دانشگاه تهران

مرکز تحقیقات بیوشیمی و بیوفیزیک

مسائل محاسباتی در شناسائی هاپلوتیپ‌ها

نگارش

سید علی کتان‌فروش

اساتید راهنما: **دکتر مهدی صادقی و دکتر حمید پزشک**

استاد مشاور: **دکتر چنگیز اصلاحچی**

رساله برای دریافت درجهٔ

دکترا در رشته‌ی بیوانفورماتیک

مهر ۱۳۸۸

مسائل محاسباتی در شناسائی هاپلوتیپ‌ها

چکیده

توالی نوکلئوتیدها در ژنوم انسان بجز برخی از جایگاه‌های معین بر روی آن، در بین همه افراد یکسان است. با این حال همین تفاوتهای اندک، عامل اصلی بروز صفات متفاوت در بین افراد جمعیت است. رایج‌ترین نوع این تفاوت‌ها، چندریختی تک نوکلئوتیدی یا اسنیپ (Single Nucleotide Polymorphism, SNP) است. در هر جایگاه اسنیپ، بیش از یک نوع نوکلئوتید بر روی ژنوم افراد مختلف جمعیت مشاهده می‌شود. یک هاپلوتیپ، نشاندهنده‌ی توالی نوکلئوتیدهای ژنوم یک فرد بر روی تعدادی از اسنیپ‌ها است. ژنوم هر فرد ترکیبی از دو هاپلوتیپ به ارث رسیده از والدین است که این ترکیب را اصطلاحا ژنوتیپ فرد می‌نامند. به نظر می‌رسد تنوع هاپلوتیپ‌ها طی نسل‌های متوالی، در نواحی معینی از ژنوم بدون تغییر باقی می‌ماند. این نواحی، ژنوم را به مجموعه‌ای از بلوک‌های هاپلوتیپی افراز می‌کنند.

در رساله‌ی پیش رو، پس از مرور روشهای توالی‌یابی ژنوم و تعیین ژنوتیپ‌ها، دو مسئله‌ی محاسباتی در شناسائی هاپلوتیپ‌ها مورد بحث قرار می‌گیرند؛ مسئله‌ی استنباط هاپلوتیپ‌ها از داده‌های ژنوتیپ، تحت مدل بیشترین پارسیمونی و مسئله‌ی تعیین بلوک‌های هاپلوتیپ. در مسئله‌ی اول، ترکیبی از یک الگوریتم ژنتیک و رده‌ای از روال‌های سودجویانه برای حل مسئله‌ی بیشترین پارسیمونی بکار گرفته می‌شوند. در مسئله‌ی دوم، برای تعیین ساختار بلوکی ژنوم به وسیله‌ی نمونه‌ای از هاپلوتیپ‌های جمعیت، ابتدا یک شاخص برای تعیین همبستگی بین اسنیپ‌ها تعریف می‌شود سپس، بلوک‌های هاپلوتیپ با حل یک مسئله‌ی بهینه‌سازی مقید بدست

می‌آیند. بر این اساس، بلوک‌های هاپلوتیپ به قسمی تعریف می‌شوند که بیشترین تعداد جفت اسنیپ‌های "همبسته" در ژنوم را در برگیرند و تعداد جفت اسنیپ‌های "مستقل" درون بلوک‌ها از کسر معینی تجاوز نکند. در این روش، از آزمون دقیق فیشر برای تعیین سطح معناداری همبستگی بین اسنیپ‌ها و برنامه‌ریزی پویا برای بدست آوردن افراز بهینه استفاده می‌شود.

در این رساله چند طرح جدید برای ارزیابی جنبه‌های مختلف ساختارهای بلوکی معرفی می‌شوند که از طریق آنها، شباهت بین ساختارهای بلوکی، ثبات بلوک‌های هاپلوتیپ و کارایی افرازهای بلوکی در شناسائی نقاط پراحتمال نوترکیبی و نیز شناسائی جایگاه ژنی مرتبط با بیماری مورد بررسی قرار می‌گیرند. برخلاف نتایج نه چندان رضایت‌بخش بدست آمده از الگوریتم پیشنهادی برای استنباط هاپلوتیپ‌ها، روش پیشنهادی برای افراز بلوکی هاپلوتیپ‌ها در بیشتر جنبه‌های مورد بررسی، برتری مطلوبی نسبت به دیگر روش‌های رایج نشان می‌دهد.

واژه‌های کلیدی: *ژنوتیپ، هاپلوتیپ، اسنیپ SNP، ژنتیک جمعیت، بازتوالی‌یابی ژنوم، آزمون همبستگی فیشر، برنامه‌ریزی خطی، برنامه‌ریزی پویا، ساختار بلوکی کروموزوم، نرخ نوترکیبی، مدلهای بیماری، شناسائی جایگاه بیماری، مطالعه‌ی case-control.*

پیشگفتار

در قرن جدید، دانش زیست‌شناسی بیشترین سهم در تحقیقات را متوجه خود می‌کند. زیست‌شناسی در مقایسه با دیگر علوم پایه مثل فیزیک و شیمی، قرنها از داشتن مدلهای نظری به قدر کافی دقیق و در عین حال جامع بی‌بهره بود و از این رو بیشتر به توده‌ای از واژگان می‌مانست که تنها برای نامگذاری پدیدارهای مرتبط با حیات ابداع می‌شدند در حالیکه در تبیین روابط علّی بین آنها نومیدکننده می‌نمود. تجربیات مندل در زمینه‌ی وراثت و نظریه‌ی تکاملی داروین تا پیش از قرن بیستم زمینه‌ی اولیه‌ی بنیان‌های تئوری پیدایش و بقای جانداران را فراهم کردند. کشف ساختار مارپیچ دو رشته‌ای DNA توسط واتسون و کریک در نیمه‌ی قرن گذشته تقریبا تمام آنچه را زیست‌شناسی به عنوان اصول موضوعه نیاز داشت فراهم کرد و مفهوم ذهنی ژن به عنوان عامل وراثت را در شکل ماده‌ای شیمیایی با ساختاری کاملا پیچیده عینیت بخشید. اکنون، در کنار این پارادایم جدید، گردآیه‌ای از دیگر علوم پایه، ریاضیات و آمار، و علوم مهندسی قرار گرفته‌اند تا درک ما از دنیای حیات را با سرعت بیشتری گسترش بخشند. از این میان، همزمان با دوران آغازین پروژه‌ی ژنوم و نیاز به ابزارهای کارآمد برای سازماندهی و جستجوی اطلاعات مرتبط با توالی‌های زیستی، رشته‌ی جدیدی تحت عنوان بیوانفورماتیک ظهور کرد. اکنون، بیوانفورماتیک با معرفی و توسعه‌ی ابزارهایی از علوم کامپیوتر برای حل مسائل زیست‌شناسی و به طور خاص زیست‌شناسی سلولی و مولکولی و ژنتیک جایگاه ویژه‌ای میان زمینه‌های تحقیقاتی در دنیای علم دارد.

رساله‌ی حاضر حاوی بخشهایی از مجموعه فعالیتهای پژوهشی نگارنده در طول دوره‌ی دکترای تخصصی بیوانفورماتیک در دانشگاه تهران از ابتدای سال تحصیلی ۱۳۸۳ است که در ارتباط با موضوع «مسائل محاسباتی در شناسائی هاپلوتیپ‌ها» به عنوان پایان‌نامه‌ی دکترا به انجام رسیده‌اند. در این رساله، دو مسئله‌ی متفاوت مورد بحث و بررسی قرار می‌گیرند: مسئله‌ی استنباط هاپلوتیپ‌ها از داده‌های ژنوتیپ توسط یک الگوریتم ژنتیک و مسئله‌ی تعریف یک ساختار بلوکی بر روی ژنوم افراد یک زیرجمعیت انسانی. هر یک از این مسائل با خاستگاه‌های متفاوتی به عنوان موضوع تحقیق مطرح می‌شدند و از اینرو پیشرفت یکسانی نداشته‌اند. ایده‌ی بکارگیری الگوریتم ژنتیک در مسئله‌ی استنباط هاپلوتیپ‌ها به عنوان اولین موضوع تحقیق، از کارآیی امیدوارکننده‌ای در مقایسه با دیگر روشهای رایج برخوردار نبود. با این حال مطالعه‌ی بخش‌های مرتبط با شیوه‌ی پیاده‌سازی الگوریتم ژنتیک برای حل مسئله‌ی استنباط هاپلوتیپ‌ها و مرور دیگر رویکردهای حل این مسئله در این رساله می‌تواند اطلاعات مفیدی از جزئیات و حدود کارآیی چنین روشهایی در اختیار کسانی‌که به این مبحث علاقمندند قرار دهد. مسئله‌ی دیگر یعنی تعریف ساختار بلوکی ژنوم در زیرجمعیت‌های انسانی

حجـم اصلی این رساله را در بر می‌گیرد. کارآیی روش پیشـنهاد شده برای این مسئله در این رساله به تفصیل با بکارگـیری در دو مسـئله‌ی دیگر مورد بررسی قرار می‌گیرد: مسـئله‌ای از ژنتیـــک آماری درباره‌ی تعییـــن نقاط پراحتمال نوترکیبی در ژنوم و دیگری مسئله‌ی شناسائی جایگاه ژنی مرتبط با بیماری در نمونه‌های case و control.

به عنوان موضوعی رایج در بیوانفورماتیک، در این رساله نیز خواننده با طیف پراکنده‌ای از ابزارها از آمار و علوم کامپیوتر تا ژنتیک و تکامل که برای حل مسائل طرح شده بکار گرفته شده‌اند مواجه می‌شود. هرچند بیشتر این رهیافتها، مثل برنامه‌ریزی خطی برای حل مسئله‌ی افراز بلوکی هاپلوتیپ‌ها، از پایه مورد بحث قرار نگرفته‌اند اما برای خواننده‌ی آشنا به چنین زمینه‌هایی می‌تواند قابل توجه باشد. مطالب در این رساله، مطابق با چارچوب استاندارد در علوم زیستی، به ترتیب در فصل‌های مقدمه ــ مواد و روشها ــ نتایج و بحث و نتیجه‌گیری قرار گرفته‌اند. بدین ترتیب لازم است، خواننده برای مطالعه‌ی هر یک از سه مسئله‌ی مورد بحث، بخشهای مجزا در هر یک از فصل‌های رساله را به طور جداگانه تعقیب کند.

قدردانی

این رساله حاصل پشتیبانی‌های فکری و روحی افراد بی‌شماری است که نگارنده بر خود لازم می‌داند در اینجا بخشی از آنها را یاد آورد شود. اصولا، بخش عمده‌ای از پژوهشگران بیوانفورماتیک در ایران، آشنایی با این رشته را مرهون تلاش‌های جناب دکتر بهرام گلیائی مدیر گروه بیوانفورماتیک دانشگاه تهران و جمعی از اساتید بین‌رشته‌ای می‌دانند که هسته‌ی اولیه‌ی این گروه را تشکیل می‌دادند. به ویژه، تلاش‌های جناب دکتر مهدی صادقی در معرفی مسائل گوناگون بیوانفورماتیک و ایجاد ارتباط علمی بین اساتید رشته‌های مختلف، شایسته‌ی قدردانی است. همچنین است راهنمایی‌های جناب دکتر حمید پزشک در مسائل آماری که جزء همیشگی مسائل بیوانفورماتیک هستند. به علاوه لازم است از اساتید محترم جناب دکتر اصلاحچی (دانشگاه شهید بهشتی) و سرکار خانم دکتر الهی به خاطر راهنمایی‌هایشان که به بهبود راه حل‌ها و پرداخت نتایج کمک کرد تشکر کنم. راهنمایی‌های جناب دکتر آرمین مددکار در مورد شیوه‌ی ارائه‌ی مطالب و همکاری بسیار صمیمانه‌ی آقای سید امیر مرعشی در نگارش مقاله نیز درخور قدردانی فراوانند. در اینجا لازم است به طور جداگانه از پژوهشگاه دانش‌های بنیادی و قطب زیست ـ ریاضی دانشگاه تهران که بخش‌هایی از پژوهش مرتبط با این رساله را مورد حمایت مالی قرار دادند تشکر شود. بسیاری دیگر از اساتید و دانشجویان مرکز تحقیقات بیوشیمی و بیوفیزیک و دانشکده‌ی علوم دانشگاه تهران، پژوهشگاه دانشهای بنیادی و نیز مرکز تحقیقات ژنتیک و زیست‌شناسی مولکولی ماکس پلانک (برلین) هستند که نگارنده را در طی دوره‌ی پژوهشی از دیدگاه‌ها و ایده‌های پرفایده‌یشان در ارتباط با موضوع این تحقیق بهره‌مند ساخته‌اند. در انتها باید از تشویق‌های امیدبخش همسرم برای ادامه این دوره و تکمیل رساله تشکر کنم.

فهرست مطالب

لیست جدول‌ها

لیست شکل‌ها

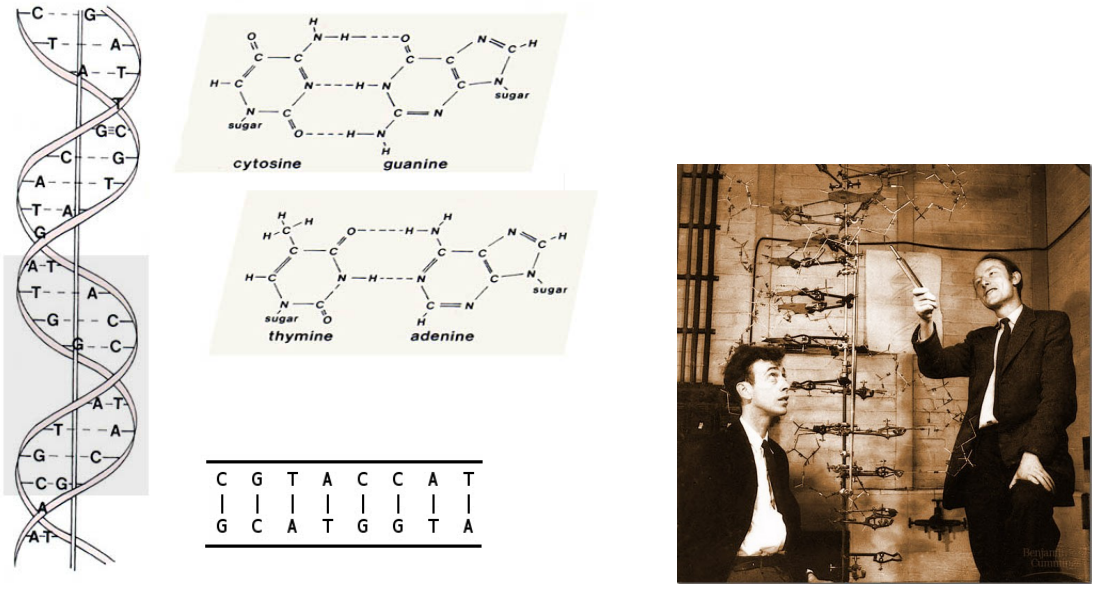
فصل ۱

مقدمه

۱۰۱ آشنایی با واژگان ژنتیک

برای هر موجود زنده، اطلاعات به مفهوم کلی کلمه، از خصیصه‌های ظاهری یک فرد در جمعیتی از جانداران گرفته تا جزئیات دستورالعمل تقسیم سلولی، در واحدهایی که به طور کلی آنها را ژن می‌نامیم نگهداری می‌شوند. ژنها طی تولیدمثل جاندار یا تقسیم سلولی، تکثیر می‌شوند و نسخه‌ی یکسانی از آنها به نسل بعد منتقل می‌شود تا بدین ترتیب موجود جدید تمام اطلاعات لازم برای حیات را در اختیار داشته باشد. تا پیش از انقلابی که با کشف واتسون و کریک در زیست‌شناسی آغاز شد معلوم شده بود که ژنها با ماده‌ای که درون هسته‌ی سلول است در ارتباطند. این ماده با نام شیمیایی اسید دزوکسی ریبونوکلئیک یا به اختصار DNA شناخته می‌شد. با کشف ساختار DNA معلوم شد که DNA مولکول پلیمری بسیار طویلی از زیرواحدهای ساده‌تری به نام نوکلئوتید است. آنها نشان دادند تنها چهار نوع نوکلئوتید در ساختمان DNA نقش دارند که اکنون آنها را به اختصار با حروف A ، C ، G و T می‌شناسیم. در واقع ترکیبهای مختلف کنار هم قرار گرفتن این چهار نوع نوکلئوتید در امتداد زنجیر DNA به طور آرمانی توانایی رمزکردن بیشمار ژن را فراهم می‌کند. زنجیر DNA جدا از ساختار نردبانی تاب‌خورده‌اش توالیی از چند صد میلیون نوکلئوتید است که می‌توان آنرا مانند یک رشته از الفبای چهار حرفی {A,C,G,T} در نظر گرفت. البته طبیعت برای حفظ پایداری بیشتر و

مصون نگهداشتن این اطلاعات از صدمات، اطلاعات را به صورت تکراری، به شکل یک نردبان دورشته‌ای شامل جفت نوکلئوتیدهای روبروی هم نگه‌می‌دارد. از آنجا که نوکلئوتیدها دارای خاصیت بازی نیز هستند، اندازه‌ی DNA را در واحد جفت‌باز(base pair) می‌شمارند. به عنوان مثال اندازه‌ی رشته‌ی DNA در یک سلول باکتری E.Coli، نزدیک به۵ میلیون جفت‌باز است.



شکل ۱۰۱: ساختمان DNA
راست) واتسون و کریک در کنار مدل ساختمان DNA. **چپ)** ساختار اتمی اسیدهای نوکلئیک و موقعیت آنها در قطعه‌ای از زنجیر DNA.

هر ژن جایگاه خاصی از این رشته‌ی عظیم را به خود اختصاص می‌دهد. با اینکه قابل تصور است اگر هر ژن رشته‌ی DNA جداگانه‌ای می‌داشت ولی تکامل، طبیعت را به این سو هدایت کرده است که تمامی ژنها بر روی یک رشته‌ی بزرگ DNA در کنار یکدیگر قرار گیرند. البته این وضعیت در موجودات عالی کمی تغییر می‌کند. بدین ترتیب که ژنها بر روی چندین قطعه‌ی بزرگ DNA در هسته‌ی سلول قرار می‌گیرند. هر یک از این قطعات بزرگ DNA یک کروموزوم را تشکیل می‌دهند. از نظر ساختار فضایی، هر کروموزوم ساختمان در هم پیچیده و فشرده‌شده‌ی یک زنجیر بسیار بلند DNA است که در زیر میکروسکپ به اشکال خاصی قابل مشاهده است. از دیدگاه سلولی، تعداد و اندازه کروموزومها خصیصه‌ای مربوط به گونه است، یعنی در تمام سلولهای افراد یک گونه‌ی خاص (تقریبا) یکسان است. مثلا این تعداد در مورد انسان ۴۶ است. البته در تمامی جاندارانی که از تولیدمثل جنسی بوجود می‌آیند، سلولها همواره حاوی دو دست کروموزوم هستند. هر دست

کروموزوم نسخه‌ی یکسانی از کروموزومهایی است که از هر یک از والدین به ارث رسیده است. بدین ترتیب هر سلول انسان دقیقا حاوی ۲۳ جفت کروموزوم است. کروموزومهای متناظر در هر جفت را کروموزومهای همسان (homologous) می‌گویند. مجموع DNA در هر دست از این ۲۳ جفت کروموزوم شامل حدود ۳ میلیارد جفت‌باز می‌شود. جاندارانی که دو نسخه‌ی یک DNA را داشته باشند اصطلاحا دیپلوئید (diploid) نامیده می‌شوند.

اصطلاح ژنوم به مجموعه‌ی همه‌ی ژنهای یک جاندار اطلاق می‌شود. در موجودات عالی نواحی بسیار وسیعی از DNA فاقد ویژگیهای عمومی ژنها هستند. به عنوان مثال بیش از سه چهارم DNA در انسان را نواحی بین ژنی تشکیل می‌دهند. با این حال امروزه باور بر این است که برای هیچ یک ازمناطق DNA به قطعیت نمی‌توان اظهار کرد که هیچ نقشی وجود ندارد. بنابراین رایج بر این است که تمامی DNA یک جاندار را ژنوم آن جاندار بنامند. در این رساله، منظور از ناحیه‌ی ژنومی، قطعه‌ای از کروموزوم است که علاقمند به مطالعه برخی ویژگی‌های آن هستیم. منظور از فنوتیپ خصیصه‌های ظاهری و به سادگی قابل تشخیص در بین افراد مختلف یک گونه از جانداران است، مثل رنگ پوست یا استعداد ابتلا به یک بیماری خاص. در واقع، وجود هر فنوتیپ مرتبط با بروز یک یا چند ژن خاص است. توالی‌های مختلفی که می‌توانند در این ژن یا ژنها دیده شوند را اصطلاحا ژنوتیپ‌های مرتبط با فنوتیپ می‌نامند.

تنوع حیات ناشی از تنوع ژنوم است. دو رویداد اصلی زمینه‌ی بروز این تنوع هستند: جهش و نوترکیبی. جهش رویدادی است که طی آن یک یا چند باز بر روی رشته‌ی DNA به باز یا بازهای دیگری تغییر می‌یابد یا در مواردی یک نوکلئوتید از رشته حذف می‌شود یا یک نوکلئوتید جدید در رشته درج می‌شود. اگر اطلاعات ژنتیکی در این توالی جدید آنچنان معیوب نشده باشد که به مرگ جاندار منجر شود، توالی جهش یافته به نسلهای بعدی منتقل می‌شود و بسته به شرایط جمعیتی و میزان سازگاری (fitness) با شرایط محیط، سهمی از ژنوتیپ جمعیتی را به خود اختصاص می‌دهد.

در متداولترین شکل رویداد نوترکیبی، بخشهایی مشابهی از دو کروموزوم همسان با یکدیگر مبادله می‌شوند که به آن cross-over می‌گویند. منظور از کروموزومهای همسان، کروموزومهای متناظر با یکدیگر در افراد مختلف یک گونه‌ی ثابت است. بر خلاف رویداد جهش، ژنومی که از چنین نوترکیبی بدست می‌آید به ندرت

ممکن است به مرگ جاندار و حتی به زایش گونه‌ی جدیدی منجر شود. از این رو نرخ رویدادهای نوترکیبی در مقایسه با جهش به مراتب بیشتر است. بین کروموزومهای همسان انسان در هر نسل تا دهها نوترکیبی می‌تواند روی دهد.

باید به یاد داشت که هنگامی که صحبت از ژنوم انسان می‌شود منظور توالی کاملی از DNA است که به طور مشترک در بین تمام افراد مختلف جمعیت انسان دیده می‌شود. در واقع توالی ژنوم در بین افراد مختلف یک گونه‌ی معین به جز اختلافهای بسیار جزئی، یکسان هستند. مثلا در مورد افراد انسان، تخمین زده می‌شود بیش از ۹۹/۹ درصد ژنوم بین هر دو فرد یکسان باشد[1]. همین اختلافهای ناچیز زمینه‌ی اصلی تفاوت‌های ظاهری یا فنوتیپی در بین افراد مختلف است. این اختلافها به اشکال گوناگونی بر روی ژنوم دیده می‌شوند. رایج‌ترین شکل تفاوت در بین ژنوم افراد مختلف یک جمعیت، چندریختی تک نوکلئوتیدی (Single Nucleotide Polymorphis) یا به اختصار SNP (اسنیپ) است. یک اسنیپ جایگاهی بر روی ژنوم است که در بین افراد مختلف جمعیت بیش از یک نوع نوکلئوتید در آن مشاهده می‌شود. در واقع، هر اسنیپ نتیجه‌ی یک رویداد جهش در زمان بسیار دور بوده است که به نسل‌های بعد منتقل شده است تا در زمان حال که ژنومِ کسر معینی از جمعیت، حامل آن است.

به طور کلی اشکال مختلفی که در یک جایگاه ژنی معین دیده می‌شوند را آلل‌های آن ژن می‌گویند. مثال ساده‌ی آن، آللهای ژن مرتبط با گروه خون است که آنها را با A، B و O می‌شناسیم. در مورد اسنیپ‌ها، نوکلئوتیدی را که فراوانتر در جمعیت مشاهده می‌شود آلل اصلی و نوکلئوتیدی که کمتر مشاهده می‌شود را آلل فرعی می‌نامیم. نرخ رویداد جهش، از آن دست که مشکلی برای بقای جاندار ایجاد نکند، بسیار پایین است، آن چنان که احتمال اینکه در یک جفت باز در بین میلیاردها جفت باز ژنوم انسان دو بار جهش روی داده باشد نزدیک به صفر است. بدین ترتیب تقریبا همه اسنیپ‌ها دو آللی هستند و به ندرت اسنیپ سه آللی مشاهده می‌شود. از دیگر سو، بر روی ژنوم هر فرد، نوکلئوتیدهای فراوانی ممکن است وجود داشته باشند که با آنچه در ژنوم دیگر افراد وجود دارد تفاوت داشته باشند. با این حال، بسیاری از آنها جز جهش‌های بدنی (somatic mutation) بیانگر اطلاعات دیگری نیستند و نمی‌توان آنها را اسنیپ به شمار آورد. از اینرو رایج آن است

[1] جالب است اگر بدانید بین ژنوم انسان و ژنوم موش قریب به ۹۰ درصد شباهت وجود دارد.

که یک کران پایین برای فراوانی مشاهده‌ی آلل فرعی[۲] در نظر می‌گیرند تا این جهش‌های نادر از اسنیپ‌های رایج در جمعیت تمایز داده شوند. مثلا در بسیاری از مطالعات، تنها جایگاه‌هایی را به عنوان اسنیپ در نظر می‌گیرند که در آنها فراوانی آلل فرعی بیشتر از یک درصد باشد.

```
A G G A C T A G A T A A T A G A C C G     0 1 1 0 0
A G G A C C A C A T T A T A G T C C G     0 0 0 1 1
A G G A C C A G A T A A T A G T C C G     0 0 1 0 1
A T G A C C A C A T T A T A G T C C G     1 0 0 1 1
A T G A C T A C A T A A T A G A C C G     1 1 0 0 0
```

شکل ۲۰۱: آلل‌های متفاوت پنج اسنیپ در قطعه‌ی یکسانی از پنج نمونه‌ی متفاوت از یک کروموزوم
توالی کروموزومی هر فرد را می‌توان مجموعه‌ای از توالی نوکلئوتیدهای یکسان در بین تمام افراد و توالی نوکلئوتیدهای فرد در جایگاه‌های اسنیپ در نظر گرفت. توالی کروموزوم در جایگاه‌های اسنیپ را می‌توان با کدهای صفر و یک برای هر نمونه نمایش داد.

وقتی ژنوتیپ یک جاندار دیپلوئید را در یک جایگاه اسنیپ بررسی می‌کنیم سه وضعیت متفاوت ممکن است مشاهده شود: یا آلل‌های بر روی دو کروموزوم، متفاوتند که یک حالت هتروزیگوت (heterozygot) نامیده می‌شود یا حالتی را داریم که هر دو کروموزوم آللهای یکسانی دارند یعنی یا هر دو آلل، آلل اصلی هستند یا هر دو، آلل فرعی هستند، که به ترتیب هموزیگوت ماژور (homozygot major) و هموزیگوت مینور (homozygot minor) نامیده می‌شوند. وقتی توالی ژنومی را تنها به جایگاه اسنیپ‌ها محدود کنیم، دنباله‌ای از وضعیتهای سه گانه بر روی اسنیپ‌ها بدست می‌آید. به این توالی اسنیپ‌ـژنوتیپ می‌گوییم. یاد آوری این نکته ضروری است که در این حالت توالی ژنوم همزمان از مجموعه‌ی دو کروموزوم همسان خوانده می‌شود و به همین جهت در مواردی برای تاکید بر این موضوع، توالی آللها در این حالت را دیپلوتیپ (diplotype) می‌گویند. اگر توالی ژنوم را تنها بر روی یک کروموزوم به طور مستقل مطالعه کنیم، توالی آللهایی که در موقعیت اسنیپ‌ها مشاهده می‌شوند اسنیپ‌ـهاپلوتیپ (SNP-haplotype) نامیده می‌شوند. در این رساله به اختصار واژگان هاپلوتیپ و ژنوتیپ را به ترتیب برای اشاره به اسنیپ‌ـهاپلوتیپ و اسنیپ‌ـدیپلوتیپ به کار می‌بریم.

[۲] Minor Allele Frequency (MAF)

۲۰۱ فنآوری‌های تعیین ژنوتیپ

توالی‌یابی، مجموعه عملیاتی است که از طریق آن ترتیب قرار گرفتن نوکلئوتیدها در یک رشته‌ی DNA تعیین می‌شود. از زمان کشف ساختار DNA تا به امروز تلاشهای زیادی برای توالی‌یابی کامل ژنوم جانداران مختلف صورت گرفته است. ماکسام و گیلبرت اولین شیوه‌ی تعیین توالی DNA را در ۱۹۷۶ ابداع کردند[۱]. رهیافت آنها مبتنی بر اجرای تعدادی واکنشهای شیمیایی است که طی آن نوکلئوتیدها به ترتیب از یک سر زنجیر DNA حذف می‌شوند. نوع نوکلئوتید انتهای زنجیر مرتبط با نوع واکنشی که برای جدا کردن آن به کار می‌رود قابل تشخیص است و با تعیین طول رشته‌ی باقیمانده، مکان قرارگیری این نوکلئوتید در زنجیر معین می‌شود. این شیوه و گونه‌های مشابه آن در زمان خود بسیار فراگیر شدند و از این طریق می‌توانستند توالی‌هایی تا چند صد باز را در نواحی ژنی متعددی از جانداران مختلف تعیین کنند. از این میان باید به اولین ژنومی که توالی DNA آن به طور کامل استخراج گردید اشاره کرد که البته مربوط به ویروسی به نام MS2 بود[۲].

شیوه‌ای که رواج بیشتری یافت و حتی برخی روشهای امروزی تعیین توالی بر پایه‌ی آن استوار هستند، روش منسوب به سنگر است [۳]. بر خلاف شیوه‌ی ماکسام و گیلبرت که مبتنی بر انجام تعدادی واکنش‌های شیمیایی بر روی DNA است و از اینرو به مراحل متعدد خالص‌سازی نیاز دارد، شیوه‌ی سنگر با طی مراحل آزمایشگاهی ساده‌تر می‌تواند رشته‌های به نسبت بلندتری را توالی‌یابی کند. با این شیوه، سنگر توانست در ۱۹۷۷ ژنوم ویروس φX174 را به طور کامل تعیین کند[۴]. شیوه‌ی سنگر مبتنی بر استفاده از آنزیم DNA ـ پلیمراز و فرایندی مشابه فرایند طبیعی همانندسازی DNA از یک رشته‌ی الگو است. سنگر ماده‌ی خاصی به نام دی‌ـ دئوکسی‌نوکلئوتید(di-deoxynucleotide) را شناسائی کرد که بکارگیری آن در محیط همانندسازی باعث می‌شود ادامه‌ی همانندسازی DNAپس از جفت شدن این ماده با نوکلئوتید متناظرش متوقف گردد. با هدف استفاده از حسگرهای نوری برای خواندن توالی DNA، در ترکیب دی‌ـ دئوکسی نوکلئوتید متناظر با هر نوع نوکلئوتید، عنصر فلورسانس با طول موج معینی مورد استفاده قرار می‌گیرد و از این طریق نوع نوکلئوتید انتهای رشته‌ی همانندسازی شده قابل تشخیص خواهد بود. این رویکرود در اولین ماشین اتوماتیک توالی‌یابی بکارگرفته شد که توسط آن ژنوم چندین جاندار ساده، این بار فراتر از ویروسها، به طور کامل تعیین گردید[۵، ۶].

هیچ یک از شیوه‌های توالی‌یابی قادر نیستند توالی کامل رشته‌های بلند DNA را با تنها یکبار اجرا بدست آورند. طول توالی‌های خوانده شده توسط ماشین‌های اتوماتیک توالی‌یابی از ۲۰۰۰ جفت‌باز تجاوز نمی‌کند. اساسا اطمینان از درستی نوکلئوتید خوانده شده با افزایش طول رشته کاهش می‌یابد. این مشکل علیرغم پیشرفت در فنآوری‌های توالی‌یابی تا به امروز نیز وجود دارد. با آنکه ماشین‌های توالی‌یابی امروزی با توانی صدها برابر همنوعان اولیه‌یشان، اطلاعات نوکلئوتیدها را استخراج می‌کنند با این حال طول توالی‌های خوانده شده در آنها از ۵۰۰ جفت‌باز تجاوز نمی‌کند. این مشکل به طور اساسی باعث می‌شود برای تعیین توالی یک ژنوم به طور کامل، ترفندهای خاصی بکار گرفته شود. رهیافت اصلی در عمده‌ی روشهای تعیین توالی کامل ژنوم آن است که ژنوم مکررا و از نقاط متفاوت به تعداد زیادی قطعات کوتاه قابل توالی‌یابی بریده شود تا پس از توالی‌یابی این قطعات، با رویهم قرار دادن آنها توالی کامل ژنوم بدست آید[۳]. در روش‌های موسوم به توالی‌یابی تفنگ‌ساچمه‌ای [۴]، ژنوم به طور تصادفی در نقاط متفاوتی بریده می‌شود. از آنجا که در این رویکرد همپوشانی قطعات با یکدیگر به طور تصادفی روی می‌دهد، لازم است هر نقطه از ژنوم توسط دست کم تعداد معینی از قطعات پوشانیده شود تا احتمال به وجود آمدن شکاف در توالی کامل ژنوم به حداقل مقدار قابل پذیرش برسد. معادلات اریک لاندر و مایکل واترمن [۷]، مبنای تئوری تحقق یک توالی‌یابی کامل از طریق این شیوه را فراهم کرد و بدین ترتیب ایده‌ی تعیین توالی ژنوم انسان به طور عملی مورد توجه قرار گرفت. به‌طور مثال، اگر ۳۰ میلیون توالی با متوسط طول ۵۰۰ جفت‌باز برای تعیین توالی ژنوم ۳ میلیارد نوکلئوتیدی انسان بکار گرفته‌شود در هر هزار نوکلئوتید به طور میانگین کمتر از ۷ نوکلئوتید با هیچ قطعه‌ای ممکن است پوشیده نشوند.

با فراهم آمدن زمینه‌های تئوری و عملی توالی‌یابی در مقیاس ژنومی، پروژه‌ی توالی‌یابی ژنوم انسان به طور رسمی در ۱۹۹۰ با پشتیبانی مالی مؤسسه‌ی NIH در ایالات متحده طرح گردید و پیش‌بینی شد در مدت ۱۵ سال این پروژه به پایان برسد. با این حال تا ۱۹۹۸ تنها پنج درصد از کل ژنوم توالی‌یابی شده بود و پیش‌بینی پایان طرح در ۲۰۰۵ کمی دور از انتظار بنظر می‌رسید. اما آغاز پروژه‌ی دیگری برای توالی‌یابی ژنوم انسان در شرکت Celera و با هزینه‌ی بخش‌های غیردولتی، رقیب تازه‌ای برای پروژه‌ی ژنوم بوجود آورد که در نتیجه‌ی

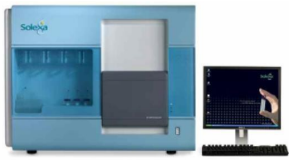
[۳] ابزار رایج برای قطعه قطعه کردن DNA، استفاده از آنزیمهای محدودکننده (restriction enzyme) در انواع و غلظت‌های متفاوت است. استفاده از دمای پایین و فشار بالا به طور ناگهانی، شیوه‌ی متداول دیگری برای این کار است.

[۴] shotgun sequencing

رقابت و نیز همکاری متقابل این دو مجموعه، سرانجام نقشه‌ی اولیه‌ی یک ژنوم توالی‌یابی شده در اواخر ۲۰۰۰ بدست آمد[۸، ۹]. سرانجام دو سال جلوتر از زمان پیش‌بینی شده، پروژه‌ی ژنوم انسان به طور رسمی پایان یافت و نقشه‌ی (نسبتاً) کاملی از توالی ژنومی و موقعیت قرارگیری قریب به ۲۶،۰۰۰ ژن روی آن بدست آمد[۱۰]. این توالی بیش از ۹۹ درصد ژنوم یوکروماتیک در انسان را پوشش می‌دهد و احتمال خطا در آن، یک در یکصدهزار نوکلئوتید است. پروژه‌ی ژنوم انسان با هدف تهیه‌ی یک توالی هاپلوئیدی مرجع برای ژنوم انسان شروع به کار کرد[۵]. پس از آن و با گسترش فنآوری‌های توالی‌یابی و بازتوالی‌یابی، ژنوم انسان به صورتی که به طور معمول در سلولهای مختلف بدن حضور دارد یعنی به شکل دیپلوئید، به طور اختصاصی برای یک فرد توالی‌یابی شد[۱۱].

ABI 3730XL
Applied Biosystems
capillary tube
1 Mb / day

Genome Sequencer FLX
Roche / 454
Pyrosequencing
100 Mb / run



Genetic Analyzer
Illumina / Solexa
Reversible terminator
2000 Mb / run (36bp reads)

SOLiD
Applied
Biosystems
Oligonucleotide Ligation
3000 Mb / run (25bp reads)

شکل ۳۰۱: فنآوری نسل جدید توالی‌یابی رشته‌های نوکلئوتیدی

امروزه، ماشین‌های نسل جدید توالی‌یابی توانایی خواندن صدها مگاباز اطلاعات ژنومی را در زمانی کمتر از یک روز دارا می‌باشند. بدین ترتیب امکان تعیین توالی کامل ژنوم یک فرد به طور خصوصی به هیچ وجه امری دور از ذهن نیست. با این وصف در بسیاری از کاربردها کافی است تا با صرف هزینه‌ی به مراتب کمتر

[۵] نمونه‌های مورد استفاده در پروژه‌ی ژنوم انسان متعلق به چهار نفرند که هویتشان ناشناس نگه‌داشته می‌شود.

تنها، اطلاعات متفاوت ژنوم فرد با ژنوم مرجع بدست آید. در این بین اسنیپ‌ها به دلیل سادگی ساختاریشان و فراوانی و پراکندگی نسبتا قابل قبولشان در سراسر ژنوم، دستاویز مناسبی برای تعیین توالی ژنومی با چنین رویکردی هستند. پروژه‌ی HapMap بر همین مبنا و در آغاز عصر مابعد پروژه‌ی ژنوم، به منظور شناسائی ژنوتیپ اسنیپ‌ها در میان جمعیت‌های گوناگون انسان با همکاری بین‌المللی شش کشور در ۲۰۰۲ آغاز گردید[۱۲]. این پروژه در اولین فاز، ژنوتیپ بیش از یک میلیون اسنیپ را در نمونه‌ای از ۲۶۹ نفر بدست آورد [۱۳]. در دومین فاز، این تعداد به بیش از سه میلیون اسنیپ در ۲۷۰ نفر رسید [۱۴]. در هر دوی این فازها، نمونه‌ها از سه پانل مختلف انتخاب می‌شدند. هر پانل شامل نمونه‌هایی از یک جمعیت نژادی خاص است که عبارتند از پانل YRI شامل ۳۰ سه‌تائی فرزند ـ پدر ـ مادر از ایبادان[۶] در نیجریه، پانل CEU شامل ۳۰ سه‌تائی فرزند ـ پدر ـ مادر مقیم ایالات متحده که منشاء اروپای شمالی و غربی دارند و پانل CHB+JPT شامل ۴۵ نفر غیرخویشاوند از ساکنین توکیو و ۴۵ نفر غیرخویشاوند از ساکنین پکن[۷]. این اطلاعات مجموعه‌ی باارزشی از داده‌های مورد نیاز را برای انجام بسیاری از تحقیقات از جمله ژنتیک بیماری‌ها، طراحی دارو، تکامل، ژنتیک جمعیت و تشخیص هویت فراهم کرده است. پایگاه اطلاعات HapMap هم اینک از طریق اینترنت در دسترس عموم قرار دارد[۸]. پایگاه داده‌ای دیگری که در این مبحث مرجع مهمی به حساب می‌آید dbSNP است. این پایگاه داده‌ای، اطلاعات مربوط به مکان و ژنوتیپ تمام اسنیپ‌هایی که در پی تحقیقات آزمایشگاهی مختلف شناسائی شده‌اند را نگهداری می‌کند. عمده‌ی اولین اطلاعات ثبت شده در این پایگاه داده‌ای از اسنیپ‌هایی تشکیل می‌شد که طی پروژه‌ی ژنوم انسان شناسائی شده بودند[۱۵، ۱۶]. اسنیپ‌های گزارش شده توسط پروژه‌ی HapMap بخش قابل توجه دیگری از مجموع ۴/۳ میلیون اسنیپ ثبت شده در dbSNP را تشکیل می‌دهند[۱۷].

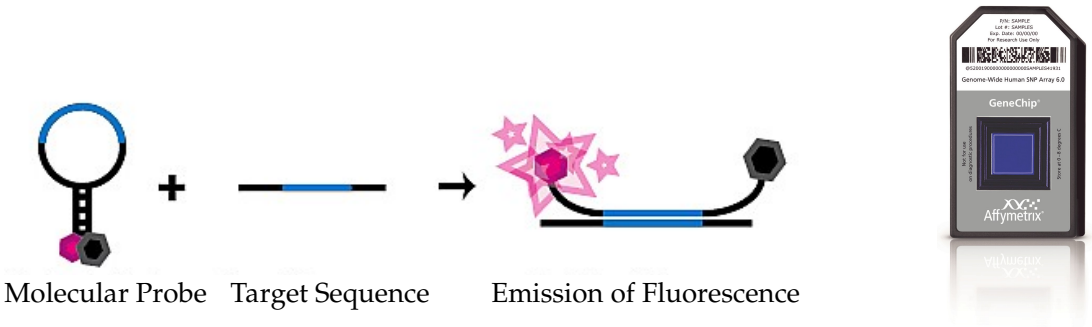
امروزه فنآوری‌های متنوعی برای تعیین ژنوتیپ اسنیپ‌ها عرضه می‌شوند. تراشه‌های تولیده شده توسط شرکت Affymetrix یکی از رایج‌ترین ابزارها در این زمینه است. به وسیله‌ی این تراشه‌ها که SNP-microarry نامیده می‌شوند می‌توان ژنوتیپ صدهاهزار اسنیپ در نمونه‌ی مورد آزمایش را طی یک اجرا

[۶] Ibadan

[۷] هر سه‌تائی تنها چهار هاپلوتیپ غیریکسان و هر فرد دو هاپلوتیپ غیریکسان دارد.

[۸] به تازگی داده‌های فاز سوم HapMap شامل اطلاعات ۱/۳ میلیون اسنیپ در ۱۰ پانل جمعیتی شامل ۱۱۸۴ فرد نمونه بر روی اینترنت قرار داده شده است.

بدست آورد. بر روی هر تراشه، آرایه‌ای از هزاران پروب (probe) قرار داده شده است. هر پروب از یک زنجیر تک رشته‌ای DNA تشکیل شده است. تعدادی از نوکلئوتیدها در دو انتهای این زنجیر مکمل یکدیگرند و یک ساختار دو سنجاق‌سری ایجاد می‌کنند. قسمت میانی هر پروب شامل یک توالی است که به طور اختصاصی مکملِ یک توالی ۲۵ نوکلئوتیدی هدف در ژنوم است. در این کاربرد منظور از یک توالی هدف در واقع همان توالی تشکیل شده از نوکلئوتیدهای قرار گرفته در دو طرف یک اسنیپ بر روی ژنوم است. بازای هر آلل اسنیپ، یک توالی اختصاصی و در نتیجه یک دسته پروب بر روی تراشه قرار دارد. در صورت وجود توالی هدف در نمونه DNA مورد بررسی که مکمل توالی قسمت میانی پروب باشد و قرار گرفتن آن در نزدیکی پروب، یک زنجیر دو رشته‌ای با جفت شدن پروب و توالی هدف بوجود می‌آید که سبب می‌شود بازهای جفت‌شده در ساقه‌ی پروب از یکدیگر جدا شوند. با این امر، عامل فلورسانسی که در انتهای پروب تعبیه شده است از عامل فرونشاننده دور می‌شود و تابش آن توسط حسگر قابل دریافت می‌شود.



شکل ۴۰۱: سازوکار شناسائی توالی نوکلئوتیدی در تراشه‌های SNP-microary
راست) تراشه‌های شرکت Affymetrix. **چپ)** بر روی هر تراشه صدهاهزار توالی کوتاه اختصاصی که در انتهایشان یک عامل فلورسانس قرار دارد نصب شده است. اگر در نمونه‌ی مورد بررسی توالی مکمل این توالی وجود داشته باشد، عامل فلورسانس از مجاورت عامل فرونشاننده دور می‌شود و تابش آن توسط حسگرها قابل دریافت می‌شود.

شیوه‌های بسیار دیگری نیز برای تعیین ژنوتیپ اسنیپ‌ها ابداع شده‌اند که هر یک ابزار متفاوتی را برای شناسائی نوکلئوتیدهای قرار گرفته در موقعیت‌های اسنیپ بکار می‌گیرند. به عنوان مثال، در حالی که در شیوه‌های معمول اسنیپ‌ها دو آللی فرض می‌شوند، تکنیک ویژه‌ای برای تعیین ژنوتیپ اسنیپ‌های سه آللی در [۱۸] معرفی شده‌است. برخی دیگر از این روشها در قالب یک محصول تجاری عرضه می‌شوند و به طور فراگیر برای تعیین ژنوتیپ اسنیپ‌ها و دیگر اشکال تنوع ژنومی در زمینه‌های مختلف تحقیقاتی به کار گرفته می‌شوند، که از آن جمله می‌توان به فنآوریهای ABI TaqMan [۱۹]، APEX-2 [۲۰] و Illumina

Beadchip [۲۱] اشاره کرد.

علاوه بر فنآوری‌های اختصاصی برای تعیین ژنوتیپ اسنیپ‌ها مثل SNP-microarry که در بالا به آن اشاره شد، می‌توان برخی ماشین‌های نسل جدید توالی‌یابی را نیز برای تعیین ژنوتیپ اسنیپ‌ها به استخدام گرفت. ماشین توالی‌یاب GA محصول شرکت Illumina/Solexa نمونه‌ی از این فنآوری‌هاست. این ماشین، در واقع وسیله‌ای برای تعیین توالی هر نوع رشته‌ی نوکلئوتیدی اعم از انواع DNA و RNA است. ایده‌ی کلیدی مورد استفاده در ماشین‌های توالی‌یاب Illumina در بکارگیری فرآیندی به نام Reversible Terminator Chemistry نهفته است. در این فرآیند، نوکلئوتیدهایی که برای همانندسازی DNA بکارگرفته می‌شوند به گونه‌ای طراحی شده‌اند که با اتصال اولین نوکلئوتید به زنجیره‌ی آغازگر الگوبرداری، ادامه‌ی همانندسازی ناممکن می‌شود. با تحریک عامل فلورسانس موجود در این نوکلئوتید طول موج معینی منتشر می‌شود که از طریق آن می‌توان نوع نوکلئوتید را شناسائی کرد(شکل ۵۰۱)[۲۲]. این ماشین قادر است در مدت سه روز میلیونها قطعه‌ی کوتاه ۳۶ نوکلئوتیدی جمعا به حجم ۱/۳ گیگاباز را برای حداکثر هشت نمونه‌ی مستقل و با هزینه‌ای در حدود ۳۰۰۰ یورو توالی‌یابی کند[۲۳]. علیرغم هزینه بالاتر و حجم بالای داده‌های خروجی در فنآوری‌های بازتوالی‌یابی، بکارگیری آنها برای تعیین ژنوتیپ‌ها در مقایسه با فنآوری‌های مبتنی بر پروب‌ها از این مزیت برخوردار است که در شیوه‌های مبتنی بر پروب این امکان وجود دارد که برای برخی اسنیپ‌های موجود در نمونه‌ی مورد مطالعه هیچ پروب مناسبی تعبیه نشده باشد و به همین دلیل اطلاعات بدست آمده از فنآوری‌های مبتنی بر پروب، زیرمجموعه‌ای از تمام اسنیپ‌های ممکن در ژنوتیپ فرد را شامل می‌شوند در حالیکه در روشهای بازتوالی‌یابی، تمام توالی ژنوم بدست می‌آید.

باید به خاطر داشت که چون نمونه‌های DNA در این نوع بررسی‌ها اساسا از سلول‌های دیپلوئید برداشته می‌شوند، نتیجه‌ی نهایی نیز به صورت اطلاعات ژنومیِ ترکیب شده از جفت کروموزوم‌ها بدست می‌آید و از این رو هاپلوتیپ‌ها به طور مستقل نامعلومند. شناسائی هاپلوتیپ‌ها با در اختیار داشتن چنین اطلاعاتی، مسئله‌ی مشترک تمام بررسی‌های ژنتیکی از این دست است. رویکرد رایج برای حل این مسئله، استفاده از روشهای محاسباتی در ترکیبیات و آمار است. بر این اساس الگوریتم‌های کارآمد زیادی توسط تیم‌های تحقیقاتی مختلف معرفی شده‌اند. توسط این الگوریتم‌ها می‌توان هاپلوتیپ‌های موجود در نمونه‌ی مورد بررسی را تنها با داشتن

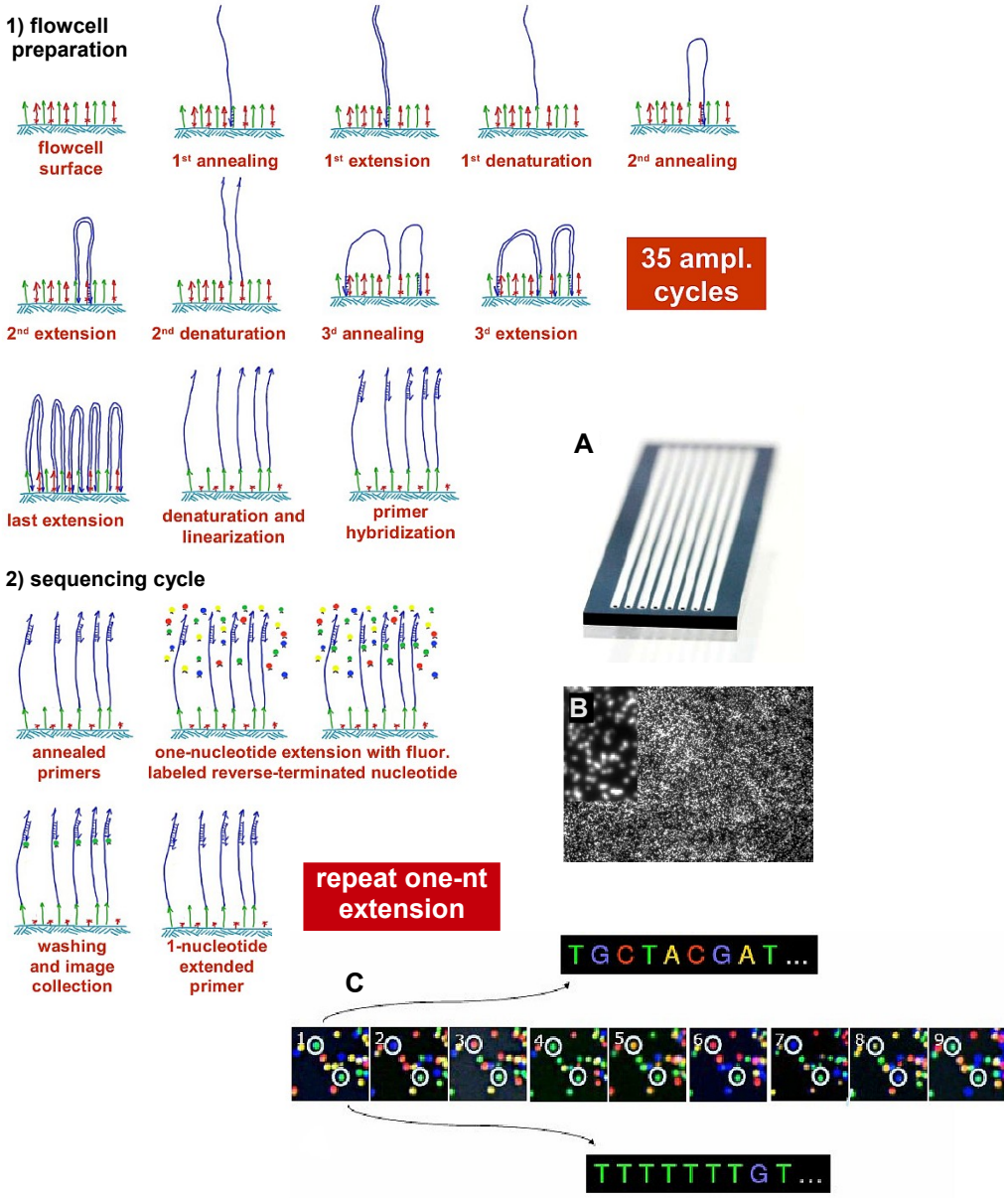
نمونه ژنوتیپ‌های دیپلوئید پیشگویی کرد و حتی تخمین کمابیش دقیقی از فراوانی آنها در جمعیت بدست آورد. بحث کاملتری درباره‌ی این مسئله در بخش ۳۰۱ آمده است.

برهمگذاری و نگاشت

مسئله‌ی برهمگذاری[۹] قطعات توالی‌یابی شده، یکی از مسائل مشترک در بین تمام روشهای توالی‌یابی است. هدف در این مسئله یافتن قطعات توالی‌یابی شده‌ای است که با یکدیگر همپوشانی دارند تا از طریق کنار هم قرار دادن آنها توالی ژنوم به طور کامل بدست آید. هر یک از فنآوری‌های توالی‌یابی، متناسب با روال عملیاتی خاص خود، رهیافت‌های متفاوتی برای حل این مسئله در پیش می‌گیرند. به عنوان مثال در پروژه‌ی ژنوم انسان و برخی دیگر از پروژه‌های مشابه، در مرحله‌ای پیش از توالی‌یابی کامل نوکلئوتیدها، نقشه‌ای از جایگاه قطعات بزرگ DNA به کمک کروموزوم‌هایی ساختگی که اختصاراً BACنامیده می‌شوند بدست می‌آید. رهیافت آزمایشگاهی دیگری نیز در انتهای پروسه‌ی برهمگذاری ژنوم، از توالی‌های بیان‌شده‌ی برچسب‌گذاری[۱۰] برای تعیین جهت و ترتیب قطعات بزرگ توالی‌یابی شده بر روی ژنوم استفاده می‌کند. به جز رهیافت‌های آزمایشگاهی معدودی از این دست که تا حدودی به تعیین نقشه‌ی قرارگیری نوکلئوتیدها بر روی ژنوم کمک می‌کنند، قسمت اصلی مسئله‌ی برهمگذاری با توسل به رویکردهای الگوریتمی مورد بررسی قرار می‌گیرد. بسته به اینکه برهمگذاری تنها بر پایه‌ی همپوشانی‌های قطعات توالی‌یابی شده و بدون در دست داشتن هیچ ژنوم توالی‌یابی شده‌ای از پیش صورت گیرد یا برعکس یک ژنوم از پیش توالی‌یابی شده به عنوان ژنوم مرجع در دسترس باشد مسئله به دو شکل مختلف تعریف می‌شود. به طور بدیهی، پروژه‌های ژنوم انسان و ژنوم دیگر جانداران با مسئله‌ای از نوع اول مواجه هستند. این مسئله را برهمگذاری توالی‌ها از مبنا[۱۱] می‌گویند. دو الگوریتم overlap-layout-consensus[۲۴—۲۶] و Euler[۲۷] از جمله رایجترین الگوریتم‌هایی هستند که با رویکردهایی متفاوت به حل این مسئله می‌پردازند.

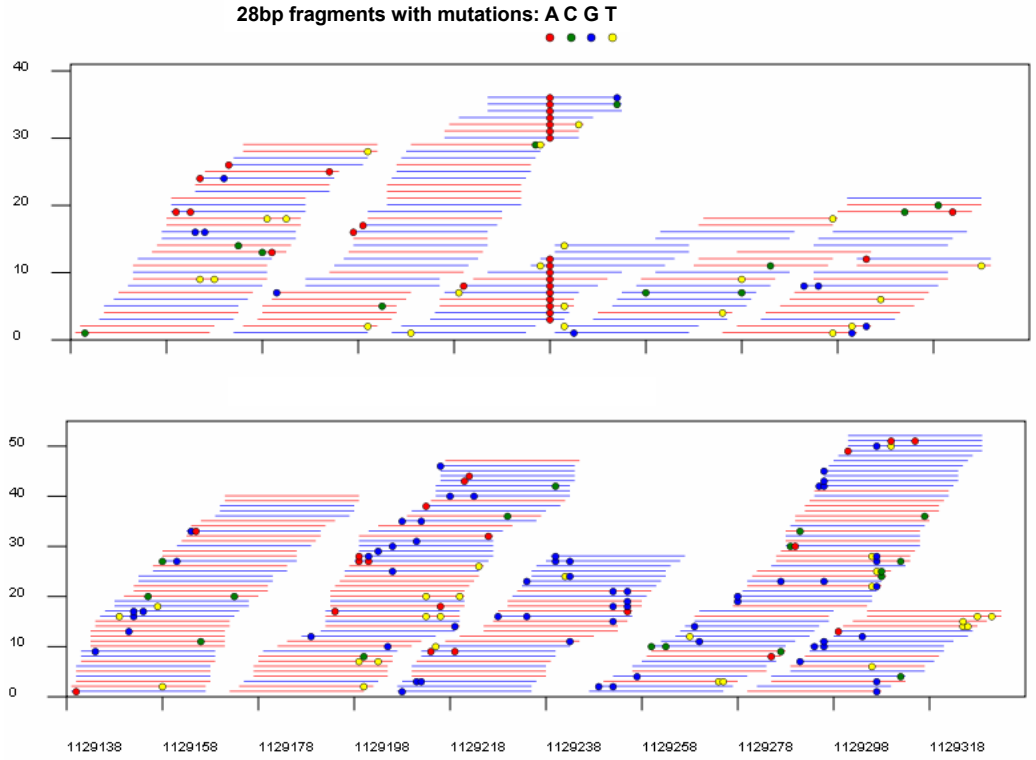
اکنون با اتمام پروژه‌ی ژنوم انسان، یک توالی توافق‌شده به عنوان ژنوم مرجع در دسترس است و بدین ترتیب مسئله‌ی برهمگذاری در بسیاری از کاربردها، از جمله تشخیص و تعیین ژنوتیپ اسنیپ‌ها، به مسئله‌ی

[۹] Assembly
[۱۰] Expressed Tagging Sequence, ETS
[۱۱] *de novo* sequence assembly



شکل ۵۰۱: فرایند آماده‌سازی و خواندن توالی‌های نوکلئوتیدی در فنآوری Illumina/Solexa
۱) مراحل آماده‌سازی نمونه بر روی flowcell. طی این فرآیند هزاران توالی یکسان از هر یک از قطعه توالی‌های کوتاه موجود در نمونه، الگوبرداری می‌شوند. با این شیوه توده‌ای از توالی‌های یکسان، پیرامون یک نقطه‌ی تصادفی بر روی flowcell بوجود می‌آید. ۲) خواندن توالی. در هر تکرار دقیقا یک نوکلئوتید همراه با عنصر فلورسانس به انتهای زنجیر الگوبرداری‌شده اضافه می‌شود. **A**) تصویری از تراشه‌ی مورد استفاده در ماشین توالی‌یاب. بر روی تراشه، هشت باند وجود دارد که هر باند شامل ۲۰۰ flowcell است. **B**) تصویر بزرگنمایی‌شده‌ای از یک flowcell. نقاط روشن نشان‌دهنده‌ی وجود عامل فلورسانسی برانگیخته شده با یک طول موج معین هستند. **C**) با اضافه شدن هر نوکلئوتید به رشته‌ی الگوبرداری‌شده، نقاط مختلف بر روی flowcell با رنگ‌های متفاوتی روشن می‌شوند.

ساده‌تر نگاشت[۱۲] قطعات توالی‌یابی شده بر روی ژنوم مرجع تقلیل می‌یابد. بر پایه‌ی چنین رویکردی می‌توان طیف گسترده‌ای از انواع چندریختی‌های ژنومی و از جمله ژنوتیپ مرتبط با اسنیپ‌ها را توسط فنآوری‌های نسل جدید بازتوالی‌یابی شناسائی کرد. به عنوان مثال، ماشین Illumina-GA ژنوتیپ نمونه‌ی مورد بررسی را به صورت قطعات ۳۶ نوکلئوتیدی توالی‌یابی می‌کند که پس از تطابق توالی‌های مشابه موجود در نمونه با یکدیگر و نگاشت آنها بر روی ژنوم مرجع می‌توان مکان و ژنوتیپ اسنیپ‌های نمونه‌ی مورد بررسی را بدست آورد (شکل ۶۰۱). حجم قابل توجهی از ژنوتیپ‌های تعیین توالی‌شده در پروژه‌های HapMap و 1000 Genomes با استفاده از فنآورهای نسل جدید بازتوالی‌یابی و همکاری شرکت‌های تولیدکننده‌ی آنها بدست آمده است.



شکل ۶۰۱: نگاشت قطعات کوتاه توالی‌یابی شده بر روی ژنوم مرجع و شناسائی جایگاه‌های اسنیپ
قطعات ۲۸ نوکلئوتیدی از ژنوم تجلی‌یافته در دو سویه‌ی مختلف موش که توسط ماشین Illumina/Solexa-GA توالی‌یابی شده‌اند در نقاط مختلف با توالی منطقه‌ی خاصی از ژنوم مرجع تطابق دارند. بازهایی که با توالی مرجع تطابق ندارند می‌توانند نشان‌دهنده‌ی یک اسنیپ در نمونه‌ی مورد بررسی باشند.

در بیشتر ماشین‌های نسل جدید بازتوالی‌یابی، ژنوم از طریق برهمگذاری تعداد بسیار زیادی از توالی‌های

[۱۲] Mapping

بسیار کوتاه بدست می‌آید. از این رو تعدادی از الگوریتم‌های برهمگذاری از مبنا، به طور خاص برای کار بر روی توالی‌های کوتاه توسعه داده شده‌اند که از آن جمله می‌توان به Velvet[۲۸]، SSAKE[۲۹] و SHARCGS[۳۰] اشاره کرد. با این حال رویکرد ذاتی مورد توجه در فنآوریهای بازتوالی‌یابی، نگاشت توالی‌های کوتاه بر روی ژنوم مرجع به جای استفاده از برهمگذاری از مبنا است. از جمله نرم‌افزارهایی که برای نگاشت توالی‌های کوتاه توسعه یافته‌اند می‌توان به ELAND بسته‌ی اختصاصی ماشین‌های توالی‌یاب Illumina، Maq، SHRiMP[۳۱]، SOAP[۳۲] و BLAT[۳۳] اشاره کرد.

دو شاخص، حوزه‌ی کاربرد هر یک از این نرم‌افزارها را محدود می‌کنند: حداکثر طول قطعات بازتوالی‌یابی‌شده و نحوه‌ی تعریف تطابق بین قطعه‌ی بازتوالی‌یابی‌شده و ژنوم مرجع. کارآیی برخی از نرم‌افزارهای نگاشت تنها به نگاشت توالی‌های کوتاه، یعنی توالی‌هایی که طولشان کمتر از ۳۶ باز است محدود می‌شود که نرم‌افزار ELAND نمونه‌ای از این دست است. برعکس نرم‌افزاری مثل BLAT به طور ذاتی توانایی نگاشت قطعاتی به طول چندین مگاباز را بر روی ژنوم‌های متختلف داراست. تطابق بین قطعات توالی‌یابی‌شده و ژنوم مرجع نیز در بین نرم‌افزارهای نگاشت به صورتهای مختلفی ارزیابی می‌شود. برخی الگوریتم‌ها فقط مکان‌هایی را از ژنوم مرجع مناسب برای نگاشت یک توالی کوتاه به شمار می‌آورند که حداکثر در دو نوکلئوتید با توالی کوتاه اختلاف داشته باشند. در این شرایط اگر تفاوت بین توالی کوتاه و ژنوم مرجع به دلیل حذف یا درج یک نوکلئوتید یا بیشتر بوجود آمده باشد، توالی کوتاه غیرقابل نگاشت گزارش می‌شود. این رویکرد در ELAND مورد استفاده قرار می‌گیرد. از سویی دیگر، می‌توان برای تعیین انطباق بین توالی کوتاه و ژنوم مرجع از الگوریتم موسوم به تطابق موضعی[۱۳] با پارامترهای دلخواه استفاده کرد. در این حالت، تعداد بازهای متفاوت بین دو رشته و طول شکاف‌ها یعنی موقعیت‌هایی که به دلیل رویداد درج یا حذف در تطابقِ بین دو رشته مشاهده می‌شوند به طور بهینه توسط یک الگوریتم مبتنی بر برنامه‌ریزی پویا بدست می‌آیند. هرچند به نظر می‌آید نتایج قابل اعتمادتری توسط شیوه‌ی دوم یعنی نگاشت توالی‌های کوتاه با استفاده از تطابق موضعی بدست می‌آید اما نباید فراموش کرد که محاسبه‌ی تطابق موضعی برای چند میلیارد رشته‌ی کوتاه بازتوالی‌یابی شده در برابر توالی‌هایی با تعدادی از همین مقیاس از نقاط مختلف ژنوم مرجع زمان بسیار زیادی را صرف می‌کند و از این رو در بسیاری موارد یک راه حل عملی به شمار نمی‌آید. در عوض، با محدود کردن تطابق‌ها به تطابق‌هایی

[۱۳] Local alignment

با حداکثر دو یا سه نوکلئوتید اختلاف، می‌توان از الگوریتم‌های مبتنی بر دَرهم‌سازی[۱۴] و جستجوی الگو[۱۵] استفاده کرد و از این طریق زمان اجرای الگوریتم نگاشت را به طور چشمگیری کاهش داد. از این رو، عمده‌ی نرم‌افزارهای نگاشت مانند SOAP و SHRiMP از الگوریتم‌هایی مرکب از هر دو ایده بهره می‌گیرند. بر خلاف روش‌های برهمگذاری از مبنا، تعداد قطعات توالی‌یابی‌شده و طول کل ژنوم مورد بررسی غالباً موضوع مشکل‌آفرینی در الگوریتم‌های نگاشت نیست.

۳۰۱ استنباط هاپلوتیپ‌ها با استفاده از داده‌های ژنوتیپ

توالی ژنومِ هر فرد انسان از مجموعه‌ای از توالی‌های دو به دو جفت شده در کنار یکدیگر تشکیل می‌گردد. در واقع، در هر جاندار دیپلوئید مانند انسان، اطلاعات ژنتیکی در مجموعه‌ای از جفت کروموزومها نگهداری می‌شوند که از هر جفت یک نماینده به نسل بعد منتقل می‌شود. توالی اسنیپ‌ها بر روی جفت کروموزوم را ژنوتیپ و بر روی یک کروموزوم منفرد را هاپلوتیپ فرد می‌نامیم.

در بخش ۲۰۱ به برخی فنآوری‌های جدید توالی‌یابی اشاره شد. تعیین ژنوتیپ به کمک این فنآوری‌ها به سهولت و با هزینه‌ی به نسبت پایین امکانپذیر است و به همین جهت پروتکل‌های خاصی برای تعیین ژنوتیپ‌ها رواج یافته‌اند. این در حالی است که نمونه‌گیری و استخراج هاپلوتیپ افراد توسط شیوه‌های آزمایشگاهی در مقایسه با شناسائی ژنوتیپ‌ها هزینه‌ی به مراتب بیشتری دارد. شیوه‌های آزمایشگاهی که تاکنون برای تعیین مستقیم هاپلوتیپ‌ها توسعه یافته‌اند مثل Long-range allele specific PCR[۳۴]، single-molecule dilution[۳۵]، dipliod-to-haploid conversion[۳۶]، Carbon nanotube probing[۳۷]، pyrosequencing[۳۸]، intracellular ligation[۳۹]، rolling-circle amplification[۴۰] و یا clone-based systematic haplotyping[۴۱] جملگی به موارد تحقیقاتی خاص خود محدود می‌شوند و از این رو به طور کاربردی فراگیر نشده‌اند. در عوض، بر پایه‌ی برخی مدلهای ریاضی درباره‌ی پیدایش هاپلوتیپ‌ها و تنوع آنها در جمعیت و تنها با استفاده از شیوه‌های محاسباتی می‌توان هاپلوتیپ‌های نمونه‌های مورد بررسی را با داشتن ژنوتیپ آنها بدست آورد. برای درک بهتر رابطه‌ی بین ژنوتیپ‌ها و هاپلوتیپ‌ها، در

[۱۴] Hashing
[۱۵] Pattern matching

ادامه به تعریف برخی اصطلاحات می‌پردازیم و پس از آن صورت کلی مسئله‌ی استنباط هاپلوتیپ‌ها و اشکال خاص آن مسئله به بیان ریاضی طرح می‌گردند.

