

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

رضا فیض‌آبادی<sup>۱</sup>، مهری باقریان<sup>۲\*</sup>

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 هاپلوتاپ‌ها از ژنوتایپ‌ها یکی از مسائل مهم حوزه ریاضیات زیستی است. اهمیت این مساله بهدلیل کاربردهای فراوان آن در تشخیص و درمان بیماری‌های ژنتیکی همچون دیابت، آلتزیمر و امراض قلبی است که موجب رقابت پژوهشگران در ارایه مدل‌های ریاضی بهتر و طراحی الگوریتم‌های کارتر 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 هاپلوتاپ‌ها تحت معیارهای متفاوتی بیان می‌شود. پارسیمونی یکی از مهم‌ترین آن هاست و در این مقاله مساله با این معیار مورد بررسی قرار گرفته است. روش‌های حل مساله استنباط هاپلوتاپ‌ها از ژنوتایپ‌ها با معیار پارسیمونی به دو دسته دقیق و تقریبی تقسیم می‌شود. اغلب روش‌های دقیق مساله را به صورت یک مساله برنامه‌ریزی با اعداد صحیح فرمولبندی می‌کنند. اخیرا در مقاله‌ای یک مدل دقیق به نام 10-HI Base ارایه شده که ابتدا به هر هاپلوتاپ و ژنوتایپ یک عدد متناظر کرده و سپس مدل را بر اساس این اعداد تشکیل می‌دهد که در آن 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 10-HI Base 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:** بیانفورماتیک، استنباط هاپلوتاپ‌ها، مدل عدد صحیح، پارسیمونی، ژنوتایپ

\* عهده‌دار مکاتبات

آدرس الکترونیکی: mbagherian@guilan.ac.ir

## ۱ مقدمه

انسان یک موجود دیپلولئید است؛ یعنی DNA انسان به ۲۳ جفت کروموزوم تقسیم می‌شود. هر زوج کروموزوم نسخه یکسانی از کروموزوم‌هایی است که از هر یک از والدین به ارث رسیده است. کروموزوم شامل مولکول DNA 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 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 اختلاف دار چندربختی تک نوکلئوتیدی<sup>۱</sup> یا به اختصار اسنیپ (SNP) گفته می‌شود. تقریباً همه اسنیپ‌ها دو آللی<sup>۲</sup> هستند؛ یعنی تنها دو نوکلئوتید از چهار حالت ممکن در یک جایگاه مشاهده می‌شود. به آلل‌های مشاهده شده با تکرار بالا در جمعیت، نوع غالب<sup>۳</sup> گفته می‌شود [۵] و این جایگاه‌ها به کدگذاری می‌شود. آلل‌های با تکرار کم به ۱ کدگذاری می‌شود و به آن‌ها نوع مغلوب<sup>۴</sup> گفته می‌شود [۵]. با صرف نظر کردن از جایگاه‌های یکسان و پشت سر هم قراردادن اسنیپ‌ها دنباله‌ای از ارقام ۰ و ۱ حاصل می‌شود که به آن هاپلوتاپ 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 شده است. با حذف جایگاه‌های حاوی نوکلئوتید یکسان و پشت هم ردیف کردن اسنیپ‌ها در شکل ۱.ب هاپلوتاپ‌ها ایجاد و در شکل ۱.پ کدگذاری شده است.

مقایسه هاپلوتاپ‌ها بین اشخاص سالم و بیمار به تشخیص و درمان بسیاری از بیماری‌ها کمک می‌کند. تخمین هاپلوتاپ‌ها در زمینه‌های مختلفی همچون طراحی دارو، عملکرد ژن‌ها، نقشه‌ی ژنی بیماری‌های پیچیده و غیره کمک می‌کند [۶]. برخی از مطالعاتی که با استفاده از هاپلوتاپ‌ها 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 هاپلوتاپ‌ها با روش‌های آزمایشگاهی بسیار گران و زمانبر است؛ اما ژنوتایپ‌ها که در واقع ترکیب هاپلوتاپ‌های پدرانه و مادرانه هستند از روش‌های آزمایشگاهی با سرعت و با هزینه کم قابل دستیابی‌اند. در آزمایشگاه برای هر فرد تنها یک ژنوتایپ به دست می‌آید که رشته‌ای به طول هاپلوتاپ‌هاست و اطلاعات آن بدین صورت است که اگر جایگاهی روی هر دو هاپلوتاپ‌های 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 هاپلوتاپ‌ها دارای یک نوکلئوتید و روی هاپلوتاپ دیگر دارای نوکلئوتید دیگری باشد، به جایگاه متناظر روی ژنوتایپ مقدار ۲

