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

2008－MPS－72 (30)
2008－BIO－15 (30)
2008／12／18

多重解像度独立性検定を用いた遺伝子ネットワークの構築

山本 隆之[†], 滝口 哲也[††], 有木 康雄[††]

[†] 神戸大学大学院工学研究科 [††] 神戸大学自然科学系先端融合研究環

現在，ベイジアンネットワークを用いた遺伝子ネットワークの構築において，確率分布を仮定する score-based approach が一般的である．しかしこの方法では確率分布を仮定することから，すべての依存関係を検出できるとは限らない．本研究では Margaritis らが提案した多重解像度独立性検定及び条件付き独立性検定を用いて確率分布を仮定せずにベイジアンネットワークを構築する方法を導入し，実際の遺伝子発現データを用いて遺伝子ネットワークの構築を行った．

Structuring Gene Network Using Multiresolution Independence Test

Takayuki Yamamoto[†], Tetsuya Takiguchi[††], Yasuo Ariki[††]

[†] Graduate School of Engineering, Kobe University
[††] Organization of Advanced Science and Technology, Kobe University

In order to struct a gene network, a score-based approach is often used, where a probability distribution is assumed. But the assumption of probability distribution prevents from finding various kinds of dependence relationships with genes. In this research, we structed a gene network from observed gene expression data using multiresolution independence test and conditional independence test, which is the method proposed by Margaritis for learning the structure of Bayesian networks without making any assumptions on the probability distribution.

1 はじめに

近年，DNA 塩基配列の解析技術の発達により，さまざまな生物種の遺伝子の配列情報が解明されているが，その機能やネットワーク構造については明らかになっていないものも多く，解明する技術の発達が求められている．遺伝子の配列情報が mRNA に転写され，タンパク質に翻訳される一連の過程は遺伝子発現と呼ばれ，それぞれの遺伝子の発現の間には依存関係が存在し，互いに複雑に影響し合ってタンパク質を生成していることがわかっている．このような遺伝子相互間の発現の依存関係をグラフィカルモデルで表したものが遺伝子ネットワークである．遺伝子ネットワークにおいては遺伝子を頂点，遺伝子の依存関係を有向辺として表す．この遺伝子ネットワークの推定問題は現在のバイオインフォティクスにおける重要課題の1 つとなっている．

DNA マイクロアレイは，発現の過程で生成される mRNA の量を観測することで遺伝子発現量を測定する器具であり，遺伝子ネットワークの推定問題は，DNA マイクロアレイによって得られる遺伝子発現量のデータを確率変数ととらえたグラフィカルモデルの推定問題として定式化される．また，DNA マイクロアレイの性質として数千種もの遺伝子の発現量のデータを一度に得ることができるという利点があるが，得られるサンプル数が少なく，ノイズや欠値を多く含むということがあげられる．すなわち，この遺伝子ネットワークの推定問題は，信頼度が低く少ないサンプル数のデータから，過適合を避けて多くの頂点集合を含むグラフの構造を求めるという非常に困難な問題であるといえる．

これを解決する方法として，ベイジアンネットワーク [3] (Friedman et al.,2000)，ブーリアンネットワーク [4] (Akutsu et al.,2000)，微分方程式系 [5] (Chen et al.,1999)，グラフィカルガウシアンモデリング [6] (Toh and Horimoto,2002) といった様々なモデルを用いて遺伝子ネットワークを推定する方法が提案されているが，本研究では誤差に強く，少ないデータ数でも構造推定可能なベイジアンネットワークによる遺伝子ネットワークの推定法に着目した．従来，ベイジアンネットワークによる遺伝子ネットワークの推定法としては，遺伝子ネットワークをグラフ構造の事後確率最大化によって推定する方法が一般的であるが，この方法では遺伝

子間の依存関係を表す確率分布を仮定しなければならず，仮定した確率分布で表される関係しか検出できないという問題がある．これに対し，本研究では確率分布を仮定せずに様々な依存関係を含む遺伝子ネットワークを推定する方法として，各遺伝子発現量の間の独立性・条件付き独立性を多重解像度で検定し，ネットワークを構築する方法を導入した．

2 遺伝子ネットワークの構築

ベイジアンネットワークをデータから構築する方法は，Score-Based Approach と Independence-Based Approach の2 種類に大別される．この章ではベイジアンネットワークの構築の2 つのアプローチについて紹介し，その特性について述べる．