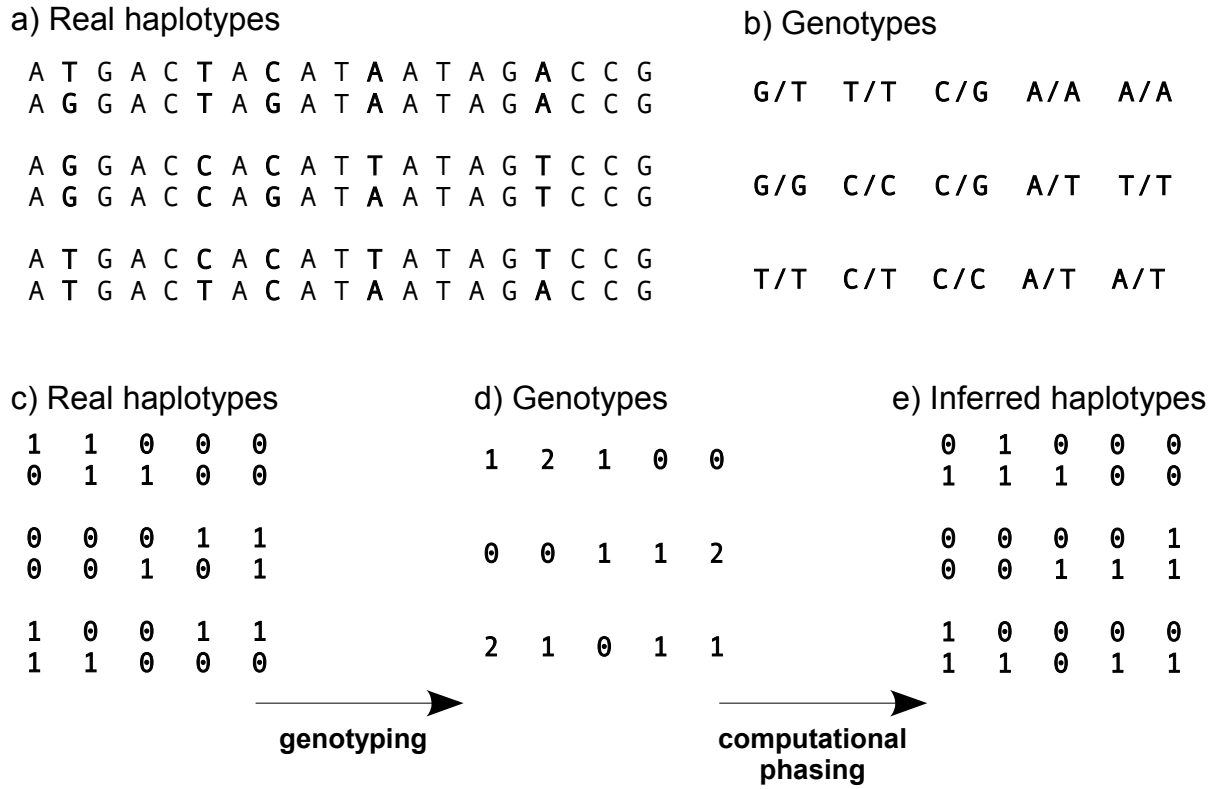
هر هاپلوتیپ را می‌توان همانند یک رشته از حروف که هر یک نمایانگر یک نوکلئوتید در یک موقعیت اسنیپ است در نظر گرفت. هر ژنوتیپ اما، توالیی از زوج نوکلئوتیدها در همان اسنیپ‌هاست به طوری که هیچ اطلاعی درباره‌ی دو هاپلوتیپ تشکیل‌دهنده‌ی این ژنوتیپ نداریم. به بیان ریاضی، یک ژنوتیپ یک توالی از زوج‌های بدون ترتیب است. به عنوان مثال ژنوتیپ آخر در شکل ۷۰۱ می‌تواند ترکیبی از هاپلوتیپ‌های TTCAT و TCCTA یا هاپلوتیپ‌های TTCAA و TCCTT باشد. با توجه به اینکه اسنیپ‌ها عموما دو آللی هستند و با فرض اینکه نوع نوکلئوتید هر آلل در هر اسنیپ مستقل از دیگر اطلاعات، از پیش مشخص است وضعیت هر آلل در هر اسنیپ را می‌توان به طور کلی، توسط اعداد 0و 1نشان داد. به طور متعارف، 0نماینده‌ی آلل فراوانتر و 1نماینده‌ی آلل نادرتر است[۱۶]. بدین ترتیب هر هاپلوتیپ به صورت برداری مثل $h = \langle a_1, \ldots, a_l \rangle$ نشان داده می‌شود که در آن هر a_i یا 0است یا 1و l تعداد اسنیپ‌هاست. به طور مشابه، وضعیت هر اسنیپ در یک ژنوتیپ را می‌توان با سه مقدار 0، 1و 2نمایش داد که در آن 0نمایانگر ژنوتیپ هموزیگوت از آلل فراوان یا زوج $\langle \mathbf{0}, \mathbf{0} \rangle$، 1نمایانگر ژنوتیپ هتروزیگوت یا زوج بدون ترتیب $\langle \mathbf{0}, \mathbf{1} \rangle$ و 2نمایانگر ژنوتیپ هموزیگوت از آلل نادر یا زوج $\langle \mathbf{1}, \mathbf{1} \rangle$ است. برای هر ژنوتیپ، هاپلوتیپ‌هایی که آللی یکسان با آلل اسنیپ‌های هموزیگوت در آن ژنوتیپ داشته باشند اصطلاحا سازگار[۱۷] با آن ژنوتیپ نامیده می‌شوند. وقتی یک ژنوتیپ مثل g از ترکیب دو هاپلوتیپ مثل h_1 و h_2 بدست آمده باشد می‌نویسیم $g = h_1 \oplus h_2$. در واقع، زوج هاپلوتیپ تشکیل‌دهنده‌ی یک ژنوتیپ، هاپلوتیپ‌هایی هستند که با ژنوتیپ مورد نظر سازگاری دارند و به علاوه هر کدام آللی متفاوت با دیگری در جایگاه‌های هتروزیگوت آن ژنوتیپ دارند. بدیهی است که برای یک ژنوتیپ با k جایگاه هتروزیگوت، 2^{k-1} زوج هاپلوتیپ متفاوت می‌توان معرفی کرد که ترکیبشان ژنوتیپ مورد نظر را پدید آورد. از این میان اگر یک زوج هاپلوتیپ معین به عنوان هاپلوتیپ‌های تشکیل‌دهنده‌ی ژنوتیپ تعیین گردند می‌گوییم ژنوتیپ تعیین فاز[۱۸] شده است یا اصطلاحا تفکیک[۱۹] شده است.

[۱۶] در سراسر این رساله، اعداد متناظر با آلل‌ها، با ارقام لاتین نمایش داده می‌شوند.

[۱۷] Compatible

[۱۸] phase

[۱۹] resolve



شکل ۷۰۱: نمونه‌هایی از ژنوتیپ‌ها و هاپلوتیپ‌های تشکیل‌دهنده‌ی هر یک از آنها
a) توالی قسمتی از کروموزوم‌ها در سه فرد نمونه. **b)** ژنوتیپ اسنیپ‌ها در همان افراد. **c)** هاپلوتیپ‌های واقعی درون نمونه. اینجا آلل‌ها با 0و 1نمایش داده شده‌اند. **d)** همان ژنوتیپ‌های بخش b. این بار، وضعیت اسنیپ‌ها با اعداد قراردادی نمایش داده شده‌اند. **e)** هاپلوتیپ‌های استنباط شده توسط یک روش محاسباتی برای تعیین فاز ژنوتیپ‌ها.

اگر در یک ژنوتیپ، بیش از یک موقعیت هتروزیگوت وجود داشته باشد تفکیک ژنوتیپ با ابهام روبرو می‌شود. در این صورت، لازم است تا قیود بیشتری در مورد ساختار هاپلوتیپ‌ها وارد مسئله گردند تا بتوان در مورد تعیین فاز ژنوتیپ تصمیم‌گیری کرد. بدون در نظر گرفتن چنین قیودی مسئله‌ی تعیین فاز ژنوتیپ‌ها در شکل کلی به صورت زیر طرح می‌شود:

مسئله‌ی استنباط هاپلوتیپ‌ها و تعیین فاز ژنوتیپ

فرض کنید نمونه‌ای مثل $G = \{g_1, \ldots, g_n\}$ شامل n ژنوتیپ بر روی l اسنیپ داده شده است. می‌خواهیم مجموعه‌ای از هاپلوتیپ‌ها مثل H را تعیین کنیم به قسمی که برای هر $g_i \in G$، دو هاپلوتیپ $h_a, h_b \in H$ وجود داشته باشند که $g_i = h_a \oplus h_b$. به عنوان مثال هر یک از مجموعه هاپلوتیپ‌ها در شکل ۷۰۱ قسمت c و قسمت e می‌توانند جوابهایی برای مسئله‌ی تعیین فاز به ازای ژنوتیپ‌های داده شده در قسمت d باشند.

قیودی که برای تعیین یک جواب یا دسته‌ای از جوابهای مطلوب بر مجموعه جوابهای ممکن اعمال می‌شوند معمولا دائر بر ارضای برخی ملاحظات رایج در ژنتیک و تکامل و مرتبط با ساختار و تنوع هاپلوتیپ‌های جواب است. عمده‌ی مدل‌هایی که تاکنون برای حل این مسئله ارائه شده‌اند را می‌توان در یکی از چهار دسته‌ی کلی زیر رده‌بندی کرد.

۱. بیشترین پارسیمونی[۲۰]. تعداد هاپلوتیپ‌های متمایز در جواب کمترین است. به بیان متفاوت، هر هاپلوتیپ در مجموعه‌ی جواب، در تفکیک تعداد بیشتری از ژنوتیپ‌های نمونه شرکت دارد.

۲. فیلوژنی کامل[۲۱]. مجموعه‌ی هاپلوتیپ‌های جواب، یک درخت فیلوژنی کامل تشکیل می‌دهند. در این درخت هر اسنیپ دقیقا متناظر یک گره درخت است و هر هاپلوتیپ منحصرا بر اثر رویدادهای جهش از یک هاپلوتیپ اجدادی واحد پدید آمده‌اند.

۳. بیشترین درستنمائی[۲۲]. توزیع احتمال هاپلوتیپ‌ها در جمعیت به قسمی تعیین می‌شود که تابع درستنمائی مشاهده‌ی ژنوتیپ‌های داده شده بیشینه شود.

۴. استنتاج بیزی[۲۳]. فراوانی هاپلوتیپ‌ها با محاسبه‌ی احتمال پسین برخی مدلهای آماری بدست می‌آید.

با در نظر گرفتن قیود مطلوب در هر یک از مدلهای فوق، مسئله‌ی استنباط هاپلوتیپ‌ها به شکل متفاوتی بیان می‌شود. این مسائل را در ادامه با جزئیات بیشتری می‌بینیم.

مدل بیشترین پارسیمونی

دیدگاه بیشترین پارسیمونی بر این باور استوار است که اصولا تنوع هاپلوتیپ‌ها در طبیعت بسیار کمتر از تنوع ژنوتیپ‌ها است. به عبارت دیگر، ملاحظات تکاملی ایجاب می‌کند که گونه‌های هاپلوتیپ بسیار خاص باشند، در حالی که چنین قیودی خیلی کمتر بر تنوع ژنوتیپی تاثیر می‌گذارند. بدین ترتیب، در مسئله‌ی بیشترین

[۲۰] Parsimony
[۲۱] Perfect phylogeny
[۲۲] Maximum likelihood
[۲۳] Bayesian inferrence

پارسیمونی به دنبال مجموعه‌ای از هاپلوتیپ‌ها برای تفکیک ژنوتیپ‌های داده شده می‌گردیم که با کمترین تعداد هاپلوتیپ‌های متمایز بتوان ژنوتیپ‌های داده شده را بازسازی کرد.

اولین رویکرد برای حل مسئله‌ی استنباط هاپلوتیپ‌ها، روشی را برای حل مسئله‌ی بیشترین پارسیمونی دنبال می‌کرد. این روش، الگوریتم منسوب به کلارک[۴۲] است که در آن از یک رهیافت سودجویانه[۲۴] برای تفکیک ژنوتیپ‌ها استفاده می‌شود. در الگوریتم کلارک، ابتدا تمام ژنوتیپ‌هایی که حداکثر یک جایگاه هتروزیگوت دارند بدون ابهام به هاپلوتیپ‌های متناظرشان تفکیک می‌شوند. سپس با شروع از این مجموعه‌ی اولیه از هاپلوتیپ‌ها، هر کدام از ژنوتیپ‌های باقی مانده که با یکی از هاپلوتیپ‌های تاکنون استنتاج شده سازگار باشند تفکیک می‌شوند و هاپلوتیپ مکمل در صورتی که در مجموعه‌ی فعلی وجود نداشته باشد به آن اضافه می‌شود. علیرغم برخی ایرادات جدی که در رهیافت سودجویانه‌ی به کار رفته در شیوه‌ی کلارک وجود دارد ولی در بسیاری از نمونه‌های واقعی، نتایج قابل قبولی بدست می‌دهد[۴۳].

برخی محدودیتهای این شیوه عبارتند از:

- دست کم یک ژنوتیپ بدون ابهام باید در اختیار باشد.
- ممکن است در پایان روال، برخی ژنوتیپ‌ها تفکیک نشده باقی بمانند.
- ترتیب انتخاب ژنوتیپ‌ها برای تفکیک و نیز انتخاب یک هاپلوتیپ از بین تمام هاپلوتیپ‌های فعلی سازگار با این ژنوتیپ، در پایان می‌تواند به نتایج متفاوتی برسد.

در مسائل عملی مشکل اول به ندرت پیش می‌آید. در مورد مشکل دوم، صورت خاصی از مسئله طرح شده است که در آن هدف یافتن ترتیب خاصی از مراحل الگوریتم کلارک است که در پایان بیشترین تعداد ممکن از ژنوتیپ‌ها را تفکیک کند. هابل این مسئله را بررسی کرده است و با تحلیل دادن مسئله‌ی صدق‌پذیری به این مسئله، نشان داده است که این شکل از مسئله یک مسئله‌ی NP-hard است[۴۴]. همانجا راه‌حلی مبتنی بر برنامه‌ریزی خطی برای حل تقریبی این حالت تحدید شده‌ی مسئله ارائه شده است که در عمل جواب‌های قابل پذیرشی بدست می‌آورد. هرچند در حالت کلی ممکن است نتواند تمام ژنوتیپ‌ها را تفکیک کند.

[۲۴] greedy

جدا از شکل خاص مسئله‌ی بیشترین پارسیمونی که در بالا به آن اشاره شده، این مسئله در حالت کلی نیز یک مسئله‌ی NP-hard است [۴۵]. با این حال، برای حالتی که تعداد جایگاه‌های هتروزیگوت در تمام ژنوتیپ‌های داده شده کمتر از سه باشد یک راه‌حل چندجمله‌ای وجود دارد[۴۶]. از دیگر سو، روشهای متعددی با رهیافت‌های اکتشافی برای حل تقریبی این مسئله ارائه شده است. در واقع نه تنها الگوریتم کلارک بلکه سایر روشهای حل این مسئله، در حالت کلی تضمینی برای رسیدن به کمترین تعداد هاپلوتیپ برای تفکیک ژنوتیپ‌های داده شده ندارد. گاسفیلد با تبدیل این مسئله به یک برنامه‌ریزی خطی، الگوریتمی کارآمد برای داده‌هایی با اندازه‌ی متوسط ارائه کرده است[۴۷]. این برنامه‌ریزی خطی برای حل مسئله‌ی بیشترین پارسیمونی به شکل زیر صورت‌بندی می‌شود:

$$
\begin{aligned}
&\min \sum_{h \in \mathcal{H}} x_h \\
s.t. \quad &\sum_{p \in \mathcal{P}_g} y_{gp} = ۱, \quad for \quad g \in G \\
&y_{gp} \leq x_h, \quad for \quad g \in G \quad and \quad p = \{h, h'\} \in \mathcal{P}_g \\
&x_h \in \{۰, ۱\}, \quad for \quad h \in \mathcal{H} \\
&y_{gp} \in \{۰, ۱\}, \quad for \quad g \in G \quad and \quad p \in \mathcal{P}_g
\end{aligned}
$$

که در آن G مجموعه ژنوتیپ‌های داده شده، $\mathcal{H}$ مجموعه‌ی تمام هاپلوتیپ‌هایی است که هر یک دست کم با یکی از ژنوتیپ‌های G سازگار هستند، $x_h = ۱$ نشانگر انتخاب هاپلوتیپ h در مجموعه‌ی جواب، $\mathcal{P}_g$ مجموعه‌ی تمام زوج هاپلوتیپ‌هایی است که ژنوتیپ g را می‌توان به آنها تفکیک کرد و $y_{gp} = ۱$ نشانگر انتخاب زوج هاپلوتیپ $p = \{h, h'\}$ به عنوان فازهای ژنوتیپ g در مجموعه‌ی جواب است. در عمل، به کمک بسته‌های نرم‌افزاری رایج برای حل مسائل برنامه‌ریزی عدد صحیح، نمونه‌های متعددی از مسئله‌ی بیشترین پارسیمونی توسط این شیوه حل می‌شوند. متاسفانه در این رویکرد، تعداد معادلات وارد در برنامه‌ریزی خطی نسبت به تعداد جایگاه‌های هتروزیگوت مرتبه‌ی نمایی دارد که مانع کارآیی این شیوه در حالت کلی می‌شود. وانگ با پیاده‌سازی یک روش توسعه و تحدید[۲۵] به مقایسه‌ی نتایج دقیق با نتایج بدست آمده از برخی ایده‌های بهبود دهنده‌ی الگوریتم کلارک پرداخته است[۴۸]. مزیت روشهای ارائه شده در [۴۹، ۵۰]،

[۲۵] branch and bound

در مقایسه با دو روش قبل در ارائه‌ی فرمول‌بندی مسئله در فضای حافظه‌ای با پیچیدگی چندجمله‌ای است.

نگرش بیشترین پارسیمونی در مسئله‌ی استنباط هاپلوتیپ‌ها، با آنکه اولین صورت‌بندی این مسئله بوده است و در بین مدل‌های دیگر، بیشترین توجه را به خود معطوف ساخته است اما مقایسه‌ی هاپلوتیپ‌های استنباط شده از طریق حل دقیق این مسئله با هاپلوتیپ‌های واقعی نشان می‌دهد که طبیعت لزوما با بیشترین پارسیمونی رفتار نمی‌کند. کلایمر و دیگران[۵۱] نشان داده‌اند که جواب‌های بیشترین پارسیمونی نه تنها لزوما با هاپلوتیپ‌های واقعی یکسان نیستند بلکه برای بسیاری از نمونه‌ها، مجموعه‌های متعددی از هاپلوتیپ‌ها مقدار بهینه برای تابع هدف مدل بیشترین پارسیمونی را بدست می‌آورند.

مدل فیلوژنی کامل

در بین ایده‌های متفاوتی که برای انتخاب مجموعه‌ی مطلوب هاپلوتیپ‌ها در مسئله‌ی تعیین فاز ژنوتیپ‌ها بکار گرفته می‌شوند، فرضیه‌ی «نیای مشترک[۲۶]» یک رویکرد طبیعی برای تبیین پدید آمدن هاپلوتیپ‌های متفاوت از جنبه‌ی تکاملی است. این فرضیه معتقد بر این است که تمام گونه‌های هاپلوتیپ، در جمعیت زمان حاضر، از یک نیای مشترک منشاء می‌گیرند و تنها عاملی که باعث پدید آمدن تمایز در بین آنها شده‌است رویدادهای جهشی بوده است که بر روی ژنوم در موقعیت کنونی اسنیپ‌ها رخ داده است. به علاوه بر اساس این فرضیه، بر روی هر موقعیت اسنیپ حداکثر یکبار امکان رویداد جهش وجود داشته است[۵۲].

اصطلاح «فیلوژنی کامل» بیان فرضیه‌ی نیای مشترک با عبارات ریاضی است[۵۳]. فرض کنید $H = \{h_1, \ldots, h_{2n}\}$ مجموعه‌ای از $2n$ هاپلوتیپ باشد که هر یک از هاپلوتیپ‌ها از l اسنیپ تشکیل شده‌اند. یک فیلوژنی کامل، یک درخت ریشه‌دار مثل T با $2n$ برگ (گره پایانی) است که در چهار ویژگی زیر صدق می‌کند:

۱. هر هاپلوتیپ از H دقیقا با یک گره پایانی درخت T تناظر داده می‌شود.

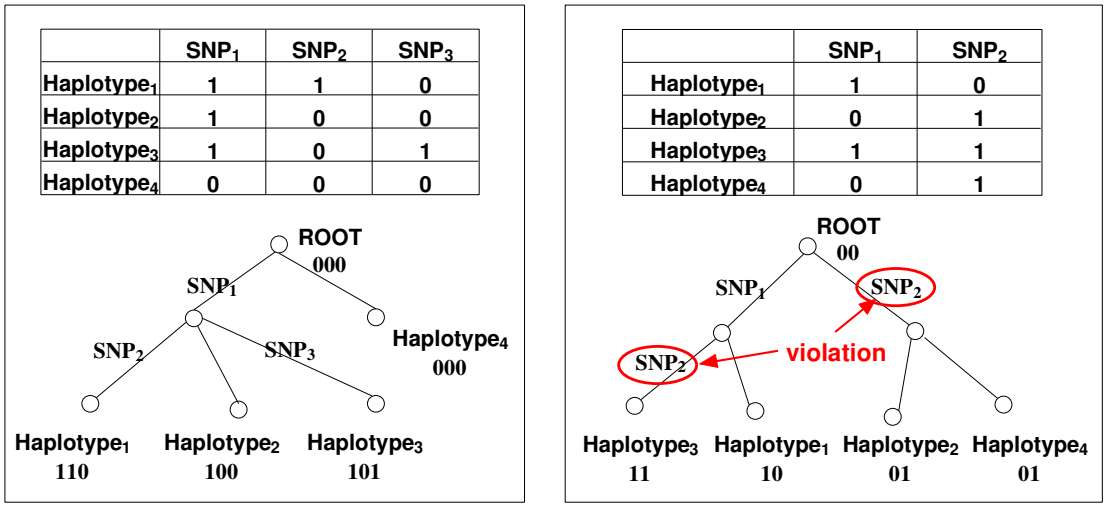
۲. هر یال در T با مجموعه‌ای از یک یا چند اسنیپ علامت‌گذاری می‌شود، به جز یال‌های متصل به گره‌های پایانی که ممکن است با هیچ اسنیپی علامت‌گذاری نشوند.

[۲۶] coalescent theory

۳. هر اسنیپ دقیقا یکبار برای علامت‌گذاری در کل درخت به کار می‌رود.

۴. هاپلوتیپ متناظر با هر گره‌ی پایانی توسط مسیر ریشه به برگ تعیین می‌شود. بدین ترتیب که با شروع از هاپلوتیپ ریشه، پس از عبور از هر یال به طرف پائین، آلل اسنیپهای مرتبط با آن یال را تغییر می‌دهیم.

در شکل ۸۰۱ نمونه‌ای از یک مجموعه هاپلوتیپ با ویژگی فیلوژنی کامل و نمونه‌ی دیگری که نمی‌توان آنرا در شرایط فیلوژنی کامل صدق داد نمایش داده شده‌اند. با فرض اینکه هاپلوتیپ ریشه به صورت $\underbrace{00\ldots0}_{l}$ است



a) Perfect Phylogeny **b) No Perfect Phylogeny**

شکل ۸۰۱: درخت فیلوژنی کامل

a) وجود ساختار درخت فیلوژنی کامل مرتبط با هاپلوتیپ‌ها نشان می‌دهد که این مجموعه در شرایط فیلوژنی کامل صدق می‌کند. **b)** در اینجا، هیچ درخت فیلوژنی کاملی نمی‌توان برای مجموعه هاپلوتیپ‌ها یافت. در درخت نمایش داده شده، رویداد جهش در جایگاه SNP_2 دو بار رخ داده است.

می‌توان ثابت کرد که شرط لازم و کافی برای صدق شرایط یک فیلوژنی کامل توسط هاپلوتیپ‌های مجموعه‌ی H آن است که برای هر جفت دلخواه از اسنیپ‌ها هیچ سه هاپلوتیپی نتوان در H یافت که ترکیب این دو اسنیپ در آنها به صورت $\{(0,1),(1,0),(1,1)\}$ باشد[۵۰]. به عبارت دیگر، در فیلوژنی کامل، با تحدید هاپلوتیپ‌ها به هر دو اسنیپ دلخواه، حداکثر سه هاپلوتیپ متمایز مشاهده می‌شود. این ویژگی به شکلی دیگر در بخش ۴۰۱ تحت عنوان آزمون چهار گامتی[۲۷] مورد استفاده قرار می‌گیرد. روشهای استنباط هاپلوتیپ‌ها تحت قیود فیلوژنی کامل در [۵۴—۵۶] تشریح شده‌اند. پیچیدگی محاسباتی آنها همگی از $O(nl^2)$ است. راه

[۲۷] Four gamete test

حل اولیه‌ی ارائه شده توسط گاسفیلد در [۵۳] یک الگوریتم از مرتبه‌ی $O(nl.\alpha(nl))$ است که در مقایسه با روشهای رایج سریع‌تر است[۲۸] ولی جزئیات پیچیده‌ای در پیاده‌سازی آن وجود دارد. الگوریتم کارآمدتر دیگری توسط Ding معرفی شده است که مرتبه‌ی آن نسبت به اندازه‌ی مسئله خطی است[۵۷].

هرچند تاکنون الگوریتم‌های کارآمدی برای حل مسئله‌ی تفکیک ژنوتیپ‌ها بر اساس فیلوژنی کامل در اختیار داریم، با این وجود در بسیاری از مسائل واقعی این روش‌ها نمی‌توانند جوابهای قابل انتظار را بدست آورند که دلیل اصلی آن این است که شرایط فیلوژنی کامل بسیار به ندرت در واقعیت محقق می‌شوند. این امر پیش از آنکه دلیلی بر نادرستی فرضیه‌ی نیای مشترک باشد، برخاسته از خطاهای آزمایشگاهی در تعیین ژنوتیپ‌ها و نیز وقوع رویدادهای نوترکیبی در مناطق ژنومی مورد مطالعه است. در روش اِسکین با حذف تعدادی از ژنوتیپ‌ها یک جواب با ویژگی فیلوژنی کامل برای باقی مانده‌ی ژنوتیپ‌ها بدست می‌آید[۵۸]. در این شکل از مسئله، مطلوب آن است که تعداد ژنوتیپ‌های حذف شده برای رسیدن به شرایط فیلوژنی کامل کمینه باشد که البته چنین مسئله‌ای NP-hard است. در شکل دیگری از مسئله‌ی فیلوژنی کامل، آللهای مفقود[۲۹] به گونه‌ای مقداردهی می‌شوند که شرایط فیلوژنی کامل برقرار شود[۵۹]. در هر دو دیدگاه اخیر، فرض بر این است که یک فیلوژنی کامل برای مجموعه ژنوتیپ‌های داده شده وجود دارد ولی به دلیل وجود خطا یا داده‌های مفقود نمی‌توان آن را بدست آورد. در مقابل، نگرش واقعی‌تری وجود دارد که در آن فرض می‌شود لزومی به برقراری شرایط فیلوژنی کامل برای تمام هاپلوتیپ‌های مجموعه جواب وجود ندارد. در این صورت‌بندی از مسئله که «فیلوژنی ناکامل»[۳۰] نامیده می‌شود اجازه داده می‌شود تا در نسبت کوچک معینی از هاپلوتیپ‌ها، رویدادهای نوترکیبی و جهش‌های چندباره رخ داده باشد. روش‌های ترکیبیاتی متعددی تاکنون برای حل حالتهای گوناگون مسئله‌ی فیلوژنی ناکامل ارائه شده است[۶۰، ۶۱].

بر اساس دو مدلی که تا اینجا معرفی شدند، یعنی مدل بیشترین پارسیمونی و مدل فیلوژنی کامل، صورت‌بندی‌های متفاوتی برای مسئله‌ی تعیین فاز ژنوتیپ‌ها طرح شدند که هدف اصلی در آنها تشخیص زوج هاپلوتیپ‌های تشکیل‌دهنده‌ی هر یک از ژنوتیپ‌های داده شده است و از این رو اطلاعی درباره‌ی فراوانی هاپلوتیپ‌ها و توزیع آنها در جمعیت بدست نمی‌دهند. در ادامه‌ی شکل تعمیم‌یافته‌ای از مسئله‌ی تعیین فاز

[۲۸] $\alpha(.)$ معکوس تابع آکرمن است.

[۲۹] missing alleles

[۳۰] Imperfect phylogeny

ژنوتیپ‌ها طرح می‌شود که هرچند هدف اصلی در آن، یافتن توزیع احتمال هاپلوتیپ‌ها در جمعیت است اما به کمک جواب بدست آمده می‌توان مجموعه‌ای از زوج هاپلوتیپ‌های تشکیل‌دهنده‌ی ژنوتیپ‌ها را نیز به نحو مطلوب انتخاب کرد.

برآورد فراوانی هاپلوتیپ‌ها

فرض کنید مجموعه‌ی $G = \{g_1, \ldots, g_n\}$ شامل n ژنوتیپ بر روی l اسنیپ داده شده است و

$$\mathcal{H} = \{h \mid \exists g \in G, h \sim g\}$$

که در آن $h \sim g$ نشاندهنده‌ی رابطه‌ی سازگاری بین هاپلوتیپ h و ژنوتیپ g است. هدف برآورد فراوانی نسبی هر یک از هاپلوتیپ‌های مجموعه‌ی $\mathcal{H}$ با استفاده از نمونه ژنوتیپ‌های داده شده است.

تعریف فوق، یک صورت کلی برای مسئله‌ی برآورد فراوانی هاپلوتیپ‌ها در جمعیت است و مشابه مسئله‌ی تعیین فاز ژنوتیپ‌ها لازم است تا برخی مدل‌ها، برای تعیین جواب مطلوب از بین جواب‌های امکان‌پذیر به کار گرفته شوند. مدل‌های بیشترین درستنمائی و استنباط بیزی دو رویکرد متفاوت برای برآورد فراوانی هاپلوتیپ‌ها هستند.

مدل بیشترین درستنمائی

تابع درستنمائی، احتمال مشاهده‌ی ژنوتیپ‌های نمونه به ازای مقادیر نامعلوم برای فراوانی نسبی هاپلوتیپ‌ها در جمعیت را نشان می‌دهد. در مدل بیشترین درستنمائی، هدف یافتن این مقادیر نامعلوم با بیشینه‌سازی تابع درستنمائی است[۳۱]. تابع درستنمائی در این مسئله به شکل زیر تعریف می‌شود:

$$L = L(\langle G, f\rangle \mid \Theta) = \prod_{i=1}^{n} P(g_i \mid \Theta)^{f_i} = \prod_{i=1}^{n} \left(\sum_{h_a \oplus h_b = g_i} P(h_a, h_b) \right)^{f_i} \qquad (۱۰۱)$$

که در آن $f = \langle f_1, \ldots, f_n\rangle$ فراوانی ژنوتیپ‌های مشاهده شده و $\Theta = \langle \theta_1, \ldots, \theta_{|\mathcal{H}|}\rangle$ بردار احتمال مشاهده‌ی هاپلوتیپ‌های سازگار با ژنوتیپ‌های داده شده است. هدف یافتن مقادیر برای Θ است که مقدار

[۳۱] رویکرد Maximum Likelihood Estimation که به اختصار MLE نامیده می‌شود برای تبیین بسیاری از پدیده‌های آماری به کار گرفته می‌شود.

L را بیشینه کند. در این معادله، رابطه‌ی بین تابع درستنمائی و فراوانی هاپلوتیپ‌ها به طور صریح مشخص نیست. با این حال با در نظر گرفتن برخی فرضیات متعارف در ژنتیک آماری، می‌توان یک رابطه‌ی صریح بدست آورد. تعادل هاردی‌ـ واینبرگ، رابطه‌ی بین احتمال مشاهده‌ی یک ژنوتیپ و فراوانی هاپلوتیپ‌های تشکیل‌دهنده‌ی آنرا تعریف می‌کند.

تعادل هاردی‌ـ واینبرگ[۳۲]

با فرض تصادفی بودن جفت‌یابی‌ها در یک جمعیت، احتمال مشاهده‌ی یک ژنوتیپ برابر است با حاصلضرب فراوانی نسبی هاپلوتیپ‌های تشکیل‌دهنده‌ی آن ژنوتیپ در جمعیت و داریم:

$$P(g^s=\{\mathbf{0},\mathbf{0}\}) = P(s=\mathbf{0})^2, \qquad P(g^s=\{\mathbf{0},\mathbf{1}\}) = 2P(s=\mathbf{0})P(s=\mathbf{1}),$$
$$P(g^s=\{\mathbf{1},\mathbf{1}\}) = P(s=\mathbf{0})^2,$$

که در آن g^s ژنوتیپ مشاهده شده در اسنیپ sام است.

برای یک جایگاه معین، اگر شرایط فوق برای تمام ژنوتیپ‌های موجود در جمعیت برقرار باشد می‌گوئیم جمعیت در تعادل هاردی‌ـ واینبرگ قرار دارد. با فرض HWE، تابع درستنمائی در رابطهٔ (۱۰۱) به صورت زیر بازنویسی می‌شود:

$$L = \prod_{i=1}^{n} \left(\sum_{h_a \oplus h_b = g_i} \theta_a \theta_b \right)^{f_i} \qquad (۲۰۱)$$

حال بر اساس رابطهٔ (۲۰۱) و با استفاده از روش‌های مرسوم آماری برای برآورد پارامتر، مانند EM[۳۳] می‌توان جواب‌هایی برای مسئله‌ی استنباط هاپلوتیپ‌ها بدست آورد. الگوریتم‌های مبتنی بر EM، از تکرارهای متوالی دو گام مستقل تا حصول شرایط همگرائی تشکیل می‌شوند. مراحل این الگوریتم به طور خلاصه، به شرح زیر انجام می‌شوند:

شروع) مقادیر دلخواهی برای احتمال هاپلوتیپ‌ها در نظر بگیر.

تا رسیدن به شرایط همگرائی مراحل زیر را تکرار کن،

[۳۲] Hardy-Weinberg Equilibrium, HWE
[۳۳] Expectation-Maximization

گام E) بر مبنای توزیع احتمال فعلی، برای هر ژنوتیپ g_i برای $i = ۱, \ldots, n$ حساب کن:

$$\lambda_i = \sum_{h_a \oplus h_b = g_i} \theta_a \theta_b$$

گام M) مقدار جدید برای θ_i برای $i = ۱, \ldots, |\mathcal{H}|$ را با رابطه‌ی زیر به هنگام کن:

$$\theta_i = \frac{۱}{۲n} \sum_{g_j \sim h_i} \lambda_j$$

روشها و نتایج بدست آمده در [۶۲–۶۴] نمونه‌ای از اولین کارهای انجام شده برای حل مسئله‌ی تخمین فراوانی هاپلوتیپ‌های جمعیت بر پایه‌ی روش فوق هستند. پیچیدگی محاسباتی تمام روشهای مبتنی بر الگوریتم EM از مرتبه‌ی $O(n۲^k)$ است که در آن، k تعداد بیشترین جایگاه‌های هتروزیگوت در ژنوتیپ‌های نمونه است. نمایی بودن ذاتی مرتبه‌ی این الگوریتم نسبت به تعداد اسنیپ‌ها باعث می‌شود این الگوریتم نتواند برای نمونه‌هایی بر روی ژنوتیپ‌های بزرگ به طور کارآمد اجرا شود. البته برای مقایر کوچک l، مثلا بین ۱۰ تا ۱۲ اسنیپ، اجراهای مکرر این الگوریتم بر روی رایانه‌های امروزی در زمان کوتاهی امکان‌پذیر است.

کِن و دیگران [۶۵] شیوه‌ای به نام «افراز و انعقاد»[۳۴] معرفی کردند که در آن ژنوتیپ‌ها در قطعاتی با طول‌های یکسان، هر یک شامل تعداد معینی از اسنیپ‌ها، مثلا ۱۰ اسنیپ، افراز می‌شوند و الگوریتم EM به طور مستقل بر روی هر یک آنها اجرا می‌شود. سپس در مرحله‌ی انعقاد، هر دو قطعه‌ی کنار هم، یکی در میان، با یکدیگر ترکیب می‌شوند و روال EM بر روی قطعات ترکیب شده تکرار می‌شود با این تفاوت که فضای جستجو به مجموعه هاپلوتیپ‌هایی محدود می‌شود که از ترکیب هاپلوتیپ‌های دارای بیشترین فراوانی در دو قطعه‌ی مجاور بدست می‌آیند. روال افراز و انعقاد تا زمانی که تمام طول ژنوتیپ‌ها پوشش داده شود ادامه پیدا می‌کند. ایده‌های مشابه دیگری بر پایه‌ی این رویکرد توسعه یافتند که عنوان کلی PL-EM را به آنها اطلاق می‌کنیم[۶۶، ۶۷]. مزیت اصلی رویکردهای مبتنی بر PL-EM در این است که اندازه‌ی فضای جستجو در روال EM، محدود و مستقل از اندازه‌ی ژنوتیپ‌ها است و از این رو الگوریتم با سرعت به مراتب بیشتری اجرا می‌شود و نتایجی نزدیک به نتایج بدست آمده از الگوریتم EM بدون استفاده از PLبدست می‌آورد[۶۸–۷۰].

با آنکه فرض برقراری شرایط HWE برای صورت‌بندی تابع درستنمائی در این مسئله ضروری است ولی مشاهدات انجام شده بر روی نمونه‌های شبیه‌سازی شده که دور از شرایط تعادل HW تولید شده‌اند نیز نشان

[۳۴] Partition-Ligation, PL

می‌دهد که الگوریتم های ارائه شده تا حد خوبی نسبت به انحراف از این تعادل نیز تحمل‌پذیر هستند و نتایج نزدیک به واقعیت را بدست می‌آورند[۶۸]. در شرایط LD بالا تا متوسط، اگر تا ۳۰ درصد آللها در کل نمونه مفقود باشند باز هم روش EM نتایج مطلوب را می‌تواند بدست آورد. این در حالی است که خطا در تعیین ژنوتیپ‌ها می‌تواند تاثیر مخرب‌تری در بدست آوردن نتایج درست توسط روش EM داشته باشد[۶۹].

با اینکه امروزه بسیاری از روش‌های رایج در مسئله‌ی استنباط هاپلوتیپ‌ها، مبتنی بر رهیافت EM هستند و عموماً نتایج کم‌خطایی بدست می‌دهند با این حال برخی اشکالات ذاتی را با خود به همراه دارند. اول آنکه همگرایی روش EM به شدت حساس به انتخاب مقادیر اولیه‌ی پارامتر است و به همین خاطر نیز تضمینی در رسیدن به ماکزیمم سراسری توسط این روش وجود ندارد. دیگر اینکه روابط تحلیلی دقیقی برای محاسبه‌ی خطای برآورد وجود ندارد و از این رو برای بررسی سطح اعتبار برآوردهای بدست آمده تنها می‌توان از روشهای عددی که متضمن اجراهای مکرر الگوریتم است استفاده کرد.

مدل بیزی

در مقایسه با روش بیشترین درستنمائی، دیدگاه بیزی از زاویه‌ی متفاوتی به مسئله‌ی برآورد فراوانی هاپلوتیپ‌ها در جمعیت نگاه می‌کند. در نگرش بیزی، پارامترهای مدل یعنی احتمال مشاهده‌ی هاپلوتیپ‌ها در جمعیت Θ، خود به عنوان یک بردار از متغیرهای تصادفی در نظر گرفته می‌شود که از یک توزیع پیشین[۳۵] پیروی می‌کند. برآورد فراوانی هاپلوتیپ‌ها در دیدگاه بیزی از طریق محاسبه‌ی توزیع پسین[۳۶] صورت می‌گیرد. احتمال پسین در حالت کلی از طریق رابطه‌ی زیر بدست می‌آید:

$$P(\Theta \mid G) \propto L(G;\Theta).P(\Theta) \qquad (۳۰۱)$$

در این رابطه $L(G;\Theta)$ تابع درستنمائی است و مشابه رابطهٔ (۲۰۱) تعریف می‌شود و $P(\Theta)$ تابع توزیع احتمال پیشین برای Θ و $P(\Theta \mid G)$ توزیع پسین را نشان می‌دهند. مزیت دیدگاه بیزی در این است که اولا از طریق توزیع پیشین می‌توان برخی پیش‌فرضها و قیود مطلوب برای هاپلوتیپ‌ها را در مدل وارد کرد و دوم آنکه بر خلاف رویکرد MLE که برآورد پارامترهای مدل به صورت نقطه‌ای صورت می‌گیرد در رویکرد بیزی، احتمال

[۳۵] prior distribution
[۳۶] posterior distribution

پسین برای هر انتخاب برای فراوانی نسبی هاپلوتیپ‌ها در دسترس است. متاسفانه با وجود آنکه این رهیافت یک مدل کلی برای تحلیل آماری مسئله‌ی تفکیک هاپلوتیپ‌ها است ولی در عمل، محاسبه‌ی تابع پسین به صورت تحلیلی ناممکن است و استفاده از شیوه‌های عددی نیز به حجم بالای محاسبات نیازمند است[۷۱].

به طور متداول، برخی از شیوه‌های آماری موسوم به MCMC[۳۷] برای برآورد عددی توزیع پسین به کار گرفته می‌شوند. استفن و دیگران با بکارگیری روش Gibbs sampling که یکی از روشهای خانواده‌ی MCMC است و با انتخاب یک توزیع پسین خاص برگرفته از برخی جنبه‌های فرضیه‌ی نیای مشترک، نسخه‌ی اولیه‌ی نرم‌افزار PHASE را که از جمله پرکاربردترین نرم‌افزارهای استنباط هاپلوتیپ‌هاست توسعه دادند[۷۲]. نرم‌افزارهای دیگری چونfastPHASE[۷۳] و Shape-IT[۷۴] نیز بر مبنای مدل بکارگرفته شده در PHASE توسعه یافته‌اند. در تحقیق مستقل دیگری، ژانگ و نیو[۷۵] از تلفیق مدل نیای مشترک و مدل بیزی برای استنباط هاپلوتیپ‌ها استفاده کرده‌اند. در رویکرد دیگری از روش بیزی که توسط نیو و دیگران معرفی گردید از توزیع دیریکله[۳۸] به عنوان توزیع پیشین و روال افراز و انعقاد برای کاهش پیچیدگی فضای نمونه‌گیری استفاده می‌شود[۷۶]. استفاده از توزیع دیریکله به عنوان پیشین، فرمول‌بندی سر راست تری برای محاسبه‌ی توزیع پسین بدست می‌دهد که بر پایه‌ی آن، روال Gibbs sampling به شکل یک الگوریتم بسیار ساده و در عین حال سریع قابل پیاده‌سازی است. این الگوریتم به طور خلاصه به صورت زیر بیان می‌شود:

ورودی: مجموعه‌ی ژنوتیپ‌ها G و توزیع پیشین برای احتمال هاپلوتیپ‌ها به صورت $\Theta \sim Dirichlet(\beta)$ داده شده است که $\beta = \langle \beta_1, \ldots, \beta_{|\mathcal{H}|} \rangle$ بردار شبه شمارنده‌های[۳۹] Θ است.

شروع) به هر ژنوتیپ یک زوج هاپلوتیپ مکمل و سازگار با آن ژنوتیپ را نسبت بده.

۱) تا زمانی که نمونه‌گیری بیشتری برای توزیع پسین نیاز است مراحل زیر را تکرار کن،

۲) به طور تصادفی یک ژنوتیپ را انتخاب کن (یا به ترتیب برای هر یک از ژنوتیپ‌ها)، مثل g_i، و یک زوج هاپلوتیپ مکمل و سازگار برای آنرا بر اساس توزیع زیر به طور تصادفی انتخاب کن،

$$P\left(\langle h_a, h_b \rangle \mid g_i = h_a \oplus h_b, G_{-i}\right) \propto (n_a + \beta_a) \,.\, (n_b + \beta_b)$$

[۳۷] Markov Chain Monte Carlo
[۳۸] Dirichlet
[۳۹] pseudo counts

که در آن G_{-i} مجموعه ژنوتیپ‌ها، به جز g_i است و n_a و n_b به ترتیب تعداد حضور هاپلوتیپ h_a و h_b در G_{-i} است.

این الگوریتم که هسته‌ی اصلی نرم‌افزار HAPLOTYPER است مشابه روشی است که پیشتر در [۷۷] برای شناسائی موتیف‌ها مورد استفاده قرار گرفته بود. روش مشابه دیگری ولی با استفاده از فرآیندهای دیریکله، در [۷۸] معرفی شده‌است. کاربرد مدل بیزی برای حل شکل دیگری از مسئله‌ی استنباط هاپلوتیپ‌ها که با فرض در اختیار داشتن شجره‌ی[۴۰] نمونه‌ی مورد مطالعه، مطرح می‌شود نیز توسعه داده شده است[۷۰]. بر خلاف روش EM، روشهای مبتنی بر دیدگاه بیزی نسبت به برقراری شرایط مورد پیش فرضشان حساسیت بیشتری دارند. برای مثال، دور بودن از شرایط فیلوژنی کامل می‌تواند مانع رسیدن روال Gibbs sampling به نتایج مطلوب شود[۷۹].

سایر روشها و الگوریتم ژنتیکی برای حل مسئله‌ی استنباط هاپلوتیپ‌ها

علاوه بر چهار رویکرد اصلی در مسئله‌ی استنباط هاپلوتیپ‌ها که در بالا به آنها اشاره شد برخی رویکردهای اکتشافی[۴۱] نیز برای حل این مسئله بکار گرفته شده‌اند که از آن میان می‌توان به روش 2SNP اشاره کرد. در این روش، تفکیک ژنوتیپ‌ها به تعیین فاز ژنوتیپ‌های دو اسنیپی کاهش داده می‌شود و مجموعه هاپلوتیپ‌های جواب با ترکیب نتایج بدست آمده و به وسیله‌ی حل یک مسئله‌ی درخت فراگیر کمینه[۴۲] بدست می‌آیند[۸۰].

الگوریتم‌های ژنتیکی[۴۳] نیز یکی دیگر از رویکردهای رایج در حل مسائل گوناگون بهینه‌سازی و از جمله مسئله‌ی استنباط هاپلوتیپ‌ها هستند. اولین بار، براتِن و همکارانش از یک الگوریتم ژنتیکی در یک مطالعه‌ی موردی برای شناسائی هاپلوتیپ‌های ناحیه‌ی ژنومی ژن LDL-receptor استفاده کردند[۸۱]. با این حال جزئیات روش به کار گرفته شده به روشنی توضیح داده نشده است و نمونه‌ی پیاده‌سازی شده الگوریتم ارائه نگردیده است. از سویی دیگر، در مسئله‌ی استنتاج هاپلوتیپ‌ها از داده‌های ژنوتیپی در یک شجره، یک الگوریتم ژنتیکی در قالب یک نرم‌افزار کارآمد پیاده‌سازی شده است[۸۲]. کاربرد الگوریتم ژنتیکی برای حل مسئله‌ی

[۴۰] pedigree
[۴۱] heuristic
[۴۲] Minimum spanning tree
[۴۳] Genetic Algorithm, GA

برآورد فراوانی هاپلوتیپ‌های جمعیت نیز توسط آزوما و همکارانش در [۸۳] مورد بررسی قرار گرفته است.

در مجموع، به نظر می‌رسد استفاده از الگوریتم ژنتیکی در حالی که یک ابزار انعطاف‌پذیر برای حل مسائل بهینه‌سازی به شمار می‌آید در حل مسئله‌ی استنباط هاپلوتیپ‌ها و در مقایسه با دیگر روشها چندان مورد توجه قرار نگرفته است. در این رساله، در بخش ۱۰۲ توانایی الگوریتم ژنتیک برای حل مسئله‌ی تفکیک هاپلوتیپ‌ها با هدف بیشترین پارسیمونی مورد مطالعه قرار می‌گیرد. به نظر می‌رسد انعطاف‌پذیریGA در جستجوی فضای جواب، امکان بررسی جوابهای بهینه‌ی نزدیک به جواب بهینه‌ی سراسری را که در دیدگاه بیشترین پارسیمونی دارای اهمیت هستند فراهم می‌کند. در همین راستا، روشهای متعددی برای تولید نمونه‌های شبیه‌سازی شده از ژنوتیپ‌ها نیز در بخش ۲۰۲ معرفی خواهند شد.

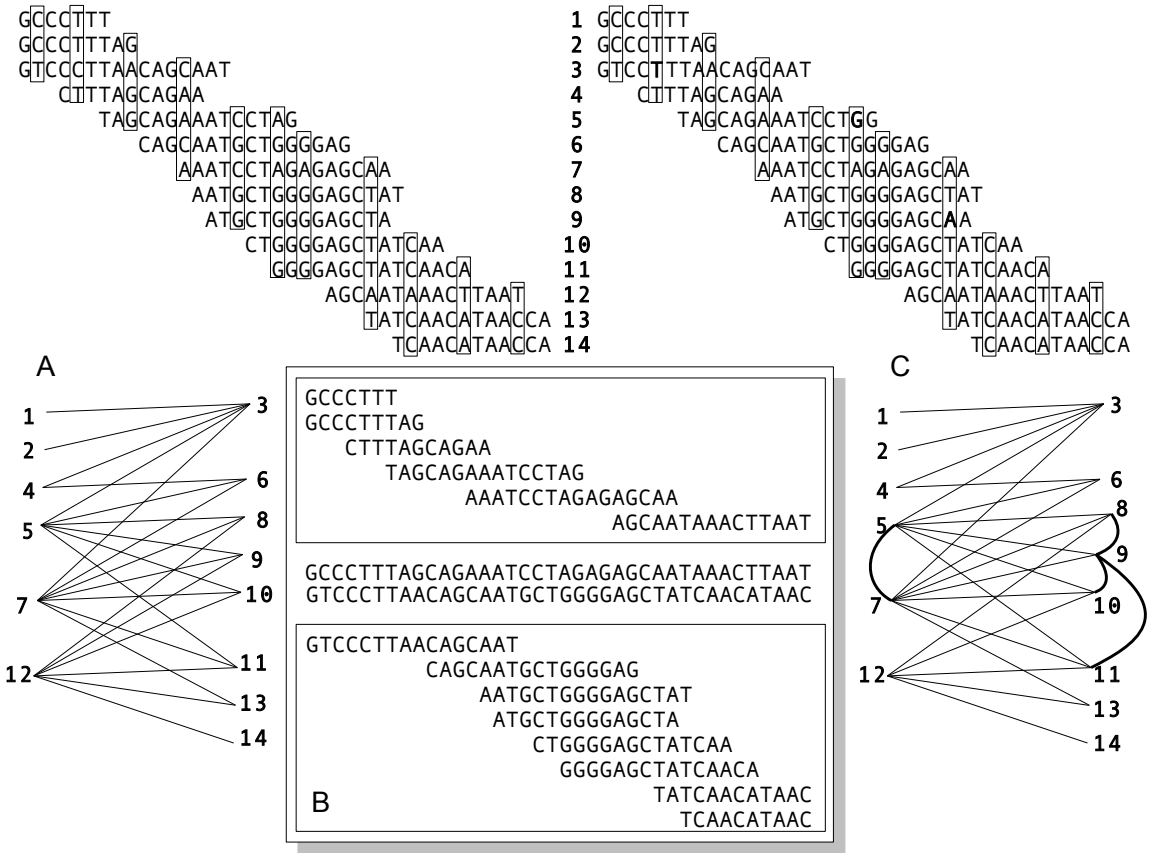
بازسازی هاپلوتیپ‌های فردی با برهمگذرای قطعات توالی‌یابی‌شده

همانطور که در بخش ۲۰۱ اشاره شد توالی‌یابی ژنوم از طریق توالی‌یابی قطعات ژنومی که از بریدن[۴۴] رشته‌های DNA همانندسازی[۴۵] شده از ژنوم در نقاط مختلف بدست می‌آیند صورت می‌گیرد. وقتی نمونه ژنوم مورد مطالعه تنها به یک فرد تعلق داشته باشد منشاء هر یک از قطعات بدست آمده پس از پروسه‌ی همانندسازی و برش، یکی از دو کروموزوم همسان فرد است. تفکیک قطعات توالی‌یابی‌شده به دو دسته و برهمگذاری آنها برای بازسازی هاپلوتیپ‌های هر یک از دو کروموزم را اختصارا مسئله‌ی برهمگذاری قطعات هاپلوتیپ می‌نامند. حل این مسئله از یک طرف، شیوه‌ای برای یافتن یک برهمگذاری برای ژنوم به کمک اسنیپ‌هاست و از طرف دیگر، در نوع خود یک مسئله‌ی شناسائی هاپلوتیپ‌ها به حساب می‌آید.

رویکرد اصلی برای حل مسئله‌ی برهمگذرای قطعات هاپلوتیپ، محاسبه‌ی گراف تضاد[۴۶] است. در گراف تضاد هر رأس نماینده‌ی یکی از قطعات است و دو رأس همسایه یکدیگرند اگر توالی قطعه‌های متناظرشان دست کم در یک اسنیپ متفاوت با دیگری باشد. از نظر تئوری، گراف تضاد یک گراف دوبخشی است و قطعات متناظر با رئوس هر بخش، متعلق به یک کروموزوم هستند. این در شرایطی است که هیچ خطایی در خواندن نوکلئوتیدها در فرآیند توالی‌یابی قطعات روی نداده باشد در غیر اینصورت، گراف بدست آمده لزوما

[۴۴] shearing
[۴۵] cloning
[۴۶] Conflict graph

دوبخشی نخواهد بود.



شکل ۹۰۱: بازسازی هاپلوتیپ‌های فردی به وسیله‌ی برهمگذاری قطعات توالی‌یابی‌شده
A بالا) برهمگذاری تعدادی از قطعه توالی‌ها بدون تفکیک آنها به هاپلوتیپ‌های فرد مورد مطالعه. **A پائین)** گراف تضاد مرتبط با قطعات برهمگذاری شده در بالا. **B)** تعیین فاز قطعات به وسیله‌ی گراف دوبخشی و برهمگذاری هاپلوتیپ‌ها. **C)** همانند قسمت (A). در اینجا به دلیل خطای توالی‌یابی در برخی از اسنیپ‌ها، دوبخشی بودن گراف تضاد از بین می‌رود.

به طور معمول، وجود خطا در توالی‌یابی نوکلئوتیدها، حتی با نرخ بسیار پائین امری اجتناب‌ناپذیر است. بر همین اساس، اشکال دیگری از مسئله‌ی برهمگذاری قطعات هاپلوتیپ‌ها مطرح می‌شوند که در ادامه به برخی از آنها اشاره می‌شود. در ساده‌ترین شکل این مسائل، با حذف حداقل تعداد ممکن از قطعات داده شده یک گراف تضاد دوبخشی به دست می‌آید[۸۴، ۸۵]. شکل دیگری از مسئله که بیشتر مورد توجه قرار می‌گیرد مسئله‌ی تغییر و تصحیح حداقل تعداد ممکن از نوکلئوتیدها در قطعات توالی‌یابی‌شده است به قسمی که با توالی‌های بدست آمده بتوان یک گراف تضاد دوبخشی بدست آورد. این مسئله را به اختصار مسئله‌ی MEC[۴۷] می‌نامند.

[۴۷] Minimum Error Correction

در حالیکه بدست آوردن جواب بهینه برای MEC، مسئله‌ای NP-hard است[۸۶]، اما روش‌های متعددی برای حل تقریبی این مسئله ارائه شده‌اند از جمله روش مبتنی بر رویکرد self-organizing map [۸۷]، روش مبتنی بر information fusion[۸۸]، روش local-exhaustive search[۸۹]، روش HASH مبتنی بر رویکرد MCMC[۹۰]، روش مبتنی بر particle swarm optimization[۹۱] و روش HapCUT[۹۲]. عمده‌ی این روش‌ها از ابزارهای متعارف در هوش مصنوعی به شمار می‌آیند. همچنین، رهیافت‌هایی مبتنی بر الگوریتم ژنتیکی برای حل این مسئله پیشنهاد شده اند[۹۳، ۹۴]. داشتن اطلاعات ژنوتیپی از دیگر نمونه‌ها نیز می‌تواند به حل مسئله‌ی MEC کمک کند. یک الگوریتم مبتنی بر شبکه‌های عصبی[۹۵] و روش دیگری مبتنی بر الگوریتم ژنتیکی[۹۶] از جمله روش‌هایی هستند که تاکنون برای حل مسئله‌ی MEC به کمک دیگر اطلاعات ژنوتیپی ارائه شده‌اند.

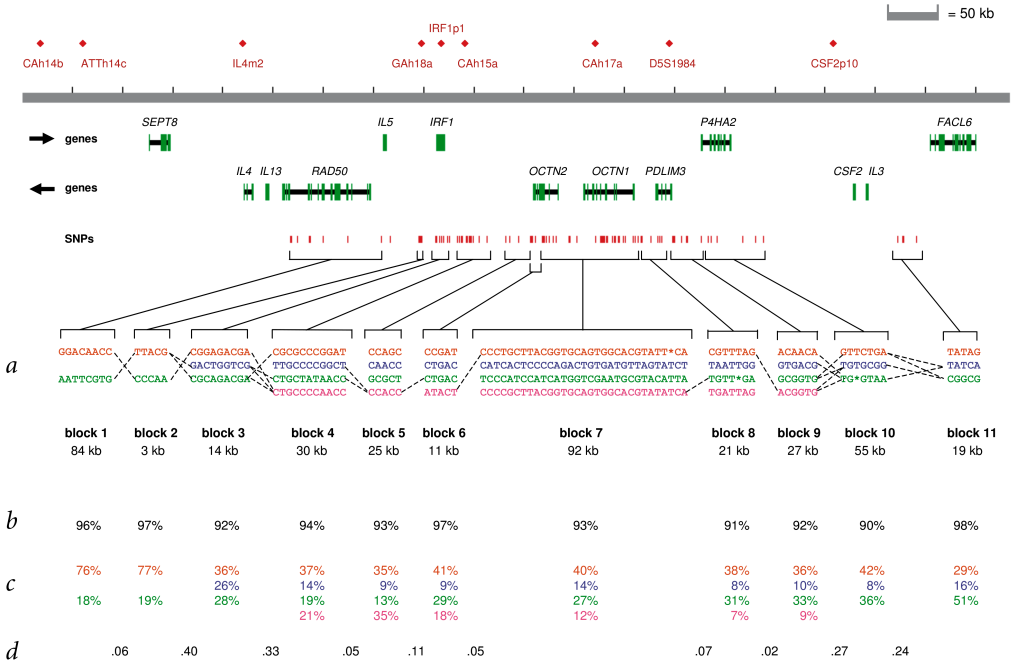
۴۰۱ بلوک‌های هاپلوتیپ

نتایج بدست آمده از مشاهدات اخیر بر روی ژنوم انسان، این باور را در بین محققان تقویت کرده است که ژنوم را می‌توان به نواحی پیوسته و متمایز از یکدیگری افراز کرد که تنوع هاپلوتیپها درون هر یک از این نواحی در طی نسل‌ها حفظ می‌شود. هر یک از این نواحی را یک بلوک هاپلوتیپی[۴۸] می‌نامیم. ساختار بلوکی ژنوم، اولین بار با مشاهدات پَتیل و همکارانش در بررسی تنوع هاپلوتیپها در کروموزوم ۲۱ در بین ۲۲ فرد مورد توجه قرار گرفت[۹۷]. در واقع به نظر می‌رسد هاپلوتیپ‌ها در قطعات معینی از ژنوم، دست نخورده از نسلی به نسل دیگر منتقل می‌شوند. از دیدگاه تئوری، تنوع هاپلوتیپ‌ها در ناحیه‌ای شامل l اسنیپ، بالقوه می‌تواند بالغ بر ۲^l هاپلوتیپ متمایز باشد اما در واقعیت، تنوع هاپلوتیپها به مراتب پائین‌تر از این مقدار است. بررسی تنوع هاپلوتیپ‌ها در نواحی مختلف ژنوم، یکی از رویکردهای رایج برای تعریف بلوک‌های هاپلوتیپ است.

یکی دیگر از اولین تحقیقاتی که وجود ساختار بلوکی بر روی ژنوم را مورد توجه قرار داد مطالعه‌ی دالی و دیگران[۹۸] بر روی ناحیه‌ای شامل ۱۰۳ اسنیپ از کروموزوم پنج (5q31) بود. تحقیقات آنها بیانگر این واقعیت بود که بین گروه‌هایی از اسنیپ‌ها که در مناطق معینی از ژنوم قرار گرفته‌اند همبستگی آماری مشاهده

[۴۸] Haplotype block

می‌شود. در این نواحی نیز به گونه‌ای دیگر، طرح‌های بلوک‌مانندی در امتداد ژنوم مشاهده می‌شدند که پیش از هر چیز نشاندهنده‌ی عدم رویداد نوترکیبی یا پایین بودن نرخ آن درون این مناطق است.



شکل ۱۰۰۱: بلوک‌های هاپلوتیپ در ناحیه‌ی 5q31
a) هاپلوتیپ‌های رایج در هر یک از بلوک‌ها. پاره‌خطها نشاندهنده‌ی موقعیت‌هایی است که در آن بیش از دو درصد از تمام کروموزومها با گذر از یک هاپلوتیپ رایج به هاپلوتیپ رایج بعدی بدست می‌آید. **b)** نسبتی از کل نمونه که توسط هاپلوتیپ‌های رایج پوشش داده می‌شوند. **c)** فراوانی هاپلوتیپ‌های رایج در هر یک از بلوک‌ها. **d)** نرخ جابجایی هاپلوتیپ‌ها در مرز بلوک‌ها. شکل از Daly et al. [98] اقتباس شده است.

مقایسه‌ی بلوک‌های هاپلوتیپ در بین نژادهای مختلف انسان و تغییراتی که در طرح بلوک‌ها با ترکیب نمونه‌های نژادهای مختلف بدست می‌آید دستیابی به یک ساختار بلوکی توافق‌شده در کل جمعیت را در حیطه‌ی شک قرار می‌دهد[۹۹–۱۰۲]. در واقع لازم است این حقیقت مورد توجه قرار گیرد که تنوع هاپلوتیپی در زیرگروه‌های جمعیتی می‌توانند متاثر از رویدادهای دیگری به جز نوترکیبی نیز باشند. به عنوان مثال در جمعیتی که نسل نخستین آن، گروه نسبتا کوچک و جدا افتاده از بقیه‌ی جمعیت باشد تنوع هاپلوتیپی کمتری وجود دارد و به تعبیری بلوک‌های هاپلوتیپی بزرگتری بر روی ژنوم دیده می‌شوند هر چند در عمل این موضوع دلیل بر عدم رویداد نوترکیبی در محدوده‌ی چنین بلوکی نیست[۱۰۳، ۱۰۴].

مسئله‌ی افراز بلوکی هاپلوتیپ‌ها[۴۹]

فرض کنید $H = \{h_1, h_2, \ldots, h_n\}$ نمونه‌ای شامل اطلاعات l اسنیپ در n هاپلوتیپ داده شده است؛ یعنی هر هاپلوتیپ مثل h_i برداری lتائی از 0و 1است که هر مؤلفه‌ی آن نشانگر آلل مشاهده شده در یک اسنیپ است. هدف تعیین بازه‌هایی مثل $\{[a_1, b_1], \ldots, [a_m, b_m]\}$ بر روی ژنوم است به قسمی که $1 \leq a_1 \leq b_1 \leq \cdots \leq a_m \leq b_m \leq l$ اندیس اسنیپ‌های قرار گرفته در لبه‌ی بازه‌ها هستند و مجموعه‌ی هاپلوتیپ‌هایی که از تحدید هاپلوتیپ‌های داده شده در این بازه‌ها بدست می‌آید شرایط مطلوب بهینگی را ارضاء می‌کند.

صورت فوق، تعریف کلی مسئله‌ی افراز بلوکی هاپلوتیپ‌هاست. با این حال، به ازای مدلهای متفاوتی که برای تبیین ساختار بلوکی ژنوم پیشنهاد می‌شوند، معیارهای بهینگی متفاوتی نیز برای تعیین بلوک‌های هاپلوتیپی در نظر گرفته می‌شوند. همانطور که پیشتر اشاره شد، علیرغم ویژگی‌های مشترک و کلی در مورد بلوک‌های هاپلوتیپی که مورد توافق عموم صاحب نظران است، یک مدل منحصربفرد برای تعریف بلوک‌های هاپلوتیپی وجود ندارد و بسته به کاربرد، شرایط متفاوتی برای افراز بلوک‌ها ارائه می‌گردد.

نکته‌ی دیگری که در ارتباط با مسئله‌ی افراز بلوکی هاپلوتیپ‌ها باید مورد توجه قرار داد این است که این مسئله به طور متعارف با این فرض که داده‌های ورودی مسئله، هاپلوتیپ‌ها هستند مورد بررسی قرار می‌گیرد در حالی که اصولا هاپلوتیپ‌ها پس از تفکیک داده‌های ژنوتیپی بدست می‌آیند. این وضعیت غالبا باعث می‌شود دو مسئله‌ی تفکیک ژنوتیپ‌ها به هاپلوتیپ‌ها و افراز بلوکی هاپلوتیپ‌ها به طور همزمان لازم و ملزوم یکدیگر باشند. از این رو رهیافت‌های مختلفی برای رفع این مشکل مطرح شده‌اند. برخی از شیوه‌ها مانند روش GERBIL[۱۰۵]، از یک مدل تلفیقی برای حل همزمان هر دو مسئله استفاده می‌کنند. از دیگر سو، روشهای تعیین فاز ژنوتیپ‌ها که از تکنیک موسوم به partition-ligation استفاده می‌کنند (بخش ۳۰۱) در عمل نیازی به پیش فرض قرار دادن یک افراز بلوکی برای ژنوم ندارند. با این حال، عمده‌ی رویکردهای رایج برای حل مسئله‌ی افراز بلوکی هاپلوتیپ‌ها به طور ابتدائی بر مبنای در اختیار داشتن داده‌های هاپلوتیپ توسعه داده می‌شوند ولی در عمل، هر یک ترفند خاصی را برای استفاده‌ی مستقیم از داده‌های ژنوتیپ نیز بکار

[۴۹] Haplotype block partitioning

می‌گیرند. بدین ترتیب، تعیین بلوک‌های هاپلوتیپ پیش از روال تعیین فاز ژنوتیپ‌ها امکان‌پذیر خواهد بود.

پس از این مقدمه و در ادامه، دو جنبه‌ی متفاوت از شیوه‌های رایج در افراز بلوکی ژنوم را مورد بحث قرار می‌دهیم. این دو جنبه عبارتند از معیار مورد استفاده برای تعریف بلوک هاپلوتیپی و ساختار افراز بهینه.

شیوه‌های افراز بلوک‌های هاپلوتیپی را می‌توان به دو گروه کلی تقسیم کرد: روشهایی که در آنها از سنجه‌های مرتبط با واگرائی هاپلوتیپ‌ها برای تعیین بلوک‌ها استفاده می‌شود و روشهایی که در آنها بلوک‌ها از طریق ارزیابی برخی آماره‌های مرتبط با همبستگی جفتی اسنیپ‌ها تعریف می‌شوند.

واگرائی هاپلوتیپ‌ها[۵۰]

در یک ناحیه‌ی ژنوم، هر اندازه تعداد هاپلوتیپ‌های متمایز در جمعیت بیشتر باشد می‌گوئیم واگرائی هاپلوتیپ‌ها در آن ناحیه بیشتر است. به بیان ساده، واگرائی هاپلوتیپ‌ها توصیف کیفی تنوع هاپلوتیپ‌ها در منطقه‌ی مورد بررسی است. از نظر کمّی، سنجه‌های متعددی برای اندازه‌گیری واگرائی هاپلوتیپ‌ها معرفی شده‌اند که در ادامه به برخی از آنها اشاره می‌کنیم.

هتروزیگوسیتی هاپلوتیپ‌ها[۵۱]

ساده‌ترین شیوه برای اندازه‌گیری واگرائی هاپلوتیپ‌ها، محاسبه‌ی تفاضل هاپلوتیپ‌ها از یکدیگر و در نظر گرفتن حاصلجمع مربع تفاضلات به عنوان یک سنجه برای واگرائی هاپلوتیپ‌هاست.

$$D_H = \sum_{i=1}^{n} \sum_{k=1}^{n} \|h_i - h_k\|^2 \qquad (۴۰۱)$$

با توجه به اینکه مقدار مؤلفه‌های بردارهای هاپلوتیپ 0یا 1است در عمل، مقدار $\|h_i - h_k\|^2$ در اینجا با فاصله‌ی همینگ[۵۲] بین دو هاپلوتیپ معادل خواهد بود. وقتی تمام هاپلوتیپها یکسان باشند، مقدار D_H برابر صفر است و با افزایش تعداد هاپلوتیپهای متمایز مقدار آن افزایش می‌یابد. برای بدست آوردن یک آماره‌ی

[۵۰] Haplotype diversity
[۵۱] Heterozygosity
[۵۲] Hamming distance

نرمال شده، مقدار واگرائی را در بین هاپلوتیپهای متمایز نمونه‌ی داده شده نیز محاسبه می‌کنند؛ یعنی،

$$D_{\mathcal{H}} = \sum_{h\in\mathcal{H}} \sum_{h'\in\mathcal{H}} \|h - h'\|^{۲} \qquad (۵۰۱)$$

که در آن $\mathcal{H}$ مجموعه‌ی هاپلوتیپ‌های متمایز در H است. نسبت $\frac{D_{\mathcal{H}}}{D_H}$ ، مقداری بین صفر و یک است که به عنوان مقدار نرمال شده‌ی واگرائی هاپلوتیپ‌ها مورد استفاده قرار می‌گیرد. بنابر قرارداد در اینجا نسبتی با صورت و مخرج صفر را برابر با یک در نظر می‌گیریم.

به سادگی می‌توان نشان داد،

$$D_H = ۲l^{۲} \sum_{j=۱}^{l} f_j(۱ - f_j) \qquad (۶۰۱)$$

که در اینجا f_j فراوانی نسبی آلل 1در اسنیپ jام است. عبارتی که جمع در رابطهٔ (۶۰۱) بر روی آن انجام می‌شود به نوعی نشاندهنده‌ی میزان ناهمگونی یا اصطلاحا هتروزیگوسیتی اسنیپ jام در نمونه است. بدین ترتیب ملاحظه می‌شود که نزدیک بودن فراوانی آلل‌های متفاوت به یکدیگر در هر یک از اسنیپ‌ها، افزایش سطح واگرائی را در پی دارد. به طور کلی، مقدار D_H به طور مطلق، تابعی از وضعیت مستقل اسنیپ‌ها در جمعیت است و از این رو توصیف‌کننده‌ی قدرتمندی برای بدست آوردن اطلاعات موجود در ترکیب اسنیپ‌ها و واگرائی هاپلوتیپ‌ها نیست. با این حال، استفاده از این رابطه دارای این مزیت است که محاسبه‌ی D_H، حتی بدون اطلاع از فاز ژنوتیپ‌ها نیز امکان‌پذیر است. به طور کلی، این سنجه و شکل نرمال شده‌ی آن، بیشتر در برخی کاربردهای مسئله‌ی انتخاب تگ‌اسنیپ‌ها مورد استفاده قرار می‌گیرند.

