社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－BIO－2 (4)
2005／10／7

異なる時系列マイクロアレイデータの比較のためのデータ補正方式

萩本健二[†,††] 宮崎純[†] 金谷重彦[†]
小笠原直毅[†] 植村俊亮[†]

異なる時系列マイクロアレイデータは，実験のサンプリング時間およびリファレンス値が異なるため，得られた時系列発現プロファイルを直接比較することができなかった．本論文では，ダイナミックタイムワーピング法ならびに発現強度の散布図を利用し，異なる時系列マイクロアレイデータを直接比較するためのデータ補正方法を提案する．

A Data Correction Scheme for Comparing Different Time-Series Microarrays

Kenji Hagimoto [†,††] Jun Miyazaki [†] Shigehiko Kanaya [†]
Naotake Ogasawara [†] and Shunsuke Uemura[†]

In this paper, we propose a method to compare multiple time-series microarray data obtained from different experiments. In order to compare these data which have different time series, time duration is corrected by applying Dynamic Time Warping algorithm (DTW) on two time-series gene expressions. We also describe an approach to directly compare gene expressions by normalizing their referential values.

1. はじめに

1953 年に J.Watson と F.Crick が DNA は二重螺旋構造であるという画期的な発見[3]をしたのち，ゲノム構造の解析技術は急速に発展した．1980 年代に始まった各種生物の全塩基配列を決定しようというゲノムシークエンスプロジェクトにより，現在では約 200 種類もの生物のゲノムが解読されている．このようなゲノムの構造解析が進んで得られる主な情報は，(1) 配列情報 (塩基配列，アミノ酸配列) と (2) 遺伝子位置の情報である．しかしながら，ゲノムの構造解析が終わっても，遺伝子の働きなど，ゲノムに書かれた情報の意味がすぐに理解できるわけではない．実際，これまでにゲノムが決定された生物種において，どのような働きをしているのか全く分かっていない遺伝子が数多く残っている．

そこで，現在の分子生物学はポストゲノムシークエンス時代と呼ばれるようにゲノムの機能解析に研究の焦点が当てられており，さまざまな実験手法が開発されつつある．その結果，ゲノム変異情報，遺伝子発現情報 (トランスクリプトーム)，タンパク質発現情報 (プロテオーム)，ならびにタンパク質間相互作用情報に関するデータが大量に産出されつつある．ゲノムの機能解析を行う主力な実験技術のひとつとして，マイクロアレイがあげられる．マイクロアレイは一度に数千から数万の遺伝子発現の状態を網羅的に解析する技術である．一枚のマイクロアレイを解析する研究はマイクロアレイの基本であり，すでに多くの研究がなされてきている．近年，その技術発展は著しく，時系列発現プロファイル解析のような複数のマイクロアレイを対象とする実験も増加している[10),11)]．実験で得られるマイクロアレイデータは共同研究者やオンラインデータベース[2)]などを通じて，その解析のために広く利用可能になってきている[4)]．このため，今後は複数のマイクロアレイデータを組合せ，比較する研究が重要である[6)]．

異なる実験から得られた時系列遺伝子発現プロファイルは，サンプル数とサンプル時刻が同じではないという問題が考えられる．また，現在のマイクロアレイでは mRNA の発現量を定量的に測定することができ

† 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, NAIST
†† 現在，日本ヒューレットパッカード株式会社
Currently, with Hewlett-Packard Japan, Ltd.

ないため，リファレンスの異なる二つの遺伝子発現プロファイルの発現量を直接比較することができないという問題もある．

本論文では，同じ生物種を用いて異なる実験条件のもと得られた複数の時系列遺伝子発現プロファイルや，異種生物において類似した実験条件のもと得られた時系列発現プロファイルを対象として，二つの時系列マイクロアレイデータにDTWを適用することで時間軸を正規化するとともに，同じ生物種間であれば同じ遺伝子ペアの発現類似性，異なる生物種間であればオーソログ遺伝子ペアの発現類似性を用いて遺伝子ペア群の発現プロファイルを解析することでリファレンスを補正し，異なる時系列マイクロアレイデータを直接比較するためのデータ補正方式を提案する．さらに，提案手法を8種類の異なる培養条件で増殖させた際の時系列発現プロファイルに対して行い，提案手法の有効性を示す．

