様式 C-19

科学研究費補助金研究成果報告書

平成２２年　６月　６日現在

機関番号：１２６０１
研究種目：若手研究(A)
研究期間：2008 ～ 2010
課題番号：２０６８００１６
研究課題名（和文）　超高次元複雑ヘテロデータ解析に基づく柔軟かつ頑健な非線形統計的モデリングの研究
研究課題名（英文）　Flexible and Robust Nonlinear Statistical Modeling Based on High-Dimensional Complex Heterogeneous Data Analysis
研究代表者
井元　清哉　（IMOTO SEIYA）
東京大学・医科学研究所・准教授
研究者番号：１０３４５０２７

研究成果の概要（和文）：多様な形式で与えられる高次元観測データから、有効に情報を抽出するための頑健でかつ柔軟な統計的データ解析手法の開発を行った。その結果、事前に定義された特徴量集合が与えられた際に、その特徴量集合が大規模観測データに対して有意に異なる分布をしているか否かを判定する一連の統計解析手法を開発した。また、大規模遺伝子ネットワークを推定するための頑健な推定方法を提案し、がんの多様性解析に応用した。

研究成果の概要（英文）：We developed robust statistical methods for extracting valuable information from various types of high-dimensional heterogeneous data. As the results, when a subset of feature variables is defined *a priori*, we develop a series of statistical methods that can evaluate whether the subset has a unique distribution by comparing with background variables in the given large dataset. We also proposed a robust estimation method for large gene network from microarray gene expression data and other types of data like transcription binding sites and applied it to the analysis of cancer heterogeneity.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	6,100,000	1,830,000	7,930,000
２００９年度	4,700,000	1,410,000	6,110,000
２０１０年度	5,000,000	1,500,000	6,500,000
年度			
年度			
総　計	15,800,000	4,740,000	20,540,000

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：高次元データ解析、異種データ統合、ベイズ統計、ロバスト

１．研究開始当初の背景

高度に発達した様々な計測技術・機器により、興味ある対象が実数、離散値、順序カテゴリー、文字(アルファベット)、グラフ構造など様々な形式の観測データにより特徴づけられることは珍しいことではなくなっていた。さらに、網羅的観測データは、その解像度・観測可能な変数は技術革新と共に飛躍的に増大し、その結果、はずれ値・欠損値・エラーの中に真に有用な情報を含む超高次元ヘテロデータが蓄積されていた。国内外において、」このようなヘテロデータを統合するためのアルゴリズムベース（手続きベース）の方法論はいくつか見られていた。しかしながら、研究開始当初において、これらを統計モデルベースの確かな理論的基盤の上で高度に統合し、様々な形式で与えられる超高次元観測データから有効な情報を抽出するための理論、および、実際面でも有用な方

法論はまだ十分に研究は進んでいなかった。

このような背景のもと、申請者は、多数の確率変数の依存関係を探るためのグラフィカルモデル推定の一例として、マイクロアレイデータ、DNA 配列データ、データベース情報などの超高次元ヘテロデータから階層ベイズモデルを用い遺伝子の依存関係を探るための解析を行った。遺伝子間の依存関係を明らかにすることは、生命のシステム的理解を目的とした Systems Biology の根幹を成す極めて重要な問題である。この研究を通して、次のことが明らかとなりつつあった：

- ✓ データの高解像度化に伴いデータにはずれ値・欠損値が増大し、その処理・評価が重要になること。
- ✓ カーネル法をベースとした情報統合では､各データに対してカーネル行列を計算し､加重平均により情報統合を行っている。この方法をベースに、さらに各データの相関構造を考慮した方法論が構築可能であること。
- ✓ 複雑な構造を有する超高次元ヘテロデータに対しては､階層ベイズモデルの階層性(変数間の非閉路有向グラフ構造)を緩めた統計的モデリングの可能性。