2.1 Score-Based Approach

遺伝的アルゴリズムや hill-climbing などの探索アルゴリズムを用いて score を上昇させるようにネットワークを変形していき，最も高い score をとるネットワーク構造を返す方法を Score-Based Approach という．この方法においてはネットワーク構造の探索手法，確率変数間の関係を表す確率分布，ネットワーク全体を評価する score を設計しなければならない．この手法の問題点として変数の数が増えると計算時間が指数関数的に増大することや変数間の関係を確率分布で仮定しなければならないこと，変数が多いほど探索が局所最適解に落ちやすいことなどが挙げられる．

2.2 Independence-Based Approach

全ての確率変数間の独立性・条件付き独立性を判定し，独立な変数間の有向辺を削除してネットワーク構造を求める方法を Independence-Based Approach という．この方法においては独立性の判定基準を設定する必要がある．本研究では Score-Based Approach における確率分布の設定によって，ネットワークの構築が多大な制約を受けていると考え，確率分布を仮定しない Independence-Based Approach による手法を提案しているが，問題点としてデータ依存性が高く，低信頼性の少数データからは信頼できるネットワーク構造が得られないということが挙げられる．

3 独立性検定

Independence-Based Approach においては独立性を判定する基準を設定しなければならない．本研究においては Margaritis らが提案した多重解像度の独立性検定[1] 及びそれに基づく条件付き独立性検定[2] を導入し，独立性の判定基準とした．この章ではその2 つの検定法について述べる．

3.1 多重解像度独立性検定

はじめに，解像度と境界を固定した場合の独立性検定を考える．まず，解像度を $R \equiv I \times J$ とし，2 つの確率変数のサンプルの存在区間を $I \times J$ の領域に分割する．各領域に含まれるサンプルの数をそれぞれ $c_1,\cdots,c_K$，$K \equiv IJ$，サンプルの総数を N，各領域でサンプルが発生する確率を $p_1,\cdots,p_K$，各領域の境界の集合を $\mathbf{B}_R$ とするとデータ $\mathbf{D}$ の尤度は次式で表される．

$$Pr(\mathbf{D}|p_{1\cdots K},\mathbf{B}_R,R) = N!\prod_{k=1}^{K}\frac{p_k^{c_k}}{c_k!} \tag{1}$$

この式においてパラメータ p_k は未知であるため，次式のように多項分布の共役事前分布であるディリクレ分布を事前分布として与える．

$$Pr(p_{1\cdots K}) = \Gamma(\gamma)\prod_{k=1}^{K}\frac{p_k^{\gamma_k-1}}{\Gamma(\gamma_k)} \tag{2}$$

式 (2) において $\gamma = \Sigma_{k=1}^{K}\gamma_k$ であり，$\gamma_{1\cdots k}$ はディリクレ分布の設定パラメータを表す．また，データ尤度 $Pr(\mathbf{D})$ をパラメータ p_k で周辺化し，式 (1)，(2) を代入すると，固定解像度・固定境界におけるデータ尤度 $Pr(\mathbf{D})$ が得られる．

$$\begin{aligned} Pr(\mathbf{D}) &= \int Pr(\mathbf{D}|p_{1\cdots K})Pr(p_{1\cdots K})dp_{1\cdots K} \\ &= \frac{\Gamma(\gamma)}{\Gamma(\gamma+N)}\prod_{k=1}^{K}\frac{\Gamma(\gamma_k+c_k)}{\Gamma(\gamma_k)} \end{aligned} \tag{3}$$

データ $\mathbf{D}$ が独立モデル M_I，依存モデル $M_{\neg I}$ のどちらか一方から生成されると仮定すると，これについて周辺化した式は次式のようになる．

$$\begin{aligned} Pr(\mathbf{D}) &= Pr(\mathbf{D}|M_I)Pr(M_I) \\ &+ Pr(\mathbf{D}|M_{\neg I})Pr(M_{\neg I}) \end{aligned} \tag{4}$$

式 (4) をベイズの定理を用いて変形し，両モデルの事前確率をそれぞれ $Pr(M_I) \equiv \rho$，$Pr(M_{\neg I}) \equiv 1-\rho$ とすると次式が得られる．

$$Pr(M_I|\mathbf{D}) = 1/\left[1+\frac{(1-\rho)}{\rho}\frac{Pr(\mathbf{D}|M_{\neg I})}{Pr(\mathbf{D}|M_I)}\right] \tag{5}$$

