「平安京ビュー」を用いた階層型遺伝子ネットワークの可視化

西山　慧子*　　伊藤　貴之**
（*）　お茶の水女子大学大学院　人間文化研究科
（**）　お茶の水女子大学　理学部情報科学科
E-mail：{nishy, itot}@itolab.is.ocha.ac.jp

概要

遺伝子ネットワークとは、各遺伝子をノードとし、遺伝子間をエッジで接続して構築されるデータである。数千単位の遺伝子群で構成される遺伝子ネットワークには、複雑な連結成分を含むことが多く、その解釈や把握が困難な場合も多い。

本報告では、遺伝子群にクラスタリングとネットワーク化を同時に適用して構築される、階層型ネットワークデータを対象とした可視化手法を提案する。本手法では、各々の遺伝子は数種類の発現率を持つと仮定し、その発現率の相関性の高さによりクラスタリングを行う。それと同時に本手法では、発現率の相関性の高い遺伝子間をエッジで連結することにより、ネットワークデータも同時に生成する。本手法は、このようにして生成されたデータを、大規模階層型データ可視化手法「平安京ビュー」を用いて可視化する。本手法を用いることにより、遺伝子学の研究者は、膨大な遺伝子群の中から、特定の遺伝子の相互関係を分析、あるいは興味深い特徴を持つ遺伝子の発見、などが容易になるものと考えられる。

なお本報告は遺伝子ネットワークの可視化を試みるものであるが、提案手法は拡大性とランダム性の高い複雑ネットワークと呼ばれるネットワーク全般に適用可能な、応用範囲の広い可視化手法である。

1. はじめに

情報可視化は世の中にある一般的な情報を可視化する研究分野である。その応用範囲は非常に広いが、最近では特に、生物情報の可視化の研究が活発に進められている。生物情報の中でも特に急速に研究が進んでいる分野に、遺伝子（ゲノム）解析が挙げられる。現在すでに、ヒトゲノム解読は完了していると言われているが、これは DNA を構成する塩基配列が解読されたというだけであり、その遺伝子の振る舞い等は、はっきり分かっていない。そこで現在その遺伝子の振舞いについての研究が必要とされている。その中でもマイクロアレイデータ[1]からの遺伝子ネットワーク同定問題は、バイオインフォマティックス分野における重要なトピックの一つであると言える。

遺伝子ネットワークとは、各遺伝子をノードとし、遺伝子間にエッジがあるようなグラフ構造で、ゲノム上での位置関係、代謝、制御パスウェイ上での隣接関係、転写時の共発現率、蛋白質相互作用など、多くの性質を表現するために用いられる。遺伝子ネットワークは多くの場合において無向グラフとして扱われるが、パスウェイ等の遷移関係を表す場合に限って有向グラフとして扱われる。このようなグラフ構造を分析することで、遺伝子が変異したときに何が起こるか予測することができる。

遺伝子ネットワークは大変膨大なものであり、複雑な連結成分を含んでおり、そのままでは解釈や把握が困難である。よって、何らかの方法でより興味深い遺伝子群を抽出し、注目すべき対象を絞り込むことが必要である。しかしながら常に目的に叶った結果を切り出せていないのが現状である。情報可視化はこのような目的において非常に有効であると考えられる。

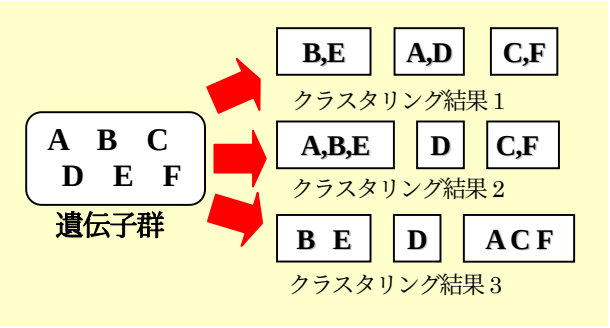

図１：クラスタ生成の一例

本報告では、マイクロアレイデータを参照して各遺伝子に数種類の発現率を仮定し、この相関性の高さで遺伝子をクラスタリングし、さらに相関性の高い遺伝子同士をエッジで接続した階層型ネットワークデータを想定する。このとき図 1 で示すように、クラスタリングの方法や実行条件によって、クラスタリング結果は様々に変化する。図 1 より、A は