هاپلوتیپ‌های رایج[۵۳]

رویکرد دیگر برای سنجش واگرائی در هاپلوتیپ‌ها، تعریف «هاپلوتیپ‌های رایج» است. منظور از هاپلوتیپ رایج، هاپلوتیپی است که فراوانی آن در نمونه‌ی مورد مطالعه، به طور معنادار بیشتر از آن باشد که یک هاپلوتیپ منفرد به حساب آید. به عبارت دقیق‌تر، هاپلوتیپی را هاپلوتیپ رایج می‌گوئیم که فراوانی آن در نمونه‌ی مورد بررسی بیش از β درصد باشد. متعارف آن است که هاپلوتیپ‌هایی که در بیش از ۵ درصد نمونه مشاهده می‌شوند را هاپلوتیپ‌های رایج به حساب می‌آورند. در این رویکرد، نسبتی از کل هاپلوتیپ‌های نمونه که به

[۵۳] Common haplotypes

مجموعه‌ی هاپلوتیپ‌های رایج تعلق دارند تحت عنوان «پوشش هاپلوتیپ‌های رایج» محاسبه می‌شود و به عنوان معیاری برای تعیین بلوک‌های هاپلوتیپ مورد استفاده قرار می‌گیرد. در اینجا، از مقدار آستانه‌ای دیگری که معمولا با α نمایش داده می‌شود استفاده می‌شود و نواحیی که پوشش هاپلوتیپ‌های رایج در آنها دست کم α درصد باشد به عنوان یک ناحیه‌ی بلوکی بالقوه در نظر گرفته می‌شود. این همان معیاری است که توسط پَتیل و همکارانش در [۹۷] مورد استفاده قرار گرفت. الگوریتم‌های دیگری از جمله HaploBlockFinder و HapBlock نیز از این رویکرد برای تعیین بلوک‌های هاپلوتیپ استفاده می‌کنند[۱۰۶–۱۰۸].

دسته‌ی دیگری از روش‌ها، برای افراز بلوکی هاپلوتیپ‌ها از یک مدل مبتنی بر نظریه‌ی اطلاعات موسوم به Minimum description length استفاده می‌کنند[۱۰۹، ۱۱۰]. این مدل نیز در واقع در رده‌ی مدل‌های مبتنی بر تنوع هاپلوتیپها قرار می‌گیرد.

عدم‌تعادلِ ناشی از بهم‌پیوستگی[۵۴]

یکی دیگر از معیارهای رایج برای تعیین بلوک‌های هاپلوتیپی، همبستگی آماری بین جفت اسنیپ‌ها در نواحی مختلف ژنوم است. به دلیل برخی سازوکارهای زیستی در نواحی خاصی از ژنوم، در بین دسته‌ای از نوکلئوتیدهای مجاور هم نوعی بهم‌پیوستگی مشاهده می‌شود؛ بدین معنی که احتمال رویداد نوترکیبی داخل این نواحی تا آن اندازه پایین است که می‌توان فرض کرد توالی کروموزم‌ها در این نواحی بدون تغییر به نسل بعد منتقل می‌شوند. برای بررسی وجود این نوع بهم‌پیوستگی، به طور متعارف، ژنوتیپ زوج اسنیپ‌هایی که در ناحیه‌ی مورد مطالعه واقع شده‌اند را در دو وضعیت بررسی می‌کنند؛ یکی وضعیتی که رویداد نوترکیبی بارها بین آنها رخ داده است و دیگری وضعیت حال ژنوتیپ‌های موجود در جمعیت یا نمونه‌ی داده شده.

بروز رویدادهای نوترکیبی فراوان بین دو اسنیپ باعث می‌شود فراوانی آلل‌های هر یک از اسنیپ‌ها مستقل از فراوانی ژنوتیپ‌های زوج اسنیپ باشد. از دیدگاه آماری، تنوع ژنوتیپ‌ها در دو جایگاه ژنی، نشاندهنده‌ی یک وضعیت تعادل است چنانچه فراوانی نسبی توأم برای ژنوتیپ معینی از دو اسنیپ برابر با حاصلضرب فراوانی نسبی هر یک از آلل‌های تشکیل‌دهنده‌ی آن ژنوتیپ باشد. وجود تفاوت بین فراوانی یک ژنوتیپ مشاهده شده

[۵۴] Linkage Disequilibrium, LD

در نمونه و فراوانی همان ژنوتیپ در وضعیت تعادل، نشانگر بهم‌پیوستگی جایگاه‌های مرتبط با آن ژنوتیپ است. اندازه‌ی این اختلاف را عدم‌تعادل ناشی از بهم‌پیوستگی یا به اختصار LD می‌گوئیم. آماره‌ی اساسی برای اندازه‌گیری LD به وسیله‌ی رابطه‌ی زیر تعریف می‌شود:

$$D = D_{XY} = P(X = \mathbf{1}, Y = \mathbf{1}) - P(X = \mathbf{1}).P(Y = \mathbf{1}) \qquad (۷۰۱)$$

که در آن X و Y متغیرهای تصادفی مرتبط با اسنیپ‌های زوج اسنیپ مورد بررسی هستند و $P(X = x)$ نشاندهنده‌ی احتمال مشاهده‌ی آلل x در اسنیپ X است. مقدار D در شرایط تعادل، صفر است و با افزایش LD، در جهت مثبت یا منفی از صفر دور می‌شود. نکته‌ی جالب توجه آن است که مقدار قدرمطلق D مستقل از نحوه‌ی انتخاب آلل 0یا 1برای تعریف D در رابطهٔ (۷۰۱) است و در واقع داریم:

$$D_{XY} = (-۱)^{(x+y)}\big(P(X = x, Y = y) - P(X = x).P(Y = y)\big)$$

که در آن x و y مستقلاً می‌توانند هر یک از مقادیر 0یا 1باشند.

ویژگی جالب دیگر، ارتباط نرخ نوترکیبی با کاهش مقدار D با گذشت زمان است. فرض کنید ρ نشاندهنده‌ی احتمال نوترکیبی بین جایگاه‌های مورد مطالعه در یک نسل است. می‌توان نشان داد:

$$D_{t+۱} = (۱ - \rho)D_t$$

که در آن D_t مقدار LD بین جایگاه‌های مورد مطالعه در نسل t و $D_{t+۱}$ همان مقدار در نسل $t + ۱$ است. در واقع با گذشت زمان، D به طور نمایی به سمت صفر میل می‌کند و نرخ این تصاعد $(۱ - \rho)$ است.

معمولا به جای D، از قدرمطلق آماره‌ی نرمال شده‌ای که آنرا با D' نشان می‌دهند استفاده می‌شود. این آماره به صورت زیر تعریف می‌شود:

$$D' = D'_{XY} = \frac{D_{XY}}{D_{MAX}} \qquad (۸۰۱)$$

$$D_{MAX} = \begin{cases} \min\big(P(X = \mathbf{0}).P(Y = \mathbf{1}), P(X = \mathbf{1}).P(Y = \mathbf{0})\big), & if\ D_{XY} > ۰ \\ \min\big(P(X = \mathbf{0}).P(Y = \mathbf{0}), P(X = \mathbf{1}).P(Y = \mathbf{1})\big), & if\ D_{XY} < ۰ \end{cases}$$

بازه‌ی تغییرات D' بین -۱ تا ۱ است. $D' = ۰$ نشاندهنده‌ی LDصفر یعنی استقلال کامل دو اسنیپ از یکدیگر است و $|D'| = ۱$ نشاندهنده‌ی همبستگی کامل بین دو اسنیپ است. وقتی $D > ۰$ است دو اسنیپ

همبستگی مثبت دارند که به بیان ساده نشاندهنده‌ی وضعیتی است که در آن همبستگی بیشتر بین آلل‌های فراوان هر اسنیپ با یکدیگر دیده می‌شود و وقتی $D' < ۰$ است همبستگی بین دو اسنیپ منفی است که نشاندهنده‌ی وضعیتی است که در آن همبستگی بیشتر بین آلل وحشی از یک اسنیپ و آلل جهش‌یافته از اسنیپ دیگر مشاهده می‌شود. ویژگی D' در آن است که تغییرات آن مستقل از تغییرات احتمال حاشیه‌ای[۵۵] هر یک از اسنیپ‌های مورد بررسی، یعنی مقادیر $P(X)$ و $P(Y)$ است. در واقع می‌توان نشان داد که $|D'|$ تنها تابعی از $P(X,Y)$ است که به ازای مقادیر مختلف آن به طور خطی بین صفر و یک تغییر می‌کند. این آماره اولین بار توسط لِونتین[۱۱۱] معرفی شد و اکنون به عنوان مقیاس رایج برای اندازه‌گیری LD بر روی ژنوم بکار گرفته می‌شود.

آماره‌ی دیگری که بعضاً برای اندازه‌گیری مقدار LD بین جفت اسنیپ‌ها مورد استفاده قرار می‌گیرد آماره‌ی $r^۲$ است که به صورت زیر تعریف می‌شود:

$$r^۲ = r^۲_{XY} = \frac{D^۲_{XY}}{P(X=\mathbf{0}).P(Y=\mathbf{0}).P(X=\mathbf{1}).P(Y=\mathbf{1})} \qquad (۹۰۱)$$

مقدار $r^۲$ بین صفر و یک است که $r^۲ = ۰$ نشاندهنده‌ی عدم وجود LD و $r^۲ = ۱$ نشاندهنده‌ی یک همبستگی کامل بین دو اسنیپ است. تغییرات $r^۲$، بر خلاف D' تابعی از احتمالات حاشیه‌ای دو اسنیپ است که در جای خود به عنوان ابزار مناسبی برای تخمین اندازه‌ی نمونه به کار گرفته می‌شود. نرم‌افزارهایی چون GOLD[۱۱۲]، goldsurfer[۱۱۳] و Haploview[۱۱۴] از این آماره‌ها برای نمایش گرافیکی گستردگی LD در ژنوم استفاده می‌کنند. با اینکه در رویکرد دالی و دیگران[۹۸] از یک مدل مارکوف پنهان[۵۶] برای تشخیص مرز بلوک‌ها استفاده می‌شود اما این مدل نیز در نهایت به شکل روابطی بر حسب D بیان می‌شوند.

برآوردگر فاصله‌ای برایLD

از نظر ریاضی، می‌توان نشان داد که D و $r^۲$ به ترتیب معادل کوواریانس و ضریب همبستگی[۵۷] متغیرهای تصادفی (X,Y) هستند که در آن هر کدام از متغیرهای X و Y، یک متغیر برنولی است. به طور معمول، برای

[۵۵] marginal probability
[۵۶] Hidden Markov Model
[۵۷] corelation coefficient

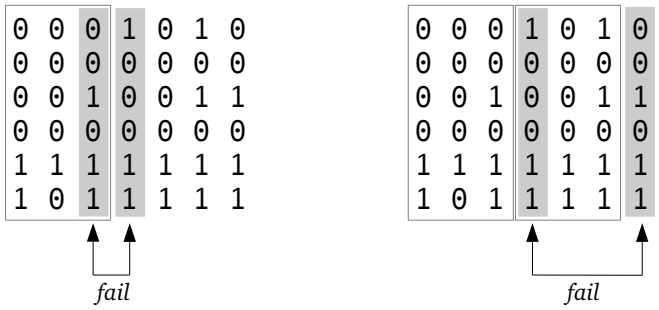
برآورد هر یک از آماره‌های فوق، برآوردی از فراوانی نسبی ژنوتیپ اسنیپ‌ها و جفت اسنیپ‌ها در نمونه‌ای از هاپلوتیپ‌های جمعیت، جایگزین مقادیر احتمال در روابط ۷۰۱، ۸۰۱ و ۹۰۱ می‌شود. از این طریق یک برآورد نقطه‌ای[۵۸] برای آماره‌ی مورد نظر بدست می‌آید که از نظر تئوری می‌تواند با مقدار آماره‌ی جمعیت تفاوت داشته باشد. اساساً این تفاوت با افزایش حجم نمونه کاهش می‌یابد و در واقع برآورد نمونه‌ای به سمت آماره‌ی جمعیت میل می‌کند. یک رویکرد رایج برای تعیین اعتبار و دقت برآورد نمونه‌ای، محاسبه‌ی بازه‌ی اطمینان[۵۹] برای برآوردگر نقطه‌ای است. برای محاسبه‌ی یک بازه‌ی اطمینان لازم است اطلاعاتی درباره‌ی تابع توزیع برآوردگر مورد نظر در دسترس باشد. به عنوان مثال، می‌توان نشان داد nr^2 به طور مجانبی دارای توزیع مربع کای با یک درجه آزادی است که در آن n همان تعداد هاپلوتیپ‌های نمونه است[۱۱۱]. در نقطه مقابل، تابع توزیع احتمال برآوردگر D' بر حسب توابع شناخته‌شده‌ی آماری قابل محاسبه نیست و برای محاسبه‌ی بازه‌ی اطمینان، از روش‌های عددی مبتنی بر نمونه‌گیری مونت کارلو استفاده می‌شود. بر مبنای چنین رویکردی، در روش منسوب به گابریل[۱۱۵] برای هر جفت از اسنیپ‌ها یک بازه‌ی اطمینان ٪۹۵ مرتبط با برآوردگر آماره‌ی D' محاسبه می‌شود و بر اساس آن هر جفت اسنیپ به یکی از سه رده‌ی «قویاً همبسته»[۶۰]، «نامعلوم»[۶۱] و «نوترکیب»[۶۲] طبقه‌بندی می‌شود.

به عبارت مشخص، مقدار LD بین جفت اسنیپ‌ها در روش گابریل، به صورت یک شاخص سه وضعیتی خلاصه‌سازی می‌شود به طوری که اگر کران‌های بالا و پائین برآورد فاصله‌ای $|D'|$ به ترتیب بیشتر از ۰/۹۸ و ۰/۷ باشند جفت اسنیپ متناظر به عنوان جفت اسنیپ قویاً همبسته و اگر کران بالای همین برآورد کمتر از ۰/۹ باشد جفت اسنیپ متناظر به عنوان یک جفت نوترکیب در نظر گرفته می‌شود و حالت‌هایی جز این دو حالت، نامعلوم در نظر گرفته می‌شوند. در ادامه، در هر منطقه از ژنوم، بزرگترین ناحیه‌ای که در آن بیش از ٪۹۵ از جفت اسنیپ‌ها قویاً همبسته باشند یک بلوک هاپلوتیپی را معین می‌کند.

[۵۸] point estimation
[۵۹] confidence interval
[۶۰] strongly associated
[۶۱] uninformative
[۶۲] recombinant

آزمون چهار گامتی[۶۳]

آزمون چهار گامتی یک معیار ساده و در عین حال کارآمد برای بررسی وقوع رویداد نوترکیبی بین دو جایگاه اسنیپ در گذشته است. اگر فرض کنیم تمام هاپلوتیپ‌های زمان حال دارای یک منشاء واحد هستند و هیچ نوکلئوتیدی وجود ندارد که تاکنون بیش از یکبار دستخوش جهش شده باشد آنگاه مشاهده‌ی چهار هاپلوتیپ متمایز در یک جفت اسنیپ، منحصراً می‌تواند نشاندهنده‌ی وقوع یک رویداد نوترکیبی بین این دو اسنیپ در گذشته باشد. البته این قید سختگیرانه، ممکن است در برخی مواقع، با واقعیت منطبق نباشد که هم می‌تواند به دلیل عدم مطابقت طبیعت با فرضیات چنین مدلی باشد و هم ممکن است ناشی از وجود خطا در توالی‌یابی و استنباط هاپلوتیپ‌ها باشد.



شکل ۱۱۰۱: آزمون چهار گامتی برای تعیین بلوک‌های هاپلوتیپ
برای هر زوج از اسنیپ‌ها، هاپلوتیپ‌هایی که از تحدید داده‌ها به این زوج اسنیپ بدست می‌آیند در نظر گرفته می‌شوند. مشاهده‌ی چهار هاپلوتیپ متمایز در یک زوج اسنیپ، شاهدی بر رویداد نوترکیبی به حساب می‌آید. با قرار دادن مرز بلوک‌ها در مکان مناسب، چنین اسنیپ‌هایی را از یکدیگر جدا می‌کنیم.

به طور کلی، در این آزمون، هر یک از چهار نوع هاپلوتیپ ممکن برای یک جفت اسنیپ، یک گامت نامیده می‌شود و آزمون تنها زمانی با موفقیت همراه است که برای جفت اسنیپ مورد مطالعه در نمونه‌ی داده شده حداکثر سه گامت متفاوت مشاهده شود. روش افراز بلوکی هاپلوتیپ‌ها که توسط وانگ و دیگران[۱۱۶] معرفی شده است در واقع برای اولین بار از آزمون چهار گامتی برای تعیین بلوک‌های هاپلوتیپی استفاده می‌کرد. برای جلوگیری از تاثیر خطاهایی که به آنها اشاره شد، آنها مرز بلوک هاپلوتیپی را تا جایی که آزمون چهار گامتی در بیش از دو درصد از جفت اسنیپ‌ها با شکست مواجه نشده است گسترش می‌دهند.

آزمون چهار گامتی دقیقا همان شرط لازم و کافی برای فیلوژنی کامل است که در بخش ۳۰۱ به آن اشاره

[۶۳] Four gamete test

شده است. از دیگر نکات جالب توجه درباره‌ی این آزمون این است که وضعیتی از دو اسنیپ که در آن $|D'| = ۱$ است دقیقا متناظر با حالتی است که در آن آزمون چهار گامتی قبول می‌شود؛ یعنی حداکثر سه گامت متمایز برای دو اسنیپ مشاهده شده است.

ساختار افراز بهینه

یکی از جنبه‌های تفاوت در بین مدل‌های مختلف افراز بلوکی هاپلوتیپ‌ها، نوع ساختار افراز بهینه و شکل پوشش ژنوم توسط بلوک‌هاست. به طور کلی می‌توان افراز بلوکی مطلوب در مدل‌های مختلف را از حیث ساختار بهینه به دو رده‌ی افراز موضعی و افراز سراسری تقسیم کرد. در افراز موضعی، هر بلوک مستقل از ترکیب سراسری بلوک‌ها بر روی ژنوم تعریف می‌شود. معمولا بلوک‌هایی که بدین ترتیب تولید می‌شوند مجموعه‌ای شبیه به جزایر جدا از هم بر روی ژنوم تشکیل می‌دهند. در افرازهای موضعی، بلوک‌ها معمولا بر مبنای یک معیار موضعی بهینگی تعریف می‌شوند. در نقطه‌ی مقابل، در افراز سراسری هدف تقسیم کل ژنوم به بلوک‌های هاپلوتیپ است. در این نوع افراز، ژنوم توسط مجموعه‌ای از بلوک‌های چسبیده به یکدیگر پوشیده می‌شود و اصطلاحا ژنوم توسط بلوک‌های هاپلوتیپ سنگفرش می‌شود. بدین ترتیب، هر نقطه از ژنوم در یکی از بلوک‌های قرار می‌گیرد. یک افراز سراسری اصولا جواب بهینه‌ی یک تابع هدف خاص، در بین تمام افرازهای ممکن در ناحیه‌ی ژنومی مورد مطالعه است.

در بین روشهای متنوعی که تاکنون برای افراز بلوکی هاپلوتیپ‌ها ارائه شده‌اند ترکیب‌های مختلفی از معیارهای تعیین‌کننده‌ی بلوک و انواع ساختار افراز بهینه می‌توان یافت. به عنوان مثال، روش پَتیل و همکارانش در [۹۷] یک روش مبتنی بر واگرائی هاپلوتیپ‌ها و مفهوم هاپلوتیپ رایج است که در آن یک افراز موضعی با استفاده از یک الگوریتم حریصانه بدست می‌آید. روش اندرسون و نومِبر[۱۰۹] نیز در واقع از سنجه‌ای مرتبط با واگرائی هاپلوتیپ‌ها استفاده می‌کند اما افراز بدست آمده با جستجو در بین تمام افرازهای ممکن روی ناحیه مورد نظر بدست می‌آید و از این رو یک افراز سراسری است. معیار تعیین‌کننده‌ی نواحی بلوکی در روش منسوب به گابریل[۱۱۵] برآورد فاصله‌ای $|D'|$ است که در رده‌ی معیارهای مبتنی بر همبستگی جفتی اسنیپ‌ها به شمار می‌رود. با این حال بلوک‌های هاپلوتیپ در روش گابریل به طور موضعی تعریف می‌شوند.

بنابر اطلاعاتی که ما در مورد روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها بدست آوردیم تاکنون هیچ روش مبتنی بر همبستگی جفتی اسنیپ‌ها که بلوک‌ها در آن از طریق یک افراز بهینه‌ی سراسری تعیین شوند مورد توجه قرار نگرفته است. در بخش ۴۰۲، به معرفی روشی جدید برای بدست آوردن یک افراز سراسری، دربرگیرنده‌ی بیشترین تعداد ممکن جفت اسنیپ‌های همبسته می‌پردازیم.

کاربردهای افراز بلوکی هاپلوتیپ‌ها

علیرغم عدم توافق بر روی یک مدل استاندارد برای تعریف بلوک‌های هاپلوتیپ، در اختیار داشتن هر افرازی از ژنوم که بلوک‌ها در آن تا حدودی، نواحی با LD بالا را معین کنند می‌تواند کاربردهای متنوعی در زمینه‌های مختلف تحقیقاتی داشته باشد. مطالعه‌ی تنوع هاپلوتیپ‌ها و گستردگی LD در بین نژادها و زیرگروه‌های جمعیتی مختلف یکی از زمینه‌های مهم تحقیقاتی در ژنتیک جمعیت است که به یافتن بسیاری از پرسش‌های رایج درباره‌ی منشاء انسان و مهاجرت نژادهای مختلف کمک می‌کند[۱۰۱، ۱۰۴، ۱۱۷، ۱۱۸].

برای تعیین فاز ژنوتیپها در مقیاس ژنومی، هم از جنبه‌ی تئوری و هم از جنبه‌ی محاسباتی لازم است یک ساختار بلوکی از ژنوم در اختیار باشد تا روال محاسباتی تفکیک ژنوتیپ‌ها، با محدود کردن آنها درون هر یک از بلوکها انجام شود. مزیت افراز ژنوتیپ‌ها پیش از اجرای محاسبات مربوط به تعیین فاز در این است که به دلیل کوتاهتر بودن طول ژنوتیپ‌ها پس از افراز، روال تعیین فاز با پیچیدگی محاسباتی کمتری مواجه خواهد بود که به نوبه‌ی خود می‌تواند باعث کاهش خطای محاسباتی در تعیین فاز نیز گردد. از این رو در بیشتر بسته‌های نرم‌افزاری مربوط به تحلیل داده‌های هاپلوتیپ، مانند Haploview[۱۱۴]، روال افراز بلوکی هاپلوتیپها اولین مرحله، پیش از دیگر مراحل تحلیل است. از جمله دیگر کاربردهای افراز بلوکی هاپلوتیپ‌ها این است که در دست داشتن اطلاعات مربوط به ساختار بلوکی هاپلوتیپ‌ها در یک زیرجمعیت معین می‌تواند به تعیین ژنوتیپ‌های دیگر افراد همان زیرجمعیت با دقتی مطلوب و هزینه‌ی کمتر کمک کند[۱۱۹].

این که اطلاعات وراثتی در نقاط مختلفی از کروموزم‌ها و با فواصل معینی از یکدیگر قرار گرفته‌اند حتی پیش از کشف ساختار DNA و به واسطه‌ی کارهای مورگان[۶۴] و شاگردانش از حدود دهه‌ی ۱۹۲۰ میلادی به

[۶۴] Morgan

بعد، موضوعی شناخته شده بود. بدیهی است احتمال نوترکیبی بین دو نقطه‌ی مفروض بر روی یک کروموزم با افزایش فاصله‌ی آنها از یکدیگر، افزایش می‌یابد. این اصل ساده، مبنای شیوه‌ی کلاسیک اندازه‌گیری فاصله‌ی جایگاه‌های ژنی از یکدیگر قرار گرفت. بدین ترتیب که با شمارش تعداد زادهای نوترکیب از دو ژن متفاوت و محاسبه‌ی نسبت آن به دیگر زادها، احتمال نوترکیبی در فاصله‌ی بین دو ژن برآورد می‌شود و مقدار مقیاس‌بندی شده‌ای بر حسب لگاریتم این احتمال به عنوان فاصله‌ی ژنتیکی[۶۵] دو ژن در نظر گرفته می‌شود. برمبنای اطلاعات جدید ما از ساختار [۶۶]، اینک می‌توانیم فاصله‌ی جایگاه‌های ژنی از یکدیگر را برحسب تعداد نوکلئوتیدهای قرار گرفته بین آنها بر روی ژنوم برآورد کنیم که به آن فاصله‌ی فیزیکی[۶۷] می‌گویند. تفاوت نرخ نوترکیبی در طول ژنوم، باعث می‌شود بین فاصله‌ی ژنتیکی و فاصله‌ی فیزیکی نسبت ثابتی برقرار نباشد. بر همین اساس، نرخ نوترکیبی در نقاط مختلف ژنوم با واحد cM/Mb[۶۸] اندازه‌گیری می‌شود که در آن cM واحد اندازه‌گیری فاصله‌ی ژنتیکی و معادل با یک رویداد نوترکیبی در یکصد نسل است.

این نظریه که در برخی جایگاه‌های خاص ژنوم به دلایل صرفا مولکولی احتمال بیشتری برای رویداد نوترکیبی وجود دارد دارای طرفداران بسیاری است[۱۲۰، ۱۲۱]. از دیدگاه ژنتیک آماری، برای چنین جایگاه‌هایی باید دست کم در بین برخی زیرگروه‌های جمعیتی شواهد معتبری نشان‌دهنده‌ی فراوانی بالای رویداد نوترکیبی در مقایسه با دیگر نواحی ژنوم وجود داشته باشد. چنین جایگاه‌هایی را اصطلاحا نقاط پراحتمال [۶۹] نوترکیبی می‌نامیم. تاکنون مدلهای آماری پیچیده‌ای برای تعیین تغییرات نرخ نوترکیبی در امتداد ژنوم و بدست آوردن نقاط پراحتمال نوترکیبی معرفی شده‌اند[۱۲۲—۱۲۹]. متاسفانه محاسبات لازم در این مدل‌ها، به ویژه برای برآورد نرخ نوترکیبی و یافتن نقاط پراحتمال در کل ژنوم، با صرف زمان بسیار طولانی به انجام می‌رسد. از این رو، غالبا نمی‌توان از این مدل‌ها به طور عملی جز برای طول‌های کوتاهی از ژنوم استفاده کرد. در عوض، مرزهای بلوکهای هاپلوتیپ می‌توانند خلاصه‌ای بسیار ساده از مکان تقریبی نقاط پراحتمال نوترکیبی را فراهم کنند. در بخش ۶۰۵۰۲، توانائی روش‌های افراز بلوکی هاپلوتیپ‌ها در این کاربرد خاص با جزئیاتی بیشتری مورد بررسی قرار می‌گیرد.

[۶۵] genetic distance
[۶۶] DNA
[۶۷] physical distance
[۶۸] centi Morgan per Mega bases
[۶۹] hotspot

بلوک‌های هاپلوتیپی کاربرد گسترده‌ای در مطالعات مربوط به شناسائی زمینه‌های ژنتیکی بیماری‌ها دارند. شرح بیشتر این کاربرد، موضوع بخش ۵۰۱ است. شیوه‌ی جدیدی برای بکارگیری اطلاعات ساختار بلوکی هاپلوتیپ‌ها در مطالعه‌ی بیمارها بر پایه‌ی داده‌های case-control در بخش ۷۰۵۰۲ مورد بحث و ارزیابی قرار می‌گیرد. مسئله‌ی انتخاب تگ‌اسنیپ‌ها که ادامه به بحث درباره‌ی آن می‌پردازیم نیز از مسائلی است که فرض در اختیار داشتن ساختار بلوکی در آن، نقش تعیین‌کننده‌ای دارد.

انتخاب تگ‌اسنیپ‌ها

با افراز ژنوم به بلوک‌هایی با LD بالا، هاپلوتیپ‌هایی تعریف می‌شوند که پیچیدگی اطلاعات در آنها به میزان قابل توجهی پائین است. بر این اساس می‌توان اطلاعات کل هاپلوتیپ‌ها در سراسر ژنوم را تنها با بررسی هاپلوتیپ‌های تعداد اندکی از اسنیپ‌ها بدست آورد. این اسنیپ‌ها را اصطلاحا تگ‌اسنیپtagSNP می‌گوئیم. تگ‌اسنیپ‌ها به کاهش هزینه و زمان لازم برای تعیین توالی ژنومی افراد کمک می‌کنند و در واقع، شیوه‌ای کارآمد برای فشرده‌سازی اطلاعات ژنومی هستند.

تگ‌اسنیپ‌ها معمولا از دو دیدگاه متفاوت مورد توجه قرار می‌گیرند که هر یک مشابه یکی از رویکردهای رایج در تعیین بلوک‌های هاپلوتیپ است. در دیدگاه مبتنی بر همبستگی جفتی اسنیپ‌ها، اسنیپی که همبستگی بالایی با دیگر اسنیپ‌ها دارد می‌تواند به عنوان نماینده‌ی دیگر اسنیپ‌ها و به جای آنها در آزمون‌هایی که برای مطالعه‌ی همبستگی بین فنوتیپ بیماری و اسنیپ‌ها مورد بررسی قرار می‌گیرند بکار گرفته شود. چنین اسنیپ‌هایی را اسنیپ ـ تگ‌اسنیپ[۷۰] می‌گوئیم. برآورد می‌شود بیش از ۹۰درصد از اسنیپ‌های کل ژنوم در زیرجمعیت‌های اروپائی و آسیایی مورد مطالعه‌ی پروژه‌ی HapMap در همبستگی بالا ($r^۲ \geq ۰/۸$) با دیگر اسنیپ‌ها هستند. بر مبنای همین مطالعات، اگر بخواهیم برای هر اسنیپ، نماینده‌ای با همبستگی $r^۲ \geq ۰/۸$ به عنوان تگ‌اسنیپ داشته باشیم از بین۳/۱ میلیون اسنیپی که در فاز دوم پروژه‌ی HapMap تعیین ژنوتیپ شده‌اند تنها ۵۰۰ هزار تگ‌اسنیپ برای زیرجمعیت‌های اروپایی و آسیایی و نزدیک به یک میلیون تگ‌اسنیپ برای زیرجمعیت آفریقایی کافی است[۱۴].

[۷۰] SNP tagging SNP, SNP-tagSNP

اصولا علاقمندیم از بین مجموعه‌های محتلف برای تگ‌اسنیپ‌ها، مجموعه‌ای را انتخاب کنیم که کمترین تعداد تگ‌اسنیپ را داشته باشد. هرچند بدست آوردن جواب بهینه به طور دقیق برای این مسئله یک الگوریتم NP-hard است با این حال رویکردهای ساده‌سازی شده گوناگونی برای این مسئله ارائه شده‌اند که معروف‌ترین آنها الگوریتم منسوب به کارلسون و دیگران[۱۳۰] است. در روش کارلسون برای انتخاب تگ‌اسنیپ‌ها، یک الگوریتم حریصانه برای نزدیک شدن به مجموعه بهینه‌ی تگ‌اسنیپ‌ها مورد استفاده قرار می‌گیرد. در این الگوریتم طی یک روال تکرارشونده، از بین اسنیپ‌های موجود، اسنیپی که با تعداد بیشتری از اسنیپ‌ها در همبستگی بالاست به عنوان تگ‌اسنیپ انتخاب می‌شود و این فرآیند پس از حذف این اسنیپ و اسنیپ‌های همبسته با آن از مجموعه‌ی اولیه تکرار می‌شود. در این الگوریتم، وجود همبستگی از طریق اعمال یک مقدار آستانه‌ای معین بر روی $r^۲$ تعریف می‌شود. استفاده از ماتریس کوواریانس و آنالیز PCA نیز یکی دیگر از روشهای رایج در انتخاب اسنیپ‌ـ تگ‌اسنیپ‌ها هستند[۱۳۱، ۱۳۲].

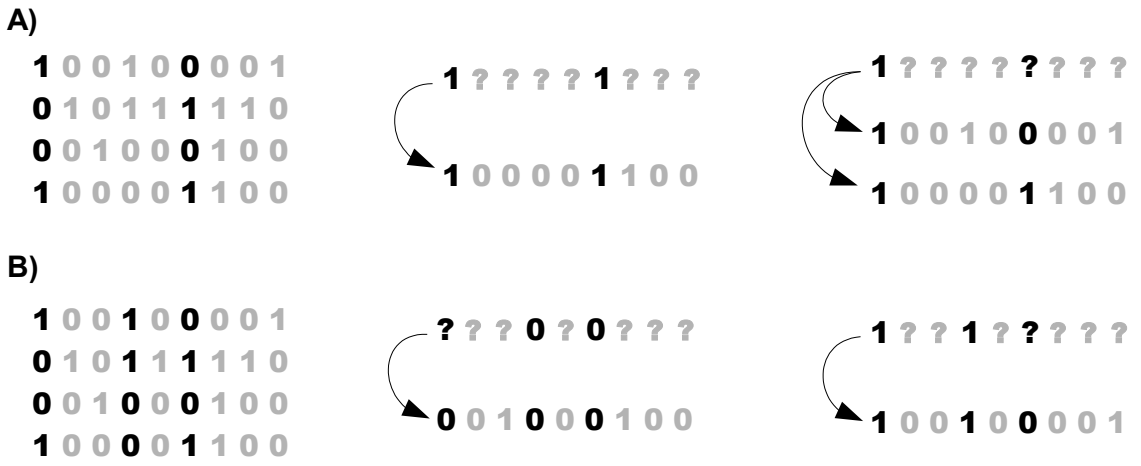
در دیگر رویکرد رایج، مفهوم هاپلوتیپ‌ـ تگ‌اسنیپ[۷۱] مورد توجه قرار می‌گیرد. اصولا در هر بلوک هاپلوتیپی، مجموعه‌ی خاصی از اسنیپ‌ها را می‌توان یافت که هاپلوتیپ‌های تحدید شده به این مجموعه، تمام تنوع هاپلوتیپی موجود در بلوک را پوشش می‌دهند. به عبارت مشخص، در اینجا هدف یافتن زیرمجموعه‌ای از اسنیپ‌ها است که موقعیت هر تفاوت یک نوکلئوتیدی بین هر دو هاپلوتیپ دلخواه در نمونه، توسط یکی از اسنیپ‌های این مجموعه نشان داده می‌شود. این مجموعه را اصطلاحا مجموعه‌ی هاپلوتیپ‌ـ تگ‌اسنیپ‌ها یا اختصارا مجموعه‌ی htSNPها می‌نامیم. مطلوب در اینجا نیز بدست آوردن کمترین تعداد htSNP لازم برای ناحیه‌ی ژنومی مورد مطالعه است. این مسئله را می‌توان به شکل یک مسئله‌ی برنامه‌ریزی خطی ۰ و ۱ به صورت زیر بیان کرد:

$$\begin{aligned} & \min \sum_{j=۱}^{l} x_j \\ s.t. \quad & \sum_{j:h_{i_۱}(j)\neq h_{i_۲}(j)} x_j \geq k, \quad ۱ \leq i_۱ < i_۲ \leq n \\ & x_j \in \{۰,۱\}, \quad j = ۱,\ldots,l \end{aligned}$$

که در آن x_j نشاندهنده‌ی انتخاب اسنیپ jام به عنوان تگ‌اسنیپ است و $h_{i_۱}(j)$ و $h_{i_۲}(j)$ به ترتیب

[۷۱] Haplotype tagging SNP, htSNP

آلل‌ها اسنیپ j در هاپلوتیپ i_1 و i_2 هستند و k یک عدد صحیح ثابت و دلخواه است. جواب این مسئله‌ی برنامه‌ریزی خطی، وقتی $k = 1$، یک جواب برای مسئله‌ی متعارف انتخاب تگ‌اسنیپ‌هاست. جواب‌هایی که به ازای $k \geqslant 2$ بدست می‌آیند در واقع تگ‌اسنیپ‌هایی را معین می‌کنند که می‌توانند بر پایه‌ی هاپلوتیپهای نمونه‌ی اولیه، هاپلوتیپ کامل یک نمونه جدید را حتی با وجود $k-1$ آلل مفقود بازسازی کنند(شکل ۱۲۰۱).



شکل ۱۲۰۱: انتخاب هاپلوتیپ‌ـ تگ‌اسنیپ‌ها

A) نمونه‌ای از هاپلوتیپ‌های متفاوت بر روی نه موقعیت اسنیپ (چپ). اگر این چهار هاپلوتیپ، نماینده‌ی تمام هاپلوتیپ‌های جمعیت مورد مطالعه باشند آنگاه دانستن ژنوتیپ تنها دو اسنیپ (اسنیپ‌های اول و ششم) برای شناسائی هاپلوتیپ کامل یک نمونه‌ی جدید کافی است که به این اسنیپ‌ها، هاپلوتیپ‌ـ تگ‌اسنیپ می‌گوئیم (وسط). اگر ضعف‌های آزمایشگاهی یا محاسباتی باعث شود ژنوتیپ یکی از تگ‌اسنیپ‌ها نامعلوم باقی بماند امکان شناسائی هاپلوتیپ کامل به طور منحصر بفرد از بین می‌رود (راست). **B)** با افزودن یک اسنیپ خاص به مجموعه‌ی قبلی تگ‌اسنیپ‌ها، بازسازی هاپلوتیپ نمونه‌ی جدید حتی با وجود حداکثر یک اسنیپ مفقود امکان‌پذیر می‌شود.

از دیدگاه تئوری، تنوع هاپلوتیپی در ناحیه‌ای شامل l اسنیپ و n هاپلوتیپ متمایز را می‌توان با تعدادی بین $\log_2 n$ تا l هاپلوتیپ‌ـ تگ‌اسنیپ پوشش داد. در عمل، در بلوک‌هایی با تنوع هاپلوتیپی پائین، تمام تنوع هاپلوتیپی موجود در بلوک توسط تعداد اندکی از htSNPها پوشش داده می‌شود. مسئله‌ی بدست آوردن کمترین تعداد htSNP لازم برای تعیین تنوع هاپلوتیپی در یک ناحیه‌ی ژنومی نیز در حالت کلی NP-hard و در عمل بدست آوردن جواب بهینه به طور دقیق، جز برای نواحیی با اسنیپ‌های کم امکان‌پذیر نیست. در واقع، می‌توان نشان داد که پیچیدگی محاسباتی این مسئله تابعی از واگرائی هاپلوتیپ‌های درون بلوک است. بنابراین، افراز ژنوم به بلوک‌هایی با واگرائی هاپلوتیپی پائین برای تعیین کمترین تعداد htSNPهای مورد نیاز برای کل ژنوم، یک پروسه‌ی ضروری است. بر پایه‌ی همین استدلال، ژانگ و دیگران[۱۰۸] قیود متنوعی را برای تعیین

بلوک‌های محتمل در نرم‌افزار خود HapBlock به کار گرفته‌اند. آنها با استفاده از یک الگوریتم برنامه‌ریزی پویا، افراز بلوکی بهینه‌ای را از بین بلوک‌های محتمل بدست می‌آورند که توسط آن، مجموع تگ‌اسنیپ‌های لازم برای تعیین هاپلوتیپ‌های سراسر ناحیه‌ی مورد بررسی کمینه می‌شود. بی‌شک، نحوه‌ی تعیین بلوک‌های هاپلوتیپ و شکل افراز بلوکی بر تعداد htSNPهای مورد نیاز در ناحیه‌ی مورد بررسی تاثیر می‌گذارد. این موضوع در تحقیق دینگ و همکارانش[۱۳۳] به بحث گذاشته شده است.

دینگ و همکارانش همچنین، بسته‌ی نرم‌افزاری جامعی با نام htSNPer[۱۳۴] را تولید کردند که در آن علاوه بر روال‌های مورد نیاز برای تعیین فاز ژنوتیپ‌ها و الگوریتم‌های مختلف افراز بلوکی هاپلوتیپ‌ها، یک الگوریتم توسعه و تحدید نیز برای محاسبه‌ی دقیق مجموعه‌ی بهینه‌ی htSNPها پیاده‌سازی شده است. با آنکه پیچیدگی محاسباتی الگوریتم انتخاب تگ‌اسنیپ‌ها در نرم‌افزار htSNPer نیز مانند دیگر الگوریتم‌های این مسئله، از مرتبه‌ی نمائی است اما در عمل سریعترین الگوریتم شناخته‌شده برای تعیین مجموعه‌ی بهینه‌ی htSNPها تا کنون است.

از دیدگاه عملی ممکن است قیود دیگری به جز ویژگی‌های مطلوبی که در بالا به آنها اشاره شد برای تعریف تگ‌اسنیپ‌های بکار گرفته شود. از این میان می‌توان به پراکندگی یکنواخت یا نسبتا یکنواخت تگ‌اسنیپ‌ها بر روی ژنوم به عنوان یک ویژگی مطلوب در برخی روشهای انتخاب تگ‌اسنیپ اشاره کرد[۱۳۵—۱۳۷]. توزیع پسینی که توسط روش‌های بیزی برای استنباط هاپلوتیپ‌ها از نمونه‌های ژنوتیپ محاسبه می‌شود نیز از جمله اطلاعات مفیدی است که می‌توان از آن برای تخمین واگرائی هاپلوتیپ‌های جمعیت و انتخاب تگ‌اسنیپ‌ها استفاده کرد[۱۳۸].

۵۰۱ شناسائی جایگاه ژنی خصیصه

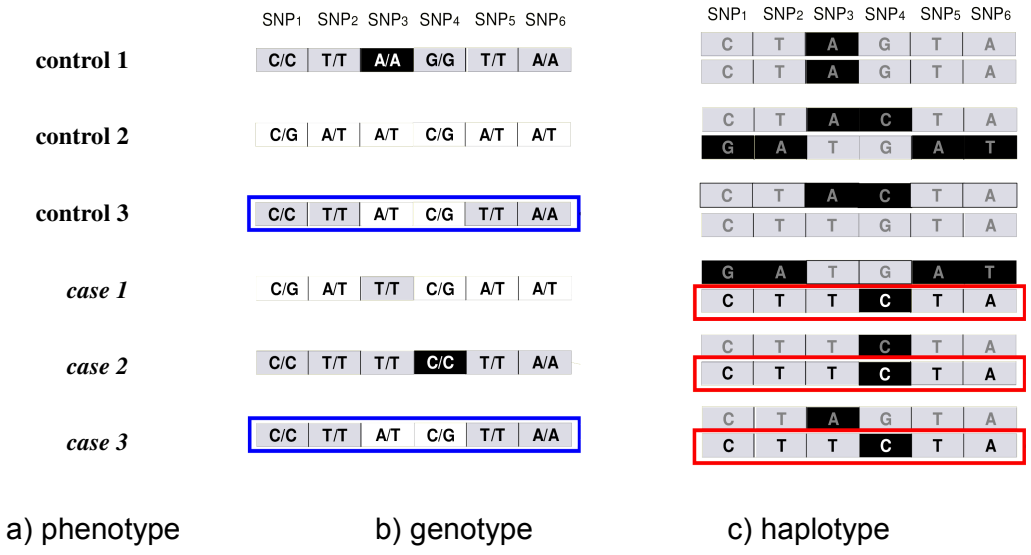
تا کنون بیماری‌های زیادی شناسائی شده‌اند که عامل اصلی بروزشان را ناشی از حضور یک آلل خاص در یک جایگاه ژنی می‌دانیم. اصول ژنتیک مندلی، مدل ساده‌ای برای کشف ژن مرتبط با این گونه بیماریها را برای ما فراهم می‌کند. بیماری‌های مثل Cystic fibrosis، Tuberous sclerosis، Sickle cell anemia و Huntington از این دست هستند. با این حال اصول نظریه‌ی ژنتیک به ما یادآوری می‌کند که نه تنها این

نوع از بیماری‌ها بلکه هر خصیصه‌ای[۷۲]، خواه پیچیده یا ساده، باید دست کم در برخی سطوح ناشی از ژنوتیپ خاص فرد باشد. بدین ترتیب، نه تنها رنگ پوست و اندازه‌ی قامت بلکه میزان استعداد ابتلا به بسیاری از بیماری‌ها، حتی آنهایی که به دلایل میکروبی بروز می‌کنند نیز مرتبط با ژنوتیپ خاص افراد است. پیچیدگی این موضوع در این است که بسیاری از بیماری‌ها و استعداد ابتلا به آنها و نیز بسیاری از فنوتیپ‌های ظاهری افراد، برآیند همزمان تاثیرات محیطی و عملکرد دسته‌ای از چندین ژن بخصوص است. مطالعه‌ی ژنتیکی بیماری‌ها [۷۳] در واقع جستجوی عامل ژنتیکی مرتبط با بیماری از طریق انجام یک سری آزمونهای آماری برای بررسی همبستگی بین خصیصه‌های بیماری و ژنوتیپ‌های شناخته شده‌ی جمعیت است.

بی‌شک مطالعه‌ی ژنوتیپ‌ها حتی در وضعیت تعیین فاز نشده[۷۴] می‌تواند اطلاعات مفیدی درباره‌ی شناخت زمینه‌های ژنتیکی خصیصه‌ها و بیماری‌ها به ما بدهد اما، از آنجاکه یک فنوتیپ در پایه‌ای‌ترین سطح، محصول کارکرد زیست‌شناسانه‌ی یک توالی ژنی هاپلوئید است مطالعه‌ی ژنوتیپ‌های تفکیک شده، یعنی هاپلوتیپ‌ها، به طور مستقیم اطلاعات بهتری برای شناخت زمینه‌ی ژنتیکی خصیصه‌ی مورد بررسی در اختیار ما می‌گذارد. به بیان دیگر، ممکن است ژنی که از طریق تنها یکی از هاپلوتیپ‌ها بیان می‌شود زمینه‌ی ابتلا به بیماری را به وجود آورده باشد. از این رو ممکن است ژنوتیپ یک فرد مبتلا و دسته‌ای دیگر از افراد غیر مبتلا در ظاهر یکسان باشند و به همین دلیل نتوان جایگاه ژنی مرتبط با بیماری را به سادگی تشخیص داد (شکل ۱۳۰۱). بنابراین، تفکیک ژنوتیپ‌ها به هاپلوتیپ‌های تشکیل‌دهنده‌یشان و استنباط هاپلوتیپ‌ها یک پروسه‌ی لازم پیش از پرداختن به مسئله‌ی مطالعه‌ی جایگاه ژنی یک بیماری است.

رایج‌ترین و در عین حال ساده‌ترین مدل ژنتیکی بیماری‌ها بر این فرضیه استوار است که هر خصیصه‌ی بیماری، تنها با یک جایگاه ژنی در ارتباط است. زمینه‌ی ژنتیکی تعدادی از بیماری‌ها مانند Cystic fibrosis به درستی توسط این مدل تبیین می‌شوند و جایگاه‌های ژنی مرتبط با آنها به خوبی شناخته شده است. پایگاه داده‌ای OMIM جامع‌ترین مرجع اطلاعاتی در زمینه‌ی جایگاه‌های ژنی شناخته شده‌ی مرتبط با بیماری‌ها است. در این پایگاه داده‌ای علاوه بر بیماری‌های تگ‌جایگاهی[۷۵]، هر مکانی از ژنوم که توسط یک تحقیق علمی، احتمال

[۷۲] trait
[۷۳] Genetic disease study
[۷۴] unphased
[۷۵] single locus



شکل ۱۳۰۱: شناسائی هاپلوتیپ مرتبط با بیماری در بین نمونه‌های case و control
a) فنوتیپ‌های مشاهده شده در شش فرد مورد مطالعه. نمونه‌های case مبتلا به بیماری هستند یا نشانه‌های آن را به همراه دارند و نمونه‌های control افرادی هستند که به عنوان غیرمبتلا به بیماری شناخته شده‌اند. **b)** بدون تفکیک ژنوتیپ‌ها به هاپلوتیپ‌ها، دو فنوتیپ متفاوت در ارتباط با دو ژنوتیپ ظاهرا یکسان مشاهده می‌شوند (ژنوتیپ‌های داخل کادر). **c)** با شناسائی هاپلوتیپ‌های واقعی تشکیل‌دهنده‌ی ژنوتیپ‌ها در هر فرد، هاپلوتیپ مرتبط با بیماری تشخیص داده می‌شود.

ارتباط آن با یک خصیصه نشان داده شده باشد نیز فهرست شده است. در واقع بروز بسیاری از خصیصه‌ها، برآیند همکاری چندین ژن با یکدیگر و تاثیر آنها بر میزان بیان دیگر ژن‌ها است. چنین خصیصه‌هایی را چندجایگاهی[۷۶] می‌نامند. در حالت کلی، ژنهای همدست در بروز یک خصیصه‌ی معین ممکن است در فواصل بسیار دور از یکدیگر بر روی ژنوم قرار گرفته باشند. نحوه‌ی همدستی ژن‌های مرتبط با یک خصیصه نیز می‌تواند بسیار پیچیده باشد. به عنوان مثال، در تعامل epistasy بین دو ژن، بیان یکی از ژن‌ها باعث خاموش شدن ژن دیگر می‌شود.

در مجموع، وجود روابط پیچیده در عملکرد و تجلی ژن‌های مرتبط با یک خصیصه، این ایده را مطرح می‌کند که بین مشاهده‌ی یک خصیصه در یک فرد و داشتن یک ژنوتیپ خاص، نه یک رابطه‌ی مستقیم و صد در صدی علت و معلولی، بلکه تنها روابطی مبتنی بر مدل‌های احتمالاتی می‌تواند وجود داشته باشد. در ادامه، به توضیح این روابط در ساده‌ترین مدل، یعنی مدل تک‌جایگاهی می‌پردازیم.

اگر برای ژن یا اسنیپ مرتبط با بیماری، دو آلل A و a را فرض کنیم که به ترتیب نشاندهنده‌ی آلل وحشی

[۷۶] multi-locus

و آلل جهش‌یافته هستند آنگاه احتمال مشاهده‌ی بیماری به شرط داشتن هر یک از سه ژنوتیپ AA، Aa و aa را اصطلاحا نفوذ[۷۷] بیماری به ازای هر یک از این ژنوتیپ‌ها می‌گویند. شیوع بیماری[۷۸] به معنای فراوانی نسبی یا احتمال مشاهده‌ی نمونه‌ی مبتلا به بیماری در کل جمعیت است. در مدل تک‌جایگاهی داریم:

$$\text{Prevalence} = P(\text{affected}) = P(AA).P_{\text{affected}|AA} + P(Aa).P_{\text{affected}|Aa} + P(aa).P_{\text{affected}|aa} \quad (۱۰۰۱)$$

که در آن هر یک از $P_{\text{affected}|AA}$، $P_{\text{affected}|Aa}$ و $P_{\text{affected}|aa}$ نشاندهنده‌ی نفوذ بیماری به ازای یک ژنوتیپ و هر یک از $P(AA)$، $P(Aa)$ و $P(aa)$ احتمال مشاهده‌ی یک ژنوتیپ در جمعیت است . فراوانی نسبی ژنوتیپ‌ها در جمعیت را معمولا می‌توان با تعیین ژنوتیپ در نمونه‌های پراکنده، برآورد کرد. همچنین برآورد شیوع بیماری از طریق نمونه‌گیری امکان‌پذیر است. برای برآورد مقادیر نفوذ، علاوه بر رابطهٔ (۱۰۰۱)، به رابطه‌ی دیگری نیاز است که از طریق آن میزان نفوذ در ژنوتیپ‌های مختلف بر حسب پارامترهای ساده‌تری بیان گردد. به عنوان مثال، دو مدل کلاسیک بیماری یعنی مدل غالب[۷۹] و مدل مغلوب[۸۰] به ترتیب متناظر با حالت‌های زیر هستند؛

Dominant. $P_{\text{affected}|AA} = ۰, \quad P_{\text{affected}|Aa} = P_{\text{affected}|aa} = \theta$

Recessive. $P_{\text{affected}|AA} = P_{\text{affected}|Aa} = ۰, \quad P_{\text{affected}|aa} = \theta$

علاوه بر آن، مدل جمعی[۸۱] و مدل ضربی[۸۲] نیز از جمله مدل‌های رایج برای تعریف تاثیر تجلی ژنوتیپ‌ها بر یکدیگر هستند. هر یک از این دو مدل بر حسب دو پارامتر آزاد و به صورت زیر تعریف می‌شوند؛

Additive. $P_{\text{affected}|AA} = \alpha, \quad P_{\text{affected}|Aa} = \alpha + \beta, \quad P_{\text{affected}|aa} = \alpha + ۲\beta$

Multiplicative. $P_{\text{affected}|AA} = \alpha, \quad P_{\text{affected}|Aa} = \alpha\beta, \quad P_{\text{affected}|aa} = \alpha\beta^{۲}$

به طور کلی با معلوم بودن میزان شیوع بیماری و فراوانی نسبی هر یک از سه ژنوتیپ ممکن و با توجه به رابطهٔ (۱۰۰۱)، در هر مدل دلخواه برای تبیین اثر ژنوتیپ‌های مختلف بر بیان یکدیگر و بروز بیماری، تنها

[۷۷] Penetrance
[۷۸] Prevalence
[۷۹] dominant
[۸۰] recessive
[۸۱] additive
[۸۲] multiplicative

دو پارامتر آزاد کافی است. بر پایه‌ی همین استدلال و در رویکردی دیگر، مفهوم «ریسک نسبی ژنوتیپ»[۸۳] برای بیان روابط احتمالاتی بین ژنوتیپ و فنوتیپ در یک بیماری تک‌جایگاهی بکار گرفته می‌شود. ریسک نسبی یک ژنوتیپ، نسبت خطر ابتلا به بیماری در افرادی دارای این نوع ژنوتیپ به خطر ابتلا به بیماری در افراد هموزیگوت با ژنوتیپ AA است. بدین ترتیب دو ریسک نسبی مستقل داریم که به ترتیب زیر تعریف می‌شوند؛

$$GRR_1 = \frac{P_{\text{affected}|Aa}}{P_{\text{affected}|AA}}, \qquad GRR_2 = \frac{P_{\text{affected}|aa}}{P_{\text{affected}|AA}} \tag{۱۱۰۱}$$

در اینجا فرض کرده‌ایم که افراد هموزیگوت با ژنوتیپ AA کم خطرترین ژنوتیپ مرتبط با ابتلای به بیماری را دارند و از این رو $GRR_2 \geqslant GRR_1 \geqslant 1$ است. برای داشتن درکی بهتر از ریسک نسبی، بد نیست بدانید که ریسک نسبی ابتلا به بیماری هانتینگتون در فرزندان یک فرد مبتلا به این بیماری تا یکهزار بار بیشتر از افرادی است که رابطه‌ی خویشاوندی با هیچ فرد مبتلا به هانتینگتون ندارند. این نسبت برای اوتیسم نزدیک به ۷۵، آسم نزدیک به ۶ و دیابت دیر–پیشرونده[۸۴] بین ۲ تا ۳ برآورد می‌شود[۱۳۹]. مدل‌های کلاسیک در توارث بیماری‌ها را نیز، می‌توان بر حسب ریسک نسبی ژنوتیپ‌ها بیان کرد. در این حالت، تنها یک پارامتر آزاد در تعریف هر یک از مدل‌ها ظاهر می‌شود (جدول ۱۰۱).

جدول ۱۰۱: برخی مدل‌های رایج در توارث بیماری‌ها بر حسب ریسک نسبی ژنوتیپ‌ها

Genetic disease model	GRR_1	GRR_2
Recessive	1	γ
Dominant	γ	γ
Additive	γ	$2\gamma - 1$
Multiplicative	γ	γ^2

جستجوی همبستگی بین فنوتیپ بیماری و ژنوتیپ‌های نواحی مختلف در سراسر ژنوم، اصطلاحا «بررسی همبستگی در مقیاس ژنومی»[۸۵] و شناسائی جایگاه ژن مرتبط با بیماری، «نگاشت ژن بیماری»[۸۶] نامیده می‌شوند. رهیافت متعارف در این نوع مطالعات، جمع‌آوری نمونه‌های متعدد از افراد مبتلا و غیرمبتلا به بیماری و استفاده از آزمون‌های آماری برای بررسی همبستگی بین ژنوتیپ‌ها و فنوتیپ بیماری است. در این رویکرد خاص که

[۸۳] Genotype Relative Risk, GRR
[۸۴] late-onset
[۸۵] Genome-wide association study
[۸۶] Disease gene mapping

اصطلاحا case-control study نامیده می‌شود، ژنوتیپ افراد در هر دو گروه، در نقاط مختلفی از ژنوم مورد بررسی قرار می‌گیرد. علیرغم پیشرفت در فنآوری‌های توالی‌یابی، هزینه‌ی تعیین توالی کامل ژنوم برای تعداد زیادی از افراد، همچنان فرآیندی پر هزینه و زمان‌بر است. از این رو به طور متعارف، تنها به تعیین ژنوتیپ در تعدادی از جایگاه‌های پراکنده اکتفا می‌شود. موقعیت این جایگاه‌ها در وهله‌ی اول بستگی به وجود ژن‌های قابل افتراق در آنها در بین افراد مختلف جمعیت و نیز امکانات آزمایشگاهی لازم برای تعیین ژنوتیپ دارد. این جایگاه‌ها را اصطلاحا «نشانگذار»[۸۷] می‌نامند. انتخاب مناسب نشانگذارها می‌تواند نقش تعیین‌کننده‌ای در رسیدن به نتایج دقیق و معتبر داشته باشد. در این شیوه، تنها زمانی موفق به یافتن یک جایگاه ژنی ناشناخته، مرتبط با بیماری مورد مطالعه خواهیم بود که این ژن جدید به تصادف در نزدیکی یکی از نشانگذارها قرار داشته باشد. از این رو، اسنیپ‌ها به دلیل فراوانی و پراکندگی گسترده‌یشان بر روی ژنوم، انتخاب مناسبی برای استفاده به عنوان نشانگذارها هستند و به سادگی امکان بدست آوردن یک نگاشت دقیق[۸۸] برای ژن مرتبط با بیماری را فراهم می‌کنند. در واقع، بخشی از اهداف پروژه‌ی HapMap و نیز ایده‌ی توسعه‌ی فنآوری‌هایی چون SNP-microarray بر پایه‌ی همین استدلال استوار است. از سویی دیگر انتخاب تگ‌اسنیپ‌ها (بخش ۴۰۱) در مناطقی با گستره‌ی بالای LD، بررسی همبستگی در مقایس ژنومی را با هزینه‌ی کمتر و در زمان کوتاهتر امکان‌پذیر می‌سازد.

علاوه بر طرح case-control برای مطالعه‌ی همبستگی ژنتیکی خصیصه که در آن فنوتیپ خصیصه یک مقدار دو ارزشی فرض می‌شود مدل دیگری، مبتنی بر خصیصه‌های کمّی[۸۹] نیز رواج دارد که در آن فنوتیپ هر فرد بر حسب یک کمیت عددی اندازه‌گیری شده بیان می‌شود. از جنبه‌ای دیگر، نمونه‌های مورد استفاده در مطالعه می‌توانند مربوط به افراد خویشاوند یا غیرخویشاوند جمعیت باشند. آنالیز رایج در مورد اول، یعنی وقتی از نمونه‌های خویشاوند در مطالعه استفاده می‌شود آنالیز بهم‌پیوستگی[۹۰] است که در آن ارتباط بین خصیصه و جایگاه ژنی از طریق بیشینه‌سازی یک تابع درستنمائی که بر اساس روابط درون شجره‌ی خانوادگی تعریف شده‌است بررسی می‌گردد. استفاده از آزمون موسوم به TDT[۹۱]، روش ساده و متداولی برای آنالیز همبستگی در

[۸۷] marker
[۸۸] fine map
[۸۹] quantitative trait
[۹۰] Linkage analysis
[۹۱] transmission-disequilibrium test, TDT

مواردی است که اطلاعات خویشاوندی تنها به صورت سه‌تائی‌های فرزند–والد در اختیار است. برای آشنائی بیشتر با دیگر روش‌ها و طرح‌های مطالعه‌ی ژنتیکی بر پایه‌ی نمونه‌های خویشاوند، مرجع [۱۴۰] را ببینید.

علیرغم برخی ویژگی‌های مطلوب در مطالعات ژنتیکی مبتنی بر نمونه‌های خویشاوند، در عمل جمع‌آوری نمونه‌های خویشاوند به ویژه به دلیل طبیعت دیر–پیشرونده‌ی بسیاری از بیمارها با مشکل روبرو است[۷۰]. از این رو روشهای مبتنی بر نمونه‌های غیرخویشاوند جمعیت بیشتر مورد توجه محققین قرار گرفته‌اند. دو رویکرد در آنالیزهای مبتنی بر نمونه‌های جمعیت[۹۲]، رایج است؛ رویکرد مبتنی بر آنالیز ژنوتیپ نشانگذارها به طور منفرد و رویکرد مبتنی بر آنالیز هاپلوتیپ‌های چندجایگاهی. تحقیقات نشان داده است که آزمون‌های همبستگی مبتنی بر هاپلوتیپ‌ها در مقایسه با آزمون‌های همبستگی تک‌جایگاهی از دقت و توان بالاتری برخوردارند[۱۴۱، ۱۴۲]. رده‌ی بزرگی از روش‌ها، آزمون‌های مبتنی بر مقایسه‌ی فراوانی آلل‌های متفاوت یک جایگاه ژنی یا هاپلوتیپ‌های متفاوت در بین نمونه‌های مبتلا (case) و غیرمبتلا (control) است[۱۴۳]. از این بین رایج‌ترین آزمون برای بررسی وجود همبستگی بین ناحیه‌ی مورد مطالعه و خصیصه، آزمون نیکوئی برازش[۹۳] است که در آن برابری توزیع هاپلوتیپ‌ها بین نمونه‌های case و نمونه‌های control مورد بررسی قرار می‌گیرد. معمولا از لگاریتم نسبت درستنمائی تحت فرض استقلال به درستنمائی تحت فرض برابری توزیع هاپلوتیپ‌ها در این دو دسته، به عنوان آماره‌ی آزمون استفاده می‌شود؛

$$LLR = ۲\left(\ln L_{\text{case}} + \ln L_{\text{control}} - \ln L_{\text{pool}}\right) \qquad (۱۲۰۱)$$

این آماره به طور مجانبی دارای توزیع مربع کای با $|\mathcal{H}| - ۱$ درجه آزادی است که در آن $\mathcal{H}$ مجموعه‌ی هاپلوتیپ‌های متمایز در کل نمونه است. اشکال اصلی استفاده از این آماره در مواردی است که تعداد هاپلوتیپهای متمایز در نمونه، بالا باشد که در آن صورت، توان آزمون با افزایش درجات آزادی کاهش می‌یابد. در این شرایط معمولا برآورد فراوانی هاپلوتیپ‌های نادر نیز با مشکل همراه است که این خود باعث عدم تبعیت آماره از توزیع مربع کای می‌شود. افراز بلوکی هاپلوتیپ‌ها، می‌تواند تا حدودی به رفع این مشکل کمک کند؛ چرا که تنوع هاپلوتیپی درون بلوک‌های هاپلوتیپ محدود شده، است. رهیافت دیگر برای بهبود این روش، خوشه‌بندی[۹۴] هاپلوتیپ‌های "مشابه" در دسته‌های یکسان است[۱۴۴–۱۴۷]. از نظر تئوری، رابطهٔ (۱۲۰۱)

[۹۲] Population-based analysis
[۹۳] Goodness-of-fit, GOF
[۹۴] clustering

تنها در صورت برقراری تعادل HW بین هاپلوتیپ‌های نمونه، معتبر است. این موضوع باعث محدودیت کاربرد این رابطه در شرایط کلی می‌شود. در بخش ۷۰۵۰۲، ضمن توسعه‌ی یک شیوه‌ی جدید بر اساس همین رویکرد، جزئیات بیشتری درباره‌ی محاسبه‌ی آماره‌ی آزمون در رابطهٔ (۱۲۰۱) و خوشه‌بندی هاپلوتیپ‌ها ارائه می‌شود. برای آشنائی بیشتر با شیوه‌های مطالعه‌ی همبستگی مبتنی بر هاپلوتیپ‌ها، [۱۴۸] را ببینید.

فصل ۲

مواد و روشها

۱۰۲ یک الگوریتم ژنتیک برای استنباط هاپلوتیپ‌ها

یکی از شیوه‌های رایج برای حل مسائل بهینه‌سازی، زمانی که روش‌های تحلیلی، کارآمدی مطلوب را ندارد الگوریتم‌های ژنتیکی[۱] هستند. این الگوریتم‌ها که در رده‌ی روشهای اکتشافی قرار می‌گیرند حالت خاصی از روشهای موسوم به روشهای تکاملی[۲] هستند. روشهای تکاملی به طور عمومی به روش‌هایی اطلاق می‌شود که در آنها جواب بهینه توسط مدلهای مبتنی بر اصول تکامل و انتخاب طبیعی جستجو می‌شود. در کلی‌ترین حالت، نقاط مختلف فضای جستجو، مانند جمعیتی از جانداران زنده تصور می‌شوند که برای بقا با یکدیگر رقابت می‌کنند. در این نوع الگوریتم‌ها، از افراد "برنده" در نسل فعلی، یک نسل جدید زاده می‌شوند تا پس از گذشت نسل‌های متعدد، جمعیتی از افراد "برتر" بوجود آید. در واقع، امکان حضور هر فرد برای تولید نسل بعد، توسط یک تابع سازگاری[۳] و متناسب با تابع هدف مورد بررسی در مسئله تعیین می‌شود. این الگوبرداری ساده از طبیعت در کنار برخی ملاحظات تئوری، شیوه‌ی کارآمدی برای بدست آوردن جواب بهینه، یا جواب‌های نزدیک به آن در بسیاری از مسائل بهینه‌سازی است. مبنای الگوریتم ژنتیک بر پایه‌ی این اصل استوار است که سازگاری یک جاندار تابعی از اطلاعات رمزشده در ژن‌های آن جاندار است. بر همین اساس، در الگوریتم

[۱] Genetic Algorithm, GA
[۲] Evolutionary Algorithms
[۳] fitness

ژنتیک از سازوکارهای زیستی متنوعی در ارتباط با ژن‌ها یا کروموزم‌ها الگوبرداری می‌شود.

یک الگوریتم ژنتیک به طور ساده، مراحل زیر را شامل می‌شود؛

۱. N جواب شدنی[۴]، یعنی جوابهایی که قیود مسئله را صدق می‌دهند به دلخواه در نظر گرفته می‌شوند. اطلاعات هر یک از این جواب‌ها به شکل یک رشته‌ی بیتی[۵] که از این پس به آن "کروموزم" می‌گوئیم نگهداری می‌شوند.

۲. دسته‌ای از "کروموزم‌ها" که بیشترین سازگاری را در بین دیگر "کروموزم‌ها" دارند انتخاب می‌شوند. تابع سازگاری در اینجا می‌تواند به سادگی، همان تابع هدف باشد.[۶]

۳. گروهی از "کروموزم‌های" انتخاب شده در گام ۲، از نقاط مختلف به طور تصادفی بریده می‌شوند و مجموعه‌ی جدیدی از "کروموزم‌ها" با اتصال قطعات بدست آمده از "کروموزم‌های" مختلف به یکدیگر بدست می‌آیند. این عمل را کراس‌اوِر[۷] می‌گوئیم.

۴. بیت‌هایی از دیگر "کروموزم‌های" انتخاب شده در گام ۲ به تصادف انتخاب می‌شوند و وضعیت آنها از 0به 1یا بالعکس تغییر داده می‌شود. این عمل را جهش[۸] می‌گوئیم.

۵. "کروموزم‌های" بدست آمده در گام‌های ۳ و ۴ نسل بعدی را تشکیل می‌دهند و گام‌های ۲ تا ۵ تا رسیدن به برخی شرایط همگرائی تکرار می‌شوند.

الگوریتم فوق یک الگوی ساده و در عین حال کلی برای بسیاری از انواع الگوریتم‌های ژنتیک است. استقلال و سادگی محاسبات، امکان اجرای موازی‌سازی شده و نیز مکرر الگوریتم را فراهم می‌کند. پارامتر N در این الگوریتم، تعداد "کروموزم‌ها" را نشان می‌دهد و ما آنرا اندازه‌ی جمعیت می‌نامیم. اصولا، این تعداد در تکرارهای الگوریتم ثابت نگهداشته می‌شود. به عبارت دیگر، از هر "کروموزم" یک نماینده به طور میانگین به نسل بعد می‌رود. بزرگ بودن اندازه‌ی جمعیت، از نظر تئوری می‌تواند در رسیدن به جواب بهینه‌ی دقیق موثر

[۴] feasible

[۵] برداری از مؤلفه‌های صفر و یک؛ bit-string.

[۶] منظور مسائل بیشینه‌سازی است. در مسائل کمینه‌سازی، سازگاری عکس تابع هدف است.

[۷] cross-over

[۸] mutation

باشد. با این حال، یک انتخاب مناسب برای اندازه‌ی جمعیت، در عمل، بیشتر به پیچیدگی فضای جستجو و نوع مسئله‌ی بهینه‌سازی وابسته است و در کاربردهای متفاوت ممکن است بین ۵۰ تا چند هزار متغیر باشد.

هرگاه مقدار تابع هدف به ازای برترین "کروموزم"، در چند نسل متوالی تغییر نکند الگوریتم را متوقف می‌کنیم. علاوه بر این شرط، برخی شرایط دیگر نیز به عنوان معیارهای همگرائی و توقف روال تکرار در الگوریتم بکار گرفته می‌شود. به عنوان مثال، در برخی کاربردها به جای بررسی همگرائی در مقدار تابع هدف، شرط قویتر همگرائی "کروموزم‌ها" در جمعیت، برای توقف الگوریتم در نظر گرفته می‌شود. ممکن است شرایط همگرائی هیچگاه محقق نشوند یا همگرائی آنها بیش از زمان برآورد شده برای اجرای برنامه به طول بیانجامد. برای جلوگیری از چنین مواردی، یک کران بالا برای تعداد تکرارهای الگوریتم در نظر گرفته می‌شود.

پارامترهای متعددی بر روند همگرائی الگوریتم ژنتیک اثرگذارند که مهمترین آنها نرخ کراس‌اوِر و جهش هستند. به عنوان مثال، اگر نرخ کراس‌اوِر صفر در نظر گرفته شود الگوریتم ژنتیک به یک الگوریتم جستجوی تصادفی تبدیل می‌شود. از طرفی اگر نرخ جهش را صفر انتخاب کنیم الگوریتم به سرعت به یک جواب بهینه‌ی موضعی همگرا می‌شود. در بسیاری از مسائل، رابطه‌ی معینی بین نرخ کراس‌اوِر و جهش وجود دارد که بر ازای آن الگوریتم همواره به جواب بهینه‌ی سراسری همگرا می‌شود؛ هرچند تعیین این رابطه خود، عملی عمدتا دشواری است.

