Genetics* **93**, 447-457.

Wu, R. L. (1998). The detection of plasticity genes in heterogeneous environments. *Evolution* **52**, 967-977.

Wu, R. L., C.-X. Ma, I. Painter and Z.-B. Zeng (2002a). Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. *Theoretical Population Biology* **61**, 349-363.

Wu, R. L., C.-X. Ma, S. S. Wu, and Z-B. Zeng (2002b). Linkage mapping of sexspecific differences. *Genetical Research* **79**, 85-96.

Wullschleger, S. D., T. M. Yin, S. P. DiFazio, T. J. Tschaplinski, L. E. Gunter, M. F. Davis, and G. A. Tuskan (2005). Genotypic variation in growth and biomass distribution for two advanced-

generation (F2) pedigrees of hybrid poplar (*Populus* spp.). *Canadian Journal of Forest Research* **35**, 1779-1789.

Xu, S. (1995). A comment on the simple regression method for interval mapping. *Genetics* **141**, 1657-1659.

Xu, S. (1998). Mapping quantitative trait loci using multiple families of line crosses. *Genetics* **148**, 517-524.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789-801.

Xu, S. and W. R. Atchley (1995). A random model approach to interval mapping of quantitative trait loci. *Genetics* **141**, 1189-1197.

Yan, J. Q., J. Zhu, C. X. He, M. Benmoussa, and P. Wu. (1998). Molecular dissection of developmental behavior of plant height in rice (*Oryza sativa* L.), *Genetics* **150**, 1257-1265.

Yin, T. M., M. R. Huang, M. X. Wang, Z.-B. Zeng and R. L. Wu (2001). Interspecific linkage maps of *Populus adenopoda* × *P. alba*. *Genome* **44**, 602-609.

Yin, T. M., S. P. DiFazio, L. E. Gunter, D. Riemenschneider, and G. A. Tuskan (2004). Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. *Theoretical and Applied Genetics* **109**, 451-463.

Yin, T. M., X. Y. Zhang, M. R. Huang, M. X. Wang, Q. Zhuge, S. M. Tu, L.-H. Zhu and R. L. Wu (2002). The molecular linkage maps of the Populus genome. *Genome* **45**, 541-555.

Zeng, Z.-B. (1992). Correcting the bias of Wright's estimates of the number of genes affecting a quantitative character—A further improved method. *Genetics* **131**, 987-1001.

Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences USA* **90**, 10972-10976.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457-1468.

Zeng, Z.-B., D. Houle and C. C. Cockerham (1990). How informative is Wright's estimator of the number of genes affecting a quantitative character? *Genetics* **126**, 235-247.

Zeng, Z.-B., J. Liu, L. F. Stam, C.-H. Kao, J. M. Mercer and C.C. Laurie (2000). Genetic architecture of a morphological shape difference between two *Drosophila* species. *Genetics* **154**, 299-310.

Zhang, W.-K., Y.-J. Wang, G.-Z. Luo, J.-S. Zhang, C.-Y. He, X.-L. Wu, J.-Y. Gai, and S.-Y. Chen (2004). QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. *Theoretical and Applied Genetics* **108**, 1131-1139.

Zou, F., J. P. Fine, J. Hu, and D. Y. Lin (2004). An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics* **168**, 2307-2316.

واژه‌نامه

Additive genetic variance	واریانس ژنتیکی افزایشی
Allocation rate	نرخ تخصیص
Association analysis	تجزیه ارتباط
Asymmetrical marker cross types	انواع تلاقی نشانگری نامتقارن
Asymptotic approximation	تقریب مجانبی
Backcross	تلاقی برگشتی
Bayesian estimation	برآورد بیزین
Bias	اریبی
Binomial distribution	توزیع دوجمله‌ای
Bootstrap	خودراه‌اندازی
Broad-sense heritability	وراثت پذیری عمومی
Chiasma interference	تداخل کیاسمایی
Chi-squared approximation	تقریب کای مربع
Chromatid interference	تداخل کروماتیدی
Cluster analysis	تجزیه خوشه‌ای
Codominant marker	نشانگر هم بارز
Cofactor	متغیر همراه
Coincidence	تطابق
Composite interval mapping	مکان‌یابی فاصله‌ای مرکب
Conditional distribution	توزیع شرطی
Conditional probability	احتمال شرطی
Confidence interval	فاصله اطمینان
Cosegregation	تفرق همزمان
Coupling	حالت جفتی
Degree of freedom	درجه آزادی
Discrete distribution	توزیع گسسته
Distorted marker	نشانگر اریب
DNA probe	کاوشگر DNA
Dominance genetic variance	واریانس ژنتیکی غالبیت
Dominant marker	نشانگر غالب

