Vol.2010-BIO-20 No.2
2010/3/4

発現量データからの相関係数による タンパク質間相互作用の推定手法

村上翔† 井上悦子†† 吉廣卓哉†† 中川優††

生命現象の仕組みを理解する方法の一つとして，タンパク質の相互作用の解析が盛んに行われている．しかし，発現量データを用いた 3 つ以上のタンパク質の複合的相互作用の推定は難しく，発展が望まれる研究課題の一つとなっている．本研究では，複数タンパク質が構成するタンパク質の複合体が別のタンパク質の発現量に影響する相互作用モデルに基づいて，相関係数を用いて，３つのタンパク質が集まったときに初めて現れる相乗的な相互作用が推定されるタンパク質の組合せを抽出する手法を提案する．提案手法を実際のタンパク質発現量データに適用し，統計的な分布に基づいて提案手法の有効性を検討する．

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

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

The interaction of proteins is actively analyzed as one of the methods of understanding the mechanism of the living creatures. However, prediction of combined interaction of three or more proteins that use the expression data is one of the research topics which is considered difficult. In this paper, we propose the technique for extracting the combination of the protein to which the interaction is predicted by using the correlation coefficient, and the technique based on the interaction model that the complex composed of two or more proteins influences the expression data of another protein. We evaluate the proposed method with real protein expression data and gine the result it based on statistical distribution.

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ための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，遺伝子を対象とする場合には，マイクロアレイによる発現定量データを用いることが多い．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試みが過去になされており，ベイジアンネットワーク[2][3][4]やブーリアンネットワーク[5]等数多くの手法が提案されている．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では例えば，各遺伝子の発現量を多・少の 2 段階，或いは多・中・少の 3 段階に離散化することで事象を定義し，単純な場合にはタンパク質 A の発現量が多の場合にタンパク質 B が多である確率，より複雑な場合には３以上の事象間での条件付き確率を計算し，これらの確率を用いて遺伝子間の相互作用ネットワークを推定する．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，また実験の手間がかかることからサンプル数を増やすことが困難で，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．

我々は過去に，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した[7]．この方法は複合体が他のたんぱく質に及ぼす影響の強さをスコア化して３タンパ

† 和歌山大学大学院システム工学研究科
Graduate School of Systems Engineering, Wakayama University
†† 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University

ク質間の複合的な相互作用を推定していたが，このスコアは２タンパク質間の相互作用の強さを含んでおり，そもそも２タンパク質間で相関が強いタンパク質が集まると相互作用が検出されやすい傾向があった．これに対して本論文では，統計処理を用いて３タンパク質が集まった場合にのみ現れる相乗的な影響のみを用いてスコア化することで，複合的な相互作用をより正確に検出する方法を提案する．本論文の構成は以下の通りである．2 章では本研究で想定するタンパク質の相互作用モデルを説明する．3 章では，このモデルに基づいた相互作用推定法を提案する．4 章では統計処理を用いて相乗的な相互作用効果をスコア化する手法について述べ，5 章ではこれを実データに適用することで本手法の評価を行う．最後に 6 章でまとめとする．

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．

ここで，本研究で想定するタンパク質の相互作用モデルついて説明する．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．A と B が影響を与えるタンパク質で，C が影響を受けるタンパク質である．この相互作用モデルでは，図 1 で示すようにタンパク質 A あるいは B 単体のタンパク質 C の発現量への影響はあるものの，タンパク質 A と B が複合体を形成した場合のタンパク質 C の発現量への影響の方が明らかに大きく，AとBの相乗的な作用が認められる場合である．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，タンパク質 A－C，B－C の分子数の間の相関係数に比べて，A と B の複合体数と C の分子数の相関係数が十分に大き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．

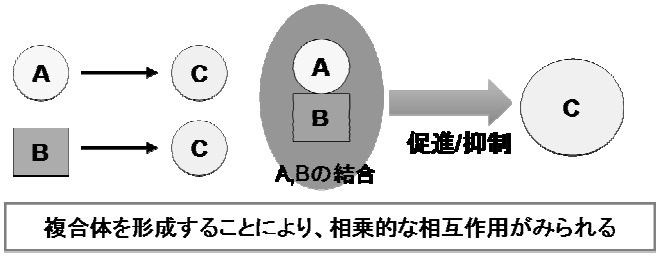

図 1　想定する相互作用モデル

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，二次元電気泳動などの生物学的な実験によって得られる．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．

