条件付確率を用いた発現量データからの複合的タンパク質間相互作用の推定法の高速化

藤木　生聖†　　吉廣　卓哉††

† 和歌山大学大学院システム工学研究科　〒 640–8510 和歌山県和歌山市栄谷 930
†† 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
E-mail: †{s101044,tac}@sys.wakayama-u.ac.jp

あらまし　生命現象の仕組みを解明するためには，タンパク質間の相互作用を推定することが重要である．近年，タンパク質の発現量を測定する技術の向上により，発現量を用いてタンパク質間の相互作用を推定する研究が盛んに行われている．我々は先行研究において，条件付確率に基づいて，発現量データから 3 タンパク質間の質相互作用を推定する手法を提案した．しかし先行研究では，相互作用推定アルゴリズムを動作させる事前処理として，シミュレーションによる確率分布の作成が必要であり，膨大な計算時間がかかる問題があった．本研究では，シミュレーションによる確率分布の作成を解析的な計算に置き換えることで，計算時間を大幅に短縮する手法を提案する．また，評価実験によって，提案する解析的計算方法の精度と計算速度を評価する．

キーワード　タンパク質間相互作用, プロテオーム解析, タンパク質発現量, データマイニング, 条件付確率

1. はじめに

生物は膨大な種類のタンパク質で構成されており，タンパク質間の複雑な相互作用によって維持されているため，タンパク質間の相互作用を明らかにすることは，生命現象を解明するための重要な問題である．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，生命現象の解明を目指した研究が盛んに行われている．これらの研究によりタンパク質間の相互作用の解明が進んできているが，生命現象の全体像を把握するためにはまだ十分とはいえない．

タンパク質間の相互作用の解明を進めるための有力な方法として，発現量データを用いた手法がある．例えば，2 タンパク質間の相互作用を推定するには，2 つのタンパク質発現量の相関係数を計算するなど単純な方法がいくつもあり，よく用いられる．しかし，単純な相互作用だけでは生命現象を把握することはできないため，次の段階として，発現量データを用いてより複雑なタンパク質間相互作用を推定する必要がある．

複雑なタンパク質間相互作用を解明する主要な方法の一つとして，発現量データから相互作用ネットワークを推定する手法がある．例えば，ブーリアンネットワーク [1] [2] やベイジアンネットワーク [3] [4] は発現量データから相互作用を推定する手法としてよく知られている．特にベイジアンネットワーク [3] [4] を用いた相互作用の推定は，発現量データから遺伝子間またはタンパク質間の条件付確率に基づいて，最適な相互作用ネットワークを構築する．しかし，推定されたネットワークには複数タンパク質間の相互作用，例えばタンパク質 A と B が C に影響を与えている相互作用の中には，A と B がそれぞれ単体で C に影響を与える相互作用が含まれているため，これらを分離して，純粋な複数タンパク質間の相互作用を推定することが必要である．

3 タンパク質間の相互作用に着目した研究は数が少ないが，Zhang らは，遺伝子 C の発現量が高い場合または低い場合において，遺伝子 A と B の発現量の相関係数を比較することで相互作用を推定する手法 [5] を提案している．茅野らは，発現量データと遺伝子型データを用いて，遺伝子型によって遺伝子 A と B の発現量の相関係数の符号が切り替わることを検出する手法 [6] を提案している．これらの相互作用は 3 遺伝子間の相互作用を推定しているが，遺伝子 A と B の発現量の相関が別の遺伝子または遺伝子型の状態によって切り替わるという相互作用を検出しているため，対象とする相互作用の働きが限定的である．

我々は発現量データから 3 タンパク質間の相互作用を推定する別の手法 [7] を提案した．この手法では，ベイジアンネットワーク [3] [4] と同様に，3 タンパク質間の相互作用を条件付確率に基づいて推定するが，統計的指標に基づいて 2 タンパク質間の相互作用の影響を取り除くことで，純粋な 3 タンパク質間の相互作用の推定が可能である．しかし，この手法ではコンピュータシミュレーションによる統計分布の作成が必要であり，膨大な計算時間がかかってしまう問題がある．本論文では，計算時間を改善するために，コンピュータシミュレーションを用いた方法に代わって，解析的な方法で統計分布を作成する手法を提案する．

本論文の構成は以下の通りである．2 章では我々が提案した 3 タンパク質間の相互作用を推定するための手法 [7] を説明する．3 章では相互作用推定手法 [7] の計算時間を改善するための高速化手法について説明する．4 章では提案手法の計算時間と精度を評価する．最後に 5 章で本研究のまとめとする．

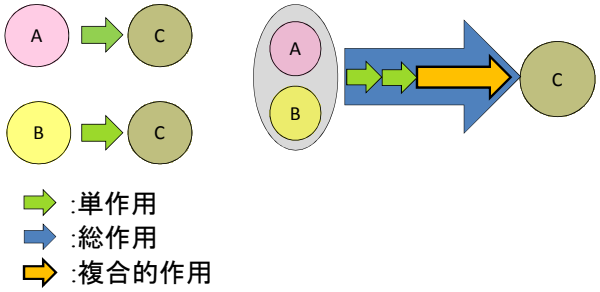

図 1　想定する相互作用のモデル

2.　複合的タンパク質間相互作用の推定手法

2.1　想定する相互作用のモデル

我々は先行研究として，条件付確率に基づいた 3 タンパク質の相互作用の推定手法を提案した [7]．この章では，提案した相互作用の推定手法の概要を説明する．その後，3 章で，膨大な時間を要するシミュレーション計算を置き換え高速化する提案手法について述べる．