ベイジアンネットワークに基づくマイクロアレイデータからの遺伝子制御ネットワークの推定において、条件付き分布の強健かつ信頼できる推定のために 2006 年に提案した枠組み (Imoto *et al.*, *Pac. Symp. Biocomput.* (2006); Imoto *et al.*, *Statistical Methodology* (2006)) は、遺伝子制御ネットワークのモデル構成プロセスにおいて、マイクロアレイデータからの遺伝子発現量の情報とデータベース化された生物学的知識の情報を結合させる一般的枠組みである。その手法の利点の一つは、マイクロアレイデータからの情報と生物学的知識間のバランスをどうとるかを情報量規準が定めることができることである。生物学的知識をベイジアンネットワークに付加することにより、マイクロアレイデータからさまざまなノイズによる影響を排除しつつ精密に遺伝子ネットワークを推定でき、結果としてより多くの情報を抽出することに成功している。これらの研究成果による方法論をさらに発展させることが必要となっていた。

２．研究の目的

複数種類の複雑な構造を有する高次元データから有効に情報を抽出するための統計的データ解析の方法論を構築することを目的とする。特に、生命科学における遺伝子制御のネットワークの推定（統計科学においては超多数の確率変数間の依存関係を明らかにするグラフィカルモデリングに相当する）をその応用として、多様なゲノム、トランスクリプトーム、プロテオームなどの生体内分子の網羅的計測データを統合的に解析し、その鍵となっている遺伝子制御ネットワークを明らかにすることの出来る方法論を構築することを目的とする。

また、生命科学においては、細胞サンプルによる腫瘍の分類に加えて、ゲノム情報によるより精緻な腫瘍の分類が重要な課題となっていた。その背景は、ゲノム情報により、より正確に分子標的薬の副作用や効果の予測が可能となり、また、がん細胞の悪性度が予測可能であることが徐々に明らかとなってきたため、個別化医療へのトランスレーショナルリサーチへと生命情報学が進んでいったためである。そのために、階層ベイズモデル、カーネル法などにより情報統合を行い、目的の多変量解析を行うための方法論の構築を行うことを研究の目的とした。

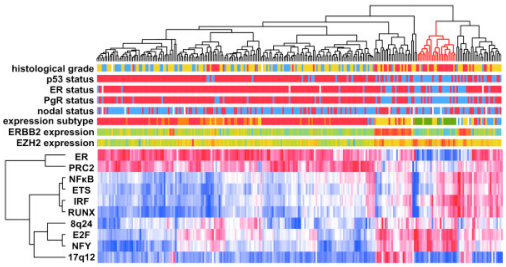
３．研究の方法

複数種類の高次元データを統合的に解析するためには、それぞれのデータについてのモデリングを精査し、データの解像度の違い、エラーの分布、内在する情報の質を明らかにする必要がある。その違いに応じた情報統合を実データを用いつつ検証している方式をとる。特に、遺伝子発現データからの遺伝子ネットワーク推定においては、マイクロアレイ遺伝子発現データにおいてはベイジアンネットワークを用いる。また、様々なデータから得られた結果をデータベースにまとめられているため、その情報を整理し、ベイズモデルにおける事前情報として使用する。その枠組みについては Imoto et al. (2003) において発表したベイズモデルを利用する。

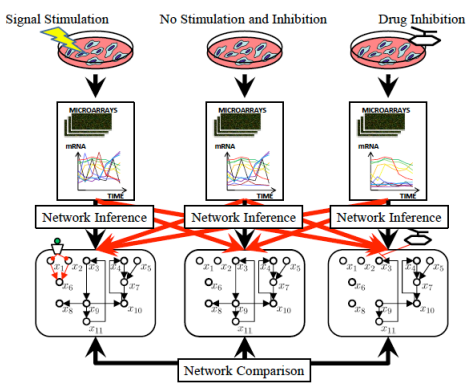
４．研究成果

1. シミュレーションモデルと観測データを融合し,シミュレーションモデルの自動構築を可能とするデータ同化技術について,ペトリネットモデルによって記述されたシミュレーションモデルを線形, 共線形方程式として表し, そのパラメータ推定を事後確率最大化法 (Maximum a posteriori) として定式化することで行う方法を開発した (Yoshida et al. (2008) Bioinformatics). また, 複数のネットワーク仮説があった場合に,シミュレーション結果のロバストネスから最適仮説を選択する情報量規準を導出し,実際のサーカディアンリズムパスウェイに対して適用した.

