

誤差を考慮した時系列マイクロアレイデータの時間軸補正方法

采 女 寛 行[†] 宮 崎 純[†] 金 谷 重 彦[†] 植 村 俊 亮[†]

本論文では、DTWを用いた誤差を考慮に入れる時間軸の補正手法を提案する。DNA マイクロアレイは遺伝子発現を網羅的に解析するための技術であるが、データの誤差が大きいため複数の時系列データの比較手法は十分に議論されていない。その中で二種類の時系列データを正しく比較できるように時間軸を補正する手法として DTW (Dynamic Time Warping) が提案されたが、非常に誤差に弱いという問題を持っている。

Error Adapting Time-Series Microarray Time Axis Correction Method

HIROYUKI UNEME, [†] JUN MIYAZAKI, [†] SHIGEHICO KANAYA [†]
and SHUNSUKE UEMURA [†]

In this paper, we propose a new error adapting comparison technique which uses DTW (Dynamic Time Warping algorithm) for the correction of the time axis in time-series microarray data. In recent years, DTW was developed and used in speech recognition research as well as in gene expression pair matching. However, this algorithm is susceptible to data errors, and can not compare entirely different time-series microarrays. Thus, we propose a method which considers such errors in the comparison of different time-series microarrays.

1. はじめに

1997 年に Fred Blattner らが大腸菌の全ゲノム配列の決定^①を行って以来、生物の全ゲノム配列はモデル生物を中心として急速に解読されている。2000 年までに約 50 種^②の生物のゲノム配列が決定され、2005 年にはそれが約 300 種^③にまで及んでいる。さまざまな生物種の全ゲノム配列が決まったことにより、今までのように遺伝子を個々として捉えるだけでなく遺伝子の機能やネットワークを網羅的に解析する方法が実現化し始めてきた。このような背景のもとに遺伝子発現を網羅的に解析する技術として

DNA マイクロアレイが誕生した。

この技術を使って得られたデータは共同研究者やオンラインデータベース^{④⑤}などを通じて広く利用可能になってきており^⑥、これからはこれらのデータが今まで以上に幅広く利用されるようになると考えられる。しかし、これらのデータには (1) DNA マイクロアレイのプラットフォームが数多く存在し、プラットフォーム間で得られたデータを直接比較することができず、(2) 得られたデータの誤差が大きいため比較すること自体が難しいという欠点がある。また、細菌の時系列データを比較する際には (3) 増殖曲線の形と観測点を合わせなければ時系列データを比較できないという問題も生じてしまう。上述 (1) に関しては使用するプラットフォームと手順が共通化されれば複数の研究室間でのマイクロアレイの結果が比較可

[†] 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

能になることが示唆されている⁷⁾が、(2) (3) に関してはこれらの問題を解決する画期的な手法はまだ実現されていない。

このようななかで上述 (3) に対応するために DTW (Dynamic Time Warping algorithm) と呼ばれる手法が注目された⁸⁾。この手法は配列アライメントに利用される動的計画法⁹⁾に似ており、動的計画法が記号の対を比較し二つの配列の最適な対応付けを得るのに対し、DTW は数値の対を比較することで二つの時系列データの最適な対応付けを得る方法である。

この DTW の手法がマイクロアレイデータの解析に導入されたのは著者らが知る限りでは John Aach らの論文⁹⁾が最初である。この論文ではクラスタリング手法を利用しないデータマイニングの方法として紹介されている。その後、萩本ら¹⁴⁾も DTW を用いて枯草菌のデータを解析しているが、どちらの論文も時間軸を正規化した観測点の対応付けが遺伝子ごとに異なっているため二つの時系列データの誤差を考慮して比較できているとは言えない。

そこで、本論文では誤差に大きく影響される従来の DTW の解析手法を改良することで、誤差を考慮し二つの時系列データにただ一つの対応付けを与えることができる新たな解析手法を提案する。また、この手法は遺伝子発現量を基準に取るため、増殖曲線と観測点の異なった測定結果どうしを比較することも可能となる。

2. マイクロアレイ

生体における遺伝子の発現は、時間的かつ空間的に調節されており、この調節機構を網羅的に解明することが今後の課題となっている¹⁵⁾。この遺伝子の発現はタンパク質の合成の有無や合成されたタンパク質の量で判断できるがタンパク質はその化学的性質が一樣でないため網羅的な観測が難しい。そのため、化学的性質が一樣で検出が容易なタンパク質合成のテンプレートとなる mRNA