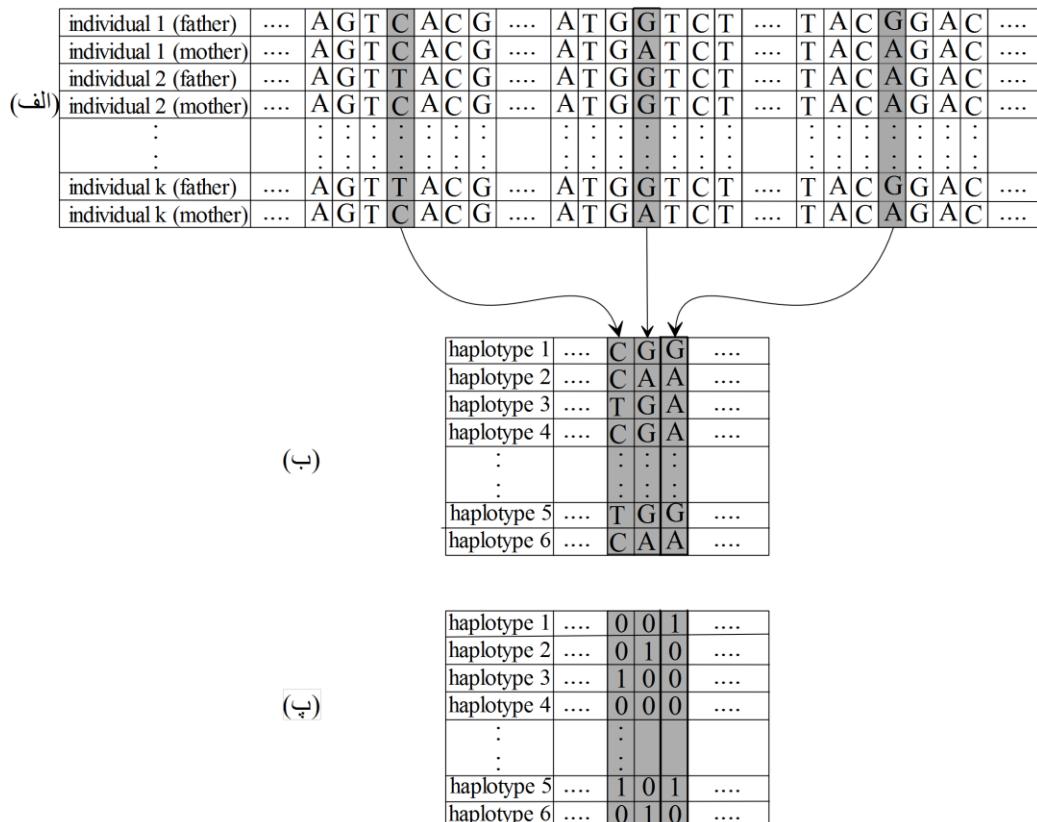
<sup>1</sup> Single Nucleotide Polymorphism (SNP)

<sup>2</sup> biallelic

<sup>3</sup> major

<sup>4</sup> minor

اختصاص می‌یابد. چنین جایگاهی هتروزیگوت نامیده می‌شود. در حالت 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.



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

به دست آوردن ژنتوتایپ‌ها در آزمایشگاه آسان و کم‌هزینه است؛ اما آن چه در کاربرد ارزشمند است، هاپلوتایپ‌ها هستند. خوشبختانه محاسبه هاپلوتایپ‌ها از روی ژنتوتایپ‌ها توسط روش‌های محاسباتی امکان‌پذیر است. به محاسبه هاپلوتایپ‌ها از روی ژنتوتایپ‌ها، استنباط هاپلوتایپ‌ها<sup>۱</sup> گفته می‌شود[۱۷]. مشکل این محاسبه در تعداد بسیار زیاد زوج هاپلوتایپ‌هایی است که می‌توانند هاپلوتایپ‌های اولیه تعدادی ژنتوتایپ باشند. این تعداد به طور دقیق  $2^{\alpha-1}$  زوج برای ژنتوتایپی با  $\alpha$  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<sup>۲</sup> و بیشترین درستنمایی<sup>۳</sup> 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 هاپلوتاپ‌هایی که قابلیت تولید ژنوتایپ‌ها را داشته باشند، باید کمترین تعداد عضو ممکن را داشته باشند. مساله تحت این معیار  $PPH$  نامیده شده است.

مساله استباط هاپلوتاپ‌ها از روی ژنوتایپ‌ها با معیار پارسیمونی NP-hard است [۱۷، ۱۸]؛ بنابراین روش‌های تقریبی زیادی برای حل این مساله ارایه شده است که برخی از آن‌ها عبارت است از [۱۹، ۲۸].

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 مرتبه 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.

اخیراً در مقاله‌ای در ۲۰۱۶ یک مدل عدد صحیح آمیخته چندجمله‌ای به نام 10- HI Base-  $PPH$  ارایه شده که هیچ متغیر و قیدی متناظر جایگاه‌های هتروزیگوت ندارد [۳۶]. در این مقاله مدل‌بندی دیگری برای این مساله ارایه می‌کنیم که برخلاف مدل قبلی هیچ متغیر و قیدی متناظر جایگاه‌های هموزیگوت ندارد. همان طور که نشان خواهیم داد، با توجه به اینکه معمولاً 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 10- HI Base-  $PPH$  تشریح می‌شود. در بخش سوم مدل جدید 10- HI Base-  $PPH$  اصلاح شده را می‌آوریم. بخش چهارم به نتایج محاسباتی بر روی نمونه‌های شبیه‌سازی شده و مقایسه 10- HI Base-  $PPH$  و 10- HI Base- 10-  $PPH$  اصلاح شده اختصاص دارد. در انتها بخش پنجم، نتیجه گیری را ارایه می‌کنند.

