مجله تحقیق در عملیات در کاربردهای آن، سال پانزدهم، شماره چهارم (پیاپی ۵۹)، زمستان ۹۷، ص ص ۷۷-۶۱
شاپا ۷۲۸۶-۲۲۵۱

یک مدل ریاضی جدید برای مساله استنباط هاپلوتایپ‌ها از ژنوتایپ‌ها با معیار پارسیمونی

رضا فیض‌آبادی[۱]، مهری باقریان[۲*]

۱- دانشجوی دکتری، گروه ریاضی کاربردی، دانشکده علوم ریاضی، دانشگاه گیلان، رشت، ایران

۲- استادیار گروه ریاضی کاربردی، دانشکده علوم ریاضی، دانشگاه گیلان، رشت، ایران

رسید مقاله: ۱۰ اردیبهشت ۱۳۹۶
پذیرش مقاله: ۲ آبان ۱۳۹۶

چکیده

مساله استنباط هاپلوتایپ‌ها از ژنوتایپ‌ها یکی از مسایل مهم حوزه ریاضیات زیستی است. اهمیت این مساله به‌دلیل کاربردهای فراوان آن در تشخیص و درمان بیماری‌های ژنتیکی همچون دیابت، آلزایمر و امراض قلبی است که موجب رقابت پژوهشگران در ارایه مدل‌های ریاضی بهتر و طراحی الگوریتم‌های کاراتر برای حل این مساله شده است. علی‌رغم پژوهش‌های فراوان، به دلیل NP-hard بودن مساله همچنان نیاز به ارایه مدل‌های جدید و یا بهبود روش‌های قبلی احساس می‌شود. استنباط هاپلوتایپ‌ها تحت معیارهای متفاوتی بیان می‌شود. پارسیمونی یکی از مهم‌ترین آن‌هاست و در این مقاله مساله با این معیار مورد بررسی قرار گرفته است. روش‌های حل مساله استنباط هاپلوتایپ‌ها از ژنوتایپ‌ها با معیار پارسیمونی به دو دسته دقیق و تقریبی تقسیم می‌شود. اغلب روش‌های دقیق مساله را به‌صورت یک مساله برنامه‌ریزی با اعداد صحیح فرمول‌بندی می‌کنند. اخیرا در مقاله‌ای یک مدل دقیق به نام HI Base -10 برای این مساله ارایه شده که ابتدا به هر هاپلوتایپ و ژنوتایپ یک عدد متناظر کرده و سپس مدل را بر اساس این اعداد تشکیل می‌دهد که درآن هیچ متغیر و قیدی متناظر جایگاه‌های هتروزیگوت به مساله تحمیل نمی‌شود. در این مقاله نیز با شیوه‌ای متفاوت به ژنوتایپ‌ها اعدادی متناظر کرده و بر اساس این اعداد یک مساله برنامه‌ریزی با متغیرهای دودویی و آمیخته می‌سازیم. در نتیجه این تبدیلات، مدل جدید، متغیر عدد صحیح نداشته و متغیرهای کم‌تری نسبت به HI Base – 10 دارد. به علاوه در مدل جدید هیچ متغیر و قیدی متناظر جایگاه‌های هموزیگوت وجود ندارد و متغیرها به جایگاه‌های هتروزیگوت اختصاص داده می‌شوند. با توجه به تعداد زیاد جایگاه‌های هموزیگوت در مقایسه با جایگاه‌های هتروزیگوت در داده‌های واقعی ارزش این مدل مشخص می‌شود.

کلمات کلیدی: بیوانفورماتیک، استنباط هاپلوتایپ‌ها، مدل عدد صحیح، پارسیمونی، ژنوتایپ

* عهده‌دار مکاتبات
آدرس الکترونیکی: mbagherian@guilan.ac.ir

۱ مقدمه

انسان یک موجود دیپلوئید است؛ یعنی DNA انسان به ۲۳ جفت کروموزوم تقسیم می‌شود. هر زوج کروموزوم نسخه یکسانی ازکروموزوم‌هایی است که از هر یک از والدین به ارث رسیده است. کروموزوم شامل مولکول پلیمری بسیار طویل از زیرواحدهای ساده‌تری به نام نوکلئوتید است. تنها چهار نوع نوکلئوتید در ساختمان DNA نقش دارند که آن‌ها را به اختصار با حروف A، T، C و G می‌شناسند. به جز در ۰.۱ درصد از جایگاه‌های نوکلئوتیدی، در بقیه جایگاه‌ها عناصر متناظر روی هر دو کروموزوم یک فرد یا بین افراد جامعه یکسان است [۱،۲،۳،۴]. به جایگاه‌های اختلاف‌دار چندریختی تک‌نوکلئوتیدی[1] یا به اختصار اسنیپ (SNP) گفته می‌شود. تقریبا همه اسنیپ‌ها دو آللی[2] هستند؛ یعنی تنها دو نوکلئوتید از چهار حالت ممکن در یک جایگاه مشاهده می‌شود. به آلل‌های مشاهده‌شده با تکرار بالا در جمعیت، نوع غالب[3] گفته می‌شود [۵] و این جایگاه‌ها به ۰ کدگذاری می‌شود. آلل‌های با تکرار کم به ۱ کدگذاری می‌شود و به آن‌ها نوع مغلوب[4] گفته می‌شود [۵].

با صرف‌نظرکردن از جایگاه‌های یکسان و پشت سر هم قراردادن اسنیپ‌ها دنباله‌ای از ارقام ۰ و ۱ حاصل می‌شود که به آن هاپلوتایپ می‌گوییم. شکل ۱.الف بخشی از کروموزوم‌های k فرد را نمایش می‌دهد. جایگاه‌هایی که با زمینه رنگی مشخص شده‌است، در تمام جمعیت دارای نوکلئوتید یکسان نیست و اسنیپ نامیده شده‌است. با حذف جایگاه‌های حاوی نوکلئوتید یکسان و پشت هم ردیف‌کردن اسنیپ‌ها در شکل ۱.ب هاپلوتایپ‌ها ایجاد و در شکل ۱.پ کدگذاری شده‌است.

مقایسه هاپلوتایپ‌ها بین اشخاص سالم و بیمار به تشخیص و درمان بسیاری از بیماری‌ها کمک می‌کند. تخمین هاپلوتایپ‌ها در زمینه‌های مختلفی همچون طراحی دارو، عملکرد ژن‌ها، نقشه‌ی ژنی بیماری‌های پیچیده و غیره کمک می‌کند [۶]. برخی از مطالعاتی که با استفاده از هاپلوتایپ‌ها به مطالعه بیماری‌ها می‌پردازند عبارتند از: [۷، ۸، ۹] در دیابت نوع ۱، [۱۰، ۱۱] در دیابت نوع ۲، [۱۲،۱۳] در آلزایمر، [۱۴] در سکته مغزی، [۱۵] در آسم و [۱۶] در بیماری عروق کرونر.

متاسفانه استخراج هاپلوتایپ‌ها با روش‌های آزمایشگاهی بسیار گران و زمان‌بر است؛ اما ژنوتایپ‌ها که در واقع ترکیب هاپلوتایپ‌های پدرانه و مادرانه هستند از روش‌های آزمایشگاهی با سرعت و با هزینه کم قابل دستیابی‌اند. درآزمایشگاه برای هر فرد تنها یک ژنوتایپ به دست می‌آید که رشته‌ای به طول هاپلوتایپ‌هاست و اطلاعات آن بدین صورت است که اگر جایگاهی روی هر دو هاپلوتایپ‌های پدرانه و مادرانه دارای نوکلئوتید یکسان باشد، جایگاه متناظر بر روی ژنوتایپ نیز همان نوکلئوتید را دارد. چنین جایگاه‌هایی هموزیگوت نامیده می‌شود و بر اساس نوع غالب (بیش‌ترین فراوانی در آن جایگاه) یا نوع مغلوب (فراوانی کم‌تر در آن جایگاه) به ترتیب مقادیر ۰ و ۱ به آن جایگاه اختصاص می‌یابد. اگر جایگاهی بر روی یکی از هاپلوتایپ‌ها دارای یک نوکلئوتید و روی هاپلوتایپ دیگر دارای نوکلئوتید دیگری باشد، به جایگاه متناظر روی ژنوتایپ مقدار۲

[1] Single Nucleotide Polymorphis (SNP)
[2] biallelic
[3] major
[4] minor

اختصاص می‌یابد. چنین جایگاهی هتروزیگوت نامیده می‌شود. درحالت اخیر این اطلاع که مقدار ۰ یا ۱ روی کدام یک از هاپلوتایپ‌های پدرانه یا مادرانه است، از دست می‌رود و تنها آگاهی ما از وجود اختلاف روی دو هاپلوتایپ در چنین جایگاهی است.