二次元電気泳動によって得られたタンパク質の発現量データの例を表 1 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も多い．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．正規化法については本稿の範囲外とする．

表 1　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，単体の影響側タンパク質の作用の強さに比べて，２つの影響側タンパク質が集まった場合の相互作用の強さが十分に大きく，これらのタンパク質間に何らかの相乗効果が見られるような３つのタンパク質の組合せを抽出するものである．すなわち，タンパク質 A と C，タンパク質 B と C の発現量の相関係数を計算し，一方でタンパク質 A,B の複合体の量とタンパク質 C の発現量の相関係数を計算し，後者の相関係数が十分大きい場合に相乗効果が見られるとして，そのようなタンパク質の組み合わせを抽出する．相関係数は，2 つのデータ系列の相関を示す統計量で，絶対値が 1 に近いほど関係が強いことを示す．相関係数の値が正である時は正の相関，マイナスの値である時は負の相

関があり，0 の時は相関がないことを示す．

ここで，影響側タンパク質 A,B の発現量から，複合体の量を求める必要がある．本研究では，タンパク質 A と B は同時に存在する場合には必ず複合体を形成すると考え，タンパク質 A と B の発現量の小さい方の値を複合体の量であると考える．図 2 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られる．（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等に依存した平衡状態にあるため，この値にある程度比例した量になると考えられる．

このように複合体の量を推測し，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，2 つのタンパク質 A と B が複合体を形成し，別のタンパク質 C の発現量に影響を与える相互作用を推定することができる．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．

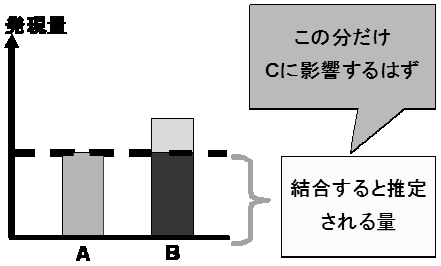

図 2　あるサンプルの A,B の発現量の棒グラフ

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)等を専用ソフトウェアにより計測して数値化する．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 3 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，複合体の数は発現量の小さいタンパク質（この場合は A）に依存するとは限らないことを表している．一見するとタンパク質 A の発現量の方が少ないため，タンパク質 C に影響を及ぼす結合体の数はタンパク質 A に依存するかのように思える．しかし，実際にはタンパク質 A と B で結合体に必要な発現量のスケールに差があるため，結果的に結合体の量はタンパク質 B に依存してしまう結果となっている．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．

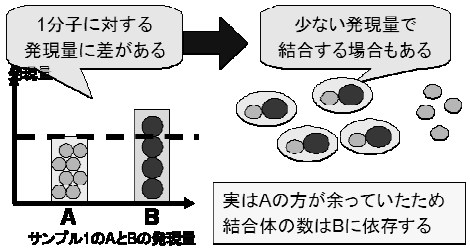

図 3　発現量にスケール差がある問題

このスケール差の問題を解決するために，一方のタンパク質のスケールを調整しつつ相関係数を計算する．具体的には，相関係数を計算する際にタンパク質 A の発現量のスケールを段階的に変化させてから 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，図 3 の状況で想定モデルのような相互作用があるのであれば，正解であるスケールにおいて min(A,B)と C の間に十分大きな相関関係が見られるはずと判断されるからである．逆に相互作用がないにもかかわらず，偶然に大きな相関関係が見られることは非常に稀であり，十分なサンプル数があればほとんど発生しないと考えられる．

ここで，スケールを調整する範囲について考えてみる．図 4 は段階的にタンパク質 A のスケールを大きくしていくうえで，スケールの調整を行う範囲を示した図である．縦軸がサンプル番号，横軸が発現量であり，三角のマーク（▲）がタンパク質 A，ひし形のマーク（◆）がタンパク質 B を表している．タンパク質 A を段階的に大きくしていったものである．①の状態のようにタンパク質 A とタンパク質 B のスケールに大きな差があった場合には全てタンパク質 A が採択される（つまり全てのサンプルにおいて B より A の発現量が小さい）．ここで，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，図の②の状態へと変化するが，こ

