基準要素集合を用いたたんぱく質の発現量データのスケーリング手法

坊木　好彦† 　井上　悦子†† 　吉廣　卓哉†† 　中川　優††

† 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 番地
†† 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: †s101048@sys.wakayama-u.ac.jp , ††{etsuko,tac,nakagawa}@sys.wakayama-u.ac.jp

あらまし　近年，たんぱく質の発現量を測定した発現量データから，たんぱく質の相互作用を推定する研究が盛んに行われている．発現量を測定する際に混入する様々な誤差や偏りを発現量データから取り除く操作を正規化と呼ぶ．本研究では，発現量のスケールを補正する “スケーリング” と呼ばれる正規化を対象とし，従来手法よりも高精度にスケーリングする手法を提案する．提案手法では，対象となるデータから，発現量の安定したたんぱく質の集合を自動的に抽出し，これら基準要素集合の発現量を基準として全体の発現量を正規化する．また，局所探索法を提案手法に適用し，提案手法を高速化する．従来の正規化手法との精度比較を通じて提案手法を評価する．

キーワード　たんぱく質，発現量，正規化，スケーリング，局所探索法

Normalization for Protein Expression Data of Using a Standard Element Set

Yoshihiko BOUKI†, Etsuko INOUE††, Takuya YOSHIHIRO††, and Masaru NAKAGAWA††

† Graduate School of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama-shi, Wakayama, 640–8510, Japan
†† Faculty of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama-shi, Wakayama, 640–8510, Japan
E-mail: †s101048@sys.wakayama-u.ac.jp , ††{etsuko,tac,nakagawa}@sys.wakayama-u.ac.jp

Abstract　In recent years, there are many trials to find interaction of proteins from protein expression data. In this process, we perform an operation called normalization to remove technical or experimental bias fromthe expression measurements. In this paper, we propose a new higher-performance scaling method, which is one of the normalization methods. In our proposal, we calculate a set of proteins which expression levels are relatively stable and normalize the whole data based on the set. We also evaluate the proposed method by comparing performance with traditional scaling methods.

Key words　Protein, Expression Data, Normalization, Scaling, Local Search

1. はじめに

生命現象が遺伝子や mRNA，たんぱく質等の相互作用により支えられていることが一般的に知られる事実となり，これらが生体内でどのように機能するのかを解明する研究が全世界で急速に進んでいる．この中で，調査サンプル中の遺伝子やたんぱく質の発現量を測定し，異なる性質を持つサンプル間（例えば，病気サンプルと通常サンプル）で比較することで，その性質に関連する遺伝子やたんぱく質を推定する手法が頻繁に用いられている．近年では網羅的に発現量を測定する方法として，遺伝子発現量はマイクロアレイ，たんぱく質発現量はプロテインアレイや 2 次元電気泳動と呼ばれる，大量の遺伝子／たんぱく質（以下，たんぱく質で統一する）の発現量を一度に測定できる技術が発展し，また手軽に利用できるようになった．これにより，多くのサンプルのたんぱく質発現量を測定し，その発現量データをコンピュータで分析することにより，たんぱく質の相互作用を解明する研究も盛んになってきた．この種のデータ解析は実験時に混入する誤差や偏りに非常に敏感であり，正規化を行うことでこれらの誤差や偏りによる影響を最小化することが必須である．

発現量データの正規化の目的は，実験時に混入する種々の誤差や偏りの影響を発現量データから取り除き，サンプル本来の発現量となるように補正することである．実験は手作業で行われ，正確かつ信頼できる実験結果を出すためには熟練を要する

ため，混入する誤差や偏りの量には実験者の個人差も反映されることがある．発現量の測定においては，サンプルの準備，マイクロアレイや二次元電気泳動の操作，或いは染色過程等，各過程で特有の性質を持つ誤差や偏りの混入が考えられる．一方で，発現量を測定するための色素の性質による誤差など，実験材料や器具による誤差も含まれる可能性がある．要するに，様々な性質の誤差や偏りに対する補正が必要となるが，これらに対応していくつかの正規化手法が実用されている．

本研究ではその中で，発現量のスケールを補正する,「スケーリング」と呼ばれる正規化手法を扱う．異なるサンプルの発現量を測定する場合，染色時の色素濃度や染色時間等の要因により，読みとられた発現量のスケールに差異が生まれる．本研究では，比較的発現量が安定したたんぱく質集合を計算し，この基準要素集合を用いて正規化することで，従来手法よりも精度の高いスケーリング手法を新たに提案し，局所探索法の導入による提案手法の高速化を行い，評価を行ったのでこれを報告する．

本稿の構成を以下に示す．2 章では既存の正規化手法について説明する．3 章では基準要素集合に基づいた新たな正規化手法を提案する．4 章では局所探索法の導入による提案手法の高速化を行う．5 章で人工データおよび実データに適用することで評価を行い，6 章でまとめる．

2. 発現量データの正規化

2.1 既存の正規化手法

発現量測定実験の方法，及び補正したい誤差の種類に応じて，いくつかの正規化手法が提案されている．ある種の正規化手法は，染色色素の特性による誤差を補正することを試みている．マイクロアレイや二次元電気泳動では，遺伝子やたんぱく質を色素で染色し，その色の強度を画像解析により読み取るのが一般的である．しかし，色素により発現強度に応じて着色効率が異なる，或いは時間とともに色素が減衰することがあり，この影響を取り除く方法が求められる．この補正のために用いられる方法の一つに Lowess 正規化がある．Lowess 正規化は，2 つのマイクロアレイを用いた発現量データに適用するのが一般的で，各マイクロアレイの発現強度に応じた発現量の偏りを，ノンパラメトリック回帰の一手法である Lowess 法を用いて補正する [1]．また，3 つ以上のマイクロアレイを用いた発現量データに対しては，組合せ的な Lowess 正規化の適用を繰り返すことで発現量データ全体を正規化する手法も提案されている [2]．

