公募研究：2005 ～ 2009 年度

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[1)]　◆藤 博幸[2)]　◆油谷 幸代[1)]
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＞

本研究は、遺伝子発現プロファイルなどの生命情報から生命現象におけるシステム間及び分子間の因果関係を推定するために、グラフィカルモデル（GM）に基づく統計因果推定法の開発し、生命システムを構成する物質間の関係性の理解を深めることを目的とする。

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。しかしながら、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種の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多種類のデータから、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。そこで、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

＜研究開始時の研究計画＞

上記研究課題に関する具体的な計画は、以下の通りである。

[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

グラフィカル連鎖モデルの適用に関する改良について、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かった。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適用法の確立を行う。具体的には、解析データにおいて状態間で異なると分類される遺伝子群に加えて、従来あまり考慮されない分類が困難な遺伝子群も含有して連鎖モデルを適用する方法と、連鎖モデルと PC アルゴリズムとの融合法の数学的な検討から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。本研究では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。この特徴を利用して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。またさらに、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の照応により､PC アルゴリズムの性能評価を行う。

[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とネットワーク構造離散化による構造評価法の開発

Path Consistency(PC) algorithm における連続量に関する偏相関係数の計算部分を、クロス表に基づき条件付き統計量を算出し検定を実行するように変更し、離散値データに関して関連性推定を可能にする。予備的な研究により変更は終了しているが、統計モデルに基づく検定では大きな自由度に関して検定力が極めて弱いことが判明している。その点を改良し公開されているタンパク質相互作用データに適用し、従来の推定法による結果と比較して有効性を評価する。さらに、データに基づく推定とは逆に、既知ネットワーク構造を離散しその構造と計測データとの照応により評価する方法を開発する。

[4] 推定ネットワーク構造に連携した動態解析法の開発

時系列データに関してネットワーク推定を実行し、その結果得られたネットワーク構造に基づき微分方程式を定式化する。次に、推定の結果得られた偏相関係数などの関連性強度を初期値としてパラメータ最適化を行う。さらに､推定構造及び最適パラメータを用いて感度解析を行い、再現性のテスト及び動態に関する知見を得る。上記にプロセスに従い、静的ネットワーク構造推定から動的振る舞いを推定する新しいアプローチの開発を試みる。開発に際しては、シミュレーションデータを利用して構造推定とパラメータ最適化に関する複数手法を実行し、それぞれの段階で精度を見積もることで最適な手法の選択を行う。

＜研究期間の成果＞

各研究計画についての成果は、以下の通りである。

[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

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。

[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。ただし、その有用性を確認するために、類似な物性特性を示す化合物データについて適用した。また、PC アルゴリズムを改良し、共発現する遺伝子を推定する方法を開発した。また、偏相関係数に基づいてアミノ酸配列の共進化情報を利用してタンパク質相互作用を推定する方法を開発した。タンパク質相互作用推定法と共発現遺伝子推定法は論文として掲載された。

さらに、GCM のように異なる状態間のネットワーク構造変化に関する推定はできないが、この点を GCM の数学的枠組み用いて 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。実際、肝硬変から肝がんへの進展に関して、2 種の細胞の発現情報データに適用しこれら 2 細胞状態間の進展関連遺伝子ネットワーク候補を推定した。

[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とネットワーク構造離散化による構造評価法の開発

ネットワーク構造と計測データとの整合性評価法については、ガウスモデルに基づく統計的方法を確立し、シミュレーション及び計測データを用いて性能評価を行いその有効性を確認した。大腸菌の既知遺伝子制御ネットワーク 30 について、嫌気性下で計測された発現プロファイルとの整合性を推定した結果、嫌気呼吸関連ネットワークと maltose 代謝関連ネットワークの 2 つが整合性を示す結果を得た。共に嫌気性下に特徴的なネットワークであり、特に嫌気呼吸関連ネットワークは 150 以上の遺伝子から構成される複雑な構造を持つネットワークである。既知ネットワークデータについて、特異的条件下で活性化される制御ネットワークを厳格に検出する方法として有効であることがわかり、今後既知ネットワークの整備と様々な状況下で計測されたデータの適用により、新たな活性化ネットワークの発見を期待している。ただし、統計的手法では適用グラフが directed acyclic graph (DAG) に限定される。その問題を解決するために、代数的な手法を開発した。現在、シミュレーションによって cyclic loop を持つグラフについて有効であることを検証した。

