

条件付き確率及び統計的指標を用いた 発現量データからの 組合せ的なタンパク質相互作用の推定

藤木生聖[†] 井上悦子^{††} 吉廣卓哉^{††} 中川優^{††}

本研究では、タンパク質や遺伝子発現量データを用いて、タンパク質間または遺伝子間の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、発現している複数タンパク質の組合せが別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、条件付き確率を用いて、3つのタンパク質が集まったときに初めて現れる組合せ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。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。

Prediction of Combinatorial Protein Interaction from Expression Data Using Conditional Probability and Statistical Indicator

Takatoshi Fujiki[†] Etsuko Inoue^{††} Takuya Yoshihiro^{††} and
Masaru Nakagawa

This paper proposes a data mining method for predicting the interaction between three proteins (or genes) from protein (or gene) expression data. We assume the model that the total effect of three proteins A and B on C is far larger than the sole effects between two proteins A-C and B-C. In this method, we retrieve the combinations of three proteins (or genes) including the combinatorial effect using the conditional probability. We apply this method to real protein expression data, and examine the effectiveness of this method 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、ポストゲノムの研究として、遺伝子やタンパク質の機能の特定や、遺伝子やタンパク質間で起こる相互作用によって生じる生命現象の解明を目指した研究が盛んに行われている。その中でも、プロテオーム解析はそのようなポストゲノム研究の一分野として数えられる。プロテオームとは生体内に存在するタンパク質の総体である。プロテオーム解析とは、タンパク質全体を網羅的に解析することである。生体内のタンパク質の相互作用やその機能の全体像を明らかにすることで、生体を構成する有機システムの全体を解明しようとする研究が行われている。

プロテオーム解析の中でも、タンパク質の相互作用の解明は代表的なテーマの一つであり、タンパク質のアミノ酸配列や文献データベースなど様々なデータに対して、単に統計的手法を用いるだけでなく、情報分野の手法を応用することでタンパク質の相互作用と機能の解明を行う研究が進んでいる。近年では、プロテオーム分析機器や技術の発展により、タンパク質の発現量を精密に計測することができるようになってきた。遺伝子の発現量とは異なり、タンパク質の発現量を測定する方法は代表的なものはいくつか存在し、複数の方法で発現量データが得られる。しかし、解析手法としては、マイクロアレイを用いた遺伝子発現量データ等に適用できたベイジアンネットワークなどの手法をそのまま適用することが可能となっている。

発現量から複合的な相互作用の推定する手法として、マイクロアレイによって得られる遺伝子の発現量を用いてネットワークの推定を行う解析が先行している。この遺伝子における相互作用ネットワークを推定する既存手法として、ベイジアンネットワーク[1][2][3]やブーリアンネットワーク[4]など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、発現量を用いた遺伝子のネットワーク推定における代表的な解析手法となっている。この手法では例えば、各タンパク質の発現量を多・少2段階、あるいは多・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の発現量が多い場合にタンパク質Bが多である割合を計算し、この割合がタンパク質間に相関関係があるといえる数値になっていれば、このタンパク質間に相関があるとしてタンパク質ノード間にリンクを生成する。これによって発現量を測定した全ての遺伝子間のリンクを生成でき、その結果、遺伝子間の相互作用ネットワークを推定することができる。また、タンパク質Aが少、Bが少

[†] 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

の時に C が多である割合、のように 3 つ以上のタンパク質が関係する場合にも適用できる。しかし、この手法では、タンパク質 A や B が単体で C に作用しているときに、A と B を組み合わせた場合の C への作用も大きく見積もられてしまう欠点がある。

よって、本研究では複数タンパク質間や複数遺伝子間の相互作用を統計処理と条件付き確率を用いて単体で作用が強い関係を取り除き、複数タンパク質や複数遺伝子が集まった場合にのみ現れる相互作用の推定とその相互作用をスコアリングする方法を提案する。

本論文の構成を次に示す。2 章では本研究で想定するタンパク質や遺伝子の相互作用モデルや統計的指標の説明をしたうえで、組合せ的な相互作用の推定手法を提案する。3 章では提案手法の実装とこれを実データに適用することで本手法の評価おこない、結果について考察する。最後に 4 章でまとめるとする。

2. 組合せ的なタンパク質相互作用の推定手法

2.1 本研究で扱う発現量データの形式

入力となる発現量データは例えば、二次元電気泳動や液体クロマトグラフィなどの生物学的な実験によって得られるタンパク質の発現量データや、マイクロアレイなどで測定された遺伝子の発現量データなどを想定する。

二次元電気泳動やマイクロアレイ等によって得られた発現量データの例を表 1 に示す。スポット ID は発現量解析によって分離された各スポットに割り振られる ID である。サンプル ID は発現量解析の実験対象となったサンプルごとに割り振られる ID である。スポット ID より下の行からは、各サンプルにおけるスポットの発現量データが実数で示されている。なお、表中にある BA05C19A0521 のスポット spot_3316 の発現量が空欄になっているが、これは欠損値である。欠損値は発現量が極端に小さい場合や、二次元電気泳動による分離が不十分でスポットを検出できなかった場合などに発生する。また、発現量データには一般的に何らかの正規化を行うことが一般的である。これは、発現量の測定実験には誤差が生じやすいため、その実験誤差を補正するために行われる。

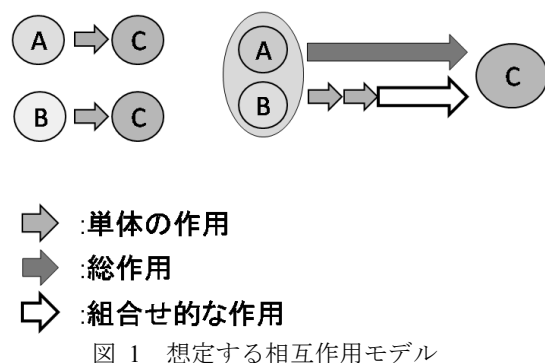
表 1 タンパク質の発現量データ

サンプルID	スポットID				
	spot_3311	spot_3316	spot_3321	spot_3328	...
BA05C19A0517	0.000582	0.000107	0.000338	0.000451	...
BA05C19A0521	0.000563		0.000475	0.000458	...
BA05C19A0522	0.000495	0.000126	0.000433	0.000565	...
BA05C19A0525	0.000553	0.000153	0.000382	0.000486	...
BA05C19A0526	0.000536	0.000134	0.000536	0.000471	...
BA05C19A0527	0.000601	0.000185	0.000457	0.000513	...
:	:	:	:	:	:

2.2 想定する組合せ的なタンパク質相互作用モデル

タンパク質は生命現象のあらゆる側面に関わっている。例えば、生物体の構造を形成や代謝などの化学反応を起こさせる触媒としての働き、生体内の情報のやり取りに関与する受容体や細胞内シグナル伝達などがある。タンパク質には単体で機能するものもあるが、先に挙げた例のように多くのタンパク質は他のタンパク質と相互作用することでその機能を果たすと考えられている。そこで本研究においては、タンパク質の相互作用をモデル化し、そのモデルに基づいて相互作用を推定する手法を提案・評価する。

ここで、本研究で想定するタンパク質や遺伝子の相互作用のモデルについて説明する。以後の説明にはタンパク質を例に説明する。本研究では、例えばタンパク質 A, B, C においてタンパク質 A, B が同時に発現している時にタンパク質 C の発現量に与える影響が、タンパク質 A や B がそれぞれ単体でタンパク質 C の発現量に与える影響に対して遥かに大きいモデルを想定する。このモデルを図 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。このモデルでは、相互作用の関係が 3 つある。1 つ目の関係は**単体の作用**である。これはタンパク質 A, B がそれぞれ単体で C の発現量に影響を与える関係である。2 つ目の関係は**総作用**である。これは A と B が共に発現している時に C の発現量に影響を与える関係であり、この中には A と C, B と C による単体の作用を含んでいる。そして 3 つ目の関係は、総作用の関係から単体の作用の関係を除外した**組合せ的な作用**である。すなわち、A, B が共に発現しているという組合せのときにだけ C に影響を与える関係である。本研究で抽出したい関係は組合せ的な作用の大きい関係である。組合せ的な作用を求めることで、単体の作用と総作用だけでは求めることができないタンパク質の組合せ的な相互作用の関係を求めることができる。この組合せ的な作用を測定するためには総作用を測定し、そこから単作用の影響分を差し引く必要がある。



2.3 単作用と総作用の測定

本節では、前節で述べた組合せ的作用を測定するために、単体の作用と総作用の測定方法について述べる。単体の作用と総作用の度合いを測定するために、本研究では条件付き確率を用いる。条件付き確率を用いた測定方法について図 2 のベン図を用いて説明する。図 2 の A, B, C はそれぞれタンパク質 A, B, C が「発現している」（つまり発現量が多い）サンプルの集合とする。発現の定義は後に与える。まず、A-C や B-C の単体の作用の度合いについて述べる。A が C に与える単体の作用の度合いは、A の発現量が多いサンプル数に対する、A と C の両タンパク質の発現量が共に多いサンプル数の割合で表される。また、A と B が C に与える総作用の度合いについても同様の計算で求められる。つまり、A と B の両タンパク質の発現量が多いサンプル数に対する、A と B と C すべてのタンパク質の発現量が多いサンプル数の割合で表される。

ここで、上記の問題に対していくつかの定義を行う。まず、タンパク質の発現の事象を定義するために、タンパク質を $i(1 \leq i \leq I)$ 、サンプルを $j(1 \leq j \leq J)$ で表す。そして、タンパク質が発現しているかを判定するための閾値 r を用意する。サンプル j のタンパク質 i の発現量が全サンプルのタンパク質 i の発現量の上位 $r\%$ に含まれているとき、「サンプル j においてタンパク質 i が発現している」と定義する。次に単体の作用と総作用の測定方法を定義する。 $\text{num}(A)$ をタンパク質 A が発現しているサンプル数とする。また、 $\text{num}(A \cap B)$ をタンパク質 A と B が共に発現しているサンプル数とする。そして、タンパク質 A が単体で C の発現量に与える単体の作用の度合いを単作用度とし、 $E_A^C = \frac{\text{num}(A \cap C)}{\text{num}(A)}$ と定義する。同様に B が単体で C の発現量に与える単

作用度を $E_B^C = \frac{\text{num}(B \cap C)}{\text{num}(B)}$ と定義する。そして、A と B が C の発現量に与える総作用の度合いを総作用度とし、 $E_{A,B}^C = \frac{\text{num}(A \cap B \cap C)}{\text{num}(A \cap B)}$ と定義する。

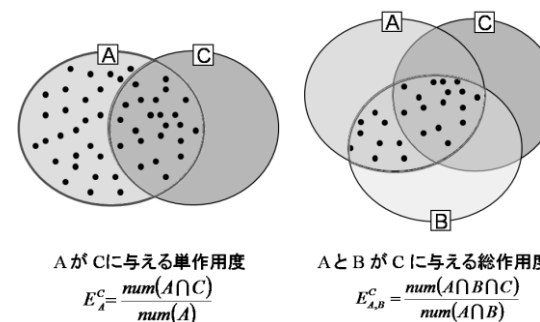


図 2 単作用度と総作用度の測定方法

2.4 組合せ的作用の測定方法

本研究で求めたいものはタンパク質 A と B, C による組合せ的作用が大きい時のタンパク質の組合せである。組合せ的作用度は総作用度と単作用度から求めることができる。図 3 に示すように、組合せ的作用度は総作用度から単作用度を差し引くことで測定できるが、単純に引くだけでは求めることはできない。それは、単作用度は二つのタンパク質間の作用、総作用度は三つのタンパク質間の作用を表すため同列として扱えないからである。そこで本研究では組合せ的作用度を求めるために、A と B, C の間に組合せ的作用がないと仮定した総作用度 $E_{A,B}^{'C}$ を定義し、コンピュータシミュレーションによって $E_{A,B}^{'C}$ の統計的分布を計算する。そして $E_{A,B}^{'C}$ の統計的分布と総作用度 $E_{A,B}^C$ から統計的な生起確率として組合せ的作用度を求める。

