دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر
گزارش پروژه نهایی درس ژنومیک محاسباتی

عنوان:

فیزینگ سریع دومرحله‌ای داده‌های توالی بزرگ‌مقیاس
Fast two-stage phasing of large-scale sequence data

نگارش:
جواد راضی
۴۰۱۲۰۴۳۵۴

استاد درس:
دکتر مطهری

زمستان ۱۴۰۱

چکیده: فیزینگ[1] هپلوتایپ، پروسه‌ایست که در طی آن هپلوتایپ‌های به ارث رسیده از هر والد در هر لوکوس[2]، برای یک موجود دیپلوید مشخص می‌شوند. هر هپلوتایپ، مجموعه‌ای از الل‌هاییست که با یکدیگر در یک کروموزوم به ارث رسیده‌اند. اهمیت پروسه فیزینگ هپلوتایپ، در تحلیل‌ها و مطالعاتی می‌باشد که به داده‌های هپلوتایپ‌ها نیازمندند. برای مثال، در برخی از بیماری‌ها، داده‌های ژنوتایپ به تنهایی برای شناسایی جهش‌های منجر به بیماری کافی نمی‌باشد. در این گزارش، مقاله با عنوان « فیزینگ سریع دومرحله‌ای داده‌های توالی بزرگ‌مقیاس» ارائه می‌گردد. این مقاله، روشی دومرحله‌ای برای فیزینگ هپلوتایپ را که به طور خاص برای داده‌های بزرگ‌مقیاس آرایه SNP و داده‌های توالی کارآمد است را ارائه می‌دهد. در این روش، نخست SNPهای با فرکانس بالا، فیز شده و در مرحله دوم با الل‌های جانهی‌شده[3] توسط این هپلوتایپ‌ها، هپلوتایپ‌های با فرکانس پایین فیز می‌گردند. مزیت این راه‌کار نسبت به روش‌های مشابه، در استفاده بهینه از حافظه، و سرعت بالاتر محاسباتی به علت نادیده‌گرفتن SNPهای نادر در مرحله نخست الگوریتم است.

واژه های کلیدی: فیزینگ هپلوتایپ، ژنوتایپ، توالی، SNP، ژنومیک محاسباتی، بیوانفورماتیک

1 مقدمه

بسیاری از آنالیزها و مطالعات، به داده‌های حاصل از فیزینگ هپلوتایپ متکی هستند. از جمله این موارد، می‌توان به تشخیص جهش‌های منجر به برخی از بیماری‌ها، جانهی ژنوتایپ، تست‌های نسبت ژنتیکی افراد، و مطالعات ژنتیک جمعیت اشاره کرد. در این پروسه، هپلوتایپ‌های به ارث‌رسیده از هر والد استنتاج آماری می‌گردد. دقت این استنتاج، با افزایش اندازه نمونه بیشتر می‌شود. با نرخ رو به رشد حجم دیتاست‌های کلان زیستی در دسترس در سال‌های اخیر، انگیزه ارائه متدهای فیزینگ هرچه بهینه‌تر و مقیاس‌پذیرتر، بیش‌تر شده‌است. روش‌های و ابزارهایی نظیر HAPI-UR، SHAPEIT و EAGLE23 در سال‌های اخیر معرفی شده‌اند که الگوریتم‌های پیاده‌سازی‌شده توسط آن‌ها، به طور خطی با اندازه جمعیت نمونه افزایش می‌یابد و این امر، آن‌ها را قادر به تحلیل دیتاست‌های بزرگ کرده‌است.

دسترس‌بودن داده‌های توالی کل ژنوم، و دیتاست‌هایی که شامل میلیون‌ها مارکر ژنتیکی هستند، چالش‌های جدیدی را برای الگوریتم‌های فیزینگ ایجاد کرده‌است. این چالش‌ها در استفاده بهینه از حافظه، زمان محاسباتی و دقت پروسه فیزینگ ایجاد شده‌اند. در این مقاله روشی دومرحله‌ای برای پروسه فیزینگ هپلوتایپ معرفی شده که در نرم‌افزار Beagle 5.2 نیز پیاده‌سازی شده‌است. این روش برای فیزکردن، از مدل HMM[4] استفاده می‌کند. در این مدل، پنل مرجع هپلوتایپ، ترکیبی از هپلوتایپ‌هایی هستند که از چندین منبع مختلف نظیر پروژه ۱۰۰۰ ژنوم آمده‌اند. در نتیجه، هپلوتایپ‌های مرجع، از قومیت‌ها و جمعیت‌های متنوعی می‌آید که گونه‌گونی پنل مرجع را تضمین می‌کند.

