رویکرد بیزی به استنتاج ساختار جمعیت

مهرداد تمیجی [۱] ، **سید محمود طاهری** [۲]

تاریخ دریافت: ۹۹/۹/۱۲
تاریخ پذیرش: ۹۹/۱۲/۲۷

چکیده:
روش‌های استنتاج ساختار جمعیت، و کاربردهای آن در شناسایی بیماری‌ها و آینده‌نگری دربارهٔ وضعیت جسمی و روانی انسان‌ها، اهمیت روزافزون یافته است. در این مقاله، ابتدا به بررسی انگیزه و اهمیت بررسی ساختار جمعیت پرداخته‌شده است. سپس کاربردهای استنتاج ساختار جمعیت در زیست‌شناسی و درمان انواع بیماری‌ها شرح داده‌شده است. آنگاه روش‌های استنتاج ساختار جمعیت و همچنین یافتن مدل بیماری متناظر با هر زیرجمعیت، برای جمعیت‌هایی که اعضای آن مخلوط یا غیرمخلوط هستند، به تفکیک، تشریح شده‌اند. دراین‌باره، بر روش‌های استنتاج ساختار جمعیت با رویکرد بیزی، تأکید شده است. و دلایل برتری روش‌های بیزی بیان‌شده‌اند.

واژه‌های کلیدی: استنتاج ساختار جمعیت، مدل گرافیکی احتمالاتی، بیوانفورماتیک، مطالعات گسترده ژنوم، زنجیره مارکوف، جمعیت مخلوط

۱ انگیزه و اهمیت کار

از دید شما چگونه می‌توان بر پایه شناخت و خوشه‌بندی ژن‌های هر انسان، به بیماری‌های کنونی و آینده وی پی برد؟

این موضوع یکی از هدف‌هایی است که در دهه‌های اخیر به آن توجه شده است. بر همین اساس شناسایی بهتر مکان‌های مؤثر در بیماری‌ها بر اساس ژنوتیپ افراد بسیار با اهمیت است. بدین منظور باید مشکلات بزرگی مثل در اختیار داشتن ژنوم افراد مختلف از نژادها و محل‌های مختلف حل می‌شد. خوشبختانه با پیشرفت دانش توالی‌یابی [۳] ژنوم، راه برای تولید داده‌های زیستی بیش‌ازپیش فراهم‌شده است. از سوی دیگر پروژه‌های بزرگ در حوزه زیستی نظیر پروژه ژنوم [۴] [۱۰] و پروژه HapMap (نقشه هاپلوتایپ) [۵] نتایج بسیار ارزشمندی مانند ژنوتیپ افراد متعلق به نژادهای مختلف برای تحقیق در اختیار ما گذاشته‌اند.

یکی از مهم‌ترین بررسی‌ها، بررسی در حوزه GWAS (مطالعهٔ ارتباطات گسترده ژنوم) [۵] [۹] است که اصلی‌ترین هدف آن پیدا کردن مکان و ارتباط ژن‌های مختلف با ویژگی فیزیکی بیان‌شده توسط آن ژن‌ها است. به‌عنوان‌مثال پیدا کردن جاهای مشخص در توالی DNA با عنوان SNP (تک نوکلئوتید چندریختی) [۶] یا Microarray ها (ریزآرایه) به‌عنوان عامل مؤثر در یک بیماری خاص یکی از وظایف اصلی GWAS است. ازآنجاکه تغییرات ژنتیکی نسبت به ژنوم مرجع در افراد بیمار نسبت به افراد سالم بیشتر اتفاق می‌افتد، پیدا کردن این تغییرات نشان‌دهنده پیدا کردن ژن‌های مؤثر در یک بیماری است.

یکی از مراحل رایج برای یک تحقیق در حوزه GWAS ،پیدا کردن ساختار جمعیت‌ها است. محققین با استفاده از داده‌های ژنوتیپ، فنوتیپ، سن، مکان زندگی و ... مربوط به هر جمعیت، ساختار آن جمعیت را برای اهدافی همچون تشخیص انتخاب طبیعی، جهش و یا درمان بیماری تشخیص می‌دهند و جمعیت را خوشه‌بندی می‌کنند.

در این زمینه دو سناریو وجود دارد.

۱. محققین می‌خواهند چندین نمونه از افراد یک جمعیت را انتخاب کرده و در مورد ویژگی‌های آن جمعیت اظهارنظر کنند. برای مثال فهمیدن تکامل انسان در یک منطقه.
۲. محققین یک جمعیت را بررسی کرده و سعی می‌کنند افراد در دست بررسی را به یک یا چند زیرجمعیت نسبت دهند تا هر زیرجمعیت جداگانه موردبررسی قرار گیرد. به‌عنوان‌مثال خوشه‌بندی و پیدا کردن

[۱] کارشناسی ارشد مهندسی کامپیوتر (الگوریتم‌ها و محاسبات) (tamiji.mehrdad@alumni.ut.ac.ir)
[۲] استاد دانشگاه تهران (نویسنده مسئول) sm_taheri@ut.ac.ir

[3] Sequencing
[4] Genome Project
[5] Genome-Wide Association Study
[6] Single Nucleotide Polymorphism

ساختار جمعیت افراد حاضر در کشورهای مختلف آسیا.

در هر دو وضعیت بالا اولین گام مهم، تعریف و تشخیص زیرجمعیت‌ها است. گرچه تعریف زیرجمعیت می‌تواند بر پایه ویژگی‌هایی مانند زبان، فرهنگ، ویژگی‌های فیزیکی، محل جغرافیایی و ... باشد، اما فهمیدن اینکه آیا خوشه‌های تعریف‌شده بر اساس این ویژگی‌ها ازنظر ژنتیکی هم درست هستند یا خیر کار سختی است زیرا در نظر گرفتن موارد تاریخی همچون مهاجرت دست جمعی، جنگ، خشک‌سالی و ... در مدل کار سختی است. همچنین خوشه‌بندی بر اساس یک ویژگی قابل‌رؤیت یا همان فنوتیپ گرچه بسیار مفید است، اما کار بسیار دشواری است زیرا عوامل مختلف ژنتیکی می‌تواند منجر به رخداد یک فنوتیپ شوند و بدون در نظر گرفتن ژنوتیپ، نتیجه خوشه‌بندی دقیق نخواهد بود. اما خوشه‌بندی این موارد (خوشه‌بندی بر اساس فنوتیپ) که با نام جمعیت پنهان شناخته می‌شوند، از طریق ژنتیکی قابل انجام است.

برای مثال خوشه‌بندی جمعیت، نتایج حاصل از تحقیق پریچارد [7] و همکاران [۱۷] که تا زمان نوشتن این مقاله بیش از ۲۹۶۰۰ ارجاع به آن شده است و به نام STRUCTURE معروف است در شکل ۱ نمایش داده‌شده است که پرنده‌هایی در سه منطقه مختلف Chawia، Ngangao و Mbololo بررسی‌شده‌اند. ژنوتیپ هر پرنده در ۷ مکان بررسی‌شده است و الگوریتم با دقت نسبتاً خوبی توانسته است خوشه‌بندی را انجام دهد. همان‌طور که در شکل مشخص است الگوریتم به‌خوبی مرز بین زیرجمعیت‌ها رو مشخص کرده و آن‌ها را به‌درستی تفکیک کرده است .دقت شود که خطاهای الگوریتم به‌صورت عددی از ۱ تا ۴ در شکل مشخص‌شده‌اند.

در مقالهٔ پیش‌رو، ابتدا چند اصطلاح زیستی ضروری برای پیگیری و درک بهتر مطالب، مرور می‌شوند. در بخش سوم به‌مرور برخی مفاهیم و روابط آماری مرتبط با رویکرد بیزی شامل توزیع دیریکله، الگوریتم زنجیره مارکوف مونت‌کارلو، نمونه‌گیری گیبز و الگوریتم متروپولیس-هستینگس پرداخته می‌شود. در بخش ۴، کاربردهای استنتاج ساختار جمعیت را بیان می‌کنیم و کارکردهای GWAS در درمان انواع بیماری‌ها را شرح می‌دهیم. در بخش ۵، روش‌های رایج در تشخیص ساختار جمعیت را شرح می‌دهیم. در بخش ۶، استنتاج ساختار جمعیت با روشی معروف به STRUCTURE که مبتنی بر رویکرد بیزی است برای دو حالت جمعیت غیر مخلوط و مخلوط شرح داده می‌شود و در ادامه استنتاج ساختار جمعیت و کشف ارتباطات هر زیرجمعیت با یک بیماری خاص برای دو حالت جمعیت غیر مخلوط و مخلوط بیان می‌شود. در بخش پایانی، نتیجه‌گیری و برتری روش‌های بیزی را بیان می‌کنیم.

۲ تعاریف و مفاهیم ژنتیکی

در این بخش مفاهیم و تعاریف مرتبط با علم ژنتیک را که در ادامهٔ مقاله به آن‌ها نیاز داریم، ارائه می‌کنیم.

DNA

دی اکسیدریبونوکلئیک (DNA) [8] یک مولکول است که یک طرح ژنتیکی را کدگذاری می‌کند. DNA شامل دو رشته خطی است که در مقابل یکدیگر قرار می‌گیرند که به‌عنوان رشته‌های غیر-موازی [9] شناخته می‌شوند؛ این رشته‌ها باهم تلاقی و یک مارپیچ دوگانه بسیار بلند را ایجاد می‌کنند. ساختار DNA بدون در نظر گرفتن مارپیچ را می‌توان به‌صورت نردبان توصیف کرد. ستون نردبان از مولکول‌های شکر و فسفات ساخته‌شده است که توسط پیوندهای شیمیایی متصل می‌شوند و پله‌های نردبان جفت واحدهایی نیتروژنی هستند که با نام‌های A، C، G و T شناخته می‌شوند. دقت شود که به‌غیراز مواردی مثل جهش، همواره ارتباط بین A و T یا بین C و G است. این جفت‌ها، جفت باز [10] نامیده می‌شوند و دو پایه فسفات قند را از طریق پیوند هیدروژنی به هم وصل می‌کنند. در شکل ۲ ظاهر DNA و ساختار شیمیایی ۴ نوع باز ذکرشده را مشاهده می‌کنید.

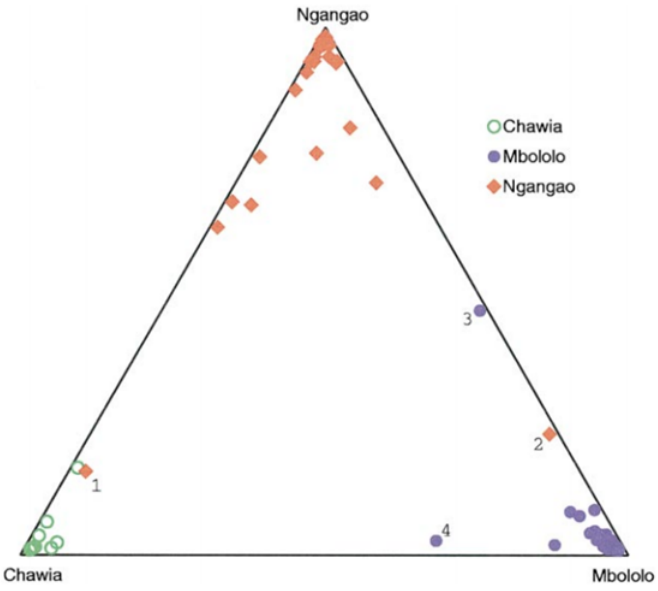
ژنوتیپ

ژنوتیپ [11] به آرایش ژنتیکی موجودات گفته می‌شود. به سخن دیگر، ژنوتیپ مجموعه کاملی از ژن‌های موجودات زنده را توصیف می‌کند. هر انسان تقریباً ۳۰۰۰۰ ژن دارد که بر روی ۲۳ جفت کروموزوم پخش‌شده‌اند. یک ژن می‌تواند بر اساس محیط، شرایط و موارد دیگر غالب یا مغلوب باشد و یا بر روی یک یا چند ویژگی تأثیر مستقیم یا غیرمستقیم داشته باشد.