コンピュータシミュレーションによって $E_{A,B}^{'C}$ の統計的分布を計算する方法について述べる。まず、入力データである発現量データから影響側のタンパク質を二つ、被影響側のタンパク質を一つ選び、それらタンパク質を順に A, B, C とする。そして A が単体で C に与える単作用度を α , B が単体で C に与える単作用度を β とする。次に、単作用度が α , β となるように人工的に作り出した発現量データを用意しこの時の総作用度を求める。ここで、人工データはタンパク質の発現量がランダムになっているため、この時求めた総作用度は組合せ的作用が含まれていない総作用度 $E_{A,B}^{'C}$ である。そして人工データの作成と組合せ的作用が含まれていない総作用度 $E_{A,B}^{'C}$ の測定を十分多くの回数行うことで、 $E_{A,B}^{'C}$ の分布を求める。次に、求めた $E_{A,B}^{'C}$ の分布から、z ス

コアを用いることで組合せ的な作用を測定する. z スコア $z_{A,B}^C$ を $z_{A,B}^C = \frac{E_{A,B}^C - \mu}{\sigma}$ と定義する. ここで, $E_{A,B}^C$ は実際の発現量データから得られるタンパク質 A と B, C による総作用度で, μ と σ はコンピュータシミュレーションによって得られた $E_{A,B}^C$ の分布の平均と標準偏差である. すなわち, z スコア $z_{A,B}^C$ は $E_{A,B}^C$ の平均 μ から $E_{A,B}^C$ がどれだけ離れたところに分布しているかを $E_{A,B}^C$ の標準偏差 σ を単位として表した値である.

z スコアを用いることで, α , β によって $E_{A,B}^C$ の分布が異なっている場合でも, 各分布の平均と標準偏差によって標準化することで, $E_{A,B}^C$ が統計的に求めた分布 $E_{A,B}^C$ 上でどれだけ現れやすいかを比較できるようになる. また z スコアを計算するためには正確には $E_{A,B}^C$ の分布が正規分布である必要があるが, 分布の全体像がおおよそ対象で単峰である場合, 関係なく使用しても構わないとされている. そして, 様々な α , β に対して $E_{A,B}^C$ の分布をコンピュータシミュレーションによって求めた結果, α , β の値によらず, $E_{A,B}^C$ の分布が上記の条件を満たしていると判断できた. シミュレーション結果の $E_{A,B}^C$ の分布の一例を図 4 に示す. 図 4 の横軸は $E_{A,B}^C$, 縦軸は出現頻度であり, 典型的な単峰の分布になっている.

また, この z スコアは絶対値が大きいほど平均から離れるので, その分だけ出現する確率が下がることを意味する. このことについて正規分布を例に考えてみる. 図 5 と表 2 で表されるように, 正規分布の平均を挟んで $\pm 1\sigma$ によって囲まれる面積は全体の 68.27%, 同様に $\pm 2\sigma$, $\pm 3\sigma$ で囲まれる面積はそれぞれ全体の 95.45%, 99.73% である. つまり平均からの距離が標準偏差の 1 倍以上, 2 倍以上, 3 倍以上となると, 統計的にはそれぞれ片側で 15.87%, 2.275%, 0.135% の確率でしか起こりえない. つまり z スコアの値が大きいほど統計的には出にくい値である. 言い換えると, z スコアが大きいほど組合せ的な作用が強く現れていると言える.