{B,E},{D},{C,F}の 3 組の遺伝子と同一のクラスタに属する可能性があると言える。このことより、A は複数の遺伝子の機能を同時に持つ遺伝子かもしれない、と予測できる。このように、2 種以上の遺伝子の機能を同時にもつ遺伝子は、マルチドメインと呼ばれ、遺伝子分析の中でも興味深い問題である。しかし 1 つのクラスタリング結果だけを可視化しても、このような特性は発見しにくい。このような現象は、遺伝子クラスタリング結果と遺伝子ネットワークを組み合わせて可視化することにより、その存在が理解しやすくなると考えられる。

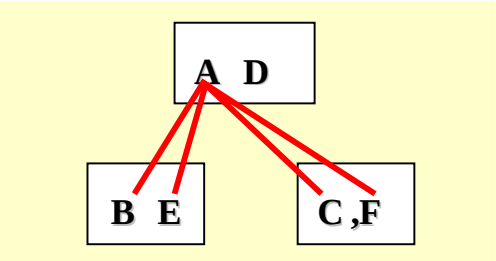

図 2：クラスタにネットワークを重ねた一例

本報告では、遺伝子群に対してクラスタリングとネットワーク化を同時に適用して生成される、階層型ネットワークデータの可視化手法を提案する。本手法は図 2 に示すように、異なるクラスタ間をまたいで相関性を有する遺伝子間をエッジで表現することで、マルチドメインに代表される遺伝子の興味深い現象の発見に貢献するものである。本手法では情報可視化手法「平安京ビュー」を用いてクラスタリング結果を階層型データとして可視化し、それにエッジを重ねて描くことにより階層型ネットワークデータを表現する。

なお本手法は 3.2 節にて後述するとおり、大規模かつランダム性の高い複雑ネットワーク全般に適用できる、極めて適用範囲の広い手法である。

2. 関連研究

2.1 大規模情報可視化手法「平安京ビュー」

日常生活に氾濫する情報の多くは、階層化された構造を持っている。計算機のファイルシステム、企業や大学の組織構造、図書館の書籍の分類、などはその典型的な例であろう。このように階層化された情報の全貌を、計算機のディスプレイで一望できたら、という要求は当然のように起こりえる。「平安京ビュー」は、そのような要求を満たす大規模情報可視化手法として提案されている。

「平安京ビュー」は、階層型データの葉ノードを長方形のアイコンで、枝ノードを長方形の枠で表現し、階層構造を 2 次元の長方形群の入れ子構造で表現し、その全体を一画面に表示する事を目標とした手法である。計算機のファイルシステムに例えるなら、葉ノードはファイルに、枝ノードはディレクトリに相当する。企業の組織構造に例えるなら、葉ノードは従業員、枝ノードは部・課・プロジェクトといった団体に相当する。

この手法は、階層型データ中の葉ノードと枝ノードの親子関係よりも、階層型データ全体に分布する葉ノード群を全て一画面に表現することに主眼をおいた視覚化手法である。

「平安京ビュー」で技術的に重要な点は、枝ノード群を表現する任意の大きさ・形状の長方形群を、限られた大きさの画面空間に有効に配置できるという点である。言い換えれば「平安京ビュー」は、計算機上の限られた大きさのウィンドウやディスプレイに、できるだけ大規模な情報を詰め込んで表現する技術、ということができる。この要件を実現するための処理手順については、文献[2]を参照して頂きたい。

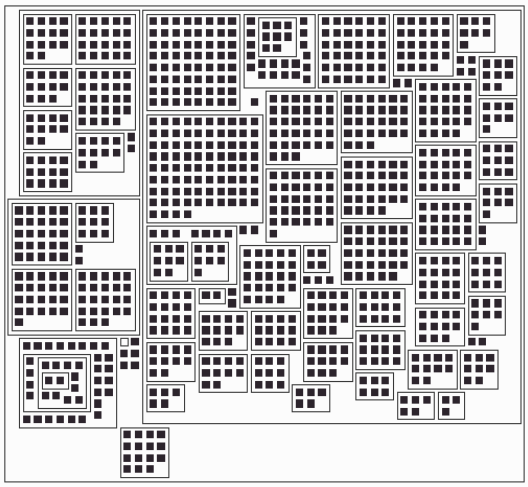

図 3. 「平安京ビュー」による大規模階層型データの可視化の例。

2.2 ネットワークデータ可視化

ネットワークの可視化手法には、すでに様々な手法が発表されている。一例として、頂点を稜線で連結したグラフ構造型データを理解する為、ウェブのリンク構造をグラフ表示した可視化手法[3]が報告されている。