本研究で想定するタンパク質の相互作用のモデルを図 1 に示す．タンパク質 A と B が影響を与える側で，タンパク質 C が被影響側である．タンパク質 A や B が単体でタンパク質 C の発現量に与える影響があっても良い．本研究では，これら単体の影響に加えて，A と B が同時に発現している場合に C の発現量に大きな影響を与えるモデルを想定する．タンパク質 A(または B) が単体で C に影響を与える作用を**単作用**，A と B が C に影響を与えている作用を**総作用**，そして A と B が共に発現している場合にだけ C に影響を与える作用を**複合的作用**と定義する．ここで注意しておきたいことは，総作用には単作用と複合的作用の両方が含まれていることである．

3 タンパク質間の相互作用の推定にあたっては，単作用を除いた複合的作用の量を見積もることが重要である．文献 [7] は，複合的作用の量を見積もるアルゴリズムを提案した．複合的作用を計算するには，まず総作用度を計算する．そして，総作用度から 2 つの単作用度 (A→C と B→C の作用度) を取り除くことで複合的作用度を求める．

2.2　本研究で扱うタンパク質の発現量データ

タンパク質の発現量とは，サンプル組織中に含まれる各タンパク質の含有量のことである．入力となるタンパク質発現量データは図 2 に示す生物学的な実験 [8] によって得られる．まず，サンプルを用意し，生物学的な実験によって個々のタンパク質に分離された二次元電気泳動画像を得る．次に，画像処理を行うことによって，分離されたタンパク質のスポットを識別し，発現量の測定を行う．最後に，二次元電気泳動画像から同じタンパク質のスポットが一致するように画像処理を行う．これ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．サンプルを $i(=1,2,3,...,I)$，タンパク質を $j(=1,2,3,...,J)$ とおいたとき，サンプル i におけるタンパク質 j の発現量 e_{ij} は実数値で与えられる．二次元電気泳動により，タンパク質の発現量を測定する場合，発現量データに含まれるタンパク質数は (生物種や部位にもよるが) 数百から数千であるといわれており，また，実験には手間もかかるため，サンプル数はせいぜい数十から多くとも数百である．

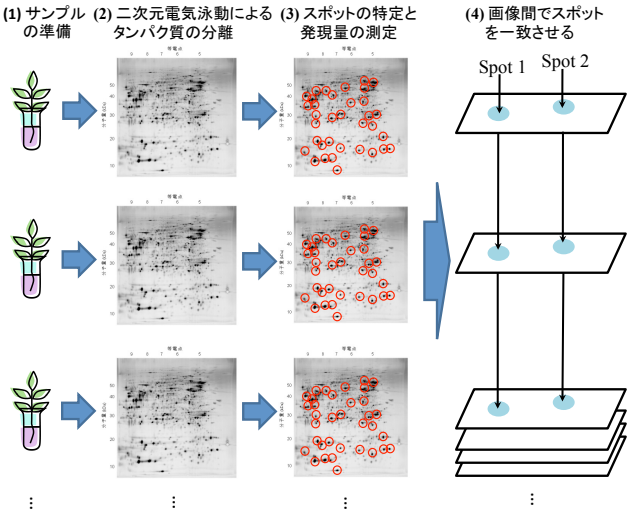

図 2　発現量データを得るための実験手順

表 1　タンパク質の発現量データ

Sample ID	Protain ID					
	$j=1$	$j=2$	$j=3$	$j=4$	...	J
$i=1$	0.000582	0.000107	0.000338	0.000451	...	...
$i=2$	0.000563	0.000142	0.000475	0.000458	...	...
$i=3$	0.000495	0.000126	0.000433	0.000565	...	...
$i=4$	0.000553	0.000153	0.000382	0.000486	...	...
⋮	⋮	⋮	⋮	⋮	⋱	
I	⋮	⋮	⋮	⋮		⋱

2.3　単作用度と総作用度の計算方法

本節では，条件付確率を用いて発現量データから単作用度と総作用度を計算する方法について図 3 を用いて述べる．タンパク質 A がタンパク質 C に単体で影響を与える度合いを表す単作用度は，A の発現量が多いサンプル数に対する，A と C の発現量が共に多いサンプル数の比として計算できる．A と B が C に影響を与える総作用度も同様に計算できる．つまり，A と B の発現量が共に多いサンプル数に対する，A と B と C の発現量が全て多いサンプル数の比として計算できる．

この考え方に基づいた単作用と総作用の正式な定義は次のようになる．まず，タンパク質が発現しているかどうかを 2 値で表現する．このために，しきい値 $r(0 \leqq r \leqq 1)$ を定義し，タンパク質 j に対して，サンプル i の発現量 e_{ij} が全ての $i(1 \leqq i \leqq I)$ の中で上位 r に含まれるとき,「サンプル i ではタンパク質 j が“発現している”」と定義する．また，タンパク質 A, B, C が発現しているサンプルの集合を A, B, C で表し $|A|$ をタンパク質 A が発現しているサンプル数，$|A \cap B|$ 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．このとき，A が C の発現量に単体で影響を与える単作用度を $E_{A \to C} = \dfrac{|A \cap C|}{|A|}$ と定義し，A と B が C の発現量に影響を与える総作用度を