(الف)

individual 1 (father)	....	A	G	T	C	A	C	G	....	A	T	G	G	T	C	T	....	T	A	C	G	G	A	C	....
individual 1 (mother)	....	A	G	T	C	A	C	G	....	A	T	G	A	T	C	T	....	T	A	C	A	G	A	C	....
individual 2 (father)	....	A	G	T	T	A	C	G	....	A	T	G	G	T	C	T	....	T	A	C	A	G	A	C	....
individual 2 (mother)	....	A	G	T	C	A	C	G	....	A	T	G	G	T	C	T	....	T	A	C	A	G	A	C	....
:		:	:	:	:	:	:	:		:	:	:	:	:	:	:		:	:	:	:	:	:	:	
:		:	:	:	:	:	:	:		:	:	:	:	:	:	:		:	:	:	:	:	:	:	
individual k (father)	....	A	G	T	T	A	C	G	....	A	T	G	G	T	C	T	....	T	A	C	G	G	A	C	....
individual k (mother)	....	A	G	T	C	A	C	G	....	A	T	G	A	T	C	T	....	T	A	C	A	G	A	C	....

(ب)

haplotype 1	....	C	G	G	....
haplotype 2	....	C	A	A	....
haplotype 3	....	T	G	A	....
haplotype 4	....	C	G	A	....
:		:	:	:	
:		:	:	:	
haplotype 5	....	T	G	G	....
haplotype 6	....	C	A	A	....

(پ)

haplotype 1	....	0	0	1	....
haplotype 2	....	0	1	0	....
haplotype 3	....	1	0	0	....
haplotype 4	....	0	0	0	....
:		:			
:		:			
haplotype 5	....	1	0	1	....
haplotype 6	....	0	1	0	....

شکل ۱. الف. بخشی از کروموزوم‌های پدرانه و مادرانه در یک جمعیت نمونه. **ب.** هاپلوتایپ‌های متناظر کروموزوم‌ها. پ. هاپلوتایپ‌های کدشده.

به‌دست آوردن ژنوتایپ‌ها در آزمایشگاه آسان و کم‌هزینه است؛ اما آن چه در کاربرد ارزشمند است، هاپلوتایپ-ها هستند. خوشبختانه محاسبه هاپلوتایپ‌ها از روی ژنوتایپ‌ها توسط روش‌های محاسباتی امکان‌پذیر است. به محاسبه هاپلوتایپ‌ها از روی ژنوتایپ‌ها، استنباط هاپلوتایپ‌ها[1] گفته می‌شود[۱۷]. مشکل این محاسبه در تعداد بسیار زیاد زوج هاپلوتایپ‌هایی است که می‌توانند هاپلوتایپ‌های اولیه تعدادی ژنوتایپ باشند. این تعداد به طور دقیق $۲^{\alpha-۱}$ زوج برای ژنوتایپی با α جایگاه هتروزیگوت است؛ لذا برای انتخاب زوج هاپلوتایپ‌های صحیح از بین زوج‌های ممکن به یک معیار نیاز داریم. دو معیار مهم که انطباق زیادی با واقعیت دارند، پارسیمونی[2] و بیش-ترین درستنمایی[3] هستند. ما در این مقاله از معیار پارسیمونی استفاده می‌کنیم.

1 Haplotype Inference
2 parsimony
3 Maximum likelihood

اگر یک مجموعه شامل n ژنوتایپ داده شده باشد براساس معیار پارسیمونی مجموعه هاپلوتایپ‌هایی که قابلیت تولید ژنوتایپ‌ها را داشته باشند، باید کم‌ترین تعداد عضو ممکن را داشته باشند. مساله تحت این معیار PPH نامیده شده است.

مساله استنباط هاپلوتایپ‌ها از روی ژنوتایپ‌ها با معیار پارسیمونی NP-hard است [۱۷، ۱۸]؛ بنابراین روش‌های تقریبی زیادی برای حل این مساله ارایه شده‌است که برخی از آن ها عبارت است از [۱۷، ۲۸-۱۹].

برخی از روش‌ها نیز در پی یافتن جواب دقیق این مساله هستند. این روش‌ها اغلب مساله را به صورت یک مدل برنامه‌ریزی عدد صحیح فرمول‌بندی می‌کنند. اولین مدل برای مساله PPH معروف به TIP توسط گاسفیلد در ۲۰۰۳ ارایه شد [۱۸] که تعداد متغیرها و قیدهای آن از مرتبه‌ی نمایی است. لانسیا و سارافینی در ۲۰۰۹ مدل نمایی دیگری ارایه کردند [۲۹]. اولین مدل‌های چندجمله‌ای در ۲۰۰۴ توسط لانسیا و همکارانش [۱۷] و همچنین توسط هالدورسون و همکارانش[۳۰] ارایه شدند. بعد از آن مدل‌های پیوندی براون و هاروور در ۲۰۰۶ [۳۱] و چندجمله‌ای کاتانزارو در ۲۰۰۷ [۳۲] و چندجمله‌ای برتولازی در ۲۰۰۸ [۳۳] ارایه شد. دو روش دقیق دیگر که به تازگی ارایه شده اند [۳۴] و [۳۵] هستند.

اخیرا در مقاله‌ای در ۲۰۱۶ یک مدل عدد صحیح آمیخته چندجمله‌ای به نام HI Base- 10 برای PPH ارایه شده که هیچ متغیر و قیدی متناظر جایگاه‌های هتروزیگوت ندارد [۳۶]. در این مقاله مدل‌بندی دیگری برای این مساله ارایه می‌کنیم که بر خلاف مدل قبلی هیچ متغیر و قیدی متناظر جایگاه‌های هموزیگوت ندارد. همان طور که نشان خواهیم داد، با توجه به اینکه معمولا تعداد جایگاه‌های هموزیگوت بیش‌تر از تعداد جایگاه‌های هتروزیگوت است، مدل جدید در بسیاری از نمونه‌ها سریع‌تر از مدل قبلی اجرا می‌شود.

ساختار مقاله در ادامه بدین صورت است که در بخش دوم مدل HI Base- 10 تشریح می‌شود. در بخش سوم مدل جدید HI Base- 10 اصلاح‌شده را می‌آوریم. بخش چهارم به نتایج محاسباتی بر روی نمونه‌های شبیه‌سازی شده و مقایسه HI Base- 10 و HI Base- 10 اصلاح‌شده اختصاص دارد. در انتها بخش پنجم، نتیجه گیری را ارایه می‌کند.

۲ مدل HI Base- 10

n ژنوتایپ داده شده است که هر کدام از m جایگاه شامل اعداد ۰، ۱ و ۲ تشکیل شده اند. جایگاه‌ها را از راست به چپ از ۱ تا m اندیس‌گذاری می‌کنیم و s امین جایگاه ژنوتایپ t ام را با $g_t(s)$ نمایش می‌دهیم.

هاپلوتایپ‌ها نیز رشته‌هایی به طول m هستند که هر جایگاه‌شان یکی از اعداد ۰ و ۱ را شامل می‌شود. اندیس‌گذاری جایگاه‌های هاپلوتایپ‌ها که آن ها را با حرف h نمایش می‌دهیم با ژنوتایپ‌ها یکسان است.

در صورتی که رابطه‌ی زیر بین هاپلوتایپ‌های h_i و h_j و ژنوتایپ g برقرار باشد h_i و h_j را هاپلوتایپ-های مولد ژنوتایپ g گوییم و این رابطه را با $g = h_i \oplus h_j$ نمایش می‌دهیم.

$$g(s)=\begin{cases}\circ & h_i(s)=h_j(s)=\circ\\ ۱ & h_i(s)=h_j(s)=۱\\ ۲ & o.w.\end{cases}$$

برای مثال $h_i=\circ\circ۱$ و $hj=۱\circ۱$ هاپلوتایپ‌های مولد $g=۲\circ۱$ هستند؛ یعنی $\circ\circ۱\oplus۱\circ۱=۲\circ۱$ است. هاپلوتایپ‌ها وژنوتایپ‌ها را با تعریف فوق هاپلوتایپ و ژنوتایپ معمولی می‌نامیم و با اعمال تبدیلاتی بر روی آن‌ها انواع دیگری را از هاپلوتایپ‌ها و ژنوتایپ‌ها به دست می‌آوریم که در زیر به تعریف آن‌ها و تشریح هرکدام با استفاده از هاپلوتایپ معمولی $h=۱\circ۱\circ۱$ و ژنوتایپ معمولی $g=۲\circ۱۲۱\circ$ می‌پردازیم.

