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

ارائه روشهاي خوشه‌بندي مبتني بر شبكه‌هاي عصبي در حل مسئله‌ي بازسازي هاپلوتيپها با استفاده از اطلاعات ژنوتيپي

محمدزاده*، جواد[1] - نوذري داليني**، عباس[1و2] - اهرابيان، هايده[1و2] - معين زاده، محمد حسين[1]

[1]قطب بيومت، دانشكده رياضي، آمار و علوم كامپيوتر، دانشگاه تهران
[2]مركز تحقيقات بيوشيمي و بيوفيزيك، دانشگاه تهران

چكيده

مسئله بازسازي هاپلوتيپ، تقسيم كردن قطعه‌هاي SNP (Single Nucleotide Polymorphism Fragments)به دو قسمت و استنتاج يك جفت هاپلوتيپ از آنهااست. مدل MEC يكي از مدل‌هاي موجود براى حل اين مسئله مي‌باشد. استفاده از اطلاعات ژنوتيپ، يك مدل توسعه‌يافته براي MEC ميباشد كه MEC/GI ناميده مي‌شود. با توجه به NP-hard بودن اين مسئله، به دنبال الگوريتم‌هايي مكاشفه‌اي و روشهاي خوشه‌بندي براي كاهش زمان اجرا مي‌باشيم. در اين مقاله به ارائه روشهايي مبتني بر شبكه‌هاي عصبي براي كاهش زمان و افزايش ميزان شباهت بين هاپلوتيپهاي واقعي و بازسازي شده و همچنين برطرف كردن كاستي‌هاي روشهاي ارائه شده قبلي مي‌پردازيم. در نهايت نتايج آزمايشات انجام گرفته بر روي پايگاه داده‌هاي واقعي مقايسه مي‌شود.

كلمات كليدي: هاپلوتيپ، قطعه‌هاي SNP، اطلاعات ژنوتيپ، خوشه‌بندي، ميزان بازسازي، شبكه‌هاي عصبي

مقدمه

با توجه به در دسترس بودن توالي ژنوم انسان[1]، امكان بررسي در تفاوت-هاي ژنتيكي و ارتباط دادن آنها به بيماري‌هاي پيچيده، ميسر شده است[5]. زيست شناسان معتقدند كه همه‌ي انسان ها عموماً در سطح DNA يكسان و تفاوت تنها در برخي مكان‌هاي آن وجود دارد كه سبب بروز بيماري‌هاي ژنتيكي مي-شوند[3و2]. اين تفاوتها در SNP ظهور ميكند. SNP، بين افراد مختلف متفاوت بوده و شايع‌ترين عامل بروز تفاوت‌هاي ژنتيكي است. بيشتر SNPها از دو نوع مختلف كه در اينجا 'A' و 'B' ناميده مي‌شوند، تشكيل شده‌اند. توالي SNPها در هر يك از دو كروموزوم مربوط به يك ژنوم، هاپلوتيپ (پدرى و مادرى) ناميده مي‌شود دو مكان SNP بر روي رشته ژنوتيپ اگر مقادير يكساني داشته باشند، آن مكان هموزيگوت و در غير اينصورت هتروزيگوت ناميده ميشود. بدليل اينكه استخراج هاپلوتيپها از طريق روشهاي آزمايشگاهي بسيار دشوار و هزينه برميباشد، استفاده از روشهاي محاسباتي مورد توجه قرار گرفته است. متاسفانه بدست آوردن قطعه‌هاي SNP (داده هاي مساله)همواره با درصدي خطا ميسر ميباشد و حل مساله را با مشكل مواجه كرده است. براي حل اين مشكل، چندين مدل از طريق روشهاي محاسباتي ارائه شده است. مدل حداقل تصحيح خطا (MEC) [4] يكي از مدلهاي متداول ميباشد. MEC به تصحيح تفاوتهاي موجود بين هاپلوتيپ و توالي SNPها مي‌پردازد تا قطعات SNPرا به دو كلاس افراز نمايد. از هر كلاس يك هاپلوتيپ بدست آورده ميشود. از طرف ديگر با توجه به اينكه MEC يك مسئله‌ي NP-Hard است[5] براي حل آن از روشهاي خوشه بندي و الگوريتم‌هاي مكاشفه‌اي استفاده شده است. استفاده از اطلاعات ژنوتيپ(كه بسيار آسان‌تر از هاپلوتيپ به دست مي‌آيند) يك مدل توسعه يافته از مدل MEC مي‌باشد كه MEC/GI. ناميده مي‌شود.