$E_{(A,B)\to C} = \dfrac{|A \cap B \cap C|}{|A \cap B|}$ と定義する．

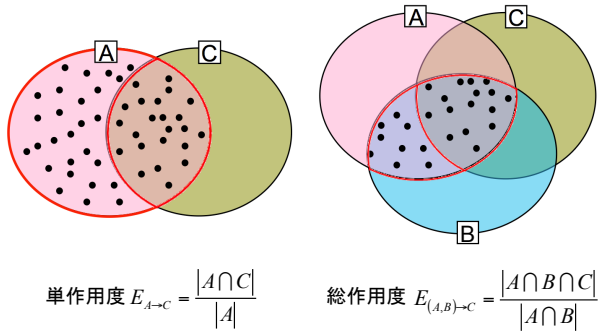

図 3　単作用度と総作用度の計算方法

2.4　複合的作用度の計算方法

本手法で推定したいものは，タンパク質 A と B が共に発現している場合にだけ C に影響を与える複合的作用である．タンパク質 A，B，C の組合せから複合的作用度を求めるには，2.1 節で説明したように，総作用度から 2 つの単作用度 (A→C と B→C の作用度) を分離する必要がある．ここで，複合的作用度を計算するために総作用度から単作用度を分離する方法について図 4 を用いて説明する．まず，2.3 節で述べた $E_{(A,B)\to C} = \frac{|A\cap B\cap C|}{|A\cap B|}$ (図 4(a)) を計算する．次に，タンパク質 A，B，C の間に複合的作用がないと仮定した場合の総作用度 $E'_{(A,B)\to C}$ の統計的分布 (図 4(b)) を計算する．この統計的分布を計算する方法は後で述べる．それから，$E_{(A,B)\to C} = \frac{|A\cap B\cap C|}{|A\cap B|}$ と $E'_{(A,B)\to C} = \frac{|A\cap B\cap C|}{|A\cap B|}$ の差 (図 4(c)) を z 値として計算する．ここで，z 値は $z = \dfrac{E_{(A,B)\to C} - \mu}{\sigma}$ により定義される．ここで，$E_{(A,B)\to C}$ は実際の発現量データから得られるタンパク質 A，B，C による総作用度，μ と σ は $E'_{(A,B)\to C}$ の分布の平均と標準偏差である．つまり，z 値は実際の総作用度 $E_{(A,B)\to C}$ が $E'_{(A,B)\to C}$ の分布の平均からどれだけ離れているかを $E'_{(A,B)\to C}$ の分布の標準偏差を単位として表した値である．そして，z 値は絶対値が大きいほど平均から離れているため，その分だけ生起確率が下がることを意味し，これは複合的作用度が大きいことを意味する．

先行研究 [7] では，$E'_{(A,B)\to C}$ の分布をコンピュータシミュレーションによって計算した．コンピュータシミュレーションによって $E'_{(A,B)\to C}$ の分布を計算する手順を以下に示す．まず，実際の発現量データから 2 つの単作用度 (A→C と B→C の作用度) を計算し，それぞれ α，β とする．次に，単作用度が α，β となるようにランダムに人工的な発現量データを作成し，総作用度 $E'_{(A,B)\to C}$ を計算する．ここで，人工的な発現量データはタンパク質の発現量がランダムになるように作成されているため，この時計算した総作用度は複合的作用が含まれていない総作用度 $E'_{(A,B)\to C}$ である．そして，人工的な発現量データの作成と総作用度 $E'_{(A,B)\to C}$ の計算の一連の手順を十分な回数だけ試行することで，$E'_{(A,B)\to C}$ の分布を求めることができる．

ここで，コンピュータシミュレーションによって $E'_{(A,B)\to C}$ の分布を計算する方法の問題点を説明する．$E'_{(A,B)\to C}$ の分布を計算するには，人工的な発現量データの作成と総作用度 $E'_{(A,B)\to C}$ の計算の一連の手順を十分な回数だけ試行する必要があるため，かなりの計算時間を要するという問題がある．

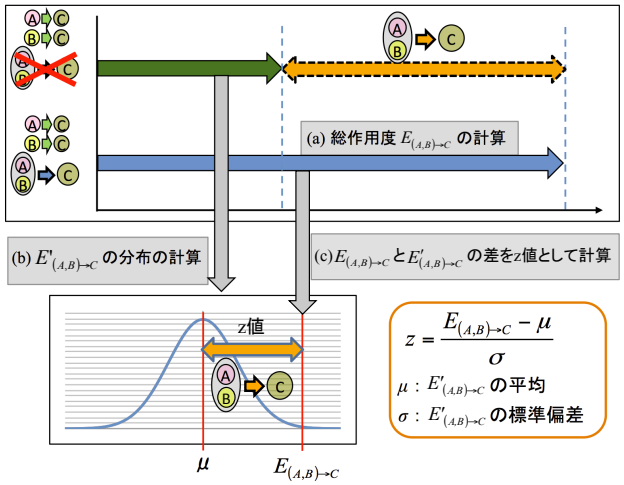

図 4　複合的作用を推定する手順の概要

3.　統計的分布作成の高速化手法

3.1　高速化手法の概要

本章では，2. 章で述べた相互作用推定手法 [7] の問題点である計算時間を改善するため，コンピュータシミュレーションによる複合的作用度の計算に代わって，解析的な方法で複合的作用度を計算する提案手法を述べる．

コンピュータシミュレーションによって複合的作用度を計算するには，膨大な時間をかけて，複合的作用がないと仮定した場合の総作用度の確率分布を求めなければならないという問題があった．本論文では，2.1 節で述べたタンパク質相互作用のモデルに従って，解析的な方法で複合的作用がないと仮定した場合の総作用度の確率分布を求めることにより，複合的作用度を計算する．総作用度は 2.3 節で述べた式 $E'_{(A,B)\to C} = \frac{|A\cap B\cap C|}{|A\cap B|}$ により計算できる．この総作用度 $E'_{(A,B)\to C}$ の確率分布を確率変数として表現する．確率変数 $X = |A \cap B \cap C|$ 及び $Y = |A \cap B \cap \overline{C}|$ を用いると（図 5 のベン図を参照のこと），総作用度は確率変数 $\frac{X}{X+Y}$ で表される．まず，2 つの確率変数 X，Y を比較的単純な確率分布モデルで表現する方法を 3.2 節で説明する．そして，複合的作用がない場合の総作用度 $\frac{X}{X+Y}$ の分布の計算方法を 3.3 節で説明する．