2. マイクロアレイ

ある遺伝子が生体内で機能しているか否かを知るためには，遺伝子発現という過程を経て，その遺伝子にコードされているタンパク質が合成されたかどうかを調べればよい．しかしながら，タンパク質は不安定であるため，一般的に扱いが難しい．そのため，タンパク質が合成される前段階であるmRNAの発現状態を観測する方法として開発された技術がマイクロアレイである．mRNAの発現状態を解析する手法としては，従来ノーザンブロット法やディファレンシャルディスプレイ法が用いられてきたが，これらの方法による解析遺伝子数は一度に100前後に過ぎない．一方，マイクロアレイを用いると数千から数万種といった規模の遺伝子発現を同時に観察すること可能である[14]．

マイクロアレイにより遺伝子の発現量を測る際は，目的実験のターゲット細胞と対照実験のリファレンス細胞由来の2種類のサンプル細胞を用意する．それぞれを異なる蛍光色素 (通常，Cy3:緑，およびCy5:赤) で標識し，同一マイクロアレイ上で競合的にハイブリダイズさせ，各プローブDNA (各スポット) のシグナル (蛍光強度) をスキャンする．蛍光色素Cy3，Cy5はそれ自体には色はないが，スキャン後，ソフトウェアにより強度の大きさに応じてそれぞれ緑色，赤色を着色する．したがって，ある遺伝子がリファレンス細胞ではほとんど発現せず，ターゲット細胞では過剰に発現しているような場合，その遺伝子に対応するスポットは赤となる．緑のスポットはリファレンス細胞でのみ発現している遺伝子であり，黄色のスポットは2つの細胞で同程度，共に発現しているものである．両方の細胞で共に発現がない遺伝子に関してはスポットは黒となる[14]．

ある遺伝子GeneAの発現量 $GeneExprA$ を表すには，ターゲットのシグナル T とリファレンスのシグナル R を用いた以下の式が一般的に用いられている．

$$GeneExprA = \log \frac{T}{R}$$

2.1 時系列マイクロアレイ

一枚のマイクロアレイ解析は，目的実験と対照実験の間で有意な発現量変化を持った遺伝子の探索 (例えば，正常な大腸の上皮細胞と大腸ガン細胞の発現パターンの違い[13]) や，目的実験の細胞全体の遺伝子発現量による特徴付けなど多くの研究がなされている．

一方，時系列で遺伝子発現を観測する場合，細胞あるいは組織レベルで個々の遺伝子がどのように協調的に機能しているかを網羅的に把握することができる[12]．時間 t_n における遺伝子GeneAの発現量 $GeneExprA(t_n)$ は，ある時刻 t_r でサンプルされた一定のリファレンス $R(t_r)$ を用いて，以下の式で示されるように，ターゲット $T(t_n)$ との発現比率の対数をとったものとする．

$$GeneExprA(t_1) = \log \frac{T_1}{R(t_r)}$$
$$GeneExprA(t_2) = \log \frac{T_2}{R(t_r)}$$
$$\cdots\cdots\cdots \quad \cdots\cdots$$
$$GeneExprA(t_n) = \log \frac{T_n}{R(t_r)}$$

3. 時系列マイクロアレイデータのデータ補正

本節では，異なる時系列マイクロアレイ実験A，Bから得られた二つの遺伝子発現プロファイルの時間軸を正規化して比較可能にし，さらに，リファレンスを補正して遺伝子発現量を直接比較する手法について述べる．

3.1 データ補正方式の概略

まず，二つの時系列マイクロアレイ実験A，Bから，BLASTを用いて配列相同性が高い遺伝子ペア群 $GenePairs(X, X')$ を抽出する．実験AとBで同じ生物が用いられている場合は，単純に同じ遺伝子を

$GenePairs(X, X')$ とする．

次に，配列情報で選別された遺伝子ペア群 $GenePairs(X, X')$ に対して，二つの波形の類似度を測る Dynamic Time Warping (DTW) アルゴリズム [5),9)] を適用し，時間軸の正規化と時系列発現プロファイルの類似度スコアの算出を行う．類似度スコアを用いることで，配列情報に加えて，さらなる遺伝子ペアのフィルタリングが可能である．

最終的に，時間軸正規化後の遺伝子発現量を $x-y$ 平面上にプロットして散布図を描き，相関係数を求めることで実験 A と B における時系列発現プロファイルの網羅的な比較ができる．また，実験 A と B でリファレンスが異なる場合，散布図の傾向からリファレンスを補正するための関数が得られ，発現量の直接比較が可能となる (図 1 参照)．

