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

يافتن موقعيت هاي داياد با الگوريتم ژنتيك

هاشمی فر، سمیه*[1] - صادقی، مهدی[2,3,4] - بهرام، گلیائی[2] - نوذری، عباس[1,2] –اهرابیان، هایده[1,2] - زارع، فاطمه**

[1]گروه علوم کامپیوتر دانشکده ریاضی، آمار و علوم کامپیوتر دانشگاه تهران

[2]مرکز تحقیقات بیوشیمی و بیوفیزیک دانشگاه تهران

[3]پژوهشکده مهندسی ژنتیک و زیست فناوری

[4]پژوهشکده کامپیوتر و پژوهشگاه دانش های بنیادی

چکیده

در این مقاله یک الگوریتم ژنتیک جدید برای یافتن موتیف هاي داياد ناشناخته در یک مجموعه از دنباله های داده شده ارائه شده است. الگوریتم بر روی یک مجموعه داده ای تست گردیده و نتایج آن با الگوریتم های داياد دیگر مقایسه شده است. نتایج مقایسه، کارآیی بسیار خوبی را براي الگوریتم نشان می دهد.

کلمات کلیدی: عوامل الگوبرداری، موتیف، الگوریتم های یافتن موتیف، الگوریتم ژنتیک.

مقدمه

يكي ازمسائل مهم دربيولوژي مولكولي، يافتن چگونگي ارتباط بين ژن هايي است كه يك فعاليت معين را انجام مي دهند و يا با يكديگر بيان مي شوند [4]. آزمايشات تجربي بر روي ناحيه پروموتر تعدادي از ژن ها نشان داده است كه براي شروع پروسه نسخه برداري، يك مولكول خاص كه (TF) فاكتور رونويسی نامیده مي شود به زير رشته هاي كوتاهي در ناحيه پروموتر ژن هاي بيان شده با هم پيوند مي خورد. به هر كدام از اين زير رشته ها يك مكان اتصال (bs) از فاكتور رونويسی مزبور مي گويند. اين مكان های اتصال داراي يك الگوي مشترك هستند كه اغلب موتيف يا سيگنال ناميده مي شود و به مساله يافتن اين الگوهايافتن موتيف مي گويند [4]. تركيبي از چندين الگوي مرتبط با هم كه همزمان رخ مي دهند يك الگوي تركيبی نام دارد. مشكل اساسي در يافتن سيگنال هاي composite اين است كه بعضي از الگوهاي موجود در تركيب ممكن است خيلي ضعيف باشند كه بر اساس ديدگاه هاي قديمي به سختي پيدا مي شوند [5]. يك نمونه از الگوهاي تركيبی سيگنال هاي داياد هستند كه تركيبي ازيك جفت الگوي تكي (موناد) مي باشند كه در فاصله ثابتي از يكديگر رخ مي دهند[2]. برخی از الگوريتم های معروف كه تا كنون براي يافتن سيگنالها ارائه شده اند عبارتند از: MEME و AlignACE. MEME بر اساس تكنيک expectation-maximization، عمل مي كند [1]، در حالی كه AlignACE [3] با بهره گيري از روش Gibbs sampling موتيف ها را مي يابد. در اين مقاله يك الگوريتم ژنتيك براي يافتن سيگنال هاي داياد ارائه مي شود. يكي از خصوصيات بهينه اين روش يافتن دو بخش سيگنال داياد به طور همزمان مي باشد كه منجر به يافتن الگوي موتيف به نحو موثرتري مي شود. تعميم اين الگوريتم برای حالت سيگنال های تركيبی نيز براحتی قابل انجام است .

مواد و روشها

مجموعه داده ها

مجموعه داده هايي كه در اين مقاله مورد آزمايش قرارگرفته است، 5 تنظيم كننده مخمر است كه در پايگاه داده scpd گزارش شده است[4]. در اين پايگاه همراه با هرموتيف ژن هاي تنظيم شده بوسيله آن و موقعيت مصداق هايش ذكرشده است. براي هرتنظيم كننده ناحيه اي به طول 850 bp از موقعيت 800- تا 50+ از ژن هاي بيان شده بوسيله آن آزمايش شده است.

الگوريتم يافتن موتيف

مجموعه $\{s_1,\ldots,s_t\}$ متشكل از t رشته به طول n داده شده است و هدف يافتن موتيف ناشناخته دربين آنها است. يك موتيف داياد E به طول L از دوبخش E_1 به طول l_1 و E_2 به طول l_2 $(l_1+l_2=L)$ تشكيل شده است كه در فاصله k $(l_{\min}\le k\le l_{\max})$ از يكديگر قرار دارند و $E=E_1+E_2$. براي يافتن مصداق هاي موتيف E، عمليات از اولين رشته s_1 آغاز شده و تمام $(n-L+1)\times(l_{\max}-l_{\min}+1)$ زير رشته L تايي داياد موجود در s_1 با E مقايسه مي گردد و در نهايت زير رشته ي با بيشترين تعداد تطابق به عنوان اولين مصداق (B_1) در نظر گرفته می شود. تعداد تطابق هاي بين E با هر كدام از زير رشته هاي $s_1[i]\ldots s_1[i+l_1+l_2-1]$ برابر با جمع تعداد تطابق هاي E_1 و E_2 با زيررشته هاي $s_1[i]\ldots s_1[i+l_1-1]$ به طول l_1 و $s_1[i+l_1+k]\ldots s_1[i+l_1+l_2-1]$ به طول l_2 مي باشد كه به صورت زير محاسبه مي گردد:

$$M(s_1[i]\ldots s_1[i+l_1+l_2-1],E[1]\ldots E[l_1+l_2])=\sum_{j=1}^{l_1+l_2}d(s_1[i+j-1],E[j]),$$