تلاقی برگشتی مضاعف	Double backcross
نوترکیبی مضاعف	Double recombination
اثر اپیستازی	Epistatic effect
صحت برآورد	Estimation accuracy
برآوردگر	Estimator
احتمال مثبت کاذب	probability False-positive
افراد پایه‌گذار	Founders
خانواده تنی	Full-sib family
نشانگر حاوی اطلاعات کامل	Fully informative marker
زنده‌مانی گامتی	Gametic viability
تابع گاما	Gamma function
برآورد تعداد ژن	Gene enumeration
معماری ژنتیکی	Genetic architecture
مکان‌یابی ژنتیکی	Genetic mapping
انتقال ژنتیکی	Genetic transmission
ارزش ژنوتیپی	Genotypic value
سری‌های هندسی	Geometric series
قانون هاردی-واینبرگ	Hardy-Weinberg law
وراثت‌پذیری	Heritability
ناهمگنی واریانس	of variance Heterogeneity
واریانس همگن	Homogeneous variance
نشانگر با تنوع بسیار زیاد	Hypervariable marker
آزمون فرض	Hypothesis testing
آلل‌های یکسان اجدادی	Identical-by-descent alleles
ماتریس تلاقی	Incidence matrix
جور شدن مستقل	Independent assortment
اثر متقابل	Interaction
تداخل	Interference
مکان‌یابی فاصله‌ای	Interval mapping
تکرار	Iteration

Least-squares analysis	تجزیه حداکثر درستنمایی
Likelihood estimation	برآورد درستنمایی
Likelihood function	تابع درستنمایی
Likelihood ratio test	آزمون نسبت درستنمایی
Linkage	پیوستگی ژنی
Locus	مکان ژنی
Locus ordering	مرتب‌سازی مکان‌های ژنی
Map construction	ایجاد نقشه ژنتیکی
Map distance	فاصله نقشه
Map function	تابع نقشه
Marker	نشانگر
Marker density	تراکم نشانگری
Maximum likelihood	حداکثر درستنمایی
Mechanistic model	مدل مکانیسمی
Meiosis	تقسیم میوز
Meiosis interference	تداخل میوزی
Mendelian segregation	تفرق مندلی
Microsatellite marker	نشانگر ریزماهواره
Misclassified markers	نشانگرهای فاقد طبقه‌بندی و بی‌شکل
Mitosis	تقسیم میتوز
Mixture likelihood	درستنمایی مختلط
Mixture model	مدل مختلط
Molecular genetics	ژنتیک مولکولی
Multialleles	حالت چندآللی
Multilocus likelihood	درستنمایی چندمکانی
Multinomial distribution	توزیع چندجمله‌ای
Multiple interval mapping	مکان‌یابی فاصله‌ای چندگانه
Multiplicative-epistatic model	مدل اپیستازی ضرب‌پذیر
Narrow-sense heritability	وراثت پذیری خصوصی
Nonadditive effect	اثر غیرافزایشی

Nuclear family	خانواده هسته
Null allele	آلل خاموش
Numerator	صورت کسر
Offspring	نتاج
Outbred cross	تلاقی آزاد
Outbred populations	جمعیت‌های غیرهمخون
Outcrossing species	گونه‌های دگربارور
Overdominance	فوق غالبیت
Partially informative marker	نشانگر حاوی اطلاعات جزیی
Pedigree	شجره
Permutation test	آزمون تبدیل یا جایگشت
Pleiotropic effect	چنداثری یک ژن
Polymerase chain reaction	واکنش زنجیره‌ای پلی‌مراز
Posterior distribution	توزیع پسین
Power transformation	تبدیل نمایی
Precision analysis	تجزیه دقت
Primer	آغازگر
Prior distribution	توزیع پیشین
Progeny	فرزند
Putative QTL	QTL مفروض
Quantitative trait	صفت کمّی
Quantitative trait loci	جایگاه‌های ژنی صفات کمّی
Random sample	نمونه تصادفی
Recessive allele	آلل مغلوب
Recombinant inbred lines	لاین‌های خویش‌آمیخته نوترکیب
Recombinant types	انواع نوترکیب
Recombination fraction	نرخ نوترکیبی
Regression analysis	تجزیه تابعیت یا رگرسیون
Repulsion	حالت ناجفت
Sample density	چگالی نمونه

Sampling error	خطای نمونه‌برداری
Segregation ratio	نسبت تفرق
Selfing	خودگشنی
Sequence polymorphism	چندشکلی توالی
Sex-linked genes	ژن‌های وابسته به جنس
Sibling mating	آمیزش برادر-خواهری
Simulation	شبیه‌سازی
Single Nucleotide Polymorphism (SNP)	چندشکلی تک‌نوکلئوتیدی
Statistical genetics	ژنتیک آماری
Structured pedigrees	شجره‌های ساختاری
Threshold	آستانه
Whole-genome marker analysis	تجزیه نشانگری کل ژنوم
Zygotic viability	زنده‌مانی زیگوتی

University of Guilan Press

Statistical Genetics of Quantitative Traits
Linkage, Maps and QTL

By:
Rongling Wu Chang-Xing Ma George Casella

Translated by:
Navid Ghavi Hossein-Zadeh, Ph. D
Babak Rabiei, Ph. D