در برخی پیاده‌سازی‌ها، چند"کروموزم" با بیشترین سازگاری، بدون اعمال هیچ یک از عملگرهای کراس‌اوِر و جهش از نسل قبل به نسل بعد منتقل می‌شوند. این "کروموزم‌ها" را "کروموزم‌های" نخبه[۹] می‌نامیم. تعداد "کروموزم‌های" نخبه از جمله دیگر پارامترهای موثر بر روند رسیدن به جواب در الگوریتم‌های ژنتیک است. از دیگر پارامترهای مهم در پیاده‌سازی یک الگوریتم ژنتیک، نحوه‌ی اندازه‌گیری سازگاری و نیز انتخاب "کروموزم‌های" برتر برای استفاده در تولید نسل بعدی است. به طور منطقی در حل یک مسئله‌ی بهینه‌سازی به وسیله‌ی الگوریتم ژنتیک، سازگاری باید تابعی از تابع هدف مسئله باشد. با این وجود، به ندرت تابع هدف به طور مستقیم، به عنوان تابع سازگاری مورد استفاده قرار می‌گیرد. در بیشتر پیاده‌سازی‌ها، تابع دیگری که تغییراتی متناسب با تغییرات تابع هدف داشته باشد به عنوان تابع سازگاری در نظر گرفته می‌شود. برخی از روابط رایج برای تعریف تابع سازگاری بر این مبنا، عبارتند از:

[۹] elite

- تابع رتبه؛ سازگاری یک "کروموزم" برابر با مکان آن در فهرست مرتب شده‌ی "کروموزم‌ها" برحسب مقدار تابع هدف است.
- تابع بالاترین؛ سازگاری نسبت معینی از "کروموزم‌ها" که کمترین مقدار تابع هدف را در بین دیگر "کروموزم‌ها" دارند برابر یک و مابقی صفر در نظر گرفته می‌شوند.
- نسبت خطی؛ سازگاری با یک تناسب خطی بر حسب تابع هدف تعریف می‌شود به قسمی که مجموع سازگاری‌ها برابر واحد باشد.
- نسبت خطی انتقال یافته؛ مشابه روش قبل، با این تفاوت که پیش از محاسبه‌ی نسبت خطی، یک ثابت معین به مقادیر تابع هدف افزوده می‌شود به قسمی که احتمال حضور بهترین "کروموزم" برای تولید نسل بعد دو برابر میانگین "کروموزوم‌ها" باشد.

روشهای متفاوتی برای انتخاب "کروموزم‌های" والد جهت استفاده در روال‌های کراس‌اوِر و جهش برای تولید نسل بعدی بکار گرفته می‌شوند. برخی از رایج‌ترین این روال‌ها عبارتند از:

- انتخاب یکنواخت؛ "کروموزم‌های" والد با احتمال یکنواخت از میان "کروموزم‌های" برتر انتخاب می‌شوند.
- انتخاب تصادفی یکنواخت؛ "کروموزم‌های" والد به طور مستقل از هم و با احتمالی متناسب با اندازه‌ی سازگاری انتخاب می‌شوند.
- انتخاب به روش چرخ رولت[۱۰]؛ "کروموزم‌های" والد با ترتیب و احتمالی متناسب با اندازه‌ی سازگاری انتخاب می‌شوند. ترتیب انتخاب، بر مبنای موقعیت "کروموزم‌ها" که به ترتیب نزولی بر حسب سازگاری در یک لیست دوری قرار گرفته‌اند تعیین می‌شود.
- انتخاب رقابتی[۱۱]؛ k "کروموزم" به طور کاملا تصادفی انتخاب می‌شوند و "کروموزمی" که بیشترین اندازه‌ی سازگاری را در بین آنها دارد به عنوان والد انتخاب می‌شود. این روال به دفعات تکرار می‌شود

[۱۰] roulette
[۱۱] tournament

تا تعداد مورد نیاز از "کروموزم‌های" والد بدست آیند. k در اینجا، نشان‌دهنده‌ی اندازه‌ی رقابت است.

برای جزئیات بیشتر در مورد روش‌های اندازه‌گیری سازگاری بر حسب تابع هدف و انتخاب "کروموزم‌ها" برای ایجاد نسل بعد به [۱۴۹] نگاه کنید.

مسئله‌ی تعیین فاز ژنوتیپ‌ها بر مبنای مدل بیشترین پارسیمونی

مسئله‌ی بهینه‌سازی مورد علاقه‌ی ما در این بخش، مسئله‌ی تفکیک ژنوتیپ‌ها به هاپلوتیپ‌ها، بر پایه‌ی مدل بیشترین پارسیمونی است. صورت عمومی مسئله‌ی تعیین فاز ژنوتیپ‌ها را به نقل از بخش ۳۰۱ مجددا در اینجا می‌آوریم:

فرض کنید نمونه‌ای مثل $G = \{g_1, \ldots, g_n\}$ شامل n ژنوتیپ بر روی l اسنیپ داده شده است. می‌خواهیم مجموعه‌ای از هاپلوتیپ‌ها مثل H را تعیین کنیم به قسمی که برای هر $g_i \in G$، دو هاپلوتیپ $h_a, h_b \in H$ وجود داشته باشند که $g_i = h_a \oplus h_b$.

در مدل بیشترین پارسیمونی، تابع هدف تعداد هاپلوتیپ‌های متمایز در جواب مسئله‌ی تفکیک ژنوتیپ‌ها است. با توجه به چارچوب کلی الگوریتم ژنتیک که شرح آن پیشتر آمد برای طراحی یک الگوریتم ژنتیک برای حل مسئله‌ی تفکیک ژنوتیپ‌ها با هدف بیشترین پارسیمونی، این عناوین را با جزئیات بیشتری مورد بحث قرار می‌دهیم: تعریف "کروموزم‌ها"؛ یعنی نمایش یک جواب به وسیله‌ی یک رشته‌ی بیتی، نحوه‌ی محاسبه‌ی تابع هدف به ازای هر "کروموزم"، تولید یک مجموعه‌ی اولیه از جواب‌های شدنی، تعریف "کراس‌اوِر"، تعریف "جهش"، انتخاب پارامترهای مناسب برای نرخ کراس‌اوِر، روال‌های مربوط به تعریف سازگاری بر حسب تابع هدف و انتخاب "کروموزم‌های" والد.

یک الگوریتم ژنتیکی ساده برای مسئله‌ی تعیین فاز با بیشترین پارسیمونی

کار را ابتدا با معـرفی یک مدل ساده[۱۲] آغاز می‌کنیم. در ساده‌ترین شکل نمـایش جواب برای مسئله‌ی تفکیـک ژنوتیپ‌ها، به ازای هر بردار $g_i = \langle g_{i1}, \ldots, g_{il} \rangle$ در مجموعه‌ی G، دو هاپلوتیپ h_{2i-1} =

[۱۲] naive

$\langle h_{۲i-۱,۱}, \ldots, h_{۲i-۱,l} \rangle$ و $\langle h_{۲i,۱}, \ldots, h_{۲i,l} \rangle$ در مجموعه‌ی جواب H داریم به قسمی که

$$g_{ij} = h_{۲i-۱,j} + h_{۲i,j}, \qquad \text{for } i = ۱, \ldots, n \text{ and } j = ۱, \ldots, l. \tag{۱۰۲}$$

می‌توان مجموعه‌ی ژنوتیپ‌های داده شده، G را به صورت یک ماتریس $n \times l$ با درآیه‌های 0، 1و 2در نظر گرفت که هر سطر آن نماینده‌ی یک ژنوتیپ است و مجموعه‌ی جواب، H را به صورت یک ماتریس $۲n \times l$ با درآیه‌های 0و 1در نظر گرفت که هر سطر آن نماینده‌ی یک هاپلوتیپ است.[۱۳] با این نمادگذاری، جمع برداری $g_i = h_{۲i-۱} + h_{۲i}$، به سادگی نشاندهنده‌ی تفکیک ژنوتیپ g_i به وسیله‌ی هاپلوتیپ‌های $h_{۲i-۱}$ و $h_{۲i}$ است.

نمایش جواب به وسیله‌ی یک رشته‌ی بیتی

بدیهی است که داشتن یکی از دو هاپلوتیپ تشکیل‌دهنده برای هر ژنوتیپ، برای بازسازی کامل جواب کافی است. به عبارت مشخص، اگر تنها $h_{۲i}$ یا $h_{۲i-۱}$ را داشته باشیم هاپلوتیپ دیگر با حل معادله‌ی (۱۰۲) برحسب هاپلوتیپ معلوم و ژنوتیپ مورد مطالعه بدست می‌آید. علاوه بر آن، مقدار مؤلفه‌های $h_{۲i-۱,j}$ و $h_{۲i,j}$ وقتی $g_{ij} = \mathbf{0}$ یا $g_{ij} = \mathbf{2}$ است بنابر تعریف، از پیش معین است. به همین خاطر تنها اطلاعات نابدیهی در جواب، مربوط به مؤلفه‌هایی است که در آنها $g_{ij} = \mathbf{1}$ است. به چنین موقعیت‌هایی، یعنی جایگاه‌های هتروزیگوت در یک ژنوتیپ، موقعیت‌های مبهم[۱۴] برای مسئله‌ی تعیین فاز می‌گوئیم. در واقع، تعیین فاز در موقعیت‌های مبهم معادل بدست آوردن یک جواب برای مسئله به طور کامل است. بر این اساس می‌توانیم جواب را به صورت فشرده در یک رشته‌ی بیتی نمایش دهیم.

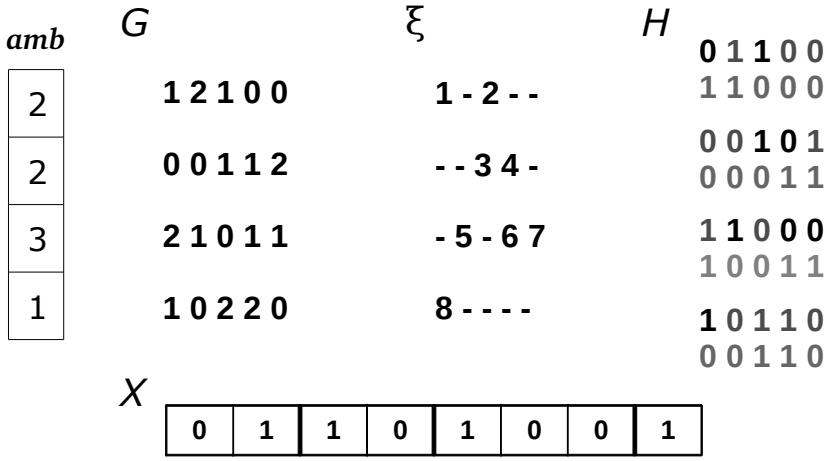
برای این کار یک رشته‌ی بیتی مثل X به طول M در نظر می‌گیریم؛

$$M = \sum_{i=۱}^{n} amb_i$$

که در آن amb_i تعداد موقعیت‌های مبهم در ژنوتیپ g_i است. در واقع این رشته حاصل اتصال n رشته‌ی

[۱۳] برای اطلاع از معنای اعداد قرارداد شده برای ژنوتیپ‌های اسنیپ بخش ۳۰۱ را ببینید.

[۱۴] ambiguous



شکل ۱۰۲: نمایش جواب مسئله‌ی تعیین فاز توسط یک رشته‌ی بیتی
ماتریس ژنوتیپ‌ها، G داده شده است. آرایه‌ی amb تعداد جایگاه‌های مبهم در هر ژنوتیپ را نشان می‌دهد. فاز جایگاه‌های مبهم برای هاپلوتیپ "اول" هر ژنوتیپ در رشته‌ی بیتی X نگهداری می‌شود. ماتریس ξ اندیس متناظر با هر موقعیت مبهم در رشته‌ی بیتی X را نشان می‌دهد.

کوچکتر، هر یک متناظر با مؤلفه‌های مبهم یک ژنوتیپ است. بر پایه‌ی این نحوه‌ی نمایش جواب، برای بدست آوردن هاپلوتیپ‌های تشکیل‌دهنده‌ی یک ژنوتیپ داریم:

$$h_{۲i-\delta,j} = \begin{cases} g_{ij}/۲ & \text{if} \quad g_{i,j} = \mathbf{0} \ \text{ or } g_{ij} = \mathbf{2}, \\ X_{\xi(i,j)} & \text{if} \quad g_{ij} = \mathbf{1} \text{ and } \delta = ۱, \\ ۱ - X_{\xi(i,j)} & \text{if} \quad g_{ij} = \mathbf{1} \text{ and } \delta = ۰. \end{cases}$$

که در آن $\xi(i,j)$ مکان بیت متناظر با موقعیت مبهم g_{ij} در رشته‌ی بیتی X است(شکل ۱۰۲).

تولید تصادفی یک مجموعه‌ی اولیه از جواب‌های شدنی

تولید تصادفی جواب‌های شدنی بر مبنای این نحوه‌ی نمایش جواب در رشته‌های بیتی، عملی بسیار ساده است. به عبارت مشخص، جمعیت اولیه‌ی "کروموزوم‌ها" در این الگوریتم ژنتیک، شامل N رشته‌ی بیتی، هر یک به طول M است که مقدار هر بیت در آن به احتمال یکسان 0یا 1است. می‌توان نشان داد تعداد جواب‌های شدنی متمایز برای یک مسئله‌ی تعیین فاز برابر است با ۲^{M-n_e} که در آن n_e نشاندهنده‌ی تعداد ژنوتیپ‌هایی

با حداقل یک جایگاه هتروزیگوت در مجموعه ژنوتیپ‌های داده شده است. بر این اساس، تعداد جواب‌های شدنی متمایز حتی در ساده‌ترین نمونه‌های عملی مسئله، عددی نجومی است در حالی که الگوریتم ژنتیک در عمل، تنها می‌تواند نقاط بسیار پراکنده‌ای از فضای جستجوی مسئله را ارزیابی کند. کارآیی الگوریتم پیشنهادی به ازای دو اندازه‌ی جمعیت مختلف، $N = ۵۰$ و $N = ۱۰۰$ مورد بررسی قرار خواهد گرفت.

محاسبه‌ی تابع هدف

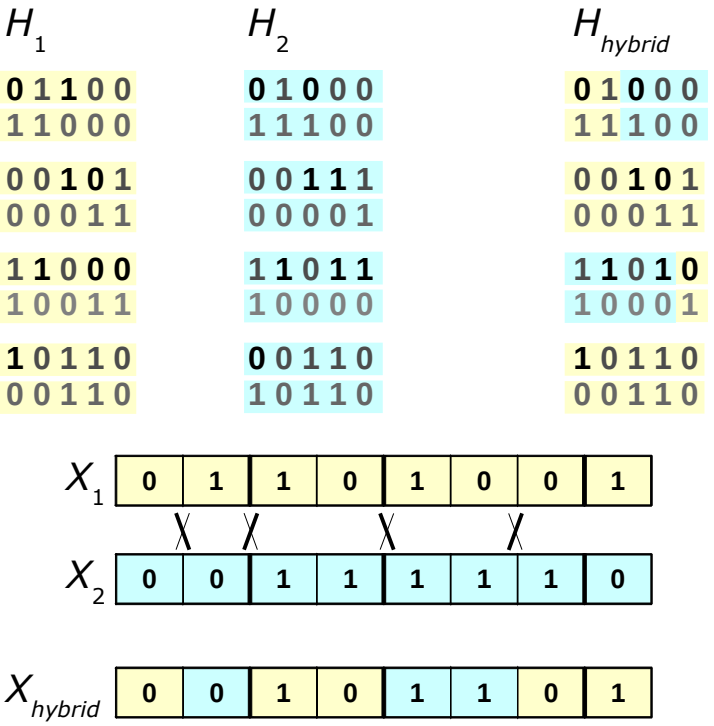
تابع هدف در مسئله‌ی بیشترین پارسیمونی، تعداد هاپلوتیپ‌های متمایز در مجموعه‌ی جواب، یعنی مجموعه‌ی هاپلوتیپ‌های تفکیک‌کننده‌ی ژنوتیپ‌ها است. به طور بدیهی، برای تعیین هاپلوتیپ‌های متمایز می‌توان هر هاپلوتیپ را با دیگر هاپلوتیپ‌های مجموعه‌ی جواب مقایسه کرد و از این طریق تعداد هاپلوتیپ‌های متمایز را بدست آورد. بدین ترتیب زمان محاسبه‌ی تابع هدف به ازای هر جواب شدنی از مرتبه‌ی $O(n^۲l)$ خواهد بود. با این حال، روش ساده و در عین حال کارآمدتری نیز برای محاسبه‌ی تعداد هاپلوتیپ‌های متمایز وجود دارد. در این روش ابتدا سطرهای ماتریس جواب با استفاده از الگوریتم مرتب‌سازی مبنائی[۱۵] مرتب می‌شوند. اکنون، با یکبار مقایسه‌ی سطرهای متوالی در ماتریس مرتب‌شده می‌توان تعداد هاپلوتیپ‌های متمایز را بدست آورد. پیچیدگی محاسباتی این روش $O(nl)$ است.

تعریف "کراس‌اوِر"

فرض کنید دو ماتریس جواب، $H_۱$ و $H_۲$ داده شده است. می‌خواهیم با ترکیب اطلاعات این دو جواب یک جواب جدید برای مسئله‌ی تفکیک ژنوتیپ‌ها بسازیم. ساده‌ترین رویکرد، تعریف ماتریس جواب جدیدی است که در آن، هر جفت از سطرهای متوالی، یعنی هر جفت از هاپلوتیپ‌های تفکیک‌کننده‌ی یک ژنوتیپ، به طور تصادفی از یکی از دو ماتریس $H_۱$ یا $H_۲$ انتخاب می‌شود. این شیوه‌ی کراس‌اوِر تضمین می‌کند که "کروموزم" بدست آمده هنوز می‌تواند ژنوتیپ‌های داده شده را تفکیک کند. شیوه‌ای که ما برای کراس‌اوِر جواب‌ها به کار می‌بریم تعمیمی از این رویکرد ساده است. در این روش، هاپلوتیپ‌های هر جفت متناظر در دو ماتریس از یک نقطه‌ی تصادفی به طور مشترک شکسته می‌شوند و با جابجائی قطعات بدست آمده بین آنها

[۱۵] radix sort

یک جواب جدید بدست می‌آید (شکل ۲۰۲). این روش نیز ویژگی جواب برای تفکیک ژنوتیپ‌های داده شده را حفظ می‌کند.



شکل ۲۰۲: "کراس‌اوِر" جواب‌ها در الگوریتم ژنتیک ساده برای حل مسئله‌ی تفکیک ژنوتیپ‌ها

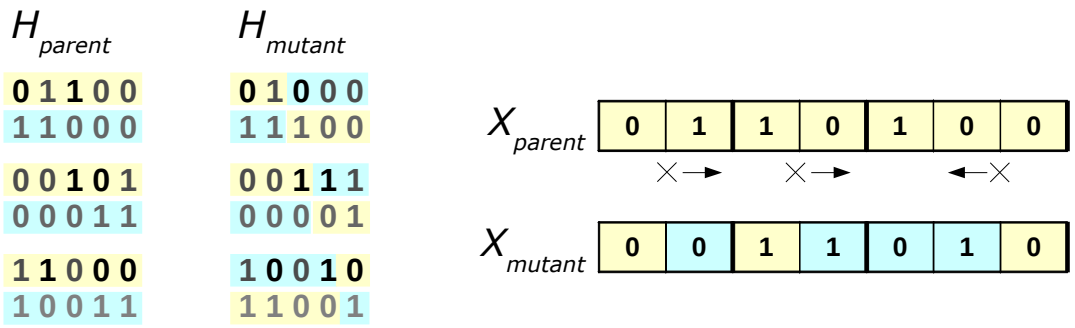
در این شیوه، از یک پارامتر که آنرا با cr_{int} نشان می‌دهیم برای کنترل میزان دورگه بودن[۱۶] "کروموزم‌های" بدست آمده استفاده می‌کنیم. این پارامتر احتمال کراس‌اوِر بین جفت هاپلوتیپ‌های متناظر در دو ماتریس جواب را نشان می‌دهد. به عبارت دیگر، احتمال اینکه یک جفت از هاپلوتیپ‌های تفکیک‌کننده‌ی یک ژنوتیپ در جواب نوترکیب تنها از یکی از والدها به ارث رسیده باشد برابر $1 - cr_{int}$ است. این الگوریتم برای اجرا بر مبنای استفاده از نمایش رشته‌ی بیتی جواب‌ها پیاده‌سازی می‌شود.

تعریف "جهش"

برای تولید یک "کروموزوم" جهش‌یافته از یک "کروموزوم" والد کافی است وضعیت برخی درآیه‌های متناظر با جایگاه‌های مبهم در ماتریس جواب را معکوس کنیم. برای این کار، یک جفت از هاپلوتیپ‌های ماتریس جواب

[۱۶] hybridity

به طور تصادفی و با احتمال mr_{int} انتخاب می‌شود سپس مؤلفه‌های این دو سطر از یک نقطه‌ی تصادفی به بعد با یکدیگر جابجا می‌شوند. این عمل معادل معکوس کردن وضعیت بیت‌های یک زیررشته‌ی پیشوندی یا پسوندی در زیررشته‌ی متناظر با یک ژنوتیپ در نمایش رشته‌ی بیتی جواب است (شکل ۳۰۲).



شکل ۳۰۲: "جهش" جواب‌ها در الگوریتم ژنتیک ساده برای حل مسئله‌ی تفکیک ژنوتیپ‌ها

یک الگوریتم ژنتیکی برای جستجوی بهترین الگوریتم سودجویانه برای حل مسئله‌ی تعیین فاز با بیشترین پارسیمونی

مدل ساده‌ای که در بالا معرفی گردید در عمل توانائی حل مسئله‌ی تعیین فاز ژنوتیپ‌ها با بیشترین پارسیمونی را جز برای نمونه‌های کوچک ندارد. در واقع در این مدل ساده، جستجوی جواب بهینه تنها از طریق جستجو در بین نمونه‌های متعددی از جواب‌های شدنی که از طریق روال‌های "کراس‌اوِر" و "جهش" الگوریتم ژنتیک تولید شده‌اند انجام می‌گیرد. این رویکرد هر چند به عنوان یک الگوریتم ژنتیکی سر راست و طبیعی به نظر می‌رسد اما همانطور که اشاره شد تکیه بر جستجوی فضای جواب با نمونه‌گیری از جواب‌های شدنی مسئله وقتی بعد فضای جواب بالا باشد نمی‌تواند به طور مؤثر به یافتن جواب بهینه کمکی کند. نتایجی که در بخش ۱۰۳ موضوع بحث قرار می‌گیرند مؤید ضعف این رویکرد است.

با استفاده از ایده‌های مطرح شده در الگوریتم ژنتیک ساده‌ای که در قسمت قبل معرفی گردید و برای بهبود آن، یک الگوریتم سودجویانه برای حل مسئله‌ی تعیین فاز با هدف بیشترین پارسیمونی را در قالب یک الگوریتم ژنتیک مورد بررسی قرار می‌دهیم. ابتدا یک صورت پارامتری برای این الگوریتم سودجویانه معرفی

می‌کنیم. سپس تعابیر[۱۷] مختلف این الگوریتم را به ازای پارامترهای مختلفی که توسط یک الگوریتم ژنتیک تعیین می‌شوند اجرا می‌کنیم.

الگوریتم *GreedyPhasing*

ورودی). مجموعه‌ی $G = \{g_1, \ldots, g_n\}$ شامل n ژنوتیپ بر روی l اسنیپ داده شده است. به علاوه، یک جایگشت از اعداد ۱ تا n، مثل $\sigma = \langle \sigma_1, \ldots, \sigma_n \rangle$ و مجموعه‌ای از n هاپلوتیپ، مثل $H_g = \{\hbar_1, \ldots, \hbar_n\}$ که از این پس آنها را هاپلوتیپ‌های راهنما می‌نامیم داده شده‌اند به قسمی که برای $i = 1, \ldots, n$ داریم، $\hbar_i \sim g_i$. مجموعه‌ی G را ورودی مسئله و جایگشت σ و مجموعه‌ی هاپلوتیپ‌های راهنما، H_g را پارامترهای الگوریتم می‌نامیم.

گام ۱). قرار بده $H = \varnothing$ و برای $i = 1, \ldots, n$ به ترتیب، گام‌های زیر را انجام بده.

گام ۱-۱). اولین هاپلوتیپ سازگار با g_{σ_i} را در لیست H جستجو کن و آنرا با h_a نشان بده. اگر چنین هاپلوتیپی وجود نداشت قرار بده $h_a = \hbar_{\sigma_i}$.

گام ۱-۲). هاپلوتیپ‌های h_a و $h_b = g_{\sigma_i} - h_a$ را به انتهای لیست H اضافه کن.

گام ۲). ماتریس جواب، با اعمال جایگشت معکوس، σ^{-1} بر روی لیست H بدست می‌آید.

به ازای پارامترهای داده‌شده، این الگوریتم یک الگوریتم قطعی[۱۸] برای بدست آوردن یک جواب شدنی برای مسئله‌ی تفکیک ژنوتیپ‌ها است. با این حال، مزیت منطق سودجویانه‌ی این الگوریتم در آن است که جوابی برای مسئله‌ی تفکیک ژنوتیپ‌ها بدست می‌دهد که تا حدودی به جواب بهینه‌ی بیشترین پارسیمونی نزدیک است. هر نمونه از این الگوریتم به ازای جایگشت داده‌شده‌ی σ و مجموعه‌ی داده‌شده‌ی H_g از هاپلوتیپ‌های راهنما را با $\mathit{GreedyPhasing}(\sigma, H_g)$ نشان می‌دهیم.

اطلاعات مجموعه‌ی H_g را می‌توان همانند روشی که در بالا برای نگهداری اطلاعات ماتریس جواب در رشته‌ی بیتی شرح دادیم در یک "کروموزم" نگهداری کرد(شکل ۱۰۲). با پشت سر هم قرار دادن نمایش

[۱۷] instances
[۱۸] deterministic

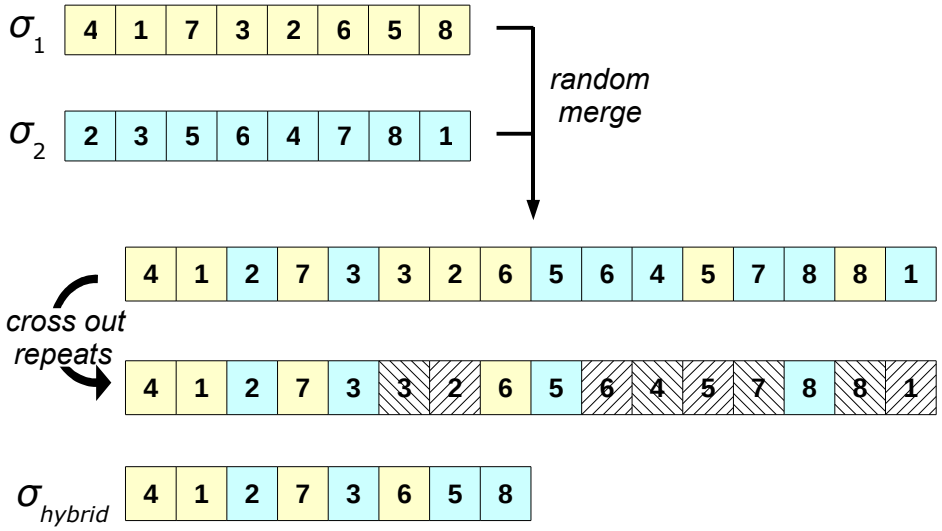
دودوئی اعداد $\langle \sigma_1, \sigma_2, \ldots, \sigma_n \rangle$، یک رشته‌ی بیتی به طول $n.d$ خواهیم داشت که از آن برای نمایش جایگشت σ در یک "کروموزم" استفاده می‌کنیم. در اینجا، $d = \lceil \log_2 n \rceil$ حداقل تعداد بیت‌های لازم برای نمایش بزرگترین عدد جایگشت است. اکنون می‌توان، هر نمونه از الگوریتم $\mathit{GreedyPhasing}(\sigma, H_g)$ را با کنار هم قرار دادن نمایش بیتی σ و H_g در یک "کروموزم" واحد نمایش داد. بدین ترتیب هر "کروموزم" از دو بخش تشکیل خواهد شد؛ یکی، بخش مربوط به نمایش جایگشت به طول $n.d$ بیت و دیگری بخش مربوط به نمایش هاپلوتیپ‌های راهنما به طول M بیت.

محاسبه‌ی تابع هدف، یعنی تعداد هاپلوتیپ‌های متمایز مجموعه‌ی جواب، پس از اجرای $\mathit{GreedyPhasing}(\sigma, H_g)$ و بدست آوردن ماتریس جواب، با روشی همانند روش شرح داده شده در الگوریتم ابتدائی انجام می‌گیرد. برای تولید جمعیت اولیه‌ی "کروموزم‌ها" همانند روش قبل می‌توان بیت‌های مربوط به اطلاعات هاپلوتیپ‌های راهنما را به طور تصادفی با مقادیر 0و 1پر کرد. اما برای بیت‌های بخش مربوط به جایگشت‌ها، ابتدا یک جایگشت تصادفی از اعداد ۱ تا n بدست می‌آید سپس نمایش دودوئی آن، در بخش مربوط قرار داده می‌شود.

"کراس‌اوِر" در الگوریتم ژنتیک بهبودیافته

در اینجا، "کراس‌اوِر" به نوعی، تلفیق دو نمونه از الگوریتم *GreedyPhasing* برای بدست آوردن نمونه‌ای دیگر از همین الگوریتم است. این کار باید بر مبنای ترکیب اطلاعات جایگشت‌ها و هاپلوتیپ‌های راهنما در دو الگوریتم صورت گیرد. برای این منظور، دو نوع "کراس‌اوِر" را در نظر می‌گیریم. در نوع اول، اطلاعات مربوط به هاپلوتیپ‌های راهنما، دقیقا همانند روش شرح داده شده در الگوریتم ابتدائی با یکدیگر ترکیب می‌شوند(شکل ۲۰۲). در این نوع "کراس‌اوِر"، اطلاعات مربوط به جایگشت، مستقیما از یکی از والدین، بدون تغییر به زاد جدید منتقل می‌شود. برای تلفیق جایگشت‌ها و در "کراس‌اوِر" نوع دوم، عناصر دنباله‌ی $\langle \sigma^1_1, \sigma^1_2, \ldots, \sigma^1_n \rangle$ با حفظ ترتیب و به طور تصادفی در بین عناصر دنباله‌ی $\langle \sigma^2_1, \sigma^2_2, \ldots, \sigma^2_n \rangle$ درج می‌شوند. بدیهی است که در دنباله‌ی بدست آمده، هر یک از اعداد ۱ تا n دقیقا دو بار ظاهر می‌شوند. در اینجا، دومین تکرار هر عدد را حذف می‌کنیم و دنباله‌ی باقیمانده را به عنوان جایگشت نوترکیب در نظر

می‌گیریم. در این نوع "کراس‌اوِر"، والد عنصر منتقل شده به جایگشت نوترکیب، منشاء هاپلوتیپ راهنما را تعیین می‌کند(شکل ۴۰۲). به تناوب یکی از این دو نوع "کراس‌اوِر" در روال "کراس‌اوِر" الگوریتم ژنتیک به کار گرفته می‌شوند.



شکل ۴۰۲: "کراس‌اوِر" جایگشت‌ها در الگوریتم ژنتیک بهبودیافته

دو جایگشت σ_1 و σ_2 با حفظ ترتیب و به طور تصادفی با یکدیگر ادغام می‌شوند. تکرار دوم از هر عدد در دنباله‌ی بدست آمده حذف می‌شود و حاصل به عنوان جایگشت نوترکیب در نظر گرفته می‌شود. منشاء هاپلوتیپ‌های راهنمایی که می‌بایست از یکی از والدین به "کروموزم" نوترکیب منتقل شود بر اساس والد مؤلفه‌ی منتقل شده به جایگشت نوترکیب تعیین می‌شود.

"جهش" در الگوریتم ژنتیک بهبودیافته

در الگوریتم ژنتیک بهبودیافته، "جهش" نیز همانند "کراس‌اوِر" به دو شکل انجام می‌شود؛ "جهش" بر روی هاپلوتیپ‌های راهنما و "جهش" بر روی جایگشت‌ها. "جهش" بر روی هاپلوتیپ‌های راهنما دقیقا به همان شیوه‌ای که در الگوریتم ژنتیک ساده شرح داده شد انجام می‌شود(شکل ۳۰۲). برای اعمال "جهش" بر روی جایگشت‌ها، ما شیوه‌ی ساده‌ای را که در آن دو مؤلفه‌ی جایگشت به تصادف با یکدیگر جابجا می‌شوند بکار می‌بریم. در اینجا نیز این دو نوع "جهش" به تناوب در روال تکرار الگوریتم ژنتیک به کار گرفته می‌شوند.

تا اینجا دو الگوریتم ژنتیک متفاوت برای حل مسئله‌ی تفکیک ژنوتیپ‌ها با هدف بیشترین پارسیمونی معرفی شدند؛ الگوریتم اول که در آن ایده‌ای ساده و ابتدائی برای جستجوی جواب بهینه بکار گرفته می‌شد و

آنرا naive-GAhap می‌نامیم و دیگری که در آن، جواب بهینه از طریق ارزیابی نمونه‌های متعددی از یک الگوریتم سودجویانه جستجو می‌گردید و از این پس آنرا GAhap خواهیم خواند. هر دو الگوریتم را تحت زبان برنامه‌نویسی MATLAB پیاده‌سازی کرده‌ایم. بسته‌ی نرم‌افزاری MATLAB علاوه بر معرفی یک زبان برنامه‌نویسی قدرتمند برای پیاده‌سازی انواع الگوریتم‌های رایج در محاسبات علمی، ابزار ویژه‌ای نیز برای پیاده‌سازی الگوریتم‌های ژنتیکی دارد. علاوه بر جزئیات شرح داده شده در مورد نحوه‌ی نمایش اطلاعات "کروموزم‌ها" به صورت رشته‌های بیتی، محاسبه‌ی تابع هدف، تولید جمعیت اولیه و روال‌های "کراس‌اوِر" و "جهش" در هر دو الگوریتم، هنوز لازم است سایر پارامترهای مؤثر در الگوریتم‌های ژنتیک، یعنی نرخ نوترکیبی، روال انتخاب "کروموزم‌های" والد و اندازه‌گیری سازگاری بر حسب تابع هدف نیز تعیین گردند تا امکان استفاده‌ی کاربردی این الگوریتم‌ها محقق گردد.

برای تعیین مقادیر مناسب برای پارامترهای مذکور و نیز ارزیابی کارایی الگوریتم‌های ژنتیکی پیشنهاد شده، نتایج حاصل از اجراهای مکرر این الگوریتم‌ها به ازای مقادیر مختلف پارامتر بر روی داده‌های ژنوتیپی شبیه‌سازی شده مورد بررسی قرار می‌گیرند. مزیت استفاده از داده‌های شبیه‌سازی شده در این است که یک جواب شدنی برای مسئله‌ی تعیین فاز در اختیار است. در واقع، بنابر روالی که برای تولید تصادفی ژنوتیپ‌ها بکار گرفته می‌شود جفت هاپلوتیپ‌های تشکیل‌دهنده‌ی هر یک از ژنوتیپ‌ها از پیش شناخته شده‌اند. از این رو یک کران بالا برای جواب مسئله‌ی بیشترین پارسیمونی در دست است.

شبیه‌سازی نمونه‌های ژنوتیپ برای استفاده در الگوریتم‌های ژنتیکی و بررسی کارایی این الگوریتم‌ها، به ترتیبی که در ادامه می‌آید انجام می‌گیرد. ابتدا نمونه‌ی به قدر کافی بزرگی از هاپلوتیپ‌ها بر روی l اسنیپ با قید $MAF > ۰/۰۵$ به طور تصادفی تولید می‌شوند. سپس با اضافه کردن نمونه‌های مشابه به طور متناسب فراوانی هر هاپلوتیپ موجود در این نمونه افزایش می‌یابد تا جائیکه "هاپلوتیپ‌های رایج" ٪۸۰ هاپلوتیپ‌های نمونه را در برگیرند؛ شیوه‌ی نسبتا پیچیده‌ی دیگری برای تولید نمونه‌های هاپلوتیپ در بخش ۲۰۲ مورد بحث قرار می‌گیرد. پس از این مرحله، به طور تصادفی و کاملا یکنواخت n جفت هاپلوتیپ از میان هاپلوتیپ‌های این نمونه‌ی بدست آمده انتخاب می‌شوند و نمونه ژنوتیپ‌های مورد نیاز با ترکیب هاپلوتیپ‌های هر یک از این جفت‌ها بدست می‌آید. تعداد هاپلوتیپ‌های متمایز در مجموعه‌ی اخیر محاسبه می‌شود و به عنوان یک کران

بالا برای جواب بهینه‌ی تفکیک ژنوتیپ‌ها با بیشترین پارسیمونی نگهداری می‌شود. برای هر انتخاب از مقادیر برای پارامترهای مورد مطالعه در الگوریتم ژنتیک، الگوریتم‌های پیشنهادی بر روی ۲۰ نمونه‌ی شبیه‌سازی شده‌ی مستقل، هر یک شامل ۴۰ ژنوتیپ و ۱۲ اسنیپ اجرا می‌شوند.

برای بررسی کارایی الگوریتم‌های ژنتیک پیشنهادی به ازای هر یک از مقادیر مورد انتخاب برای پارامترهای الگوریتم ژنتیک، مقدار یکی از پارامترها، مثلا نرخ جهش در هاپلوتیپ‌ها، ثابت نگه داشته می‌شود و الگوریتم با ازای تمام ترکیب‌های ممکن از موارد انتخاب دیگر پارامترها، طبق روشی که در بالا شرح داده شد اجرا می‌شود. این فرایند برای هر پارامتر و هر مقدار مورد مطالعه برای آن تکرار می‌شود. به طور خلاصه، مقادیر و گزینه‌های مورد مطالعه برای هر یک از پارامترها به این شرح می‌باشند:

- ۰/۲، ۰/۸ و ۰/۹ برای نرخ نوترکیبی در الگوریتم ژنتیک، cr.
- ۰/۱، ۰/۵ و ۰/۹ برای نرخ نوترکیبی بین هاپلوتیپ‌های یک ژنوتیپ، cr_{int}.
- ۰/۱، ۰/۵ و ۰/۹ برای نرخ جهش در هاپلوتیپ‌ها، mr_{int}.
- انتخاب یکنواخت، انتخاب تصادفی یکنواخت، انتخاب به روش چرخ رولت و انتخاب رقابتی به عنوان شیوه‌های انتخاب "کروموزم‌های" والد هر نسل.
- تابع رتبه، تابع بالاترین، نسبت خطی و نسبت خطی انتقال‌یافته به عنوان روش‌های تبدیل تابع هدف به تابع سازگاری.

خلاصه‌ی نتایج بدست آمده و بحث درباره‌ی آنها را در بخش ۱۰۳ پی گیرید.

پس از تعیین کارآمدترین انتخاب برای پارامترها، الگوریتم GAhap را بر روی نمونه‌ای واقعی از ژنوتیپ‌ها اجرا خواهیم کرد. این نمونه‌ی واقعی، ژنوتیپ‌های مربوط به ژن هورمون رشد، GH1 در منطقه‌ی کروموزمی 17q23 است که توسط هورَن و همکارانش از بین ۱۵۴ فرد نمونه در انگلستان بدست آمدند[۱۵۰]. آنها توانستند فراوانی ۳۶ هاپلوتیپ متمایز موجود در این مجموعه را به کمک روش‌های آزمایشگاهی تعیین کنند. در این ناحیه‌ی ژنی ۱۶ اسنیپ شناخته شده وجود دارد. در مطالعه‌ی هورَن و همکارانش، داده‌های مربوط به ۱۵ اسنیپ از کیفیت لازم برای مطالعه برخوردار بودند. ما توانستیم در مجموع، اطلاعات مربوط به ۱۵۰

ژنوتیپ را بدون داشتن داده‌ی مفقود از بین داده‌های در دسترس قرار داده شده‌ی این مجموعه هاپلوتیپ‌ها بدست آوریم.

به طور متعارف، برای برآورد دقت یک الگوریتم در حل مسئله‌ی تعیین فاز ژنوتیپ‌ها، تعداد موارد اختلاف بین زوج هاپلوتیپ استنباط شده و زوج هاپلوتیپ واقعی به ازای هر یک از ژنوتیپ‌های داده شده محاسبه می‌شود و میانگین این تعداد به عنوان مقدار خطای استنباط در نظر گرفته می‌شود. در مجموعه‌ی هورَن، هرچند فراوانی و نوع هاپلوتیپ‌های واقعی در دسترس‌اند اما فاز مرتبط با ژنوتیپ‌های این مجموعه معین نیست. از این رو نمی‌توان از روابط متعارف برای محاسبه‌ی خطا استفاده کرد. به همین جهت ما روابط دیگری تنها مبتنی بر فراوانی هاپلوتیپ‌های شناخته شده در جمعیت بدست می‌آوریم که از طریق آن می‌توان خطای استنباط را به طور میانگین برآورد کرد. فرض کنید مجموعه‌ی هاپلوتیپ‌های واقعی $H_o = \{\hbar_1, \ldots, \hbar_k\}$ و فراوانی هر یک از این هاپلوتیپ‌ها در جمعیت، یعنی $f(\hbar_i)$ برای $i = 1, \ldots, k$ داده شده است. بر پایه‌ی این مفروضات می‌توان، میانگین خطای استنباط هاپلوتیپ‌ها را نسبت به تمام تعیین فازهایی که از توزیع f پیروی می‌کنند بدست آورد. در این ارتباط، معمولا دو نوع خطا در استنباط هاپلوتیپ‌ها مورد توجه قرار می‌گیرند که با توجه به مفروضات فوق به صورت زیر محاسبه می‌شوند:

$$e_{haplotype} = \frac{1}{n} \sum_{i=1}^{n} HE(g_i)/P(g_i) \ ,$$

$$e_{switch} = \frac{1}{M - n_e} \sum_{i=1}^{n} SE(g_i)/P(g_i) \ ,$$

که در آن

$$HE(g_i) = \sum_{\hbar_a \oplus \hbar_b = g_i} f(\hbar_a) f(\hbar_b) \ HE_{\langle \hbar_a, \hbar_b \rangle, \langle h_{2i-1}, h_{2i} \rangle} \ ,$$

$$SE(g_i) = \sum_{\hbar_a \oplus \hbar_b = g_i} f(\hbar_a) f(\hbar_b) \ SE_{\langle \hbar_a, \hbar_b \rangle, \langle h_{2i-1}, h_{2i} \rangle} \ ,$$

$$P(g_i) = \sum_{\hbar_a \oplus \hbar_b = g_i} f(\hbar_a) f(\hbar_b) \ .$$

در اینجا $HE_{\langle \hbar_a,\hbar_b\rangle,\langle h_{۲i-۱},h_{۲i}\rangle}$ تابع مشخصه‌ی برابر بودن هاپلوتیپ‌های زوج بدون ترتیب $\langle \hbar_a, \hbar_b\rangle$ و هاپلوتیپ‌های زوج بدون ترتیب $\langle h_{۲i-۱}, h_{۲i}\rangle$ در تمام اسنیپ‌ها است و $SE_{\langle \hbar_a,\hbar_b\rangle,\langle h_{۲i-۱},h_{۲i}\rangle}$ نشان‌دهنده‌ی کمترین تعداد جابجائی‌های لازم بین اسنیپ‌های مبهم در جفت هاپلوتیپ‌های $\langle h_{۲i-۱}, h_{۲i}\rangle$ برای رسیدن به $\langle \hbar_a, \hbar_b\rangle$ است. به بیان ساده، $e_{haplotype}$ نشان‌دهنده‌ی متوسط نرخ خطا در شناسائی یک هاپلوتیپ و e_{switch} نشان‌دهنده‌ی احتمال به اشتباه جابجا شدن فاز در امتداد کروموزم توسط الگوریتم است.

نتایج بدست آمده از تفکیک ژنوتیپ‌های مجموعه‌ی هورَن توسط روش GAhap و دیگر الگوریتم‌های رایج در تفکیک ژنوتیپ‌ها و خطای استنباط در آنها در بخش ۱۰۳ با یکدیگر مقایسه می‌شوند.

۲۰۲ تولید نمونه‌های تصادفی هاپلوتیپ تحت مدل فیلوژنی کامل

ارزیابی کارایی و دقت روش‌هایی که در حل مسائل محاسباتی مرتبط با هاپلوتیپ‌ها بکار می‌روند از جمله، مسئله‌ی تفکیک ژنوتیپ‌ها و استنباط هاپلوتیپ‌ها و مسئله‌ی افراز بلوکی هاپلوتیپ‌ها نیازمند بررسی نتایج بدست آمده از اجرای مکرر این روش‌ها بر روی نمونه‌های متعدد است. تولید نمونه‌های تصادفی از طریق روش‌های محاسباتی در کامپیوتر را اصطلاحا شبیه‌سازی نمونه می‌گویند. مزیت کلی استفاده از روش‌های محاسباتی برای تولید نمونه، انعطاف‌پذیری آنها در انتخاب ساختار آماری نمونه و حجم آن و نیز تکرارپذیری نمونه‌های تولید شده توسط آنها است. ما برای ارزیابی الگوریتم‌های مورد بحث در این رساله به دفعات ابزارهای مختلف شبیه‌سازی را برای تولید نمونه‌هایی از ژنوتیپ‌ها و هاپلوتیپ‌ها به کار می‌گیریم.

در بسیاری از مطالعات کاربردی، دسته‌ای از افراد غیرخویشاوند جمعیت به تصادف انتخاب می‌شوند و ژنوتیپ‌ها یا هاپلوتیپ‌های این افراد به عنوان نمونه‌ی مورد مطالعه بکار گرفته می‌شوند. ممکن است در ابتدا این طور به نظر آید که ترکیب آلل‌ها در اسنیپ‌های مختلف چنین نمونه‌هایی تابع هیچ رابطه یا قید به خصوصی نیستند اما در واقع، همانطور که در بخش ۴۰۱ اشاره گردید حتی در هاپلوتیپ‌های افراد غیرخویشاوند جمعیت نیز ساختار معناداری از نظر توزیع آماری هاپلوتیپ‌های غیریکسان مشاهده می‌شود که نشان‌دهنده‌ی تنوع اندک هاپلوتیپ‌های متمایز درون جمعیت در مقایسه با آنچه از نظر تئوری می‌توان تصور کرد است. عامل دیگری که باعث می‌شود نتوان به سادگی نمونه هاپلوتیپ‌های غیرخویشاوند را کاملا تصادفی فرض کرد نابرابر

بودن فراوانی آلل‌های مختلف یک اسنیپ در جمعیت و تفاوت آن در بین اسنیپ‌ها است که معمولا نتیجه‌ی سازوکارهای انتخاب طبیعی، رانش ژنی[۱۹] و نوترکیبی است. علاوه بر عوامل فوق، فرایند پیدایش هاپلوتیپ‌ها در افراد یک گونه خود باعث به وجود آمدن انواع کاملا خاصی از هاپلوتیپ‌ها می‌شود که لزوماً شباهتی با توالی‌های تصادفی از اعداد 0و 1ندارند. در این بخش، روش ساده‌ای را برای تولید نمونه‌های تصادفی از هاپلوتیپ‌ها ارائه می‌کنیم و در آن برخی از این ویژگی‌ها را در قالب قیود حاکم بر ساختار هاپلوتیپ‌های شبیه‌سازی شده، مورد توجه قرار می‌دهیم.

رایج‌ترین نظریه درباره‌ی پیدایش هاپلوتیپ‌ها، در بین دانشمندان حال حاضر علوم زیستی، نظریه‌ی "نیای مشترک" یا همانطور که در بخش ۳۰۱ به آن اشاره شد مدل "فیلوژنی کامل" است. این مدل برپایه‌ی دو اصل اولیه بنا می‌شود؛ اول اینکه، منشاء تمام هاپلوتیپ‌های موجود در جمعیت، یک هاپلوتیپ واحد در ابتدای حیات است که هاپلوتیپ‌های زمان حال نتیجه‌ی روی داد جهش‌های متعدد بر روی این هاپلوتیپ اولیه هستند و دوم اینکه، ژنوم در مقایسه با نرخ جهش‌های پایدار، طویل‌تر از آن است که برای یک نوکلئوتید، احتمال رویداد جهش بیش از یکبار وجود داشته باشد. این مدل ساده‌سازی شده‌ای است که در آن، از نقش رویداد نوترکیبی در پدید آوردن هاپلوتیپ‌های جدید صرف نظر شده است و از این رو، با عنوان "مدل نیای مشترک بدون نوترکیبی"[۲۰] شناخته می‌شود. در واقع، ساختار هاپلوتیپ‌های طبیعی در مناطقی از ژنوم که LD در آنها بالاست، نزدیک به ساختار نمونه‌هایی است که بر مبنای این مدل شبیه‌سازی می‌شوند.

فرض کنید می‌خواهیم بر پایه‌ی مدل فوق، نمونه‌ای تصادفی شامل n هاپلوتیپ برای l اسنیپ تولید کنیم. یک ویژگی ذاتی در هاپلوتیپ‌های مبتنی بر مدل فیلوژنی کامل که می‌تواند تولید تعداد دلخواه هاپلوتیپ‌های تصادفی را با محدودیت مواجه سازد این قضیه است که تعداد هاپلوتیپ‌های متمایز برای l اسنیپ در هر نمونه‌ی پدید آمده بر پایه‌ی مدل فیلوژنی کامل، حداکثر $l+1$ هاپلوتیپ است. برای اثبات ریاضی این قضیه [۵۰] را ببینید. کلیت اثبات به بیان ساده، بر این گزاره متکی است که در درخت فیلوژنی کامل، هر یک از یال‌های درخت نماینده‌ی یک جهش در یک اسنیپ و هر یک از هاپلوتیپ‌های متمایز در تناظر با یکی از راس‌های درخت‌اند. بر این اساس بدیهی است برای تولید نمونه‌ای با بیش از $l+1$ هاپلوتیپ بر روی l اسنیپ

[۱۹] Genetic drift
[۲۰] Coalescent model without recombination

و برای حفظ شرایط فیلوژنی کامل، لازم است تعداد معینی از نسخه‌های یکسان از هر یک از هاپلوتیپ‌های متمایز، در نمونه تکرار گردند.

تولید نمونه‌های تصادفی به روشی که ما در اینجا معرفی می‌کنیم در دو مرحله انجام می‌شود. ابتدا، یک نمونه‌ی تصادفی، شامل m هاپلوتیپ غیریکسان بر روی l اسنیپ بدست می‌آوریم که در آن $m \leqslant l + ۱$. سپس با تکرار نسخه‌های یکسان از هاپلوتیپ‌های بدست آمده، حجم نمونه را به n هاپلوتیپ می‌رسانیم. در این مرحله، فراوانی هاپلوتیپ‌ها در نمونه به گونه‌ای تعیین می‌شود که برخی قیود مرتبط با فراوانی آلل فرعی در اسنیپ‌ها ارضا گردد.

در الگوریتم زیر، روش ما برای تولید هاپلوتیپ‌های تصادفی غیریکسان (مرحله‌ی اول)، تحت قیود فیلوژنی کامل، تشریح می‌گردد.

الگوریتم *RandPerfectHap*

ورودی). l تعداد اسنیپ‌ها و m تعداد هاپلوتیپ‌ها ($m \leqslant l + ۱$) و یک هاپلوتیپ اولیه، $h_۱$ داده شده‌اند.

گام ۱). $m - ۲$ عدد به تصادف از میان اعداد ۱ تا $l - ۱$ انتخاب کن و آنها را، همراه با ۰ و l در یک مجموعه قرار بده. اعداد این مجموعه را به ترتیب صعودی در لیست $s = \langle s_۱, \ldots, s_m \rangle$ قرار بده.

گام ۲). یک جایگشت تصادفی از اعداد ۱ تا l انتخاب کن و آنرا با $\sigma = \langle \sigma_۱, \ldots, \sigma_l \rangle$ نشان بده.

گام ۳). برای $i = ۲, \ldots, m$ به ترتیب گام‌های زیر را انجام بده.

گام ۳-۱). یک عدد تصادفی از بین اعداد ۱ تا $i - ۱$ انتخاب کن و آنرا با π_i نشان بده.

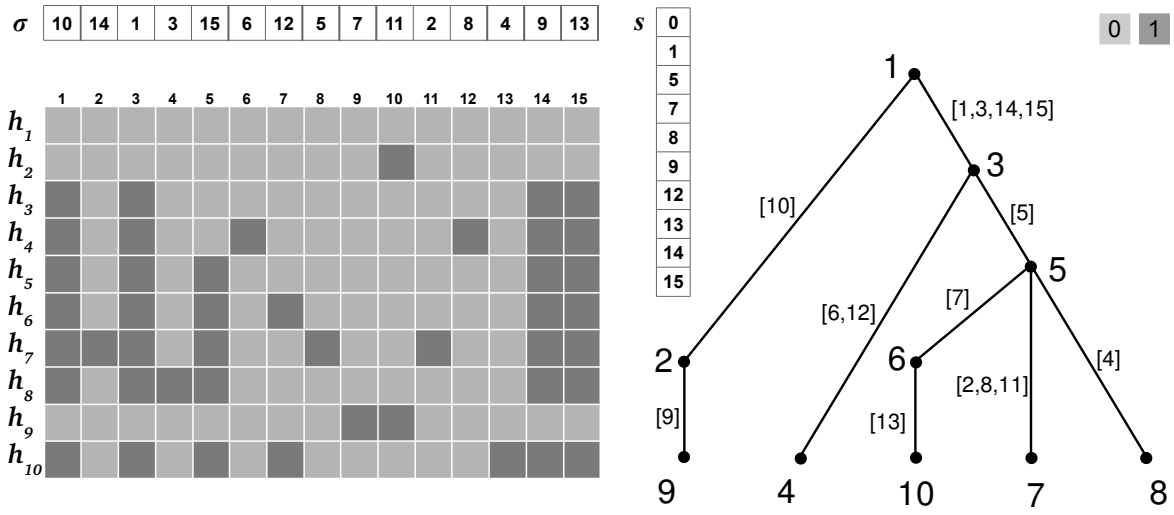
گام ۳-۲). با اعمال جایگشت σ بر روی اعداد $s_{i-۱} + ۱$ تا s_i، قرار بده،

$$M_i = \sigma \langle s_{i-۱} + ۱, s_{i-۱} + ۲, \ldots, s_i \rangle.$$

گام ۳-۳). هاپلوتیپ h_i با روابط زیر تعریف می‌شود؛

$$h_{ij} = \begin{cases} h_{\pi_i j} & \text{if} \quad j \notin M_i \\ ۱ - h_{\pi_i j} & \text{if} \quad j \in M_i \end{cases}$$

به کمک اعداد ذخیره شده در مجموعه‌های π و M در الگوریتم *RandPerfectHap*، می‌توان درخت فیلوژنی مرتبط با نمونه‌ی تصادفی تولید شده را نیز ترسیم کرد. در واقع، π_i نشاندهنده‌ی والد گره متناظر با هاپلوتیپ h_i و M_i مجموعه‌ی اسنیپ‌هایی است که در انتقال از h_{π_i} به h_i جهش داده می‌شوند. ما در این الگوریتم اجازه می‌دهیم هر ترکیب دلخواهی از آللهای 0و 1به عنوان هاپلوتیپ ریشه، h_1 قرار داده شود. هرچند انتخاب یک هاپلوتیپ خاص به عنوان هاپلوتیپ اولیه، تاثیری در درستی الگوریتم و مدل فیلوژنی کامل ندارد اما انتخاب $h_1 = \langle \mathbf{0}, \mathbf{0}, \ldots, \mathbf{0} \rangle$ متعارف است. شکل ۵۰۲ نمونه‌ای از اجرای الگوریتم و مجموعه هاپلوتیپ‌های تولید شده توسط آنرا نمایش می‌دهد.



شکل ۵۰۲: تولید هاپلوتیپ‌های تصادفی تحت مدل فیلوژنی کامل توسط الگوریتم *RandPerfectHap*

مرحله‌ی دوم شبیه‌سازی نمونه‌های تصادفی، تعیین فراوانی هر یک از هاپلوتیپ‌های تولید شده توسط *RandPerfectHap*، برای شرکت در مجموعه‌ی نهایی نمونه‌های هاپلوتیپ است. فرض کنید f_i تعداد دفعات تکرار هاپلوتیپ h_i در نمونه‌ی نهایی باشد. از بین قیودی که دیدگاه‌های مختلف درباره‌ی ساختار آماری هاپلوتیپ‌های جمعیت پیشنهاد می‌دهند، برای سادگی در اینجا، ما تنها، قید مرتبط با حداقل فراوانی آلل فرعی در اسنیپ‌ها را در تعیین فراوانی هاپلوتیپ‌های نمونه‌ی نهائی مورد توجه قرار می‌دهیم. بر اساس این قید، فراوانی نسبی آلل فرعی در نمونه‌ی تولید شده، در هیچ یک از اسنیپ‌ها نباید کمتر از مقدار داده شده‌ی μ_o باشد. در عمل، بسیاری از کاربردهای واقعی تنها جایگاه‌هایی را به عنوان اسنیپ به شمار می‌آورند که چندریختی بودن در آنها به طور معنادار قابل مشاهده باشد. از این رو، صرف نظر از این قید می‌تواند باعث از

بین رفتن اشکال چندریختی در برخی از l اسنیپ مورد مطالعه گردد.

توجه کنید که در اینجا ما نگران کران بالا برای فراوانی نسبی آلل فرعی نیستیم آنطور که لزوما کمتر از ۵۰ درصد باشد چرا که "آلل فرعی" عبارتی قراردادی برای آللی است که فراوانی کمتری در جمعیت (نمونه) داشته باشد و از این روی همواره می‌توان با تغییر آلل‌های 0به 1و بلعکس در اسنیپ‌هایی که فراوانی 1در آنها بیشتر از 0است این مشکل را بر طرف کرد. با در نظر گرفتن همین موضوع، قیود مطلوب برای فراوانی هاپلوتیپ‌های نمونه را می‌توان به صورت زیر خلاصه کرد؛

$$\begin{aligned}
\sum_{i=1}^{m} f_i &= n, \\
\sum_{i=1}^{m} h_{ij} f_i &\geq \mu_o n, \qquad j = 1, \ldots, l \\
\sum_{i=1}^{m} h_{ij} f_i &\leq (1 - \mu_o) n, \qquad j = 1, \ldots, l \\
f_i &\geq 1, \qquad i = 1, \ldots, m
\end{aligned}$$

روابط فوق، یک دستگاه نامعادلات خطی تشکیل می‌دهند که در آن $2l + m + 1$ نامعادله برای تعیین m مجهول وجود دارد. در این دستگاه، n تعداد هاپلوتیپ‌های نمونه و $\mu_o \leqslant 0/5$ مقادیری مفروض هستند. در حالت کلی، این دستگاه می‌تواند بدون جواب باشد. در مقابل، می‌توان نشان داد با حذف معادله‌ی اول از دستگاه، همواره یک جواب وجود خواهد داشت. بر همین اساس، ما این دستگاه را به شکل یک مسئله‌ی برنامه‌ریزی خطی به صورت زیر تغییر می‌دهیم؛

$$\begin{aligned}
N^* = \min_{f} N &= \sum_{i=1}^{m} f_i \\
s.t. & \\
\sum_{i=1}^{m} h_{ij} f_i &\geq \mu_o N\,, \qquad j = 1, \ldots, l \\
\sum_{i=1}^{m} h_{ij} f_i &\leq (1 - \mu_o) N\,, \qquad j = 1, \ldots, l \\
f_i &\geq 1\,, \qquad i = 1, \ldots, m
\end{aligned}$$

نکته قابل تامل در استفاده از این رویکرد برای تولید نمونه‌های تصادفی، اختلاف بین تعداد هاپلوتیپ‌های مورد نیاز، n و مقدار جواب بهینه‌ی این مسئله‌ی برنامه‌ریزی خطی، N^* است. اگر $N^* < n$ باشد با بزرگ

کردن فراوانی تمام هاپلوتیپ‌ها به طور یکسان توسط ضریبی متناسب با نسبت n/N^* می‌توان بدون نقض قیود مسئله، جواب مطلوب را بدست آورد. می‌توان نشان داد در حالتی که $N^* > n$ است دستگاه نامعادلات اولیه جواب ندارد. در این حالت می‌توان روال مرحله‌ی اول، یعنی الگوریتم *RandPerfectHap* را مجددا اجرا کرد و مسئله‌ی برنامه‌ریزی خطی دیگری را بر حسب مقادیر جدید h_{ij} تشکیل داد. از آنجا که هر اجرای الگوریتم *RandPerfectHap* به طور تصادفی نمونه‌های متفاوتی تولید می‌کند تکرار این روال می‌تواند به یک مسئله‌ی دارای جواب منتهی گردد. انتخاب مقادیر متفاوت برای تعداد هاپلوتیپ‌های متمایز، m و نیز تغییر هاپلوتیپ اولیه h_1 در هر فراخوانی *RandPerfectHap*، می‌تواند به یافتن جواب کمک کند. ما این روش را با بکارگیری روال‌های حل مسائل برنامه‌ریزی خطی در نرم‌افزار MATLAB، پیاده‌سازی کرده‌ایم و از طریق آن نمونه‌ی نهایی هاپلوتیپ‌های شبیه‌سازی شده را تولید می‌کنیم.

تولید نمونه‌های تصادفی برای ژنوتیپ‌ها

برای شبیه‌سازی نمونه‌های ژنوتیپ ابتدا، نمونه‌ی بزرگی از هاپلوتیپ‌ها را توسط روشی که در بالا شرح داده شد تولید می‌کنیم. حال کافی است جفت‌های تصادفی از این هاپلوتیپ‌ها را ترکیب کنیم تا نمونه‌ای از ژنوتیپ‌های تصادفی در اختیار داشته باشیم. با این حال نباید فراموش کرد که از این طریق ممکن است نمونه‌هایی بدست آیند که تنوع هاپلوتیپ‌ها در آنها نمونه‌ی مناسبی از آنچه در واقعیت در نواحی LD بالا یا بلوک‌های هاپلوتیپی مشاهده می‌شود نباشد. از این روی در روال تولید ژنوتیپ‌های تصادفی ما به جای در نظر گرفتن قیود مربوط به حداقل فراوانی آلل فرعی در اسنیپ‌ها، به سادگی پس از بدست آوردن نمونه‌ای از هاپلوتیپ‌های متمایز توسط الگوریتم *RandPerfectHap*، فراوانی α درصد از هاپلوتیپ‌ها را به طور تصادفی با مقادیری بزرگتر از β نسبت‌دهی می‌کنیم. مقادیر $\alpha = ۰/۸$ و $\beta = ۰/۰۵$ مقادیری هستند که معمولا در تعریف بلوک‌های هاپلوتیپ مبتنی بر مفهوم "هاپلوتیپ‌های رایج" مورد استفاده قرار می‌گیرند (بخش ۴۰۱).

از نظر آماری تنوع ژنوتیپ‌هایی که بدین ترتیب شبیه‌سازی می‌شوند ساختاری نزدیک به ژنوتیپ‌های واقعی در نواحی LD بالا دارد. بر همین اساس، ما برای شبیه‌سازی ژنوتیپ‌ها در مقیاس ژنومی، نمونه‌های تصادفی مستقلی را با تعداد ژنوتیپ‌های یکسان و با طول‌هایی بین ۵ تا ۳۰ اسنیپ تولید می‌کنیم که با در کنار هم

قرار دادن آنها، دسته‌ای از ژنوتیپ‌ها در امتداد ژنوم بدست می‌آیند. در نمونه‌هایی که بدین ترتیب بدست می‌آیند استقلال هاپلوتیپ‌های شبیه‌سازی شده در قطعات مجاور هم، تنها ابزاری است که می‌تواند باعث ایجاد ساختار بلوکی در نمونه‌های تولید شده گردد. در بخش ۶۰۵۰۲، ما مدل کامل‌تری را برای شبیه‌سازی هاپلوتیپ‌ها در امتداد ژنوم بکار می‌گیریم که در آن، تغییرات نرخ نوترکیبی در امتداد ژنوم به عنوان عامل ایجاد ساختار بلوکی در هاپلوتیپ‌های شبیه‌سازی شده مورد توجه قرار می‌گیرد.

۳۰۲ یک شاخص برای تعیین وجود همبستگی بین اسنیپ‌ها

در این بخش، ما به بحث درباره‌ی نحوه‌ی استفاده از آزمون دقیق فیشر[۲۱] برای کمّی‌سازی مفهوم اسنیپ‌های "قویاً همبسته"[۲۲] می‌پردازیم. طی این بحث، شاخص جدیدی برای تعیین میزان همبستگی بین جفت اسنیپ‌ها معرفی می‌شود. برای محاسبه‌ی این شاخص، کمیت D' برای سنجش LD بین جفت اسنیپ‌ها و آزمون دقیق فیشر برای تعیین معناداری مقدار مشاهده شده‌ی LD، مورد استفاده قرار می‌گیرند. استفاده از آزمون دقیق فیشر در علوم زیستی، موضوعی کاملا فراگیر است. با این وجود، درباره‌ی بکارگیری آن در زمینه‌ی مطالعه‌ی همبستگی بین اسنیپ‌ها، تلاش‌های بسیار پراکنده‌ای صورت گرفته است[۱۵۱، ۱۵۲].

فرض کنید نمونه‌ای شامل n هاپلوتیپ برای l اسنیپ، در یک ناحیه‌ی کروموزومی که علاقمند به بررسی همبستگی بین اسنیپ‌های آن هستیم داده شده است. کار را ابتدا با تشکیل یک جدول توافقی برای هر دو اسنیپ واقع در این ناحیه آغاز می‌کنیم. فراوانی مشاهده شده‌ی هر یک از چهار هاپلوتیپ ممکن برای هر جفت اسنیپ را در نمونه‌ی داده شده بدست می‌آوریم. فرض کنید n_{00}، n_{01}، n_{10} و n_{11} به ترتیب فراوانی هاپلوتیپ‌های 00، 01، 10و 11در یک جفت اسنیپ باشند. بر این اساس برای هر جفت اسنیپ یک جدول توافقی همانند جدول زیر بدست می‌آید؛

[۲۱] Fisher's exact test
[۲۲] strongly associated

	$Y=\mathbf{0}$	$Y=\mathbf{1}$	
$X=\mathbf{0}$	$n_{\mathbf{00}}$	$n_{\mathbf{01}}$	$n-n_a$
$X=\mathbf{1}$	$n_{\mathbf{10}}$	$n_{\mathbf{11}}$	n_a
	$n-n_b$	n_b	n

که در آن X و Y نشان‌دهنده‌ی متغیرهای تصادفی متناظر با هر یک از دو اسنیپ مورد بررسی و $n_a = n_{\mathbf{10}} + n_{\mathbf{11}}$ و $n_b = n_{\mathbf{01}} + n_{\mathbf{11}}$ فراوانی آلل **1**، به ترتیب در اسنیپ‌های X و Y هستند. این دو عدد، یعنی n_a و n_b را اصطلاحا فراوانی حاشیه‌ای اسنیپ‌ها در جفت اسنیپ مورد بررسی می‌نامیم. به وضوح ملاحظه می‌شود که هر جدول توافقی را می‌توان، تنها با داشتن مقادیر n، n_a، n_b و $n_{\mathbf{11}}$ به طور یکتا تعیین کرد. به کمک این جدول و بر حسب نمادگذاری قرارداد شده، به سادگی می‌توان آماره‌های D و r^2 را در نمونه‌ی داده شده برآورد کرد. داریم؛

$$D = (n_{\mathbf{11}}n_{\mathbf{00}} - n_{\mathbf{10}}n_{\mathbf{01}})/n^2 \qquad (۲۰۲)$$

و

$$r^2 = \frac{(nn_{\mathbf{11}} - n_a n_b)^2}{n_a(n-n_a)n_b(n-n_b)}$$

این آماره‌ها را پیشتر در بخش ۴۰۱، بر حسب روابطی دیگر و به عنوان کمیت‌های رایج برای اندازه‌گیری LD معرفی کرده‌ایم. به طور مشابه می‌توان با استفاده از رابطهٔ (۸۰۱)، مقدار D' را نیز بر حسب فراوانی مشاهده‌ی ژنوتیپ‌ها در یک جفت اسنیپ بدست آورد. اما آماره‌ی دیگری که در این بخش، برای ارزیابی همبستگی بین اسنیپ‌ها بکار خواهیم گرفت، آماره‌ی آزمون دقیق فیشر است[۱۵۳]. در واقع، انجام آزمون دقیق فیشر، منوط به محاسبه‌ی مقدار این آماره، برای جدول توافقی مرتبط با داده‌های مشاهده است و بر اساس رابطه‌ی زیر تعریف می‌شود؛

$$F_{ex} = \frac{n_a! n_b! (n-n_a)! (n-n_b)!}{n! n_{\mathbf{00}}! n_{\mathbf{01}}! n_{\mathbf{10}}! n_{\mathbf{11}}!} \qquad (۳۰۲)$$

ایده‌ی را که فیشر برای بررسی وجود ارتباط معنادار بین دو پدیده‌ی آماری بکار می‌گرفت، در اینجا، در قالب مثالی از اسنیپ‌ها شرح می‌دهیم. فرض کنید نمونه‌ای شامل ۱۰ فرد غیرخویشاوند را از جمعیت انتخاب کرده‌ایم و ژنوتیپ‌های آنها را در دو اسنیپ X و Y مورد مطالعه قرار می‌دهیم. به عنوان مثال فرض کنید در اسنیپ X، ۳ فرد هموزیگوت با آلل وحشی، ۵ فرد هتروزیگوت و ۲ فرد هموزیگوت با آلل جهش‌یافته داریم؛ یعنی در مجموع $۲ \times ۳ + ۵ = ۱۱$ هاپلوتیپ با آلل 0و $۲ \times ۲ + ۵ = ۹$ هاپلوتیپ با آلل 1در اسنیپX مشاهده کرده‌ایم. به همین ترتیب در اسنیپ Y فرض کنید در مجموع ۱۵ هاپلوتیپ با آلل 0و ۵ هاپلوتیپ با آلل 1مشاهده کرده‌ایم. پس از تفکیک ژنوتیپ‌ها به هاپلوتیپ‌ها می‌توانیم فراوانی هر یک از چهار ترکیب 00، 01، 10و 11را در این دو اسنیپ، در نمونه‌های داده شده بدست آوریم. جدول توافقی زیر، فراوانی مشاهده‌ی هر یک از هاپلوتیپ‌ها در این نمونه را نشان می‌دهد.