また，各マイクロアレイの発現量の分布は等しくなるはずであるという仮定に基づいて，分布を乱すような誤差を補正する Quantile 正規化が提案されている [2]．Quantile 正規化は，まず各マイクロアレイ内で遺伝子の発現量の順位を求め，次にマイクロアレイ間で同順位の発現量の平均値を求める処理により，分布の平滑化を行う．

一方，マイクロアレイ実験や二次元電気泳動に用いられたサンプル量の誤差や，色素の投与量，着色時間等に起因するスケール誤差を補正するスケーリングと呼ばれる正規化も頻繁に適用される．スケーリングは本研究で対象とする正規化の種類

表 1　たんぱく質の発現量データ

	たんぱく質					
サンプル	$j=1$	$j=2$	$j=3$	$j=4$	$\ldots$	J
$i=1$	10.4553	18.1541	3.70123	42.1293	$\ldots$	$\ldots$
$i=2$	7.28141	12.7985	4.85014	12.0288	$\ldots$	$\ldots$
$i=3$	32.8747	23.8471	7.78416	31.5857	$\ldots$	$\ldots$
$i=4$	26.1472	31.6816	10.3989	23.9821	$\ldots$	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
I	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\ddots$

である．スケーリングに含まれる正規化法としては，

i) 内部標準を用いた正規化 [3],
ii) グローバル正規化 [3],

などが良く用いられるが，これらについて次節で紹介する．

2.2 スケーリング正規化

スケーリングの一手法として，内部標準と呼ばれるたんぱく質を基準にスケールを補正する手法が知られている [3]．内部標準となるたんぱく質は，常に一定の発現量を持つと仮定される．このため，人為的にサンプルに導入される場合や，ハウスキーピングタンパク質と呼ばれる，細胞の維持に必須で常に安定して発現していると見なされるたんぱく質が用いられる．しかし内部標準たんぱく質の発現量が変動してしまうこともあり，また内部標準たんぱく質への実験誤差が全体に大きく影響してしまう問題がある．

この手法とは対照的に，グローバルスケーリングと呼ばれる手法が提案されている [3]．これは，サンプル内の全たんぱく質の発現量の平均値が一定であると仮定して，各サンプルの発現量を線形に補正する．平均値ではなく，中央値を用いる場合もあり，また発現量の上下位 2%を取り除いたうえで処理することもある．グローバルスケーリングの欠点は，たんぱく質数が比較的少ない場合に発現量の大きさ，ばらつきともに大きいたんぱく質の影響を受けやすいことである．一般にたんぱく質の発現量解析では，遺伝子の発現量解析よりも要素数が少ないため，誤差が大きくなりやすい．

本研究では，グローバルスケーリングのように全てのたんぱく質を用いて正規化するのではなく，比較的発現量が安定したたんぱく質（要素）の集合を自動的に検出し，これら基準要素集合の発現量の平均を用いて正規化することで，要素数が少ない場合にもより精度の高い正規化手法を提案する．

3. 基準要素集合を用いたスケーリング手法

3.1 たんぱく質の発現量データ

本研究で扱うたんぱく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，含まれる各たんぱく質の発現量が数値として表されている．発現量データの例を表 1 に示す．サンプルを $i(=1,2,3,\ldots,I)$，要素を $j(=1,2,3,\ldots,J)$ とおいたとき，サンプル i における要素 j の発現量を実数 $x_{ij}(>0)$ で表す．二次元電気泳動によりたんぱく質の発現量測定を行う場合には，たんぱく質数は数

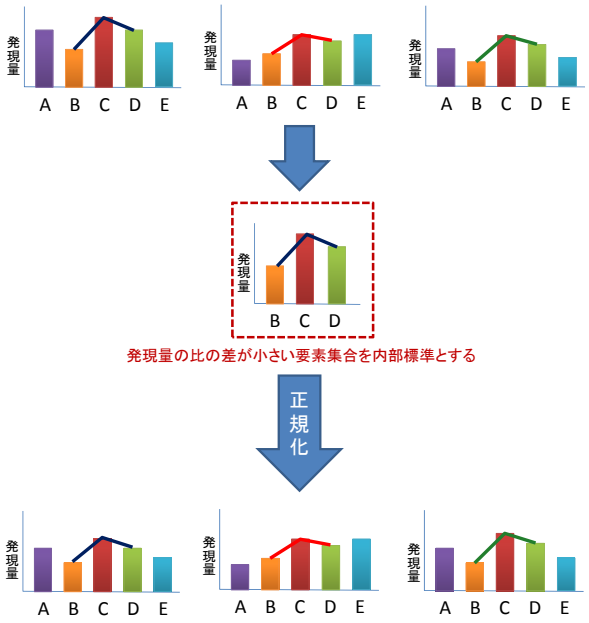

図 1　基準要素集合の抽出とスケーリング

百〜数千，また，実験は熟練を要するうえ手間もかかるため，サンプル数も数十程度の場合が多い．一方，マイクロアレイで遺伝子の発現量を測定する場合には，遺伝子数（要素数）J は数千〜数万の規模になる場合があり，サンプル数も多い時には数百程度の場合がある．

