تخمین هاپلوتایپ با استفاده از ریلکس‌سازی بهینه‌سازی چندجمله‌ای

سینا مجیدیان[۱]، دانشجوی دکتری؛ محمدحسین کهائی[۲]، دانشیار

۱- دانشکده مهندسی برق- دانشگاه علم و صنعت ایران- تهران- ایران– s_majidian@elec.iust.ac.ir
۲- دانشکده مهندسی برق- دانشگاه علم و صنعت ایران- تهران- ایران – kahaei@iust.ac.ir

چکیده: این مقاله به بررسی تخمین هاپلوتایپ با استفاده از داده‌های توالی DNA می‌پردازد. الگوریتم پیشنهادی با استفاده از ریلکس‌سازی بهینه‌سازی چندجمله‌ای به روش Lasserre با نام HapLas معرفی می‌شود. این الگوریتم برپایه استفاده از ساختار گسسته مساله بهینه‌سازی تخمین هاپلوتایپ می‌باشد که با استفاده از تئوری اندازه به یک فضای پیوسته نگاشت می‌گردد. سپس با استفاده از خواص ماتریس ممان، ریلکس‌سازی انجام می‌گیرد. نتایج شبیه‌سازی نشان می‌دهد که استفاده از الگوریتم پیشنهادی منجر به بهبود نرخ بازسازی هاپلوتایپ در مقایسه با الگوریتم‌های متداول SDhaP و RefHap در حدود ۵ درصد می‌گردد. این بهبود به‌ازای افزایش قابل ملاحظه زمان اجرا و پیچیدگی محاسبات حاصل می‌شود به‌طوری که در کاربردهای پزشکی قابل صرف‌نظرکردن است.

واژه‌های کلیدی: هاپلوتایپ، تخمین، بهینه‌سازی، ریلکس‌سازی، ماتریس مثبت معین، تئوری اندازه.

Haplotype Estimation Using Polynomial Optimization Relaxation

Sina Majidian[1], PhD student; Mohammad Hossein Kahaei[2], Associate Professor

1- School of Electrical Engineering, Iran University of Science & Technology, Tehran, Iran, Email: s_majidian@elec.iust.ac.ir
2- School of Electrical Engineering, Iran University of Science & Technology, Tehran, Iran, Email: kahaei@iust.ac.ir

Abstract: This paper is dedicated to investigating haplotype estimation based on DNA sequences. The proposed algorithm named as (HapLas) is introduced using relaxation of polynomial optimization based on the Lasserre technique. This algorithm is based on the discrete structure of optimization problem of haplotype estimation which is mapped to a continuous space using the measure theory. Then, relaxation is performed via the properties of the moment matrix. Simulation results show that the proposed algorithm improves the reconstruction rate of the haplotype about five percent in comparison to the SDhaP and RefHap. This is achieved at the cost of increasing the running time and computational complexity which can practically be ignored in medicine.

Keywords: Haplotype, estimation, optimization, relaxation, positive definite matrix, measure theory.

تاریخ ارسال مقاله: ۱۳۹۷/۰۳/۳۱
تاریخ اصلاح مقاله: ۱۳۹۷/۰۸/۱۹
تاریخ پذیرش مقاله: ۱۳۹۷/۱۱/۲۱
نام نویسنده مسئول: محمدحسین کهائی
نشانی نویسنده مسئول: دانشکده مهندسی برق، دانشگاه علم و صنعت ایران، تهران، ایران.

۱- مقدمه

در سال‌های اخیر حل مسائل پزشکی با استفاده از داده‌های زیستی بسیار متداول شده‌است. باتوجه‌به حجم عظیم داده‌های دردسترس و پیچیدگی ذاتی آن‌ها، تحقیقات بین رشته‌ای توسط محققین از رشته‌های مهندسی از جمله علوم ریاضی، برق و کامپیوتر سهم به‌سزایی در حل چالش‌های این حوزه دارند [۲-۱]. در ابتدا، مقدمه‌ای زیستی، برای آشنایی با مفاهیم پایه‌ای در این حوزه مطرح می‌گردد.

موجودات زنده از لحاظ ظاهری و اندازه بسیار متفاوت هستند، اما تمام آن‌ها از واحدهایی به نام سلول ساخته شده‌اند. در سلول، چهار نوع ماده‌ی شیمیایی اصلی وجود دارد، که عبارتند از کربوهیدرات‌ها، لیپیدها (چربی‌ها)، پروتئین‌ها و نوکلئیک اسیدها. پروتئین‌ها از واحدهایی به نام آمینواسید ساخته می‌شوند. آمینواسیدها در طبیعت در ۲۰ نوع مختلف ظاهر شده‌اند که همگی شامل گروه اسیدی کربوکسیل و یک گروه آمینو می‌باشند. این گروه‌ها با یک کربن به‌یک‌دیگر متصل هستند. آمینواسیدها در پروتئین توسط پیوندی به نام پیوند پپتیدی به‌یک‌دیگر متصل هستند. نقش پروتئین در حیات بسیار ضروری است. ازآنجایی‌که پروتئین می‌تواند با مولکول‌های دیگر پیوند ایجاد کند، عمل‌کردهای متنوعی در سلول دارد. از آن جمله می‌توان به کاتالیزور (تسریع‌کننده‌ی واکنش‌های شیمیایی و فرآیندهای زیستی)، دریافت‌کننده سیگنال، راه‌اندازی واکنش شیمیایی و هم‌چنین نقش موتوری نام برد [۳].

۱-۱- معرفی DNA

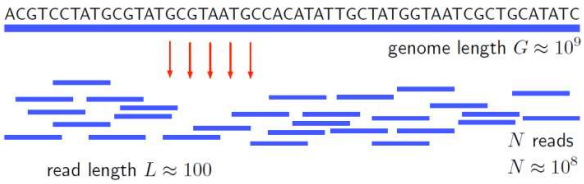
نوکلئیک اسید[۱] یکی از اجزا تشکیل‌دهنده هسته‌ی سلول انسان است که بر دو نوع ریبونوکلئیک اسید (RNA) و دئوکسی‌ریبونوکلئیک اسید (DNA) یافت می‌شود. DNA وظیفه‌ی ذخیره‌ی اطلاعات ژنتیکی را برعهده دارد. درواقع، عاملی که باعث انتقال خصوصیات و ویژگی‌های یک نوع جاندار، از نسلی به نسل دیگر می‌شود، ماده ژنتیک نام دارد. این ماده، شامل اطلاعات و دستورالعمل‌های تمام فعالیت‌های جاندار است و شامل تمامی دستورها برای ایجاد پروتئین‌ها به‌منظور انجام تمام وظایف سلولی می‌باشد. به بیان دیگر، اطلاعات بنیادی در رابطه با نحوه‌ی زندگی یک موجود زنده در DNA قرار دارد. از نظر ساختار فضایی، DNA به صورت کروموزم در سلول قرار دارد. هر کروموزم ساختمان درهم پیچیده و فشرده‌شده‌ی یک پلی‌مر بسیار بلند است، که در انسان‌ها تعداد دو سری ۲۳تا (۴۶ عدد) می‌باشد. به جاندارانی مانند انسان که دونسخه کروموزم دارند، دیپلوئید می‌گویند. هرکدام از این دونسخه کروموزم طی تولیدمثل از هر یک از والدین به ارث رسیده‌است که به آن‌دو، نسخه پدری و مادری می‌گویند. کروموزم‌های متناظر از هر جفت را کروموزم‌های همسان[۲] گویند. این همسانی، به معنی دقیقاً یکسان نیست، بلکه بدین معنی است که ممکن است قسمت کوچکی از یک کروموزم خاص پدری در یک فرد با نسخه‌ی مادری همان فرد متفاوت باشد [۳].