依存モデルにおいては，1 つの多項分布を仮定すれば良いが，独立モデルにおいては，2 つの確率変数が独立であるため両軸方向に多項分布を仮定しなければならない．よってデータ尤度 $Pr(\mathbf{D}|M_{\neg I})$，$Pr(\mathbf{D}|M_I)$ は式 (3) を用いて次のように表される．

$$\begin{aligned} Pr(\mathbf{D}|M_{\neg I}) &= \frac{\Gamma(\gamma)}{\Gamma(\gamma+N)}\prod_{k=1}^{K}\frac{\Gamma(\gamma_k+c_k)}{\Gamma(\gamma_k)} \\ &\equiv \Upsilon(\mathbf{C}_K,\gamma_K) \end{aligned} \tag{6}$$

$$
\begin{aligned}
Pr(\mathbf{D}|M_I) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha+N)}\prod_{i=1}^{I}\frac{\Gamma(\alpha_i+c_{i+})}{\Gamma(\alpha_i)} \\
&\times \frac{\Gamma(\beta)}{\Gamma(\beta+N)}\prod_{j=1}^{J}\frac{\Gamma(\beta_j+c_{+j})}{\Gamma(\beta_j)} \\
&\equiv \boldsymbol{\Upsilon}(\mathbf{C}_{I+},\alpha_{I+})\boldsymbol{\Upsilon}(\mathbf{C}_{+J},\beta_{+J})
\end{aligned} \tag{7}
$$

式 (6), (7) において $\gamma = \Sigma_{k=1}^{K}\gamma_k$, $\alpha = \Sigma_{i=1}^{I}\alpha_i$, $\beta = \Sigma_{j=1}^{J}\beta_j$ とし，事前分布を一様分布とするため，$\alpha_i = \beta_j = \gamma_k = 1$ と設定する．式 (6), (7) を式 (5) に代入し，次式を得る．

$$
Pr(M_I|\mathbf{D}) = \frac{1}{\left[1+\frac{(1-\rho)\boldsymbol{\Upsilon}(\mathbf{C}_K,\gamma_K)}{\rho\boldsymbol{\Upsilon}(\mathbf{C}_{I+},\alpha_{I+})\boldsymbol{\Upsilon}(\mathbf{C}_{+J},\beta_{+J})}\right]} \tag{8}
$$

式 (8) により固定解像度・固定境界でのモデル尤度を計算することができる．これを多重解像度に拡張することを考える．

$$
\begin{aligned}
Pr(M_I|R_{max},\mathbf{D}) = \int & Pr(M_I|\mathbf{B}_R,R_{max},\mathbf{D}) \\
& Pr(\mathbf{B}_{R_{max}}|R_{max},\mathbf{D})d\mathbf{B}_{R_{max}}
\end{aligned} \tag{9}
$$

式 (9) において，R_{max} は依存性を最大とする解像度であり，解像度を上げて繰り返し独立性を計算し，これが最小となる解像度を R_{max} とする．また B_R は，解像度 R における境界集合であり，解像度 R において取り得る全ての境界の組み合わせに関して $Pr(M_I|\mathbf{D})$ を計算し，これの重み付き平均を解像度 R での独立性とする．また境界には各データ間の中点を通る直線を用い，$Pr(\mathbf{B}_R|R,\mathbf{D})$ にはデータ間の距離をデータの存在区間の長さで割った値を用いる．

しかし解像度 R において取り得る全ての境界の組み合わせは R が増加するにつれて大幅に増加するため，境界は依存性を最大とする軸を各軸 1 本ずつ加えていき，加えた軸に関しては除外して考える．

3.2 条件付き独立性検定

3.1 節で紹介した多重解像度独立性検定による条件付き確率の計算は次の手順で行う．

1. 変数 Z についてデータを分割する
2. 分割したデータについて X と Y の独立性を検定する
3. それぞれの結果を統合する

この方法において，データを分割する理由について述べる．

定理 1. 条件 $(X, Y \perp Z)$ が与えられた時，次の条件が成り立つ．

$$
(X \perp Y|Z) \Leftrightarrow (X \perp Y)
$$

上述の定理 1 によれば，$(X, Y \perp Z)$ の条件下において，X と Y の独立性を計算することで X と Y の条件付き独立性が求められることになる．しかし条件付き独立性を調べたい X, Y, Z においてはこの条件は成り立たない．この条件が成り立つ状況を作り出すためにデータを分割し，Z と X, Y の独立性を高めている．しかし，この方法を用いても $(X, Y \perp Z)$ の条件が成り立たなければ条件付き独立性を判定することはできない．以上を踏まえた上で $(X, Y \perp Z) \equiv U$, $(X \perp Y) \equiv I$, $(X \perp Y|Z) \equiv CI$ とし，分割データ $\mathbf{D}_i$ における条件付き独立確率 $Pr(CI|\mathbf{D}_i)$ を表すと，次式のようになる．

$$
\begin{aligned}
Pr(CI|\mathbf{D}_i) &= Pr(CI|U,\mathbf{D}_i)Pr(U|\mathbf{D}_i) + \\
&\quad Pr(CI|\neg U,\mathbf{D}_i)Pr(\neg U|\mathbf{D}_i) \\
&= Pr(I|\mathbf{D}_i)Pr(U|\mathbf{D}_i) + \rho(1-Pr(U|\mathbf{D}_i))
\end{aligned} \tag{10}
$$