ネットワークの可視化において重要な点に、ノードをどこに配置するか、という点がある。これを解くために、力学モデルを用いてノード位置を算出する可視化手法[4]が報告されている。

また木(tree)構造に特化した可視化手法も、すでに多数発表されている。Hyperbolic Tree[5]は、木構造を双曲面上に放射状に配置し、注目ノード周辺を対話的に拡大表示するようなビューを実現する。Cone Tree[6]は、円錐を用いて木構造を 3 次

元的に表現する。Tree-Maps[7]は、画面空間の 2 次元的な分割により、木構造を構成するデータ要素を一括表示する。

またネットワークをクラスタ化し、階層的に表示させる可視化手法も多く報告されている。一例として、階層型データを 3 次元で表し、あるクラスタから深さの同じクラスタを同じ高さで表現する可視化手法[8]、クラスタごとにズーム値を変えた階層型 2 次元可視化手法[9]などがある。

注目部分を強調表示する操作機能を備えたネットワーク可視化手法も、すでに多く発表されている。一例として、データ中の注目部分を、魚眼レンズでみたように拡大表示し、その周囲部分を画面の端によせる可視化手法が報告されている[10]。この手法の課題は、拡大表示部分の外にあるノードは、エッジで連結されていても画面の端に追いやられてしまうという点である。

大規模ネットワークを選択的に部分表示する手法として、ウォークスルーするイメージの可視化手法[11]や、インクリメンタルなグラフレイアウト手法[12]なども報告されている。しかし、ネットワーク全体を把握する事ができず、注目しているノードの全体での位置が把握しにくい。

また、データ中の注目部分を 3 次元的に引き上げ、それとエッジで連結されているノードも連鎖的に引き上げることで、注目ノード周辺相互関係を解りやすく表示する「納豆ビュー」という手法が報告されている[13]。本報告の提案手法は、納豆ビューに類似した手法であるといえる。

3. 提案内容

3.1 本報告の概要

本報告では 1 章にて述べたとおり、遺伝子群にクラスタリングとネットワーク化の両方を適用した階層型ネットワークデータの可視化手法を提案する。提案手法は以下の機能性を重視した手法である。

(1) できるだけ多くの遺伝子を一画面に、クラスタ単位で表示できること。

(2) 注目したい任意の遺伝子を強調でき、その遺伝子と相関性の高い遺伝子の分布を理解しやすいこと。

まず(1)を満たすために本手法では、平安京ビュー[13]を用いてクラスタごとにノードを画面配置する。このとき、各々のノードは遺伝子に対応する。これらのノードは画面上でクリッカブルな状態で表示されており、クリック操作によって遺伝子の詳細情報を提示できる。

続いて(2)を満たすために本手法では、注目ノードを 3 次元的に引き上げることにより、その注目ノードとエッジで連結されているノードの接続性を強調表示する。筆者らの実装では、注目ノードをクリックするか、または検索エンジンのようなキーボード入力による GUI で注目ノードを指定すると、その注目ノード、およびそのノードとエッジで連結されたノードを、3 次元的に引き上げて表示する。

一般的に、膨大な遺伝子群の中から、注目すべき興味深い遺伝子を視覚的に発見することは容易ではない。ここで本研究の目的において、クラスタ間をまたぐエッジを多く持つ遺伝子は、マルチドメインなどの興味深い現象をもつ遺伝子である可能性が高い。そこで本手法では、クラスタ間をまたぐエッジを一定以上有するノードを、あらかじめ所定の色で表示する。これにより、特殊な反応のありそうな遺伝子群を発見しやすくできる。

