

تخمین هاپلوتایپ با استفاده از ریلکس سازی بهینه سازی چندجمله ای

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 داده های توالی DNA می پردازد. الگوریتم پیشنهادی با استفاده از ریلکس سازی بهینه سازی چندجمله ای به روش Lasserre با نام HapLas معرفی می شود. این الگوریتم برپایه استفاده از ساختار گسسته مساله بهینه سازی تخمین هاپلوتایپ می باشد که با استفاده از تئوری اندازه به یک فضای پیوسته نگاشت می گردد. سپس با استفاده از خواص ماتریس ممان، ریلکس سازی انجام می گیرد. نتایج شبیه سازی نشان می دهد که استفاده از الگوریتم پیشنهادی منجر به بهبود نرخ بازسازی هاپلوتایپ در مقایسه با الگوریتم های متداول SDhaP و RefHap در حدود ۵ درصد می گردد. این بهبود به ازای افزایش قابل ملاحظه زمان اجرا و پیچیدگی محاسبات حاصل می شود به طوری که در کاربردهای پزشکی قابل صرف نظر کردن است.

واژه های کلیدی: هاپلوتایپ، تخمین، بهینه سازی، ریلکس سازی، ماتریس مثبت معین، تئوری اندازه.

Haplotype Estimation Using Polynomial Optimization Relaxation

Sina Majidian¹, PhD student; Mohammad Hossein Kahaei², 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. باتوجه به حجم عظیم داده‌های در دسترس 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 [۱-۲]. 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 به یکدیگر متصل هستند. آمینواسیدها در پروتئین توسط پیوندی به نام پیوند پپتیدی به یکدیگر متصل هستند. نقش پروتئین در حیات بسیار ضروری است. از آنجایی که 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

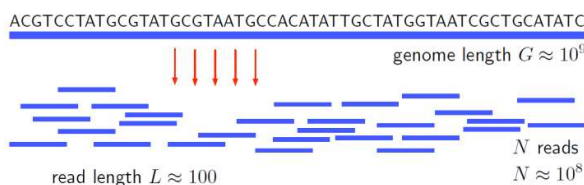
نوکلئیک اسید^۱ یکی از اجزا تشکیل‌دهنده هسته‌ی سلول انسان است که بر دو نوع ریبونوکلئیک اسید (RNA) و دئوکسی‌ریبونوکلئیک اسید (DNA) یافت می‌شود. DNA وظیفه‌ی ذخیره‌ی اطلاعات ژنتیکی را برعهده دارد. درواقع، عاملی که باعث انتقال خصوصیات و ویژگی‌های یک نوع جاندار، از نسلی به نسل دیگر می‌شود، ماده ژنتیک نام دارد. 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 کروموزوم در سلول قرار دارد. هر کروموزوم ساختمان درهم پیچیده و فشرده‌شده‌ی یک پلی‌مر بسیار بلند است، که در انسان‌ها تعداد دو سری ۲۳ تا (۴۶ عدد) می‌باشد. به جاندارانی مانند انسان که دونسخه کروموزوم دارند، دیپلوئید می‌گویند. هرکدام از این دونسخه کروموزوم طی تولیدمثل از هر یک از والدین به ارث رسیده‌است که به آن دو، نسخه پدری و مادری می‌گویند. کروموزوم‌های متناظر از هر جفت را کروموزوم‌های همسان^۲ گویند. این همسانی، به معنی دقیقاً یکسان نیست، بلکه بدین معنی است که ممکن است قسمت کوچکی از یک کروموزوم خاص پدری در یک فرد با نسخه‌ی مادری همان فرد متفاوت باشد [۳].

هر قسمتی از کروموزوم متناظر با عمل‌کرد یا ویژگی ظاهری افراد مانند رنگ چشم، اندازه، قد و ... است، که ژن نام دارد. DNA از دو رشته‌ی 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 تیمین^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 قابل‌ملاحظه‌ای را در جهان به وجود آورده‌است. برای مثال یک رشته ATCCCGGACCCT می‌تواند توالی بخشی از DNA یک انسان باشد.

۲-۱- 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 تیمین 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 سَنگِر^۷، سَنگِر 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 روی هم قرار دادن بخش‌های هم‌پوشان توالی کل رشته‌ی 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.

محاسباتی تخمین هاپلوتاایپ گفته می‌شود. دانستن هاپلوتاایپ ضروری و بسیار کارگشا می‌باشد؛ درمان اختصاصی مبتنی بر هاپلوتاایپ هر فرد به‌عنوان یکی از شاخه‌های پزشکی آینده بسیار مورد توجه قرار گرفته است. با استفاده از هاپلوتاایپ، می‌توان اطلاعات قابل توجهی در باره ژنتیک جمعیتی، مهاجرت و انتخاب طبیعی، روند تکاملی موجودات، نرخ باز ترکیب ژنتیکی نسل‌ها به دست آورد. دو راهبرد کلی برای تخمین هاپلوتاایپ وجود دارد [۸]:

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

ب) گردهم‌آوری هاپلوتاایپ: این راهبرد بر پایه داده‌های NGS است که برخلاف راهبرد جمعیتی، می‌تواند هاپلوتاایپ یک فرد را تخمین بزند. از الگوریتم‌های معروف این راهبرد می‌توان به HapTree [۹]، 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 هاپلوتاایپ به‌صورت ریاضی مدل می‌گردد. در بخش سوم الگوریتم پیشنهادی مطرح خواهد شد. نتایج شبیه‌سازی و نتیجه‌گیری به ترتیب در بخش ۴ و ۵ ارائه خواهد شد.