こでも min(kA,B)では全てタンパク質Aが採択される．さらに k を大きくして図の③の状態になると初めてタンパク質 B が採択される様になり，min(kA,B)と C の相関係数が変化する．さらに k を大きくしていくと，図の④の状態までは相関係数が変化するが，図の⑤の状態になると min(kA,B)において全てタンパク質 B が採択されてしまうため，min(kA,B)と C の相関係数が変化しないことがわかる．以上の考察より，スケールを調整する範囲を求める．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）．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 の最小値 Nmin より大きくなった時初めて B が採択される．また，k が N の最大値 Nmax より大きくなると，全てのサンプルに対して A が採択される．これより，Nmin と Nmax の間でしか相関係数が変化しない．この範囲の中で，q 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．

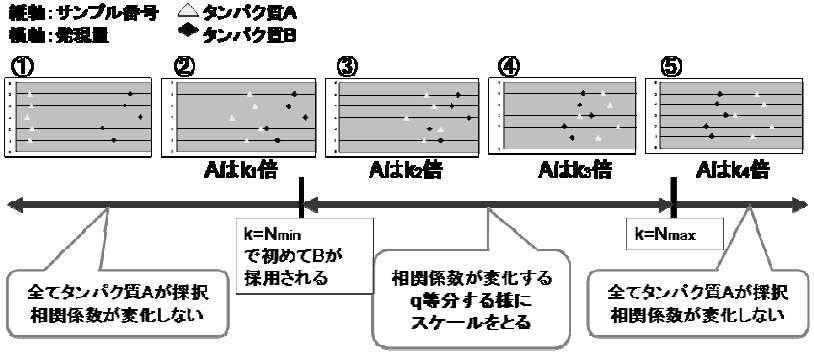

図 4　スケール調整を行う範囲

3.4 相互作用推定アルゴリズム

本節では，複合体と単体の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 $\mathrm{Cor}(e_a,e_b)$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，$k_p= N_{min} + p(N_{max} - N_{min})/m$ ($0 \leqq p \leqq q$，p は実数)に対して，min(a,b)とタンパク質 c の相関係数，すなわち $\mathrm{Cor}(k_p e_m, e_c)$を計算し，その最大値を相互作用スコア S_{abc} 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う．以上のアルゴリズムにより，複合体と単体の相互作用を測ることができる．

しかしながら，この S_{abc} は A と B の相互作用と A と C の相互作用，つまり 1 対 1 の相互作用の効果を含んでいるため S_{abc} の値が高くても相互作用があるとは言い切れない．1 対 1 の関係を取り除き，複合体と単体の間に相互作用があると判断するためのスコア z を求め，その値順にランキングを作成する． z スコアの求め方は第 4 章で説明する．

4. 相互作用の有無を判定する統計的指標

4.1 複合的な相互作用の検出

これまでに，複合体による相互作用を推定するための相互作用スコア S_{abc} を求めるアルゴリズムを示した．しかしながら，このスコアにはタンパク質 A と C，B と C の 1 対 1 の相互作用の効果を含んでおり，この効果によりスコア S_{abc} に影響が出ることが想定される．本研究で求めたいのは，1 対 1 の相互作用に比べて，複合体を形成したときの相互作用が十分に大きく，相乗的な相互作用が認められるような A と B, C の組合せである．実際に，相互作用がないと仮定した人工データを用いた計算機シミュレーションにより，1 対 1 の関係が強いほどスコア S_{abc} が高くなることを確認した．この結果を図 5 に示す．この図は，正規分布に従ったサンプル数 200 個の人工データ A，B，C について，A－C 間，B－C 間の相関係数を共に 0.2，0.3，0.4 と変化させた場合に 3.4 節で示したアルゴリズムを適用し，試行を 300 万回行った結果のスコア S_{abc} の分布である．この結果より，1 対 1 の相関係数が高いほど S_{abc} の値が高くなることが読み取れる．本節では，1 対 1 の効果と複合的な効果を分離する方法を説明する．

