

## グラフィカル・モデルに基づく生命情報からの因果・関連性解析

●堀本 勝久<sup>1)</sup> ◆藤 博幸<sup>2)</sup> ◇油谷 幸代<sup>1)</sup>

1) 産業技術総合研究所生命情報工学研究センター 2) 九州大学生体防御医学研究所

### <研究の目的と進め方>

これまでに、グラフィカル・モデルに基づいた網羅的データからのネットワーク構造推定法の開発を主に行い、幸い順調に推定法開発が進み着実に成果を上げることができた。この解析法は、遺伝子発現プロファイルなどの類似パターンが頻出する生命情報データへの適用のために、類似パターンを示すデータをクラスター解析により前処理をする改良により大量データ解析を実現した。一方、この前処理の結果、推定ネットワークは遺伝子群間の関連性を示し、個々の遺伝子の関連性を示すものではないために、解像度の点で問題があった。またさらに、大量ではあるが1条件下で計測されたデータを解析する方法であるため、異なる条件下で計測されたデータから、ネットワーク構造の変化を推定することは困難であった。2008年度は、これらの欠点を克服するために、遺伝子間関連推定のためのPath Consistency(PC) algorithmに基づくネットワーク推定法及びネットワーク構造変化を推定するためのグラフィカル連鎖モデルに基づいた推定法を開発・確立する。これらの開発によって、これまで開発した大量データから推定される全体的な関連性の枠組みと細部の関連性との二面から、また関連性の推移の観点からのネットワーク推定が可能になる。

具体的な研究課題は、以下の4つである。[1] グラフィカル連鎖モデルに基づく時系列発現データ解析法の確立。[2] Path Consistency (PC) アルゴリズムに基づく連続変量についてのネットワーク解析法の開発。[3] PC アルゴリズムに基づく離散変量についてのネットワーク解析法の開発。[4] 推定ネットワーク構造に連携した動態解析法の開発

### <2008年度の研究の当初計画>

研究目的[1]、[2]について実行する。

[1] グラフィカル連鎖モデルに基づく時系列発現データ解析法の確立

酵母の細胞周期と肝がん進展過程に関する発現プロファイルへの適用によって、既に予備的な研究は終了しているが、推定の前処理の段階である特異的に発現する遺伝子群の同定及びそれら遺伝子群の分類について十分な検討がなされていない。これらの問題点について重点的に改良し、時系列データから多状態間の因果推定法の確立を行う。具体的には、解析データにおいて現実状態に加えて現実状態への分類が困難な遺伝子群から成る仮想的な状態を仮定した条件の下で連鎖モデルを適用する方法と、数学的な検討による連鎖モデルとPC アルゴリズムとを融合した前処理を必要としない方法との2つのアプローチを試みる。

[2] Path Consistency (PC) アルゴリズムに基づく連続変量についてのネットワーク解析法の開発

PC アルゴリズムの遺伝子発現プロファイルなどの実数データへの適用法は既に基礎部分の実装は終了し、予備的な解析として、原核生物のオペロン構成遺伝子の発現プロファイルへの適用を行い、良好な結果を得ている。本年度多様な遺伝子発現プロファイルへの適用を行い適用限界を把握して、最終的な手法の確立を行う。また、PC アルゴリズムは生物学的知見を解析の際の束縛

条件として導入が容易であるという特徴をもつ。遺伝子のゲノム上の位置などの様々な生物学的束縛条件の導入とそれによる推定結果との照応によって、遺伝子発現に影響を及ぼす生物学的な条件の探索も同時に行う。さらに、物性データなど揺らぎの比較的少ないデータに適用し、データの揺らぎに依存しない条件下でPC アルゴリズムの関連性推定の性能を把握する。そのために、創薬対象である化合物及びその物性属性のデータについて適用し、化合物の薬理活性の有無と推定された関連性とを照応する。

### <2008年度の成果>

グラフィカル連鎖モデル(GCM)の適用法については、細胞周期及び肝癌進展過程の解明のため、それぞれの遺伝子発現プロファイルに適用した。細胞周期については隣接する細胞状態の関係性に加え時間的に離れた細胞状態間の遺伝子発現の関連性を推定することができた。肝癌進展過程の解析では、臨床及び病理の知見と一致した関係性が確認された。ただし、現解析法では異なる状態間で変数の重複が許されないという欠点がある。これは、同一分子が異なる細胞状態間で重要な役割を担う現象が解析不能であることを意味している。この欠点を克服するために、数学的な枠組みから逸脱せず、データ入力に関して工夫し仮想状態を仮定して解析する。これにより、重複変数の問題のみならず、変数の選別(特異的発現遺伝子の抽出)の必要がなくなり、すべての変数(全遺伝子)についてネットワーク構造変化及びそれらの寄与を推定することができた。

