

表現型による遺伝子ネットワークの推定

理学専攻 情報科学コース 阿部奈津美 (指導教員：瀬々 潤)

1 はじめに

疾病の原因や治療補助に向け、生命のシステムの理解研究が活発に行われている。この研究として、最たる物はマイクロアレイによる網羅的遺伝子発現データからの遺伝子ネットワーク推定である [1]。しかし、遺伝子発現によって推定できるネットワークは生命システムの極一部であり、代謝パスウェイを含む、様々な生体内ネットワークの構築が望まれている。本研究では、観測技術の高まりによって今後増加の見込まれる上、非常に容易に観察可能な顕微鏡画像に着目し、顕微鏡画像から観察される細胞の表現型変異を基に、形態を制御する遺伝子のネットワークを考察する。特に, Ohya *et. al.* [2] で網羅的に観測されている出芽酵母の 1 遺伝子破壊株の形態変異データを利用し、ネットワークの構築を行う。

1 遺伝子の破壊により、その遺伝子が影響を及ぼしている様々な細胞内プロセスは停止し、結果として、形態に変異が起こることがある。1 遺伝子の破壊による形態情報の例を図 1 に示す。左側のイラストは野生型の形態を模擬しており、大きな楕円は母細胞、小さい楕円は出芽した芽を示している。また右側のイラストは遺伝子破壊によって変化した後の形態を示している。たとえば、図 1 の一番上の図は、遺伝子 A の破壊により母細胞の大きさ、芽の大きさ、母細胞の形、母細胞と芽の角度という 4 つの異常が観測された事を表している。以降、 $B(A) = \{ \text{母細胞の大きさ, 芽の大きさ, 母細胞の形, 母細胞と芽の角度} \}$ のように、破壊した遺伝子 x に対する形態変異集合を $B(x) = \{ p | p \in \text{変異した形態} \}$ と表記する。イラストでは太線もしくは点線で変化のあった形態部位を示した。

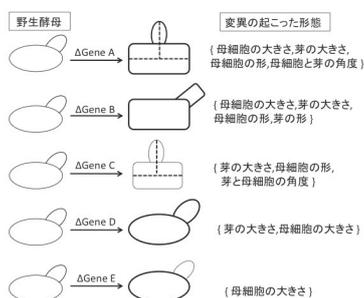


図 1: 1 遺伝子破壊に対する変異の起こった形態の例

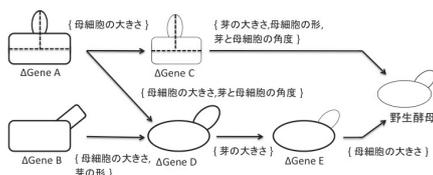


図 2: 表現型の観測により推測されるネットワーク

この観測結果から、遺伝子間のネットワークを考える。遺伝子 A と D の破壊株について、A の破壊により母細胞の大きさ、芽の大きさ、母細胞の形、母細胞と芽の角度の 4 つの形態変異が、D の破壊により芽の大

きさと母細胞の大きさの 2 つの形態変異が観測されている。D の破壊で起こる 2 つの形態変異は、A の破壊で起こる 4 つの形態変異に包含されている、すなわち $B(A) \supset B(D)$ であることに注意して、遺伝子破壊によって引き起こされる生体内プロセスの停止について考えると、D の破壊が形態に及ぶ影響は、A の破壊が形態に及ぶ影響の一部であると考えられ、遺伝子ネットワーク上で D は A の下流にあると考えられるだろう。同様に、A と C を考えると、C の破壊で起こる 3 つの変異は A で起こる変異に包含されている、すなわち、 $B(A) \supset B(C)$ であるので、C は A の下流であると考えられる。しかし、C と D を比べると、いずれの形態にも包含関係は無く、上流、下流の判断はできない。以上の考察を A から E の 5 つの遺伝子間で考えると図 2 に示すネットワークが構築できる。

しかしながら、多くの場合観測ノイズによる異常形態の誤検出を含むため、必ずしも上述の通りの理想的なネットワーク構築はできない。本論文では、以上の考察を元に観測ノイズ耐性を持つ形態ネットワーク構築を行う。