の発現状態を網羅的に解析する手法としてマイクロアレイが開発された。マイクロアレイの原理は基板上に多数の DNA プローブを高密度に固定し、そこに標識した核酸 (ターゲット) をハイブリダイズさせることで結合した核酸の量を調べるというものである。この手法が提案された当初は 48 個の遺伝子¹⁰⁾しか解析できなかったが、現在では数万から数十万遺伝子を網羅的に解析することが可能となった。

マイクロアレイにより遺伝子の発現量を測る際には、目的実験のターゲットと対照実験のリファレンス由来の 2 種類のサンプルを用意する必要がある。それらの遺伝子を異なる蛍光色素 (通常、Cy3: 緑 "ターゲット"、および Cy5: 赤 "リファレンス") 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 の蛍光強度を測定することで目的の遺伝子の発現量を知ることができる。例えば、ある遺伝子がリファレンスではほとんど発現せず、ターゲットでは過剰に発現している場合、その遺伝子に対応するスポットは赤色となる。緑色のスポットはリファレンスのみで発現している遺伝子であり、黄色のスポットは二つで同程度、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はスポットが黒色となる¹⁵⁾。

ある遺伝子 GeneX の発現量 $GeneExprX$ を表すには、ターゲットの蛍光強度 T とリファレンスの蛍光強度 R を用いた以下の式が一般的に用いられている¹⁴⁾。

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

2.1 時系列マイクロアレイ

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¹⁵⁾。時間 t_n における遺伝子 GeneX の発現量 $GeneExprX(t_n)$ 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取ったものとする¹⁴⁾。

$$\begin{aligned} GeneExprX(t_1) &= \log \frac{T(t_1)}{R(t_r)} \\ GeneExprX(t_2) &= \log \frac{T(t_2)}{R(t_r)} \\ &\dots \dots \dots \\ GeneExprX(t_n) &= \log \frac{T(t_n)}{R(t_r)} \end{aligned}$$

3. 時系列マイクロアレイデータの補正

本節では、二つの時系列マイクロアレイ実験 A, B (以下、時系列 A, B と呼ぶ) から得られた二つの遺伝子発現プロファイルを時間軸の正規化をして比較可能にする手法について述べる。

3.1 データ補正方式の概略

二つの時系列 A と B から同じ遺伝子ペア群 $GenePairs(X_A, X_B)$ を抽出する。抽出した遺伝子ペア間の全ての観測点の差から類似度を求める。

次に、得られた全ての類似度に DTW アルゴリズムを適用し、時系列 A と B の各観測点の最適な対応付けの候補を得る。

最終的に得られた候補について相関係数を計算する。計算された相関係数の値が最も高いものを最適な対応付けと判断する。

3.2 類似度の算出

はじめに時系列 A もしくは B のある観測点の遺伝子の発現量 $GeneExprA, B(t_n)$ を遺伝子の総数 n と誤差により測定できないスポットの数 (Incapable Measurement), IM を用いて遺伝子 X のある時刻 (t_n) での発現量の合計に対する発現比率 (Proportion), P に変換する。なお、時系列 A の観測点の数を a 個、時系列 B の観測点の数を b 個とする。このとき、発現量から発現比率への変換式は次のように表すこと

ができる。

$$\begin{aligned} TotalX(t_i) &= \sum_{i=1}^n GeneExprX_i(t_i) \\ Mean &= \frac{\left(\sum_{i=1}^a TotalA(t_i) + \sum_{i=1}^b TotalB(t_i) \right)}{(a+b)} \\ P(t_i) &= GeneExprX(t_i) + \frac{(Mean - TotalX(t_i))}{(n - IM)} \end{aligned}$$

この結果をもとにして時系列 A のある点と時系列 B のある点の類似度 (Similarity), $Sim(t_i)$ を時系列 A のある遺伝子 i の発現比率 P_{Ai} と時系列 B のある遺伝子 i の発現比率 P_{Bi} を用いて以下の式で算出する。

$$Sim(t_i) = \sum_{i=1}^n \sqrt{(P_{Ai}(t_i) - P_{Bi}(t_i))^2}$$

つまり、この式から算出される類似度の値が低いほうが類似しているということになる。そして、この類似度を時系列 A と B の間の全てのマイクロアレイの組み合わせで算出する。

この方法を利用すると類似度がもつ誤差は正規分布の形で与えられる誤差の平均値を取るようになるため、誤差を相殺すると考えられる。

3.3 DTW の適用

前節で算出した類似度 $Sim(t_i)$ を観測点ごとに並べ類似度行列 $Sim(x, y)$ を作成する (図 1 参照)。類似度行列内の各値は見出しとなっている整数が観測点を示し、0 以外の整数が観測点間の類似度を示している。例えば、時系列 A の観測点 1 と時系列 B の観測点 1 の類似度は図 1 では 410 となる。