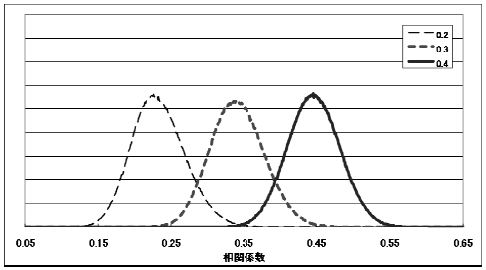

図 5　1 対 1 の相関係数を変化させた S_{abc} の分布

1 対 1 の効果と複合的な効果を分離するために，統計処理技術を用いる．タンパク

質の発現量が正規分布に従うと仮定する．本提案手法は相関係数を基礎としているので，タンパク質 A と B の 1 対 1 の相互作用の強さは，A と B の相関係数で測る．ここで, A と C の相関係数を α , B と C の相関係数を β とした時の S_{abc} の分布を考える．いくつかの α ， β の時の S_{abc} からランダムにサンプルを抽出し，S_{abc} が正規分布に従うかの検定を行った．仮説を「S_{abc} が正規分布ではない」とし，Jaque-Bera の検定を行った結果, 有意水準 5%で仮説を棄却できた．これより S_{abc} は正規分布でないとは言えないことが確認できた．従って，相互作用の有無を測る方法として S_{abc} の z スコアを用いる．z スコアは平均から標準偏差がどれくらい離れているかを表した数値である．分布の平均を $\mu_{\alpha\beta}$，標準偏差を $\sigma_{\alpha\beta}$ とおくと z スコアは

$$z = \frac{S_{abc} - \mu_{\alpha\beta}}{\sigma_{\alpha\beta}} \cdots (1)$$

と表わされる．z スコアが大きい程，低い有意水準で複合体と単体の間に相互作用がないとは言えないことになり，すなわち相互作用があることを示唆している．z スコアを用いることにより， S_{abc} がどれだけ起こりにくい値かを比較することができる．この原理に基づいて相互作用を示唆している可能性が高いもの順にランキングするために，統計分布に基づいて $\mu_{\alpha\beta}$ と $\sigma_{\alpha\beta}$ を決定する方法が必要である．

では，A と C，B と C の相関係数が与えられたときに，S_{abc} がどんな分布になるのかを考える．複合的な効果が存在しないと仮定した場合に，S_{abc} に影響する要因は，A, B, C の分布（つまり各平均値と標準偏差）と A－C，B－C 間の相関係数である．A と C，B と C の相関係数を固定した場合に，A, B, C の分布により S_{abc} がどのように変化するかを調査した．ここで，A と B の分布を固定して C の分布を変化させても S_{abc} の分布が変化しないことに注意したい．なぜなら，min(kA, B)と C の相関係数は，C の平均値や標準偏差を変化させても変化しないからである．よって，A と B の分布のみを考えればよい．さらに，提案アルゴリズムでは，変数 p を用いて A と B のスケールを変化させて最大値を求めている．つまり，B の平均が変化すれば，その分だけ分布が引き延ばされるように B の標準偏差を変化させることで，S_{abc} の分布が等しくなる．つまり，A と B の平均か分散のどちらかの影響のみを調べればよい．以上より，上記のアイデアを実現するためには，A と B の分散（または平均）の変化に応じて S_{abc} の分布がどのように変化するかを調べることが必要である．

4.2 相互作用スコアの分布

本節では，タンパク質 A と B の分布によって，相互作用スコア S_{abc} がどのように変化するかを調査した結果を示す．この影響は，理想的には数式を用いて理論的に議論すべきところであるが，本手法は不連続な min 関数を用いており，また A と B のスケール調整を行っているため，理論的な解析が難しい．そこで，計算機シミュレーションにより S_{abc} の分布の挙動を確認した．

まず，A と C の相関係数を α，B と C の相関係数を β と固定した条件下で，A の平均を変化させた時の S_{abc} の分布を調査した．図 6 は，分布 A と B の分散を 1 とし，B の平均を 10，A の平均を 10，20，30 と段階的に増加させた場合の S_{abc} の分布である．なお，この図は， $\alpha = \beta = 0.4 \sim 0.5$ とし，サンプル数を 200，試行を 1000 万回行った結果である．図 6 の分布をみると，平均が同一である時の S_{abc} の値が最も大きく，平均の差が広がるほど値が小さくなる分布になっていることが読み取れる． α， β や分散値を変化させて同様の実験を試みた結果，同様の傾向が読み取れた．

以上の結果より，相関係数 α と β を固定した場合には，A と B の分布が同一である場合に S_{abc} が最も高くなる分布となることがわかった．この結果から，タンパク質の組合せ ABC を処理するときには，相関係数が α と β で，かつ A と B の分布が等しい場合の分布を作成し，平均 $\mu_{\alpha\beta}$ と標準偏差 $\sigma_{\alpha\beta}$ を求めればよいことがわかる．求めた $\mu_{\alpha\beta}$ と $\sigma_{\alpha\beta}$ から(1)式より z スコアを算出し，z スコアが高いものから「A，B，C の複合的な相互作用がない」とは言えないことになり，より強く相互作用を示唆している．