2. 転写因子の結合するシス制御配列モチーフ，ChIP-chip（ChIP-seq）データ，遺伝子の DNA 上の位置情報とマイクロアレイによるトランスクリプトームデータを統合し，機能的に有意な遺伝子モジュールを抽出する EEM 法を提案した（Niida et al. (2009) BMC Bioinformatics）．この方法により，多種多様なゲノム情報を統合し，遺伝子モジュールとしての単位で原料を理解することが可能となった．提案した EEM 法は乳がんのマイクロアレイデータに適用し，これまで知られていたいくつかの乳がんバイオマーカー，および原因遺伝子の関連を予測した．下図は、EEM 法を用いて推定した乳がんにおいてマスター制御因子的役割をになう遺伝子モジュールにより乳がんのサブタイプを同定した結果を示している。



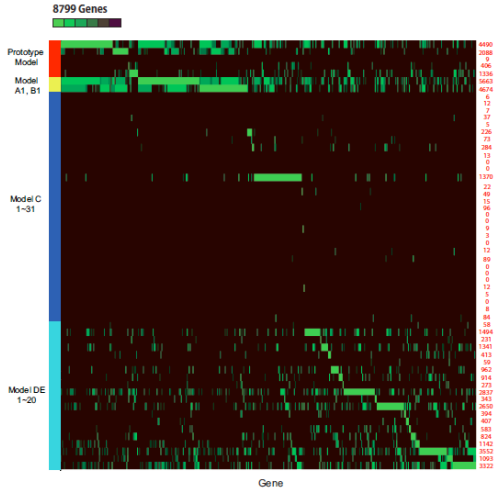
3. 薬剤応答時系列マイクロアレイデータとタンパク質間相互作用データを融合し，薬剤が転写レベルで細胞に与える影響の時間変化を推定し，その影響を伝えるシグナル伝達経路を予測する方法をダイナミックベイジアンネットワークと統計的メタアナリシスを用いて開発した．提案した方法は，ヒト血管内皮細胞における高脂血症薬 Fenofibrate の薬剤作用機序を明らかにするための解析に適用し，それまで仮説であった薬剤のオートクライン的２次作用を検証した．その結果，オートクライン的２次作用に関わっていると考えられる統計的に有意なシグナル伝達経路とそのターゲット遺伝子を同定した．
4. 現在、Genome Expression Omnibus などの公共データベースには様々ながん細胞の遺伝子発現データが大量に蓄積されている。本研究において、機能的遺伝子発現モジュールを網羅的に同定し、それらに対してメタアナリシスを行うことにより、さまざまながんにおいてドライバー的役割を果たす機能モジュールの同定を行うための方法を開発した。開発した手法を、１２２個のがんに関する研究において公開されたマイクロアレイデータセットに対して適用し、実際の機能モジュールを同定し、それらの生物学的ながんとの関係を検証した。その結果、発見した多くの機能モジュールはすでにがんとの関係が既知のものであったが、レアな機能モジュールに関しては、がんとの関連が未知なものも多く含まれていた。
5. 遺伝子発現データからベイジアンネットワークを用いて遺伝子ネットワークを推定する手法は、マイクロアレイが安価になり大量のデータが利用可能になったこともあり、ゲノムデータ解析の標準的な手法になりつつある。しかしながら、ベイジアンネットワークの構造学習の計算量が膨大なこともあり、事後確率最大化法に基づく方法では、最適なネットワークは探索できず、局所解として得られるネットワーク構造では、精度はそれほど高くないというのが現状である。本研究の成果により、ネットワークの事前情報としてラフなネットワーク（無向グラフでよい）が得られたとき、そのネットワークをクラスタリングして、各クラスタ上で最適学習を行い、クラスタ同士を結んでもサイクルのできないための必要十分条件を数学的に求めたアルゴリズムを用いることにより、これまでは５０程度の遺伝子でしか達成できなかった条件付き最適学習が、約１０倍の規模（５００遺伝子）で行えるようになった。
6. 実験条件の異なる複数の時系列データを統合し、遺伝子ネットワークを推定する問題は、時系列マイクロアレイデータの時点数の少なさから、遺伝子ネットワーク推定の精度を上げるための方法として近年注目を集めている。しかしながら、これまでは単純にデータを重ねるという方法をとっており、かえってノイズを増幅される原因となっていた。本研究の成果により、ネットワークの似ている部分では積極的に情報を統合し、似てい