生成された類似度行列 $Sim(x, y)$ をもとに、次式によって類似度スコア行列 (Score matrix), $Sco(x, y)$ に変換する。

$$Sco(x, y) = Sim(x, y) + \min \begin{cases} Sco(x-1, y-1) \\ Sco(x-1, y) \\ Sco(x, y-1) \end{cases}$$

この類似度スコア行列 $Sco(x, y)$ から始点を

類似度行列 $Sim(x, y)$ の局所値と同じ座標の類似度スコア行列 $Sco(a, y)$ と $Sco(x, b)$ (つまり、図 1 の場合、始点は 6 列目の局所値を取る点 617 と 881、9 行目の局所値を取る点 1354 と 1099 の 4 点に対応する場所にある類似度スコア行列の各点となる) で取り、終点を $Sim(1, y)$ と $Sim(x, 1)$ の中の最小値で取る点となる $Sim(x, y)$ の最小値を通る経路を決定する (図 2 参照)。

		時系列 A の観測点					
		1	2	3	4	5	6
時系列 B の観測点	1	410	523	879	660	800	847
	2	506	516	788	539	643	747
	3	646	586	641	429	557	687
	4	741	675	770	393	545	681
	5	719	666	669	414	397	617
	6	1050	980	936	806	761	704
	7	1285	1282	1201	993	842	973
	8	1142	1143	1092	855	743	881
	9	1354	1418	1434	1188	1099	1200

図 1 類似度行列 (6×9 行列)

	1	2	3	4	5	6
1	410	933	1813	2473	3274	4121
2	916	927	1715	2254	2898	3645
3	1563	1503	1568	1997	2555	3242
4	2304	2179	2273	1961	2507	3189
5	3024	2845	2848	2376	2359	2976
6	4074	3825	3782	3182	3120	3064
7	5359	5108	4983	4176	3963	4037
8	6502	6251	6076	5032	4706	4845
9	7856	7670	7511	6220	5805	5906

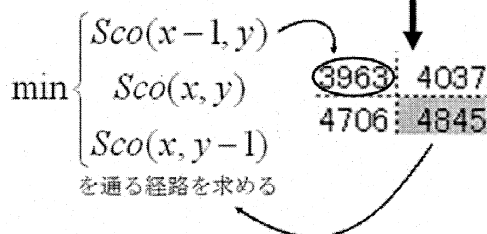


図 2 類似度スコア行列と最適経路の決定

最後に、得られた経路に対応する時系列 A と B の発現比率のペア $GenePairs(P_A(t))$,

$P_B(t_i)$ の相関係数を計算し、その値が最も高いものを最適な経路とする。

4. 実験

本論文では、まず、筆者らが採取した同じ実験条件で増殖曲線の形と観測点の観測時間と観測数を合わせた *E.coli* W3110 の 2 つ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 *E.coli* MG1655 の 2 段階増殖の時系列マイクロアレイデータ^{2) 11)} をオクラホマ大学のデータベース²⁾ より取得し、それを用いて実験した。

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理を施すことを目的としている。

本論文で扱う *E.coli* W3110 のデータの正規化には著者らが開発した Micro-Array Infomatics システム¹²⁾ を利用し、データベースから得られた *E.coli* MG1655 のデータの正規化には lowess 法¹³⁾ が用いられている。

本論文では 10 を底とする対数で発現量を計算しており、値が -100 から 100 の間に収まらないものは全て測定不能 *IM* として計算に関与しないよう値を 0 にした。

4.2 実験手順

まずはじめに、実験条件が同じで観測点と増殖曲線を合わせた *E.coli* W3110 の二つの時系列 A, B を比較し、提案手法を検証した。

次に、時系列 A と任意の観測点を欠落させた時系列 B を比較した。

ここで算出した最適経路を萩本ら¹⁴⁾ の手法と比較し、本研究の有意性を証明した。

最後に、実験手法を応用して得られた知見

をオクラホマ大学のデータベースから取得した *E.coli* MG1655 を用いて試す。

4.3 実験結果と考察

4.3.1 *E.coli* W3110

本論文ではアレイから得られた全遺伝子である 4309 セットの遺伝子ペアを利用した。データの増殖曲線と観測点を図 3 に示し、時系列 A, B から得られた最適経路と相関係数の計算結果を図 4 に示す。