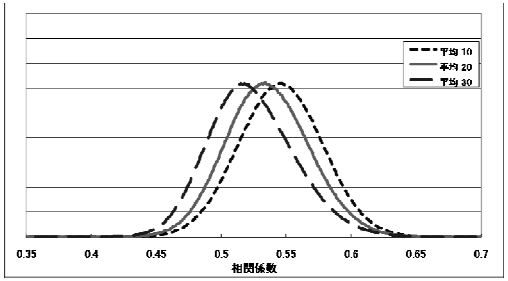

図 6　分布 A の平均の変化による S_{abc} の分布の変化

4.3 相互作用の有無を判定する z スコア算出表の作成

これまでに, 相互作用の有無を判定するための $\mu_{\alpha\beta}$ と $\sigma_{\alpha\beta}$ を決める方法を述べたが, この値を求めるためには，長時間の計算機シミュレーションが必要であり，タンパク質の組合せ全てに対して毎回計算することは現実的ではない．このため，予め様々な α と β の値に対して $\mu_{\alpha\beta}$ と $\sigma_{\alpha\beta}$ を計算しておき，z スコア算出表として用意しておく．z スコア算出表を計算する手順は以下のようになる．

z スコア算出表の計算手順

1. $0 < \alpha < 1$ の範囲を d 等分し， α_1， α_2， …, α_d を決める．同様に β_1， β_2, …， β_d を決める．
2. α_s と β_t（$1 \leqq s \leqq d, 1 \leqq t \leqq d$）の全ての組合せに対して 3 以下を実行する．

3. 平均，分散が同じ正規分布に従う分布 A,B,C を用意する．
4. Ａの発現量値をランダムに 3 つ選び入れ替る．ＡとＣの相関係数が上がればそのままとし，下がれば 2 つの発現量値を元に戻す．
5. ４を A と C の相関係数が α_s になるまで繰り返す．
6. Ｂについても 4，5 と同様にしてＢとＣの相関係数を β_t にする．
7. 提案手法に基づいて S_{abc} を求める．
8. 3～7 を十分な回数試行し，S_{abc} の分布を作成する．
9. 作成した分布の平均 $\mu_{s,t}$ と標準偏差 $\sigma_{s,t}$ を表データの α_s と β_t が対応する箇所の値とする．

上記の手順に従い，z スコア算出表を作成した．相関係数を 0.05 刻み（d=20）で計算機シミュレーションを 300 万回行うことにより作成した．この z スコア算出表を表 2 に示す．（上段が平均，下段が標準偏差である．）

作成された z スコア算出表と 3.4 節のアルゴリズムを実行することで求めた S_{abc} から z スコアを求め，ランキングを作成する．ランキングが上位のもの程「相互作用が無いとは言えない」タンパク質の組合せであり，より強く複合体と単体の相互作用の可能性を示唆しているタンパク質の組合せである．

