科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　９　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2012～2013
課題番号：　２４７００２９２

研究課題名（和文）多種薬剤応答パネルデータからの知識抽出法の開発

研究課題名（英文）Development of methods to extact kowledge from drug-stimulated gene expression time-coure compendium data

研究代表者
山口　類（Yamaguchi, Rui）

東京大学・医科学研究所・講師

研究者番号：９０３８０６７５

交付決定額（研究期間全体）：（直接経費）　　3,100,000円、（間接経費）　　930,000円

研究成果の概要（和文）：本研究では、多種薬剤刺激応答パネルデータからの知識抽出のための統計的データ解析手法の開発を行った。薬剤構造と薬剤刺激に対する生体の応答プロファイルの異種情報統合の検討を進め、薬剤刺激下の遺伝子発現時系列データから薬剤応答の特徴量抽出法を開発し、その特徴量を利用した薬剤構造から薬剤効果プロファイルを予測するための統計的時系列モデルに基づく手法を開発した。

研究成果の概要（英文）：In this study, we investigated a statistical methodology to extract useful knowledge from gene-expression time-course compendium data of rats under multiple-drug stimulation. We studied methods to integrate heterogeneous information from drug structures and reactions of biological systems to drug stimulation and invented a method to obtain summarized features of the reactions using time-series model. Finally, we developed a method to predict the summarized biological reactions from the drug structures based on a statistical time-series model.

研究分野：　総合領域

科研費の分科・細目：　情報学・生体生命情報学

キーワード：　バイオインフォマティクス　統計的時系列モデル　薬剤応答　薬剤構造　時系列遺伝子発現データ

１．研究開始当初の背景

近年、実験・観測技術の発達で遺伝子発現等の生体に関する網羅的なデータを比較的容易にかつ大量に得ることができるようになった。これからは、それらの膨大なデータから生体に関する有用な情報を抽出し実際の医療や創薬につなげていくためのデータ解析技術の開発が重要な課題である。

研究代表者は、これまで薬剤投与下の時系列遺伝子発現データから、生体システムに対する薬剤作用機序に関わる情報を抽出するために統計的時系列モデル（状態空間モデル、ベクトル自己回帰モデル）を用いて遺伝子制御構造ネットワークの推定および、システムの動的特性の差異に関わる遺伝子群を同定する手法を開発してきた。本質的に動的システムである生体の応答を知る上で、動的統計モデルを用いるアプローチは有効である。研究代表者は適切な次元縮約、パラメータ制約や正則化を考案、適用することにより高次元遺伝子発現時系列データからのシステム推定を可能としてきた。それらの手法を用いた例として、抗がん剤 Gefitinib (Iressa)のオフターゲットに関わる遺伝子候補群の探索等がある。

そのような経験の中で薬剤刺激下の遺伝子発現時系列データから得られる生体システム情報の重要性を実感するとともに、より詳細な薬剤作用機序や生体応答に関わる情報を抽出するためには、複数の薬剤刺激からの情報および薬剤の構造の情報を統合的に解析することのできる手法が必要であると考えた。そこで今回研究を行った手法の原型を着想するに至った。しかしながら当時そのような手法を開発するために必要な、動的遺伝子発現情報を含む薬剤反応パネルデータが公には存在しなかった。ところが 2011 年に、医薬基盤研究所から、そのような情報を含むデータベースが公開された（Open TG-GATEs http://toxico.nibio.go.jp/）。このことにより手法の開発が現実的なものとなった。このデータベースには 131 種類の化合物をラットに対して、それぞれ複数用量条件で作用させ経時的に遺伝子発現を取得したデータセットが含まれる。これは通常の薬剤反応パネルデータ（NCI60, CMAP, Cancer Cell Line Encyclopedia 他）が、薬剤投与後もしくは無投与状態での一時点での発現情報しか含まないことに比べて類例のないものであった。