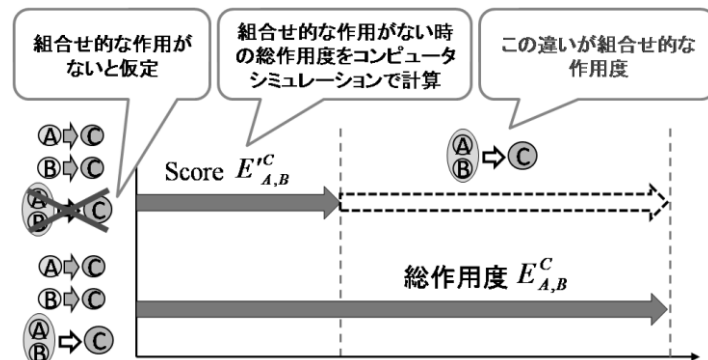


図 3 総作用から単体の作用と組合せ的な作用の分離

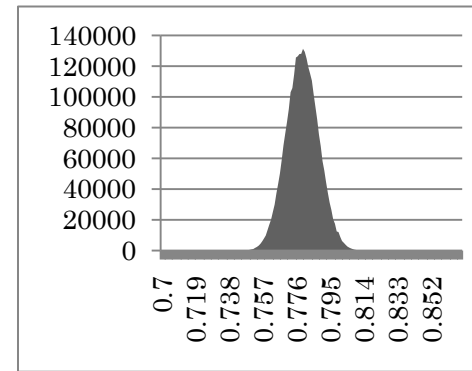


図 4 $E_{A,B}^C$ の分布（一例）

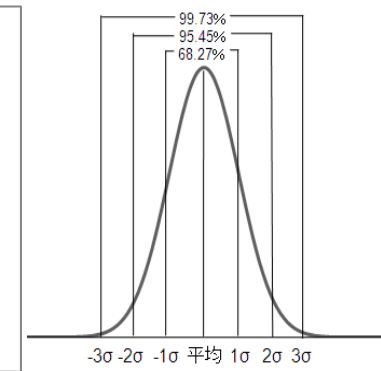


図 5 正規分布と確率

表 2 標準正規分布確率表

z	0	1	2	3	4	5	6	7	8	9
0.0	0.000000	0.007979	0.015957	0.023933	0.031907	0.039878	0.047844	0.055806	0.063763	0.071713
0.1	0.079656	0.087591	0.095517	0.103434	0.111340	0.119235	0.127119	0.134990	0.142847	0.150691
0.2	0.158519	0.166332	0.174129	0.181908	0.189670	0.197413	0.205136	0.212840	0.220522	0.228184
0.3	0.235823	0.243439	0.251032	0.258600	0.266143	0.273661	0.281153	0.288618	0.296055	0.303463
0.4	0.310843	0.318194	0.325515	0.332804	0.340063	0.347290	0.354484	0.361645	0.368773	0.375866
0.5	0.382925	0.389949	0.396936	0.403888	0.410803	0.417681	0.424521	0.431322	0.438085	0.444809
0.6	0.451494	0.458138	0.464742	0.471305	0.477827	0.484308	0.490746	0.497142	0.503496	0.509806
0.7	0.516073	0.522296	0.528475	0.534610	0.540700	0.546745	0.552745	0.558700	0.564609	0.570472
0.8	0.576289	0.582060	0.587784	0.593461	0.599092	0.604675	0.610211	0.615700	0.621141	0.626534
0.9	0.631880	0.637177	0.642427	0.647629	0.652782	0.657888	0.662945	0.667954	0.672914	0.677826
1.0	0.682689	0.687505	0.692272	0.696990	0.701660	0.706282	0.710855	0.715381	0.719858	0.724287
1.1	0.728668	0.733001	0.737286	0.741524	0.745714	0.749856	0.753951	0.757999	0.762000	0.765954
1.2	0.769861	0.773721	0.777535	0.781303	0.785025	0.788700	0.792331	0.795915	0.799455	0.802949
1.3	0.806399	0.809804	0.813165	0.816482	0.819755	0.822984	0.826170	0.829313	0.832413	0.835471
1.4	0.838487	0.841460	0.844392	0.847283	0.850133	0.852941	0.855710	0.858438	0.861127	0.863776
1.5	0.866386	0.868957	0.871489	0.873983	0.876440	0.878858	0.881240	0.883585	0.885893	0.888165
1.6	0.890401	0.892602	0.894768	0.896899	0.898995	0.901057	0.903086	0.905081	0.907043	0.908972
1.7	0.910869	0.912734	0.914568	0.916370	0.918141	0.919882	0.921592	0.923273	0.924924	0.926546
1.8	0.928139	0.929704	0.931241	0.932750	0.934232	0.935686	0.937114	0.938516	0.939892	0.941242
1.9	0.942567	0.943867	0.945142	0.946393	0.947620	0.948824	0.950004	0.951162	0.952296	0.953409
2.0	0.954500	0.955569	0.956617	0.957643	0.958650	0.959636	0.960601	0.961548	0.962474	0.963382
2.1	0.964271	0.965142	0.965994	0.966828	0.967645	0.968445	0.969227	0.969993	0.970743	0.971476
2.2	0.972193	0.972895	0.973581	0.974253	0.974909	0.975551	0.976179	0.976792	0.977392	0.977979
2.3	0.978552	0.979112	0.979659	0.980194	0.980716	0.981227	0.981725	0.982212	0.982687	0.983152
2.4	0.983605	0.984047	0.984479	0.984901	0.985313	0.985714	0.986106	0.986489	0.986862	0.987226
2.5	0.987581	0.987927	0.988265	0.988594	0.988915	0.989228	0.989533	0.989830	0.990120	0.990402
2.6	0.990678	0.990946	0.991207	0.991462	0.991709	0.991951	0.992186	0.992415	0.992638	0.992855
2.7	0.993066	0.993272	0.993472	0.993667	0.993856	0.994040	0.994220	0.994394	0.994564	0.994729
2.8	0.994890	0.995046	0.995198	0.995345	0.995489	0.995628	0.995764	0.995895	0.996023	0.996148
2.9	0.996268	0.996386	0.996500	0.996610	0.996718	0.996822	0.996924	0.997022	0.997118	0.997210
3.0	0.997300	0.997388	0.997472	0.997554	0.997634	0.997712	0.997787	0.997859	0.997930	0.997998

2.5 z スコア測定の高速度化

前節で説明した方法で z スコアを求めるためには長時間のシミュレーションが必要であり、膨大なタンパク質の組合せに対して毎回計算することは効率的ではない。このため様々な単作用度 α , β に対して総作用度 $E'_{A,B}$ の分布の平均と標準偏差を求め、それらを一覧表として用意することで高速化を行った。以下に分布に対する平均と標準偏差の計算手順を示す。

1. $0 \leq \alpha \leq 1$ の範囲を n 等分し, $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_n$ を決める。同様に $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ を決める。
2. α_s と β_t ($0 \leq s \leq n, 0 \leq t \leq n$) のすべての組合せに対して 3 以下を実行する。
3. A, B, C の発現量をランダムに作成する。
4. A の発現量値を E_A^C が α_s に近づくような組合せからランダムに選び入れ替える。
5. 4 を E_A^C が α_s になるまで繰り返す。
6. B についても 4, 5 を同様にして E_B^C を β_t にする。
7. $E'_{A,B}$ を計算する。
8. 3~7 を十分な回数施行し, $E'_{A,B}$ の分布を作成する。
9. 作成した分布の平均 μ と標準偏差 σ を表データの α_s と β_t に対応する箇所の値とする。

上記の手順に従い, z スコア算出のための平均の一覧表と標準偏差の一覧表を作成した。一覧表の作成に当たり, 各 α, β の組に対してサンプル数を 10,000, 試行回数を 3,000,000 回とした。作成した平均と標準偏差の一覧表を表 3, 表 4 に示す。これらの表は α と β の各値に対する $E'_{A,B}$ の分布の平均値と標準偏差を示している。今回は $n=20$ とし, 5%区切り作成することとした。