表 2　z スコア算出表

B−Cの相関係数 \ A−Cの相関係数	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
0.05	0.069	0.101	0.143	0.188	0.233	0.280	0.326	0.373	0.418	0.467	0.514	0.562	0.609	0.657	0.705	0.753	0.800	0.850	0.898	0.955
	0.031	0.032	0.032	0.032	0.032	0.032	0.031	0.032	0.031	0.031	0.030	0.030	0.029	0.029	0.028	0.027	0.026	0.024	0.023	0.017
0.10		0.122	0.156	0.197	0.240	0.286	0.331	0.379	0.423	0.471	0.518	0.566	0.612	0.660	0.707	0.755	0.802	0.851	0.899	0.955
		0.034	0.034	0.033	0.033	0.032	0.031	0.032	0.030	0.031	0.030	0.030	0.029	0.028	0.027	0.027	0.026	0.024	0.023	0.017
0.15			0.178	0.211	0.249	0.294	0.337	0.384	0.428	0.475	0.521	0.569	0.616	0.663	0.710	0.757	0.804	0.852	0.899	0.954
			0.035	0.035	0.034	0.034	0.032	0.032	0.031	0.031	0.030	0.030	0.029	0.028	0.027	0.026	0.026	0.024	0.023	0.018
0.20				0.233	0.263	0.303	0.344	0.390	0.433	0.480	0.525	0.573	0.618	0.665	0.712	0.759	0.804	0.852	0.898	0.953
				0.036	0.036	0.035	0.033	0.033	0.031	0.031	0.030	0.030	0.029	0.028	0.027	0.026	0.026	0.024	0.023	0.018
0.25					0.285	0.318	0.354	0.397	0.439	0.485	0.529	0.578	0.622	0.668	0.714	0.761	0.806	0.852	0.898	0.951
					0.037	0.037	0.035	0.034	0.032	0.032	0.030	0.030	0.029	0.028	0.027	0.026	0.026	0.024	0.023	0.019
0.30						0.341	0.371	0.409	0.447	0.492	0.535	0.581	0.626	0.672	0.717	0.763	0.807	0.853	0.898	0.950
						0.037	0.036	0.035	0.033	0.032	0.030	0.030	0.028	0.028	0.027	0.026	0.025	0.024	0.023	0.019
0.35							0.392	0.423	0.457	0.499	0.541	0.586	0.630	0.675	0.720	0.765	0.809	0.854	0.898	0.948
							0.036	0.036	0.033	0.033	0.031	0.030	0.028	0.028	0.027	0.026	0.025	0.024	0.023	0.020
0.40								0.446	0.474	0.511	0.550	0.593	0.636	0.681	0.725	0.770	0.812	0.857	0.899	0.948
								0.035	0.034	0.033	0.031	0.030	0.028	0.027	0.026	0.025	0.025	0.023	0.022	0.020
0.45									0.493	0.524	0.559	0.600	0.641	0.685	0.728	0.772	0.814	0.858	0.899	0.946
									0.033	0.033	0.031	0.030	0.028	0.027	0.025	0.025	0.024	0.023	0.022	0.020
0.50										0.544	0.573	0.609	0.648	0.689	0.732	0.775	0.816	0.858	0.897	0.942
										0.032	0.030	0.030	0.028	0.027	0.026	0.025	0.024	0.024	0.024	0.023
0.55											0.593	0.622	0.657	0.696	0.737	0.779	0.819	0.859	0.897	0.939
											0.029	0.029	0.027	0.026	0.025	0.024	0.024	0.023	0.024	0.025
0.60												0.642	0.669	0.704	0.743	0.784	0.822	0.861	0.897	0.935
												0.028	0.027	0.026	0.025	0.024	0.024	0.024	0.025	0.026
0.65													0.688	0.716	0.751	0.790	0.827	0.865	0.898	0.933
													0.025	0.024	0.024	0.023	0.023	0.023	0.025	0.028
0.70														0.733	0.760	0.796	0.832	0.867	0.898	0.927
														0.023	0.022	0.022	0.022	0.022	0.024	0.028
0.75															0.778	0.806	0.838	0.871	0.898	0.923
															0.020	0.020	0.021	0.022	0.025	0.027
0.80																0.822	0.846	0.876	0.900	0.924
																0.018	0.019	0.020	0.021	0.020
0.85																	0.860	0.884	0.907	0.933
																	0.017	0.018	0.018	0.015
0.90																		0.900	0.922	0.949
																		0.016	0.015	0.011
0.95																			0.942	0.969
																			0.012	0.009
1.00																				0.996
																				0.002

※この表は対角線に対して対象であるため，左下の値が空欄となっている

5. 評価

5.1 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[8]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．得られたデータは実験誤差が生じることがあるため，同一サンプルにつき複数回実験を行うことがある．複数回実験を行ったサンプルの発現量について再現性の確認を行い，再現性のあるデータのみ複数回の実験データの平均値を発現量データとして分析に用いた．本研究で用いたデータのサンプル数は 195，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[9]を行なったものを用いた．総インテンシティ正規化とは，1 つのサンプル中に含まれている全てのスポットの面積を合計した値で正規化を行った上で，各タンパク質のスポットの面積がその内でどの程度の割合を占めているかによって発現量を定量化するものである．つまり，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のデータである．

また，標準偏差 3 つ以上離れているデータをはずれ値とし，はずれ値を除去したタンパク質の発現量データが正規分布に従ったものかどうかの確認を行った．仮説を「発現量データは正規分布ではない」とし，Jaque-Bera の検定を行った結果，有意水準 5%で仮説を棄却できたタンパク質は，発現量が正規分布に従う可能性が示唆される．Jarque-Bera 検定とは歪度と尖度から計算する正規性の検定方法である．これより発現量データは正規分布でないとは言えないと判断できたタンパク質は半数以上の 454 個であった．

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組合せ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破棄することとした．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±2.5 σ（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，その組合せは破棄することとした．A と B のスケール調整は 10 段階で行った．すなわち，q=10 とした．