هر قسمتی از کروموزم متناظر با عمل‌کرد یا ویژگی ظاهری افراد مانند رنگ چشم، اندازه، قد و ... است، که ژن نام دارد. DNA از دو رشته ی طولانی تشکیل شده‌است، که حول محوری به دور یک‌دیگر پیچیده شده‌اند. به این ساختار مارپیچ یا نردبانی نیز گفته می‌شود. هر یک از رشته‌ها، به‌صورت پلی‌مر است، بدین‌معنا که از واحدهای کوچک‌تر تکرارشونده‌ای به نام نوکلئوتید ساخته شده‌است. نوکلئوتید، یکی از چهار ماده آلی به نام‌های آدنین[۳]، سیتوزین[۴]، گوانین[۵] و تمین[۶] می‌باشد، که به اختصار و به ترتیب با حروف A، C، G و T نمایش داده می‌شود. ترکیب مختلف کنار هم قرارگرفتن این چهار نوکلئوتید، تنوع زیستی قابل ملاحظه‌ای را در جهان به وجود آورده‌است. برای مثال یک رشته ATCCCCGGACCCT می‌تواند توالی بخشی از DNA یک انسان باشد.

۱-۲- مروری بر روش‌های توالی‌یابی

از کل سه میلیارد نوکلئوتید در هر سلول انسان، در حدود ۹۹/۹ درصد آن‌ها، در بین همه‌ی انسان‌ها یکسان است. در ساختار مارپیچی مولکول DNA، نحوه‌ی قرارگرفتن نوکلئوتیدهای دو رشته مقابل هم به‌گونه‌ای است که همیشه آدنین مقابل تمین و گوانین مقابل سیتوزین قرار بگیرد. در سال ۲۰۰۱ اولین بار کل توالی DNA انسان توسط پژوهشگران به دست آمده‌است. بعد از استخراج DNA از یک سلول، به مشخص‌نمودن توالی نوکلئوتیدهای رشته‌ی DNA توالی‌یابی گویند.

از جمله روش‌های مختلف توالی‌یابی می‌توان از سنگر[۷]، سنگر خودکار و نسل بعدی توالی‌یابی (NGS[۸]) نام برد. هم‌اکنون روش NGS بسیار متداول می‌باشد. در این روش باتوجه به طولانی‌بودن رشته‌ی DNA، برای توالی‌یابی می‌بایست تکه‌هایی به طول ۱۰۰ الی ۱۰۰۰ نوکلئوتید از کل رشته DNA توالی‌یابی شود که به توالی هرکدام از این تکه‌ها، خوانش[۹] می‌گویند. سپس با استفاده از هم‌پوشانی بین این خوانش‌ها، کل رشته به‌دست می‌آید. به بیان دیگر در این روش، به‌طور تصادفی از نقاط مختلف کل رشته DNA، قطعات کوتاه هم‌پوشان خوانده می‌شود. سپس با روی‌هم قراردادن بخش‌های هم‌پوشان توالی کل رشته ی DNA به‌دست می‌آید [۴]. همان‌طور که در شکل ۱ مشاهده می‌گردد برای به‌دست‌آوردن کل رشته DNA به طول تقریبی $G = 10^9$ از تعداد $N = 10^8$ خوانش، هرکدام به طول $L = 100$ استفاده شده‌است.



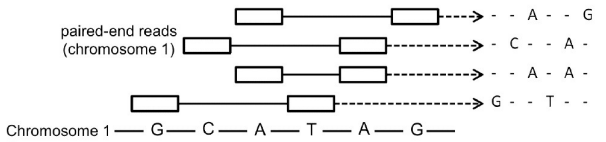
شکل ۱: مفهوم توالی‌یابی [۵].

همان‌طور که ذکر شد، هدف از توالی‌یابی، یافتن رشته DNA است. بدین‌منظور از الگوریتم‌های گردهم‌آوری[۱۰] استفاده می‌شود. ورودی این الگوریتم، خوانش‌ها و خروجی الگوریتم، رشته DNA می‌باشد.

الگوریتم‌های گردهم‌آوری به دو دسته کلی با نام‌های گردهم‌آوری از پایه[۱۱] و گردهم‌آوری نگاشتی تقسیم می‌شوند.

الگوریتم‌هایی که در دسته گردهم‌آوری نگاشتی قرار می‌گیرند، از یک رشته‌ی DNA به‌عنوان رشته مرجع استفاده می‌کنند و هر خوانش به قسمتی از رشته مرجع تطبیق داده می‌شود. این موضوع به علت شباهت بالای رشته‌ی DNA در بین انسان‌ها است. هنگامی که رشته‌ی DNA مرجع در اختیار نباشد، از الگوریتم‌های گردهم‌آوری از پایه استفاده می‌شود.

برای افزایش دقت توالی‌یابی، روش جدیدی به نام توالی‌یابی جفت انتها[۱۲] معرفی شده‌است. در این روش خوانش‌ها دارای دو بخش هستند که هرکدام توالی از نوکلئوتیدها را به دست می‌دهد. علاوه‌بر این دو بخش، خوانش جفت‌انتها، فاصله‌ی بین این دو بخش را نیز مشخص می نماید. بدین‌ترتیب یک خوانش با طول بزرگ‌تری در دسترس قرار خواهدگرفت که ابتدا و انتهای آن مشخص است ولی میان آن مشخص نیست. بدین‌ترتیب خوانش به طول بسیار بزرگ‌تر دردسترس خواهدبود که منجر به افزایش دقت در نگاشت به مرجع می‌شود. در شکل (۲) یک مثال ساده از خوانش جفت‌انتها که هر انتهای آن یک نوکلئوتید خوانده شده‌است، ارائه شده است.



شکل ۲: خوانش جفت انتها تک نوکلئوتیدی [۶].

در شکل ۲ چهار خوانش از کروموزم یک شامل نوکلئوتیدهای $GCATAG$ ارائه شده‌است. منظور از "-" آن است که آن خوانش اطلاعاتی درباره‌ی آن جایگاه نوکلئوتید ارائه نکرده‌است.

۱-۳- معرفی مفهوم هاپلوتایپ

متداول‌ترین اختلاف در جمعیت به‌صورت تکی در بین نوکلئوتیدها می باشد که چندریختی تک‌نوکلئوتیدی[۱۳] (اسنیپ) خوانده می‌شود. یک اسنیپ جایگاهی از ژن است که در بین افراد جامعه، نوکلئوتیدهای مختلفی مشاهده می‌شود. رشته‌ی اسنیپ برروی هر یک از جفت کروموزم همسان، هاپلوتایپ نام دارد. به‌حالتی‌که اسنیپ برای یک فرد برروی جفت کروموزم، یکسان باشد هموزایگت[۱۴] نامیده می‌شود، در غیر از این صورت هتروزایگت[۱۵] نام دارد [۷].

هنگامی‌که توالی نوکلئوتید در اسنیپ‌ها برروی جفت کروموزم (نه به طور مجزا) مورد نظر باشد، به آن ژنوتایپ[۱۶] گفته می‌شود. اکثر روش های آزمایشگاهی توالی‌یابی به یافتن ژنوتایپ منجر می‌گردد که بسیار فراگیر و با هزینه‌ی بسیار کم دردسترس قرار دارد. البته توانایی توالی‌یابی هاپلوتایپ با روش‌های هزینه‌بر امکان‌پذیر هست. هنگامی‌که داده ژنوتایپ دردسترس است، به یافتن هاپلوتایپ با استفاده از روش‌های محاسباتی تخمین هاپلوتایپ گفته می‌شود. دانستن هاپلوتایپ ضروری و بسیار کارگشا می‌باشد؛ درمان اختصاصی مبتنی بر هاپلوتایپ هر فرد به‌عنوان یکی از شاخه‌های پزشکی آینده بسیار موردتوجه قرار گرفته است. با استفاده از هاپلوتایپ، می‌توان اطلاعات قابل‌توجهی در باره ژنتیک جمعیتی، مهاجرت و انتخاب طبیعی، روند تکاملی موجودات، نرخ بازترکیب ژنتیکی نسل‌ها به دست آورد. دو راه‌برد کلی برای تخمین هاپلوتایپ وجود دارد [۸]:

الف) جمعیت: این راه‌برد براساس داده‌های جمعیتی (داشتن توالی ژنوتایپ چندین فرد) با استفاده از فرض‌هایی انجام می‌پذیرد که در چهار دسته‌ی کلی جای می‌گیرند: ۱. بیشینه ایجاز: تعداد هاپلوتایپ‌های متمایز در جواب، کمترین است. ۲. درخت فیلوژنی کامل: مجموعه‌ی هاپلوتایپ‌های جواب، یک درخت فیلوژنی کامل تشکیل دهند. یعنی هر هاپلوتایپ از یک جهش نسبت به هاپلوتایپ نیای (اجداد) افراد پدید آمده باشد. ۳. بیشینه درست‌نمایی: توزیع احتمال هاپلوتایپ‌ها در جمعیت به طوری باشد که تابع درست‌نمایی مشاهده‌ی ژنوتایپ‌های داده شده بیشینه شود. ۴. استنتاج بیزی: محاسبه احتمال پسین در مدل‌های آماریِ ازپیش‌تعیین شده به دست آید.