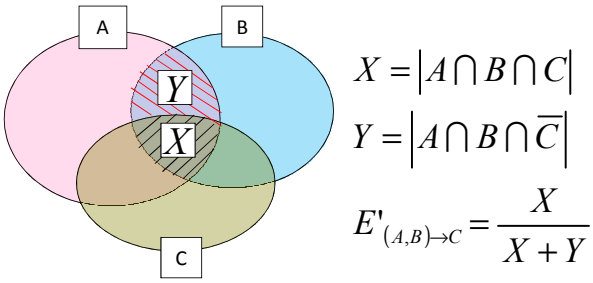

図 5　求めたい確率密度関数の確率変数

3.2 超幾何分布の適用とその近似

本節では，確率変数 X と Y の確率密度関数を求める方法を説明する．まず，前提を整理する．サンプル数を I とする．定義より，タンパク質 A の発現量の上位 r に入るサンプル s に対して $s \in A$ であり，それ以外のサンプル t に対して $t \notin A$ である．即ちタンパク質 A が発現しているサンプル数は $|A| = rI$ である．タンパク質 B, C に対しても同様である．ここで，本節で扱う，複合的作用がないと仮定した場合の総作用度の確率分布を求めるにあたっては，A-C 間及び B-C 間の単作用度 $E_{A\to C} = \alpha$ 及び $E_{B\to C} = \beta$ のみが制約として与えられ，その他は全てランダムに決まると仮定する．ここで単作用度 α と β は，現在複合的作用度を計算しているタンパク質 A, B, C に対して実データから求められた値である．定義 $\alpha = \frac{|A\cap C|}{|A|}$ より，C に含まれるサンプルで A にも含まれるサンプルの集合 $A\cap C$ の要素数は $|A\cap C| = \alpha|C|$ である．同様に，$|B\cap C| = \beta|C|$ である．α, β 以外の制約はなくランダムであるから，確率変数 X は，$|C|$ 個の要素からランダムに $\alpha|C|$ 個の要素を選んだときに，その中に B が発現しているサンプルが $x = |A\cap B\cap C|$ 個含まれる確率分布 $f(x)$ として表される．

これは超幾何分布である．超幾何分布 $f(w)$ は，全体集合 N の中に特定の性質があるものの集合を M として，N から k 個の要素を選択した時に，M に含まれる要素が w 個含まれる確率として定義される．従って，$N = rI$, $M = |A\cap C| = \alpha|C| = \alpha rI$, $k = \beta|C| = \beta rI$ とおくと，確率変数 X の確率密度関数 $f(x)$ は，$f(x) = \frac{{}_{\alpha rI}C_x \cdot {}_{(1-\alpha)rI}C_{(\beta rI - x)}}{{}_{rI}C_{\beta rI}}$ で表される．同様に，確率変数 Y の確率密度関数は，$y = |A\cap B\cap \overline{C}|$, $N = (1-r)I$, $M = |\overline{A}\cap C| = (1-\alpha)|C| = (1-\alpha)rI$, $k = (1-\beta)|C| = (1-\beta)rI$ とおくと，$g(y) = \frac{{}_{(1-\alpha)rI}C_y \cdot {}_{\alpha rI}C_{((1-\beta)rI-y))}}{{}_{(1-r)I}C_{(1-\beta)rI}}$ で表される．

次に，確率変数 X と Y が独立であることを示す．超幾何分布は，N, M, k の値が決まれば確率分布が決まる．確率変数 X, Y のいずれも，これらの値は I, α, β, r の 4 変数に依存し，いずれも入力される固定値である．従って，X と Y は互いに依存することなく，独立である．

超幾何分布は組合せの計算をする必要があるため，サンプル数が多くなると計算時間が爆発的に大きくなる．そのため，超幾何分布を正規分布に近似する．超幾何分布は，N が十分大きい仮定の下で，超幾何分布と同じ平均・標準偏差の正規分布に近似できる．ここで，$f(x)$ の平均 μ_x，標準偏差 σ_x および $g(y)$ の平均 μ_y，標準偏差 σ_y はそれぞれ以下のように表される．

$$
\begin{aligned}
\mu_x &= \alpha\beta rI \\
\mu_y &= \frac{(1-\alpha)(1-\beta)r^2 I}{1-r} \\
\sigma_x &= \frac{\alpha\beta(1-\alpha)(1-\beta)r^2 I^2}{rI-1} \\
\sigma_y &= \frac{(1-\alpha)(1-\beta)(1-2r+\alpha r)(1-2r+\beta r)r^2 I^2}{(1-r)^2((1-r)I-1)}
\end{aligned}
$$

3.3 総作用度 $E'_{(A,B)\to C}$ の分布の計算

この節では，複合的作用がない場合の総作用度 $E'_{(A,B)\to C}$ の分布を計算する方法について述べる．求めたい分布は確率変数 $\frac{X}{X+Y}$ の確率密度関数で表される．確率変数 $\frac{X}{X+Y}$ は分子と分母で互いに独立ではないため，まず確率変数の逆数 $1+\frac{Y}{X}$ の分布を求め，最後に逆数を戻すことで確率変数 $\frac{X}{X+Y}$ の確率密度関数を求める．

確率変数同士の除算 $\frac{Y}{X}$ の確率密度関数は，Hinkley が提案した互いに独立な正規分布に従う確率変数 X，Y の除算の確率密度関数 [9] を用いると以下のように表すことができる．ここで，確率変数 $M = \frac{Y}{X}$ とおき，M の確率密度関数を $\phi(m)$ とする．また，$P = m^2{\sigma_x}^2 + {\sigma_y}^2$，$Q = \mu_x{\sigma_y}^2 + m\mu_y{\sigma_x}^2$，$R = {\mu_x}^2{\sigma_y}^2 + {\mu_y}^2{\sigma_x}^2$，$S = 2{\sigma_x}^2{\sigma_y}^2$ とおく．さらに，$erf()$ は誤差関数であり，$erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt.$ で表される．

$$
\phi(m) = \frac{e^{\frac{Q^2}{PS}-\frac{R}{S}}}{2\pi\sigma_x\sigma_y P}\left\{Se^{-\frac{Q^2}{PS}} + Q\sqrt{\frac{S\pi}{P}}\mathrm{erf}\left(\frac{Q}{\sqrt{PS}}\right)\right\} \quad (1)
$$

