2008－BIO－14 (9)
2008／9／19

表現型による遺伝子ネットワークの推定

阿部奈津美, 瀬々潤
お茶の水女子大学

概要

近年, 科学技術の進歩により, 細胞の形態 (表現型) を大規模に調査することが可能になった.Ohya *et al.* では, 出芽酵母について, ある遺伝子を破壊したときの表現型の変化を 501 のパラメータについて調べている. 本研究ではこの網羅的に調べられた表現型のデータを用い, 表現型による遺伝子ネットワークを推定し, 表現型と遺伝子との関係を調べた.172 個の転写制御因子を用いて遺伝子ネットワークを描くことで 63 の遺伝子グループとその遺伝子グループを 183 のエッジでつなぐネットワークが推定された. また, 本手法は種によらない為, 酵母以外の種における表現型の網羅的なデータからその種の遺伝子ネットワークも構築できる.

Uncovering a Phenotypic Gene Network Using Morphological Inclusion Relations

Natsumi Abe and Jun Sese. Ochanomizu University.

Abstract

Recent technologies enable us to make systematical observations of noteworthy traits such as morphological data. Ohya *et al.* observed morphological changes that occurred in 501 parameters as a result of one gene deletion in budding yeast. In this study, using the comprehensive phenotype data, we construct a phenotypic gene network for revealing the genetic effort on phenotype. We design a new method to build the network based on inclusion relations of abnormal phenotypes. This method generates a network having 63 groups and 183 edges from 172 transcription related genes' phenotypes. The network tells us which genes each phenotype is dependent on.

1 研究動機

疾病の原因や治療補助に向け, 生命のシステム的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 [5, 6]. 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, 代謝パスウエイを含む, 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.* 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.

1.1 形態ネットワークの例

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 A 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 (出芽酵母が出芽の際出芽元となる細胞を母細胞と呼ぶ) , 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 A 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. 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.

この観測結果から, 遺伝子間のネットワークを考えよう. 遺伝子 A と D の破壊株を考える. A の破壊により 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事に注意し, 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

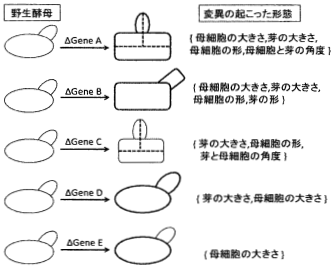

図 1 1 遺伝子破壊に対する変異の起こった形態の例

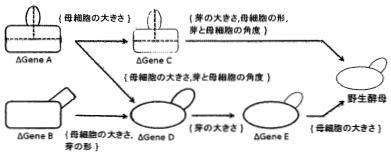

図 2 表現型の観測により推測されるネットワーク

B 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.

1.2 関連研究

遺伝子ネットワーク構築の研究は, 今まで遺伝子発現量を元にして行われてきた. 解析手法としては, ブーリアンネットワーク [6] やベイジアンネットワーク [5] が用いられてきた. しかし, 遺伝子発現によるネットワークは, 遺伝子が働いた事による下流の影響を見ているが, 本研究では遺伝子が破壊された事による影響を見ていること, また, 遺伝子間の発現調節は比較的直接的であるのにたいし, 形態は遺伝子からの影響が間接的であることから, 必ずしもこれらの手法が適用できるわけではない. 本研究では, より形態の変異観測に適したモデルとして, 形態の包含関係に基づくネットワーク構築を行う.

本論文は, 以下,2 章で形態ネットワーク構成の定義を行った後, ノイズ耐性を持つネットワーク構築を行う.3 章で,2 章で導入したネットワークに含まれるパラメータの決定と, 構築されるネットワークの考察を行い,4 章で結論と今後の課題について述べる.

2 遺伝子ネットワークモデル

2.1 形態変異からのネットワーク構成

まず表 1 から第 2 章の考察をもとに, ネットワークを構築しよう. ここで x の破壊で起きた形態変異集合を $B(x)$ と表す. この表 1 では遺伝子 o_i を破壊した際に変異した形態が表されている. 例えば, 遺伝子 o_B を破壊すると, 形態 $B(o_B) = \{p_1, p_2, p_3, p_5\}$ が変異していることを示している.