ب) گردهم‌آوری هاپلوتایپ: این راه‌برد بر پایه دادهای NGS است که برخلاف راه‌برد جمعیتی، می‌تواند هاپلوتایپ یک فرد را تخمین بزند. از الگوریتم‌های معروف این راه‌برد می‌توان به HapCUT [۹]، HapTree [۱۰]، SDhaP [۶] و RefHap [۱۱] نامبرد. شایان ذکر است که روش پیشنهادی این مقاله نیز بر پایه همین راه‌برد می‌باشد.

در ادامه مقاله در بخش ۲، مساله تخمین هاپلوتایپ به‌صورت ریاضی مدل می‌گردد. در بخش سوم الگوریتم پیشنهادی مطرح خواهدشد. نتایج شبیه‌سازی و نتیجه‌گیری به ترتیب در بخش ۴ و ۵ ارائه خواهدشد.

۲- مدل مساله

فرض کنید تعداد n خوانش دردسترس باشد. خوانش iام را با نماد r_i نمایش می‌دهیم. مجموعه‌ی r شامل خوانش‌ها است که به‌صورت $r = \{r_1, r_2, \dots, r_n\}$ خواهدبود. هم‌چنین، فرض بر این است که ابتدا خوانش‌ها جایابی شده‌اند؛ این بدین معنا است که اطلاعات مربوط به محل‌قرارگفتن هر یک از r_iها در رشته DNA دردسترس می‌باشد. هم‌چنین، بخشی از خوانش که شامل اسنیپ‌ها است، جدا شده‌اند. برای هر اسنیپ، به آلل فراوان عدد 1 و آلل نادر عدد -1 نسبت داده شده‌است. سپس ماتریس خوانش ($\boldsymbol{R}$) تشکیل داده می‌شود: هر خوانش را به‌عنوان یک سطر ماتریس در نظر می‌گیریم به‌طوری‌که اگر خوانش اطلاعی درباره یک درایه ارائه ندهد، صفر قرار می‌دهیم. برای تخمین هاپلوتایپ به طول m با استفاده از n خوانش، ابعاد ماتریس خوانش $n \times m$ خواهدبود [۱۲].

ازآنجایی‌که ممکن است تعدادی از اسنیپ‌ها هموزایگت باشند، ستون‌هایی از ماتریس که تمام درایه‌های یکسان (همه 1 و یا همه -1) هستند، حذف می‌گردد. سپس با ذخیره‌کردن مکان آن، آن بخش از

هاپلوتایپ مشخص می‌گردد؛ زیرا ابهامی ندارد و نیاز به تخمین نیست. در ادامه مدل‌سازی ریاضی رشته خوانش و هاپلوتایپ مطرح می‌گردد. بدین‌منظور برای تخمین جایگاه‌های هتروزایگت مثالی عددی ساده بدون نویز مطرح می‌گردد دو رشته هاپلوتایپ به‌صورت زیر در نظر بگیرید:

$$\begin{matrix}\boldsymbol{h}_1 = \begin{bmatrix} 1 & 1 & 1 & -1 \end{bmatrix}^T \\ \boldsymbol{h}_2 = \begin{bmatrix} -1 & -1 & -1 & 1 \end{bmatrix}^T \end{matrix} \quad (۱)$$

در ادامه، مثالی از خوانش‌ها ارائه شده است:

- Read 1 from $\boldsymbol{h}_1 \rightarrow \begin{bmatrix} 1 & \sim & \sim & -1 \end{bmatrix}$
- Read 2 from $\boldsymbol{h}_1 \rightarrow \begin{bmatrix} \sim & 1 & 1 & \sim \end{bmatrix}$
- Read 3 from $\boldsymbol{h}_2 \rightarrow \begin{bmatrix} \sim & -1 & -1 & \sim \end{bmatrix}$
- Read 4 from $\boldsymbol{h}_2 \rightarrow \begin{bmatrix} -1 & \sim & \sim & 1 \end{bmatrix}$

در نمایش فوق، اگر درایه‌ای به‌صورت $\sim$ باشد، بدین معنا است که این درایه در خوانش مربوطه قرار نداشته‌است. بدین‌ترتیب ماتریس خوانش به‌صورت زیر خواهدبود:

$$\boldsymbol{R} = \begin{bmatrix} 1 & \sim & \sim & -1 \\ \sim & 1 & 1 & \sim \\ \sim & -1 & -1 & \sim \\ -1 & \sim & \sim & 1 \end{bmatrix} \quad (۲)$$

پس مدل‌سازی ریاضی به‌صورت زیر خواهدبود [۶]:

$$\boldsymbol{M} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix} \quad (۳)$$

اما ازآنجایی که در دیپلوئیدها خوانش‌ها از هاپلوتایپ مادری ($\boldsymbol{h}_m$) و یا پدری ($\boldsymbol{h}_p$) توالی‌یابی شده‌اند، برای اسنیپ‌های باقی‌مانده (هتروزایگت) مقادیر قرینه یک‌دیگر هستند؛ به بیان دیگر $\boldsymbol{h}_p = -\boldsymbol{h}_m$ برقرار است. پس برای مثال فوق به صورت زیر خواهدبود:

$$\boldsymbol{M} = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & -1 \end{bmatrix} \quad (۴)$$

بدین ترتیب مدل برداری زیر ارائه می‌گردد:

$$\boldsymbol{M} = \boldsymbol{u}\boldsymbol{h}^T \quad (۵)$$

که در آن بردار $\boldsymbol{u}$ با ابعاد $n \times 1$ با درایه‌های ± 1 است. هنگامی‌که درایه $+1$ باشد حاکی از آن است که این خوانش مربوط به توالی هاپلوتایپ $\boldsymbol{h}_p$ است هم‌چنین درایه -1، به معنای خوانش از $\boldsymbol{h}_m$ است. هم‌چنین بردار $\boldsymbol{h}$، برداری با ابعاد $m \times 1$ است، که شامل درایه‌های ± 1 می‌باشد. بدین‌ترتیب ماتریس $\boldsymbol{R}$ را می‌توان به‌صورت زیر نوشت:

$$\boldsymbol{R} = P_\Omega(\boldsymbol{M}) = \begin{cases} \boldsymbol{M}_{ij}, & (i,j) \in \Omega \\ 0, & o.w. \end{cases} \quad (۶)$$

که در آن Ω مجموعه‌ای است که خوانش‌ها در آن مکان اطلاعات ارائه کرده‌است. $P_\Omega(\boldsymbol{M})$ به‌عنوان عمل‌گر نمونه‌بردار است، پس ماتریس مشاهدات $\boldsymbol{R}$ برابر با نسخه‌ی نمونه‌برداری‌شده از $\boldsymbol{M}$ خواهدبود. یکی از نوآوری‌های این مقاله ارائه صورت مساله تخمین هاپلوتایپ به‌صورت بهینه‌سازی رتبه یک است که در حالت توالی‌یابی دقیق (بدون نویز) به‌صورت زیر است:

$$\text{Find } \boldsymbol{M} \text{ s.t.} \begin{cases} P_\Omega(\boldsymbol{M}) = P_\Omega(\boldsymbol{R}) \\ \text{rank}(\boldsymbol{M}) = 1 \end{cases} \quad (۷)$$

که در آن $\boldsymbol{R}$ ماتریس داده‌های دردسترس و $\boldsymbol{M}$ مجهول بهینه‌سازی است. بدین‌ترتیب باتوجه‌به ساختار متغیرها در رابطه‌ی (۵)، مساله‌ای معادل برای (۷) به‌صورت بهینه‌سازی چندجمله‌ای ارائه می‌گردد:

$$\text{Find } \boldsymbol{u}, \boldsymbol{h} \quad \text{s.t. } u_i h_j - \boldsymbol{R}_{ij} = 0 \ \ \forall (i,j) \in \Omega \quad (۸)$$

