یک الگوریتم ژنتیک برای استنباط هاپلوتیپ‌ها

سید علی کتان‌فروش، مهدی صادقی، چنگیز اصلاحچی و حمید پزشک

چکیده. در این مقاله، مسئله‌ی استنباط هاپلوتیپ‌های جمعیت بر پایه‌ی نمونه‌های ژنوتیپی بدست آمده ازSNP ها مورد توجه قرار می‌گیرد. مسئله‌ی استنباط هاپلوتیپ‌ها از داده‌های ژنوتیپ، تحت مدل بیشترین پارسیمونی از مسائل دشوار در زیست‌شناسی محاسباتی است. هرچند رهیافت‌های متفاوتی برای حل این مسئله معرفی شده‌اند تا کنون، توسعه‌ی روشی مبتنی بر الگوریتم‌های ژنتیک برای حل این مسئله مورد توجه قرار نگرفته است. در این مقاله، تلفیقی از یک الگوریتم ژنتیک و رده‌ای از روال‌های سودجویانه، برای حل مسئله‌ی استنباط هاپلوتیپ‌ها با فرض بیشترین پارسیمونی بکار گرفته می‌شوند.

۱. مقدمه

توالی نوکلئوتیدها در ژنوم انسان، تقریبا در بین همه افراد بشر یکسان است. تخمین زده می‌شود بین ژنوم هر دو فرد انسان بیش از ۹۹/۹ درصد شباهت وجود داشته باشد. همین اختلاف ناچیز، زمینه‌ی اصلی تفاوت در خصیصه‌های فردی است. در سطح ژنومی، رایج‌ترین شکل تفاوت در بین افراد مختلف یک جمعیت، چندریختی تک نوکلئوتیدی (Single Nucleotide Polymorphism) 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. تقریبا همه اسنیپ‌ها دو آللی هستند بنابراین به سادگی می‌توان وضعیت هر اسنیپ در یک فرد را با **0** و 1 نشان داد. در جانداران دیپلوئید از جمله انسان، سلولهای بدنی همواره حاوی دو نسخه‌ی همسان از DNA هستند که هر نسخه از یکی از والدین به ارث می‌رسند. در روش‌های آزمایشگاهی متعارف، مطالعه‌ی ژنوم افراد با استفاده از خوانش توالی‌های نوکلئوتیدی استخراج شده از سلول‌های دیپلوئید صورت می‌گیرد. بدین ترتیب اطلاعات بدست آمده، حاصل جمع توالی‌های والدی است. این ترکیب را اصطلاحا ژنوتیپ می‌نامیم. ژنوتیپ یک اسنیپ در هر فرد، می‌تواند یکی از وضعیت‌های هموزیگوت، یعنی **0/0** یا **1/1**و یا هتروزیگوت، یعنی **0/1** را داشته باشد. اگر ژنوم را به طور مستقل بر روی هر یک از توالی‌های والدی مطالعه کنیم، توالی بدون ابهامی از وضعیت اسنیپ‌ها بدست می‌آوریم که آنرا اصطلاحا اسنیپ-هاپلوتیپ (SNP-haplotype) و به اختصار هاپلوتیپ می‌نامند. هاپلوتیپ‌ها داده‌های مناسبی برای شناسائی زمینه‌های ژنتیکی خصیصه‌های فردی به ویژه استعداد ابتلا به بیماری‌ها، مطالعه‌ی ژنتیکی ساختارهای جمعیتی و تشخیص هویت هستند.

برای هر ژنوتیپ، هاپلوتیپ‌هایی که آللی یکسان با آلل اسنیپ‌های هموزیگوت آن ژنوتیپ داشته باشند اصطلاحا سازگار با آن ژنوتیپ نامیده می‌شوند. وقتی یک ژنوتیپ مثل g از ترکیب دو هاپلوتیپ مثل h_1و h_2بدست آمده باشد می‌نویسیم $g = h_1 \oplus h_2$. 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 شود می‌گوییم ژنوتیپ، تعیین فاز شده است یا اصطلاحا تفکیک شده است.

با اینکه برخی شیوه‌های آزمایشگاهی مثل clone-based systematic haplotyping[1] برای تعیین مستقیم هاپلوتیپ‌ها توسعه یافته‌اند اما بسیار پرهزینه‌اند. در عوض با استفاده از شیوه‌های محاسباتی می‌توان هاپلوتیپ‌های نمونه‌های مورد بررسی را با داشتن ژنوتیپ‌های تعیین فاز نشده، بدست آورد.