ない部分ではデータを独立に取り扱うことが自動的に可能な数学的な枠組みを構築し、実データに対して適用し、従来法を大きく超える性能を示すことができた。上手は、手法の概念図を示している。

7. 遺伝子発現データと遺伝子機能、共通の転写因子からの制御、共通のパスウェイ、染色体上の位置情報、共通の疾患における関連などさまざまな情報から定義される遺伝子セットを合わせた解析として、遺伝子発現データの解釈を行い、機能的な遺伝子セットを探索する Gene Set Analysis は、近年着目を集めている。しかしながら、従来の方法は、サンプルラベルを利用する教師有り学習で定式化されたものが大部分を占め、また、教師なし学習においては特定のモデルを仮定したものであった。サンプルラベルに依らない複雑な遺伝子発現の協調性を捉えるため、（１）BEEM と名付けた Biclustering に基づく部分サンプル共発現遺伝子探索手法を開発した。これにより、大規模かつヘテロなデータにおいてサンプルラベルの信頼性を問題にすることなくサンプルサブタイプにおける機能的遺伝子セットの自動抽出が可能となった。また、（２）BEEM は正規分布のような楕円分布を発現データに仮定するモデルと同値な計算プロセスにより解析を行うが、特異値分解を利用したモデルを仮定しない機能的遺伝子セット探索手法 MIEA を開発した。これにより、より複雑な相関構造をモデルの仮定無く探索することが可能となった。

8. 細胞内システムのシミュレーションモデルの構築は、システム生物学における重要な研究課題の一つであるが、現在は、文献により裏付けられた、堅いネットワークにおいて反応パラメータを遺伝子発現データやタンパク質発現データを利用し推定することに研究の力点は置かれている。しかしながら、文献から構成されるネットワークは完全ではない場合がほとんどである。そこで、あらかじめ生物学的な知識を用いて、プロトタイプである文献ネットワークから可能性のあるネットワーク構造を自動的に生成し、それぞれのネットワークにおいてパラメータ推定を行いプロトタイプモデルよりも良いモデルを探索する手法を開発した。下図は、５つのプロトタイプのシミュレーションモデルから生成した６３個のモデルを用いて、８７９９個の遺伝子がそれぞれどのモデルによって説明されるかを網羅的に探索した結果である。縦軸にモデルが並び、横軸には遺伝子が並ぶ。上の赤で示したモデルがプロトタイプの５モデルである。緑が当てはまりが良いことを示している。プロトタイプは多くの遺伝子の発現プロファイルを旨く説明しているが、プロトタイプよりも生成したモデルがよりよく発現プロファイルを説明できている場合も多く見られる。



５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１０件）

① T. Hasegawa, R. Yamaguchi, M. Nagasaki, S. Imoto, S. Miyano (2011) Comprehensive pharmacogenomic pathway screening by data assimilation, Springer-Verlag, LNCS, in press.

② K. Kojima, E. Perrier, S. Imoto and S. Miyano (2010) Optimal search on clustered structural constraint for learning Bayesian network structure, Journal of Machine Learning Research, 11, 285-310.

③ T. Shimamura, S. Imoto, R. Yamaguchi, M. Nagasaki, S. Miyano (2010) Inferring dynamic gene networks under varying conditions for transcriptomic network comparison. Bioinformatics, 26(8), 1064-1072.