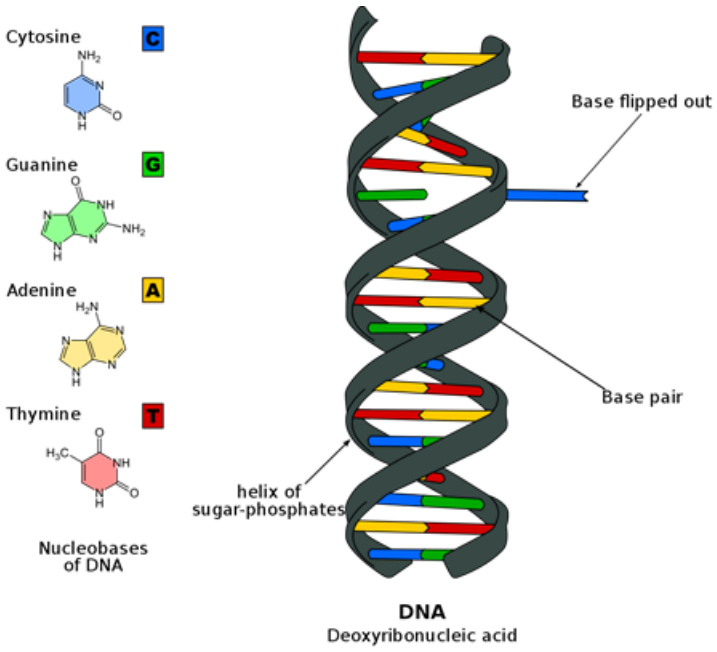
فنوتیپ

فنوتیپ [12] به ویژگی‌های فیزیکی مشاهده‌پذیر هر موجود گفته می‌شود و شامل ظهور، توسعه و رفتار موجود است. فنوتیپ‌های هر موجود عمدتاً توسط ژنوتیپ آن و همچنین تأثیرات محیطی بر روی موجود تعیین می‌شوند. برای مثال، ارتفاع و طول بال در پرندگان و همچنین رنگ مو در موجودات مختلف

[7] Pritchard
[8] Deoxyribonucleic
[9] Anti-Parallel
[10] Base Pair
[11] Genotype
[12] Phenotype



شکل ۱: خوشه‌بندی ژنوتیپ سه نوع پرنده در سه منطقه مختلف Chawia، Ngangao و Mbololo با استفاده از روش بیزی STRUCTURE. الگوریتم موفق شده است خوشه‌بندی را با دقت خوبی انجام دهد اما موارد شماره گذاری شده از ۱ تا ۴ به‌عنوان موارد خطای الگوریتم در تشخیص زیرجمعیت پرنده‌ها در نظر گرفته می‌شود.

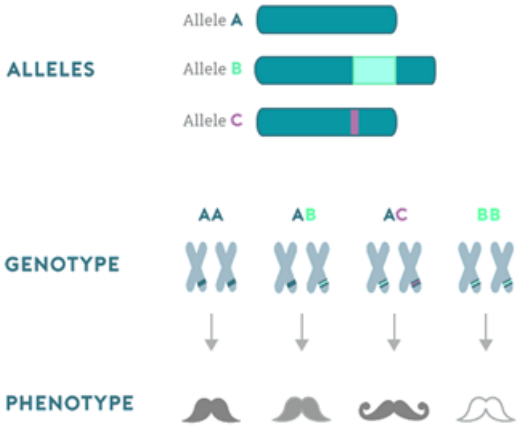


شکل ۲: تصویر DNA که در آن جفت‌بازها که هریک به شکل یکی از چهار حالت ممکن هستند، به‌صورت دوبه‌دو به یکدیگر متصل شده‌اند.

فنوتیپ هستند.

آلل

آلل صورت‌های مختلف از یک ژن است. ژن‌هایی که در یک مکان یا موقعیت ژنتیکی بر روی یک کروموزوم واقع‌شده‌اند، دارای صورت‌های مختلفی هستند. انسان‌ها دیپلوئیدی نامیده می‌شوند زیرا در هر مکان ژنتیکی دارای دو آلل هستند که هر یک از آن‌ها را از یک والد به ارث برده‌اند. هر جفت آلل نشان‌دهنده ژنوتیپ یک ژن خاص است. آلل‌ها بر اساس بسیاری عوامل مثل محیط و وراثت می‌توانند غالب، مغلوب و یا ترکیبی از هردو باشند. هنگامی که موجود یک هتروزیگوت باشد معمولاً یک آلل غالب و یک آلل



شکل ۳: تعامل ژنوتیپ، فنوتیپ و آلل: تغییر در آلل که در واقع تغییر در ژنوتیپ است باعث تغییراتی در فنوتیپ مانند حالت و رنگ سبیل می‌شود.

مغلوب دارد که عضو فنوتیپ آلل غالب را بیان می‌کند. آلل‌ها همچنین می‌توانند تغییرات جزئی توالی DNA را نشان دهند که لزوماً بر فنوتیپ ژن تأثیر نمی‌گذارند. در تصویر ۳ نوع تعامل ژنوتیپ، فنوتیپ و آلل را می‌بینید [۱].

فراوانی آلل

فراوانی آلل [۱۳] با تقسیم تعداد دفعات مشاهدهٔ آلل موردنظر در یک مکان مشخص از ژن بر کل دفعات رخ دادن حالات مختلف آلل محاسبه می‌شود. در هر جمعیت، فراوانی آلل‌ها بیانگر تنوع ژنتیکی آن جمعیت است. تغییرات فراوانی آلل‌ها در طول زمان می‌تواند نشان‌دهنده این باشد که جهش‌های جدید در جمعیت رخ‌داده است.

SNP

SNP تغییرات در یک مکان مشخص در توالی DNA بین افراد است. توالی DNA از زنجیره‌ای از چهار نوع نوکلئوتید A، C، G و T تشکیل‌شده است. اگر بیش از ۱٪ جمعیت یک نوکلئوتید مشابه را در مکانی مشخص در توالی DNA نداشته باشند، این تفاوت می‌تواند به‌عنوان یک SNP طبقه‌بندی شود. SNPها ممکن است منجر به تغییرات توالی اسیدآمینه‌ها شوند چراکه بعد از رونویسی از □□□ و تولید رشته ،□□□ رشته حاصل برای تولید پروتئین به بیرون از سلول رفته و ریبوزوم با متصل شدن به مکان‌های خاصی(کدن آغاز که □□□ است) شروع به تولید پروتئین بر اساس تکه‌های ۳ نوکلئوتیدی می‌کند. حال اگر یک □□□ تغییر کند، ممکن است اسیدآمینه متفاوتی تولید شود و در نتیجه پروتئین متفاوتی با کارکرد متفاوتی تولید شود. SNP ها فقط مربوط به ژن نیستند، بلکه می‌توانند در ناحیه‌های غیر هسته DNA نیز رخ دهند. یک SNP خاص ممکن است باعث هیچ اختلالی نباشد اما برخی SNPها با بیماری‌های خاصی مرتبط هستند. در شکل ۴ SNP مربوط به سه شخص که در مکان چهارم متفاوت هستند و باقی ژنوم آن‌ها یکسان است نشان داده‌شده است .

فراوانی آلل کوچک‌تر

فراوانی آلل کوچک‌تر (MAF) [۱۴] به دومین فراوان‌ترین آلل که در یک جمعیت معین رخ می‌دهد، گفته می‌شود. برای مثال، اگر ۳ آلل با فراوانی‌های ۰٫۵۰ ، ۰٫۴۹ و ۰٫۰۱ وجود داشته باشد، مقدار ۰٫۴۹ به‌عنوان MAF گزارش خواهد شد. دقت شود فراوان‌ترین آلل در واقع رایج‌ترین است و اکثر افراد آن حالت را دارند و اطلاعات چندانی ندارد. اما دومین فراوان‌ترین آلل برای ما مهم است زیرا افراد کمتری آن را دارند و در نتیجه آن افراد به خاطر این آلل‌ها متمایز می‌شوند. MAF به‌طور گسترده در مطالعات ژنتیک جمعیتی استفاده می‌شود زیرا اطلاعات مهمی برای تشخیص دگرگونی‌های رایج و نادر در جمعیت فراهم می‌کند. برای مثال، صفحه refSNP برای گزارش rs222 به شکل زیر است:

$$MAF/MinorAlleleCount : G = ۰٫۲۴۹/۵۴۲$$

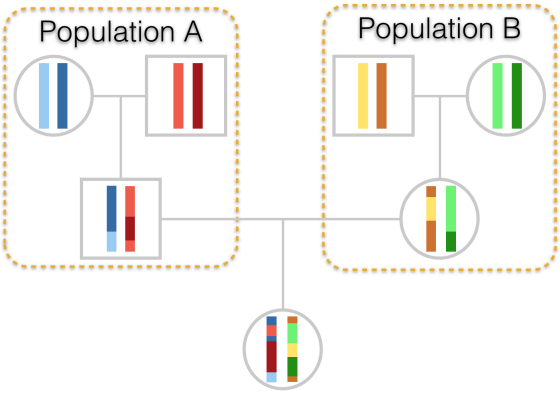
این بدان معنی است که برای rs222، MAF برابر 'G' است و دارای فراوانی ٪۲۴٫۹ در جمعیت است و G به تعداد ۵۴۲ بار در جمعیت نمونه شامل ۱۰۸۸ نفر شامل ۲۱۷۶ کروموزوم مشاهده‌شده است.

جمعیت مخلوطی و غیرمخلوطی

جمعیت غیرمخلوطی به جمعیتی گفته می‌شود که تمام نشانگرهای هر فرد متعلق به یک زیرجمعیت باشد. اگرچه این فرض در واقعیت درست نیست،

[13] Allele Frequency
[14] Minor Allele Frequency

شکل ۴: در این تصویر ژنوتیپ مربوط به سه شخص که در یک SNP متفاوت هستند نشان داده‌شده است که همان SNP باعث بروز تمایز بین افراد شده است.



شکل ۵: در این تصویر ترکیب ژنوم مخلوطی یک فرد که از ترکیب دو زیرجمعیت حاصل‌شده است نمایش داده‌شده است. هر بخش ژنوم از یک زیرجمعیت آمده است.

اما باعث سادگی در مدل‌سازی می‌شود. جمعیتی مخلوطی است که هر فرد متعلق به چندین زیرجمعیت باشد و برخلاف حالت غیرمخلوطی، هر نشانگر هر فرد متعلق به یک زیرجمعیت باشد. در شکل ۵ به ژنوم به‌دست‌آمده از دو زیرجمعیت دقت کنید. این ژنوم هر بخش خود را از یک زیرجمعیت گرفته است و کل ژنوم مخلوطی از هر دو زیرجمعیت است.