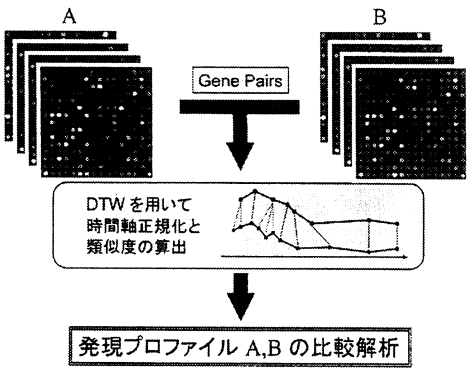

図 1　データ補正の処理手順

3.2　遺伝子ペアの選別

実験 A で用いた全遺伝子セットを $AllX = (x^1, x^2, ...x^a)$ とし，発現量のサンプル数を n とする．同様に，実験 B についても全遺伝子セットとサンプル数を，$AllX' = (x'^1, x'^2, ..., x'^b)$ と m とする．実験 A における時系列マイクロアレイデータの例を表 1 に示す．

表 1　実験 A の時系列発現プロファイル

X	t_1	t_2	...	t_{n-1}	t_n
x^1	0.12	0.07	...	1.99	-0.12
x^2	2.11	1.88	...	1.31	-0.21
.....	....	....	...	....	....
x^{a-1}	1.88	1.19	...	0.32	-0.02
x^a	2.1	1.42	...	-0.79	-1.02

実験 A で用いられた遺伝子 a 個，実験 B で用いられた遺伝子 b 個から，遺伝子ペアを s 個を選別する．実験 A と B が同じ生物種で行われた実験か，異なる生物種で行われたかによって，以下のように場合分けされる．

(1)　同じ生物種の場合: 同じ遺伝子をペア (x, x') とする．

(2)　異なる生物種の場合:BLAST を用いて，配列相同性が高い遺伝子をペア (x, x') とする．

結果として得られた s 個の遺伝子ペア群 $GenePairs(X, X')$ は，

$$GenePairs(X, X') = \{(x^1, x'^1), (x^2, x'^2), ..., (x^s, x'^s)\}$$

となり，ある遺伝子ペア (x^p, x'^p) の時系列発現量 $\{GeneExpr(x^p), GeneExpr(x'^p)\}$ は，次のように表される．

$$GeneExpr(x^p) = \{x_1^p, x_2^p, \cdots, x_{n-1}^p, x_n^p\}$$
$$GeneExpr(x'^p) = \{x'^p_1, x'^p_2, \cdots, x'^p_{m-1}, x'^p_m\}$$

3.3　DTW の適用

前節で選別した遺伝子ペア群 $GenePairs(X, X')$ の各遺伝子に対して，DTW を適用し，時間軸の正規化を行い，時系列発現量変化の類似度スコアを求める．ここでは，ある遺伝子ペア (x, x') を例に用いて説明する．

3.3.1　発現量の振幅の正規化

まず，遺伝子発現量が取りうる最大値が 1，最小値が -1 となるように発現量の振幅を $-1 \sim 1$ の間に正規化する．振幅を正規化するのは，発現の振幅ではなく，変化量が類似した遺伝子ペアを抽出するためであり，次に行う距離マトリクスの生成の際にもこの前処理が必要である．時系列遺伝子発現量を波形グラフとしてみなすと，図 2 に示すような，同じような波形を示すグラフに注目していることになる．

発現量の振幅の正規化は，観測された発現量から最大値 Max と最小値 Min を見つけ，それらの平均値 Ave が横軸に重なるように波形グラフを上下に平行移動する．そして，各時点の遺伝子発現量を $|Max - Ave|$ で割ることで，図 2 のように発現量が $-1 \sim 1$ に収まるように正規化を行う．正規化された (x, x') の発現量を $GeneExprN(x)$，$GeneExprN(x')$ とする．

3.3.2　距離マトリクスの生成

$GeneExprN(x)$，$GeneExprN(x')$ の発現量変化の類似度を算出するため，発現量の差分である

$$\Delta x_1 = x_{i+1} - x_i, \Delta x'_j = x'_{j+1} - x'_j$$

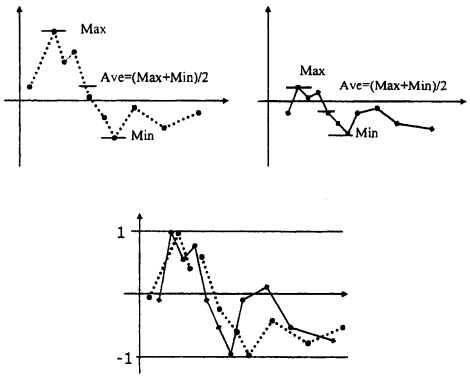

図 2　発現量の振幅の正規化 (横軸:時間, 縦軸:発現量)

