様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成　25　年　6　月　7　日現在

機関番号：62603
研究種目：若手研究(B)
研究期間：2011～2012
課題番号：23700346
研究課題名（和文）　最先端トランスクリプトーム解析のための統計的バイオモデリング

研究課題名（英文）Statistical biomodeling for advanced transcriptome research

研究代表者
吉田　亮（YOSHIDA RYO）
統計数理研究所・モデリング研究系・准教授
研究者番号：70401263

研究成果の概要（和文）：生化学反応システムのシミュレーションモデルを開発するための統計科学の方法論の研究を行った。本研究の開発手法は、代謝工学や合成生物学の分野において将来的な応用が期待される。また、応用研究として、抗癌剤の薬剤動態を対象に遺伝子発現ネットワークのモデル開発を行った。薬剤耐性癌の全遺伝子の発現変化を観測し、データのパターンに基づきシミュレーションモデルを開発した。

研究成果の概要（英文）：This research has established statistical methods for making a range of biochemical reaction simulators. The developed methods can be utilized in the applied studies on synthetic biology and metabolic engineering. We also conducted the development of gene regulation simulators for predicting drug responses of drug tolerant lung tumors.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	3,100,000	930,000	4,030,000

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：ベイズ統計学　状態空間モデル　システム生物学

１．研究開始当初の背景

システムズバイオロジーのマイルストーンは、生体内分子間相互作用ネットワークを同定し、そこに書き込まれた制御プログラムを明らかにすることである。近年は、DNA マイクロアレイによる遺伝子発現プロファイリングやゲノムワイドなメタボローム計測など、多種多様な分子計測技術がハイスループット化され、オミックスレベルのデータと統計手法の協働により、システムの特徴を生み出す中心的なメカニズムを一気に見つけ出すという接近法が定着しつつある。本研究は、このような多面網羅的なオミックスデータから生体内分子相互作用系の動作原理を解明することを目的とし、状態空間モデルに基づく統計解析技術を確立することを目指した。

２．研究の目的

本研究はトランスクリプトームに特化したかたちでプロジェクトを遂行する。転写制御システムの大域的理解を実現すべく、トランスクリプトームとベイズ統計理論を融合した多次元オミックス横断型データ解析の方法論を構築する。トランスクリプトーム研究は、DNA マイクロアレイの発明以来、産学の垣根を越えた品質改善努力が実を結び、今や成熟期と言える段階に移行している。しがしながら、生命科学の最前線では、膨大なデータと対象系の多面網羅的な観察事象を知識循環過程に組み込むための合理的な術を持たず、依然としてエキスパートがごく限られ

た部分を不完全に考察しているのが現状である。本研究では、ゲノム、定量オミックス、転写因子・パスウェイ、de novo 転写因子結合モチーフ、エピジェネティックスを介した転写調節など、利用可能な多面的オミックス情報を活用し、ヒトの全遺伝子を網羅する転写調節系のような超巨大システムを巨視的な視点から読み解くことに挑戦した。

より具体的に述べると、遺伝子発現の量的なデータと転写因子・プロモータ相互作用データのような質的情報が与えられたもとで、これらをいかに情報統合するかという問題である。転写という事象を異なる側面から観察した二つのデータが与えられたとき、動的潜在変数モデルの形式で制御プログラムの表現を行い、細胞特性に関連する転写モジュールを網羅的に同定する。本研究は、この問題を統計科学的観点から厳密に精査し、汎用的解析手法に仕上げ、さらに、癌研究における実質的な貢献を目指した。

３．研究の方法

(1)本研究で開発されるモデルは、統計科学の分野でスパース動的潜在変数モデル（あるいは状態空間モデル）と呼ばれる。データドリブンなスパース学習によって、動的に変動する超高次元観測データベクトルと潜在確率（状態）変数間のグラフ構造を同定し、潜在変数の逆解析によってデータ生成の背景シグナルを復元する。本研究の文脈では、データは遺伝子発現量、潜在変数は転写制御シグナル、スパース性は制御シグナルと遺伝子の相互作用関係に相当する。

