社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－BIO－6 (2)
2006／9／15

遺伝子発現データからの遺伝子間因果関係ネットワーク推定

安富祖 仁+ 岡崎 威生++
+:琉球大学理工学研究科情報工学専攻, ++:琉球大学工学部情報工学科

遺伝子発現データには，遺伝子間の直接的影響と間接的影響に関する情報が含まれ，観測データ上では区別できない．そのため，遺伝子発現データから遺伝子間因果関係ネットワークを推定するためには，データ中から直接的影響のみを抽出する必要がある．本研究は，以下の２ステップからなる遺伝子間因果関係ネットワークの推定手法を提案した．まず偏相関係数を用いた共分散選択により直接的な相関を表現する無向グラフを推定する．次に各辺の因果方向を決定するため，遺伝子発現データとの適合度を示す評価関数を用いた探索を行い，因果関係ネットワークを推定する．本提案手法によって，ネットワークの構造に対する制約のない推定を行うことが可能となる．

Genetic Causal Network Estimation from the Gene Expression Data

Hitoshi AFUSO+ Takeo OKAZAKI++
+: Information Engineering, Graduate School of Engineering and Science, University of the Ryukyus
++: Information Engneering, Faculty of Engineering, University of the Ryukyus

Gene expression data contains direct and indirect information about the effect between genes. To estimate the genetic causal network, we need to extract the only direct effect from the data. In this study, we proposed the method for estimation of genetic causal network from gene expression data. This method consists two parts. (1) Estimate the undirected graph using the partial correlation among genes. (2) Determine the orientation of effects using the evaluation function for fitness.

1 はじめに

遺伝子間の因果関係が解明できれば，新薬の開発や病気の治療に応用できることなどから，遺伝子間因果関係ネットワークの解析に関する研究は盛んに行われている．このような遺伝子間依存関係の推定のため，多数の遺伝子の発現量変化を観測できる DNA マイクロアレイ [1] が開発された．DNA マイクロアレイから得られる遺伝子発現データによって，ある遺伝子の制御によって生じた他の遺伝子の発現量の変動を観測することができる．しかし，遺伝子発現データ中では，遺伝子の発現変動が遺伝子間の直接的な影響の結果なのか，直接的な影響の伝搬によって生じた間接的な影響によって生じたものかを，判別することができない．本研究では，遺伝子発現データを用いて，遺伝子間の因果関係を表すネットワークを推定することを目指す．

2 問題のモデル化と提案手法

遺伝子発現データからの遺伝子間因果関係ネットワーク推定問題は次のように表現できる.

ネットワーク推定問題
入力: ある遺伝子から他の遺伝子への直接的または間接的な影響の結果生じた発現量変化情報を含んだデータ
出力: 各遺伝子間の直接的な因果関係のみで表現されたネットワーク

この問題に対する従来研究として,ブーリアンネットワーク [2] や線形モデル [3], ニューラルネットワーク [4] によるアプローチなどがある. その中でも近年注目を集めている手法として, ベイジアンネットワーク [5] があげられる. ベイジアンネットワークは, 変数を頂点で, 変数間の依存関係を向きを持った辺で表現した有向グラフと, 依存状態を表した条件付確率分布によって構成される確率モデルである. しかし, ベイジアンネットワークは推定可能なグラフがサイクルを持たないものに限るという制約がある. そのため, 実際に遺伝子間の依存関係を推定するには, 適用上の問題がある.

本研究の目標は, 推定するネットワークの構造についての制約のないネットワーク推定手法の提案である. そこで, 遺伝子発現データからまず直接的な相関関係を表す無向グラフを推定し, 得られた無向グラフの各辺について因果方向を推定することで, 遺伝子間因果関係ネットワークを推定する手法を開発した.

3 直接的な相関を表す無向グラフ推定

DNA マイクロアレイ実験によって, ある遺伝子 g_i 制御の影響が伝搬した結果生じた遺伝子 g_j の発現量変化を観測したデータ w_{ij} の集合 $D = \{w_{ij} : i, j \in G\}$ が得られる. ここで G は実験で発現量を観測した遺伝子の集合である. 今各 w_{ij} の対数比を正規化 [6] したデータ w'_{ij} の集合 $D' = \{w'_{ij} : i, j \in G\}$ を遺伝子発現データとして用いる. 遺伝子発現データには, 直接的な影響による変動と, 直接的な影響の伝搬によって生じた間接的な影響による変動が含まれている.