を特徴ベクトル (q_i, c_j) として用い，距離マトリクスを生成する．実験 A のサンプル数は n 個，実験 B のサンプル数は m 個であるため，それぞれの特徴ベクトル $q_i = \Delta x_i, c_j = \Delta x'_j$ は，以下のように $n-1$ 個，$m-1$ 個作られる．

$$Q = q_1, \cdots, q_i, \cdots, q_{n-1}$$

$$C = c_1, \cdots, c_j, \cdots, c_{m-1}$$

この二つの特徴ベクトル間の距離を $d(q_i, c_j) = |q_i - c_j|$ として，図 3 のように距離マトリクスが生成される．

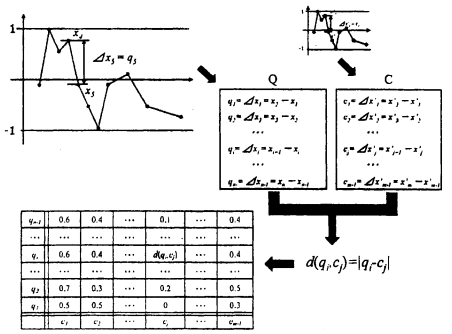

図 3　距離マトリクス

3.3.3　類似度スコアの算出

生成された距離マトリクスを基に，次式によって二つの距離スコア $g(i,j)$ が動的に求まる．初期条件を $g(1,1) = d(1,1)$ とし，終点における距離スコア $g(n-1, m-1)$ を，$GeneExprN(x)$ と $GeneExprN(x')$ の類似度スコアとする．

$$g(i,j) = d(q_i, c_i) + min \begin{cases} g(i-1, j-1) \\ g(i-1, j) \\ g(i, j-1) \end{cases}$$

3.3.4　時間軸の正規化

マトリクス上を通ったパスの情報を基に，各遺伝子ペアの遺伝子発現量のサンプル数が同じとなるように時間軸を正規化する．それぞれのサンプル数が n, m 個の $\{GeneExpr(x^p), GeneExpr(x'^p)\}$ は，次式から分かるように，ともに同じサンプル数 $dtw \leq n+m$ 個となる．

$$GeneExprDTW(x^p) = \{x^p_1, ..., x^p_n, ..., x^p_{dtw}\}$$

$$GeneExprDTW(x'^p) = \{x'^p_1, ..., x'^p_m, ..., x'^p_{dtw}\}$$

3.4　リファレンスの補正

前節までの結果，各遺伝子ペアの発現量のサンプル数が正規化されて比較可能となった．例えば，異なる実験 A と B から得られた s 個の遺伝子の発現量 $\{GeneExprDTW(x), GeneExprDTW(x')\}$ を，全て $x-y$ 平面上にプロットして相関係数を求めることで，二つの実験間の網羅的な発現プロファイルの比較ができる．

また，s 個の遺伝子ペア群 $DTWedGenePairs(X, X')$ は，時系列発現量プロファイルの類似度スコアで順位付けされている．s 個の遺伝子ペアから，上位 top 個を任意に選び，それらの遺伝子ペアの発現量 $\{GeneExprDTW(x), GeneExprDTW(x')\}$ を $x-y$ 座標にプロットし，発現プロファイルの傾向を見ることで，実験 A と B におけるリファレンスの違いを補正し，発現量を直接比較可能である．

一般に，ある遺伝子ペア (x, x') のサンプル時点 i におけるターゲットを T_i T'_i とすると，時間軸正規化後 (サンプル数が dtw 個に正規化されたとする) の時系列発現量 $x_1, x_2, ..., x_{dtw}$ と $x'_1, ..., x'_{dtw}$ は次の式で表せる．

$$x_1 = \log \frac{T_1}{R_{const}}, \cdots, x_{dtw} = \log \frac{T_n}{R_{const}}$$

$$x'_1 = \log \frac{T'_1}{R'_{const}}, \cdots, x'_{dtw} = \log \frac{T'_n}{R'_{const}}$$

この式から，リファレンス $R_{const} = R'_{const}$ であれば，発現量 x_i と x'_i を比べることは，ターゲットのシグナル T_i と T'_i を比べることであり，直接比較できることがわかる．仮に，$R_{const} \neq R'_{const}$ であれば，これらの発現量は，

$$x_i = \log \frac{T_i}{R_{const}} = \log T_i - \log R_{const}$$
$$x'_i = \log \frac{T'_i}{R'_{const}} = \log T'_i - \log R'_{const}$$