مواد و روشها

براي حل اين مساله از مجموعه داده ACE(Angiotensin Converting Enzyme) [7] استفاده شده است. تغييرات انجام شده بر روي داده ها همانند ژنگ، داده ها را به يك ماتريس $M_{20\times52}$ تبديل كرده است[5]. اين ماتريس با توجه به نرخ گپ كه همان از دست دادن اطلاعات ميباشد، $g=0.25,0.5,0.75$، و نرخ خطا در خواندن $e=0.2,0.3,0.4$ در 9 مجموعه بدست مي آيند. فرض كنيم m (m=20) قطعه SNP داريم (رشته-هايي از 'A'، 'B' يا '-' (شكاف)) كه هر يك با يكي از دو هاپلوتيپ به طول n (n=52) متناظر ميباشد. پارتيشن P قطعه‌ها را به دو كلاس C_1 و C_2 تقسيم مي-كند. براي بازسازي هاپلوتيپها، قطعه‌هاي موجود در هر كلاس، با يكديگر تركيب ميشوند. اين عمل كه با فرمول 1 انجام ميگردد، راي گيري ناميده مي-شود. $N_A^j(C_i)$ تعداد 'A'هاي ستون j از قطعه‌هاي مربوط به كلاسC_i ميباشد. هر ژنوتيپ كه از تركيب دو هاپلوتيپ بدست ميآيد، رشته‌ايست كه از 'A'، 'B' يا '-' تشكيل شده است(فرمول 2). ميزان بازسازي (RR) كه براي مقايسه كارايي الگوريتم‌هاي طراحي شده به كار رفته و بر اساس فاصله هامينگ بين قطعه‌ها محاسبه ميشود (فرمول 3). در اين رابطه، $HD(h_i,h_k)=\sum_{j=1}^{n} d(h_{ij},h_{kj})$ فاصله‌ي همينگ و $d(x,y)$ مقدار تفاوت موجود بر روي دو قطعه در يك مكان خاص ميباشد.

$$V_{ij}=\begin{cases} A & N_A^j(C_i)>N_B^j(C_i) \\ B & otherwise \end{cases}$$
$$i=1,2$$
$$0\le j\prec n$$

(فرمول 1)

$$g_{i,(1\le i\le n)}=\begin{cases} A & (h_{1i}=h_{2i}=A \\ B & (h_{1i}=h_{2i}=B \\ - & otherwise \end{cases}$$

(فرمول 2)

$$RR(h,h')=1-\frac{\min(r_{11}+r_{22},r_{12}+r_{21})}{2n},$$
$$r_{ij}=HD(h_i,h'_j),\ i,j\in\{1,2\}$$
$$d(m_{ij},m_{kj})=\begin{cases} 1 & (m_{ij}\ne m_{kj}\ne -) \\ 0 & otherwise \end{cases}$$

(فرمول 3)

ژنگ شبكه عصبي نظارتي دو لايه‌اي (شكل1) را بر مبناي مدل MCIH (نام ديگرMEC/GI) را طراحي كرده است. اين شبكه عصبي پيشرو از انتشار خطاي گره‌هاي خروجي تنها در لايه‌ي اول استفاده مينمايد، پس داراى m گره ورودي (هر كدام براي يك قطعه)، دو گره‌ي مياني(هاپلوتيپها و جواب نهايي مسئله) و يك گره‌ي خروجي به عنوان ژنوتيپ ميباشد. روش ارائه شده در [5]، داراى مشكلاتي از قبيل عدم استقلالِ شبكه از تعداد قطعه‌ها و زمان اجراي زياد ميباشد. از طرف ديگر محدودِ بودن به اطلاعات ژنوتيپي باعث عدم كارايي آن در مدل MEC شده است.براي رفع اين مشكلات، دو روش خوشه-بندي به ازاي دو مدل MEC (داراي دو قسمت شبكه‌ي عصبي و پس‌پردازش) و MEC/GI توسط شبكه‌هاي عصبي در اين مقاله ارائه شده است.