ژنوتایپ حسابی: از جایگزین‌کردن مقادیر ۱ با ۲ و بالعکس در ژنوتایپ معمولی حاصل می‌شود. در ژنوتایپ معمولی نمونه، $g=۲\circ۱۲۱\circ$، جایگاه‌های دوم و چهارم از سمت راست حاوی مقدار ۱ هستند که مقدار آن ها را به ۲ تغییر می‌دهیم و جایگاه‌های سوم و ششم دارای مقدار ۲ هستند، لذا مقدار این جایگاه‌ها به ۱ تغییر می‌یابد. حاصل ژنوتایپ حسابی $\hat{g}=۱\circ۲۱۲\circ$ است.

ژنوتایپ دودویی: با جایگزین کردن هر مقدار ۲ با جمع دو مقدار ۱ در ژنوتایپ حسابی و به‌طور همزمان در نظرگرفتن ژنوتایپ به صورت عددی در مبنای ۲ حاصل می‌شود. به عبارتی چنانچه جایگاه sام ژنوتایپ حسابی مقدار ۲ داشته باشد، مقدار آن را به ۰ تغییر داده و مقدار ۱ به جایگاه $s+۱$ام این ژنوتایپ اضافه می‌کنیم؛ بنابراین دنباله حاصل فقط شامل اعداد ۰ و ۱ خواهد بود. ژنوتایپ دودویی متناظر با ژنوتایپ حسابی $\hat{g}=۱\circ۲۱۲\circ$ برابر $\tilde{g}=۱۱۱\circ\circ\circ$ می‌باشد و بدین صورت محاسبه شده است: با توجه به این که مقدار جایگاه اول ۰ است هیچ تغییری در این جایگاه نداریم. مقدار ۲ در جایگاه دوم $\hat{g}$ با مجموع دو مقدار ۱ جایگزین می‌شود که درنتیجه مقدار ۰ به این جایگاه اختصاص یافته و مقدار ۱ به جایگاه بعدی؛یعنی سوم اضافه می‌شود. جایگاه سوم از قبل دارای مقدار ۱ بوده است و در نتیجه افزایش یک واحد به آن مقدار این جایگاه ۰ خواهد شد و یک واحد به مقدار جایگاه بعدی (چهارم) اضافه می‌شود. اکنون جایگاه چهارم باتوجه به مقدار قبلی‌اش حاوی مجموع سه مقدار ۱ است که یکی از آن ها در این جایگاه قرار گرفته و دو مقدار ۱ دیگر، یک واحد به جایگاه پنجم اضافه می‌کنند. درنتیجه مقدار جایگاه پنجم از ۰ به ۱ افزایش می‌یابد. مقدار جایگاه ششم تغییری نمی‌کند. در نتیجه‌ی تبدیل ژنوتایپ حسابی به دودویی، ممکن است طول ژنوتایپ یک واحد افزایش یابد؛ ولی این امر خللی در روند محاسبات در ادامه به وجود نمی‌آورد.

ژنوتایپ عددی: با در نظر گرفتن هر ژنوتایپ دودویی به صورت عددی در مبنای ۲ و انتقال آن به مبنای ۱۰ حاصل می‌شود؛ بنابراین برای محاسبه ژنوتایپ عددی از $\tilde{g}$ آن را به صورت عددی در مبنای ۲ در نظر می‌گیریم، سپس این عدد را به مبنای ۱۰ انتقال می‌دهیم. واضح است که برای این انتقال باید جایگاه jام ژنوتایپ دودویی را در $۲^{j-۱}$ ضرب کرده و مقادیر حاصل را با هم جمع کنیم؛ لذا مقدار ژنوتایپ عددی برای نمونه تحت بررسی به صورت زیر به دست می‌آید:

$$z=\circ\times۲^{\circ}+\circ\times۲^{۱}+\circ\times۲^{۲}+۱\times۲^{۳}+۱\times۲^{۴}+۱\times۲^{۵}=۵۶$$

هاپلوتایپ عددی: با درنظرگرفتن هر هاپلوتایپ معمولی به صورت عددی در مبنای ۲ و انتقال آن به مبنای ۱۰حاصل می‌شود. مقدار هاپلوتایپ عددی متناظر با هاپلوتایپ نمونه،$h=۱۰۱۰۱$ به صورت زیر به دست می-آید:

$$z=۱\times۲^{۰}+۰\times۲^{۱}+۱\times۲^{۲}+۰\times۲^{۳}+۱\times۲^{۴}=۲۱$$

جداول ۱ و ۲ انواع ژنوتایپ‌ها و هاپلوتایپ‌ها و نمونه‌ای از این تبدیلات را برای چند مثال دیگر نمایش می‌دهند.

قضیه زیر ارتباط بین ژنوتایپ‌های عددی و هاپلوتایپ‌های عددی را بیان می‌کند.

قضیه ۱. [۳۶] ژنوتایپ عددی z متناظر با ژنوتایپ معمولی g مفروض است. اگر هاپلوتایپ‌های عددی $h_۱$ و $h_۲$ در شرایط (۱)- (۵) صدق کنند، آن‌گاه هاپلوتایپ‌های معمولی مولدشان ژنوتایپ g را تولید می‌کنند.

جدول ۱. چند مثال از ژنوتایپ های معمولی، حسابی، دودویی و عددی

اندیس ژنوتایپ	ژنوتایپ معمولی	ژنوتایپ حسابی	ژنوتایپ دودویی	ژنوتایپ عددی
۱	۱۲۲۱	۲۱۱۲	۱۱۰۰۰	۲۴
۲	۲۰۲۲	۱۰۱۱	۱۰۱۱	۱۱
۳	۲۰۱۱	۱۰۲۲	۱۱۱۰	۱۴
۴	۱۰۲۲	۲۰۱۱	۱۰۰۱۱	۱۹
۵	۱۲۰۲	۲۱۰۱	۱۰۱۰۱	۲۱

جدول ۲. چند مثال از هاپلوتایپ های معمولی و عددی

اندیس هاپلوتایپ	هاپلوتایپ معمولی	هاپلوتایپ عددی
$h_۱$	۱۰۱۱	۱۱
$h_۲$	۱۱۰۱	۱۳
$h_۳$	۰۰۱۱	۳
$h_۴$	۱۰۰۰	۸

$$h_۱+h_۲=z \qquad (۱)$$

$$h_۱=۲^{p}(q_{۱,p}+r_{۱,p}) \qquad \forall p:\ g(p)\neq ۲ \qquad (۲)$$

$$h_۲=۲^{p}(q_{۲,p}+r_{۲,p}) \qquad \forall p:\ g(p)\neq ۲ \qquad (۳)$$

$$۰\le r_{i,p}<۰/۵ \qquad i=۱,۲, \qquad \forall p:\ g(p)=۰ \qquad (۴)$$

$$۰/۵\le r_{i,p}<۱ \qquad i=۱,۲, \qquad \forall p:\ g(p)=۱ \qquad (۵)$$

همان‌طورکه از روابط (۱) تا (۵) مشخص است برای تولید هاپلوتایپ‌های مولد یک ژنوتایپ، جایگاه‌های هتروزیگوت هیچ محدودیتی اعمال نمی‌کنند. تنها به اعمال محدودیت روی جایگاه‌های هموزیگوت (جایگاه هایی که $g\neq ۲$) نیاز داریم و این محدودیت‌ها با تقسیم هاپلوتایپ عددی بر ۲^{p} در (۲) و (۳) ساخته

می‌شود. با کنترل باقیمانده این تقسیم در (۴) و (۵) مقدار جایگاه هموزیگوت p ام برای هاپلوتایپ مشخص می‌شود. به طوری که اگر جایگاه p ام هموزیگوت از نوع غالب باشد باید باقیمانده کم‌تر از ۰/۵ باشد (قید ۴) و اگر هموزیگوت از نوع مغلوب باشد، باقیمانده بزرگ‌تر یا مساوی ۰/۵ باشد (قید ۵).