## ۲ مدل 10- HI Base-

$n$  ژنوتایپ داده شده است که هر کدام از  $m$  جایگاه شامل اعداد  $0$ ،  $1$  و  $2$  تشکیل شده اند. جایگاه‌ها را از راست به چپ از  $1$  تا  $m$  اندیس‌گذاری می‌کنیم و  $s$  امین جایگاه ژنوتایپ  $t$  ام را با  $(s)_t$  نمایش می‌دهیم. هاپلوتاپ‌ها نیز رشته‌هایی به طول  $m$  هستند که هر جایگاه‌شان یکی از اعداد  $0$  و  $1$  را شامل می‌شود. اندیس‌گذاری جایگاه‌های هاپلوتاپ‌ها که آن‌ها را با حرف  $h$  نمایش می‌دهیم با ژنوتایپ‌ها یکسان است. در صورتی که رابطه‌ی زیر بین هاپلوتاپ‌های  $h_i$  و  $h_j$  و ژنوتایپ  $g$  برقرار باشد  $h_i \oplus h_j = g$  نمایش می‌دهیم.

$$g(s) = \begin{cases} 0 & h_i(s) = h_j(s) = 0 \\ 1 & h_i(s) = h_j(s) = 1 \\ 2 & \text{o.w.} \end{cases}$$

برای مثال  $h_1 = 101$  و  $h_2 = 100$  ها پلوتایپ‌های مولد  $g = 201$  هستند؛ یعنی  $101 \oplus 100 = 201$  است.  
هایپلوتایپ‌ها و ژنوتایپ‌ها را با تعریف فوق هایپلوتایپ و ژنوتایپ معمولی می‌نامیم و با اعمال تبدیلاتی بر روی آن‌ها انواع دیگری را از هایپلوتایپ‌ها و ژنوتایپ‌ها به دست می‌آوریم که در زیر به تعریف آن‌ها و تشریح هر کدام با استفاده از هایپلوتایپ معمولی  $h = 101210$  و ژنوتایپ معمولی  $g = 201210$  می‌پردازیم.

ژنوتایپ حسابی: از جایگزین کردن مقادیر ۱ با ۲ و بالعکس در ژنوتایپ معمولی حاصل می‌شود. در ژنوتایپ معمولی نمونه،  $g = 201210$ ، جایگاه‌های دوم و چهارم از سمت راست حاوی مقدار ۱ هستند که مقدار آن‌ها 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  $\hat{g} = 102120$  است.

ژنوتایپ دودویی: با جایگزین کردن هر مقدار ۲ با جمع دو مقدار ۱ در ژنوتایپ حسابی و به طور همزمان در نظر گرفتن ژنوتایپ به صورت عددی در مبنای ۲ حاصل می‌شود. به عبارتی چنانچه جایگاه ۳ام ژنوتایپ حسابی مقدار ۲ داشته باشد، مقدار آن را به ۰ تغییر داده و مقدار ۱ به جایگاه  $+1$  این ژنوتایپ اضافه می‌کنیم؛ بنابراین دنباله حاصل فقط شامل اعداد ۰ و ۱ خواهد بود. ژنوتایپ دودویی متناظر با ژنوتایپ حسابی  $\hat{g} = 102120$  برابر  $g = 11100$  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 قبلی اش حاوی مجموع سه مقدار ۱ است که یکی از آن‌ها 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 ۰ حاصل می‌شود؛ بنابراین برای محاسبه ژنوتایپ عددی از  $\hat{g}$  آن را به صورت عددی در مبنای ۲ در نظر می‌گیریم، سپس این عدد را به مبنای ۰ انتقال می‌دهیم. واضح است که برای این انتقال باید جایگاه ۳ام ژنوتایپ دودویی را در  $z^j$  ضرب کرده و مقادیر حاصل را با هم جمع کنیم؛ لذا مقدار ژنوتایپ عددی برای نمونه تحت بررسی به صورت زیر به دست می‌آید:

$$z = 0 \times z^0 + 0 \times z^1 + 0 \times z^2 + 1 \times z^3 + 1 \times z^4 + 1 \times z^5 = 56$$

هاپلوتاپ عددی: با درنظرگرفتن هر هاپلوتاپ معمولی به صورت عددی در مبنای ۲ و انتقال آن به مبنای حاصل می‌شود. مقدار هاپلوتاپ عددی متناظر با هاپلوتاپ نمونه،  $h = 10101$  به صورت زیر به دست می‌آید:

$$z = 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 21$$

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

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

قضیه ۱. [۳۶] ژنوتایپ عددی  $z$  متناظر با ژنوتایپ معمولی  $g$  مفروض است. اگر هاپلوتاپ‌های عددی  $h_1$  و  $h_2$  در شرایط (۱)–(۵) صدق کنند، آنگاه هاپلوتاپ‌های معمولی مولدشان ژنوتایپ  $g$  را تولید می‌کنند.

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

اندیس ژنوتایپ	ژنوتایپ معمولی	ژنوتایپ حسابی	ژنوتایپ دودویی	ژنوتایپ عددی
۱	۱۲۲۱	۲۱۱۲	۱۱۰۰	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 هاپلوتاپ‌های معمولی و عددی