[４] 推定ネットワーク構造に連携した動態解析法の開発

動態解析の新規なアプローチとして、計算機代数に基づいた手法を考案し、その有効性を調査した。その結果、従来の数値解析アプローチでは解析解が非常に複雑になるためパラメータ推定不能である場合にも、ラプラス変換によって代数方程式に変換することで解析可能になる場合が発見された。ただし、一般的な方法として欠点があるためさらに改良を加え、新規な動態解析法を開発した。一般的な生体ネットワーク動態解析では、まず実験解析結果などの生物学的知見に基づき、分子反応モデルを構築する。次に、分子反応の様式に基づいて微分方程式を定式化する。そして最後に、実験計測データに基づいて、反応パラメータの数値解析を実行する。しかしながら、特に実験計測データが少数である場合において、反応パラメータを一意に同定できないことがある。我々は、この問題を克服するための試みとして、代数的アプローチを導入した。微分方程式モデルから、代数手法の一つである、Differential Elimination によって反応パラメータ間の束縛条件を導出し、この束縛条件をパラメータの数値最適化における評価関数の一部に採用した。4 分子の内 1 分子のみが測定可能なネットワークを想定し、そのシミュレーションデータを用いて、束縛条件を考慮する場合としない場合で反応パラメータの最適化を実行した。このとき、最適化の手法として実数値遺伝的アルゴリズムを用いた。その結果、束縛条件を考慮しない場合はパラメータすべてについて正しく推定できなかったが、考慮した場合はすべて推定できた。初歩的な解析ではあるが、Differential Elimination による束縛条件を数値最適化における評価関数に導入することは、少数の実験計測データのみからネットワーク動態解析を行うための有用な手法の一つと考えられる。

＜国内外での成果の位置づけ＞

グラフィカルモデルに基づくネットワーク解析法は、解析法の数理が簡潔である、利用に際して実装が容易である、パソコンレベル以上の計算機性能を要求しない、解析結果の直感的理解が容易である、などの利点があり、様々な実験研究の解析に適用されている。しかしながら、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適用条件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。適用例はまだ少数であるが本研究で新規な生物学的知見が発見され、適用有用性を実証された。

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、発現データを利用したタンパク質相互作用及びパ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ヴェルで振興している。これまで国際会議 Algebraic Biology を３回主催し、その論文集を Springer 社から刊行し好評を得ている。

＜達成できなかったこと、予想外の困難、その理由＞

全般的に、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。

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、従来のパ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、将来的に生物種・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。現在代数算法の一つである Differential Elimination を利用した方法を考案し、予備的な計算実験で極めて良好な結果を得ているが、さらに十分な性能評価を行う必要がある。

＜今後の課題、展望＞

全般的には、方法自体の開発に関する報告は済んでいるので、様々な生命情報データへの適用によって、生物学的知見発見のために有用であることを示す報告を行う。具体的な課題は以下の通りである。