بنابراین مدل ارایه شده توسط روابط (۶)- (۱۴) هاپلوتایپ‌های عددی را محاسبه می‌کند که به راحتی می‌توان از آن ها هاپلوتایپ‌های معمولی را تولید کرد.

$$\text{Min} \quad \sum_{j=1}^{2n} x_j \tag{۶}$$

$$s.t. \quad h_{2t-1} + h_{2t} = z_t \qquad t = 1,\dots,n \tag{۷}$$

$$h_{2t-1} - 2^{m-s+1}(q_{2t-1,s} + r_{2t-1,s}) = 0 \qquad t = 1,\dots,n \qquad \forall s:\ g_t(s) \neq 2 \tag{۸}$$

$$h_{2t} - 2^{m-s+1}(q_{2t,s} + r_{2t,s}) = 0 \qquad t = 1,\dots,n \qquad \forall s:\ g_t(s) \neq 2 \tag{۹}$$

$$r_{2t-1,s}, r_{2t,s} \leq 0/5 - \varepsilon \qquad t = 1,\dots,n \qquad \forall s:\ g_t(s) = 0 \tag{۱۰}$$

$$0/5 \leq r_{2t-1,s}, r_{2t,s} \leq 1 - \varepsilon \qquad t = 1,\dots,n \qquad \forall s:\ g_t(s) = 1 \tag{۱۱}$$

$$h_i - h_j \leq (2^m - 1)y_{ij} \qquad i, j = 1,\dots,2n\ ;\ i \neq j \tag{۱۲}$$

$$x_j \geq 2n - j + 1 - \sum_{i>j}(1 - y_{ij}) - \sum_{i>j}(1 - y_{ji}) \qquad j = 1,\dots,2n \tag{۱۳}$$

$$h_j \leq 2^m - 1 \qquad j = 1,\dots,2n \tag{۱۴}$$

$$x_j, h_j, q_{j,s}, r_{j,s} \geq 0 \qquad j = 1,\dots,2n \qquad s = 1,\dots,m$$

$$h_j, q_{j,s} \in Z, \qquad r_{j,s} \in R \qquad j = 1,\dots,2n \qquad s = 1,\dots,m$$

$$x_j, y_{ij} \in \{0,1\} \qquad i, j = 1,\dots,2n$$

که در آن فرض شده نمونه تحت بررسی حاوی n ژنوتایپ است و جایگاه sام ژنوتایپ tام با $g_t(s)$ نمایش داده شده است. ژنوتایپ عددی z_t متناظر g_t فرض شده است. از آن جا که در بدترین حالت به $2n$ هاپلوتایپ برای تولید n ژنوتایپ نیازمندیم، $2n$ متغیر عدد صحیح متناظر هاپلوتایپ‌های عددی در نظر گرفته شده‌اند که مانند قبل از حرف h برای نمایش آن ها استفاده شده است و فرض شده h_{2t} و h_{2t-1} هاپلوتایپ‌های عددی مولد ژنوتایپ tام باشند. در مدل $2n$ متغیر دودویی x_j متناظر هاپلوتایپ‌ها وجود دارد که مقدار ۱ در آن‌ها بیانگر حضور h_j درجواب بهینه و مقدار ۰ برای آن نشان از عدم حضور h_j در مجموعه هاپلوتایپ‌های جواب دارد. متغیرهای $q_{j,s}$ و $r_{j,s}$ به ترتیب حاوی مقادیر صحیح و اعشاری حاصل از تقسیم هاپلوتایپ عددی h_j بر 2^{m-s+1} برای مشخص‌کردن مقدار جایگاه sام در هاپلوتایپ معمولی h_j می‌باشند. قیود (۷) تا (۱۱) مشابه روابط (۱) تا (۵) تضمین می‌کنند که هاپلوتایپ‌های به دست آمده از حل مدل به درستی ژنوتایپ‌ها را تولید می‌کند. ε یک عدد بسیار کوچک است و تنها برای تبدیل قیود از نوع کوچک‌تری اکید به کوچک‌تر مساوی از مقدار سمت راست قیود (۱۰) و (۱۱)کم شده است. مقدار $1/2^m$ برای کوچکی ε کفایت می‌کند [۳۶].

تعریف متغیرهای $y_{i,j}$ به گونه‌ای است که اگر $h_i = h_j$ باشد بین $y_{i,j}$ و $y_{j,i}$ یکی مقدار ۱ و دیگری مقدار ۰ داشته باشد. این شرط توسط دسته قیود (۱۲) و ماهیت مینیمم‌سازی تابع هدف برقرار است. M یک عدد بسیار بزرگ است و $2^m - 1$ برای بزرگی آن کفایت می‌کند [۳۶]. دلیل اضافه‌کردن M این است که $y_{i,j}$ها در سمت راست دسته محدودیت‌های (۱۲) دودویی هستند در حالی که سمت چپ می‌تواند مقادیر بزرگ‌تر از ۱

داشته باشد. با توجه به مقادیر به دست آمده برای $y_{i,j}$ها هر هاپلوتایپ h_j با هاپلوتایپ‌هایی که اندیس بالاتر از خودش دارند مقایسه می‌شود. چنانچه لااقل با یکی از آن ها برابر باشد x_j مقدار ۰ می‌گیرد و درغیراین صورت x_j مقدار ۱ خواهد داشت و بدین صورت مجموع $\sum_{j=1}^{2n} x_j$ تعداد هاپلوتایپ‌های متمایز را شمارش می‌کند که با توجه به معیار پارسیمونی باید مینیمم باشد و این منجر به تعریف تابع هدف (۶) می‌شود. بزرگ‌ترین مقداری که یک هاپلوتایپ عددی می‌تواند اختیار کند، زمانی است که تمام جایگاه‌های هاپلوتایپ معمولی متناظرش هموزیگوت از نوع مغلوب باشد. دراین صورت مقدار هاپلوتایپ عددی $2^m - 1$ خواهدبود؛ لذا $2^m - 1$ یک کران بالا برای هاپلوتایپ‌های عددی است. این محدودیت بزرگی h_j که در دسته قیود (۱۴) آمده است، قابل حذف از مدل می‌باشد.

۳ مدل HI Base- 10 اصلاح شده

مدل HI Base- 10 هر هاپلوتایپ و ژنوتایپ را یک عدد در مبنای ۲ در نظر می‌گیرد. سپس این اعداد را به مبنای ۱۰ برده و مدل را بر اساس این اعداد بنا می‌کند. در نتیجه این تغییر، مدل نسبت به مدل‌های قبلی کاراتر می‌شود، چون هیچ متغیر و قیدی متناظر جایگاه‌های هتروزیگوت ندارد.

در این بخش برای بهبود مدل، با شیوه‌ای متفاوت به ژنوتایپ‌ها یک عدد متناظر می‌کنیم و مدل جدیدی روی این اعداد بنا می‌کنیم. برای درک بهتر، تناظر بین ژنوتایپ‌ها و اعداد را با یک مثال توضیح می‌دهیم. ژنوتایپ حسابی $g = 2120.1210$ و شکل ۲ را متناظر آن درنظر بگیرید. همان طور که در شکل ۲ مشاهده می‌شود، مانند قبل اندیس‌گذاری برای جایگاه‌ها از سمت راست به چپ صورت می‌گیرد.

ارزش هر جایگاه:	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
ژنوتایپ g:	۲	۱	۲	۰	۱	۲	۱	۰
اندیس هر جایگاه:	۸	۷	۶	۵	۴	۳	۲	۱

شکل ۲. ژنوتایپ تحت بررسی.

جایگاه‌ها برای هر ژنوتایپ حسابی از جمله ژنوتایپ g به سه دسته تقسیم می‌شوند:

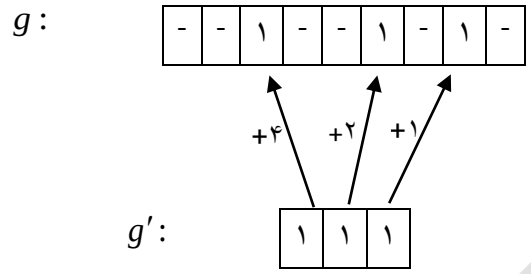
۱- جایگاه‌های هموزیگوت از نوع غالب: این جایگاه‌ها مقدار ۰ دارند و اندیس این جایگاه‌ها را در مجموعه $Homw$ نگهداری می‌کنیم. برای ژنوتایپ تحت بررسی $Homw = \{1,5\}$ است.

۲- جایگاه‌های هموزیگوت از نوع مغلوب: این جایگاه‌ها مقدار ۲ دارند و اندیس این جایگاه‌ها را در مجموعه $Homm$ قرار می‌دهیم. برای ژنوتایپ تحت بررسی $Homm = \{3,6,8\}$ است.

۳- جایگاه‌های هتروزیگوت: این جایگاه ها مقدار ۱ دارند.

مقدار جایگاه‌های نوع اول و دوم در هاپلوتایپ‌های مولد g از قبل مشخص است. به‌طوری‌که اگر h_1 و h_2 هاپلوتایپ‌های مولد g باشند، باید در جایگاه‌های اول و پنجم مقدار ۰ و در جایگاه‌های سوم، ششم و هشتم

مقدار ۱ داشته باشند؛ لذا این جایگاه‌ها را می‌توانیم در بررسی در نظر نگیریم. با حذف جایگاه‌های هموزیگوت از g و پشت سر هم قراردادن جایگاه‌های هتروزیگوت ژنوتایپ $g'=۱۱۱$ به دست می‌آید که جایگاه‌های اول، دوم و سوم آن به ترتیب متناظر جایگاه‌های دوم، چهارم و هفتم g می‌باشند. این تناظر در شکل ۳ به تصویر کشیده شده است.