مسئله‌ی استنباط هاپلوتیپ‌ها و تعیین فاز ژنوتیپ. فرض کنید نمونه‌ای مثل $G = \{g_1, \ldots, g_n\}$ شامل nژنوتیپ بر روی l اسنیپ داده شده است. می‌خواهیم مجموعه‌ای از هاپلوتیپ‌ها مثل Hرا تعیین کنیم به قسمی که برای هر $g_i \in G$، دو هاپلوتیپ $h_a, h_b \in H$وجود داشته باشند که $g_i = h_a \oplus h_b$. علاوه بر این، قیود دیگری برای انتخاب یک جواب یا دسته‌ای از جوابهای مطلوب از بین مجموعه جوابهای ممکن به مسئله اضافه می‌شوند. این قیود معمولا دائر بر ارضای برخی ملاحظات در ژنتیک و تکامل و مرتبط با ساختار و تنوع هاپلوتیپ‌های جواب هستند. عمده‌ی مدل‌هایی که تاکنون برای حل این مسئله ارائه شده‌اند را می‌توان در یکی از چهار دسته‌ی بیشترین پارسیمونی(maximum parsimony) ، فیلوژنی کامل(perfect phylogeny) ، بیشترین درستنمائی و مدل‌های

واژه های کلیدی: الگوریتم ژنتیک، بیشترین پارسیمونی، ژنوتیپ، هاپلوتیپ، Single Nucleotide Polymorphism.

بیزی قرار داد. در این مقاله، روشی مبتنی بر الگوریتم ژنتیک برای حل مسئله‌ی استنباط هاپلوتیپ‌ها با قید بیشترین پارسیمونی معرفی می‌گردد.

دیدگاه بیشترین پارسیمونی بر این باور استوار است که اصولا تنوع هاپلوتیپ‌ها در طبیعت بسیار کمتر از تنوع ژنوتیپ‌ها است. بدین ترتیب، در مسئله‌ی بیشترین پارسیمونی به دنبال مجموعه‌ای از هاپلوتیپ‌ها می‌گردیم که با کمترین تعداد هاپلوتیپ‌های متمایز بتوان ژنوتیپ‌های داده شده را بازسازی کرد. فرضیه بیشترین پارسیمونی اولین مدلی بود که برای صورت‌بندی مسئله‌ی استنباط هاپلوتیپ‌ها مورد استفاده قرار گرفت و به عنوان اولین رهیافت حل، الگوریتمی سودجویانه توسط کلارک پیشنهاد شد**[2]**. 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 به این رهیافت وارد می‌شود ولی در بسیاری از نمونه‌های واقعی، نتایج قابل قبولی از این روش بدست می‌آید**[3]**.

به طور کلی، بکارگیری روش کلارک با محدودیتهایی مواجه است. اول اینکه دست کم یک ژنوتیپ بدون ابهام باید در اختیار باشد. مشکل دوم تفکیک نشده باقی ماندن برخی ژنوتیپ‌ها در پایان الگوریتم است و سوم اینکه، ترتیب انتخاب ژنوتیپ‌ها برای تفکیک و نیز انتخاب یک هاپلوتیپ از بین تمام هاپلوتیپ‌های فعلی سازگار با این ژنوتیپ، می‌تواند به جواب‌های متفاوتی منجر شود. 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**[4]**. جدا از این شکل خاص، مسئله‌ی بیشترین پارسیمونی در حالت کلی نیز یک مسئله‌ی NP-hard است**[5]**. روشهای متعددی با رهیافت‌های اکتشافی برای حل تقریبی این مسئله ارائه شده‌اند. 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 مسئله‌ی برنامه‌ریزی خطی، الگوریتمی کارآمد برای داده‌هایی با اندازه‌ی متوسط ارائه کرده است**[6]**. وانگ با پیاده‌سازی یک روش توسعه و تحدید به مقایسه‌ی نتایج دقیق با نتایج بدست آمده از برخی ایده‌های بهبود دهنده‌ی الگوریتم کلارک پرداخته است**[7]**.

الگوریتم‌های ژنتیکی (**GA**) 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**[8]**. 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**[9]**. کاربرد الگوریتم ژنتیکی برای حل مسئله‌ی برآورد فراوانی هاپلوتیپ‌های جمعیت نیز توسط آزوما و همکارانش در **[10]** مورد بررسی قرار گرفته است.

با وجود اینکه الگوریتم ژنتیک، ابزاری انعطاف‌پذیر برای حل مسائل بهینه‌سازی به شمار می‌آید به نظر می‌رسد در حل مسئله‌ی استنباط هاپلوتیپ‌ها و در مقایسه با دیگر روشها بخوبی مورد توجه قرار نگرفته است. در ادامه، پس از معرفی یک الگوریتم ژنتیک، کارآیی آن برای حل مسئله‌ی استنتاج هاپلوتیپ‌ها با هدف بیشترین پارسیمونی مورد بررسی قرار می‌گیرد.