اندیس هاپلوتاپ	هاپلوتاپ معمولی	هاپلوتاپ عددی
$h_1$	۱۰۱۱	۱۱
$h_2$	۱۱۰۱	۱۳
$h_3$	۰۰۱۱	۳
$h_4$	۱۰۰۰	۸

$$h_1 + h_2 = z \quad (1)$$

$$h_1 = 2^p (q_{1,p} + r_{1,p}) \quad \forall p: g(p) \neq 2 \quad (2)$$

$$h_2 = 2^p (q_{2,p} + r_{2,p}) \quad \forall p: g(p) \neq 2 \quad (3)$$

$$\leq r_{i,p} < 0 / 5 \quad i = 1, 2, \quad \forall p: g(p) = 0 \quad (4)$$

$$0 / 5 \leq r_{i,p} < 1 \quad i = 1, 2, \quad \forall p: g(p) = 1 \quad (5)$$

همان‌طور که از روابط (۱) تا (۵) مشخص است برای تولید هاپلوتاپ‌های مولد یک ژنوتایپ، جایگاه‌های هتروزیگوت هیچ محدودیتی اعمال نمی‌کنند. تنها به اعمال محدودیت روی جایگاه‌های هموزیگوت (جایگاه‌هایی که  $2 \neq g$ ) نیاز داریم و این محدودیت‌ها با تقسیم هاپلوتاپ عددی بر  $2^p$  در (۲) و (۳) ساخته

می شود. با کنترل باقیمانده این تقسیم در (۴) و (۵) مقدار جایگاه هموزیگوت  $p$  ام برای هاپلوتاپ مشخص می شود. به طوری که اگر جایگاه  $p$  ام هموزیگوت از نوع غالب باشد باید باقیمانده کمتر از  $5/0$  باشد (قید ۴) و اگر هموزیگوت از نوع مغلوب باشد، باقیمانده بزرگتر یا مساوی  $5/0$  باشد (قید ۵). بنابراین مدل ارایه شده توسط روابط (۶) - (۱۴) هاپلوتاپ های عددی را محاسبه می کند که به راحتی می توان از آن ها هاپلوتاپ های معمولی را تولید کرد.

$$\begin{aligned} \text{Min} \quad & \sum_{j=1}^n x_j && (6) \\ \text{s.t.} \quad & h_{\gamma_{t-1}} + h_{\gamma_t} = z_t & t = 1, \dots, n & (7) \\ & h_{\gamma_{t-1}} - 2^{m-s+1}(q_{\gamma_{t-1},s} + r_{\gamma_{t-1},s}) = 0 & t = 1, \dots, n \quad \forall s: g_t(s) \neq 2 & (8) \\ & h_{\gamma_t} - 2^{m-s+1}(q_{\gamma_t,s} + r_{\gamma_t,s}) = 0 & t = 1, \dots, n \quad \forall s: g_t(s) \neq 2 & (9) \\ & r_{\gamma_{t-1},s}, r_{\gamma_t,s} \leq 0 / \Delta - \varepsilon & t = 1, \dots, n \quad \forall s: g_t(s) = 0 & (10) \\ & 0 / \Delta \leq r_{\gamma_{t-1},s}, r_{\gamma_t,s} \leq 1 - \varepsilon & t = 1, \dots, n \quad \forall s: g_t(s) = 1 & (11) \\ & h_i - h_j \leq (2^m - 1)y_{ij} & i, j = 1, \dots, 2n ; i \neq j & (12) \\ & x_j \geq 2n - j + 1 - \sum_{i>j} (1 - y_{ij}) - \sum_{i>j} (1 - y_{ji}) & j = 1, \dots, 2n & (13) \\ & h_j \leq 2^m - 1 & j = 1, \dots, 2n & (14) \\ & x_j, h_j, q_{j,s}, r_{j,s} \geq 0 & j = 1, \dots, 2n \quad s = 1, \dots, m & \\ & h_j, q_{j,s} \in Z, \quad r_{j,s} \in R & j = 1, \dots, 2n \quad s = 1, \dots, m & \\ & x_j, y_{ij} \in \{0, 1\} & i, j = 1, \dots, 2n & \end{aligned}$$

که در آن فرض شده نمونه تحت بررسی حاوی  $n$  ژنوتایپ است و جایگاه  $s$  ام ژنوتایپ  $t$  م با  $g_t(s)$  نمایش داده شده است. ژنوتایپ عددی  $g_t$  متاظر  $z_t$  فرض شده است. از آن جا که در بدترین حالت به  $2n$  هاپلوتاپ برای تولید  $n$  ژنوتایپ نیازمندیم،  $2n$  متغیر عدد صحیح متاظر هاپلوتاپ های عددی در نظر گرفته شده اند که مانند قبل از حرف  $h$  برای نمایش آن ها استفاده شده است و فرض شده  $h_{\gamma_t}$  و  $h_{\gamma_{t-1}}$  هاپلوتاپ های عددی مولد ژنوتایپ  $t$  م باشند. در مدل  $2n$  متغیر دودویی  $x_j$  متاظر هاپلوتاپ ها وجود دارد که مقدار ۱ در آن ها بیانگر حضور  $h_j$  در جواب بهینه و مقدار ۰ برای آن نشان از عدم حضور  $h_j$  در مجموعه هاپلوتاپ های جواب دارد. متغیرهای  $q_{j,s}$  و  $r_{j,s}$  به ترتیب حاوی مقادیر صحیح و اعشاری حاصل از تقسیم هاپلوتاپ عددی  $h_j$  بر  $2^{m-s+1}$  برای مشخص کردن مقدار جایگاه  $s$  ام در هاپلوتاپ معمولی  $j$  می باشند. قیود (۷) تا (۱۱) مشابه روابط (۱) تا (۵) تضمین می کنند که هاپلوتاپ های به دست آمده از حل مدل به درستی ژنوتایپ ها را تولید می کنند. ۶ یک عدد بسیار کوچک است و تنها برای تبدیل قیود از نوع کوچک تر اکید به کوچک تر مساوی از مقدار سمت راست قیود (۱۰) و (۱۱) کم شده است. مقدار  $2^m / 1$  برای کوچکی ۶ کفايت می کند [۳۶].