۲- مدل مساله

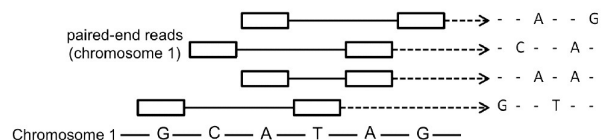
فرض کنید تعداد n خوانش در دسترس باشد. خوانش i ام را با نماد r_i نمایش می‌دهیم. مجموعه‌ی r شامل خوانش‌ها است که به‌صورت $r = \{r_1, r_2, \dots, r_n\}$ خواهد بود. هم‌چنین، فرض بر این است که ابتدا خوانش‌ها جابجایی شده‌اند؛ این بدین معنا است که اطلاعات مربوط به محل قرارگرفتن هر یک از r_i ها در رشته DNA در دسترس می‌باشد. هم‌چنین، بخشی از خوانش که شامل اسنیپ‌ها است، جدا شده‌اند. برای هر اسنیپ، به آل فراوان عدد 1 و آل نادر عدد -1 نسبت داده شده‌است. سپس ماتریس خوانش (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 هاپلوتاایپ به طول m با استفاده از n خوانش، ابعاد ماتریس خوانش $n \times m$ خواهد بود [۱۲].

از آنجایی که 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

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 هر کدام توالی از نوکلئوتیدها را به دست می‌دهد. علاوه بر این دو بخش، خوانش جفت‌انتهای، فاصله‌ی بین این دو بخش را نیز مشخص می‌نماید. بدین ترتیب یک خوانش با طول بزرگ‌تری در دسترس قرار خواهد گرفت که ابتدا و انتهای آن مشخص است ولی میان آن مشخص نیست. بدین ترتیب خوانش به طول بسیار بزرگ‌تر در دسترس خواهد بود که منجر به افزایش دقت در نگاشت به مرجع می‌شود. در شکل (۲) یک مثال ساده از خوانش جفت‌انتهای 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 هاپلوتاایپ

متداول‌ترین اختلاف در جمعیت به‌صورت تکی در بین نوکلئوتیدها می‌باشد که چندریختی تک‌نوکلئوتیدی^{۱۳} (اسنیپ) خوانده می‌شود. یک اسنیپ جایگاهی از ژن است که در بین افراد جامعه، نوکلئوتیدهای مختلفی مشاهده می‌شود. رشته‌ی اسنیپ بر روی هر یک از جفت کروموزم همسان، هاپلوتاایپ نام دارد. به‌حالتی که اسنیپ برای یک فرد بر روی جفت کروموزم، یکسان باشد هموزایگت^{۱۴} نامیده می‌شود، در غیر این صورت هتروزیگت^{۱۵} نام دارد [۷].

هنگامی که توالی نوکلئوتید در اسنیپ‌ها بر روی جفت کروموزم (نه به طور مجزا) مورد نظر باشد، به آن ژنوتاایپ^{۱۶} گفته می‌شود. اکثر روش‌های آزمایشگاهی توالی‌یابی به یافتن ژنوتاایپ منجر می‌گردد که بسیار فراگیر و با هزینه‌ی بسیار کم در دسترس قرار دارد. البته توانایی توالی‌یابی هاپلوتاایپ با روش‌های هزینه‌بر امکان‌پذیر هست. هنگامی که داده ژنوتاایپ در دسترس است، به یافتن هاپلوتاایپ با استفاده از روش‌های

بهینه سازی رتبه یک است که در حالت توالی یابی دقیق (بدون نویز) به صورت زیر است:

$$\text{Find } \mathbf{M} \text{ s.t. } \begin{cases} P_{\Omega}(\mathbf{M}) = P_{\Omega}(\mathbf{R}) \\ \text{rank}(\mathbf{M}) = 1 \end{cases} \quad (7)$$

که در آن $\mathbf{R}$ ماتریس داده های در دسترس و $\mathbf{M}$ مجهول بهینه سازی است. بدین ترتیب با توجه به ساختار متغیرها در رابطه ی (۵)، مساله ای معادل برای (۷) به صورت بهینه سازی چندجمله ای ارائه می گردد:

$$\text{Find } \mathbf{u}, \mathbf{h} \text{ s.t. } u_i h_j - R_{ij} = 0 \quad \forall (i, j) \in \Omega \quad (8)$$

از آنجایی که متغیرهای مساله به صورت اعداد ± 1 تعریف شده است، می توان به صورت زیر بهینه سازی را تکمیل نمود:

$$\text{Find } \mathbf{u}, \mathbf{h} \text{ s.t. } \begin{cases} u_i h_j - R_{ij} = 0 \quad \forall (i, j) \in \Omega \\ h_j \in \{1, -1\} \quad j = 1, \dots, m \\ u_i \in \{1, -1\} \quad i = 1, \dots, n \end{cases} \quad (9)$$

به بیان معادل، قید را نیز به صورت چندجمله ای نوشت:

$$\text{Find } \mathbf{u}, \mathbf{h} \text{ s.t. } \begin{cases} u_i h_j - R_{ij} = 0 \quad \forall (i, j) \in \Omega \\ h_j^2 = 1 \quad j = 1, \dots, m \\ u_i^2 = 1 \quad i = 1, \dots, n \end{cases} \quad (10)$$