ازآنجایی‌که متغیرهای مساله به‌صورت اعداد ± 1 تعریف شده‌است، می توان به‌صورت زیر بهینه‌سازی را تکمیل نمود:

$$\text{Find } \boldsymbol{u}, \boldsymbol{h} \text{ s.t.} \begin{cases} u_i h_j - \boldsymbol{R}_{ij} = 0 \ \ \forall (i,j) \in \Omega \\ h_j \in \{1, -1\} \ \ j = 1, \dots, m \\ u_i \in \{1, -1\} \ \ i = 1, \dots, n \end{cases} \quad (۹)$$

به بیان معادل، قید را نیز به‌صورت چندجمله‌ای نوشت:

$$\text{Find } \boldsymbol{u}, \boldsymbol{h} \text{ s.t.} \begin{cases} u_i h_j - \boldsymbol{R}_{ij} = 0 \ \ \forall (i,j) \in \Omega \\ h_j^2 = 1 \ \ j = 1, \dots, m \\ u_i^2 = 1 \ \ i = 1, \dots, n \end{cases} \quad (۱۰)$$

که یک مساله چندجمله‌ای کلاسیک است که برای حل مساله از پکیج‌های نرم‌افزاری بهینه‌سازی می‌توان استفاده نمود. هم‌چنین در حالت نویزی می‌توان مساله بهینه‌سازی را به شرح زیر نوشت:

$$\min_{\boldsymbol{u}, \boldsymbol{h}} \ \| P_\Omega(\boldsymbol{R} - \boldsymbol{u}\boldsymbol{h}^T) \|_F \quad (۱۱)$$

برای حل آن از روش‌های عددی از جمله روش تکراری گرادیانی استفاده شده‌است [۴، ۹]. این روش حل از ساختاری که از قبل دانسته فرض‌شده استفاده نکرده‌است. به‌عنوان یکی از نوآوری‌های این مقاله ارائه مساله بهینه‌سازی چندجمله‌ای ساختاریافته برای تخمین هاپلوتایپ است که به‌صورت زیر می‌باشد:

$$\min_{\boldsymbol{u}, \boldsymbol{h}} \sum_{(i,j) \in \Omega} \left(u_i h_j - \boldsymbol{R}_{ij} \right)^2 \text{ s.t.} \begin{cases} h_j^2 = 1 & j = 1, \dots, m \\ u_i^2 = 1 & i = 1, \dots, n \end{cases} \quad (۱۲)$$

به‌منظور ساده‌سازی، بردار $\boldsymbol{x}$ به طول $m + n$ با کنارهم‌گذاشتن دو بردار مجهول به‌صورت زیر تعریف می‌گردد:

$$\boldsymbol{x} = [u_1, \dots, u_n, h_1, \dots, h_m] \quad (۱۳)$$

بدین‌ترتیب مساله بهینه‌سازی برای دو حالت دقیق و نویزی به‌صورت زیر خواهدبود:

$$\text{Find } \boldsymbol{x} \text{ s.t.} \begin{cases} x_i x_{j+n} - \boldsymbol{R}_{ij} = 0 \ \ \forall (i,j) \in \Omega \\ x_j^2 = 1, j = 1, \dots, m+n \end{cases} \quad (۱۴)$$

$$\min_{\boldsymbol{x} \in \boldsymbol{R}^{m+n}} \sum_{(i,j) \in \Omega} \left(x_i x_{j+n} - \boldsymbol{R}_{ij} \right)^2 \quad (۱۵)$$
$$\text{s.t. } x_j^2 = 1, \qquad j = 1, \dots, m+n$$

۳- ریلکس‌سازی بهینه‌سازی چندجمله‌ای

در این بخش، ریلکس‌سازی یک مساله چندجمله‌ای به یک مساله بهینه‌سازی مثبت نیمه‌معین درحالت کلی معرفی می‌گردد. مقدماتی از مفهوم اندازه[۱۷] و ماتریس ممان در بخش پیوست آورده شده‌است. برای ریلکس‌سازی بهینه‌سازی چندجمله‌ای به یک گام میانی نیاز است. این گام، تبدیل مساله بهینه‌سازی چندجمله‌ای به یک مساله بهینه‌سازی معادل برروی فضای اندازه می‌باشد.

در ابتدا، چندجمله‌ای به‌صورت دقیق معرفی می‌گردد. یک چندجمله‌ای n بعدی $p(x)=\sum_\alpha p_\alpha x^\alpha$ با نماد $p(x):\mathbb{R}^n\to\mathbb{R}$، استفاده خواهدشد، که در آن $\alpha=(\alpha_1,\alpha_2,\dots\alpha_n)$ و $x^\alpha=x_1^{\alpha_1}x_2^{\alpha_2}\dots x_n^{\alpha_n}$ است. هم‌چنین $t=\sum_i\alpha_i$ درجه‌ی چندجمله‌ای خواهدبود. مساله بهینه‌سازی چندجمله‌ای به‌صورت زیر می‌باشد:

$$\min p(x) \quad \text{s.t.} \quad \begin{matrix} g_i(x)\geq 0\ i=1,\dots,I, \\ h_l(x)=0\ \ l=1,\dots L \end{matrix} \qquad (۱۶)$$

که در آن $p(x)$، $g_i(x)$ها و $h_l(x)$ها چندجمله‌ای می‌باشند. هم‌چنین مجموعه‌ی شدنی[۱۸] (مجموعه‌ی xهایی که در قید صدق می‌کنند) برای مساله بهینه‌سازی فوق به‌صورت

$$K\triangleq\{x|g_1(x)\geq 0,\dots,g_I(x)\geq 0,h_1(x)=0,\dots,h_L(x)=0\ \}$$

تعریف می‌شود. در [۱۳] ثابت شده‌است که مساله بهینه‌سازی (۱۶) معادل مساله چندجمله‌ای بهینه‌سازی زیر است:

$$\min_{\mu\in MS(K)}\int p(x)d\mu \qquad (۱۷)$$

که در آن $MS(K)$[۱۹] مجموعه‌ی فضای تمام اندازه‌های احتمال ممکن تعریف‌شده روی مجموعه‌ی K می‌باشد. در ابتدا حالتی‌که مساله بهینه سازی چندجمله‌ای بدون قید، مدنظر قرار می‌گیرد. با استفاده از تعریف ممان می‌توان تابع هدف مساله بهینه‌سازی چندجمله‌ای را به‌صورت زیر نوشت:

$$\begin{aligned}\int p(x)d\mu=\int\sum_\alpha p_\alpha x^\alpha\,d\mu=\sum_\alpha p_\alpha\int x^\alpha d\mu\\ =\sum_\alpha p_\alpha\,y_\alpha\end{aligned} \qquad (۱۸)$$

بدین‌ترتیب مساله بهینه‌سازی معادل جدید به‌صورت زیر می‌باشد:

$$\min_y\sum_\alpha p_\alpha\,y_\alpha \quad \text{s.t.} \quad y \text{ is a moment sequence.} \qquad (۱۹)$$

آن‌گاه رابطه پاسخ مساله (۱۹) با مجهول‌های مساله اصلی به‌صورت $y=(1,x_1^*,x_2^*,x_1^{*2},x_1^*x_2^*,x_2^{*2},\dots)$ است. باتوجه‌به این‌که حل مساله بهینه‌سازی فوق با قید ذکرشده ممکن نیست. می‌بایست قید را ریلکس نمود. بدین‌منظور، باتوجه‌به ویژگی اول ماتریس ممان، مساله جدید بهینه‌سازی به‌صورت زیر خواهدشد:

$$\min_y\sum_\alpha p_\alpha\,y_\alpha \quad \text{s.t.} \quad \boldsymbol{M}_t(y)\succcurlyeq 0 \qquad (۲۰)$$

ازآنجایی‌که برای مقادیر مختلف t مساله حل می‌شود به این راه‌برد که توسط آقای Lasserre معرفی شده‌است Lasserre Hierarchy گویند. این یک مساله استاندارد مثبت نیمه‌معین (SDP) است که شامل تابع هدف خطی و یک قید به‌صورت شرط مثبت‌نیمه‌معین‌بودن ماتریس می‌باشد.

