

بهینه سازی و تلفیق الگوریتم‌های k-means و PSO جهت بازسازی هاپلوتیپ‌ها

با استفاده از اطلاعات نشانگرهای SNP

Optimizing and integrating K-means and PSO algorithms for haplotype reconstruction using SNP maker information

مصطفی قادری زفره‌ای^{۱*}، صمد شریفی^۲، ماشاء... عباسی دزفولی^۳، احسان عسگریان^۴، محمدحسین بناء بازی^۴

۱- استادیار زیست سامانه‌های محاسباتی، دانشگاه یاسوج، ایران

۲- به ترتیب دانش‌آموخته کارشناسی ارشد، استادیار، علوم رایانه، دانشگاه آزاد اسلامی اهواز

۳- دانشجوی دکتری رایانه، دانشگاه فردوسی مشهد، ایران

۴- استادیار، ژنتیک و اصلاح نژاد دام، بخش بیوتکنولوژی، موسسه تحقیقات علوم دامی کشور، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران

Ghaderi Zefrhei M^{*1}, Sharifi S², Abbassi Dezfouli M², Asgarian E³, Banabazi MH⁴

1- Assistant Professor in Computational Systems Biology, University of Yasouj, Yasouj, Iran

2- Graduate MSc Student, Assistant Professor, Computer Science, Islamic Azad University of Ahwaz, Ahwaz, Iran

3- PhD candidate in Computer Science, Ferdowsi University oh Mashhad, Mashhad, Iran

4- Research Assistant, Department of Biotechnology, Animal Science Research Institute of IRAN (ASRI), Agricultural Research, Education & Extension Organization (AREEO), Karaj, Iran

* نویسنده مسئول مکاتبات، پست الکترونیکی: mghaderi@yu.ac.ir

(تاریخ دریافت: ۹۵/۳/۲۴ - تاریخ پذیرش: ۹۵/۶/۲۸)

چکیده

بازسازی هاپلوتیپ یکی از موضوعات مهم در مطالعات ژنتیکی و بیوانفورماتیکی است. تشکیل هاپلوتیپ‌ها به صورت مستقیم با استفاده از روش‌های زیستی و آزمایشگاهی بسیار دشوار و پرهزینه است. از این رو، بازسازی هاپلوتیپ‌ها معمولاً با استفاده از روش‌های محاسباتی بر روی اطلاعات ژنوتیپی و نشانگرهای ژنتیکی انجام می‌شود. در این مقاله، دو الگوریتم بر اساس مدل حداقل تصحیح خطا برای بازسازی هاپلوتیپ‌ها از روی ژنوتیپ افراد برای چندشکلی‌های تک نوکلئوتیدی (SNP) ارائه می‌شود. الگوریتم اول حالتی از الگوریتم K-Means است با این تفاوت که این الگوریتم بجای استفاده از مراکز اولیه تصادفی، این مراکز را با شرایط ویژه انتخاب می‌کند. الگوریتم دوم نیز ترکیبی از الگوریتم PSO بهبودیافته و K-Means است که IPSOKM نام‌گذاری شد. الگوریتم‌های بهینه‌سازی شده بر روی داده‌های شبیه‌سازی شده و واقعی به کار گرفته شدند. نتایج نشان داد که دقت بازسازی هاپلوتیپ‌ها با استفاده از الگوریتم‌های ارائه شده در مقایسه با برخی از الگوریتم‌های مرسوم استفاده شده جهت بازسازی هاپلوتیپ‌ها، به خصوص در حالت‌های وجود خطا و حفره، به طور قابل ملاحظه‌ای افزایش یافت. از این رو، این الگوریتم‌ها می‌توانند به طور مؤثری در مطالعات ژنتیک انسانی و به نژادی گیاه و دام به کار گرفته شوند.

واژه‌های کلیدی

الگوریتم

بازسازی هاپلوتیپ

قطعات SNP

IPSOKM

K-Means

مقدمه

با پیشرفت علم زیست‌شناسی و توانایی بشر در تعیین توالی DNA، زمینه برای به‌دست آوردن منشا بسیاری از بیماری‌های ژنتیکی بوجود آمده است. از طرفی دیگر کاربرد داده‌های متکی بر DNA در بهبود و اصلاح نژاد گیاهان و حیوانات، به یک واقعیت روز تبدیل شده‌است. یکی از مهم‌ترین داده‌های متکی بر DNA که هم در پزشکی انسانی و هم در اصلاح نژاد حیوانات و گیاهان جایگاه ویژه‌ای برای خود باز کرده، نشانگرهای مبتنی بر چند شکلی‌های تک نوکلئوتیدی^۱ یا SNP می‌باشد. در صورتی که فراوانی الی مبنای کدگذاری ژنوتیپ‌ها باشد، ال‌های این نشانگرها در موجودات دیپلوئید می‌تواند به یکی از سه حالت زیر باشد: (۱) هر دو ال از نوع بیشینه باشند که با صفر در ژنوتیپ مشخص می‌شود؛ (۲) هر دو ال از نوع کمینه باشند که با ۱ در ژنوتیپ مشخص می‌شود و (۳) یک ال بیشینه و دیگری کمینه باشد که این حالت با ۲ در ژنوتیپ مشخص می‌شود (Chakravarti 1998). جدول ۱ ژنوتیپ و هاپلوتیپ و محل SNP‌ها را بر روی یک جفت کروموزوم نشان می‌دهد.

جدول ۱- نمایش SNP‌ها، هاپلوتیپ‌ها و ژنوتیپ‌ها بر روی یک جفت کروموزوم (SNP‌ها بر روی کروموزوم با حروف بزرگ نشان داده شده‌اند).

