様式 C-19

科学研究費補助金研究成果報告書

平成２２年５月３１日現在

研究種目：若手研究（Ｂ）
研究期間：２００８～２００９
課題番号：２０７００２６９
研究課題名（和文）　生物ネットワーク構造に基づく統合的データマイニング手法の構築

研究課題名（英文）　Integrative data mining for analyzing biological networks

研究代表者
志賀　元紀（SHIGA MOTOKI）
京都大学・化学研究所・助教
研究者番号：２０４３７２６３

研究成果の概要（和文）：
本研究では、生物ネットワークと関連情報を統合的に解析するデータマイニング法を開発した。新たに開発した手法は、複数の相互作用ネットワーク上の遺伝子（ノード）のクラスタリング法および機能予測法、遺伝子発現量とネットワークを統合するクラスタリングに基づく遺伝子機能のアノテーション法、また、統計的なアプローチと頻出パターンの抽出手法を組み合わせることにより相互作用の要素を高速に予測する手法である。

研究成果の概要（英文）：
For analyzing heterogeneous biological networks and related information sources, I developed integrative data mining methods. My developed new methods are clustering nodes on multiple networks, an annotation method of gene functions, and fast mining methods by combining frequent pattern mining and a statistical hypothesis test.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,300,000	390,000	1,690,000
２００９年度	800,000	240,000	1,040,000
総　計	2,100,000	630,000	2,730,000

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、生物ネットワーク、遺伝子機能解析、クラスタリング、教師有り学習

１．研究開始当初の背景

昨今のゲノム研究・疾病研究・創薬研究では、遺伝子配列のみならず、遺伝子発現量、タンパク質相互作用、立体構造、薬物の化学的性質などの多種多様な形式のデータが観測され、データベースが急速に蓄積されている。これらのデータを有功に利用するためには、膨大なデータから必要な情報を高速に抽出できる高度なマイニング手法が必要不可欠である。こうした手法を構築するに当たり、従来の統計学の範疇で扱うことができない、配列の文字列、低分子化合物の構造式グラフ、タンパク質の相互作用ネットワークなどの非構造化データの取り扱いが困難な問題であり、特に、生物ネットワークの解析は、生物のシステムを理解するために今後も重要な課題である。

生物ネットワークを含む実世界のネットワークは、非常に大規模であるため、その計

算規模などを考慮した研究が様々に行われている。例えば、ネットワーク上のノードのクラスタリングにおいては、Newman によるモジュラリティー指標のような単純な統計量に基づく手法や高速なアルゴリズムに基づくマイニング法によりエッジが密な部分ネットワーク構造を抽出する手法が提案されている。

本研究代表者は、前年度の研究において、モジュラリティー指標を改良した正規化モジュラリティーという新しい指標を提案した。さらに、ネットワーク形式データを数値ベクトル形式データと最適に組み合わせるクラスタリング手法を開発した。そして、開発手法を、遺伝子発現量と代謝パスウェイを使用する遺伝子クラスタリングの問題に応用し、代表者の開発した統合的な手法の優位性を報告した。そして、個々のデータを解析した場合よりも安定した性能を向上できることを確認している。

このように高精度な手法を用いることで、データ解析の結果の信頼性が高まる。そこで、今後、生物ネットワークを深く理解するためには、多種多様な非構造化データを統合的に取り扱うデータマイニング法を開発することが重要である。

２．研究の目的

本研究の内容および成果を理解しやすくするため、具体的に従事した研究ごとに分けて記載をした。２．研究の目的～４．研究成果において、小節に番号を付けて、研究項目を対応付けている。

（1）ネットワーク上のノードのクラスタリングの精密化

ネットワーク上のノードのクラスタリングの精度向上を目指すために、まず、代表者が開発している手法、および、様々にある既存手法の性能を評価する必要がある。本研究では、実ネットワークを特徴付ける性質を考慮したいくつかの生成モデルに対する性能を評価した。その際、代表者が開発した正規化モジュラリティーの性質およびスペクトラルクラスタリングの計算量およびその優位性を数値実験により既存手法と比較し、これらの性能の考察に基づき、提案するアプローチの改良を目的とする。