تعریف متغیرهای  $y_{i,j}$  به گونه ای 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 مینیمم سازی تابع هدف برقرار است.  $M$  یک عدد بسیار بزرگ است و  $1 - 2^m$  برای بزرگی آن کفايت می کند [۳۶]. دلیل اضافه کردن  $M$  این است که  $y_{i,j}$  ها در سمت راست دسته محدودیت های (۱۲) دودویی هستند در حالی که سمت چپ می تواند مقادیر بزرگتر از ۱

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

### ۳ مدل 10 HI Base- اصلاح شده

مدل 10 HI Base- هر هاپلوتاپ و ژنوتایپ را یک عدد در مبنای ۲ در نظر می‌گیرد. سپس این اعداد را به مبنای ۱۰ برد 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  $= g$  و شکل ۲ را متناظر آن درنظر بگیرید. همان طور که در شکل ۲ مشاهده می-

شود، مانند قبل اندیس گذاری برای جایگاه‌ها از سمت راست به چپ صورت می‌گیرد.

۲۰ ۲۱ ۲۲ ۲۳ ۲۴ ۲۵ ۲۶ ۲۷ : ارزش هر جایگاه

2	1	2	0	1	2	1	0
---	---	---	---	---	---	---	---

۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ : اندیس هر جایگاه

شکل ۲. ژنوتایپ تحت بررسی.

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

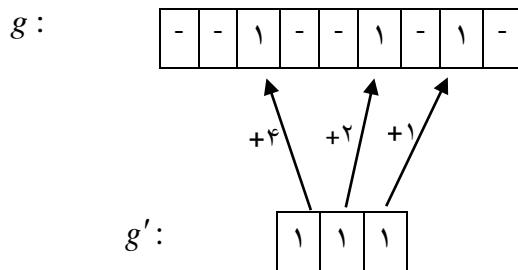
۱- جایگاه‌های هموزیگوت از نوع غالب: این جایگاه‌ها مقدار ۰ دارند و اندیس این جایگاه‌ها را در مجموعه  $Homw$  نگهداری می‌کنیم. برای ژنوتایپ تحت بررسی  $\{1,5\}$  است.

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  $\{3,6,8\}$  است.

۳- جایگاه‌های هتروزیگوت: این جایگاه‌ها مقدار ۱ دارند.

مقدار جایگاه‌های نوع اول و دوم در هاپلوتاپ‌های مولد  $g$  از قبل مشخص است. به طوری که اگر  $h_1$  و  $h_2$  هاپلوتاپ‌های 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 ۸ و پشت سر هم قراردادن جایگاه‌های هتروزیگوت ژنوتایپ  $= 111'g$  به دست می‌آید که جایگاه‌های اول، دوم و سوم آن به ترتیب متناظر جایگاه‌های دوم، چهارم و هفتم ۸ می‌باشند. این تناظر در شکل ۳ به تصویر کشیده شده است.



شکل ۳. بیان تصویری تناظر بین  $g'$  و  $g$ .

با اضافه کردن 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 مود نظر  $(1, 2, 4) = c$  است. ارزش عددی ژنوتایپ جدید  $g'$  (ژنوتایپ عددی متناظر  $g$ ) که با درنظر گرفتن  $g$  به صورت یک عدد در مبنای ۲ و تبدیل مبنای این عدد به  $10^0$  حاصل می‌شود. برابر  $2^p - 1 = 7$  است که  $p$  تعداد جایگاه‌های این ژنوتایپ و مساوی تعداد جایگاه‌های هتروزیگوت  $g$  است. بنابر آنچه در بخش قبل بیان شد، هاپلوتایپ‌های عددی  $l_1$  و  $l_2$  در صورتی می‌توانند مولد  $g'$  باشند که رابطه زیر برقرار باشد:

$$l_1 + l_2 = 2^p - 1 \quad (15)$$

برای مثال هاپلوتایپ‌های عددی  $4 = l_1$  و  $3 = l_2$  در رابطه (1) صدق کرده و می‌توانند مولد  $g'$  باشند. هاپلوتایپ‌های معمولی متناظر آن‌ها به ترتیب برابر  $100$  و  $11$  می‌باشد. واضح است که این هاپلوتایپ‌ها، مولد  $g$  نمی‌باشند؛ زیرا تنها جایگاه‌های هتروزیگوت  $g$  را دارند. ولی می‌توان از روی آن‌ها و با توجه به بردار تناظر  $c$  هاپلوتایپ‌های  $h_1$  و  $h_2$  مولد  $g$  را ساخت. به این منظور باید در هر کدام تعداد مناسبی  $0$  و  $1$  بین جایگاه‌های یاشان اضافه شود. تعداد و نوع غالب یا مغلوب بودن جایگاه‌های اضافه شده توسط بردار تناظر  $c$  و مجموعه‌های  $Homm$  و  $Homw$  مشخص می‌شود. آن‌چه مد نظر ماست، محاسبه هاپلوتایپ‌های عددی  $l_1$  و  $l_2$  متناظر با هاپلوتایپ‌های عددی  $h_1$  و  $h_2$  است. به این منظور ارزش جایگاه‌های هر دسته از  $h_i$  ( $i = 1, 2$ ) را به طور جداگانه محاسبه و سپس جمع می‌کنیم.