ここで, 遺伝子発現データから, 遺伝子を頂点, 発現量変化の直接的な相関関係を辺で表現した無向グラフを推定したい. この問題に対して, グラフが部分的に既知である場合, 従来研究として Support Vector Machine[7] が提案されている. しかし, 事前情報の前提は汎用的でないため, 本研究では, そのような情報を仮定しない偏相関係数を用いた共分散選択 [8] をとりあげた.

連続値をとる変数 $X_1, X_2, \cdots, X_n$ に対して, 変数 X_i, X_j 間の偏相関係数とは, 着目している 2 変数以外の変数の影響を取り除いた相関係数である. 変数 X_i, X_j 間の偏相関係数 $r_{ij.1\cdots n}$ は次のように定義される.

$$r_{ij.1\cdots n} = \frac{r^{ij}}{\sqrt{r^{ii}r^{jj}}} \tag{1}$$

ここで, r^{ij} は変数 X_i, X_j 間の相関係数 r_{ij} を並べた相関行列 $\mathbf{R} = [r_{ij}]$ の逆行列 $\mathbf{R}^{-1}$ の ij 要素である.

遺伝子発現データから偏相関係数 $\mathbf{R}'$ を計算したとき, これらの要素内には, データに含まれる誤差によって生じた相関が含まれる. それらを除去するために共分散選択を用いる. 共分散選択では, ある 2 要素間の偏相関係数が統計的に有意であるかを, χ^2 分布に従う逸脱度を用いて推定する.

偏相関係数を用いた共分散選択のステップを以下に示す.

1. 遺伝子発現データ $D' = \{w'_{ij}\}$ から偏相関係数行列 $R = \{r_{ij.1\cdots n}\}$ を求める.
2. 偏相関係数行列 R のある要素を 0 とおいた行列 R' を求める.
3. R' に対応する分散共分散行列 Σ' を求める

4. 遺伝子発現データから得られる分散共分散行列 Σ(Full Model に対応する) と Σ'(Reduced Model に対応する) を用いて逸脱度 $dev(RM) = n \log \frac{|\Sigma'|}{|\Sigma|}$ を計算する.

5. 逸脱度 $dev(RM)$ をとる確率 p を求め, p が 0.5 以上ならば Reduced Model を採択し ($R = R'$ とする), ステップ 2. へ戻る. そうでないならそのままステップ 2. へ戻る.

6. 0 にすることができる R の要素がなくなったら終了する.

共分散選択によって得られた偏相関係数行列から, 相関が 0 でない遺伝子に対応するノードどうしを辺で結ぶことによって, 直接的な相関を表す無向グラフが得られる.

4 無向グラフの因果方向決定

因果関係ネットワークを得るため, 無向グラフの各辺について因果の向きを決定する. このとき, 偏相関係数を用いて無向グラフを推定したことによって向き付けに関する制約が生じている [8]. そのため, 観測された遺伝子発現データに最も適合した向き付けを得るためには, 複数の制約を満たした向き付けを, 発現データとの適合を示す指標で評価しながら, 最も評価が高い向き付けを探索する方法が必要である.

4.1 適合度評価関数

一般的に, 真の因果関係ネットワークを観測することは困難であるため, 真の因果関係ネットワークと候補となるネットワークの直接的な比較は難しい. そこで遺伝子発現データから推定可能な特徴量に着目した. マイクロアレイ実験から得られる遺伝子発現データは, ある遺伝子 i から他の遺伝子 j に伝搬した影響 u_{ij} によって生じる発現量の変化 $w_{ij} = f(u_{ij}) =$ を観測している. すなわち, 各遺伝子の発現量の平均によって, 各遺伝子に伝わった影響の平均を推定することができる. このとき影響は因果関係ネットワーク上をランダムウォークしながら伝搬するとみなすことができるので, 各遺伝子に伝わった影響の平均は, 因果関係ネットワークの隣接関係によって定まる. つまり, 遺伝子発現量の各遺伝子について平均は, ネットワークの隣接関係に対応して決定される各遺伝子の特徴量とみなすことができる. すなわち, ある遺伝子 i を制御したときに他の遺伝子 j に生じた発現量の変化を w_{ij} とすると, 遺伝子 g の特徴量 $\hat{\mathbf{v}}_g$ は,