۱.۲ اپیستازیس

اصطلاح اپیستازیس [15] را نخستین بار در سال ۱۹۰۳ بتسون [16] برای توصیف پدیده‌ای که یک آلل از یک ژن، اثر ژن دیگری را مخفی می‌کند، معرفی کرد. توجه شود که به این حالت غالب و مغلوب گفته نمی‌شود زیرا حالت غالب و مغلوب برای یک ژن یکسان در مکان یکسان در کروموزوم‌های مختلف است اما اپیستازیس در بین ژن‌ها است و ممکن است هر دو ژن غالب باشند اما اثر فنوتیپی یک ژن توسط ژن دیگری پنهان شود. فرض کنید دو مکان، B و G وجود دارد که بر یک ویژگی مانند رنگ مو در موش تأثیر می‌گذارد. مکان B دارای دو آلل B یا b است و مکان □ دارای دو آلل G یا g است. نتایج حاصل از فنوتیپ‌های ممکن (موهای سفید، سیاه و یا خاکستری) بر اساس ژنوتیپ‌های متفاوت در جدول ۱ نشان داده‌شده است. ملاحظه می‌شود که بدون در نظر گرفتن ژنوتیپ در مکان B، افراد با هر نسخه از آلل G موی خاکستری دارند، آلل G به آلل g غالب است، به‌عبارت‌دیگر G اثر آلل g را پنهان می‌کند. همچنین می‌بینید که اگر ژنوتیپ در مکان G حالت g/g باشد، موشی با هر نسخه‌ای از آلل B دارای مو سیاه است، به‌طوری‌که در مکان B آلل B به b غالب است. البته اگر ژنوتیپ در مکان G حالت g/g نباشد، اثر مکان B قابل‌مشاهده نیست، چراکه افراد با هر نسخه از آلل G موی خاکستری دارند. اثر مکان B توسط مکان G پوشیده شده است و گفته می‌شود مکان G به مکان B، اپیستاتیک است یا به‌طور دقیق‌تر، آلل G در مکان G به آلل B در مکان B، اپیستاتیک است.

[15] Epistasis
[16] Bateson

جدول ۱: پدیده اپیستازیس در دو مکان مؤثر در رنگ موی موش

□/□	□/□	□/□	
خاکستری	خاکستری	سفید	□/□
خاکستری	خاکستری	سیاه	□/□
خاکستری	خاکستری	سیاه	□/□

۳ مفاهیم و روابط آماری

در این بخش توضیح مختصری در مورد توزیع دیریکله، الگوریتم MCMC (زنجیره مارکوف مونت‌کارلو)، نمونه‌گیری گیبز و الگوریتم متروپولیس-هستینگس ارائه می‌شود.

توزیع دیریکله

توزیع دیریکله معمولاً به‌عنوان توزیع پیشین برای فراوانی نسبی آلل‌ها در مدل‌سازی بیزی استفاده می‌شود. این توزیع گسترش‌یافته توزیع بتا در حالت چندمتغیره است. خاصیت مهم توزیع دیریکله این است که پیشین مزدوج [۱۷] برای توزیع چندجمله‌ای است. این خاصیت در محاسبه توزیع تمام‌شرطی [۱۸] فراوانی نسبی آلل‌ها در گام‌های نمونه‌گیری گیبز و درصد تعلق هر فرد به هر زیرجمعیت بدین‌صورت استفاده می‌شود که ضرب یک توزیع چندجمله‌ای و یک توزیع دیریکله، یک توزیع دیریکله با ابرمتغیرهای متفاوت می‌شود و درنتیجه تولید نمونه از توزیع تمام شرطی راحت‌تر می‌شود.

تابع چگالی احتمال توزیع دیریکله به‌صورت زیر است

$$f(x_1,\ldots,x_{K-1};\alpha_1,\ldots,\alpha_K)=\frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)}\prod_{k=1}^{K}x_i^{\alpha_i-1} \qquad (1)$$

که در آن $0<x_1,\ldots,x_{K-1}<1$ و $x_K=1-x_1-\ldots-x_{K-1}$. در حالت خاص که α_iها برابرند آن را توزیع دیریکله متقارن [۱۹] می‌نامند. توزیع دیریکله متقارن در استنتاج ساختار جمعیت برای درصد تعلق هر فرد به هر زیرجمعیت استفاده می‌شود. تابع چگالی احتمال توزیع دیریکله متقارن به این صورت است

$$f(x_1,\ldots,x_{K-1};\alpha)=\frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K}\prod_{i=1}^{K}x_i^{\alpha-1} \qquad (2)$$

زنجیره مارکوف مونت‌کارلو (MCMC)

یک دشواری در رویکردهای بیزی، محاسبه توزیع پسین یا $\pi(x)$ است. اگر صورت دقیق تابع توزیع پسین $\pi(x)$ شناخته‌شده نباشد نمی‌توان توزیع $\pi(x)$ را به‌صورت مستقیم تولید کرد و حتی اگر تابع توزیع احتمال هم مشخص باشد باز ممکن است به دلیل سختی تولید مستقیم توزیع یا به دلیل پیچیدگی تابع توزیع احتمال $\pi(x)$ و ابعاد بزرگ x، تولید توزیع غیرممکن باشد. یکی از راه‌حل‌های ارائه‌شده برای این مشکل استفاده از روش MCMC است. روش MCMC به دلیل انعطاف‌پذیری و عمومیت در حال حاضر به‌طور گسترده‌ای در تمام زمینه‌های آمار استفاده می‌شود. در روش MCMC برای حل این مشکلات از زنجیره مارکوف که راهی برای تولید توزیع به‌صورت غیرمستقیم است، استفاده می‌شود [۱۱]. دو مزیت عملی روش MCMC بدین شرح است:

۱. عملکرد مناسب در همگرا شدن به توزیع پسین.

۲. پیاده‌سازی ساده.

در میان الگوریتم‌های MCMC، دو روش رایج وجود دارد:

۱. روش نمونه‌گیری گیبز که مسائل با ابعاد بزرگ را به روش تولید متوالی نمونه از زیرمجموعه‌های مختلف x ساده می‌کند. به‌عبارت‌دیگر نمونه‌گیری گیبز برای یافتن توزیع احتمال داده‌ها با بعد بالا مناسب است.

۲. روش متروپولیس-هستینگس که دارای یک قاعده برای پذیرش یا رد نمونه برای اصلاح یک زنجیره دلخواه مارکوف است تا تولید نمونه‌های توزیع $\pi(x)$ با دخالت توزیع پیشنهادی [۲۰] که در بسیاری از موارد یک توزیع نرمال با میانگین و واریانس مشخص است، صورت پذیرد.

نمونه‌گیری گیبز

برای توزیع پسین چندمتغیره $\pi(x)=\pi(x_1,\ldots,x_p)$، روش نمونه‌گیری گیبز به‌طور متوالی نمونه‌هایی را برای هر یک از متغیرهای تصادفی X_i از توزیع

[17] Conjugate Prior
[18] Full conditional
[19] Symmetric Dirichlet Distribution
[20] Proposal Distribution

شرطی کامل $(X_i|X_1,\ldots,X_{i-1},X_{i+1},,X_p)$ به دست می‌آورد. این نمونه‌ها، تحت شرایط کلی، به توزیع مانا [۲۱] $\pi(x_1,\ldots,x_p)$ همگرا می‌شوند. بنابراین برای تعداد کافی تکرارها، N تا نمونه، $(X^{(0)},\ldots,X^{(N)})$، می‌تواند به‌عنوان توزیعی واقعی از $\pi(x)$ تلقی شود. مراحل الگوریتم به شرح زیر است.

۱. مقدار اولیه را تنظیم کنید، $X^{(0)}=(X_1^{(0)},\ldots,X_p^{(0)})$

۲. برای تکرار k ام از ۱ تا N، تولید نمونه $X_i^{(k)}$ از $(X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_p)$ برای i از ۱ تا p را انجام دهید.

۳. مقادیر $(X^{(0)},\ldots,X^{(N)})$ را به‌عنوان خروجی تولید کنید.

پس از اجرای مدل باید مقداری از نمونه‌های تولیدشده در مراحل ابتدایی را حذف کرد زیرا آن نمونه‌ها دارای توزیع پسین نیستند و به آن‌ها دوره سوخت [۲۲] گفته می‌شود [۱۲].

الگوریتم متروپولیس-هستینگس

الگوریتم متروپولیس-هستینگس الگوریتمی است که یک توزیع پیشنهادی با توجه به متغیرهای تصادفی $X^{(k)}$ داده می‌شود و سپس مقادیر تولیدشده از طریق توزیع پیشنهادی با توجه به توزیع پسین پذیرش یا رد می‌شود. الگوریتم متروپولیس-هستینگس به‌صورت زیر اجرا می‌شود [۱۲]:

۱. مقدار اولیه را تنظیم کنید، $X^{(0)}=(X_1^{(0)},,X_p^{(0)})$

۲. برای k از ۱ تا N، مراحل زیر را انجام دهید:

(آ) از توزیع پیشنهادی، یک مقدار را به‌شرط $X^{(k)}$ تولید کنید،
$$X^* \sim q(.,X^{(k)})$$

(ب) مقدار بعدی $X^{(k+1)}$ را به‌صورت زیر تولید کنید:

$$X^{(k+1)}=\begin{cases} X^* & \square\square\square\square(0,1)\le\alpha(X^{(k)},X^*) \\ X^{(k)} & \text{جاهای دیگر} \end{cases} \quad (۳)$$

که $\alpha(X^{(k)},X^*)$ از رابطهٔ زیر محاسبه می‌شود

$$\alpha(X^{(k)},X^*)=$$
$$\begin{cases} \min\{\frac{\pi(X^*)q(X^{(k)},X^*)}{\pi(X^k)q(X^*),X^{(k)}},1\} & \text{اگر } \pi(x^{(k)})q(X^*,x^{(k)})>0 \\ 1 & \text{جاهای دیگر} \end{cases}$$
(۴)

۳. مقادیر $(X^{(0)},\ldots,X^{(N)})$ را به‌عنوان خروجی تولید کنید.

تابع $\alpha(X^{(k)},X^*)$ احتمال پذیرش مقدار پیشنهادی است در حالتی که تابع پیشنهاددهنده متقارن باشد. اگر مقدار پیشنهادی رد شود، مقدار فعلی به‌عنوان مقدار بعدی در زنجیره مارکوف در نظر گرفته می‌شود. پس از اجرای مدل باید دوره سوخت را حذف کرد.

۴ کاربردهای استنتاج ساختار جمعیت و GWAS در تشخیص بیماری‌ها

در پی، به کوتاهی، چند کاربرد استنتاج ساختار جمعیت را ذکر می‌کنیم.

۱. بررسی انتخاب طبیعی رخ‌داده در طول حیات یک جمعیت. برای مثال بررسی گونه‌ای از یک حیوان که به علت عوامل ژنتیکی و محیطی منقرض‌شده است و یا تکامل پیداکرده است تا تعامل بهتری با محیط داشته باشد.

۲. بررسی جهش‌ها در نژاد یک جمعیت و عامل‌هایی که باعث این جهش شده‌اند و همچنین بررسی تأثیراتی که جهش‌ها بر تعداد افراد جمعیت گذاشته و یک یا چند ویژگی فیزیکی را در جمعیت تغییر داده است. برای مثال جهشی که باعث شود که گروهی از جامعه قدبلندتری داشته باشند.

۳. تولید دارو برای زیرجمعیت‌ها به‌صورت اختصاصی به‌نحوی‌که تولید دارو صرفهٔ اقتصادی داشته باشد. دقت دارید که هزینه تولید دارو برای هر فرد بسیار زیاد است، درحالی‌که تولید دارو برای هر خوشه جمعیتی بسیار ارزان‌تر است. علاوه بر این، چون افراد درون یک زیرجمعیت دارای شباهت‌های بسیار ژنتیکی هستند بنابراین اگر یک دارو برای برخی از افراد زیرجمعیت عوارض جانبی کمتری داشته باشد، احتمال داشتن عوارض جانبی برای باقی اعضای زیرجمعیت کاهش می‌یابد.