また、PC アルゴリズムを類似な物性特性を示す化合物データに適用し、化合物薬理活性データに基づいて性能評価を行い、有用であることを確認した。さらに、PC アルゴリズムに基づく異なる状態間のネットワーク構造変化を推定する方法を開発した。GCMの数学的枠組みを用いてPC アルゴリズムの拡張を行い、GCMによる遺伝子群間のネットワーク構造変化の推定と同様に、より詳細な遺伝子間ネットワーク構造変化の推定が可能になった。GCM及び拡張PC アルゴリズムを組み合わせることで、マクロとミクロの二つの観点から階層的にネットワーク構造変化を追跡できることを確認した。

上記のデータからネットワーク構造を推定するアプローチとは逆のアプローチとして、既知ネットワーク構造についてデータがどの程度整合性を示すかを推定するアプローチを開発した。このアプローチはネットワーク構造変化推定に有用である。すなわち、特定の条件下で計測されたデータについて、複数のネットワーク構造についてそれらの整合性を推定することで、有意な整合性を示すネットワーク構造はその条件下で活性化しているとみなすことが可能である。様々な条件下で計測されたデータについてこの手法を適用することで、それぞれの条件の活性化ネットワーク群が推定され、条件に応じたネットワーク構造の変化が追跡できる。整合性推定については、統計アプローチと記号-数値計算アプローチの2つのアプローチを開発した。前者は、E.coliの8以上の遺伝子で構成される29制御ネットワークと嫌気性条件下で計測された発現プロファイルデータにより、後者は、同じくE.coliのSOSシステムとその構成遺伝子発現プロファイルデータによ

り、有効性を示した。

特に、後者の研究は、代数学の生命情報解析への適用を主題として2004年に創設した研究分野 Algebraic Biology の一環でもある。2008年度、第3回国際代数学生物学会議をオーストリア・リンツのヨハネスケプラー大学記号計算研究所で開催し、特に記号計算手法の生命情報解析への適用を促進している。その論文集“Algebraic Biology: Lecture Notes in Computer Science 5147”は、Springer社から出版された。

#### <国内外での成果の位置づけ>

グラフィカル連鎖モデル及びPCアルゴリズムの生命情報データへの適用例は稀である。実際、これまでネットワーク解析に採用したグラフィカル・ガウシアン・モデルは、国内外の研究者による適用法の開発やそれらの適用例はあるが、同じグラフィカル・モデルである上記2つの方法に関する研究例は、特に生命情報への適用研究は極めて少ない。これは、その他のネットワーク解析手法に比べ他分野での利用状況が限定されていることもあるが、生命情報への適用に際してデータに関する生物学的知見を必要とするためでもある。新規な生物学的私見の発見を積み重ねることで、適用有用性をさらに示したい。

既知ネットワーク構造の評価手法に関しては、2つの計測データを利用して判別分析の延長として開発された方法が、発見データを利用したタンパク質相互作用及びバスウェイ解析において複数知られる。しかし特異的条件下で計測された1つのデータのみからネットワーク構造を評価する方法は、本手法のみである。実験条件とネットワーク構造が1対1対応で評価可能であることは、広範囲な適用可能性を保証する。

代数アプローチによる生命情報解析手法の開発は、現在そのアプローチの有用性を世界的なレベルで振興している。残念ながら、少数の例外はあるが生命情報解析に有用な本格的な研究は少ない。2009年度第4回国際代数学生物学会議を米国ノースカロライナ州 Statistical and Applied Mathematical Sciences Institute (SAMSI) で開催することで、従来欧州に偏りがちであった参加者に加え米国の研究者が積極的に参加することが期待される。