	$Y=0$	$Y=1$	
$X=0$	۱۱	۰	۱۱
$X=1$	۴	۵	۹
	۱۵	۵	۲۰

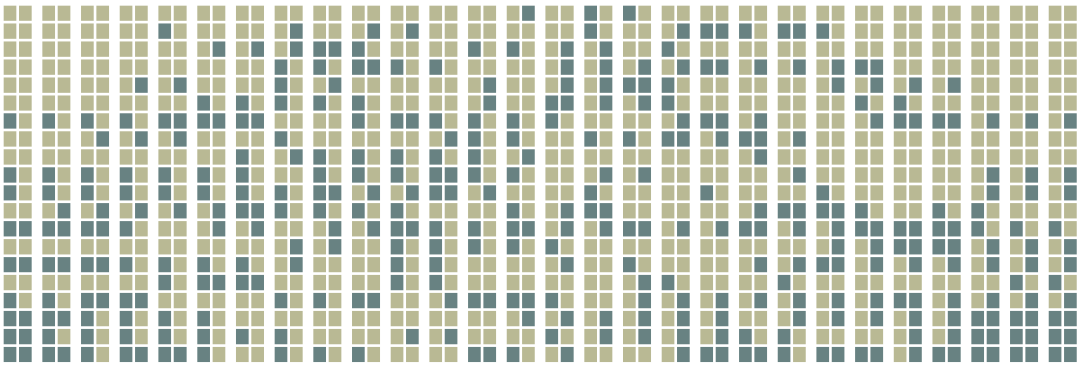
جدول توافقی فراوانی مشاهدات در مثال ۱

	$Y=0$	$Y=1$	
$X=0$	۹	۲	۱۱
$X=1$	۶	۳	۹
	۱۵	۵	۲۰

جدول توافقی فراوانی مشاهدات در مثال ۲

برای چنین نمونه‌ای، حتی بدون نیاز به محاسبه، وجود ارتباط بین اسنیپ‌های X و Y را به سختی می‌توان نادیده گرفت؛ چرا که در تمام موارد، آلل 1در اسنیپ Y فقط در کنار آلل 1در اسنیپ X مشاهده می‌شود. حال نمونه‌ی دیگری را در نظر بگیرید که در آن $n_{11} = ۳$ است (مثال ۲). در اینجا به سادگی نمی‌توان درباره‌ی وجود یا عدم وجود ارتباط بین دو اسنیپ قضاوت کرد.

بدیهی است انتظار نداریم ترتیب انتخاب این افراد، تاثیری در نتیجه‌گیری ما درباره‌ی وجود یا عدم وجود همبستگی بین این دو اسنیپ داشته باشد. در واقع، با هر جایگشت تصادفی بر روی ترتیب افراد در نمونه‌ی داده شده، باز همان جدول توافقی قبل بدست می‌آید. اگر بتوان فرض کرد پراکندگی آلل‌های مختلف در جمعیت، در هر یک از دو اسنیپ نیز مستقل از دیگری است آنگاه می‌توانیم ترتیب نمونه‌ها را در یک اسنیپ، در برابر اسنیپ دیگر بُر بزنیم(شکل ۶۰۲). این فرض چندان دور از واقعیت نیست. در واقع، این فرض تعبیر ساده شده‌ای از وقوع رویدادهای متعدد نوترکیبی بین دو اسنیپ است. این دیدگاه به طور دقیق‌تر، به صورت یک فرایند تصادفی در پیوست الف تشریح شده است. با کمی محاسبه می‌توان نشان داد، احتمال مشاهده‌ی



5	4	3	2	2	2	3	0	2	2	2	2	2	2	1	1	2	1	4	3	2	3	3	4	4	3	3	3
1	9	30	38	38	38	30	3	38	38	38	38	38	38	19	19	38	19	9	30	38	30	30	9	9	30	30	30
.41	.16	.03	0	0	0	.03	.27	0	0	0	0	0	0	.08	.08	0	.08	.16	.03	0	.03	.03	.16	.16	.03	.03	.03
1	.64	.27	.11	.11	.11	.27	1	.11	.11	.11	.11	.11	.11	.56	.56	.11	.56	.64	.27	.11	.27	.27	.64	.64	.27	.27	.27

شکل ۶۰۲: مشاهده‌ی نمونه‌هایی با فراوانی‌های حاشیه‌ای یکسان در جمعیت

هر جفت از ستون‌ها، نشان‌دهنده‌ی هاپلوتیپ‌های دو اسنیپ مورد مطالعه، X و Y، در نمونه‌ای از ۱۰ فرد جمعیت هستند. در نمونه‌ی سمت چپ، ۵ هاپلوتیپ **11**و در نمونه‌ی سمت راست، ۳ هاپلوتیپ **11**مشاهده کرده‌ایم. در هر دو نمونه، تعداد افراد دارای آلل **1**در X یکسان است و برابر ۹ است. همچنین، تعداد افراد دارای آلل **1**در Y نیز در هر دو نمونه، یکسان و برابر ۵ است. فرض کنید در این جمعیت، نوترکیبی بین X و Y، هاپلوتیپ‌های جمعیت را دستخوش تحول می‌کند. اگر از بین افراد جمعیت، نمونه‌های ۱۰ نفره‌ی دیگری انتخاب کنیم که تعداد افراد دارای آلل **1**در آنها دقیقا همانند دو نمونه‌ی اولیه باشد آنگاه نمونه‌هایی بدست خواهیم آورد که در ۳۰ درصد موارد، همان ترکیب نمونه‌ایی مثال سمت راست را خواهند داشت. بر خلاف آن، نمونه‌هایی همانند نمونه‌ی سمت چپ بسیار به ندرت ممکن است مشاهده شوند. اعدادی که در پائین هر جفت از ستون‌ها نوشته شده‌اند، به ترتیب از بالا به پائین عبارتند از: $n_{\mathbf{11}}$، F_{ex} بر حسب درصد، r^2 و $|D'|$.

نمونه‌ای شامل $n_{\mathbf{11}}$ هاپلوتیپ **11**در بین نمونه‌هایی که در آنها دو اسنیپ X و Y به ترتیب با n_a و n_b آلل **1**، "به طور تصادفی در کنار یکدیگر" قرار گرفته‌اند به صورت زیر بدست می‌آید؛

$$P(n_{\mathbf{11}}|n_a, n_b, n) = \frac{\binom{n_a}{n_{\mathbf{11}}}\binom{n-n_a}{n_b-n_{\mathbf{11}}}}{\binom{n}{n_b}} \qquad (۴۰۲)$$

این احتمال در ریاضیات، با نام احتمال فوق‌هندسی[۲۳] شناخته می‌شود و مقدار آن، همان آماره‌ی آزمون دقیق فیشر، F_{ex} را نشان می‌دهد. رابطهٔ (۴۰۲) معادل رابطهٔ (۳۰۲) است؛ از این رو، ارزش آن با جابجا کردن نقش متغیرهای X و Y تغییر نمی‌کند. با ارزیابی رابطهٔ (۴۰۲)، مقدار احتمال برای مثال ۱، برابر با $P(n_{\mathbf{11}} = ۵) = ۰/۰۰۸۱$ و برای مثال ۲، برابر $P(n_{\mathbf{11}} = ۳) = ۰/۲۹۸۰$ است. این بدان معنا است که

[۲۳] Hypergeometric

اگر نوترکیبی بین دو اسنیپ مورد مطالعه، به طور گسترده در بین هاپلوتیپ‌های افراد مختلف جمعیت رواج داشته باشد و مثلا ۱۰۰ بار نمونه‌های مختلف ۱۰ نفری، همانند مثال‌های فوق، از جمعیت انتخاب کنیم در نزدیک به ۳۰ مورد، مشاهداتی همانند مثال دوم خواهیم داشت در حالیکه، به سختی می‌توان یک نمونه همانند مثال اول بدست آورد. از این رو، اگر در جمعیت مورد مطالعه، مقدار این احتمال در مشاهدات بدست آمده، پائین باشد این موضوع خود می‌تواند نشانه‌ای دال بر عدم احتمال نوترکیبی یا پائین بودن نرخ آن بین دو اسنیپ به حساب آید. البته، باز هم یادآوری می‌کنیم که سازوکارهای تکوینی در طبیعت می‌توانند به مراتب پیچیده‌تر از مدل ساده‌ای باشند که ما در اینجا مثال زدیم.

رویکردی را که فیشر برای بررسی وجود ارتباط معنادار بین دو پدیده‌ی آماری مورد توجه قرار می‌دهد، مشابه آنچه در مثال قبل ملاحظه کردید، مستقل از هیچ گونه پیش فرضی درباره‌ی نوع توزیع احتمال متغیرهای تصادفی موضوع مورد مطالعه است. در واقع، ایده‌ی اصلی آزمون دقیق فیشر، در نظر گرفتن احتمال مشاهده‌ی نمونه در بین ترکیب‌های تصادفی دو متغیر مورد مطالعه با فرض ثابت بودن فراوانی حاشیه‌ای آنها، به عنوان نشانه‌ای منطقی، دال بر عدم وجود ارتباط معنادار بین آنها است. یادآور می‌شود که در دیگر آزمون‌های رایج برای همبستگی مثل آزمون مربع کای، وجود ارتباط بین دو متغیر، به طور غیرمستقیم توسط آماره‌های آزمون تقریب زده می‌شود. در این روش‌ها معمولا، متغیرهای تصادفی مورد مطالعه یا آماره‌های مرتبط با فراوانی نسبی آنها، دارای توزیع نرمال فرض می‌شوند. هر چند از نظر تئوری می‌توان توزیع احتمالاتی آماره‌های نسبت را به ازای مقادیر بالای حجم نمونه، تقریبا نرمال فرض کرد اما در حالت کلی، استفاده از آزمون‌های مبتنی بر آماره‌ای که دارای توزیع مجانبی، مثل مربع کای است، خالی از اشکال نیست. آزمون همبستگی فیشر از این روی دقیق نامیده می‌شود که آماره‌ی مورد استفاده در آن درگیر با این نوع تقریب‌ها نیست. به همین خاطر، کاربرد این آزمون در مواردی که اندازه‌ی نمونه کوچک است برای بررسی وجود همبستگی بین متغیرهای مورد مطالعه بسیار رایج است. به ویژه، استفاده از آزمون دقیق فیشر به جای آزمون مربع کای در جداولی که تعداد مشاهدات در یکی از خانه‌های آن کمتر از ۵ باشد به عنوان یک قاعده‌ی متعارف، ممکن است برای شما آشنا باشد.

محاسبه‌ی pـ مقدار و استفاده از آزمون دقیق فیشر برای تعیین معناداری همبستگی

یکی از نکات درخور توجه درباره‌ی آماره‌ی F_{ex}، وابستگی آن به اندازه‌ی نمونه است. به مثال‌های بالا باز می‌گردیم. فرض کنید اندازه‌ی نمونه را از ۱۰ نفر به ۱۰۰ نفر افزایش داده‌ایم و هر یک از چهار هاپلوتیپ متمایز برای دو اسنیپ، در نمونه‌هایی به همان نسبت بیشتر مشاهده شده‌اند؛ یعنی، $n = ۲۰۰, n_a = ۹۰, n_b = ۵۰$ و $n_{11} = ۳۰$. مقدار آماره‌های D، D' و $r^۲$ که بر حسب مقادیر فراوانی نسبی تعریف می‌شوند، با تغییر اندازه‌ی نمونه تغییر نمی‌کنند و برای این مثال، داریم، $D = ۰/۰۳۷۵, D' = ۰/۲۷۲۷, r^۲ = ۰/۰۳۰۳$. اما مقدار احتمال فوق هندسی با تغییر اندازه‌ی نمونه تغییر می‌کند و در این مثال، داریم، $F_{ex} = ۰/۰۰۶۵$ که به مراتب، کوچکتر از مقدار آن به ازای $n = ۱۰$ است.

ما این ویژگی را به عنوان ابزاری برای تعیین سطح معناداری آزمون همبستگی مورد استفاده قرار می‌دهیم. بدیهی است نتیجه‌ی یک آزمون، وقتی برای برآورد آماره‌های آن، از نمونه‌ای با حجم بالا استفاده می‌کنیم قابل اعتمادتر از آزمونی است که با استفاده از نمونه‌های کوچک اجرا می‌شود. میزان اعتبار نتایج بدست آمده از یک آزمون را به بیان ریاضی سطح معناداری[۲۴] آن آزمون می‌نامیم.

آماره‌ی آزمون دقیق فیشر، از این جهت که وابسته به اندازه‌ی نمونه است به طور مستقیم نمی‌تواند معیار مناسبی برای ارائه‌ی یک تعریف معین برای "اسنیپ‌های قویاً همبسته" باشد و به همین دلیل، ما از $|D'|$ که مستقل از اندازه‌ی نمونه است برای تعریف شاخص همبستگی استفاده خواهیم کرد. با این حال در نقطه‌ی مقابل، به کمک آزمون دقیق فیشر می‌توانیم معیار مناسبی برای تعیین معناداری همبستگی ادعا شده توسط $|D'|$، ارائه کنیم.

در مثال ۱، مقدار D' برابر ۱ است که بیشترین مقدار ممکن برای این آماره در حالت کلی است. بنابراین می‌توان ادعا کرد که بین دو اسنیپ مورد مطالعه از نظر آماری همبستگی وجود دارد یا به تعبیری دیگر در وضعیت LD کامل هستند. به یاد آورید که اندازه‌ی نمونه در این مثال $n = ۱۰$ است. سئوال اینجاست که تا چه حد ممکن است این ادعا با مشاهدات بیشتر، دچار تغییر گردد. یا در مثال ۲، داریم، $D' = ۰/۲۷۲۷$ که بنابر مقادیر متعارف آستانه، به عنوان نشانه‌ای از وجود نوترکیبی بین دو اسنیپ تلقی می‌شود. اما آیا با توجه

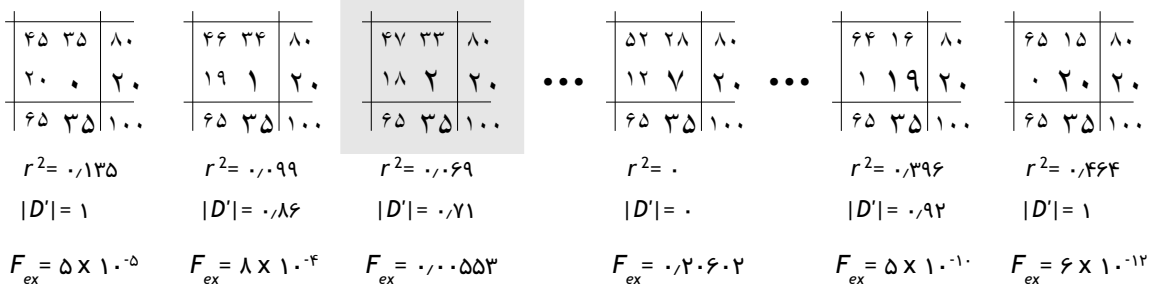
[۲۴] Level of significance

به حجم نسبتا کم نمونه در این مثال می‌توان به این نتیجه‌گیری اعتماد کرد. تفاوت آماره‌ی F_{ex} در این دو نمونه و نمونه‌ای با فراوانی‌های ۱۰ برابر بیشتر، راهنمای ما برای تصمیم‌گیری درباره‌ی این دو ادعا است. در مثال اول، با اینکه حجم نمونه کوچک است اما احتمال مشاهده‌ی هر نمونه‌ی دیگر با مقدار $D' = ۱$ کمتر از یک درصد (۰/۰۰۸۱) است و بنابراین قضاوت درباره‌ی همبسته بودن دو اسنیپ چندان بی‌اعتبار نیست. بر عکس، حجم نمونه‌ایی $n = ۱۰$ ، برای دومین مثال کمتر از آن است که بتوان ادعای محکمی درباره‌ی مستقل بودن دو اسنیپ داشت، چرا که برای نزدیک به ۳۰ درصد از دیگر نمونه‌ها، با همان فراوانی حاشیه‌ای، مقادیری بزرگتر یا برابر با مقدار بدست آمده برای D' در این نمونه بدست می‌آید(شکل ۶۰۲). در نمونه‌ای با تعداد مشاهدات ۱۰ برابر قبل، با اینکه مقدار D' با این افزایش مشاهدات، تغییری نمی‌کند اما به طور معنادار می‌تواند نشانه‌ی استقلال دو اسنیپ باشد، چرا که در این حالت، در بین نمونه‌های مختلف در کمتر از یک درصد موارد (با احتمال ۰/۰۰۷۶) می‌توان نمونه‌ای با $D' > ۰/ ۲۷۲۷$ یافت.

به طور رسمی، سطح معناداری در آزمون دقیق فیشر، با محاسبه‌ی pـ مقدار جدول توافقی مرتبط با مشاهده، تعیین می‌گردد. pـ مقدار، احتمال مشاهده‌ی نمونه‌ای با فراوانی حاشیه‌ایی همانند نمونه‌ی مورد بررسی است که مقدار سنجه‌ی همبستگی در آن، بزرگتر یا برابر با مقدار سنجه‌ی همبستگی در نمونه‌ی مورد بررسی باشد. اگر تمام حالت‌های ممکن برای جدول‌های توافقی ۲×۲ با فراوانی‌های حاشیه‌ای یکسان را در یک ردیف به ترتیب از چپ به راست، بر حسب ترتیب صعودی برای مقادیر n_{11} بنویسیم (مانند شکل ۷۰۲) آنگاه می‌توان نشان داد که مقدار سنجه‌های متعارف برای آزمون همبستگی، مثل $r^۲$ یا $|D'|$، در این جدول‌ها، ابتدا با افزایش مقدار n_{11} کاهش می‌یابند تا جائیکه $n\, n_{11} - n_a\, n_b$ برابر صفر شود یا مقدار آن تغییر علامت دهد و پس از آن سیر افزایشی می‌گیرند. بدین ترتیب برای محاسبه‌ی pـ مقدار کافی است حاصلجمع مقدار آماره‌ی F_{ex} را متناظر با هر یک از جدول‌های دو طرف لیست، تا جائیکه به جدول نمونه‌ی مورد مطالعه‌ی یا جدول متناظر با سنجه‌ی همبستگی برابر آن، در طرف دیگر لیست می‌رسیم، بدست آوریم. pـ مقداری که با این روش بدست می‌آید را اصطلاحا pـ مقدار دو طرفه[۲۵] می‌نامند.

در مسئله‌ی بررسی همبستگی بین اسنیپ‌ها، استفاده از pـ مقدار یکطرفه[۲۶] منطقی‌تر به نظر می‌رسد. برای

[۲۵] Two-tailed p-value
[۲۶] One-tailed p-value



شکل ۷۰۲: نحوه‌ی محاسبه‌ی pـ مقدار در آزمون دقیق فیشر
تمام جدول‌های توافقی با فراوانی‌های حاشیه‌ای برابر با نمونه‌ی مشاهده شده را بر حسب ترتیب صعودی n_{11} در یک ردیف قرار می‌دهیم. مجموع مقادیر F_{ex} در جدول‌های واقع در یک طرف جدول مورد بررسی، pـ مقدار آزمون همبستگی مرتبط با آن جدول را تعیین می‌کند. جزئیات دقیق‌تر را در متن مطالعه کنید.

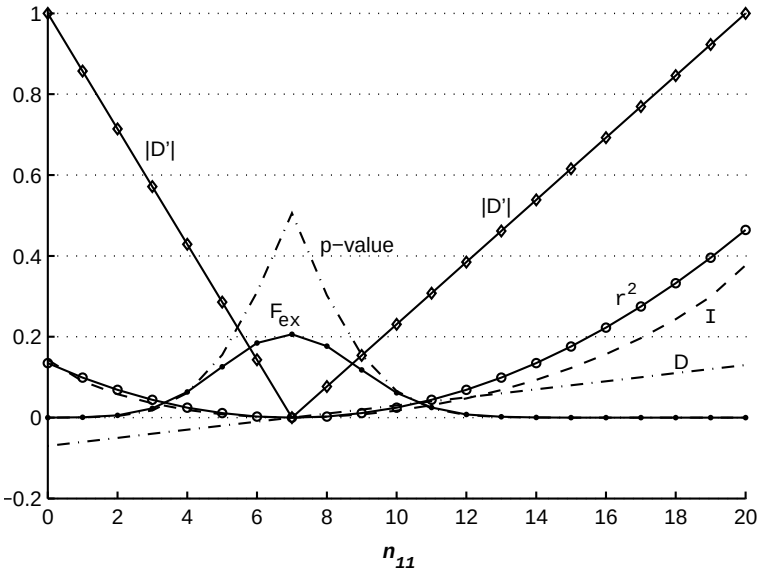
محاسبه‌ی pـ مقدار یکطرفه، مجموع مقادیر F_{ex} را تنها به ازای جدول‌های همان سمتی از لیست که جدول نمونه‌ی مورد مطالعه‌یمان در آن قرار دارد بدست می‌آوریم. این رویکرد از این رو ارجح است که از نظر زیستی، همبستگی آلل‌ها بین دو اسنیپ، یا در بین تمام نمونه‌های به تصادف انتخاب شده از جمعیت، به طور مثبت مشاهده می‌شود، یعنی ترکیب 00بیشترین فراوانی را در آن دارد و یا در تمام نمونه‌های به تصادف انتخاب شده از جمعیت، منفی است، یعنی بیشترین فراوانی در یکی از ترکیب‌های 01یا 10دیده می‌شود. آخرین نکته‌ای که ما در ارتباط با محاسبه‌ی pـ مقدار مورد توجه قرار می‌دهیم، نحوه‌ی شمول F_{ex} مرتبط با جدول نمونه‌ی مورد مطالعه در حاصلجمعی است که به عنوان pـ مقدار تعریف می‌کنیم. از جنبه‌ی ریاضی، دیدگاه‌های متعددی برای این موضوع وجود دارد که از این بین، استفاده از قاعده‌ی موسوم به mid-p از نظر تئوری، آماره‌ی تصمیم‌گیری با ریسک پائین‌تری را تعریف می‌کند[۱۵۴، ۱۵۵]. بر این اساس، pـ مقداری که ما برای بررسی سطح معناداری همبستگی در بین اسنیپ‌ها بکار می‌بریم، بر حسب روابط زیر تعریف می‌شود؛

$$p_{\text{left}} = \frac{۱}{۲} F_{ex}(n_{11}) + \sum_{i<n_{11}} F_{ex}(i)\,,$$

$$p_{\text{right}} = \frac{۱}{۲} F_{ex}(n_{11}) + \sum_{i>n_{11}} F_{ex}(i)\,,$$

$$p\_value_{\text{one-tailed}} = \begin{cases} p_{\text{left}} , & \text{if } n_{11} < n_{\max} \\ p_{\text{right}} , & \text{if } n_{11} > n_{\max} \\ max(p_{\text{left}}, p_{\text{right}}) , & \text{if } n_{11} = n_{\max} \end{cases} \quad (۵۰۲)$$

در رابطهٔ (۵۰۲)، $F_{ex}(i) = P(n_{11} = i | n_a, n_b, n)$ آماره‌ی آزمون دقیق فیشر مرتبط با جدول توافقیی است که در آن $n_{11} = i$ است و $n_{\max} = argmax_i F_{ex}(i)$ معادل مقداری است که به ازای آن، D تغییر علامت می‌دهد و در واقع "متوازن‌ترین" جدول را در بین جدول‌های لیست، معین می‌کند(شکل ۸۰۲).



شکل ۸۰۲: نمودار تغییرات آماره‌های همبستگی و pـ مقدار بر حسب فراوانی هاپلوتیپ **11**
n_{11} فراوانی مشاهده‌ی هاپلوتیپ **11**در نمونه‌های مختلف است. تعداد کل هاپلوتیپ‌ها در هر یک از این نمونه‌ها ۱۰۰ است. فراوانی‌های حاشیه‌ای در آنها به طور ثابت، برابرند با $n_a =$ ۳۵ و $n_b =$ ۲۰ برای اسنیپ X و برای اسنیپ Y. منحنی I، تغییرات مقدار اطلاع دوطرفه[۲۷] بین دو اسنیپ را به ازای مقادیر مختلف n_{11} نمایش می‌دهد. منحنی p-value در این شکل، pـ مقدار یکطرفه را نشان می‌دهد.

تعریف شاخص همبستگی بین اسنیپ‌ها

آزمون دقیق فیشر را برای بررسی همبستگی بین هر جفت از اسنیپ‌ها، در ناحیه‌ی کروموزومی مورد مطالعه، ارزیابی می‌کنیم و pـ مقدار متناظر با آزمون را به وسیله‌ی رابطهٔ (۵۰۲) بدست می‌آوریم. همچنین برآورده نقطه‌ای آماره‌ی D' را نیز به ازای هر یک از جفت اسنیپ‌های واقع در ناحیه‌ی مورد مطالعه بدست می‌آوریم

(روابط ۲۰۲ و ۸۰۱).

رویکرد ما برای مطالعه‌ی پراکندگی الگوی LD در یک ناحیه‌ی کروموزومی، همانند روش گابریل است که در آن، وجود یا عدم وجود همبستگی بین اسنیپ‌ها، بر حسب یکی از سه وضعیت "قویا همبسته"، "نوترکیب" و یا "نامعلوم" بیان می‌شود[۱۱۵]. ما نیز در اینجا، نتایج بدست آمده از آزمون دقیق فیشر و برآورد آماره‌ی D' را در قالب یک شاخص سه وضعیتی برای هر جفت از اسنیپ‌ها خلاصه می‌کنیم. شاخص پیشنهادی ما به طور دقیق به صورت زیر تعریف می‌شود؛

> فرض کنید مقادیر آستانه‌ای D'_o و p_o داده شده‌اند و $\tilde{D}'$ و $p\_val_{one-tailed}$ به ترتیب نشان‌دهنده‌ی مقدار برآورد شده‌ی آماره‌ی D' و pـ مقدار یکطرفه‌ی آزمون دقیق فیشر برای یک جفت از اسنیپ‌ها، مثل (X, Y) در نمونه‌ی داده شده باشند. جفت اسنیپ (X, Y) را "همبسته" می‌نامیم اگر $p\_val_{one-tailed} \leqslant p_o$ و $|\tilde{D}'| \geqslant D'_o$ و "مستقل" می‌نامیم اگر $|\tilde{D}'| < D'_o$ باشد. اگر $p\_val_{one-tailed} > p_o$ باشد وضعیت جفت اسنیپ را "از نظر آماری غیر معنادار"[۲۸] به حساب می‌آوریم.

در قیاس با روش گابریل، در رویکرد ما برای تعریف شاخص همبستگی بین اسنیپ‌ها، به جای استفاده از برآورد فاصله‌ای، آزمون دقیق فیشر برای سنجش معناداری برآورد D' بکار گرفته می‌شود. انتخاب مقادیر مناسب برای آستانه‌های D'_o و p_o بیشتر به شرایط مسئله و مشخصات ژنتیکی ناحیه‌ی کروموزومی مورد مطالعه بستگی دارد. یک انتخاب متعارف برای D'_o، معمولا عددی نزدیک به ۱ در بازه‌ی $[۰, ۱]$ است. به طور کلی، D'_o باید کوچکترین مقدار شناخته شده برای $|\tilde{D}'|$ در بین جفت اسنیپ‌هایی باشد که از نظر تجربی، همبسته به شمار می‌روند. در سوی دیگر، p_o، به نوعی میزان سختگیری ما برای تصمیم در مورد "همبسته" یا "مستقل" انگاشتن جفت اسنیپ‌ها را نشان می‌دهد. به عبارت دقیق، p_o نشان‌دهنده‌ی مقدار تحملی است که می‌توانیم نسبت به تشخیص اشتباه وضعیت همبستگی یک جفت اسنیپ روا داریم. متعارف آن است که مقادیر کوچک نزدیک به صفر برای p_o انتخاب شوند. به ویژه در بحث ما، انتخاب مقادیر سختگیرانه‌تر برای p_o باعث بزرگتر شدن فضای جفت‌های "نامعلوم" و کم اهمیت شدن نقش D'_o می‌شود. انتخاب ما برای این دو آستانه، به

[۲۸] Not statistically significant

طور پیش فرض $D'_o = ۰/۸$ و $p_o = ۰/۰۱$ است. شاخصی که به ازای این مقادیر آستانه‌ای تعریف می‌شود مبنای روش ما در این رساله، برای مطالعه‌ی ساختار بلوکی در هاپلوتیپ‌ها خواهد بود.

بررسی همبستگی در جفت اسنیپ‌های ناحیه‌ای شامل l اسنیپ با ارزیابی نمونه‌ای شامل n هاپلوتیپ نیازمند اجرای $O(nl^۲)$ عمل اصلی است؛ $O(n)$ عمل اصلی برای محاسبه‌ی pـ مقدار و برآورد $\tilde{D}'$ به ازای هر جفت از اسنیپ‌ها و $O(l^۲)$ تعداد جفت اسنیپ‌ها. در عمل، ما روشی را پیاده‌سازی کرده‌ایم که نتیجه از طریق آن، می‌تواند با محاسبات کمتری بدست آید. برای این کار، جدولی را شامل pـ مقدارها و مقادیر $\tilde{D}'$ به ازای سه‌تائی‌های (n_a, n_b, n_{11})، که در آن $n_a = ۱, \ldots, \lfloor n/۲ \rfloor$ ، $n_b = ۱, \ldots, n_a$ و $n_{11} = ۱, \ldots, n_b$ است، تشکیل می‌دهیم. با در نظر گرفتن تقارن در تعریف pـ مقدار و D' و استفاده از مقادیر از پیش محاسبه شده در جدول، ارزیابی شاخص همبستگی برای تمام جفت اسنیپ‌های ناحیه‌ی مورد مطالعه، با اجرای $O(n^۳ + l^۲)$ عمل اصلی انجام می‌شود. این شیوه در نمونه‌هایی که تعداد اسنیپ‌ها در آن، بیشتر از تعداد هاپلوتیپ‌ها است سریعتر از رویکرد اول است. به عنوان مثال، در نمونه‌ای شامل ۲۰۰ هاپلوتیپ بر روی ۵۰۰ اسنیپ، این شیوه چهار برابر سریعتر از رویکرد اول است.

فرض کنید w بر حسب تعداد اسنیپ، طول بزرگترین فاصله‌ای باشد که هیچ همبستگی معناداری بین دو اسنیپ، خارج از این فاصله، از نظر تجربی قابل تصور نباشد. با این فرض، تنها $w(l - w)$ جفت اسنیپ برای تعیین شاخص همبستگی، مورد ارزیابی قرار خواهند گرفت.

برآورد فراوانی هاپلوتیپ‌ها بر روی دو اسنیپ

در اینجا به طور خلاصه، روشی را شرح می‌دهیم که با استفاده از آن می‌توان فراوانی هاپلوتیپ‌های متفاوت بر روی دو اسنیپ را با داشتن نمونه‌های ژنوتیپ، که تعیین فاز نشده‌اند، برآورد کرد. محاسبه‌ی شاخص همبستگی در یک جفت اسنیپ، بنا بر تعریف آن، نیازمند محاسبه‌ی روابطی بر حسب فراوانی هاپلوتیپ‌ها در آن جفت اسنیپ است. فراوانی هاپلوتیپ‌ها در جفت اسنیپ‌ها را به طور مستقیم نمی‌توان در داده‌های ژنوتیپی تعیین فاز نشده بدست آورد، مگر آنکه پیش از آغاز محاسبات مربوط به بررسی همبستگی بین اسنیپ‌ها، برخی روش‌های تعیین فاز برای تفکیک داده‌های ژنوتیپی به هاپلوتیپ‌ها، بکار گرفته شوند.

استفاده از روشی که در ادامه معرفی می‌شود، این امکان را به ما می‌دهد که بدون نیاز به استفاده از روال‌های پیچیده برای تفکیک ژنوتیپ‌ها و استنباط هاپلوتیپ‌ها، بتوانیم شاخص همبستگی بین اسنیپ‌ها را با داشتن داده‌های تعیین فاز نشده‌ی ژنوتیپ، نیز بدست آوریم. این روش، یک نسخه‌ی ساده‌سازی شده از الگوریتم EM، برای استنباط فراوانی هاپلوتیپ‌ها بر روی دو اسنیپ است. فرض کنید، m_{ij} فراوانی ژنوتیپ ij در جفت اسنیپ مورد مطالعه است که در آن ij یکی از شش ترکیب زیر است.

$$\{(\mathbf{0},\mathbf{0}),(\mathbf{0},\mathbf{1}),(\mathbf{0},\mathbf{2}),(\mathbf{1},\mathbf{0}),(\mathbf{1},\mathbf{1}),(\mathbf{1},\mathbf{2}),(\mathbf{2},\mathbf{0}),(\mathbf{2},\mathbf{1}),(\mathbf{2},\mathbf{2})\}$$

برای برآورد فراوانی هاپلوتیپ‌های مورد بحث، یعنی n_{00}، n_{01}، n_{10} و n_{11}، ابتدا مقادیر دلخواهی برای آنها در نظر می‌گیریم و بر اساس روابط زیر، مقادیر جدیدی برای فراوانی هر یک از هاپلوتیپ‌ها بدست می‌آوریم.

$$\begin{aligned}
\tilde{n}_{00} &= ۲m_{00} + m_{01} + m_{10} + m_{11}n_{00}n_{11}/z \\
\tilde{n}_{01} &= ۲m_{02} + m_{01} + m_{12} + m_{11}n_{01}n_{10}/z \\
\tilde{n}_{10} &= ۲m_{20} + m_{21} + m_{10} + m_{11}n_{01}n_{10}/z \\
\tilde{n}_{11} &= ۲m_{22} + m_{21} + m_{12} + m_{11}n_{00}n_{11}/z
\end{aligned}$$

که در آن، $\tilde{n}_{ij}$ نشان‌دهنده‌ی فراوانی برآورد شده برای هاپلوتیپ ij در هر تکرار و $z = n_{00}n_{11}+n_{01}n_{10}$ است. پس از ارزیابی هر چهار رابطه‌ی فوق، مقادیر بدست آمده برای $\tilde{n}_{ij}$ را مجددا به جای فراوانی‌های اولیه قرار می‌دهیم و محاسبات را تا رسیدن به شرایط همگرائی تکرار می‌کنیم. پس از حصول شرایط همگرائی، مقادیر بدست آمده را به نزدیکترین عدد صحیح گرد می‌کنیم.

۴۰۲ روش GPMAP برای افراز بلوکی هاپلوتیپ‌ها

فرض کنید n نمونه از هاپلوتیپ‌های یک ناحیه‌ی کروموزومی، شامل l اسنیپ، از افراد غیرخویشاوند جمعیت در اختیار داریم. هر جفت از اسنیپ‌های واقع در این ناحیه را بر اساس شاخص معرفی شده در بخش ۳۰۲، در یکی از سه رده‌ی "همبسته" ، "مستقل" و "از نظر آماری غیرمعنادار" طبقه‌بندی می‌کنیم. جهت رسیدن به یک مدل عینی برای تعریف بلوک‌های هاپلوتیپ، جمعیتی را در نظر بگیرید که تنوع هاپلوتیپ‌ها در آن کم و بیش از تاثیر عوامل محیطی چون رانش ژنی، انتخاب طبیعی و مهاجرت مصون است. در چنین جمعیتی،

یک بلوک هاپلوتیپ در نهایت، ناحیه‌ای را بر روی ژنوم معین می‌کند که هر جفت از اسنیپ‌های واقع در آن در "همبستگی" با یکدیگر قرار دارند و بلعکس، اسنیپ‌های واقع در دو بلوک مختلف، مستقل از یکدیگرند. ایده‌ای مشابه این رویکرد، اخیرا توسط پاتارو و همکارانش[۱۵۶]، مورد توجه قرار گرفته است. در آنجا نیز، فرض بر این است که الگوی تغییرات LD، در بیرون از بلوک‌های هاپلوتیپ و در داخل آنها توسط دو تابع توزیع احتمال مختلف مدل می‌شوند و افراز بلوکی هاپلوتیپ‌ها از طریق محاسبه‌ی بیشترین درستنمائی این مدل بدست می‌آید.

شرایطی را که ما به عنوان فرضیات ایده‌آل برای مطالعه‌ی هاپلوتیپ‌ها در جمعیت، تحت مدل پیشنهادی بالا نیاز داریم در واقعیت به ندرت محقق می‌شوند. علاوه بر آن، وجود خطا در برآورد میزان همبستگی بین اسنیپ‌ها که در نمونه‌های کوچک، معضلی دور از انتظار نیست، می‌تواند سبب آن شود که در نواحی LD بالا، برخی جفت اسنیپ‌ها به اشتباه به عنوان جفت اسنیپ‌های مستقل تشخیص داده شوند. با توجه به این ملاحظات، در مسئله‌ای که ما به عنوان تعریف بلوک‌های هاپلوتیپ در نظر می‌گیریم، در جستجوی افرازی هستیم که در آن، بلوکها بیشترین تعداد ممکن از جفت اسنیپ‌های همبسته را در بر می‌گیرند در حالیکه تعداد جفت اسنیپ‌های مستقل درون بلوکها، تا سقف معینی محدود نگه داشته می‌شود.

همانند دیگر مسائل چند_هدفه[۲۹] در بهینه‌سازی، در اینجا نیز، بین بدست آوردن بلوک‌هایی با بیشترین تعداد از جفت اسنیپ‌های همبسته و بدست آوردن بلوک‌هایی با کمترین تعداد جفت اسنیپ‌های مستقل، یک "موازنه‌ی[۳۰] هزینه_فایده‌ای" برقرار است؛ برای حالت اول، بلوک‌هایی به بزرگی طول کل کروموزم مورد مطالعه و برای حالت دوم، بلوک‌های تک اسنیپی، جواب‌های بهینه‌اند. برای رفع این مشکل، مسئله‌ی بهینه‌سازی مقیدی را بکار می‌گیریم که در آن، هر دو هدف مطلوب، به نوعی مورد توجه قرار گرفته‌اند. این مسئله به وسیله‌ی روابط زیر، صورت‌بندی می‌شود؛

$$\max_{S} \sum_{i=1}^{k} B[s_{i-1}, s_i]$$

$$s.t. \qquad \sum_{i=1}^{k} A[s_{i-1}, s_i] \;<\; \alpha\, N_{ind} \qquad (۶۰۲)$$

[۲۹] multi-objective
[۳۰] trade-off

که در آن، $A[a,b]$ و $B[a,b]$ به ترتیب، نشان‌دهنده‌ی تعداد جفت اسنیپ‌های "مستقل" و تعداد جفت اسنیپ‌های "همبسته" ، در ناحیه‌ی محدود بین اسنیپ‌های$a+1$ و b هستند. بیشینه‌سازی بر روی تمام افرازهای $\mathcal{S} = \langle s_o, \ldots, s_k \rangle$ صورت می‌گیرد که در آن $s_o = s_0 < s_1 < \cdots < s_k = l$ است. در واقع، اگر اسنیپ‌ها را از سمت چپ به راست به ترتیب شماره‌گذاری کنیم، s_i شماره‌ی اسنیپی را نشان می‌دهد که مرز سمت راست بلوک iام را در افراز $\mathcal{S}$ معین می‌کند. تعداد بلوک‌ها، k، و موقعیت اسنیپ‌های نشان‌دهنده‌ی مرز بلوک‌ها، به جز اسنیپ‌های اول و آخر، متغیرهای آزاد این مسئله‌ی بهینه‌سازی مقید هستند. N_{ind} در رابطهٔ (۶۰۲)، تعداد جفت اسنیپ‌های "مستقل" در کل ناحیه‌ی مورد مطالعه است و α یک ثابت حقیقی دلخواه بین صفر و یک است که به وسیله‌ی آن، مجموع جفت اسنیپ‌های "مستقل" در کل بلوک‌ها را محدود می‌کنیم.

برای حل مسئله‌ی ۶۰۲، آن را با استفاده از ضرایب لاگرانژ، به یک مسئله‌ی بهینه‌سازی نامقید تبدیل می‌کنیم و داریم،

$$\max_{\mathcal{S}} \sum_{i=1}^{k} B[s_{i-1}, s_i] - \lambda\, A[s_{i-1}, s_i]\,. \qquad (۷۰۲)$$

در اینجا، λ یک پارامتر حقیقی مثبت و مجهول، مرتبط با α است. با ثابت گرفتن هر مقدار داده شده برای λ، افراز بهینه برای مسئله‌ی ۷۰۲ را می‌توان، با رهیافتی مبتنی بر شیوه‌ی برنامه‌ریزی پویا، به صورت زیر بدست آورد.

$$\begin{aligned} S^{opt}(0) &= 0\,, \\ S^{opt}(i) &= \max_{1 \leqslant d \leqslant \min(w,i)} \{S^{opt}(i-d) + S(i;d)\}\,, \quad for\ i = 1, \ldots, l\,. \end{aligned} \qquad (۸۰۲)$$

در اینجا، $S(i;d) = B[i-d,i] - \lambda A[i-d,i]$ نشان‌دهنده‌ی امتیازی است که به بازه‌ی ژنومی منتهی به اسنیپ iام و مشتمل بر d اسنیپ داده می‌شود و $S^{opt}(i)$، امتیاز افراز بلوکی بهینه برای اولین i اسنیپ سمت چپ، داخل ناحیه‌ی مورد مطالعه است.

امتیاز بهترین افراز بلوکی برای کل ناحیه‌ی مورد مطالعه، یعنی $S^{opt}(l)$، با استفاده از روابط بازگشتی فوق بدست می‌آید. این کار با محاسبه‌ی مقادیر بهینه‌ی $S^{opt}(i)$ به ترتیب به ازای $i = ۱, \ldots, l$ و ذخیره‌ی آنها در یک جدول، موسوم به جدول برنامه‌ریزی پویا، صورت می‌گیرد. پس از کامل کردن جدول و بدست آوردن امتیاز بهترین افراز بلوکی، شماره‌ی اسنیپ‌هایی که مرز بلوک‌های افراز بهینه‌ی مورد نظر ما را تعیین می‌کنند با دنبال کردن روابط بازگشتی زیر و به کمک مقادیر ذخیره شده در جدول برنامه‌ریزی پویا، بدست می‌آیند.

$$\begin{aligned} s_k^* &= l\,, \\ s_{j-۱}^* &= s_j^* - \underset{۱ \leqslant d \leqslant \min(w, s_j^*)}{argmax} \{S^{opt}(s_j^* - d) + S(s_j^*; d)\}\,, \quad for\ \ j = k, \ldots, ۱\,. \end{aligned} \tag{۹۰۲}$$

شماره‌ی اسنیپ‌هایی را که توسط رابطهٔ (۹۰۲) بدست می‌آیند، به ترتیب در یک پشته[۳۱] وارد می‌کنیم تا زمانی که یک اسنیپ با شماره‌ی صفر بدست آید. طول این پشته، تعداد بلوک‌های افراز بهینه را تعیین می‌کند و محتوای آن، افراز بلوکی بهینه برای مسئله‌ی ۷۰۲ است که آنرا با $\mathcal{S}^*(\lambda)$ نشان می‌دهیم.

مقدار w در رابطهٔ (۸۰۲)، بر حسب تعداد اسنیپ، نشان‌دهنده‌ی حداکثر طول دلخواه برای هر یک از بلوک‌های افراز بهینه است. به طور معمول، محدودیتی برای طول بلوک‌ها در نظر نمی‌گیریم و w را برابر با تعداد کل اسنیپ‌های موجود در ناحیه‌ی مورد مطالعه، l، قرار می‌دهیم. مطالعات تجربی بر روی ژنوم، مؤید آن است که مقدار LD بین جایگاه‌هایی که فاصله‌یشان از حد معینی بیشتر است بسیار به صفر نزدیک است. بنابراین منطقی‌تر آن است که مقداری متناسب با این فاصله، برای w در نظر بگیریم. مثلا اگر متوسط فاصله‌ی اسنیپ‌ها را یکهزار باز در نظر بگیریم و فرض کنیم هیچ همبستگی معناداری نمی‌تواند بین دو اسنیپ که در فاصله‌ی ۵۰۰ کیلوباز یا بیشتر از یکدیگر واقع شده‌اند وجود داشته باشد آنگاه منطقی خواهد بود اگر قرار دهیم $w = ۵۰۰$. بر این اساس، محاسبه‌ی ماتریس‌های A و B و نیز هر اجرای کامل روال برنامه‌ریزی پویا (روابط ۸۰۲ و ۹۰۲)، نیازمند اجرای $O(w.l)$ عمل اصلی است.

تا اینجا، شیوه‌ی حل مسئله‌ی بهینه‌سازی تعریف شده در رابطهٔ (۷۰۲)، با فرض ثابت بودن λ، توسط یک روال برنامه‌ریزی پویای تشریح گردید. برای حل مسئله‌ی اصلی مورد بحث در این بخش، لازم است مقدار

[۳۱] stack

ضریب لاگرانژ به قسمی تعیین شود که قید مسئله‌ی (۶۰۲) به ازای مقدار داده شده‌ی α برقرار گردد.

به یاد داشته باشید که پس از بدست آوردن افراز بهینه به ازای یک مقدار داده شده برای λ، توسط روال برنامه‌ریزی پویا و جایگذاری جواب بدست آمده در قید مسئله‌ی (۶۰۲) با دو حالت ممکن است روبرو شویم؛ یا نابرابری به طور اکید در جهت قید مورد نظر برقرار است که در این صورت جواب بدست آمده از مسئله‌ی نامقید، یک کران پائین برای مقدار بهینه‌ی مسئله‌ی مقید خواهد بود و از این رو با انتخاب مقادیر کوچکتری برای λ در همسایگی مقدار فعلی آن، مسئله‌ی نامقید جدیدی تعریف می‌شود که جواب بهینه‌ی آن ضمن حفظ قید مسئله‌ی اصلی، مقدار هدف مسئله را نیز افزایش می‌دهد، یا نابرابری به طور اکید در خلاف جهت قید مطلوب برقرار است که در این حالت، لازم است مقدار λ را برای مسئله‌ی نامقید تا جائی افزایش دهیم که جواب بهینه‌ی بدست آمده، قید مسئله‌ی اصلی را ارضا کند.

به بیان دیگر، متناظر با هر مقدار داده شده برای λ، با جایگذاری افراز بدست آمده از روال برنامه‌ریزی پویا در قید مسئله‌ی (۶۰۲)، یک مقدار عملی برای α، مثل $\hat{\alpha}$ بدست می‌آید که به ازای آن، قید مطلوب برقرار است. افزایش λ به طور پیوسته، $\hat{\alpha}$ را کاهش می‌دهد و بلعکس کاهش λ به طور پیوسته، مقدار $\hat{\alpha}$ را افزایش می‌دهد. بر این اساس، می‌توان مقدار ایده‌آل برای ضریب لاگرانژ را، متناظر با مقدار داده شده برای پارامتر α توسط یک روال جستجوی دودوئی بدست آورد. این روال در دو مرحله اجرا می‌شود. ابتدا با شروع از یک مقدار اولیه‌ی دلخواه، مثلا $\lambda_o = ۱$، و افزایش مقدار آن با گام‌های بزرگ، روال برنامه‌ریزی پویا را مکرراً اجرا می‌کنیم تا زمانی که در این دنباله، دو مقدار $\lambda_۱$ و $\lambda_۲$ بدست آوریم به قسمی که نابرابری قید مسئله‌ی اصلی به ازای $\mathcal{S}^*(\lambda_۱)$ در جهت بزرگتر و به ازای $\mathcal{S}^*(\lambda_۲)$ در جهت کوچکتر برقرار باشد. در مرحله‌ی دوم، که در واقع همان الگوریتم جستجوی دودوئی است، در هر تکرار، ابتدا روال برنامه‌ریزی پویا به ازای یک مقدار میانی در بازه‌ی $[\lambda_۱, \lambda_۲]$ اجرا می‌شود، سپس بر اساس وضعیتی که از صدق دادن جواب، در قید مسئله‌ی اصلی بدست می‌آید، بازه‌ی تغییرات λ را کوچکتر می‌کنیم تا در نهایت مقدار مورد جستجو، با دقت مطلوب بدست آید. در عمل، این روال جستجو، مقدار مطلوب برای پارامتر λ متناظر با $\alpha = ۰/۰۱$ را به طور متوسط با ۱۰ تکرار بدست می‌آورد.

پیاده‌سازی روش GPMAP و الگوریتمی دیگر مبتنی بر شاخص گابریل

در واقع، رویکرد ما می‌تواند برای توسعه‌ی هر شیوه‌ای که در آن بلوک‌های هاپلوتیپ با آنالیز LD بین جفت اسنیپ‌ها تعریف می‌شوند، به کار رود. به طور مثال، می‌توان از شاخص گابریل به عنوان معیار تشخیص همبستگی بین اسنیپ‌ها، همراه با الگوی بهینه‌سازیی که در بالا تشریح شد برای تعیین یک افراز سراسری استفاده کرد.

روش شرح داده شده در این بخش را، تحت عنوان کلی "افراز سراسری هاپلوتیپ‌ها برای بیشترین جفت اسنیپ‌های همبسته"، یا به اختصار GPMAP (Globlal Block Partitioning for Maximal Associated SNP Pairs) می‌نامیم. ما، دو گونه‌ی متفاوت از این الگوریتم را تحت زبان جاوا پیاده‌سازی کردیم؛ یکی مبتنی بر شاخص همبستگی معرفی شده در بخش ۳۰۲ که در آن، از آزمون دقیق فیشر برای بررسی سطح معناداری همبستگی بین اسنیپ‌ها استفاده می‌شود و آنرا به اختصار GPMAPF می‌نامیم و دیگری، مبتنی بر شاخص همبستگی منسوب به گابریل که آنرا به اختصار GPMAPG می‌نامیم. ما این دو روش را در قالب دو گزینه‌ی جدید به مجموعه‌ی روش‌های افراز بلوکی هاپلوتیپ‌ها، در نرم‌افزار کد آزاد Haploview, ver. 4.1، اضافه کردیم. نرم‌افزار حاصل، از طریق وب سایت `http://bioinf.cs.ipm.ac.ir/gpmap` در دسترس عموم است[۱۵۷].

۵۰۲ مقایسه‌ی الگوریتم‌های افراز بلوکی هاپلوتیپ‌ها

مقایسه‌ای جامع بین الگوریتم‌های مختلف افراز بلوکی هاپلوتیپ‌ها، نیازمند ارزیابی جنبه‌های گوناگونی از بلوک‌های هاپلوتیپی بدست آمده و بررسی کارآیی آنها در زمینه‌های کاربردی دیگر است. در این بخش، در کنار برخی طرح‌های رایج برای مقایسه‌ی روش‌های مختلف افراز بلوکی، چند طرح مقایسه‌ایی جدید نیز معرفی می‌شود که از طریق آنها، ثبات روش‌های مختلف در بازتولید بلوک‌های هاپلوتیپی یکسان به ازای نمونه‌های نوترکیب، مطابقت بین نقاط پراحتمال نوترکیبی و مرز بلوک‌ها و کارآیی ساختار بلوکی بدست آمده، برای شناسائی جایگاه ژنی مرتبط با یک بیماری، مورد مطالعه قرار می‌گیرد.

از بین روش‌های متعددی که توسط دیگر پژوهشگران برای تعریف بلوک‌های هاپلوتیپ معرفی شده‌اند، شش روش متداول را برای ارزیابی و مقایسه با روش افراز بلوکی پیشنهادی در این رساله انتخاب می‌کنیم. روش‌های گوناگون افراز بلوکی را می‌توان از حیث ساختار افزار و معیاری که در آنها، برای تعریف بلوک‌ها مورد توجه قرار می‌گیرد، در گروه‌هایی از روش‌های مشابه یکدیگر، دسته‌بندی کرد. هر یک از شش روش انتخاب شده برای ارزیابی در این رساله، نماینده‌ی یکی از این گروه‌ها است. جدول ۱۰۲، خلاصه‌ای از مشخصات اصلی آنها را نمایش می‌دهد.

توجه به اینکه نکته ضروری است که روش HOT، مورد متفاوتی در بین دیگر روش‌های جدول ۱۰۲ است. این روش در واقع، شیوه‌ای برای استنباط "نقاط پراحتمال" نوترکیبی بر روی ژنوم است. نتایج بدست آمده از اجرای این روش بر روی هاپلوتیپ‌های HapMap، در کنار دیگر اطلاعات پایگاه داده‌ای HapMap و از طریق اینترنت قابل دسترس برای عموم است. این نتایج، نرخ تغییرات LD را در امتداد ژنوم انسان بر حسب Mb/cM و نیز موقعیت و عرض نواحی "پراحتمال" نوترکیبی را در سراسر ژنوم تعیین می‌کند. از آنجا که عرض این نواحی در مقیاس ژنومی و در مقایسه با فواصل پراکندگی اسنیپ‌ها، کم است (به طور متوسط ۲ کیلو باز)، این نواحی را "نقاط پر احتمال" نوترکیبی می‌نامند. این نقاط، مکان‌های مناسبی هستند تا به عنوان مرز بلوک‌ها در نظر گرفته شوند و از این رو ناحیه‌ی محدود بین دو نقطه‌ی پراحتمال متوالی را به عنوان یک بلوک، تحت عنوان روش HOT در نظر می‌گیریم. مقایسه‌ی این روش، در کنار دیگر روش‌های افراز بلوکی هاپلوتیپ‌ها، به ما کمک می‌کند تا ایده‌ایی اولیه درباره‌ی امکان استفاده از نتایج بدست آمده از روش‌های افراز بلوکی هاپلوتیپ‌ها برای شناسائی نقاط پراحتمال نوترکیبی، بدست آوریم.

جدول ۲.۱: روش‌های افراز بلوکی هاپلوتیپ‌های مورد ارزیابی در این رساله

اختصار	روش	ساختار افراز	معیار تعیین‌کننده‌ی بلوک	قید تعریف	نرم‌افزار	مرجع
HOT	Hotspot	موضعی*	نقاط پراحتمال نوترکیبی	–	"نتایج از پیش محاسبه شده، قابل دسترس در HapMap"	[۱۲۲]
MB	کمترین تعداد بلوک	سراسری	کمینه‌سازی تعداد بلوک‌ها	واگرائی هاپلوتیپ‌ها	HapBlock v.3	[۱۰۸]
HB	HapBlock	سراسری	کمینه‌سازی مجموع تگ‌اسنیپ‌ها	واگرائی هاپلوتیپ‌ها	HapBlock v.3	[۱۰۸]
MDL	MDBlock	سراسری	minimum description length	واگرائی هاپلوتیپ‌ها	MDBlock v.1	[۱۰۹]
GAM	آزمون چهار گامتی	موضعی	نشانه‌ی نوترکیبی	گامت چهارم	Haploview v.4	[۱۱۶]
GAB	روش گابریل	موضعی	نشانه‌ی نوترکیبی	اسنیپ‌های "قویاً همبسته"	Haploview v.4	[۱۱۵]
GPG	GPMAPG	سراسری	بیشینه‌سازی جفت اسنیپ‌های "همبسته" بر اساس شاخص گابریل	اسنیپ‌های "مستقل"	Haploview + GPMAP	بخش ۴۰۲
GPF	GPMAPF	سراسری	بیشینه‌سازی جفت اسنیپ‌های "همبسته" بر اساس آزمون دقیق فیشر	اسنیپ‌های "مستقل"	Haploview + GPMAP	بخش ۴۰۲

* رویکردی که در الگوریتم اصلی این شیوه بکار رفته است، مبتنی بر آنالیز موضعی تغییرات نرخ نوترکیبی در امتداد ژنوم است.

روش دیگری که به اختصار با نام MB در جدول ۱۰۲ نشان داده شده است، به طور صریح در هیچ مرجع مشخصی معرفی نشده است و در واقع، گزینه‌ی خاصی از مجموعه‌ی روش‌های پیاده‌سازی شده در نرم‌افزار HapBlock است. با انتخاب این گزینه در HapBlock، دقیقا یک تگ‌اسنیپ برای هر بلوک در نظر گرفته می‌شود. بدین ترتیب، کمترین تعداد بلوک‌ها برای افراز سراسری هاپلوتیپ‌های داده شده با قید محدود نگه داشتن واگرائی هاپلوتیپ‌ها در بلوک‌ها بدست می‌آید.

۱۰۵۰۲ نمونه‌گیری از داده‌های HapMap

اولین موضوعی را که در مقایسه‌ی روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها، مورد بررسی قرار می‌دهیم، ویژگی‌های کلی بلوک‌های بدست آمده از اجرای این روشها بر روی داده‌های واقعی است. برای این کار، هاپلوتیپ‌های پانل CEU از پایگاه داده‌ای HapMap, release 22 را در ده ناحیه‌ی ENCODE انتخاب کردیم. این ده ناحیه، توسط پروژه‌ی Encyclopedia of DNA Elements به عنوان موضوع فاز اول پروژه، برای مطالعه در زمینه‌ی اجزاء عمل‌کننده‌ی[۳۲] ژنوم انسان انتخاب شدند[۱۵۸]. طی پروژه‌ی HapMap، ژنوتیپ‌های این ده ناحیه به طور خاص، با چگالی اسنیپی بالاتری تعیین شدند. جدول ۲۰۲، برگرفته از وب سایت HapMap، خلاصه‌ای از اطلاعات این ده ناحیه و تعداد اسنیپ‌های تشخیص داده شده در پانل CEU را نشان می‌دهد.

حدود ۲۰۰۰ اسنیپ در هر یک از نواحی ENCODE وجود دارند. با این حال ما تنها، زیرمجموعه‌ای از آنها که بین هر سه پانل CEU، YRI و JPT+CHB مشترکند را مورد بررسی قرار می‌دهیم. علاوه بر آن، در هر ناحیه، ژنوتیپ‌های مورد مطالعه‌یمان را به۴۰۰ اسنیپ با بیشترین هتروزیگوسیتی محدود می‌کنیم. برای این کار، ناحیه مورد نظر را به ۲۰ بازه‌ی مساوی تقسیم می‌کنیم و از هر بازه، ۲۰ اسنیپ با بیشترین هتروزیگوسیتی را از بین اسنیپ‌های مشترک بین هر سه پانل نمونه، انتخاب می‌کنیم. بدین ترتیب، یک توزیع نسبتا یکنواخت از اسنیپ‌هایی حاوی بیشترین "اطلاعات" بدست می‌آید. این کاهش اولیه‌ی داده‌ها ضروری است چون، اجرای بسیاری از روش‌های افراز بلوکی هاپلوتیپ‌ها بر روی حجم بالایی از اسنیپ‌ها، در مدت

[۳۲] functional elements

جدول ۲۰۲: اطلاعات کلی ژنوتیپ‌های HapMap در نواحی ENCODE

Region name	Chromosome band	Genomic interval (NCBI B36)	Genotyped SNPs*
ENr112	2p16.3	Chr2:51512208..52012208	2,601
ENr131	2q37.1	Chr2:234156563..234656627	2,214
ENr113	4q26	Chr4:118466103..118966103	2,538
ENm010	7p15.2	Chr7:26924045..27424045	1,830
ENm013	7q21.13	Chr7:89621624..90121624	1,770
ENm014	7q31.33	Chr7:126368183..126865324	3,343
ENr321	8q24.11	Chr8:118882220..119382220	2,128
ENr232	9q34.11	Chr9:130725122..131225122	1,909
ENr123	12q12	Chr12:38626477..39126476	2,189
ENr213	18q12.1	Chr18:23719231..24219231	1,990

برگرفته از `http://hapmap.org/downloads/encode1.html` در ۲۰ ژوئن ۲۰۰۸

* اسنیپ‌های تشخیص داده شده در پانل CEU

زمانی معقول به نتیجه نمی‌رسد. در ادامه، معیارهای مختلفی برای ارزیابی نتایج حاصل از اجرای روش‌های مندرج در جدول ۱۰۲ بر روی این نمونه‌ها، معرفی می‌شوند.

۲۰۵۰۲ تنوع هاپلوتیپ‌ها

اولین موضوعی را که در مقایسه‌ی روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها مورد توجه قرار می‌دهیم تنوع هاپلوتیپ‌ها در بلوک‌های هاپلوتیپی بدست آمده از اجرای روش‌های مختلف بر روی ژنوتیپ‌های نواحی ENCODE است. برای ارزیابی میزان واگرائی هاپلوتیپ‌ها در هر بلوک، ما از روشی برای خوشه‌بندی[۳۳] هاپلوتیپ‌ها و تعریف "هاپلوتیپ‌های رایج" استفاده می‌کنیم. این روش، تعمیمی ساده از رویکرد پَتیل و همکارانش[۹۷]، برای تعریف "هاپلوتیپ‌های رایج" است. در این روش، نمونه هاپلوتیپ‌های "مشابه یکدیگر" در هر بلوک را در یک خوشه قرار می‌دهیم. در اینجا، ما دو هاپلوتیپ را "مشابه یکدیگر" در نظر می‌گیریم اگر حداکثر در چهار درصد از اسنیپ‌ها، تفاوت داشته باشند. در نظر گرفتن این تفاوت کوچک اما قابل تحمل در تعریف هاپلوتیپ‌های مشابه، اثر اختلالات تصادفی موجود در خوانش اسنیپ‌ها را در برآورد واگرائی هاپلوتیپ‌ها در بلوک‌های بلند، کاهش می‌دهد. خوشه‌های حاوی شش نمونه هاپلوتیپ یا بیشتر را به عنوان

[۳۳] clustering

نشانه‌ایی از چندریختی‌های[۳۴] رایج در جمعیت، خوشه‌های "معنادار" می‌نامیم[۳۵]. نسبت نمونه هاپلوتیپ‌های متعلق به خوشه‌های "معنادار" به کل نمونه را «پوشش هاپلوتیپ‌های رایج» می‌نامیم. این کمیت را به عنوان یک معیار پایه‌ای برای مقایسه‌ی روش‌های افراز بلوکی هاپلوتیپ‌ها مورد توجه قرار می‌دهیم(بخش ۳۰۳).

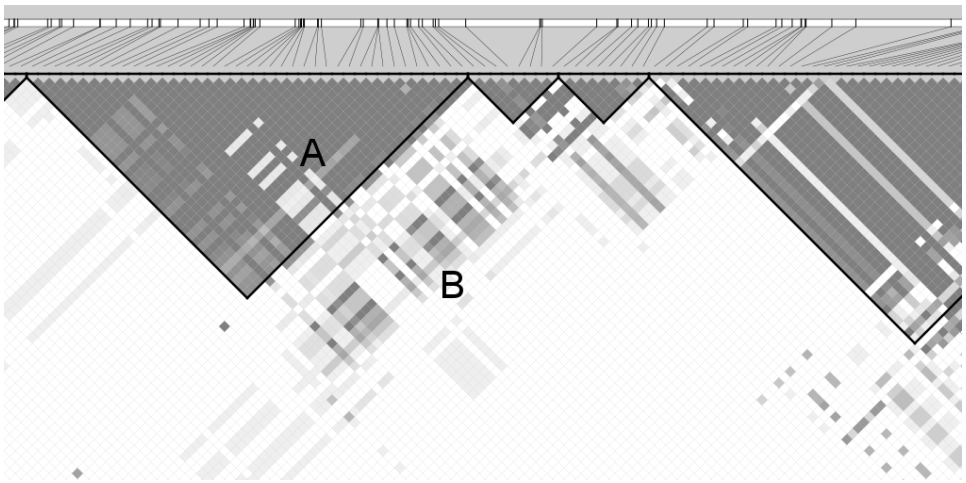
سازگاری بین ساختار بلوکی بدست آمده از یک روش افراز بلوکی و الگوی تغییرات LD در امتداد ناحیه‌ی مورد مطالعه بر روی ژنوم، می‌تواند موضوعی جالب توجه برای مقایسه بین روش‌های مختلف باشد. به طور شهودی، یک "حفره" در یک بلوک، جائی است که در آن یک اسنیپ بدون هیچ همبستگی معناداری با دیگر اسنیپ‌های همان بلوک واقع شده است و به طور مشابه، هر موقعیتی را که در آن، اسنیپی از یک بلوک با تعدادی از اسنیپ‌های بلوکی دیگر در همبستگی باشد، "جزیره" نامیده می‌شود. به طور دقیق، ما موقعیت یک اسنیپ را به عنوان یک حفره در یک بلوک تلقی می‌کنیم اگر میانگین $|D'|$ بین آن و دیگر اسنیپ‌های همان بلوک کمتر از ۰/۸ باشد و آنرا یک جزیره به حساب می‌آوریم اگر میانگین $|D'|$ بین آن و اسنیپ‌های دیگر بلوک‌ها بیشتر از ۰/۸ باشد(شکل ۹۰۲). ما فراوانی موقعیت‌های "حفره" و "جزیره" را به ازای بلوک‌های هاپلوتیپی بدست آمده از هر یک از روش‌های مندرج در جدول ۱۰۲، به عنوان معیاری دیگر برای مقایسه، مورد بررسی قرار می‌دهیم(بخش ۳۰۳).

۳۰۵۰۲ محاسبه‌ی تگ‌اسنیپ‌ها

همانطور که در مقدمه اشاره شد، یکی از کاربردهای افراز بلوکی هاپلوتیپ‌ها، افراز ژنوم به بلوک‌هایی است که تنوع هاپلوتیپ‌ها درون این بلوک‌ها را به طور کامل بتوان با در اختیار داشتن داده‌های تعداد معدودی از اسنیپ‌ها بازسازی کرد. این اسنیپ‌ها را اصطلاحا با عنوان «هاپلوتیپ‌ـ‌تگ‌اسنیپ‌ها» می‌شناسیم و به اختصار، htSNP می‌گوئیم. تعیین زیرمجموعه‌ای با کمترین تعداد ممکن از اسنیپ‌ها در یک بلوک که بازسازی کامل هاپلوتیپ‌های بلوک، با استفاده از داده‌های تحدید شده به این اسنیپ‌ها امکانپذیر باشد، تحت عنوان «مسئله‌ی انتخاب تگ‌اسنیپ‌ها» شناخته می‌شود. بدیهی است شکل افراز بلوکی بر مجموع تگ‌اسنیپ‌های مورد نیاز در یک ناحیه‌ی ژنومی تاثیرگذار است. برای بررسی این موضوع در بین بلوک‌های هاپلوتیپی بدست

[۳۴] polymorphism

[۳۵] حجم نمونه‌ی مورد بررسی در اینجا، ۱۲۰ هاپلوتیپ است. بنابراین یک خوشه‌ی "معنادار"، دست کم پنج درصد از کل نمونه را در برمی‌گیرد.



شکل ۹۰۲: نقشه‌ی LD بین جفت اسنیپ‌ها در قسمتی از ناحیه‌ی 4q26 (ENr113)
در این شکل که به طور عمومی نقشه‌ی LD نامیده می‌شود، همانند جدول مسافت بین شهرها، هر خانه‌ی مربعی شکل کوچک، اندازه‌ی $|D'|$ را بین جفت اسنیپ متناظر با ردیف‌های متقاطع در آن خانه توسط رنگ خاصی نشان می‌دهد (سیاه: $|D'| = ۱$ و سفید: $|D'| = ۰$). نقاط LD پائین، درون بلوک‌ها را "حفره" می‌نامیم (مثل A) و نقاط LD بالا، بیرون بلوک‌ها را "جزیره" می‌نامیم (مثل B).

آمده از اجرای روش‌های مندرج در جدول ۱۰۲ بر روی هاپلوتیپ‌های نواحی ENCODE، ما از نرم‌افزار htSNPer[۱۳۴] برای تعیین کمترین تعداد تگ‌اسنیپ‌های لازم برای بازسازی تنوع هاپلوتیپی در این نواحی، استفاده می‌کنیم.

ما از تنظیمات پیش‌فرض در htSNPer، برای اجرای آن بر روی نمونه هاپلوتیپ‌های درون هر بلوک استفاده کردیم. بر اساس این تنظیمات، htSNPer تگ‌اسنیپ‌ها را به گونه‌ای انتخاب می‌کند که تنوع هاپلوتیپ‌ها در بین دست کم ٪۸۰ از "هاپلوتیپ‌های رایج"، به طور یکتا توسط تگ‌اسنیپ‌های انتخاب شده برای این بلوک، قابل بازسازی است. به طور پیش فرض، htSNPer هاپلوتیپ‌هایی را در زمره‌ی "هاپلوتیپ‌های" رایج به شمار می‌آورد که فراوانی آنها در نمونه‌ی مورد بررسی بیش از۰/۰۵ باشد. با این حال، این نرم‌افزار به واسطه‌ی یکی از گزینه‌های پیش‌فرض خود، پیش از اجرای روال مربوط به انتخاب تگ‌اسنیپ‌ها، نمونه هاپلوتیپ‌های داده شده را با استفاده از یک روش خاص، به بلوک‌های هاپلوتیپ افراز می‌کند. از آنجا که ما این نرم‌افزار را تنها به منظور تعیین تگ‌اسنیپ‌ها در بلوک‌های بدست آمده از روش‌های دیگر به طور مستقل مورد توجه قرار داده‌ایم و در واقع، نمی‌خواهیم در بلوک‌های مورد مطالعه‌یمان تغییری پدید آورد، بخشی از پارامترهای پیش‌فرض برای اجرای htSNPer را تغییر دادیم به قسمی که این برنامه از

اجرای روال افراز بلوکی صرفنظر می‌کند. دستور زیر، تنظیمات انتخابی ما برای فراخوانی برنامه‌ی htSNPer را نشان می‌دهد.

```
htSNPer -T2 -F1 -W10 -S1 -D2 -H1 -C0.05 -L0 -M0.01 -A0.8 -B0.8
```

اجرای htSNPer با این تنظیمات، به ما این امکان را می‌دهد که بتوانیم مجموع تگ‌اسنیپ‌های مورد نیاز برای پوشش تنوع هاپلوتیپی موجود در ناحیه‌ی مورد مطالعه را به طور مستقل برای هر یک از افرازهای تعیین شده توسط روش‌های مختلف بدست آوریم. به یاد داشته باشید که htSNPer با حل یک مسئله‌ی کمینه‌سازی به طور محاسباتی تضمین می‌دهد که تعداد تگ‌اسنیپ‌های گزارش شده، به ازای نمونه هاپلوتیپ‌های داده شده و برای ساختار بلوکی داده شده کمترین است.