I. これまでの共同研究は、実験終了後のデータ解析に関する共同研究であったが、実験計画の段階から参加する共同研究を成功させたいと考える。

II. 新規方法の開発を行う。特にネットワーク構造解析と動態解析の橋渡しになる、ネットワーク構造変化を推定が可能な時系列データの解析手法を開発する。

III. 開発方法の統合を行う。構造推定法による静的ネットワークについて動態解析法を有機的に連携し、データ入力により動的ネットワーク解析が可能なシステムを構築する。

IV. 既開発法の公開を行う。現在、グラフィカル・ガウシアン・モデルの適用法のみ公開しているが、本研究で開発した解析法を公開し、解析データの性質や解析目標など利用状況に応じた包括的なネットワーク解析サイトを構築する。

V. 既開発法と生物情報データとの連携を行う。実験研究者が所有するデータの解析だけでなく、既知データの有効利用を実行できるように関連データベースとの連携が可能なシステムの構築を行う。

VI. 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適用可能性を探る。

＜研究期間の全成果公表リスト＞

1）論文／プロシーディング

1. 0801291230
   Yoshida, H., Horimoto, K.,, Anai, H.: Inference of probabilities over a stochastic IL-system by quantifier elimination. Math. Compu. Sci., 1, 2008, 473-485.
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, 2008, 806-813.
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, 2008, 385-399.
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, 2008, 84.
5. 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),2008, pp. 109-114.
6. 0801291226
   Aburatani, S., Sun, F., Saito, S., Honda, M., Kaneko, S. and Horimoto, K., Gene systems network inferred from expression profiles in hepatocellular carcinogenesis by graphical Gaussian model. EURASIP J. Bioinfo. Systems Biol., 2007, 47214.
7. 0702131228
   Yoshida, H., Anai, H. and Horimoto, K., Derivation of rigorous conditions for high cell–type diversity by algebraic approach. BioSystems, 90, 2007, 486–495.
8. 0801291238
   Sato, T., Yamanishi, Y., Horimoto, K., Kanehisa, M. and Toh, H., Inference of Protein-Protein Interactions by Using Co-evolutionary Information. In Anai, H., Horimoto, K. and Kutsia, T. (eds), Algebraic Biology 2007, Lecture Notes in Computer Science 4545, 2007, p.322-333, Springer, Heidelberg.
9. 0801291246
   Aburatani, S., Inference of Complex Regulatory Network for Cell Cycle System in Saccharomyces Cerevisiae. In Anai, H., Horimoto, K. and Kutsia, T. (eds), Algebraic Biology 2007, Lecture Notes in Computer Science 4545, 2007, p.350-364, Springer, Heidelberg.
10. 0801291251
    Yoshida, H., Nakagawa, K., Anai, H. and Horimoto, K. Exact parameter determination for Parkinson' s disease diagnosis with PET using an algebraic approach. In Anai, H., Horimoto, K. and Kutsia, T. (eds), Algebraic Biology 2007, Lecture Notes in Computer Science 4545, 2007, p.110-124, Springer, Heidelberg.
11. 0801291257
    Yoshida, H., Nakagawa, K., Anai, H. and Horimoto, K., An Algebraic-Numeric Algorithm for the Model Selection in Kinetic Networks. Proceedings of 10th CASC, Lecture Notes in Computer Science 4770, 2007, p.433-447, Springer, Heidelberg.
12. 0801291303
    Hayashida, M., Sun, F., Aburatani, S., Horimoto, K. and Akutsu, T., Integer Programming-based Approach to Allocation of Reporter Genes for Cell Array Analysis. Proceedings of 10th OSB, Lecture Notes in Operation

Research, 2007, p.288-301, World Publishing Corporation, Beijing.
13. 0801291306
Katoh, K., Toh, H., PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. Bioinformatics, 23, 2007, 372-374.
14. 0801291308
Standley, D.M., Toh, H., Nakamura, H., ASH structural alignment package: Sensitiviry and selectivity in domain classification. BMC Bioinformatics, 8, 2007, 116.
15. 0702131228
Yoshida, H., Anai, H. and Horimoto, K.: Derivation of rigorous conditions for high cell–type diversity by algebraic approach. BioSystems, 90, 2007, 486–495.
16. 601271217
Aburatani, S., Saito, S., Toh, H. and Horimoto, K.: A graphical chain model for inferring regulatory system networks from gene expression profiles. Statistical Methodology, 3, 2006, 17-28.
17. 0702131213
Anai, H., Orii, S. and Horimoto, K.: Symbolic-numeric estimation of parameters in biochemical models by quantifier elimination. J. Bioinfo. Comput. Biol. 4, 2006, 1097-1117.
18. 0702131220
Sato, T., Yamanishi, Y., Horimoto, K., Kanehisa, M. and Toh, H.: Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions. Bioinformatics, 22, 2006, 2488-2492.
19. 0702131223
Yoshida, H., Anai, H., Orii, S. and Horimoto, K.: On relationship between proliferation and transition rates of multicells. 数理解析研究所講究録 , 1514, 2006, 59–65.
20. 0702131446
Aburatani, S. and Horimoto, K.: Serial Network Inference in Cell Cycle Regulation on Yeast. 10th World Multiconference on Systems, Cybernetics and Informatics, 4, 2006, 1-6.
21. 0702131450
Yoshida, H., Anai, H. and Horimoto, K.: On Rigorous Conditions for Cell-type Diversity by Algebraic Approach. Proceedings of the First International Conference on Mathematical Aspects of Computer and Information Sciences: MACIS, 2006, 3–14.
22. 0702131454
Yoshida, H., Anai, H. and Horimoto, K.: Derivation of Rigorous Relationships between Proliferation and Transition Rates of Multiple Cells by Algebraic Approach. 10th World Multiconference on Systems, Cybernetics and Informatics, 4, 2006, 23-28.

