LA-008

遺伝子の機能分類を利用した遺伝子制御ネットワーク推定手法
Inference method of gene regulatory networks using gene functional classification

瀧 浩平 * 寺本 礼仁 † 竹中 要一 * 松田 秀雄 *
Kohei Taki Reiji Teramoto Yoichi Takenaka Hideo Matsuda

1. はじめに

遺伝子制御ネットワークの解明のため, cDNA マイクロアレイによって測定された発現データを解析し, 多数の遺伝子について網羅的に制御ネットワークを推定する研究が行われてきた. その中でも大量の遺伝子を対象とする場合には, モジュールネットワークモデルに基づく推定が適する事が報告されている [1]. このモデルには, 同じ遺伝子によって制御される遺伝子の集合をモジュールとしてまとめて扱う事で, 組合せ爆発の問題を緩和できる利点がある. しかし, このモデルに基づく推定の先行研究では, 発現データのみを利用して推定が行われており, 発現データに含まれる測定誤差が推定に大きな影響を与える問題点がある. マイクロアレイから得られる発現データには大きな測定誤差が含まれており, 測定誤差によって遺伝子間に偽の関係が推定される恐れがある.

そこで本論文では, モジュールネットワークモデルに基づく推定に, 遺伝子の機能分類の情報を取り入れる手法を提案する. 機能分類と発現データの双方を考慮した, 遺伝子の間の類似性に基づいてモジュールを推定する事で, 生物学的知見との間に矛盾の少ないモジュールの推定を可能にする. 提案手法を出芽酵母の細胞周期の制御ネットワークの推定に適用し, その評価を行う事で, 測定誤差の影響の軽減について有効性を示す.

2. モジュールネットワークモデル

モジュールネットワークは, モジュールの集合と, その間を結ぶ有向辺によって表される. モジュールは同じ遺伝子に制御される遺伝子の集合として定義される. 有向辺は遺伝子からモジュールへ結ばれ, その遺伝子がモジュールに含まれるすべての遺伝子を制御する事を表す. 図 1 は例えば, 遺伝子 CLN1, CLN2 が同じモジュールに含まれ, 遺伝子 SWI4, SWI6 に制御される事を表す.

モジュールネットワークの推定は図 1 に示す様に, 1) 各遺伝子が含まれるモジュールの推定, 2) 各モジュールの間の制御関係の推定, の 2 つの段階に分かれる. 評価値が収束するまで 2 つの段階を交互に繰り返す事で, モジュールネットワークは推定される.

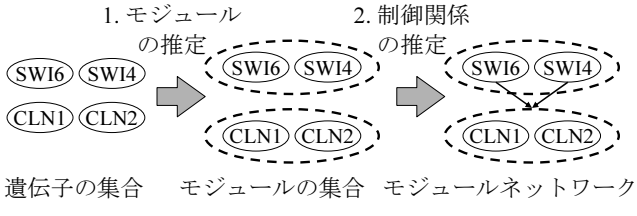

図 1: モジュールネットワークモデルにおける推定手順

3. 遺伝子の機能分類を利用したモジュールネットワーク推定手法の提案

3.1 機能分類の利用

遺伝子の機能分類とは, 遺伝子を機能に応じてカテゴリへ分類した情報である. 本論文ではこの様なカテゴリの体系として Gene Ontology (GO) を用いる. GO では, カテゴリは GO term として定義され, GO term 間には上位 · 下位概念の関係が階層的に定義されている.

提案手法では, 遺伝子の発現データに加えて, 機能分類もモジュール集合の推定に用いる. 同じ遺伝子に制御される遺伝子は似た役割を果たす場合が多く, 遺伝子間の機能の類似性の利用はモジュールの推定に有用だと考えられる. 機能分類による遺伝子間の類似性の評価では, GO の階層を考慮する GO term 間の類似度を用いる事によって, 機能分類の情報の効果的な利用を試みる. 本論文で利用する semantic similarity[2] $sim(c1, c2)$ は, 階層上で GO term $c1, c2$ に共通の上位概念となる GO term の情報量によって, $c1, c2$ 間の類似度を評価する. GO term c の情報量 $I(c)$ は,GO term c やその下位概念に分類されている遺伝子が少ないほど大きくなる. $sim(c1, c2)$ の定義は以下の様に与えられる.

$$sim(c1, c2) = - \max_{c \in \mathbf{S}(c1,c2)} I(c)$$

関数 $\mathbf{S}(c1, c2)$ は,$c1, c2$ に共通の上位概念の集合を与える.