كه در آن

$$d(a,b)=\begin{cases}1 & if \quad a=b\\ 0 & otherwise.\end{cases}$$

به ازاي ديگر رشته هاي s_i $(2\le i\le t)$ اين عمل تكرارمي شود تا در نهايت مصداق هاي $\{B_1\ldots B_t\}$ تشكيل گردد. با توجه به مطالب فوق الگوريتم ژنتيك يافتن موتيف هاي داياد به صورت زير می باشد:

1. تشكيل جمعيت اوليه p_0 به سايز $Psize$.
2. ارزش دهي به اعضاي p_0 بر اساس تابع برازندگي F و مرتب كردن آن ها.
3. مقدار دهي $k=1$ و انجام عمليات زير به ازاي N نسل :

- انتقال $Psize\times(1-pc)$ اعضاي بهتر جمعيت P_{k-1} به P_k .
- انجام عمل crossover روي $pc\times Psize/2$ عضو از P_{k-1} و انتقال آنها به P_k.
- مرتب كردن P_k بر اساس تابع برازندگي.
- انجام عمل mutation روي $pm\times Psize$ عضو از P_k و افزايش 1 واحد به k.

4. انتخاب 10 موتيف با رتبه بالاتر به عنوان نتيجه نهايي.

براي تشكيل جمعيت اوليه p_0 الگوريتم، از هر رشته s_i $(2\le i\le t)$، $Psize/t$ موتيف به صورت زير انتخاب مي گردد: به ازاي موقعيت تصادفي j از رشته s_i $(1\le j\le n-L+1)$ تمام $(l_{\max}-l_{\min}+1)$ موتيف داياد ممكن كه ازدو زير رشته E_1 در موقعيت j ام و E_2 درموقعيت $(j+l_1+k)$ تشكيل شده است، بررسي مي شود. اگر تعداد تطابق ها از يك حد آستانه α بيشتر باشد و موتيف تصادفي نباشد به عنوان يكي از افراد جمعيت اوليه p_0 در نظر گرفته می شود. اين عمل تا زماني انجام مي شود كه $Psize/t$ موتيف مناسب از هر رشته s_i انتخاب شود. در نهايت $Psize$ عضو در جمعيت اوليه p_0 به دست خواهد آمد. چون در حالت ايده آل تعداد كل تطابق ها برابر با $t\times(l_1+l_2)$ خواهد بود مي توان $\alpha=\lfloor t\times(l_1+l_2)\times(3/4)\rfloor+1$ در نظر گرفت. براي بررسي تصادفي بودن موتيف، 20 ناحيه پروموتر از مخمر به طور تصادفي انتخاب مي شود. اگر موتيف مورد نظرحداقل با β زير رشته از اين 20 ناحيه به تعداد $(l_1+l_2)/2$ تطابق داشته باشد، تصادفي قلمداد مي شود. برای محاسبه β، C_x برابر با تعداد رخدادهاي رشته اي به طول L كه از كاراكتر x تشكيل شده است،

در نظر گرفته می شود که حداقل $(l_1+l_2)/2$ تطابق را با نواحي تصادفي انتخاب شده داشته باشد. آنگاه $\beta=(C_A+C_C+C_G+C_T)/4$.

براي ارزشگذاري هر موتيف از مجموع سه تابع برازندگي F_{SP}، F_{NM} و F_{IC} استفاده مي شود: $F=F_{SP}+F_{NM}+F_{IC}$.
براي محاسبه F_{SP} كه بيانگر *sum of pairs* است، ابتدا تمامي مصداق هاي الگوي E يعني $\{B_1\ldots B_t\}$ به دست آمده وسپس مجموع تعداد تطابق هاي بين هر جفت ازمصداق هاي مطابق فرمول زير محاسبه مي گردد: $F_{SP}=\sum_{j=1}^{t}\sum_{r=1}^{t}M(B_j,B_r)$.
F_{NM} برابر با مجموع كل تعداد تطابق هاي بين الگوي E با همه مصداق هاي آن مي باشد: $F_{NM}=\sum_{j=1}^{t}M(E,B_j)$.
براي محاسبه F_{IC} كه بيانگر Information content از ماتريس وزنی موقعيت مربوط به مصداق هاي موتيف طبق فرمول زير استفاده می شود: $F_{IC}=\sum_{i=1}^{4}\sum_{j=1}^{t}p_{ij}\times\log(p_{ij}/p_0)$ كه p_{ij} عناصر ماتريس وزنی موقعيت و p_0 احتمال رخداد نوكلئوتيد در زمينه است. سپس مقادير حاصل از اين سه تابع با تقسيم بر مقدار ماكزيمم هر ارزش، به مقداري دربازه $[0,1]$ تبديل مي گردد. مقدار ماكزيمم براي F_{NM} برابر با $t\times(l_1+l_2)$، براي F_{IC} برابر با $2\times(l_1+l_2)$ و براي F_{SP} برابر با $t\times(t-1)/2\times(l_1+l_2)$ مي باشد.

