Vol.2009-BIO-17 No.14
2009/5/26

スペクトル法によるタンパク質相互作用ネットワークのモジュール分解

井上健太郎[†]　Weijiang Li[††]　倉田博之[†]

概要：タンパク質相互作用(PPI)ネットワークをモジュール分解するためにさまざまクラスタリング法が提案されているが、高速に正確なモジュールを発見する手法が確立されていない。本研究では、従来のスペクトル法に新規のベキ乗因子を加えることで、複雑なネットワークのわずかな幾何学的差を見いだし、クラスタに分解することができた。PPI に存在する生体機能モジュールを位相幾何学的に見つけ出すための手法を開発した。

Modular decomposition based spectrum for protein-protein interaction network

Kentaro Inoue[†], Weijiang Li[††] and Hiroyuki Kurata[†]

Abstract: A method for the division of a protein-protein interaction (PPI) network into functional modules is proposed, since there have been few methods for discovering functional modules at a high speed. In this study, we develop the new spectral analysis method that implements a power factor to identify the cluster structures of complex networks. It can separate the PPI networks into topologically meaningful modules with biological functions.

1. はじめに

タンパク質相互作用(PPI)ネットワークは細胞内の代謝機能や生物学的プロセスに関する重要な情報を担っている。ネットワークトポロジーから機能モジュールに分解することは PPI の全体的なイメージを理解するために役に立つ。PPI は一般的にタンパク質をノード、相互作用をエッジとするグラフに置き換えられる。ほとんどの PPI は次数の高いノードが少なく、次数の低いノードが多く，スケールフリー性を持つ。このようなネットワークをモジュールに分けるさまざまなクラスタリング法が提案されているが、高速に正確なモジュールを発見する手法が確立されていない。これまでのクラスタリング法では、多数のノードで構成される巨大クラスタが少数作られる一方で、少数のノードから構成されるクラスタが多数表れるという結果となっていた。

スペクトルクラスタリング法（以下、スペクトル法とする）は、対象となるデータに対して、そのスペクトルを解析するために、クラスタの数に応じた次元の固有ベクトルを求め、クラスタリングを行う。クラスタ構造が明確でない対象では、この固有ベクトルを角距離に置き換えることで、改善される[1]。本研究では PPI を分子がネットワーク中を拡散するという拡散方程式に基づいたスペクトル法に[2]、新規のベキ乗因子を加えることによって、複雑なネットワークのわずかな幾何学的差を見いだし、従来のスペクトル法より明確に生体機能モジュールに分解する手法を開発した。

2. 方法

大規模な生体分子ネットワークのクラスタリングでは、計算速度、クラスタの疎密性、機能分類評価が重要となってくる。提案した手法と比較をするために、Markov Clustering(MCL) [3]と Shortest Path Betweenness(SPB) [4]、従来のスペクトル法を用いる。MCL は大規模なネットワークに対しても計算速度が速いことで知られている。SPB は分割的手法でエッジを取り除く際に、ネットワークの疎密性を計算しながら、最も密となるような状態を導き出すアルゴリズムである。MCL とスペクトル法は MATLAB で、SPB は C 言語で計算を行った。

[†]九州工業大学大学院
Department of Bioscience and Bioinformatics, Kyushu Institute of Technology
[††]江南大学
School of Biotechnology, Southern Yangtze University

Vol.2009-BIO-17 No.14
2009/5/26

2.1 スペクトル法

分子はネットワークを拡散すると考えることができる。分子の拡散方程式は以下のようになる。

$$\frac{d}{dt}\mathbf{p}(t) = -\Gamma\mathbf{p}(t) \quad \cdots(1)$$

$\mathbf{p}(t)=(p_1(t),p_2(t),\ldots,p_n(t))^T$ はネットワークのノード上に分子が存在する確率分布である。Γ は拡散係数行列であり、グラフラプラシアンを以下のように変形する。

$$\mathbf{V} = \mathbf{D}^{\beta/2}\Gamma\mathbf{D}^{-\beta/2} = \Delta - \mathbf{D}^{-\beta/2}\mathbf{A}\mathbf{D}^{-\beta/2} \cdots(2)$$