3.2 機能分類に基づいたモジュール評価関数

モジュールネットワークの評価関数は, ベイズ統計に基づいて, 発現データ D 観測後の, モジュール集合 $\mathbf{M}$ と有向辺 $B_{\mathbf{M}}$ の事後確率 $p(B_{\mathbf{M}}, \mathbf{M}|D)$ として定義される. 事後確率をベイズの定理により分解し, 推定中に不変な係数を無視する事で, 評価関数 $score(B_{\mathbf{M}}, \mathbf{M} : D)$ は以下の様に定義される.

$$\begin{aligned} score(B_{\mathbf{M}}, \mathbf{M} : D) &= \log p(\mathbf{M}) + \log p(B_{\mathbf{M}}|\mathbf{M}) \\ &+ \log p(D|B_{\mathbf{M}}, \mathbf{M}) \end{aligned}$$

$p(D|B_{\mathbf{M}}, \mathbf{M})$ は Segal らによって与えられている定義 [1] を利用し, $p(B_{\mathbf{M}}|\mathbf{M})$ は均一に分布すると想定する. $p(\mathbf{M})$ はモジュール集合 $\mathbf{M}$ の事前分布である.

機能分類は発現データに依存しない事を考慮し, 本論文では機能分類に基づいて事前分布 $p(\mathbf{M})$ を定義した. 上記の議論より, 互いに似た機能に分類される遺伝子が同じモジュールに含まれる事が多いほど, モジュール集合の推定は尤もらしく, その事前確率は高くなると考えられる. モジュールに含まれる遺伝子の組すべてについての, semantic similarity に基づいた遺伝子間の類似度の和に比例する様に, 事前分布 $p(\mathbf{M})$ を定義した.

$$p(\mathbf{M}) = C \cdot \sum_{\mathbf{m} \in \mathbf{M}} L(\mathbf{m}) - \log Z(C)$$

*大阪大学大学院情報科学研究科バイオ情報工学専攻
†住友製薬株式会社研究本部ゲノム科学研究所

$$p(\mathbf{m}) = \frac{1}{|\mathbf{C}(\mathbf{m})|} \sum_{g1 \in \mathbf{m}} \sum_{g2 \in \mathbf{m}} l(g1, g2)$$

$$l(g1, g2) = \sum_{c1 \in \mathbf{C}(g1)} \sum_{c2 \in \mathbf{C}(g2)} sim(c1, c2)$$

$\mathbf{C}(g)$ は遺伝子 g が分類されている GO term の集合とし，$\mathbf{C}(\mathbf{m}) = \bigcup_{g \in \mathbf{m}} \mathbf{C}(g)$ とする．C はパラメータであり，$Z(C)$ は C に依存する $p(\mathbf{M})$ の正規化項を表す．

この事前確率の定義により提案手法では，似た発現パターンを示す遺伝子に加えて，似た機能に分類されている遺伝子をより多く含むモジュールが推定される．

4. 細胞周期の制御ネットワークの推定実験

4.1 実験条件と結果

出芽酵母の細胞周期の発現データ [3] に対して従来手法と提案手法を適用し，遺伝子数 800 個の制御ネットワークを推定した．ここで従来手法とは，発現データのみを用いて，モジュールネットワークに基づいた推定を行う手法であるとする．提案手法では，*Saccharomyces* Genome Database で公開されている機能分類を利用した．

細胞周期において発現する時期が既知の 95 個の遺伝子を，発現する時期で 5 つのモジュール G_1(47), S(8), S/G_2(6), G_2/M(15), M/G_1(19) に分け，推定精度の評価のため，これらを正しいモジュール集合と仮定した．括弧内の数字は各モジュールが含む遺伝子の数を表す．従来手法と提案手法それぞれを用いて推定したモジュール集合の，精度の評価結果を表 1 に示した．提案手法は従来手法と比較して高い精度を示しており，特に sensitivity に大きな改善が見られる．表 1 の最後の行は，機能分類のみを用いた推定結果の精度を示しており，機能分類だけから推定できる自明な問題ではない事を示唆している．

