

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[‡]

1. はじめに

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。

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く、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うことが挙げられる。

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。本研究では 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。

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 $c_1 \dots c_{I \times J}$ 、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ディリクレ分布で仮定し、各領域の境界の集合を B_R とする。データ D の尤度は次式で表される。

$$\Pr(D) = \frac{\Gamma(\gamma)}{\Gamma(\gamma + N)} \prod_{k=1}^R \frac{\Gamma(\gamma_k + c_k)}{\Gamma(\gamma_k)}$$

$\gamma = \sum \gamma_i$ であり γ_i はディリクレ分布のパラメータを表す。上の式を $r(C_{IJ}, \gamma_a)$ とおく。

データが依存モデルか独立モデルのどちらか一方から発生すると仮定すると次式が成り立つ。

$$\Pr(D) = \Pr(D | M_I) \Pr(M_I) + \Pr(D | M_{\neg I}) \Pr(M_{\neg I})$$

さらに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 C_{IJ} の情報を I 方向と J 方向の成分に分解し、それぞれのデータ尤度を求める。よって依存モデル、

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

独立モデルのデータ尤度は $r(C_{IJ}, \gamma_a)$ を用いてそれぞれ次のように表せる。

$$\Pr(\mathbf{D} | M_{\neg I}) = r(C_{IJ}, \gamma_J)$$

$$\Pr(\mathbf{D} | M_I) = r(C_I, \alpha_I)r(C_J, \beta_J)$$

これを前述の式に代入し、ベイズの定理で変形することによって次式を得る。

$$\Pr(M_I | \mathbf{D}) = \frac{1}{1 + \frac{1-\rho}{\rho} \frac{r(C_{IJ}, \gamma_J)}{r(C_I, \alpha_I)r(C_J, \beta_J)}}$$

解像度を上げながらそれぞれの解像度でこの値を計算し、依存度が最大になる解像度における独立性をこのデータの独立性とする。

3.2. 条件付き独立検定

3つの変数集合 X, Y, Z において $(X \perp Y | Z)$ が成り立つならば、 $(X \perp Y | Z) \Leftrightarrow (X \perp Y)$ であるという定理を利用するため、 $(X, Y \perp Z)$ の平均値を最大にするようにデータを変数 Z について分割する。この分割には Z の中央値で分割し、独立性の計算を再帰的に行う recursive-median アルゴリズムを用いる。データを分割した後、分割領域内において $(X \perp Y)$ を計算することで $(X \perp Y | Z)$ を求め、これの相乗平均をとることで全体の条件付き独立性とする。

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。遺伝子ネットワークの構築法として Score-Based Approach (探索法は Hill Climbing, 親子間の関係は線形, スコアは BIC とする。), Independence-Based Approach (独立性の検定には相関係数を用いる), Independence-Based Approach に多重解像度独立性検定を導入したもの 3 手法について行い、ターゲットのエッジ数に対する獲得エッジの正解数である Sensitivity とターゲットの非エッジ数に対する獲得非エッジの正解数である Specificity を用

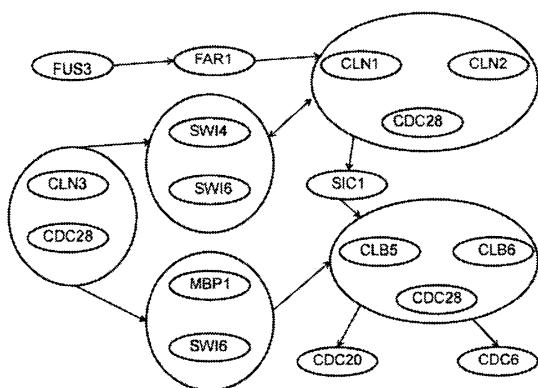


図 1. ターゲットネットワーク

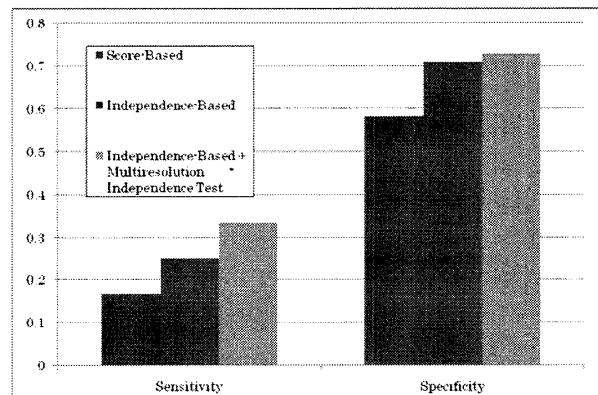


図 2. 実験結果

いて 3 手法を比較し、多重解像度独立性検定を導入することで精度の改善がみられた。その結果を図 2 に示す。

5. まとめ

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、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.