となり，x_i と T_i，x'_i と T'_i は対応しておらず，発現量の直接比較は行えない．

しかしながら，類似度スコアの高い遺伝子ペア群は，生命活動の根元をなす遺伝子であり，マイクロアレイ実験の条件にあまり依存せずに，それらの時系列遺伝子発現量の波形は非常に類似していると考えられる．類似した波形は $x-y$ 座標上にプロットすれば，ある近似直線上に集まる．二つのマイクロアレイ実験において，リファレンスが同じであれば，発現量 0 の位置は双方とも一致するはずであり，この近似直線は原点を通ることとなる．もし，近似直線が原点を通らなければ，これがすなわち二つの実験のリファレンスの差異となる．従って，リファレンスを補正するためには，近似直線を原点に通るように x 軸方向もしくは y 軸方向に平行移動させれば良い．この移動量がリファレンスの補正値に相当する．

4. 実　　験

本論文では，まず 8 種類の培養条件からなる枯草菌の時系列マイクロアレイデータを用いて実験を行い，次に同じ培地条件で増殖させた枯草菌と大腸菌の時系列マイクロアレイデータを用いて実験を行った．

4.1 実験データ

本論文では，2 で説明した二蛍光標識法を用いたマイクロアレイの実験結果を対象にしたものである．あるスポットに対して得られた二色の蛍光色素の強度の比を正規化したデータを用いる．正規化はある実験におけるコントロール細胞とターゲット細胞での発現比の平均値が各実験間で等しくなるような処置を施すことを目的としている．本論文で扱っている実験データの正規化には，著者らが開発した MicroArrayInfomatics システム [8] を用いた．

枯草菌の時系列マイクロアレイ実験は，ほぼ全遺伝子数に近い 4220 個の遺伝子を用いて行われた [1]．ターゲット (目的実験) は，8 種類の培養条件において経時的に合計で 81 個の測定点から抽出した mRNA の発現量である．また，全てのターゲットに対し，リファレンス (対照実験) は DSM 培地で $OD600_{nm}$ における濁度が 0.4 のときの mRNA の発現量 R_{const} である．

図 4 はこのときの増殖曲線で，横軸は時間，縦軸は菌体数，曲線の脇にある数字はサンプルポイントを表す．8 種類の培養条件では，それぞれのサンプル数，サンプル時刻は同じとは限らず，むしろほとんどが異なっている (表 2，図 4)．

以下，サンプル数が 8 個の LB 培地由来のデータを LB8，サンプル数が 5 個の CM 培地由来のデータを CM5 というように，培地名とサンプル数を用いて各培地由来のデータを DGG6，DSM19，GS5，CSM13，PS6，MGM8 と略記する．

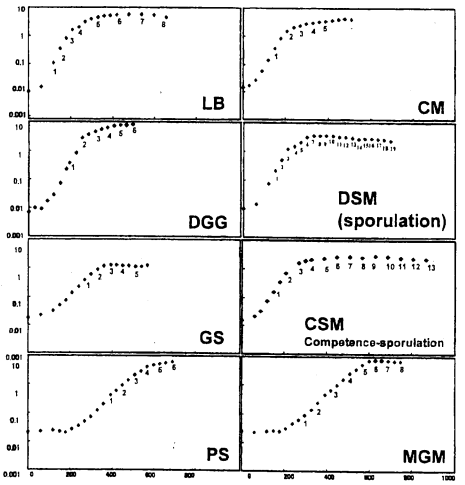

図 4　培養条件の異なる枯草菌の増殖曲線

表 2　8 種類の培地とサンプル数

培地名	サンプル数	培地の説明
LB	8	通常の生育に必要な成分を含んだ培地
CM	5	コンピーテンス培地
DGG	6	DSM にグルコースとグルタミンを添加 (胞子形成が起こらない)
DSM	19	胞子形成培地
GS	5	グルコース飢餓状態
CSM	13	コンピーテンス胞子形成培地
PS	6	リン酸飢餓状態
MGM	8	最小グルコース培地

枯草菌の時系列マイクロアレイデータは培養条件の違いにより 8 種類あるため，二つの異なるデータの選び方には ${}_8C_2 = 27$ 通りあり，その全組合わせに対して実験を行なった．また，前実験として LB8 と，LB8 から 6 番目 (図 4 参照) における全遺伝子の発現

量を欠落させたデータ (サンプル数 7) である LB7 を用いて実験を行った.

同じ生物種である枯草菌を用いているため, 単純に同じ遺伝子 (x, x') を遺伝子ペア群 (X, X') として選択した (すなわち, $x = x'$ である). この際 2 節で説明したように遺伝子発現をしていない (スポットのシグナルが黒色である) ものや実験上データが得られなかった遺伝子は全て取除いた.