$$\hat{\mathbf{v}}_g = \begin{pmatrix} \hat{v}_1 \\ \hat{v}_2 \\ \vdots \\ \hat{v}_n \end{pmatrix} \tag{2}$$

$$= \frac{1}{|G| - 1} \sum_{i \in G \setminus \{g\}} w_{ig} \tag{3}$$

と推定される.

遺伝子発現から推定可能なネットワークノードの特徴量として, これまでに PageRank[9] が提案されている. PageRank による各ノードの特徴量は, ネットワークの隣接行列を $\mathbf{A}$, その各列ベクトルを $\mathbf{a}_i$, 第 i 列ベクトルの成分を a_{ij} としたとき,

$$\mathbf{R} = \left(\frac{1}{\sum_j a_{ij}} \mathbf{a}_i \right) \tag{4}$$

で求められる行列 $\mathbf{R}$ の最大固有値に対する固有ベクトルの成分として表される. PageRank は元来, Web のハイパーリンクをリンク元からリンク先への推薦行為とみなし, 各ページがどの程度他のページから推

薦されているかを点数化したものである．各辺が推薦行為であるため，その推薦行為自体の信頼性すなわち辺の重みが，各ページの出次数によって決定される．このため，一般的な影響を表現したネットワークにおいては適用が難しい．

そこで，影響を表現したネットワーク中の各ノードが他のノードからどの程度影響を受けているかを示す特徴量の提案を目的として，PageRank を拡張した Influence Measure[10] を提案した．Influence Measure による各ノードの特徴量は，

$$\mathbf{R} = \left(\frac{1}{\max_i \sum_j a_{ij}} \mathbf{a}_i \right) \tag{5}$$

で求められる行列 $\mathbf{R}$ の最大固有値に対する固有ベクトルの成分として表される．Influence Measure は各要素が他の要素からどの程度影響を受けているかを点数化したものである．遺伝子発現データは，影響が因果関係ネットワーク上を伝搬した結果生じた発現量を観測したものであるため，本手法では Influence Measure を特徴量として用いる．遺伝子発現データから推定された各遺伝子の特徴量 $\hat{v}_g$ は，遺伝子間の影響の強さを考慮したものである．候補ネットワークは隣接関係のみで表記されているので影響の大きさをもたない．そこで，候補となるネットワークから特徴量を計算する際に，各遺伝子間の影響の強度を推定する必要がある．ここでは，各遺伝子間の影響の強度を，対応する遺伝子間の偏相関係数により推定値とする．

以上より，候補となる隣接関係 $\mathbf{A} = (a_i)$ と偏相関係数行列 $\mathbf{R} = (r_{ij})$ とするとき，遺伝子 g の特徴量 $\mathbf{v}_g$ を，

$$\mathbf{R}' = \left(\frac{r_{ij}}{\max_i r_{ij}} a_{ij} \right) \tag{6}$$

で決定される行列 $\mathbf{R}'$ の最大固有値に対応する固有ベクトルの成分で定める．

これらの遺伝子発現データから推定された $\hat{\mathbf{v}}_g$ と候補となる隣接関係 $\mathbf{A}'$ から計算された $\mathbf{v}'_g$ 間の距離が近いならば，対応する真の隣接関係 $\mathbf{A}$ と $\mathbf{A}'$ も類似していると考えられる．ここで，各特徴量間のスケールの違いを考慮する必要があるため，特徴量を並べたベクトルの方向に着目して，評価関数 $Score(\mathbf{A}')$ を以下のように定義した．

$$Score(\mathbf{A}') = \frac{\sum_{g \in G} (v_g \times v'_g)}{\sqrt{\sum_{g \in G} v_g^2} \sqrt{\sum_{g \in G} (v')_g^2}} \tag{7}$$

