論文の内容の要旨

Modeling Methods and Efficient Algorithms for Gene Network Estimation and Network Layout

（遺伝子ネットワーク推定とネットワークレイアウトのための

モデリング法と効率的なアルゴリズム）

氏名　小　島　　要

（本文）細胞内システムの解明は，癌細胞メカニズムの理解，薬剤応答の推定，新規薬剤の開発に有用である．近年のマイクロアレイ技術の発展から大規模な遺伝子発現データの計測が可能であり，異なる条件下での遺伝子発現量の変化の比較により，条件に関連した遺伝子の同定が行われてきた．現在では，より複雑な機構の解明のため，遺伝子発現量から遺伝子間の制御関係の推定が行われており，大規模な遺伝子ネットワークの推定と理解が細胞内システムを解明するための基盤技術となっている．

遺伝子の発現データは大きく，遺伝子ノックダウンマイクロアレイデータと時系列マイクロアレイデータの二つに分けられる．ノックダウンマイクロアレイデータでは，いくつかの遺伝子をノックダウンし，定常状態になったところで遺伝子発現量が計測される．このため，定常状態での遺伝子間の制御関係の推定に用いられる．一方，時系列マイクロアレイデータでは，薬剤の投与などの刺激に反応した状態での遺伝子間の制御関係が推定される．制御関係の推定は統計モデルにより推定された因果関係を制御関係とみなすことで行われる．ノックダウンマイクロアレイデータと時系列マイクロアレイデータでは，原理となる因果関係が異なり，ノックダウンマイクロアレイデータでは，Pearl の因果律により確率変数間の条件付き独立性から因果律が推定され，ベイジアンネットワークとして因果ネットワークの実現がなされる．一方，時系列発現データでは，Granger や Pearl らの提唱により，前の時点に発生した事象により現在の事象が予測される場合に因果関係が

考えられている．

推定された遺伝子ネットワークから細胞内システムの理解を行うためには，対応するタンパク質の細胞内局在情報をはじめとする既存の生物学的情報を共に考える必要がある．しかしながら，１００以上の遺伝子の間での制御関係と対応する細胞内局在情報のリストから知見を見出すことは非常に難しく，実際の細胞内での制御関係を抽象化したグラフ表現などの可視化技術の開発が必要不可欠である．

ノックダウンマイクロアレイデータからの遺伝子ネットワーク推定はベイジアンネットワークの構造学習により行われる．これまで，条件付き独立性の検定による構造推定法と情報量規準をスコアとした構造探索を行う方法が提案されている．しかしながら，条件付き独立性の検定による手法では，有向閉路を避けるために，検定結果と異なるエッジの方向付け行われるなど理論的な問題がある．また，情報量規準をスコアとしたベイジアンネットワークの最適構造推定問題は NP 困難であり，最適構造の探索は動的計画法による効率計算手法と超並列による計算を行った場合でも３２ノードからなるネットワークの推定が限界であった．しかしながら，小さな系で見た場合，外部からの制御を無視した形で制御関係の推定が行われるため，誤った制御関係を導いてしまう可能性が高く，また外部からの制御が考慮されないことから，細胞内システムの理解に不十分である．このため，正確な制御関係の推定と細胞内システムの理解のためには，より大きな系での推定が不可欠となる．Perrier らは上位構造と呼ばれる条件付き独立性の検定やその他の統計的手法により推定された無向グラフが与えられたもとで，上位構造の部分グラフとなるような最適構造を探索する方法を提案している．これは，条件付き独立性の検定による手法の欠点を解決し，かつ構造探索空間を絞ることで，より大きな系での構造推定を可能にするものである．Perrier らのアルゴリズムは動的計画法を用いた手法であり，高い精度での５０ノードからなるベイジアンネットワーク構造の推定が可能だが，平均次数が２前後の上位構造までしか扱うことができない限界があった．そこで本研究では，上位構造が与えられたもとでの最適ベイジアンネットワーク構造学習において，上位構造を密につながった部分グラフに分割し，各部分グラフでの最適ベイジアンネットワーク構造を結合する方法を提案した．しかしながら，単純に各部分グラフでの最適ベイジアンネットワークを結合した場合，有向閉路が発生し，非有向閉路性を保つことができない可能性がある．そこで，最適ネットワークを得るための数学的な必要十分条件を導き，必要な制約条件のもとで部分グラフにおける最適構造を探索し，結合するアルゴリズムを構築した．構築したアルゴリズムでは平均次数４程度の上位構造で数百ノードからなる最適ネットワークの探索が可能であり，定常状態での遺伝子ネットワークを大きな系で推定することが可能となった．推定アルゴリズムをヒト臍帯静脈内皮細胞でのノックダウン遺伝子発現データに適用することで，９６遺伝子からなる遺伝子ネットワークの推定を行い，文献データとの比較を行った．