اندازه‌گیری پوشش htSNPها[۳۶]

به طور شهودی، منظور از پوشش یک htSNP یا مجموعه‌ای از htSNPها، وسعت ناحیه‌ای بر روی ژنوم مورد مطالعه است که تنوع هاپلوتیپ‌ها در محدوده‌ی آن، توسط htSNP یا htSNPهای مورد بحث قابل بازسازی است. برای ارزیابی پوشش htSNPها در بین روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها، بر اساس تگ‌اسنیپ‌های بدست آمده از اجرای htSNPer برای هر یک از افرازهای بلوکی مندرج در جدول ۱۰۲، "پوشش kـ تگ‌اسنیپ‌ها" را به شرحی که در ادامه می‌آید محاسبه می‌کنیم. ابتدا، بازه‌هایی را بر روی ژنوم تعیین می‌کنیم که تنوع هاپلوتیپ‌ها در آنها را می‌توان توسط یک htSNP پوشش داد. مجموع طول این بازه‌ها را پوشش ۱ ـ تگ‌اسنیپ‌ها می‌نامیم. همین روال را دوباره، پس از حذف بازه‌های پوشش داده شده در مراحل قبل، بر روی ژنوم باقی‌مانده تکرار می‌کنیم. در تکرار kام، مجموع طول‌های پوشش داده شده در این گام و طول‌های پوشش داده شده در گام‌های قبل را پوشش kـ تگ‌اسنیپ‌ها می‌نامیم. در بخش ۴۰۳، نتایج بدست آمده از محاسبه‌ی پوشش kـ تگ‌اسنیپ‌ها را در بین افرازهای بلوکی متفاوت، مورد بررسی قرار می‌دهیم.

[۳۶] htSNP coverage

۴۰۵۰۲ کمیتی برای اندازه‌گیری شباهت بین ساختارهای بلوکی

یک پرسش رایج درباره‌ی روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها، این است که بلوک‌های بدست آمده از روش‌های متفاوت تا چه اندازه شبیه به یکدیگرند. رویکرد متداول برای اندازه‌گیری شباهت بین دو ساختار بلوکی متفاوت، در نظر گرفتن مجموع فاصله‌ی بین مرزهای "متناظر"، در دو افراز بلوکی است. کمیتی که بدین ترتیب بدست می‌آید با اینکه شباهت بین دو افراز بلوکی را به شیوه‌ای کاملا ساده و صریح بیان می‌کند اما تعیین اینکه کدام بلوک‌ها را می‌توان در بین دو افراز، در تناظر با یکدیگر قرار داد، خود پرسش دیگری بوجود می‌آورد که ارائه‌ی پاسخی فراگیر به آن، بغرنج به نظر می‌رسد. پیچیدگی این مسئله از آنجا ناشی می‌شود که در بسیاری از موارد، روش‌های متفاوت، افرازهایی با تعداد بلوک‌های متفاوت تولید می‌کنند. به عنوان مثال، افرازی را در نظر بگیرید که بلوک‌های آن از تظریف[۳۷] بلوک‌های افرازی دیگر بدست آمده باشند. تعریفی که ما در اینجا برای اندازه‌گیری شباهت بین دو افراز بلوکی ارائه می‌دهیم کاملا خالی از ابهام است و بکارگیری آن با مشکلات مذکور روبرو نیست.

بر اساس مدل پیشنهادی در بخش ۴۰۲، واقع شدن دو اسنیپ در یک بلوک، تنها عامل تعیین‌کننده‌ی وجود همبستگی بین آنها است. به عبارت دیگر، هر ساختار بلوکی، مدلی متفاوت برای وجود یا عدم وجود همبستگی بین جفت اسنیپ‌ها پیشنهاد می‌دهد. با این مقدمه، ما یک زوج اسنیپ را "همبسته فرض شده" توسط یک افراز بلوکی می‌نامیم اگر بلوکی در این افراز وجود داشته باشد که هر دو اسنیپ این زوج را در برگیرد. مبنای ما برای اندازه‌گیری شباهت بین دو افراز بلوکی، شمارش تعداد زوج اسنیپ‌هایی است که توسط هر دو افراز "همبسته فرض شده" باشند. نسبت این زوج اسنیپ‌ها به زوج اسنیپ‌هایی که دست کم توسط یکی از دو افراز "همبسته فرض شده" باشند را به عنوان کمیت سنجش شباهت بین دو افراز بلوکی در نظر می‌گیریم. بر این اساس داریم:

$$similarity(\mathcal{S},\mathcal{T}) = \frac{\sum_{i=1}^{k_1} u_i^2 + \sum_{i=1}^{k_2} v_i^2}{l \; + \; \sum_{i=1}^{k_1} u_i^2 + \sum_{i=1}^{k_2} v_i^2 - \sum_{i=1}^{k_o} w_i^2} - 1 \qquad (۱۰۰۲)$$

که در آن $\mathcal{S} = \langle s_o, s_1, \ldots, s_{k_1} \rangle$ و $\mathcal{T} = \langle t_o, t_1, \ldots, t_{k_2} \rangle$، دو افراز بلوکی بر روی یک ناحیه‌ی ژنومی

[۳۷] Refinement

واحد هستند و l تعداد اسنیپ‌های موجود در این ناحیه است. $u_i = s_i - s_{i-1}$ و $v_i = t_i - t_{i-1}$ به ترتیب نشاندهنده‌ی طول بلوک iام در افراز $\mathcal{S}$ و $\mathcal{T}$ هستند. w_iها طول بلوک‌های حاصل از اجتماع دو افراز، $\mathcal{S} \cup \mathcal{T}$ هستند. مقدار بدست آمده از رابطهٔ (۱۰۰۲)، برابر است با احتمال اینکه یک زوج اسنیپ که توسط یکی از دو افراز "همبسته فرض شده" است، توسط افراز دیگر نیز "همبسته فرض شود".

۵۰۵۰۲ شیوه‌ای برای ارزیابی ثبات الگوریتم‌های افراز بلوکی

همانطور که در مقدمه اشاره شد، یک توضیح ساده درباره‌ی پیدایش ساختار بلوکی در هاپلوتیپ‌های ژنوم انسان، وقوع رویدادهای نوترکیبی در هاپلوتیپ‌های اجدادی، به طور برجسته در موقعیت نوکلئوتیدهای اطراف مرز بلوک‌ها، در مقایسه با نواحی داخل بلوک‌ها است. بر پایه‌ی این مدل، یک الگوریتم افراز بلوکی را "با ثبات"[۳۸] می‌نامیم اگر بلوک‌های هاپلوتیپی یکسانی، خواه با اجرا بر روی هاپلوتیپ‌های اجدادی، خواه با اجرا بر روی هاپلوتیپ‌هایی از نسل‌های اخیر بدست آورد. پایداری مرزها در روش‌های مختلف افراز بلوکی را می‌توان با بررسی اختلاف بین مرز بلوک‌های بدست آمده از داده‌های نسل اول و مرز بلوک‌های بدست آمده از داده‌های چندین نسل بعد، اندازه‌گیری کرد. برای این منظور، ۱۲۰ نمونه هاپلوتیپ HapMap در ناحیه‌ی 9q34.11 را طی شبیه‌سازی در ده نسل متوالی، با فرض وقوع رویداد نوترکیبی در مرز بلوک‌ها با احتمال ۰/۵ در هر نسل و ثابت بودن اندازه‌ی جمعیت، مورد مطالعه قرار می‌دهیم. این فرایند را ۵۰ بار برای هر روش افراز تکرار می‌کنیم و ساختار بلوک‌های بدست آمده از داده‌های هر نسل را برای برآورد میزان "ثبات" در الگوریتم‌های افراز ثبت می‌کنیم. نتایج بدست آمده در بخش ۶۰۳ مورد بررسی قرار می‌گیرند.

تا اینجا، چند طرح مقایسه‌ایی برای بررسی نتایج بدست آمده از الگوریتم‌های گوناگون افراز بلوکی هاپلوتیپ‌ها معرفی شدند. در این طرح‌ها، مقایسه‌ی روش‌های افراز بر پایه‌ی ارزیابی جنبه‌های مختلفی از ساختار بلوک‌های هاپلوتیپی بدست آمده از اجرای این روش‌ها بر روی برخی نمونه‌های واقعی صورت می‌گیرد. علاوه بر آن، ما توان روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها را بر اساس کارایی ساختارهای بلوکی بدست

[۳۸] robust

آمده از آنها در دو مسئله‌ی کاربردی مرتبط با مفهوم بلوک، مورد بررسی قرار می‌دهیم. این مسائل عبارتند از: شناسائی نقاط پراحتمال نوترکیبی در ژنوم و شناسائی جایگاه ژن مرتبط با خصیصه‌ی بیماری در نمونه‌های case و control. این دو بررسی را ما با استفاده از نمونه‌های شبیه‌سازی شده انجام می‌دهیم.

۶۰۵۰۲ سنجش توان شناسائی نقاط پراحتمال نوترکیبی

اولین موضوعی را که به عنوان کاربردی از روش‌های افراز بلوکی هاپلوتیپ‌ها مورد توجه قرار می‌دهیم، مسئله‌ی شناسائی نقاط پراحتمال[۳۹] نوترکیبی در ژنوم است(بخش ۴۰۱). از آنجا که هیچ توافق علمی فراگیری در رابطه با وجود و موقعیت نقاط پراحتمال نوترکیبی وجود ندارد، ما از داده‌های شبیه‌سازی شده، تحت مدلی مبتنی بر سازوکار نقاط پراحتمال نوترکیبی استفاده می‌کنیم. ما برای این منظور، نرم‌افزار msHOT[۱۵۹] را مورد استفاده قرار می‌دهیم. مدل بکارگرفته شده در این نرم‌افزار، تعمیمی از مدل مورد استفاده در الگوریتم هودسون[۱۶۰] است که در آن توالی‌های نوکلئوتیدی به طور تصادفی تحت مدل «نیای مشترک با نوترکیبی» تولید می‌شوند. در الگوریتم هودسون، نرخ نوترکیبی بین نوکلئوتیدهای مجاور در ژنوم، ثابت در نظر گرفته می‌شود اما msHOT این امکان را به کاربر می‌دهد که نرخ نوترکیبی و موقعیت نقاط پراحتمال نوترکیبی را به دلخواه انتخاب نماید.

توسط نرم‌افزار msHOT، ما ۱۰۰ مجموعه از نمونه‌های تصادفی، هر یک شامل ۴۰ هاپلوتیپ بر روی ۳۰۰ اسنیپ تولید کردیم. به طور مشابه، ۱۰۰ مجموعه‌ی دیگر از نمونه‌های تصادفی، هر یک شامل ۱۰۰ هاپلوتیپ بر روی ۳۰۰ اسنیپ با استفاده از msHOT بدست آوردیم. از این طریق، می‌توانیم تاثیر حجم نمونه، یعنی تعداد نمونه‌های هاپلوتیپ را بر دقت و کارایی روش‌های افراز در شناسائی نقاط پراحتمال نوترکیبی بررسی کنیم. ما پارامترهای مدل شبیه‌سازی مورد استفاده در msHOT را به گونه‌ای انتخاب کردیم که نواحی پراحتمال نوترکیبی در نمونه‌های تولید شده، عرضی حداکثر برابر با $2kb$ داشته باشند و حداکثر شش موقعیت پراحتمال نوترکیبی در یک ناحیه‌ی ۳۰۰ کیلوبازی قرار گیرند و نرخ نوترکیبی در آنها بین ۵۰ تا ۴۰۰ برابر نرخ نوترکیبی پس زمینه باشد. هدف از انتخاب این شرایط برای تولید نمونه‌های تصادفی، بدست آوردن نمونه‌هایی

[۳۹] hotspots

است که ویژگی‌هایی نزدیک به ویژگی‌های واقعی ژنوم انسان داشته باشند[۱۲۲]. ما موقعیت نقاط پراحتمال نوترکیبی را برای هر یک از مجموعه نمونه‌های تولید شده برای انجام محاسبات بعدی ثبت می‌کنیم.

هر یک از روش‌های مندرج در جدول ۱۰۲ را به جز HOT، بر روی مجموعه هاپلوتیپ‌های شبیه‌سازی شده توسط روال فوق، اجرا می‌کنیم. برای ارزیابی توان روش‌های افراز بلوکی هاپلوتیپ‌ها در شناسائی نقاط پراحتمال نوترکیبی، ما تعداد نقاطی را می‌شماریم که در هر یک از آنها، مرز یک بلوک هاپلوتیپ و یک ناحیه‌ی پراحتمال نوترکیبی بر یکدیگر منطبق هستند. دقت یک روش افراز بلوکی در شناسائی نقاط پراحتمال نوترکیبی، تابعی است از تعداد مرزهایی که بیرون نواحی پراحتمال قرار می‌گیرند که آنها را موارد false positive می‌نامیم و تعداد نواحی پراحتمالی که در نزدیکی مرز هیچ یک از بلوک‌ها قرار نمی‌گیرند که آنها را موارد false negative می‌نامیم. یک ناحیه‌ی پراحتمال را در نزدیکی مرز یک بلوک به حساب می‌آوریم اگر فاصله‌ی یکی از لبه‌های آن با مرز بلوک کمتر از kb۲ باشد. نسبت تعداد موارد false positive به تعداد کل بلوک‌ها را نرخ false positive و نسبت تعداد موارد false negative به تعداد کل نقاط پراحتمال را نرخ false negative می‌نامیم و مجموع آنها را به عنوان خطای کل در شناسائی نقاط پراحتمال نوترکیبی مورد بررسی قرار می‌دهیم. نسبت مرزهای بلوکی منطبق با نواحی پراحتمال نوترکیبی به تعداد کل نواحی پراحتمال نوترکیبی را به عنوان معیار کارایی در این مطالعه مورد بررسی قرار می‌دهیم(بخش ۷۰۳).

۷۰۵۰۲ سنجش توان شناسائی جایگاه ژنی یک خصیصه

به عنوان یکی دیگر از کاربردهای افراز بلوکی هاپلوتیپ‌ها، در این بخش، به معرفی شیوه‌ای مبتنی بر استفاده از ساختارهای بلوکی، برای شناسائی جایگاه ژنی یک خصیصه می‌پردازیم و از این طریق کارایی روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها را از زاویه‌ای دیگر در این رساله، مورد بررسی قرار می‌دهیم. برای این کار ما از طرح مطالعه‌ایی case-control و نمونه‌های شبیه‌سازی شده استفاده خواهیم کرد.

فرض کنید n_{case} هاپلوتیپ از افراد مبتلا به بیماری مورد مطالعه و $n_{control}$ هاپلوتیپ از دسته‌ای دیگر از افراد که اطلاعی از وجود خصیصه‌ی بیماری در آنها نداریم، داده شده‌اند. هدف، یافتن موقعیت ژن مرتبط با زمینه‌ی بیماری، به کمک این داده‌ها است. ما در اینجا دو روش برای بررسی همبستگی بین خصیصه‌ی مورد

مطالعه و یک جایگاه بر روی ژنوم معرفی می‌کنیم: روش «آزمون تک اسنیپ»[۴۰] که آنرا به اختصار SS می‌نامیم و روشی عمومی، مبتنی بر بلوک‌های هاپلوتیپ که نمونه‌هایی از آنرا به ازای چند مورد از روش‌های افراز بلوکی هاپلوتیپ‌ها، مندرج در جدول ۱۰۲ و با همان اختصار، مورد بررسی قرار می‌دهیم.

در «آزمون تک اسنیپ»، یک جدول توافقی مانند شکل ۱۰۰۲ برای هر اسنیپ مورد بررسی در ژنوم، مثل X تشکیل می‌دهیم که در آن n_{ij} نشاندهنده‌ی تعداد نمونه‌هایی با برچسب j است که آلل X در آنها i است. استفاده از این جدول نیز بسیار شبیه به جدول مورد استفاده در بخش ۳۰۲ است با این تفاوت که در اینجا

	$X=\mathbf{0}$	$X=\mathbf{1}$	
$case$	$n_{\mathbf{0},case}$	$n_{\mathbf{1},case}$	n_{case}
$control$	$n_{\mathbf{0},control}$	$n_{\mathbf{1},control}$	$n_{control}$
	$n_{\mathbf{0}}$	$n_{\mathbf{1}}$	n

شکل ۱۰۰۲: جدول توافقی برای مطالعه‌ی همبستگی بین یک اسنیپ و خصیصه

ما از آزمون مربع کای پیرسون استفاده می‌کنیم. بر این اساس، آماره‌ی آزمون برای روش SS، به صورت زیر تعریف می‌شود؛

$$\chi^{۲}_{ss} = n\, r^{۲} = \frac{n\left(n_{\mathbf{0},case}\, n_{\mathbf{1},control} - n_{\mathbf{1},case}\, n_{\mathbf{0},control}\right)^{۲}}{n_{\mathbf{0}}\, n_{\mathbf{1}}\, n_{case}\, n_{control}} \qquad (۱۱۰۲)$$

تحت فرض استقلال بین خصیصه و اسنیپ مورد مطالعه، مقدار $\chi^{۲}_{ss}$، دارای توزیع مربع کای با یک درجه آزادی است. فرض استقلال در این آزمون رد می‌شود اگر $\chi^{۲}_{ss} > \chi^{۲}_{۱,\alpha}$ که در آن $\chi^{۲}_{۱,\alpha}$ آستانه‌ی متناظر با ناحیه‌ی رد α درصدی در توزیع مربع کای با یک درجه آزادی است. اصولا $\alpha = ۰/۰۵$ استاندارد رایج در آزمون‌های فرض احتمال است. در اینجا، مقدار آستانه‌ای برای رد استقلال، متناظر با $\alpha = ۰/۰۵$ برابر است با $\chi^{۲}_{o} = ۳/۸۵$. بر این اساس، می‌گوئیم آزمون، یک اسنیپ را در همبستگی با خصیصه‌ی بیماری تشخیص داده است اگر $\chi^{۲}_{ss} > \chi^{۲}_{o}$.

اجرای این آزمون بر روی تک اسنیپ‌ها بسیار کم توان و پرخطا ظاهر می‌شود. به عنوان مثال، در بسیاری از جایگاه‌ها، به دلیل پائین بودن هتروزیگوسیتی، توان آزمون پائین می‌آید در حالیکه ممکن است در واقعیت

[۴۰] Single SNP Association Test

بین اسنیپ مورد مطالعه و خصیصه همبستگی وجود داشته باشد و یا بر عکس در بسیاری از جایگاه‌ها، به دلیل بالا بودن هتروزیگوسیتی و وجود "قشربندی"[۴۱] در جمعیت مورد مطالعه، تصمیم‌گیری بر اساس این آزمون پر خطا ظاهر می‌شود یعنی اسنیپ‌هایی نامرتبط با خصیصه به عنوان اسنیپ‌های مرتبط با خصیصه شناسائی می‌شوند. رویکردی رایج برای بهبود این روش، ایده‌ی «بیشترین امتیاز»[۴۲] است[۱۶۱]. در این رویکرد، به منظور کاهش اثر شرایط فوق بر توان آزمون، آماره‌ی مورد استفاده برای آزمون مربع کای برابر با بیشترین مقدار بدست آمده برای این آماره در بازه‌ای شامل w اسنیپ در دو طرف اسنیپ مورد مطالعه، در نظر گرفته می‌شود. این رویکرد بر این ایده تکیه دارد که تشخیص همبستگی بین خصیصه و اسنیپ‌هایی که در نزدیکی اسنیپ مرتبط با بیماری قرار گرفته‌اند به دلیل وجود LD می‌تواند معنادار باشد. با این حال افزایش عرض پنجره، از سویی دیگر می‌تواند باعث افزایش خطا در آزمون گردد. با برآوردی که ما از اجرای این روش به ازای مقادیر مختلف w بر روی بخشی از نمونه‌های شبیه‌سازی شده بدست آوردیم، $w = ۳$ بهترین انتخاب از حیث بیشترین توان و کمترین خطا در بین انتخاب‌های دیگر برای این پارامتر است.

خوشه‌بندی سلسله مراتبی برای تشخیص رابطه بین خصیصه و هاپلوتیپ‌ها در یک بلوک

در روش دوم برای جستجوی جایگاه ژنی مرتبط با یک خصیصه، ما تنوع هاپلوتیپ‌ها را در بین نمونه‌های case و control، به منظور تشخیص وجود رابطه بین خصیصه و ناحیه‌ی ژنومی محدود به یک بلوک، مورد بررسی قرار می‌دهیم. همانطور که در بخش ۵۰۱ اشاره شد، ابزار رایج برای سنجش همبستگی بین خصیصه و ژنوتیپ‌های ناحیه‌ی کروموزومی تحت بررسی، استفاده از «آزمون نسبت درستنمائی»[۴۳] (رابطه ۱۲۰۱) است. اصولا، مدل‌های ریاضی بسیار متنوعی برای تعریف درستنمائی بر حسب داده‌های case و control معرفی شده‌اند که کم و بیش در تمامی آنها لازم است برای محاسبه‌ی درستنمائی، هاپلوتیپ‌های داده شده خوشه‌بندی شوند[۱۴۸]. برای سادگی در اینجا، بدون وارد شدن به بحث درباره‌ی انتخاب یک مدل درستنمائی، ما از آماره‌ی مربع کای پیرسون که در واقع تقریب مناسبی برای رابطهٔ (۱۲۰۱) است، استفاده می‌کنیم.

فرض کنید هاپلوتیپ‌های نمونه‌ی داده شده در یک بلوک را در m دسته، مثل $\mathcal{C}_۱, \mathcal{C}_۲, \ldots, \mathcal{C}_m$ خوشه‌بندی

[۴۱] Stratification
[۴۲] Max score
[۴۳] Likelihood ratio test

کرده‌ایم. بر این اساس، برای هر بلوک یک جدول توافقی مانند شکل ۱۱۰۲ تشکیل می‌دهیم.

	$\mathcal{C}_1$	$\mathcal{C}_2$	...	$\mathcal{C}_m$	
$case$	A_1	A_2	...	A_m	n_{case}
$control$	B_1	B_2	...	B_m	$n_{control}$
	C_1	C_2	...	C_m	n

شکل ۱۱۰۲: جدول توافقی برای مطالعه‌ی همبستگی بین یک بلوک از هاپلوتیپ‌ها و خصیصه

در شکل ۱۱۰۲، A_i و B_i به ترتیب فراوانی هاپلوتیپ‌های case و control در خوشه‌ی $\mathcal{C}_i$ هستند و C_i تعداد کل هاپلوتیپ‌ها در خوشه‌ی $\mathcal{C}_i$ است. آماره‌ی مربع کای برای این جدول بر اساس رابطه‌ی زیر تعریف می‌شود:

$$\chi^2_{block} = \sum_{i=1}^{m-1} \frac{(A_i/n_{case} - B_i/n_{control})^2}{C_i/n_{case}n_{control}} . \qquad (۱۲۰۲)$$

تحت فرض استقلال، مقدار بدست آمده از رابطهٔ (۱۲۰۲) دارای توزیع مربع کای با $m-1$ درجه آزادی است. همانند آزمون تک اسنیپی، در اینجا نیز اگر $\chi^2_{block} > \chi^2_{m-1,\alpha}$، آنگاه رد فرض استقلال در سطح $1-\alpha$ معنادار است. همانطور که ملاحظه می‌کنید، در محاسبات مربوط به این آزمون، بر خلاف روش‌های مبتنی بر آزمون نسبت درستنمائی، نیازی به دانستن مدل آماری حاکم بر بیماری نیست.

روشی که تا اینجا شرح دادیم، به ازای هاپلوتیپ‌های برچسب‌گذاری شده با case و control و به وسیله‌ی یک خوشه‌بندی داده شده، درباره‌ی وجود یا عدم وجود همبستگی آماری بین بیماری و ناحیه‌ی کروموزمی محدود به یک بلوک از هاپلوتیپ‌ها تصمیم می‌گیرد. در ادامه، روشی برای تعیین یک خوشه‌بندی مناسب برای اجرای این آزمون، پیشنهاد می‌کنیم که به واسطه‌ی آن توان و دقت آزمون فوق، از برخی جهات بهبود می‌یابد.

ساده‌ترین رویکرد برای دسته‌بندی نمونه‌ای از هاپلوتیپ‌ها، قرار دادن نمونه‌های یکسان در یک دسته است؛ مشابه ایده‌ی مورد استفاده در تعریف "هاپلوتیپ‌های رایج". در این روش، حتی نمونه‌هایی که اختلاف بین هاپلوتیپ‌هایشان ناچیز است در خوشه‌های جداگانه‌ای قرار داده می‌شوند که باعث می‌شود استفاده از این رویکرد در کنار آزمون مربع کای پیرسون برای ارزیابی همبستگی، با اشکالاتی همراه شود. به عنوان مثال، در بلوک‌هایی که تنوع هاپلوتیپ‌ها بالا است، تعداد خوشه‌های بدست آمده، یعنی m بالا خواهد بود و وقتی

درجات آزادی بالا باشد، آزمون مربع کای اصطلاحا با "محافظه‌کاری" بیشتری تصمیم می‌گیرد و در واقع توان آن کاهش می‌یابد. به علاوه، استفاده از این آزمون برای جدولی که مقادیر برخی خانه‌های آن کمتر از پنج است، بنابر دلایل ریاضی، از دقت و اعتبار کافی برخوردار نیست.

با توجه به ملاحظات فوق، ما یک روش مبتنی بر خوشه‌بندی سلسله‌مراتبی[۴۴] را برای تعیین یک خوشه‌بندی کارآمد برای اجرای آزمون مربع کای در بلوک‌های هاپلوتیپی بکار می‌بندیم. در این روش ابتدا، فاصله‌ی همینگ[۴۵] بین هاپلوتیپ‌های هر جفت از نمونه‌های داده شده را محاسبه می‌کنیم. جفت نمونه‌ای را که فاصله‌ی همینگ بین هاپلوتیپ‌هایشان بیشترین است انتخاب می‌کنیم و هر یک را در خوشه‌ای جدا قرار می‌دهیم و به آنها هاپلوتیپ‌های مرجع می‌گوئیم. سپس، هر یک از دیگر نمونه‌ها را در یکی از این دو خوشه قرار می‌دهیم به طوری که فاصله‌ی هر هاپلوتیپ از هاپلوتیپ مرجع در خوشه‌ای که در آن قرار دارد کمتر از فاصله‌ی آن هاپلوتیپ تا هاپلوتیپ مرجع در خوشه‌ی دیگر باشد. در این مرحله، یک جدول توافقی بر اساس خوشه‌بندی بدست آمده تشکیل می‌دهیم و آماره‌ی مربع کای مربوط به آنرا با عنون امتیاز خوشه‌بندی سطح یک، ثبت می‌کنیم. از بین خوشه‌های بدست آمده، خوشه‌ای را که فاصله‌ی دورترین اعضایش از یکدیگر بیشترین است انتخاب می‌کنیم و با اجرای روال مشابه بر روی آن، آنرا به دو خوشه‌ی کوچکتر افراز می‌کنیم و در کنار سایر خوشه‌ها قرار می‌دهیم. این روال را تا زمانی که خوشه‌ای با بیش از پنج هاپلوتیپ وجود داشته باشد که اختلاف بین متفاوت‌ترین اعضایش در بیش از چهار درصد از اسنیپ‌ها باشد ادامه می‌دهیم. در هر سطح از روال سلسله‌مراتبی فوق، آماره‌ی مربع کای متناظر با خوشه‌بندی بدست آمده در آن سطح را محاسبه و ثبت می‌کنیم. به یاد داشته باشید که درجه آزادی آماره‌ی مربع کای، در سطح kام از روال سلسله‌مراتبی فوق برابر با k است. از این رو، برای داشتن یک معیار مستقل برای اجرای آزمون همبستگی، ما pـ مقدار مرتبط با آماره‌ی مربع کای را در هر سطح محاسبه می‌کنیم. در بین pـ مقدارهای بدست آمده، کمترین pـ مقدار را به عنوان نتیجه‌ی نهایی الگوریتم خوشه‌بندی سلسله‌مراتبی در نظر می‌گیریم.

با استفاده از الگوریتم فوق، یک خوشه‌بندی مناسب برای محاسبه‌ی آماره‌ی χ^2_{block} (رابطه ۱۲۰۲)، بدست

[۴۴] Hierarchical clustering
[۴۵] Hamming distance

می‌آید. بدیهی است که افرازهای بلوکی متفاوت، هاپلوتیپ‌های متفاوتی بر روی ژنوم تعریف می‌کنند که به تبع آن، روال خوشه‌بندی و نتیجه‌ی آزمون همبستگی بین خصیصه و هاپلوتیپ‌ها ممکن است تغییر کند. هدف ما این است که دریابیم کدام یک از روش‌های افراز بلوکی در تشخیص جایگاه ژنی مرتبط با خصیصه دقیق‌تر است. برای این منظور، ما ساختارهای بلوکی پیشنهاد شده توسط روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها را در کنار الگوریتم خوشه‌بندی و آزمون همبستگی شرح داده شده در بالا، بر روی نمونه‌های شبیه‌سازی شده case و control، ارزیابی می‌کنیم.

طرح مقایسه‌ایی برای سنجش کارایی روش‌های افراز بلوکی در شناسائی جایگاه ژنی مرتبط با بیماری

طرح ما برای ارزیابی کارایی و دقت روش‌های افراز بلوکی در زمینه‌ی شناسائی جایگاه ژنی مرتبط با بیماری شامل مراحل زیر است:

۱. با استفاده از هاپلوتیپ‌های موجود در پایگاه داده‌ای HapMap، ساختار بلوکی کروموزم‌ها در یک جمعیت معین تعیین می‌شوند.

۲. با توجه به هزینه و محدودیت‌های فنی در تعیین ژنوتیپ اسنیپ‌ها، تعداد مناسبی از اسنیپ‌ها از بین مجموعه اسنیپ‌های شناخته شده بر روی ژنوم، به عنوان نشانگذار انتخاب می‌شوند تا ژنوتیپ نمونه‌های case و control در آنها تعیین گردند.

۳. آزمون همبستگی بر روی هاپلوتیپ نمونه‌های case و control درون هر بلوک اجرا می‌شود. از این طریق تعدادی از بلوک‌ها به عنوان جایگاه‌های مرتبط با بیماری تعیین می‌شوند.

۴. با مقایسه‌ی بلوک‌های تشخیص داده شده به عنوان جایگاه‌های مرتبط با بیماری و موقعیت واقعی اسنیپ مسبب بیماری، توان و دقت ساختار بلوکی پیشنهاد شده را ارزیابی می‌کنیم.

تاکید می‌کنیم که افراز بلوکی تنها یکبار و آن هم بر اساس داده‌های HapMap تعیین می‌شود سپس از همان ساختار بلوکی برای اجرای آزمون همبستگی در نمونه‌های case و control استفاده می‌شود. در مرحله‌ی اول از طرح فوق، ما از ساختارهای بلوکی پیشنهاد شده توسط شش روش مختلف افراز بلوکی هاپلوتیپ‌ها، در

ده ناحیه‌ی ENCODE استفاده کردیم؛ این روش‌ها عبارتند از MB، HB، MDL، GAB، GPG و GPF (جدول ۱۰۲).

در مرحله‌ی دوم از طرح فوق، ما از نرم‌افزار gs [۱۶۲] برای تولید نمونه‌های تصادفی case و control تحت یک مدل بیماری تک جایگاهی استفاده کردیم. هاپلوتیپ‌های شبیه‌سازی شده توسط این نرم‌افزار بر پایه‌ی مجموعه‌ای از هاپلوتیپ‌های از پیش تعیین شده بدست می‌آیند و از این رو، از نظر تنوع هاپلوتیپ‌ها و ساختار LD مشابه نمونه‌ی داده شده‌اند. این قابلیت نرم‌افزار برای ما بسیار مهم است چون برای بررسی توان روش‌های افراز بلوکی در مطالعه‌ی case-control لازم است نمونه‌های مورد مطالعه از جمعیتی باشند که بلوک‌های هاپلوتیپ بر پایه‌ی نمونه‌های همان جمعیت تعریف شده‌اند. از آنجا که ما از هاپلوتیپ‌های پانل CEU در HapMap برای تعیین بلوک‌های ناحیه‌های ENCODE استفاده کرده‌ایم، در اینجا نیز برای تولید نمونه‌های case و control توسط نرم‌افزار gs از همان داده‌ها استفاده می‌کنیم. نرم‌افزار gs، دو گزینه‌ی متفاوت برای تولید نمونه‌های تصادفی با استفاده از مجموعه‌ای از هاپلوتیپ‌های داده شده ارائه می‌دهد که ما از گزینه‌ی `extension model` با پارامترهای پیش‌فرض آن استفاده کردیم. در گزینه‌ی دیگر این نرم‌افزار، نمونه‌ها بر پایه‌ی افراز بلوکی هاپلوتیپ‌های مرجع تولید می‌شوند که استفاده از آن در این بررسی، می‌تواند باعث تولید نتایج سوگیری شده شود.

مدل ژنتیکی حاکم بر ژنوتیپ‌های مستعد بیماری را نیز می‌توان به دلخواه برای این نرم‌افزار تعیین کرد. ما از دو مدل جمعی، یکی با $GRR_1 = 3$ و دیگری با $GRR_1 = 5$ به عنوان مدل‌های آماری توارث بیماری برای تولید نمونه‌های تصادفی case و control استفاده کردیم. دیگر پارامتر قابل توجه در شبیه‌سازی نمونه‌های case و control، تحت مدل تک جایگاهی بیماری، فراوانی نسبی آلل مرتبط با بیماری در جمعیت است که به اختصار با DAF نشان می‌دهیم و البته، تنها یکی از عوامل موثر بر شیوع بیماری در جمعیت است. به ازای مقادیر مختلف DAF، نمونه‌های تصادفی متعددی توسط gs تولید کردیم و آنها را به طور جداگانه در دو گروه، یکی با مقادیر پایین DAF (بین ۰/۰۵ تا ۰/۱۵) و دیگری با مقادیر بالای DAF (بین ۰/۲۰ تا ۰/۳۰) برای استفاده در مطالعات case-control، مورد بررسی قرار دادیم.

به ازای هر یک از تنظیمات فوق و هر یک از ده ناحیه‌ی ENCODE، ۵۰۰ مجموعه از نمونه‌هایی شامل ۵۰ نمونه‌ی case و ۵۰ نمونه‌ی control، با استفاده از هاپلوتیپ‌های HapMap در پانل CEU تولید کردیم. در هر مجموعه، مکان اسنیپ مسبب بیماری که توسط الگوریتم شبیه‌سازی gs به طور تصادف از بین اسنیپ‌های مختلف انتخاب می‌شود را یادداشت می‌کنیم تا در مرحله‌ی بعد برای سنجش توان و دقت نتایج بدست آمده از آزمون همبستگی بکار گیریم. پیش از آغاز مرحله‌ی بعد، در هر یک از مجموعه‌های شبیه‌سازی شده، ستون مربوط به اطلاعات اسنیپ مسبب بیماری را به طور کامل حذف می‌کنیم.

روش خوشه‌بندی و آزمون‌هایی که در ابتدای این بخش برای بررسی همبستگی بین خصیصه و جایگاه‌های ژنی، معرفی گردیدند، جملگی با فرض در دست داشتن اطلاعات تعیین فاز شده، اجرا می‌شوند و از این رو، برای استفاده از این روش‌ها در مسائل عملی لازم است روال جداگانه‌ای برای تفکیک ژنوتیپ‌ها به هاپلوتیپ‌های تشکیل‌دهنده، پیش از آغاز مطالعه‌ی case-control بکار گرفته شود. خوشبختانه، در طرح مقایسه‌ایی ما، نمونه‌های تصادفی تولید شده توسط gs،نمونه‌های هاپلوتیپ هستند و بنابراین در اینجا، مسئله‌ی تعیین فاز ژنوتیپ‌ها به میان نمی‌آید.

راهبردهای انتخاب اسنیپ‌ها به عنوان نشانگذار در مطالعه‌ی case-control

یکی دیگر از جنبه‌های شایان توجه در بررسی توان روش‌های مختلف افراز بلوکی در «مطالعه‌ی همبستگی در مقایس ژنومی»[۴۶] (GWAS)، نقش چگالی نشانگذارها بر کارایی این روش‌ها و دقت نتایج بدست آمده از آنها است. همانطور که در مقدمه اشاره شد، اسنیپ‌ها گزینه‌های بسیار متداولی برای ایفای نقش نشانگذارها در GWAS هستند. بدیهی است، با در اختیار داشتن نشانگذارهایی با پراکندگی یکنواخت و به قدر کافی نزدیک به هم بر روی ژنوم، تعیین جایگاه مرتبط با بیماری با دقت بیشتری امکان‌پذیر است. متاسفانه، داشتن چنین داده‌هایی نیازمند تعیین ژنوتیپ تعداد بسیار زیادی از اسنیپ‌ها است. توجه به همبستگی‌های موجود بین اسنیپ‌ها، توجیه اقتصادیِ تعیین ژنوتیپ تمام اسنیپ‌های شناخته‌شده را زیر سئوال می‌برد. در اینجا ما دو راهبرد برای انتخاب اسنیپ‌ها به عنوان نشانگذارها در GWAS را با مطالعه بر روی نمونه‌های شبیه‌سازی شده‌ی بالا، مورد بررسی قرار می‌دهیم.

[۴۶] Genome-wide association study

در راهبرد اول، از میان هر k اسنیپ شناخته‌شده‌ی متوالی بر روی ژنوم، یکی را به عنوان نشانگذار انتخاب می‌کنیم. این راهبرد را انتخاب یکنواخت نشانگذارها می‌نامیم. بدیهی است، به ازای $k = ۱$ یک نقشه‌ی تا حد ممکن دقیق از موقعیت‌های افتراقی[۴۷] ژنوم برای انجام GWAS در اختیار خواهیم داشت. ما حالت‌های دیگری را نیز به ازای $k = ۲, ۳, \ldots, ۱۰$، برای انتخاب یکنواخت اسنیپ‌ها به عنوان نشانگذار مورد بررسی قرار دادیم که از طریق آن می‌توانیم روند کاهش توان تشخیص جایگاه مرتبط با بیماری را در برابر کاهش چگالی نشانگذارها، در بین روش‌های مختلف مقایسه کنیم.

برای انتخاب نشانگذارها در راهبرد دوم، از شیوه‌ای مشابه الگوریتم کارلسون[۱۳۰] استفاده می‌کنیم (بخش ۴۰۱). در الگوریتم کارلسون برای انتخاب تگ‌اسنیپ‌ها، از یک روال تکرارشونده استفاده می‌شود که در هر تکرار آن اسنیپی که در همبستگی با بیشترین تعداد از دیگر اسنیپ‌ها است به عنوان تگ اسنیپ انتخاب می‌شود و این روال با تکرار بر روی اسنیپ‌های باقی مانده ادامه می‌یابد. بر اساس همین روش، ما اسنیپ‌ها را در هر بلوک بر حسب شماره‌ی مرحله‌ای که در الگوریتم کارلسون مورد ارزیابی قرار می‌گیرند، رتبه‌بندی می‌کنیم. برای انتخاب نشانگذارها در راهبرد دوم، اسنیپ‌ها را بر حسب رتبه‌یشان، به ترتیب از هر بلوک انتخاب می‌کنیم تا زمانی که نشانگذارها به تعداد مطلوب بدست آیند. به عبارت دقیق‌تر، در این روش ابتدا، اسنیپ‌های با رتبه‌ی یک از تمام بلوک‌ها انتخاب می‌شوند. اگر تعداد آنها کمتر از تعداد مورد نیاز برای نشانگذارها باشد همین روال با انتخاب اسنیپ‌های رتبه‌ی دو، سه و ...ادامه داده می‌شود. در این راهبرد نیز، تعداد نشانگذارهای مورد بررسی را بین ۱ : ۱ تا ۱۰ : ۱ تعداد کل اسنیپ‌های شناخته‌شده در ناحیه‌ی مورد مطالعه تغییر می‌دهیم. نتایج اتخاذ این راهبرد را با عنوان انتخاب اولویت داده شده‌ی نشانگذارها در بخش ۱۰۸۰۳ بررسی می‌کنیم.

برآورد خطای نوع اول و توان آزمون

تا اینجا، جزئیات نسبتا کاملی از طرح ما برای مطالعات case-control به کمک اطلاعات ساختار بلوکی ژنوم و نیز تهیه‌ی نمونه‌های شبیه‌سازی شده برای این مطالعه ارائه گردید. در انتهای این بخش، اشاره‌ای خواهیم داشت به نحوه‌ی برآورد خطا و اندازه‌گیری توان آزمون در نتایج بدست آمده از بررسی‌های فوق. ابتدا برخی

[۴۷] segregation sites

واژگان مورد نیاز در این مبحث را مرور می‌کنیم.

در هر بلوک، نتیجه‌ی آزمون همبستگی بین خصیصه و هاپلوتیپ‌های بلوک، یکی از چهار حالت زیر را تعریف می‌کند:

true positive . وجود ارتباط بین خصیصه و تنوع هاپلوتیپ‌های درون بلوک توسط آزمون پذیرفته شده است و اطلاعات از پیش ثبت شده‌ی ما نیز نشان می‌دهد که اسنیپ مسبب بیماری درون همین بلوک واقع شده است.

false positive . آزمون وجود ارتباط بین خصیصه و بلوک را پذیرفته است اما در واقع، اسنیپ مسبب بیماری در مکانی خارج از این بلوک قرار دارد.

false negative . آزمون وجود ارتباط بین خصیصه و بلوک را رد می‌کند اما اسنیپ مسبب بیماری در همین بلوک قرار دارد.

true negative . آزمون وجود ارتباط بین خصیصه و بلوک را رد می‌کند و اطلاعات از پیش ثبت شده‌ی ما نیز مؤید آن است که اسنیپ مسبب بیماری در این بلوک قرار ندارد.

برای هر یک از ساختارهای بلوکی پیشنهاد شده توسط روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها، تعداد کل نتایج true positive بدست آمده از اجرای آزمون بر روی نمونه‌های شبیه‌سازی شده، در بلوک‌های نواحی ENCODE را با TP نشان می‌دهیم. تعداد کل هر یک از موارد false positive، false negative و true negative را تحت همین شرایط بدست می‌آوریم و به ترتیب با FP، FN و TN نشان می‌دهیم. بر این اساس، خطای نوع اول و توان در تشخیص جایگاه ژنی مرتبط با بیماری، به صورت زیر تعریف می‌شود:

$$\text{Type } I \text{ error} = \frac{FP}{TN + FP}, \qquad \text{Power} = \frac{TP}{TP + FN} \tag{۱۳۰۲}$$

به بیان ساده، خطای نوع اول نشاندهنده‌ی احتمال آن است که آزمون به اشتباه، بلوکی را که اسنیپ مسسب بیماری را شامل نمی‌شود به عنوان بلوک مرتبط با بیماری تشخیص دهد و توان نشاندهنده‌ی نسبت تشخیص‌های

درست آزمون به کل مواردی است که بلوک اسنیپ مسبب بیماری را شامل می‌شود.

مقایسه‌ی خطای نوع اول بین روش‌های مختلف، ابزار مناسبی است که نشان می‌دهد کدام یک از روش‌ها در تشخیص جایگاه مرتبط با بیماری از دقت بالاتری برخوردارند. اما برای مقایسه‌ی توان بین روش‌های مختلف برخی ملاحظات را باید در نظر گرفت. یک روش ممکن است تعداد زیادی از بلوک‌ها را به عنوان بلوک‌های مرتبط با بیماری تشخیص دهد و از این طریق توان قابل قبولی در تشخیص بلوک‌های مرتبط با بیماری بدست آورد اما باید در نظر داشت که میزان خطای نوع اول در این روش به همان میزان می‌تواند بالا باشد. از طرف دیگر، یک روش ممکن است نسبت به وجود ارتباط بین بیماری و موقعیت محدوده شده به یک بلوک، بسیار "محافظه‌کارانه" تصمیم بگیرد یعنی به ندرت یک بلوک را به عنوان بلوک مرتبط با بیماری تشخیص دهد تا خطای نوع اول کمتری را مرتکب شود اما این رویکرد به طور همزمان توان روش را نیز کاهش می‌دهد. اصولا، بین توان یک آزمون و دقت تشخیص در آن، یک موازنه بر قرار است به قسمی که همواره افزایش یکی با کاهش دیگری همراه است. از این رو، مقایسه‌ی بین توان روش‌های مختلف، تنها زمانی معنادار است که خطای نوع اول در تمام آنها در سطح یکسانی باشد.

بنابر توضیحات فوق، هدف ما مقایسه‌ی توان روش‌های مختلف، با فرض ثابت نگهداشتن خطای نوع اول در آنها است. نتایج اولیه‌ی اجرای آزمون همبستگی بر روی نیمی از نمونه‌های شبیه‌سازی شده (۲۵۰ نمونه)، به ما نشان داد که خطای نوع اول در بین روش‌های مختلف متفاوت است(جدول ۹۰۳). از نظر تئوری، اگر آماره‌ی χ^2_{block}، دقیقا دارای توزیع مربع کای باشد، سطح معناداری انتخاب شده برای اجرای آزمون مربع کای، از نظر عددی با خطای نوع اول بدست آمده برابر است. در واقع، به دلیل مدل ژنتیکی بیماری و تاثیر ضمنی نحوه‌ی افراز بلوک‌های هاپلوتیپ بر خوشه‌بندی هاپلوتیپ‌ها در نمونه‌های case و control، آماره‌ی χ^2_{block} به طور دقیق از توزیع مربع کای پیروی نمی‌کند و از این رو یکسان نگه داشتن سطح معناداری آزمون مربع کای برای تمام روش‌ها، به تفاوت میزان خطای نوع اول در آنها منجر شود. با توجه به این ملاحظات ما سطح معناداری آزمون مربع کای را متناسب با هر روش به نحوی انتخاب می‌کنیم که میزان خطای نوع اول در تمام روش‌ها به یک اندازه برسد. به این منظور، برای هر روش به طور جداگانه، pـ مقدار آستانه‌ایی برای رد فرض استقلال در آزمون مربع کای را به تدریج افزایش می‌دهیم تا جائیکه نرخ خطای نوع اول در آن ـ برآورد شده

از ۲۵۰ نمونه‌ی اول ــ به حدود ۱ ٪ برسد(جدول ۱۰۰۳).

مشابه روال فوق را برای روش «آزمون تک اسنیپی»، نیز اجرا می‌کنیم. به یاد داشته باشید که در اینجا، حالتی به عنوان true positive در نظر گرفته می‌شود که در آن، آزمون وجود ارتباط بین بیماری و اسنیپ را پذیرفته باشد و اسنیپ مسبب بیماری در بازه‌ای به شعاع سه اسنیپ اطراف آن قرار داشته باشد. پس از بدست آوردن pــ مقدار آستانه‌ای مناسب برای هر یک از روش‌های تشخیص جایگاه ژنی مرتبط با بیماری، شامل روش‌های مبتنی بر ساختار بلوکی و روش SS، توان این روش‌ها را با بررسی نتایج بدست آمده از اجرای آنها بر روی نیمه‌ی دوم از ۵۰۰ نمونه‌ی شبیه‌سازی شده برآورد می‌کنیم. تمامی روال فوق، به ازای هر یک از مدل‌های بیماری و مقادیر مختلف DAF و هر یک از راهبردهای انتخاب نشانگذار به طور جداگانه اجرا می‌شود.

فصل ۳

نتایج و بحث

۱۰۳ کارآیی الگوریتم ژنتیک در استنباط هاپلوتیپ‌ها

در این بخش، ابتدا نتایج بدست آمده از اجرای الگوریتم‌های naive-GAhap و GAhap بر روی داده‌های شبیه‌سازی شده مورد بحث قرار می‌گیرد و در انتها، کارایی الگوریتم GAhap بر روی نمونه‌ای واقعی از ژنوتیپ‌ها با دیگر روش‌های رایج در حل مسئله‌ی تفکیک ژنوتیپ‌ها و استنباط هاپلوتیپ‌ها مقایسه می‌گردد.

همانطور که در بخش ۱۰۲ شرح داده‌ایم لازم است الگوریتم‌های naive-GAhap و GAhap به ازای انتخاب‌های متفاوت برای پارامترهای درگیر در شکل عمومی الگوریتم‌های ژنتیکی، اجرا گردند تا برآوردی از کارآمدترین انتخاب برای این پارامترها بدست آید. پنج پارامتر مورد بحث در اینجا، بر اساس واژگان مورد استفاده در الگوریتم GAhap عبارتند از: نرخ نوترکیبی در الگوریتم ژنتیک cr، نرخ نوترکیبی بین هاپلوتیپ‌های راهنمای یک ژنوتیپ cr_{int}، نرخ جهش در هاپلوتیپ‌های راهنما mr_{int}، شیوه‌ی انتخاب "کروموزم‌های" والد هر نسل و شیوه‌ی تبدیل تابع هدف به تابع سازگاری. هر یک از الگوریتم‌های naive-GAhap و GAhap به ازای انتخاب‌های مختلف برای این پارامترها و بر روی ۲۰ نمونه‌ی شبیه‌سازی شده‌ی متفاوت، هر یک شامل ۴۰ ژنوتیپ و ۱۲ اسنیپ به طور مستقل اجرا شدند. با توجه به تعداد گزینه‌های مورد بررسی برای هر یک از این پارامترها، در مجموع ۸۶۴۰ اجرای مختلف از هر یک از این دو الگوریتم به انجام

رسید(بخش ۱۰۲).

در هر اجرا، تعداد هاپلوتیپ‌های متمایز در جواب بدست آمده را با تعداد هاپلوتیپ‌های متمایز در هاپلوتیپ‌هایی که برای تولید ژنوتیپ‌های شبیه‌سازی شده بکار گرفته شده‌اند مقایسه می‌کنیم. از آنجا که هدف الگوریتم بدست آوردن جوابی با کمترین تعداد هاپلوتیپ متمایز است الگوریتم را "موفق" به حساب می‌آوریم اگر مقدار گزارش شده توسط الگوریتم کوچکتر از، یا برابر با تعداد هاپلوتیپ‌های متمایز داده‌های مبداء باشد. کارآمدی تنظیمات مختلف برای پارامترهای الگوریتم ژنتیک را با شمارش تعداد موفقیت‌های الگوریتم در بین ۲۰ اجرای مختلف آن بر روی نمونه‌های شبیه‌سازی شده اندازه می‌گیریم. به طور کلی، یک انتخاب از مقادیر پارامتر را کارآمدتر از دیگری می‌دانیم اگر تعداد موفقیت‌های الگوریتم در این ۲۰ اجرای به ازای این مقادیر پارامتر بیشتر از دیگر انتخاب باشد.

نکته‌ی مهم، عدم توانایی الگوریتم naive-GAhap در رسیدن به جواب بهینه است. نتایج بدست آمده از naive-GAhap به ازای هیچ یک از تنظیمات پارامتر به مقدار بهینه‌ی جواب نمی‌رسد و تعداد هاپلوتیپ‌های متمایز در جواب‌های بدست آمده از این الگوریتم به طور متوسط حدود ۱۰ هاپلوتیپ بیشتر از تعداد از پیش شناخته شده در ژنوتیپ‌های شبیه‌سازی شده است. از این رو، بحث بر روی نتایج را تنها با بررسی الگوریتم بهبود یافته، یعنی GAhap ادامه خواهیم داد.

جدول ۱۰۳ نتایج بهترین انتخاب برای پارامترهای مختلف الگوریتم ژنتیک را نشان می‌دهد. هر سطر جدول، بهترین تنظمیات برای الگوریتم ژنتیک را وقتی مقدار یکی از پارامترها ثابت نگه داشته شده است نشان می‌دهد. پیشرفت الگوریتم GAhap در مقایسه با روش ابتدائی naive-GAhap بسیار چشمگیر است به قسمی که اجرای آن در ٪۵۲ از کل ۸۶۴۰ نمونه‌ی مورد بررسی، با "موفقیت" همراه بوده است. این موضوع می‌تواند نشان‌دهنده‌ی تاثیر معنادار استفاده از ایده‌ی الگوریتم‌های سودجویانه در این الگوریتم باشد. همانطور که در جدول ۱۰۳ مشاهده می‌شود بهترین نتایج این الگوریتم با انتخاب نرخ ۰/۸ برای "کراس‌اوِر" جوابها، ۰/۹ برای "کراس‌اوِر" هاپلوتیپ‌های راهنما، همین مقدار برای نرخ "جهش" در هاپلوتیپ‌های راهنما و استفاده از انتخاب تصادفی یکنواخت به عنوان شیوه‌ی انتخاب "کروموزم‌های" والد در هر نسل و استفاده از نسبت خطی انتقال‌یافته برای تبدیل تابع هدف به تابع سازگاری، بدست آمده است. بر همین اساس، این

تنظیمات را از این پس به عنوان تنظیمات پیش‌فرض برای روش پیشنهادی GAhap در نظر می‌گیریم.

جدول ۱۰۳: بهترین تنظیمات برای پارامترهای الگوریتم ژنتیک GAhap

گزینه‌ی مورد انتخاب*	بهترین تنظیمات	تعداد موفقیت‌ها در ۲۰ آزمایش
cr		
۰/۲	$cr_{int} = ۰/۵, mr_{int} = ۰/۵$, *stochastic, rank*	۱۶
۰/۸	$cr_{int} = ۰/۹, mr_{int} = ۰/۹$, *stochastic, shift linear*	۱۸
۰/۹	$cr_{int} = ۰/۹, mr_{int} = ۰/۹$, *tournament, rank*	۱۴
cr_{int}		
۰/۱	$cr = ۰/۲, mr_{int} = ۰/۹$, *uniform, rank*	۱۵
۰/۵	$cr = ۰/۲, mr_{int} = ۰/۵$, *stochastic, rank*	۱۶
۰/۹	$cr = ۰/۸, mr_{int} = ۰/۹$, *stochastic, shift linear*	۱۸
mr_{int}		
۰/۱	$cr = ۰/۲, cr_{int} = ۰/۵$, *roulette, shift linear*	۱۶
۰/۵	$cr = ۰/۲, cr_{int} = ۰/۵$, *uniform, top*	۱۶
۰/۹	$cr = ۰/۸, cr_{int} = ۰/۹$, *stochastic, shift linear*	۱۸
انتخاب "کروموزوم‌ها"		
stochastic	$cr = ۰/۸, cr_{int} = ۰/۹, mr_{int} = ۰/۹$, *shift linear*	۱۸
uniform	$cr = ۰/۲, cr_{int} = ۰/۵, mr_{int} = ۰/۵$, *top*	۱۶
roulette	$cr = ۰/۲, cr_{int} = ۰/۵, mr_{int} = ۰/۱$, *shift linear*	۱۶
tournament	$cr = ۰/۲, cr_{int} = ۰/۹, mr_{int} = ۰/۱$, *linear*	۱۶
تابع سازگاری		
rank	$cr = ۰/۲, cr_{int} = ۰/۵, mr_{int} = ۰/۵$, *stochastic*	۱۶
top	$cr = ۰/۲, cr_{int} = ۰/۵, mr_{int} = ۰/۵$, *uniform*	۱۶
linear	$cr = ۰/۲, cr_{int} = ۰/۹, mr_{int} = ۰/۱$, *tournament*	۱۶
shift linear	$cr = ۰/۸, cr_{int} = ۰/۹, mr_{int} = ۰/۹$, *stochastic*	۱۸

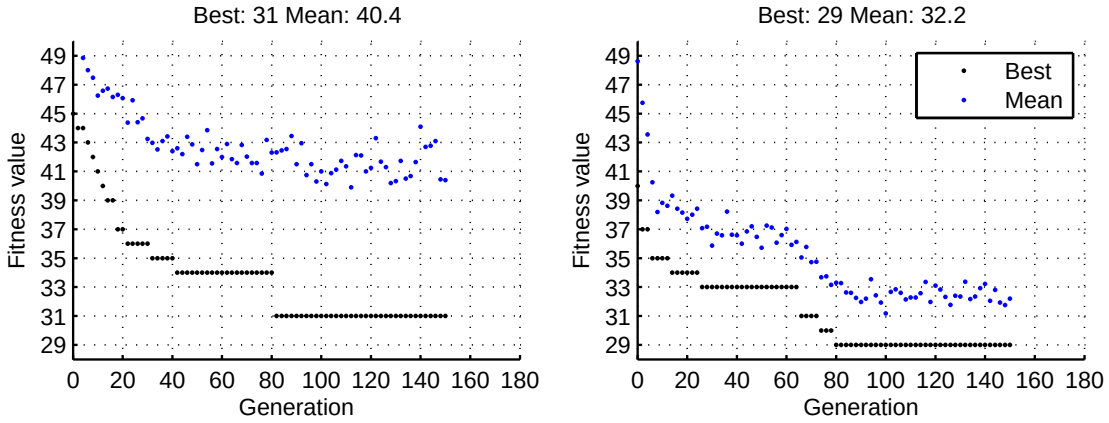
* شرح اختصارات را در بخش ۱۰۲ ببینید.

برخی گزینه‌ها در بین بهترین انتخاب‌های پارامتر در الگوریتم ژنتیکی GAhap، با فراوانی بیشتری مشاهده می‌شوند. این موضوع می‌تواند نشان از تاثیر معنادار نحوه‌ی انتخاب این پارامترها بر عملکرد الگوریتم GAhap باشد. به طور مشخص، استفاده از روش تصادفی یکنواخت برای انتخاب "کروموزم‌های" والد، گزینه‌ی نسبتاً مناسبی در کنار هر انتخاب دلخواه برای دیگر پارامترها است. کارایی عمومی این گزینه علاوه بر موارد فهرست شده در جدول ۱۰۳، در نتایج بدست آمده از سایر اجراهای الگوریتم GAhap نیز مشاهده می‌شود. مشابه همین ویژگی را ما در مورد پارامترهای cr_{int} و mr_{int} نیز مشاهده کرده‌ایم. در اینجا، انتخاب مقادیر بالا برای نرخ "کراس‌اوِر" و "جهش" در هاپلوتیپ‌های راهنما، عمدتاً به حصول نتایج نزدیکتر به جواب بهینه

کمک می‌کند.

نکته‌ی شایان توجه در ارتباط با کارآمدترین مقادیر پارامتر در الگوریتم ژنتیکی GAhap، کارآیی قابل قبول این الگوریتم به ازای مقادیر پائین cr است. همانطور که در جدول ۱۰۳ مشاهده می‌شود در بین بهترین تنظیمات پارامتر، انتخاب $cr = ۰/۲$ در کنار چند ترکیب مختلف از دیگر پارامترها می‌تواند در ٪۸۰ نمونه‌های مورد ارزیابی با "موفقیت" همراه باشد. مشاهده‌ی چنین حالتی در یک الگوریتم ژنتیک، معمولا می‌تواند نشان‌دهنده‌ی ضعف عملگر "کراس‌اوِر" برای تولید زادهایی باشد که می‌بایست به طور میانگین کارآمدتر از والدین خود باشند. به یاد آورید که در الگوریتم ژنتیک، "کروموزم‌ها" در هر نسل، با نسبت $cr : ۱ - cr$ با اعمال "کراس‌اوِر" و "جهش" بر روی "کروموزم‌ها" منتخب در نسل قبل بدست می‌آیند. از این رو، کارآیی قابل قبول یک الگوریتم ژنتیک به ازای مقادیر پائین نرخ نوترکیبی می‌تواند دال بر اهمیت عملگر دیگر، یعنی "جهش"، برای جستجو در فضای جواب باشد. یادآوری این نکته که تابع هدف در مسئله‌ی بیشترین پارسیمونی، دارای رفتاری ناپیوسته در فضای جواب است می‌تواند راهنمای مناسبی برای توضیح این ویژگی باشد. در واقع، می‌توان تصور کرد الگوریتم ژنتیکی ما، برای یافتن جواب بهینه به مراتب بیش از آنچه بر همگرائی سریع عده‌ای از "کروموزم‌های" برتر در یک کمینه‌ی موضعی متکی باشد نیازمند نمایندگان متعددی از نقاط پراکنده‌ی فضای جواب است. با این استدلال ممکن است این سئوال طرح شود که آیا اصولا عملگر "کراس‌اوِر" هیچ تاثیر معناداری بر روند یافتن جواب بهینه دارد یا نه. شکل ۱۰۳ روند تغییرات مقادیر تابع هدف را طی تکرارهای الگوریتم GAhap در حل یک نمونه‌ی واحد از مسئله‌ی تفکیک ژنوتیپ‌ها به ازای انتخاب مقادیر متفاوت برای cr نشان می‌دهد. همانطور که در شکل ۱۰۳ مشاهده می‌شود، عدم استفاده از عملگر "کراس‌اوِر" ($cr = ۰$) باعث می‌شود پراکندگی جمعیت "کروموزم‌ها" طی نسل‌های متوالی الگوریتم ژنتیک افزایش یابد و در واقع الگوریتم، یک جستجوی تصادفی را برای یافتن جواب بهینه در پیش گیرد. در این حالت، اطلاعات متناظر با بهترین جواب منحصراً از طریق معدودی از "کروموزوم‌های" نخبه به نسل‌های بعد منتقل می‌شود. در مقابل، استفاده از عملگر "کراس‌اوِر" حتی با نرخ پائین ($cr = ۰/۲$) باعث می‌شود جمعیت "کروموزم‌ها" پس از چند نسل، حول نقاط نزدیک به بهینه جمع شوند. در این حالت، ظهور یک "کروموزم" جدید با تعداد هاپلوتیپ‌های متمایز کمتر در بین سایر "کروموزم‌ها" باعث می‌شود مابقی جمعیت

نیز به سرعت به سمت این جواب حرکت کنند و بدین ترتیب همواره بخشی قابل ملاحظه‌ای از جمعیت "کروموزم‌ها"، اطلاعات متناظر با جواب تقریبی بهینه را در اشکال متنوع با خود حمل می‌کنند.



شکل ۱۰۳: روند همگرائی به جواب در الگوریتم GAhap به ازای مقادیر مختلف cr
روند تغییرات پائین‌ترین و میانگین تعداد هاپلوتیپ‌های متمایز در جمعیت "کروموزم‌های" جواب با پیشرفت نسل‌های جدید در الگوریتم GAhap بر روی نمونه‌ای از ۴۰ ژنوتیپ شبیه‌سازی شده از ۲۹ هاپلوتیپ متمایز. الگوریتم به ازای دو مقدار متفاوت برای نرخ "کراس‌اوِر"، cr اجرا شده است: $cr = ۰$ (چپ) و $cr = ۰/۲$ (راست).

ارزیابی کارایی GAhap برای حل مسئله‌ی بیشترین پارسیمونی بر روی نمونه‌های شبیه‌سازی شده

حال که گزینه‌های مناسب برای پارامترهای عمومی الگوریتم ژنتیک GAhap بدست آمده‌اند، به اختصار به مطالعه‌ی کارایی این الگوریتم در رسیدن به جواب‌های بهینه می‌پردازیم. این کار را با ادامه‌ی ارزیابی الگوریتم بر روی نمونه‌های شبیه‌سازی شده انجام می‌دهیم. سنجه‌ی مورد بررسی برای اندازه‌گیری کارایی الگوریتم، نسبت تعداد "موفقیت‌های" الگوریتم در رسیدن به جوابی با تعداد هاپلوتیپ‌های برابر یا کمتر از تعداد از پیش شناخته شده در نمونه‌ی مورد مطالعه است. در ادامه، کارایی و دقت الگوریتم GAhap، بر پایه‌ی نتایج بدست آمده از ارزیابی الگوریتم بر روی نمونه‌های شبیه‌سازی شده به ازای مقادیر مختلف حجم نمونه مورد بحث قرار می‌گیرند.

تغییرات کارایی و دقت الگوریتم با افزایش حجم نمونه‌ی مورد بررسی را می‌توان از دو جنبه‌ی متفاوت مورد ارزیابی قرار داد. در ارزیابی اول، تعداد ژنوتیپ‌های نمونه ثابت نگه داشته می‌شود و حجم نمونه با بالا بردن تعداد اسنیپ‌ها، افزایش داده می‌شود. در شیوه‌ای دیگر برای ارزیابی کارایی الگوریتم، برعکس، تعداد

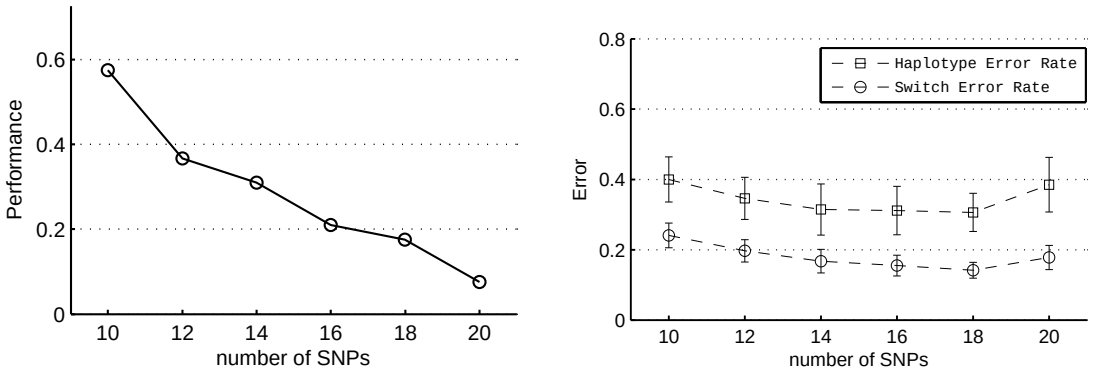
اسنیپ‌ها را ثابت نگه می‌داریم و حجم نمونه را با افزایش تعداد ژنوتیپ‌های نمونه افزایش می‌دهیم. در هر دو ارزیابی، ۴۰ نمونه‌ی تصادفی مستقل، هر یک شامل n ژنوتیپ بر روی l اسنیپ شبیه‌سازی می‌شوند. این کار به ازای مقادیر مختلف برای هر یک از دو مقدار n و l تکرار می‌گردد. به طور مشخص، ما ۲۴۰ نمونه‌ی مستقل از ۴۰ ژنوتیپ را به ترتیب بر روی ۱۰ تا ۲۰ اسنیپ، برای ارزیابی اول (تعداد ژنوتیپ‌ها ثابت) و ۱۶۰ نمونه‌ی مستقل دیگر را به ترتیب شامل ۲۰، ۴۰، ۶۰ و ۸۰ ژنوتیپ بر روی ۱۵ اسنیپ، برای ارزیابی دوم (تعداد اسنیپ‌ها ثابت) به کمک الگوریتم مولد نمونه‌های تصادفی بخش ۲۰۲ تولید کردیم.

شکل ۲۰۳ روند تغییرات کارایی و دقت الگوریتم را به ازای افزایش تعداد اسنیپ‌های نمونه‌ی مورد بررسی نشان می‌دهد. نمودار کارایی به طور کلی یک روند نزولی را در امتداد محور متناظر با تعداد اسنیپ‌ها نشان می‌دهد و به نظر می‌رسد با افزایش بیشتر تعداد اسنیپ‌ها، کارایی آن ممکن است به صفر برسد. افزایش نمائی حجم فضای جستجو با افزایش تعداد اسنیپ‌ها می‌تواند یک توضیح منطقی برای وجود این پدیده باشد. از آنجا که ما به طور یکسان و بدون رعایت ویژگی‌های خاص هر نمونه، به الگوریتم اجازه می‌دهیم تا جواب را با انجام حداکثر تعداد معینی از تکرارها بدست آورد، پیچیده‌تر شدن فضای جستجو سبب می‌شود الگوریتم نتواند در تعداد گام‌های محدود به اندازه‌ی کافی به جواب بهینه نزدیک گردد. البته، به یاد داشته باشید که اعمال این محدودیت کاملا ضروری است چون در غیر اینصورت، مثلا با انتخاب حداکثر تعداد تکرارهای الگوریتم بر حسب حجم فضای جستجو، الگوریتم در عمل زمان بسیار زیادی را برای رسیدن "احتمالی" به جواب بهینه صرف خواهد کرد.

استراتژی متعارف برای حل این مشکل، همانطور که به اشکال گوناگون در دیگر روش‌های تفکیک ژنوتیپ‌ها نیز بکار گرفته می‌شود، شکستن ژنوتیپ‌های داده شده در امتداد کروموزم به قطعات کوتاه، هر یک شامل تعداد معدود و معینی از اسنیپ‌ها، اجرا الگوریتم بر روی هر یک از این قطعات و سپس ترکیب نتایج است که انجام آن، فرصت تحقیقی دیگر را می‌طلبد.

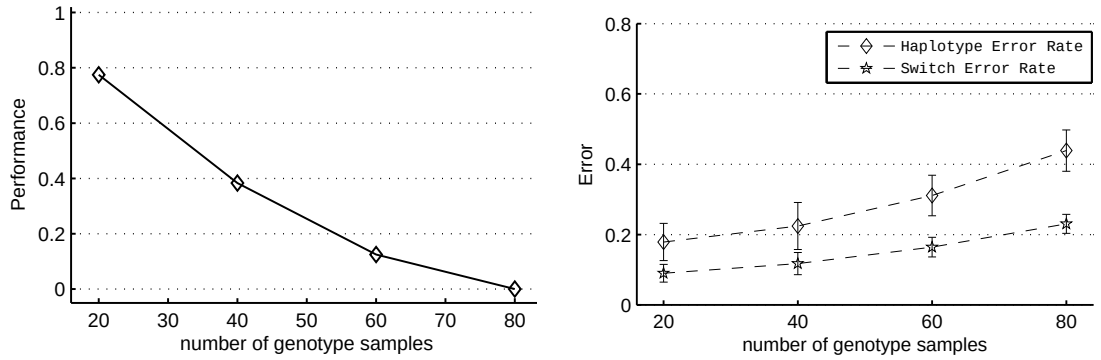
نمودار تغییرات دقت، روند متفاوتی را نشان می‌دهد(شکل ۲۰۳). افزایش نرخ خطا در انتهای نمودار، یعنی به ازای نمونه‌هایی با ۱۸–۲۰ اسنیپ، همنوا با کاهش کارایی الگوریتم، نشانه‌ی ناکافی بودن تعداد نقاط ارزیابی شده در فضای جواب‌های شدنی توسط الگوریتم است. اما کاهش میزان خطا در میانه‌ی نمودار، یعنی

به ازای نمونه‌هایی با ۱۴–۱۸ اسنیپ می‌تواند جالب توجه باشد؛ به خصوص آنکه الگوریتم در این محدوده تنها در ۲۰ تا ۳۰ درصد موارد توانسته است به کران بالای شناخته شده برای جواب بهینه دست یابد. هرچند این پدیده می‌تواند زائیده‌ی ساختار درونی نمونه‌های شبیه‌سازی شده و ارتباط آنها با مقادیر خطا باشد اما بی‌شک نشانه‌ای از این واقعیت است که قید بیشترین پارسیمونی لزوماً نمی‌تواند در تناظر با درست‌ترین جواب باشد.