（2）統合的クラスタリングによる遺伝子アノテーション法

DNA マイクロアレイは細胞内の膨大な数の遺伝子の活性を同時に計測できるハイスループットな技術である。しかしながら、観測値に含まれる雑音が問題になりやすいためにロバストな解析法が要求される。ところで、最近では，多種多様なゲノム情報が採取され，データベースとして世の中に提供され，遺伝子機能のデータベース Gene Ontology (GO)，代謝情報の豊富さが特徴的な統合データベース Kyoto Encyclopedia of Genes and Genomes (KEGG) のように様々なゲノム情報が容易に入手可能である。そこで，発現量のみではなく，こうした既知のゲノム情報を利用することで機能解析の精度を向上させることを目的とする。

（3）複数ネットワークを統合する遺伝子ネットワークの解析

遺伝子に関する相互作用（ネットワーク）は、タンパク質相互作用、遺伝子発現の制御関係のように様々な形式が存在する。遺伝子は、これらの全ての複雑な相互作用に基づいて、個々の機能を働かせることができる。こうしたことから、関係する全ての種類の相互作用を用いて遺伝子機能の解析（クラスタリング・機能予測）を実行する必要がある。しかしながら、こうした統合的な技術が十分に整備されていないために、個々のネットワークを解析した出力に基づく２段階的なアドホックな手法が用いられてきた。そこで、複数のネットワークを統合できる遺伝子ネットワーク解析法を構築する。

（4）多様なゲノム情報に基づく相互作用の解析

ゲノム情報には、遺伝子発現量のように数値ベクトル形式データのみではなく、カテゴリ変数のような質的なデータ、化合物のようなグラフ形式データなど様々に存在する。また、多様な情報が混在するゲノム情報解析は、データ規模が膨大であるために、効率的なデータ解析法が求められる。そこで、こうした多様なゲノム情報から、重要な部分を抽出するマイニング手法および予測における重要度に基づくランキング手法の構築を目的とする。

３．研究の方法

（1）ネットワーク上のノードのクラスタリングの精度向上を目指すために、まず、様々な既存手法の性能を評価する必要がある。既存手法の評価には、非常に単純なモデルとして知られるランダムグラフが頻繁に使用されてきた。しかしながら、ランダムグラフは、実世界のネットワークとは、かけ離れた性質をもつために、実問題に則した十分な性能評価をできていなかった。しかしながら、近年、

スケールフリー性、階層性、モジュラー構造などの実ネットワークを特徴付ける性質を考慮したモデルが多数提案されており、本研究では、これらを用いることによって、より実ネットワークに則した生成モデルにおいて性能を評価した。

（2）遺伝子の機能をアノテーションするために、遺伝子発現量と遺伝子配列の情報を統合する。遺伝子配列の情報は、そのままの形式では遺伝子発現量データと統合することが難しいために、まず、遺伝子ペアの配列類似性を用いて、遺伝子ネットワークを構築した。そして、遺伝子発現量のクラスタリングのコスト関数、および、遺伝子ネットワークのクラスタリングのコスト関数を導出し、これらを統合した。そして、統合されたコスト関数を最小化するように、クラスタリングのアルゴリズムを導出した。通常、こうしたデータ解析では、本質的に高次元データとなるために、コスト関数における低次元スペクトラル空間にマッピングして、その次元の上でクラスタリングすることにより、最適化過程で局所解に陥ることが少なくなるようにアルゴリズムを設計した。

（3）単一のネットワークに対してノードをクラスタリングする場合、隠れ変数をもつネットワークの確率的な生成モデルが使用される。このモデルを拡張して、複数のネットワークに対する新しい確率モデルを設計した。そのモデルでは、ノードに仮定されるクラスターラベルは全ネットワークにわたって共通であるが、ネットワークごとのエッジ生成に関する特徴が異なることを仮定している。こうしたモデルの仮定によって、従来のモデルのように複数のネットワークの特徴が平均化され、重要な個別ネットワークの特徴が相殺されることがない。また、確率モデルのパラメータ学習には、最尤法のような点推定よりもロバストなベイズ学習法を採用した。

