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

2003－MPS－46　(19)
2003／9／19

生物学的知見を利用したModule Bayesian Networkによる遺伝子制御ネットワークの推定

瀧　浩平[†]　竹中要一[†]　松田秀雄[†]

遺伝子間の制御関係は遺伝子の制御ネットワークとしてグラフによって表現する事ができる．性質の似た変数の集合をモジュールとして扱う Module Bayesian Network を用いて，遺伝子の制御ネットワークを推定する研究が行われているが，遺伝子の数に対して測定結果の数が不足しているため正確な推定を行えていない．この様な場合，推定に事前知識を採り入れる事が有効とされるが，Module Bayesian Network を用いた過去の研究では，モジュールの推定に十分な事前知識を採り入れていない．そこで本研究では，Gene Ontology を用いた遺伝子に対する注釈情報を利用して，注釈の似た遺伝子を同一のモジュールに集めるモジュール評価関数を提案する．本手法を実際の遺伝子間の制御関係の推定に適用し，性能評価を行った．

Adapting the biological prior knowledge to the gene regulation network inference using Module Bayesian Network

Kohei Taki,[†] Yoichi Takenaka[†] and Hideo Matsuda[†]

The gene regulation networks are directed graph expression of gene regulations. Because the amount of expression profile's samples is rarely enough to robustly learn dependencies of a large number of genes, any methods have not succeeded in a precise inference of gene regulation network. In this research a gene regulation network is infered by modeling to Module Bayesian Networks, which are introduced the notion of module in which variables set has the same dependency, and a module scoring method is proposed so that infered module's genes have the similar annotation of prior knowledge. A performance of proposed method is evaluated on the actual gene expression profile.

1. はじめに

生体の機能を実現している遺伝子のほとんどは，その働きが他の遺伝子により制御されており，遺伝子間の制御関係は遺伝子の制御ネットワークとして，頂点を遺伝子，辺を制御関係とするグラフによって表現することができる．遺伝子の制御ネットワークの構造を解明する事は，生命現象を理解するのに重要な役割を果たすが，膨大な数の遺伝子が存在するため，遺伝子間の制御関係を実験のみによって解明するのには限界があり，大部分の制御関係は不明なままである．

このため，計算機を用いて遺伝子の発現プロファイルの情報を解析し，遺伝子の制御ネットワークを推定する研究が精力的に行われている．遺伝子の発現プロファイルとは，研究対象の遺伝子群の発現量を様々な条件下で測定した結果である．制御される側の遺伝子の発現量は，制御する側の遺伝子の状態を入力とする関数によって決定する事が出来ると考えられており，遺伝子の発現プロファイルと遺伝子の制御ネットワークの関係は，真理値表と Boolean Network の関係に対応させて考える事が可能である．

この様な特性を利用し，各遺伝子の発現の状態を Boolean Network の変数として，発現プロファイルから Boolean Network を推定する研究がされてきた．その発展手法として，Boolean Network に確率理論を採り入れ，共通の変数に依存する変数の集合であるモジュールの概念を採り入れた Module Bayesian Network(以下 Module BN) による関数関係の推定手法が提案されており，遺伝子の制御ネットワークの推定にも適用されている．

しかし，Module BN による遺伝子の制御ネットワークの推定には，推定結果の信頼性がまだ十分ではないという問題がある．推定結果の信頼性を低下させる要因の一つは，対象とする遺伝子の数が膨大なのに対して，発現プロファイルが正確な推定に十分なサンプル数を与えていない事にある．一般の Bayesian Network の推定においてはこの様な場合，事前知識を用いて情報の不足を補う事が有効だが，Module BN による遺

† 大阪大学 大学院情報科学研究科 バイオ情報工学専攻
Department of Bioinformatic Engineering, Graduate School of Engineering Science, Osaka University

伝子の制御ネットワーク推定では, モジュールの推定に対して事前知識を十分に採り入れていない. そこで本研究では, 遺伝子に対する注釈情報を事前知識として利用して, 遺伝子の制御ネットワークの推定を行う.

2. 遺伝子の発現プロファイルと Module Bayesian Network

2.1 遺伝子の発現プロファイル

生体の機能を実現しているタンパク質は, その構成の情報が DNA 塩基配列によって遺伝子にコードされている. 生体内で特定のタンパク質が必要になると, その遺伝子の DNA 塩基配列が mRNA に転写されてタンパク質が合成される. 遺伝子の DNA 塩基配列が mRNA に転写された量を遺伝子の発現量と呼び, その測定にはマイクロアレイや GeneChip が用いられる.