このような可視化を実現するために本手法では、遺伝子に対してクラスタリングとネットワーク化を同時に適用した階層型ネットワークデータを構築する。

3.2 階層型遺伝子ネットワークデータの構築

本手法は、M 個のマイクロアレイ上にそれぞれにある N 個の遺伝子、M×Nのマトリクス型データを対象とし、N個の遺伝子のうち発現率傾向の近いものをクラスタリングし、平安京ビューにおける階層型データに変換する。筆者らの実装では、これにネットワークデータを重ねることで、階層型ネットワークデータを可視化する。本手法におけるクラスタリングおよびネットワーク化の手順の概要を図 4 に示す。

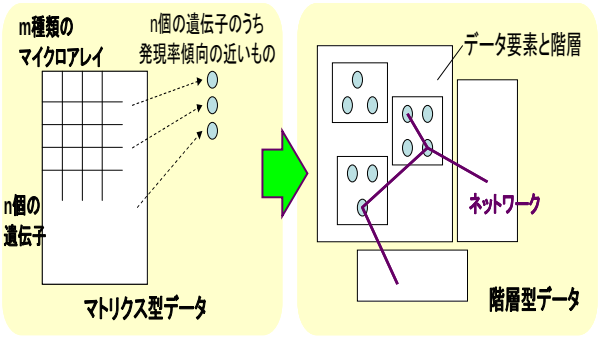

図 4：階層型データへの変換

筆者らの実装では、Cluster 3.0[14]というクラスタリングソフトウェアに実装されている階層的クラスタリングアルゴリズムを適用して、階層型データを構築する。図 5(上)において、クラスタを k_1～k_9 とすると、本手法では距離が近いクラスタに対して併合処理を反復することで、階層的クラスタリングを実現する。このときクラスタ間距離に複数の閾値を設け、この閾値より距離の小さいクラスタを一階層に収める、という処理を反復することで階層型データを構築する。仮に図 5(上)に示す S_1,S_2 の 2 つの閾値を設けたとすると、平安京ビュ

ーによる階層型データ可視化結果は図5(下)のようになる。

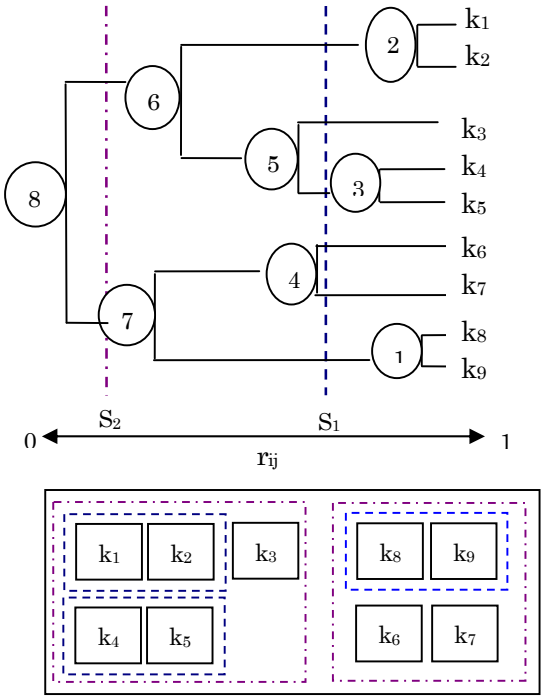

図 5. (上)階層的なクラスタリング。(下)平安京ビューにより表示されたクラスタ。

続いてネットワーク化の手順について説明する。任意の 2 個のノード（遺伝子）を nodeA, nodeB とし、m 種類のマイクロアレイに対する発現率が与えられているとする。さらに、

nodeA の発現率を $A=\{a_1, a_2, ...a_m\}$

nodeB の発現率を $B=\{b_1, b_2, ...b_m\}$

とする。このとき nodeA と nodeBの発現率同士の相関性 r_{ab} は、以下の式で算出する。

$$r_{ab} = 1.0 - \frac{d_{ab}}{D_{\max}} \quad (1)$$

ただし d_{ab} は A,B 間のユークリッド距離の 2 乗で、

$$d_{ab} = \sum_{i=1}^{m} (a_i - b_i)^2 \quad (2)$$

で示される。$D_{\max}$ は、すべてのノードの組み合わせにおける d_{ab} の最大値である。以上の算出式は、クラスタリングに使用したソフトウェア Cluster3.0[14] にも用いられている。