また、転写因子・プロモータ相互作用のデータを準備するために、プロモータ領域から転写因子結合部位を予測する解析ツールを開発した。プロモータに埋め込まれた「モチーフ」と呼ばれる短い保存配列を発見する「モチーフ発見問題」を解くための 効率的な手法を提案するモチーフ発見問題は遺伝子の制御機構解明のために非常に重要な問題であり、生物情報学の創生期からの研究対象である 。これまでの提案手法を大別すると二つの系統に分けることができる。一つの系統は Weeder [Pavesi G et.al: Nucleic Acids Res, 32, 199-203, 2004] を初めとする単語数え上げ型アルゴリズムであり、もう一つの系統は、MEME [Bailey TL et.al: Proc Int Conf Intell Syst Mol Biol, 3, 21-29, 1995] などの統計モデルに基づく手法である。しかしながら、近年のデータの大規模化により、これら第一世代のモチーフ検出手法はその機能を果たせなくなりつつある。そこで、第二世代アルゴリズムの開発競争が始まることとなる。次世代シーケンサ以降の第二世代アルゴリズムの中で、真に実用レベルに達するものは今のところ存在しない。本研究では、タスク分割型モチーフサンプラーという独自性の高いベイジアンモデリング及び並列コンピューティングの技術を開発した。

さらに、生化学物質の濃度変化測定値から、システムの動作を規定する反応速度係数や初期状態、生化学反応ネットワークの構造を推定することを目的とし、新しい推定アルゴリズムの開発を推進した。開発手法の最も大きな特徴は、ロバストなシステムモデルを自動設計する機能にある。ノイズやパルスによる位相揺らぎ、ネットワーク構造の部分的な破壊等をシステムに与えながら、パラメータ推定やネットワークの構造改変を行う。時系列データのパターンを再現しながら、同時に、このような摂動の影響を抑制・緩和することができる頑強なシステムを、パラメータやネットワーク構造の改変によって実現する。この問題をベイズ統計学の手法で解くことができることを示した。本研究の開発手法は、代謝工学や合成生物学の分野において将来的な応用が期待される。また、開発手法を用いて、抗癌剤の薬剤動態を対象に遺伝子発現ネットワークのモデル開発を行った。他機関の共同研究者の協力のもと、薬剤耐性癌の全遺伝子の発現変化を観測し、データのパターンに基づきシミュレーションモデルを開発した。

４．研究成果

生化学反応システムを対象とするベイジアンモデリングに関する研究開発を実施した。状態空間モデルをもとに、システムの動作を決定する生化学パラメータや生体内分子ネットワークの構造を観測データから逆推定する手法を開発した。とりわけ、生化学反応システムに備わるロバスト性に着目し、モデリングの過程で、ノイズによる位相の揺らぎやパルス、ネットワーク構造の部分的な破壊など、摂動に対して頑強なシステムモデルを自動設計することを可能とする極めてユニークな方法論を構築することに成功した。

また、開発したモデリング技術を活用して抗癌剤の薬剤動態を対象とした遺伝子発現ネットワークのモデリングを行った。他機関の共同研究者による協力のもと、薬剤耐性を獲得した肺癌細胞の全遺伝子の発現状態を観測して、得られたデータのパターンを高精度に再現できるシミュレーションモデルを構築した。この癌細胞の転写制御に関しては、現時点で解明されていないメカニズムがかなり多いため、モデル開発において多くの課題が残されているが、予測という観点ではある程度の精度を持つモデルが出来上がった。現在は未知のメカニズムの解明へ向けたトランスクリプトーム・データ解析を実施しながら、モデルのさらなる性能強化を目指して研究を行っている。

本研究の副産物として、極めて高性能なモチーフ発見プログラムの開発に成功した。従来法の致命的な欠陥は、局所解（モチーフ配列）へのトラップにある。アルゴリズムの初期条件を変え繰り返し計算を実行しても、情報量過多の同じ疑似モチーフに何度もトラップされてしまう。疑似モチーフは、反復配列や GC 含有量が異常に高い、医学的応用上意味のないものである。このような欠陥は全ての既存法に共通するもので、これまでに非常に多くの知識が見過ごされてきたに違いない。本研究で開発されたタスク分割型モチーフサンプラーは、単純かつ画期的なアイデアで、この欠陥を克服した。基本設計概念は、以下のように説明される：複数のアルゴリズムを並列実行する際、探索軌道間に「反発作用」を与え、各々が異なるモチーフ配列に到達するように作業分担させる。このタスク分割機能によって、多様なモチーフ配列を重複なく、たった一回の並列計算で検出できるようになった。従来法のいずれにも、このような設計概念はなく、多様なモチーフ配列の列挙というタスクにおいて他を凌駕する。さらに計算の並列度と検出性能が直に結びついており、今後超高並列計算機との連携によって、潜在的統計性能は飛躍的に増大する可能性が高い。

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　3 件）

(1) Yamauchi, M., Yamaguchi, R., Nakata, A., Kohno, T., Nagasaki, M., Shimamura, T., Imoto, S., Saito, A., Ueno, K., Hatanaka, Y., Yoshida, R., Higuchi, T., Nomura, M., Beer, D., Yokota, J., Miyano, S., Gotoh, N., Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma, PLoS One, 7(9), e43923, doi:10.1371/journal.pone.0043923, 2012（査読有）