ACGTAGAGAA...CAGACAT...GACTTCAC	کروموزوم ۱
ACGTACAGAA...CAGACAT...GACATCAC	کروموزوم ۲
G A T	هاپلوتیپ ۱
C A A	هاپلوتیپ ۲
G/C A/A T/A	ژنوتیپ

اگر فرض شود h_1 و h_2 دو هاپلوتیپ با طول n باشند، مجموع آن‌ها $g = h_1 \oplus h_2$ به‌عنوان ژنوتیپ در نظر گرفته می‌شود. رابطه هاپلوتیپ و ژنوتیپ به‌صورت الگوریتمی به صورت زیر نشان داده می‌شود:

$$g[i] = \begin{cases} h_1[i] & \text{if } h_1[i] = h_2[i] \\ 2 & \text{otherwise} \end{cases} \quad (i = 1, \dots, n)$$

¹ Single Nucleotide Polymorphisms

چندین هاپلوتیپ کشف شده‌اند که بر باروری نژادهای مختلف گاو شیری از جمله هلشتاین، براون سویس و جرزی موثرند (Van Raden et al. 2011). علی‌رغم فراوانی معنی‌دار این هاپلوتیپ‌ها در جمعیت (۴/۵ تا ۲۵ درصد برای افراد ناقل)، این هاپلوتیپ‌ها به‌واسطه عدم وجود حالت هموزیگوت در جمعیت شناسایی شده‌اند. این امر نشان می‌دهد که هاپلوتیپ‌های هموزیگوت کشته شده‌اند. از این‌رو آزمون‌های هاپلوتیپی می‌تواند به اصلاح‌گران کمک نماید تا از آمیزش‌هایی که چنین نقایصی را انتقال می‌دهند، پرهیز نموده و فراوانی آن‌ها را در آینده کاهش دهند. یکی از دیگر کاربردهای مهم تعیین هاپلوتیپ در زمینه تجویز داروی مناسب برای افراد بیمار است. هر شخص بیمار با توجه به هاپلوتیپ خود ممکن است یک واکنش خاص نسبت به یک نوع دارو نشان دهد. مثلاً کسانی که به بیماری آسم مبتلا هستند هر کدام با توجه به هاپلوتیپ‌های خود نسبت به یک اسپری ویژه واکنش متفاوتی از خود نشان می‌دهند. بنابراین یک نوع داروی خاص برای تمام بیماران مفید نیست. به‌عنوان مثال، شکل ۱ اطلاعات مربوط به جایگاه‌های SNP، ژنوتیپ و هاپلوتیپ مربوط به شش فرد که سه نفر از آن‌ها مبتلا به سرطان ریه هستند را نشان می‌دهد. در این شکل ال‌های بیشینه با رنگ سیاه و ال‌های کمینه با رنگ خاکستری نشان داده شده‌اند. فرض کنید که افراد ۴، ۵ و ۶ مبتلا به سرطان ریه هستند. حال با تحلیل کدام اطلاعات می‌توان سرطان را به‌راحتی تشخیص داد؟ با توجه به اطلاعات ارائه شده، واضح است که هیچ‌کدام از تک جایگاه-های SNP، به‌طور مستقل توانایی تفکیک افراد بیمار و سالم را ندارند. برای مثال SNP شماره ۱ در فرد اول و ششم مشابه هم است در حالی‌که فرد اول سالم و فرد ششم بیمار می‌باشند. با استفاده از اطلاعات ژنوتیپی نیز امکان پیش‌بینی وجود سرطان فراهم نیست. برای مثال ژنوتیپ فرد سوم و ششم کاملاً مشابه هم می‌باشند؛ در حالی که یکی سالم و دیگری بیمار است. با استفاده از اطلاعات هاپلوتیپی مشخص می‌شود که همه افراد بیمار دارای هاپلوتیپی به‌صورت CTTCTA هستند پس می‌توان استدلال کرد در هر فردی که ترکیب ال‌های این SNP‌های خاص در قالب هاپلوتیپی به‌صورت CTTCTA بروز نماید، استعداد ابتلا به بیماری سرطان ریه وجود دارد (Zhang and Rajapakse 2009).

یک کروموزوم، جهت تشکیل یک هاپلوتیپ کامل است (Zhao et al. 2007). علاوه بر این، به منظور کاهش مسائل محاسباتی در بازسازی هاپلوتیپ‌ها، همواره سعی بر این است که SNP‌های حاوی اطلاعات بیشتر (برای تشخیص بیماری‌ها) انتخاب شوند. پیچیدگی مسائل محاسباتی در تحلیل هاپلوتیپ‌ها عمدتاً از نوع NP-Hard و APX-Hard است (Clark 1990; Alimoti and Kann 1997; Patil et al. 2001; Bafna et al. 2003; Bafna et al. 2005).

اولین گام در روش‌های بازسازی هاپلوتیپ از روی قطعات SNP تعیین توالی کروموزوم‌های یک فرد و تولید قطعات SNP آن فرد است. قطعات SNP به قطعه‌های کروموزومی گفته می‌شود که فقط شامل SNP‌های موجود بر روی آن کروموزوم باشد. اشتباه در تعیین ژنوتیپ یک جایگاه را خطا^۵ می‌گویند. برای مثال، ممکن است مقدار یک SNP به جای صفر، یک ثبت شود و برعکس. اگر ژنوتیپ یک فرد در یک تک SNP خوانده نشود، در نمایش قطعه به جای آن - گذاشته می‌شود و به آن ژنوتیپ نامشخص حفره می‌گویند.

به دلیل اهمیت فراوان تحلیل هاپلوتیپ‌ها در مطالعه بیماری‌های ژنتیکی، روش‌های تجربی زیادی برای استخراج اطلاعات هاپلوتیپی ابداع شده‌است. ولی علاوه بر هزینه بالا، این قبیل آزمایش‌ها بسیار زمان‌بر بوده و بازدهی پایینی دارند که جوابگوی تحلیل هاپلوتیپ در سطح و مقیاس بالا نیستند. روش‌های محاسباتی می‌توانند تا حد زیادی هزینه و زمان شناسایی هاپلوتیپ‌ها را کاهش دهند (Asgarian et al. 2007). تا کنون، دو روش محاسباتی برای استخراج اطلاعات هاپلوتیپی ارائه شده است (Gusfield 2001; Gusfield 2003; Bonizzoni et al. 2006; Wang et al. 2003; Zhang et al. 2003). در روش اول هاپلوتیپ‌ها از روی مجموعه‌ای از ژنوتیپ‌های یک جمعیت به دست می‌آیند. این روش را استنتاج هاپلوتیپ^۱ می‌نامند. در روش دوم هاپلوتیپ‌های هر فرد از روی قطعات حامل SNP در آن فرد به دست می‌آید. این روش به بازسازی هاپلوتیپ از روی قطعات SNP^۲ یا سرهم نمودن هاپلوتیپ^۳ موسوم است. در این روش هدف اصلی، تصحیح و ادغام قطعه‌های SNP روی هم افتاده^۴ از

