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 پيدا كردن موتيف هاي توالي DNA بر مبناي جبر خطّي

شيخ عطّار عليرضا[1و2*] ، اصلاح چي چنگيز[1و2**] ، پزشك حميد[3] ، صادقي مهدي[4]
1-دانشكدة رياضي ، دانشگاه شهيد بهشتي 2-پژوهشكدة كامپيوتر ، پژوهشگاه دانشهاي بنيادي 3- گروه آمار دانشكدة رياضي آمار و علوم كامپيوتر دانشگاه تهران 4-پژوهشگاه ملّي مهندسي ژنتيك و زيست شناسي

چكيده

بررسي الگوهاي تقريباً مشابه روي توالي هاي DNA ، يا همان موتيف ها (Motifs) ، يك مسئلة بسيار مهم در زيست شناسي محاسباتي است كه در سالهاي اخير، تحقيقات و كوشش هاي فراواني به منظور دست يابي به روش مناسب تر براي يافتن آن صورت گرفته است. در اين راستا ده ها روش محاسباتي متفاوت ارائه شده است كه اخيراً تلاشهايي براي مقايسة نه همه، بلكه تعداد اندكي از اين روش ها به عمل آمده است امّا هنوز به سختي مي توان با استفاده از اين كوشش ها يكي از روش هاي مذكور را برجسته تلقّي كرد. ما با ديدگاه جبر خطّي، بر آنيم تا روش جديدي براي يافتن موتيف ها ارائه دهيم.

كلمات كليدي: موتيف ، رديف سازي توالي ، پروفايل ، نظريه گراف

مقدّمه

يكي از مسئلة هاي بسيار مهمي كه زيست شناسان را با چالشي بزرگ مواجه كرده است ، دانستن ساز و كاري است كه ناظر بر قاعده مندي بيان ژن مي باشد. گام اساسي براي مو شكافي دقيق تر چنين نظام حاكم بر ژن ها، روشن ساختن مناطق قاعده مندي روي DNA است. اين مطلب منجر به مسئلة يافتن موتيف ها مي شود:" با در نظر گرفتن تعدادي از توالي هاي DNA داده شده، مناطقي از آنها را شناسايي كنيد كه نامزد خوبي براي موتيف ها باشند." انتظار مي رود اين مناطق با توجّه به قاعده مندي بيان ژن، همگي الگوي تقريباً مشابهي داشته باشند. اين تشابه باعث مي شود مسئلة يافتن موتيف ها كه شالوده اي زيستي دارد به مسئله اي كاملاً محاسباتي تبديل شود.

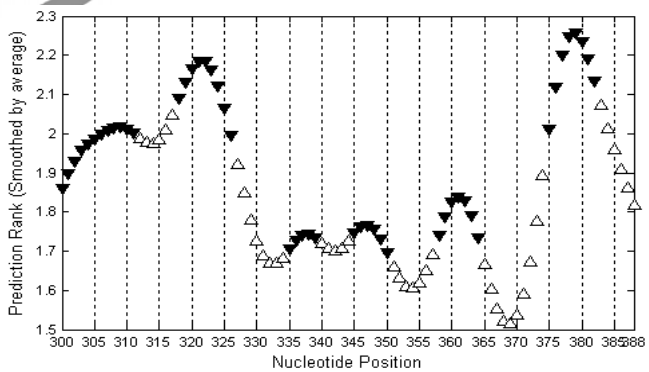
تئوري و روش

در اكثر روش هاي موجود براي يافتن موتيف ها، هدف پيدا كردن زيررشته هايي هم اندازه به طول l از توالي هاي DNA است كه حداكثر d نوكلئوتيدِ غيرمنطبق داشته باشند. در روش ارائه شده خود را محدود به يافتن موتيف هايي به طول خاص نكرده ايم و علاوه بر آن به جاي حداكثر d نوكلئوتيدِ غيرمنطبق، از مفهومي به نام همرديفي محلّي بهره برده ايم. در اين روش ابتدا هر توالي DNA طبق نظام خاصي قطعه قطعه مي شود. در واقع به هر نوكلئوتيد بر اساس ميزان پيش بيني پذيري اش نسبت به چند نوكلئوتيد قبلي، يكي از اعداد 1، 2، 3 ويا 4 را نسبت مي دهيم. فرض كنيد f نگاشتي است كه نوكلئوتيدهاي C,T,A و G را به ترتيب به چهار پاية استاندارد مربوط به فضاي برداري R^4 ، يعني e_1,e_2,e_3 و e_4 مي نگارد. اگر $N_{a-3}N_{a-2}N_{a-1}N_a$ چهار نوكلئوتيد پشت سرهم روي يك توالي DNA باشد، منظور از يك تركيب خطّي براي آن عبارت است از $\sum_{n=0}^{3} c_n f(N_{a-n})$ كه در آن c_n ها چهار عدد حقيقي در فاصلة $[-1,1]$ هستند. هدف يافتن c_n هايي