3.2　提案手法のアイデア

提案する正規化手法は，入力された要素の中から比較的発現量が安定した要素の集合を自動的に抽出し，この基準要素集合の発現量を用いてスケーリングを行うものである．これは，内部標準を用いたスケーリングとグローバルスケーリングそれぞれの利点を生かした方法と言える．グローバルスケーリングでは，発現量の変動が大きい要素の影響で精度が落ちる可能性があるが，提案手法では「内部標準」となり得る要素を自動的に抽出することで，変動が比較的小さい要素のみを用いてスケーリングを行う．

しかしここで，サンプル間の発現量のスケールが異なる状況で発現量が安定した要素を抽出する必要があることに注意が必要である．本研究ではこの問題を，基準要素集合内で発現量の比がほぼ等しい場合，つまり「比の差」が小さい場合に，これらの要素の発現量が安定していると見なすことで解決する．つまり，図 1 に示すように，発現量の比の差が小さい要素 B, C, D を発見することが必要となる．「比の差」をうまく定義し，これをできるだけ小さくする基準要素集合を抽出することが本提案の鍵となるが，この説明は 3.4 節に譲る．

3.3　提案手法の手順

本節では，基準要素集合を用いた正規化手法の手順を説明する．以下に，手順の概略を示す．

i)　基準要素集合の候補として，要素の全組合せを作成する．

ii)　各候補に対して，「比の差」を求める．

iii)　「比の差」が最小になるものを基準要素集合とする．

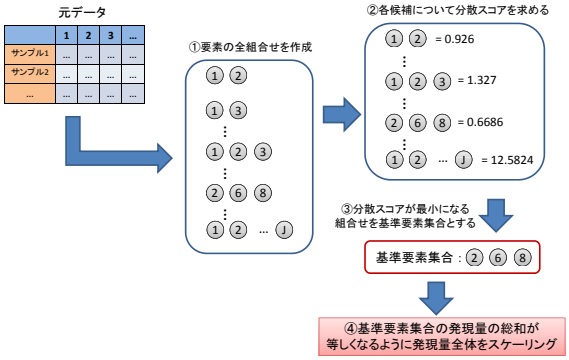

図 2　提案手法の概要

iv)　基準要素集合の発現量の平均が等しくなるようにスケーリングする．

上記の手順を 2 を用いて具体的に説明する．まず，基準要素集合を求めるために，全ての要素の組合せを作成し，それぞれについて 3.4 節で述べる方法で「比の差」を計算する．この値が最小になる組合せを基準要素集合として決定し，最後にグローバルスケーリングと同様の方法で，但し基準要素集合のみを用いて，サンプルあたりの発現量の平均値が等しくなるようにスケーリングを行う．

3.4　基準要素集合の選択

本節では，基準要素集合の選択問題を定式化する．まず，3.2 節で述べた「比の差」を定義する．発現量データではサンプル毎のスケールが異なることから，「比の差」は，サンプルのスケールを様々に変化させた場合の，発現量のばらつきの最小値と見ることができる．ここでばらつきは，各要素について，各サンプルの発現量とその要素の発現量の平均の差の二乗，即ち分散とすれば十分であろう．つまり，各組合せに対して，スケールを様々に変化させ，要素毎に分散を求め分散の和の最小値をとるスケールを決定する．この値をその組合せの分散スコアとする．しかし，要素間で発現量の大きさに差があることから，このままでは，発現量の大きい要素ほど分散スコアへの影響が大きくなる．そこで，発現量の大きさが影響しないように，上記の計算を対数領域で行うこととする．以上より，本研究では，各サンプル i のスケールを定める変数を α_i としたとき，以下の問題 RatioDiff の最適化関数 f の最小値を，求める分散スコアとする．

問題 RatioDiff

任意の j に対して，

$$\bar{X}_j^{(log)} = \bar{X}_j^{(\alpha,log)}$$

が成立する制約の下で次の目的関数 $f(\alpha_1, \alpha_2, \ldots, \alpha_I)$ を最小化する $\alpha_1, \alpha_2, \ldots, \alpha_I$ を定める．

$$f(\alpha_1, \alpha_2, \ldots, \alpha_I) = \frac{1}{J}\sum_{i=1}^{I}\sum_{j=1}^{J}(\log \alpha_i x_{ij} - \bar{X}_j)^2$$

但し，

表 2　**アルゴリズム ComptRatioDiff-Sub**

1	**for** $i = 1$ to I
2	$\sum_{j=1}^{J}(\log \alpha_i x_{ij}) = 0$ となるような α_i を定める
3	**end for**

$$\bar{X}_j^{(log)} = \frac{1}{I}\sum_{i=1}^{I}\log x_{ij},$$

$$\bar{X}_j^{(\alpha,log)} = \frac{1}{I}\sum_{i=1}^{I}\log \alpha_i x_{ij}$$

とする．　□

ここで制約 $\bar{X}_j^{(log)} = \bar{X}_j^{(\alpha,log)}$ は，$\alpha_1, \alpha_2, \ldots, \alpha_I$ が唯一に定まるための制約であり，α を乗じた前後で各要素についての発現量の平均が等しいことを意味する．但し，唯一解を持つためには α_i のうち 1 つが固定されていれば良いので，それよりも少し厳しい条件になっている．