遺伝子の発現プロファイルは, 研究対象とする遺伝子それぞれについて, 様々な条件下で測定した発現量の系列であり, (遺伝子×条件) の行列で表現される. ここでは, ある条件下における各遺伝子の発現量のベクトルをサンプルと呼ぶ.

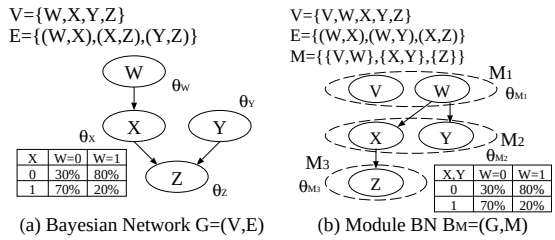
遺伝子の発現量は, 特定の遺伝子が発現する事で生成されるタンパク質の量に依存して増減する事が知られており, この現象は遺伝子の転写制御と呼ばれる. 遺伝子の発現プロファイルの各サンプルにおいて, 遺伝子の発現量の間に依存関係がある事を検出できれば, 遺伝子の間の制御関係があると推定できる. 本研究では, 発現量の間の依存関係の検出に, Module BN による推定を利用する.

2.2 Bayesian Network と Module BN

Bayesian Network は Directed Acyclic Graph(以下 DAG) と Conditional Probability Table(以下 CPT) によって構成され, 変数の間の依存関係を表現するのに利用される (図 1(a)). 変数を DAG の頂点と 1 対 1 で対応付け, 変数間に依存関係がある事を DAG の対応する頂点間に有向辺を引いて表現する. 変数 X_i の頂点から変数 X_j の頂点へ有向辺を引く事は, 変数 X_j の取る値が変数 X_i の取る値に依存する事を表している. 依存関係のある変数同士がどの様な関数関係で依存するかは, 各変数 X_i に対応する CPT θ_{X_i} で確率を用いて表現する.

図 1(a) は Bayesian Network の例であり, 変数 W,X,Y,Z の間の依存関係を表現している. X は W に, Z は X,Y にそれぞれ依存している事を, DAG の接続関係で表現している. 変数 X の左の表は変数 X の CPT θ_X を表しており, 変数 W が 0 の場合には 70%の確率で変数 X が 1 になり, 変数 W が 1 の場合には 80%の確率で 0 になる事を表現している.

Module BN は, 変数の集合であるモジュールの概念を Bayesian Network に採り入れた手法である (図 1(b)). Module BN はモジュールの集合 $\mathbf{M}$ と



(a) Bayesian Network G=(V,E) (b) Module BN BM=(G,M)

図 1 Bayesian Network と Module BN の例
Fig. 1 example of Bayesian Network and Module BN

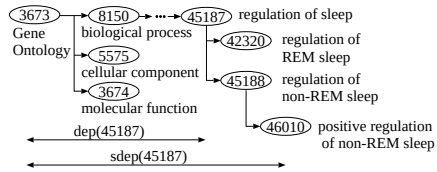


図 2 Gene Ontology の階層の例
Fig. 2 example of hierarchy of Gene Ontology

Bayesian Network から成り, 各変数はモジュール集合の内のどれか 1 つのモジュール $\mathbf{M}_k$ に含まれる (今後, 集合を表す変数にはドットを付けるものとする). モジュールは共通の変数に依存する変数の集合であり, Bayesian Network の DAG の頂点の接続関係は, 同一のモジュールに属する頂点が共通の親頂点を持つ制約を受ける.

Module BN の推定は, Module BN の評価関数を極大化するため, Greedy Hill Climbing アルゴリズムにより, グラフの接続関係とモジュールの結合解空間を探索する事によって行われる. Module BN の評価関数は, Segal らが与えている定式化 [1] を Bayesian Network の評価関数 BayesFactor[2] に適用したものを利用する.

3. Gene Ontology[3]

GO は遺伝子の特徴を注釈する語彙の体系である. GO の各語彙は GO term と呼ばれ, それぞれに GO ID が割り振られている. ここでは, GO ID i を割り当てられた GO term を T_i で表し, 全 GO term の集合を $\mathbf{T} = \{T_{i_1}, \cdots, T_{i_{|\mathbf{T}|}}\}$ で定義する. 各 GO term は注釈の具体性の程度が異なり, 例えば, 図 2 の GO term "*regulation of sleep*" はより具体的な GO term "*regulation of REM sleep*" の上位概念である. GO term 間の上位・下位概念の関係によって, GO term の階層が構成されており, GO term の階層は図 2 の様に, GO term "*Gene Ontology*" を根とする DAG を用いて表現される.