است كه فاصلة اقليدسي $\sum_{n=0}^{3} c_n f(N_{a-n})$ تا $f(N_{a+1})$ كمتر يا مساوي با فاصلة اقليدسي $\sum_{n=0}^{3} c_n f(N_{a-n})$ تا هر پاية استاندارد ديگري شود. درواقع ضرايبي را خواهيم يافت كه بتوانيم نوكلئوتيد پنجم را پيش بيني كنيم. اگر منظور از $\|V_1, V_2\|$ فاصلة اقليدسي دو بردار V_1 و V_2 باشد، نقطه اي كه در آن تابع زير كمينه مي شود، گزينة خوبي براي c_n هاي مورد انتظار خواهند بود:

$$P(c_0, c_1, c_2, c_3) = \sum_{t=0}^{3} \left\| f(N_{a+1+t}) \,,\, \sum_{n=0}^{3} c_n f(N_{a-n+t}) \right\|^2$$

يعني در طول هشت نوكلئوتيد، مجموع مربّعات فاصلة هر چهار نوكلئوتيد تا نوكلئوتيد پنجم كمينه شده است. حال رتبة $\left\| f(N_{a+1+4}) \,,\, \sum_{n=0}^{3} c_n f(N_{a-n+4}) \right\|$ در بين مجموعة $\left\{ \left\| e_x \,,\, \sum_{n=0}^{3} c_n f(N_{a-n+4}) \right\| \right\}_{x=1,2,3,4}$ را به N_{a+1+4} نسبت مي دهيم و از نو همين كار را تكرار مي كنيم تا نموداري بدست آوريم كه به هر نوكلئوتيد عددي از مجموعة $\{1,2,3,4\}$ نسبت مي دهد. البته اين نمودار را با روش ميانگين گيري هموار مي كنيم (شكل1). دامنة اين نمودار مي تواند به چهار ناحية متفاوت تقسيم شود: 1)تقعّر هاي بالا 2)تقعّرهاي پايين 3) شيب هاي مثبت 4) شيب هاي منفي. در واقع مي توان از روي هر توالي DNA چهار نسخة قطعه قطعه شده بدست آورد كه هر نسخه شامل قطعه هايي است كه از تقسيمات ناحيه اي 1 تا 4 بدست آمده اند (شكل1). به طور شهودي هر قطعه گوياي ميزان پيچيدگي ترتيب ظاهر شدن نوكلئوتيد ها در آن ناحيه روي توالي DNA است. چون $Motif$ ها داراي الگوي تقريباً مشابهي هستند، انتظار مي رود فراواني آنها در يكي از چهار ناحية مذكور بيشتر باشد. حال يك قطعه از DNA_1 به نام seg_1 را با قطعه اي از DNA_2 به نام seg_2 به طور محلّي همرديف دوگانه مي كنيم. اگر هر قطعة بدست آمده روي هر DNA را به صورت يك رأس از يك گراف نمايش دهيم، seg_1 را به seg_2 متّصل مي كنيم اگر و تنها اگر امتياز همرديفي بيش از 16 باشد كه اين امتياز از آزمايش هاي فراوان بدست آمده است. اين كار را براي هر دو قطعه از دو توالي DNA متمايز انجام مي دهيم تا گرافي چند بخشي بدست آوريم (شكل2). در واقع قطعه هاي ايجاد شده از يك DNA همگي در يك بخش قرار دارند. زير گراف هاي كامل ماكسيمال در اين گراف را پيدا مي كنيم. به اين زيرگراف هاي كامل، ميانگين امتياز همرديفي دوگانه رأس هايشان را نسبت مي دهيم و آنها را بر حسب اين امتياز در يك ليست رتبه بندي مي كنيم. براي يافتن يك موتيف اولين زير گراف كامل را از ليست انتخاب مي كنيم. هر قطعه را به صورت يك

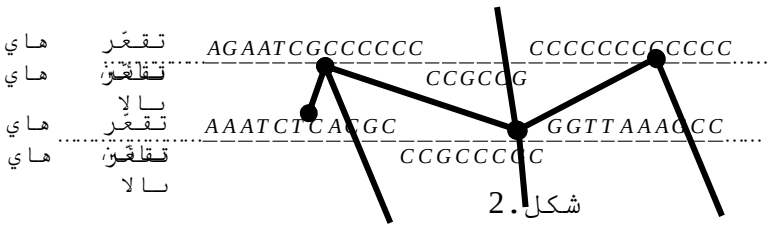


پروفايل (*profile*)