جایگاه‌های نوع اول، هموزیگوت غالب، که اندیس آن‌ها با  $Homw$  مشخص شده است، هیچ سهمی در مقدار هاپلوتایپ‌های عددی  $h_i$  ( $i = 1, 2$ ) ندارند و لذا هیچ نقشی در محاسبات ما ندارند.

هر جایگاه نوع دوم، هموزیگوت مغلوب، با اندیس  $j$  به مقدار  $2^{j-1}$  در مقدار هاپلوتاپ عددی  $(i=1,2)h_i$  نقش دارد؛ بنابراین این نوع جایگاه‌ها در مجموع سهمی برابر  $\sum_{j \in Homm} 2^{j-1}$  دارند. این مقدار برای مثال تحت بررسی ما برابر  $2^5 + 2^6 + 2^7 = 128$  است.

برای به دست آوردن ارزش متناظر جایگاه‌های نوع سوم؛ یعنی هتروزیگوت، ابتدا باید هاپلوتاپ‌های معمولی  $i$  ( $i=1,2$ ) را با تبدیل مبنای مقدار آن از  $10$  به  $2$  به دست آورد. این کار را می‌توان توسط قید زیر انجام داد.

$$l_i = b_{i,1} \cdot 2^0 + b_{i,2} \cdot 2^1 + b_{i,3} \cdot 2^2 + \dots + b_{i,p} \cdot 2^{p-1} \quad i=1,2 \quad (16)$$

که  $b_{i,j}$  ها متغیرهای دودویی بوده و مقدار جایگاه زام را در هاپلوتاپ معمولی  $i$  ( $i=1,2$ )  $l_i$  نمایش می‌دهند. ارزش این جایگاه برای هاپلوتاپ  $i$  ( $i=1,2$ )  $l_i$  مولد  $g'$  برابر  $2^{j-1}$  است؛ ولی ارزش این جایگاه در هاپلوتاپ  $i$  ( $i=1,2$ )  $h_i$  مولد  $g$ ، با توجه به بردار تناظر  $c$  برابر  $2^{j+c_{j-1}}$  است. در نتیجه سهم کل جایگاه‌های از

$$\text{این نوع در } h_i \text{ برابر } \sum_{j=1}^p b_{i,j} \cdot 2^{j+c_{j-1}} \text{ است.} \quad (i=1,2) \quad (17)$$

لذا مقدار هاپلوتاپ‌های عددی  $i$  ( $i=1,2$ )  $h_i$  از رابطه زیر به دست می‌آید:

$$h_i = \sum_{j \in Homm} 2^{j-1} + \sum_{j=1}^p 2^{j+c_{j-1}} b_{i,j} \quad i=1,2 \quad (17)$$

بنابراین قضیه زیر اثبات شد.

قضیه ۲. اگر  $g'$  ژنوتایپ حاصل از حذف جایگاه‌های هموزیگوت  $g$  با بردار تناظر  $c$  باشد و  $p$  تعداد جایگاه‌های هتروزیگوت  $g$  باشد. و چنانچه  $i=1, \dots, p$  ( $i=1,2, \dots, p$ )  $b_{i,j}, l_i, h_i$  در روابط (۱۵)–(۱۷) صدق کنند، آن‌گاه  $h_i$  و  $h_i$  هاپلوتاپ‌های عددی مولد  $g$  خواهند بود.

با توجه به قضیه فوق مدل طراحی شده براساس روابط (۱۸)–(۲۳) را برای مساله *PPH* ارایه می‌کنیم.

$$\text{Min} \quad \sum_{j=1}^n x_j \quad (18)$$

$$\text{s.t.} \quad l_{\gamma_{t-1}} + l_{\gamma_t} = 2^{p_t} - 1 \quad t=1, 2, \dots, n \quad (19)$$

$$l_i = \sum_{j=1}^{p_t} b_{i,j} \cdot 2^{j-1} \quad t=1, 2, \dots, n \quad i=\gamma_t - 1, \gamma_t \quad (20)$$

$$h_i = \sum_{j \in Homm_i} 2^{j-1} + \sum_{j=1}^{p_t} 2^{j+c_{i,j}-1} b_{i,j} \quad t=1, 2, \dots, n \quad i=\gamma_t - 1, \gamma_t \quad (21)$$

$$h_i - h_j \leq My_{ij} \quad i, j = 1, 2, \dots, \gamma_n \quad (22)$$

$$x_j \geq \gamma_n - j + 1 - \sum_{i>j} (1 - y_{ij}) + \sum_{i>j} (1 - y_{ji}) \quad j=1, 2, \dots, \gamma_n \quad (23)$$

$$l_i, h_i \in \mathbb{Z} \quad i=1, 2, \dots, \gamma_n$$

$$y_{ij}, x_i \in \{0, 1\} \quad i, j = 1, 2, \dots, \gamma_n$$

$$b_{i,j} \in \{0, 1\} \quad i=1, 2, \dots, \gamma_n \quad j=1, 2, \dots, p_t \quad t=1, 2, \dots, n$$