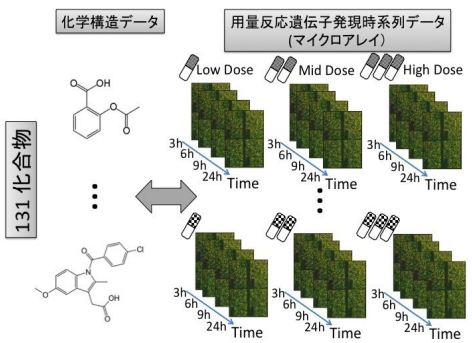

図　1 多種薬剤反応パネルデータ（Open TG-GATEs イメージ）

２．研究の目的

上述の背景のもと　本研究では、薬剤構造データおよび動的生体反応モデルから抽出された生体システムの情報を統合することにより、薬剤構造から生体の反応を予測し、また逆に、生体反応プロファイルから薬剤構造を予測する手法を開発することを目的とした。前者は薬剤バーチャルスクリーニングへ、後者はバーチャルドラッグデザインへの応用が大きく期待される。

本研究では薬剤の化学構造情報と、薬剤投与下の時系列遺伝子発現データから動的モデリングにより抽出される生体応答システムの情報（制御構造および動的特性）を鍵に異種情報を統合することにより、上記の予測を可能にする手法の確立を目指した。

３．研究の方法

本研究では化学構造と生体システムの異種情報を統合し予測モデルを得る枠組みとして、カーネル正準相関分析およびベイズ型エミュレーション法に基づく方法を検討し開発を進めた。カーネル正準相関分析では、高次元空間特徴量間の正準相関構造を利用し、入力に対して最も近傍の出力例を提示することにより上記の予測を行うことができる。モデルの構造は対称的であり、薬剤構造および発現プロファイルの予測を、入力変数に応じて一つのモデルで行うことが可能である。化合物構造の類似度を反映したカーネル行列としては様々なものが提案されている。しかしながら、薬剤刺激に対する生体反応プロファイルに関しては、生体の動的システムの、どのような情報をどのようにカーネル行列に投入するかは自明ではない。本研究では、時系列遺伝子発現情報から動的モデルにより推定される遺伝子制御構造、薬剤標的情報および、薬剤応答プロファイルを用いる方法について研究を進めた。このことにより単なる発現プロファイル間の相関係数に基づく類似情報に比べ、より生体システムに関する高次情報（制御構造および動的特性）を用いた情報抽出が期待される。

またベイズ型エミュレーション法では、通常非常に計算量の大きなシミュレーションモデルの挙動を、シミュレーションへの入力パラメータと出力の情報から、より構造のシンプルな統計モデルで模倣することを狙いとするが (Liu and West, 2009)、本研究では生体自身を、複雑なシミュレーションモデルと見なし、その挙動を動的統計モデルで模倣するという観点で手法の開発を進めた。ここ

での入力パラメータとしては薬剤構造となる。

本研究では、薬剤構造と薬剤刺激に対する生体の応答プロファイルの異種情報統合に基づく情報抽出および予測を目的とする。ここで薬剤刺激に対する応答としては、観測遺伝子群の発現値そのものを利用することも

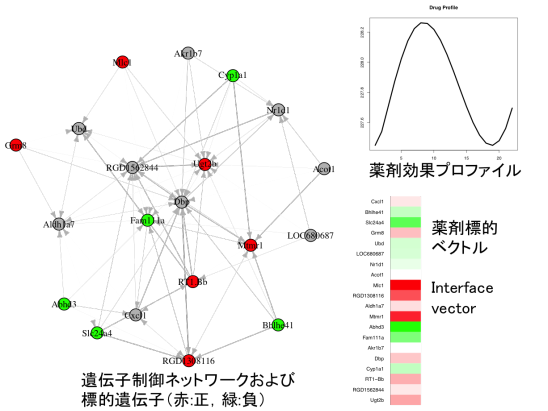

図 2　遺伝子ネットワーク、薬剤効果プロファイル、薬剤標的ベクトル（アセトアミノフェン）

考えられるが、そのデータは一般に高次元の時系列であり、そこからシステム全体の応答を把握するのは困難である。本研究では薬剤に対する生体の応答プロファイルとして高次元遺伝子発現時系列データから動的モデルを用いて、縮約抽出された低次元の薬剤効果プロファイルを用いる着想を得て、そのための時系列モデル推定手法および薬剤ターゲット同定手法の開発も進めた。また上記の動的モデルからは、同時に、生体内制御構造を反映したネットワーク情報および薬剤標的情報情報も推定できる。それらの薬剤効果プロファイルおよび、薬剤標的プロファイルは上記のカーネル正準相関分析に基づく方法と、ベイズ型エミュレーション法に基づく方法の入力および出力情報として用いた。

