



انتخاب SNP های شاخص با استفاده از الگوریتم ژنتیک

قاسم مهدور^{۱*}، هایده اهرابیان^{۲،۱*}، مهدی صادقی^{۳،۲}، عباس نوزری^{۲،۱}

۱- قطب بیومت، دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران ۲- مرکز تحقیقات بیوشیمی بیوفیزیک، دانشگاه تهران ۳- پژوهشگاه مهندسی ژنتیک و زیست فن آوری ۴- پژوهشگاه علوم کامپیوتر، پژوهشگاه دانش های بنیادی

امروزه یکی از مهمترین مسایل مطرح در ژنتیک جمعیت پیش بینی استعداد مبتلا شدن به یک بیماری خاص است. چند ریختی های تک نوکلئوتیدی (SNP - Single Nucleotide Polymorphisms)، به عنوان نشانگر، کمک شایانی به حل این مساله نموده اند، زیرا به راحتی می توان یک بیماری را به حالتی خاص از آنها منتسب نمود. این برتری SNP ها به دلیل فراوانی آنها در طول ژنوم و بنابراین در هر ژن مشکوک به بیماری را بودن می باشد. از طرفی دیگر این فراوانی باعث غیر کاربردی شدن آنها نیز شده است، زیرا بدست آوردن مقدار SNP هزینه بالای دارد و با زیاد شدن تعداد SNP تحلیل اطلاعات آنها دشوار می شود. البته می توان با انتخاب تعداد اندکی SNP به گونه ای که از کیفیت اطلاعات کاسته نشود هر دو نقص را مرتفع نمود. به SNP های که انتخاب می شوند SNP های شاخص و به این پروسه، پروسه انتخاب SNP های شاخص می گویند. به هر حال نشان داده شده است که این مساله خود غیر قابل حل در زمان چند جمله ای است. در این مقاله ما روشی جدید برای حل این مساله طرح و برای نشان دادن قدرت روش آن را با برخی روش های معروف مقایسه می نمایم.

مقدمه

چند ریختی های تک نوکلئوتیدی (SNP ها) موقعیت های در توالی DNA هستند که در گذشته جهشی در آنها رخ داده و به ارث رسیده است. اکثر SNP ها دو مقدار هستند، یعنی فقط دو باز از چهار باز ممکن در هر یک از آنها دیده می شود و حدوداً هشت میلیون از آنها در ژنوم انسان موجود است. مقدار SNP های که در فاصله کمی از هم قرار گرفته اند معمولاً به هم وابسته می باشند و بنابراین تعداد ترکیبات متفاوت از SNP های مجاور، هپلوتیپ ها (Haplotype)، کسری کوچک از کل تعداد حالات ممکن است. لذا یافتن زیر مجموعه ای از SNP به گونه ای که امکان تعیین مقدار تمام SNP ها فراهم شود راه حل مشکل پیش گفته است. اگر بتوان یک چنین زیر مجموعه ای را یافت، می توان هر هپلوتیپ را با داشتن تعداد اندکی SNP به صورت منحصر به فرد مشخص نمود و در ادامه با داشتن اطلاعات کامل هپلوتیپ ها سعی در یافتن علل بوجود آورنده بیماری ها نمود. در شکل 1 نمونه ای واقعی از مساله و حل بهینه آن نشان داده شده است.

	SNP ₁	SNP ₂	SNP ₃	SNP ₄	SNP ₅	تعداد
H ₁	1	1	0	0	1	46
H ₂	1	0	0	1	0	13
H ₃	0	1	1	1	0	52
H ₄	0	1	1	0	0	9

شکل 1: چهار گونه هپلوتیپ با پنج SNP مشاهده شده در کروموزوم اول 120 فرد اروپایی؛ رابطه مکملی بین SNP₃ و SNP₁ وجود دارد، همچنین SNP₄ = SNP₂ ⊕ SNP₁ و SNP₅ = SNP₂ × SNP₁. بنابراین داشتن مقدار دو SNP₁ و SNP₂ برای ساخت هر یک از هپلوتیپ ها کافی است.

تا کنون محققین تلاش زیادی برای یافتن کوچک ترین مجموعه SNP های شاخص نموده اند. برای نمونه کلیتون (2) با آزمایش تمامی زیر مجموعه های ممکن سعی در یافتن مجموعه بهینه، مجموعه ای که تمامی گوناگونی ها را در بر گیرد، نمود - البته با توجه به پیچیدگی نمایی این کار، کاربرد این روش به مسایلی با اندازه کوچک محدود شده است. کارلسون و همکاران (1) با استفاده از روش حریمانه کوچکترین زیر مجموعه ای را که هر یک از SNP های خارج از آن مجموعه وابستگی آماری قابل قبولی با اعضای آن دارند را یافتند - البته با توجه به حریمانه بودن روش آنها یافتن حل بهینه تضمین نشده است. در این مقاله ما الگوریتمی ژنتیک برای مساله انتخاب SNP های شاخص، با نام GTAGER طراحی خواهیم نمود.