هم‌چنین در حالت مقید، باتوجه‌به تعریف مجموعه K، مساله جدید به‌صورت زیر خواهدبود [۱۳]:

$$\min_y\sum_\alpha p_\alpha\,y_\alpha \quad \text{s.t.} \quad \begin{cases}\boldsymbol{M}_t(y)\succcurlyeq 0\\ \boldsymbol{M}_{t-d_{g_i}}(g_iy)\succcurlyeq 0\\ \boldsymbol{M}_{t-d_{h_l}}(h_iy)=0\end{cases} \qquad (۲۱)$$

بدین‌ترتیب مساله چندجمله‌ای که در حالت کلی محدب نیست، به یک مساله محدب ریلکس شده‌است. این مساله با نرم افزار MATLAB ابزار CVX و هم‌چنین GloptiPoly [۱۴] حل می‌گردد.

۴- الگوریتم پیشـــنهادی: الگوریتم تخمین هاپلو تا یپ با استفاده از ریلکس سازی بهینه سازی چندجمله‌ای به روش Lasserre

مساله بهینه‌سازی مطرح‌شده برحسب متغیر $\boldsymbol{x}$ در رابطه (۱۴) را درنظر بگیرید. در این مساله تعداد $|\Omega|+m+n$ قید تساوی وجود دارد. منظور از $|.|$ تعداد اعضای یک مجموعه است. بدین‌ترتیب باتوجه‌به مقدمات مطرح‌شده در بخش قبل، مساله ریلکس‌شده در حالت دقیق به‌صورت زیر خواهدبود:

$$\text{Find } y \quad \text{s.t.} \quad \begin{cases}\boldsymbol{M}_t(y)\succcurlyeq 0\\ \boldsymbol{M}_{t-d_{h_i}}(h_iy)=0\ i=1,\dots,|\Omega|+m+n\end{cases} \qquad (۲۲)$$

هم‌چنین در حالت نویزی به‌صورت زیر خواهدبود:

$$\min_y\sum_\alpha p_\alpha\,y_\alpha \quad \text{s.t.} \quad \begin{cases}\boldsymbol{M}_t(y)\succcurlyeq 0\\ \boldsymbol{M}_{t-d_{h_i}}(h_iy)=0\ i=1,\dots,m+n\end{cases} \qquad (۲۳)$$

که‌درآن p_α متناظر با ضرایب چندجمله‌ای زیر است:

$$p(x)=\sum_\alpha p_\alpha\,x^\alpha=\sum_{(i,j)\in\Omega}x_{ij}^2-2x_{ij}R_{ij} \qquad (۲۴)$$

عبارت فوق با حذف‌ترم ثابت R_{ij}^2 از $\sum_{(i,j)\in\Omega}\left(x_ix_{j+n}-\boldsymbol{R}_{ij}\right)^2$ به دست آمده‌است. بدین ترتیب این بهینه‌سازی بخش اصلی الگوریتم معرفی‌شده در این مقاله است که در شکل ۳ آورده شده‌است.

Algorithm 1: Haplotype estimation using Lasserre hierarchy relaxations (HapLas).

Input: n aligned reads

Output: Haplotype $\widehat{\boldsymbol{h}}$

Read Matrix Preparation

1 Convert the sequences of nucleotides (reads) to the sequences of numbers
2 Add zeros to each read to construct $\boldsymbol{r}_i$ with the length of m
3 Construct the read matrix $\boldsymbol{R}$ $(n\times m)$
4 Construct the constraints from $\boldsymbol{R}$

** Optimization part **

5 Solve the following optimization
$\min_y\sum_\alpha p_\alpha\,y_\alpha \quad \text{s.t.} \quad \begin{cases}\boldsymbol{M}_t(y)\succcurlyeq 0\\ \boldsymbol{M}_{t-d_{h_i}}(h_iy)=0\ i=1,\dots,m+n\end{cases}$
6 Extract the x_is from moment sequence by
$y^*=(1,x_1^*,x_2^*,\dots,x_n^*,x_1^*x_2^*,\dots)$

Haplotype finding

7 Find $\widehat{\boldsymbol{h}}$ through $\widehat{\boldsymbol{h}}=[x_{n+1},\dots,x_{n+1}]$
8 Convert the entries of $\widehat{\boldsymbol{h}}_m$ and $\widehat{\boldsymbol{h}}_p$ to the nucleotides.

شکل ۳: الگوریتم پیشنهادی HapLas

۵- شبیه‌سازی

برای بررسی عمل‌کرد روش پیشنهادی از معیار جذر میانگین مربع خطا (RMSE) استفاده می‌گردد که به‌صورت زیر تعریف شده‌است:

$$RMSE^2 = \frac{1}{mK}\sum_{k=1}^{K}\sum_{i=1}^{m}\left(\hat{h}_i^{(k)} - h_i\right)^2 \qquad (۲۵)$$

که در آن K تعداد تکرار است که ۱۰ درنظر گرفته شده‌است. هم‌چنین $\hat{h}_i^{(k)}$ هاپلوتایپ تخمین‌زده‌شده در جای‌گاه iام و تکرار kام می‌باشد. علاوه‌براین از معیار نرخ بازسازی (rr) که در زیر تعریف شده‌است، نیز استفاده می‌شود [۱۵]:

$$rr = \frac{1 - hd(\hat{\boldsymbol{h}}, \boldsymbol{h})}{m} \qquad (۲۶)$$

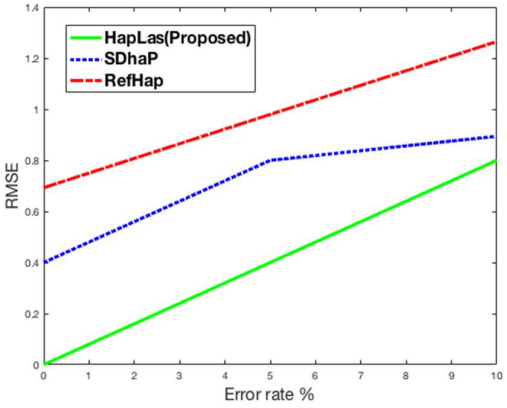
$$hd(\hat{\boldsymbol{h}}, \boldsymbol{h}) = \frac{1}{K}\sum_{k=1}^{K} d\left(\hat{h}_i^{(k)}, h_i\right) \qquad (۲۷)$$

که در آن فاصله $d(.,.)$ به‌صورت زیر تعریف می‌گردد:

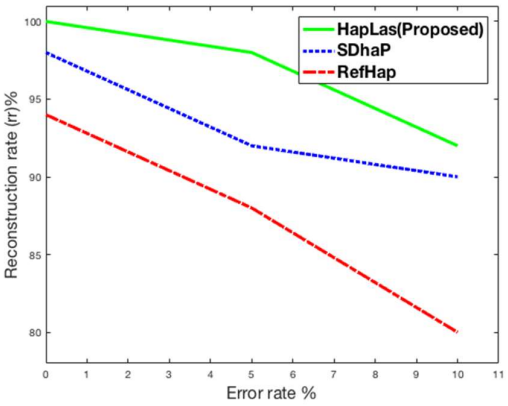
$$d(a,b) = \begin{cases} 0, & a = b \\ 1, & \text{o. w.} \end{cases} \qquad (۲۸)$$

هرچه این معیار بزرگ‌تر باشد، حاکی از آن است که تخمین بهتری انجام شده‌است. هم‌چنین زمان اجرای الگوریتم به‌عنوان معیاری برای پیچیدگی محاسباتی نیز برای مقایسه الگوریتم‌ها نیز مورد توجه قرار گرفته‌است. در این مقاله، الگوریتم پیشنهادی با دو الگوریتم معروف در این حوزه با نام‌های SDhaP و RefHap مقایسه شده‌است.