شکل ۳. بیان تصویری تناظر بین g' و g.

با اضافه‌کردن یک مقدار ثابت به اندیس هر جایگاه در g'، اندیس جایگاه متناظر در g به دست می‌آید. این مقادیر را در آرایه c نگهداری می‌کنیم. که برای ژنوتایپ مورد نظر $c=(۱,۲,۴)$ است. ارزش عددی ژنوتایپ جدید g' (ژنوتایپ عددی متناظر g' که با درنظرگرفتن g' به صورت یک عدد در مبنای ۲ و تبدیل مبنای این عدد به ۱۰حاصل می‌شود.) برابر $۲^p-۱=۷$ است که p تعداد جایگاه‌های این ژنوتایپ و مساوی تعداد جایگاه‌های هتروزیگوت g است. بنابرآنچه در بخش قبل بیان شد، هاپلوتایپ‌های عددی $l_۱$ و $l_۲$ در صورتی می‌توانند مولد g' باشند که رابطه زیر برقرار باشد:

$$l_۱+l_۲=۲^p-۱ \qquad (۱۵)$$

برای مثال هاپلوتایپ‌های عددی $l_۱=۴$ و $l_۲=۳$ در رابطه (۱) صدق کرده و می‌توانند مولد g' باشند. هاپلوتایپ‌های معمولی متناظر آن ها به ترتیب برابر ۱۰۰ و ۰۱۱ می‌باشد. واضح است که این هاپلوتایپ‌ها، مولد g نمی‌باشند؛ زیرا تنها جایگاه‌های هتروزیگوت g را دارند. ولی می‌توان از روی آن ها و با توجه به بردار تناظر c هاپلوتایپ‌های $h_۱$ و $h_۲$ مولد g را ساخت. به این منظور باید در هرکدام تعداد مناسبی ۰ و ۱ بین جایگاه-هایشان اضافه شود. تعداد و نوع غالب یا مغلوب بودن جایگاه‌های اضافه شده توسط بردار تناظر c و مجموعه‌های $Homw$ و $Homm$ مشخص می‌شود. آن چه مد نظر ماست، محاسبه هاپلوتایپ‌های عددی $h_۱$ و $h_۲$ متناظر با هاپلوتایپ‌های عددی $l_۱$ و $l_۲$ است. به این منظور ارزش جایگاه‌های هر دسته از $h_i\,(i=۱,۲)$ را به طور جداگانه محاسبه و سپس جمع می‌کنیم.

جایگاه‌های نوع اول، هموزیگوت غالب، که اندیس آن ها با $Homw$ مشخص شده است، هیچ سهمی در مقدار هاپلوتایپ‌های عددی $h_i\,(i=۱,۲)$ ندارند و لذا هیچ نقشی در محاسبات ما ندارند.

هر جایگاه نوع دوم، هموزیگوت مغلوب، با اندیس j به مقدار $۲^{j-۱}$ در مقدار هاپلوتایپ عددی $h_i\,(i=۱,۲)$ نقش دارد؛ بنابراین این نوع جایگاه‌ها در مجموع سهمی برابر $\sum_{j\in Homm} ۲^{j-۱}$ دارند. این مقدار برای مثال تحت بررسی ما برابر $۲^۷+۲^۵+۲^۲$ است.

برای به دست‌آوردن ارزش متناظر جایگاه‌های نوع سوم؛ یعنی هتروزیگوت، ابتدا باید هاپلوتایپ‌های معمولی $l_i\,(i=۱,۲)$ را با تبدیل مبنای مقدار آن از ۱۰ به ۲ به دست آورد. این کار را می‌توان توسط قید زیر انجام داد.

$$l_i = b_{i,۱}۲^{۰}+b_{i,۲}۲^{۱}+b_{i,۳}۲^{۲}+....+b_{i,p}۲^{p-۱} \qquad i=۱,۲ \qquad (۱۶)$$

که $b_{i,j}$ ها متغیرهای دودویی بوده و مقدار جایگاه j ام را در هاپلوتایپ معمولی $l_i\,(i=۱,۲)$ نمایش می‌دهند. ارزش این جایگاه برای هاپلوتایپ $l_i\,(i=۱,۲)$ مولد g' برابر $۲^{j-۱}$ است؛ ولی ارزش این جایگاه در هاپلوتایپ $h_i\,(i=۱,۲)$ مولد g ، با توجه به بردار تناظر c برابر $۲^{j+c_j-۱}$ است. در نتیجه سهم کل جایگاه‌های از این نوع در $h_i\,(i=۱,۲)$ برابر $\sum_{j=۱}^{p} b_{i,j}۲^{j+c_j-۱}$ است.

لذا مقدار هاپلوتایپ‌های عددی $h_i\,(i=۱,۲)$ از رابطه زیر به دست می‌آید:

$$h_i = \sum_{j\in Homm} ۲^{j-۱}+\sum_{j=۱}^{p}۲^{j+c_j-۱}b_{i,j} \qquad i=۱,۲ \qquad (۱۷)$$

بنابراین قضیه زیر اثبات شد.

قضیه۲ . اگر g' ژنوتایپ حاصل از حذف جایگاه‌های هموزیگوت g با بردار تناظر c باشد و p تعداد جایگاه‌های هتروزیگوت g باشد. و چنانچه $h_i\,,l_i\,,b_{i,j}\,(i=۱,۲ \quad j=۱,\dots p)$، در روابط (۱۵)- (۱۷) صدق کنند، آن گاه $h_۱$ و $h_۲$ هاپلوتایپ‌های عددی مولد g خواهند بود.

با توجه به قضیه فوق مدل طراحی‌شده براساس روابط (۱۸)- (۲۳) را برای مساله PPH ارایه می‌کنیم.

$$\text{Min} \quad \sum_{j=۱}^{۲n} x_j \qquad (۱۸)$$

$$s.t. \quad l_{۲t-۱}+l_{۲t}=۲^{p_t}-۱ \qquad t=۱,۲,...,n \qquad (۱۹)$$

$$l_i=\sum_{j=۱}^{p_t} b_{i,j}۲^{j-۱} \qquad t=۱,۲,...,n \quad i=۲t-۱,۲t \qquad (۲۰)$$

$$h_i=\sum_{j\in Homm_i} ۲^{j-۱}+\sum_{j=۱}^{p_t}۲^{j+c_{i,j}-۱}b_{i,j} \qquad t=۱,۲,...,n \quad i=۲t-۱,۲t \qquad (۲۱)$$

$$h_i-h_j \le My_{ij} \qquad i,j=۱,۲,...,۲n \qquad (۲۲)$$

$$x_j \ge ۲n-j+۱-\sum_{i>j}(۱-y_{ij})+\sum_{i>j}(۱-y_{ji}) \qquad j=۱,۲,...,۲n \qquad (۲۳)$$

$$l_i,h_i\in\mathbb{Z} \qquad i=۱,۲,...,۲n$$

$$y_{ij},x_i\in\{۰,۱\} \qquad i,j=۱,۲,...,۲n$$

$$b_{i,j}\in\{۰,۱\} \qquad i=۱,۲,...,۲n \qquad j=۱,۲,...,p_t \qquad t=۱,۲,...,n$$

که در آن n تعداد ژنوتایپ‌ها و p_t تعداد جایگاه‌های هتروزیگوت ژنوتایپ t ام می‌باشد. تعداد $2n$ هاپلوتایپ در نظر گرفته شده است و فرض شده است که هاپلوتایپ‌های l_{2t-1} و l_{2t} مولد ژنوتایپ g'_t از حذف جایگاه‌های هموزیگوت g_t و پشت سرهم قراردادن جایگاه‌های هتروزیگوت آن به دست آید و متناظر آن‌ها، h_{2t-1} و h_{2t} هاپلوتایپ‌های مولد ژنوتایپ g_t باشند. متغیر دودویی $b_{i,j}$ مقدار j امین جایگاه هاپلوتایپ l_i تعریف شده است. C ماتریس تناظر بین مجموعه ژنوتایپ‌های g' و g است. قیود (۱۹)- (۲۱) چنانچه در بالا نشان داده شد، تضمین می‌کنند که تمام ژنوتایپ‌ها توسط هاپلوتایپ‌ها تولید شوند.