2 遺伝子ネットワークモデル

表 1: 形態変異データの例

| 破壊した遺伝子 x | 変異した形態 $B(x)$ |
|-------------|----------------------------|
| o_A | $\{ p_1, p_2, p_3, p_4 \}$ |
| o_B | $\{ p_1, p_2, p_3, p_5 \}$ |
| o_C | $\{ p_2, p_3, p_4 \}$ |
| o_D | $\{ p_1, p_2 \}$ |
| o_E | $\{ p_1 \}$ |

2.1 形態変異からのネットワーク構成

まず表 1 から第 1 章の形態ネットワークの例をもとに、ネットワークを構築する。

1 章から $B(o_1) \supset B(o_2)$ であれば、遺伝子 o_1 は遺伝子 o_2 の遺伝子ネットワーク上で上流に存在すると考えられるので、表 1 では $B(o_A) = \{ p_1, p_2, p_3, p_4 \}$, $B(o_C) = \{ p_2, p_3, p_4 \}$ から $B(o_A) \supset B(o_C)$ なので o_A は o_C の上流にあると推定できる。ここで o_A が o_C の上流にある場合、 $o_A \succ o_C$ と書き、遺伝子間に順序を定める。同様に、順序関係を全ての遺伝子対に対し計算できる。

しかし、生物の実験はしばしば観測ノイズを含む。単純なブリアンネットワークでは、このノイズに敏感に反応してネットワークを推定してしまう。例えば、 $B(o_A) = \{ p_1, p_2, p_3, p_4 \}$ にノイズが発生して、 $B(o_A) = \{ p_1, p_2, p_3 \}$ と観測されたとすると、 $o_A \succ o_C$ が成立しなくなる。そこで我々は、1 遺伝子破壊に対する変異した形態情報をもとに遺伝子間の距離を定義し、ノイズを考慮したネットワークの推定を拡張する。

2.2 ノイズを考慮したネットワークの為の距離の定義

ノイズを考慮するために遺伝子間に形態変異集合の順序関係を拡張した次のような距離を定義する。

$$d_{sub}(o_1, o_2) = |B(o_2) \setminus B(o_1)|$$

ここでは“\”は集合の差を表している。

しかし、この距離は $|B(o_1)|$ と $|B(o_2)|$ を考慮に入れていない。形態変異の数が多くなればなるほど、偶然一つの観測誤差が起こる確率は高くなるため、このような状況を解決する為に距離を $d(o_1, o_2)$ を定義する。

$$d(o_1, o_2) = \frac{|B(o_2) \setminus B(o_1)|}{(|B(o_1)| + |B(o_2)|)^k}$$

k は正規化の為にパラメータで、 k が大きい程、形態変異の大きな破壊株でノイズを許す指標となっている。

2.3 遺伝子の距離とネットワークの推定

我々は 2.2 節で求めた距離をもとに次の 3 つの手順でネットワーク推定を行う。

ここで、今まで定義した 1 遺伝子破壊による形態変異間の距離を拡張して、遺伝子グループ間にも形態距離を導入する。 C_i, C_j を遺伝子のグループとして C_i と C_j の遺伝子グループの距離は

$$d(C_i, C_j) = \max\{d(o_1, o_2) \mid o_1 \in C_i \text{ and } o_2 \in C_j\}$$

と計算する。

このグループ間距離の定義を用いると、閾値 θ と $d(C_i, C_j)$ との間に次の 3 つの関係が存在する。ここで、一般性を失うことなく $d(C_i, C_j) \leq d(C_j, C_i)$ と仮定できる。 θ は、観測ノイズ耐性を示すパラメータであり、値が大きいほどノイズ耐性の高いネットワークが形成される。

- 1, $d(C_i, C_j) \leq d(C_j, C_i) \leq \theta$
- 2, $d(C_i, C_j) \leq \theta < d(C_j, C_i)$
- 3, $\theta < d(C_i, C_j) \leq d(C_j, C_i)$