４．研究成果

まず各種薬剤刺激に対する多次元の遺伝子発現時系列情報を低次元に縮約した動的薬剤効果、遺伝子間制御ネットワーク、薬剤標的情報を遺伝子発現データから抽出するために、ベクトル自己回帰型状態空間モデルを基に、遺伝子制御関係を表す隣接行列および薬剤標的を表すインターフェース行列の L1 正則化に基づく疎学習の手法の開発を進めた。隣接行列の疎学習に関してはこれまでも手法が提案されていたが、システムへの制御シグナルのインターフェース行列の疎学習の定式化はなされていなかった。これにより遺伝子ネットワーク上での薬剤標的および効果の推定が行えるようになった。これは研究代表者がこれまで開発してきた状態ベクトルの次元縮小に基づく薬剤標的探索モデルに比べて解釈可能性が高く、応用範囲広いものである。TG-GATEs から得られた時系列データ(3h，6h，9h，24 時間)のデータは、スプライン関数を用いて補完し 3 時間刻みの等間隔データとした。しかし補完したとしても、一般に遺伝子数は観測時点数より大きく、通常の時系列モデルのパラメータ推定を行うことは困難である。しかし上述の疎学習に基づく手法を用いることでパラメータの推定が可能となった。

上記手法を、TG-GATEs から得られた実データに適用し薬剤効果プロファイルおよび薬剤標的、遺伝子間制御ネットワークの推定を図った。図２に実際に得られた遺伝子ネットワークおよび薬剤効果プロファイルおよび薬剤標的ベクトルの例を示す。ここで遺伝子ネットワークは薬剤刺激なしの条件で得られた時系列遺伝子発現データから推定されたものであり、基本となる制御構造である。また薬剤効果プロファイルは、アセトアミノフェン刺激下の遺伝子発現時系列データと上記の無刺激下薬剤ネットワークから推定されたプロファイルである。また薬剤標的ベクトルで薬剤効果プロファイルにかかる係数の符号に応じて色がついており、赤が正の符号、緑が負の符号がある。ネットワーク上では、その符号の色に応じて、遺伝子を表すノードに色が着いている。これを見ると薬剤効果の方向が異なるもの、また制御を強く受けるものがあることがわかる。薬剤標的ベクトルは、疎学習を行っており、薬剤標的をよ

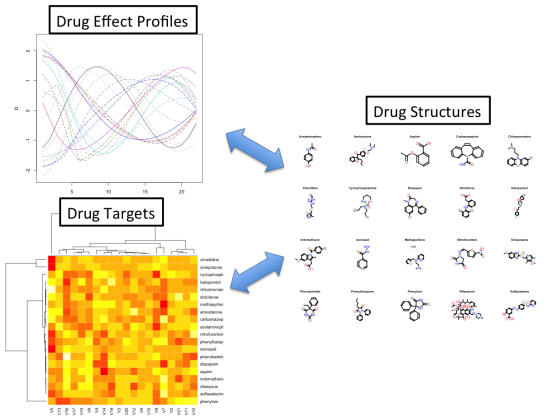

図 3 多種薬剤効果プロファイル、薬剤標的行列、薬剤構造

り明瞭に予測出来ている。図３に薬剤刺激パネルデータで用いられている薬剤の構造、薬剤刺激下の時系列遺伝子発現データから推定された、薬剤効果プロファイルおよび薬剤標的ベクトルをまとめた薬剤標的ベクトルである。推定された、薬剤プロファイルおよび薬剤標的は、薬剤刺激に反応して変動する多数の遺伝子の動きを、遺伝子制御構造を基に縮約した特徴量であり。それぞれ薬剤構造から予測ができると、未知の薬剤構造からの生体反応および薬剤標的の予測ができ有用である。

次に上記の二つの特徴量およびカーネル

正準相関分析に基づく手法の検討開発を行った。薬剤構造のカーネルとしては、これまでに提案されている種々の化合物カーネル（Kashima et al.2013 等）に基づくものや、薬剤フィンガープリントから構成した多項式カーネルを適用し検討を行った。推定された隠れシグナルとしての薬剤構造シグナルおよび薬剤プロファイルシグナルの間には高い相関が見られたが、過学習が起きがちであった。正則化により過学習を押さえることもできるが、これは時点数が少ないことによる薬剤プロファイルの情報量不足も原因と考えられる。他の情報を取り込むことによるより精度の高い予測手法の開発が期待される。