3.5　基準要素集合の選択アルゴリズム

本節では，3.4 節で定式化した問題を解くためのアルゴリズムを説明し，アルゴリズムが正しいことを証明する．

まず，問題 RatioDiff を解く前に，対数をとった発現量の平均が一定であるという制約を課した問題 RatioDiff-Sub を検討する．問題 RatioDiff-Sub を解くためのアルゴリズム ComptRatioDiff-Sub を表 2 に示す．

問題 RatioDiff-Sub

$\frac{1}{I}\sum_{i=1}^{I}\log x_{ij} = 0$ であると仮定する．任意の j に対して

$$\bar{X}_j^{(\alpha,log)} = \frac{1}{I}\sum_{i=1}^{I}(\log \alpha_i x_{ij}) = 0$$

が成り立つという制約の下で次の目的関数 $f(\alpha_1, \alpha_2, \ldots, \alpha_I)$ を最小化する $\alpha_1, \alpha_2, \ldots, \alpha_I$ を定める．

$$f(\alpha_1, \alpha_2, \ldots, \alpha_I) = \frac{1}{J}\sum_{i=1}^{I}\sum_{j=1}^{J}(\log \alpha_i x_{ij} - \bar{X}_j^{(\alpha,log)})^2$$

□

［定理 1］ アルゴリズム ComptRatioDiff-Sub は問題 RatioDiff-Sub の解を与える．

証明：アルゴリズム ComptRatioDiff-Sub が実行するように，$\sum_{j=1}^{J}(\log \alpha_i x_{ij}) = 0$ が成り立つように α_i を定める．このとき，

$$\log \alpha_i = -\frac{1}{J}\sum_{j=1}^{J}\log x_{ij}$$

が成り立つ．ここで，

$$\begin{aligned}\frac{1}{I}\sum_{i=1}^{I}\log \alpha_i x_{ij} &= \frac{1}{I}\sum_{i=1}^{I}\log x_{ij} + \frac{1}{I}\sum_{i=1}^{I}\log \alpha_i \\ &= \frac{1}{I}\sum_{i=1}^{I}\log x_{ij} - \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J}\sum_{j=1}^{J}\log x_{ij} \\ &= \frac{1}{I}\sum_{i=1}^{I}\log x_{ij} - \frac{1}{J}\sum_{j=1}^{J}\frac{1}{I}\sum_{i=1}^{I}\log x_{ij} \\ &= 0\end{aligned}$$

つまり，アルゴリズム ComptRatioDiff-Sub によって求めた α_i を乗ずる前後で，各要素における（対数をとった）発現量の平均は変化せず，

$$\bar{X}_j = \frac{1}{I}\sum_{i=1}^{I}(\log x_{ij}) = 0$$

が保証される．この条件の下で $g_i(\alpha_i) = \sum_{j=1}^{J}(\log \alpha_i x_{ij})^2$ とおくと，関数 f は，

$$f(\alpha_1, \alpha_2, \ldots, \alpha_I) = \sum_{i=1}^{I} g_i(\alpha_i)$$

と表すことができる．任意の $i, j (i \neq j)$ に対して，関数 $g_i(\alpha_i)$ と $g_j(\alpha_j)$ は独立であるため，アルゴリズム ComptRatioDiff-Sub で定めた α_i が $g(\alpha_i)$ を最小化するならば，アルゴリズム ComptRatioDiff-Sub で定めた α_i は関数 $f(\alpha_1, \alpha_2, \ldots, \alpha_I)$ を最小化する．

よって，証明の仕上げとして，アルゴリズム ComptRatioDiff-Sub で定めた α_i が $g_i(\alpha_i)$ を最小化することを示す．

$\alpha_i' = \log \alpha_i$, $x_{ij}' = \log x_{ij}$ とおくと，

$$\begin{aligned}g_i(\alpha_i) &= \sum_{j=1}^{J}(\log \alpha_i + \log x_{ij})^2 \\ &= \sum_{j=1}^{J}(\alpha_i' + x_{ij}')^2 \\ &= \sum_{j=1}^{J}({\alpha_i'}^2 + 2\alpha_i' x_{ij}' + {x_{ij}'}^2) \\ &= J{\alpha_i'}^2 + 2\alpha_i' \sum_{j=1}^{J} x_{ij}'\end{aligned}$$

微分すると，

$$g_i'(\alpha_i) = 2J\alpha_i' + 2\sum_{j=1}^{J} x_{ij}'$$

$g_i'(\alpha_i) = 0$ の時に $g_i(\alpha_i)$ が最小値をとり，そのときの α_i の値は，次の式を満たす．

$$\log \alpha_i = -\frac{1}{J}\sum_{j=1}^{J}\log x_{ij}$$

一方で，アルゴリズム ComptRatioDiff-Sub で定めた値からも同じ式が導かれるため，アルゴリズム ComptRatioDiff-Sub は $g_i(\alpha_i)$ を最小化する α_i を求めることが示された．　□

次に，アルゴリズム ComptRatioDiff-Sub を用いて問題 RatioDiff を解くためのアルゴリズムを検討する．表 3 にアルゴリズム ComptRatioDiff を示す．

アルゴリズム ComptRatioDiff は，問題 RatioDiff を RatioDiff-Sub に帰着して解く．アルゴリズム ComptRatioDiff-Sub の処理を行う前に，全ての要素について，対数領域での発現量の平均が一定になるように，発現量を平均値 β_j で除する．アルゴリズム ComptRatioDiff-Sub は平均値を保持するため，β で除した後に（α を計算してから）β を乗ずることで，問題 RatioDiff でも平均値を保持できる．