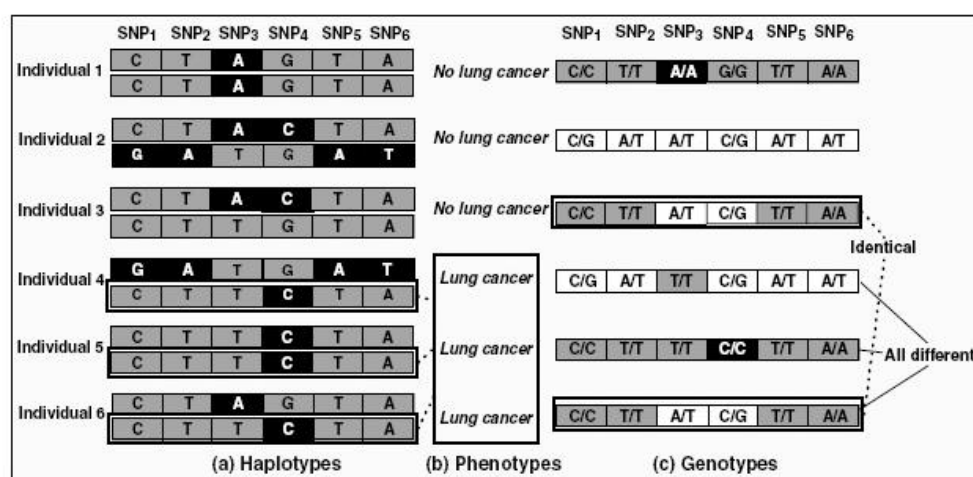
¹ Haplotype Inferring

² Haplotype Reconstruction

³ Haplotype Assembly

⁴ Overlap

⁵ Noise



شکل ۱- نمایش شماتیک وضعیت تک جایگاه‌های SNP، ژنوتیپ و هاپلوتیپ در شش فرد در رابطه با زمینه ژنتیکی ابتلا به بیماری سرطان ریه (Zhang Y and Rajapakse J. 2009)

f_1	A	T	C	G	A	T	G	C	A	T	G
f_2		T	C	G	A	T	G	G	A	T	G
f_3			A	T	G	C	A	T	G	C	A
f_4				A	T	G	A	A	T	G	C
f_5							C	A	T	G	C

شکل ۲- قطعات SNP ترازبندی شده

همراه حفره های موجود در آن ها، با طول هاپلوتیپ آن ها برابر است (s_i). حال با فرض این که $x, y \in \{0, 1, -\}$ دو SNP باشند، یک معیار برای شباهت و فاصله دو قطعه SNP می تواند به صورت ذیل باشد: $d(x, y) = \begin{cases} 1 & (x \neq y \neq -) \\ 0 & \text{otherwise} \end{cases}$ پس فاصله دو قطعه SNP و یک هاپلوتیپ نیز به همین صورت محاسبه می شود.

شباهت و یا فاصله برای دو قطعه SNP به نام های S_i و S_k از رابطه $D(S_i, S_k) = \sum_{j=1}^n d(S_{ij}, S_{kj})$ به دست می آید. فاصله بین یک قطعه SNP و یک هاپلوتیپ نیز به همین صورت محاسبه می شود. دو قطعه SNP به نام های S_i و S_k با هم ناسازگارند اگر

$D(S_i, S_k) > 0$ باشد. در غیر این صورت گفته می شود که دو قطعه سازگار هستند. بنابراین یک هاپلوتیپ بازسازی شده از روی قطعات SNP را می توان مجموعه از قطعه های SNP ناسازگار (که از روی یک جفت کروموزوم به دست آمده اند) دانست. در این مطالعه، هدف، پیدا نمودن و تصحیح خطای درون داده ها در طی فرآیند بازسازی هاپلوتیپ از روی قطعات SNP است تا بتوان یک زوج هاپلوتیپ با بیشینه سازگاری را از روی قطعه های SNP تصحیح شده به دست آورد. در مسأله بازسازی هاپلوتیپ، هدف، پیدا نمودن بهترین جفت هاپلوتیپی است که کمترین میزان تضاد را با قطعه های SNP داده شده داشته باشد.

مدل های مبتنی بر حذف قطعه خطا شامل موارد ذیل می باشند: (الف) مدل حذف حداقل قطعه ^۲ (MFR): این مدل فرض می کند که قطعه های "بد" باعث ایجاد خطا می شوند (این ها قطعاتی از سایر ارگانیزم ها هستند که به اشتباه وارد مسأله شده اند). بنابراین، MFR سعی می نماید این قطعه های بد را حذف کند. در صورتی-

وجود حفره ها در داخل یک قطعه ایجاد مشکل می کند. بر روی هر قطعه SNP (که از الفبای -، صفر، یک تشکیل شده است) مفهومی به نام شکاف^۱ تعریف می شود. شکاف، یک یا چند حفره متوالی در داخل قطعه هاست بطوری که دو طرف آن مقادیر غیرحفره ای ("صفر"، یا "یک") قرار داشته باشد. برای مثال قطعه -01110- بدون شکاف است، -0---0- فقط یک شکاف دارد و -1-101---01- دارای دو شکاف است. معمولاً مدل های مختلف برای بازسازی هاپلوتیپ ها از روی قطعات SNP، تنها در مواردی که قطعات SNP بدون شکاف و یا با یک شکاف باشند، به کار می روند. پس از تولید قطعات SNP آن ها را ترازبندی می کنند. فرض کنید قطعه های تولید شده و ترازبندی شده به صورت شکل ۲ باشد. در شکل ۲ عناصری که در کادر مشخص شده اند، SNP هستند. آن گاه بعد از جداسازی SNP ها، ماتریس SNP به صورت زیر نمایش داده می شود:

$$S = \begin{bmatrix} 1 & - & - \\ 0 & 1 & - \\ 1 & 0 & 1 \\ - & 1 & 0 \\ - & - & 0 \end{bmatrix}$$