در الگوريتم ژنتيك ارائه شده، از crossover يك نقطه اي استفاده مي شود. به اين منظور دو موتيف x_i و x_j ازجمعيت را به عنوان دو والد منتخب توسط roulette wheel درنظرگرفته ومصداق هاي $\{B_{i_1},\ldots,B_{i_t}\}$ و $\{B_{j_1},\ldots,B_{j_t}\}$ يافته می شود. سپس اين دو مجموعه t تايي در نقطه تصادفي r $(.2\times t\le r\le(t-.2\times t))$ با هم crossover شده ودو فرزند جديد به دست مي آيد. براي mutation ابتدا ماتريس وزنی موقعيت مربوط به مصداق هاي موتيف محاسبه مي شود. هر ستون j از ماتريس كه مقدار Information content آن كمتر از 1 است به عنوان يك ستون غير conserve در نظر گرفته می شود. سپس روي هرستون غيرconserve، با توجه به عدد تصادفي q_i به صورت زيرعمل مي شود:

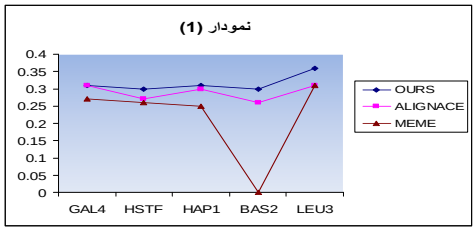
$$\begin{cases} B_{ij}=C & if \quad q_i\le .3\\ B_{ij}=G & if \quad .3<q_i\le .6\\ B_{ij}=T & if \quad .6<q_i\le .8\\ B_{ij}=A & otherwise.\end{cases}$$

براي يافتن مصداق هاي نهايی هرموتيف، ماتريس وزنی موقعيت آن با رشته هاي $\{s_1,\ldots,s_t\}$ متوازن مي گردد. براي توازن ماتريس وزنی موقعيت به هر زير رشته داياد $x=x[1],\ldots,x[l_1+l_2]$ كه در فاصله k از يكديگر قراردارند رتبه اي به صورت $score(x)=\sum_{i=1}^{4}\sum_{j=1}^{l_1+l_2}p_{ij}\times S_{ij}$ اختصاص داده مي شود كه $S_{ij}=\begin{cases}1 & if \quad x[j]=ith\ base\ in\ \{A,C,G,T\}\\ 0 & otherwise\end{cases}$.
سپس در هر رشته s_i زير رشته هاي با بهترين رتبه به عنوان مصداق انتخاب مي شوند.

نتايج

براي مقايسه الگوريتم ارائه شده با دو روش AlignACE 3.0 و MEME 3.5.4 ازتركيب پنج مقياس زيراستفاده شده است : 1) positive (5 sensitivity (4 specificity (3 correlation coefficient (2 performance coefficient

prediction . اين مقياس ها در سطح نوكلئوتيد موتيف ها ارزيابي مي شوند . نمودار 1 نتايج مقايسه سه الگوريتم براي داده هاي ذكر شده را نشان مي دهد.



تشكرو قدردانی

بخشی از هزينه های اين پروژه توسط دانشگاه تهران و بخشی از آن توسط پژوهشگاه دانش های بنيادی $(CS-1385-1-02)$ تامين می شود.

منابع

1. Bailey T.,Elkan C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning 21:51-80.
2. Helden J., Rios A.F., Collado-Vides j. (2000) Discovering Regulatory Elements in Non-Coding Sequences by Analysis of Spaced Dyads. Nucleic Acids Research 28:1808-18.
3. Hughes J.D.,et al. (2000) Computational identification of cisregulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. Molecular biology 296:1205–14.
4. Sinha S.,Tompa M. (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Research 31(13):3586-8
5. Tompa M.,Li N.,et al. (2005) Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites.Nature Biotechnology 23:137-144.

Abstract

A major challenge in molecular biology is to understand the mechanisms that regulate the expression of genes. An important step in this challenge is the ability to identify regulatory elements, notably the binding sites in DNA for transcription factors. These binding sites are called motifs.

In this paper a new genetic algorithm for finding unknown dyad motifs in a given set of sequences is presented. Not many algorithms for finding dyad motif are presented in the literature and no genetic algorithm is given till now. In our genetic algorithm a novel multi-objective fitness function is employed for the selection of best individuals. The fitness function is designed based on instance consensus motifs and position weight matrix. The algorithm is tested on the different sets of real data. The result of algorithm is compared with the result of two well known algorithms MEME and AlignACE.

WWW.IBP.IR iranian bioinformatics portal