この結果から、図 4 の 0 以外の整数で示された最適経路が増殖曲線から得られる観測点の対応と一致しており、その相関係数の値からかなりの相関があることが分かる。

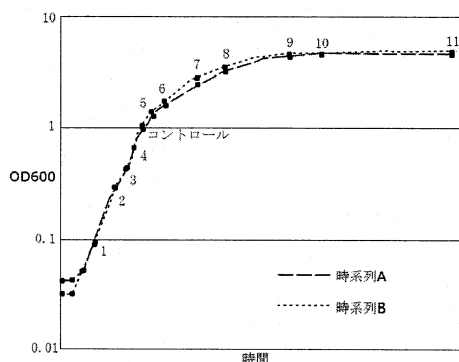


図 3 *E.coli* W3110 の増殖曲線と観測点

		時系列Aの観測点										
時 系 列 B の 観 測 点		1	2	3	4	5	6	7	8	9	10	11
	1	410	0	0	0	0	0	0	0	0	0	0
	2	0	927	0	0	0	0	0	0	0	0	0
	3	0	0	1568	0	0	0	0	0	0	0	0
	4	0	0	0	1961	0	0	0	0	0	0	0
	5	0	0	0	0	2359	0	0	0	0	0	0
	6	0	0	0	0	0	3064	0	0	0	0	0
	7	0	0	0	0	0	0	3801	0	0	0	0
	8	0	0	0	0	0	0	0	4565	0	0	0
	9	0	0	0	0	0	0	0	0	5383	0	0
	10	0	0	0	0	0	0	0	0	0	6325	0
	11	0	0	0	0	0	0	0	0	0	0	7353

図 4 最適経路 (相関係数 : 0.62)

次に、時系列 A と任意の観測点を欠如させた時系列 B の最適経路を示す (図 5a, 5b 参照)。

図 5a, b ではそれぞれ観測点 6 と観測点 2, 5, 8 を削除した。その結果、どちらの場合も図中の灰色の部分で 1 つの観測点に対し複数の観測点に対応付けられ、最適経路が正し

く補間されていることが判明した。この実験結果から、提案手法を用いることで高い精度でデータの欠落した点を補間することができることが分かる。また、そのときの相関係数の結果もかなりの相関を示していることが分かる。紙面の都合上、本論文には掲載していないが他の観測点を除いた場合でも高い精度でデータの欠落した観測点の補間ができた。

5a		時系列Aの観測点										
		1	2	3	4	5	7	8	9	10	11	
時 系 列 B の 観 測 点	1	408	0	0	0	0	0	0	0	0	0	
	2	0	922	0	0	0	0	0	0	0	0	
	3	0	0	1563	0	0	0	0	0	0	0	
	4	0	0	0	1955	0	0	0	0	0	0	
	5	0	0	0	0	2350	0	0	0	0	0	
	6	0	0	0	0	3110	0	0	0	0	0	
	7	0	0	0	0	0	3845	0	0	0	0	
	8	0	0	0	0	0	0	4604	0	0	0	
	9	0	0	0	0	0	0	0	5417	0	0	
	10	0	0	0	0	0	0	0	0	6357	0	
	11	0	0	0	0	0	0	0	0	0	7382	

5a. 観測点 6 の欠落 (相関係数 : 0.62)

5b

		時系列Aの観測点								
		1	3	4	6	7	9	10	11	
時系列Bの観測点	1	408	0	0	0	0	0	0	0	
	2	912	0	0	0	0	0	0	0	
	3	0	1552	0	0	0	0	0	0	
	4	0	0	1844	0	0	0	0	0	
	5	0	0	2357	0	0	0	0	0	
	6	0	0	0	3059	0	0	0	0	
	7	0	0	0	0	3793	0	0	0	
	8	0	0	0	0	4569	0	0	0	
	9	0	0	0	0	0	5401	0	0	
	10	0	0	0	0	0	0	6341	0	
	11	0	0	0	0	0	0	0	7365	

5b. 観測点 2, 5, 8 の欠落 (相関係数 : 0.62)

図 5 観測点を欠落させた最適経路

最後に、萩本らの手法を用いた結果の分析 (図 6a, b を参照) と本研究の図 4 の最適経路を比較した。

図 6a, b は横軸が同じ経路を通る遺伝子の数、縦軸が全体の数に占める遺伝子の数の割合としている。例えば、同じ経路を通る遺伝子の総数が一つのデータは全体の 40%以上に及ぶが、これは経路が一致する遺伝子が自分以外で一つもないものが全体の 40%の遺伝子数を占めるということである。図 6a では 768 遺伝子を通る経路と提案手法の示した最適経路が一致した。