اگر فرض شود m قطعه SNP وجود دارد که از یک جفت کروموزوم گرفته شده اند و با دو هاپلوتیپ به طول n متناظر

هستند، آنگاه، $S = [s_{ij}]_{m \times n}$ ماتریسی متشکل از قطعه های SNP

است که هر عضو s_{ij} مقداری برابر "صفر"، (اگر ال زام SNP در قطعه i بیشینه باشد)، "۱" (اگر کمینه باشد) یا "-" (اگر حفره باشد) دارد. هر سطر ماتریس (s_i)، متناظر با یک قطعه SNP به-

دست آمده از روی یک کروموزوم است و هر ستون آن (s_j)، متناظر با یک جایگاه SNP است. طول همه قطعه های SNP به

^۱ Gap

^۲ Minimum Fragment Removal

برخورد استفاده می‌کردند. ولی این مدل‌ها با استفاده از الگوریتم‌ها و نظریه گراف به صورت مستقیم از گراف برخورد، برای حل مسأله بازسازی هاپلوتیپ استفاده می‌کنند. (Reed et al. 2004) و Huffner (2005) الگوریتمی برای دوبخشی کردن گراف ارائه نمودند. الگوریتم آن‌ها کوچک‌ترین مجموعه از رئوس را که گراف برخورد آن‌ها دوبخشی باشد، پیدا نمود. علاوه بر استفاده از قطعه‌های SNP، می‌توان از اطلاعات ژنوتیپ نیز استفاده کرد (Reed et al. 2004; Huffner 2005). اطلاعات ژنوتیپ بسیار آسان‌تر و کم هزینه‌تر از هاپلوتیپ به دست می‌آیند. بر این اساس، چند روش ترکیبی از جمله روش حداقل برخورد هاپلوتیپ⁵ (MCIH) (Zhang 2006) و روش حداقل خطای تصحیح با استفاده از اطلاعات ژنوتیپ⁶ (MEC-GI) (Asgarian et al. 2007) ابداع شده‌اند.

مواد و روش‌ها

الگوریتم K-Means بهبود یافته⁷ (IKM)

هدف K-Means قرار دادن مجموعه‌ای از داده‌ها $X = \{x_1, x_2, \dots, x_n\}$ در تعدادی خوشه است بطوری‌که داده‌های موجود در یک خوشه تا حد ممکن مشابه و داده‌های موجود در خوشه‌های متفاوت تا حد ممکن متمایز باشند. در خوشه‌بندی قطعه‌های SNP، ابتدای دو قطعه که بیش‌ترین فاصله را نسبت به هم دارند به عنوان مرکز خوشه‌ها در نظر گرفته می‌شوند. در مرحله بعد، براساس فاصله داده‌ها از مراکز تعیین شده، هر داده در یکی از خوشه‌ها قرار می‌گیرد. سپس مراکز داده‌های هر خوشه به کمک الگوریتم رأی‌گیری⁸ محاسبه شده و دوباره با توجه به فاصله هر داده تا مرکز جدید، خوشه‌بندی تکرار می‌شود. این عمل تا زمانی که مرکز خوشه‌ها در دو بار تکرار الگوریتم تغییر نکند، ادامه خواهد یافت. (Zhang et al. 2007) جزئیات بیش‌تر این روش را شرح داده‌اند. در مطالعه حاضر، از الگوریتم‌های خوشه‌بندی متفاوتی استفاده شد. اولین الگوریتم مورد استفاده برای این منظور را الگوریتم K-Means بهبود یافته (IKM) نام

که ماتریس SNP بدون شکاف باشد مدل MFR دارای حل چندجمله‌ای است (Lancia et al. 2001). اما این مدل برای ماتریس SNP که قطعه‌های آن دارای شکاف هستند NP-hard است (Bafna et al. 2005).

(ب) مدل بازسازی بزرگترین هاپلوتیپ¹ (LHR): منظور از LHR حذف مجموعه‌ای از قطعه‌هاست به صورتی‌که تشکیل ماتریس SNP امکان‌پذیر باشد و مجموع طول هاپلوتیپ‌های حاصل از ماتریس SNP بیشینه باشد. این مسأله در صورتی‌که قطعه‌های SNP بدون شکاف باشند، دارای یک الگوریتم چندجمله‌ای است. ولی در مورد قطعه‌های یک شکاف APX-hard است (Cilibiasi 2005).

(ج) مدل حذف حداقل SNP² (MSR): در این مدل فرض می‌شود که همه قطعه‌ها از یک ارگانیزم هستند اما در توالی‌یابی داده‌ها خطا وجود دارد. در واقع در ماتریس قطعات، یک یا چند SNP خطا وجود دارد. این فرض کمک می‌کند تا جایگاه‌های SNP مشکل‌ساز را حذف نموده و یک ماتریس از قطعات SNP سازگار بسازیم. MSR برای یک ماتریس SNP که هر قطعه SNP حداقل دو شکاف داشته باشد، مسأله‌ای NP-hard است. (Lancia et al. 2001) و Bafna et al. (2005) اثبات نمودند که مدل MSR مسأله‌ای APX-Hard است. پس برای این مدل الگوریتم تقریبی خوبی وجود ندارد.

(د) مدل حداقل تصحیح خطا³ (MEC³ / MLF⁴): این مدل برای حالتی مناسب است که در خواندن ال هر SNP، احتمال رخداد اشتباه وجود دارد. پس باید درایه‌های ماتریسی که اشتباه هستند، تصحیح شوند تا ماتریس SNP بدون خطا شود. تفاوت اصلی این مدل با MSR بر سر تصحیح مقادیر SNP به جای حذف جایگاه‌های ناسازگار است. سعی بر این است که درایه‌های ماتریس SNP را طوری تغییر دهیم که ماتریس SNP بدون خطا شده و مجموع وزن تغییر یافته‌ها کمینه باشد. مدل MLF حتی با ماتریس SNP بدون شکاف مسأله‌ای NP-hard است (Zhao and Zhang 2005). مدل‌های قبلی به طور غیر مستقیم از گراف