۴. ممکن ساختن مطالعات آماری دربارهٔ یک جمعیت به‌نحوی‌که تنوع ژنتیکی و بیماری‌های رایج را در جمعیت بررسی کرد. برای مثال بررسی تأثیر بیماری کرونا بر نژادهای مختلف جمعیتی در ایران جهت ارائه خدمات بهتر و منصفانه‌تر.

۵. بررسی یک جمعیت باهدف بررسی یک بیماری خاص با هزینه بسیار کمتر. زیرا می‌توان به‌جای اینکه کل افراد را بررسی کرد، از هر زیرجمعیت چند نمونه را انتخاب کرد و بر روی همان افراد مطالعات را انجام داد.

در ادامه به چند مورد از موفقیت‌های GWAS در زمینه تشخیص انواع بیماری‌ها می‌پردازیم.

[21] Stationary Distribution

[22] burn-in

۱.۴ تشخیص دیابت

دیابت نوع یک، بیماری رایجی است که بر اثر عوامل متعدد ژنتیکی و محیطی به وجود می‌آید. بارت و همکاران [۲] تعدادی از مکان‌های مؤثر برای دیابت نوع یک را شناسایی و گزارش یافته‌های مطالعه مرتبط با هموگلوبین دیابت نوع یک را منتشر کرده‌اند. نمونه کل شامل ۷۵۱۴ فرد مبتلا و ۹۰۴۵ فرد کنترل بوده که شواهدی برای ارتباط چهل‌ویک مکان ژنتیکی با دیابت نوع یک ارائه داده است. این پژوهش پس از حذف ارتباطات گزارش‌شده قبلی، ۲۷ مکان را به‌صورت مستقل از ۴۲۶۷ مورد مبتلا، ۴۴۶۳ مورد کنترل بر روی ۲۳۱۹ خانواده بررسی کرده است. از این تعداد ۱۸ مکان تکراری بوده و ۶ مکان دیگر عواملی مؤثر در بیماری بودند. ژن‌های جدید کشف‌شده در این تحقیق شامل IL10، IL19، IL20، GLIS3، CD69 و IL27 می‌شوند.

۲.۴ تشخیص سرطان

روش‌های GWAS بر روی بیماری‌های پیچیده‌تری مانند انواع مختلف سرطان نیز به کار گرفته‌شده است [۷] . در تحقیق [۶] عامل‌های مؤثر در سرطان سینه شناسایی‌شده‌اند. این تحقیق بر روی ۵۲۸۱۳۷ SNP در ۱۱۴۵ زن یائسه اروپایی مبتلا به سرطان سینه و ۱۱۴۲ فرد کنترل انجام‌شده است که چهار SNP را در اینترون [۲۳] شماره دو از FGFR2 شناسایی کرده که به‌شدت با سرطان سینه ارتباط دارند.

در پژوهش توماس و همکاران [۲۱] برخی ژن‌های مؤثر در سرطان پروستات کشف‌شده است. این تحقیق بر روی ۵۲۷۸۶۹ SNP برای ۱۱۷۲ فرد مبتلابه سرطان پروستات و ۱۱۵۷ فرد کنترل در اروپا انجام‌شده است. در تحلیل نتایج، سه مکان که قبلاً گزارش‌شده شامل دو SNP مستقل در 8q24 و HNF1B تأییدشده‌اند و علاوه بر این، مکان‌هایی در کروموزوم‌های ۷، ۱۰ و ۱۱ که بسیار مهم هستند کشف‌شده است. یکی از مکان‌های روی کروموزوم ۱۰، MSMB است که یک نشانگر [۲۴] پیشگیری از سرطان پروستات است و دیگری CTBP2 است که عاملی تأثیرگذار در ابتلا به سرطان پروستات است.

۳.۴ تشخیص پارکینسون

مطالعه‌ای درباره بیماری پارکینسون با استفاده از مجموعه‌ای شامل ۷۸۹۳۲۷۴ نشانگر بر روی ۱۳۷۰۸ مورد مبتلا به پارکینسون و ۹۵۲۸۲ فرد کنترل انجام شد که نشان داد ۲۶ مکان ارتباط معنی‌داری با پارکینسون دارند [۱۴] . این ۲۶ مکان و ۶ مکان به‌دست‌آمدهٔ قبلی، بر روی مجموعه‌ای مستقل از ۵۳۵۳ فرد مبتلا و ۵۵۵۱ فرد کنترل مورد آزمایش قرار گرفتند و از ۳۲ SNP موردبررسی،۶ مورد جدید بودند. تجزیه‌وتحلیل مکان‌ها نشان داد که چهار مکان GBA، GAK-DGKQ، SNCA و HLA مکان‌هایی مؤثر در پارکینسون هستند.

۴.۴ تشخیص دوقطبی

در تحقیقی درباره بیماری دوقطبی ۷۴۸۱ نفر مبتلابه بیماری دوقطبی و ۹۲۵۰ فرد کنترل بررسی شدند [۱۹]. در این بررسی ۳۴ SNP در ۴۴۹۶ نمونه مستقل با بیماری دوقطبی و ۴۲۴۲۲ فرد کنترل مستقل آزمایش شدند و کشف شد که ۱۸ تا از ۳۴ SNP دارای تأثیر زیادی هستند. در این تحقیق مسیری که باعث تقویت احتمال بروز بیماری دوقطبی می‌شوند، شناسایی شد و درنهایت، تجزیه‌وتحلیل ترکیبی GWAS از اسکیزوفرنی و دوقطبی نشان داد که بین SNP های درون CACNA1C و مکان‌های NEK4، ITIH1، ITIH3 و ITIH4 ارتباطی معنادار وجود دارد.

۵.۴ تشخیص اسکیزوفرنی

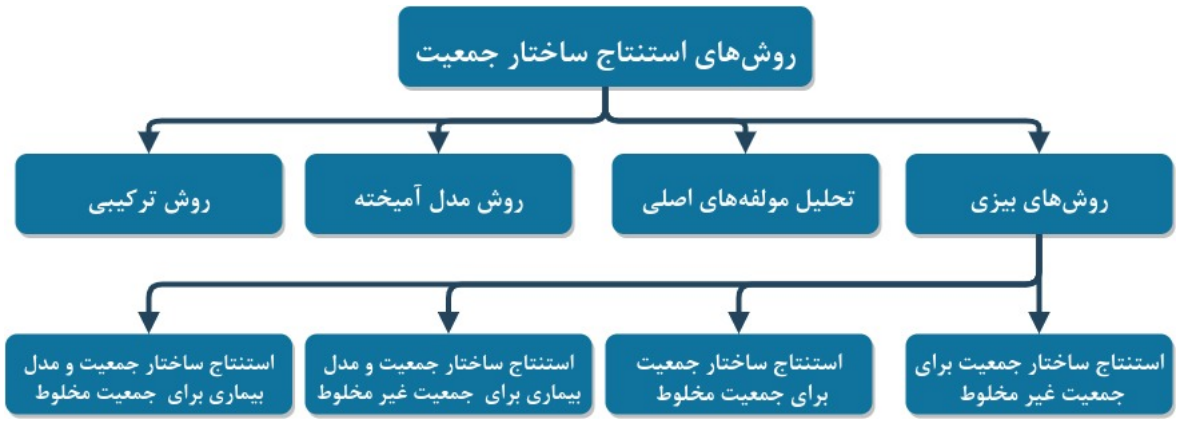
در مرجع [۱۸] گزارشی از یک تحقیق دربارهٔ بیماری اسکیزوفرنی با استفاده از روش‌های GWAS ارائه‌شده است. در این تحقیق نقش گوناگونی ژنتیکی در اسکیزوفرنی در دو مرحله بررسی‌شده است. مرحله اول ۲۱۸۵۶ فرد از نژاد اروپایی و مرحله دوم ۲۹۸۳۹ فرد مستقل مطالعه شدند. ترکیب مرحله ۱ و ۲ منجر به کشف ارتباط معنی‌دار بین اسکیزوفرنی و ۷ مکان شد که پنج مورد 1p21.3، 2q32.3، 8p23.2، 8q21.3 و 10q24.32-q24.33 جدید بودند و دو مورد 6p21.32-p22.1 و 18q21.2 قبلاً کشف‌شده بودند.

از ژن‌های کشف‌شده که در بالا به آن‌ها اشاره کردیم برای درمان بهتر یا پیشگیری از ابتلا به بیماری استفاده می‌شود. برای مثال، با تشخیص ژن‌های مرتبط با اسکیزوفرنی یا سرطان در کودکان، می‌توان با محدود کردن عوامل محیطی و استفاده از دارو، بیان آن ژن را محدود کرده تا بعداً شخص به آن بیماری مبتلا نشود. همچنین از نتایج حاصل از این تحقیقات می‌توان با استفاده از بررسی جنین‌ها قبل از تولد، از تولد نوزادانی که به‌احتمال بالا در آینده دچار بیماری می‌شوند جلوگیری کرد.

۵ روش‌های رایج در تشخیص ساختار جمعیت

جمعیت را می‌توان با روش‌های مختلفی خوشه‌بندی کرد. در شکل ۶ نمودار مربوط به روش‌های مختلف را مشاهده می‌کنید. روش‌های استنتاج ساختار جمعیت با رویکرد بیزی به‌طور کامل در بخش ۶ شرح داده خواهد شد. در

[23] Intron
[24] Marker



شکل ۶: نمودار دسته‌بندی انواع روش‌های استنتاج ساختار جمعیت.

این بخش برخی روش‌های دیگر را مرور می‌کنیم.

۱.۵ تشخیص ساختار جمعیت با روش تحلیل مؤلفه‌های اصلی

پرایس و همکاران [۱۶] روشی بر مبنای تحلیل مؤلفه‌های اصلی ارائه داده‌اند که اصلی‌ترین رقیب روش‌های بیزی است. این روش که EIGENSTRAT نام دارد باهدف رفع محدودیت‌های روش STRUCTURE برای ژنوتیپ‌های بزرگ پیشنهادشده است. این الگوریتم شامل سه مرحله است:

۱. روش تجزیه مؤلفه‌های اصلی بر داده‌های ژنوتیپ به‌منظور استنتاج محورهای پیوستهٔ تنوع ژنتیکی به کار می‌رود. در عمل، محور تنوع ژنتیکی، ابعاد داده‌ها را به تعداد کمی بعد کاهش می‌دهد. محور تنوع ژنتیکی به‌عنوان بردار ویژه ماتریس کوواریانس‌های بین نمونه‌ها تعریف می‌شود. محورهای تنوع ژنتیکی در مجموعه داده‌هایی که تفاوت‌های نژادی بین نمونه‌های آن زیاد است، اغلب تفسیر جغرافیایی دارند. برای مثال محور تنوع مربوط به یک منطقه از شمال غربی تا جنوب شرقی در اروپا دارای مقادیری است که به‌تدریج برای نمونه‌های شمال غربی اروپا مثبت می‌شود و سپس در مرکز اروپا نزدیک به صفر شده و در آخر در جنوب شرقی اروپا منفی می‌شود.

۲. ژنوتیپ‌ها و فنوتیپ‌ها با محورهای تنوع ژنتیکی تنظیم می‌شوند. این کار که یک مجموعه از موردها و کنترل‌ها را ایجاد می‌کند، از طریق محاسبه مجدد رگرسیون‌های خطی انجام می‌شود.

۳. ارتباطات [۲۵] با استفاده از ماتریس ژنوتیپ‌ها و فنوتیپ‌ها تنظیم می‌شود.