که در آن  $n$  تعداد ژنوتاپ‌ها و  $p_t$  تعداد جایگاه‌های هتروزیگوت ژنوتاپ  $t$  ام می‌باشد. تعداد  $2n$  هاپلوتاپ در نظر گرفته شده است و فرض شده است که هاپلوتاپ‌های  $l_{ij}$  و  $l_{ji}$  مولد ژنوتاپ  $g_i$  از حذف جایگاه‌های هموزیگوت  $g_j$  و پشت سرهم قراردادن جایگاه‌های هتروزیگوت آن به دست آید و متانظر آن‌ها،  $l_{ij}$  و  $h_{ij}$  هاپلوتاپ‌های مولد ژنوتاپ  $g_t$  باشند. متغیر دودویی  $b_{i,j}$  مقدار زامین جایگاه هاپلوتاپ  $l_i$  تعریف شده است.  $C$  ماتریس تناظر بین مجموعه ژنوتاپ‌های  $g_i$  و  $g_j$  است. قیود (۲۱)- (۱۹) چنانچه در بالا نشان داده شد، تضمین می‌کنند که تمام ژنوتاپ‌ها توسط هاپلوتاپ‌ها تولید شوند.

$M$  کران هاپلوتاپ‌های عددی  $h_i$  است و همان‌طور که در [۳۶] نشان داده شده است، مقدار  $1 - 3^m$  برای بزرگی آن کفايت 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 ۰ و در صورت عدم تساوی  $h_i$  و  $h_j$  یکی از  $y_{ij}$  و  $y_{ji}$  مقدار ۰ و دیگری مقدار ۱ داشته باشد. این محدودیت توسط مجموعه قیود (۲۲) و ماهیت مینیمم‌سازی مساله تضمین می‌شود. با استفاده از  $y_{ij}$  ها هر  $h_i$  با هاپلوتاپ‌های با اندیس بالاتر از خودش در مجموعه قیود (۲۳) مقایسه می‌شود. چنانچه لاقل با یکی از آن‌ها برابر باشد  $x_i$  مقدار ۰ و در غیراین صورت مقدار ۱ را اختیار می‌کند. در نهایت تعداد هاپلوتاپ‌های متمایز توسط تابع هدف شمارش شده و بر اساس معیار پارسیمونی مینیمم مقدار ممکن آن‌ها بازگردانده می‌شود.

$g_i$  ها هاپلوتاپ‌های عددی مولد  $g$  باید مقادیر عدد صحیح اختیار کنند؛ اما عدد صحیح 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  $h_i$  ها به دلیل وجود دسته قیود (۲۱) و  $x_i$  ها به دلیل دسته قیود (۲۳) و نوع مینیمم‌سازی تابع هدف برقرار است؛ بنابراین می‌توانیم این متغیرها را پیوسته تعریف کنیم.

تاكید می‌شود که در مدل جدید جایگاه‌های هاپلوتاپ‌ها و ژنوتاپ‌ها از راست به چپ اندیس گذاري شده‌اند. در صورت اندیس گذاري از چپ به راست توآن‌های متفاوتی از ۲ در مدل ظاهر خواهند شد.

با فرض اینکه تعداد ژنوتاپ‌ها  $n$  است، مدل جدید دو متغیر دودویی  $b_{i,j}$  متانظر هر جایگاه هتروزیگوت و تعداد  $2n$  متغیر دودویی  $y_{i,j}$  دارد. متغیرهای پیوسته متانظر هر ژنوتاپ عبارت از دو  $l_i$ ، 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 Base-10 HI سریع‌تر اجرا شود. جدول ۳ به مقایسه تعداد متغیرها و محدودیتها بین HI Base-10 و مدل جدید می‌پردازد که در آن  $\sum$  تعداد جایگاه‌های هموزیگوت است.

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

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

HI Base-10	HI Base-10 اصلاح شده	متغیر بازیزی
$4n^3 - 2f$	$4n^3 - 2n$	متغیر عدد صحیح
-	$2n + 2f$	متغیر پیوسته
$6n$	$2n + 2f$	محدودیت
$4n^3 + 7n$	$\leq 4n^3 + 6f + 3n$	

#### ۴ نتایج محاسباتی

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

برای پیاده‌سازی مدل‌های HI Base-10 و اصلاح شده 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 8 GB RAM اجرا شدند.

برای تولید نمونه‌ها ابتدا تعدادی بردار تصادفی دودویی تولید کردیم که هر کدام  $m$  مولفه دارد و متناظر هاپلوتاپی با  $m$  اسینپ است. سپس یکی از هاپلوتاپ‌ها را به طور تصادفی انتخاب کرده در  $k$  جایگاه تصادفی آن جهش می‌دهیم. از تلفیق دو هاپلوتاپ 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 هاپلوتاپ تولید شده است که هر کدام ۱۵ اسینپ دارند. برای گروه دوم تعداد ژنوتایپ‌ها، هاپلوتاپ‌ها و اسینپ‌ها به ترتیب برابر ۱۶، ۱۰ و ۱۸ می‌باشند و این مقادیر برای گروه سوم به ترتیب 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، احتمالاً 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 ۱۰ دسته ۱۰ تایی 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 مشاهده شده 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 بر حسب ثانیه نشان می‌دهند. همان‌طور 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 جایگاه‌های هموزیگوت و متغیرهای مدل HI Base-10 اصلاح شده متناظر جایگاه‌های هتروزیگوت هستند، در نمونه‌های ابتدای جدول با درصد هتروزیگوت کمتر از 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 حداقل ۵ جایگاه هتروزیگوت در ژنوتایپ‌ها) HI Base-10 اصلاح شده در ۸۰ درصد نمونه‌ها سریع‌تر از HI Base-10 به جواب بهینه رسیده است در حالی که با افزایش درصد جایگاه‌های هتروزیگوت در دسته‌های دیگر این روند به تدریج تغییر می‌کند تا جایی که در دسته آخر جدول ۴ (حاوی نمونه‌هایی با حداقل ۸ جایگاه هتروزیگوت در ژنوتایپ‌ها) HI Base-10 در نیمی از نمونه‌ها سریع‌تر به جواب بهینه دست یافته است. روند نسبتاً مشابهی در جداول ۵ و ۶ دیده می‌شود.