⁵ Minimum Conflict Individual Haplotyping

⁶ Minimum Error Correction with Genotype Inference

⁷ Improved K-Means

⁸ Voting algorithm

¹ Longest Haplotype Reconstruction

² Minimum SNP Removal

³ Minimum Error Correction

⁴ Minimum Letter Flips

گذاری شد. در روش IKM، K-Means غیر تصادفی است و از نتایج حاصل از ماتریس فاصله قطعات از همدیگر، برای انتخاب هوشمندانه نمونه های اولیه در الگوریتم K-Means استفاده شد. انتخاب نمونه های اولیه در الگوریتم K-Means در نتایج نهایی بسیار مؤثر است. بنابراین بهترین انتخاب برای نمونه های اولیه انتخاب نمونه هایی است که به خوشه های متفاوت تعلق داشته و تقریباً در مرکز خوشه ها باشند. با توجه به این دو ویژگی، می توان بهترین مراکز اولیه را برای الگوریتم پیدا نمود. به عبارت دیگر، برای پیدا کردن مراکز اولیه از دو اصل زیر استفاده شد: (۱) مراکز اولیه می بایست متعلق به خوشه های متفاوت باشند. بنابراین در ماتریس فاصله قطعات، این مراکز می بایست بیش ترین فاصله را از هم داشته باشند. (۲) بعد از آن که بر اساس اصل اول دو قطعه به عنوان مرکز مشخص شد، بر اساس نزدیکی قطعات به هر کدام از مراکز، خوشه ها را تشکیل می دهیم. آنگاه از روی یک حداقل آستانه برای سهم هر خوشه از کل قطعات که بر اساس تجربه تعیین می شود، خوشه هایی که این آستانه را تامین نکنند کنار گذاشته می شوند. اگر سهم یک خوشه کمتر از این آستانه باشد، این دو قطعه برای مرکز بودن مناسب نیستند زیرا یکی از خوشه ها تراکم زیاد شده و دیگری تراکم خیلی کمتری دارد. سپس مجدداً از روی ماتریس فاصله، دو قطعه ای را که فاصله زیادی از هم دارند را به عنوان مرکز انتخاب می کنیم (بیشینه دوم) و دوباره اصل دوم را تکرار می کنیم تا به هدف مورد نظر، که تعیین مراکز اولیه برای الگوریتم K-Means است، دست یابیم.

الگوریتم خوشه بندی ترکیبی بهبود یافته IPSOKM دومین الگوریتم پیشنهادی سعی می کند با ترکیب الگوریتم های KM و PSO^۱ با هم، از مزایای هر دو آن ها بهره مند شود. از این رو این الگوریتم خوشه بندی ترکیبی بهبود یافته، IPSOKM نام گذاری شد که ترکیبی از دو الگوریتم K-Means و PSO است. در واقع الگوریتم K-Means دیگر بطور تصادفی مرکز اولیه را انتخاب نمی کند بلکه به نوعی هوشمندانه عمل می کند و از بهترین نتیجه در الگوریتم IPSO به عنوان مرکز اولیه استفاده می شود. جدول ۲ این الگوریتم را نشان می دهد. در جدول ۲، w وزن اینرسی حرکتی در جهت بردار سرعت فعلی ذره و همچنین c_1 و c_2 ضرایب ثابت شتاب ذره هستند. برای درک بهتر این فراسنجه ها خواننده به الگوریتم روش بهینه سازی اجتماع ذرات (PSO) ارجاع داده می شود.

برای آزمون الگوریتم های پیشنهادی برای بازسازی هاپلوتیپ از روی قطعات SNP و بررسی کارایی و راندمان آن ها، از داده های زیستی واقعی و شبیه سازی شده موجود در چندین پایگاه داده شد. در مطالعه حاضر، داده های واقعی از پایگاه های ACE^۲ و Daly و داده های شبیه سازی شده از پایگاه های SIM0 و SIM50 به دست آمدند.

¹ Particle Swarm [Optimization (PSO)

² Angiotensin Converting Enzyme

جدول ۲- شبه کد برای الگوریتم پیشنهادی ترکیبی IPSOKM

Input: SNP fragments	
Output: two haplotypes	
Step1: Set the population size N, w, c_1 and c_2 .	
Step2: Initialize a population of size N.	
Step2: Select k pair fragment with maximum distance and maximum nigh to add the N.	
Step3: Set iterative count Gen1=0.	
Step4: Set iterative count Gen2 =0.	
Step5: (PSO Method)	
Step5.1: Apply the PSO operator to update the N+k Particles.	Based upon PSO algorithm
Step5.2: Gen1=Gen1+1.if Gen1<8, go to Step 5.1.	
Step6: (KM Method) For P_k	
Step6.1: Recalculate P_k using the KM algorithm.	Based upon K-Means algorithm
Step6.2: Gen2=Gen2+1. If Gen2<4, go to step 6.1.	
Step6: based on voting algorithm, it is produced two haplotypes from current two clusters.	

جدول ۳- نرخ بازسازی الگوریتم‌های پیشنهادی با و بدون استفاده از اطلاعات ژنوتیپ بر روی پایگاه داده ACE و مقایسه با الگوریتم‌های AGC، PAM، K-Means و PSO

نرخ خطا	نرخ خطا	(MEC/GI) IPSOKM	(MEC/GI) IKM	(MEC) IKM	PSO	AGC	PAM	K-Means
0.25	0.1	1	0.999	1	0.994	0.99	0.948	0.998
0.2	0.2	0.997	0.9	0.94	0.924	0.924	0.941	0.952
0.3	0.3	0.935	0.861	0.823	0.834	0.834	0.84	0.814
0.4	0.4	0.86	0.816	0.683	0.662	0.662	0.669	0.65
0.5	0.1	0.997	0.98	0.972	0.965	0.965	0.965	0.977
0.2	0.2	0.985	0.915	0.9	0.883	0.883	0.883	0.898
0.3	0.3	0.896	0.836	0.76	0.752	0.752	0.754	0.756
0.4	0.4	0.865	0.803	0.663	0.655	0.655	0.63	0.638
0.75	0.1	0.983	0.928	0.925	0.883	0.883	0.893	0.885
0.2	0.2	0.882	0.793	0.75	0.774	0.774	0.758	0.733
0.3	0.3	0.862	0.76	0.726	0.696	0.696	0.677	0.669
0.4	0.4	0.842	0.717	0.645	0.635	0.635	0.605	0.628