表 3　**アルゴリズム ComptRatioDiff**

1	**for** $j = 1$ to J
2	$\frac{1}{I}\sum_{i=1}^{I} \log \beta_j x_{ij} = 0$ となるような β_j を定める
3	**end for**
4	**for** $i = 1$ to I
5	$\beta_j x_{ij} = x'_{ij}$ とおくとき， $\sum_{j=1}^{J}(\log \alpha_i x'_{ij}) = 0$ となるような α_i を定める
6	**end for**

［定理 2］ アルゴリズム ComptRatioDiff は問題 RatioDiff の解を与える．

証明： β_j を定めると，x'_{ij} は $\frac{1}{I}\sum_{i=1}^{I} \log x'_{ij} = 0$ を満たす．よって，アルゴリズム ComptRatioDiff の 4〜6 行目は，アルゴリズム ComptRatioDiff-Sub の x_{ij} を x'_{ij} に置き換えたときの問題の解を与える．ここで，最適化関数の一部である式を変形すると，

$$
\begin{aligned}
&\log \alpha_i x_{ij} - \bar{X}_j^{(\alpha,log)} \\
&= \log \alpha_i \beta_j x'_{ij} - \frac{1}{I}\sum_{i=1}^{I} \log \alpha_i \beta_j x'_{ij} \\
&= (\log \alpha_i x'_{ij} + \log \beta_j) - (\frac{1}{I}\sum_{i=1}^{I} \log \alpha_i x'_{ij} + \log \beta_j) \\
&= \log \alpha_i x'_{ij} - \frac{1}{I}\sum_{i=1}^{I} \log \alpha_i x'_{ij} \\
&= \log \alpha_i x'_{ij} - \bar{X'}_j^{(\alpha,log)}
\end{aligned}
$$

となり，問題 RatioDiff と問題 RatioDiff-Sub の最適化関数は恒等であることがわかる．

また，アルゴリズム ComptRatioDiff-Sub より，

$$
\begin{aligned}
\bar{X}_j^{(\alpha,log)} &= \bar{X'}_j^{(\alpha,log)} + \log \beta_j \\
&= \bar{X'}_j^{(log)} + \log \beta_j \\
&= \bar{X}_j^{(log)}
\end{aligned}
$$

であり，アルゴリズム ComptRatioDiff の解は問題 RatioDiff の制約を満たしている．よって，アルゴリズム ComptRatioDiff が問題 RatioDiff を解くことが示された． □

4. 局所探索法の導入

4.1 提案手法の高速化

これまでに，最適に基準要素集合を選択するアルゴリズムを示した．しかし，要素集合の組合せすべてについて分散スコアを求めると，基準要素数が 5 個であっても組合せ数が膨大となり，相応の計算時間がかかる．基準要素数をさらに上げると組合せ数が指数関数的に増加するため現実的とは言えず，提案手法の高速化が必要となる．組合せ最適化問題について，高速化を目的としたアルゴリズムが多数提案されている．その中でも，単純かつ代表的であり，解の探索が高速である局所探索法を導入する．

4.2 局所探索法 [4]

局所探索法は，近傍を探索しながら解を求める近似解法の一つである．近傍とは，ある解が与えられたとき，その解の部分集合を持つ解の集合である．一般的な局所探索法の手順を以下に示す．

i) 初期解をランダムに生成する．
ii) 現在の解の近傍をすべて探索する．
iii) 近傍の中により良い解があればその近傍に移動し，ii) に戻る．
iv) 近傍の中により良い解が見つからなければ探索を終了する．

上記の手順を具体的に説明する．図 3 に示すように，まず，初期解をランダムに生成し，現在の解の近傍を探索する．近傍の中により良い解があるならばその近傍に移動し，その近傍を現在の解として探索を行う．探索を繰り返し，近傍の中により良い解が見つからなければ，現在の解を局所最適解として決定し探索を終了する．

局所探索法は，すべての解を求める厳密解法と比べ，高速に解を求めることが出来るが，局所探索法で得られた解は，最適性の保証をもたない．そこで，局所探索法を初期解をかえて繰り返し行い，その中から最良の解を選択することでより最適性の高い解が得られるようにする必要がある．

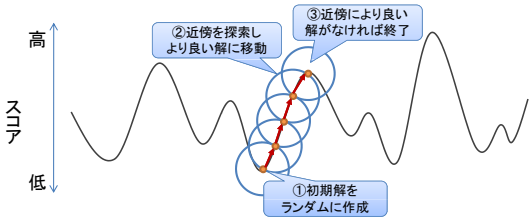

図 3　局所探索法

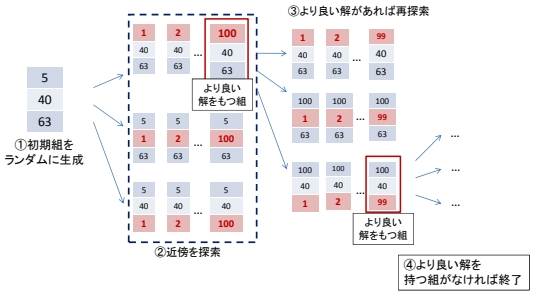

図 4　局所探索法を導入した提案手法の概要

4.3 提案手法への局所探索法の導入方法

本節では，提案手法への局所探索法の導入方法を説明する．図 4 に示すように，まず，要素の組合せをランダムに一組生成し初期解とする (①)．現在の組合せから要素を一つ入替えた組合せを近傍と定義し，すべての近傍に対して分散スコアを求める (②)．分散スコアが現在の組合せよりも小さい組合せがあれば，その組合せに移動し繰り返し探索を行う (③)．より小さい組合せが見つからなければ探索を終了し，現在の組合せを基準要素集合の候補とする (④)．この処理を十分に繰り返し，基準要素集合の候補の中で分散スコアが最小になる組合せを基準要素集合として決定する．