作成した表 3 の各値を $\mu_{s,t}$ (単作用度が α_s と β_t の時の $E'_{A,B}$ の分布の平均) とし, 表 4 の各値は $\sigma_{s,t}$ (同様に計算された標準偏差) とする。このデータは単作用度を 0~1 で n 等分しているので, 実際のデータから得られる単作用度とは誤差が生じる。よって, 単作用度 E_A^C と E_B^C が $\alpha_i \leq E_A^C \leq \alpha_{i+1}, \beta_j \leq E_B^C \leq \beta_{j+1}$ ($0 \leq i, j \leq n$) で与えられたとき, 平均, 標準偏差をそれぞれ $\mu_{ij}, \mu_{i+1,j}, \mu_{ij+1}, \mu_{i+1,j+1}$, と $\sigma_{ij}, \sigma_{i+1,j}, \sigma_{ij+1}, \sigma_{i+1,j+1}$ を用いて近似する。近似の方法は平均と標準偏差ともに同じなので, ここでは平均を例にして近似の方法を述べる。 x, y, z を 3 軸とする 3 次元空間上の 4 点を $P_1 (\alpha_i, \beta_i, \mu_{ij}), P_2 (\alpha_{i+1}, \beta_j, \mu_{i+1,j}), P_3 (\alpha_i, \beta_{j+1}, \mu_{ij+1}), P_4 (\alpha_{i+1}, \beta_{j+1}, \mu_{i+1,j+1})$ とする。この 4 点が同一平面上にあるとは言えないので, 直線による近似を行う。

まず, zx 平面と平行な線分 P_1P_2, P_3P_4 上で x 座標が E_A^C のときの z 座標をそれぞれ μ'_j, μ'_{j+1} とし, この時の座標をそれぞれ $P_5 (E_{A \rightarrow C}, \beta_i, \mu'_j), P_6 (E_{A \rightarrow C}, \beta_{i+1}, \mu'_{j+1})$ とする。次に線分 $P_5 P_6$ 上で y 座標が E_B^C のときの z 座標を $\mu'_{A,B}$ とする。先程は zx 平面と平行な線分 P_1P_2, P_3P_4 から求めていったが, 次は yz 平面と平行な線分 $P_1 P_3, P_2 P_4$ から同様にして $P_7 (\alpha_i, E_B^C, \mu'_i), P_8 (\alpha_{i+1}, E_B^C, \mu'_{i+1})$ を求め, 線分 $P_7 P_8$ 上で x

座標が E_A^C の時の z 座標を $\mu''_{A,B}$ とする。そして $\mu'_{A,B}$ と $\mu''_{A,B}$ の中点を単作用度が E_A^C, E_B^C の時の近似された平均 μ とする。

同様にして, 単作用度が E_A^C, E_B^C の時の近似された標準偏差 σ を求め, この μ と σ を用いて z スコアを求める。

表 3 平均の一覧表

		B-Cの単作用度 (β)																					
A-Cの単作用度 (α)		0.00	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	
	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
	0.05	0.0000	0.0064	0.0135	0.0212	0.0298	0.0393	0.0500	0.0620	0.0757	0.0913	0.1094	0.1305	0.1556	0.1858	0.2228	0.2694	0.3296	0.4107	0.5255	0.7005	1.0000	
	0.10	0.0000	0.0135	0.0280	0.0437	0.0609	0.0795	0.1000	0.1225	0.1473	0.1750	0.2059	0.2406	0.2800	0.3251	0.3770	0.4377	0.5093	0.5953	0.7005	0.8315	1.0000	
	0.15	0.0000	0.0212	0.0437	0.0678	0.0933	0.1207	0.1500	0.1815	0.2154	0.2520	0.2917	0.3348	0.3819	0.4334	0.4901	0.5528	0.6224	0.7002	0.7877	0.8868	1.0000	
	0.20	0.0000	0.0298	0.0609	0.0933	0.1273	0.1628	0.2000	0.2390	0.2800	0.3231	0.3685	0.4163	0.4668	0.5201	0.5766	0.6365	0.7001	0.7679	0.8401	0.9173	1.0000	
	0.25	0.0000	0.0393	0.0795	0.1207	0.1628	0.2059	0.2500	0.2952	0.3415	0.3889	0.4376	0.4874	0.5385	0.5910	0.6449	0.7001	0.7569	0.8152	0.8751	0.9367	1.0000	
	0.30	0.0000	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5001	0.5501	0.6001	0.6501	0.7001	0.7501	0.8001	0.8501	0.9001	0.9500	1.0000	
	0.35	0.0000	0.0620	0.1225	0.1815	0.2390	0.2952	0.3500	0.4035	0.4559	0.5069	0.5569	0.6057	0.6534	0.7001	0.7457	0.7904	0.8341	0.8769	0.9188	0.9598	1.0000	
	0.40	0.0000	0.0757	0.1473	0.2154	0.2800	0.3415	0.4000	0.4559	0.5091	0.5601	0.6088	0.6554	0.7001	0.7429	0.7841	0.8236	0.8616	0.8982	0.9334	0.9673	1.0000	
	0.45	0.0000	0.0913	0.1750	0.2520	0.3231	0.3889	0.4500	0.5069	0.5601	0.6097	0.6563	0.7001	0.7412	0.7801	0.8167	0.8514	0.8843	0.9154	0.9450	0.9732	1.0000	
	0.50	0.0000	0.1094	0.2059	0.2917	0.3685	0.4376	0.5001	0.5569	0.6088	0.6563	0.7000	0.7405	0.7779	0.8126	0.8449	0.8750	0.9033	0.9297	0.9546	0.9780	1.0000	
	0.55	0.0000	0.1305	0.2406	0.3348	0.4163	0.4874	0.5501	0.6057	0.6554	0.7001	0.7405	0.7770	0.8106	0.8412	0.8699	0.8954	0.9194	0.9417	0.9625	0.9819	1.0000	
	0.60	0.0000	0.1556	0.2800	0.3819	0.4668	0.5385	0.6001	0.6534	0.7001	0.7412	0.7779	0.8106	0.8399	0.8667	0.8909	0.9131	0.9334	0.9520	0.9692	0.9852	1.0000	
	0.65	0.0000	0.1858	0.3251	0.4334	0.5201	0.5910	0.6501	0.7001	0.7429	0.7801	0.8126	0.8412	0.8667	0.8894	0.9100	0.9286	0.9455	0.9609	0.9750	0.9880	1.0000	
	0.70	0.0000	0.2228	0.3770	0.4901	0.5766	0.6449	0.7001	0.7457	0.7841	0.8167	0.8449	0.8694	0.8909	0.9100	0.9269	0.9423	0.9561	0.9686	0.9800	0.9904	1.0000	
	0.75	0.0000	0.2694	0.4377	0.5528	0.6365	0.7001	0.7501	0.7904	0.8236	0.8514	0.8750	0.8954	0.9131	0.9286	0.9423	0.9545	0.9655	0.9754	0.9844	0.9925	1.0000	
	0.80	0.0000	0.3296	0.5093	0.6224	0.7001	0.7569	0.8001	0.8341	0.8616	0.8843	0.9033	0.9194	0.9334	0.9455	0.9561	0.9655	0.9738	0.9815	0.9882	0.9944	1.0000	
	0.85	0.0000	0.4107	0.5953	0.7002	0.7679	0.8152	0.8501	0.8769	0.8982	0.9154	0.9297	0.9417	0.9520	0.9609	0.9686	0.9754	0.9815	0.9868	0.9917	0.9960	1.0000	
	0.90	0.0000	0.5255	0.7005	0.7877	0.8401	0.8751	0.9001	0.9188	0.9334	0.9450	0.9546	0.9625	0.9692	0.9750	0.9800	0.9844	0.9882	0.9917	0.9947	0.9975	1.0000	
	0.95	0.0000	0.7005	0.8315	0.8868	0.9173	0.9367	0.9500	0.9598	0.9673	0.9732	0.9780	0.9819	0.9852	0.9880	0.9904	0.9925	0.9944	0.9960	0.9975	0.9988	1.0000	
	1.00	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	