また，データ分割・統合の方法として recursive-median アルゴリズムを用いる．このアルゴリズムは以下の手順で行われる．

1. 変数 Z の中央値でデータを分割する
2. 分割したそれぞれのデータに対し $Pr(U|\mathbf{D}_i)$, $Pr(I|\mathbf{D}_i)$ を計算し相乗平均をとる
3. $Pr(U|\mathbf{D}_i)$ の平均が分割する前の $Pr(U|\mathbf{D})$ よりも大きければ分割を行う

以上の手順を再帰的に繰り返すことで，$Pr(I|\mathbf{D})$ を最大にするデータの分割が行われ，信頼度の高い $Pr(CI|\mathbf{D})$ の値を得ることができる．

4 実験

Gene Expression Omnibus(http://www.ncbi.nlm.nih.gov/projects/geo/) から入手した出芽酵母の細胞周期を計測した遺伝子発現量のデータ (GSE4987) から PC(Path Consistency) アルゴリズム [7] (Spirtes, Glymour, Scheines 1993) を用いて遺伝子ネットワークの構築を行った．実験には図 1 に示した KEGG(Kyoto Encyclopedia of Genes and Genomes) データベースの細胞周期パスウェイの 1 部に関係のある遺伝子のみを用いた．実験結果を図 2 に示す．

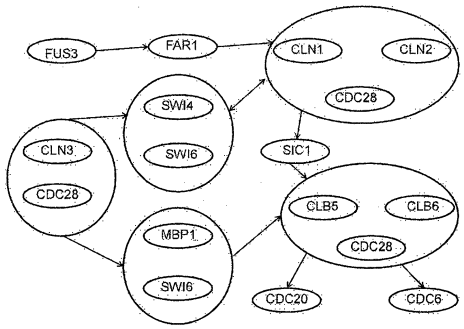

図 1: ターゲット ネットワーク

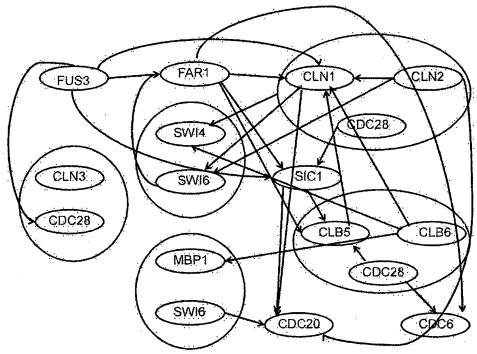

図 2: 実験結果

実験では，ターゲット ネットワーク内に存在するエッジが 12 本，ターゲット ネットワークに存在しないエッジが 13 本検出された．依存関係の検出の精度を向上させるためには，同じ実験データの統合によるサンプルの追加，遺伝子発現データのノイズ除去や欠損値補完といったデータの信頼度を高める処理を行う必要がある．また生物学の知識に基づいた制約条件の導入の検討や他手法との比較なども今後行う必要がある．

5 おわりに

確率分布を仮定しない多重解像度独立性検定，条件付き確率検定を用いたベイジアンネットワークの構築法を導入し，PCアルゴリズムを用いて実際の遺伝子発現データから遺伝子ネットワークの構築を行った．

参考文献

1) Margaritis D., Thrun S., "A Bayesian multiresolution independence test for continuous variables," Uncertainty in Artificial Intelligence(UAI), 346-353, 2001.

2) Margaritis D., "Distribution-free learning of Bayesian network structure in continuous domains," Proceedings of the Twentieth National Conference on Artificial Intelligence(AAAI), 825-830, 2005.

3) Friedman N., Linial M., Nachman I., Pe'er D., "Using Bayesian networks to analyze expression data," Journal of Computational Biology, 7, 601-620, 2000.

4) Akutsu T., Miyano S., Kuhara S., "Algorithm for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function," Journal of Computational Biology, 7, 331-343, 2000.

5) Chen T., He H. L., Church G. M., "Modeling gene expression with differential equations," Proceedings of Pacific Symposium on Biocomputing, 29-40, 1999.

6) Toh H., Horimoto K., "Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling," Bioinformatics, 18, 287-297, 2002.

7) MSpirtes P., Glymour G., Scheines R., "Causation,Prediction,and Search," New York: Springer-Verlag, 1993.