ازآنجاکه داده‌ها تنها شامل یک بعد اضافه‌شده نسبت به ساختار جمعیت هستند، تنها محور اصلی تنوع ژنتیکی اهمیت دارد. شبیه‌سازی‌ها نشان می‌دهد که با مدل‌سازی موردها و کنترل‌ها، EIGENSTRAT نرخ مثبت کاذب [۲۶] مساوی و یا پایین‌تری را نسبت به دیگر روش‌ها به دست می‌آورد. گفتنی است که نرخ مثبت کاذب نسبتی از تعداد موارد منفی است که به‌اشتباه مثبت رده‌بندی می‌شوند. [۱۶].

۲.۵ تشخیص ساختار جمعیت با روش مدل آمیخته

در روش‌های مبتنی بر مدل آمیخته [۲۷] می‌توان ساختار جمعیت و ارتباط پنهان [۲۸] را مدل کرد [۳]. رویکرد اصلی این روش‌ها بدین‌صورت است که فنوتیپ‌ها بر اساس ترکیبی از اثرات ثابت و اثرات تصادفی مدل‌سازی می‌شوند. اثرات ثابت شامل SNPها با متغیرهای کمکی اختیاری، مانند جنسیت یا سن است. اثرات تصادفی بر اساس یک ماتریس کوواریانس فنوتیپی است که به‌عنوان مجموع تغییرات تصادفی ارثی و غیر ارثی مدل می‌شوند تا اثر ژنوتیپ و محیط هر دو برای خوشه‌بندی افراد استفاده شود. قابل‌ذکر است مدل‌های آمیخته ازلحاظ نظری دارای دقت مناسبی هستند اما از نظر محاسباتی مقرون‌به‌صرفه نیستند [۳].

[25] Associations
[26] False Positive
[27] Mixture Model
[28] Hidden Association

۳.۵ تشخیص ساختار جمعیت با روش‌های ترکیبی

توکر [29] و همکاران [۲۲]، روش تحلیل مؤلفه‌های اصلی را با مدل‌های آمیخته خطی [30] ترکیب کردند و شیوه‌ای جدید برای تشخیص ساختار جمعیت پیشنهاد دادند. هدف از انجام این کار بهره بردن از قدرت هر دو روش برای پیدا کردن دقیق‌ترین زیرجمعیت است. مراحل انجام کار در این روش به این صورت است:

۱. استخراج مؤلفه‌های اصلی: پنج مؤلفه اصلی از محور ارتباطات ژنتیکی با استفاده از داده‌های ژنوتیپ، استخراج می‌شوند که با عنوان ویژگی‌ها معرفی می‌شوند.

۲. رتبه‌بندی SNP ها با رگرسیون خطی: SNP ها با استفاده از آزمون مبتنی بر رگرسیون خطی رتبه‌بندی می‌شوند.

۳. تعیین ماتریس ارتباطات ژنتیکی: مؤلفه‌های اصلی از زیرمجموعه‌ای از SNP ها که احتمال بیشتری برای مکان مؤثر شدن دارند، انتخاب می‌شوند.

۴. محاسبه ارتباط آماری [31] : با استفاده از تعدادی از SNP های برتر که برای تعیین محور ارتباطات ژنتیکی استفاده‌شده‌اند، ارتباط آماری برای هر SNP محاسبه می‌شود.

برای بررسی دقت روش ترکیبی، داده‌های تصادفی SNP برای ۵۰۰ مورد و ۵۰۰ کنترل برای ۱۰۰۰۰۰ SNP تولیدشده که ۶۰٪ موردها و ۴۰٪ از کنترل‌ها به‌عنوان زیرجمعیت ۱ و باقی‌مانده موردها و کنترل‌ها به‌عنوان زیرجمعیت ۲ طبقه‌بندی‌شده‌اند. شبیه‌سازی‌ها نشان داده است که با مدل‌سازی موردها و کنترل‌ها، این روش نرخ مثبت کاذب [32] مساوی و یا پایین‌تری را نسبت به دیگر روش‌ها به دست آورده است. الگوریتم توکر و همکاران دارای این مزیت است که سریع‌تر از روش‌های بیزی است و برای اجرا بر ژنوتیپ‌های بزرگ مناسب است. اما ساختار جمعیت حاصل از اجرای این روش دارای توجیه بیولوژیکی پایین‌تری نسبت به روش‌های بیزی است زیرا فرضیات این روش منطبق بر واقعیات بیولوژیکی نیستند. در نتیجه موارد به‌دست‌آمده از این روش ممکن است اصلاً ارتباطی با شبکه ژنی نداشته باشند.

۶ رویکردهای بیزی به استنتاج ساختار جمعیت

روش‌های استنتاج ساختار جمعیت که مبتنی بر استنتاج بیزی است تأثیر بسیاری بر GWAS داشته است. برخی از مهم‌ترین رویکردهای بیزی را در ادامه توضیح می‌دهیم.

۱.۶ استنتاج ساختار جمعیت با رویکرد بیزی با فرض غیرمخلوطی بودن جمعیت

در پژوهش پریچارد و همکاران [۱۷] الگوریتمی مدل-مبنا با رویکرد بیزی که با نام STRUCTURE شناخته می‌شود، برای خوشه‌بندی ژنوتیپ ارائه‌شده است تا به کمک آن بتوان ساختار مربوط به جمعیت را استنتاج کرد. ورودی این الگوریتم مجموعه‌ای از آلل‌ها برای مکان‌های مختلف چندین نفر است. هدف ایجاد مدلی مبتنی بر آمار بیزی برای خوشه‌بندی این افراد است.
علت استفاده کردن از خوشه‌بندی مدل-مبنا [33] به‌جای خوشه‌بندی فاصله-مبنا [34]، مشکلات خوشه‌بندی فاصله-مبنا است که در ادامه برخی از اصلی‌ترین آن‌ها را ذکر کرده‌ایم.

۱. خوشه‌بندی فاصله-مبنا معمولاً از یک فرمول محاسبه فاصله مانند فاصله اقلیدسی استفاده می‌کند که به‌شدت به فاصله و نمایش گرافیکی داده‌ها وابسته است که باعث می‌شود مدل ارائه‌شده وابستگی بین پارامترها را به‌درستی مدل نکند.

۲. فهمیدن اینکه خوشه به‌دست‌آمده از روش فاصله-مبنا قابل‌اعتماد است، مشکل است زیرا فرضیات مبتنی بر تأثیر فاصله ژن‌ها بر روی عملکردشان چندان قابل‌اعتماد نیستند.

۳. مرتبط کردن اطلاعات اضافه مانند مکان جغرافیایی افراد به مدل فاصله-مبنا کار دشواری است زیرا ممکن است اطلاعاتی مثل تأثیر آب‌وهوا بر نژاد افراد در این مدل که مبتنی بر فاصله است قابل ارزیابی نباشند.

فرض کنید که ژنوتیپ مربوط به N شخص دیپلوید در L مکان موجود است و K زیرجمعیت وجود دارد. در حالت غیرمخلوط فرض بر این است که کل ژنوم هر شخص متعلق به یکی از K تا زیرجمعیت است که هر زیرجمعیت با

[29] Tucker
[30] Linear Mixed Models
[31] Statistical Association
[32] False Positive
[33] Model-based
[34] Distance-based

استنتاج ساختار جمعیت اهمیت فراوانی در علوم مختلف به‌ویژه در بیولوژی و پزشکی دارد. از استنتاج ساختار جمعیت می‌توان در بررسی انتخاب طبیعی، جهش‌ها، و تولید داروهای اختصاصی متناسب با زیرجمعیت‌های متفاوت استفاده کرد. افزون این‌که استنتاج ساختار جمعیت گامی ضروری در مطالعه بسیاری از بیماری‌ها مانند سرطان، پارکینسون و بیماری‌های روانی است.

فراوانی آلل‌ها برای آن زیرجمعیت شناخته می‌شود. پارامترهای مسئله به شرح زیر است.

$(X_l^{(n,1)}, X_l^{(n,2)})$: ژنوتیپ مربوط به n امین نفر در مکان l ام که $n = 1,2,\ldots,N$ و $l = 1,2,\ldots,L$. توجه شود که ۱ و ۲ بیانگر شماره کروموزوم هستند. برای مثال $X_4^{(3,2)} = 1$ به این معنی است که در مکان چهارم کروموزوم دوم شخص سوم آلل شماره ۱ وجود دارد.

$Z^{(n)}$: نشان‌دهنده زیرجمعیتی است که فرد n ام متعلق به آن است که $n = 1,2,\ldots,N$. برای مثال $Z^{(3)} = 1$ به این معنی است که شخص سوم متعلق به زیرجمعیت اول است.

$P_{k,l,j}$: نشان‌دهنده فراوانی آلل با نمایه j در مکان l در زیرجمعیت k ام است که $k = 1,2,\ldots,K$ و $j = 1,2,\ldots,J_l$ و $l = 1,2,\ldots,L$ و J_l نشان‌دهنده تعداد آلل‌های منحصربه‌فرد مشاهده‌شده در مکان l است. برای مثال، $P_{2,4,1} = 0.3$ به این معنی است که فراوانی آلل شماره ۱ در مکان چهارم زیرجمعیت دوم برابر ۰٫۳ است. به‌عبارت‌دیگر به‌صورت میانگین ۰٫۳ افرادی که در زیرجمعیت دوم هستند در مکان چهارم خود آلل شماره ۱ را دارند.

در حالت غیرمخلوط [35] مدلی با پارامترهای Z و P برای داده‌های X با رویکرد بیزی ارائه‌شده است که توزیع پسین آن به شکل زیر است:

$$\mathbb{P}(Z,P|X) = \mathbb{P}(X|Z,P)\mathbb{P}(Z)\mathbb{P}(P)$$

عملاً بعد از ایجاد مدل و خوشه‌بندی می‌توان هر فرد جدید را به یک زیرجمعیت نسبت داد. به همین دلیل سه عبارت $\mathbb{P}(X|Z,P)$، $\mathbb{P}(Z)$ و $\mathbb{P}(P)$ را به‌طور مختصر شرح می‌دهیم.

۱. گزاره $\mathbb{P}(X|Z,P)$ تابع درستنمایی برای X است که به‌صورت زیر محاسبه می‌شود.

$$\mathbb{P}(X_l^{(n,a)} = j|Z,P) = P_{Z^{(n)},l,j}$$

این تابع درستنمایی بیان می‌کند که احتمال اینکه کروموزوم a ام فرد n ام در مکان l ام برابر با یک آلل خاص باشد برابر است با فراوانی آن آلل در مکان l مربوط به زیرجمعیتی که فرد n ام عضو آن است.

۲. دربارهٔ احتمال $\mathbb{P}(Z)$ فرض بر این است که ما هیچ اطلاعاتی در مورد زیرجمعیت اولیه اشخاص نداریم و برای همین با احتمال $\frac{1}{K}$ هر فرد متعلق به زیرجمعیت k ام است. البته اگر اطلاعاتی در مورد زیرجمعیت افراد داشته باشیم می‌توان با تعریف یک تابع پیشین مناسب اطلاعات اضافه افراد را به مدل اضافه کرد.