جدول ۴. مقایسه عملکرد بین ۱۰ و ۱۰-۱۱ اصلاح شده روی نمونه‌هایی با ۱۴ ژنوتایپ، ۸ هاپلوتایپ و ۱۵ استایپ.

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	۴/۱۹	۰/۶۱	۲۳		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	۷	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		۱	۵/۰۳	۸/۵۲	۴۱
	۲	۷/۰۵	۱۸/۱۹	۳۶		۲	۴۴/۰۵	۵۱/۸۳	۴۰
	۳	۱/۹۹	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	۸	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 ۱۰ HI Base و ۱۰ HI اصلاح شده روی نمونهای با ۱۶ ژنوتایپ، ۱۰ هاپلوتاپ 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	۵۱/۴۴	۲۰		۱	۱۴/۳۳	۱۶۸/۳۸	۳۲
	۲	۶/۹۱	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	۱۲۷/۷۷	۲۸
	۶	۳۳۶/۲۰	۲۹/۳۳	۲۱	V	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		۱	۱۰۷/۸۳	۳۹/۹۵	۳۵
	۲	۲/۷۷	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	A	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	-	۹۳/۱۷	۲۶		۹	۱۰۹/۰۰	۹۴/۴۲	۳۶
	۱۰	۵/۰۲	۲/۶۱	۲۶		۱۰	۸/۳۸	۱۱۱/۱۷	۳۷

جدول ۶. مقایسه عملکرد بین HI Base-10 و HI Base-10 اصلاح شده روی نمونهای با ۱۹ ژنوتایپ، ۱۲ هاپلوتاپ 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 10 Time	Hetero percent
۶	۱	۱۱۱۸/۲۳	۸۶۷/۴۰	۲۶		۱	۱۱۵۵/۹۸	۱۰۹۵/۴۴	۳۳
	۲	۱۹۵/۵۸	۸۵/۸۹	۲۳		۲	۱۱۴۵/۸۱	۲۲۶/۱۶	۳۲
	۳	۲۴۱/۷۳	۱۱۴۸/۵۸	۲۶		۳	۱۱۸۹/۶۱	۲۹/۴۱	۲۶
	۴	۸۲۰/۸۳	۴۵/۰۵	۲۳		۴	۳۷۲/۳۸	۴۹۶/۹۷	۳۳
	۵	۷/۷۴	۸۷/۸۳	۲۲		۵	۲۵۵/۰۸	۲۹۹/۷۲	۳۲
	۶	۵۳۹/۸۸	۸۰/۲۲	۲۲	۸	۶	۱۱۵/۸۶	۱۱۷/۵۳	۳۶
	۷	۴۰۰/۵۹	۱۲۱/۲۷	۲۳		۷	۵۹۸/۲۷	۱۳۵/۲۳	۳۲
	۸	۲۱۵/۴۷	۲۰۲/۰۸	۲۶		۸	-	۸۷/۸۶	۳۶
	۹	-	۱۸۰/۹۵	۲۳		۹	۱۶۵/۳۳	۱۶/۱۱	۳۶
	۱۰	۲۰/۳۸	۸/۴۲	۲۶		۱۰	۳۷۷/۵۲	۲۰۲/۶۳	۳۶
۷	۱	۳/۲۳	۱۲۴/۱۱	۲۸		۱	۱۰۰۸/۷۰	۷۱۹/۴۴	۳۸
	۲	۲۴۲/۳۸	۱۷۴/۷۷	۲۸		۲	-	۱۱۹۸/۴۳	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	-	۱۱۷/۵۶	۳۶
	۶	۳۹۰/۷۵	۱۶/۶۴	۲۷	۹	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 ارزیابی شده HI Base-10 در ۴۵ نمونه و HI Base-10 اصلاح شده در ۷۵ نمونه سریع تر از دیگری به جواب بهینه رسیده اند.

## 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 می‌پردازنند، اغلب مساله را به صورت یک مدل برنامه‌ریزی عدد صحیح فرمول‌بندی می‌کنند که حل آن‌ها برای نمونه‌های بزرگ با دشواری‌هایی مواجه است. ما اخیراً در مقاله‌ای با تخصیص مقادیر عددی به هاپلوتاپ‌ها و ژنوتایپ‌ها یک مدل عدد صحیح آمیخته به نام HI Base-10 *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 اصلاح شده و در سایر نمونه‌ها مدل HI Base-10 سریع تر به

مقدار بهینه می‌رسد؛ بنابراین می‌توان قبل از حل *PPH* روی یک نمونه، ابتدا درصد جایگاه‌های هتروزیگوت آن را 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. Labb   (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 mellitus." *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] Nistic  , 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 PPAR  2 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 PPAR   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 Erdewegh, P., et al. (2002). "Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness." *Nature* **418**(6896): 426-430.
- [16] Tr  gou  t, 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.