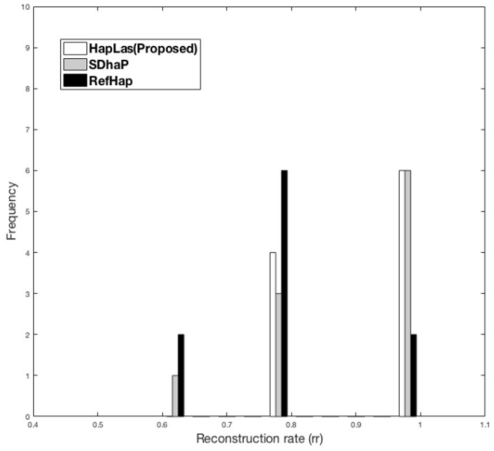
طی درخواستی از نویسنده مقاله [۱۵]، داده‌های آن دریافت شد و با استفاده از آن شبیه‌سازی‌های مختلفی انجام شده‌است. در ابتدا یک آزمایش برای هاپلوتایپی به طول ۵ با درنظرداشتن ۵ خوانش، با پوشش برابر با ۳ (تعداد درایه معلوم در هر ستون) در نظرگرفته شده‌است. در شکل‌های ۴ و ۵ به‌ترتیب نمودار جذر میانگین مربع خطا (RMSE) و نمودار نرخ بازسازی (rr) برحسب درصد نرخ خطا در آزمایش ذکرشده رسم شده‌است. نتایج همه تکرارهای آزمایش اول، به‌صورت هیستوگرام برای نرخ خطا ۱۰، ۵ و بدون خطا به‌ترتیب در شکل‌های ۶، ۷ و ۸ رسم شده‌است.



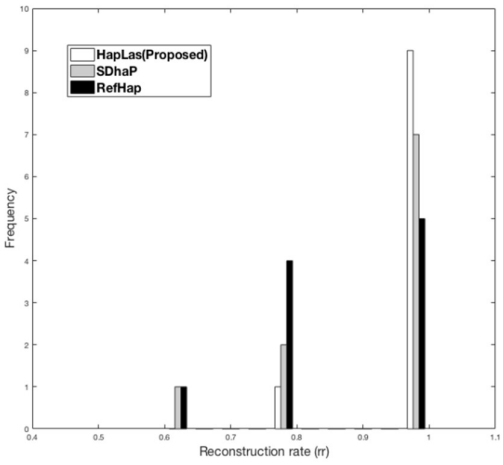
شکل ۴: نمودار RMSE برحسب درصد نرخ خطا در آزمایش اول.



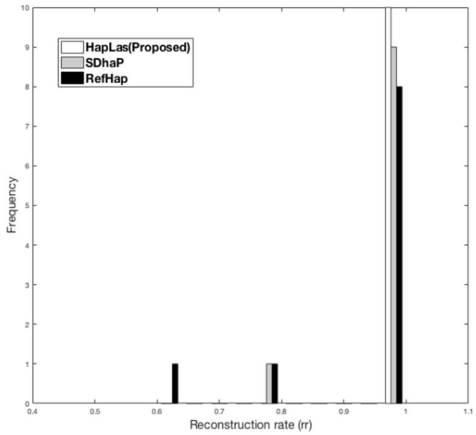
شکل ۵: نمودار درصد نرخ بازسازی (rr) برحسب درصد نرخ خطا در آزمایش اول.



شکل ۶: هیستوگرام نرخ بازسازی برای نرخ خطا ۱۰درصد



شکل ۷: هیستوگرام نرخ بازسازی برای نرخ خطا ۵درصد.



شکل ۸: هیستوگرام نرخ بازسازی برای حالت بدون خطا.

متوسط زمان اجرای الگوریتم در جدول ۱ ارائه شده‌است. همان‌طور که در شکل‌های فوق مشاهده می‌شود با افزایش خطای داده میزان خطای تخمین RMSE افزایش یافته و میزان بازسازی rr کاهش می‌یابد. هم‌چنین، الگوریتم HapLas از دو الگوریتم دیگر SDhaP و RefHap تخمین بهتری از نظر MSE و هم‌چنین rr در ازای زمان بیش‌تر ارائه می‌دهد. علاوه‌برآن، باتوجه‌به نمودارهای هیستوگرام، نتایج آزمایش برای HapLas پراکندگی کمتری دارد و در نقطه بیشینه متمرکز است. نکته قابل توجه در نتایج این است که در حالت بدون نویز انتظار این است که الگوریتم به‌راحتی بتواند هاپلوتایپ را تخمین بزند، درحالی‌که روش RefHap و SDhaP در موارد معدودی دچار خطا می‌شوند.

جدول ۱: زمان اجرای الگوریتم پیشنهادی در مقایسه با الگوریتم‌های دیگر برحسب ثانیه برای آزمایش اول

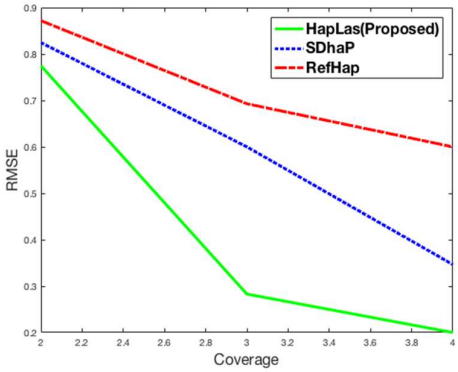
الگوریتم	RefHap	SDhaP	HapLas
زمان اجرا (ثانیه)	۰/۱	۰/۰۱	۰/۷

در آزمایش دیگری که انجام پذیرفت، هدف تخمین هاپلوتایپی به طول ۱۰ است که ۷ خوانش با نرخ خطا به‌صورت ثابت ده درصد دردسترس است. در شکل ۹ میزان RMSE برحسب میزان پوشش و هم‌چنین در شکل ۱۰، میزان rr برحسب میزان پوشش رسم شده‌است.

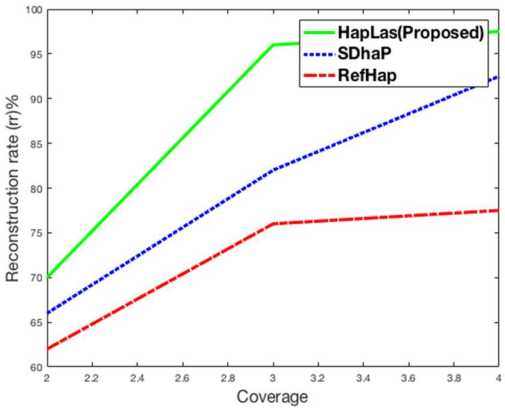
باتوجه‌به شکل‌های ۹ و ۱۰ همان‌طورکه انتظار می‌رود با افزایش مقدار پوشش، میزان خطای تخمین RMSE کاهش‌یافته و میزان بازسازی rr افزایش می‌یابد. هم‌چنین برای روش پیشنهادی میزان خطا کم‌ترین و میزان نرخ بازسازی بیش‌ترین به دست می‌آید. هم‌چنین باتوجه‌به جدول ۲، متوسط زمان اجرای الگوریتم ارائه‌شده نسبت به الگوریتم‌های موجود بیش‌تر است که افزایش سرعت آن در حال حاضر دردست تحقیق است.

جدول ۲: زمان اجرای الگوریتم پیشنهادی در مقایسه با الگوریتم‌های دیگر برحسب ثانیه برای آزمایش دوم

الگوریتم	RefHap	SDhaP	HapLas
زمان اجرا (ثانیه)	۰/۲	۰/۰۲	۱۵۷



شکل ۹: نمودار RMSE برحسب میزان پوشش در آزمایش دوم.



شکل ۱۰: درصد نرخ بازسازی (rr) برحسب میزان پوشش در آزمایش دوم

۶- نتیجه‌گیری

در این مقاله، الگوریتم تخمین هاپلوتایپ با استفاده از ریلکس‌سازی بهینه‌سازی چندجمله‌ای (HapLas) معرفی شد. بدین‌منظور از خاصیت مثبت‌نیمه‌معین‌بودن ماتریس ممان برای ریلکس‌سازی یک مساله بهینه‌سازی غیرقابل‌حل استفاده شد. نتایج شبیه‌سازی مشخص نمود که روش پیشنهادی منجربه افزایش نرخ بازسازی هاپلوتایپ در حدود ۵ درصد می‌شود. بدین‌ترتیب با افزایش دقت تخمین هاپلوتایپ، بهبود کیفیت در درمان بیماری‌های ژنتیکی موردانتظار است. درازای این بهبود محاسبات بیشتری موردنیاز است. ازآنجایی‌که روش پیشنهادی دارای دقت مناسب اما دارای محاسبات بیشتری است، یکی از پیشنهادات به منظور ادامه پژوهش، کاهش محاسبات الگوریتم همراه با حفظ نسبی دقت مدنظر می‌باشد.

