社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－BIO－5 (2)
2006／6／15

生体ネットワーククラスタの可視化に関する研究

辻尚,Md.Altaf-Ul-Amin(*),有田正規,西尾泰和(**),真保陽子,黒川顕,金谷重彦(*)
(*)奈良先端科学技術大学院大学(**)東京大学

タンパク質相互作用のネットワーク,ゲノムネットワーク等の生体ネットワークは,その複雑さゆえに解析に困難を伴うことが多い.特にネットワーク全体の構造の特徴を捉えることは難しく,ネットワーク解析のために有用な可視化手法の提案が望まれている. 本研究では, エッジが相対的に密になっている,クラスタと呼ばれる領域に着目した大規模ネットワークの可視化手法を提案する.さらに,その手法を生物学分野へ応用して様々なネットワークのトポロジーを特徴づけることを目指す.

Visualization of clusters in bio-networks

H. Tuji, M.A. Amin(*),M. Arita,H. Nishio(**), Y. Shinbo, K. Kurokawa, S. Kanaya(*)
(*) Nara Institute of Science and Technology (**) Tokyo University

Many large scale networks in biology, for example, protein-protein interaction networks or genome networks pose many difficulties in their analysis, due to their complexity. It demands an efficient visualizing method to allow us to analyze large scale networks. In this paper, we proposed a visualizing method focusing cluster structure. In this work, a cluster is defined as relatively densely connected nodes. Finally, we demonstrate that this method has potential to application in biology. The goal of this work is to facilitate analyzing different network classes by means of visualization.

1. はじめに

近年コンピュータの性能の向上により,これまで解析が困難であった大規模かつ複雑なネットワークの研究が盛んに行われるようになった.大規模ネットワークの種類は様々で,コンピュータの世界では World Wide Web を代表とするインターネット,生物学の世界ではタンパク質相互作用のネットワーク,ゲノムネットワーク,他にも食物ネットワーク,論文引用ネットワーク,社会ネットワーク等多岐に渡る[1-6].これらのネットワークはしばしばノードとエッジからなるグラフとして表現される.本研究の研究対象の１つであるタンパク質相互作用のネットワークはタンパク質をノード,タンパク質間の相互作用関係をエッジに対応させて表現される.タンパク質相互作用

のネットワークは生体タンパク質機能予測の有力なツールになり得ると期待されており,ネットワーク構造解析のために有用な可視化手法の考案が必要と考えられている.本研究室では,互いに密に連結したノードの集合をクラスタと定義し,この集合体を抽出(クラスタリング)するプログラムを開発しつつある[7].一方,ネットワーク描画における審美性についての研究も以前から行われている.例えばBatiniらは人が「良い」と判断するネットワーク描画基準を規定している[8].しかし,大規模ネットワークはノード,エッジが多数存在し,ネットワーク全体を描画するときに審美性を考えることは難しい.そこで,本研究ではクラスタリング技術を応用したネットワーク可視化手法を提案する.様々なクラスタリングの手法が提案されている[7].ここでは,クラスタリング手法としてNewmanのアルゴリズム[9]を用いる.Newmanのアルゴリズムはネットワークノード数Nに対して$O(N^2)$の計算時間でクラスタリングを行う高速手法である.可視化を行うソフトウェアとしてGRINEditCを用いる.これは当研究室で開発されたGRINEdit[10]を大規模ネットワーク描画用に改良したものである. この論文ではGRINEditCの大規模ネットワークトポロジー解析への応用性及びタンパク質相互作用のネットワークの機能予測ツールとしての可能性を示す.

2. 諸定義

本研究では,グラフをノード集合及びノードの非順序対$\{u,v\}$で規定されるエッジの集合で定義する.ノード間にエッジは多くとも1本存在する.クラスタリングとは,グラフのノード集合をエッジが密に張られたクラスタと呼ばれる領域ごとに分割することである(図.1参照).なお,本研究ではエッジの向き,セルフループ,多重エッジは考慮しない.すなわち単純グラフを扱う.