次に, 選択された各遺伝子ペアに対して DTW を適用し, 時間軸の正規化を行った. 例えば, サンプル数がもっとも少ない CM5(または GS5) とサンプル数がもっとも多い DSM19 を用いた場合, サンプル数 n は $19 \leq n \leq 24 (= 5 + 19)$ 個に正規化される.

DTW により時間軸正規化された遺伝子発現量を用いて, 散布図を描いて相関係数と近似直線を求めた. 枯草菌におけるマイクロアレイデータは全て同じリファレンス R_{const} を用いているので, 補正の必要がないことを確認する.

4.2 実験結果と考察

4.2.1 LB 培地由来のデータ (LB8-LB7)

遺伝子ペアの選別により得られた遺伝子数は, 枯草菌が持つ遺伝子の約半分である 2606 個であった. これらに DTW を適用して, 時間軸正規化を行った各遺伝子ペアの発現量を $x - y$ 平面にプロットした (図 5 参照). x 軸は LB8, y 軸は LB7 の遺伝子発現量を表している.

LB8 と LB7 において, 欠落させたデータ以外は全て同じであるため, 発現量のプロットはほぼ直線 $y = x$ 上に乗っている. 直線 $y = x$ からのばらつきは DTW で時間軸正規化されたポイントである. 相関係数が 1 に近く, かなりの強い相関があることがわかる.

この結果から, サンプル数の異なる時系列マイクロアレイデータに対して, 提案手法の DTW による時間軸の正規化が有効であることが確認できる.

4.2.2 培地の全組合せによる結果

全ての培養条件の組合せによる 27 通りのデータを用いて, 選択された遺伝子数と, それらの発現量に DTW を適用したものを $x - y$ 平面にプロットし, 得られた相関係数, 近似直線の式を表 3 に示す. 図 6 は, 相関係数が特に高かった PS6-MGM8, GS5-MGM8, GS5-PS6 と最も低かった LB8-PS6 の散布図である. また, PS6-MGM8 と LB8-PS6 において, それぞれ

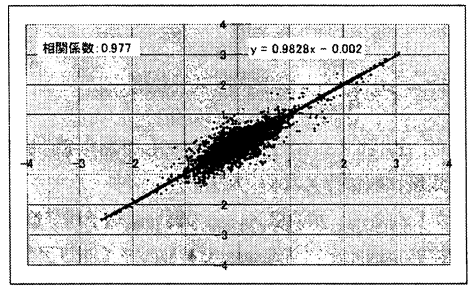

図 5 DTW 適用後の LB8 と LB7 の遺伝子発現量の散布図

発現が観測された全ての遺伝子と, その中から DTW で類似度が上位 100 番目までの遺伝子の発現量で散布図を生成した (図 7 参照).

特に高い相関係数を示した培地の組合せは PS6-MGM8 と GS5-MGM8 で, それぞれ 0.798 と 0.786 であった. GS5-PS6 の相関係数も 0.674 と高い. PS6, MGM8, GS5 の三つの培地はそれぞれリン酸飢餓培地, MGM8 は最小グルコース培地, GS5 はグルコース飢餓培地と栄養条件が悪いことが共通している. また, もっとも低い相関係数を示しているのは LB8-PS6 で, 0.279 であった. LB8 は生育に必要な栄養が豊富な培地であり, 一方, PS6 は栄養飢餓培地である.

類似した培養条件由来のデータの相関は高く, その逆の場合は相関が低くなっていることから, 提案手法による, 異なるサンプル数, サンプル時刻の正規化が有効であることが確認された. また, 相関係数を見ることで培地条件の違いによる, 遺伝子発現の網羅的な変化を把握できることが確認できた.

また, PS6-MGM8 と LB8-PS6 において, 発現が観測された全ての遺伝子の発現量では相関が低い場合でも, DTW で類似度が上位の遺伝子を選択し発現量を見ると相関があがっている (図 7). この結果から, 培地の違いに依存せずに発現量が変化する遺伝子が, 類似度の上位に抽出できていることが考えられる.

本実験において, 全てのデータにおいてリファレンス R_{const} は一定である. 表 3 を見ると, 全 27 通りの組合せにおいて, 近似直線の切片がほぼ 0 となり原点に近接する. このことから, リファレンスが同じデータに対して本提案手法を用いると, 散布図から得られる近似直線の切片が 0 となり, 3.4 で述べたようにリファレンスの補正の必要がないことが確認できる.

