2009－BIO－16 (2)
2009／3／5

発現量データを用いた 相関係数によるタンパク質の複合的な相互作用の推定

村上 翔† 吉廣 卓哉‡ 井上 悦子‡ 中川 優‡

† 和歌山大学大学院システム工学研究科
‡ 和歌山大学システム工学部
〒640-8510 和歌山県和歌山市栄谷 930 番地

E-mail: † s091057@sys.wakayama-u.ac.jp ‡ {tac, etsuko, nakagawa}@ sys.wakayama-u.ac.jp

概要 本研究では，タンパク質の発現量データを用いて，タンパク質の複合的な相互作用を推定するデータマイニング手法を提案する．提案手法では，複数のタンパク質が複合体を作り別のタンパク質へ作用する相互作用モデルを想定し，相関係数を用いて相互作用するタンパク質の組合せを抽出する．ベイジアンネットワーク等の従来手法と比較して，サンプル数が少なくても適用可能である特徴がある．また，提案手法を実際のタンパク質発現量データに適用し有用性を評価する．

Predicting Combinatorial Interaction of Proteins using Correlation Coefficient from Protein Expression Data

Sho Murakami† Takuya Yoshihiro‡ Etsuko Inoue‡ Masaru Nakagawa‡

† Graduate School of Systems Engineering, Wakayama University
‡ Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama, 640-8510 Japan

E-mail: † s091057@sys.wakayama-u.ac.jp ‡ {tac, etsuko, nakagawa}@ sys.wakayama-u.ac.jp

Abstract. In this paper, we propose a data mining technique to retrieve combinatorial interactions of proteins from expression data of proteins. In our proposal, we suppose the interaction model that two or more proteins are unified into one complex protein and it acts on another protein, and we predict the interaction using correlation coefficient. Our method has a feature that we can apply it even if the number of available expression samples is not so large, compared with other existing methods such as Bayesian networks. We evaluate our method with real protein expression data and report about it.

1. はじめに

近年，ヒトゲノムプロジェクトに代表されるゲノム解読プロジェクトが完了し，ポストゲノム研究として，遺伝子やタンパク質の機能や，その複雑な相互作用の結果として生じる生命現象の解明を目指した研究が活盛んに行われている．中でもタンパク質全体としての作用や機能を解明する解析をプロテオーム解析と呼び，配列や立体構造など様々な視点からタンパク質の機能解明を行う研究が進んでいる．本研究ではこのうち，タンパク質の発現量を定量し，その定量データからタンパク質の機能を解明するアプローチ[1]を対象とし，タンパク質の発現量データから複合的なタンパク質の作用を推定することを目的としている．

発現量から複合的な相互作用を推定する手法としては，マイクロアレイによる遺伝子の発現定量データを用いる手法があり，この発現量データから遺伝子の複合的な相互作用を推定する手法が数多く提案されている．特にベイジアンネットワークを用いた推定手法[2][3]は，事象の発生確率に基づいて複数のタンパク質間の相互作用を推定できる手法として注目されている．ベイジアンネットワークは，遺伝子数が数千～数万と非常に多い場合にも比較的高速に計算可能であり，マイクロアレイのように遺伝子数，サンプル数ともに多くのデータを効率的に生成できる場合には有用である．

一方，タンパク質の発現定量にあたっては，各サンプルに対して２次元電気泳動を行い，この結果を画像解析して定量する方法が一般的である[1]．しかしこの方法では，定量できるタンパク質数が数百～数千と遺伝子に比べて少

なく，また実験に非常に手間がかかり，サンプル数を増やせないため，ベイジアンネットワーク等の既存手法の適用に向かない面がある．

本研究では，タンパク質の複合的な相互作用として，複数のタンパク質が複合体を作り，この複合体が他のタンパク質の発現量に影響する相互作用モデルを想定し，この相互作用を比較的少ないサンプル数のデータからでも推定できる手法を提案する．

2. 想定する相互作用モデル

生命活動は主にタンパク質の相互作用により維持されていると考えられているが，各タンパク質の相互作用は，タンパク質が単体で，或いは複合体を形成して，別のタンパク質分子に作用すると考えられている．

本研究では，複数のタンパク質が複合体を形成して別のタンパク質の発現量を促進，或いは抑制するモデルを想定する．このモデルの模式図を図1に示す．タンパク質A, Bは単体ではタンパク質Cに作用せず，Cの発現量に影響しないが，AとBが複合体を形成した場合には，これがCの発現量に影響するモデルである．本モデルでは，AとBが形成する複合体の数と C の分子数の間に何らかの関係があるはずである．本研究では，AとBの複合体数とCの分子数の相関係数をとり，この絶対値が1に近い場合に本モデルで示す相互作用があるものと推定する．