本手法では、r_{ab} 値が一定値より大きい時、この 2 つのノードを接続するエッジを表示する。

3.3 本手法の応用例

世の中には、様々なネットワークデータが存在する。本報告では、遺伝子をノードとし、相関性の高い遺伝子をエッジで連結するネットワークを対象としているが、このネットワークは近年注目されている「複雑ネットワーク」の一種であると考えられる。

複雑ネットワークとは、際限ない拡大性を有し、ランダム度の高いリンク構造を有するネットワークの総称である。特に近年では情報技術の発達により、例えば以下のような分野において複雑ネットワークが見られている。

- 文書データベースに出現するキーワード間の相関性から構築したネットワーク。
- 計算機のアクセス履歴、コンピュータウィルスの感染経路などのログから構築したネットワーク。
- ニューロンやタンパク質の情報伝達経路から構築したネットワーク。
- 会社や社会の人間関係における様々な人間関係のネットワーク。
- ウェブのリンク構造のネットワーク。

本報告の提案手法は、遺伝子ネットワークに限らず、上記のような複雑ネットワーク全般に適用可能な、応用範囲の広い手法であると考えられる。

4. 実行結果

4.1 結果画像

本報告による可視化の例を図 6 に示す。本報告の可視化例では、以下の URL に公開されている、イースト遺伝子発現率データを用いた。

http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/demo.txt

この可視化結果では、ある 1 つのノードを注視ノードとして指定したとき、そのノードと連結しているエッジだけを表示している。

図 6 では、クラスタ内のノードすべてと注視ノードが連結している、というクラスタを赤丸で囲んだ。この赤丸で示すようなクラスタが存在するということは、注視ノードは現在属するクラスタの他に、赤丸で示すクラスタに属していてもおかしくない、ということを示している。つまり、この注視ノードが示す遺伝子はマルチドメインかもしれない、ということが推測できる。