در مرحله نخست روش ارائه‌شده توسط این مقاله، ابتدا مارکرهای ژنتیکی پرفرکانس با رویکرد روش فیزینگ پیشرو، فیز می‌شوند که به صورت تکرارشونده، مجموعه الل‌های فیزشده را توسعه می‌دهد. هپلوتایپ‌های فیزشده مرحله نخست، در مرحله دوم برای پروسه جانهی ژنوتایپ، و استنتاج مارکرهای ژنتیکی کم‌فرکانس از الل‌های جانهی‌شده استفاده می‌شوند. با نادیده گرفتن مارکرهای نادر در مرحله نخست که از نظر پیچیدگی محاسباتی سنگین است، این روش قادر به صرفه‌جویی در حافظه، و کاهش زمانی محاسباتی شده‌است.

برای ارزیابی عملکرد مدل پیاده‌سازی‌شده توسط این مقاله، عملکرد Beagle 5.2 در مقابل نسخه 4.2.1 ابزار SHAPEIT مقایسه شده‌است. SHAPEIT یک ابزار دیگر برای فیزینگ هپلوتایپ است که به صورت گسترده مورد استفاده بوده و برای داده‌های حجیم، عملکرد مناسبی را از خود نشان داده‌است. دو دیتاست مختلف در این مقاله برای ارزیابی عملکرد استفاده شده‌اند؛ یکی داده آرایه SNP پایگاه داده UK Biobank، و دیگری داده‌های توالی پایگاه داده TOPMed[5] می‌باشد. در نتایج ارزیابی، دو ابزار از نظر دقت، و مقیاس‌پذیری نسبت به اندازه جمعیت نمونه، مشابه بودند.

---

[1] Phasing

[2] Locus

[3] Imputed

[4] Hidden Markov Model

[5] The Trans-Omics for Precision Medicine

از نظر زمان محاسباتی نیز برای داده‌های SNP پایگاه داده Biobank، دو روش تفاوت محسوسی نداشتند. اما در فیزینگ با داده‌های توالی TOPMed، ابزار Beagle حدودا ۲۰ برابر سریع‌تر از SHAPEIT بود.

2 روش‌ها

Beagle 5.2، برای فیزینگ از مدل HMM استفاده می‌کند. احتمال حالت‌های گذر این HMM، به مقدار «اندازه جمعیت موثر» که ورودی مدل است بستگی دارند. از آنجایی که مقدار صحیح این پارامتر، برای برخی از گونه‌ها نامعلوم است، Beagle توسط HMM این مقدار را به صورت تکرارشونده بروزرسانی می‌کند. در نتیجه، حتی در صورتی که تخمین اولیه دور از مقدار واقعی باشد، مقدار «اندازه جمعیت موثر» به مقدار واقعی نزدیک‌تر شده و نتیجه تا حد زیادی مستقل از مقدار اولیه می‌شود.

در روش ارائه‌شده توسط این مقاله، از یک پنجره لغزنده مارکرها در پروسه فیزینگ استفاده می‌شود. این پنجره، سایز معینی دارد که به صورت پیش‌فرض ۴۰ سنتی‌مورگان است. ابزار Beagle، توالی را به بخش‌هایی با اندازه‌ای به نسبت اندازه پنجره مارکر تقسیم کرده، و ژنوتایپ‌های هر پنجره را به صورت مستقل فیز می‌کند. برای پنجره‌ها، اندازه‌ای به عنوان بازه هم‌پوشانی در نظر گرفته شده که به صورت پیش‌فرض ۲ سنتی‌مورگان است. با در نظر گرفتن یک بازه هم‌پوشانی، می‌توان از اینکه هپلوتایپ‌هایی که در مرز پنجره‌های مارکر هستند فیز نشوند، تا حدی اطمینان یافت.

Beagle 5.2 از یک الگوریتم تکرارشونده (Iterative) «پیشرو» استفاده می‌کند تا به مرور فیز هر ژنوتایپ هتروزایگوس با در نظرگرفتن هتروزایگوت قبلی (در پنجره مارکر قبلی) مشخص شود. در الگوریتم استفاده‌شده توسط بیگل، هر ژنوتایپ هتروزایگوس، در یکی از حالت‌های «در حال پیشرفت»، یا «اتمام‌یافته» است. در پایان هر چرخه از الگوریتم فیزینگ، آن هتروزایگوسی که در حالت «در حال پیشرفت» است و بیش از بقیه به درستی فیزشدن آن اطمینان داریم، به حالت «اتمام‌یافته» تغییر می‌کند.