M کران هاپلوتایپ‌های عددی h_i است و همان‌طورکه در [۳۶] نشان داده شده است، مقدار 2^m-1 برای بزرگی آن کفایت می‌کند. متغیرهای باینری y_{ij} به گونه‌ای هستند که اگر h_i با h_j مساوی باشد آن‌گاه y_{ij} و y_{ji} هر دو مقدار ۱ و در صورت عدم تساوی h_i و h_j یکی از y_{ij} و y_{ji} مقدار ۱ و دیگری مقدار ۰ داشته باشد. این محدودیت توسط مجموعه قیود (۲۲) و ماهیت مینیمم‌سازی مساله تضمین می‌شود. با استفاده از $y_{i,j}$ ها هر h_j با هاپلوتایپ‌های با اندیس بالاتر از خودش در مجموعه قیود (۲۳) مقایسه می‌شود. چنانچه لااقل با یکی از آن ها برابر باشد x_j مقدار ۱ و درغیراین‌صورت مقدار ۰ را اختیار می‌کند. در نهایت تعداد هاپلوتایپ-های متمایز توسط تابع هدف شمارش شده و بر اساس معیار پارسیمونی مینیمم مقدار ممکن آن ها بازگردانده می‌شود.

l_jها هاپلوتایپ‌های عددی مولد g باید مقادیر عدد صحیح اختیار کنند؛ اما عددصحیح بودن آن‌ها با توجه به دسته قیود (۲۰) برقرار است و نیازی به اعمال محدودیت در تعریف نوع این متغیرهای تصمیم وجود ندارد. همین مطلب در مورد h_jها به دلیل وجود دسته قیود (۲۱) و x_jها به دلیل دسته قیود (۲۳) و نوع مینیمم‌سازی تابع هدف برقرار است؛ بنابراین می‌توانیم این متغیرها را پیوسته تعریف کنیم.

تاکید می‌شود که در مدل جدید جایگاه‌های هاپلوتایپ‌ها و ژنوتایپ‌ها از راست به چپ اندیس‌گذاری شده‌اند. در صورت اندیس‌گذاری از چپ به راست توآن های متفاوتی از ۲ در مدل ظاهر خواهند شد.

با فرض اینکه تعداد ژنوتایپ‌ها n است، مدل جدید دو متغیر دودویی $b_{i,j}$ متناظر هر جایگاه هتروزیگوت و تعداد $2n^2$ متغیر دودویی $y_{i,j}$ دارد. متغیرهای پیوسته متناظر هر ژنوتایپ عبارت از دو l_j، دو h_j و دو x_j است که در کل تعداد $6n$ متغیر پیوسته مدل را تشکیل می‌دهند. مهم‌تر اینکه مدل جدید هیچ متغیر عدد صحیحی ندارد، که موجب شده مدل جدید در مقایسه با HI Base -10 سریع‌تر اجرا شود. جدول ۳ به مقایسه تعداد متغیرها و محدودیت‌ها بین HI Base-10 و مدل جدید می‌پردازد که درآن f تعداد جایگاه‌های هموزیگوت است.

HI Base-10 متناظر هر جایگاه هموزیگوت تعدادی متغیر عددصحیح و پیوسته دارد. در حالی که مدل جدید هیچ متغیری متناظر جایگاه‌های هموزیگوت ندارد و به جای آن فقط دو متغیر دودویی متناظر هر جایگاه هتروزیگوت دارد؛ بنابراین در نمونه‌هایی که تعداد جایگاه‌های هتروزیگوت آن‌ها در مقابل جایگاه‌های هموزیگوت زیاد است، مناسب می‌باشد. این وضعیت تقریبا در تمام نمونه‌های واقعی رخ می‌دهد.

جدول ۳. مقایسه تعداد متغیرها و محدودیت ها بین HI Base-10 و HI Base-10 اصلاح شده

	HI Base-10	HI Base-10 اصلاح شده
متغیر باینری	$۴n^۲-۲n$	$۴n^۲-۲f$
متغیر عدد صحیح	$۲n+۲f$	-
متغیر پیوسته	$۲n+۲f$	$۶n$
محدودیت	$\le ۴n^۲+۶f+۳n$	$۴n^۲+۷n$

٤ نتایج محاسباتی

در این بخش عملکرد مدل جدید ارزیابی می‌شود. بدین منظور زمان اجرای HI Base-10 و HI Base-10 اصلاح شده را برروی تعداد زیادی نمونه شبیه‌سازی شده مقایسه می‌کنیم.

برای پیاده‌سازی مدل‌های HI Base-10 و اصلاح‌شده آن از نرم افزار GAMS ورژن 24.1.2 و برای تولید نمونه‌های شبیه‌سازی شده از متلب ورژن 7.14.0.739 استفاده کردیم. هر دو مدل بر روی سیستمی با پردازنده Core i5-2430M, 2.40 GHz و با حافظه 2.69 GByte RAM اجرا شدند.

برای تولید نمونه‌ها ابتدا تعدادی بردار تصادفی دودویی تولید کردیم که هر کدام m مولفه دارد و متناظر هاپلوتایپی با m اسنیپ است. سپس یکی از هاپلوتایپ‌ها را به‌طور تصادفی انتخاب کرده در k جایگاه تصادفی آن جهش می‌دهیم. از تلفیق دو هاپلوتایپ یک ژنوتایپ می‌سازیم. این عمل را n بار تکرار می‌کنیم.

ما سه گروه هر کدام شامل ۴۰ نمونه تولید کردیم. در هر نمونه گروه اول، ۱۴ ژنوتایپ از تلفیق ۸ هاپلوتایپ تولید شده است که هر کدام ۱۵ اسنیپ دارند. برای گروه دوم تعداد ژنوتایپ‌ها، هاپلوتایپ‌ها و اسنیپ‌ها به ترتیب برابر ۱۶ ، ۱۰ و ۱۸ می‌باشند و این مقادیر برای گروه سوم به‌ترتیب برابر ۱۹، ۱۲ و ۲۰ است.

از آنجا که نسبت جایگاه‌های هتروزیگوت به هموزیگوت و بیش‌ترین تعداد جایگاه هموزیگوتی که یک ژنوتایپ در یک نمونه دارد بر روی تعداد متغیرها و کران آن‌ها در مدل‌ها تاثیر می‌گذارند، احتمالا بر روی زمان اجرا نیز اثرگذار هستند؛ لذا ما هر گروه را بر حسب بیش‌ترین تعداد جایگاه هتروزیگوت مشاهده شده در ژنوتایپ‌های یک نمونه به ۴ دسته ۱۰تایی تقسیم کردیم.

جداول ۴، ۵ و ۶ نتایج حاصل از اجرای دو مدل به ترتیب روی گروه‌های اول، دوم و سوم را نمایش می‌دهند. در ستون اول بیش‌ترین تعداد جایگاه هتروزیگوت مشاهده‌شده در نمونه درج شده است. ستون دوم اندیس نمونه تحت بررسی و ستون پنجم درصد جایگاه‌های هتروزیگوت تمام ژنوتایپ‌های یک نمونه به کل جایگاه‌های آن نمونه را دربردارند. ستون‌های سوم و چهارم به ترتیب زمان دستیابی به مقدار بهینه توسط دو مدل HI Base-10 و HI Base-10 اصلاح شده در نرم افزار GAMS را برحسب ثانیه نشان می‌دهند. همان طور که انتظار می‌

رود با افزایش جایگاه‌های هتروزیگوت فضای جواب بزرگ‌تر شده و لذا زمان دستیابی به جواب طولانی‌تر می‌شود؛ اما از آن جا که متغیرهای مدل HI Base-10 متناظر جایگاه های هموزیگوت و متغیرهای مدل HI Base-10 اصلاح‌شده متناظر جایگاه‌های هتروزیگوت هستند، در نمونه‌های ابتدای جدول با درصد هتروزیگوت کم‌تر مدل HI Base-10 اصلاح شده سریع‌تر از مدل HI Base-10 اجرا شده است؛ اما با افزایش درصد جایگاه‌های هتروزیگوت در انتهای جدول این برتری تغییر کرده است. به طوری که در دسته اول جدول ۴ (حاوی نمونه‌هایی با حداکثر ۵ جایگاه هتروزیگوت در ژنوتایپ ها) HI Base-10 اصلاح شده در ۸۰ درصد نمونه ها سریع‌تر از HI Base-10 به جواب بهینه رسیده است در حالی‌که با افزایش درصد جایگاه‌های هتروزیگوت در دسته‌های دیگر این روند به تدریج تغییر می‌کند تا جایی‌که در دسته آخر جدول ۴ (حاوی نمونه‌هایی با حداکثر ۸ جایگاه هتروزیگوت در ژنوتایپ ها) HI Base-10 در نیمی از نمونه‌ها سریع‌تر به جواب بهینه دست یافته است. روند نسبتا مشابهی در جداول ۵ و ۶ دیده می‌شود.