④ A. Niida, S. Imoto, R. Yamaguchi, M. Nagasaki, S. Miyano (2010) Gene set-based module discovery decodes cis-regulatory codes governing diverse gene expression across human

multiple tissues, PLoS One, 5(6), e10910.

⑤ Niida, S. Imoto, R. Yamaguchi, M. Nagasaki, A. Fujita, T. Shimamura, S. Miyano (2010) Model -free unsupervised gene set screening based on information enrichment in expression profiles, Bioinformatics, 3090-3097.

⑥ S. Kawano, T. Shimamura, A. Niida, S. Imoto, R. Yamaguchi, M. Nagasaki, R. Yoshida, C. Print, and S. Miyano (2010) Discovering functional gene pathways associated with cancer heterogeneity via sparse supervised learning, IEEE Bioinform atics and Biomedicine, 253-258.

⑦ A. Niida, A.D. Smith, S. Imoto, S. Tsutsumi, H. Aburatani, M.Q. Zhang and T. Akiyama (2009) Gene set-based module discovery in the breast cancer transcriptome, BMC Bioinformatics, 10:71.

⑧ Y. Tamada*, H. Araki*, S. Imoto*, M. Nagasaki, A. Doi, Y. Nakinishi, Y. Tomiyasu, K. Yasuda, B. Dunmore, D. Sanders, S. Humphries, C. Print, D.S. Charnock-Jones, K. Tashiro, S. Kuhara, S. Miyano (2009) Unraveling dynamic activities of autoacine pathways that control drug-response transcriptome networks, Pacific Symposium on Biocomputing, 14, 251-263. (*Contributed equally to this work)

⑨ A. Niida, S. Imoto, M. Nagasaki, R. Yamaguchi and S. Miyano (2009) A novel meta-analysis approach of cancer transcriptomes reveals prevailing transcriptional networks in cancer cells, Genome Informatics, 22, 121-131.

⑩ R. Yoshida, M. Nagasaki, R. Yamaguchi, S. Imoto, S. Miyano, T. Higuchi (2008) Bayesian learning of biological pathways on genomic data assimilation, Bioinformatics, 24(22), 2592-2601.

〔学会発表〕(計５件)

① S. Imoto, "Statistical Modeling of Dynamic Gene Networks in Cancer", Winter School in Mathematical and Computational Biology (University of Queensland, Australia), 2010 年 7 月 (招待講演)

② S. Imoto, "Transcriptomic Network Comparison Unravels Key Differences in Gene Regulatory Networks Characterizing Cell Conditions", Probability and Statistics with Applications in Systems Biology (Taiwan), 2010 年 7 月 (招待講演)

③ S. Imoto, "Computational Biomarker Identification by Analyzing Time-Course Microarray Gene Expression Dat a", International Symposium of Case Studies involving Statistics and Operations Research for Decision Making: Solving Human Problems in Business, Society, and Scientific Areas, Institute of Mathematical Statistics, Japan, 2009 年 3 月 (招待講演)

④ S. Imoto, "Statistical inference of gene networks for computational drug-target pathway discovery", Winter School in Mathematical and Computational Biology (University of Queensland, Australia), 2009 年 7 月 (招待講演)

⑤ S. Imoto, "Unraveling dynamics of biological networks towards computational drug target discovery", Statistical Computation and Visualization 2008, (Academia Sinica, Taiwan), 2008 年 12 月 (招待講演)

〔図書〕(計１件)

① S. Imoto, Y. Tamada, H. Araki and S. Miyano (2010) Computational Drug Target Pathway Discovery: A Bayesian Network Approach. in H. Lu, B. Schokop, H. Zhao (Eds.), *Handbook of Computational Statistics: Statistical Bioinformatics*, Springer-Varlag. in press.

〔産業財産権〕
○出願状況 (計０件)

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況 (計０件)

名称：
発明者：

権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
井元　清哉（IMOTO SEIYA）
東京大学・医科学研究所・准教授
研究者番号：１０３４５０２７

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：