また、複数ネットワークを用いた遺伝子機能の分類法を構築した。関連する既存手法は、各ネットワークに一様な重みを付けて一つのネットワークに統合する手法であるが、こうした手法では、予測に関して重要な部分データとノイズとなる部分データが相殺されてしまう。そこで、ネットワークの部分構造を予測への寄与度によって、振るい分けるアプローチを採用した。

（4）糖鎖は，発生・分化，免疫系，癌の転移など、多彩な生命現象や疾患に重要なシグナル伝達ネットワークと関係する。膨大な糖鎖構造から、こうした生命現象に関係する部分構造を抽出するために、頻出パタンマイニングと統計的検定に基づく手法を開発した。

また、遺伝子発現量と一塩基多型（SNP）の関係を調べるための統計的検定法を開発した。ヒトの SNP の数は約２００万と膨大なために、厳密な統計的検定を実行する前に、高速なスクリーニング処理をするという２段階処理をするアルゴリズムにより計算コストの軽減を図った。

４．研究成果

（1）ネットワーク上のノードのクラスタリングをスペクトラル法によって実装し、既存手法と性能を比較した。比較手法には、カーネル k-means のように新しい手法を含んでいる。人工データを用いた数値実験によって、提案するアプローチは計算時間および精度の面で従来の手法よりも優れていることを示し、さらに、従来法の問題点を理論的に考察した。また、一般に公開されているイースト（出芽酵母）の遺伝子ネットワーク（代謝パスウェイおよびタンパク質相互作用）を用いた大規模データの実験においても、計算時間および解析精度の観点から良い結果を得られることを確認した。現在、この研究で得られた研究成果を論文にまとめて学術雑誌に投稿している。

（2）前章３-２で述べた手法の遺伝子機能クラスタリングに基づくアノテーションを評価するための数値実験を行った。アノテーションは、開発法の出力クラスターと既知の遺伝子機能のラベルの間のオーバーラップを測定し、オーバーラップが最も大きい機能ラベルを各クラスターに割り当てる手法である。

イーストのゲノム情報を使用して、こうした遺伝子アノテーション法の性能を既存手法と比較した。比較実験において、開発法は、サポートベクターマシーンのような一般的に高性能として知られる予測法よりも良い性能を確認できた。また、本手法は、超幾何分布の p-value に基づくアノテーション手法 GO Term Finder と比較しても優れていることを確認できた。GO Term Finder は、クラスター割り当てを確率的な事象として捉え、スコアを補正するものであるが、本研究の問題に対しては、直接的にオーバーラップを図る方が良いことを意味している。これらの一連の研究成果をまとめ、International Workshop on Bioinformatics and Systems Biology において発表し、そして、学術論文誌 Genome Informatics に掲載している。

（3）ベイズ学習法によるクラスタリングは、クラスターのラベルを推定するのではなく、クラスターのラベルの事後分布を推定する

ものである。そこで、観測データの関数として、ラベルの事後分布を導出する必要があるが、本研究で採用するような隠れ変数を仮定する確率モデルに対しては、ラベルの事後分布を解析的に導出することができない。そこで、変分ベイズ学習法と呼ばれる近似的なアプローチにより、事後分布を近似的に導出した。

ここで得られたアルゴリズムの性能を人工データおよび実データより検証した。人工データの実験では、観測されるグラフ数の増加と性能の関係、および、ノイズ成分に対応するエッジに対するロバスト性を検証した。比較には、論理演算や線形演算によりネットワークを統合して単一ネットワーク用のアルゴリズムを採用する手法、また、複数ネットワーク向けのスペクトラルクラスタリングなど様々な手法と比較した。上述した２つの観点における比較では、開発手法の性能が優位であることを確認できている。また、イーストの複数のネットワーク情報を使用した数値実験により提案法の性能を検証した。使用したネットワークは、遺伝子相互作用、タンパク質相互作用、発現制御関係、配列類似性などを含んでいる。クラスリングの性能を、アルゴリズムより出力されたクラスターと既知の遺伝子機能ラベルとのオーバーラップの度合いにより測った。その結果として、提案法が上述の比較法よりも優れていることを確認できた。また、出力されたクラスター構造を詳細に調べることにより、性能が良くなる理由は、ネットワークごとに異なるグラフ生成過程を仮定することに起因していることを確認できた。

