WWW.IBP.IR iranian bioinformatics portal

پنجمين همايش ملي بيوتكنولوژي جمهوري اسلامي ايران

3-5 آذر ماه 1386، سالن اجلاس سران

The 5th National Biotechnology Congress of Iran

24-26 Nov, 2007, Summit Meeting Conference Hall, Tehran- Iran

الگوريتمي براي بلاك بندي هاپلوتيپ ها بر مبناي درخت هاي فيلوژنتيك كامل

اصلاحچي چنگيز [1,3,*,**]، پزشك حميد [2] ، صادقي مهدي [4]، افضلي نرجس [1]

1- دانشكده رياضي، دانشگاه شهيد بهشتي
2- گروه بيوانفورماتيك، دانشكده رياضي، آمار و علوم كامپيوتر، پرديس علوم، دانشگاه تهران
3- گروه بيوانفورماتيك، پژوهشكده علوم كامپيوتر، پژوهشگاه دانش هاي بنيادين
4- پژوهشگاه ملی مهندسي ژنتيك و زيست فناوري

چكيده فارسي:

در اين مقاله ما روشي براي بلوك بندي هاپلوتايپ هاي يك جمعيت ارائه داده ايم كه بر اساس دو پارامتر اين كار را انجام مي دهد.پارامتر اول ميزان نزديكي بلوك را به بلوك كامل نشان مي دهد و پارامتر دوم ميزان هاپلوتايپ هاي تكراري را معين مي كند.همچنين با گرفتن پارامتر هاي سومي را كه درصد هاپلوتايپ هايي را كه كاربر مي خواهد توسط Tag SNPs شناسايي شوند ميگيريم و روشي براي يافتن Tag SNPs ارائه مي دهيم.

كلمات كليدي:

بازتوليد هاپلوتايپ، مساله NP-سخت، الگوريتم خوشه بندي

مقدمه:

تفاوتهاي ژنتيكي بين انسان ها در نواحي كوچكي از نوكلئوتيدها به نام اسنيپ(SNP) رخ مي دهد. در واقع SNP ها مهمترين عامل بروز تفاوتهاي ژنتيكي در انسانهاي مختلف هستند و در تشخيص بيماريها و ساخت داروهاي جديد از اهميت بالا برخوردار هستند. به دلايلي كه هنوز معلوم نيست در هر موقعيت اسنيپ به طور معمول دو نوع نوكلئوتيد ظاهر ميشود. در هر موقعيت اسنيپ ميتوان علل با فراواني بيشتر را با 0 و ديگري را با 1 نشان داد.به نظر ميآيدكه انتقال كروموزومها از پدر و مادر به بچه در قطعات معيني صورت ميگيرد كه به آن بلوك گفته ميشود يك بلوك را Perfect مينناميم اگر در آن در هر موقعيت SNP بيش از يك بار جهش در طول تاريخ رخ نداده باشد.از طرفي بلوك بندي را ميتوان بر اساس پارامترهاي ديگر بنا كرد از جمل ميزان تكرار هاپلوتايپها (Common) و تحليل پيوست اسنيپها (Linkage Analysis) . ..... بلوكها دقيقاً Prefect نيستند ولي از ميزان Prefect بودن بالايي برخوردار هستند. همچنين ميزان Common بودن در هر بلوك هم ميتواند متغير باشد.در اين مقاله امكان بلوك بندي با مقادير مختلف براي اين پارامترها وجود دارد.نكته ديگر در مورد tag SNP هاست يعني زير مجموعه اي از SNP هاي يك بلوك كه به كمك آنها تمام اسنيپهاي ديگر در هاپلوتايپها قابل شناسايي باشد

مواد و روش ها:

بخش اول: براي مجموعه داده‌هاي معين كه شامل چندين هاپلوتايپ باشد. ما ابتدا هاپلوتايپ‌هاي تكراري را حذف كرده و سپس روي مجموعه‌ي جديد بلوك بندي مي‌كنيم. براي مجموعه داده‌هايي كه در آن يك سري از داده‌ها مخدوش هستند (missing data) مفهوم جديدي به‌نام Prefect تعريف كرده‌ايم كه با يكي كردن تمام هاپلوتايپ‌هايي كه در داده‌هاي غير مخدوش مشترك هستند بدست مي‌آيد. مي‌توان ثابت كرد اگر مجموعه Prefect هاي حاصل Prefect باشد مجموعه‌ي هاپلوتايپ‌هاي اوليه نيز Prefect است.ما دو پارامتر (α,σ) را كه اعدادي بين 0 و 1 هستند مي‌گيريم كه α در صد Common بودن σ ميزان نزديكي بلوك به بلوك Prefect را نشان مي‌دهد. به طوري كه با حداكثر η تغيير در ورودي‌ها بلوكPrefect شود . ηرا به صورت زير تعريف مي‌كنيم.:

$$\eta=\begin{cases}\sigma & \sigma=1\\ (a-\sigma)mn+1 & \sigma\neq 1\end{cases}$$

كه در آن m تعداد هاپلوتايپ‌ها و n تعداد اسنيپ‌هاست.همچنين بلوك بندي داراي اين خاصيت است كه $\alpha\%$ از هاپلوتايپ‌ها Common باشد. كه اين كار را به صورت حريصانه با شروع از 2 ستون اول انجام مي‌دهيم. اگر دو ستون در اين شرايط صدق كند كه به سراغ ستون بعدي مي‌رويم و گرنه ستون اول به عنوان بلوك در نظر گرفته مي‌شود. در مرحله‌ي k ام ، k+1 ستون اول را در نظر مي‌گيريم. به كمك دو روش حريصانه و برنامه سازي پويا ، تقريب مناسبي براي كمترين تعداد تغييرات لازم براي Prefect كردن بلوك بدست مي‌آوريم و سپس چك مي‌كنيم اگر در شرايط صدق كند به مرحله‌ي k+1 مي‌رويم و در غير اينصورت k ستون اول را به عنوان بلوك در نظر مي‌گيريم و بلوك‌هاي بعدي را نيز به همين شيوه بلوك بندي مي‌كنيم.

نتايج و بحث:

بعد از بلوك بعدي براي هر بلوك تقريب مناسبي براي كمترين تعداد tag SNP ها بدست مي‌آوريم. ثابت كرده‌ايم كه اگر بلوك با m هاپلوتايپ و n اسنيپ، Prefect و بدون داده‌هاي مخدوش باشد، آنگاه كمترين تعداد tag SNP ها دقيقاً برابر m-1 است كه الگوريتم‌ها آن را در زمان O(mn) مي‌يابد. در جدول1 الگوريتم خود را با پارامترهاي $(\alpha,\sigma)=(0,1)$ و $(\alpha,\sigma)=(0,0.98)$ با الگوريتم حريصانه‌ي پاتيل (Patil et al.) و الگوريتم برنامه سازي پويا ژانگ (Zhang et al) كه هر دو بر پايه‌ي Common بودن هستند مقايسه كرده‌ايم. همانطور كه مشاهده مي‌شود. تمام اين نتايج مربوط به كروموزم 21 انسان است كه از 20 هاپلوتايپ و 24407 اسنيپ تشكيل شده است كه در 4

بخش مختلف قرار دارند. پارامترهاي $(\alpha,\sigma)=(0,1)$ بيان مي‌دارد كه بلوك بايد كاملاً Perfect باشد ولي روي ميزان Common بودن قيدي نمي‌گذارد. پارامترهاي $(\alpha,\sigma)=(0,0.98)$حداكثر 2% در ايده‌ها مجاز به تغيير مي‌دانند تا بلوك Perfect گردد و دو بار روي ميزان Common بودن قيدي نمي‌گذارد. در حالت $(\alpha,\sigma)=(0,0.98)$نسبت به پارامترهاي (1،0) تعداد بلوك‌ها به طور قابل ملاحظه‌اي از 3737 به 1860 در حالت(0/98، 0) كاهش مي‌يابد كه اين مقدار در الگوريتم پاتيل برابر 4135 و در الگوريتم ژانگ 2575 مي‌باشد. در حالت $(\alpha,\sigma)=(0,1)$. 657 بلوك بيشتر از 10 اسنيپ دارند كه در حالت Co,0.98 به 744 مي‌رسد. در پاتيل اين مقدار برابر 589 و ژانگ 742 مي‌باشد. همچنين تعداد بلوك‌هاي با كمتر از 3 اسنيت در حالت (0.1) برابر1116 و در حالت (0.0/98) برابر 97 ، در پاتيل برابر2138 و در ژانگ برابر 924 مي‌باشد. متوسط تعداد هاپلوتايپ‌هاي پترن در حالت (0.1) برابر 3/49 در حالت (0.0/98) برابر 4/11 و در پاتيل برابر 2072 و در ژانگ برابر 3/05 مي‌باشد. همچنين تعداد كل tag SNP ها در حالت (0.1) برابر 11271 و در حالت (0.0/98) برابر 7873 مي‌باشد.