表 4 標準偏差の一覧表

		B-Cの単作用度 (β)																						
A-Cの単作用度 (α)		0.00	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		
	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		
	0.05	0.0000	0.0024	0.0032	0.0040	0.0046	0.0053	0.0059	0.0065	0.0070	0.0076	0.0082	0.0088	0.0095	0.0103	0.0111	0.0122	0.0137	0.0156	0.0183	0.0209	0.0000		
	0.10	0.0000	0.0032	0.0048	0.0055	0.0064	0.0071	0.0078	0.0084	0.0090	0.0096	0.0101	0.0107	0.0112	0.0118	0.0125	0.0133	0.0141	0.0150	0.0155	0.0164	0.0000		
	0.15	0.0000	0.0040	0.0055	0.0069	0.0076	0.0084	0.0090	0.0096	0.0101	0.0105	0.0109	0.0114	0.0117	0.0121	0.0125	0.0129	0.0132	0.0132	0.0127	0.0108	0.0000		
	0.20	0.0000	0.0046	0.0064	0.0076	0.0087	0.0092	0.0098	0.0103	0.0107	0.0110	0.0113	0.0115	0.0117	0.0119	0.0120	0.0121	0.0120	0.0116	0.0106	0.0085	0.0000		
	0.25	0.0000	0.0053	0.0071	0.0084	0.0092	0.0099	0.0103	0.0107	0.0110	0.0112	0.0113	0.0114	0.0115	0.0115	0.0114	0.0112	0.0109	0.0102	0.0091	0.0070	0.0000		
	0.30	0.0000	0.0059	0.0078	0.0090	0.0098	0.0103	0.0107	0.0109	0.0111	0.0112	0.0112	0.0112	0.0112	0.0111	0.0109	0.0107	0.0103	0.0098	0.0090	0.0078	0.0059	0.0000	
	0.35	0.0000	0.0065	0.0084	0.0096	0.0103	0.0107	0.0109	0.0111	0.0111	0.0111	0.0110	0.0108	0.0106	0.0104	0.0100	0.0095	0.0089	0.0080	0.0068	0.0050	0.0000		
	0.40	0.0000	0.0070	0.0090	0.0101	0.0107	0.0110	0.0111	0.0111	0.0113	0.0109	0.0107	0.0104	0.0101	0.0098	0.0093	0.0087	0.0080	0.0071	0.0060	0.0044	0.0000		
	0.45	0.0000	0.0076	0.0096	0.0105	0.0110	0.0112	0.0112	0.0111	0.0109	0.0112	0.0103	0.0100	0.0096	0.0092	0.0086	0.0080	0.0073	0.0064	0.0053	0.0038	0.0000		
	0.50	0.0000	0.0082	0.0101	0.0109	0.0113	0.0113	0.0112	0.0110	0.0107	0.0103	0.0109	0.0095	0.0091	0.0085	0.0080	0.0073	0.0066	0.0057	0.0047	0.0033	0.0000		
	0.55	0.0000	0.0088	0.0107	0.0114	0.0115	0.0114	0.0112	0.0108	0.0104	0.0100	0.0095	0.0090	0.0085	0.0079	0.0073	0.0067	0.0060	0.0050	0.0042	0.0029	0.0000		
	0.60	0.0000	0.0095	0.0112	0.0117	0.0117	0.0115	0.0111	0.0106	0.0101	0.0096	0.0091	0.0085	0.0080	0.0073	0.0067	0.0061	0.0054	0.0046	0.0037	0.0026	0.0000		
	0.65	0.0000	0.0103	0.0118	0.0121	0.0119	0.0115	0.0109	0.0104	0.0098	0.0092	0.0085	0.0079	0.0073	0.0065	0.0058	0.0051	0.0044	0.0034	0.0023	0.0010	0.0000		
	0.70	0.0000	0.0111	0.0125	0.0125	0.0120	0.0114	0.0107	0.0100	0.0093	0.0086	0.0080	0.0073	0.0067	0.0061	0.0055	0.0049	0.0043	0.0036	0.0029	0.0020	0.0000		
	0.75	0.0000	0.0122	0.0133	0.0129	0.0121	0.0112	0.0103	0.0095	0.0087	0.0080	0.0073	0.0067	0.0061	0.0055	0.0049	0.0062	0.0038	0.0032	0.0025	0.0017	0.0000		
0.80	0.0000	0.0137	0.0141	0.0131	0.0120	0.0109	0.0098	0.0089	0.0080	0.0073	0.0066	0.0059	0.0054	0.0048	0.0043	0.0038	0.0030	0.0022	0.0021	0.0015	0.0000			
0.85	0.0000	0.0156	0.0150	0.0132	0.0116	0.0102	0.0090	0.0080	0.0077	0.0066	0.0057	0.0051	0.0046	0.0041	0.0036	0.0032	0.0027	0.0029	0.0018	0.0012	0.0000			
0.90	0.0000	0.0183	0.0155	0.0127	0.0106	0.0091	0.0078	0.0068	0.0060	0.0053	0.0047	0.0042	0.0037	0.0033	0.0029	0.0025	0.0021	0.0018	0.0025	0.0010	0.0000			
0.95	0.0000	0.0209	0.0144	0.0108	0.0085	0.0070	0.0059	0.0050	0.0044	0.0038	0.0033	0.0029	0.0026	0.0023	0.0020	0.0017	0.0015	0.0012	0.0010	0.0013	0.0000			
1.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000			