شکل ۲۰۳: نمودارهای دقت و کارایی الگوریتم GAhap بر حسب تعداد اسنیپ‌های نمونه
ارتفاع نقاط در نمودار سمت چپ، نشان‌دهنده‌ی نسبت تعداد مواردی از ۴۰ اجرای مستقل الگوریتم بر روی نمونه‌های تصادفی است که در آنها الگوریتم توانسته است نتیجه‌ای با تعداد هاپلوتیپ‌های متمایز برابر یا کمتر از کران بالای شناخته شده بدست آورد. در نمودار سمت راست، ارتفاع نقاط نشان‌دهنده‌ی میانگین نرخ خطا در اجرای الگوریتم بر روی ۴۰ نمونه‌ی شبیه‌سازی شده‌ی مستقل است. بازه‌ی مشخص شده در اطراف هر نقطه، اندازه‌ی خطای استاندارد میانگین محاسبه شده را نشان می‌دهد. هر یک از نمونه‌های از ۴۰ ژنوتیپ شبیه‌سازی شده تشکیل شده‌اند.

رفتار الگوریتم به ازای تغییرات حجم نمونه بر حسب تعداد ژنوتیپ‌ها، روند ساده‌تری دارد(شکل ۳۰۳). افزایش تعداد ژنوتیپ‌های نمونه، با کاهش همزمان کارایی و دقت الگوریتم همراه است. البته در اینجا نیز حتی زمانی که الگوریتم به مرز هیچ یک از کران‌های شناخته شده برای تعداد هاپلوتیپ‌های متمایز تشکیل‌دهنده‌ی ژنوتیپ‌های نمونه نرسیده است، یعنی در $n = ۸۰$، جواب‌های بدست آمده به طور میانگین بیش از ۵۰ درصد هاپلوتیپ‌های تشکیل‌دهنده‌ی ژنوتیپ‌های نمونه را به درستی شناسائی می‌کنند. نتایج در مجموع بیانگر آن است که استفاده از الگوریتم ژنتیک برای استنباط هاپلوتیپ‌ها در نمونه‌هایی با حجم بالای داده‌های ژنوتیپ، بر پایه‌ی مدل بیشترین پارسیمونی و از طریق الگوریتم GAhap چندان قابل اطمینان نیست. البته، خطای میانگین در نتایج بدست آمده برای نمونه‌هایی با ۲۰–۴۰ ژنوتیپ به مقادیر خطا در دیگر الگوریتم‌های رایج در تفکیک ژنوتیپ‌ها به ازای نمونه‌های در همین اندازه، نزدیک است.



شکل ۳۰۳: نمودارهای دقت و کارایی الگوریتم GAhap بر حسب تعداد ژنوتیپ‌های نمونه
شرح نمودارها همانند شکل ۲۰۳ است. تمامی ژنوتیپ‌های مورد ارزیابی در اینجا بر روی ۱۵ اسنیپ تعریف شده‌اند.

دقت الگوریتم GAhap در استنباط هاپلوتیپ‌های یک مجموعه از ژنوتیپ‌های واقعی و مقایسه‌ی آن با دیگر روش‌ها

در ادامه به مقایسه‌ی نتایج حاصل از اجرای الگوریتم GAhap و تعدادی دیگر از الگوریتم‌های رایج در تفکیک ژنوتیپ‌ها، بر روی ژنوتیپ‌های مجموعه‌ی هورَن می‌پردازیم. همانطور که در بخش ۱۰۲ شرح داده شد داده‌های را که ما از مجموعه‌ی هورَن برای بررسی کارایی الگوریتم‌های تعیین فاز مورد استفاده قرار می‌دهیم ژنوتیپ‌های ۱۵۰ فرد نمونه در ناحیه‌ی ژنی GH1 به طول ۱۵ اسنیپ است.

جدول ۲۰۳ نرخ خطای تشخیص هاپلوتیپ‌ها و خطای جابجائی فاز و نیز تعداد هاپلوتیپ‌های متمایز استنباط شده در ژنوتیپ‌های مجموعه‌ی هورَن را به ازای الگوریتم‌های مختلف تعیین فاز نشان می‌دهد. مقایسه‌ی تعداد هاپلوتیپ‌های متمایز استنباط شده توسط GAhap و دیگر روش‌ها گویای آن است که این الگوریتم در رسیدن به کمترین تعداد هاپلوتیپ متمایز ناکام بوده است هرچند فاصله‌ی آن با کوچکترین مورد، یعنی ۳۲ هاپلوتیپ استنباط شده توسط PHASE چندان زیاد نیست. نکته‌ی جالب توجه آن است که هیچ یک از دیگر الگوریتم‌ها، مدل بیشترین پارسیمونی را به عنوان رویکرد مورد استفاده در استنباط هاپلوتیپ‌ها مورد توجه قرار نمی‌دهند اما ظاهرا، نتایج بسیار نزدیکی به جواب بهینه در مدل بیشترین پارسیمونی بدست می‌آورند. به جز روش 2SNP، تعداد هاپلوتیپ‌های متمایزی که توسط دیگر روش‌ها بدست آمده است بین ۳۲ تا ۳۵ هاپلوتیپ است حال آنکه در واقعیت، ۳۶ هاپلوتیپ ترکیب ژنوتیپ‌های مورد مطالعه را تشکیل داده‌اند. این واقعیت مؤید آن است که طبیعت لزوماً بر پایه‌ی مدل بیشترین پارسیمونی رفتار نمی‌کند. خطای

روش GAhap در تشخیص هاپلوتیپ‌ها و نیز خطای آن در جابجائی فازها در مقایسه با دیگر روش‌ها چندان دلگرم کننده نیست که البته، عدم توافق هاپلوتیپ‌های واقعی با مدل بیشترین پارسیمونی می‌تواند تا اندازه‌ای وجود چنین خطایی را توجیه نماید. با این حال، اشاره به این نکته ضروری است که الگوریتم GAhap به دلیل تصادفی بودن محاسباتی که در آن انجام می‌شود و نیز پیچیدگی فضای جستجو در مدل بیشترین پارسیمونی به سادگی می‌تواند جواب‌هایی از نظر تعداد هاپلوتیپ‌های متمایز نزدیک به بهینه اما از نظر ترکیب آلل‌ها کاملا متفاوت با آن را بدست آورد.

جدول ۲۰۳: خطای استنباط و تعداد هاپلوتیپ‌های متمایز در نتایج بدست آمده از اجرای الگوریتم‌های مختلف بر روی ژنوتیپ‌های مجموعه‌ی هورَن

روش	رویکرد اصلی مورد استفاده برای حل مسئله	$\|\mathcal{H}\|$ [a]	$e_{haplotype}$ [b*]	e_{switch} [c*]
HAPLOTYPER	استنباط بیزی برپایه‌ی پیشین دیریکله[۷۶]	۳۳	۵/۴	۳/۰
PHASE	استنباط بیزی برپایه‌ی مدل فیلوژنی کامل[۷۲]	۳۲	۵/۶	۳/۱
fastPHASE	تقریب و ساده‌سازی روش PHASE برای افزایش سرعت[۷۳]	۳۵	۷/۳	۴/۵
2SNP	تعیین فاز جفت اسنیپ‌ها و درخت فراگیر کمینه	۴۰	۱۰/۴	۵/۶
GAhap	بیشترین پارسیمونی و الگوریتم ژنتیک (بخش ۱۰۲)	۳۴	۹/۷	۵/۷

[a] تعداد هاپلوتیپ‌های استنباط شده‌ی متمایز ـ [b] خطای شناسائی هاپلوتیپ‌ها ـ [c] خطای جابجائی فازها ـ * مقادیر خطا بر حسب "درصد" هستند.

۲۰۳ نمونه‌های از افرازهای بلوکی در ناحیه‌های ENCODE

از این بخش تا انتهای فصل، به بحث بر روی نتایج بدست آمده از روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها می‌پردازیم و آنها را از جنبه‌های متفاوت با روش پیشنهادی در این رساله، یعنی «افراز بلوکی سراسری برای

بیشترین جفت اسنیپ‌های همبسته» (GPMAP) مقایسه می‌کنیم. شرح کاملی از جزئیات روش GPMAP در بخش ۴۰۲ و مشخصات سایر روش‌های تحت مقایسه، در بخش ۵۰۲ (جدول ۱۰۲) آمده‌اند.

اجرای هر یک از الگوریتم‌های افراز بلوکی هاپلوتیپ‌ها بر روی نمونه هاپلوتیپ‌های HapMap از پانل CEU، در ده ناحیه‌ی ENCODE، زمانی معادل مقادیر مندرج در جدول ۳۰۳ را صرف کرد. همانطور که ملاحظه می‌کنید، زمان صرف شده برای اجرای الگوریتم MDL بر روی ده ناحیه‌ی ۵۰۰ کیلوبازی، هر کدام شامل ۴۰۰ اسنیپ، بیشتر از آن است که بتوان این روش را به سادگی بر روی داده‌هایی در مقیاس ژنوم اجرا کرد. طول زمان اجرا در روش HB، هر چند در مقایسه با روش MDL، امیدوار کننده است اما برای اجرای آن نیز در مقیاس ژنومی باید اندکی تأمل کرد. پیاده‌سازی ما از الگوریتم‌های GPG و GPF در نرم‌افزار Haploview، باعث شده‌است زمان اجرای این دو روش بسیار نزدیک به زمان اجرای دو روش دیگر پیاده‌سازی شده در این نرم‌افزار، یعنی GAB و GAM باشد[۱]. توسط همین نرم‌افزار می‌توان ساختار بلوکی کل ژنوم انسان را با استفاده از هاپلوتیپ‌های HapMap در یک زیرجمعیت معین، در مدت زمانی کمتر از ۲۰ ساعت بدست آورد.

جدول ۳۰۳: طول زمان اجرا در روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها

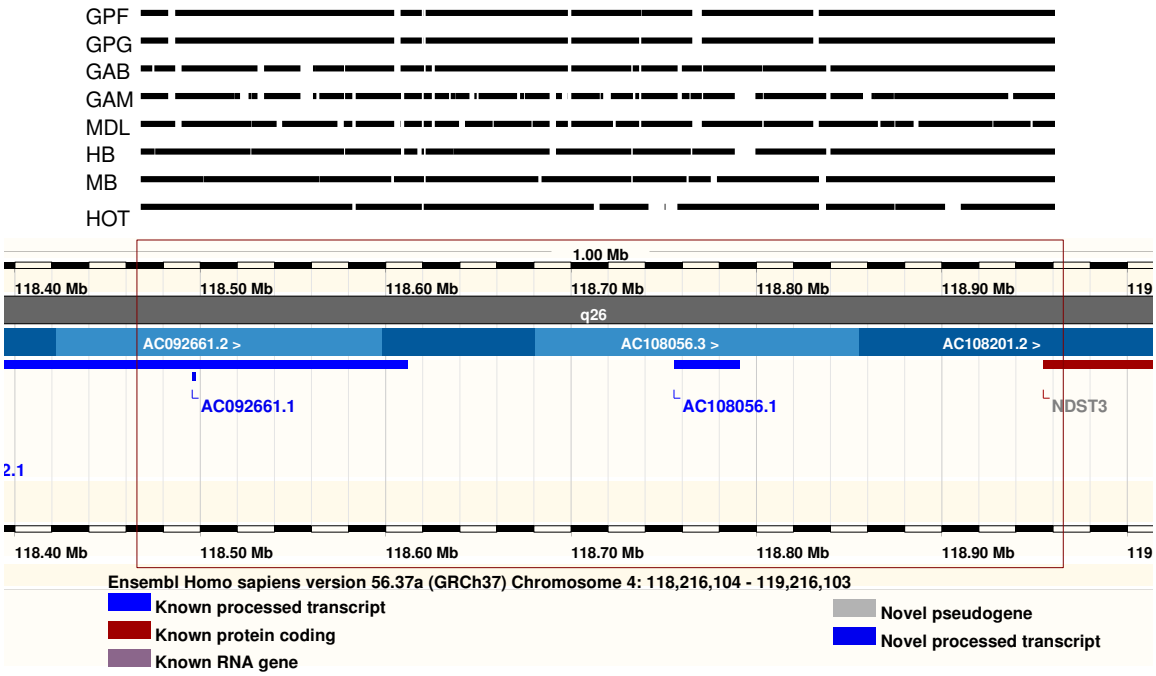
روش افراز	MB	HB	MDL	GAM	GAB	GPG	GPF
مجموع زمان اجرا (ثانیه)	۲۲	۷۴۳	۳۲۹۵	۱۱۲	۱۴۳	۱۳۱	۱۲۸

پیش از بررسی جنبه‌های مختلف افرازهای بلوکی به وسیله‌ی کمیت‌های عددی، برای کسب درک بهتری از ساختارهای بلوکی در ژنوم، تصویری از افرازهای بلوکی بدست آمده توسط روش‌های مختلف را به طور نمونه در سه ناحیه از نواحی ENCODE نمایش می‌دهیم. شکل‌های ۴۰۳، ۵۰۳ و ۶۰۳ ساختارهای بلوکی بدست آمده توسط روشهای مختلف جدول ۱۰۲ را به ترتیب در نواحی 4q26، 7p15.2 و 2q37.1، در کنار نقشه‌ای از موقعیت ژن‌های موجود در آن نواحی، نشان می‌دهند. نقشه‌ی موقعیت‌های ژنی، از مجموعه‌ی اطلاعات ژنی ENSEMBL[۲] گرفته شده است.

افرازهای بلوکی در هر سه شکل، بر روی هاپلوتیپ‌های HapMap در پانل CEU تعریف شده‌اند. توجه

[۱] بیش از ۹۵ درصد این زمان، صرف خواندن اطلاعات ژنوتیپ‌ها، تخصیص حافظه برای اطلاعات مربوط به جفت اسنیپ‌ها و اجرای الگوریتم EM برای تفکیک ژنوتیپ‌ها می‌گردد.

[۲] http://www.ensembl.org/Homo_sapiens

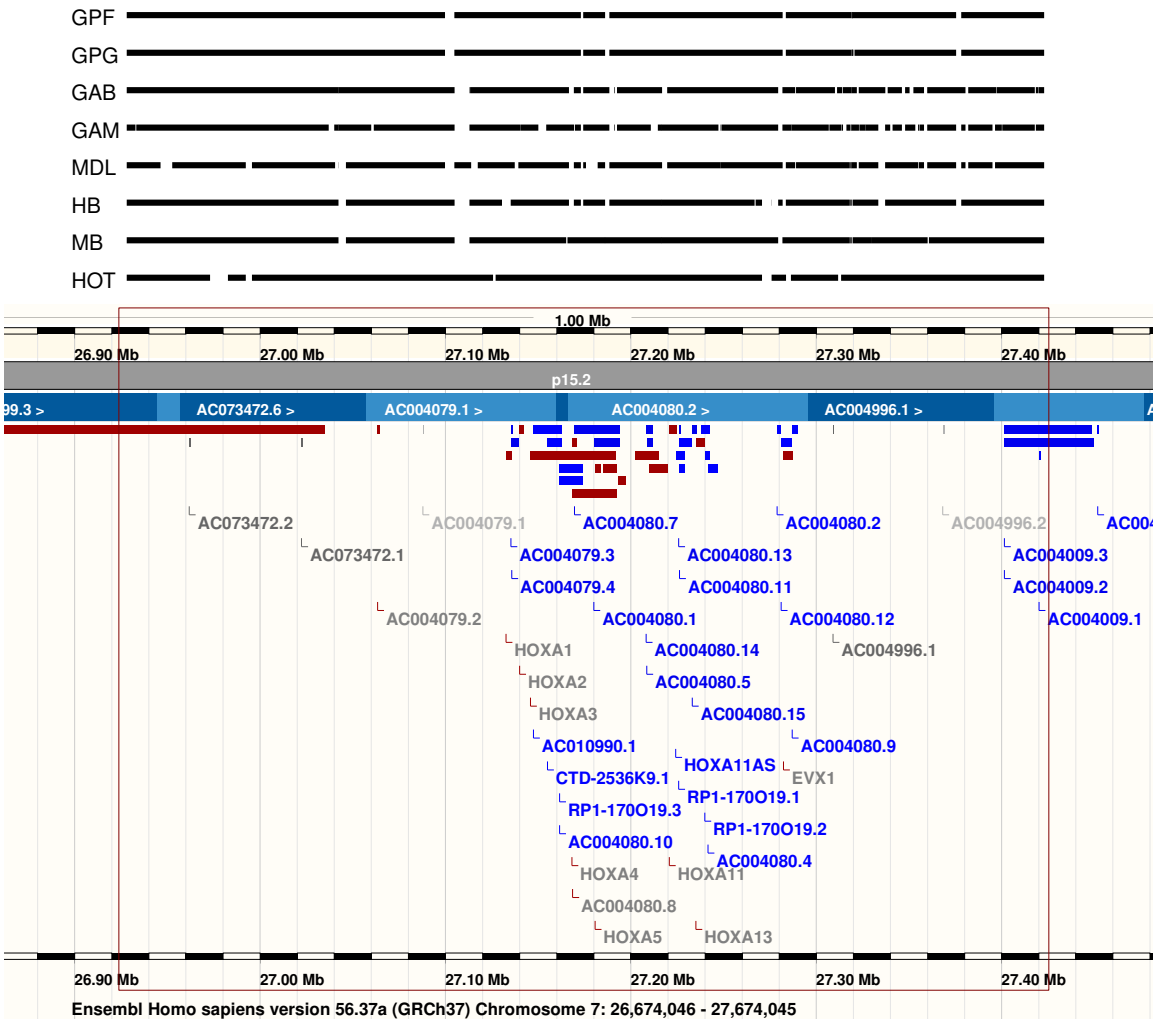


شکل ۴۰۳: افرازهای بلوکی مختلف در ناحیه‌ی 4q26 (ENr113)
نقشه‌ی موقعیت‌های ژنی برگرفته از ENSEMBL است.

کنید که وسعت شکاف‌ها بین بلوک‌های متوالی، تنها نشاندهنده‌ی فاصله‌ی فیزیکی بین دو اسنیپ، یکی در انتهای بلوک سمت چپ و دیگری در ابتدای بلوک سمت راست است و از این رو در بین افرازهای سراسری مثل HB و GPF نیز دیده می‌شوند.

با مقایسه‌ی افرازهای بلوکی مختلف در این اشکال می‌توان به برخی نتایج ساده رسید؛ ۱) روش‌های مختلف، نواحی متفاوتی را به عنوان بلوک شناسائی می‌کنند و از این رو نمی‌توان از یک افراز بلوکی استاندارد سخن گفت، ۲) با این حال، برخی موقعیت‌ها بر روی ژنوم در بین تمام افرازها بر روی مرز بلوک‌ها قرار گرفته‌اند. ۳) طول بلوک‌ها در افرازهای بدست آمده از روش‌های افراز سراسری بلوک‌ها، یعنی MB، HB، GPG و GPF، به طور نسبی بیشتر از طول بلوک‌ها در افرازهای بدست آمده از روش‌های افراز موضعی بلوک‌ها است. ۴) شباهت بین افرازهای بدست آمده از دو روش GPG و GPF بیشتر از شباهت بین دیگر روش‌ها است. به یاد آورید که این دو روش، گونه‌های مختلفی از الگوریتم GPMAP هستند که در هر یک از آنها، شاخص متفاوتی برای تعیین همبستگی در جفت اسنیپ‌ها، مورد استفاده قرار می‌گیرد(بخش ۴۰۲).

به سختی می‌توان ارتباط روشنی بین موقعیت ژنها و موقعیت بلوک‌های هاپلوتیپ در نتایج بدست آمده در



شکل ۵۰۳: افرازهای بلوکی مختلف در ناحیه‌ی 7p15.2 (ENm010)

این سه ناحیه‌ی ژنومی مشاهده کرد. بنابر ملاحظات عملی و کاربردی، اسنیپ‌های بیشتری در نزدیکی و درون نواحی رمزگردان ژنی شناسائی شده‌اند. از این رو، نمونه‌های هاپلوتیپ در این نواحی، ماهیتاً کمی سوگیری شده هستند. چگالی بالاتر اسنیپ‌ها در نواحی ژنی و اطراف آن می‌تواند سبب افراز هاپلوتیپ‌ها به بلوک‌های کوچکتر در این نواحی شود. این موضوع، تاثیر شدیدتری بر روش‌های موضعی افراز بلوکی هاپلوتیپ‌ها دارد. با این وجود، نقاط انفصال بین برخی بلوک‌های مجاور یکدیگر که به طور مشترک در نتایج بدست آمده از بیشتر روش‌های افراز مشاهده می‌شوند می‌بایست مورد توجه قرار گیرند. از جمله، انفصال بلوک‌ها در فاصله‌ی حدود ۱۰۰ کیلوبازی، بالادست خوشه‌ی ژنی HOXA1-HOXA5 و نیز انفصال بین بلوک‌ها در دو طرف خوشه‌ی ژنی HOXA11-EVX1 در ناحیه‌ی 7p15.2 (شکل ۵۰۳). این نقاط انفصال می‌توانند نشانه‌ی نقاط

شکل ۶۰۳: افرازهای بلوکی مختلف در ناحیه‌ی 2q37.1 (ENr131)

پراحتمال نوترکیبی و یا تاثیر سازوکارهای جمعیتی مثل مهاجرت و رانش ژنی باشند. از سوی دیگر، نقاط انفصال مورد توافق بین مدل‌های متفاوت افراز، لزوما در خارج ناحیه‌ی رمزگردان قرار نمی‌گیرند. به عنوان مثال، چنین نقطه‌ای را می‌توان در میانه‌ی ژن DGKD در ناحیه‌ی 2q37.1 مشاهده کرد (شکل ۶۰۳).

۳۰۳ تنوع هاپلوتیپ‌ها در بلوک‌ها

در این بخش، نتایج حاصل از اجرای روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها بر روی برخی نمونه‌های واقعی از هاپلوتیپ‌ها را مورد بحث قرار می‌دهیم. در جدول ۴۰۳ خلاصه‌ای از جنبه‌های متفاوت یک افراز بلوکی را به

ازای روش‌های مختلف، ملاحظه می‌کنید. هر یک از مقادیر نمایش داده شده در جدول ۴۰۳، میانگین کمیت متناظر در بلوک‌های بدست آمده از اجرای هر روش بر روی داده‌های HapMap در ده ناحیه‌ی ENCODE است. میانگین طول بلوک‌ها بر حسب تعداد نوکلئوتید در بین روش‌های مختلف، از ۱۳/۳ تا ۴۶/۸ کیلوباز متغیر است. میانگین فاصله‌ی بین نقاط پراحتمال نوترکیبی، یعنی طول بلوک‌ها در روش HOT، حدود ۶۹ کیلوباز است که تفاوتی چشمگیر با طول بلوک‌ها در دیگر روش‌ها دارد. این موضوع می‌تواند نشاندهنده‌ی آن باشد که قیود رایج برای تعریف بلوک‌ها، یعنی افزایش تنوع هاپلوتیپ‌ها یا کاهش میزان LD بین اسنیپ‌ها لزوما نمی‌تواند نشانه‌ای دال بر وقوع نوترکیبی با نرخ بالا باشد و در واقع نباید نقش عوامل دیگری چون رانش ژنی را در تشکیل الگوهای بلوکی در ژنوم نادیده گرفت.

جدول ۴۰۳: مشخصات بلوک‌های هاپلوتیپی در نواحی ENCODE به ازای روش‌های مختلف

روش افراز بلوکی	HOT	MB	HB	MDL	GAM	GAB	GPG	GPF
متوسط طول بلوک برحسب kb	۶۸/۹	۴۶/۸	۳۶/۲	۱۷/۱	۱۳/۳	۲۳/۳	۳۵/۷	۳۹/۷
متوسط طول بلوک برحسب اسنیپ	۵۲	۳۶	۲۷	۱۳	۱۰	۱۸	۲۷	۳۰
پوشش هاپلوتیپ‌های رایج	۰/۶۷	۰/۸۹	۰/۹۱	۰/۹۶	۰/۹۶	۰/۹۳	۰/۸۸	۰/۸۷
فراوانی "حفره‌ها"	۰/۵۰	۰/۲۳	۰/۱۴	۰/۰۶	۰/۰۴	۰/۰۴	۰/۱۵	۰/۱۸
فراوانی "جزیره‌ها"	۰/۰۸	۰/۱۰	۰/۰۹	۰/۱۷	۰/۱۹	۰/۱۱	۰/۰۶	۰/۰۷

همانطور که در بخش قبل، با نمایش افرازهای بلوکی مختلف در کنار یکدیگر مشاهده گردید، طول بلوک‌ها در افرازهای سراسری، مثل MB و GPG به طور میانگین بیشتر از طول بلوک‌ها در افرازهای موضعی، مثلا GAB و GAM است. این تفاوت بر حسب متوسط تعداد اسنیپ‌های درون بلوک‌ها نیز مؤید همین رابطه است.

پوشش هاپلوتیپ‌های رایج، همانطور که در بخش ۲۰۵۰۲ توضیح دادیم، نشاندهنده‌ی نسبتی از هاپلوتیپ‌های نمونه است که در خوشه‌هایی با بیش از شش عضو قرار می‌گیرند. همانطور که در جدول ۴۰۳ مشاهده می‌شود، پوشش هاپلوتیپ‌های رایج در تمام روش‌ها، البته به جز روش HOT، بالاتر از کران متعارف برای این کمیت، یعنی ۰/۸۰ است. به یاد آورید که در روش‌های مبتنی بر واگرائی هاپلوتیپ‌ها، این کران به عنوان قیدی برای تعیین بلوک‌های محتمل استفاده می‌شود. نکته جالب توجه این است که پوشش هاپلوتیپ‌های رایج، در

روش‌هایی که قید واگرائی را به طور صریح برای تعریف بلوک‌ها بکار نمی‌برند، در عمل، همان شرایط را صدق می‌دهد. در روش‌های پیشنهادی ما، یعنی GPG و GPF نیز علیرغم اینکه ژنوم با بلوک‌هایی به نسبت عریض‌تر پوشش داده می‌شود اما همچنان هاپلوتیپ‌های رایج، بخش عمده‌ای از تنوع هاپلوتیپ‌های درون بلوک‌ها را در بر می‌گیرند.

نکته‌ی جالب توجه دیگر، در رابطه با متوسط تعداد اسنیپ‌ها در هر بلوک است. همانطور که در جدول ۴۰۳ ملاحظه می‌کنید در کوچکترین بلوک‌ها، یعنی بلوک‌های افراز بدست آمده از روش آزمون چهار گامتی، GAM، ده اسنیپ به طور میانگین در هر بلوک قرار می‌گیرند. این تعداد و تعداد اسنیپ‌های بیشتر در بلوک‌های دیگر افرازها، به طور کلی نشاندهنده‌ی آن است که میزان LD بین اسنیپ‌ها، دست کم در بخش‌های خاصی از نواحی ENCODE به میزان قابل توجهی بالا است.

سازگاری بین بلوک‌ها و الگوی تغییرات LD را می‌توان با بررسی میزان فراوانی "حفره‌ها" و "جزیره‌ها" در افرازهای بدست آمده از روش‌های مختلف مورد مطالعه قرار داد. همانطور که در جدول ۴۰۳ ملاحظه می‌کنید، افرازهایی با بلوک‌های عریض‌تر به طور میانگین "حفره‌های" بیشتری را شامل می‌شوند و افرازهایی با بلوک‌های کوچکتر "جزیره های" بیشتری از اسنیپ‌های همبسته را از دست می‌دهند. بر این اساس، می‌توان گفت افرازهای بلوکی GPG، GPF و HB، در مواردی که می‌خواهیم بیشترین اطلاعات از الگوی LD بر روی ژنوم را در بلوک‌ها حفظ کنیم قابل اعتمادترند و در مقابل، روش‌های GAB، GAM و MDL به ترتیب، یکپارچه‌ترین بلوک‌ها را تولید می‌کنند؛ به این معنی که اندازه‌ی LD تقریبا بین تمام اسنیپ‌های درون بلوک‌های تعریف شده توسط این روش‌ها بالا است.

۴۰۳ تعداد و پوشش htSNPها در بلوک‌ها

پس از تعیین بلوک‌های هاپلوتیپ توسط یک روش افراز بلوکی برای هر یک از ده ناحیه‌ی ENCODE، کمترین تعداد تگ‌اسنیپ‌های لازم برای بازسازی تنوع هاپلوتیپ‌های درون هر یک از بلوک‌ها را با استفاده از نرم‌افزار htSNPer بدست می‌آوریم. جدول ۵۰۳ تعداد htSNPهای لازم برای بازسازی هاپلوتیپ‌های CEU را به تفکیک، به ازای هر روش افراز بلوکی و هر ناحیه‌ی ENCODE نشان می‌دهد. به یاد داشته باشید که نحوه‌ی

تعیین بلوک‌های هاپلوتیپ بر میزان واگرائی هاپلوتیپ‌های درون آن و به تبع آن بر تعداد htSNPها تاثیرگذار است.

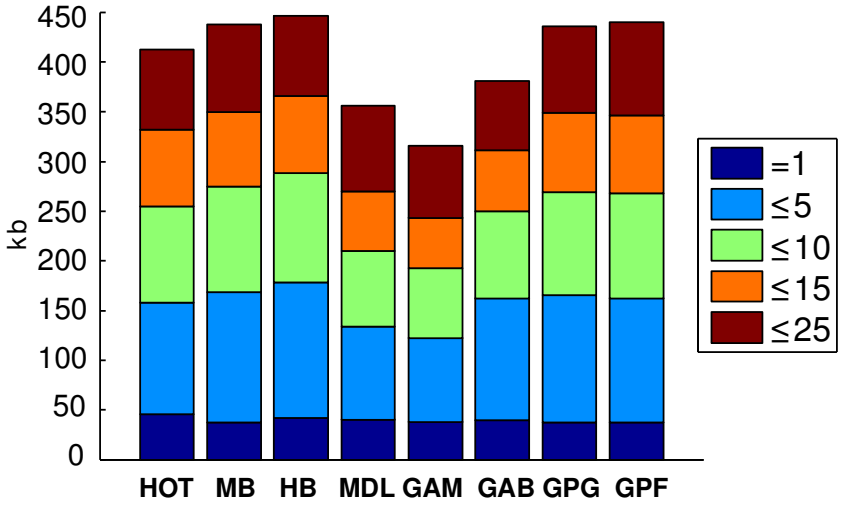
جدول ۵۰۳: تعداد htSNPها برای هر یک از نواحی ENCODE به ازای افرازهای بلوکی مختلف

اختصار ENCODE	منطقه‌ی کروموزومی	HOT	MB	HB	MDL	GAM	GAB	GPG	GPF
ENr112	2p16.3	۳۷	۳۵	۳۳	۵۱	۹۶	۷۷	۳۲	۳۳
ENr131	2q37.1	۳۲	۴۸	۴۰	۶۰	۱۰۱	۷۹	۴۷	۴۲
ENr113	4q26	۳۴	۳۷	۳۰	۴۶	۶۷	۴۷	۳۱	۳۱
ENm010	7p15.2	۳۶	۳۷	۳۴	۴۹	۸۷	۷۸	۳۷	۳۶
ENm013	7q21.13	۱۷	۱۶	۱۵	۳۴	۶۹	۲۹	۲۳	۲۵
ENm014	7q31.33	۳۸	۲۷	۲۵	۴۸	۶۷	۴۷	۲۷	۲۷
ENr321	8q24.11	۲۷	۳۵	۲۶	۴۸	۶۳	۴۹	۳۱	۳۰
ENr232	9q34.11	۵۱	۴۷	۴۲	۶۳	۷۰	۷۵	۴۹	۵۲
ENr123	12q12	۱۴	۳۳	۲۹	۴۸	۷۹	۵۹	۳۷	۳۸
ENr212	18q12.1	۳۱	۳۶	۳۰	۵۲	۶۹	۴۶	۲۸	۳۳

نکته‌ی قابل تأمل در ارتباط با نتایج نشان داده شده در جدول ۵۰۳، این است که جایگزین کردن قید واگرائی با برخی معیارهای دیگر از جمله، همبستگی بین جفت اسنیپ‌ها، می‌تواند بلوک‌هایی را تعیین کند که تعداد htSNPهای لازم در آنها همچنان اندک باشد. مقایسه‌ی تعداد htSNPهای لازم برای بلوک‌های هر یک از افرازهای GPG و GPF و تعداد htSNPهای لازم برای بلوک‌های بدست آمده از روش HB، مؤید همین موضوع است. به یاد داشته باشید که روش HB به طور خاص، روش تعیین بهترین افراز بلوکی برای بدست آوردن کمترین تعداد تگ‌اسنیپ‌های لازم برای کل ناحیه‌ی مورد بررسی است. از این رو، نزدیک بودن مقادیر بین تعداد htSNPهای افرازهای بلوکی GPF و GPG و افراز بلوکی HB می‌تواند نشان از کارآیی قابل توجه روش GPMAP برای انتخاب کوچکترین مجموعه‌ی ممکن از htSNPها باشد. نزدیکی نتایج، بین HB و MB در جدول ۵۰۳ نیز نکته‌ی جالب توجه دیگری است. همانطور که در بخش ۲۰۳ (جدول ۳۰۳) ملاحظه کردید زمان اجرای MB در مقایسه با HB بسیار کوتاه‌تر است و اصولا، تابع هدف در روش MB در مقایسه با تابع هدف روش HB، به مراتب از پیچیدگی کمتری برخوردار است اما در نهایت، از حیث تعداد htSNPها، نتایج کاملا نزدیکی به روش بهینه در این مسئله، یعنی HB بدست می‌آورد.

در ادامه، برای داشتن قضاوتی دقیق‌تر درباره‌ی htSNPها، کارایی آنها را از دو جنبه‌ی دیگر مورد بررسی قرار می‌دهیم که عبارتند از: پوشش htSNPها و توان بازسازی هاپلوتیپ‌های نمونه. شکل ۷۰۳ پوشش

htSNPها را بر اساس میانگین بدست آمده از این کمیت بر روی ده ناحیه‌ی ENCODE را نمایش می‌دهد. همانطور که در بخش ۳۰۵۰۲ توضیح دادیم، با اضافه شدن بلوک‌هایی که برای بازسازی تنوع هاپلوتیپ‌های درون آنها به تعداد بیشتری htSNP نیاز است به تدریج پوشش کاملی از ناحیه‌ی کروموزمی مورد مطالعه بدست می‌آید. همانطور که در شکل ۷۰۳ ملاحظه می‌کنید، بیشترین پوشش در تمامی روش‌ها توسط مناطقی بدست می‌آید که بین ۲ تا ۵ htSNP برای بازسازی تنوع هاپلوتیپ‌ها در آنها کافی است. به طور کلی، هر اندازه پوشش htSNPها وسیع‌تر باشد امکان طراحی شیوه‌های کم هزینه‌تری برای خواندن ژنوتیپ‌ها فراهم می‌شود. در اینجا، ملاحظه می‌کنید که HB بهترین پوشش htSNP را بین دیگر روش‌ها دارد که البته موضوعی دور از انتظار نیست چون این الگوریتم به طور اختصاصی به منظور بهینه‌سازی همین کمیت، طراحی شده است.



شکل ۷۰۳: پوشش htSNPها به ازای افرازهای بلوکی متفاوت برای نواحی ENCODE
ارتفاع هر بخش از یک ستون، نشاندهنده‌ی مجموع طول نواحیی از ژنوم است که تنوع هاپلوتیپ‌ها در هر یک از آنها را می‌توان با حداکثر k تگ‌اسنیپ پوشش داد. نتایج با محاسبه‌ی میانگین این کمیت بر روی ده ناحیه‌ی ENCODE بدست آمده است. یادآور می‌شود، عرض هر یک از نواحی ENCODE، $۵۰۰kb$ است.

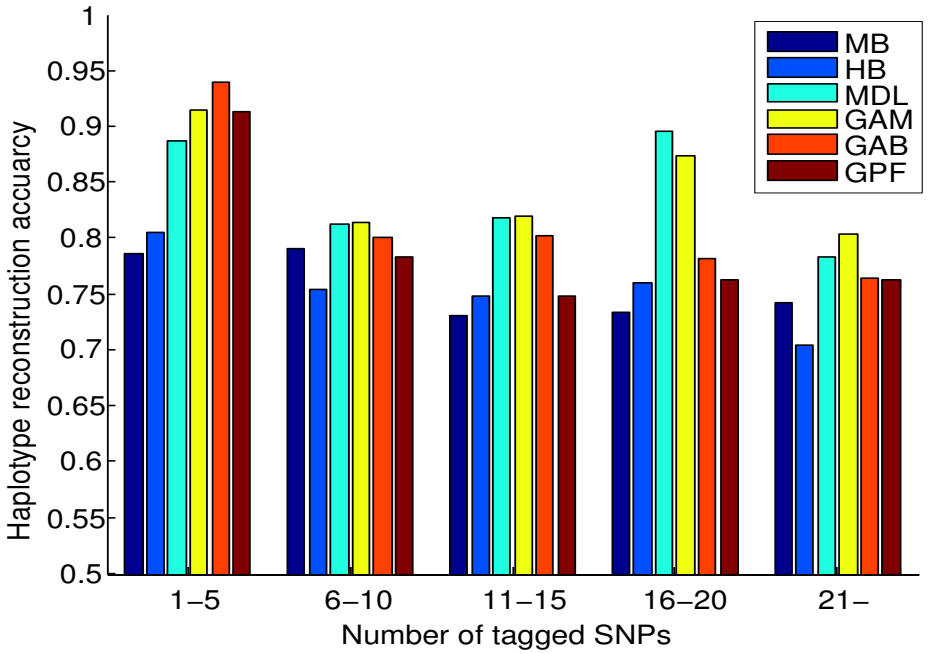
همانطور که در شکل ۷۰۳، در بالاترین سطح از پوشش htSNPها در روش‌های GAM، MDL و GAB ملاحظه می‌کنید، این روش‌ها تنوانسته‌اند به همان میزان پوششی که دیگر روش‌ها با ۲۵ htSNP یا کمتر بدست آورده‌اند، نزدیک شوند و از این به نظر می‌رسد برای پوشش کامل ناحیه‌ی مورد مطالعه بر روی ژنوم به htSNPهای بیشتری نیاز دارند. این نکته نیز جالب توجه است که میزان پوشش htSNPها در اولین

سطح، یعنی مجموع طول ژنومی که با یک htSNP پوشش داده می‌شود در بین روش‌های مختلف چندان متفاوت نیست و تفاوت در واقع، با افزایش تعداد htSNPها بیشتر نمایان می‌شود. به طور خلاصه می‌توان نتیجه گرفت، روش‌های GAM و MDL تعداد htSNPهای بیشتری برای پوشش کامل ناحیه‌ی کروموزمی مورد مطالعه نیاز دارند، روش GAB در بین روش‌های موضعی افراز بلوکی، بیشترین پوشش را حتی با تعداد پائین htSNPها بدست می‌آورد و کارایی بلوک‌های MB در مقایسه با بلوک‌های روش "بهینه"، یعنی HB، از نظر پوشش htSNPها شایان توجه است چون همانطور که در بحث راجع به نتایج جدول ۵۰۳ اشاره شد، تعریف بلوک‌ها در این روش با مدل به مراتب ساده‌تری در مقایسه با HB صورت می‌گیرد.

مسئله‌ای که در بسیاری از تحقیقات، در زمینه‌ی استفاده از htSNP موضوع بحث قرار می‌گیرد، این نکته است که تا چه اندازه می‌توان برای تعیین هاپلوتیپ یک نمونه‌ی جدید، به اطلاعاتی که توسط مجموعه‌ی htSNPهای منتخب حمل می‌شوند اتکا کرد[۱۶۳]. برای بررسی این موضوع، در اینجا ما کارایی htSNPهای بدست آمده برای افرازهای بلوکی مورد مطالعه در این رساله را در بازسازی یک هاپلوتیپ جدید به طور صحیح، مورد مطالعه قرار می‌دهیم تا از این طریق معین شود آیا تعداد به نسبت پائین htSNPهای بدست آمده برای بلوک‌های بزرگ، نوعی خطای برآورد است یا نه. برای این کار، ما دقت بازسازی هاپلوتیپ‌های نمونه را به ازای هر یک از روش‌های افراز بلوکی، در قالب یک روال cross-vlidation کامل بر روی تمام هاپلوتیپ‌های HapMap از پانل CEU در تمام ده ناحیه‌ی ENCODE مورد بررسی قرار دادیم. طی این روال، هاپلوتیپ هر یک از ۱۲۰ نمونه‌ی مورد بررسی به ترتیب، از مجموعه‌ی داده‌ها کنار گذاشته می‌شود و روال مربوط به محاسبه‌ی htSNPها بر روی مجموعه‌ی باقی‌مانده اجرا می‌شود. سپس میانگین اختلاف بین اسنیپ‌های هاپلوتیپ کنار گذاشته و اسنیپ‌های نمونه‌های برابر با این هاپلوتیپ در htSNPها، به عنوان خطای بازسازی هاپلوتیپ محاسبه می‌شود.

شکل ۸۰۳ نتایج بدست آمده از این بررسی را نشان می‌دهد. همانطور که ملاحظه می‌کنید، htSNPهای بدست آمده برای هر یک از روش‌ها، می‌توانند دست کم در ٪۷۰ موارد، به طور دقیق تمام اطلاعات لازم برای بازسازی هاپلوتیپ کنار گذاشته را فراهم کنند. این مقدار نزدیک به نسبتی است که الگوریتم htSNPer برای پوشش "هاپلوتیپ‌های رایج" توسط تگ‌اسنیپ‌ها، تضمین می‌کند. دقت بازسازی هاپلوتیپ‌ها، در

بین روش‌های مختلف متفاوت است، اما در مجموع برای هاپلوتیپ‌هایی که تعداد بیشتری htSNP برای بازسازی‌شان لازم است، اندکی کاهش می‌یابد. مقدار این تغییرات در روش‌هایی که ژنوم را به بلوک‌های کوچکتری افراز می‌کنند، مثل MDL و GAM بیشتر است و از این رو به نظر می‌رسد نمی‌توان یک روند مشخص از کاهش دقت بازسازی هاپلوتیپ‌ها برای این دو روش، در شکل ۸۰۳ مشاهده کرد. اما در دیگر روش‌ها، دقت بازسازی هاپلوتیپ‌ها با استفاده از htSNPها، وقتی نسبت اسنیپ‌های درون بلوک به htSNPهای آن به شش برابر یا بیشتر می‌رسد تقریبا ثابت باقی می‌ماند.



شکل ۸۰۳: دقت بازسازی هاپلوتیپ‌ها توسط htSNPها

ارتفاع هر ستون، احتمال بازسازی دقیق یک هاپلوتیپ جدید را با استفاده از اطلاعات موجود در htSNPها نشان می‌دهد. این نتایج از طریق اجرای یک روال cross-validation بر روی ۱۲۰ هاپلوتیپ از نمونه‌های HapMap در ده ناحیه‌ی ENCODE برآورد شده‌اند. برای نمایش بهتر تفاوت‌ها، مبداء محور عرض‌ها ۰/۵ انتخاب شده است. در اینجا، منظور از تعداد "tagged SNP"ها، نسبت اسنیپ‌های بلوک به htSNPها است.

۵۰۳ شباهت بلوک‌های هاپلوتیپی در بین روشهای متفاوت

کمیت معرفی شده در بخش ۴۰۵۰۲، معیاری برای سنجش شباهت بین دو افراز بلوکی مختلف بر روی یک ناحیه‌ی واحد ارائه می‌کند. جدول ۶۰۳، نتایج بدست آمده از محاسبه‌ی این سنجه بین افرازهای بلوکی بدست

آمده از اجرای روش‌های مختلف بر روی هاپلوتیپ‌های ده ناحیه‌ی ENCODE را نشان می‌دهد. یادآوری می‌شود که مقدار این سنجه برابر است با احتمال "همبسته فرض شدن" یک جفت اسنیپ توسط یکی از افرازها زمانی که توسط افراز دیگر "همبسته فرض شده است". مقادیر مندرج در جدول ۶۰۳ میانگین این سنجه‌ی بر روی ده ناحیه‌ی ENCODE است. می‌توان دید که نتایج بدست آمده در این جدول، تائیدی بر بحث صورت گرفته بر روی نمایش افرازهای بلوکی در چند ناحیه‌ی نمونه، در بخش ۲۰۳ است.

جدول ۶۰۳: شباهت افرازها بین روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها

روش افراز بلوکی	HOT	MB	HB	MDL	GAM	GAB	GPG	GPF
HOT	۱/۰۰	۰/۳۹	۰/۳۷	۰/۲۰	۰/۱۸	۰/۳۲	۰/۴۵	**۰/۴۶**
MB	۰/۳۹	۱/۰۰	**۰/۶۷**	۰/۳۶	۰/۳۱	۰/۵۸	۰/۶۰	۰/۵۷
HB	۰/۳۷	**۰/۶۷**	۱/۰۰	۰/۴۲	۰/۳۶	۰/۶۰	۰/۵۸	۰/۵۵
MDL	۰/۲۰	۰/۳۶	۰/۴۲	۱/۰۰	**۰/۵۴**	۰/۴۴	۰/۳۱	۰/۲۹
GAM	۰/۱۸	۰/۳۱	۰/۳۶	**۰/۵۴**	۱/۰۰	۰/۴۴	۰/۲۶	۰/۲۴
GAB	۰/۳۲	۰/۵۸	۰/۶۰	۰/۴۴	۰/۴۴	۱/۰۰	۰/۵۷	۰/۵۳
GPG	۰/۴۵	۰/۶۰	۰/۵۸	۰/۳۱	۰/۲۶	۰/۵۷	۱/۰۰	**۰/۸۹**
GPF	۰/۴۶	۰/۵۷	۰/۵۵	۰/۲۹	۰/۲۴	۰/۵۳	**۰/۸۹**	۱/۰۰

همانطور که ملاحظه می‌کنید، نمی‌توان هیچ توافق عمومی معناداری بین روش‌های مختلف افراز بلوکی مشاهده کرد. با این حال، می‌توان شبیه‌ترین افرازها به یکدیگر را به ترتیب زیر نام برد: افرازهای روش GPG و روش GPF، افرازهای روش MB و روش HB، افرازهای روش GAB و روش HB و افرازهای روش GAM و روش MDL. وجود برخی از این شباهت‌ها، قابل انتظار است. به عنوان مثال، روش‌های GPG و GPF، گونه‌های متفاوتی از روش پیشنهادی ما در این رساله، یعنی GPMAP هستند. رابطه‌ی بین روش‌های MB و HB نیز بر همین روال است چون در هر دوی آنها به طور مشابه، یک الگوریتم برنامه‌ریزی پویا برای تعیین افراز بهینه، مقید به شرط واگرائی هاپلوتیپ‌ها بکار گرفته می‌شود. با این حال، شباهت بین روش‌های MDL و GAM و نیز شباهت بین روش‌های GAB و HB، با اینکه چندان زیاد نیست اما می‌تواند قابل تأمل باشد. به طور میانگین افرازهای بدست آمده از روش‌های پیشنهادی ما، یعنی GPG و GPF، به طور عمومی بیشترین شباهت را با دیگر افرازهای بدست آمده از دیگر روش‌ها دارند که در جای خود می‌تواند یک مزیت محسوب شود. در بین روش‌های موضعی افراز بلوکی هاپلوتیپ‌ها نیز، افرازهای بدست آمده از روش GAB از شباهت به نسبت قابل تأملی با تمام افرازهای دیگر برخوردار است. بر پایه‌ی سنجه‌ی شباهت مورد بررسی

در این بخش، ساختارهای بلوکی بدست آمده از روش‌های GPG و GPF، سازگارترین ساختار با بلوک‌های تعریف شده بر اساس روشHOT، یعنی موقعیت نقاط پراحتمال نوترکیبی بر روی ژنوم هستند.

۶۰۳ مقایسه‌ی ثبات مدلهای متفاوت در تعریف بلوک‌های هاپلوتیپی

یک ویژگی مطلوب برای روش‌های افراز بلوکی هاپلوتیپ‌ها، ثبات این روش‌ها در بازتولید بلوک‌های هاپلوتیپی یکسان به ازای هاپلوتیپ‌های یک جمیعت معین در نسل‌های متوالی است. در بخش ۵۰۵۰۲، روشی برای بررسی ثبات روش‌های افراز بلوکی ارائه گردید که در آن ثبات مرزها در افرازهای بلوکی بدست آمده از هاپلوتیپ‌های شبیه‌سازی شده در ده نسل متوالی مورد بررسی قرار می‌گیرد. ما از هاپلوتیپ‌های HapMap در ناحیه‌ی 9q34.11 به عنوان هاپلوتیپ‌های اجدادی در این شبیه‌سازی استفاده کردیم. در این شبیه‌سازی، تنها عامل تکوین هاپلوتیپ‌ها طی نسل‌های متوالی، رویداد نوترکیبی با نرخ ۰/۵ بر روی موقعیت‌های مرزی بلوک‌های تعیین شده در نسل اول، هستند.

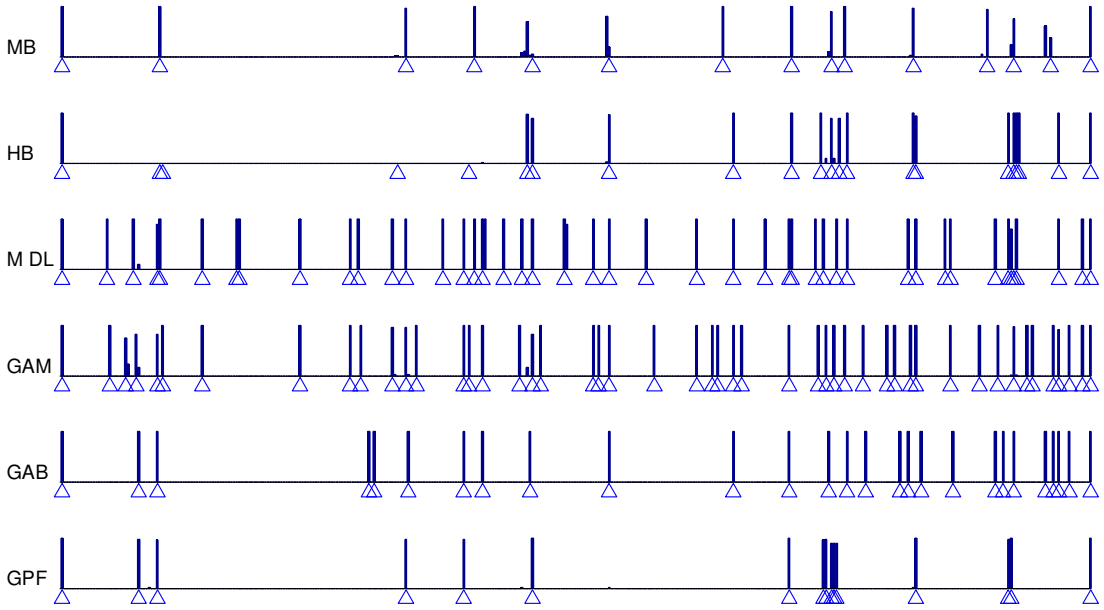
شکل ۹۰۳، موقعیت مرز بلوک‌ها در افرازهای مختلف ناحیه‌ی 9q34.11 را همراه با نموداری از احتمال وقوع مرز بلوک‌ها در هر یک از نقاط ناحیه‌ی ژنومی مورد مطالعه، در دهمین نسل از نمونه‌های شبیه‌سازی شده نشان می‌دهد. مقدار این احتمال، از طریق ۵۰ بار تکرار روال شبیه‌سازی شرح داده شده در ذیل شکل ۹۰۳، برآورد شده است. جدول ۷۰۳ مقدار برآورد شده به عنوان سنجه‌ی ثبات را به ازای روش‌های مختلف نشان می‌دهد.

جدول ۷۰۳: مقایسه‌ی ثبات در تعریف بلوک‌های هاپلوتیپ در بین روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها

روش افراز بلوکی	MB	HB	MDL	GAM	GAB	GPG	GPF
ثبات	۹۲/۰	۶۹/۲	۹۹/۴	۹۹/۷	۱۰۰	۱۰۰	۹۷/۶

مقادیر بر حسب درصد است.

همانطور که ملاحظه می‌کنید، تعریف بلوک‌ها در تمام روش‌ها به جز HB از ثبات قابل قبولی برخوردارند. عملکرد ضعیف روش HB می‌تواند به تاثیر ناخواسته‌ی رویکرد بهینه‌سازی مورد استفاده در این روش بازگردد. ناپایداری و حساسیت بالای جواب به شرایط معمولا می‌تواند نتیجه‌ای غیر قابل اجتناب از تحمیل شرایط بهینگی بر جواب باشد؛ به ویژه اینکه تابع هدف در HB، تعداد تگ‌اسنیپ‌های مورد نیاز برای پوشش تنوع



شکل ۹۰۳: ثبات روش‌های مختلف افراز در تعریف ساختار بلوکی ناحیه‌ی 9q34.11 (ENr232)
موقعیت مرز بلوک‌های افراز در نمونه‌ی اجدادی با مثلث‌های کوچک نشان داده شده است. برای هر یک از روش‌های افراز، نمونه‌ی دیگری از هاپلوتیپ‌ها با اعمال فرایند نوترکیبی بر روی مرز بلوک‌های پیشنهاد شده توسط همان روش در نسل اول، بدست می‌آید. پس از تکرار این فرایند تا ده نسل، افراز بلوکی مجددا به ازای نمونه‌های بدست آمده در نسل دهم بدست می‌آید و با افراز اولیه مقایسه می‌شود. ارتفاع ستون‌ها بر روی هر نقطه از ژنوم مورد مطالعه، احتمال وقوع مرز بلوک در آن موقعیت را به ازای هاپلوتیپ‌های دهمین نسل نشان می‌دهد.

هاپلوتیپ‌های درون بلوک‌ها است که به نظر می‌رسد مدل مناسبی برای تعریف "طبیعی" بلوک‌ها نیست.

از سوی دیگر، ثبات روش GAB برای تعریف بلوک‌ها، کاملا شگفت‌آور است؛ در شبیه‌سازی‌هایی که ما انجام دادیم هر دو روش GAB و GPG از ثبات صد درصد برخوردار بودند، یعنی، تکوین هاپلوتیپ‌ها تحت شرایط مدل فوق هیچ تغییری در موقعیت مرزهای بلوک در این دو روش ایجاد نمی‌کرد. این موضوع نشان‌دهنده‌ی، ثبات بالای شاخص گابریل در تعیین همبستگی بین اسنیپ‌ها است.

۷۰۳ توان شناسائی نقاط پراحتمال نوترکیبی

با استفاده از برنامه‌ی مولد توالی‌های تصادفی، دو مجموعه از نمونه هاپلوتیپ‌های شبیه‌سازی شده، تحت مدلی که شرح آن در بخش ۶۰۵۰۲ آمده است، بدست آوردیم. در هر یک از این دو مجموعه، ۱۰۰ نمونه‌ی

مستقل تصادفی وجود دارد که هر یک از آنها در مجموعه‌ی اول، شامل ۴۰ هاپلوتیپ و در مجموعه‌ی دوم، شامل ۱۰۰ هاپلوتیپ هستند. برنامه‌ی مولد توالی‌های تصادفی این امکان را به ما می‌داد که موقعیت و وسعت نواحی پراحتمال نوترکیبی و شدت نوترکیبی در آنها را به دلخواه انتخاب کنیم تا شبیه‌سازی نمونه‌های تصادفی بر اساس همین انتخاب‌ها صورت گیرد. بر این اساس، مکان نقاط پراحتمال نوترکیبی در هر یک از نمونه‌های تولید شده، برای ما شناخته شده است. دقت روش‌های افراز بلوکی در تشخیص نقاط پراحتمال نوترکیبی، با مقایسه‌ی موقعیت مرزی بلوک‌های افراز و مکان از پیش‌شناخته شده‌ی نقاط پراحتمال نوترکیبی و تعیین موارد false positive و false negative در آنها مورد بررسی قرار می‌گیرد (برای تعاریف بخش ۶۰۵۰۲ را ببینید).

جدول ۸۰۳ نرخ خطا در تشخیص‌های false positive ، نرخ خطا در تشخیص‌های false negative و نرخ خطای کل به ازای روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها و دو مقدار متفاوت حجم نمونه نشان می‌دهد. به جز روش HB، نرخ خطای کل در دیگر روش‌ها منطقی و قابل قبول است. ساختار مدل و چارچوب اصلی الگوریتم در هر دو روش MB و HB یکسان است و تنها تفاوت آنها در این است که در MB هدف کمینه‌سازی تعداد بلوک‌های افراز و در HB کمینه‌سازی مجموع htSNPها در سراسر ناحیه است. از این رو می‌توان نتیجه گرفت که رویکرد مبتنی بر کمترین htSNP، چندان با الگوی نوترکیبی بر روی ژنوم سازگار نیست. در نقطه‌ی مقابل، روش‌های GPG و GPF در بین سایر روش‌های افراز بلوکی هاپلوتیپ‌ها، کمترین خطا را در تشخیص نقاط پراحتمال نوترکیبی دارند.

جدول ۸۰۳: خطای روش‌های افراز بلوکی هاپلوتیپ‌ها در تشخیص نقاط پراحتمال نوترکیبی

حجم نمونه*	نرخ خطا**	MB	HB	MDL	GAM	GAB	GPG	GPF
$n = ۴۰$	false positive rate	۲/۶	۱۵/۴	۳/۸	۵/۳	۳/۰	۱/۹	۱/۶
	false negative rate	۲/۵	۰/۹	۱/۲	۰/۷	۳/۱	۰/۹	۱/۰
	total error rate	۵/۱	۱۶/۳	۵/۰	۶/۰	۶/۱	۲/۸	۲/۶
$n = ۱۰۰$	false positive rate	۳/۸	۱۲/۲	۲/۶	۴/۹	۳/۴	۱/۱	۱/۰
	false negative rate	۲/۲	۱/۱	۰/۹	۰/۹	۲/۲	۰/۸	۱/۱
	total error rate	۶/۰	۱۳/۳	۳/۵	۵/۸	۵/۶	۱/۹	۲/۰

* n تعداد هاپلوتیپ‌های نمونه است. ** نرخ خطا بر حسب درصد است.

با مطالعه‌ی نرخ خطا در جدول ۸۰۳، بین موارد false positive و false negative ، به نظر می‌رسد احتمال واقع شدن مرز یک بلوک بیرون از نواحی پراحتمال نوترکیبی، بیشتر از احتمال آن است که یک ناحیه‌ی

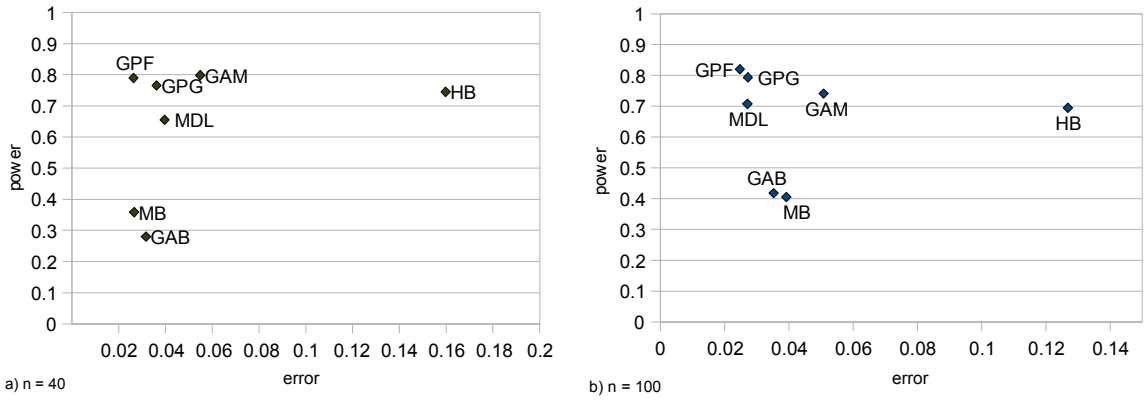
پراحتمال نوترکیبی توسط مرز هیچ بلوکی، شناسائی نشود. این موضوع به بیان ساده می‌تواند نشان‌دهنده‌ی تعداد بیشتر بلوک‌ها در مقایسه با نقاط پراحتمال نوترکیبی باشد علاوه بر آنکه، الگوی LD و تنوع هاپلوتیپ‌ها آنچنان در اطراف نواحی پراحتمال نوترکیبی دچار تغییر می‌شوند که بر اساس هیچ تعریفی، چنین مناطقی نمی‌توانند درون یک بلوک واقع شده باشند.

برای داشتن تصویری روشن‌تر از میزان دقت و کارایی روش‌های مختلف افراز بلوکی در تشخیص نقاط پراحتمال نوترکیبی، در شکل ۱۰۰۳، توان شناسائی نقاط پراحتمال نوترکیبی را در برابر خطای تشخیص، به ازای روش‌های مختلف و مقادیر متفاوت حجم نمونه نشان می‌دهیم. به طور کلی، افزایش تعداد هاپلوتیپ‌ها، نرخ خطا را در تمام روش‌ها کاهش داده است؛ به جز MB که در عوض، با افزایش توان همراه بوده است (جدول ۸۰۳ و شکل ۱۰۰۳). توان شناسائی نقاط پراحتمال نوترکیبی در روش‌های GPG، MDL، GAB و MB با افزایش حجم نمونه، افزایش یافته است. بر عکس، در روش‌های GPF، HB و GAM، افزایش حجم نمونه، تنها با افزایش دقت پیشگوئی همراه است. روش‌های پیشنهادی ما در این رساله، یعنی روش‌های GPG و GPF، کمترین نرخ خطا را در بین سایر روش‌های افراز بلوکی هاپلوتیپ‌ها، برای شناسائی نقاط پراحتمال نوترکیبی دارند و در مرز بلوک‌ها در آنها، در بیش از ۷۵ درصد موارد منطبق بر موقعیت نقاط پراحتمال نوترکیبی است. بلوک‌های روش GAM، اما با دقتی کمتر، در قریب به ۸۰ درصد از موارد منطبق بر نقاط پراحتمال نوترکیبی هستند که می‌تواند نتیجه‌ای از معیار تعریف بلوک‌ها در این روش و ارتباط آن با مدل نیای مشترک مورد استفاده در برنامه‌ی شبیه‌سازی نمونه‌ها باشد.

۸۰۳ توان شناسائی جایگاه ژنی یک خصیصه

در بخش ۷۰۵۰۲، طرح ساده‌ای برای سنجش کارایی روش‌های افراز بلوکی هاپلوتیپ‌ها در مطالعه‌ی جایگاه ژنی خصیصه، بر مبنای بررسی نمونه‌های شبیه‌سازی شده‌ی case و control ارائه کردیم. در این بخش، به بحث درباره‌ی نتایج بدست آمده از این ارزیابی می‌پردازیم.

جدول ۹۰۳ خطای نوع اول در تشخیص بلوک دربردارنده‌ی اسنیپ مسبب بیماری را به ازای روش‌های مختلف افراز بلوکی هاپلوتیپ‌ها نشان می‌دهد. مقدار این خطا نشاندهنده‌ی احتمال این است که آزمون، بلوکی



شکل ۱۰۳: کارایی روش‌های افراز بلوکی در شناسائی نقاط پراحتمال نوترکیبی
power نشان‌دهنده‌ی نرخ وقوع نقاط پراحتمال نوترکیبی در موقعیت‌های مرزی بلوک‌ها است. error نشان‌دهنده‌ی احتمال آن است که مرز یک بلوک در نقاطی بیرون از نواحی پراحتمال نوترکیبی قرار بگیرد. در نمودار سمت چپ (سمت راست)، بلوک‌ها بر روی داده‌هایی با $n = ۴۰$ ($n = ۱۰۰$) هاپلوتیپ تعیین شده‌اند.

را به عنوان بلوک مرتبط با خصیصه شناسائی کند در حالی که در واقع جایگاه بیماری در این بلوک نیست. این نتایج، بر اساس اجرای آزمون مربع کای در سطح معناداری $\alpha = ۰/۰۵$ بر روی ۲۵۰ مجموعه از نمونه‌های شبیه‌سازی شده‌ی case و control بدست آمده است. همانطور که جدول ۹۰۳ نشان می‌دهد، به ازای ریسک نسبی بالاتر ($GRR_۱ = ۵$)، نتایج در تمام روش‌ها با خطای نوع اول بیشتری همراه است. تاثیر فراوانی نسبی آلل مرتبط با بیماری بر افزایش خطای نوع اول، کمتر از تاثیر ریسک نسبی ژنوتیپ اول است.

جدول ۹۰۳: خطای نوع اول در شناسائی جایگاه ژنی بیماری بین روش‌های مختلف

پارامترهای مدل بیماری / روش	SS	MB	HB	MDL	GAB	GPG	GPF
DAF=۵٪ـ۱۵٪ , $GRR_۱ = ۳$	۰/۲۶	۰/۲۰	۰/۱۷	۰/۱۶	۰/۱۷	۰/۱۳	۰/۱۳
DAF=۵٪ـ۱۵٪ , $GRR_۱ = ۵$	۰/۳۲	۰/۲۴	۰/۲۱	۰/۲۰	۰/۲۲	۰/۱۶	۰/۱۷
DAF=۲۰٪ـ۳۰٪ , $GRR_۱ = ۳$	۰/۲۹	۰/۲۱	۰/۱۸	۰/۱۷	۰/۱۹	۰/۱۴	۰/۱۵
DAF=۲۰٪ـ۳۰٪ , $GRR_۱ = ۵$	۰/۳۴	۰/۲۵	۰/۲۲	۰/۲۲	۰/۲۳	۰/۱۷	۰/۱۹

مقادیر خطا در جدول ۹۰۳، بیانگر آنند که آزمون مربع کای در سطح $\alpha = ۰/۰۵$، به ازای افرازهای بلوکی بدست آمده از گونه‌های متفاوت روش پیشنهادی ما، GPG و GPF، خطای نوع اول کمتری به نسبت دیگر روش‌ها تولید می‌کند. با این حال، بر اساس بحثی که در بخش ۷۰۵۰۲ طرح شد، لازم است پیش از برآورد توان و قضاوت درباره‌ی کارایی روش‌ها، pـ مقدار آستانه‌ای مناسب هر روش به قسمی تعیین شود که خطای نوع اول یکسانی توسط هر یک از روش‌ها تولید شود. با ارزیابی روش مطالعه‌ی همبستگی بر روی نیمی از

مجموعه نمونه‌ها، مقادیر آستانه‌ایی مناسب برای هر یک از روش‌ها را بدست آوردیم(جدول ۱۰۰۳). از این مقادیر برای اجرای آزمون همبستگی مربع کای بر روی دیگر نمونه‌های شبیه‌سازی شده استفاده می‌کنیم و توان روش‌های مختلف را بر اساس برآورد می‌کنیم.