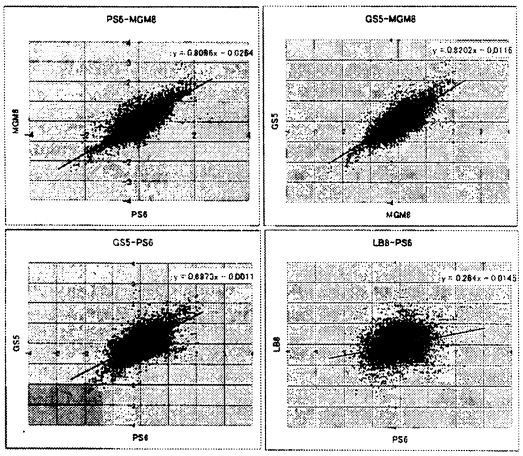

図 6　PS6-MGM8,GS5-MGM8,GS5-PS6,LB8-PS6 の散布図

表 3　27 通りの培地の組合せの結果

培地 (x 軸)	培地 (y 軸)	遺伝子ペア数	相関係数	近似直線
LB8	CM5	2352	0.361	$y = 0.3753x - 0.0044$
LB8	DGG6	1868	0.390	$y = 0.3678x - 0.0145$
LB8	DSM19	1865	0.502	$y = 0.5469x - 0.025$
LB8	GS5	2483	0.379	$y = 0.3447x - 0.0147$
LB8	CSM13	1875	0.402	$y = 0.4643x - 0.0572$
LB8	PS6	2096	0.279	$y = 0.264x - 0.0145$
LB8	MGM8	2167	0.348	$y = 0.3397x - 0.0225$
CM5	DGG6	1970	0.654	$y = 0.6126x - 0.0029$
CM5	DSM19	1860	0.466	$y = 0.4678x - 0.0342$
CM5	GS5	2588	0.587	$y = 0.5155x - 0.0158$
CM5	CSM13	1898	0.419	$y = 0.4651x - 0.0692$
CM5	PS6	2222	0.691	$y = 0.6343x + 0.0011$
CM5	MGM8	2280	0.663	$y = 0.6132x - 0.0183$
DGG	DSM19	1569	0.420	$y = 0.446x - 0.0752$
DGG6	GS5	2028	0.595	$y = 0.5616x - 0.0375$
DGG6	CSM13	1604	0.355	$y = 0.4281x - 0.12$
DGG6	PS6	1880	0.645	$y = 0.6235x - 0.0154$
DGG6	MGM8	1888	0.603	$y = 0.5997x - 0.0435$
DSM19	GS5	2024	0.485	$y = 0.4303x + 0.008$
DSM19	CSM13	1615	0.592	$y = 0.6544x - 0.0223$
DSM19	PS6	1802	0.389	$y = 0.3566x - 0.0059$
DSM19	MGM8	1764	0.479	$y = 0.4436x - 0.016$
GS5	CSM13	1974	0.529	$y = 0.6779x - 0.0627$
GS5	PS6	2329	0.674	$y = 0.6973x - 0.0011$
GS5	MGM8	2378	0.786	$y = 0.8202x - 0.0116$
CSM13	PS6	1785	0.400	$y = 0.3267x + 0.0166$
CSM13	MGM8	1806	0.506	$y = 0.4273x + 0.0158$
PS6	MGM8	2095	0.798	$y = 0.8096x - 0.0264$

5. 関連研究

時系列マイクロアレイデータの解析を行う研究として，本研究のように異なる時系列マイクロアレイデータを直接比較するためのデータ補正を行う研究は筆者らの知る限り存在しない．しかしながら，時系列マイクロアレイデータから細胞周期を解析する研究は幾つか存在する．

一般的な問題として，マイクロアレイデータ解析に

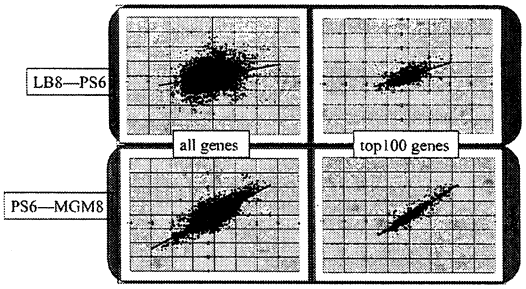

図 7　PS6-MGM8 と LB8-PS6 の散布図

は高レベルでノイズが発生することと，各遺伝子あたりの実験数 (N) が少ないことがあげられる．また，細胞周期に関連する遺伝子は，全遺伝子の中でごく一部であり，周期的に発現しているシグナルを見つけようとする際，非周期的な (細胞周期に関連していない) 遺伝子のシグナルに左右されてしまうという問題がある．