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 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. 我々は同様に, 半順序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. 例えば, 我々は [1] の,1 遺伝子を破壊した時の酵母の形態の変異を 501 種類について調べた実験の結果を用いているが, この実験では, 実験にかかる時間や実験のコストの問題により, 野生株の形態に対しては 127 回の観測,1 遺伝子破壊株については各遺伝子に対しそれぞれ 1 回ずつしか実験していない. その為, 実験でどの程度の観測ノイズが生じたか推定が困難である. しかも, 単純なブーリアンネットワークでは, このエラー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.

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

ここでは"\"は集合の差を表している. 例えば, 表 1 において,

$B(o_A) = \{p_1, p_2, p_3, p_4\}$ と $B(o_C) = \{p_2, p_3, p_4\}$

の距離は

$$d_{sub}(o_A, o_C) = |\{p_2, p_3, p_4\} \backslash \{p_1, p_2, p_3, p_4\}| = |\{\}| = 0$$

となり, また

$$d_{sub}(o_C, o_A) = |\{p_1, p_2, p_3, p_4\} \backslash \{p_1, p_2, p_3\}| = |\{p_4\}| = 1$$

である.

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ため, 変異の多い破壊株の形態変異を不当に大きく見る傾向にある. 遺伝子 o_A と o_B の破壊株間の距離, と遺伝子 o_C と o_D の破壊株間の距離を考えよう.$d(o_A, o_B) = |\{p_1, p_2, p_3, p_5\} \backslash \{p_1, p_2, p_3, p_4\}| = 1$.$d(o_D, o_E) = |\{p_1, p_2\} \backslash \{p_1\}| = 1$ となり,$d(o_A, o_B) = d(o_D, o_E)$ である. しかし,$d(o_A, o_B)$ は,$|B(o_A)| = 4$ であることから,4 つの形態変異中 1 つの変異が o_B の変異と異なる事を示し, 一方,$|B(o_D)| = 2$ であることから,2 つの形態変異中 1 つが o_E と異なる事をしめしている. 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,$B(o_A), B(o_B)$ 間の距離は,$B(o_D), B(o_E)$ 間の距離よりは短く設定したい. 我々は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.

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

k は正規化の為のパラメータで,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.$k = 0$ の時,$d(o_1, o_2) = d_{sub}(o_1, o_2)$ である.k の値の検証については,4 章で行う.

$k = 1$ とし実際の距離を計算する.

$d(o_B, o_A) = |\{p_4\}|/(|\{p_1, p_2, p_3, p_4\}|\{p_1, p_2, p_3, p_5\}) = 1/(4 \times 4) = 0.0625$

$d(o_E, o_D) = |\{p_2\}|/(|\{p_1, p_2\}|\{p_1\}) = 1/(1 \times 2) = 0.5$

以上より $d(o_B, o_A) < d(o_E, o_D)$ となり,$B(o_A)$ から $B(o_B)$ への変異はノイズに近いものとみなされ,$B(o_E)$ から $B(o_D)$ への変異は確かに起こった形態変異とみなされやすくなる. 全ての遺伝子間距離の計算結果を表 2 に示す.o_A, o_E の関係を考えよう.$d(o_A, o_E) = 0, d(o_E, o_A) = 0.75$ より,$d(o_A, o_E) < d(o_E, o_A)$ であることから,$B(o_A)$ が $B(o_E)$ に変異する可能性の方が $B(o_E)$ が $B(o_A)$ に変異する可能性より高く,o_A は o_E の上流にあると考えられる.