5. 評　　価

5.1 人工データを用いた評価実験

5.1.1 グローバルスケーリングとの性能比較

提案手法を人工的に作成したデータに適用することで，どの程度元のスケールに戻せるかを測定し，提案手法とグローバルスケーリングの精度と比較する．

本評価実験では、発現量データを次のようにモデル化し、これに従って人工データを生成する。発現量データはサンプル i、要素 j に対して、x_{ij} という発現量観測値を持つ。この発現量観測値にはいくつかの「偏り」要因が含まれていると仮定する。本研究では、まず実験誤差が混入し、その後に線形にスケール誤差が混入するモデルを考える。即ち、サンプル i、要素 j の真の発現量を $\check{x}_{ij}$ とすると、

$$x_{ij} = s_{ij}(\check{x}_{ij} + e_{ij})$$

と書ける．ここで、e_{ij} は実験誤差、s_{ij} はスケール誤差である．

人工データを生成する際には、$\check{x}_{ij}$ 及び e_{ij} は正規分布に従うと仮定する．また、s_{ij} が従う分布に実験結果は依存しないため、s_{ij} は一定範囲内でランダム値をとることとする．上記のモデルに従って、次のように人工データを生成した．人工データのサンプル数を 100，要素数を 100 とし、各要素に対してランダムに $m(1 \leqq m \leqq 100)$ を選び，平均が m，標準偏差が $\sigma(\frac{m}{10} \leqq \sigma \leqq \frac{m}{4})$ の正規分布に従うように発現量をランダム生成した．

ランダム seed を変えることで正解データを 10 セット生成し，各サンプルのスケールをランダムに変化させた．それぞれについて基準要素数を 5〜50 と変化させて提案手法を適用し，グローバルスケーリングの結果と精度を比較した．つまり，基準要素数を決めて基準要素集合の候補を網羅的に作成し，その中から基準要素集合を決定する．精度の比較のために，正規化後のデータと正解データの両方に対して，全発現量の平均が 1 になるように線形にスケールを変化させた後，各要素について正規化後の発現量の分散から正解データの発現量の分散を減算した値の絶対値をとり，この値の全ての要素についての総和を求めた．この値を**評価値**と呼ぶことにする．評価値が小さいほど，正規化後のデータのスケールを正解データのスケールに戻せていることになる．

5.1.2 結果と考察

用意した 10 セットの人工データのそれぞれについて，提案手法とグローバルスケーリングを適用させた時の評価値の平均を計算した結果を表 4 に示す．この結果によると，基準要素数が 50 の場合には提案手法の値が小さく，正規化後の発現量はグローバルスケーリングよりも提案手法の方が正解データに近いことがわかる．また，基準要素数が増えるとより正解データに近づくことがわかる．全体としてグローバルスケーリングよりも提案手法の方が正解データに近い値となる結果が得られた．

さらに，基準要素数を 5〜50 の間で変化させた場合とグローバルスケーリングの，評価値の推移の一例を図 5，図 6 に示す．縦軸が基準要素数毎の評価値，横軸が基準要素数の値であり，

表 4　実験結果 1

基準要素数	評価値の平均
10	3.4921
20	3.4147
30	3.3833
40	3.3648
50	3.3543
GS	3.3564

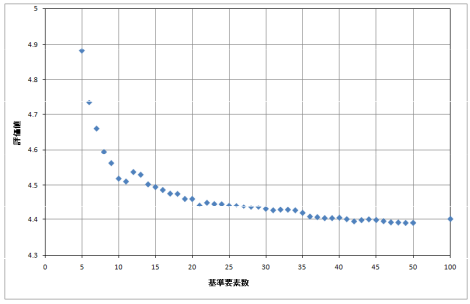

図 5　基準要素数を変化させた時の評価値の分布

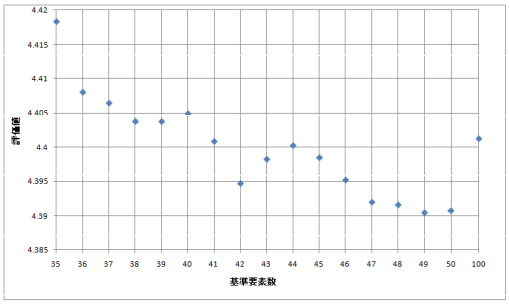

図 6　評価値の分布の詳細（基準要素数 35〜50 の分布）

基準要素数が 100 の値は，グローバルスケーリングの値を表している．これによると，基準要素数が増えるとより正解データに近づき，ある一定数を越えると正解データから遠ざかっている．正規化後の分散から正解データの分散を減じた値の絶対値の総和が大きくなるということから，必要以上にスケール誤差を補正していることがわかり，グローバルスケーリングは過矯正の傾向があることがわかる．一方で提案手法に関しては，求めた分散の差は概ね正解データよりも大きいものの，差の幅はグローバルスケーリングよりも小さくなっている．基準要素数を増やすことで正解データとの差の幅が小さくなり，精度が高くなることがわかる．以上より，提案手法はグローバルスケーリングよりも精度の高いスケーリング性能があることを確認できた．