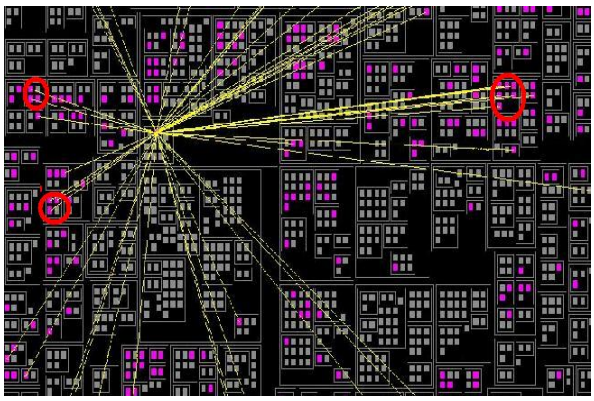

図６：本報告を用いた、注視ノードが一つの実行例。

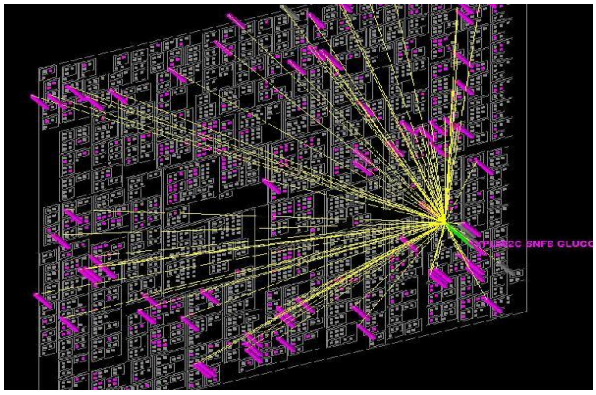

図７：注視ノードを１段階引き上げた表示画像。

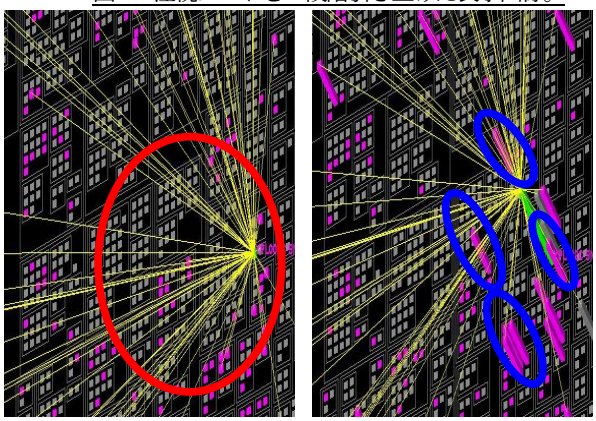

図８：（左）注視ノードを引き上げてない結果画像。
（右）注視ノードを引き上げた結果画像。

また、図６を詳しく調べてみると、所定の色（紫）で表示されたノードを両端とするエッジが多く存在していることが解る。このことより、図 6 に示す遺伝子ネットワークは、マルチドメインの可能性のある遺伝子同士が複雑に絡み合ったネットワークである、といえる。

図 7,8 は、本手法により、注視ノードと相関性の高いノードを３次元的に引き上げた結果画像である。図 8(左)の注視ノードを引き上げていない画像では、どのノードが注視ノードとエッジ連結されているのか、一目には理解しにくい。それに対して図 8(右)では、注視ノードを引き上げることにより、注視ノードとエッジで連結されたノードを一目瞭然に発見できる。これらの結果画像より、ネットワークの注視部分を 3 次元的に引き上げることにより、ノード間の連結関係が理解しやすくなると言える。

4.2 既存ソフトウェアとの比較

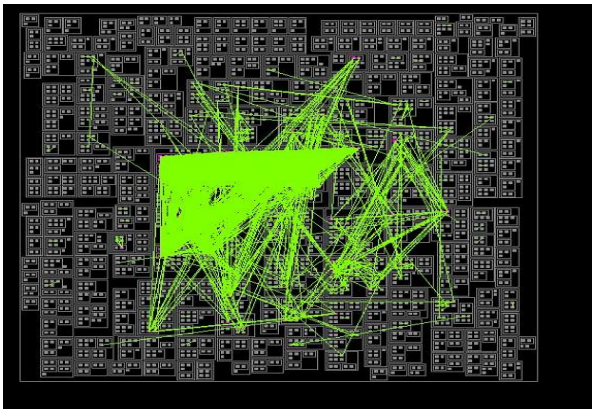

図９：平安京ビューにおける表示結果。

マイクロアレイデータから得られる遺伝子発現率情報の可視化ソフトウェアの中の多くは、ノード間の相互関係をエッジで結ぶ古典的なネットワーク 2 次元可視化手法や、TreeView[15]と呼ばれるクラスタリング結果の可視化手法を搭載しており、遺伝子分析に携わる多くの研究者がこれらを利用している。以下、これらの手法に対する提案手法の優位性について論じる。

まず前者の方法では、発現率の相関性の高いノードをエッジで結んで表示する事から、遺伝子間の関連性は一目瞭然である。しかし、一画面に表示するノード数は数十～数百程度に留まっている。またクラスタリング結果を同時に表示してはいない。それに対して本手法（図９参照）には、

- 整然と構造化された形で遺伝子群を表示する。
- 数千、数万といった膨大な量の遺伝子の分布の全貌を、一画面に一括表示できる。

といった点で利点があると考えられる。

続いて後者の TreeView は、N個の遺伝子に関する発現率を、N×N のマトリクスデータとして表現する。この手法は全ての

ノードの組み合わせに対する相関性を網羅的に表現できる利点がある。しかし、その組み合わせの多くは相関性が低いものであり、必ずしも画面空間を有効に利用した可視化結果を提示しているとは限らない、という問題がある。また、クラスタを単位とした概略的な傾向を掴み難い、という問題もある。それに対して本手法には、

- 入れ子構造による階層型データ表示により、遺伝子群をクラスタ単位で概略的に可視化できる。
- 相関性の高い 2 ノード間のみをエッジで表現することにより、相関性の高いノードにのみ注視した可視化を実現できる。