و در انتها نیز GTAGER و چند روش دیگر را روی داده‌های شبیه سازی شده و داده‌های زیستی اجرا خواهیم نمود، نتیجه این آزمایشات کارایی روش ما را اثبات خواهد نمود.

الگوریتم‌های ژنتیک در واقع بر پایه نظریه تکامل داروین بنا شده‌اند. در یک الگوریتم ژنتیک، ما با یک مجموعه (یا جمعیت) اولیه از جواب‌ها (یا افراد) شروع می‌کنیم. در ادامه با این امید که مطلوبیت جمعیت جدید بیشتر از جمعیت فعلی خواهد شد، افراد جمعیت فعلی (والدین) با توجه به برازندگی که دارند با هم زوج می‌شوند و جواب‌های جدید (فرزندان) را، با اعمال باز ترکیب تولید می‌کنند. معمولاً بعد از اینکه فرزندان تولید شدند جهش، که به سادگی اعمال تغییری کوچک است، رخ می‌دهد. با تکرار این مراحل الگوریتم ژنتیک ادامه خواهد یافت و در انتها نیز بهترین فرد ایجاد شده به عنوان جواب بازگردانده می‌شود.

مواد و روش

برای انجام آزمایشات ما از داده‌های کاملاً تصادفی، داده‌های ساخته شده با استفاده از نرم افزار MS (4) و داده‌های کاملاً زیستی (3) استفاده نموده‌ایم. مراحل محاسباتی GTAGER در شکل 2 نشان داده شده است.

GTAGER (matrix \mathbf{H} ; integer $p_f, p_m, p_r, P_{Size}, N_G$)

Start:

- Optimizing:** Compress the dataset;
- Initializing:** Generate a random population \mathbf{P} ;
- Evaluation:** Evaluate fitness of each individual in \mathbf{P} .

Evolution: Repeat following steps N_G times:

Generation:

- Selection:** Select two parent from population \mathbf{P} according to their fitness;
- Recombination:** With a recombination probability, p_r , combine the parents to form a new children, then add it in to \mathbf{P} ;
- Mutation:** With a mutation probability, p_m , mutate individuals;
- Refining:** With a refining probability, p_f , refine children;
- Evaluation:** Evaluate fitness of each individual in \mathbf{P} ;
- Survival:** Let parents and children compete for survive to next generation

End: Return the best individual in \mathbf{P} .

شکل 2: مراحل محاسباتی الگوریتم GTAGER

در ابتدای الگوریتم، صرفاً برای بالا بردن سرعت و بدون انجام محاسباتی خاص می‌توان تعداد ستون‌های ماتریس ورودی (=تعداد SNP ها) و تعداد سطرهای آن (=تعداد هپلوتیپ‌ها) را کاهش داد. برای انجام این کار ما از دو SNP هم معنی، دو SNP ی که در تمام هپلوتیپ‌ها با هم برابرند یا تعیض هم هستند یکی را حذف می‌کنیم. با استفاده از تعریفی مشابه می‌توان هپلوتیپ‌های هم معنی را نیز حذف نمود. اعمال همین روش ساده روی ژن TLR7 حجم محاسبات را تا 40٪ کاهش می‌دهد.

هر یک از افراد جمعیت ژنتیک GTAGER برداری دودویی به طول تعداد SNP ها، n است: یک بودن هر درایه به معنی شاخص بودن آن می‌باشد. در مساله انتخاب SNP های شاخص هدف ما رسیدن به حلی با کمترین تعداد SNP شاخص است. بنابراین هر چقدر تعداد SNP های شاخص یک فرد، $N(\cdot)$ ، کمتر و تعداد SNP های قابل پیش بینی، $Q(\cdot)$ ، بیشتر باشد، برازندگی (مطلوبیت) آن فرد، $f(\cdot)$ ، بیشتر باید باشد. آزمایشات نشان داده‌اند که بهترین جمعیت اولیه، جمعیتی است که احتمال شاخص بودن هر درایه در اعضای آن $1/n$ است. واضح است که $N(X)=||X||$ همچنین تابع $Q(\cdot)$ ، با استفاده از یک



3. Hapmap Project public datasets (2007). Available at www.hapmap.org.
4. Hudson R. R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18: 337-338.

Abstract

Single nucleotide polymorphisms (SNPs) information provides valuable information on human evolutionary history and may lead us to identify genetic variants that are responsible for human complex diseases, because of their high frequency in human genome. Unfortunately, molecular haplotyping methods are costly, laborious, and time consuming; therefore, algorithms for constructing full haplotype pattern with small available data through computational methods, Tag SNP Selection Problem, are convenient and attractive. This problem has been proved to be an NP-hard problem, so heuristic methods may be useful. In this paper we design a heuristic method based on genetic algorithm to obtain good solutions within acceptable. The algorithm is tested on a variety of simulated data and biological data. In comparison with the exact algorithm, which is based on brute force approach, experiment results show that the method can obtain optimal solutions in almost all cases and runs much faster than exact algorithm when the number of SNP sites is large.

WWW.IBP.IR

iranian bioinformatics portal