(2) Kawano, S., Shimamura, T., Niida, A., Imoto, S., Yamaguchi, R., Nagasaki, M., Yoshida, R., Print, C., Miyano, S., Identifying gene pathways associated with cancer characteristics via sparse statistical methods, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4), 966-972, doi:10.1109/TCBB.2012.48, 2012（査読有）

(3) Tamada, Y., Yamaguchi, R., Imoto, S., Hirose, O., Yoshida, R., Nagasaki, M., Miyano, M., SiGN-SSM: open source parallel software for estimating gene networks with state space models, Bioinformatics, 27(8), 1172-1173, doi: 10.1093/bioinformatics/btr078, 2011（査読有）

〔学会発表〕（計　17 件）

(1) 吉田亮，ベイズ統計学に基づく医薬品化合物/生化学反応システムの設計，設計情報学研究会，京都，2013 年 3 月

(2) Yoshida, R., Bayesian Robust Networking, Bayesian Inference and Stochastic Computation 2012 workshop, 東京，2012 年 6 月

(3) Yamashita, H., Yoshida, R., Iba, Y. Preimage analysis of chemical structures, Bayesian Inference and Stochastic Computation 2012 workshop, 東京，2012 年 6 月

(4) Yoshida, R., Bayesian sparse reconstruction: Latent factor analysis of gene regulatory programs, ISBA2012 World Meeting，京都，2012 年 6 月

(5) 池端久貴，吉田亮，ベイジアン・コンピューティングと DNA 配列からのプロモータ部位の予測，2012 年度統計関連学会連合大会，北海道，2012 年 9 月

(6) 吉田亮，山下博史，伊庭幸人，分子設計のカーネル逆像問題について：医薬品開発への応用，2012 年度統計関連学会連合大会，北海道，2012 年 9 月

(7) 吉田亮，データ同化法にもとづくデータ統合・生体シミュレーション技術，第 50 回日本生物物理学会年会，名古屋，2012 年 9 月

(8) 山下博史，吉田亮，伊庭幸人，樋口知之，創薬を支援するデータ駆動型化合物設計，第 15 回情報論的学習理論ワークショップ(IBIS2012)，東京，2012 年 11 月

(9) 池端久貴，吉田亮，タスク分割型ベイジアンモデリングに基づく DNA モチーフ配列の探索，第 15 回情報論的学習理論ワークショップ(IBIS2012)，東京，2012 年 11 月

(10) 吉田亮，大規模化合物データに基づく創薬情報学の新しいかたち，情報・システム研究機構シンポジウム 2012　生命科学のビッグデータ革命　－仮想から現実へ，東京，2012 年 11 月

(11) 吉田亮，ビッグデータ時代のバイオサイエンスと機械学習，ビッグデータとスマートな社会　第 5 回ビッグデータに立ち向かう機械学習，東京，2012 年 11

月

(12) Yoshida, R., Bayesian methods for making systems, molecules and others, Workshop on Applied Physics and Statistics for Quantitative Biology, 東京, 2012年11月

(13) Yoshida, R., Bayesian statistical modeling and computational technologies in systems biology and chemoinformatics, 理化学研究所ゲノム医科学研究センター RCAI Seminar Series 2012, 横浜, 2012年6月

(14) 吉田亮, 組織的メタ遺伝子解析の多重検定問題, 2011年度統計関連学会連合大会, 福岡, 2011年9月

(15) Yoshida, R., Bayesian supercomputing tackles cancers, Joint Meeting of The 2011 Taipei International Statistical Symposium and 7th Conference of The Asian Regional Section of the IASC, Taipei, Taiwan, 2011年12月

(16) Yoshida, R., Nagao, H., Saito, M.M., Higuchi, T., Dynamic Bayesian modelling of biological pathways and decoupling of hidden regulatory signals using nonlinear state space models, Yeditepe International Research Conference on Bayesian Learning (YIRCoBL 2011), Istanbul, Turky. 2011年6月

(17) Nagao, H., Yoshida, R., Higuchi, T., Hybrid Bayesian filter algorithm for multivariate time-series modellings on cloud computing systems, Yeditepe International Research Conference on Bayesian Learning (YIRCoBL 2011), Istanbul, Turky, 2011年6月

〔図書〕(計1件)

[分担執筆] 樋口知之『データ同化入門』 朝倉書店、2011 (分担執筆「遺伝子発現調節モデルのデータ同化」)

６．研究組織

(1)研究代表者

吉田　亮 (YOSHIDA RYO)

統計数理研究所・モデリング研究系・准教授

研究者番号：70401263