確率変数 $\frac{X}{X+Y}$ の確率密度関数を求めるために，$1+\frac{Y}{X} = 1+M$ の逆数を考える．ここで，確率変数 $V = \frac{X}{X+Y} = \frac{1}{1+M}$ の確率密度関数を $h(v)$，確率分布関数を $H(v)$，確率変数 M の確率分布関数を $\Phi(m)$ とおくと，確率分布関数 $H(v)$ を $\Phi(m)$ を用いると以下のように表される．また，$P(V \leqq v)$ は確率変数 V が v 以下の値をとる確率を表す．

$$
\begin{aligned}
H(v) &= P(V \leqq v) = P\left(\frac{1}{1+M} \leqq v\right) \\
&= P\left(\frac{1}{v} - 1 \leqq M\right) = 1 - \Phi\left(\frac{1}{v} - 1\right) \qquad (2)
\end{aligned}
$$

(2) 式の両辺を v について微分すると，確率変数 $V = \frac{X}{X+Y}$ の確率密度関数 $h(v)$ が次の式で表される．

$$
h(v) = \frac{dH(v)}{dv} = -\frac{d\Phi\left(\frac{1}{v}-1\right)}{dv} = \frac{1}{v^2}\phi\left(\frac{1}{v}-1\right) \qquad (3)
$$

(1) 式，(3) 式より，$h(v)$ は次の式で表される．ここで，$P' = (1-v)^2{\sigma_x}^2 + v^2{\sigma_y}^2$，$Q' = (1-v)\mu_y{\sigma_x}^2 + v\mu_x{\sigma_y}^2$ とおく．

$$
h(v) = \frac{e^{\frac{Q'^2}{P'S}-\frac{R}{S}}}{2\pi P'}\left\{\sqrt{2S}e^{-\frac{Q'^2}{P'S}} + Q'\sqrt{\frac{2\pi}{P'}}erf\left(\frac{Q'}{\sqrt{P'S}}\right)\right\}
$$

以上のようにして，複合的作用がない場合の総作用度の確率密度関数 $h(v)$ が求まる．

4. 評　　価

4.1 分布の作成にかかる計算時間の比較

提案手法と既存手法（シミュレーションによる方法）の計算速度と精度を比較し，評価を行った．まず，計算時間の比較方法について説明する．先行研究 [7] では，タンパク質 A, B, C の全組合せについてシミュレーションを行うと膨大な時間がかかるため，予め，入力となる単作用度 α と β の値の組に対して $E'_{(A,B)\to C}$ の分布の平均値と標準偏差を求めた z 値算出表を作成する．本評価実験では，この z 値算出表を求めるために要する時間を比較する．本評価実験では，α，β をそれぞれ 0 から

1 の間で 0.05 毎に区切った．つまり，それぞれ 21 の値をとる．また，α と β の値が入れ替わっても同じ値をとるため，z 値算出表は $\frac{21\times22}{2}=231$ のセルを含む．z 値算出表の例を表 2 に示す．シミュレーションにおけるパラメータとして，発現を判定する閾値を $r=0.5$，サンプル数を $I=500$，$1,000$，$10,000$ とした．また，分布を作成するための試行回数として，10,000 回，100,000 回，1,000,000 回，5,000,000 回の 4 種類を用いた．

結果を表 3 に示す．表 3 には，z 値算出表の作成時間と，1 セルあたりの計算時間が示されている．いずれの値からも，提案手法による計算が圧倒的に速いことがわかる．シミュレーションを用いた既存手法ではサンプル数や試行回数に依存して計算時間が増加するが，提案手法ではそれらに依存せずほぼ一定の時間で計算できる強みがある．

表 3　z 値算出表の作成時間

	サンプル数	コンピュータシミュレーション				高速化手法
		10,000回	100,000回	1,000,000回	5,000,000回	(提案手法)
z値算出表作成時間	500	11分15秒	17分14秒	1時間1分45秒	5時間10分58秒	38秒95
	1,000	24分57秒	35分48秒	2時間2分37秒	9時間49分41秒	
	10,000	12時間39分42秒	14時間17分4秒	28時間1分42秒	103時間28分31秒	
1セルあたりの計算時間	500	3秒55	5秒44	19秒50	1分38秒20	0秒205
	1,000	7秒87	11秒30	38秒72	3分6秒21	
	10,000	3分59秒90	4分30秒65	8分51秒06	32分40秒58	

4.2　分布精度の比較

次に精度を評価した．前節の評価実験で作成した一覧表の各セルに対して，試行回数が 5,000,000 回の場合のシミュレーション結果を正解として，提案手法により得られた分布の平均と標準偏差における相対誤差を求めた．ただし，シミュレーションで得た結果を a, 提案手法で得た結果を b としたとき，相対誤差は $\frac{b-a}{a}$ で計算される．提案手法では，計算された分布の平均と標準偏差のみを用いて組合せ的作用の大きさを推定するため，この 2 値の誤差が小さいほど正確な推定が可能である．