GO term T_j がより具体的な GO term T_i の上位概念である事を, 半順序関係 $T_i \prec_{R_{GO}} T_j$ で表すことにする. GO term T_i の上位概念である GO term の集

合 $\{T_j|T_i \prec_{R_{GO}} T_j\}$ の内で, 最も下位の概念の GO term が T_j の場合, T_i, T_j の関係を $T_i\ R_{GO_{spc}}\ T_j$ で表す. このとき, GO term の階層を表現する DAG G_{GO} を以下の様に定義する.

$$\mathbf{V}_{GO} = \{i|T_i \in \mathbf{T}\}$$
$$\mathbf{E}_{GO} = \{(j,i)|T_i\ R_{GO_{spc}}\ T_j\}$$
$$G_{GO} = (\mathbf{V}_{GO}, \mathbf{E}_{GO})$$

4. Gene Ontology を用いたモジュール評価アルゴリズム

Module BN を推定する際に構成したモジュール集合は, 制御関係の推定結果に大きな影響を及ぼす. 従って, 推定されるモジュール集合がこれまでの知見と一致する事は, 制御関係の推定結果の精度の向上に有利に働くと考えられる. 現在, 多数の遺伝子が GO term によって注釈されており, モジュール集合がこれまでの知見と一致するかの判断基準として, GO を利用する事が出来る.

4.1 モジュール評価の指針

GO は GO term の階層構造を持つため, 各 GO term は互いに独立ではない. 従って, GO term が完全に一致しない場合も, 高く評価する必要がある場合がある. 今回 GO term T_i, T_j の一致については, 次の 3 つの場合のみ高い評価を与える.

(1) $T_i = T_j$ の場合
(2) $T_i \prec_{R_{GO}} T_j$ の関係にある場合
(3) $(T_i \not\prec_{R_{GO}} T_j) \cap (T_j \not\prec_{R_{GO}} T_i)$ の場合に, 共通の祖先 $T_k \succ_{R_{GO}} T_i, T_j$ が存在する場合

(2),(3) を高く評価する根拠について述べる. (2) は,T_i, T_j の間に概念的な包含関係があり, 互いに関連が強い. (3) も同様で, T_i, T_j のどちらも同じ T_k の下位概念だと考えられ, 互いに関連が強い. 例えば, 図 2 の GO term "*regulation of REM sleep*"(T_{42322}), "*regulation of non-REM sleep*"(T_{45188}) は異なる GO ID を持ち, 祖先・子孫の関係にもないが, $T_{42332}, T_{45188} \prec_{R_{GO}} T_{45187}$ であり, どちらも睡眠の制御についての GO term なので関連が強い.

上記の 3 つの場合を同様に評価する事は現実的ではないので, GO term 間の一致の強度を決定するために GO term 間の距離を定義する.

4.2 GO term 間の距離

前節 (1) の距離を 0 と定義し, これを基準として (2),(3) の距離の定義を以下で行う. (1),(2),(3) 以外の場合の距離は ∞ と定義する.

$$dist(T_i, T_i) = 0 \quad : (1) \text{ の場合}$$
$$dist(T_i, T_j) = \infty \quad : (1), (2), (3) \text{ 以外の場合}$$

(2) の場合, 2 つの GO term の間で異なるのは注釈の具体性の程度である. 従って, GO の階層 G_{GO} において頂点 i,j の間に存在する頂点の数が少ない程, 2 つの GO term の間の具体性の程度に差は無い. (3) の場合, 2 つの GO term で共通なのは GO term T_k に包含されている事である. GO term T_k と T_i, T_j の間に具体性の差が無い程, T_i, T_j の間の距離は近く, G_{GO} においては頂点 i,j と頂点 k の間に存在する頂点の数が少ない事に対応する. 以上より, GO term T_i, T_j の間の距離は, DAG G_{GO} 上の対応する頂点 i,j の間の距離を用いるのが適当であると考えられる. グラフ G における頂点 i,j の間の有向パスの長さを $d_G(i,j)$ で表し, 以下の様に GO term 間の距離を定義する.