جدول 1.

روش	تعداد اسنيپ‌ها در بلوك	تعداد بلوك‌ها	تعداد هاپلوتايپ‌ها Pattern	تعداد tag SNP
(α,σ), (0.1)	$n>۱۰$	657	4/45	3070
	$۳\le n\le ۱۰$	1964	3/80	6539
	$n>۳$	1116	2/39	1608
	Total	3737	3/49	11217
(α,σ), (0.0/98)	$n>۱۰$	744	4/36	4077
	$۳\le n\le ۱۰$	1019	3/98	3604
	$n>۳$	97	3/27	192
	Total	1866	4/11	7873
Patil (Greedy)	$n>۱۰$	589	3/75	
	$۳\le n\le ۱۰$	1408	2/92	
	$n>۳$	2138	2/30	
	Total	4135	2/72	
Zhang	$n>۱۰$	474	4/23	
	$۳\le n\le ۱۰$	909	3/03	

(Dynamic Programming)	$n > ۳$	924	2/12	
	Total	2575	3/05	

تشكر و قدرداني: بخشي از هزينه اين كار از محل طرح شماره CS1385-1-02 پژوهشگاه دانشهاي بنيادي (IPM) تامين شده است .

منابع:

1. Chou, P.Y., Fasman, G.D. (1974). Prediction of protein conformation. Biochemistry 13: 222–245.

2. Garnier J, *et al.*. (1978). Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins. J. Mol. Biol. 120: 97-120.

3. J. Randy Macdonald, *et al.*. (2001). Environmental features are important in determining protein secondary structure. Protein science 10: 1172-1177.

4. Zhan-Yang Zhu, *et al*. (1996). The use of amino acid patterns of classified helices and strands in secondary structure prediction. J. mol. biol. 260: 261-276.

پـنـجمين هـمـايـش مـلـي بـيـوتـكـنـولـوژي جمهـوري اسلامـي ايـران

3-5 آذر مـاه 1386، سـالـن اجلاس سـران

The 5th National Biotechnology Congress of Iran

24-26 Nov, 2007, Summit Meeting Conference Hall, Tehran- Iran

An Algorithm for Haplotype Block Partitioning based on Perfect Phylogeny

Changiz Eslahchi [1,2,*,], Hamid Pezeshk[4], Mehdi Sadeghi[3], Narjess Afzaly[1]**

[1] Faculty of Mathematics, Shahid-Beheshti University, Tehran, Iran.

[2]Bioinformatics Group, School of Computer Science, IPM, Tehran, Iran.

[3]National Institute of Genetic Engineering and Biotechnology, Tehran-Karaj Highway, Tehran, Iran.

[4]Center of Excellence in Biomathematics, School of Mathematics, Statistics and Computer Sciences, College of Science, University of Tehran, Tehran, Iran.

Abstract:

Construction of two haplotypes from a set of Single Nucleotide Polymorphism (SNP) fragments is referred to as haplotype reconstruction problem. One of the most popular computational model for this problem is Minimum Error Correction (MEC). Since MEC is an NP-hard problem, here we propose a novel heuristic algorithm based on clustering analysis in data mining for haplotype reconstruction problem. In contrast to MEC model, our algorithm has polynomial time complexity and could be applied to large datasets. Based on maximum normalized hamming distance, our iterative algorithm produces two clusters of fragments; then, in each iteration, the algorithm assigns a fragment to one of the clusters. Our results suggest that the algorithm has less reconstruction error rate in comparison with other algorithms.

WWW.IBP.IR iranian bioinformatics portal