求めた分布の平均と標準偏差の相対誤差の統計量を表 4 に示す．統計量として，最大値，最小値，平均，分散の 4 値を求めた．まず，分布の平均と標準偏差に共通して，サンプル数が少なくなるにつれて相対誤差が大きくなっていることがわかる．この原因は，コンピュータシミュレーションと提案手法で複合的作用のない総作用度 $E'_{(A,B)\to C}$ の計算の扱いが異なるからだと考えられる．コンピュータシミュレーションでは条件付確率によって $E'_{(A,B)\to C}$ を計算しているため，サンプル数が十分多ければ連続型として考えられるが，そうでなければ離散型となってしまう．それに対し，提案手法では $E'_{(A,B)\to C}$ の分布を連続型として解析的に求めている．このため，サンプル数が少なくなるほど相対誤差が大きくなったと考えられる．

次に誤差が最大であったサンプル数が 500 のときの分布の平均に注目すると，相対誤差の最大値が 2.83×10^{-4} と小さい値であることがわかる．平均値は 1.36×10^{-5} であるので，この値は十分に小さく，実用的に問題ない範囲であると言える．同様にサンプル数が 500 のときの標準偏差の相対誤差は，平均値は 2.44×10^{-3} と小さく，その最大値は 3.42×10^{-2} と比較的小さいと言える結果であった．5% 以下の誤差であることを考えると平均値と同様に実用的に問題のない範囲であるといえる．ここで，各単作用度 α，β に対する標準偏差の相対誤差のばらつきを図 6 に示す．この図は，提案手法で求めた一覧表の各セルに対して，相対誤差の値により着色した図である．この図から，単作用度 α（または β）が 0.05 に，β（または α）が 0.95 に近いほど相対誤差が大きくなることがわかる．この誤差の原因は特定できていないが，超幾何分布から正規分布に近似するときの近似誤差であると考えられる．

以上の評価結果より，提案手法は計算時間が圧倒的に短く，またコンピュータシミュレーションとほぼ同等の分布が作成できることがわかった．

表 4　シミュレーションと提案手法の相対誤差

	サンプル数	相対誤差(絶対値)			
		最大値	最小値	平均	標準偏差
分布の平均	500	2.83×10^{-4}	1.63×10^{-8}	1.36×10^{-5}	2.92×10^{-5}
	1,000	2.76×10^{-4}	4.82×10^{-8}	1.44×10^{-5}	3.00×10^{-5}
	10,000	9.96×10^{-5}	1.64×10^{-8}	4.50×10^{-6}	9.56×10^{-6}
分布の標準偏差	500	3.42×10^{-2}	1.36×10^{-5}	2.44×10^{-3}	3.96×10^{-3}
	1,000	1.73×10^{-2}	5.30×10^{-6}	1.23×10^{-3}	1.96×10^{-3}
	10,000	1.76×10^{-3}	3.86×10^{-6}	2.24×10^{-4}	2.23×10^{-4}

4.3　複合的タンパク質間相互作用の推定に必要な計算時間の見積もり

この節では，相互作用の推定に必要な時間も含めた，相互作用推定手法全体の計算時間の見積もりを行う．まず，z 値算出表の作成時間について，4.1 節では単作用度 α と β をそれぞれ 0 から 1 の間で 0.05 毎に区切ったが，より精度を上げるため，今回はそれぞれ 0 から 1 の間で 0.01 毎に区切った．このときの z 値算出表は $\frac{101\times102}{2}=5151$ のセルを含む．よって，z 値算出表の作成時間はわずか $5151\times0.205=1055.955$(秒)，つまり約 17 分である．

次に，2. 章で述べた複合的タンパク質間相互作用の推定手法による計算時間の見積もりを行う．見積もりを行うために，サンプル数を 500，1,000，5,000，10,000 とし，それぞれについてタンパク質数が 500，1,000，1,500 の計 12 個の発現量データを人工的に作成した．作成した発現量データを用いて相互作用推定手法の計算時間を測定した結果を表 5 と図 7 に示す．表 5 の各値はそれぞれのサンプル数とタンパク質数における計算時間である．図 7 の横軸は 3 つのタンパク質 A，B，C の組み合わせ数，縦軸は計算時間 (分) を表している．結果より，それぞれのサンプル数において，タンパク質の組合せ数と計算時間は比例の関係にある事がわかる．同様に，それぞれのタンパク質数において，サンプル数と計算時間も比例の関係にある．このことから，サンプル数とタンパク質数がわかれば，おおよその計算時間を見積もることができる．