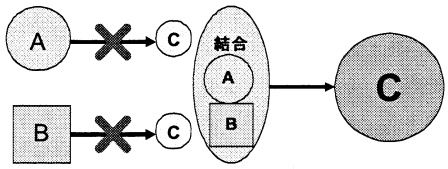

図1. 想定する相互作用モデル

3. 相互作用の推定手法

3.1. タンパク質の発現量データ

入力となるタンパク質の発現量データは，各サンプルに対して，含まれる各タンパク質の発現量が数値として表現されたものを想定する．

発現量データの例を表 2 に示す．各サンプルに対して，含まれる各タンパク質の発現量が数値として表わされている．一般的に，二次元電気泳動を用いる場合には，抽出できるタンパク質数は（生物種や部位にもよるが）数百～数千と言われており，また，実験は熟練を要するうえ手間もかかるため，サンプル数もせいぜい数十程度が限界になることが多い．この点で，マイクロアレイによる遺伝子発現量(数千～数万遺伝子，実験の手間も少ない)とは規模が異なる．また，タンパク質発現量データは，遺伝子発現量データと同様に，通常は何らかの正規化処理が行われた後に分析に適用されるが，正規化法については本稿の範囲外とする．

表2. タンパク質の発現量データ

サンプルID	タンパク質ID				
	1	2	3	4	…
1	0.003144	0.001562	0.001363	0.000572	…
2	0.005048	0.002316	0.001558	0.000781	…
3	0.00364	0.001842	0.00157	0.000656	…
4	0.005834	0.002258	0.001733	0.000837	…
5	0.005237	0.002325	0.001858	0.000876	…
6	0.001622	0.003075	0.002357	0.000505	…
：	：	：	：	：	

3.2. 相互作用推定手法のアイデア

提案する相互作用推定手法は，2 章で説明した相互作用モデルに基づき，タンパク質A,Bの複合体数とタンパク質Cの分子数の相関係数を計算し，高い値が得られたタンパク質の組み合わせを抽出するというものである．

複合体の数はタンパク質AとBの分子数の小さい方の値であると考える．図3に模式図を示す．タンパク質AとBの発現量が棒グラフで表わされている．単純に考えると，発現量に対する結合割合が1:1であれば，タンパク質AとBの結合量は，発現量の少ない方の値であると考える．(以後，この値をmin(A,B)と表記する．）実際にはタンパク質の種類により，結合状態の分子と非結合状態の分子が混在していると考えられるが，その場合にも結合状態の分子の量は濃度に依存した平衡状態にあるため，この値に比例した量になると考える．

このように複合体の数が推測されるため，min(A,B)とタンパク質Cの発現量の相関係数を計算することで，想定するモデルに基づいた相互作用を推定できる．相関係数を計算した結果，高い正の値が得られれば，タンパク質 A，B の結合体はタンパク質Cの発現量を促進すると言える．逆に高い負の値が得られればタンパク質AとBの結合体がタンパク質Cの発現を抑制していると言える．

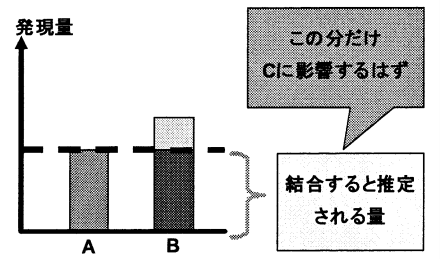

図3. あるサンプルのA,Bの発現量の棒グラフ

3.3. スケール差による問題と解決方法

3.2 節では相互作用推定手法のアイデアを述べたが，本

手法にはまだ問題があり，解決が必要である．それは，タンパク質の分子量を見積もるために発現量を用いるときの問題である．本節ではその解決方法を述べる．

タンパク質の発現量の測定基準にもよるが，例えば二次元電気泳動により定量した場合には，発現量は泳動画像中の各スポットの面積や容積(濃度の積分値)等を用いて数値化する．また，電気泳動結果の画像化にあたっては何らかの色素を用いており，この濃度をスキャナが認識することで画像化される．つまり，1 分子あたりの発現量はタンパク質によって異なることになる．よって，複数のタンパク質が関与する複合体の数を，単純に発現量の小さい方を用いて表現する時には，タンパク質により分子量に対する発現量の比（スケール）が異なる問題が発生する．図 4 はこの問題を説明した図であり，タンパク質 A と B で 1 分子あたりの発現量に差がある場合，必ずしも発現量の小さいタンパク質（この場合は A）が複合体数のボトルネックにならないことを表している．さらに，必ずしも 1 分子同士が結合して複合体を形成するわけではないこともこの問題の要因の一つである．