۳. در مورد $\mathbb{P}(P)$ ازآنجاکه برای هر زیرجمعیت چندین مکان وجود دارد و برای هر مکان نیز مجموعه‌ای از فراوانی آلل‌ها موجود است، بنابراین برای فراوانی آلل‌ها از توزیع دیریکله که پیشین مزدوج توزیع چندجمله‌ای [36] است استفاده می‌شود. برای فراوانی آلل‌ها در توزیع دیریکله از یک توزیع یکنواخت برای λ_j ها استفاده می‌شود زیرا فرض شده در ابتدا اطلاعاتی در مورد فراوانی آلل‌ها در هر مکان هر زیرجمعیت نداریم.

$$p_{k,l,*} \sim \mathscr{D}(\lambda_1, \lambda_2, ..., \lambda_{J_l})$$

برای اجرای مدل و برآورد توزیع پسین، از روش MCMC و نمونه‌گیری گیبز استفاده می‌شود. برای انجام نمونه‌گیری گیبز ابتدا مقدار $Z^{(0)}$ تعیین‌شده و سپس برای m گام، مراحل زیر انجام می‌شود:

۱. نمونه‌گیری $P^{(m)}$ بر اساس $\mathbb{P}(P|X,Z^{(m-1)})$.

۲. نمونه‌گیری $Z^{(m)}$ بر اساس $\mathbb{P}(Z|X,P^{(m)})$.

برای فهم بهتر، به‌صورت خلاصه چگونگی محاسبه توزیع تمام شرطی $\mathbb{P}(P|X,Z)$ و $\mathbb{P}(Z|X,P)$ را شرح می‌دهیم. برای توزیع $\mathbb{P}(P|X,Z)$ داریم:

$$\begin{aligned}\mathbb{P}(P|X,Z) &= \frac{\mathbb{P}(X|P,Z)\mathbb{P}(P)\mathbb{P}(Z)}{\mathbb{P}(X|Z)\mathbb{P}(Z)} \\ &= \frac{\mathbb{P}(X|P,Z)\mathbb{P}(P)}{\mathbb{P}(X|Z)} \propto \mathbb{P}(X|P,Z)\mathbb{P}(P)\end{aligned} \quad (۵)$$

ازآنجاکه توزیع پیشین P یک توزیع دیریکله است و توزیع تابع درستنمایی چندجمله‌ای است، به خاطر پیشین مزدوج، توزیع $\mathbb{P}(P|X,Z)$ نیز دارای توزیع دیریکله است. بنابراین

$$P_{k,l,*} \propto Dir(\lambda_1 + a_{k,l,1}, \cdots, \lambda_{J_l} + a_{k,l,J_l}) \quad (۶)$$

که

$$a_{k,l,j} = \#\{(n) : X_{l,n} = j \square\square\square Z_n = k\} \quad (۷)$$

[35] Without-Admixture

[36] Multinomial

به‌عبارت‌دیگر $a_{k,l,j}$ نشان‌دهنده تعداد افرادی است که در مکان l ام آن‌ها آلل j است و شخص متعلق به زیرجمعیت k ام است. به‌روزرسانی P برای هر l و k به‌صورت مستقل انجام می‌شود زیرا در فرضیات مسئله مکان‌ها از یکدیگر و زیرجمعیت‌ها از یکدیگر مستقل فرض شده‌اند.

برای توزیع $\mathbb{P}(Z|X,P)$ داریم:

$$\begin{aligned}\mathbb{P}(Z|X,P) &= \frac{\mathbb{P}(X|P,Z)\mathbb{P}(P)\mathbb{P}(Z)}{\mathbb{P}(X|P)\mathbb{P}(P)} \\ &= \frac{\mathbb{P}(X|P,Z)}{\mathbb{P}(X|P)}\end{aligned} \qquad (۸)$$

توجه شود که علت حذف $\mathbb{P}(Z)$ در رابطه بالا این است که توزیع پسین Z مقدار $1/K$ فرض شده و چون یک عدد ثابت است حذف‌شده است. بنابراین

$$\mathbb{P}(Z^{(n)}=k|X,P) = \frac{\mathbb{P}(X^{(n)}|P,Z^{(n)}=k)}{\sum_{k'=1}^{K}\mathbb{P}(X^{(n)}|P,Z^{(n)}=k')} \qquad (۹)$$

که

$$\mathbb{P}(X^{(n)}|P,Z^{(n)}=k) = \prod_{l=1}^{L} p_{k,l,X^{(n,1)}} \times p_{k,l,X^{(n,2)}} \qquad (۱۰)$$

به‌عبارت‌دیگر، احتمال این‌که فرد n ام متعلق به زیرجمعیت k باشد به فراوانی آلل زیرجمعیت k ام و انطباق آن با ژنوتیپ فرد n ام ربط دارد و فرد عضو آن زیرجمعیتی خواهد بود که ژنوتیپش بیشترین انطباق را با فراوانی آلل در آن زیرجمعیت داشته باشد. به‌روزرسانی Z برای هر نفر به‌صورت مستقل انجام می‌شود زیرا نمونه‌گیری از افراد به‌صورت مستقل فرض شده است.

پس از اجرای مدل باید دوره سوخت را حذف کرد زیرا دوره سوخت زمان موردنیاز برای همگرا شدن مدل است و در واقع نمونه‌های تولیدشده در این دوره را نمی‌توان نمونه‌هایی از توزیع پسین در نظر گرفت. پس از اتمام استنتاج، مجموعه‌ای از زیرجمعیت‌ها موجود است که هر فرد متعلق به یکی از این زیرجمعیت‌ها است.

۲.۶ استنتاج ساختار جمعیت با رویکرد بیزی با فرض مخلوطی بودن جمعیت

برای حالت مخلوطی الگوریتم بیزی STRUCTURE فرض می‌شود که هر نشانگر هر فرد متعلق به یک زیرجمعیت است و کل ژنوتیپ هر فرد متعلق به چندین زیرجمعیت است. بنابراین یک متغیر Q نیاز است تا مشخص کند که هر فرد به چه میزان متعلق به کدام زیرجمعیت است. بدین منظور متغیر Q برای هر شخص دارای توزیع دیریکله با پارامتر α است. α نشان‌دهنده میزان پراکندگی ژنوتیپ در میان زیرجمعیت‌ها است که در اینجا برابر فرض شده‌اند زیرا فرض شده است اطلاعات اولیه‌ای در مورد میزان تعلق هر فرد به هر زیرجمعیت وجود ندارد.

$$q_n \sim \mathscr{D}(\alpha,\alpha,...,\alpha)$$

در این حالت تعریف پارامتر Z نسبت به حالت قبل به علت مخلوطی بودن جمعیت تغییر می‌کند زیرا هر شخص متعلق به چندین زیرجمعیت است و هر آلل هر شخص متعلق به یک زیرجمعیت است. بنابراین بجای اینکه Z نشان‌دهنده عضویت هر شخص در هر زیرجمعیت باشد، نشان می‌دهد که مکان l ام از کروموزوم a ام فرد n ام عضو کدام زیرجمعیت است و توزیع احتمال آن به شکل زیر است.

$$\mathbb{P}(z_l^{(n,a)}=k|Q,P) = q_k^n$$

توزیع پسین در این حالت به شکل زیر است.

$$\mathbb{P}(Z,P,Q|X) = \mathbb{P}(X|Z,P,Q)\mathbb{P}(Z|P,Q)\mathbb{P}(P)\mathbb{P}(Q)$$

برای اجرای این مدل از روش MCMC، نمونه‌گیری گیبز و الگوریتم متروپولیس-هستینگس استفاده می‌شود. ابتدا مقدار $Z^{(0)}$ را به‌طور تصادفی مشخص می‌کنیم بدین معنی که هر مکان هر شخص را به یک زیرجمعیت نسبت می‌دهیم. سپس برای m بار گام‌های زیر را انجام می‌دهیم.

۱. نمونه‌گیری $P^{(m)}$ بر اساس $\mathbb{P}(P|X,Z^{(m-1)})$

۲. نمونه‌گیری $Q^{(m)}$ بر اساس $\mathbb{P}(Q|X,Z^{(m-1)})$

۳. نمونه‌گیری $Z^{(m)}$ بر اساس $\mathbb{P}(Z|X,P^{(m)},Q^{(m)})$

۴. به‌روزرسانی α با استفاده از الگوریتم متروپولیس-هستینگس

پس از اتمام استنتاج نمونه‌های تولیدشده بر اساس سه پارامتر Z، Q و P نشان‌دهنده توزیع پسین هستند و همچنین مجموعه‌ای از زیرجمعیت‌ها موجود است که هر فرد متعلق به چندین زیرجمعیت است و درصد عضویت هر فرد به هر زیرجمعیت بر اساس پارامتر Q مشخص است.

۳.۶ استنتاج ساختار جمعیت و کشف ارتباطات هر زیرجمعیت با یک بیماری خاص برای جمعیت غیر مخلوط با رویکرد بیزی

در پژوهش نجفی و همکاران [۱۳] مدلی بیزی برای ترسیم نقشهٔ ارتباطات یک ساختار جمعیت و کشف ارتباطات هر زیرجمعیت با یک بیماری خاص برای جمعیت غیرمخلوط ارائه‌شده است. این مدل آماری بر داده‌های X و Y که X نشان‌دهنده ژنوتیپ و Y نشان‌دهنده فنوتیپ است، اعمال‌شده است. در این مدل سه پارامتر P (بیانگر فراوانی آلل‌ها)، Z (بیانگر عضویت هر شخص در یک زیرجمعیت) و M (بیانگر مدل بیماری)، وجود دارد.

در این مسئله از رویکرد بیزی استفاده‌شده که توزیع پسین آن به شکل زیر است:

$$\mathbb{P}(P,Z,M|X,Y) = \mathbb{P}(X,Y|P,Z,M)\mathbb{P}(P)\mathbb{P}(Z)\mathbb{P}(M)$$

در این مدل سه پارامتر P، Z و M از هم مستقل فرض می‌شوند. این فرض بر اساس آزمایش‌های زیستی نیز درست است و شواهدی بر وجود ارتباط بین P، Z و M وجود ندارد. در ابتدا فرض می‌شود همهٔ K تا زیرجمعیت دارای تعداد افراد برابرند، پس هر شخص با احتمال $\frac{1}{K}$ متعلق به زیرجمعیت kام است. همچنین فرض می‌شود که هر شخص فقط یک نوع بیماری ژنتیکی دارند و بیماری موردنظر تحت تأثیر چندین ژن است و همچنین چندین حالت مختلف آلل ممکن است به یک بیماری منجر شوند بنابراین زیرجمعیت‌های مختلف با عامل‌های مؤثر مختلف بیماری مرتبط هستند. مدل بیماری را به K تا زیرمدل تجزیه کرده که هر زیرمدل با یک زیرجمعیت مستقل در ارتباط است. بیشینه‌سازی توزیع پسین (MAP) [۳۷] با روش نمونه‌گیری گیبز در سه مرحله انجام می‌شود:

۱. نمونه‌گیری $P^{(m+1)}$ از طریق $\mathbb{P}(P|D,M^{(m)},Z^{(m)})$

۲. نمونه‌گیری $Z^{(m+1)}$ از طریق $\mathbb{P}(Z|D,M^{(m)},P^{(m+1)})$

۳. یافتن $M^{(m+1)}$ با بیشینه‌سازی $\mathbb{P}(M|D,Z^{(m+1)},P^{(m+1)})$

در پایان مجموعه‌ای از زیرجمعیت‌ها داریم که هر فرد فقط می‌تواند متعلق به یکی از این زیرجمعیت‌ها باشد و هر زیرجمعیت با توجه به افراد متعلق به آن زیرجمعیت، مدل بیماری را برای آن افراد استنتاج کرده است. پژوهش [۱۳] برتری قابل‌توجهی در دقت استنتاج ساختار جمعیت غیر مخلوط نسبت به STRUCTURE دارد. برای مثال برای ۱۰۰ مکان دقت استنتاج ساختار جمعیت این الگوریتم ۷۷٪ است درحالی‌که دقت استنتاج ساختار جمعیت STRUCTURE برای ۱۰۰ مکان ۶۵٪ است. دلیل این برتری استفاده از فنوتیپ در پژوهش نجفی و همکاران است. به‌عبارت‌دیگر، فنوتیپ اطلاعات بیشتری در مورد جمعیت ارائه می‌دهد که باعث می‌شود پژوهش نجفی و همکاران با تعداد مکان‌های کمتر نیز قادر باشد ساختار جمعیت را با دقت مناسبی استنتاج کند. مزیت دوم این پژوهش استنتاج عوامل بیماری برای هریک از زیرجمعیت‌ها با استفاده از جستجوی جامع است که این کار در STRUCTURE پوشش داده نشده است.