A は隣接行列。D はノードの次数 d_i の対角行列、Δ は $d_i^{-\beta+1}$ の対角行列である。β は本研究で新しく導入したべき乗因子である。β=1 のとき、従来のグラフラプラシアンとなる。β の値を $1<\beta<2$ の範囲で変化させることにより、明確なクラスタ構造を持つ距離空間を実現する。V について、スペクトル分解し、変形することで、拡散方程式(1)は

$$\mathbf{p}(t) = \sum_{i=1}^{n} e^{-\lambda_i t}\mathbf{u}_i\mathbf{u}_i^T\mathbf{D}^{\beta}\mathbf{p}(0) \quad \cdots(3)$$

となる。求めた固有値を昇順に並べる。

$$0 = \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n \quad \cdots(4)$$

クラスタを k に分けるとき、

$$\mathbf{x}_i = (u_{2i}, u_{3i}, \ldots, u_{ki})^T, \forall i \quad \cdots(5)$$

となる k-1 次元ベクトル $\mathbf{x}_i$ を角距離に置き換えて、完全連結法と k-means によりクラスタリングをする。

2.2 MCL

グラフ内のランダムウォークにおいて、expansion と inflation を繰り返したときに各ノードを通る遷移確率を利用してクラスタを見つけ出す。

2.3 SPB

すべてのエッジに対して、edge betweenness を計算し、最も edge betweenness が高いエッジを取り除く。取り除かれたネットワークから、再度 edge betweenness を計算し、エッジがなくなるまで繰り返す。クラスタ構造が最も密になったときのネットワークをクラスタリング結果とする。

2.4 評価法

評価は、計算速度、クラスタの疎密性、機能分類の有意性で行う。クラスタがどの程度密になっているかは Modularity で評価する[4]。

$$Modularity = \sum_i (e_{ii} - (\sum_j e_{ij})^2) \cdots(6)$$

e_{ij} はエッジの両端のノード(a,b)がクラスタ(i,j)に存在している数である。

クラスタが機能モジュールに分かれていることを Gene Ontology の注釈(Biological Process、Molecular Function、Cellular Component)に対する p 値で評価する[5]。

$$p - value = \sum_{i=m}^{n} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \quad \cdots(7)$$

N はノードの総数、M は注釈 A を持つ分子(ノード)の数、n はクラスタ内のノード数、m はクラスタ内の注釈 A のノードの数である。

また、ネットワーク全体の評価として以下を定義する。

$$ClusteringScore = 1 - \frac{\sum_{i=1}^{n_S} \min(p_i) + (n_I \times cutoff)}{(n_S + n_I) \times cutoff} \cdots(8)$$

n_S は p<cutoff となるクラスタの数、n_I は p>cutoff となるクラスタの数、cutoff=0.05 とする。

2.5 検証モデル

Database of Interacting Proteins (DIP)から出芽酵母(*Saccharomyces cerevisiae*)のコアネットワークであるノード数 4902、エッジ数 17246 を用いた[6]。出芽酵母の PPI はスケールフリー性を持つ(Fig.1)。平均次数は 7.04、クラスタ係数は 0.126 である。