2.6 タンパク質の組合せ的な作用を抽出する手順

この節では、本研究の提案手法の手順についてまとめる。タンパク質の発現量データからタンパク質を3つ選ぶ組合せをすべて考え、その中でさらに影響側タンパク質 A と B、被影響側タンパク質 C の組合せを網羅的に作る。この手順で得られた全てのタンパク質の組合せについて単作用度 E_A^C と E_B^C 、総作用度 $E_{A,B}^C$ を計算する。平均と標準偏差の一覧表から単作用度 E_A^C と E_B^C に対応する $E_{A,B}^C$ の分布の平均と標準偏差の近似解を計算によって求める。求めた平均と標準偏差と総作用度 $E_{A,B}^C$ を基に z スコアを求める。z スコアは値が大きい程、組合せ的な作用を示唆しているので、z スコアで降順ソートを行うことで組合せ的な作用が示唆されるタンパク質の組合せのランキングを作成する。

3. 評価

3.1 評価方法

提案手法を C 言語により実装し、実際のタンパク質発現量データに対して実行することで評価を行った。用いたデータは、和歌山県地域結集型共同研究事業[5]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。この発現量データは文献[6]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。

得られた発現量データには実験誤差が生じることがあるため、同じサンプルに対して複数回実験を行っており、このため同じサンプルに対してタンパク質発現量の再現性があるのかを確認する。複数回同じ実験を行っても同じ結果が得られた場合に、その結果には再現性があると判断できる。つまり、再現性を確認することで、発現量データの信頼性を測ることができる。以下に再現性の確認方法について述べる。再現性を確認するということは、複数回の実験を行っているサンプルを対象とするため、一回しか実験が行われていないサンプルについては除外する。次に、同じサンプルでタンパク質の発現量誤差が±3 割に収まらないタンパク質は再現性がないと判断して除外する。それ以外のタンパク質は再現性があると判断して全ての発現量を平均してこれを発現量とする。このように再現性を確認した結果、評価に用いる発現量データはサンプル数が 195、タンパク質数が 897 となった。

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う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ない」とし、Jarque-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。これに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。また、タンパク質の発現量データには欠損もあるため、全サンプル中半数以上で発現量が測定されているタンパク質を用いることとした。このときタンパク質は 405 個となった。本研究ではこの再現性があり、タンパク質の欠損値がサンプル数の半数未満であり、発現量が正規分布でない仮説が棄却できたタンパク質の発現量データを用いて評価実験を行った。

評価実験を行うにあたりフィルタリングのためのパラメータを以下のように設定した。ここで、影響側タンパク質を A と B、被影響側タンパク質を C とする。

評価実験におけるパラメータと条件

1. タンパク質が発現しているかを決定する閾値 r : 30(%)
2. 組合せ作用度の計算に用いる分母 ($\text{num}(A \cap B)$) の下限: 20
3. 組合せ作用度の計算に用いる分子 ($\text{num}(A \cap B \cap C)$) の下限: 10
4. z スコアの下限: 4
5. タンパク質 3 つの発現量がいずれも欠損値でない組合せのみを使用

各パラメータの設定について説明する。1 つ目のパラメータによりタンパク質の発現量が上位の 30%に含まれているサンプルについてタンパク質が発現しているとする。2 つ目のパラメータは総作用度の計算で分母が小さいときに得られた値の信頼性が低いため、これを除外するために設定した。この値の設定方法について説明する。タンパク質 A、B が発現している事象が互いに独立であると仮定したとき同一サンプルでタンパク質 A、B が共に発現している確率はパラメータ 1 より $0.3 \times 0.3 = 0.09$ となる。サンプル数が 195 なので、タンパク質 A、B が共に発現しているサンプル数 $\text{num}(A \cap B)$ は 195×0.09 より 17.55 となる。実際同一サンプルでタンパク質 A、B の発現事象は独立とは言えないこともあるため、今回は $\text{num}(A \cap B)$ の下限を 20 とした。3 つ目のパラメータも同様に統計的有意性を確実にするために設定した。4 つ目のパラメータは、実データの組合せ作用度と統計的指標の平均との距離が統計的指標の標準偏差の 4 倍以上となるタンパク質の組合せを抽出するように設定した。5 つ目のパラメータは欠損値の影響を考慮し、選んだ 3 つのタンパク質のいずれにも欠損値が含まれていないサンプルを用いるためのものである。

3.2 結果と考察

実データに適用し、z スコアによるランキングを行った結果、タンパク質の組合せ的な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。本節では分析結果の詳細と考察について述べる。

Jarque-Bera 検定により発現量の分布が正規分布でないと言えないタンパク質に対して本手法を適用した結果のヒストグラムを図 6 に示す。横軸に z スコアの階級をとり、

縦軸に抽出されたタンパク質の組合せ数をとっている。抽出されたタンパク質の組合せ数が 116,943 であり、考えられるタンパク質の全組合せ数が 32,969,430 なので（用いたタンパク質数が 405 個であり、そこから 3 つ選ぶ組合せ数からさらに被影響側タンパク質を 1 つ選ぶので、組合せ数= ${}_{405}C_3 \times 3$ ）、確率が 0.003547 である。本来 z スコアが 4 以上となるタンパク質の組合せが抽出される確率は統計的に 0.00003167 となるので、組合せ的な作用により全体的に z スコアが上昇しており、評価に用いたデータが相応の組合せ的作用を含むことが確認された。