در همه پایگاه‌های داده موجود ۱۲ ترکیب متفاوت از نرخ حفرة و خطا وجود دارد (نرخ‌های خطای ۰/۱، ۰/۲، ۰/۳، ۰/۴ و نرخ‌های حفرة ۰/۲۵، ۰/۵ و ۰/۷۵). پیاده‌سازی برنامه تحلیل و بازسازی هاپلوتیپ‌ها (با امکان استفاده از اطلاعات ژنوتیپ)، به‌وسیله زبان برنامه‌نویسی Matlab2011a انجام شد. در نهایت، نتایج حاصل از الگوریتم‌های پیشنهادی با الگوریتم‌های K-Means (Zhang et al. 2007)، PAM^۱ (Moeinzadeh et al. 2007)، خوشه‌بندی سلسه مراتبی یا جمع شونده^۲ (AGC) (Asgarian et al. 2007) و PSO مقایسه شد.

نتایج و بحث

در الگوریتم ترکیبی بهبود یافته IPSOKM پیشنهادی تعداد افراد جمعیت ۵۰ و تعداد افرادی که به‌صورت هوشمندانه به جمعیت اضافه شد پنج در نظر گرفته شد. بقیه پارامترها به‌صورت $w=0.8$ و $C_1=C_2=0.7$ فرض شد. برای مقایسه با روش‌های مختلف دیگر، پارامترهای آن‌ها در بهترین حالت پیاده‌سازی شد. هم‌چنین برای دقت بیشتر هر الگوریتم چند بار اجرا و برای هر نرخ خطا هشت بار نرخ حفرة تکرار شد. در نهایت، از همه نتایج میانگین گرفته شد که در جداول ۳ تا ۶ ارایه شده‌اند. این جداول،

^۱ Partitioning Around Medoids

^۲ Agglomerative

نتایج حاصل از آزمون الگوریتم‌های جدید پیشنهادی را بر روی داده‌های حاصل از پایگاه‌های داده ACE، SIM0، SIM50 و Daly با دو مدل حداقل تصحیح خطا و مدل حداقل تصحیح خطا با استفاده از اطلاعات ژنوتیپ نشان داده و با نتایج حاصل از چهار الگوریتم دیگر مقایسه می‌نمایند.

در مجموع، براساس جداول مذکور معلوم می‌شود که الگوریتم‌های پیشنهادی از عملکرد بسیار بهتری نسبت به الگوریتم‌های مشابه برخوردار بوده‌اند. به عبارتی دقت بازسازی هاپلوتیپ با استفاده از الگوریتم‌های ارایه شده، بهتر از سایر الگوریتم‌های مورد مقایسه است.

در مطالعه حاضر، از پایگاه‌های داده‌ای شبیه‌سازی و واقعی استفاده شد. در پایگاه‌های داده‌ای شبیه‌سازی، داده‌ها براساس میزان شباهت دو هاپلوتیپ پدری و مادری شبیه‌سازی شده‌اند. به عبارت دیگر، این پایگاه‌های داده براساس درصد ژنوتیپ‌های هتروزیگوت تشکیل شده‌اند. به درصد شباهت بین یک جفت هاپلوتیپ (درصد ژنوتیپ‌های هتروزیگوت) نرخ شبیه‌سازی می‌گویند. برای مثال هیچ شباهتی بین دو هاپلوتیپ به‌دست آمده از قطعات SNP در پایگاه داده SIM0 وجود ندارد (در واقع، نرخ شبیه‌سازی در این پایگاه داده صفر است). بنابراین تمام جایگاه‌ها، ژنوتیپ هتروزیگوت دارند (ارزش عددی آن‌ها برابر "۲" است. این در حالی است که نیمی از مکان‌های دو هاپلوتیپ به‌دست آمده

داده وجود دارد. در پایگاه داده واقعی ACE نیز برای هر نرخ خطا و حفره، ۸ نمونه (در کل ۹۶ نمونه برای تمام ترکیبات مختلف خطا و حفره) وجود دارد (Rieder 1999). در هر نمونه، ۲۰ قطعه SNP با طول ۵۲ جایگاه SNP و ژنوتیپ مربوط به جفت هاپلوتیپ نهایی وجود داشت.

از قطعات SNP در پایگاه داده SIM50 به صورت تصادفی مشابه هم هستند. در مرحله بعد، ابتدا ۱۰ جفت هاپلوتیپ به طول ۵۲ جایگاه SNP و با توجه به فراسنجه نرخ شبیه‌سازی، به صورت تصادفی تولید می‌شوند. سپس از هر جفت هاپلوتیپ، ۱۲ نمونه کپی برداشته می‌شود و با توجه به نرخ‌های خطا و حفره مختلف، ۲۰ قطعه SNP در هر جفت هاپلوتیپ تولید می‌شوند. پس در کل ۱۲۰ نمونه با نرخ‌های خطا و حفره مختلف در این نوع پایگاه

جدول ۴- نرخ بازسازی الگوریتم‌های پیشنهادی با و بدون استفاده از اطلاعات ژنوتیپ بر روی پایگاه داده SIM0 و مقایسه با الگوریتم‌های K-Means، PAM، AGC و PSO

نرخ خطا	نرخ حفره	(MEC/GI) IPSOKM	(MEC) IPSOKM	(MEC/GI) IKM	(MEC) IKM	PSO	AGC	PAM	K-Means
0.1	0.25	0.999	0.993	0.999	0.99	0.981	0.996	0.996	0.996
0.2		0.972	0.939	0.952	0.938	0.937	0.965	0.965	0.965
0.3		0.951	0.853	0.9	0.821	0.819	0.846	0.846	0.849
0.4		0.862	0.719	0.805	0.652	0.576	0.601	0.607	0.618
0.1	0.5	0.995	0.986	0.99	0.981	0.955	0.989	0.989	0.9884
0.2		0.986	0.926	0.95	0.912	0.861	0.916	0.911	0.919
0.3		0.864	0.793	0.81	0.775	0.754	0.752	0.701	0.7524
0.4		0.793	0.709	0.75	0.638	0.569	0.58	0.587	0.5556
0.1	0.75	0.941	0.921	0.935	0.901	0.851	0.885	0.913	0.9022
0.2		0.856	0.764	0.8	0.741	0.73	0.686	0.799	0.6955
0.3		0.763	0.693	0.71	0.661	0.617	0.643	0.646	0.6266
0.4		0.704	0.599	0.654	0.571	0.562	0.528	0.542	0.5445