$$dist(T_i, T_j) = min(d_{G_{GO}}(i,j), d_{G_{GO}}(j,i)) \quad : (2) \text{ の場合}$$

しかし, (3) の場合, 2 つの GO term が G_{GO} 上で位置する深さによって, 2 つの GO term の関連の強さが異なると考えられる. 例えば, 図 2 の T_{42320}, T_{45188} と T_{5575}, T_{8150} の間の距離はどちらも 2 だが, 後者の GO term の間の関連性は弱い. 後者の場合,G_{GO} における距離は近いが, 共通の祖先が抽象的過ぎるため, 祖先が共通である事に重要な意味がないためだと考えられる. 従って, (3) は G_{GO} における頂点 i,j の深さを考慮して, T_i, T_j の距離を以下の様に定義し直す.

$$dist(T_i, T_j) = d_{G_{GO}}(i,j) * max\left(\begin{array}{l} sdep_{G_{GO}}(i)/dep_{G_{GO}}(i), \\ sdep_{G_{GO}}(j)/dep_{G_{GO}}(j) \end{array} \right) \quad : (3) \text{ の場合}$$

但し, $dep_G(i)$ はグラフ G における頂点 i の深さを表し, $sdep_G(i)$ はグラフ G において頂点 i から到達可能な最も深い頂点の深さであるとする (図 2 参照).

上と同様の理由で, G_{GO} における共通の祖先の深さが 2 以下の場合には, 距離を ∞ として定義する.

4.3 モジュール評価関数

ここでは本研究で用いるモジュール評価関数の定義を行う. まず, GO term による遺伝子の注釈の標記について諸定義を行う.

GO term による遺伝子 g_i に対する注釈の集合を $\mathbf{A}_{g_i} = \{A_{(g_i,1)}, \cdots, A_{(g_i,|\mathbf{A}_{g_i}|)}|A_{(g_i,j)} \in \mathbf{T}\}$ で表し, モジュール $\mathbf{M}_k$ に属する遺伝子の注釈の集合を $\mathbf{A}_{\mathbf{M}_k} = \{A_{(g_i,j)} \in \mathbf{A}_{g_i}|X_i \in \mathbf{M}_k\}(\subseteq \mathbf{T})$ で表すとする. 今後, $\mathbf{A}_{\mathbf{M}_k}$ の i 番目の要素を $A_{(\mathbf{M}_k,i)}$ で参照するものとする.

モジュール $\mathbf{M}_k$ に属する遺伝子に対する任意の 2 つの注釈の組 $(A_{(\mathbf{M}_k,i)}, A_{(\mathbf{M}_k,j)})$ を考える. GO term 間の距離 $dist(A_{(\mathbf{M}_k,i)}, A_{(\mathbf{M}_k,j)})$ が小さい組の数が多いモジュールは, 注釈のまとまったモジュールであるため高い評価を与える. モジュール $\mathbf{M}_k$ において, $(A_{(\mathbf{M}_k,i)}, A_{(\mathbf{M}_k,j)})$ が一致する組の数を, モジュール $\mathbf{M}_k$ に属する遺伝子対する注釈の一致数 (以下, 単に一致数) として定義し, $A_{(\mathbf{M}_k,i)} = A_{(\mathbf{M}_k,j)}$ となる組を一致数に 1 組分として数える. GO の階層を考慮して一致数を数えるために, $A_{(\mathbf{M}_k,i)} \neq A_{(\mathbf{M}_k,j)}$ だが $dist(A_{(\mathbf{M}_k,i)}, A_{(\mathbf{M}_k,j)})$ が小さい組も, モジュール $\mathbf{M}_k$ の一致数に $f(dist(A_{(\mathbf{M}_k,i)}, A_{(\mathbf{M}_k,j)}))$ 組分として数え

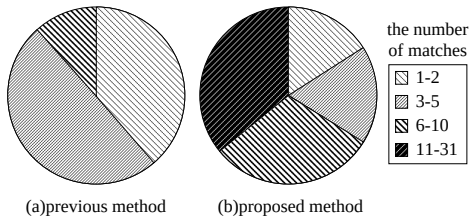


図 3 同一モジュール内で GO term が一致する回数の割合
Fig. 3 fraction of the number of the cases of matching GO terms belonging to the same module

る. 関数 f(d) は $R \rightarrow [0,1]$ の単調減少の関数であるとし, 今回は $f(d) = 1/(1+d)$ として用いる.