۲. روش

الگوریتم‌های ژنتیکی که در رده‌ی روشهای اکتشافی قرار می‌گیرند حالت خاصی از روشهای موسوم به روشهای تکاملی هستند. در روشهای تکاملی 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 (fitness function) و متناسب با تابع هدف مورد بررسی در مسئله تعیین 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. در بیشتر پیاده‌سازی‌ها به جای استفاده‌ی مستقیم از تابع هدف ، تابع دیگری که تغییراتی متناسب با تغییرات تابع هدف داشته باشد به عنوان تابع سازگاری در نظر

گرفته می‌شود؛ مثلا رتبه‌ی "کروموزم" بر حسب مقدار تابع هدف به عنوان کمیت سازگاری استفاده می‌شود. برای جزئیات بیشتر در مورد روش‌های اندازه‌گیری سازگاری بر حسب تابع هدف و انتخاب "کروموزم‌ها" برای ایجاد نسل بعد به[**11**] رجوع کنید.

در مدل بیشترین پارسیمونی، تابع هدف تعداد هاپلوتیپ‌های متمایز در جواب مسئله‌ی تفکیک ژنوتیپ‌ها است. کار را ابتدا با معرفی یک مدل ساده (naive) آغاز می‌کنیم. در ساده‌ترین شکل نمایش جواب برای مسئله‌ی تفکیک ژنوتیپ‌ها، به ازای هر بردار $g_i = \langle g_{i,1},\ldots,g_{i,l}\rangle$ در مجموعه‌ی G، دو هاپلوتیپ $h_{2i-1} = \langle h_{2i-1,1},\ldots,h_{2i-1,l}\rangle$ و $h_{2i} = \langle h_{2i,1},\ldots,h_{2i,l}\rangle$ در مجموعه‌ی جواب H داریم به قسمی که

$$g_{ij} = h_{2i-1,j} + h_{2i,j}, \quad \text{for} \quad i = 1,\ldots,n \quad \text{and} \quad j = 1,\ldots,l. \qquad (۱)$$

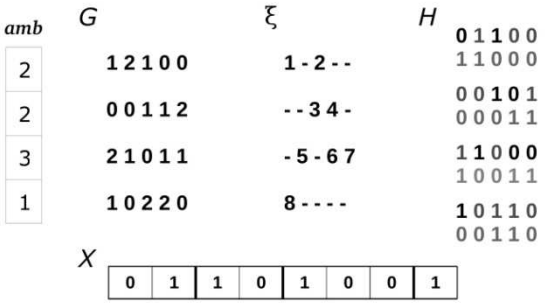
می‌توان مجموعه‌ی ژنوتیپ‌های داده شده، G را به صورت یک ماتریس $n \times l$ با درآیه‌های 0، 1 و 2 در نظر گرفت که هر سطر آن نماینده‌ی یک ژنوتیپ است و مجموعه‌ی جواب، H را به صورت یک ماتریس $n \times l$ با درآیه‌های 0 و 1 در نظر گرفت که هر سطر آن نماینده‌ی یک هاپلوتیپ است. با این نمادگذاری، جمع برداری جفت سطرهای متوالی H مثل $h_{2i-1} + h_{2i}$، به سادگی تفکیک ژنوتیپ g_i به وسیله‌ی هاپلوتیپ‌ها را نشان می‌دهد.

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_{2i} یا h_{2i-1} را داشته باشیم هاپلوتیپ دیگر با حل معادله‌ی (**1**) برحسب هاپلوتیپ معلوم و ژنوتیپ مورد مطالعه بدست می‌آید. علاوه بر آن، مقدار مؤلفه‌های $h_{2i-1,j}$ و $h_{2i,j}$ وقتی $g_{ij} = 0$ یا $g_{ij} = 2$ باشد بنابر تعریف معین هستند. به همین خاطر تنها اطلاعات نابدیهی در جواب، مربوط به مؤلفه‌هایی است که در آنها $g_{ij} = 1$ است. به چنین موقعیت‌هایی، یعنی جایگاه‌های هتروزیگوت در یک ژنوتیپ، موقعیت‌های مبهم (ambiguous) برای مسئله‌ی تعیین فاز می‌گوئیم. 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؛ که در آن $M = \sum amb_i$ و amb_i تعداد موقعیت‌های مبهم در ژنوتیپ g_i است. در واقع این رشته، حاصل اتصال n رشته‌ی کوچکتر، هر یک متناظر با مؤلفه‌های مبهم یک ژنوتیپ است. بر پایه‌ی این نحوه‌ی نمایش جواب، برای بدست آوردن هاپلوتیپ‌های تشکیل‌دهنده‌ی یک ژنوتیپ داریم:

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

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

در زیر، نمونه‌ای از این شیوه‌ی نمایش جواب تصویر شده است.