که یک مساله چندجمله ای کلاسیک است که برای حل مساله از پکیج های نرم افزاری بهینه سازی می توان استفاده نمود. هم چنین در حالت نویزی می توان مساله بهینه سازی را به شرح زیر نوشت:

$$\min_{\mathbf{u}, \mathbf{h}} \|P_{\Omega}(\mathbf{R} - \mathbf{u}\mathbf{h}^T)\|_F \quad (11)$$

برای حل آن از روش های عددی از جمله روش تکراری گرادینتی استفاده شده است [۹، ۴]. این روش حل از ساختاری که از قبل دانسته فرض شده استفاده نکرده است. به عنوان یکی از نوآوری های این مقاله ارائه مساله بهینه سازی چندجمله ای ساختاریافته برای تخمین هاپلوتاایپ است که به صورت زیر می باشد:

$$\min_{\mathbf{u}, \mathbf{h}} \sum_{(i,j) \in \Omega} (u_i h_j - R_{ij})^2 \text{ s.t. } \begin{cases} h_j^2 = 1 \quad j = 1, \dots, m \\ u_i^2 = 1 \quad i = 1, \dots, n \end{cases} \quad (12)$$

به منظور ساده سازی، بردار $\mathbf{x}$ به طول $m+n$ با کنارهم گذاشتن دو بردار مجهول به صورت زیر تعریف می گردد:

$$\mathbf{x} = [u_1, \dots, u_n, h_1, \dots, h_m] \quad (13)$$

بدین ترتیب مساله بهینه سازی برای دو حالت دقیق و نویزی به صورت زیر خواهد بود:

$$\text{Find } \mathbf{x} \text{ s.t. } \begin{cases} x_i x_{j+n} - R_{ij} = 0 \quad \forall (i, j) \in \Omega \\ x_j^2 = 1, \quad j = 1, \dots, m+n \end{cases} \quad (14)$$

$$\min_{\mathbf{x} \in \mathbb{R}^{m+n}} \sum_{(i,j) \in \Omega} (x_i x_{j+n} - R_{ij})^2 \quad (15)$$

$$\text{s.t. } x_j^2 = 1, \quad j = 1, \dots, m+n$$

هاپلوتاایپ مشخص می گردد؛ زیرا ابهامی ندارد و نیاز به تخمین نیست. در ادامه مدل سازی ریاضی رشته خوانش و هاپلوتاایپ مطرح می گردد. بدین منظور برای تخمین جایگاه های هتروزیگت مثالی عددی ساده بدون نویز مطرح می گردد دو رشته هاپلوتاایپ به صورت زیر در نظر بگیرد:

$$\mathbf{h}_1 = \begin{bmatrix} 1 & 1 & 1 & -1 \end{bmatrix}^T \quad (1)$$

$$\mathbf{h}_2 = \begin{bmatrix} -1 & -1 & -1 & 1 \end{bmatrix}^T$$

در ادامه، مثالی از خوانش ها ارائه شده است:

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

در نمایش فوق، اگر درایه ای به صورت $\sim$ باشد، بدین معنا است که این درایه در خوانش مربوطه قرار نداشته است. بدین ترتیب ماتریس خوانش به صورت زیر خواهد بود:

$$\mathbf{R} = \begin{bmatrix} 1 & \sim & \sim & -1 \\ \sim & 1 & 1 & \sim \\ \sim & -1 & -1 & \sim \\ -1 & \sim & \sim & 1 \end{bmatrix} \quad (2)$$

پس مدل سازی ریاضی به صورت زیر خواهد بود [۶]:

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix} \quad (3)$$

اما از آنجایی که در دیپلوتیدها خوانش ها از هاپلوتاایپ مادری ($\mathbf{h}_m$) و یا پدری ($\mathbf{h}_p$) توالی یابی شده اند، برای اسنپ های باقی مانده (هتروزیگت) مقادیر قرینه یک دیگر هستند؛ به بیان دیگر $\mathbf{h}_p = -\mathbf{h}_m$ برقرار است. پس برای مثال فوق به صورت زیر خواهد بود:

$$\mathbf{M} = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & -1 \end{bmatrix} \quad (4)$$

بدین ترتیب مدل برداری زیر ارائه می گردد:

$$\mathbf{M} = \mathbf{u}\mathbf{h}^T \quad (5)$$

که در آن بردار $\mathbf{u}$ با ابعاد $n \times 1$ با درایه های ± 1 است. هنگامی که درایه $+1$ باشد حاکی از آن است که این خوانش مربوط به توالی هاپلوتاایپ $\mathbf{h}_p$ است هم چنین درایه -1 ، به معنای خوانش از $\mathbf{h}_m$ است. هم چنین بردار $\mathbf{h}$ برداری با ابعاد $m \times 1$ است، که شامل درایه های ± 1 می باشد. بدین ترتیب ماتریس $\mathbf{R}$ را می توان به صورت زیر نوشت:

$$\mathbf{R} = P_{\Omega}(\mathbf{M}) = \begin{cases} \mathbf{M}_{ij}, & (i, j) \in \Omega \\ 0, & o.w. \end{cases} \quad (6)$$