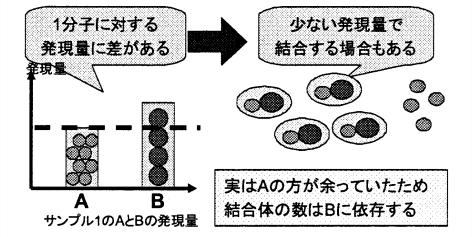

図 4. 発現量にスケール差がある問題

このスケール差の問題を解決するために，タンパク質 A の発現量のスケールを段階的に変化させてタンパク質 min(A,B)とタンパク質 C の相関係数を計算し，値が最大となったスケールを採用する手法をとる．これは，上記の状況では結局，複合体を形成するための A と B の発現量の比がわかれば十分であり，想定モデルのような相互作用があるのであれば，最も相関係数が大きくなるような発現量の比が求める値であると考えられるからである．

ここで，スケールを調整する範囲について考えてみる．複数存在するサンプルの各々に対して，N=B÷A を求める（A,B はそれぞれのタンパク質の発現量）．スケール比を k とおき，min(kA,B)について考えると，段階的に k を大きくしていった場合，k が各サンプルについて求めた N が最も小さい値 N_{min} より大きくなった時初めて B が採択される．また，k が N_{max} より大きくなってしまうと，全て A が採択される．これより，N_{min} と N_{max} の間でしか相関係数が変化しない．この範囲の中で，10 等分する様にスケールを選択し，それぞれのスケールにおいてタンパク質 A の発現量を調節しながらタンパク質 AB と C の相関係数を計算し，最大となったスケールを採用する．

3.4. 相互作用推定アルゴリズム

本節では，相互作用推定アルゴリズムの手順を改めて形式的にまとめる．タンパク質 i($1 \leqq i \leqq m$)，サンプル j($1 \leqq j \leqq n$)とおき，タンパク質 i の発現量を $e_i=(e_{i1}, e_{i2}, \ldots, e_{in})$ とベクトルにより表現する．タンパク質 a と b の発現量の小さい方をとった集合 min(a,b)の発現量を，$e_m=(e_{m1}, e_{m2}, \ldots, e_{mn})$ ($e_{mi}=\min(e_{ai}, e_{bi})$)と定義する．タンパク質 a と b の相関係数を $Cor(e_a,e_b)$ で表す．全てのタンパク質の中から，2 つの影響側タンパク質 a,b と，1 つの被影響側タンパク質 c を選ぶ全ての組み合わせについて，次の処理を行う．まず，$N_{min}=\min(e_{bj} / e_{aj})$, $N_{max}=\max(e_{bj} / e_{aj})$ ($1 \leqq j \leqq n$)を計算する．次に，$k_p= N_{min} + p(N_{max} - N_{min})/10$ ($0 \leqq p \leqq 10$)に対して，min(a,b)とタンパク質 c の相関係数，すなわち $Cor(k_p e_m, e_c)$ を計算し，その最大値を M とする．この計算を全ての a,b,c の組み合わせについて行った後，$Cor(e_a,e_c)$及び $Cor(e_b,e_c)$ がしきい値 T_1 以下であり，M がしきい値 T_2 以上であるものを抽出して出力する．しきい値 T_1, T_2 は，アルゴリズムの実行時に指定することとする．

以上のアルゴリズムにより，タンパク質 A,B の単体では C に影響せず，その複合体が C に影響するようなタンパク質 A,B,C が抽出できる．

4. 評価

4.1. 評価方法

提案手法を実際のタンパク質発現量データに適用することで評価を行った．適用データは，和歌山県地域結集型共同研究事業[4]により得られたウシのタンパク質発現量データを用いた．文献[1]に記載されているプロテオーム解析支援システムにより得られたものである．本データのサンプル数は 255，タンパク質数は 879 であり，適用にあたっては総インテンシティ正規化[5]を行い，タンパク質の総発現量に対する各タンパク質の発現量の割合を用いた．

実験にあたっては，提案アルゴリズムを C 言語により実装した．また，min(A,B)を計算するにあたり，発現量が小さい方の値として A または B のサンプルに選択が偏った結果は有用と判断できないため，片方への依存度が 3 割以下である組合せは抽出しないこととした．また相関係数は，はずれ値に影響されて大きく値が変動するため，相関係数の計算時に，2 つの各発現量ベクトルのいずれかに対して，発現量が $\pm 2.5\sigma$（σ は標準偏差）の範囲外であるサンプルははずれ値として扱い，相関係数の計算に用いなかった．また，データには欠損値が見られたため，相関係数の計算時にいずれかのデータが欠損しているサンプルの割合が 20%を超える場合には，その組合せは抽出しないこととし

た．また，組合せを抽出するためのしきい値は，それぞれ T_1=0.40, T_2=0.65 とした．

4.2. 結果と考察

データ適用の結果，抽出された組合せ数は 2696 であった．min(A,B)とタンパク質 C の相関係数の最大値は 0.78 であったが，得られた組合せを散布図表示したところ，明らかに期待した傾向が読み取れるものは存在しなかった．