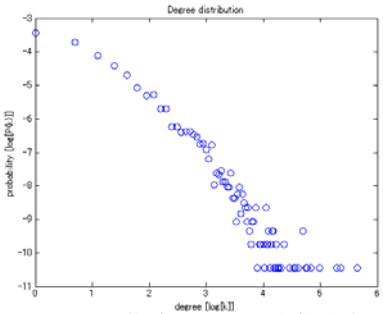

Fig. 1:出芽酵母 PPI の次数分布

Vol.2009-BIO-17 No.14
2009/5/26

3. 結果・考察

従来のスペクトル法とβを導入した新規手法では、βを導入した手法のほうが高い Modularity をもつクラスタを生成している(Fig.2)。βの値を変えることによって、ネットワークの疎密性の差をより明確にする距離空間が作られたと考えられる。クラスタ数を３３としたとき､完全連結法を用いた新規スペクトル法､従来のスペクトル法、SPB のクラスタサイズを比較した。SPB ではクラスタサイズに広い分布が生じた。新規スペクトル法は従来の手法に比べ、クラスタサイズがほぼ同じなることがわかった(Fig.3)。各クラスタリング法のクラスタサイズの変動係数(CV)を計算すると、SPB が最もばらつきが大きく、新規スペクトル法が最もばらつきが小さかった (Table 1)。MCL と SPB では１つのノードからなるクラスタが多数存在した。１つのノードをクラスタでは、モジュールの機能注釈づけの議論はできない。スペクトル法ではクラスタサイズにばらつきが少なく､機能注釈に対しても有意性を示さない(p>0.05)クラスタは現れなかった。ネットワーク全体の機能注釈としての評価 Clustering Score の値に対してもスペクトル法は最もよい値を示した。

Table 1: クラスタリング結果の比較

	計算時間	Modularity	クラスタ数	Clustering Score	n_s	n_l	クラスタサイズ
MCL	1750秒	0.275	1233	BP 0.913 MF 0.838 CC 0.732	1196 1110 1054	37 123 179	1~83 (CV=1.22)
SPB	約1ヶ月	0.506	319	BP 0.943 MF 0.914 CC 0.779	316 308 286	3 11 33	1~753 (CV=2.49)
Spectrum (β=1.4)	125秒	0.502	33	BP 0.997 MF 0.994 CC 0.999	33 33 33	0 0 0	48~452 (CV=0.48)
(β=1)		0.436	33	BP 0.994 MF 0.984 CC 0.973	33 33 33	0 0 0	31~591 (CV=0.83)