جدول ٤. مقایسه عملکرد بین HI Base-10 و HI Base-10 اصلاح شده روی نمونه‌هایی با ۱۴ ژنوتایپ، ۸ هاپلوتایپ و ۱۵ استیپ.

Max Hetero	Ins	HI Base 10 Time	Revised HI Base 10 Time	Hetero percent	Max Hetero	Ins	HI Base10 Time	Revised HI Base10 Time	Hetero percent
۵	۱	۴/۰۹	۰/۶۱	۲۳	۷	۱	۸/۴۴	۱۴/۴۱	۳۶
	۲	۴/۸۱	۰/۸۶	۲۳		۲	۳۳/۴۵	۷/۷۴	۳۵
	۳	۳۴/۳۴	۲/۷۷	۲۵		۳	۳/۰۵	۴۷/۶۶	۳۸
	۴	۳/۶۴	۶/۰۰	۲۵		۴	۸۸/۶۴	۲۷/۵۲	۳۸
	۵	۴/۵۳	۵/۷۵	۲۵		۵	۳/۸۶	۱۲/۸۹	۳۷
	۶	۶/۰۲	۲/۴۲	۲۵		۶	۳/۸۴	۷/۵۶	۳۷
	۷	۳۰/۳۳	۰/۳۶	۲۵		۷	۸۶/۵۸	۱۶/۷۰	۳۶
	۸	۱۱/۱۱	۲/۳۶	۲۲		۸	۴۵/۸۴	۱۱/۶۶	۳۶
	۹	۹/۱۷	۰/۸۴	۲۳		۹	۴۸۶/۷۰	۲۱/۸۹	۳۶
	۱۰	۲/۴۵	۲/۴۲	۲۴		۱۰	۵/۰۹	۶/۲۰	۳۶
۶	۱	۲۸/۱۱	۵/۴۱	۳۵	۸	۱	۵/۰۳	۸/۵۲	۴۱
	۲	۷/۰۵	۱۸/۱۹	۳۶		۲	۴۴/۰۵	۵۱/۸۳	۴۰
	۳	۱/۶۶	۳/۷۸	۳۴		۳	۴/۲۵	۴۰/۶۴	۴۱
	۴	۵/۷۷	۴/۴۷	۳۴		۴	۵۹/۴۵	۰/۷۵	۴۱
	۵	۱۸/۸۸	۲۴/۵۸	۳۵		۵	۴/۳۹	۳/۶۹	۴۱
	۶	۲۳/۳۴	۰/۴۱	۳۱		۶	۳۱/۰۶	۱۹/۷۲	۴۳
	۷	۱۰۶/۱۱	۲۷/۸۹	۳۲		۷	۹۰/۶۹	۱۴/۷۸	۴۰
	۸	۱۱۸/۶۶	۵/۷۲	۳۳		۸	۲/۵۹	۳۶/۴۴	۴۳
	۹	۳/۳۱	۱۲/۶۱	۳۶		۹	۳/۰۵	۳۸/۹۵	۴۵
	۱۰	۴۰/۱۶	۵/۱۹	۳۳		۱۰	۸۰/۴۷	۶۵/۲۸	۴۳

جدول ۵ . مقایسه عملکرد بین HI Base-10 و HI Base-10 اصلاح شده روی نمونه‌هایی با ۱۶ ژنوتایپ، ۱۰ هاپلوتایپ و ۱۸ استیپ

Hetero	Ins	HI Base 10 Time	Revised HI Base 10 Time	Hetero percent	Max Hetero	Ins	HI Base 10 Time	Revised HI Base 10 Time	Hetero percent
۵	۱	۳۵۰/۳۰	۵۱/۳۴	۲۰	۷	۱	۱۴/۳۳	۱۶۸/۳۸	۳۲
	۲	۶/۶۱	۲۲/۷۸	۲۰		۲	۱۰۷/۲۸	۳/۹۸	۳۱
	۳	۸/۲۳	۲/۵۶	۲۰		۳	۹۸/۹۲	۱/۳۴	۳۱
	۴	۱۰/۰۵	۷/۰۸	۱۹		۴	۱۱۳۳/۷۵	۱۵/۳۳	۲۸
	۵	۷/۳۳	۶/۸۶	۲۰		۵	۸۵/۰۵	۱۲۷/۷۲	۲۸
	۶	۳۳۶/۲۰	۲۹/۳۳	۲۱		۶	۳/۵۶	۹/۵۰	۳۱
	۷	۹/۶۶	۳/۸۹	۲۰		۷	۵۶/۶۱	۳۵/۲۳	۲۹
	۸	۱۲۴/۸۱	۴۶/۶۱	۱۹		۸	۷۱۹/۶۴	۵۲/۰۹	۳۲
	۹	۸۸/۲۷	۲/۳۱	۱۹		۹	۸/۱۴	۴/۰۶	۳۰
	۱۰	۷/۵۵	۱۴/۹۲	۱۹		۱۰	–	۱۳۶/۳۱	۳۱
۶	۱	۵۴/۲۷	۸۶/۰۸	۲۶	۸	۱	۱۰۷/۸۳	۳۹/۹۵	۳۵
	۲	۳/۷۷	۴/۵۶	۲۶		۲	۱۱۰/۰۰	۱۴۶/۱۷	۳۶
	۳	۴۷/۰۰	۳۶/۰۵	۲۶		۳	۲/۲۵	۱۶/۳۱	۳۵
	۴	۱۰۹۸/۲۰	۴۴/۴۲	۲۶		۴	۴۸۲/۷۳	۷۱۹/۵۸	۳۷
	۵	۵۶/۱۷	۷/۱۳	۲۶		۵	۶/۵۱	۱۴/۹۲	۳۶
	۶	۱۰/۰۸	۵۰/۰۳	۲۶		۶	۸/۹۴	۵۳/۹۲	۳۳
	۷	۱۱۹۸/۱۹	۹۹/۰۳	۲۷		۷	۳/۶۳	۲۲۸/۶۱	۳۷
	۸	۴/۹۸	۱۸/۱۹	۲۵		۸	۵۰/۷۰	۱۰۰/۱۶	۳۸
	۹	–	۶۳/۱۷	۲۶		۹	۱۰۹/۰۰	۹۴/۴۲	۳۶
	۱۰	۵/۰۲	۲/۶۱	۲۶		۱۰	۸/۳۸	۱۱۱/۱۷	۳۷

جدول ٦. مقایسه عملکرد بین HI Base-10 و HI Base-10 اصلاح شده روی نمونه‌هایی با ۱۹ ژنوتایپ، ۱۲ هاپلوتایپ و ۲۰ استیپ.

Max Hetero	Ins	HI Base 10 Time	Revised HI Base 10 Time	Hetero percent	Max Hetero	Ins	HI Base 10 Time	Revised HI Base 1۰ Time	Hetero percent
۶	۱	۱۱۱۸/۲۳	۸۶۷/۳۰	۲۴	۸	۱	۱۱۵۵/۹۸	۱۰۹۵/۴۴	۳۳
	۲	۱۹۵/۵۸	۸۵/۸۹	۲۳		۲	۱۱۴۵/۸۱	۲۲۶/۱۶	۳۲
	۳	۳۴۱/۷۳	۱۱۴۸/۵۸	۲۴		۳	۱۱۸۹/۶۱	۲۹/۴۱	۳۴
	۴	۸۲۰/۸۳	۴۵/۰۵	۲۳		۴	۳۷۲/۳۸	۴۹۶/۶۷	۳۳
	۵	۷/۷۴	۸۷/۸۳	۲۲		۵	۲۵۵/۰۸	۲۹۹/۷۲	۳۲
	۶	۵۳۹/۸۸	۸۰/۲۲	۲۲		۶	۱۱۵/۸۶	۱۱۷/۵۳	۳۴
	۷	۴۰۵/۵۹	۱۲۱/۲۷	۲۳		۷	۶۹۸/۲۷	۱۳۵/۲۳	۳۲
	۸	۲۱۵/۴۷	۲۰۲/۰۸	۲۴		۸	-	۸۷/۸۶	۳۴
	۹	-	۱۸۰/۹۵	۲۳		۹	۱۶۵/۳۳	۱۴/۱۱	۳۴
	۱۰	۲۰/۳۸	۸/۴۲	۲۴		۱۰	۳۷۷/۵۲	۲۰۲/۶۳	۳۴
۷	۱	۳/۲۳	۱۲۴/۱۱	۲۸	۹	۱	۱۰۰۸/۷۰	۷۱۹/۴۴	۳۸
	۲	۲۴۲/۳۸	۱۷۴/۷۷	۲۸		۲	-	۱۱۹۸/۳۳	۳۸
	۳	۲۵۷/۸۴	۱۰۸۶/۷۰	۲۷		۳	۴۴۵/۱۳	۴۹۴/۲۵	۳۷
	۴	۱۱۷۷/۳۸	۱۱۹۱/۴۷	۲۸		۴	۱۰۶۵/۸۱	۴۷۹/۸۳	۳۷
	۵	۱۰۸۹/۸۶	۲۶۱/۴۲	۲۷		۵	-	۱۱۰۷/۵۶	۳۴
	۶	۳۹۰/۷۵	۱۴/۶۴	۲۷		۶	-	۲۸۸/۱۴	۳۶
	۷	۱۶۲/۲۷	۴۴/۸۶	۲۸		۷	۸۰/۱۱	۲۴۷/۷۶	۳۷
	۸	۶۴۳/۳۱	۱۱۱۵/۲۳	۲۸		۸	۵۷/۴۵	۸۶۸/۶۶	۳۷
	۹	۶۰۴/۴۵	۵۹/۶۷	۲۷		۹	۳۰۳/۹۱	۱۰۸/۷۰	۳۶
	۱۰	-	۸۸/۸۶	۲۸		۱۰	۱۱۵۶/۹۵	-	۳۷