3 نتایج

برای ارزیابی، نسخه 5.2 ابزار Beagle با نسخه 4.2.1 ابزار SHAPEIT مقایسه شده‌است. تمام آزمون‌های ارزیابی بر روی کامپیوتری با پردازنده ۲۰ هسته‌ای ۲.۴ گیگاهرتز، مدل Intel Xeon E5-2640 و حافظه ۲۵۶ گیگابایتی اجرا شده‌اند.

نرخ خطای پروسه فیزینگ، با «نرخ خطای سوییچ (SER)» سنجیده شده‌است. SER، سنجه‌ای رایج برای اندازه‌گیری خطا در پروسه فیزینگ هپلوتایپ است. خطای سوییچینگ، زمانی اتفاق می‌افتد که الگوریتم فیزینگ، اختصاص هپلوتایپ‌ها برای SNPهای مجاور هم را برعکس انجام می‌دهد. SER، حاصل تقسیم تمام خطاهای سوییچ، بر تمام SNPهای فیزشده است.

تصویر ۳.۱ نرخ خطای سوییچ، و زمان محاسباتی را برای اجرای پروسه فیزینگ با هریک از ابزارها نمایش می‌دهد. این پروسه فیزینگ، با استفاده از داده‌های آرایه‌های SNP دیتاست UK Biobank انجام یافته‌است. مطابق تصویر، بر روی این دیتاست، هر دو ابزار از نظر نرخ خطا و سرعت محاسباتی، عملکرد مشابهی را از خود به نمایش گذاشته‌اند.

تصویر ۳.۲، نرخ خطای سوییچ و زمان محاسباتی حاصل از اجرای الگوریتم فیزینگ را، این بار با داده‌های توالی نمونه از دیتاست TOPMed نمایش می‌دهد. این توالی‌ها، توالی‌های کروموزوم ۲۰ می‌باشند. همانند داده‌های Biobank، بر روی این دیتاست نیز دو روش نرخ خطای مشابهی را دارند. اما از نظر زمان محاسباتی، سرعت عملکرد ابزار Beagle، ۲۳ تا ۲۶/۷ برابر SHAPEIT بود.

تغییر اندازه پنجره مارکر در ابزار Beagle، این امکان را فراهم می‌کند که حافظه مورد نیاز برای اجرای پروسه فیزینگ را کنترل نمود. کاهش حافظه مورد نیاز با کوچک‌نمودن اندازه این پنجره، امکان فیز کردن دیتاست‌های بزرگ‌تر را به ما می‌دهد. از سوی دیگر، این کار پیچیدگی محاسباتی را تحت تاثیر قرار داده و زمان مورد نیاز برای اتمام پروسه فیزینگ را افزایش می‌دهد. اندازه پیش‌فرض پنجره مارکر در فیزنمودن هپلوتایپ‌ها، ۴۰

سـنتی‌مورگان[6] می‌باشـد. تصـویر ۳.۳، نتایج فیزینگ را برای ابعاد مختلف ۵، ۱۰، ۲۰ و ۴۰ cM نمایش می‌دهد. از نظر دقت فیزینگ، اندازه پنجره تاثیری در این میزان نداشـت و دقت، مسـتقل از این پارامتر بود. اما دو سـنجه دیگر، میزان حافظه مصـرفی و زمان محاسـباتی، با تغییر اندازه پنجره، دچار تغییر محسوسی می‌شوند. همان‌گونه که از تصویر ۳.۳ پیداست، با کاهش اندازه پنجره، حافظه مصرفی کاهش یافته، اما زمان محاسباتی افزایش می‌یابد.

پارامتر دیگری که تاثیر آن در دقت فیزینگ در این مقاله تحلیل شد، «اندازه جمعیت موثر» است. جمعیت موثر، در این همبافت، اندازه‌ای از جمعیت است که از لحاظ گونه‌گونی ژنتیکی، مشابه کل جمعیت واقعی است. این اندازه در محاسبه احتمالات گذر[7] مدل HMM استفاده می‌شود. تاثیر میزان اولیه این پارامتر که به عنوان ورودی به مدل داده می‌شـود، از در بازه‌ای که میانه آن از نظر مرتبه بزرگی، ۳ واحد از ابتدا و انتها فاصـله دارد، ارزیابی شد. مطابق شکل ۳.۴، در مدل Beagle، این مقدار اولیه تاثیری در دقت عملکرد مدل ندارد. اما در مدل SHAPEIT، نرخ خطا بسته به فاصله مقدار اولیه از اندازه ایده‌آل، نوسانات قابل توجهی دارد. مستقل بودن نرخ خطای سوییچ از میزان اولیه «اندازه جمعیت موثر» در مدل Beagle، یک مزیت دیگر این مدل به شمار می‌رود.