2）学会発表

1. Horimoto, K.: “Network Screening” - Detection of Activated Pathways from the Data Measured in a Finite Condition. Computational Systems Biology Workshop, Institute of Systems Biology, Sep. 19, 2009, Shanghai University.
2. 齊藤秀、堀本勝久 :「遺伝子発現プロファイルによるネットワーク構造評価」, 8 月 6 日 2009 年 , 電気通信学会
3. 中津井雅彦、堀本勝久 :「代数的アプローチと遺伝的アルゴリズムの組み合わせによる計測不能変数を含むネットワークのパラメータ最適化」, 8 月 6 日 2009 年 , 電気通信学会
4. Nakatsui, M. and Horimoto, K.: Parameter Optimization in the network dynamics including unmeasured variables by the symbolic-numeric approach. Proceedings of the Third International Symposium on Optimization and Systems Biology (OSB’ 09), Sep. 19, 2009, China
5. Akutsu, T., Tamura, T. and Horimoto, K.: Complementing Networks Using Observed Data. The 20th International Conference on Algorithmic Learning Theory, October 3 - 5, 2009, University of Porto, Portugal.
6. Saito, S. and Horimoto, K.: Co-Expressed Gene Assessment Based on the Path Consistency Algorithm: Operon Detention in Escherichia coli. IEEE SMC, Oct. 11, 2009, San Antonio, USA
7. 堀本勝久 :「細胞内分子ネットワーク構造変化の追跡」, 数理生物学会 , 2009 年 9 月 11 日 , 東京大学
8. 堀本勝久 : 2つアプローチによる時間発展するネットワーク構造変化解析 , BMB2008（第 31 回日本分子生物学会年会・第 81 回日本生化学会大会 合同大会）シンポジウム 4 S8「動的ネットワーク構造探索の計算イニシアティブ」, 2008 年 12 月 12 日 , 神戸ポートピアホテル・南館
9.Nakatsui, M., Yoshida, H. and Horimoto, K.: An Algebraic-Numeric Algorithm for the Model Selection in Network Motifs in Escherichia coli. The Second International Symposium on Optimization and Systems Biology (OSB’ 08), October 31–November 3, 2008, Lijiang, China
10. Horimoto, K.: Dynamical analysis of biological networks by using reporter gene expressions in tranfected cell microarray, DMHF2007, Oct. 1, 2007, Kyushu University

3）図書

1. Aburatani, S., Saito, S., and Horimoto, K., Humana Press, New Jersey, ASIAN: Automatic System for Inferring A Network. In Krawetz, S. (ed): Bioinformatics for Systems Biology: Second Edition, Introduction to Informatics, p.563-577, 2009, p639.
2. Anai, H. and Horimoto, K. (eds), Springer, Symbolic Computation in Biology, Math. Comput. Sci., 2, 399-556, 2009, p157.
3. Horimoto, K., Regensburger, G., Rosenkranz, M. and Yoshida, H. (eds), Springer, Heidelberg, Algebraic Biology, Lecture Notes in Computer Science 5147, 2008, p245.
4. Anai, H., Horimoto, K. and Kutsia, T. (eds), Springer, Heidelberg, Algebraic Biology, Lecture Notes in Computer Science 4545, 2007, p379
5. Anai, H. and Horimoto, K. (eds), Universal Academy Press, Tokyo, Algebraic Biology – Computer Algebra in Biology, 2005, p173

4）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よる網羅的ネットワーク推定
http://eureka.cbrc.jp/asian