#### <達成できなかったこと、予想外の困難、その理由>

一般的に、基本的な解析法の実装は順調に終了し解析もほぼ終了したが、成果の発表が部分的に遅れ気味である。特にGCMとPCアルゴリズムの融合及びPCアルゴリズム自体のネットワーク構造変化推定のための拡張について、適用結果の公表が不十分である。ただし、遅れの一因は、開発手法の適用限界を見積もるために、複数の生命現象への適用を慎重に実行しているからでもある。

本年度開発したネットワーク評価に基づく構造変化推定手法について、統計アプローチによる解析に際して、制御ネットワークの構造を生物機能と対応させて複数収集する必須である。しかしながら、従来のバスウェイデータベース等にこれらデータが予想外に整備されていなかった。現在限定された生命現象に関してのみのネットワーク構造を整備することで解析を進めているが、将来、生物種・細胞種毎に制御ネットワーク構造データを整備したい。また、代数アプローチにおいては、ラプラス空間への変換による記号計算を適用するアプローチを試みた。このアプローチは厳密解を得られる可能性がありその場合は非常に頑強な結果が得られるが、そうでない場合ラプラス空間上での数値解析による不安定性のみが目立つ。ラプラス空間への変換による厳密解の探索に固執することなく、パラメータ間の代数関係を導出することで従来の数値最適化の脆弱性を補完するために記号計算を利用するアプローチも採用する必要がある。

#### <今後の課題>

離散値に対応したネットワーク推定手法の開発を行う。この際、関連性有意検定に際して自由度の扱いが重要であることが、予備的な解析によって判明している。また、離散値の解析は、代数アプローチの馴染み易いことから、ネットワーク評価以外の課題について代数的手法の適用可能性を探る。

ネットワーク構造変化の推定手法の開発は順調であるので、さらに一歩進めて、推定された構造変化を動的に表現する手法の開発を行う。

#### <成果公表リスト>

##### 1) 論文/プロシーディング (査読付きのものに限る)

1. 0801291230  
Yoshida, H., Horimoto, K., Anai, H.: Inference of probabilities over a stochastic IL-system by quantifier elimination. *Math. Compu. Sci.*, 1, 473-485 (2008).
2. 0806181237  
Nishino, R., Honda, M., Yamashita, T., Takatori, H., Minato, H., Zen, Y., Sasaki, M., Takamura, H., Horimoto, K., Ohta, T., et al.: Identification of novel candidate tumour marker genes for intrahepatic cholangiocarcinoma. *J. Hepatol.*, 37, 806-813 (2008)
3. 0901151259  
Hayashida, M., Sun, F., Aburatani, S., Horimoto, K. and Akutsu, T.: Integer Programming-based Approach to Allocation of Reporter Genes for Cell Array Analysis. *Int. J. Bioinformatics Research and Applications*, 4, 385-399, (2008)
4. 0901151303  
Saito, S., Aburatani, S. and Horimoto, K.: Network evaluation from the consistency of the graph structure with the measured data. *BMC Sys. Biol.* 2, 84, (2008).
5. 0901151309  
Nakatsui, M., Yoshida, H. and Horimoto, K.: An Algebraic-Numeric Algorithm for the Model Selection in Network Motifs in *Escherichia coli*. *Proceedings of the Second International Symposium on Optimization and Systems Biology (OSB' 08)*, pp. 257-264, (2008).
6. 0901151316  
Morioka, R., Arita, M., Sakamoto, K., Kawaguchi, S., Tei, H. and Horimoto, K.: Phase Shifts of Circadian Transcripts in Rat Suprachiasmatic Nucleus. *Proceedings of the Second International Symposium on Optimization and Systems Biology (OSB' 08)*, pp. 109-114, (2008).
7. 0901151325  
Zhang, Z.Y., Horimoto, K. and Liu, Z.: Time Series Segmentation for Gene Regulatory Process with Time-Window-Extension Technique. *Proceedings of the Second International Symposium on Optimization and Systems Biology (OSB' 08)*, pp. 198-203, (2008).
- 2) データベース/ソフトウェア  
1.0602211524  
階層型クラスタリングとグラフィカル・ガウシアン・モデリングとの組み合わせによる網羅的ネットワーク推定  
<http://eureka.cbrc.jp/asian>
- 3) 職員間での共同研究  
ネットワーク評価に関して、阿久津達也教授(京都大学)と共同で開発している。また、時計遺伝子の関連性解析に関して、有田正規準教授(東京大学)、程肇主任研究員(三菱化学生命科学研究所)と共同研究している。