次にベイズ型エミュレーターに基づく薬剤構造からの薬剤プロファイルの予測を図った。ベイズ型エミュレーターの定式化および推定方法は、Liu and West, 2009 にならった。エミュレーター中で入力パラメータとして扱われる、薬剤構造の特徴量はフィンガープリントからなるベクトルを用いた。図４は、ある１０組の薬剤構造と薬剤効果プロファイルをそれぞれ入力、出力とするトレーニングセットに対し、ベイズ型エミュレーターを適用しパラメータを学習し、トレーニングとは別の薬剤からなるテストセットに対して、薬剤構造からの薬剤効果プロファイルの予測をおこなったものである。青線が真値、黒線が予測値である。予測値は、MCMC によりサンプルされた 1000 組の薬剤効果プロファイル実現値の中央値である。結果、よく予測がうまくいっている薬剤がある。これらは、この方式により未知の薬剤に対する、生体の反応の予測が可能であることを示している。一方、大きく予測が外れているものもある。一つの原因としては特徴量ベクトルがうまく薬剤構造の特徴を表現しきれていない可能性が上げられる。実際、フィンガープリントの特徴量ベクトルをみると情報量の乏しい要素が多く、実効的な次元数は小さい、方向性としてはより大きな特徴量を生成するフィンガープリントを用い、かつ PCA などにより得られた高次元の特徴量ベクトルを縮約することにより、情報量が高く低次元の特徴量ベクトルを用いることが考えられる。また現在の方式では、サンプリングの定式化の関係上、特徴量間の類似度表現として特定のものしか用いることができなくなっているためと考えられる。今後、これまでに提案されている、化合物のカーネル構造を柔軟にプラグインすることのできるベイズ型エミュレーターの定式化を進める予定である。

このような手法で薬剤構造からの生体反応の予測を行った例はなく、今後、薬剤構造だけではなく、付随する薬理情報などの統合を図ることで、更なる性能向上が期待される、また投与される側の差異も加味したモデル化も重要である。現在、多種細胞株に対する薬剤刺激パネルデータはあるが、時系列で取得されたそのようなデータセットは公開されていない。そのようなデータセットと、上述のモデルを発展させることで、薬剤耐性機構に関する情報および、個別化医療に資する情報抽出が実現することが期待される。

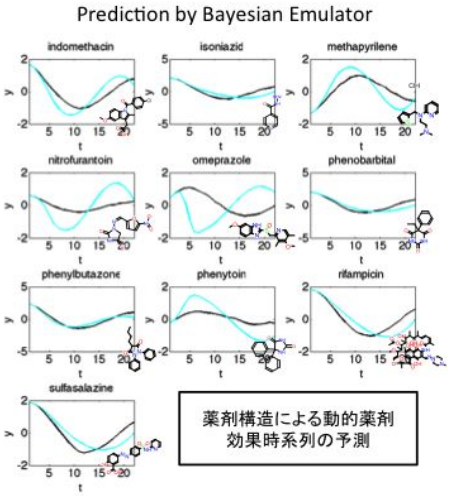

図4 薬剤構造を入力としたベイズ型エミュレータによる薬剤効果予測（青：真値, 黒：予測値)

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　１　件）
① Ogami K, Yamaguchi R, Imoto S, Tamada Y, Araki H, Print C, Miyano S Computational gene network analysis reveals TNF-induced angiogenesis, BMC Systems Biology, 査読有, 2012, 6(Supple2)

〔学会発表〕（計　２　件）
① 尾上健太郎、山口類、井元清也、玉田嘉紀、荒木啓充、Cristin Pring、宮野悟
Computational gene network analysis reveals TNF-induced angiogenesis、
The 23rd International Conference on Genome Informatics、2012 年 12 月 12 日、台湾
2. Maruyama Y, Yamaguchi R, Imoto S, Miyano S. Modeling of dynamic drug effect on gene networks
The 13th Annual International Workshop on Bioinformatics and Systems Biology、2013 年 8 月 1 日、京都大学

〔図書〕（計　０　件）

〔産業財産権〕
○出願状況（計　　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　　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
山口　類（YAMAGUCHI RUI）
東京大学・医科学研究所・講師
研究者番号：90380675

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：