که در آن Ω مجموعه ای است که خوانش ها در آن مکان اطلاعات ارائه کرده است. $P_{\Omega}(\mathbf{M})$ به عنوان عملگر نمونه بردار است، پس ماتریس مشاهدات $\mathbf{R}$ برابر با نسخه ی نمونه برداری شده از $\mathbf{M}$ خواهد بود. یکی از نوآوری های این مقاله صورت مساله تخمین هاپلوتاایپ به صورت

۳- ریلکس سازی بهینه سازی چندجمله ای

در این بخش، ریلکس سازی یک مساله چندجمله ای به یک مساله بهینه سازی مثبت نیمه معین در حالت کلی معرفی می گردد. مقدماتی از مفهوم اندازه^{۱۲} و ماتریس ممان در بخش پیوست آورده شده است. برای ریلکس سازی بهینه سازی چندجمله ای به یک گام میانی نیاز است. این گام، تبدیل مساله بهینه سازی چندجمله ای به یک مساله بهینه سازی معادل بر روی فضای اندازه می باشد.

در ابتدا، چندجمله ای به صورت دقیق معرفی می گردد. یک چندجمله ای n بعدی $p(x): \mathbb{R}^n \rightarrow \mathbb{R}$ با نماد $p(x) = \sum_{\alpha} p_{\alpha} x^{\alpha}$ استفاده خواهد شد، 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. هم چنین $t = \sum_i \alpha_i$ درجه ی چندجمله ای خواهد بود. مساله بهینه سازی چندجمله ای به صورت زیر می باشد:

$$\min p(x) \quad \text{s.t.} \quad \begin{cases} g_i(x) \geq 0 & i = 1, \dots, I, \\ h_l(x) = 0 & l = 1, \dots, L \end{cases} \quad (16)$$

که در آن $g_i(x)$ و $h_l(x)$ چندجمله ای می باشند. هم چنین مجموعه ی شدنی^{۱۸} (مجموعه ی x هایی که در قید صدق می کنند) برای مساله بهینه سازی فوق به صورت

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

تعریف می شود. در [۱۳] ثابت شده است که مساله بهینه سازی (۱۶) معادل مساله چندجمله ای بهینه سازی زیر است:

$$\min_{\mu \in MS(K)} \int p(x) d\mu \quad (17)$$

که در آن $MS(K)$ ^{۱۹} مجموعه ی فضای تمام اندازه های احتمال ممکن تعریف شده روی مجموعه ی K می باشد. در ابتدا حالتی که مساله بهینه سازی چندجمله ای بدون قید، مدنظر قرار می گیرد. با استفاده از تعریف ممان می توان تابع هدف مساله بهینه سازی چندجمله ای را به صورت زیر نوشت:

$$\begin{aligned} \int p(x) d\mu &= \int \sum_{\alpha} p_{\alpha} x^{\alpha} d\mu = \sum_{\alpha} p_{\alpha} \int x^{\alpha} d\mu \\ &= \sum_{\alpha} p_{\alpha} y_{\alpha} \end{aligned} \quad (18)$$

بدین ترتیب مساله بهینه سازی معادل جدید به صورت زیر می باشد:

$$\min_y \sum_{\alpha} p_{\alpha} y_{\alpha} \quad \text{s.t.} \quad y \text{ is a moment sequence.} \quad (19)$$

آن گاه رابطه پاسخ مساله (۱۹) با مجهول های مساله اصلی به صورت $y = (1, x_1^*, x_2^*, x_1^{*2}, x_1^* x_2^*, x_2^{*2}, \dots)$ است. باتوجه به این که حل مساله بهینه سازی فوق با قید ذکر شده ممکن نیست. می بایست قید را ریلکس نمود. بدین منظور، باتوجه به ویژگی اول ماتریس ممان، مساله جدید بهینه سازی به صورت زیر خواهد شد:

$$\min_y \sum_{\alpha} p_{\alpha} y_{\alpha} \quad \text{s.t.} \quad M_t(y) \geq 0 \quad (20)$$

از آنجایی که برای مقادیر مختلف t مساله حل می شود به این راهبرد که توسط آقای Lasserre معرفی شده است Lasserre Hierarchy گویند. این یک مساله استاندارد مثبت نیمه معین (SDP) است که شامل تابع هدف خطی و یک قید به صورت شرط مثبت نیمه معین بودن ماتریس می باشد.

هم چنین در حالت مقید، باتوجه به تعریف مجموعه K ، مساله جدید به صورت زیر خواهد بود [۱۳]:

$$\min_y \sum_{\alpha} p_{\alpha} y_{\alpha} \quad \text{s.t.} \quad \begin{cases} M_t(y) \geq 0 \\ M_{t-d_{g_i}}(g_i y) \geq 0 \\ M_{t-d_{h_l}}(h_l y) = 0 \end{cases} \quad (21)$$

بدین ترتیب مساله چندجمله ای که در حالت کلی محدب نیست، به یک مساله محدب ریلکس شده است. این مساله با نرم افزار MATLAB ابزار CVX و هم چنین GloptiPoly [۱۴] حل می گردد.

