WWW.IBP.IR iranian bioinformatics portal

نجمين 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

الگوریتمِ بهینه برای پیدا کردنِ موتیف[1] در شبکه‌های زیست‌شناسی و کاربردِ آن

اهرابیان[1,2]، رزاقی مقدم*[1]، مسعودی‌نژاد[1]، نوذری[1,2]**

1.مرکز تحقیقاتِ بیوشیمی و بیوفیزیك دانشگاه تهران،

2.گروه علوم کامپیوتر دانشکده ریاضی، آمار و علومِ کامپیوتر دانشگاه تهران.

چکیده

درك فعل و انفعالات در شبکه‌های زیست‌شناسی مانند شبکه‌های تنظیم بیان ژن، در زیست‌شناسی مولکولی حایز اهمیت است. در این رابطه، توجه بالایی به شناخت الگوهای معنی‌دار از شبکه‌های زیست‌شناسی می‌شود. به چنین الگوهایی، موتیف شبکه می‌گویند که احتمالاً نقش اساسی در شبکه‌ها ایفا می‌کنند. در این مقاله الگوریتم نوینی برای پیدا کردن موتیف ارایه می‌شود که در مقایسه با الگوریتمهای پیشین دارای برتری زمانی و حافظه است. این الگوریتم بر روی چندین شبکه زیست‌شناسی اجرا شده و نتایج مطلوبی به دست آمده است.

کلماتِ کلیدی: شبکه زیست‌شناسی، موتیف، شبکه تصادفی.

مقدمه

شبکه‌های زیست‌شناسی، که عموماً شبکه‌هایی پیچیده و وسیع هستند، حاوی اطلاعات مهمی می‌باشند. نظر به پیشرفتِ قابل ملاحظه‌ای که امروزه در این شبکه‌ها صورت گرفته است، اهمیت بررسی عمیق آن‌ها، برای استخراج اطلاعاتی که هر کدام از آن‌ها دربردارند، روزافزون است. چنین بررسی‌هایی، عموماً به‌وسیله‌ی عملیاتی بر روی شبکه انجام می‌گیرد. یکی از مهم‌ترین عملیاتی که بر روی این گونه شبکه‌ها اِعمال می‌شود، می‌توان پیدا کردن موتیف‌های شبکه اشاره کرد(1). در ادامه روشی برای پیدا کردن موتیف‌های شبکه ارایه می‌گردد. براساس این نظریه که «تکامل مدل‌هایی را که نمایانگر عملکردهای خاصی هستند، حفظ می‌کند»(9)، میلو[2] و همکارانش طرحی ارایه کردند که بر مبنای آن، در یك شبکه‌ی زیست‌شناسی باید به دنبالِ زیرشبکه‌هایی باشیم که با فراوانی بیشتری نسبت به آن‌چه در شبکه‌های تصادفی اتفاق می‌افتد، مشاهده شوند(6). به این‌چنین زیرشبکه‌های کوچك هم‌بند، که فراوانی آن‌ها بالاتر از میزان تصادفی است، موتیف اطلاق می‌شود. یافتنِ موتیف‌های شبکه نقش مهمی در درکِ عملکرد و طرح قاعده‌ای کلی برای ارتباطات مولکولی ایفا می‌کند. تحقیقات نشان داده است که در شبکه‌هایی که فرآیندهای مشابه را نشان می‌دهند، موتیف‌های مشابهی مشاهده می‌گردد، هرچند که این شبکه‌ها به سیستم‌های مختلف زیست‌شناسی مربوط باشند(9). به همین خاطر، موتیف‌ها می‌توانند معرفی کننده‌ی کلاس‌های جامعی از شبکه‌ها باشند.

پیدا کردنِ موتیف‌های شبکه در حقیقت مساله‌ی پیدا کردن زیرگراف‌هایی است که از نظر آماری دارای اهمیت‌اند. برای نیل به این هدف، روش‌های محاسباتی برای شمارش الگوهایی با اندازه‌ی مشخص در شبکه، که مساله‌ی محاسباتی پیچیده‌ای است، مورد نیاز می‌باشد. تاکنون، الگوریتم‌های مختلفی برای پیدا کردن زیرگراف‌های با فراوانی بالا در مجموعه شبکه‌ها یا در یك گراف ارایه شده است(4، 7، 8، 10 و 11). این الگوریتم‌ها یا تنها به شمارش انواع خاصی از زیرگراف‌ها می‌پردازند و یا شرطی بر روی اندازه‌ی زیرگراف دارند. این مساله شایان توجه است که یافتن موتیف‌ها می‌بایست به شمارش زیرگراف‌هایی محدود شود که دارای یال‌های مشترك نیستند، چرا که در غیر این صورت ویژگی انحطاطپذیری (فراوانی زیرگراف‌ها، با افزایش اندازه آن‌ها، کاهش می‌یابد) نخواهند داشت و مساله از نظر محاسباتی قابل حل نخواهد بود.

