

# ベイジアンネットワークによる遺伝子制御ネットワークの推定

島田 公敬<sup>\*1</sup> 相場 亮

<sup>\*1</sup> 芝浦工業大学大学院 工学研究科<sup>†</sup>

## 1 はじめに

ゲノム資源の増加と共に、生命科学全体で分子・要素レベルからシステムネットワークレベルへと理解の枠組みが大きく変化しつつある。その一例として近年マイクロアレイ技術の発展に伴い、数千から数万の遺伝子の発現状態を同時に観測することが可能となった。また大規模な遺伝子発現の構造だけでなく、細胞周期や環境変化に対する応答、体内時計、発生・組織特異的な発現制御などの生命現象を支える複雑なネットワーク構造も明らかになってきた[1]。その結果機能未知の遺伝子も含んだ膨大な相互作用ネットワークのデータが蓄積している。この大量かつ多種類の情報を用いた遺伝子ネットワークの推定問題は、現在バイオインフォマティクスにおいて最も精力的に研究が進められている分野の一つである[2]。

本研究は世界的に広く使われているマイクロアレイデータを用いて遺伝子制御ネットワークの推定を行うものである。しかし、これにはノイズが非常に大きいという欠点がある。そこで実験誤差に強いベイズ統計に基づくベイジアンネットワーク(以下 BN)を用いて遺伝子制御ネットワークの推定を行う。また情報量の不足問題に対し、生物学的データを組み合わせ、エージェント間交渉により効率的な推定を実現し、最適なネットワークを生成することをねらう。

## 2 マイクロアレイ

マイクロアレイは主に以下の2つに大別される。

- Affymetrix 社の光リソグラフィー法に基づくマイクロアレイ (GeneChip)
- Stanford 大学のスポット法に基づくマイクロアレイ (cDNA マイクロアレイ)

GeneChip は、cDNA マイクロアレイの 20 倍近い密度を持ち、規格製品であるため入手も比較的容易である。しかし、研究目的に合わせた注文生産が困難となる。

一方 cDNA マイクロアレイは、研究目的にあわせてマイクロアレイをデザインでき、その簡便さから広く用いられている。実験方法は 2 種類の細胞を用意する(通常細胞・サンプル細胞)。正常細胞とサンプル細胞から全遺伝子に関して mRNA を抽出し、それを鋳型として cDNA を生成する。cDNA のコピーを稠密にスポットしたガラス表面に、細胞から抽出し蛍光標識(Cy3, Cy5)した遺伝子転写産物を結合

させることで、遺伝子発現の変化を測定できる。測定サンプルにより発現の絶対量は異なるため、正常例と異常例のそれぞれを異なる色で蛍光標識し、標準となるスポットで正規化した後に、二色の相対強度を遺伝子発現量として測定する(GeneChip の場合ガラス表面への DNA の植え付け法が既知配列のスポットティングではなく、光リソグラフィーを用いた固相合成であることが多い)[3]。

またこの手法は、発現パターンが似た遺伝子は機能的にも関連するという前提に基づいているが、これが生物学的に正しい保証はない。そのため、遺伝子の依存関係そのものをデータから学習する試みとしてブーリアンモデル、線形モデル、ニューラルネットワークが提案されてきた。しかしこれらのモデルは、実際に大量データを処理するには至っていない。初めて 800 遺伝子という大量データの解析に踏み込んだのが BN モデルである[4]。

マイクロアレイ解析の問題点として常に挙げられるのは、以下の3つである。

- 遺伝子発現は細胞内メカニズムの一面にすぎない
- 数千の遺伝子に対し統計的に有意な結果を出せるほどデータが揃わない
- データのノイズが非常に大きい

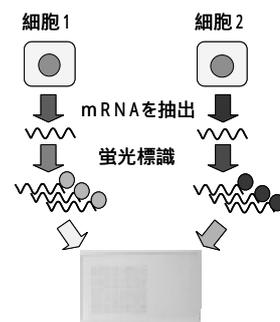


図 1: cDNA マイクロアレイ

## 3 ベイジアンネットワーク

BN とは 1988 年 Pearl によって最初に提案された確率推論のための枠組みであり、不確実性を含む対象領域における予測や意思決定を定量的に取り扱うことができる手法として注目を浴びた。

BN は設計変数間の依存関係を有向グラフにより表現する。BN の例を図 2 に示し、以下のような性質を有するグラフ構造として提示される。

- ネットワーク内のノード：確率変数
- 有効リンク(矢印)：依存関係
- 非循環グラフ：Directed Acyclic Graph

Estimation of Genetic Networks by Bayesian Networks

<sup>\*</sup> Kimitaka Shimada

<sup>†</sup> Graduate School of Engineering, Shibaura Institute of Technology

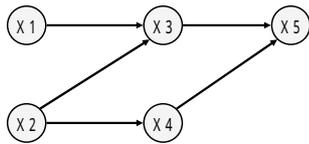


図 2: ベイジアンネットワークの例

図 2 において, ネットワークを構築する X1~X5 という 5 つのノードがそれぞれ変数を表しており, 矢印は依存関係を示す. つまり X3 は X1, X2 に依存し, X5 は X3 に依存することを示す. 影響を与える変数(矢印の始点となるノード)を親ノードと呼び, 影響を受ける変数(矢印の終点となるノード)を子ノードと呼ぶ.