۴- الگوریتم پیشنهادی: الگوریتم تخمین هاپلوتاایپ با استفاده از ریلکس سازی بهینه سازی چندجمله ای به روش Lasserre

مساله بهینه سازی مطرح شده بر حسب متغیر x در رابطه (۱۴) را در نظر بگیرید. در این مساله تعداد $|\Omega| + m + n$ قید تساوی وجود دارد. منظور از $|\Omega|$ تعداد اعضای یک مجموعه است. بدین ترتیب باتوجه به مقدمات مطرح شده در بخش قبل، مساله ریلکس شده در حالت دقیق به صورت زیر خواهد بود:

$$\text{Find } y \quad \text{s.t.} \quad \begin{cases} M_t(y) \geq 0 \\ M_{t-d_{h_i}}(h_i y) = 0 & i = 1, \dots, |\Omega| + m + n \end{cases} \quad (22)$$

هم چنین در حالت نویزی به صورت زیر خواهد بود:

$$\min_y \sum_{\alpha} p_{\alpha} y_{\alpha} \quad \text{s.t.} \quad \begin{cases} M_t(y) \geq 0 \\ M_{t-d_{h_i}}(h_i y) = 0 & i = 1, \dots, m + n \end{cases} \quad (23)$$

که در آن p_{α} متناظر با ضرایب چندجمله ای زیر است:

$$p(x) = \sum_{\alpha} p_{\alpha} x^{\alpha} = \sum_{(i,j) \in \Omega} x_{ij}^2 - 2x_{ij} R_{ij} \quad (24)$$

عبارت فوق با حذف ترم ثابت R_{ij}^2 از $\sum_{(i,j) \in \Omega} (x_{ij} x_{j+n} - R_{ij})^2$ به دست آمده است. بدین ترتیب این بهینه سازی بخش اصلی الگوریتم معرفی شده در این مقاله است که در شکل ۳ آورده شده است.

Algorithm 1: Haplotype estimation using Lasserre hierarchy relaxations (HapLas).	
Input: n aligned reads	
Output: Haplotype $\hat{h}$	
Read Matrix Preparation	
1	Convert the sequences of nucleotides (reads) to the sequences of numbers
2	Add zeros to each read to construct r_i with the length of m
3	Construct the read matrix R ($n \times m$)
4	Construct the constraints from R
** Optimization part **	
5	Solve the following optimization
$\min_y \sum_{\alpha} p_{\alpha} y_{\alpha} \quad \text{s.t.} \quad \begin{cases} M_t(y) \geq 0 \\ M_{t-d_{h_i}}(h_i y) = 0 & i = 1, \dots, m + n \end{cases}$	
6	Extract the x_i s from moment sequence by
$y^* = (1, x_1^*, x_2^*, \dots, x_n^*, x_1^* x_2^*, \dots)$	
Haplotype finding	
7	Find $\hat{h}$ through $\hat{h} = [x_{n+1}, \dots, x_{n+1}]$
8	Convert the entries of $\hat{h}_m$ and $\hat{h}_p$ to the nucleotides.

شکل ۳: الگوریتم پیشنهادی HapLas

۵- شبیه سازی

برای بررسی عملکرد روش پیشنهادی از معیار جذر میانگین مربع خطا (RMSE) استفاده می گردد که به صورت زیر تعریف شده است:

$$RMSE^2 = \frac{1}{mK} \sum_{k=1}^K \sum_{i=1}^m (\hat{h}_i^{(k)} - h_i)^2 \quad (25)$$

که در آن K تعداد تکرار است که ۱۰ در نظر گرفته شده است. همچنین $\hat{h}_i^{(k)}$ هاپلوتاایپ تخمین زده شده در جایگاه i ام و تکرار k ام می باشد. علاوه بر این از معیار نرخ بازسازی (rr) که در زیر تعریف شده است، نیز استفاده می شود [۱۵]:

$$rr = \frac{1 - hd(\hat{h}, h)}{m} \quad (26)$$

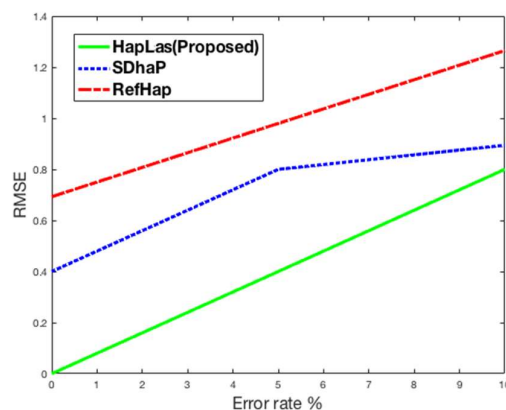
$$hd(\hat{h}, h) = \frac{1}{K} \sum_{k=1}^K d(\hat{h}_i^{(k)}, h_i) \quad (27)$$

که در آن فاصله $d(.,.)$ به صورت زیر تعریف می گردد:

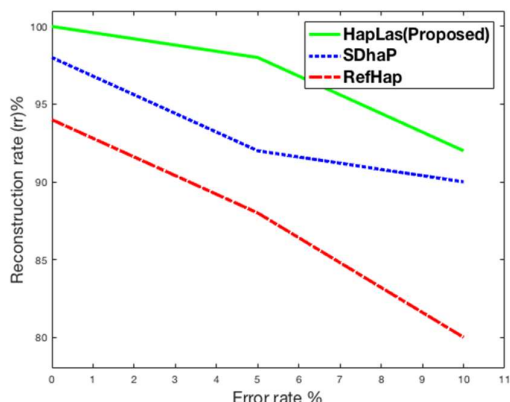
$$d(a, b) = \begin{cases} 0, & a = b \\ 1, & o. w. \end{cases} \quad (28)$$