Wichert らは，average periodogram と呼ばれる周期的なデータを検出する手法を適用した後，得られたデータが周期性を持つのか，あるいは単なる誤差なのかを区別するために Fisher's test に基づく検定を行う方法を提案している [11]．シミュレーションデータと実際に公開されているデータに対してこの手法を用いたところ，時系列を扱う古典的な手法と異なり，各遺伝子あたりの実験数 (N) が少なくても有効であることを示した．

一方，Ota らは真核生物である *S.cerevisiae, H.sapiens, A.thaliana* の時系列遺伝子発現プロファイルに対して，フーリエ解析を行い周期的に発現している細胞周期関連遺伝子を抽出し，抽出された遺伝子がオーソログであるかを調べている [7]．その結果，三種間で共通な遺伝子を見ると，細胞周期の基本に関わるチェックポイントタンパク質や DNA 調節タンパク関連の遺伝子が見つかり，真核生物全般にわたって細胞周期関連遺伝子が進化的に強く保存されていることを確認している．

上述した研究は，複数のマイクロアレイデータにおいて，いずれもある機能を持つ遺伝子セットを対象にした研究である．本研究では，時系列発現プロファイルが類似したペアを抽出しているが，最終的な目標は全遺伝子の網羅的な比較であることが，これらの関連研究とは異なる．

6. おわりに

本稿では時系列マイクロアレイデータに注目し，複数の時系列発現プロファイルのサンプル数，サンプル時点が異なっていても，時間軸正規化を行い比較可能にする手法を提案した．実際に，枯草菌において 8 種類の異なる培養条件で得られた時系列発現プロファイルに提案手法を適用することで，提案手法の正しさを示しただけでなく，二つの培養条件がもたらす網羅的な遺伝子発現の比較を行うことができた．

本稿では，枯草菌において 28 通りの実験しか行っていないため，今後はより多くのデータを用いて提案手法の検証を行い信頼性を高める必要がある．特に，現在のマイクロアレイ実験には多くの誤差やノイズが含まれているため，これらの誤差やノイズの除去や，結果の検定が今後の課題である．

参考文献

1) Subtilist. http://genolist.pasteur.fr/SubtiList/. (2005 年 2 月現在).
2) Arrayexpress at the ebi. http://www.ebi.ac.uk/arrayexpress/. (2005 年 2 月現在).
3) J.D. and F.H.C. Watson Crick. A Structure for Deoxyribose Nucleic Acid. *Nature*, Vol. 171, pp. 737–738, 1953.
4) C.J. Stoeckert Jr., H.C. Causton, and C.A. Ball. Microarray databases: standards and ontologies. *Nature Genetics*, Vol. 32, pp. 469 – 473, 2002.
5) Sang-Wook Kim, Sanghyun Park, Wesley W., and E. Chu. An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases. In *Proc. of the 17th International Conference on Data Engineering*, pp. 607 – 614, 2001.
6) See-Kiong Ng, Soon-Heng Tan, and V.S. Sundararajan. On Combining Multiple Microarray Studies for Improved Functional Classification by Whole-Dataset Feature Selection. *Genome Informatics*, Vol. 14, pp. 44–53, 2003.
7) Koji Ota, Susumu Goto, and Minoru Kanehisa. Comparative analysis of transcriptional regulation in eukaryotic cell cycles. In *Proc. of IBSB 2004*, pp. 26 – 27, 2004. (Poster).
8) S. Kanaya S, J. Ohtani, Y. Wada, K. Wada, M.A. Amin, and Nakamura Y. Integrated analytical tool for genome and transcriptome informatics (MicroArrayInformatics). In *Escherichia coli Conference Towards New Biology in the 21th Century*, 2003.
9) H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust., Speech, Signal Process*, Vol. 26, pp. 43–49, 1978.
10) PT Spellman and G. Sherlock et al. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, Vol. 9, pp. 3273 – 3297, 1998.
11) Sofia Wichert and KonstantinosFokianos and-Korbinian Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, Vol. 20, No. 1, pp. 5–20, 2004.
12) Lap Kun Yeung, Hong Yan, Alan Wee-Chung Liew, Lap Keung Szeto, Michael Yang, and Richard Kong. Measuring correlation between microarray time-series data using dominant spectral component. In *Proc. of the second conference on Asia-Pacific bioinformatics*, Vol. 29, pp. 309–314, 2004.
13) L Zhang, W Zhou, VE Velculescu, SE Kern, RH Hruban, SR Hamilton, B Vogelstein, and KW Kenzler. Gene expression in normal and cancer cells. *Science*, Vol. 276, pp. 1268–1272, 1997.
14) 松村正明, 那波宏之. DNA マイクロアレイと最新 PCR 法. 秀潤社, 2000.