ここで，G は発現量を観測した遺伝子の集合とする．

4.2 因果合流制約

因果関係ネットワークにおいては，向き付けが満たすべき制約が存在する [8]．偏相関係数を用いて，各変数間の直接的な影響を推定するとき，因果の合流がある要素の親の間には，必ず相関があると推定される．因果の合流とは，複数の要素が同一の要素に影響を及ぼしていることを示す．そこで，ある向き付けにおいて因果の合流が存在するならば，その親の間には相関があると推定されていなければならない．すなわち，推定された相関関係を表現した無向グラフの隣接行列を $\mathbf{A} = (a_{ij})$，候補となる向き付けに対応する隣接行列を $\mathbf{A}' = (a'_{ij})$ とするとき，候補となる向き付けは次の制約を満たさなければならない．

$$\text{if } a'_{ij} = 1 \text{ and } a'_{kj} = 1, \text{ then } a_{ik} = 1$$

よって向き付けの探索においては，因果合流制約を満たす因果関係ネットワークのみを考慮すればよい．

4.3 最適な向き付け探索手法

最適な向き付け決定する際に，向き付けの全組み合せに対する評価は計算量の観点から困難である．そこで，ある初期向き付けを与え，そこから評価関数値を最小にするような向き付けの探索を行うこととする．

探索には，探索の深さを表す近傍の定義，探索の初期値設定，探索方向を決定するルール (ネットワークの向き付け変更方法) の 3 つが必要となる．

近傍の定義のために，ネットワーク間の距離を定義する．本研究では，ネットワーク間の距離を，あるネットワークと他のネットワークの，互いに状態が異なる辺の数をネットワーク間の距離とした．すなわち，二つのネットワークに対応する隣接行列 $\mathbf{A}$ と $\mathbf{A}'$ が与えられたとき，それらの距離 $d(\mathbf{A},\mathbf{A}')$ を，

$$d(\mathbf{A},\mathbf{A}') = \sum_i \sum_j \mathbf{A}_{ij} \oplus \mathbf{A}'_{ij} \tag{8}$$

と定義した．このとき，$\mathbf{A}_{ij}$ を隣接行列 $\mathbf{A}$ の ij 要素，$\oplus$ を排他的論理和とする．ここでは，ネットワーク $\mathbf{A}$ の近傍 $\epsilon(\mathbf{A})$ を，

$$\epsilon(\mathbf{A}) = \{\mathbf{A}' : d(\mathbf{A},\mathbf{A}') \leq f(|E|)\} \tag{9}$$

と定義した．ここで，$f(|E|)$ は推定された無向グラフの辺数 $|E|$ によって決定される定数である．

初期向き付けは次のように設定する．

初期向き付け設定ステップ

1. 無向グラフ中から次数が最も高い頂点を選ぶ
2. 選ばれた頂点から深さ優先探索を行い，到達した頂点から順番に番号をつける
3. 番号の小さい頂点から大きい頂点へ方向をつける

ネットワークの向き付け探索は，ネットワークの辺の向きを反転することで実現できる．ネットワークの向き付け変更ステップを以下に示す．

向き付け変更ステップ

1. 変更前のネットワーク内の特徴量が最も大きい頂点へ着目ラベルをつける．
2. 着目ラベルのついた頂点を終端とする辺について以下を繰り返す．
3. 辺の向きを反転させる．
4. 着目ラベルがついている頂点に軌跡ラベルをつけて，反転させた辺の終端の頂点に着目ラベルをつける．
5. 得られたネットワークが因果合流制約を満たしているか調べ，満たしているなら 6. へ進む．満たしていないなら，制約を満たしていない部分の着目ラベルがついている頂点を終端とする辺で，軌跡ラベルがついた頂点以外からの辺の向きを反転させて 4. へ戻る．
6. 得られたネットワークと変更前のネットワークの距離を計算し，結果が近傍に含まれないなら，3. へ戻り別の辺を反転させる．
7. 選ばれた頂点の辺反転で得られるネットワークのすべてが近傍にないなら，2. に戻り二番目に大きな特徴量を持つ頂点を選ぶ．