2.3 遺伝子の距離とネットワークの推定

表2 表1から得られたそれぞれの遺伝子間の距離行列

$A\backslash B$	o_A	o_B	o_C	o_D	o_E
o_A	0.00	0.0625	0.00	0.00	0.00
o_B	0.0625	0.00	0.0833	0.00	0.00
o_C	0.083	0.167	0.00	0.167	0.333
o_D	0.25	0.25	0.333	0.00	0.00
o_E	0.75	0.75	1.00	0.5	0.00

表3 表2のグループ化後の遺伝子グループ間の距離行列 ($\theta = 0.1, k = 1$)

$A\backslash B$	$\{o_A,o_B\}$	$\{o_C\}$	$\{o_D\}$	$\{o_E\}$
$\{o_A,o_B\}$	0.0625	0.0833	0.00	0.00
$\{o_C\}$	0.167	0.00	0.167	0.333
$\{o_D\}$	0.25	0.333	0.00	0.00
$\{o_E\}$	0.75	1.00	0.5	0.00

我々は次の 1.2.3. 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 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

$d(C_i, C_j) = \max\{d(o_1, o_2) | o_1 \in C_i \text{ and } o_2 \in C_j\}$ と計算する. 例えば, 表 2 において,$C_1 = \{o_A, o_B\}$,$C_2 = \{o_C\}$ とすると, $d(C_1, C_2) = \max\{d(o_A, o_C), d(o_B, o_C)\} = \max\{0, 0.0833\} = 0.0833$ である.

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グループ化する. $\theta = 0.1, k = 1$ と仮定すると, 表 2 では,$d(\{o_A\}, \{o_B\}) = d(\{o_B\}, \{o_A\}) = 0.0625 < \theta$ が 1. に該当する. 遺伝子 o_A と遺伝子 o_B をグループ化した後の距離行列は表 3 のようになる.

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. 表 3 では $\theta = 0.1$ として,$d(\{o_B\}, \{o_C\}) = 0.167$, $d(\{o_C\}, \{o_B\}) = 0.083$. $d(\{o_B\}, \{o_C\}) \leq \theta \leq d(\{o_C\}, \{o_B\})$ であるので,$\{o_B\} \succ \{o_C\}$ が成立する.

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. 表 3 の例では $\theta \leq d(\{o_C\}, \{o_D\}) \leq d(\{o_D\}, \{o_C\})$ であり, 表 2 から 1.2.3. の順で, グループ化, 半順序の決定を行った結果, 求まった半順序を次に示す.

$\{o_A, o_B\} \succ \{o_C\}, \{o_A, o_B\} \succ \{o_D\}$,

$\{o_A, o_B\} \succ \{o_E\}, \{o_D\} \succ \{o_E\}$

この半順序から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. このように重複した関係は省略する. すると表 3 から次の半順序が求まる.

$\{o_A, o_B\} \succ \{o_C\}, \{o_A, o_B\} \succ \{o_D\}, \{o_D\} \succ \{o_E\}$

このように重複を省いた半順序から我々は遺伝子グループ間の関係を示す遺伝子ネットワークを推定することが出来る. 上に表記した半順序から得られるネットワークを図 3 に示す.

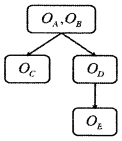

図3 推定されたネットワーク

3 閾値 k と θ の決定と結果の考察

今までの章ではネットワークの推定をする為に, 距離をの定義及びネットワーク推定の為の閾値を設定した. この章では, 提案手法を用い, 形態変異に大きな影響を及ぼしていると考えられる転写制御因子の形態変異ネットワークを構築する. 転写制御因子として, 遺伝子オントロジー (GO)[3] の biological process で転写制御に関連していると注釈づけられた遺伝子のうち,[1] で変異が観測されている 172 の遺伝子から, ネットワークを推定し, 閾値の決定と推定されたネットワークの結果の考察を行う. 転写制御因子に遺伝子を絞ってネットワークを推定したのは転写制御因子は形態と関連が強いと予想されるからである.

3.1 閾値の決定

まず我々は形態変異の適切な距離とネットワーク推定の為の閾値 k と θ を様々な閾値についてにネットワークを作成し,GO のデータと比較した. 比較の方法として次の精度を定義した.