5.2 結果と考察

実データ適用し，z スコアによるランキングを行った結果，複合体による相互作用があることを示唆するタンパク質の組合せが多数見つかった．本節ではこの分析結果

Vol.2010-BIO-20 No.2
2010/3/4

の詳細について述べる．

全てのタンパク質を適用した結果（以降結果１と呼ぶ）と，正規分布に従うタンパク質（前述の Jaque-Bera の検定により，正規分布ではないと言えないと判断されたタンパク質）を適用した結果（以降結果２と呼ぶ）をヒストグラム化した．これを図７，図 8 に示す．ともに横軸に z スコアの階級をとり，縦軸に組合せ数をとった，z スコアの階級毎の組合せ数のヒストグラムである．図７の全てのタンパク質を用いた結果１では，正規分布に従わない分布を持つタンパク質を含む組合せ（抽出したくないパターン）が上位を占め，抽出したい組合せが上位に来るのを阻んでいることがわかる．一方で，正規分布に従ったデータのみを用いた結果２では，分布の差異によって抽出したくないパターンが上位に来ることなく，効率よく抽出したい組合せを抽出できていた．

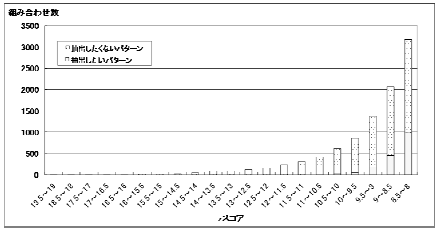

図７　結果１のヒストグラム

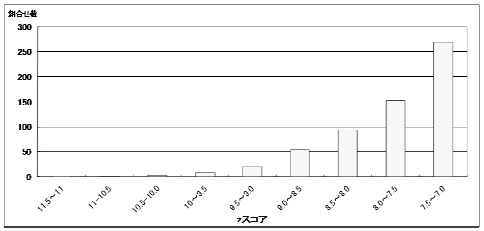

図８　結果２のヒストグラム

では，正規分布に従わない分布を持つタンパク質による，抽出したくないパターンについて説明する．上位の組合せについて散布図を作成してみたところ，結果１では抽出したくないものが見られた．このような散布図の例を図 9 に示す．縦軸がタンパク質 C の発現量をとり，横軸にタンパク質 A，B，min(A,B)の発現量をとったものである．min(A,C)とタンパク質 C の散布図が▲で表されている．なお，□がタンパク質 A と C の散布図であり，○がタンパク質 B と C の関係である．また，直線は min(A,B)と C の回帰直線である．この図を見るとわかるように，min(A,C)とタンパク質 C の分布（△で表されている）は，ある程度一直線上に並んでおり相関係数が高いことがわかる．しかし，タンパク質 B と C である○を見てもほとんどのサンプルが一直線上に並んでいる．ほんの一部のサンプルが直線から離れたところにあるため，タンパク質 B と C の相関係数が本来よりも低く計算されたため，複合的な相互作用があると判定されてしまったと考えられる．つまり，実際には 1 対 1 の関係において相互作用が見られるが，一部のサンプルがあるために相互作用が低く見積もられた例であり，このようなパターンでは複合体による相互作用があると判断できない．このようなパターンを，図 7 では抽出したくないパターンと呼んでいる．

一方抽出したいパターンを図 10 に示した. 図では 1 対 1 の関係である■や○は広域に広がっているが，複合体である min(A,B)と C の△ではある程度一直線上になっていることが読み取れる．