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

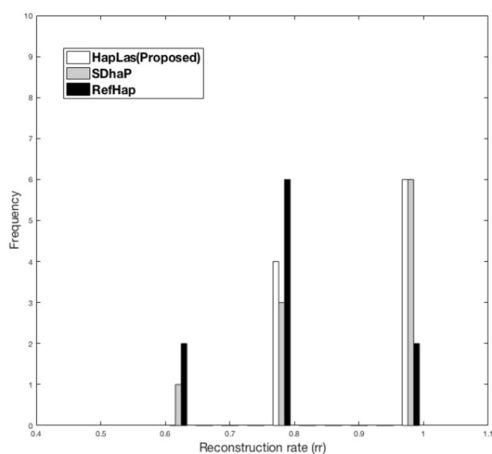
طی درخواستی از نویسنده مقاله [۱۵]، داده های آن دریافت شد و با استفاده از آن شبیه سازی های مختلفی انجام شده است. در ابتدا یک آزمایش برای هاپلوتاایپ به طول ۵ با در نظر داشتن ۵ خوانش، با پوشش برابر با ۳ (تعداد درایه معلوم در هر ستون) در نظر گرفته شده است. در شکل های ۴ و ۵ به ترتیب نمودار جذر میانگین مربع خطا (RMSE) و نمودار نرخ بازسازی (rr) بر حسب درصد نرخ خطا در آزمایش ذکر شده رسم شده است. نتایج همه تکرارهای آزمایش اول، به صورت هیستوگرام برای نرخ خطا ۱۰، ۵ و بدون خطا به ترتیب در شکل های ۶، ۷ و ۸ رسم شده است.



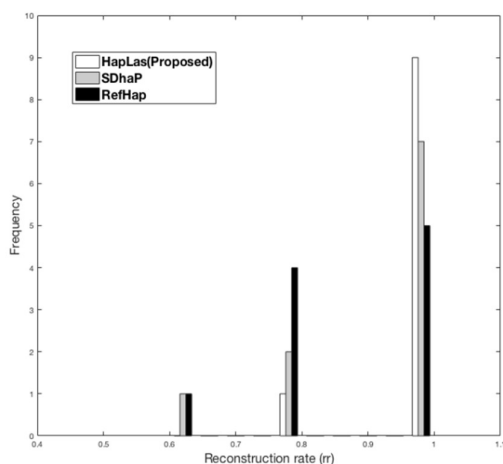
شکل ۴: نمودار RMSE بر حسب درصد نرخ خطا در آزمایش اول.



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



شکل ۶: هیستوگرام نرخ بازسازی برای نرخ خطا ۱۰ درصد

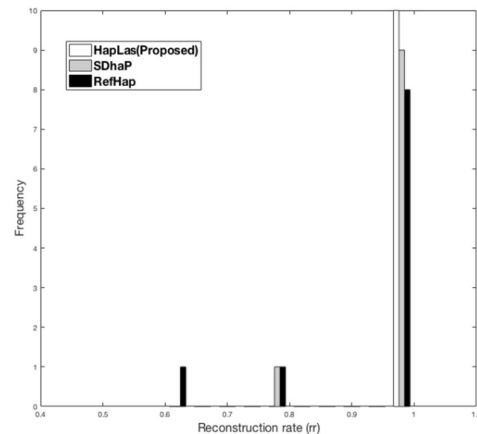
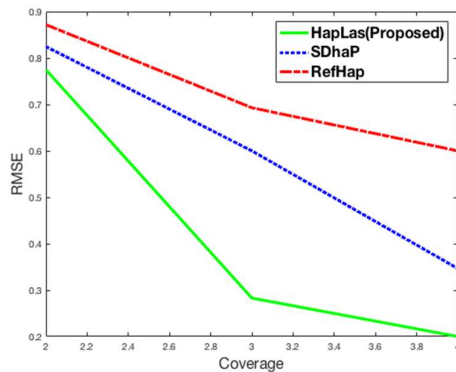


شکل ۷: هیستوگرام نرخ بازسازی برای نرخ خطا ۵ درصد.

جدول ۲: زمان اجرای الگوریتم پیشنهادی در مقایسه با الگوریتم های

دیگر برحسب ثانیه برای آزمایش دوم

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



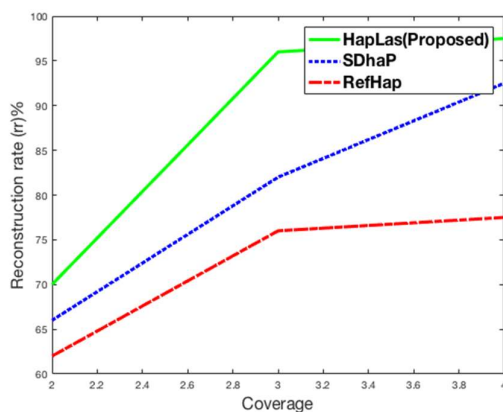
شکل ۸: هیستوگرام نرخ بازسازی برای حالت بدون خطا.

متوسط زمان اجرای الگوریتم در جدول ۱ ارائه شده است. همان طور که در شکل های فوق مشاهده می شود با افزایش خطای داده میزان خطای تخمین RMSE افزایش یافته و میزان بازسازی rr کاهش می یابد. همچنین، الگوریتم HapLas از دو الگوریتم دیگر SDhaP و RefHap تخمین بهتری از نظر MSE و همچنین rr در ازای زمان بیش تر ارائه می دهد. علاوه بر آن، باتوجه به 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 به راحتی بتواند هاپلوتاایپ را تخمین بزند، در حالی که روش SDhaP و RefHap در موارد معدودی دچار خطا می شوند.