---

[1] Motif
[2] Milo

به طور کلی، هر الگوریتمِ شمارشِ موتیف‌های از اندازه‌ی k در یک گراف مفروض نیازمندِ طی کردنِ سه مرحله‌ی زیر است:

1. شمارش تمام زیرگراف‌های از اندازه‌ی k که در گراف وجود دارد.
2. تعیین زیرگراف‌های ایزومورف در بین این زیرگراف‌ها و شمارش تعداد اعضای گروه‌های ایزومورفی.
3. مقایسه تعداد زیرگراف‌ها با تعدادِ مورد نظر در یک گرافِ تصادفی.

انجامِ مراحلِ اول و دوم با افزایشِ تعدادِ انواع زیرگراف‌ها از اندازه‌ی مشخص، بسیار زمان‌بر است. به همین خاطر، الگوریتم‌های مختلفی برای این مساله مطرح شده است، به طور مثال، برخی از این الگوریتم‌ها به صورت نمونه‌گیری بخشی از کلِ زیرگراف‌ها را موردِ شمارش قرار می‌دهند که واضح است نتیجه‌ی چنین الگوریتمی دقیق نخواهد بود(4). به این منظور ارایه الگوریتمی دقیق که این مساله را در زمانِ مناسبی حل نماید از اهمیتِ بالایی برخوردار است.

مواد

همان‌طور که مشخص است، داده‌های این الگوریتم، شبکه‌های گوناگون خصوصاً شبکه‌های زیست‌شناسی است. این شبکه‌ها را می‌توان از پایگاه‌داده‌های مختلفی مانندِ *KEGG*(2) استخراج کرد. عموماً شبکه‌هایی که تا به امروز موردِ آزمایش قرار گرفته‌اند، شبکه‌های تنظیمِ ژن است که بیشتر برای گونه‌هایی چون *ecoli* و مخمر انجام شده است. الگوریتمِ ارایه شده در این مقاله بر روی شبکه‌ی تنظیمِ ژنِ گونه‌ی مخمر و یک شبکه‌ی الکتریک موردِ آزمایش قرار گرفته شده است. همچنین، برای مقایسه‌ی بهترِ آن با الگوریتم‌های پیشین، آن را بر شبکه‌ی کاملِ تنظیمِ ژنِ *ecoli* اجرا کردیم که بخشی از نتایج در انتهای مقاله ذکر شده‌است.

الگوریتم

در این‌جا الگوریتمی ارایه می‌کنیم که مراحلِ اول و دوم را به طور دقیق با شمارشِ یک و تنها یک بارِ هر زیرگراف انجام می‌دهد. این الگوریتم بر اساسِ تجزیه‌ی شبکه، به روشِ حذفِ راس عمل می‌کند. در این الگوریتم، تمامِ موتیف‌هایی که شاملِ راس مفروضی هستند، شمارش می‌شوند و سپس راس مفروض از شبکه حذف می‌گردد.

از آن‌جایی که شبکه‌های زیست‌شناسی عموماً *scale free* هستند، حذفِ ریوسی که بیشترین درجه‌ها را دارند، به سرعت باعثِ تجزیه‌ی شبکه می‌شوند و این امر روندِ شمارشِ موتیف‌ها را سرعت می‌دهد. بر این اساس برای هر راس، با توجه به ریوسی که حداکثر k-1 یال با آن فاصله دارند، درختی به روشی مشابه با روشِ پیمایشِ عرضی گراف تولید می‌کنیم. شباهت در پیمایش شبکه است که همسایه‌های هر راس مشابه روش پیمایشِ عرضی در تولیدِ درختِ مذکور پیموده می‌شوند. همچنین به منظورآن‌که بتوانیم زیرگراف‌ها را تنها یک بار شمارش کنیم، از "الگوی شمارشی" استفاده می‌کنیم. در این "الگوی شمارشی"، ابتدا تمام زیرگراف‌هایی که به جز ریشه، بقیه‌ی ریوس در سطح اول هستند، در نظر گرفته می‌شود. سپس، تمام زیرگراف‌هایی که شاملِ k-2 راس در سطح اول و یک راس در سطح دوم هستند، شمرده می‌شوند و به همین ترتیب ادامه پیدا می‌کند. در این روش با قیدِ شروطی خاص تمام زیرگراف‌ها تنها یک بار شمرده می‌شوند. پس از آن‌که تمامِ زیرگراف‌های شاملِ راس مفروض شمرده شد، نیاز به پیدا کردنِ زیرگراف‌های ایزومورف و شمارشِ تعداد زیرگراف‌های موجود در هر گروه ایزومورفی است که این بخش به کمک الگوریتم‌های موجود انجام می‌شود.

پس از آن‌که تعدادِ زیرگراف‌های تمامِ گروه‌های ایزومورفی یک راس تعیین شد، این راس و تمامِ یال‌های متصل به آن از شبکه حذف می‌شود و روندِ شمارشِ زیرگراف‌ها برای ریوسِ دیگر تکرار می‌شود. به این ترتیب، تمامِ زیرگراف‌های از اندازه‌ی k در این شبکه شمرده می‌شود.