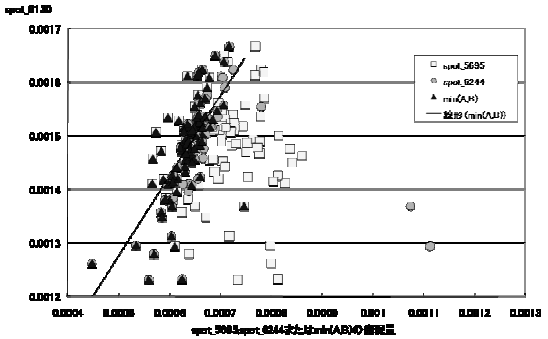

図 9　抽出したくないパターンの散布図例

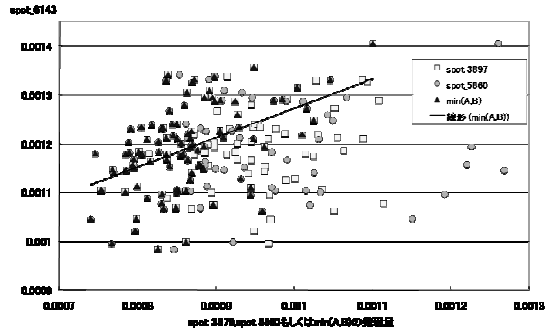

図 10　抽出したいパターンの散布図例

図 11 は図 8 と同様に縦軸と横軸をとり, 今回用いた実データと同じタンパク質数の正規分布に従った人工データを用いた場合に，z スコア毎の期待値を表したグラフである．図 8 と図 11 を比べると，図 11 では 6 以上のどの z スコアの階級においても期待値は 1 を下回る非常に低い値となっているが，図 8 を見ると，実データを適用した結果では多くのタンパク質の組合せが抽出されていることがわかる．人工データを用いた結果である図 11 は，「複合体と単体の間に相互作用がない」としたときに抽出される組合せ数の期待値を表している．実データを適用したときにはこの場合よりもは

Vol.2010-BIO-20 No.2
2010/3/4

るかに多くの組合せが抽出されたことから，実際のデータにおいては，複合体による相互作用の影響が強く見られることが示唆される．

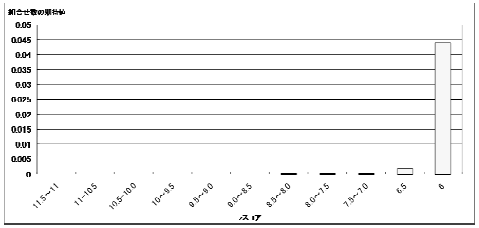

図 11　実データと同じ試行回数の z スコア毎の期待値

また，我々が以前提案した分析手法[7]では抽出することができなかった，相関係数が低くても相互作用があることを示唆するタンパク質の組合せも抽出できた．以前提案した手法は 1 対 1 の相関係数が 0.4 以下，かつ複合体と単体の相関係数が 0.65 以上のタンパク質の組合せを抽出するというものである．表 3 は正規分布に従った実データを適用したときの結果を z スコアが高い順にランキングしたものである．以前提案した手法では 3 位や 5 位などの２つのタンパク質の相関が低い組合せは抽出されないことから，本論文の提案手法によって，統計的な根拠に従って以前よりも精度の高い組合せの抽出が可能になったことがわかる．

表 3　正規分布に従った実データを適用したときの z スコアランキング表（一部抜粋）

順位	A（スポット番号）	B（スポット番号）	C（スポット番号）	Sabc	Cor(A,C)	Cor(B,C)	z
1	5148	6239	4470	0.674	0.092	0.317	11.02
2	6154	6239	5418	0.674	0.071	0.339	11.01
3	2572	4292	6239	0.561	0.137	0.293	10.94
4	5146	6239	1468	0.728	0.173	0.371	10.29
5	5661	6281	5342	0.504	−0.007	0.390	10.20
6	5146	6239	4478	0.648	0.089	0.315	10.19
7	5661	6281	5730	0.613	0.058	0.434	10.17
8	5026	6239	1333	0.560	0.052	0.360	10.15
9	5026	6239	3626	0.470	0.029	0.314	10.14
10	5695	6143	6042	0.640	0.148	0.444	10.12

6. 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，複合体と単体の相互作用を示唆するタンパク質の組合せを抽出することができた．今後は相互作用が示唆されたタンパク質の組合せの中に確認されているタンパク質間の相互作用が含まれていないかを確認し，本手法の実用性を裏付けたい．

謝辞　本研究の一部は生研センターイノベーション創出基礎的研究推進事業の支援により実施されたものである．

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) S. Imoto, T. Goto and S. Miyano, “Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression,Pacific Symposium on Biocomputing,7,175-186,2002.
5) T. Akutsu, a, S. Kuhara, b, O. Maruyama c and S. Miyano, Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model Theoretical Computer Science 298, 235-251,2003.
6)　S. Imoto, T. Goto and S. Miyano, “Estimation of genetic networks by strategic geno disruptions and gene overexpressions under a
7)　村上翔, 吉廣卓哉, 井上悦子, 中川優, 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, 情報処理学会研究報告（バイオ情報学）, 2009-BIO-16, pp.5-8, 2009.
8)　和歌山県地域結集型共同研究事業, http://www.wakayama-kessyu.com/
9)　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.