تولید تصادفی یک مجموعه‌ی اولیه از جواب‌های شدنی. 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 0 یا 1است.

محاسبه‌ی تابع هدف. برای تعیین تعداد هاپلوتیپ‌های متمایز می‌توان هر هاپلوتیپ را با دیگر هاپلوتیپ‌های مجموعه‌ی جواب مقایسه کرد. برای این کار، ابتدا سطرهای ماتریس جواب با استفاده از مرتب‌سازی مبنائی(radix sort) مرتب می‌شوند و با مقایسه‌ی سطرهای متوالی در ماتریس مرتب‌شده، تعداد هاپلوتیپ‌های متمایز بدست می‌آید. پیچیدگی این روش $O(nl)$ است.

عملگر "کراس‌اوِر". فرض کنید دو ماتریس جواب، H_1 و H_2 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

هاپلوتیپ‌های تفکیک‌کننده‌ی یک ژنوتیپ، به طور تصادفی از یکی از دو ماتریس H_1 یا H_2 انتخاب می‌شود. 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 یک جواب جدید بدست می‌آید. این شیوه‌ی کراس‌اوِر تضمین می‌کند که "کروموزم" بدست آمده هنوز می‌تواند ژنوتیپ‌های داده شده را تفکیک کند. در این شیوه، از یک پارامتر که آنرا با cr_{int} نشان می‌دهیم برای کنترل میزان دورگه بودن "کروموزم‌های" بدست آمده استفاده می‌کنیم. 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.

عملگر "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 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.

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 ادامه، 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 معرفی خواهیم کرد. ابتدا یک صورت پارامتری برای این الگوریتم معرفی می‌کنیم. سپس تعابیر 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 $\sigma=\langle\sigma_1,\sigma_2,\ldots,\sigma_n\rangle$ 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.

گام۱) 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 H جستجو کن و آنرا با h_a نشان بده. اگر چنین هاپلوتیپی وجود نداشت قرار بده $h_a=\hbar_{\sigma_i}$.

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

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

به ازای پارامترهای داده‌شده، این الگوریتم یک الگوریتم قطعی برای بدست آوردن یک جواب شدنی برای مسئله‌ی تفکیک ژنوتیپ‌ها است. با این حال، مزیت منطق سودجویانه‌ی این الگوریتم در آن است که جوابی نزدیک به جواب بهینه‌ی مسئله‌ی بیشترین پارسیمونی بدست می‌آورد. هر نمونه از این الگوریتم به ازای جایگشت داده‌شده‌ی σ و مجموعه‌ی هاپلوتیپ‌های راهنمای داده‌شده‌ی H_g را با $GreedyPhasing(\sigma,H_g)$ 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. با پشت سر هم قرار دادن نمایش دودوئی اعداد $\langle\sigma_1,\sigma_2,\ldots,\sigma_n\rangle$، یک رشته‌ی بیتی خواهیم داشت که از آن برای نمایش جایگشت σ در یک "کروموزم" استفاده می‌کنیم. اکنون می‌توان، هر نمونه از الگوریتم مثل $GreedyPhasing(\sigma,H_g)$ را با کنار هم قرار دادن نمایش بیتی σ و H_g در یک "کروموزم" واحد نمایش داد. برای تولید جمعیت اولیه‌ی "کروموزم‌ها" همانند روش قبل می‌توان بیت‌های مربوط به اطلاعات هاپلوتیپ‌های راهنما را به طور تصادفی با مقادیر 0 و 1 پر کرد. اما برای بیت‌های بخش مربوط به جایگشت‌ها، ابتدا یک جایگشت تصادفی از اعداد ۱تا 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. 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 $\langle\sigma_1^1,\sigma_2^1,\ldots,\sigma_n^1\rangle$ با حفظ ترتیب و به طور تصادفی در بین عناصر دنباله‌ی $\langle\sigma_1^2,\sigma_2^2,\ldots,\sigma_n^2\rangle$ 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 عنصرِ منتقل شده به جایگشت نوترکیب، منشاء هاپلوتیپ راهنما را تعیین می‌کند. به تناوب یکی از این دو نوع "کراس‌اوِر" در روال "کراس‌اوِر" الگوریتم ژنتیک به کار گرفته می‌شوند.