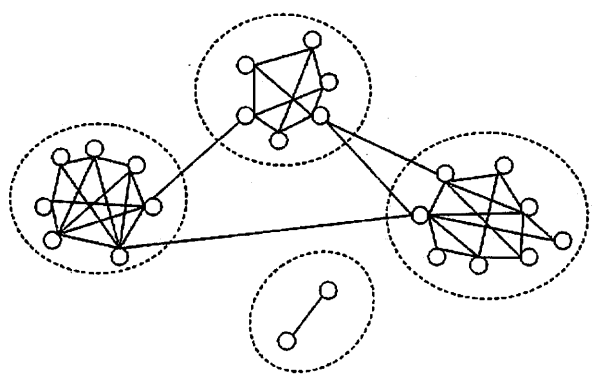

図1.グラフのクラスタリング

3. Newmanのクラスタリングアルゴリズムの概要

本研究ではクラスタリングアルゴリズムとしてNewmanの手法を用いる.Newman法はモジュール度の考え方を基本とし,クラスタリングの良さを表すモジュール度関数としてQを定めている.Qの具体的な定義は後述する.はじめに,グラフの全てのノードをそれぞれ1個のクラスタとする.e_{ij}を,クラスタiとクラスタjの間に存在するエッジ数が,グラフ全体のエッジ数に占める割合(辺分率)の半分の値とする.クラスタi,j間に存在するエッジの辺分率は$e_{ij}+e_{ji}$で表される.ただし,対角成分のe_{ii}については,クラスタi内のエッジの辺分率に等しいとする.Newmanのアルゴリズムは現在のQの値が増加するようにクラスタの結合を繰り返し,最適クラスタリングを探索するgreedyなアルゴリズムである.Σe_{ii}は全てのクラスタ内に含まれるエッジの辺分率を表し,最大値は1,最小値は0である.直感的にはモジュール度関数Qとしてこの指標を用いることが良いと考えられるが,全てのノードが1つのクラスタに含まれるとしたとき,これは明らかに良いクラスタリングではない.そこで,クラスタ分割はそのままで,全エッジをランダムに張り直した場合の全クラスタ内のエッジの辺分率をΣe_{ii}から引くことにより,この問題は解決できる.a_iをクラスタi内にあるエッジの終端数が,全体のエッジの終端数に占める割合とする.このとき,エッジをランダムに張り直した場合のクラスタiに属するエッジの辺分率はa_i^2で表される.すなわちモジュール度関数Qは以下の形で表される.

$$Q=\sum(e_{ii}-a_i^2)$$

Qの最適化には指数オーダー以上の時間が必要となることが知られている[9].そこでNewmanらはクラスタリング更新時のQの増加量ΔQが以下の式で表されることを示し,この増加量が0になるまで最大のΔQを与えるクラスタの結合を繰り返す手法を提案した.

$$\Delta Q=e_{ij}+e_{ji}-2a_ia_j=2(e_{ij}-a_ia_j)$$

この式はネットワークノード数Nに対して$O(N)$の時間で計算できる.クラスタの結合回数は最大で$N-1$回であるから,疎なネットワークではクラスタリングは$O(N^2)$時間で可能である.

4. 大規模ネットワーク可視化手法

大規模ネットワークをクラスタリングすることにより,1個のクラスタを1個のノード(クラスタノード)とする.あるクラスタ内の任意1個のノードと他のクラスタ内の任意1個のノードからな

るエッジ(クラスタ間エッジ)が1本以上存在する場合,そのクラスタ間を結ぶ1本のエッジ(クラスタエッジ)を定める.これにより,クラスタノード集合とクラスタエッジ集合からなるグラフ(クラスタグラフ)を定義することができる.クラスタノードの内部についてさらにクラスタリングを行い,クラスタグラフを作成すれば,大規模ネットワークの階層描画が可能となる.本研究では,クラスタグラフ描画及び階層描画に特化した GRINEditC による大規模ネットワーク描画を試みる.GRINEditC の具体的な説明を以下で述べる.