1. は変異した形態が似ている遺伝子グループの場合で、形態変異の差は観測ノイズの可能性があると考えられるので同一のグループとみなす。

2. では 2 つのグループ間の距離に差がある場合である。定義した距離は下流に存在しやすい程度を示していたので、この場合 C_i は C_j の下流に存在しやすく、 $C_j \succ C_i$ が成立するといえる。

3. では、どちらからの距離も閾値以上の大きい値であるので、遺伝子グループ C_i と遺伝子グループ C_j の引き起こす形態はかなり異なるといえる。ゆえに、 C_i と C_j は何の関係ももたない。

以上の方法により、遺伝子ネットワークが推定できる。しかし、順序関係に意味が重複しているものが存在する。例えば $\{o_A, o_B\} \succ \{o_D\}, \{o_D\} \succ \{o_E\}$ があれば、 $\{o_A, o_B\} \succ \{o_E\}$ は導かれる。上流の遺伝子は下流の遺伝子の制御に関わることが前提なので、このように重複した関係は省略する。

3 閾値 k と θ の決定

今までの章でネットワークの推定をする為に、距離の定義及びネットワーク推定の為に閾値を設定した。本研究では、この閾値を、遺伝子オントロジー (GO)[3] における遺伝子機能の一致度を遺伝子グループ内で調査することにより決定した。これは細胞に対する遺伝子の役割が似ている (同じ遺伝子グループに属する) ならば、起こった形態変異も類似しているはずという考えに基づいている。その結果、遺伝子機能の一致度を最大にする k と θ' はそれぞれ 0.46875, 38 と求めた。ここで、 $\theta' = \theta/42.28$ である。42.28 とは 1 遺伝子破壊時の形態変異数の平均を表している。

4 推定されたネットワークの検証

本手法で得たネットワークの信頼性を調査する為、ノイズ耐性の調査と、既存のネットワークとの比較を行った。

4.1 ノイズ耐性

利用した形態情報に人工的に 0.5 % のノイズを加え、ノイズを加える前と後で、存在する順序関係についての比較を行った。本研究ではこの手順を 10 回繰り返した。その結果、ノイズを加える前と加えた後での順序関係の一致度は、平均 96.1 %、標準偏差は 1 % となった。

4.2 既存のネットワークとの比較

推定された本手法によるネットワークと、KEGG パスウェイとの比較を行った。利用した 172 の遺伝子中 16 の遺伝子が KEGG パスウェイ上に存在した。さらに、その中の遺伝子のペアで KEGG パスウェイ上でパスが存在したのは 43 ペアであった。その中で本手法によるネットワークと順序関係が一致したものは 18 ペアだけであった。また ChIP-Chip データセットとの比較も行なったが、本手法によるネットワークとの関係はほとんどなかった。この結果は、本手法で推定したネットワークが KEGG パスウェイのネットワークと別のネットワークであることを示唆している。

5 まとめと結論

今まで網羅的な遺伝子の発現量観測から遺伝子ネットワークが描かれてきたが、本研究では遺伝子破壊による形態の変異から遺伝子ネットワークを推定した。その結果である本手法により推定された遺伝子ネットワークの全体図については、<http://lab.se-se.jp/~abe/MasterThesis/GNetwork.html> に掲載した。推定されたネットワークの、ノイズ耐性と既存のネットワークとの比較、生物学的な検証 (スペースの都合上省略) から、まだ発見されていない遺伝子同士の関係が求まっている可能性があるといえる。しかし、ノイズ耐性のさらなる向上はこれからの課題である。

参考文献

- [1] T.Akutsu, S.Miyano, and S.Kuhara. *Inferring qualitative relations in genetic networks and metabolic pathways*. Bioinformatics, 16, 727-734, 2000.
- [2] Y.Ohya, et.al. *High-dimensional and large-scale phenotyping of yeast mutants*. Proc. Natl. Acad. Sci. Vol. 102. No.52. 19015-19020, 2005.
- [3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et al. *Gene ontology: tool for the unification of biology. the gene ontology consortium*. Nat Genet, 25(1):25-29, May 2000.
- [4] M. Kanehisa and S. Goto. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Nucleic Acids Res, 28: 27-30, 2000. Proc. Natl. Acad. Sci. 98, 4569-4574, 2001.