در نظر مي گيريم كه تعداد ظاهر شدن هر نوكلئوتيد در مكان ها يش را نشان مي دهد. حال با روش رديف سازي پروفايل دو قطعه اي را بدست مي آوريم كه بيشترين امتياز همرديفي پروفايل را داشته باشند. منظور از ادغام دو پروفايل، ماتريسي است كه از مجموع درايه هاي فاصله اي كه دو پروفايل با هم همريف شده اند، بدست مي آيد. اين دو پروفايل را حذف مي كنيم و به جاي آن اين دو را با هم ادغام مي كنيم. سپس اين كار را تا زمانيكه يك پروفايل بدست آيد ادامه مي دهيم. اين پروفايل را به يك ماتريس احتمال تبديل مي كنيم و آن را روي توالي *DNA* حركت مي دهيم تا ناحيه اي از توالي *DNA* را بيابيم كه با توجّه به پروفايل بيشينه امتياز را كسب كرده باشد. حاصل يك پيشنهاد براي موتيف است. اين كار را روي *DNA* هاي ديگر نيز انجام مي دهيم تا ديگر پيشنهاد ها بدست آيد.

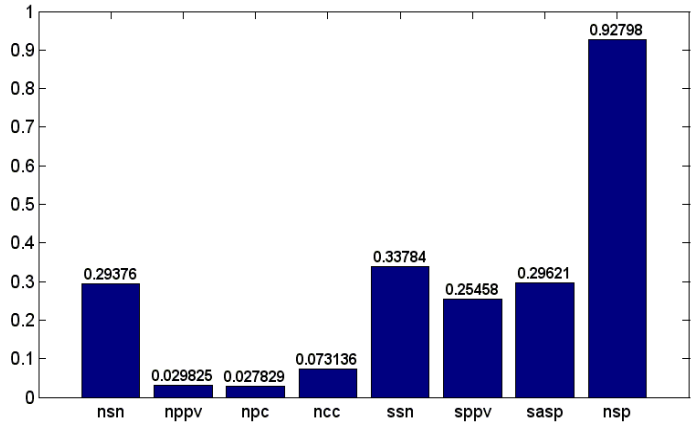
نتايج و بحث

شكل.1 مقدار عددي رتبة پيش بيني پذيري هر نوكلئوتيد را نشان مي دهد كه با روش ميانگين گيري هموار شده است. در اين نمودار مي توان تقعّر هاي بالا و پائين را از هم تشخيص داد.



شكل.2

شكل.2 قسمتي از گرافي را نشان مي دهد كه در آن تنها رئوس مربوط به تقعّر ها ديده مي شود.



شكل.3

مقدار عددي آماره هاي *Tompa*[1] براي 50 توالي *DNA* واقعي مربوط به نتايج حاصل از الگوريتم در نمودار روبرو نشان داده شده است. اين توالي ها در http://tare.medisin.ntnu.no/ قابل دسترسي است.

تشكّر و قدرداني

بخشي از هزينة اين كار از محل طرح شمارة (CS 1385-1-02) پژوهشگاه دانشهاي بنيادي (IPM) تأمين شده است.

منابع

1.Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, FavorovAV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS,Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van HeldenJ, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 2005, 23:137-44.

2.Geir Kjetil Sandve, Osman Abul, Vegard Walseng and Finn Drabløs(2007) Improved benchmarks for computational motif discovery. BMC Bioinformatics 2007, 8:193

3. Elena Zaslavsky and Mona Singh(2006) A combinatorial optimization approach for diverse motif finding Applications. Algorithms for Molecular Biology 2006, 1:13

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

An algorithm based on linear algebra for DNA sequence motif finding

Sheikh Attar Alireza [1,2*], Eslahchi Changiz[1,2], Pezeshk Hamid[3], Sadeghi Mehdi[4]**

1. Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran
2. School of Computer Science, Institute for Studies in Theoretical Physics and Mathematics(IPM), Tehran
3. Center of Excellence in Biomathematics, School of Mathematics, Statistics and Computer Sciences, University College of Science, University of Tehran, Tehran, Iran,
4. National Institute of Genetic Engineering and Biotechnology, Tehran, Iran

Abstract
One of the most important problems in molecular biology is identification of repeated patterns in DNA sequences that are presumed to have a biological function. These patterns are called Motifs. Many computational methods have been proposed and most of themhave been designed to find motifs with the fix length (d) and with at most L nucleotides mismatch between each pairs. However, we can get better results by considering sequences that are not restricted on d and L. We introduce a versatile linear algebra approach for the motif finding problem. We assign e_1, e_2, e_3 and e_4 ,four standard basis of R^4, to A, T, C and G respectively. So we have sequences of vectors instead of nucleotides. By means of linear algebra concepts we define the predictability of each nucleotide in its appearance order. This allows us to cut the DNA to subsequences called segments. Each segment is in fact one of the four kinds of regions in predictability plot. If we show any segments as a vertex of a graph, two vertexes are connected if and only if their pairwise local alignment score is more than 16. So each maximal cliques in this graph derives a profile for a motif.

WWW.IBP.IR iranian bioinformatics portal