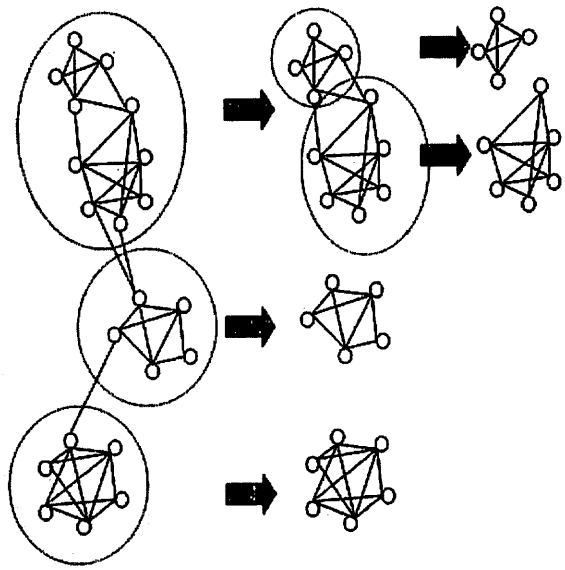

図 2. クラスタリングによる階層描画のイメージ

5. GRINEditC

GRINEditC は大規模ネットワークを階層的に描画することでその全体像を可視化するソフトウェアである.実装には Java を用いている.現在はタンパク質相互作用のネットワークの解析に焦点を当てているため,タンパク質相互作用のネットワークの描画に特化した仕様となっている.GRINEditC の特徴について述べる.

1. ノード間の隣接関係と,ノードに対する付属情報を含むデータにより入力ネットワークを規定する.

2. Newman のアルゴリズムを参考に入力ネットワークをクラスタリングし,クラスタグラフを作成して,ウィンドウに描画する.クラスタノードはそのクラスタ内に含まれるノード数に比例して描画サイズが大きくなるように視覚化される.また,クラスタエッジについても,含まれるクラスタ間エッジの本数に比例して太くなるように視覚化される.
3. クラスタノードにはクラスタノード間の反発力を,クラスタエッジにはバネの性質を与え,それぞれの配置位置の再計算及び再描画を繰り返す.これにより,クラスタグラフが物理的に最も安定した配置へと移動する過程を動画として確認できる.
4. メニュー項目"reclustering"を選択し,クラスタノードをクリックすると,そのクラスタ内でクラスタリングアルゴリズムが適用され,そのクラスタのクラスタグラフが別ウィンドウで描画される.なお,リクラスタリング計算の際に,そのクラスタ内のノードと接続するクラスタ間エッジは無視される.
5. 入力ネットワークが生体ネットワークであるときに"reclustering"メニュー選択後クリックしたクラスタに含まれるノードが 1 個である場合,その 1 個のノードが示す遺伝子又はタンパク質等のデータベース ID や機能といった生物学的情報がポップアップされる.
6. 入力ネットワークが生体ネットワークである場合,クラスタ内の全ノードについて機能数をカウントし,クラスタノードは機能数比率の円グラフとして描画される.なお,1 個のノードに複数の機能が存在する場合,全てカウントし,機能未知のものは"Unknown"としてカウントする.さらに,クラスタノードの円グラフの中心に"クラスタ内未知機能ノード数／クラスタ内全ノード数"を描き,クラスタ内に含まれる未知機能のノード数をわかりやすいように表示しておく.

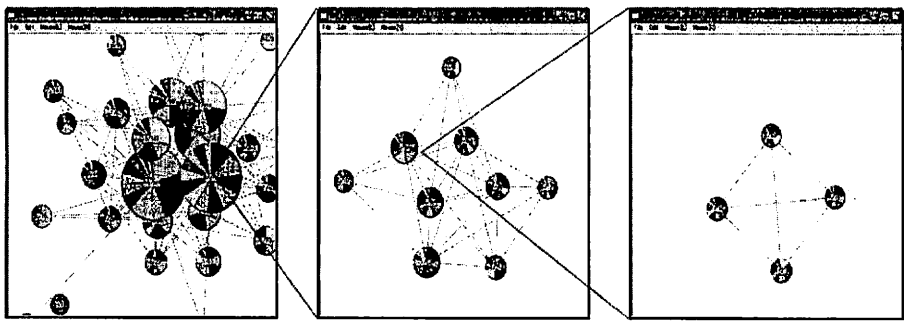

図 3.GRINEditC.左の図の円グラフはクラスタを表し,その大きさはクラスタに含まれるノード数に比例した大きさになる.クラスタノードをクリックするとそのクラスタノード内の要素の結合関係が視覚化される（中央図）,さらに,クラスタノードをクリックするとそのクラスタノード内の要素の結合関係が視覚化される(右図)という形式でネットワーク構造を把握することができる.