5.2 実データを用いた評価実験

実データを用いて，提案手法に局所探索法を導入する前後で計算時間を計測し，局所探索法により計算時間をどの程度削減できたかを比較する．また，どの程度元のスケールに戻せるかを測定し，提案手法とグローバルスケーリングの精度と比較する．

5.2.1 データの準備

実験に用いるデータは，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たんぱく質発現量データを用いた．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っている場合もあり，複数枚のゲルが存在するサンプルもある．サンプル数は 254，ゲル数は 555，要素数は 879 である．得られたデータに対し欠損値が多いものを除外する処理を行った．これは，欠損値が多いゲルまたは要素を信頼性が低いものとし，欠損値が多いスポットを基準要素として選択した場合に，正しくスケーリング処理が行えない問題を防ぐための処理である．今回は，全ゲル数の 1 割以上が欠損値であるたんぱく質，全たんぱく質数の 1 割以上が欠損値であるゲルを除外した．また，実験回数が 1 回のサンプルを除外し，複数回実験しているサンプルについては，発現量の相関係数が最も高い 2 枚のゲルセットを選抜した．これより，サンプル数は 124，ゲル数は 248，たんぱく質数は 723 となった．以上の実データから，たんぱく質をランダムに 100 個選択したデータを作成し，このデータに提案手法を適用することで評価を行った．

5. 2. 2　局所探索法の導入結果

提案手法に局所探索法を導入する前後で計算時間を計測し，局所探索法により計算時間をどの程度削減できたかを比較する．比較するにあたり，提案アルゴリズムを C 言語により実装した．5. 2. 1 項で述べたデータを用いて，基準要素数を様々に変化させて提案手法（局所探索法を導入したものと導入していないもの）を適用し，それぞれの計算時間を比較した．つまり，局所探索法を導入した手法は，基準要素数を決めて 4. 3 節で説明した局所探索処理を 100 回繰り返し，その中から基準要素集合を決定する．局所探索法を導入していない提案手法は，基準要素数を決めて基準要素集合の候補を網羅的に作成し，その中から基準要素集合を決定する．それぞれの手法で基準要素集合を決定するまでの時間を計測した．実験に用いた計算機の性能は，OS：CentOS 5.4，プロセッサ：Xeon 2.40GHz，メモリ：16GB である．

それぞれの手法について計測した結果を図 7 と図 8 に示す．いずれも縦軸が計算時間，横軸が基準要素数の値である．この結果によると，局所探索法を導入する前は（図 7），基準要素数が 5 個の場合には約 1 時間，基準要素数が 6 個の場合には約 5 時間かかっている．それに対し，局所探索法を導入した場合は（図 8），基準要素数が 5 個の場合には数秒，基準要素数が 50 個の場合でも約 20 分しかかかっていない．これより，局所探索法を導入することで計算時間が大幅に短縮されたことを確認できた．

また，図 9〜図 11 は，基準要素数を 5 個，25 個，50 個と変化させた時の，1 回の局所探索で選択した基準要素集合の候補の分散スコアの分布である．いずれも縦軸が基準要素集合の候補の分散スコア，横軸が試行回数である．局所探索法を 100 回試行した時の，基準要素集合の候補としてどの分散スコアの要素集合を選択したかを表している．基準要素数が 5 個の場合は同じ分散スコアであった．これは，常に同じ要素集合を基準要素集合として選択していることを示す．しかし，基準要素数が 25 個や 50 個の場合，同じ要素集合が選択されるとは限らず，最適であるとはいえない基準要素集合を選択する場合があるという結果が得られた．つまり，試行回数が少ない時には最適解が得られない可能性があり，十分な回数の局所探索を試行する必要がある．