از بین ۱۲۰ نمونه ارزیابی‌شده HI Base-10 در ۴۵ نمونه و HI Base-10 اصلاح‌شده در ۷۵ نمونه سریع‌تر از دیگری به جواب بهینه رسیده‌اند.

۵ نتیجه‌گیری

مساله *PPH* به دلیل کاربرد فراوانش در تشخیص و درمان بسیاری از بیماری‌ها از اهمیت ویژه‌ای برخوردار است. روش‌های دقیقی که به حل این مساله می‌پردازند، اغلب مساله را به صورت یک مدل برنامه‌ریزی عدد صحیح فرمول‌بندی می‌کنند که حل آن ها برای نمونه‌های بزرگ با دشواری‌هایی مواجه است. ما اخیرا در مقاله‌ای با تخصیص مقادیر عددی به هاپلوتایپ‌ها و ژنوتایپ‌ها یک مدل عدد صحیح آمیخته به نام HI Base-10 برای *PPH* ارایه کردیم که هیچ متغیر و قیدی متناظر جایگاه‌های هتروزیگوت ندارد. در این مقاله با تخصیص مقادیر عددی به ژنوتایپ‌ها با شیوه‌ای متفاوت مدل دودویی آمیخته جدیدی ارایه کردیم که بر خلاف مدل قبلی هیچ متغیر و قیدی متناظر جایگاه‌های هموزیگوت ندارد. به طور کلی در نمونه‌هایی که درصد جایگاه‌های هتروزیگوت پایین است، مدل HI Base-10 اصلاح‌شده و در سایر نمونه‌ها مدل HI Base-10 سریع‌تر به

مقدار بهینه می‌رسد؛ بنابراین می‌توان قبل از حل PPH روی یک نمونه، ابتدا درصد جایگاه‌های هتروزیگوت آن‌را مورد ارزیابی قرار داده و براساس آن مدل مناسب را به کار برد.

منابع

[1] Cargill, M., et al. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature genetics 22(3), 231-238.
[2] Li, W.-H. and L. A. Sadler (1991). Low nucleotide diversity in man. Genetics 129(2), 513-523.
[3] Wang, D. G., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280(5366), 1077-1082.
[4] Halushka, M. K., et al. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nature genetics 22(3), 239-247.
[5] Zhang, X.-S., et al. (2006). Models and algorithms for haplotyping problem. Current Bioinformatics 1(1), 105-114.
[6] Catanzaro, D. and M. Labbé (2009). The pure parsimony haplotyping problem, Overview and computational advances. International Transactions in Operational Research 16(5), 561-584.
[7] Bell, G. I., et al. (1984). A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes melliitus. Diabetes 33(2), 176-183.
[8] Dorman, J. S., et al. (1990). Worldwide differences in the incidence of type I diabetes are associated with amino acid variation at position 57 of the HLA-DQ beta chain. Proceedings of the National Academy of Sciences 87(19), 7370-7374.
[9] Nisticò, L., et al. (1996). The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Human molecular genetics 5(7), 1075-1080.
[10] Deeb, S. S., et al. (1998). A Pro12Ala substitution in PPARγ2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. Nature genetics 20(3), 284-287.
[11] Altshuler, D., et al. (2000). The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nature genetics 26(1), 76-80.
[12] Strittmatter, W. J. and A. D. Roses (1996). Apolipoprotein E and Alzheimer's disease. Annual review of neuroscience 19(1), 53-77.
[13] Chapuis, J., et al. (2009). Transcriptomic and genetic studies identify IL-33 as a candidate gene for Alzheimer's disease. Molecular psychiatry 14(11), 1004-1016.
[14] Gretarsdottir, S., et al. (2003). The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. Nature genetics 35(2), 131-138.
[15] Van Eerdewegh, P., et al. (2002). Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. Nature 418(6896), 426-430.
[16] Trégouët, D.-A., et al. (2009). Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nature genetics 41(3), 283-285.
[17] Lancia, G., et al. (2004). Haplotyping populations by pure parsimony: Complexity of exact and approximation algorithms. INFORMS Journal on computing 16(4), 348-359.
[18] Gusfield, D. (2003). Haplotype inference by pure parsimony. Annual Symposium on Combinatorial Pattern Matching, Springer.
[19] Wang, L. and Y. Xu (2003). Haplotype inference by maximum parsimony. Bioinformatics 19(14), 1773-1780.
[20] Godi, A., et al. (2004). Haplotype inference by parsimony for large datasets, Technical Report 616, IASI, Istituto di Analisi dei Sistemi ed Informatica–CNR, Rome.
[21] Huang, Y.-T., et al. (2005). An approximation algorithm for haplotype inference by maximum parsimony. Journal of computational biology 12(10), 1261-1274.
[22] Kalpakis, K. and P. Namjoshi (2005). Haplotype phasing using semidefinite programming. Bioinformatics and Bioengineering, 2005. BIBE 2005. Fifth IEEE Symposium on, IEEE.
[23] Li, Z., et al. (2005). A parsimonious tree-grow method for haplotype inference. Bioinformatics 21(17), 3475-3481.
[24] Wang, R.-S., et al. (2005). Haplotype inference by pure parsimony via genetic algorithm. Operations Research and Its Applications, the Fifth International Symposium (ISORA'05), Tibet, China, August.
[25] Lancia, G. and R. Rizzi (2006). A polynomial case of the parsimony haplotyping problem. Operations Research Letters 34(3), 289-295.

[26] Di Gaspero, L. and A. Roli (2008). Stochastic local search for large-scale instances of the haplotype inference problem by pure parsimony. Journal of Algorithms 63(1-3), 55-69.
[27] Do, D. D., et al. (2013). ACOHAP, an efficient ant colony optimization for the haplotype inference by pure parsimony problem. Swarm Intelligence 7(1), 63-77.
[28] Wei, B. and J. Zhao (2014). Haplotype inference using a novel binary particle swarm optimization algorithm. Applied Soft Computing 21, 415-422.
[29] Lancia, G. and P. Serafini (2009). A set-covering approach with column generation for parsimony haplotyping. INFORMS Journal on computing 21(1), 151-166.
[30] Halldórsson, B. V., et al. (2003). Combinatorial problems arising in SNP and haplotype analysis. Discrete Mathematics and Theoretical Computer Science, Springer**,** 26-47.
[31] Brown, D. G. and I. M. Harrower (2006). Integer programming approaches to haplotype inference by pure parsimony. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 3(2), 141-154.
[32] Catanzaro, D., Godi, A., Labbe´ , M., 2007. A class representative model for pure parsimony haplotyping. *Technical Report, G.O.M. – Computer Science Department – Universite´ Libre de Bruxelles (U.L.B.).*
[33] Bertolazzi, P., et al. (2008). Solving haplotyping inference parsimony problem using a new basic polynomial formulation. Computers & Mathematics with Applications 55(5), 900-911.
[34] Dal Sasso, V., et al. (2016). A Column Generation Approach for Pure Parsimony Haplotyping. OASIcs-OpenAccess Series in Informatics, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
[35] Jäger, G., et al. (2016). The complete parsimony haplotype inference problem and algorithms based on integer programming, branch-and-bound and Boolean satisfiability. Journal of Discrete Algorithms 37, 68-83.
[36] Feizabadi, R., et al. (2016). A new mathematical modeling for pure parsimony haplotyping problem. Mathematical Biosciences 281, 92-97.