表 2　z 値算出表

A→Cの単作用度(α) ＼ B→Cの単作用度(β)	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 0.0000	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	- - -	0.0064 0.0024	0.0135 0.0032	0.0212 0.0040	0.0298 0.0046	0.0393 0.0053	0.0500 0.0059	0.0620 0.0065	0.0757 0.0070	0.0913 0.0076	0.1094 0.0082	0.1305 0.0088	0.1556 0.0095	0.1858 0.0103	0.2228 0.0111	0.2694 0.0122	0.3296 0.0137	0.4107 0.0156	0.5255 0.0183	0.7005 0.0209	1.0000 0.0000
0.10	- - -	- - -	0.0280 0.0048	0.0437 0.0055	0.0609 0.0064	0.0795 0.0071	0.1000 0.0078	0.1225 0.0084	0.1473 0.0090	0.1750 0.0096	0.2059 0.0101	0.2406 0.0107	0.2800 0.0112	0.3251 0.0118	0.3770 0.0125	0.4377 0.0133	0.5093 0.0141	0.5953 0.0150	0.7003 0.0155	0.8315 0.0144	1.0000 0.0000
0.15	- - -	- - -	- - -	0.0678 0.0069	0.0933 0.0076	0.1207 0.0084	0.1500 0.0090	0.1815 0.0096	0.2154 0.0101	0.2520 0.0105	0.2917 0.0109	0.3348 0.0114	0.3819 0.0117	0.4334 0.0121	0.4901 0.0125	0.5528 0.0129	0.6224 0.0132	0.7002 0.0132	0.7877 0.0127	0.8868 0.0108	1.0000 0.0000
0.20	- - -	- - -	- - -	- - -	0.1273 0.0087	0.1628 0.0092	0.2000 0.0098	0.2390 0.0103	0.2800 0.0107	0.3231 0.0110	0.3685 0.0113	0.4163 0.0115	0.4668 0.0117	0.5201 0.0119	0.5766 0.0120	0.6365 0.0121	0.7001 0.0120	0.7679 0.0116	0.8401 0.0106	0.9173 0.0085	1.0000 0.0000
0.25	- - -	- - -	- - -	- - -	- - -	0.2059 0.0099	0.2500 0.0103	0.2952 0.0107	0.3415 0.0110	0.3889 0.0112	0.4376 0.0113	0.4874 0.0114	0.5385 0.0115	0.5910 0.0115	0.6449 0.0114	0.7001 0.0112	0.7569 0.0109	0.8152 0.0102	0.8751 0.0091	0.9367 0.0070	1.0000 0.0000
0.30	- - -	- - -	- - -	- - -	- - -	- - -	0.3000 0.0107	0.3500 0.0109	0.4000 0.0111	0.4500 0.0112	0.5001 0.0112	0.5501 0.0112	0.6001 0.0111	0.6501 0.0109	0.7001 0.0107	0.7501 0.0103	0.8001 0.0098	0.8501 0.0090	0.9001 0.0078	0.9500 0.0059	1.0000 0.0000
0.35	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.4035 0.0111	0.4559 0.0111	0.5069 0.0111	0.5569 0.0110	0.6057 0.0108	0.6534 0.0106	0.7001 0.0104	0.7457 0.0100	0.7904 0.0095	0.8341 0.0089	0.8769 0.0080	0.9188 0.0068	0.9598 0.0050	1.0000 0.0000
0.40	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.5091 0.0113	0.5601 0.0109	0.6088 0.0107	0.6554 0.0104	0.7001 0.0101	0.7429 0.0098	0.7841 0.0093	0.8236 0.0087	0.8616 0.0080	0.8982 0.0071	0.9334 0.0060	0.9673 0.0044	1.0000 0.0000
0.45	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.6097 0.0112	0.6563 0.0103	0.7001 0.0100	0.7412 0.0096	0.7801 0.0092	0.8167 0.0086	0.8514 0.0080	0.8843 0.0073	0.9154 0.0064	0.9450 0.0053	0.9732 0.0038	1.0000 0.0000
0.50	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.7000 0.0109	0.7405 0.0095	0.7779 0.0091	0.8126 0.0085	0.8449 0.0080	0.8750 0.0073	0.9033 0.0066	0.9297 0.0057	0.9546 0.0047	0.9780 0.0033	1.0000 0.0000
0.55	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.7770 0.0104	0.8106 0.0085	0.8412 0.0079	0.8694 0.0073	0.8954 0.0067	0.9194 0.0059	0.9417 0.0051	0.9625 0.0042	0.9819 0.0029	1.0000 0.0000
0.60	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.8399 0.0096	0.8667 0.0073	0.8909 0.0067	0.9131 0.0061	0.9334 0.0054	0.9520 0.0046	0.9692 0.0037	0.9852 0.0026	1.0000 0.0000
0.65	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.8894 0.0085	0.9100 0.0061	0.9286 0.0055	0.9455 0.0048	0.9609 0.0041	0.9750 0.0033	0.9880 0.0023	1.0000 0.0000
0.70	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.9269 0.0075	0.9423 0.0049	0.9561 0.0043	0.9686 0.0036	0.9800 0.0029	0.9904 0.0020	1.0000 0.0000
0.75	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.9545 0.0062	0.9655 0.0038	0.9754 0.0032	0.9844 0.0025	0.9925 0.0017	1.0000 0.0000
0.80	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.9738 0.0050	0.9815 0.0027	0.9882 0.0021	0.9944 0.0015	1.0000 0.0000
0.85	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.9868 0.0039	0.9917 0.0018	0.9960 0.0012	1.0000 0.0000
0.90	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.9947 0.0025	0.9975 0.0010	1.0000 0.0000
0.95	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	0.9988 0.0013	1.0000 0.0000
1.00	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	1.0000 0.0000