برای انجامِ مرحله‌ی سوم(یعنی مقایسه‌ی تعداد زیرگراف‌ها با شبکه‌های تصادفی)، نیازمندِ آن هستیم که به تعدادِ قابلِ ملاحظه‌ای شبکه‌ی تصادفی تولید کنیم. شبکه‌های تصادفی تولید شده می‌بایست دارای دنباله‌ی درجاتی مشابهِ شبکه‌ی اصلی باشند تا دارای همان ویژگی *scale free* باشند. پس از تولیدِ این شبکه‌ها، الگوریتمِ شمارشِ تعدادِ زیرگراف‌ها را بر روی آن‌ها اجرا می‌کنیم. حال با در اختیار داشتنِ تعدادِ زیرگراف‌ها در شبکه‌ی اصلی و شبکه‌های تصادفی، می‌توانیم معیارِ آماری مناسبی به‌کار برده و آن دسته از زیرگراف‌ها را که در شبکه‌ی اصلی نسبت به شبکه‌های تصادفی به میزانِ قابلِ توجه‌ای بیشتر مشاهده شد، به عنوانِ موتیف‌های پیدا شده، معرفی می‌کنیم. معیارِ آماری ذکر شده، با توجه به فرمولِ زیر محاسبه می‌شود:

$$Z_{Score}(Subgraph_i) = \frac{N_{real}(Subgraph_i) - \langle N_{random}(Subgraph_i)\rangle}{STD(N_{random}(Subgraph_i))}$$

نتایج

این الگوریتم را بر چندین شبکه‌ی تنظیمِ ژن که موردِ آزمایشِ الگوریتم‌های پیشین نیز قرار گرفته‌اند، اجرا کرده و به نتایجِ مشابهی در زمانِ مناسب رسیدیم. این نتایج منطبق بر نتایجِ الگوریتمِ FANMOD(11) در حالتی که به صورتِ دقیق شمارش انجام دهد بود. از آن‌جایی که در این الگوریتم، پس از شمارشِ زیرگراف‌های شاملِ هر راس، آن راس حذف می‌شود، این امر کمکِ شایانی به کاهشِ تخصیصِ حافظه می‌کند و در به‌کارگیری آن در شبکه‌های بزرگ بسیار کارا خواهد بود.

همان‌طور که گفته شد، این الگوریتم را بر شبکه‌های پیچیده‌تری مانند شبکه‌ی تنظیمِ ژنِ *ecoli*، موردِ آزمایش قرار دادیم و به نتایجِ مناسبی دست یافتیم. نتایجِ این آزمایش برای موتیف‌های از اندازه‌ی 3 در جدولِ زیر قرار دارد که *Zscore* آن‌ها نیز به عنوانِ معیارِ آماری آمده است.

جدول 1. نتایجِ آزمایش برای موتیف‌های از اندازه‌ی 3 بر شبکه‌ی تنظیمِ ژنِ *ecoli*

موتیف							
Zscore	748	26.52	31.75	11	0.5	286	652

منابع

[1] F.d'Alch´e–Buc, V.Schachter: Modeling and identification of biological networks. Proc. Intl. Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France, 2005, pp.167-179.

[2] Kanehisa, http://www.genome.jp/kegg/, 1995.

[3] N.Kashtan, S.Itzkovitz, R.Milo, and U.Alon, Mfinder tool guide. Technical report, Department of Molecular Cell Biology and Computer Science and Applied Mathematics, Weizman Institute of Science, Israel. 2002.

[4] N.Kashtan, Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics, 20, 2004, 1746–1758.

[5] S.Maslov, K.Sneppen, and U.Alon, Correlation profiles and motifs in complex networks. In Bornholdt,S. and Schuster,H.G. (eds), *Handbook of Graphs and Networks: From the Genome to the Internet.* Wiley-VCH, Berlin, 2003.

[6] R.Milo, S.Shen-Orr, S.Itzkovitz, N.Kashtan, U.Alon, and D.Chklovskii, Network motifs: simple building blocks of complex networks. Science, 298, 2002, 824–827.

[7] F.Schreiber, and H.Schwöbbermeyer, Towards motif detection in networks: frequency concepts and flexible search. In *Proc. Intl. Wsh. Network Tools and Applications in Biology (NETTAB'04)*, 2004, pp. 91–102.

[8] F.Schreiber, and H.Schwo¨bbermeyer, Mavisto: a tool for the exploration of network motifs. Bioinformatics, 21, 2005, 3572–3574.

[9] S.Wernicke, A faster algorithm for detecting network motifs. In Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI ’05), Lecture Notes in Bioinformatics. Vol. 3692, 2005, pp. 165–177.

[10] S.Wernicke, A faster algorithm for detecting network motifs. In Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI ’05), Lecture Notes in Bioinformatics. Vol. 3692, 2005, pp. 165–177.

[11] S.Wernicke, and R.Florian, FANMOD: a tool for fast network motif detection. Bioinformatics, 22, 2006, 1152-1153.

WWW.IBP.IR iranian bioinformatics portal