شبكه‌ي عصبي: مسئله‌ي بازسازي هاپلوتيپها جزء مسائل غير نظارتي طبقه بندي ميشود. شبكه طراحي شده ما دو لايه و غير نظارتي ميباشد (شكل 1- ب). قطعات SNP به عنوان ورودى در نظر گرفته ميشوند. لايه اول اين شبكه، داراي n گره ورودي، كه هر گره به يك SNP مربوط ميشود. در لايه‌ي دوم دو گره، كه هر گره مربوط به يك هاپلوتيپ ميشود. هاپلوتيپها در طي آموزش به تدريج بر روي وزن‌هاي شبكه، به صورت اعداد اعشاري شكل ميابند و در نهايت از اين وزن‌ها استخراج ميشوند (الگوريتم 1).

پس‌پردازش: ابتدا با استفاده از فرمول 4 هاپلوتيپهاي پيشنهادي اوليه كه آنها را شبه هاپلوتيپ (*semi-haplotype*)مينماميم، را بدست ميآوريم. سپس با فرض اينكه P قطعه‌هاي SNP را به دو خوشهC_1 وC_2 تقسيم ميكند (فرمول 4)، هاپلوتيپهاي نهايي را از طريق رأي گيري (فرمول 5) محاسبه ميكنيم.

روش‌هاي دوم (مبتني بر MEC/GI): اضافه كردن اطلاعات ژنوتيپ به داده‌هاي ورودي، بازسازي هاپلوتيپ‌هاي صحيح‌تري را نتيجه مي‌دهد. در اين قسمت، از روش‌خوشه‌بندي در مدل MEC، به همراه استفاده ازاطلاعات ژنوتيپ استفاده شده است. اگر g_i برابر با 'A' ('B') باشد، نشان‌دهنده اين است كه همه عناصر موجود بر روي ستون i در M بايد 'A' ('B') بوده و مقدار 'B' ('A') در اين ستون‌ها نشان‌دهنده وجود خطاست. با اين روش تقريباً نيمي از ستون‌ها حذف خواهند شد. پس از اين مرحله، شبكه‌ي عصبي براي خوشه-بندي قطعه‌ها، اعمال مي‌شوند. براي استنتاج هاپلوتيپ‌ها از دو دسته حاصل، تابع رأي‌گيري جديدي تعريف مي‌كنيم (فرمول 6).

Algorithm: Pseudo code for new approach
Input: SNP fragments **Output:** two haplotypes
Step1: Initialize weights ($w_{i,j}$=0).
Set topological neighborhood parameters (R = 0).
Set learning rate parameters ($\alpha = 1/n$).
Step2: While stopping condition is false do, step 3-7.
Step3: For each input vector **x**, do step 4-6.
Step4: For each l = {1, 2}, compute: $D(l) = \sum_i (w_{ij}.x_i)$
Step5: Find index L such that D(L) is minimum.
Step6: For unit L, and for all i: $w_{iL}(new) = w_{iL}(old) + \alpha.x_i$
Step7 Test stopping condition.
Post-Processing.

الگوريتم 1

$$semi-haploype_{ik} = \begin{cases} B & (w_{ik} > 0) \\ A & (w_{ik} < 0) \\ Random & otherwise \end{cases}$$

$k = \{1,2\}$

$C_1 = \{ f_i : HD(h_1, m_i) < HD(h_2, m_i)\}$
$C_2 = \{ f_i : HD(h_2, m_i) \le HD(h_1, m_i)\}$
$i = 1, 2, 3, ..., m$

(فرمول 4)

$$v_{ik} = \begin{cases} A & N_A^i(C_k) > N_B^i(C_k) \\ B & N_A^k(C_k) < N_B^k(C_k) \\ Random & otherwise \end{cases}$$

$k = \{1,2\}$

(فرمول 5)

$$V_{ij} = \begin{cases} A & N^j{}_A(C_i)/N^j{}_B(C_i) > N^j{}_A(C_{i'})/N^j{}_B(C_{i'}) \\ B & N^j{}_A(C_i)/N^j{}_B(C_i) < N^j{}_A(C_{i'})/N^j{}_B(C_{i'}) \\ K_{ij} & otherwise \end{cases}$$

$$K_{ij} = \begin{cases} A & N^j{}_A(C_i) - N^j{}_A(C_{i'}) \ge N^j{}_B(C_i) - N^j{}_B(C_{i'}) \\ B & otherwise \end{cases}$$

$i, i' \in \{1,2\}, i \ne i'$
$0 \le j \prec n$

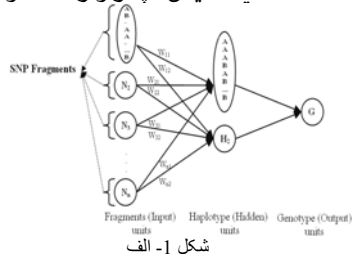
(فرمول 6)