جدول ۱: زمان اجرای الگوریتم پیشنهادی در مقایسه با الگوریتم های

دیگر برحسب ثانیه برای آزمایش اول

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



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

۶- نتیجه گیری

در این مقاله، الگوریتم تخمین هاپلوتاایپ با استفاده از ریلکس سازی بهینه سازی چندجمله ای (HapLas) معرفی شد. بدین منظور از خاصیت مثبت نیمه معین بودن ماتریس ممان برای ریلکس سازی یک مساله بهینه سازی غیر قابل حل استفاده شد. نتایج شبیه سازی مشخص نمود که روش پیشنهادی منجر به افزایش نرخ بازسازی هاپلوتاایپ در حدود ۵ درصد می شود. بدین ترتیب با افزایش دقت تخمین هاپلوتاایپ، بهبود کیفیت در درمان بیماری های ژنتیکی مورد انتظار است. درازای این بهبود محاسبات بیشتری مورد نیاز است. از آنجایی که 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 می باشد.

در آزمایش دیگری که انجام پذیرفت، هدف تخمین هاپلوتاایپ به طول ۱۰ است که ۷ خوانش با نرخ خطا به صورت ثابت ده درصد در دسترس است. در شکل ۹ میزان RMSE برحسب میزان پوشش و همچنین در شکل ۱۰، میزان rr برحسب میزان پوشش رسم شده است. باتوجه به شکل های ۹ و ۱۰ همان طور که انتظار می رود با افزایش مقدار پوشش، میزان خطای تخمین RMSE کاهش یافته و میزان بازسازی rr افزایش می یابد. همچنین برای روش پیشنهادی میزان خطا کم ترین و میزان نرخ بازسازی بیش ترین به دست می آید. همچنین باتوجه به جدول ۲، متوسط زمان اجرای الگوریتم ارائه شده نسبت به الگوریتم های موجود بیش تر است که افزایش سرعت آن در حال حاضر در دست تحقیق است.

پیوست

معرفی ماتریس ممان

حلقه‌ی چندجمله‌ای چندمتغیره، بر روی متغیرهای n بعدی $x = (x_1, x_2, \dots, x_n)$ با نماد $\mathbb{R}[x]$ نمایش داده می‌شود. هر عضو این حلقه مانند p یک چندجمله‌ای است که به صورت $p(x) = \sum_{\alpha} p_{\alpha} x^{\alpha}$ نمایش داده می‌شود. همچنین $\mathbb{N}^n$ مجموعه‌ی تمام α های n بعدی مانند $\alpha = (\alpha_1, \dots, \alpha_n)$ است که مقادیر هر بعد، از مجموعه‌ی اعداد صحیح نامنفی $(\mathbb{N})$ آمده‌است. حال برای $t \in \mathbb{N}$ 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]:

$$\mathbb{N}_t^n = \left\{ \alpha \in \mathbb{N}^n \mid |\alpha| = \sum_{i=1}^n \alpha_i \leq t \right\} \quad (35)$$

مجموعه‌ی تمام تک جمله‌ای‌های با درجه‌ی کمتر از t به صورت $\mathbb{T}_t^n = \{x^{\alpha} \mid \alpha \in \mathbb{N}_t^n\}$ خواهد بود. همچنین، درجه‌ی چندجمله‌ای p به صورت $\deg(p) = \max\{t \mid p_{\alpha} \neq 0 \text{ for some } \alpha \in \mathbb{N}_t^n\}$ 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 $y_{i,j}$ به صورت $y_{i,j} = \int x_1^i x_2^j d\mu$ تعریف می‌گردد. برای جلوگیری از ابهام می‌توان به صورت $d\mu = \mu(dx) = \mu(dx_1, dx_2)$ هم نوشت. در حالت کلی ممان به صورت زیر تعریف می‌شود:

$$y_{\alpha} = \int x^{\alpha} d\mu = \int x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n} d\mu \quad (36)$$

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

$$gy \triangleq M(y)g \in \mathbb{R}^{\mathbb{N}^n} \quad (37)$$

$$(gy)_{\alpha} = \sum_{\beta} g_{\beta} y_{\alpha+\beta} \quad (38)$$

برای رشته‌ی ممان $y = (y_{\alpha})_{\alpha \in \mathbb{N}^n} \in \mathbb{R}^{\mathbb{N}^n}$ ماتریس ممان با نماد $M(y)$ نمایش داده می‌شود. این ماتریس با $\mathbb{N}^n$ مقیاس می‌شود و درایه‌ی (α, β) ماتریس برابر با $y_{\alpha+\beta}$ می‌باشد. به صورت مشابه می‌توان برای رشته‌ی ممان تکه شده $y = (y_{\alpha})_{\alpha \in \mathbb{N}_{2t}^n} \in \mathbb{R}^{\mathbb{N}_{2t}^n}$ ماتریس ممان تکه شده از درجه‌ی t ، با نماد $M_t(y)$ تعریف نمود. در حالت دوبعدی ماتریس ممان $M_t(y)$ به صورت بلوکی از $\{M_{i,j}(y)\}_{0 \leq i \leq t, 0 \leq j \leq t}$ است، به طوری که هر بلوک به صورت یک ماتریس متقارن مانند زیر است:

$$M_{i,j}(y) = \begin{bmatrix} y_{i+j,0} & y_{i+j-1,1} & \dots & y_{i,j} \\ y_{i+j-1,1} & y_{i+j-2,2} & \dots & y_{i-1,j+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{i,j} & y_{i+j-1,1} & \dots & y_{0,i+j} \end{bmatrix}$$