"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. 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. این الگوریتم ژنتیکی را GAhap می‌نامیم.

۳. نتایج و بحث

برای تعیین مقادیر مناسب برای پارامترهای الگوریتم ژنتیک و نیز ارزیابی کارایی، نتایج حاصل از اجراهای مکرر این الگوریتم‌ها به ازای مقادیر مختلف پارامتر بر روی داده‌های ژنوتیپی شبیه‌سازی شده مورد بررسی قرار گرفتند. برای هر انتخاب از مقادیر برای پارامترهای مورد مطالعه در الگوریتم ژنتیک، الگوریتم پیشنهادی بر روی ۲۰ نمونه‌ی شبیه‌سازی شده‌ی مستقل، هر یک شامل ۴۰ ژنوتیپ و ۱۲ اسنیپ اجرا شد. ارزیابی‌های مقدماتی بیانگر پیشرفت چشمگیر کارآیی الگوریتم GAhap در مقایسه با الگوریتم ساده است به قسمی که اجرای آن در ۵۲٪ از کل ۸۶۴۰ نمونه‌ی مورد بررسی، با "موفقیت" همراه بود. (مقایسه کنید با ۷٪ برای الگوریتم ساده).
جدول **1** نتایج موفق‌ترین انتخاب برای پارامترهای مختلف این الگوریتم ژنتیک را نشان می‌دهد. هر سطر جدول، بهترین تنظیمات برای الگوریتم ژنتیک را وقتی مقدار یکی از پارامترها ثابت نگه داشته شده است نشان می‌دهد. همانطور که در جدول **1** 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. ما این تنظیمات را به عنوان تنظیمات پیش‌فرض برای اجرای روش GAhap در ادامه‌ی ارزیابی‌ها در نظر گرفتیم.

جدول ۱. موفق‌ترین مقادیر پارامتر برای الگوریتم ژنتیک **GAhap**

گزینه	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
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
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

پس از تعیین کارآمدترین انتخاب برای پارامترها، الگوریتم GAhap را بر روی نمونه‌ای واقعی از ژنوتیپ‌ها اجرا کردیم. 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[**12**]. آنها فراوانی ۳۶ هاپلوتیپ متمایز موجود در این مجموعه را به کمک روش‌های آزمایشگاهی تعیین کردند. در این ناحیه‌ی ژنی، ۱۵ اسنیپ از کیفیت لازم برای مطالعه برخوردار بودند که اطلاعات مربوط به ۱۵۰ ژنوتیپ بدون داده‌ی مفقود را از بین داده‌های این مجموعه مورد استفاده قرار دادیم.

جدول**2** 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. 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 در عمل، نتایج بسیار نزدیکی به جواب بیشترین پارسیمونی بدست می‌آورند. به جز روش2SNP، 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 امیدوارکننده 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 الگوریتمGAhap به دلیل ویژگی تصادفی محاسبات در آن 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 **2**. ارزیابی روش‌های متفاوت برای استنباط هاپلوتیپ‌های مجموعه داده‌های هورَن*

روش	رویکرد اصلی مورد استفاده برای حل مسئله	[1]	[2]	[3]
HAPLOTYPER	استنباط بیزی برپایه‌ی پیشین دیریکله[13]	۳۳	۵.۴	۳.۰
PHASE	استنباط بیزی برپایه‌ی مدل فیلوژنی کامل[14]	۳۲	۵.۶	3.1
fastPHASE	ساده‌سازی روش PHASE برای افزایش سرعت[15]	۳۵	۷.۳	4.5
2SNP	تعیین فاز جفت اسنیپ‌ها و درخت فراگیر کمینه[16]	۴۰	۱۰.۴	5.6
GAhap	بیشترین پارسیمونی و الگوریتم ژنتیک	۳۴	۹.۷	5.7

[1] تعداد هاپلوتیپ‌های استنباط شده‌ی متمایز [2] خطای شناسائی هاپلوتیپ‌ها [3] خطای جابجائی فازها * مقادیر خطا بر حسب درصد هستند.

مراجع

[1] 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.

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

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

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

[5] 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.

[6] 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.

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

[8] 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.

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

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

[11] Goldberg D.E . *Genetic Algorithms in Search*, Optimzation and Machine Learning. Addison Wesley, 1989.

[12] Horan M, Millar D.S , Hedderich J, Lewis G, Newsway V, Mo N, et al. *Human Mutation* , 21 (4): 408--423, 2003.

[13] 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.

[14] 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.

[15] 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.

[16] 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.