以上をまとめて，探索ステップを以下に示す．

探索ステップ
入力: 相関関係を表現した無向グラフ
出力: データに最も適合した因果関係ネットワーク

1. 初期ネットワークを設定する
2. 候補ネットワーク中の各要素の特徴量を計算する．
3. 推定された各要素の特徴量と計算された特徴量を比較し，特徴量の順位が一致していない頂点のうち，最大の特徴量をもつ頂点を選ぶ．
4. 与えられた向き付けが因果合流制約を満たすか調べる
5. もし満たしているなら，評価関数の値を計算する
6. 向き付けを変更し，ある近傍内の評価関数の値を計算する
7. 評価関数の値が最も改善された向き付けから 3. を行う．改善されないなら終了する

この探索手法においては，因果合流制約を満たしているかどうかの判定回数が多くなるが，特徴量を計算する回数が少なくてすむ．遺伝子ネットワークは一般的に大規模であるため，固有値問題を解く必要がある特徴量の計算回数を抑えることで，全体の計算速度の向上につながると考えられる．

探索を進めていく上で必要となる，与えられたネットワークが因果合流制約を満たすかどうか判定するステップを示す．

因果合流制約判定ステップ

1. 向き付けを与える前の無向グラフに対応する隣接行列を $\mathbf{A}$，現在候補となっている有向グラフに対応する隣接行列を $\mathbf{A}'$ とする．
2. $\mathbf{A}'$ の 1 つの行 $\mathbf{A}'_i$ を選ぶ．
3. $\mathbf{A}'_i$ のうち，値が 1 である添字を，配列 $list$ に格納する．
4. $\sum_{k,w \in list} \mathbf{A}_{kw}$ を計算する．
5. 計算結果が $|list|^2$ より小さいならば因果合流制約をみたしていない．それ以外なら満たしている．

以上の評価関数と探索手法により，観測された遺伝子発現データに最も適合した向き付けを得ることができる．このようにして遺伝子間の直接的な因果関係を表現したネットワークの推定が可能となる．

5 まとめと課題

遺伝子発現データから直接的な相関を表す無向グラフを推定し，その各辺に対して因果方向を決定することによって，遺伝子間因果関係ネットワークを推定する手法を提案した．提案手法によって，従来のベイジアンネットワークでは推定不可能なサイクルを含む因果ネットワークの推定も行うことができる．

今後の課題として，シュミレーションによる最適な近傍と評価関数の決定と，実データを用いた検証実験があげられる．

参考文献

[1] Stanford Microarray Database, http://brownlab.stanford.edu/

[2] T. Akutsu, S. Kuhara, O. Mruyama, S.Miyano, "Identification of gene regulatory networks by strategic gene disruptions and gene over-expressions", Proc 9th ACM-SIAM SODA, pp. 695-702, 1998

[3] P.D'Haeseleer, X. Wen, Fuhrnman, R. Somogyi, "Linear modeling of mRNA expression levels during CNS development and injury", Proc 4th Paci c Symposium on Biocomputing(PSB), vol.4, pp.41-52, 1999

[4] D.C. Weaver, C.T. Workman, G.D. Stormo, "Modelling Regulatory Networks with Weight Matrices", Proc 4th Pacific Symposium on Biocomputing(PSB), vol.4, pp. 112-123, 1999

[5] 木村陽一, "IT&バイオ入門", オーム社, 2002

[6] Takeo OKAZAKI, Madoka KAMIYA, Hitoshi AFUSO, Morikazu NAKAMURA, "Normalization of DNA Microarray data with BIC Model Comparison", ISPJ SIG Technical Reports, 2006, pp.

[7] 福水健次, "正定値カーネルを用いたデータ解析", 公開講座「機械学習の最近の話題」, 2004, http://www.ism.ac.jp/ fukumizu/ISM_lecture_2004/Lecture2004_kernel_method.pdf

[8] 宮川雅巳, "グラフィカルモデリング", 統計ライブラリー, 朝倉書店, 1997

[9] Lawrence Page, Sergey Brin, "the PageRank Citation Ranking: Bringing Order to the WEB", Stanford University, 1998

[10] Hitoshi AFUSO, Takeo OKAZAKI, "An Influence Character of Network Objects using Link Structure", 20th ITC-CSCC, 2005, pp. 37-38