#### 4 本システムについて

本研究は, 遺伝子をエージェントとする. 遺伝子エージェントには保持するデータと振舞が決めている. 本研究の遺伝子エージェントが保持する情報と振舞は主に以下のようなものがある.

- 保持する情報  
エージェント番号, 遺伝子名, 生物学的データ, エージェントの種類, 親ノード, 座標
- 振舞  
スコアの計算, 落札する, 交渉を受け取る, 交渉を持ちかける, 交渉相手を探す, 交渉の返事をする, 親ノードの追加, 親ノードの消去

ここで用いる生物学的データは, Kyoto

Encyclopedia of Genes and Genomes(以下 KEGG) プロジェクト[5]の代謝系やシグナル伝達系のデータである.

本システムの概要を以下の図 3 に示す.

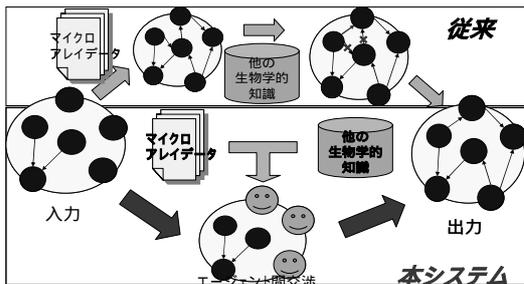


図 3: 本システムの概観

入力: マイクロアレイデータ  
出力: ベイジアンネットワーク  
方法:

ネットワーク構造の仮定  
事前確率分布の作成  
事前確率分布から事後確率分布を計算

$$p(\theta | d) = \frac{p(\theta)p(d | \theta)}{p(d)}$$

$p(\theta)$  : 事前確率  $d$  : データ

$p(\theta | d)$  : 事後確率

エージェント間交渉によりネットワーク構造を逐次変化させ, スコアが最大になるものを選択

従来研究で事前分布に既に決定しているネットワークを与える方法や, ネットワーク構築後に生物学的データを用いて修正する方法がある. 本システムは, エージェントの保持する情報として生物学的データを持たせることで, 計算速度を向上させることができるのではないかと考えた.

また従来では同じスコアである場合, どちらが最適なネットワークか判断できなかったが, 生物学的データを用いることで判断できるようになると考える.

#### 5 遺伝子制御ネットワークの推定実験

本研究で製作したシステムを用いて遺伝子制御ネットワークの推定を行った. 本システムと従来の BN との計算速度の比較を行った. 同じマイクロアレイデータから推定を行った場合本システムのほうが速いことが示すことができた.

計算精度においては, KEGG で公開されているネットワークを正解とした. その際, 従来の BN よりも本システムの有効性を示すことができた.

#### 6 むすび

本研究では知識データとして生物学的データを組み込み, 学習機構としてベイジアンネットワークにエージェント間交渉を用いるという新しい提案を行った. マイクロアレイデータだけでなく生物学的データを用いることで計算速度が向上した. しかし, 計算精度は従来の BN よりは向上したが, 他の手法であるダイナミックベイジアンネットワーク[6]と比較すると検出できなかったエッジがある.

今後の展望としては, ダイナミックベイジアンネットワークに組み込むことを考える. また他の生物学的データを組み込むことも有効である. そしてより有意なネットワークを生成し, 専門家に新しい可能性を示していければと思う.

#### 参考文献

- [1] 上田 泰己: 遺伝子発現ネットワークダイナミクス: 人工知能学会誌: Vol. 20, No. 3, 2005/5.
- [2] 井元 清哉: マイクロアレイデータ解析における統計的方法論の開発
- [3] J.L.DeRisi, C.R.Iyer, P.O.Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," Science, col. 278, pp. 680-686, 1997.
- [4] T.A.Brown, "Genomes," BIOS Scientific Publishers, 1999.
- [5] <http://www.genome.ad.jp/kegg/>
- [6] S.Y.Kim, S.Imoto, S.Miyano: "Dynamic Bayesian Network and Nonparametric Regression for Nonlinear Modeling of Gene Networks from Time Series Gene Expression Data", CMSB2003