といった点で利点があると考えられる。

5. まとめと今後の課題

本報告では、遺伝子発現率情報に対してクラスタリングとネットワーク化の両方を適用して得られる、階層型ネットワークデータの可視化手法を提案した。

本手法はネットワークとクラスタを同時表示することにより、遺伝子学的に興味深いマルチドメインなどの特性の発見に貢献できると考えられる。また、クラスタをまたぐエッジを多く持つノードに特定の色をつけることにより、興味深い遺伝子の早期発見に貢献できると考えられる。

今後の課題として、

- 有向グラフを構成する遺伝子ネットワークの可視化。
- オントロジーなどの情報を加味した、より遺伝子の研究に貢献できる可視化ソフトウェアとしての開発。
- 結果画像から発見された現象が、本当に遺伝子学的に興味深い特性なのか否か、の検証。
- 複数のノードを 3 次元的に引き上げた時、あるいは多段階にわたってノードを 3 次元的に引き上げた時、の効果的なネットワークの表現手法の確立。
- 各クラスタの画面上の位置に関する最適化問題、クラスタリングの適切な閾値の発見、などの考察。
- 遺伝子ネットワークに限らず、複雑ネットワーク全般に応用できる階層型ネットワーク可視化手法の確立、および遺伝子ネットワーク以外のデータでの検証。

などを考慮したいと考えている。

謝辞

ソフトウェア Cluster 3.0 の開発者であるコロンビア大学 Michael De Hoon 氏には、クラスタリング技術に関して貴重なご助言を賜ったことを感謝致します。遺伝子ネットワークに関して、東京大学宮野教授、中谷助教授、渋谷講師、瀬々助手、井本助手から貴重な情報を賜ったことを感謝いたします。本研究の一部は、日本学術振興会科学研究費補助金の助成に関するものです。

参考文献

[1] 有田正規, 遺伝子ネットワークと確率モデル Genetic Networks and Probabilistic Models, 2001 年ベイジアンネットチュートリアル, pp. 50-53, 2001.

[2] Itoh T., Takakura H., Sawada A., and Koyamada K., Hierarchical Visualization of Network Intrusion Detection Data in the IP Address Space, IEEE Computer Graphics and Applications, Vol. 26, No. 2, pp. 40-47, 2006.

[3] Mukherjea, S., J. Foley and S. Hudson, Visualizing Complex Hypermedia Networks through Multiple Hierarchical Views, Proceedings of ACM SIGCHI '95, Denver, Colorado, pp. 331−337, May 1995.

[4] Eades, P., "A Heuristic for Graph Drawing," Congressus Numerantium, Vol. 42, pp. 149-160, 1984.

[5] Lamping, J. and Rao, R., "The Hyperbolic Browser: A Focus + Context Technique for Visualizing Large Hierarchies," Journal of Visual Languages and Computing, Vol. 7, No. 1, pp. 33-55, 1996.

[6] J. Carrire and R. Kazman, "Research Report: Interacting with Huge Hierarchies: Beyond Cone Trees," Proceedings of the IEEE Conference on Information Visualization '95, IEEE CS Press, pp. 74-81, 1995.

[7] B. Johnson, et al., Tree-Maps: A Space-Filing Approach to the Visualization of Hierarchical Information Space, IEEE Visualization ’91, pp. 275-282, 1991.

[8] P. Eades, et al., Multilevel Visualization of Clustered Graphs, Graph Drawing ’96, pp. 101-112, 1996.

[9] D. Schaffer, et al., Navigating Hierarchically Clustered Networks through Fisheye and Full-Zoom Methods, ACM Trans. Computer-Human Interaction, Vol. 3, No. 2, pp. 162-188, 1996.

[10] M. Sarcar, M. H. Brown, Graphical Fisheyes Views of Graphs, Communication of the ACM, Vol. 37, pp. 73-83, March 1994.

[11] M. L. Huang, et al., WebOFDAV−Navigatingand Visualizing the Web On-Line with Animated Context Swapping, 7th WWW Conf, pp. 636-638, 1998.

[12] S. North, Incremental Layout in DynaDAG, Graph Drawing ’95, pp. 409-418, 1995.

[13] 塩澤他, 「納豆ビュー」の対話的な情報視覚化における位置付け, 情報処理学会論文誌, Vol. 38, No. 11, pp. 2331-2342, 1997.

[14] Open Source Clustering Software (Cluster 3.0), http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/

[15] TreeView, http://www.gmod.org/node/91