遺伝子機能の予測法に関しても、同様の遺伝子ネットワークを用いて、その性能を検証した。前述したように、本研究で開発した手法は、遺伝子機能予測に重要となる部分的なネットワークを強調する手法であった。そこで、遺伝子機能の予測精度を図るだけではなく、強調された部分ネットワークを詳細に調べ、本手法の有用性を確認した。

これらの研究成果をそれぞれ論文に執筆し、現在、論文誌に投稿している。

（４）糖鎖構造の重要部分構造の予測に関しては、共通する頻出部分構造の中で重要なもののみを取り出すために、部分構造の大きさと統計的な意味での有用性の両方を満たすように、部分構造をランキングするアプローチを提案した。そして、実データを使用した実機能の予測実験によって、木カーネルを用いたサポートベクターマシーンの予測よりも優れていることを確認できた。

また、遺伝子発現量と一塩基多型（SNP）の交互作用の高速検定法を開発した。全ゲノムの膨大情報に対して、交互作用検定することは、非線形最適化のステップがボトルネックとなり計算量が膨大になるために、まず、２次までの統計量を用いた高速スクリーニング法をほどこした。このスクリーニングにより厳密な統計的検定の結果により否定されるほとんどのデータが振るい落とされ、全体的な計算量を大幅に削減できる。実際、ヒトゲノム情報を用いた数値実験においても十倍以上高速になることを確認できた。また、検出された遺伝子ペアの中で、疾患にかかわる重要な遺伝子ペアを検出でしている。

それぞれの研究成果をまとめて、学術論文誌 Bioinformatics に掲載している。

一連の研究で得られた手法は、生物ネットワーク解析に関する基本的な技術であるものの、その応用は、ほとんどが遺伝子機能解析に限られていた。今後は、疾患メカニズムを解明するために開発手法を発展させ、ゲノムワイドな情報から疾患状態診断や創薬の応用に役立つようなデータ解析手法を研究してゆきたいと考えている。

５．主な発表論文等

〔雑誌論文〕（計３件）

① Limin Li, Motoki Shiga, Wai-ki Ching, Hiroshi Mamitsuka, "Annotating Gene Functions with Integrative Spectral Clustering on Microarray Expressions and Sequences", Genome Informatics, 査読あり, vol.22, p.95-120, 2009.

② Mitsunori Kayano, Ichigaku Takigawa, Motoki Shiga, Koji Tsuda, Hiroshi Mamitsuka, "Efficiently finding genome-wide three-way gene interactions from transcript- and genotype-data", Bioinformatics, 査読あり, vol.25, p.2735-2743, 2009.

③ Kosuke Hashimoto, Ichigaku Takigawa, Motoki Shiga, Minoru Kanehisa, Hiroshi Mamitsuka, "Mining significant tree patterns in carbohydrate sugar chains", 査読あり, Bioinformatics, vol.24, i167-i173, 2008.

〔学会発表〕（計７件）

① Limin Li, Motoki Shiga, Wai-ki Ching, Hiroshi Mamitsuka, "Annotating Gene Functions by Spectral Clustering for Combining Gene Expressions and Sequences", The 20th International Conference on Genome Informatics (GIW), Yokohama, Japan, 14-16 Dec., 2009.

② Limin Li, Motoki Shiga, Wai-ki Ching, Hiroshi Mamitsuka, "Annotating Gene

Functions with Integrative Spectral Clustering on Microarray Expressions and Sequences″, The 9th Annual International Workshop on Bioinformatics and Systems Biology, Boston, MA, USA, 27-29 July, 2009.

６．研究組織

(1)研究代表者

志賀　元紀（SHIGA MOTOKI）

京都大学・化学研究所・助教

研究者番号：２０４３７２６３