پیوست

مفهوم اندازه

اندازه با نماد μ به‌صورت مقابل تعریف می‌شود: به هر زیرمجموعه‌ای از یک مجموعه مانند T یک عدد نامنفی نسبت دهد به‌گونه‌ای که دو شرط زیر برقرار باشد [۱۷]:

I. جمع‌پذیری شمارا[۲۰]: برای زیرمجموعه‌های مجزا که $E_i \cap E_j = \emptyset$ برای $i \neq j$ است، رابطه‌ی زیر برقرار باشد:

$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k) \qquad (۲۹)$$

II. پایا نسبت به انتقال[۲۱]: برای مجموعه‌ی E و نقطه y، رابطه‌ی $\mu(E+y) = \mu(E)$ برقرار باشد، به طوری‌که:

$$E + y = \{x + y | \, x \in E\} \qquad (۳۰)$$

به‌بیان‌دیگر مجموعه‌ی $E+y$ از جانشین‌کردن هرنقطه مانند x از مجموعه‌ی E به $x+y$ به دست آمده‌است.

یک اندازه معروف، اندازه دیراک[۲۲] در نقطه‌ی x با نماد $\mu = \delta_x$ است به‌صورت مقابل تعریف می‌شود: اندازه آن در x یک است $(\mu(\{x\}) = 1)$ و در سایر نقاط صفر است. به بیان ریاضی:

$$\delta_x(E) = \begin{cases} 1, & x \in E \\ 0, & o.w. \end{cases} \qquad (۳۱)$$

در ادامه انتگرال یک تابع نسبت به اندازه تعریف می‌گردد. انتگرال تابع f نسبت به اندازه μ روی مجموعه‌ی E با نماد $\int_E f d\mu$ نمایش داده می شود. بدین‌منظور در ابتدا، مجموعه E به زیرمجموعه‌هایی مانند E_k افراز می‌شود. منظور از افراز، تقسیم‌کردن اعضای یک مجموعه به چندین زیرمجموعه به‌گونه‌ای‌که دارای دو خاصیت زیر باشند:

$$\bigcup_{k=1}^{K} E_k = E \qquad (۳۲)$$

$$E_i \cap E_j = \emptyset \; \forall i \neq j = 1, \dots, K \qquad (۳۳)$$

این افراز باید به‌گونه‌ای باشد که تابع f در آن زیرمجموعه دارای مقدار ثابت باشد که به بیان ریاضی $E_k = \{x \mid f(x) = c_k\}$ برقرار باشد. حال انتگرال به‌صورت زیر تعریف می‌شود [۱۷]:

$$\int_E f d\mu = \sum_{k=1}^{K} c_k \mu(E_k) \qquad (۳۴)$$

که در آن $\mu(E_k)$ اندازه هرکدام از افرازهاست. هنگامی‌که اندازه کل فضا یک باشد، به آن اندازه احتمال گویند. به‌بیان‌دیگر، اگر اندازه برروی مجموعه‌ی T تعریف شود، هنگامی‌که $\mu(T) = 1$ باشد، اندازه احتمال نام دارد. شایان ذکر است که این تعریف در حالت خاص است و در این مقاله نیازی به جزئیات تعریف در حالت کلی نیست.

معرفی ماتریس ممان

حلقه‌ی چندجمله‌ای چندمتغیره، برروی متغیرهای n بعدی $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$، با نماد $\mathbb{R}[\boldsymbol{x}]$ نمایش داده می‌شود. هر عضو این حلقه مانند p یک چندجمله‌ای است که به‌صورت $p(x) = \sum_\alpha p_\alpha \boldsymbol{x}^\alpha$ نمایش داده می‌شود. هم‌چنین $\mathbb{N}^n$ مجموعه‌ی تمام αهای nبعدی مانند $\alpha = (\alpha_1, .., \alpha_n)$ است که مقادیر هر بعد، از مجموعه‌ی اعداد صحیح نامنفی $(\mathbb{N})$ آمده‌است. حال $\mathbb{N}_t^n$ برای $\mathrm{t} \in \mathbb{N}$ مجموعه‌ی تمام n بعدی‌هایی از مجموعه‌ی $\mathbb{N}^n$ هستند که جمع مقادیر بعدها، کمتر از عدد t باشد. به بیان ریاضی می‌توان نوشت [۱۳]:

$$\mathbb{N}_t^n = \left\{ \alpha \in \mathbb{N}^n \mid |\alpha| = \sum_{\mathrm{i}=0}^{\mathrm{n}} a_i \leq \mathrm{t} \right\} \qquad (۳۵)$$

مجموعه‌ی تمام تک‌جمله‌ای‌های با درجه‌ی کمتر از t به‌صورت $\mathbb{T}_t^n = \left\{ \boldsymbol{x}^\alpha \mid \alpha \in \mathbb{N}_t^n \right\}$ خواهدبود. هم‌چنین، درجه‌ی چندجمله‌ای p به‌صورت $\deg(p) = \max\left\{ t \mid p_\alpha \neq 0 \;\; \text{for some } \alpha \in \mathbb{N}_t^n \right\}$ تعریف می‌گردد. هم‌چنین نماد $d_p \triangleq \lceil \deg(p)/2 \rceil$ تعریف می‌گردد.

برای حالت دوبعدی ممان (i,j)ام برای اندازه‌ی احتمال μ با نماد $y_{i,j}$ به‌صورت $y_{i,j} = \int x_1^i x_2^j d\mu$ تعریف می‌گردد. برای جلوگیری از ابهام می‌توان به‌صورت $d\mu = \mu(d\boldsymbol{x}) = \mu(d(x_1, x_2))$ هم نوشت. در حالت کلی ممان به‌صورت زیر تعریف می‌شود:

$$y_\alpha = \int x^\alpha d\mu = \int x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n} d\mu \qquad (۳۶)$$

اگر y_α را برای αهای مختلف به‌صورت رشته نوشته شود، رشته‌ی ممان حاصل می‌شود و به‌صورت $(y_\alpha)_{\alpha \in N^n}$ نمایش داده می‌شود. برای چندجمله‌ای $g \in \mathbb{R}[\boldsymbol{x}]$ و رشته‌ی $y \in \mathbb{R}^{\mathbb{N}^n}$، رشته‌ی شیفت انتقالی با نماد gy به‌صورت زیر تعریف می‌گردد [۱۳]:

$$gy \triangleq \boldsymbol{M}(y) g \in \mathbb{R}^{\mathbb{N}^n} \qquad (۳۷)$$

$$(gy)_\alpha = \sum_\beta g_\beta y_{\alpha+\beta} \qquad (۳۸)$$

برای رشته‌ی ممان $y = (y_\alpha)_{\alpha \in \mathbb{N}^n} \in \mathbb{R}^{\mathbb{N}^n}$، ماتریس ممان با نماد $\boldsymbol{M}(y)$ نمایش داده می‌شود. این ماتریس با $\mathbb{N}^n$ مقیاس می‌شود و درایه ی (α, β) ماتریس برابر با $y_{\alpha+\beta}$ می‌باشد. به‌صورت مشابه می‌توان برای رشته‌ی ممان تکه‌شده $y = (y_\alpha)_{\alpha \in \mathbb{N}_{2t}^n} \in \mathbb{R}^{\mathbb{N}_{2t}^n}$، ماتریس ممان تکه‌شده از درجه‌ی t، با نماد $\boldsymbol{M}_t(y)$ تعریف نمود. در حالت دوبعدی ماتریس ممان $\boldsymbol{M}_t(y)$ به‌صورت بلوکی از $\{\boldsymbol{M}_{i,j}(y)\}_{\substack{0 \leq j \leq t \\ 0 \leq i \leq t}}$ است، به‌طوری‌که هر بلوک به‌صورت یک ماتریس متقارن مانند زیر است:

$$\boldsymbol{M}_{i,j}(y) = \begin{bmatrix} y_{i+j,0} & y_{i+j-1,1} & \dots & y_{i,j} \\ y_{i+j-1,1} & y_{i+j-2,2} & \dots & y_{i-1,j+1} \\ \dots & \dots & & \dots \\ y_{i,j} & y_{i+j-1,1} & \cdots & y_{0,i+j} \end{bmatrix}$$