事前分布 $p(\mathbf{M})$ のパラメータ C と，提案手法による推定精度の間の関係を，図 2 に示した．$C = 0.75$ で 2 つの精度はピークを示しており，ここで示した提案手法による推定結果ではこの値を用いた．

表 1: 利用データの相違による推定結果の比較評価

利用データ	selectivity(%)	sensitivity(%)
両方 (提案手法)	**82**	**52**
発現データのみ	77	39
機能分類のみ	52	49

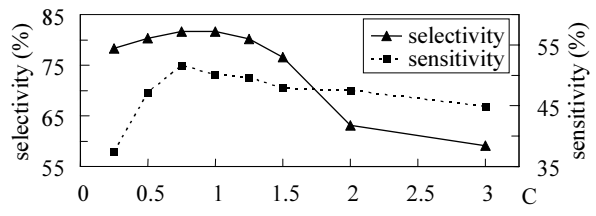

図 2: 精度とパラメータ C の関係

4.2 考察

表 1 に示す様に，発現データのみ，機能分類のみを用いた推定の双方と比較して，提案手法は高い精度を示した．これは，発現データと機能分類を同時に利用する事で，双方のデータの欠点を補いあった推定が実現された事を示唆していると考えられる．この推察を確認するため，図 3 の対称行列によって，表 1 の 3 種類の推定結果を G_1 期の遺伝子について比較した．行列の行と列は正解のモジュール集合の各遺伝子に対応し，黒色の要素は，その行と列の遺伝子が推定結果において同じモジュールに含まれる事を表す．図 3 の比較で，A の枠内の発現データから推定できない部分は機能分類からは正しく推定できており，B の枠内で機能分類から誤って分類される部分は発現データからはほとんど分類されていない．更に，提案手法による結果が双方の結果の長所を受け継いだ事が，図 3 から分る．これは，発現データの測定誤差が推定に与える影響が軽減された事を示しており，機能分類を取り入れる事の有効性を示す事実であると考えられる．

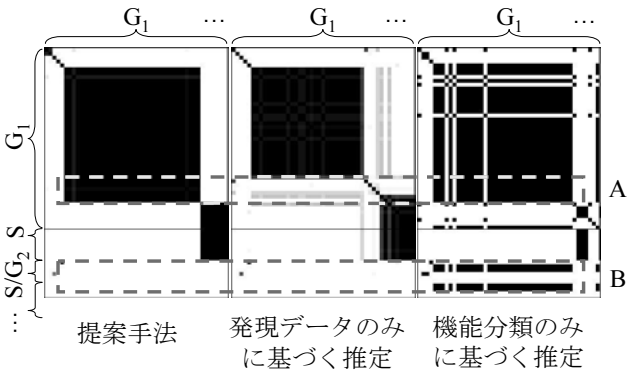

図 3: 各手法の推定結果のモジュールの構成

5. おわりに

モジュールネットワークモデルに基づき，発現データに加えて機能分類を利用して制御ネットワークを推定する手法を提案した．発現データのみによる推定は測定誤差の影響を大きく受けるのに対して，提案手法は機能分類を取り入れる事でこの影響を軽減できる．本論文では推定結果を分析し，発現データと機能分類の互いの問題点を補い合った推定を，提案手法が行った事を確認した．この事から，発現データの測定誤差が推定に与える影響の軽減について，提案手法が有効であると考えられる．

6. 謝辞

本研究は，一部，文部科学省 IT プログラム，科学研究費特定領域研究「統合ゲノム」によっている．

参考文献

[1] E.Segal, M.Shapira, et al., “Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data”, *Nat. Genet.*, Vol. 34, No. 2, pp. 166–176 (2003).

[2] P.W.Lord, R.D.Stevens, et al., “Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation”, *Bioinformatics*, Vol. 19, No. 10, pp. 1275–1283 (2003).

[3] P.T.Spellman, G.Sherlock, et al., “Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization”, *Mol. Biol. Cell*, Vol. 9, No. 12, pp. 3273–3297 (1998).