جدول ۵- نرخ بازسازی الگوریتم‌های پیشنهادی با و بدون استفاده از اطلاعات ژنوتیپ بر روی پایگاه داده SIM50 و مقایسه با الگوریتم‌های K-Means، PAM، AGC و PSO

نرخ خطا	نرخ حفره	(MEC/GI) IPSOKM	(MEC) IPSOKM	(MEC/GI) IKM	(MEC) IKM	PSO	AGC	PAM	K-Means
0.1	0.25	1	1	1	0.999	0.991	0.999	0.999	0.999
0.2		0.996	0.957	0.992	0.938	0.94	0.96	0.938	0.952
0.3		0.948	0.85	0.94	0.786	0.819	0.824	0.786	0.817
0.4		0.828	0.739	0.806	0.654	0.663	0.622	0.654	0.651
0.1	0.5	1	0.983	0.998	0.978	0.964	0.977	0.975	0.9778
0.2		0.976	0.905	0.963	0.882	0.904	0.884	0.884	0.91
0.3		0.914	0.821	0.864	0.751	0.782	0.75	0.75	0.7673
0.4		0.816	0.699	0.814	0.638	0.621	0.642	0.642	0.6395
0.1	0.75	989	0.923	952	0.905	0.873	0.88	0.88	0.9066
0.2		0.883	0.784	0.872	0.744	0.755	0.737	0.737	0.7433
0.3		0.843	0.763	0.833	0.692	0.669	0.675	0.675	0.6776
0.4		0.785	0.687	0.777	0.62	0.595	0.606	0.606	0.6101

جدول ۶- نرخ بازسازی الگوریتم‌های پیشنهادی با و بدون استفاده از اطلاعات ژنوتیپ بر روی پایگاه داده Daly و مقایسه با الگوریتم‌های AGC، PAM، K-Means و PSO

K-Means	PAM	AGC	PSO	(MEC) IKM	(MEC/GI) IKM	(MEC) IPSOKM	(MEC/GI) IPSOKM	نرخ	هزینه
0.999	0.999	0.978	0.997	0.999	1	1	1	0.1	0.25
0.99	0.989	0.933	0.974	0.989	0.999	0.991	0.999	0.2	
0.928	0.902	0.836	0.897	0.901	0.984	0.942	0.993	0.3	
0.721	0.72	0.713	0.726	0.721	0.888	0.787	0.901	0.4	
0.997	0.996	0.982	0.985	0.996	0.999	0.998	0.999	0.1	0.5
0.971	0.961	0.939	0.945	0.96	0.993	0.975	0.998	0.2	
0.853	0.832	0.818	0.837	0.832	0.94	0.863	0.97	0.3	
0.693	0.7	0.685	0.695	0.7	0.882	0.751	0.884	0.4	
0.973	0.963	0.956	0.941	0.965	0.98	0.981	0.997	0.1	0.75
0.888	0.867	0.863	0.869	0.87	0.936	0.896	0.976	0.2	
0.762	0.766	0.752	0.764	0.769	0.891	0.803	0.913	0.3	
0.657	0.663	0.651	0.66	0.665	0.873	0.727	0.878	0.4	

نتیجه‌گیری کلی

بازسازی هاپلوتیپ یکی از موضوعات مهم در مطالعات ژنتیکی است. هم‌چنین، این موضوع یکی از مسائل مهم محاسباتی در حوزه بیوانفورماتیک است و اخیراً فعالیت‌های گسترده‌ای در این زمینه انجام شده‌است. استخراج هاپلوتیپ‌ها و تعیین ژنوتیپ یک جایگاه به‌صورت مستقیم از طریق روش‌های زیستی بسیار دشوار و پرهزینه است. بنابراین، استخراج اطلاعات هاپلوتیپ‌ها عمدتاً با استفاده از اطلاعات ژنوتیپ و قطعات SNP و با کمک روش‌های محاسباتی انجام می‌شود. در مطالعه حاضر، دو الگوریتم بهینه شده برای بازسازی هاپلوتیپ‌ها بر اساس مدل حداقل تصحیح خطا ارائه شد. الگوریتم اول (IKM) حالتی از الگوریتم K-Means است با این تفاوت که در این الگوریتم به‌جای مراکز اولیه تصادفی، مراکز اولیه با شرایط خاصی انتخاب می‌شوند. الگوریتم دوم (IPSOKM) نیز یک الگوریتم ترکیبی بهبود یافته است که از ادغام دو الگوریتم PSO و K-Means بدست آمد. از آنجاکه به‌دست آوردن اطلاعات ژنوتیپ بسیار آسان و کم هزینه است، بنابراین از اطلاعات ژنوتیپ نیز جهت افزایش دقت بازسازی استفاده شد. نتایج نشان داد که الگوریتم‌های جدید پیشنهاد شده (به‌ویژه الگوریتم ترکیبی IPSOKM) دقت بازسازی را به‌خصوص در شرایط وجود خطا و حفره به‌طور چشمگیری افزایش می‌دهند.