BP: Biological Process MF: Molecular Function CC: Cellular Component

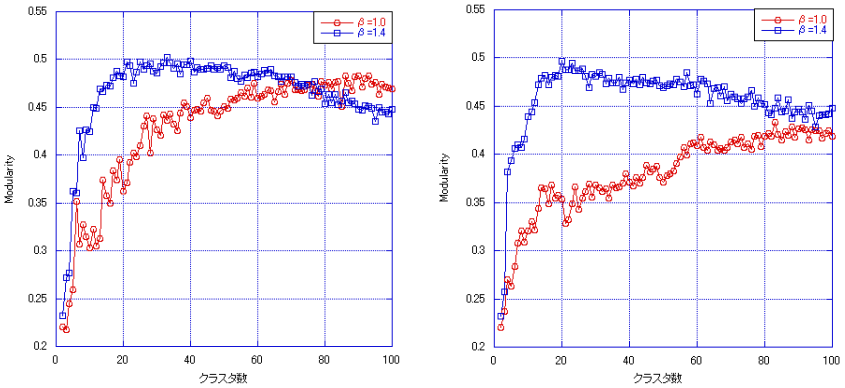

Fig.2 従来のスペクトル法(赤)と新規スペクトル法(青)の Modularity 比較
― (左) 完全連結法 (右) kmeans 法

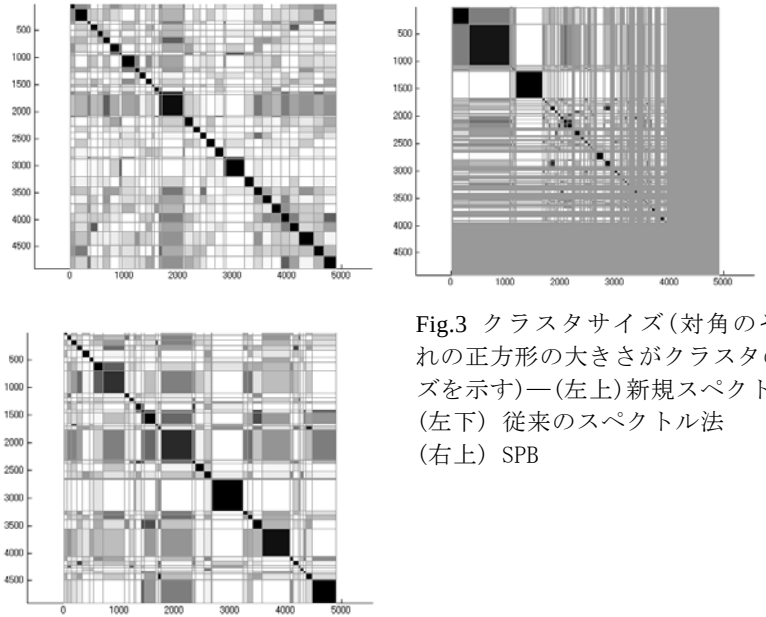

Fig.3 クラスタサイズ(対角のそれぞれの正方形の大きさがクラスタのサイズを示す)―(左上)新規スペクトル法
(左下) 従来のスペクトル法
(右上) SPB

Vol.2009-BIO-17 No.14
2009/5/26

4. 結論

従来、PPI をほぼ均等なサイズをもつクラスタに分割することは困難であったが、スペクトル法によって得られる幾何学的ノード座標を角距離でクラスタリングすることによって、その問題はかなり解決された。さらに、β を導入することにより、通常のスペクトル法(β=1)より、明確なクラスタ構造が見つけ出され、生物学的評価や計算速度においても、MCL や SPB より良い結果となった。このことから、クラスタ構造が明確でない PPI ネットワークにおいて、提案したスペクトル法は有効である。今回、提案した手法は PPI に限らず、さまざまなネットワークにも応用が期待される。

5. おわりに

スペクトル法を用いてクラスタ数を決めるための指標が提案されているが [2] 、PPI のような複雑なネットワークではクラスタ数を明確に決定することは難しい。スペクトル法は、PPI ネットワークをいくつのクラスタに分けても、クラスタサイズのばらつきが少なく、Modularity が高い (Fig.2,Fig.4)。本稿ではクラスタ数を Modularity に最も高い値を与えるクラスタ数と考えたが、今後、クラスタ数を決めるための手法の開発が課題となる。

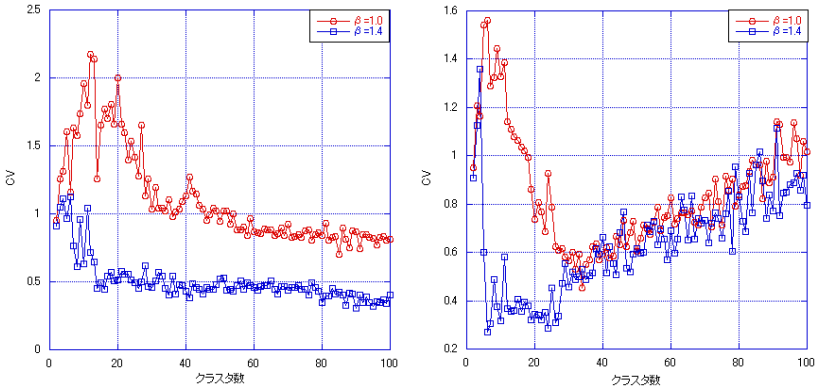

Fig. 4 スペクトル法のクラスタ数と変動係数の関係
— (左) 完全連結法 (右) kmeans

参考文献

1) Fischer I. (2004) New methods for spectral clustering. IDSIA-12-04.
2) Nadler B., Lafon S, Coifman RR, Kevrekidis IG (2005) Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck operators. Arxiv preprint math: NA/0506090.
3) van Dongen,S. (2000) Graph clustering by flow simulation. Centers for mathematics and computer science (CWI), University of Utrecht. Amsterdam, 371-382.
4) Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys 69: 026113
5) Sitaram Asur, Duygu Ucar and Srinivasan Parthasarthy (2007) An ensemble framework for clustering protein-protein interaction networks. Bioinformatics 23: i29-40
6) Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins. Nucleic Acids Res 32: D449-451.