時系列遺伝子発現データからの遺伝子ネットワーク推定は薬剤刺激やヒートショック反応時の制御関係の解析に有効であり，ダイナミックベイジアンネットワーク，ベク

トル自己回帰モデル，状態空間モデルといったモデルを用いた手法が提案されている．既存の解析手法では，通常，遺伝子制御関係の線形性が仮定され、またデータ計測時点が等間隔であることが仮定されている．しかしながら，実際には非線形制御関係の重要性が報告されており，また時系列遺伝子発現データは観測点が等間隔でない場合が多い．そこで，本研究では非線形関係まで抽出できるようなノンパラメトリックモデルをスパース学習により推定する方法を提案した．また，推定量の漸近正規性を導き，推定された遺伝子制御関係の p 値を求めることを可能にした．また，不等間隔に計測されたデータからも正確な遺伝子ネットワークの推定を行うために，ベクトル自己回帰モデルを状態空間表現することで，不等間隔データを扱うことが可能なモデルを提案し，スパース学習法を用いることでパラメータの推定を行った．提案手法を正常ヒト小気道上皮細胞からの時系列遺伝子発現データと EGF 受容体への刺激と抗がん剤イレッサとして知られるゲフィティニブの投与を行ったヒト小気道上皮細胞からの時系列遺伝子発現データに適用した．その結果，２つの状態の細胞において推定されたネットワークを比較することで EGF 受容体を介さずにイレッサから影響を受ける遺伝子の候補が同定された．

また，遺伝子制御ネットワークはパスウェイに関連したモジュールから構成されることが知られている．そこで，この知見からモジュール内では比較的密に，モジュール間では疎に制御関係が存在することを仮定し，時系列遺伝子発現データから遺伝子の制御関係と遺伝子モジュールを同時に推定する手法を変分ベイズ法の枠組みから開発した．提案した手法をヒトの子宮頸癌細胞由来の HeLa 細胞から測定された９４の癌関連遺伝子の時系列遺伝子発現データに適用することで，遺伝子ネットワークと２つの遺伝子モジュールが同時に同定された．

遺伝子ネットワークや生物パスウェイを対象とした可視化技術として，いくつかのネットワークレイアウト手法が提案されている．しかしながら，タンパク質の細胞内局在情報を考慮した手法はほとんどなく，考慮している手法も複雑な局在情報が扱えないなど不十分なものである．グリッドレイアウトは生物ネットワーク描画アルゴリズムの一つであり，ノードの位置がグリッドポイントに制限され，エッジ間交差，ノードエッジ間交差，ノード間距離を考慮した損失関数を最小化するようにノードの位置が探索される．さらにノードの配置を局在情報を満たすグリッドポイントへ制限することで，タンパク質の局在情報を考慮することができる．既存のグリッドレイアウトアルゴリズムでは各ノードを空きグリッドポイントへ移動した時の差分コストを計算し，貪欲法の枠組みから最も損失を減らすようにノード一つを空きグリッドポイント移動することで探索が行われる．しかしながら，差分コストの計算は前回の探索ステップにおける結果をキャッシュすることで効率化が図られているものの，実時間アプリケーションとしては遅いという問題があった．さらに，得られた結果が局所最適解であることから，より良い局所最適解を得るための探索手法が望まれていた．そこで，本研究ではより良い局所最適解を得るために，既存の探索手法において時間計算量を増加させることなくノードの位置の交換についても考慮

したアルゴリズムを開発した．さらに，特にスパースなグラフ構造において有効な sweep calculation と呼ばれるアルゴリズムを開発することで時間計算量を削減し，実時間アプリケーションに耐え得るグリッドレイアウトを実現した．また，細胞局在情報を表すコンパートメントの拡大・縮小と移動を探索ステップにおいて考慮することで，より良いレイアウトを得ることが可能となった．提案レイアウトアルゴリズムをヒト小気道上皮細胞のデータから推定された遺伝子ネットワークに適用し，各遺伝子に対応したタンパク質の細胞内局在情報を用いることで，細胞内局在情報を考慮した遺伝子ネットワークのレイアウトが可能となった．得られたレイアウトは細胞内局在情報を考慮にいれることで遺伝子ネットワークの特徴を捉えたものであった．