4 جمع‌بندی

فیزینگ هپلوتایپ، پروسه‌ایست که در طی آن هپلوتایپ‌های به ارث رسیده از هر والد در هر لوکوس[8]، برای یک موجود دیپلوید مشخص می‌شوند. در این مقاله، روشی کارآمد از نظر زمان محاسباتی و حافظه مصرفی، برای فیزینگ هپلوتایپ داده‌های بزرگ‌مقیاس SNP و داده‌های توالی ژنوم معرفی شده‌است. پیاده‌سازی پارامتر قابل کنترل «اندازه پنجره مارکر»، پنل رفرنس مرکب، و الگوریتم دو مرحله‌ای فیزینگ، ایده‌های ارائه‌شده در این مقاله برای کاهش حافظه مصـرفی و افزایش سـرعت محاسـباتی بودند. در الگوریتم دومرحله‌ای، نخسـت هپلوتایپ‌های مارکرهای ژنتیکی پرفرکانس فیز می‌شـوند و با اسـتفاده از آن‌ها، هپلوتایپ مارکرهای کم‌فرکانس اسـتنتاج می‌گردند. این روش، برای دیتاسـت‌های بزرگ که دارای تعداد زیادی مارکر کم‌فرکانس هستند، نظیر داده‌های توالی کل ژنوم، مناسب و کارآمد است.

الگوریتم فیزینگ معرفی‌شده در این مقاله، در نسخه 5.2 ابزار Beagle پیاده‌سازی و استفاده شده‌است. در راستای ارزیابی عملکرد این ابزار، نسخه SHAPEIT 4.2.1 برای مقایسـه عملکردها انتخاب شـده‌است. ارزیابی عملکرد، بر روی دو دیتاسـت مختلف انجام شـد. در فیزینگ با دیتاسـت نمونه UK Biobank که شامل داده آرایه‌های SNP بود، دو ابزار عملکرد مشابهی از نظر نرخ خطا، زمان محاسباتی و حافظه مصرفی داشتند. اما در ارزیابی با داده‌های توالی دیتاسـت TOPMed، ابزار Beagle حدود ۲۰ برابر سـریع‌تر از SHAPEIT بود، در حافظه مصـرفی بهینه‌تر بوده و در اندازه‌های بزرگ‌تر داده ورودی، مقیاس‌پذیری بهتری داشت.

5 کارهای آتی و چالش‌های پیش رو

با حجم بزرگی که داده‌های توالی مورد اسـتفاده در فیزیگ هپلوتایپ دارند، حافظه مصـرفی همچنان دغدغه مهمی به شـمار می‌رود. بهره‌وری بهتر در اسـتفاده از حافظه، می‌تواند امکان تحلیل دیتاسـت‌های بزرگ‌تر را بر روی رایانه‌های با حافظه کم‌تر فراهم کند. نسخه 5.2 ابزار Beagle، با زبان Java پیاده‌سازی شده‌است که کنترل تخصیص حافظه در لایه پایین‌تر را توسط ماشین مجازی جاوا (JVM) پیاده‌سازی کرده‌است. پیاده‌سازی این ابزار با زبان‌های سـطح پایین‌تر که کنترل و آزادی عمل بیشـتری در تخصـیص حافظه می‌دهند، می‌تواند تاثیر قابل توجهی در اسـتفاده این ابزار از حافظه داشته‌باشد.

---

[6] Centimorgan یا به اختصار cM، واحدی برای اندازه‌گیری نرخ بازترکیبی ژنتیکی است. هر یک واحد cM نمایانگر یک درصد شانس برای رویداد بازترکیبی میان دو مارکر ژنتیکی است. این سنجه، توسط ابزار Beagle، برای تخمین اندازه فیزیکی دو مارکر ژنتیکی استفاده شده‌است.

[7] Transition

[8] Locus

Beagle 5.2، تولنایی فیزینگ ده‌ها هزار توالی افراد را دارد. برای پاســـخ‌دهی به دیتاســـت‌های با حجم رو به رشـــد، فیزینگ هپلوتایپ نیازمند پیشروی‌های بیشتری در روش‌ها دارد. امکان فیزینگ دیتاست‌هایی که شامل توالی صدها هزار یا میلیون‌ها نفر هستند، از چالش‌های آینده روش‌های فیزینگ به شمار می‌رود.

6 منابع و مراجع

[1] Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 2021 Oct 7;108(10):1880-1890. doi: 10.1016/j.ajhg.2021.08.005. Epub 2021 Sep 2. PMID: 34478634; PMCID: .PMC8551421