次に z スコアによるランキング表を表 5 に示す。今回抽出された結果を見てみると、 E_A^C , E_B^C が比較的低く、A, C または B, C の単体では発現を抑制させているタンパク質の組合せが数多く見られる。その中で $E_{A,B}^C$ が比較的大きく、A と B が C の発現を促進しているようなタンパク質の組合せが多数抽出されており、本手法で推定したいタンパク質の組合せ的な相互作用を含むタンパク質の組合せが抽出されていることがわかる。

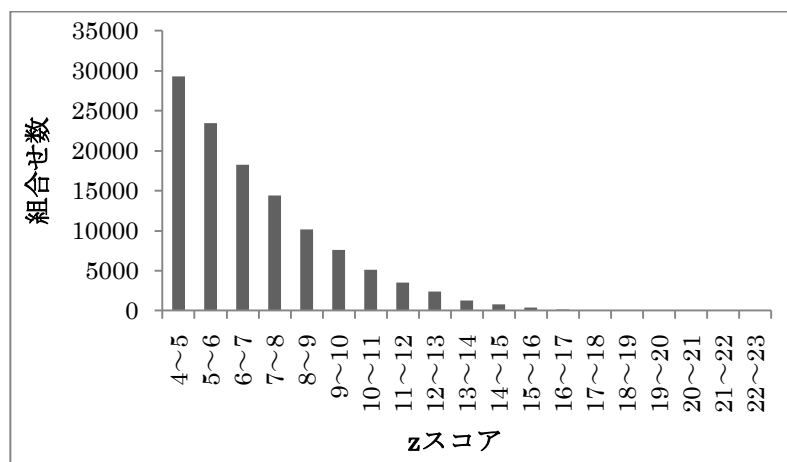


図 6 z スコアのヒストグラム

表 5 z スコアのランキング表 (上位 30 件)

順位	spot_A	spot_B	spot_C	E_A^C	E_B^C	$E_{A,B}^C$	z スコア	num(A ∩ B)	num(A ∩ B ∩ C)
1	955	4794	6152	0.3721	0.3609	0.5784	22.4383	21	12
2	4794	5955	6152	0.2609	0.3778	0.5714	21.8874	21	12
3	6143	226	4926	0.3030	0.2703	0.5000	21.6050	20	10
4	6131	4478	4976	0.3158	0.2619	0.5000	21.2306	20	10
5	6147	5198	6131	0.5750	0.5714	1.0000	20.7721	20	20
6	5159	6150	4794	0.3000	0.3023	0.5238	20.6926	21	11
7	5812	6144	4976	0.3659	0.3500	0.6500	20.6841	20	13
8	5214	6150	4794	0.4091	0.3111	0.6500	20.5788	20	13
9	4201	4794	4976	0.3333	0.2727	0.5238	20.5155	21	11
10	6236	2949	5955	0.5556	0.6316	1.0000	20.4895	20	20
11	5198	4794	4976	0.3333	0.2955	0.5500	20.4615	20	11
12	6015	6131	6152	0.3902	0.3750	0.7000	20.3798	20	14
13	5940	6035	4976	0.2791	0.3529	0.5500	20.3012	20	11
14	5577	2949	4699	0.2703	0.2941	0.4762	20.2556	21	10
15	5573	5727	930	0.5435	0.6512	1.0000	20.2155	20	20
16	4125	4794	4976	0.2857	0.2791	0.4762	20.1907	21	10
17	4284	2595	3995	0.5556	0.6429	1.0000	20.0748	20	20
18	4748	5573	5953	0.6667	0.5333	1.0000	19.9484	20	20
19	4794	5715	6152	0.2889	0.3659	0.5714	19.9375	21	12
20	2718	5572	6065	0.4783	0.7083	1.0000	19.8860	20	20
21	4926	4936	6241	0.3500	0.3750	0.6500	19.8302	20	13
22	5940	5975	4976	0.2791	0.3846	0.5714	19.2133	21	12
23	932	4395	4793	0.2821	0.3611	0.5500	19.1822	20	11
24	4284	4292	5989	0.4872	0.6098	0.9500	19.1582	20	19
25	6022	239	5848	0.5750	0.6429	1.0000	19.1408	20	20
26	5573	5812	1076	0.2727	0.3250	0.5000	19.1172	22	11
27	6022	3036	4013	0.6571	0.5676	1.0000	19.1010	20	20
28	6241	2949	5953	0.6923	0.5238	1.0000	19.0873	20	20
29	4015	4436	4795	0.6757	0.5476	1.0000	18.9992	21	21
30	6233	1333	6154	0.3095	0.3095	0.5238	18.9807	21	11

4. おわりに

本研究では、タンパク質や遺伝子の発現量データを用いて、タンパク質や遺伝子 A, B, C の組合せにおいて A, B が同時に発現している時に C の発現量に与える影響が、A や B がそれぞれ単体で C の発現量に与える影響に対して遥かに大きいモデルを想定し、発現量データから条件付き確率と統計的指標を用いたタンパク質間や遺伝子間の組合せ的な相互作用の推定手法を提案した。本研究の説明でタンパク質を例に出しているが、本研究ではタンパク質に限らず遺伝子の発現量データにも適用可能である。本研究では、条件付き確率と統計的指標により、タンパク質単体での影響に比べて、2 つのタンパク質が発現しているときに他のタンパク質の発現量により大きな影響を与えるようなタンパク質の組合せの抽出が可能である。これが他の研究との大きな違いである。実データに本研究の提案手法を適用することで評価を行った。その結果、2 つのタンパク質の組合せが他のタンパク質に組合せ的な相互作用を及ぼしていると示唆されるタンパク質の組合せが抽出された。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、実際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) 玉田嘉紀, 井元清哉, 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, 333-356.
- 2) 阿久津達也, パ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.
- 3) S. Imoto, T. Goto, S. Miyano, Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. Pacific Symposium on Biocomputing. 7,175-186, 2002.
- 4) T. Akutsu, S. Kuhara, O. Maruyama, S. Miyano, Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model *Theoretical Computer Science* 298, 235-251, 2003.
- 5) 和歌山県地域結集型共同研究事業, <http://www.wakayama-kessyu.com/>
- 6) 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.