۴.۶ استنتاج ساختار جمعیت و کشف ارتباطات هر زیرجمعیت با یک بیماری خاص برای جمعیت مخلوط با رویکرد بیزی

در بررسی انجام‌شده توسط تمیجی و همکاران [۲۰]، درباره استنتاج ساختار جمعیت مخلوطی و تشخیص مدل بیماری برای هر زیرجمعیت سعی شده است تا ایرادات پژوهش نجفی و همکاران [۱۳] برطرف شود. دو تا از اصلی‌ترین ایرادات به شرح زیر است.

۱. در پژوهش نجفی و همکاران، نوع جمعیت غیرمخلوطی فرض شده است که فرض غیرمعقولی است. زیرا به علت عوامل طبیعی مثل جهش و انتخاب طبیعی و همچنین زادوولد بین نژادهای مختلف از مکان‌های مختلف، در طبیعت انسانی که صددرصد ژنوتیپ آن متعلق به یک زیرجمعیت باشد وجود ندارد و به‌عبارت‌دیگر همه‌ی انسان‌ها دارای ژنوتیپ مخلوطی هستند. در مدل بیزی ارائه‌شده توسط تمیجی و همکاران نوع ژنوتیپ افراد کاملاً مخلوطی فرض شده است.

۲. ضعف دیگر در پژوهش نجفی و همکاران، استفاده از جستجوی جامع برای پیدا کردن عوامل بیماری در یک زیرجمعیت است. استفاده از جستجوی جامع برای مدل کردن اپیستازیس بسیار زمان‌بر است و برای ژنوتیپ‌های بزرگ این کار غیرممکن می‌شود زیرا الگوریتم باید تمام ترکیبات ممکن برای کل مکان‌های موجود را بررسی کند. این مشکل در پژوهش تمیجی و همکاران با استفاده از روشی بیزی حل‌شده است.

در الگوریتم تمیجی و همکاران بر پایه آمار بیزی، وابستگی بین متغیرها به نحوی تعریف‌شده است که بتوان مدل بیماری برای افراد با ژنوتیپ مخلوطی را برآورد کرد. در این الگوریتم، که متعلق به رده مدل‌های گرافیکی و شبکه‌های بیزی است، بر اساس وابستگی پارامترها و با استفاده از الگوریتم MCMC، نمونه‌گیری گیبز و الگوریتم متروپولیس-هستینگس مدلی ارائه‌شده است تا تابع پسین الگوریتم حداکثر شود. از جمله مواردی که این تحقیق پوشش داده است عبارت‌اند از:

۱. تشخیص مدل بیماری برای هر زیرجمعیت که با کمک آن بتوان مکان‌های مؤثر در بیماری را برای افراد درون هر زیرجمعیت تشخیص داد. به این منظور از روشی مبتنی بر روش STRUCTURE استفاده‌شده که در آن با استفاده از شبیه‌سازی زنجیره مارکوف مونت‌کارلو مدل بیماری استنتاج می‌شود.

۲. خوشه‌بندی افراد با استفاده از فنوتیپ با فرض مخلوطی بودن افراد به‌گونه‌ای که پدیده اپیستازیس [۴] در نظر گرفته شود.

[37] Maximum A Posteriori Probability

معمولاً در داده‌های ژنوتیپ به خاطر مشکلاتی مثل خطاهای انسانی یا عدم دقت دستگاه‌های اندازه‌گیری و همچنین گران بودن آزمایش‌های ژنتیکی، عدم قطعیت بسیار زیادی وجود دارد. بنابراین باید عدم قطعیت‌ها با یک مدل احتمالاتی معقول مدل شوند. یکی از بهترین ابزارها برای مدل‌سازی چنین عدم قطعیتی مدل‌های گرافیکی بیزی هستند.

توجه شود که این الگوریتم مانند سایر الگوریتم‌های ذکرشده پیش‌بینی بیماری برای یک شخص جدید انجام نمی‌دهد و فقط افرادی که موجود هستند را خوشه‌بندی می‌کند. برای مثال فرض کنید ژنوتیپ جمعیتی مخلوطی با فنوتیپ بیماری مربوط به هر نفر در این جمعیت موجود باشد. این الگوریتم می‌تواند ساختار جمعیت و مدل بیماری هر زیرجمعیت را برای افراد موجود استنتاج کند و نمی‌تواند برای فرد جدید که ژنوتیپ و فنوتیپ آن در گام‌های استنتاج نبوده است اظهارنظر خاصی کند.

فرض کنید که ژنوتیپ N شخص دیپلوییدِ ($a = ۲$) در L مکان از جمعیتی مخلوط گرفته‌شده که برخی از این افراد فنوتیپ یک بیماری خاص را بروز داده‌اند. هدف تخصیص هر مکان هر شخص به K زیرجمعیت است به‌گونه‌ای که جمعیت به‌صورت مخلوطی خوشه‌بندی شود و مدل بیماری برای هر خوشه به دست بیاید. هر زیرجمعیت با فراوانی نسبی آلل‌ها برای آن زیرجمعیت تعیین می‌شود.

ژنوتیپ داده‌های مربوط به اشخاص با X و فنوتیپ متناظر با هر شخص با Y نشان داده‌شده است. ژنوتیپ افراد به‌صورت $(X_l^{(n,۱)}, X_l^{(n,۲)})$ است که $X_l^{(n,a)}$ بیانگر ژنوتیپ مربوط به n امین نفر در مکان l ام برای کروموزوم a ام است که $n = ۱,۲,\ldots,N$ ، $l = ۱,۲,\ldots,L$ و برای اشخاص دیپلوئید، $a = ۱,۲$ است و همچنین $X \in \{۱,۲,\ldots,J\}^{(N\times L)}$. برای مثال، $X_l^{(n,a)} = A$ بدین معنی است که آلل مکان ام بر روی کروموزوم ام شخص ام، است.

فنوتیپ افراد نیز با Y^n نشان داده می‌شود که نشان می‌دهد فرد nام فنوتیپ را بروز داده است یا خیر. توجه شود که $n = ۱,۲,\ldots,N$ و $Y \in \{۰,۱\}^N$ نشان‌دهنده وجود فنوتیپ در شخص است. برای مثال $Y_n = ۱$ بدین معنی است که شخص ام فنوتیپ متناظر با بیماری را بروز داده است. بدین ترتیب، هدف از برچسب بیماری این است که وجود یک فنوتیپ را در هر فرد نشان دهد.

وابستگی پارامترهای الگوریتم پیشنهادی را در شکل ۷ مشاهده می‌کنید. در این تحقیق برای هر خوشه k، مکان‌های ژنتیکی متناظر با اعضای آن خوشه برای فنوتیپ مربوطه محاسبه می‌شود و به‌عنوان عامل مؤثر در بروز آن فنوتیپ بیان می‌شود. در این پژوهش فرض شده است که اندازه K که نشان‌دهنده تعداد زیرجمعیت‌ها است، مشخص است. در ادامه پارامترهای مسئله معرفی می‌شوند.

پارامتر Q

ازآنجاکه جمعیت مخلوطی است، برای اینکه بتوان درصد تعلق هر فرد به هر زیرجمعیت را مدل کرد پارامتر Q برای هر فرد در نظر گرفته می‌شود. برای هر فرد، پارامتر Q نشان‌دهنده این است که این فرد به چه نسبتی به کدام زیرجمعیت تعلق دارد. برای مثال $q^n = (۰/۷, ۰/۳)$ بدین معنی است که نسبت عضویت فرد nام در زیرجمعیت اول برابر ۰/۷ و برای زیرجمعیت دوم برابر با ۰/۳ است.

پارامتر Z

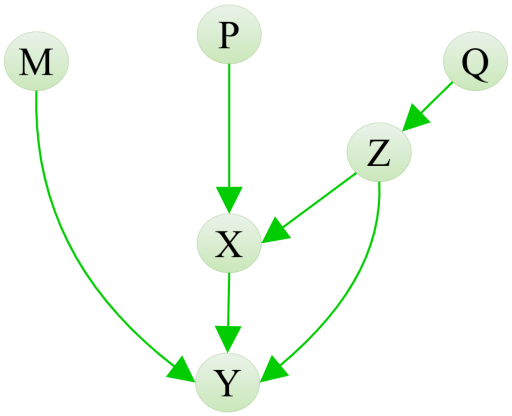
برای این‌که بتوان تخصیص هر فرد به چند زیرجمعیت را مدل کرد نیاز است که پارامتری تعریف شود که بیانگر عضویت هر آلل هر فرد در هر زیرجمعیت باشد که این پارامتر را با Z نشان می‌دهیم. پارامتر Z نشان می‌دهد که آلل هر فرد در مکان lام بر روی کروموزوم aام عضو کدام زیرجمعیت است. بنابراین $z_l^{(n,a)} = ۱$ نشان‌دهنده این است که آلل شخص n ام در مکان l ام بر روی کروموزوم a ام عضو زیرجمعیت یک است.

پارامتر P

هر زیرجمعیت با مجموعه‌ای از فراوانی نسبی آلل‌ها در هر مکان مشخص می‌شود یعنی برای هر مکان در هر زیرجمعیت، فراوانی نسبی آلل‌های موجود برای آن مکان، با توجه به افراد درون آن زیرجمعیت تعیین می‌شود. برای مدل کردن فراوانی نسبی آلل‌ها از پارامتر P استفاده می‌شود که نشان‌دهنده MAF در تمام زیرجمعیت‌ها است به‌گونه‌ای که $P \in \mathbb{R}^{J\times L\times K}$. پارامتر $P_{k,l,j}$ فراوانی نسبی آلل jام در مکان l در زیرجمعیت k ام است که $k = ۱,۲,\ldots,K$ و $j = ۱,۲,\ldots,J_l$ و $l = ۱,۲,\ldots,L$ است و J_l بیانگر تعداد آلل‌های منحصربه‌فرد مشاهده‌شده در مکان l برای کل جمعیت است. برای مثال $p_{k,l,j} = ۰/۳$ یعنی فراوانی افراد عضو زیرجمعیت ام که در مکان ام خود آلل را دارند ۰/۳ است.

پارامتر M

همان‌طور که قبلاً ذکر شد برای هر زیرجمعیت مجموعه مکان‌های مؤثر ژنتیکی منحصربه‌فردی در نظر گرفته می‌شود که با M_k نشان داده می‌شود. M_k نشان‌دهنده مدل بیماری در زیرجمعیت k ام است و مدل‌کننده ارتباط بین برچسب‌های بیماری و پارامترهای مسئله است. برای مثال M_k شامل چند مکان



شکل ۷: وابستگی پارامترهای الگوریتم تمیجی و همکاران

اگرچه روش‌های مدل گرافیکی بیزی از لحاظ زمانی طولانی‌تر هستند، اما به دلیل قدرت این روش‌ها در مدل کردن وابستگی‌های بیولوژیکی و عدم قطعیت، این روش‌ها دارای دقت تشخیص و عملکرد بهتری نسبت به دیگر روش‌ها در تشخیص زیرجمعیت‌ها و شناسایی ساختار نهان جمعیت مخلوطی هستند.