به عنوان مثال، در حالت دوبعدی $(n=2)$ و از درجه یک $(t=1)$ ماتریس $M_1(y)$ ماتریس به صورت زیر است:

مفهوم اندازه

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

I. جمع پذیری شمارا^{۲۱}: برای زیرمجموعه‌های مجزا که $E_i \cap E_j = \emptyset$ برای $i \neq j$ است، رابطه‌ی زیر برقرار باشد:

$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k) \quad (29)$$

II. پایا نسبت به انتقال^{۲۲}: برای مجموعه‌ی E و نقطه y ، رابطه‌ی $\mu(E+y) = \mu(E)$ برقرار باشد، به طوری که:

$$E+y = \{x+y \mid x \in E\} \quad (30)$$

به بیان دیگر مجموعه‌ی $E+y$ از جانشین کردن هر نقطه مانند x از مجموعه‌ی E به $x+y$ به دست آمده‌است.

یک اندازه معروف، اندازه دیراک^{۲۳} در نقطه‌ی x با نماد $\mu = \delta_x$ است به صورت مقابل تعریف می‌شود: اندازه آن در x یک است $\mu(\{x\}) = 1$ و در سایر نقاط صفر است. به بیان ریاضی:

$$\delta_x(E) = \begin{cases} 1, & x \in E \\ 0, & o.w. \end{cases} \quad (31)$$

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 می‌شود. بدین منظور در ابتدا، مجموعه E به زیرمجموعه‌هایی مانند E_k افراز می‌شود. منظور از افراز، تقسیم کردن اعضای یک مجموعه به چندین زیرمجموعه به گونه‌ای که دارای دو خاصیت زیر باشند:

$$\bigcup_{k=1}^K E_k = E \quad (32)$$

$$E_i \cap E_j = \emptyset \quad \forall i \neq j = 1, \dots, K \quad (33)$$

این افراز باید به گونه‌ای 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 به صورت زیر تعریف می‌شود [۱۷]:

$$\int_E f d\mu = \sum_{k=1}^K c_k \mu(E_k) \quad (34)$$

که در آن $\mu(E_k)$ اندازه هر کدام از افرازاهاست. هنگامی که اندازه کل فضا یک باشد، به آن اندازه احتمال گویند. به بیان دیگر، اگر اندازه بر روی مجموعه‌ی T تعریف شود، هنگامی که $\mu(T) = 1$ 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.

(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.

$$M_2(y) = \begin{bmatrix} 1 & | & y_{1,0} & y_{0,1} & | & y_{2,0} & y_{1,1} & y_{0,2} \\ \hline y_{1,0} & | & y_{2,0} & y_{1,1} & | & y_{3,0} & y_{2,1} & y_{1,2} \\ y_{0,1} & | & y_{1,1} & y_{0,2} & | & y_{2,1} & y_{1,2} & y_{0,3} \\ \hline y_{2,0} & | & y_{3,0} & y_{2,1} & | & y_{4,0} & y_{3,1} & y_{2,2} \\ y_{1,1} & | & y_{2,1} & y_{1,2} & | & y_{3,1} & y_{2,2} & y_{1,3} \\ y_{0,2} & | & y_{1,2} & y_{0,3} & | & y_{2,2} & y_{1,3} & y_{0,4} \end{bmatrix}$$

در حالت سه بعدی، ماتریس با استفاده از بلوک‌هایی از $\{M_{i,j,k}(y)\}_{0 \leq i,j,k \leq t}$ 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 $y \in \mathbb{R}^{N^{2t}}$ تعریف‌شده با اندازه‌ی μ ، ماتریس $M_t(y)$ مثبت نیمه‌معین است $(M_t(y) \succeq 0)$.

ویژگی دوم ماتریس ممان: برای رشته $y \in \mathbb{R}^{N^{2t}}$ تعریف‌شده با اندازه‌ی μ بر مجموعه‌ی $\{x \in \mathbb{R}^n | g(x) \geq 0\}$ با فرض $t, t \geq d_g$ آن‌گاه

ماتریس $M_{t-d_g}(gy)$ مثبت نیمه معین است $M_{t-d_g}(gy) \succeq 0$ هم‌چنین هنگامی که $h(x) = 0$ باشد، باتوجه به رابطه‌ی فوق، $M_{t-d_h}(hy) = 0$ خواهد بود.

بدین ترتیب، هنگامی که اندازه‌ی μ بر مجموعه $\{x \in \mathbb{R}^n | g_1(x) \geq 0, \dots, g_L(x) \geq 0\}$ تعریف شده‌باشد، شرط لازم برای این‌که y یک رشته‌ی ممان باشد، به شرح زیر است:

$$M_t(y) \succeq 0 \quad (39)$$

$$M_{t-d_{g_i}}(gy) \succeq 0 \quad \text{for } i = 1, \dots, L \quad (40)$$

مراجع

- [۱] مهری ملالو، فاطمه زارع میرک آباد. «پیدا کردن موتیف در نواحی بالادست ژن‌های هم بیان بر اساس الگوریتم بهینه سازی فاخته و سرمایه‌ی 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

زیر نویس‌ها

- 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

- 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