ところで，抽出された組合せを散布図にして確認したところ，3 種類のパターンに分類できることが示唆された．得られたパターンの代表例を図 5 に示し，以下にその特徴を述べる．

パターン 1 : タンパク質 A-C の散布図（A を x 軸、C を y 軸にとる），B-C の散布図ではデータが広域に広がって分布しているが，min(A,B)-C の散布図ではある程度直線状に分布している．抽出したいデータパターンである．

パターン 2 : はずれ値と思われるサンプルの影響で A-C 或いは B-C の相関係数が低くなり，抽出されたパターンである．はずれ値の影響がなかった場合の相関係数が十分に小さければ，興味深いデータパターンである可能性がある．

パターン 3 : A-C または B-C の散布図がある程度直線状に分布し，一部のサンプルのみが外れた分布をするパターンである．この一部のサンプルが min(A,B)に選ばれない確率が高く，この分布は誤って抽出される危険性が高い．

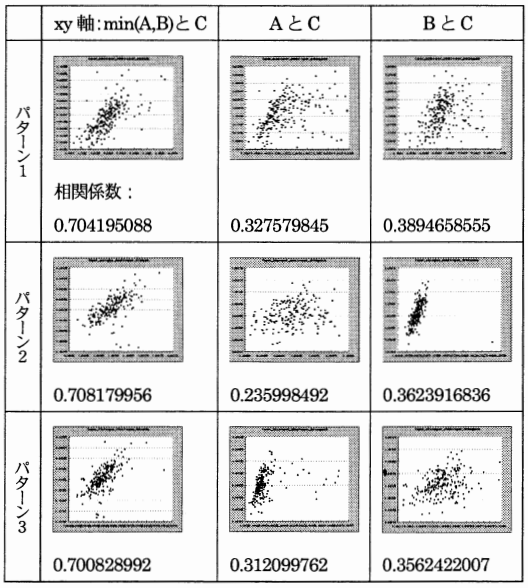

図 5. パターン別散布図例

図 6 に，min(A,B)と C の相関係数毎に抽出された組合せのヒストグラムと，そのパターンの内訳を示す．今回の実験では，min(A,B)と C の相関係数が 0.79 以上に該当する組合せは抽出されず，相関係数 0.72 以上の組み合わせもごく少数であった．このグラフから，有害なパターン 3 が数多く抽出されるために，興味深いパターン 1 の結果が効率よく抽出されていないことがわかる．また，はずれ値による悪影響が表れるパターン 2 がほとんど抽出されていないが，これは±2.5σのしきい値を用いてはずれ値を計算から除く処理を行った影響であろうと考えられる．

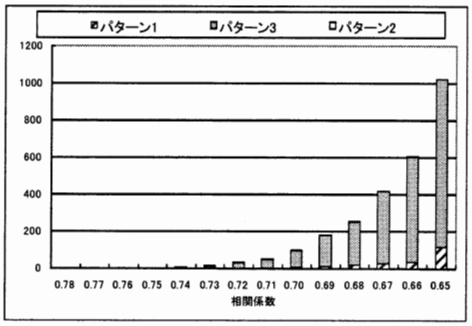

図 6. 抽出された組み合わせ数

5. おわりに

本稿では，タンパク質の発現量データから複合的な相互作用を推定する新たな手法を提案し，実データへの適用を通じて評価した．その結果，明らかに相互作用が読み取れる組合せは抽出できなかったものの，いくつかの組合せはその可能性を示唆するものであった．今後はパターン 3 の悪影響を除外する手法を考案すると同時に，興味深さを何らかの確率的指標で表現する手法など，相互作用の有無を客観的に判断する手法を確立したい．

参考文献

[1] 永井宏平，吉廣卓哉，井上悦子，池上春香，園陽平，川路英哉，小林直彦，松橋珠子，大谷健，森本康一，中川優，入谷明，松本和也，黒毛和種肥育牛の枝肉形質バイオマーカーの探索Ⅰ：大規模プロテオーム解析情報と血統・枝肉形質情報の統合情報管理システムの構築，日本畜産学会報，Vol.79, No.4, 2008.

[2] 玉田嘉紀，井本清哉，宮野悟，異種ゲノムデータの統合による遺伝子ネットワーク推定手法，統計数理，Vol. 54, No. 2, pp.333-356, 2006.

[3] 阿久津達也，バイオインフォマティクスの数理とアルゴリズム，共立出版，pp.183-186, 2007.

[4] 和歌山県地域結集型共同研究事業，http://www.wakayama-kessyu.com/ .

[5] John Quackenbush，マイクロアレイデータの正規化と変換，Nature Genetics – The Chipping Forecast II, Vol.32, pp.496-501, 2002.