و آلل متناظر با آن مکان‌ها برای زیرجمعیت □ است.

$$M = \{M_1, M_2, \ldots, M_K\} \qquad (11)$$

توزیع پیشین و درستنمایی پارامترها

در مسئلهٔ در دست بررسی، در پی مدلی بر پایه Q، Z، P و M برای ژنوتیوپ X و فنوتیپ Y هستیم. در رابطهٔ زیر توزیع پسین مدل را مشاهده می‌کنید.

$$\begin{aligned} &\mathbb{P}(Z,P,Q,M|X,Y) \propto \mathbb{P}(Y|X,M,Z) \\ &\mathbb{P}(X|Z,P)\mathbb{P}(Z|P,Q)\mathbb{P}(M)\mathbb{P}(P)\mathbb{P}(Q) \end{aligned} \qquad (12)$$

در این الگوریتم فرض می‌شود که فراوانی نسبی آلل‌های موجود در ژنوم هر فرد، از مدل داده‌ها مستقل است. همچنین فرض می‌شود که تعادل هاردی-وینبرگ برقرار است به این معنی که به‌شرط نبود عامل خارجی، فراوانی نسبی آلل‌ها در نسل بعد ثابت باقی می‌ماند. درنتیجه پارامترهای P و Z از هم مستقل فرض می‌شوند. بیشینه‌سازی توزیع پسین $\mathbb{P}(P,Z,M,Q|Y,X)$ با استفاده از نمونه‌گیری گیبز به شکل زیر انجام می‌شود:

۱. نمونه‌گیری $P^{(m+1)}$ از طریق $\mathbb{P}(P|X,Y,M^{(m)},Z^{(m)},Q^{(m)})$

۲. نمونه‌گیری پارامتر $Q^{(m+1)}$ از طریق توزیع تمام‌شرطی $\mathbb{P}(Q|X,Y,M^{(m)},Z^{(m)},P^{(m+1)})$

۳. نمونه‌گیری پارامتر $Z^{(m+1)}$ از طریق توزیع تمام‌شرطی $\mathbb{P}(Z|X,Y,M^{(m)},Q^{(m+1)},P^{(m+1)})$

۴. پیدا کردن $M^{(m+1)}$ با بیشینه‌سازی توزیع تمام‌شرطی $\mathbb{P}(M|X,Y,Z^{(m+1)},Q^{(m+1)},P^{(m+1)})$

۵. به‌روزرسانی α با استفاده از الگوریتم متروپولیس-هستینگس

نتیجهٔ حاصل از اجرای الگوریتم بالا، استنتاج ساختار جمعیتی مخلوطی است که برخی اشخاص دارای بیماری بوده و مدل بیماری با توجه به آلل‌های آن افراد برای هر زیرجمعیت متفاوت است. به سخن دیگر، هر فرد می‌تواند متعلق به چندین زیرجمعیت باشد و هر زیرجمعیت با توجه به آلل‌های متعلق به آن زیرجمعیت، مدل بیماری مخصوص خود را دارد. مزیت روش تمیجی و همکاران این است که استفاده نکردن از جستجوی جامع باعث سرعت بسیار بیشتر الگوریتم شده است. همچنین این روش با فرض اینکه کل ژنوتیپ افراد مخلوطی است، مشکل فرض غلط جمعیت غیرمخلوطی را حل کرده است. نتایج حاصل نشان داده است این الگوریتم با سرعتی چندین برابر روش نجفی و همکاران مدل بیماری را برای جمعیتی مخلوطی با دقتی نزدیک به روش جستجوی جامع پیدا می‌کند.

۷ بحث و نتیجه‌گیری

در این مقاله سعی شد استنتاج ساختار جمعیت مخلوط و پیدا کردن مدل بیماری متناسب با هر زیرجمعیت با تأکید بر رویکرد بیزی شرح داده شود.

دلایل انتخاب روش بیزی برای حل مسئلهٔ استنتاج ساختار جمعیت شامل موارد زیر است.

۱. برای مدل کردن پدیده اپیستازیس باید این پدیده را به‌صورت احتمالاتی در نظر گرفت، زیرا معمولاً درون پایگاه داده‌های ژنوتیپ، اطلاعات دقیقی در مورد اینکه کدام مکان‌ها بر دیگر مکان‌ها غالب هستند موجود نیست، و باید از مفروضات و مدل‌های احتمالاتی مناسب استفاده کرد. در این زمینه مدل‌های گرافیکی بیزی کارگشا هستند [۱۳].

۲. روش تحلیل مؤلفه‌های اصلی علی‌رغم بهبود نسبی در نتایج، لزوماً نشان‌دهندهٔ ساختارهای واقعی ژنتیکی نیست، زیرا نتایج آن دارای قابلیت اطمینان بیولوژیکی مناسب نیست و فرضیات آن تعبیر واقعی بیولوژیکی ندارند. در این روش، برخلاف روش‌های بیزی، ممکن است به علت از دست دادن بخش زیادی از اطلاعات مثل اثر پنهان اما مهم یک ژن، دقت خوشه‌بندی کاهش یابد [۱۳].

۳. روش‌های آمیخته در مقایسه با روش‌های بیزی و تحلیل مؤلفه‌های اصلی قدرت کمتری دارند. بااینکه مدل‌های آمیخته ازلحاظ نظری معقول به نظر می‌رسند اما ازنظر محاسباتی برای استنتاج ساختار جمعیت مقرون‌به‌صرفه نیستند [۱۵].

۴. از روش‌های فاصله-مبنا نمی‌توان برای استنتاج ساختار جمعیت استفاده کرد، زیرا خوشه‌های به‌دست‌آمده از این روش‌ها به‌شدت به فاصله وابسته هستند و فهمیدن معنی‌دار بودن یک خوشه کار سختی است. همچنین اضافه کردن اطلاعات اضافی مانند مکان جغرافیایی افراد به خوشه‌ها کار سخت و گاهی غیرممکن است [۱۷].

۵. روش‌های بیزی برای جمعیت‌های مخلوط و غیر مخلوط نتایج بسیار خوبی ارائه می‌دهند و یکی از کارآمدترین روش‌ها هستند. نرخ مثبت کاذب در روش بیزی STRUCTURE کمتر از دیگر روش‌ها است. اگرچه این روش از لحاظ زمانی طولانی‌تر است، اما از لحاظ دقت تشخیص زیرجمعیت‌ها، عملکرد بهتری دارد. همچنین زیرجمعیت‌های حاصل از این روش دارای توجیه بیولوژیکی بوده و برای شناسایی ساختار نهان جمعیت مخلوطی مناسب است [۱۳].

۶. یکی از مزیت‌های روش‌های بیزی که آن را برای کارهای زیستی مناسب می‌سازد این است که به کمک آن به‌راحتی می‌توان عدم قطعیت را مدل کرد. این مزیت مدل‌های بیزی باعث می‌شود این مدل‌ها برای کار در زمینهٔ ژنوتیپ که در آن انواع عدم قطعیت مانند جهش، اپیستازیس و انتخاب طبیعی وجود دارد مناسب باشد [۱۷].

پیشنهادهایی برای آینده

پیشنهادهای زیر برای کارهای آینده مطرح می‌شود:

۱. یک مشکل رایج در استنتاج ساختار جمعیت، مشخص نبودن تعداد زیرجمعیت‌ها است و بنابراین باید الگوریتم‌ها را برای تعداد زیرجمعیت‌های مختلف اجرا کرد. بهتر است قبل از اجرای استنتاج ساختار جمعیت، تعداد زیرجمعیت‌ها با روشی مثل فرآیند رستوران چینی [۳۸] برآورد شود و سپس الگوریتم را با استفاده از تعداد زیرجمعیت مشخص اجرا کرد. بدین ترتیب الگوریتم به ازای تعداد زیرجمعیت‌های مختلف اجرا نمی‌شود و جواب بهینه سریع‌تر حاصل می‌شود.

۲. با تشخیص مکان‌هایی که احتمال بروز بیماری در آن‌ها صفر است می‌توان الگوریتم‌های بیزی را هوشمندتر کرد. زیرا الگوریتم‌های بیزی با بررسی نکردن آن مکان‌ها، به بررسی بهتر مکان‌های محتمل‌تر برای بیماری می‌پردازند. این کار باعث سریع‌تر شدن الگوریتم می‌شود. توجه شود این کار باید به نحوی انجام شود که در پیدا کردن مکان‌های فعال در بیماری خللی وارد نشود. برای این کار می‌توان از نتایج کارهای GWAS استفاده کرد.

۳. بررسی داده‌ها برای برآورد مقادیر پیش‌فرض نیز راه مناسبی برای استفادهٔ بهینه‌تر از الگوریتم‌های بیزی است. زیرا این الگوریتم‌ها همواره دارای پارامترهایی در توزیع پیشین خود هستند که با برآورد اولیه آن‌ها ساختار جمعیت سریع‌تر به دست می‌آید. برای مثال، می‌توان از بررسی جمعیت برای تخمین اولیه پارامترهای توزیع دیریکله استفاده‌شده در فراوانی آلل‌ها استفاده کرد تا سرعت همگرایی بالا برود.

[38] Chinese restaurant process

مراجع

[1] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2015).*Molecular biology of the cell*. 6th edition, New York: Garland Science. **6**.

[2] Barrett, J. C. and et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet*, **41(6)**, 703–707.

[3] Bouaziz, M., Ambroise, C., and Guedj, M. (2011). Accounting for population stratification in practice: A comparison of the main strategies dedicated to genome-wide association studies. *PLoS One*, **6(12)**, 98–113.

[4] Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet*, **47(3)**, 291–295.

[5] Consortium, I. H. (2005). A haplotype map of the human genome. *Nature*, **437**, 1299–320.

[6] Easton, D. F., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447(7148)**, 1087–93.

[7] Easton, D. F. and Eeles, R. A. (2008). Genome-wide association studies in cancer. *Hum. Mol. Genet*, **17(R2)** .

[8] Hara, K. and et al. (2014). Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum Mol Genet*, **23(1)**, 239–246.

[9] Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6**, 95–108.

[10] International Human Genome Sequencing Consortium, Human, I., and Sequencing, G. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409(6822)**, 860–921.

[11] Jing, L. (2010). Hastings-within-Gibbs Algorithm : Introduction and Application on Hierarchical Model. *Stoch. Process. their Appl*, **2**, 1–13.

[12] Lynch, S. (2007). Introduction to Applied Bayesian Statistics & Estimation for Social Scientists. *Springer*.

[13] Najafi, A., Janghorbani, S., Motahari, S. A., and Fatemizadeh, E. (2019). Statistical Association Mapping of Population-Structured Genetic Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **16(2)**, 638-649.

[14] Nalls, M. A. and et al. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet*, **46(9)**, 989–993.

[15] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev*, **11(1)**, 459–463.

[16] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*, **38(8)**, 904–909.

[17] Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *The American Journal of Human Genetics*, **67(1)**, 170–81.

[18] Ripke, S. and et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet*, **43(10)**, 969–976.

[19] Sklar, P. and et al. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet*, **43(10)**, 977–983.

[20] Tamiji, M., Taheri, S., and Motahari, S. (2019). Stratification of Admixture Population: A Bayesian Approach. *2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), Bojnord, Iran*, 1-4.

[21] Thomas, G. and et al. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet*, **40(3)**, 310–315.

[22] Tucker, G., Price, A. L., and Berger, B. (2014). Improving the power of gwas and avoiding confounding from population stratification with pc-select. *Genetics*, **197(3)**, 1045–1049.