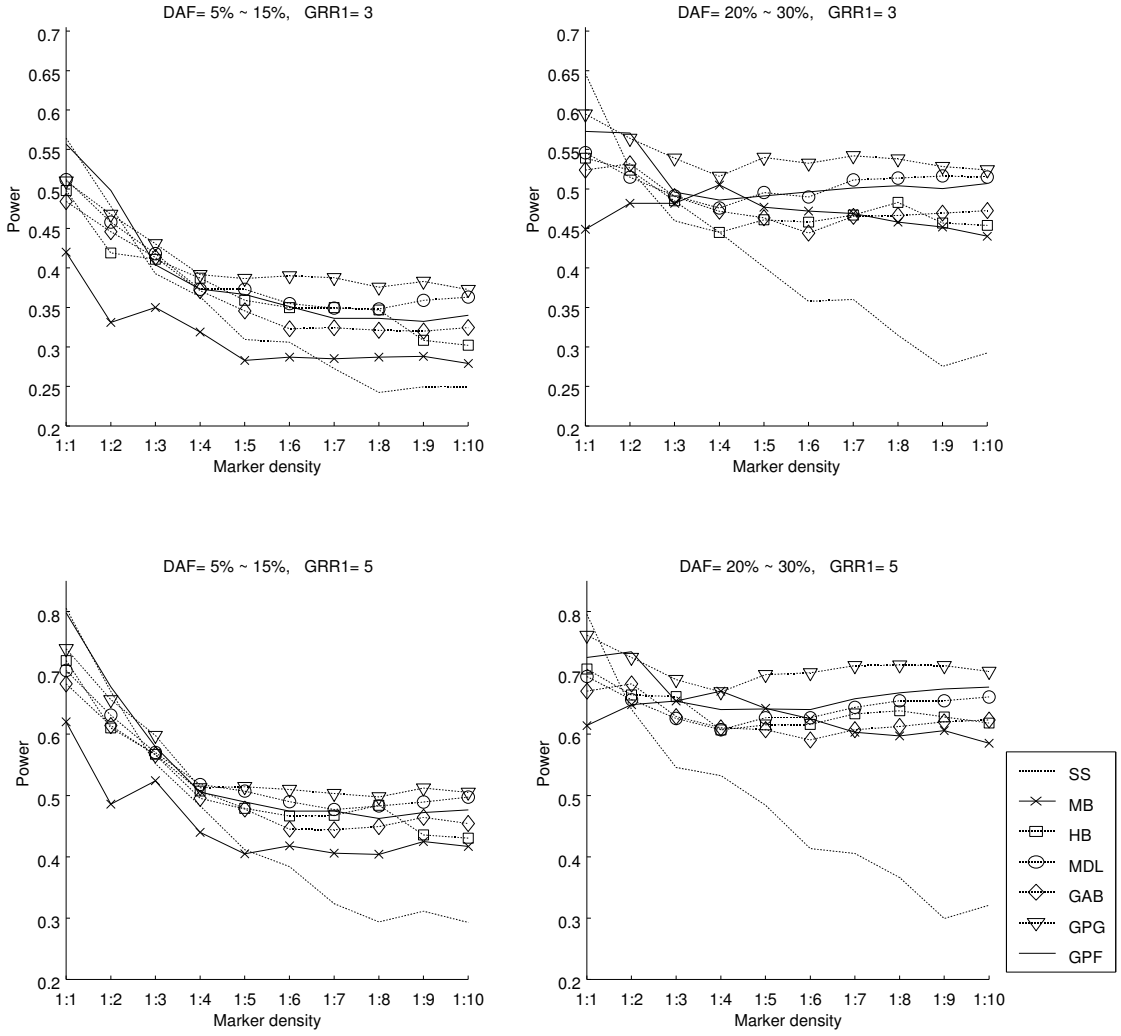
جدول ۱۰۰۳: pـ مقدار آستانه‌ای آزمون همبستگی مربع کای برای بدست آوردن خطای نوع اول ثابت بین روش‌های مختلف در مطالعه‌ی همبستگی

روش پارامترهای مدل بیماری	SS	MB	HB	MDL	GAB	GPG	GPF
DAF=۵٪ـ۱۵٪ $GRR_1 = ۳$	۱/۰e-۲	۱/۰e-۲	۱/۶e-۲	۱/۸e-۲	۱/۳e-۲	۳/۰e-۲	۲/۷e-۲
DAF=۵٪ـ۱۵٪ $GRR_1 = ۵$	۲/۹e-۳	۲/۴e-۳	۵/۰e-۳	۴/۱e-۳	۲/۹e-۳	۱/۵e-۲	۱/۱e-۲
DAF=۲۰٪ـ۳۰٪ $GRR_1 = ۳$	۷/۱e-۳	۸/۰e-۳	۱/۳e-۲	۱/۲e-۲	۹/۸e-۳	۲/۴e-۲	۱/۷e-۲
DAF=۲۰٪ـ۳۰٪ $GRR_1 = ۵$	۱/۷e-۳	۲/۲e-۳	۳/۲e-۳	۲/۴e-۳	۱/۸e-۳	۱/۱e-۲	۷/۷e-۳

۱۰۸۰۳ تاثیر نحوه‌ی انتخاب نشانگذارها بر توان

توان آزمون همبستگی در شناسائی جایگاه خصیصه، بین روش‌های مبتنی بر ساختار بلوکی ژنوم و روش مبتنی بر آزمون تک اسنیپی (بخش ۷۰۵۰۲) در شکل‌های ۱۱۰۳ و ۱۲۰۳ نمایش داده شده است. هر دو شکل، نمودار تغییرات توان را برحسب چگالی نشانگذارهای مطالعه‌ی همبستگی بر روی ژنوم، نشان می‌دهند با این تفاوت که نمودار توان در شکل ۱۱۰۳، بر اساس نتایج بدست آمده از انتخاب یکنواخت اسنیپ‌ها به عنوان نشانگذار بدست آمده است و در شکل ۱۲۰۳، بر اساس نتایج بدست آمده از انتخاب اولویت داده شده‌ی نشانگذارها است.

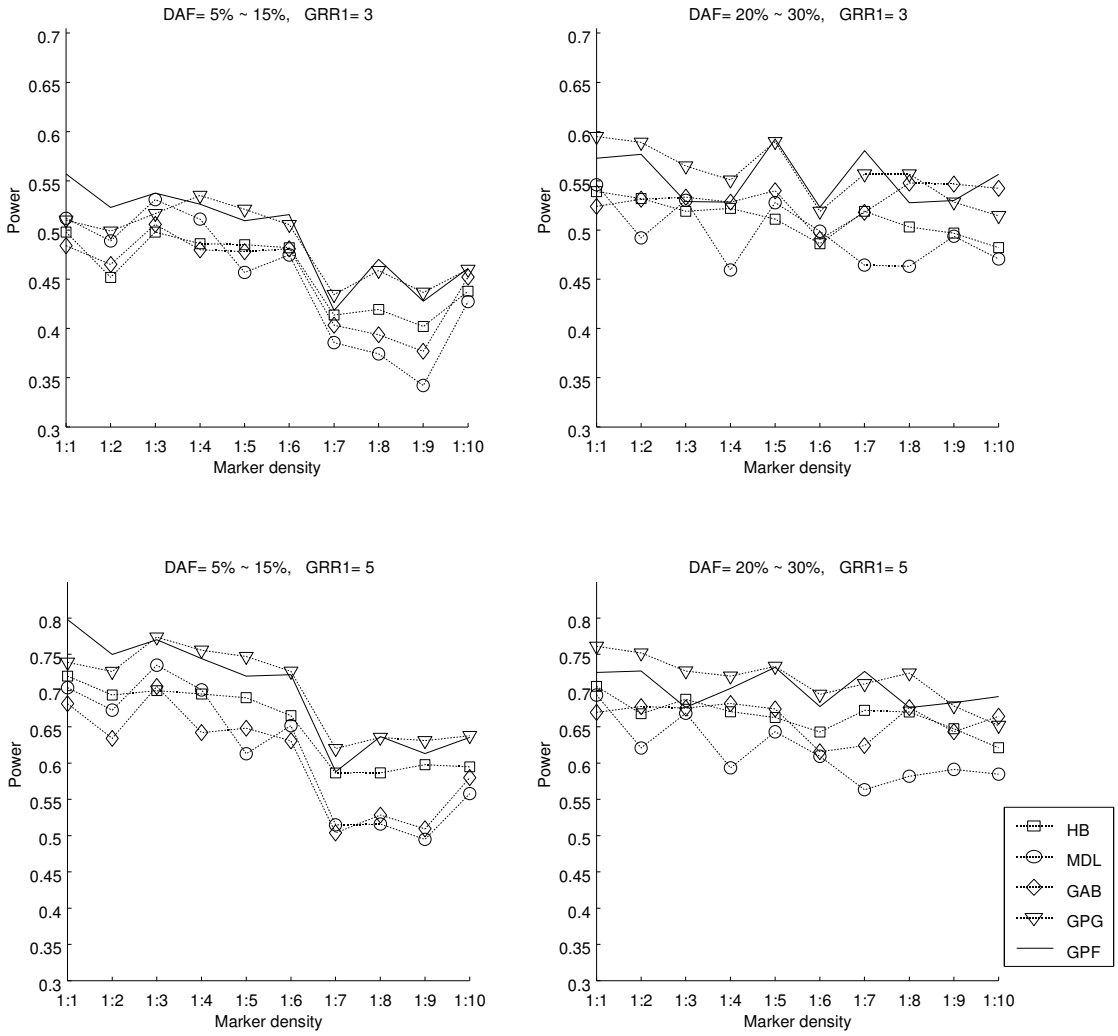
به طور کلی، توان شناسائی جایگاه خصیصه، با کاهش چگالی نشانگذارها، کاهش می‌یابد. این کاهش توان، در روش تک اسنیپی بیشتر از روش‌های مبتنی بر ساختار بلوکی ژنوم است. در واقع، مطالعه‌ی همبستگی با استفاده از اطلاعات هاپلوتیپ‌های درون بلوک، حتی در شرایطی که چگالی نشانگذارها پائین باشد از کارایی



شکل ۱۱۰۳: کارایی روش‌های افراز بلوکی هاپلوتیپ‌ها در شناسائی جایگاه ژنی مرتبط با خصیصه، با انتخاب یکنواخت نشانگذارها

نسبتاً مطلوبی برخودار است به قسمی که توان آزمون در روش‌های مبتنی بر بلوک‌های هاپلوتیپ، در شرایطی که فراوانی نسبی آلل مسبب بیماری بین ٪۲۰ تا ٪۳۰ باشد، حتی با کاهش چگالی نشانگذارها تقریبا ثابت باقی می‌ماند. البته، پائین بودن هتروزیگوسیتی در اسنیپ مسبب بیماری (٪۱۵ـ٪۵=DAF)، می‌تواند باعث کاهش LD بین اسنیپ مسبب بیماری و سایر اسنیپ‌ها شود که این موضوع خود، کاهش توان آزمون همبستگی در روش‌های مبتنی بر ساختار بلوکی را در پی دارد.

از سوی دیگر، در تمام روش‌ها، توان آزمون عمدتا با افزایش ریسک نسبی ژنوتیپ اول، GRR_1، افزایش می‌یابد. به طور مشابه، در غالب موارد، افزایش فراوانی نسبی آلل مرتبط با بیماری نیز تاثیری مثبت بر افزایش



شکل ۱۲۰۳: کارایی روش‌های افراز بلوکی هاپلوتیپ‌ها در شناسائی جایگاه ژنی مرتبط با خصیصه، با انتخاب اولویت داده شده‌ی نشانگذارها

توان آزمون‌های همبستگی دارد. به بیان دیگر، تشخیص جایگاه ژنی یک بیماری در طرح case و control، با افزایش شیوع و نفوذ بیماری با کارآیی بیشتری امکانپذیر خواهد بود اما، شناسائی ژن مرتبط با یک بیماری با نفوذ بالا، ساده‌تر از شناسائی ژن مرتبط با یک بیماری با شیوع بالا است.

همانطور که در شکل ۱۱۰۳ ملاحظه می‌کنید، در حالتی که نشانگذارها به طور یکنواخت از بین اسنیپ‌ها انتخاب می‌شوند روش‌های مبتنی بر ساختارهای بلوکی بدست آمده از GPG، GPF و MDL، از برتری مختصری نسبت به دیگر روش‌ها برخوردارند.

اگر هر اسنیپ را بر حسب حجم اطلاعاتی که حمل می‌کند رتبه‌بندی کنیم و نشانگذارها را بر اساس این

رتبه‌بندی انتخاب کنیم آنگاه در مقایسه با راهبرد انتخاب یکنواخت نشانگذارها، کاهش چگالی نشانگذارها، تاثیر بطور محسوس کمتری بر کاهش توان آزمون‌های همبستگی دارد(شکل ۱۲۰۳). با انتخاب اولویت داده شده‌ی نشانگذارها، توان شناسائی جایگاه ژنی مرتبط با خصیصه، وقتی تنها یک پنجم از اسنیپ‌های موجود در ناحیه‌ی کروموزومی مورد مطالعه، به عنوان نشانگذار در نظر گرفته می‌شوند، همچنان مقدار مطلوبی دارد. در این راهبرد نیز همانند دیگر راهبرد انتخاب نشانگذارها، گونه‌های مختلف روش افراز بلوکی پیشنهادی در این رساله، GPG و GPF، از کارایی بیشتری نسبت به سایر روش‌ها برای استفاده در مطالعات همبستگی برخوردارند. پس از این دو روش، می‌توان HB را به عنوان کارآمدترین روش برای تعیین افراز بلوکی مناسب در مطالعات همبستگی به شمار آورد.

فصل ۴

نتیجه‌گیری

۱۰۴ الگوریتم ژنتیک برای استنباط هاپلوتیپ‌ها

مسئله‌ی تفکیک ژنوتیپ‌ها برپایه‌ی مدل بیشترین پارسیمونی، یک مسئله‌ی NP-hard است. از این حیث، تلاش در بکارگیری رویکردهای اکتشافی رایج در حل مسائل بهینه‌سازی، برای حل این مسئله نیز می‌تواند ایده‌ای درخور توجه باشد. الگوریتم‌های ژنتیک یکی از ابزارهای رایج و کارآمد در حل چنین مسائلی هستند. در بخش ۱۰۲، ما دو پیاده‌سازی از این الگوریتم را برای حل مسئله‌ی تفکیک ژنوتیپ‌ها با هدف بیشترین پارسیمونی توسعه دادیم. در نسخه‌ی اول، روال‌های نزدیک به ذهن و ساده‌ای برای تعریف عملگرهای "کراس‌اوِر" و "جهش" در الگوریتم ژنتیک معرفی گردیدند. نتایجی که از ارزیابی این روش بر روی نمونه‌های شبیه‌سازی بدست آمد بیانگر آن بود که کارایی این روش ابتدائی در رسیدن به جواب‌هایی حتی نزدیک به جواب بهینه‌ی مدل بیشترین پارسیمونی، چندان امیدوارکننده نیست(بخش ۱۰۳). در مقابل، پیاده‌سازی رده‌ای از الگوریتم‌های سودجویانه برای حل مسئله‌ی بیشترین پارسیمونی، در قالب نسخه‌ی دوم الگوریتم ژنتیک پیشنهادی، توانست به طور چشمگیر به بهبود نتایج کمک کند به قسمی که در بیش از نیمی از نمونه‌های مورد ارزیابی، این نسخه از الگوریتم جواب‌هایی با تعداد هاپلوتیپ‌های متمایز برابر یا کمتر از کران بالای از پیش شناخته شده برای این مقدار بدست آورد.

در مقایسه با دیگر روش‌های رایج در حل مسئله‌ی تفکیک ژنوتیپ‌ها، الگوریتم‌های ژنتیکی ارائه شده در این رساله، به خصوص از حیث دقت در استنباط هاپلوتیپ‌ها، به سختی می‌توانند در بین الگوریتم‌های متوسط این مبحث قرار گیرند. نتایج حاصل از ارزیابی روش‌های مختلف استنباط هاپلوتیپ‌ها بر روی نمونه‌ای از ژنوتیپ‌های واقعی علاوه بر آن نشان می‌داد که برخی الگوریتم‌های مبتنی بر مدل‌های بیزی، جواب‌هایی با تعداد هاپلوتیپ‌های متمایز کمتر از الگوریتم ژنتیکی پیشنهادی ما که مستقیما هدف بیشترین پارسیمونی را تعقیب می‌کند بدست می‌آورند. این موضوع می‌تواند از دو دیدگاه مورد بررسی بیشتر قرار گیرد. اول، تلاش برای بهبود بیشتر رویکرد مبتنی بر الگوریتم‌های ژنتیک برای حل مسئله‌ی تفکیک ژنوتیپ‌ها با بیشترین پارسیمونی است. وارد کردن برخی دیگر از رویکردهای رایج در مسئله‌ی تفکیک ژنوتیپ‌ها مثل مدل فیلوژنی کامل یا مدل‌های استنباط آماری هاپلوتیپ‌ها در الگوریتم ژنتیک طرح شده در این رساله می‌تواند زمینه‌ی تحقیقات دیگری در آینده برای این موضوع باشد. اما برداشت دوم درباره‌ی ضعف الگوریتم ژنتیک مطرح شده در این رساله، به سادگی بیش از حد مدل بیشترین پارسیمونی برای تبیین ساختار آماری ژنوتیپ‌ها و هاپلوتیپ‌های موجود در طبیعت بر می‌گردد. بخشی از نتایج ما در بخش ۱۰۳ مؤید آن است که لزوما ارتباطی بین دقت الگوریتم در تشخیص هاپلوتیپ‌های واقعی و کمینه بودن تعداد هاپلوتیپ‌های متمایز وجود ندارد. این واقعیت به اشکالی دیگر توسط برخی دیگر از محققین مورد توجه قرار گرفته است[۵۱].

۲۰۴ الگوریتم GPMAP برای تعیین بلوک‌های هاپلوتیپی

در بخش ۴۰۲، روشی بر پایه‌ی آنالیز همبستگی جفت اسنیپ‌ها، برای افراز سراسری هاپلوتیپ‌ها ارائه گردید. در این روش، که آنرا به اختصار GPMAP نامیدیم، بلوک‌های هاپلوتیپ به گونه‌ای تعیین می‌شوند که تعداد کل جفت اسنیپ‌های همبسته‌ی درون بلوک‌ها، بیشترین باشد و بلوک‌ها تنها، تعداد معین و اندکی از جفت اسنیپ‌های مستقل را شامل شوند. ما از آماره‌ی نرمال شده‌ی D' برای سنجش LD بین جفت اسنیپ‌ها و به عنوان معیار استقلال یا همبستگی بین جفت اسنیپ‌ها استفاده کردیم و برای تعیین سطح معناداری برآورد این آماره، از آزمون دقیق فیشر استفاده کردیم. بر این اساس، شاخص جدیدی برای تعیین همبستگی در جفت اسنیپ‌ها معرفی شد(بخش ۳۰۲). در چارچوب الگوریتم پیشنهادی برای افراز سراسری هاپلوتیپ‌ها

و با استفاده از این شاخص جدید و شاخص گابریل، دو روش مختلف برای تعیین بلوک‌های هاپلوتیپ ارائه شدند.

از زمان اولین مشاهداتی که مبین وجود ساختار بلوکی در ژنوم انسان بودند تا کنون، تلاش‌های گوناگونی برای ارائه‌ی یک تعریف معین از بلوک‌های هاپلوتیپ صورت گرفته است که توجه به برخی قیود بر روی واگرائی هاپلوتیپ‌ها درون بلوک‌ها، در بسیاری از آنها مشترک است. نتایجی که ما با اجرای روش پیشنهادی بر روی نمونه‌هایی از هاپلوتیپ‌های واقعی نواحی ENCODE بدست آوردیم، بیانگر آن است که میزان واگرائی هاپلوتیپ‌ها درون بلوک‌های بدست آمده از روش ما، با قیود رایج در دیگر تعاریف سازگار است. این نتیجه از این رو قابل توجه است که در روش ما، هیچ قید مرتبط با واگرائی هاپلوتیپ‌ها مورد استفاده قرار نمی‌گیرد.

نتایج بدست آمده از تعیین htSNPها در بلوک‌های هاپلوتیپی تعریف شده بر روی نواحی ENCODE، به ما نشان داد که کارایی افرازهای بلوکی سراسری برای بدست آوردن کمترین تعداد تگ‌اسنیپ مورد نیاز برای پوشش تنوع هاپلوتیپ‌ها، به طور غیر قابل انکار به نتایج بدست آمده از الگوریتم بهینه در این مبحث، هم از حیث تعداد تگ‌اسنیپ‌ها و هم از حیث پوشش نوکلئوتیدی تگ‌اسنیپ‌ها نزدیک است و از این رو می‌توان آنها را به عنوان جایگزین ساده‌ای برای مدل پیچیده‌ی مورد استفاده در روش بهینه بکار گرفت(بخش ۴۰۳). این نتیجه‌گیری به شکلی دیگر در [۱۶۳] مورد تأیید قرار گرفته است. شایان ذکر است که در روش ما هیچ تابع هدفی به طور صریح برای رسیدن به کمترین تعداد htSNP یا بالاترین میزان پوشش نوکلئوتیدی htSNPها مورد توجه قرار نمی‌گیرد.

با ارزیابی یک سنجه‌ی شباهت بین افرازهای بلوکی متفاوت، این نتیجه حاصل شد که شباهت بین افرازهای بلوکی مختلف، اصولا کمتر از آن است که بتوان، از یک ساختار بلوکی استاندارد بر روی ژنوم صحبت کرد. با این وجود، ممکن است بتوان در بین افرازهای بدست آمده از برخی روش‌ها شباهتی۵۰ درصدی را مشاهده کرد.

پایداری موقعیت بلوک‌های تعریف شده بر روی ژنوم طی نسل‌های متوالی، از طریق شبیه‌سازی یک مدل تکوینی ساده بر روی نمونه‌ای از هاپلوتیپ‌های واقعی مورد مطالعه قرار گرفت(بخش ۶۰۳). بر اساس این مدل، فرض می‌کنیم ترکیب هاپلوتیپ‌ها طی هر نسل، با وقوع رویدادهای نوترکیبی بر روی مرز بلوک‌ها به طور

تصادفی تغییر می‌کند. بر پایه‌ی نتایج حاصل از این شبیه‌سازی، تنها بلوک‌های تعریف شده بر اساس شاخص گابریل، در تمام طول شبیه‌سازی ثابت باقی ماندند هرچند تغییر در بلوک‌های دیگر روش‌ها نیز چندان بالا نبود. ثبات شاخص همبستگی مبتنی بر آزمون دقیق فیشر و روش GPMAP نیز در این مطالعه قابل قبول بود.

ملاحظات زیست‌شناسی بر این نکته اشاره می‌کنند که مرز بلوک‌ها در افرازهایی که توسط روش‌های افراز بلوکی هاپلوتیپ‌ها معرفی می‌شوند می‌بایست تا حدودی با نقاط پراحتمال نوترکیبی بر روی ژنوم توافق داشته باشند. از این رو، کارایی روش‌های گوناگون افراز بلوکی هاپلوتیپ‌ها را در زمینه مورد ارزیابی قرار دادیم. برای این کار، ما از یک نرم‌افزار مولد توالی‌هایی تصادفی تحت مدل داده شده برای نقاط پراحتمال نوترکیبی استفاده کردیم. هر دو روش افراز سراسری هاپلوتیپ‌ها که در این رساله معرفی شدند، بهترین کارایی را هم از حیث دقت و هم از حیث توان در شناسائی نقاط پراحتمال نوترکیبی نشان دادند. از این رو، روش ما به دلیل سادگی نسبی محاسبات در آن، به عنوان روشی سریع می‌تواند جایگزینی برای برخی روش‌های پیچیده‌ی استنباط نقاط پراحتمال نوترکیبی باشد.

بسیاری از روش‌های رایج در مطالعات case-control بر استفاده از ایده‌ی پنجره‌ی لغزان متکی هستند که در آن، آزمون همبستگی بین خصیصه و هاپلوتیپ‌های محدود به پنجره‌ای با عرض ثابت که بر روی ژنوم مورد مطالعه حرکت داده می‌شود، ارزیابی می‌گردد. انتخاب مقادیر متفاوت برای عرض پنجره، بر دقت و توان نتایج بدست آمده از آزمون‌های همبستگی موثر است که در بیشتر موارد تعیین اندازه‌ی مناسب برای آن تنها از طریق اجراهای مکرر به ازای مقادیر متفاوت برای عرض پنجره امکان‌پذیر است. مزیت اولیه‌ی روش‌های مبتنی بر اطلاعات ساختار بلوکی ژنوم در مطالعات case-control، عدم برخورد با چنین مشکلی است. در واقع ساختار بلوکی، معیار روشنی برای تعیین محدوده‌ی نواحی مورد نیاز برای بررسی آزمون‌های همبستگی فراهم می‌کند. در این میان، بلوک‌های بدست آمده از روش افراز بلوکی پیشنهادی ما، به واسطه‌ی نحوه‌ی تعریف بلوک‌های هاپلوتیپ در آن، با نتایجی به نسبت قویتر و کم خطاتر در شناسائی جایگاه ژنی مرتبط با بیماری ظاهر گردید.

در بخش ۷۰۵۰۲، شیوه‌ی جدیدی مبتنی بر مطالعه‌ی هاپلوتیپ نمونه‌های case و control درون بلوک‌ها، برای شناسائی جایگاه خصیصه در ژنوم ارائه شد. در این روش، از آزمون مربع کای و خوشه‌بندی سلسله‌مراتبی

برای دسته‌بندی هاپلوتیپ‌ها به ازای افراز بلوکی داده شده برای بررسی همبستگی بین خصیصه و جایگاه‌های مختلف بر روی ژنوم، در کنار اطلاعات ساختار بلوکی هاپلوتیپ‌ها استفاده شد. علاوه بر آن، با استفاده از نمونه‌های شبیه‌سازی شده‌ی case و control تحت برخی مدل‌های توارثی برای یک بیماری تک جایگاهی، طرح ساده‌ای برای مقایسه‌ی دقت و توان ساختارهای بلوکی مختلف در این کاربرد ارائه شد. نتایج این ارزیابی‌ها نشان‌دهنده‌ی تاثیر مثبت و معنادار استفاده از ساختار بلوکی ژنوم در شناسائی جایگاه ژنی خصیصه است. در بخش ۸۰۳ نشان دادیم با استفاده از اطلاعات هاپلوتیپ‌ها در ساختار بلوکی و انتخاب مجموعه‌ی مناسبی از نشانگذارها برای مطالعه‌ی همبستگی بر روی ژنوم، می‌توان نتایجی با دقت مطلوب را، حتی با اطلاعات یک پنجم از اسنیپ‌ها نیز بدست آورد.

نتایج بدست آمده از تمام بررسی‌ها در کنار هم، بیانگر آن است که مطالعه‌ی همبستگی آللی بین اسنیپ‌ها، به خوبی می‌تواند بخشی از ویژگی‌های تنوع ژنومی در جمعیت انسان را تبیین نماید.

۳۰۴ تحقیقات آتی

۱. در مسئله‌ی تفکیک ژنوتیپ‌ها با بیشترین پارسیمونی، نتایج بدست آمده از الگوریتم پیشنهادی در این رساله مؤید آن است که این الگوریتم برای نمونه‌هایی با حجم کمتر از ۳۰ ژنوتیپ بر روی حداکثر ۱۵ اسنیپ از دقت و کارایی نسبتا قابل قبولی برخوردار است. البته مشابه چنین محدودیتی در بسیاری دیگر از روش‌های مقدماتی تفکیک ژنوتیپ‌ها نیز وجود دارد که عموما با بکارگیری برخی پردازشهای دیگر، مثل «افراز و انعقاد»، و تلفیق آن با رویکرد اولیه، کارایی روش به طور معنادار افزایش می‌یابد. در اینجا نیز استفاده از روش «افراز و انعقاد» و ترکیب آن با الگوریتم ژنتیکی مطرح شده در بخش ۱۰۲، یعنی GAhap می‌تواند دامنه‌ی کاربرد عملی این روش را به تفکیک ژنوتیپ‌ها در مقیاس ژنومی گسترش دهد. علاوه بر آن، مدل‌های پیچیده‌تر دیگری از الگوریتم‌های ژنتیکی رواج دارند که ممکن است از کارایی بیشتری برای جستجوی فضای جواب در مسئله‌ی تفکیک ژنوتیپ‌ها تحت مدل بیشترین پارسیمونی برخوردار باشند. از جمله می‌توان به اعمال فرایندهای "مهاجرت" در کنار دو عملگر ژنتیکی "کراس‌اوِر" و "جهش" اشاره کرد که از طریق آن، بخشی از جمعیت "کروموزوم‌های" نسل جاری، از سایرین جدا می‌شوند و به جمعیت‌های دیگر می‌پیوندند. بررسی

دقت هاپلوتیپ‌های بدست آمده از تلفیق این رویکردها و دستیابی به جواب بیشترین پارسیمونی، مناسب تحقیق دیگری در توسعه‌ی روش‌های اکتشافی در بهینه‌سازی و کاربرد آنها در حل مسائل بیوانفورماتیک و زیست‌شناسی است.

۲. در این رساله، روشی برای افراز سراسری ژنوم به بلوک‌های هاپلوتیپی، با در اختیار داشتن نمونه‌هایی از ژنوتیپ‌های افراد غیرخویشاوند جمعیت ارائه شد. نتایج بدست آمده از شبیه‌سازی بیانگر آن است که این روش ضمن سادگی، کارایی مطلوبی در شناسائی نقاط پراحتمال نوترکیبی بر روی ژنوم دارد. از جمله پرسش‌هایی که می‌تواند زمینه‌ی تحقیق دیگری را در ژنتیک مولکولی باز کند مسئله‌ی بررسی وجود و ماهیت توالی‌های نوکلئوتیدی مرتبط با فرایندهای مولکولی نوترکیبی و نیز بررسی ساختار فضایی کروموزوم در نواحی پراحتمال نوترکیبی است.

۳. ساختار بلوکی ژنوم در بین نژادهای گوناگون انسان متفاوت است و در واقع، بخشی از اطلاعات مربوط به تاریخچه‌ی تکاملی دسته‌های مختلف جمعیت انسانی را به نمایش می‌گذارد. ما شاخصی برای اندازه‌گیری میزان شباهت بین ساختارهای بلوکی معرفی کردیم و توسط آن، شباهت بین نتایج بدست آمده از اجرای روش‌های مختلف افراز بلوکی بر روی نمونه‌های یکسانی از هاپلوتیپ‌ها را مورد بررسی قرار دادیم. این شاخص می‌تواند معیار مناسبی برای اندازه‌گیری شباهت بین ساختارهای بلوکی ژنوم بین نژادهای مختلف انسان باشد. تعیین بلوک‌های هاپلوتیپی در نژادهای متفاوت توسط روش‌های مختلف افراز و استفاده از شاخص شباهت برای تحلیل این اطلاعات، از دیگر موضوعات شایان تحقیق در آینده است.

۴. نوترکیبی تنها دلیل شکل‌گیری بلوک‌های هاپلوتیپ نیست. ترکیبی از سازوکارهای رانش ژنی و انتخاب طبیعی، تحت عنوان Selective sweep یا Gene hitchhiking، از جمله فرایندهایی هستند که باعث شکل‌گیری بلوک‌های هاپلوتیپ می‌شوند. رده‌بندی و مطالعه‌ی ژن‌های واقع در این نواحی و تمایز آنها از ژن‌های واقع بر نواحی پراحتمال نوترکیبی، نیز می‌تواند زمینه‌ی تحقیقی دیگر در ژنتیک مولکولی قرار گیرد.

۵. اخیرا، موسسه‌ی Wellcome Trust با همکاری مراکز تحقیقاتی دیگر، کنسرسیومی تحت عنوان Wellcome Trust Case Control Consortium تشکیل داده‌است که هدف آن، جمع‌آوری ژنوتیپ‌های مربوط به نمونه‌های case و control در هفت بیماری که منشاء ژنتیکی کمابیش ناشناخته‌ای دارند است. اجرای روش‌های رایج در مطالعه‌ی case-control بر روی این داده‌ها تاکنون نتایج چندان برجسته‌ای به همراه نداشته است. مطالعه‌ی case-control بر مبنای بلوک‌های هاپلوتیپ و استفاده از روش خوشه‌بندی سلسله‌مراتبی برای اجرای آزمون همبستگی، نتایج امیدوارکننده‌ای در استفاده از این روش‌ها برای تشخیص جایگاه ژنی مرتبط با بیماری‌ها بر روی نمونه‌های شبیه‌سازی شده نشان می‌داد. بکارگیری روش پیشنهادی این رساله بر روی نمونه‌های واقعی case-control که توسط این دو موسسه منتشر شده‌اند از دیگر موضوعات مورد تحقیق در آینده است.

پیوست الف

تکوین هاپلوتیپ‌ها بر روی دو اسنیپ، تحت رویدادهای نوترکیبی با نرخ پائین

در این پیوست، شرح یک مدل ساده برای بررسی تغییر تنوع هاپلوتیپ‌ها بر روی دو اسنیپ ارائه می‌شود. این مدل را به طور دقیق می‌توان "مدل تکوینی برای هاپلوتیپ‌های دو جایگاهی با رویدادهای نوترکیبی با نرخ پائین و اندازه‌ی جمعیت ثابت، بدون انقراض ژن"[1] نامید.

فرض کنید دو اسنیپ (دو ژن) بر روی یک کروموزوم در نزدیکی یکدیگر قرار دارند و فرض کنید برای هر یک، دو آلل وجود دارد. می‌خواهیم فراوانی هر یک از چهار هاپلوتیپ متمایز برای دو این دو اسنیپ را در نسلهای متوالی، هنگامی که نوترکیبی بین آنها عمل می‌کند بدست آوریم. یک مدل تئوری ساده با شرایط زیر را در نظر می‌گیریم؛

- اندازه‌ی جمعیت در نسل‌های متوالی تغییر نمی‌کند.
- هر یک از هاپلوتیپ‌های افراد، دقیقا به یک فرد در نسل بعد منتقل می‌شود. به عبارت دیگر، سهم تمام هاپلوتیپ‌ها برای تولید نسل بعد، کاملا برابر است و هیچ هاپلوتیپی در انتقال به نسل بعد از بین نمی‌رود.
- نرخ نوترکیبی در هر نسل در مقایسه با اندازه‌ی جمعیت کوچک است، به طوری که احتمال تشکیل هاپلوتیپ نوترکیب از این دو اسنیپ، در بین تمام زادهای جدید صفر است بجز یک زاد به طور تصادفی،

[1] Two-loci evolutionary model with low rate recombination and fixed population size without gene extinction

که در آن، نوترکیبی به احتمال ρ ممکن است روی دهد.

- هیچ جهشی روی نمی‌دهد.

مطابق با نمادگذاری بخش ۳.۲، فرض کنید در جمعیت اولیه، فراوانی هر یک از چهار هاپلوتیپ 00، 01، 10و 11به ترتیب برابر با n_{00}، n_{01}، n_{10} و n_{11} باشد؛ بدیهی است $n = n_{00} + n_{10} + n_{01} + n_{11}$ تعداد کل هاپلوتیپ‌های موجود در جمعیت است. با مفروضات بالا، نوترکیبی تنها بین یک جفت از افراد جمعیت اولیه به تصادف روی می‌دهد و برای مابقی جمعیت، هاپلوتیپ‌ها بدون نوترکیبی به نسل بعد منتقل می‌شوند. در این جفت، اگر گامت فرد اول 01و گامت فرد دوم 10 (یا بلعکس) باشد در نسل بعد، به احتمال $\frac{۲\rho n_{01} n_{10}}{n(n-۱)}$ یک هاپلوتیپ جدید 00و یک هاپلوتیپ جدید 11داریم و از تعداد هاپلوتیپ‌های 01و 10هر کدام، یک واحد کاسته می‌شود. به طور مشابه، احتمال اینکه در نسل بعد به دلیل نوترکیبی، از تعداد 00و 11هر کدام، یک واحد کاسته شود و یک هاپلوتیپ جدید 01و یک هاپلوتیپ جدید 10بوجود آید برابر است با $\frac{۲\rho n_{00} n_{11}}{n(n-۱)}$. توجه کنید که با شرایط یادشده، تنها ممکن است فراوانی هر یک از هاپلوتیپ‌ها در نسل بعد، حداکثر یک واحد با فراوانی اولیه متفاوت باشد و نه بیشتر، هرچند این تفاوت در نسل‌های بعدتر، می‌تواند بیشتر گردد. بنابراین احتمال اینکه فراوانی هاپلوتیپ‌ها در نسل جدید بدون تغییر بماند برابر است با

$$۱ - \frac{۲\rho}{n(n-۱)}(n_{00}n_{11} + n_{10}n_{01})\,.$$

به سادگی می‌توان دید که تحت این مدل، فراوانی آلل‌ها در هر یک از اسنیپ‌ها در طی نسلهای متوالی، ثابت باقی می‌مانند یعنی مقادیر $n_a = n_{10} + n_{11}$ و $n_b = n_{01} + n_{11}$ در نسل‌های متوالی ثابت هستند. بنابراین با معلوم بودن فراوانی یکی از هاپلوتیپ‌ها در بین چهار هاپلوتیپ 00، 01، 10و 11، می‌توان فراوانی هر یک از سه هاپلوتیپ دیگر را بدست آورد. بدین ترتیب ما علاقمندیم احتمال مشاهده‌ی $P(n_{11} = x)$ را در نسل Tام بدست آوریم. برای این کار، یک ماتریس انتقال احتمال به صورت زیر تعریف می‌کنیم؛

$$A = \left[a_{ij}\right], \qquad i = ۰, \ldots, \min(n_a, n_b)\,, \quad j = ۰, \ldots, \min(n_a, n_b)$$

که در آن، a_{ij} نشان دهنده‌ی احتمال وجود i هاپلوتیپ از نوع 11در نسل بعدی است به شرط آنکه در نسل فعلی، j هاپلوتیپ از همان نوع وجود داشته باشد. به وضوح، تعداد هاپلوتیپ‌های 11هیچگاه نمی‌تواند بیشتر از فراوانی هیچ یک از آلل‌های a و b باشد. با توجه به توضیحات فوق، ماتریس احتمال انتقال مورد بحث، یک ماتریس سه‌قطری است و به عبارت دقیق داریم؛

$$a_{i,i-۱} = \frac{۲\rho(n_a - i + ۱)(n_b - i + ۱)}{n(n-۱)}$$

$$a_{i,i} = ۱ - \frac{۲\rho}{n(n-۱)}\big((n - n_a - n_b + i)i + (n_a - i)(n_b - i)\big)$$

$$a_{i,i+۱} = \frac{۲\rho(n - n_a - n_b + i + ۱)(i+۱)}{n(n-۱)}$$

$$a_{i,j} = ۰ \qquad if \quad |i-j| > ۱$$

مثلا، به ازای $n = ۲۰, n_a = ۸, n_b = ۳$ و $\rho = ۰/۰۱$ داریم؛

$$A = \begin{bmatrix} ۰/۹۹۸۷ & ۰/۰۰۰۵ & ۰ & ۰ \\ ۰/۰۰۱۳ & ۰/۹۹۸۷ & ۰/۰۰۱۲ & ۰ \\ ۰ & ۰/۰۰۰۸ & ۰/۹۹۸۵ & ۰/۰۰۱۹ \\ ۰ & ۰ & ۰/۰۰۰۳ & ۰/۹۹۸۱ \end{bmatrix}.$$

اگر $\pi_T = \langle P(n_{11} = ۰), P(n_{11} = ۱), \ldots, P(n_{11} = \min(n_a, n_b))\rangle$ بردار احتمال مشاهده‌ی هر یک از مقادیر ممکن برای فراوانی هاپلوتیپ 11در نسل T باشد، آنگاه داریم؛

$$\pi_T = A^T \pi_o$$

می‌توان اثبات کرد که در زمان ایستایی، احتمال‌های قرار گرفتن در هر وضعیت این فرایند، مستقل از پارامتر

ρ، یعنی نرخ نوترکیبی است و احتمال مشاهده‌ی هاپلوتیپ 11در جمعیت حدی، برابر است با احتمال فوق هندسی یا آماره‌ی آزمون دقیق فیشر در رابطهٔ (۳۰۲). برای مثال بالا، داریم؛

$$A^* = \lim_{T\to\infty} A^T = \begin{bmatrix} ۰/۱۹۲۹۸ & ۰/۱۹۲۹۸ & ۰/۱۹۲۹۸ & ۰/۱۹۲۹۸ \\ ۰/۴۶۳۱۶ & ۰/۴۶۳۱۶ & ۰/۴۶۳۱۶ & ۰/۴۶۳۱۶ \\ ۰/۲۹۴۷۴ & ۰/۲۹۴۷۴ & ۰/۲۹۴۷۴ & ۰/۲۹۴۷۴ \\ ۰/۰۴۹۱۲۳ & ۰/۰۴۹۱۲۳ & ۰/۰۴۹۱۲۳ & ۰/۰۴۹۱۲۳ \end{bmatrix}.$$

مراجع

[1] Maxam A.M and Gilbert W. A new method for sequencing DNA. *PNAS*, 74 (2):560–564, 1977.

[2] Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551): 500–507, 1976. ISSN 0028-0836. URL `http://www.ncbi.nlm.nih.gov/pubmed/1264203`. PMID: 1264203.

[3] Sanger F, Nicklen S, and Coulson A.R. DNA sequencing with chain-terminating inhibitors. *PNAS*, 74(12):5463–5467, 1977. ISSN 0027-8424. URL `http://www.ncbi.nlm.nih.gov/pubmed/271968`. PMID: 271968.

[4] Sanger F, Air G.M, Barrell B.G, Brown N.L, Coulson A.R, Fiddes C.A, et al. Nucleotide sequence of bacteriophage phi x174 DNA. *Nature*, 265 (5596):687–695, 1977. ISSN 0028-0836. URL `http://www.ncbi.nlm.nih.gov/pubmed/870828`. PMID: 870828.

[5] Fleischmann R.D, Adams M.D, White O, Clayton R.A, Kirkness E.F, KerlavageA.R, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, 1995. ISSN 0036-8075. URL `http://www.ncbi.nlm.nih.gov/pubmed/7542800`. PMID: 7542800.

[6] *C. Elegans* Sequencing Consortium. Genome sequence of the nematode c. elegans: A platform for investigating biology. *Science*, 282:2012–18, 1998.

[7] Lander E and Waterman M. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2:231–239, 1988.

[8] Lander E, Linton L, Birren B, Nusbaum C, Zody M, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. ISSN 0028-0836. doi: 10.1038/35057062. URL `http://www.ncbi.nlm.nih.gov/pubmed/11237011`. PMID: 11237011.

[9] Venter J.C, Adams M.D, Myers E.W, Li P.W, Mural R.J, Sutton G.G, et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001. ISSN 0036-8075. doi: 11181995. URL `http://www.ncbi.nlm.nih.gov/pubmed/11181995`. PMID: 11181995.

[10] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004. ISSN 1476-4687. doi: 10.1038/nature03001. URL http://www.ncbi.nlm.nih.gov/pubmed/15496913. PMID: 15496913.

[11] Levy S, Sutton G, Ng P, Feuk L, Halpern A, Walenz B, et al. The diploid genome sequence of an individual human. *PLoS Biol*, 5(10):e254, 2007. doi: 10.1371/journal.pbio.0050254. URL http://dx.doi.org/10.1371/journal.pbio.0050254.

[12] The International HapMap Consortium. The international hapmap project. *Nature*, 426:789–796, 2003.

[13] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005. ISSN 0028-0836. doi: 10.1038/nature04226. URL http://dx.doi.org/10.1038/nature04226.

[14] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, 2007. ISSN 0028-0836. doi: 10.1038/nature06258. URL http://dx.doi.org/10.1038/nature06258.

[15] Wang D.G, Fan J.B, Siao C.J, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, 1998. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/9582121. PMID: 9582121.

[16] Altshuler D, Pollara V.J, Cowles C.R, Van Etten W.J, Baldwin J, Linton L, et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803):513–516, 2000. ISSN 0028-0836. doi: 10.1038/35035083. URL http://www.ncbi.nlm.nih.gov/pubmed/11029002. PMID: 11029002.

[17] Sherry S.T, Ward M.H, Kholodov M, Baker J, Phan L, Smigielski E.M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29 (1):308–311, 2001. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/11125122. PMID: 11125122.

[18] Morita A, Nakayama T, Doba N, Hinohara S, Mizutani T, and Soma M. Genotyping of triallelic SNPs using TaqMan® PCR. *Molecular and Cellular Probes*, 21(3):171–176, 2007.

[19] Stevens J, Livak K.J, Williams P.M and Heid C.A. Real time quantitative pcr. *Genome Research*, 6(10):986–994, 1996. ISSN 10549803.

[20] Krjutškov K, Andreson R, Mägi R, Nikopensius T, Khrunin A, Mihailov E, et al. Development of a single tube 640-plex genotyping method for detection of nucleic acid variations on microarrays. *Nucleic Acids Research*, 36(12), 2008.

[21] Gunderson K.L, Steemers F.J, Ren H, Ng P, Zhou L, Tsan C, et al. *Whole-Genome Genotyping*, volume 410. 2006.

[22] Bentley D.R, Balasubramanian Sh, Swerdlow H.P, Smith G.P, Milton J, Brown C.G, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008. doi: 10.1038/nature07517. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2581791`. PMC2581791.

[23] *An update to the 2008 Illumina Product Guide*. URL `http://www.solexa.com/downloads/Illumina2008ProductGuideUpdate.pdf`.

[24] Denisov G, Walenz B, Halpern A, Miller J, Axelrod N, Levy S, et al. Consensus generation and variant detection by Celera assembler. *Bioinformatics*, 24 (8):1035–1040, 2008.

[25] Green P. PHRAP documentation: Algorithms. Phred/Phrap/Consed System Home Page, 2002. URL `http://www.phrap.org`.

[26] Batzoglou S, Jaffe D.B, Stanley K, Butler J, Gnerre S, Mauceli E, et al. Arachne: a whole-genome shotgun assembler. *Genome Research*, 12(1): 177–89, 2002. ISSN 1088-9051. doi: 11779843. URL `http://www.ncbi.nlm.nih.gov/pubmed/11779843`. PMID: 11779843.

[27] Pevzner P.A, Tang H, and Waterman M.S. An Eulerian path approach to DNA fragment assembly. volume 98, pages 9748–9753, 2001.

[28] Zerbino D.R and Birney E. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–9, 2008. ISSN 1088-9051. doi: gr.074492.107. URL `http://www.ncbi.nlm.nih.gov/pubmed/18349386`. PMID: 18349386.

[29] Warren R.L, Sutton G.G, Steven J, Jones M, and Holt R.A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23(4):500–1, 2007. ISSN 1460-2059. doi: btl629. URL `http://www.ncbi.nlm.nih.gov/pubmed/17158514`. PMID: 17158514.

[30] Dohm J.C, Lottaz C, Borodina T, and Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Research*, 17(11):1697–706, 2007. ISSN 1088-9051. doi: gr.6435207. URL `http://www.ncbi.nlm.nih.gov/pubmed/17908823`. PMID: 17908823.

[31] Rumble S.M, Lacroute P, Dalca A.V, Fiume M, Sidow A, and Brudno M. SHRiMP: accurate mapping of short color-space reads. *PLoS Computational Biology*, 5(5):e1000386, 2009.

[32] Li R, Li Y, Kristiansen K, and Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[33] Kent W.J . BLAT - the BLAST-like alignment tool. *Genome Research*, 12(4): 656–664, 2002.

[34] Tishkoff S.A, Bentley K.L, Kidd K.K, Ruano G, Michalatos B.S. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range pcr. *Nucleic Acids Research*, 24(23):4841–4843, 1996.

[35] Rogers J, Ruano G, Stephens J.C. Theoretical underpinning of the single-molecule-dilution (smd) method of direct haplotype resolution. *American Journal of Human Genetics*, 46(6):1149–1155, 1990.

[36] Boehnke M, Gillanders E, Trent J.M, Gruber S.B, Douglas J.A. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nature Genetics*, 28(4):361–364, 2001.

[37] Guillemette C, Cheung C.L, Housman D.E, Lieber C.M, Woolley A.T. Direct haplotyping of kilobase-size dna using carbon nanotube probes. *Nature Biotechnology*, 18(7):760–763, 2000.

[38] Holmberg K, Eriksson P, Uhlén M, Odeberg J. Molecular haplotyping by pyrosequencing™. *BioTechniques*, 33(5):1104–1108, 2002.

[39] Krynetski E.Y, Evans W.E, McDonald O.G. Molecular haplotyping of genomic dna for multiple single-nucleotide polymorphisms located kilobases apart using long-range polymerase chain reaction and intramolecular ligation. *Pharmacogenetics*, 12(2):93–99, 2002.

[40] Lizardi P.M, Huang X.H, Bray-Ward P.L, Ward D.C, Zhong X.B. Visualization of oligonucleotide probes and point mutations in interphase nuclei and dna fibers using rolling circle dna amplification. *PNAS*, 98(7):3940–3945, 2001.

[41] Kepper P, Hoehe M, Schmitt C, Reinhardt R, Lehrach H, Sauer S and Burgtorf C. Clone-based systematic haplotyping (csh): A procedure for physical haplotyping of whole genomes. *Genome Research*, 13(12):2717–2724, 2003.

[42] Clark A.G. Inference of haplotypes from pcr-amplified samples of diploid populations. *Molecular Biology and Evolution*, 7(2):111–122, 1990.

[43] Gusfield D. Inference of haplotypes from samples of diploid populations: Complexity and algorithms. *Journal of Computational Biology*, 8(3):305–323, 2001.

[44] Hubbell E. Finding a parsimony solution to haplotype phase is NP-hard, 2002. Personal communication.

[45] Pinotti M.C, Rizzi R. and Lancia G. Haplotyping populations by pure parsimony: Complexity of exact and approximation algorithms. *INFORMS Journal on Computing*, 16(4):348–359, 2004.

[46] Rizzi R. and Lancia G. A polynomial case of the parsimony haplotyping problem. *Operations Research Letters*, 34(3):289–295, 2006.

[47] Gusfield D. Haplotype inference by pure parsimony. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2676:144–155, 2003.

[48] Xu Y. and Wang L. Haplotype inference by maximum parsimony. *Bioinformatics*, 19(14):1773–1780, 2003.

[49] Steel M. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9(1):91–116, 1992.

[50] Gusfield D. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1 edition, 1997. ISBN 0521585198.

[51] Climer Sh, Jager G, Templeton A.R. and Zhang W. How frugal is mother nature with haplotypes? *Bioinformatics*, 25(1):68–74, 2008. doi: 10.1093/bioinformatics/btn572. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btn572v1`.

[52] Hedrick. *Genetics of Populations*. Jones & Bartlett Publishing Co, 3 edition, 2004. ISBN 0763747726.

[53] Gusfield D. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. *Proceedings of the sixth annual international conference on Computational biology*, pages 166–175, 2002.

[54] Gusfield D, Lancia G. and Bafna Y.S. Haplotyping as perfect phylogeny: A direct approach. *Journal of Computational Biology*, 10(3-4):323–340, 2003.

[55] Gusfield D, and Chung R.H. Perfect phylogeny haplotyper: Haplotype inferral using a tree model. *Bioinformatics*, 19(6):780–781, 2003.

[56] Gusfield D. and Orzack Sh. *CRC Handbook in Bioinformatics, chapter 1. Haplotype Inference*. CRC Press, 2005.

[57] Filkov V, Gusfield D. and Ding Z. A linear-time algorithm for the perfect phylogeny haplotyping problem. *Journal of Computational Biology*, 13 (2):522–553, 2006.

[58] Halperin E, Karp R.M and Eskin E. Large scale reconstruction of haplotypes from genotype data. *Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB*, pages 104–113, 2003.

[59] Karp R.M and Halperin E. Perfect phytogeny and haplotype assignment. *Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB*, 8:10–19, 2004.

[60] Eskin E. and Halperin E. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–1849, 2004.

[61] Halperin E, Karp R.M and Eskin E. Efficient reconstruction of haplotype structure via perfect phylogeny. *J Bioinform Comput Biol*, 1(1):1–20, 2003.

[62] Slatkin M. and Excoffier L. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5):921–927, 1995.

[63] Kidd K.K and Hawley M.E. Haplo: a program using the em algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, 86(5): 409–411, 1995.

[64] Williams R.C, Urbanek M. and Long J.C. An e-m algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, 56(3):799–810, 1995.

[65] Qin Z.S, Niu T. and Liu J.S. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics*, 71(5):1242–1247, 2002.

[66] Cutler D.J, Zwick M.E, Chakravarti A. and Lin S. Haplotype inference in random population samples. *American Journal of Human Genetics*, 71(5): 1129–1137, 2002.

[67] Zhao Y, Xu Y, Wang Zh, Zhang H. and Chen G. A better block partition and ligation strategy for individual haplotyping. *Bioinformatics*, 24 (23):2720–2725, 2008. doi: 10.1093/bioinformatics/btn519. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btn519v1`.

[68] Schork N.J and Fallin D. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics*, 67(4):947–959, 2000.

[69] Kelly E.D, Sievers F. and McManus R. Haplotype frequency estimation error analysis in the presence of missing genotype data. *BMC Bioinformatics*, 5 (188), 2004. URL `http://www.biomedcentral.com/1471-2105/5/188`.

[70] Niu T. Algorithms for inferring haplotypes. *Genetic Epidemiology*, 27(4): 334–347, 2004.

[71] Rannala B. and Beaumont M.A. The bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–261, 2004.

[72] Smith N.J, Donnelly P. and Stephens M. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68(4):978–989, 2001.

[73] Scheet P. and Stephens M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, 78(4):629–644, 2006.

[74] Delaneau O, Coulonges C. and Zagury J. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, 9(540), 2008.

[75] Zhang Y, Niu T. and Liu J.S. A coalescence-guided hierarchical Bayesian method for haplotype inference. *American Journal of Human Genetics*, 79: 313–322, 2006.

[76] Qin Z.S, Xu X, Liu J.S and Niu T. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, 70(1):157–169, 2002.

[77] Lawrence C.E, Altschul S.F, Boguski M.S, Liu J.S, Neuwald A.F and Wootton J.C. Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[78] Xing E.P, Jordan M.I and Sharan R. Bayesian haplotype inference via the dirichlet process. *Journal of Computational Biology*, 14(3):267–284, 2007.

[79] Donnelly P. and Stephens M. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, 73(5):1162–1169, 2003.

[80] Brinza D. and Zelikovsky A. 2SNP: scalable phasing method for trios and unrelated individuals. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(2):313–318, 2008.

[81] Braaten O, Rødningen O.K, Nordal I. and Leren T.P. The genetic algorithm applied to haplotype data at the LDL receptor locus. *Computer Methods and Programs in Biomedicine*, 61(1):1–9, 2000.

[82] Tapadar P, Ghosh S. and Majumder P.P. Haplotyping in pedigrees via a genetic algorithm. *Human Heredity*, 50(1):43–56, 2000.

[83] Azuma R, Sakamoto M. and Furutani H. Haplotype estimation from genotypical data by genetic algorithm. *Artificial Life and Robotics*, 13(2):535–537, 2009.

[84] Xu X, Ma J. and Wang J. A hopfield-type neural network for haplotype assembly problem. In *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, volume 5, pages 8–12, 2008.

[85] Lippert R, Schwartz R, Lancia G. and Istrail S. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in bioinformatics*, 3(1):23–31, 2002.

[86] Cilibrasi R, Van Iersel L, Kelk S. and Tromp J. *On the complexity of several haplotyping problems*, volume 3692 LNBI, pages 128–139. 2005.

[87] Wu L.Y, Li Z, Wang, R.S, Zhang X.S. and Chen L. Self-organizing map approaches for the haplotype assembly problem. In *Mathematics and Computers in Simulation*, 2009.

[88] Asgarian E, Moeinzadeh M.H, Habibi J, Sharifian S.R. and Najafi A. Solving haplotype reconstruction problem in MEC model with hybrid information fusion. In *Proceedings - EMS 2008, European Modelling Symposium, 2nd UKSim European Symposium on Computer Modelling and Simulation*, pages 214–218, 2008.

[89] Wang Y, Feng E, Wang R. and Zhang D. The haplotype assembly model with genotype information and iterative local-exhaustive search algorithm. *Computational Biology and Chemistry*, 31(4):288–293, 2007.

[90] Bansal V, Halpern A.L, Axelrod N. and Bafna V. An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Research*, 18(8):1336–1346, 2008.

[91] Qian W, Yang Y, Yang N. and Li C. Particle swarm optimization for SNP haplotype reconstruction problem. *Applied Mathematics and Computation*, 196 (1):266–272, 2008.

[92] Bansal V. and Bafna V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, 24(16), 2008.

[93] Kang S.H, Jeong I.S, Choi M.H. and Lim H.S. *Haplotype assembly from weighted SNP fragments and related genotype information*, volume 5059 LNCS, pages 45–54. 2008.

[94] Wu J, Wang J. and Chen J. A genetic algorithm for single individual SNP haplotype assembly. In *Proceedings of the 9th International Conference for Young Computer Scientists, ICYCS 2008*, pages 1012–1017, 2008.

[95] Moeinzadeh M.H, Asgarian E, Sharifian S.R, Najafi-Ardabili A. and Mohammadzadeh J. Neural network based approaches, solving haplotype reconstruction in MEC and MEC/GI models. In *Proceedings - 2nd Asia International Conference on Modelling and Simulation, AMS 2008*, pages 934–939, 2008.

[96] Wang R.S, Wu L.Y, Li Z.P. and Zhang X.S. Haplotype reconstruction from SNP fragments by minimum error correction. *Bioinformatics*, 21(10): 2456–2462, 2005.

[97] Patil N, Berno A.J, Hinds D.A, Barrett W.A, Doshi J.M, Hacker C.R, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723, 2001.

[98] Daly M.J, Rioux J.D, Schaffner S.F, Hudson T.J. and Lander E.S. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29 (2):229–32, 2001. ISSN 1061-4036. doi: 11586305. URL `http://www.ncbi.nlm.nih.gov/pubmed/11586305`. PMID: 11586305.

[99] Schwartz R, Halldórsson B.V, Bafna V, Clark A.G. and Istrail S. Robustness of inference of haplotype block structure. *Journal of Computational Biology*, 10(1):13–19, 2003.

[100] Indap A.R, Marth G.T, Struble C.A, Tonellato P. and Olivier M. Analysis of concordance of different haplotype block partitioning algorithms. *BMC Bioinformatics*, 6, 2005.

[101] Evans D.M and Cardon L.R. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *American Journal of Human Genetics*, 76(4):681–687, 2005.

[102] Gu S, Pakstis A.J, Li H, Speed W.C, Kidd J.R and Kidd K.K. Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations. *European Journal of Human Genetics*, 15(3): 302–312, 2007.

[103] Thompson E. and Chapman N. *Haplotype blocks in small populations*, volume 2983 of *Lecture Notes in Computer Science*, pages 74–83. Springer Berlin / Heidelberg, 2004. doi: 10.1007/b96286.

[104] Schmegner C, Hoegel J, Vogel W. and Assum G. Genetic variability in a genomic region with long-range linkage disequilibrium reveals traces of a bottleneck in the history of the european population. *Human Genetics*, 118 (2):276–286, 2005.

[105] Kimmel G. and Shamir R. GERBIL: genotype resolution and block identification using likelihood. *PNAS*, 102(1):158–162, 2005.

[106] Lin Y.L. *Efficient algorithms for SNP haplotype block selection problems*, volume 5092 LNCS. 2008.

[107] Zhang K. and Jin L. Haploblockfinder: Haplotype block analyses. *Bioinformatics*, 19(10):1300–1301, 2003.

[108] Zhang K, Qin Z, Chen T, Liu J.S, Waterman M.S. and Sun F. Hapblock: Haplotype block partitioning and tag snp selection software using a set of dynamic programming algorithms. *Bioinformatics*, 21(1):131–134, 2005.

[109] Anderson E. and Novembre J. Finding haplotype block boundaries by using the minimum-description-length principle. *The American Journal of Human Genetics*, 73(2):336–354, 2003.

[110] Mannila H, Koivisto M, Perola M, Varilo T, Hennah W, Ekelund J, et al. Minimum description length block finder, a method to identify haplotype blocks and to compare the strength of block boundaries. *American Journal of Human Genetics*, 73(1):86–94, 2003.

[111] Lewontin R. and Kojima K. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472, 1960. ISSN 00143820. URL `http://www.jstor.org/stable/2405995`.

[112] Abecasis G.R and Cookson W.O. GOLD - graphical overview of linkage disequilibrium. *Bioinformatics*, 16(2):182–183, 2000.

[113] Pettersson F, Morris A.P, Barnes M.R. and Cardon L.R. Goldsurfer2 (Gs2): a comprehensive tool for the analysis and visualization of genome wide association studies. *BMC Bioinformatics*, 9, 2008.

[114] Barrett J.C, Fry B, Maller J. and Daly M.J. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005. doi: 10.1093/bioinformatics/bth457. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/2/263`.

[115] Gabriel S.B, Schaffner S.F, Nguyen H, Moore J.M, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. *Science*, 296(5576): 2225–2229, 2002.

[116] Wang N, Akey J.M, Zhang K, Chakraborty R. and Jin L. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *American Journal of Human Genetics*, 71(5):1227–34, 2002. ISSN 0002-9297. doi: 12384857. URL `http://www.ncbi.nlm.nih.gov/pubmed/12384857`. PMID: 12384857.

[117] Hinds D.A, Stuve L.L, Nilsen G.B, Halperin E, Eskin E, Ballinger D.G, et al. Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–1079, 2005.

[118] Conrad D.F, Jakobsson M, Coop G, Wen X, Wall J.D, Rosenberg N.A, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics*, 38(11):1251–1260, 2006.

[119] Yu Z, Garner C, Ziogas A, Anton-Culver N.A and Schaid D.J. Genotype determination for polymorphisms in linkage disequilibrium. *BMC Bioinformatics*, 10(63), 2009.

[120] Smith A.V, Thomas D.J, Munro H.M and Abecasis G.R. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Research*, 15 (11):1519–1534, 2005.

[121] Myers S, Spencer C.C, Auton A, Bottolo L, Freeman C, Donnelly P, et al. The distribution and causes of meiotic recombination in the human genome. *Biochemical Society Transactions*, 34(4):526–530, 2006.

[122] Myers S, Bottolo L, Freeman C, McVean G. and Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310 (5746):321–324, 2005.

[123] Wiuf C. and Posada D. A coalescent model of recombination hotspots. *Genetics*, 164(1):407–17, 2003. ISSN 0016-6731. doi: PMC1462539. URL `http://www.ncbi.nlm.nih.gov/pubmed/12750351`. PMID: 12750351.

[124] Greenspan G. and Geiger D. High density linkage disequilibrium mapping using models of haplotype block variation. *Bioinformatics*, 20(SUPPL. 1): i137–i144, 2004.

[125] McVean G, Myers S, Hunt S, Deloukas P, Bentley D. and Donnelly P. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–4, 2004. ISSN 1095-9203. doi: 15105499. URL `http://www.ncbi.nlm.nih.gov/pubmed/15105499`. PMID: 15105499.

[126] Fearnhead P, Harding R.M, Schneider J.A, Myers S. and Donnelly P. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, 167(4):2067–2081, 2004.

[127] Fearnhead P. and Smith N. A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *American Journal of Human Genetics*, 77(5):781–794, 2005.

[128] Fearnhead P. Sequenceldhot: Detecting recombination hotspots. *Bioinformatics*, 22(24):3061–3066, 2006.

[129] Li J, Zhang M.Q and Zhang X. A new method for detecting human recombination hotspots and its applications to the HapMap ENCODE data. *American Journal of Human Genetics*, 79(4):628–639, 2006.

[130] Carlson C.S, Eberle M.A, Rieder M.J, Yi Q, Kruglyak L. and Nickerson D.A. Selecting a maximally informative set of Single-Nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics*, 74(1):106–120, 2004.

[131] Meng Z, Zaykin D, Xu C.F, Wagner M. and Ehm M. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *American Journal of Human Genetics*, 2003.

[132] Horne D.B. and Camp N.J. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genetic Epidemiology*, 26:11–21, 2004.

[133] Ding K, Zhou K, Zhang J, Knight J, Zhang X. and Shen Y. The effect of haplotype-block definitions on inference of haplotype-block structure and htsnps selection. *Molecular Biology and Evolution*, 22(1):148–159, 2005.

[134] Ding K, Zhang J, Zhou K, Shen Y. and Zhang X. htSNPer1.0: software for haplotype block partition and htsnps selection. *BMC Bioinformatics*, 6:38, 2005. ISSN 1471-2105. doi: 1471-2105-6-38. URL `http://www.ncbi.nlm.nih.gov/pubmed/15740612`. PMID: 15740612.

[135] Halldórsson B.V, Istrail S. and De La Vega F. Optimal selection of SNP markers for disease association studies. *Human Heredity*, 58(3-4):190–202, 2004.

[136] Hampe J, Schreiber S. and Krawczak M. Entropy-based SNP selection for genetic association studies. *Human Genetics*, 114:36–43, 2003.

[137] Edlund C, Lee W, Li D, Van Den Berg D.J and Conti D. Snagger: A user-friendly program for incorporating additional information for tagsnp selection. *BMC Bioinformatics*, 9:174, 2008. doi: 10.1186/1471-2105-9-174. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2375134.

[138] Stram D.O, Haiman C.A, Hirschhorn J.N, Altshuler D, Kolonel L.N, Henderson B.E, et al. Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Human Heredity*, 55(1):27–36, 2003.

[139] Mailund T. Association mapping: Fundamental principles and applicationsd. adapted from Lon Cardon's website. URL http://psb.stanford.edu/psb06/presentations/association_mapping.pdf.

[140] Laird N. and Lange C. Family-based designs in the age of large-scale gene-association studies. *Nature Review Genetics*, 7:385–394, 2006.

[141] Akey J, Jin L. and Xiong M. Haplotypes vs single marker linkage disequilibrium tests: What do we gain? *European Journal of Human Genetics*, 9(4): 291–300, 2001.

[142] Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, et al. Genetic analysis of case/control data using estimated haplotype frequencies: Application to APOE locus variation and alzheimer's disease. *Genome Research*, 11(1):143–151, 2001.

[143] Schaid D.J. Evaluating associations of haplotypes with traits. *Genetic Epidemiology*, 27(4):348–364, 2004.

[144] Van der Meulen M. and Te Meerman G.J. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genetic Epidemiology*, 14:915–920, 1997.

[145] Tzeng J.Y, Devlin B, Wasserman L. and Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *American Journal of Human Genetics*, (4):891–902, 2003.

[146] Durrant C, Zondervan K.T, Cardon L.R, Hunt S, Deloukas P. and Morris A.P. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *The American Journal of Human Genetics*, 75(1):35–43, 2004. doi: 10.1086/422174. URL http://www.sciencedirect.com/science/article/B8JDD-4RDBHSS-6/2/e0f77477d006b22756300a72f991d8c6.

[147] Bardel C, Darlu P. and Genin E. Clustering of haplotypes based on phylogeny: how good a strategy for association testing? *Eur J Hum Genet*, 14 (2):202–206, 2005. ISSN 1018-4813. URL http://dx.doi.org/10.1038/sj.ejhg.5201501.

[148] Liu N, Zhang K. and Zhao H. *Haplotype-Association Analysis*, volume 60, pages 335–405. 2008.

[149] Goldberg D.E. *Genetic Algorithms in Search, Optimzation and Machine Learning*. Addison Wesley, 1989.

[150] Horan M, Millar D.S, Hedderich J, Lewis G, Newsway V, Mo N, et al. *Human Mutation*, 21(4):408–423, 2003. doi: 10.1002/humu.10167. URL `http://dx.doi.org/10.1002/humu.10167`.

[151] Zapata C. and Alvarez G. On fisher's exact test for detecting gametic disequilibrium between DNA polymorphisms. *Annals of Human Genetics*, 61(01): 69–75, 1997. doi: 10.1017/S0003480096005969.

[152] Hasselblad V. and Lokhnygina Y. Tests for 2 × 2 tables in clinical trials. *Journal of Modern Applied Statistical Methods*, 6(2):456–468, 2007.

[153] Yates F. Test of significance for 2 x 2 contingency tables. *Journal of the Royal Statistical Society*, 147(3):426–463, 1984.

[154] Hwang J.T.G and Yang M.C. An optimality theory for mid p-values in 2 × 2 contingency tables. *Statistica Sinica*, 11(3):807–826, 2001.

[155] Lydersen S. and Laake P. Power comparison of two-sided exact tests for association in 2 × 2 contingency tables using standard, mid p, and randomized test versions. *Statistics in Medicine*, 22(24):3859–3871, 2003.

[156] Pattaro C, Ruczinski I, Fallin D. and Parmigiani G. Haplotype block partitioning as a tool for dimensionality reduction in snp association studies. *BMC Genomics*, 9(1):405, 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-405. URL `http://www.biomedcentral.com/1471-2164/9/405`.

[157] Katanforoush A, Sadeghi M, Pezeshk H. and Elahi E. Global haplotype partitioning for maximal associated SNP pairs. *BMC Bioinformatics*, 10(1): 269, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-269. URL `http://www.biomedcentral.com/1471-2105/10/269`.

[158] Feingold E.A, Good P.J, Guyer M.S, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*, 306 (5696):636–640, 2004.

[159] Hellenthal G. and Stephens M. msHOT: modifying hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, 23(4): 520–1, 2007. ISSN 1460-2059. doi: btl622. URL `http://www.ncbi.nlm.nih.gov/pubmed/17150995`. PMID: 17150995.

[160] Hudson R. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–8, 2002. PMID: 11847089.

[161] Li Q, Yu K, Li Z. and Zheng G. Max-rank: A simple and robust genome-wide scan for case-control association studies. *Human Genetics*, 123(6):617–623, 2008.

[162] Li J. and Chen Y. Generating samples for association studies based on HapMap data. *BMC Bioinformatics*, 9:44, 2008.

[163] Coulonges C, Delaneau O, Girard M, Do H, Adkins R, Spadoni JL, et al. Computation of haplotypes on snps subsets: advantage of the "global method". *BMC genetics*, 7:50, 2006. PMID: 17067372.

Computational Problems in Haplotype Recognition

Abstract

Recently, modern technologies enable us to get access to the large amount of data of Single Nucleotide Polymorphism (SNP). Haplotypes, as the SNP genotypes in haploid phase, provide useful materials for genetic analyses. Two primary processes in computational haplotype study are to phase genotype data into haplotypes and to partition a chromosome into blocks based on haplotype samples of unrelated individuals.

For problem of genotype phasing, we introduce a family of greedy procedures as a general approach finding nearly the best optimal solutions for the phasing problem under the model of maximum parsimony. Then we develop a hybrid method to infer haplotypes from genotype data by incorporating the greedy procedure into a Genetic Algorithm.

Global partitioning based on pairwise associations of SNPs has not previously been used to define haplotype blocks within genomes. Here, we define an association index based on LD between SNP pairs. We use the Fisher's exact test to assess the statistical significance of the LD estimator. By this test, each SNP pair is characterized as associated, independent, or not-statistically-significant. We set limits on the maximum acceptable proportion of independent pairs within all blocks and search for the partitioning with maximal proportion of associated SNP pairs. Essentially, this model is reduced to a constrained optimization problem, the solution of which is obtained by iterating a dynamic programming algorithm.

We also introduce new assessment protocols to evaluate performance of haplotype block partitioning methods for different aspects and applications, including a definition of similarity for two block partitionings, a simulation process to assess the robustness of a block definition, the application in hotspots detection and an application in disease association studies.

Results of the proposed Genetic Algorithm cannot compete neither with the

number of inferred haplotypes obtained by previously developed algorithms nor with the inference accuracy, but they are quite near to parsimonious haplotypes, just for the case of small SNP samples. Instead, performance of the proposed haplotype block partitioning algorithm is quite comparable with and even better than other methods.

In comparison with other block partitioning methods, our algorithm reports blocks of larger average size. Nevertheless, the haplotype diversity within the blocks is captured by a small number of tagSNPs. Resampling HapMap haplotypes under a block-based model of recombination shows that our algorithm is robust in reproducing the same partitioning for recombinant samples. Our algorithm performed better than previously reported models in a case-control association study aimed at mapping a single locus trait, based on simulation results that were evaluated by a block-based statistical test. Compared to methods of haplotype block partitioning, our algorithm performs best on detection of recombination hotspots.

Keywords: *Genotype, Haplotype, SNP, Statistical Genetics, Population Genetics, Fisher's association test, Linear programming, Dynamic programming, Block-like structure of chromosome, Recombination rate, Disease models, Disease locus recognition, Case-control study.*

University of Tehran

Institute of Biochemistry and Biophysics

Computational Problems in Haplotype Recognition

by

Ali Katanforoush

Under supervision of

Dr. Mehdi Sadeghi AND ***Dr. Hamid Pezeshk***

A thesis submitted to the Graduate Studies Office
In partial fulfillment of the requirements for
the degree of

Doctor of Philosophy in Bioinformatics

Oct 2009