به‌عنوان مثال، در حالت دوبعدی $(n=2)$ و از درجه یک $(t=1)$ ماتریس $\boldsymbol{M}_1(y)$ ماتریس به‌صورت زیر است:

$$M_2(y) = \left[\begin{array}{c|cc|cc|cc} 1 & y_{1,0} & y_{0,1} & y_{2,0} & y_{1,1} & y_{0,2} \\ \hline y_{1,0} & y_{2,0} & y_{1,1} & y_{3,0} & y_{2,1} & y_{1,2} \\ y_{0,1} & y_{1,1} & y_{0,2} & y_{2,1} & y_{1,2} & y_{0,3} \\ \hline y_{2,0} & y_{3,0} & y_{2,1} & y_{4,0} & y_{3,1} & y_{2,2} \\ y_{1,1} & y_{2,1} & y_{1,2} & y_{3,1} & y_{2,2} & y_{1,3} \\ y_{0,2} & y_{1,2} & y_{0,3} & y_{2,2} & y_{1,3} & y_{0,4} \end{array}\right]$$

در حالت سه بعدی، ماتریس با استفاده از بلوک‌هایی از $\{\boldsymbol{M}_{i,j,k}(y)\}_{0\le i,j,k\le t}$ به‌طور مشابه تعریف می‌گردد. شایان ذکر است که ماتریس ممان برای رشته‌ی شیفت انتقالی qy نیز به‌صورت مشابه تعریف می‌شود. در ادامه دو ویژگی مهم ماتریس ممان مطرح می‌گردد [۱۸].

ویژگی اول ماتریس ممان: برای رشته $y \in \mathbb{R}^{\mathbb{N}_{2t}}$ تعریف‌شده با اندازه‌ی μ، ماتریس $M_t(y)$ مثبت نیمه‌معین است $(M_t(y) \succcurlyeq 0)$.

ویژگی دوم ماتریس ممان: برای رشته $y \in \mathbb{R}^{\mathbb{N}_{2t}}$ تعریف‌شده با اندازه‌ی μ بر مجموعه‌ی $\{\boldsymbol{x} \in \mathbb{R}^n \mid g(\boldsymbol{x}) \ge 0\}$ با فرض $t \ge d_g$، آن‌گاه ماتریس $\boldsymbol{M}_{t-d_g}(gy)$ مثبت نیمه معین است $\boldsymbol{M}_{t-d_g}(gy) \succcurlyeq 0$. هم‌چنین هنگامی‌که $h(\boldsymbol{x}) = 0$ باشد، باتوجه‌به رابطه‌ی فوق، $\boldsymbol{M}_{t-d_h}(hy) = 0$ خواهدبود.

بدین‌ترتیب، هنگامی‌که اندازه‌ی μ بر مجموعه $\{\boldsymbol{x} \in \mathbb{R}^n \mid g_1(\boldsymbol{x}) \ge 0, \dots, g_L(\boldsymbol{x}) \ge 0\}$ تعریف شده‌باشد، شرط لازم برای این‌که y یک رشته‌ی ممان باشد، به شرح زیر است:

$$\boldsymbol{M}_t(y) \succcurlyeq 0 \qquad (۳۹)$$

$$\boldsymbol{M}_{t-d_{g_i}}(gy) \succcurlyeq 0 \quad \text{for } i = 1, \dots, L \qquad (۴۰)$$

مراجع

[۱] مهری ملالو، فاطمه زارع میرک آباد. «پیدا کردن موتیف در نواحی بالادست ژن‌های هم بیان بر اساس الگوریتم بهینه سازی فاخته و سرمایش تدریجی» مجله مهندسی برق دانشگاه تبریز. ۴۶، ۳۳۳-۳۴۴،پاییز ۱۳۹۵.

[۲] رسول صادقی، فردین ابدالی محمدی. «ارائه یک روش یادگیری ویژگی ترکیبی مبتنی بر الگوریتم شبیه‌سازی تبرید و برنامه‌نویسی ژنتیک (مطالعه موردی: تشخیص بدخیمی سرطان سینه) » مجله مهندسی برق دانشگاه تبریز. ۴۸، ۱۲۷-۱۳۶ بهار ۱۳۹۷.

[3] B. Alberts, K. Roberts, J. Lewis, D. Bray, K. Hopkin, A. Johnson, P. Walter, and M. Raff,. Essential Cell Biology. Garland Science, 2013.

[4] J. Shendure, S. Balasubramanian, G. Church, W. Gilbert, J. Rogers, J. Schloss, and R. Waterston, "DNA sequencing at 40: past, present and future." Nature, Vol. 550, 2017.

[5] A. Motahari, G. Bresler and D. Tse, "Information theory of DNA shotgun sequencing." IEEE Transactions on Information Theory. Vol. 59, pp.6273-6289, 2013.

[6] C. Cai, S. Sanghavi, and H. Vikalo "Structured low-rank matrix factorization for haplotype assembly." IEEE Journal of Selected Topics in Signal Processing Vol. 10.4, pp. 647-657, 2016.

[7] M. Snyder, A. Adey, J. Kitzman, and J. Shendure, "Haplotype-resolved genome sequencing: experimental methods and applications". Nature Reviews. Vol. 16(6), 2015.

[8] G. Klau, and T. Marschall. "A guided tour to computational haplotyping." Conference on Computability in Europe. Springer, 2017.

[9] V. Bansal and V. "Hapcut: an efficient and accurate algorithm for the haplotype assembly problem" Bioinformatics, Vol. 24(16), 2008.

[10] E. Berger, D. Yorukoglu, J. Peng, and B. Berger, "Haptree: A novel bayesian framework for single individual polyplotyping using NGS data" PLoS computational biology, Vol. 10(3), 2014.

[11] J. Duitama, G. McEwen, T. Huebsch, S. Palczewski, S. Schulz, K. Verstrepen, E. Suk, and M. Hoehe, "Fosmid-based whole genome haplotyping of a hapmap trio child: evaluation of single individual haplotyping techniques" Nucleic acids research, Vol. 40(5), 2011.

[12] H. Si, H. Vikalo, and S. Vishwanath. "Information-theoretic analysis of haplotype assembly" IEEE Transactions on Information Theory, Vol. 63(6), 2017.

[13] J. Lasserre, "Global optimization with polynomials and the problem of moments." SIAM Journal on Optimization Vol. 11.3, pp796-817, 2001.

[14] D. Henrion, J. Lasserre "GloptiPoly: Global optimization over polynomials with Matlab and SeDuMi." ACM Transactions on Mathematical Software, Vol. 29.2, pp165-194, 2003.

[15] F. Geraci, "A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem," Bioinformatics, Vol. 26.18, pp.2217–2225, 2010.

[16] M. Laurent, "Sums of squares, moment matrices and optimization over polynomials." Emerging applications of algebraic geometry. Springer New York, pp. 157-270, 2009.

[17] H. Royden, and P. Fitzpatrick. Real Analysis. Pearson. 2010.

[18] P. Parrilo, "Semidefinite programming relaxations for semialgebraic problems." Mathematical programming Vol. 96.2, pp. 293-320, 2003.

زیرنویس‌ها

[1] Nucleic acid
[2] Homologous
[3] Adenine
[4] Cytosine
[5] Guanine
[6] Thymine
[7] Frederick Sanger
[8] Next Generation Sequencing (NGS)
[9] Read
[10] Assembly
[11] De novo sequence assembly
[12] Paired-end
[13] Single Nucleotide Polymorphism (SNP)
[14] Homozygote
[15] Heterozygote
[16] Genotype
[17] Measure
[18] Feasible set
[19] Moment space
[20] Countably additivity
[21] Translation invariant
[22] Dirac measure