グループ化をし, ネットワークを作成した結果, 互いの距離が θ 未満の遺伝子群 o_A,o_B,o_C があるとする. この 3 つの遺伝子について GO を調べる.o_A と o_B が GO の同じタームに関連づけられ, 一方 o_C が違うタームに含まれたとするとその遺伝子グループについて精度 $2/3 = 66$ % と定義する. 様々なパラメータ値での結果を表 4 に示す. 表において横軸の表示に使った θ' は $\theta' = \theta/42.28^{2k}$ である. この 42.48 とは変異した形態パラメータ数の平均値である. 例えば, 表 4 において $k = 0.375, \theta' = 32$ についての精度は 0.841 だが, これは次の手順で計算している. まず, 指定した k, θ' においてネットワークを構築する. このネットワーク上で,3 つ以上遺伝子が含まれるノードを抽出する. 各ノードについて精度を計算する. ここで, 精度計算の際に用いる GO は, 全 3 カテゴリの内,Cellular Component 及び Molecular Function の 2 つである.Biological Process は遺伝子選定の際に用いているので利用していない. 抽出した全グループそれぞれ計算した精度の平均を求める. この平均が

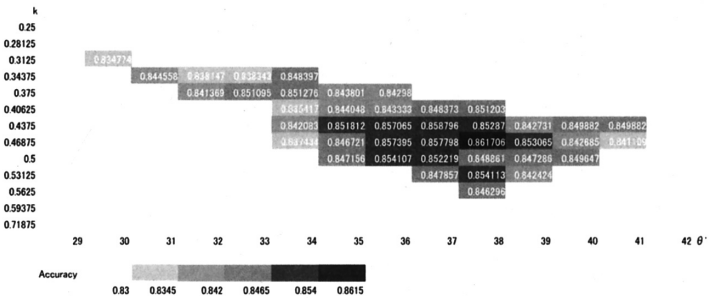

図 4　推定したネットワークの閾値対する精度

図で書かれている数値である. この表から GO による精度を最大にする閾値の値は $\theta' = 38, k = 0.46875$ である. この値でネットワークを作成した結果,63 の遺伝子グループと 183 のエッジでネットワークが作成された.

3.2　結果の考察

我々は 3.1 で決定した閾値 $\theta' = 38, k = 0.46875$ において, 形態変異に基づく遺伝子間のネットワークを作成した. その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/homepage/GNetwork.html に掲載した. 本手法で得たネットワークの信頼性を調査する為, 推定された我々のネットワークと,KEGG パスウェイとの比較を行った. 我々が使った 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. その中で我々の推定したネットワークの半順序が一致したものは 18 ペアだけであった. 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.

4　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の構築をこころみた. そして表現型の変異データに含まれる誤差に耐性のあるネットワークを構築するため, 変異した形態の包含関係に基づき, 形態変異間に半順序を定義した上で, 変異間距離を定義した. 距離およびネットワーク構築において距離の正規化パラメータおよび表現型のノイズ耐性を調整する閾値の 2 つのパラメータを定義し, 閾値決定の為に GO を利用し, 自動的に遺伝子ネットワークを推定した. しかし, 本手法で得たネットワークは既にわかっている KEGG パスウェイとはあまり一致していなかった. 本手法で推定されたネットワークに KEGG においてまだ発見されてないパスウェイが存在すると推測すると同時に,KEGG に存在したネットワークは本手法でも推定されるように本手法のネットワーク推定の精度を向上する必要がある. また, ネットワークの安定性についても, 調査し改善する必要がある. そして, 形態変異を距離化し, 数値に情報を落としたことで, どの形態が変異したかの情報が弱くなってしまったので, 形態変異の原因因子を求めることも難しくなった. 表現型による遺伝子ネットワーク推定では, 遺伝子に直接起因する形態変異も求めることが出来ることも利点なので, その点についてもさらに改良を加えたい.

参考文献

[1] Y.Ohya,*et.al High-dimensional and large-scale phenotyping of yeast mutants.* Proc. Natl. Acad. Sci. Vol. 102. No.52. 19015-19020, 2005

[2] T.Akutsu,S.Miyano, and S.Kuhara. *Inferring qualitative relations in genetic networks and metabolic pathways.* Bioinformatics,16,727-734 ,2000

[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.

[5] E.Segal,M.Shapira,A.Regev,D.Pe're,*et.al Module network:identifying regulatory modules and their condition-specific regulators.* Nature Genetics,34(2):166-176,2003.

[6] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang. *Probabilistic boolean networks:a rule based uncertainty model for gene regulatory networks.* Bioinformatics,18:261-274, 2002.