Rieder (1999) توانست توالی ژن *DCP1* (که بر روی کروموزوم ۲۲ قرار دارد) را در ۱۱ فرد استخراج کند. علاوه بر آن ژنوتیپ‌هایی که بیش از یک جایگاه هتروزیگوت داشتند را کنار گذاشت. سپس از بقیه آن‌ها، ۸ جفت هاپلوتیپ استخراج نمود و از هر کدام از آن‌ها ۱۲ نمونه تهیه کرد. در نهایت، از هر نمونه ۲۰ SNP با فراسنجه‌های مختلف نرخ خطا (۰/۱، ۰/۲، ۰/۳ و ۰/۴) و نرخ حفره (۰/۲۵، ۰/۵ و ۰/۷۵) ایجاد نمودند. در پایگاه داده واقعی Daly، برای هر نرخ خطا و حفره، ۱۲۸ نمونه (در مجموع، ۱۵۳۶ نمونه برای تمام حالت‌های خطا و حفره) موجود بود (Daly and Rioux 2001). در هر نمونه، ۴۰ قطعه SNP با طول ۸۶ یا ۹۸ جایگاه SNP و ژنوتیپ مربوط به جفت هاپلوتیپ نهایی، وجود داشت. (Daly and Rioux (2001) تحلیل بسیار دقیقی بر روی ساختار هاپلوتیپ کروموزوم 5q31 (که در موقعیت 500kb از این کروموزوم قرار دارد) در جمعیتی از مردم اروپا ارائه دادند. سپس هاپلوتیپ‌هایی که چند الل نامشخص داشتند یا ژنوتیپ تعیین کننده آن‌ها بیش از یک مکان هتروزیگوت داشت را کنار گذاشتند و بقیه هاپلوتیپ‌ها را برای پایگاه داده انتخاب نمودند و از هر کدام از آن‌ها ۱۲ نمونه تهیه کردند. در نهایت، از هر نمونه ۴۰ SNP با پارامترهای مختلف نرخ خطا (۰/۱، ۰/۲، ۰/۳ و ۰/۴) و نرخ حفره (۰/۲۵، ۰/۵ و ۰/۷۵) ایجاد نمودند.

سپاسگزاری

مولفین بر خود لازم می‌دانند از همکاری علمی و فنی شرکت رایان زیست فناوری پارس آوید (www.avidbiotech.com) در اجرای این تحقیق تشکر و قدردانی نمایند.

منابع

- Alimoti P, Kann V (1997) Hardness of approximation problems on cubic graph. In: Proceedings of third Italian Conference on Algorithms and Complexity (CIAC'97), Italy, Rome, 288-298.
- Asgarian E, Moeinzadeh MH, Najafi A, Sharifian S, Habibi J, Mohammadzadeh J (2007) Solving MEC and MEC/GI Problem Models Using Information Fusion and Multiple Classifiers. In: Proceedings of 4th IEEE International Conference on Innovations in Information Technology (Innovations'07). United Arab Emirates, Dubai, 397-401.
- Bafna V, Halldorsson B, Schwartz R (2003) Haplotypes and informative SNP selection algorithms: don't block out information. In: Proceedings of The Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB). USA, New York, 19-27.
- Bafna V, Istrail S, Lancia G, Rizzi R (2005) Polynomial and APX-hard cases of the individual haplotyping problem. Theoretical Computer Science 335:109-125.
- Bonizzoni P, Vedova GD, Dondi R, Li J (2003) The haplotyping problem: an overview of computational models and solutions. Journal of Computer Science and Technology 18:675-688.
- Chakravarti A (1998) It's raining, hallelujah? Nature Genetics 19:216-217.
- Cilibiasi R, Iersel LV, Kelk S, Tromp J (2005) On the complexity of several haplotyping problems. In: Proceedings of 5th International Workshop on Algorithms in Bioinformatics (WABI). Spain, Mallorca, 128-139.
- Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Molecular Biology and Evolution 7:111-122.
- Daly MJ, Rioux JD, Schaffner SF (2001) High-resolution Haplotype structure in the Human genome. Nature Genetics 29:229-232.
- Gusfield D (2001) Inference of haplotypes from samples of diploid populations: Complexity and algorithms. Journal of Computational Biology 8:305-323.
- Gusfield D (2003) Haplotyping by pure parsimony. In: Proceedings of the 14th Symposium on Combinatorial Pattern Matching (CPM). Mexico, Morelia, 144-155.
- Huffner F (2005) Algorithm engineering for optimal graph bipartization. In: Proceedings of the 4th International Workshop of Efficient and Experimental Algorithms (WEA). Greece, Santorini Island, 240-252.
- Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks 4:1942-1948.
- Lancia G, Bafna V, Istrail S (2001) SNP problems complexity and algorithms. In: auf der Heide FM (Ed.), Algorithms-ESA. Springer, Berlin, 182-193.
- Moeinzadeh MH, Asgarian E, Najafi-Ardabili A, Sharifian-R S, Sheikhaei MS, Mohammadzadeh J (2007) Three Heuristic Clustering Methods for Haplotype Reconstruction Problem with Genotype Information. In: Proceedings of 4th IEEE International Conference on Innovations in Information Technology (Innovations'07). United Arab Emirates, Dubai, 402-406.
- Patil N, Berno AJ, Hinds DA (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719-1723.
- Qian W., Yang Y, Yang N, Li C (2008) Particle swarm optimization for SNP haplotype reconstruction problem. Applied Mathematics and Computation 196:266-272.
- Reed B, Smith K, Vetta A (2004) Finding odd cycle transversals. Operations Research Letters 32:299-301.
- Rieder MJ, Taylor SL, Clark AG, Nickerson DA (1999) Sequence variation in the human angiotensin converting enzyme. Nature Genetics 22:59-62.
- VanRaden PM, Olson KM, Null DJ, Hutchison JL (2011) Harmful recessive effects on fertility detected by absence of homozygous haplotypes. Journal of Dairy Science 94:6153-6161.
- Wang K, Huang L, Zhou CG, Pang W (2003) Particle swarm optimization for traveling salesman problem. In: Proceedings of the Second International Conference on Machine Learning and Cybernetics. China, Xi'an, 1583-1585.
- Wang Y, Feng E, Wang R (2007) A clustering algorithm based on two distance functions for MEC model. Computational Biology and Chemistry 31:148-150.
- Zhang XS, Wang RS, Wu LY, Zhang W (2006) Minimum Conflict Individual Haplotyping from SNP Fragments and Related Genotype. Evolutionary Bioinformatics Online 2:261-270.
- Zhang Y, Rajapakse J (2009) Machine Learning in bioinformatics. John Wiley and Sons, New Jersey, USA, 462.
- Zhao Y, Xu Y, Zhang Q, Chen G (2007) An overview of the haplotype problems and algorithms. Frontiers of Computer Science in China 1:272-282.
- Zhao YY, Wu LY, Zhang JH, Wang RS, Zhang XS (2005) Haplotype assembly from aligned weighted SNP fragments. Computational Biology and Chemistry 29:281-287.