نتايج و بحث

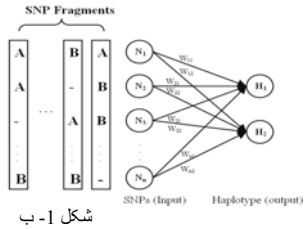
نتايج حاصل كه بر روى پايگاه داده ACE آزمايش شده بود در نمودار 3 نشان داده شده است. اين نمودار حاكيست كه روش ما در خطاي بالا (0.3 و 0.4) داراي RR بيشتر ميباشد و به مراتب بهتر از روش ارائه شده در[5] عمل مي‌كند. از طرفى وارد كردن قطعات SNP به صورت سرى، باعث عدم وابستگي اين روش به تعداد قطعات ورودى مي شود. نتايج حاصله در مدل MEC نيز، با توجه به عدم وجود اطلاعات ژنوتيپي قابل قبول مي‌باشد (نمودار1).

تشكر و قدرداني

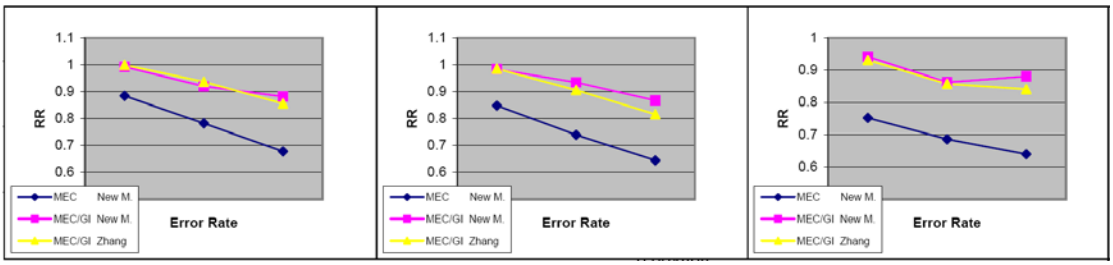
بخشي از هزينه هاي اين پروژه توسط دانشگاه تهران تامين گرديده است.



شكل 1- الف



شكل 1- ب



(a) ACE, *g* = 0.75 **(b) ACE, *g* = 0.5** **(c) ACE, *g* = 0.25**

نمودار 1: مقايسه‌ي RR در روش‌ها و مدل‌هاي مختلف

مراجع

[1] Venter, J.C. and Adams, M.D. et al. 2001. The sequence of the human genome. *Science*, 291(5507):1304–1351.
[2] Terwilliger, J. & Weiss, K. (1998). Linkage disquilibrium mapping of complex disease: Fantasy and reality? Curr. Opin. Biotechnology, 579–594.
[3] Chakravarti, A. 1998. It's raining, hallelujah? Nature Genetics, 19:216–217.
[4] Wang, R., Wu, L., Li, Z. and Zhang, X. 2005. Haplotype reconstruction from SNP fragments by Minimum Error Correction. Bioinformatics, 21(10):2456–2462.
[5] X.Zhang, R.Wang, L.Wu, W.Zhang. 2006. Minimum conflict individual Haplotyping from SNP fragments and related Genotype, Bioinformatics 271-280.
[6] X.Zhang, R.Wang, L.Wu, W.Zhang. 2007. A clustering algorithm based on two distance functions for MEC model, Computational biology and chemistry:148,150.
[7] Rieder, M., Taylor, S., Clark, A. and Nickerson, D. 1999. Sequence variation n the human an giotensim converting enzyme. Nature genetics,22:59–62.

Abstract

Most positions of the human genome are typically invariant (99%) and only some positions (1%) are commonly invariant which are associated with complex genetic diseases. Haplotype reconstruction is to divide aligned SNP fragments, which is the most frequent form of difference to address genetic diseases, into two classes, and thus inferring a pair of haplotypes from them. Minimum error correction (MEC) is an important model for this problem but only effective when the error rate of the fragments is low. MEC/GI as an extension to MEC employs the related genotype information besides the SNP fragments and so results in a more accurate inference. The haplotyping problem, due to its NP-hardness, may have no efficient algorithm for exact solution. In this paper, we design an unsupervised neural network based on MEC and MEC/GI model. As numerical results on real biological data, this neural network approach works well and an increase in the rate of similarity between the real haplotypes and the reconstructed ones is gained.

WWW.IBP.IR iranian bioinformatics portal