6. GRINEditC の有用性

GRINEditC の最大の利点は,大規模ネットワークの構造特性を見やすく可視化できることである.このソフトウェアによってネットワークトポロジーの特性を捉えるアプローチが可能になる.本研究で提案した手法は,クラスタリングを繰り返す過程を木構造ととらえ, バイオネットワークをはじめとした大規模ネットワークを構造から分類することが可能になると期待される.また,GRINEditC はタンパク質相互作用のネットワークにおいても非常に有用なツールになりうる.まず,クラスタリングを繰り返した場合に同一クラスタに属するタンパク質ノード群は,互いに密接に相互作用することを示しており,このことから,細胞内のタンパク質の機能単位の抽出ができる可能性がある.

本研究では,全体のノードの結合関係を基にクラスタを定義しているため,全体におけるタンパク質間の密接な結合状態を抽出できる.このことは,クラスタリングを繰り返した場合のノードとノードの相互作用関係の「切り離し」に注目すると,「切り離し」がおこらない要素間は,互いに密に連結されていることを示しており,二つのタンパク質のみの結合の有無だけでの妥当性に比べて,全体の構造で妥当させる強い関係性を見出すことができる.

本システムでは,タンパク質相互作用のネットワークのクラスタグラフを描画及びクラスタ内のタンパク質の分類される機能の比率による円グラフを描画することができる.円グラフにより,クラスタノードの機能分類との関係が把握できる. 例えば,あるクラスタ内で 9 割のタンパク質が機能 A をもち,残り 1 割のタンパク質が機能未知であった場合に,この 1 割のタンパク質は機能 A を持つ可能性があると考えることは自然であり,これは機能比率の円グラフから視覚的に読み取れる可能性が高い.参考に DIP[11]の酵母菌のタンパク質相互作用のネットワークの描画を GrinEditC で行った際に,大部分のタンパク質が物質輸送の機能をもつクラスタが存在することが確認された.これは,このクラスタ内の未知タンパク質も物質輸送の機能をもつことを示唆する.

7. 今後の課題

様々な大規模ネットワークデータを収集し,トポロジーによるネットワーク分類の研究へ展開し,バイオネットワーク構造のそれぞれの特異性ならびに共通性を検討することを計画している.

参考文献

[1] S.H. Strogatz.: Exploring complex networks. *Nature(London)* Vol.410,pp. 268-276 (2001).

[2] M.E.J. Newman.:The Structure and Function of Complex Networks. *SIAM, Rev.* Vol.45, No. 2 pp.167-256(2003).

[3] R. Albert and A.-L. Barabasi.:Statistical mechanics of complex networks. *Rev. Mod. Phys.*. Vol.74, No. 1,pp.44-97(2002).

[4] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas.:Self-similar community structure in a network of human interactions. *Phys. Rev. E* Vol.68 065103(2003).

[5] D. Wilkinson and B. A. Huberman.: A method for finding communities of related genes. *Proc. Natl. Acad. Sci. U.S.A.* Vol.101, pp.5241-5248(2004).

[6] P. Holme, M. Huss, and H. Jeong.: Subnetwork hierarchies of biochemical pathways. *Bioinformatics* Vol.19, No. 4 pp.532-538(2003).

[7] M.A. Amin, Y. Shinbo, K. Mihara, K. Kurokawa, S.Kanaya. :Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*,Vol 7(2006).

[8] Batini C. Furlani L. and Nardelli E.: What is a Good Diagram?-A Pragmatic Approach, Proc. 4th Int. Conf. on the Entity Relationship Approach, Chicago, pp.312-319(1985).

[9] M.E.J. Newman.: Fast algorithm for detecting community structure in networks. *Phys. Rev. E* Vol.69, 066133(2004).

[10] GRINEdit. http://www.nishiohirokazu.org/grinedit/.

[11] DIP. http://dip.doe-mbi.ucla.edu/.