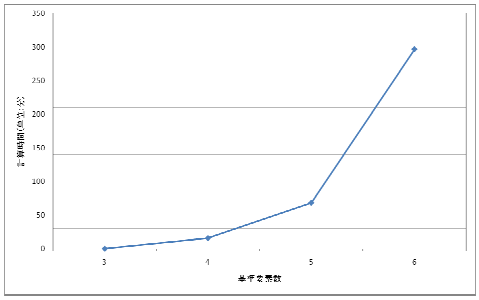

図 7　局所探索法を導入する前の計算時間

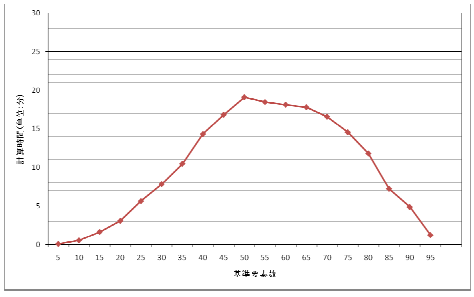

図 8　局所探索法を導入した場合の計算時間

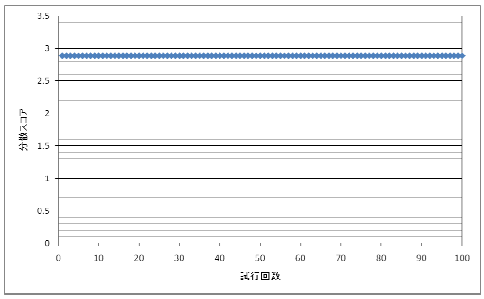

図 9　分散スコアの分布（基準要素数が 5 個の場合）

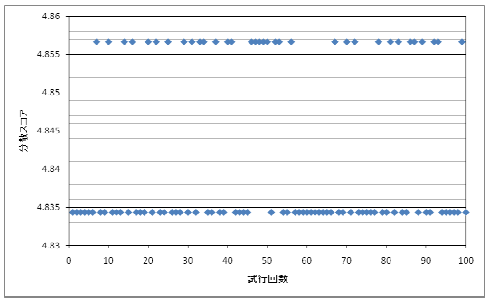

図 10　分散スコアの分布（基準要素数が 25 個の場合）

5. 2. 3　既存手法とのスケーリング性能の比較

提案手法を 5. 2. 1 項で述べた発現量データに適用することで，提案手法とグローバルスケーリングの精度と比較する．評価実験にあたっては，提案アルゴリズムおよびグローバルスケーリングを C 言語により実装した．

人工データによる評価実験と異なり，実データには正解データが存在しないため，同一サンプルから得られた 2 枚のゲルの発現量について，要素毎に発現量の比を求めることで評価を行

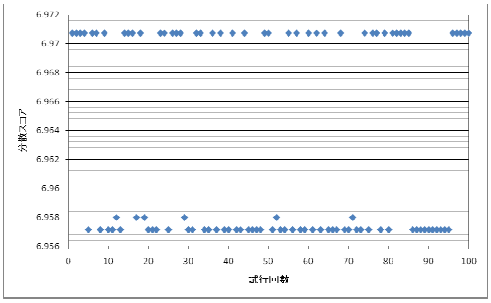

図 11　分散スコアの分布（基準要素数が 50 個の場合）

表 5　実験結果 2

基準要素数	再現性の平均	再現性の分散
20	1.1691	0.0583
40	1.1363	0.0408
60	1.1244	0.0356
80	1.1177	0.0332
90	1.1158	0.0324
95	1.1149	0.0321
GS	1.1146	0.0319

う．2 枚のゲル間の発現量の比が 1 に近づくほど，スケールによる誤差を補正していることになると考えた．各サンプルの 2 枚のゲルを比較し，要素毎に “（発現量の大きい方の値）/（発現量の小さい方の値）” を求める．この値を**再現性評価値**と呼び，すべてのサンプルについて再現性評価値を求めた．

5. 2. 4　結果と考察

表 5 は，5. 2. 3 項で述べた値の平均及び分散を，基準要素数を 5〜95 と変化させた場合とグローバルスケーリングの場合で比較したものである．この結果によると，基準要素数がどの場合であってもグローバルスケーリングの値よりも大きくなっており，提案手法よりもグローバルスケーリングの方が良い値となる結果が得られた．

評価結果について考察する．今回の評価実験では，同サンプルの 2 枚のゲル間についての発現量の再現性を求めることで正規化の精度を比較しているが，グローバルスケーリングは，すべての要素の発現量が等しいと仮定して要素全体を対象にスケーリングしているため，すべての要素の発現量の分散を小さくしている．それに対し提案手法は，基準要素集合内の発現量が等しいと仮定して一部の要素を対象にスケーリングしている．つまり，基準要素集合内の発現量の分散は小さくしているが，それ以外の要素については発現量の分散を大きくする傾向がある．

スケーリングにより再現性評価値を小さくする（再現性を良くする）ことは，同じサンプルから得られた発現量データからスケール誤差を取り除き，実験誤差のみを残すことであると理解できる．しかし，5. 1. 2 項で述べたように，グローバルスケーリングは過矯正の傾向があり，スケール誤差だけでなく，実験誤差に起因する分散や真の発現量の分散も縮小する．実験結果からはグローバルスケーリングの方が再現性評価値が低くなる結果が得られたが，これは，グローバルスケーリングの方が実験誤差に起因する分散をより縮小しており，これにより見かけ上は再現性を向上したように見えるためである．つまり，グローバルスケーリングは必ずしもスケール誤差を正しく修正しておらず，真の発現量に戻すことをスケーリングの目的とするのであれば，本実験の結果は両者の精度を正確に比較することはできない．正規化の精度を正しく比較できる指標を検討する必要がある．

6. おわりに

本研究では，基準要素集合を用いた発現量データの新たなスケーリング手法を提案した．また，局所探索法を提案手法に導入し，提案手法の高速化を行った．実際のたんぱく質の発現量の分布を想定した人工データに適用することで，提案手法とグローバルスケーリングの精度を比較した．その結果，提案手法の方が精度が高いスケーリング性能である結果が得られた．実際のたんぱく質発現量データにも適用し評価を行った結果，グローバルスケーリングは必ずしもスケール誤差を正しく修正していないという知見が得られた．今後，正規化の精度を正しく比較できる指標を検討する必要がある．

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を受けたものである．

文　　献

[1] W. S. Cleveland and S. J. Delvin, “Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting,” Journal of the American Statistical Association, **83**, pp. 596–610, 1988.

[2] B. M. Bolstand, R. A. Irizarry, M. Astrand and T. P. Speed, “A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias,”, Bioinformatics, **19**(2), pp. 185–193, 2003.

[3] 村上康文, 古谷利夫, “バイオインフォマティクスの実際”, 講談社, 2003.

[4] 久保幹雄，J. P. Pedroso，“メタヒューリスティクスの数理”，共立出版，2009.

[5] 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.