ここで定義した一致数は数えているだけなので, モジュールに遺伝子を加える場合に, 遺伝子を注釈する GO term とモジュールの間の関連の強さに関係なく一致数は増加する. モジュールに含まれる遺伝子を注釈する GO term の種類の数と, モジュールの評価値は, 反比例の関係にある事が望ましいと考えられるので, 一致数をそのまま評価値とするのは相応しくない. そこで, 以下の様にモジュール評価関数 $moduleScore(\mathbf{M}_k)$ を定義した.

$$moduleScore(\mathbf{M}_k) = C \cdot \frac{\sum_{A_i \in \mathbf{A}_{\mathbf{M}_k}} \sum_{A_j \in \mathbf{A}_{\mathbf{M}_k}} f(dist(A_i, A_j))}{\left| \bigcup_{i|X_i \in \mathbf{M}_k} \mathbf{A}_{g_i} \right|}$$

5. 実　　験

提案手法を用いた場合と, 提案手法を適用しない Module BN(以下, 従来手法) を用いた場合それぞれで, 遺伝子の制御ネットワークを推定し, その推定結果を比較した. 遺伝子の発現プロファイルは, Gasch らが行った出芽酵母の環境ストレスに対する反応実験から得られた [4], サンプル数 173 個のものを用いた. 発現量の変化が最大で 2 以上の遺伝子 2473 個を制御関係の推定対象とし, Module BN で推定するモジュールの数は 50 個に固定した [1].

まず, 提案したモジュール評価関数を用いる事で, 注釈の関連の強い GO term を多数含むモジュール集合を推定できる事を確認する.

図 3 は, 従来手法と提案手法それぞれについて, 各 GO term が同一モジュール内の他の GO term と完全に一致する回数の分布を調べ, 円グラフにまとめた結果である. 図 3 によれば, 従来手法では各 GO term に対して, 同一モジュール内に同一の GO term が 5 個以下しか存在しない割合が 90%近いが, 提案手法ではその割合は 30%程度に留まっており, 同一モジュール内に同一の GO term が多数存在する割合が高くなっている事が確認できる.

GO term が最も抽象的な概念 (biological process などの 1 つ下の概念) でも一致しない場合, GO term の間に関連は無いと考えられる. 同一モジュールに属する GO term の間に関連が無い割合の平均は, 従来手法が 48.2%なのに対して提案手法では 37.7%であり, 提案手法の方が有為に割合が低い (表 1). 従って, 提案手法では同一モジュールに, 関連がある程度強い GO term が集まっていると言える.

表 1 GO term 間に関連の無い割合
Table 1 fraction of the cases of no relations between GO terms

	従来手法	提案手法
同一モジュールの GO term の間に関連の無い割合	48.2%	**37.7%**
モジュールの親集合の GO term の間に関連の無い割合	50.5%	**47.4%**

以上より, 提案したモジュール評価関数を用いる事で, 関連の強い GO term を多数含むモジュール集合を推定できる事が確認できる.

次に, 推定した Module BN の各モジュールを制御する, 遺伝子の集合を注釈する GO term について関連の強さを調べた. 上述と同様に, 制御遺伝子の GO term の間に関連が無い割合の平均は従来手法で 50.5%, 提案手法で 47.4%となり, 提案手法によって従来手法が改善されている事が確認できる (表 1).

6. おわりに

Module BN による遺伝子制御ネットワークの推定において, モジュールの推定に GO を利用する手法を提案し, 互いに関連の強い遺伝子を含むモジュールから構成される Module BN を, 推定する事が出来る事を確認した. 本手法により, 遺伝子の発現データと遺伝子の注釈情報という異種の情報を組み合わせた推定が可能になる. 今後の方向性として, 他の生物学的情報も考慮に入れた推定を行う事が考えられる.

参 考 文 献

1) Segal,E. et al. “Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data”, *Nature Genetics*, Vol. 34, pp.. 166–176 (2003).
2) Heckerman,D. “A tutorial on learning with Bayesian networks”, Technical Report, MSR-TR-95-06. (1995).
3) The Gene Ontology Consortium “Creating the Gene Ontology Resource: Design and Implementation”, *Genome Research*, Vol. 11, pp.. 1425–1433 (2001)
4) Gasch,A.P. et al. “Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes”, *Molecular Biology of the Cell*, Vol. 11, pp.. 4241–4257 (2000).