※1. 対角線に対して対称であるため，左下の値は空欄となっている
※2. 各行の上段は分布の平均，下段は分布の標準偏差である

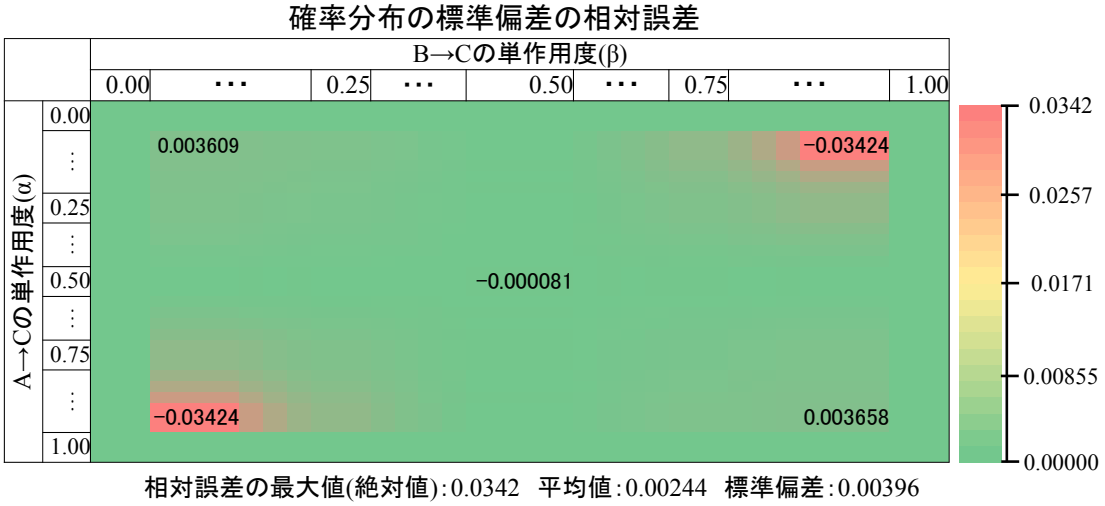

図 6　α，β に対する標準偏差の相対誤差のばらつき

次に，実際の発現量データの性質から，本手法の計算時間が実用的かどうかについて考察する．実際の発現量データは，2.2 節でも述べたように，タンパク質数は数百から数千，また，サンプル数はせいぜい数十から多くとも数百である．例えば，サンプル数が 500，タンパク質数が 10,000 における本手法の計算時間を見積もると，およそ 22 日 8 時間である．つまり，1ヶ月もかからずにタンパク質数 10,000（組合せ数が約 5.00×10^{11}）における相互作用が推定できると考えれば，十分に実用的であ

ると考えられる．

表 5　相互作用推定手法の計算時間

		タンパク質数(組合せ数)		
		500(6.21×10^7)	1,000(4.99×10^8)	1,500(1.68×10^9)
サンプル数	500	4分0秒	32分2秒	1時間50分20秒
	1,000	8分1秒	1時間4分24秒	3時間52分8秒
	5,000	40分8秒	5時間16分49秒	17時間30分34秒
	10,000	1時間17分26秒	10時間33分48秒	35時間0分5秒

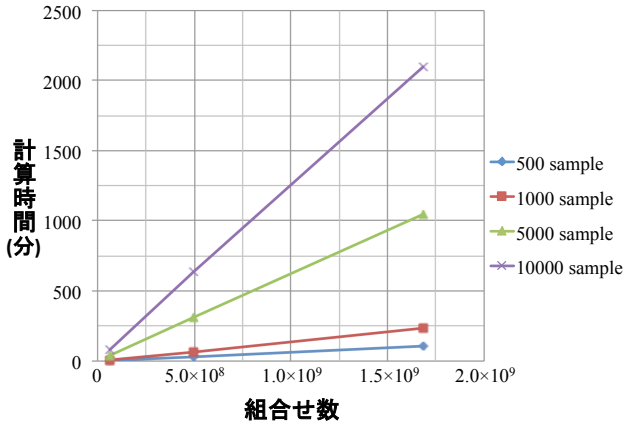

図 7　タンパク質数（組合せ数）と計算時間の関係

5. おわりに

本研究では，我々が提案した 3 タンパク質間の相互作用を推定する手法 [7] で用いたコンピュータシミュレーションによる統計分布の作成方法に代わって，解析的な方法で統計分布を作成する方法を提案した．提案手法の性能評価として，複合的作用がない場合の総作用度の分布の作成をコンピュータシミュレーションによる作成方法と提案手法とで分布の作成時間を比較した結果，提案手法の方が圧倒的に速いことがわかった．また，２つの方法で作成した分布精度を比較した結果，分布の平均と標準偏差の相対誤差が実用的に問題のない範囲で十分に小さい事がわかった．また，相互作用推定手法全体の計算時間を見積もったところ，十分実用的であることがわかった．

今後の課題として，現在の手法ではタンパク質の発現量を決定する閾値がすべてのタンパク質で一律であるが，これをタンパク質毎に設定できるようにすることで，より柔軟に 3 タンパク質間の相互作用を推定できる手法を提案したい．また，本手法で推定した 3 タンパク質間の相互作用の中に，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るかを確認し，本手法の実用性を裏付けたい．

文　献

[1] Liang, S., Fuhrman, S., Somogyi, R, “REVEAL, a General Reverse Engineering Algorithm for Inference of Genetic Network Architectures”, In: Proc. of Pacific Symosium on Biocomputing ’98, pp18–29, 1998

[2] Shmulevich, I., Dougherty, E.R., Kim, S., Zhang, W, “Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks”, Bioinformatics **18**(2), pp261–274, 2002

[3] Friedman, N., Linial, M., Nachman, I., Pe ’er, D, “Using Bayesian Networks to Analyze Expression Data”, Journal of Computational Biology **7**(3/4), pp.601–620, 2000

[4] D. Husmeier, “Sensitivity & Specificity of Inferring Genetic Regulatory Interactions From Microarray Experiments with Dynamic Bayesian Networks”, Bioinformatics, Vol. 19, No. 17, pp. 2271–2282, 2003

[5] Zhang, J., Ji, Y., Zhang, L, “Extracting Three-way Gene Interactions from Microarray Data”, Bioinformatics **23**(21), pp2903–2909, 2007

[6] Kayano, M., Takigawa, I., Shiga, M., Tsuda, K, “Efficiently Finding Genome-wide Three-way Gene Interactions from Transcript- and Genotype-data”, Bioinformatics **25**(21), pp2735–2743, 2009

[7] Fujiki, T., Inoue, E., Yoshihiro, T., Nakagawa, M, “Prediction of Combinatorial Protein-Protein Interaction from Expression Data Based on Conditional Probability”, In: Protein-Protein Interactions - Computational and Experimental Tools, InTech Web Press, pp131–146, 2012

[8] Nagai, K., Yoshihiro, T., Inoue, E., Ikegami, H., Sono, Y., Kawaji, H., Kobayashi, N., Matsuhashi, T., Ohtani, T., Morimoto, K., Nakagawa, M., Iritani, A., Matsumoto, K.: Developing an Integrated Database System for the Large-scale Proteomic Analysis of Japanese Black Cattle. Animal Science Journal **79**(4) (2008) (in Japanese)

[9] Hinkley, D.V, “On the Ratio of Two Correlated Normal Random Variables”, Biometrika **56**(3), pp635–639, 1969