しかし、図 6b で示す観測点 6 を欠損させ

たデータを用いた場合は本研究の提案手法と一致する経路を通る遺伝子は萩本らの結果には存在しなかった。図 6b の萩本らの結果で 633 遺伝子を通る経路は図 7 の灰色部分 (0 以外の値の部分) は提案手法の示す最適経路) のようになり、106 遺伝子と 176 遺伝子を通る経路はどちらも測定不能の値が大部分を占める経路であることが判明した。すなわち、本手法は誤差に強いことが示された。

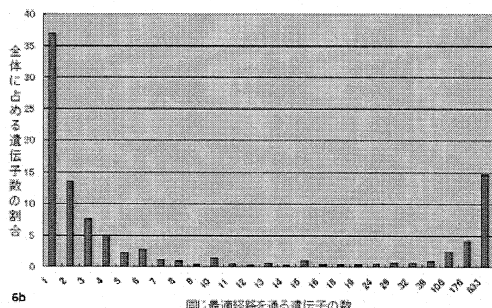
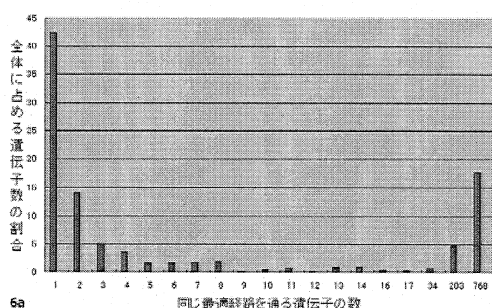


図 6 萩本らの手法を用いた経路の分岐

		時系列Aの観測点										
		1	2	3	4	5	7	8	9	10	11	
時系列Bの観測点	1	408	0	0	0	0	0	0	0	0	0	0
	2	0	922	0	0	0	0	0	0	0	0	0
	3	0	0	1563	0	0	0	0	0	0	0	0
	4	0	0	0	1955	0	0	0	0	0	0	0
	5	0	0	0	0	2350	0	0	0	0	0	0
	6	0	0	0	0	0	3110	0	0	0	0	0
	7	0	0	0	0	0	0	3845	0	0	0	0
	8	0	0	0	0	0	0	0	4604	0	0	0
	9	0	0	0	0	0	0	0	0	5417	0	0
	10	0	0	0	0	0	0	0	0	0	6357	0
	11	0	0	0	0	0	0	0	0	0	0	7362

図 7 633 個の遺伝子に共通した経路

4.3.2 *E.coli* MG1655

本データもアレイから得られた全遺伝子である 4289 セットの遺伝子ペアを利用した。データの増殖曲線と観測点を図 8a, b に示す。

これら二つのデータから提案手法を用いて最適経路を求めた結果を図 9 に示す。

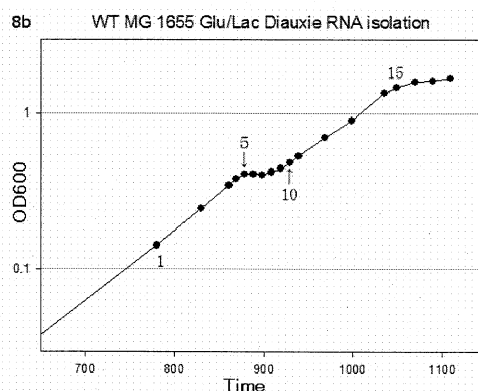
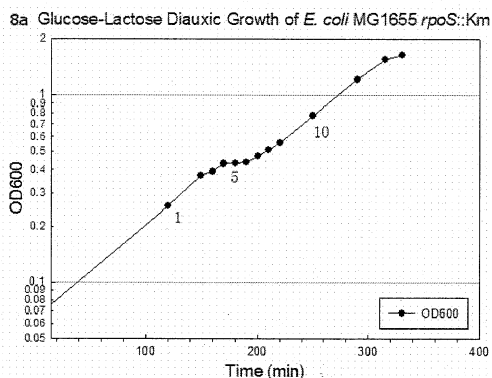


図 8 *E.coli* MG1655 の増殖曲線と観測点

**図 7a, b はオクラホマ大学のデータベース 2) より引用

図 9 の結果を観察すると三カ所に対応付けの重複した部分 (灰色) があることが分かる。そのなかの矢印の部分を図 8a, b と対応させると二段階増殖期の二段階目の対数増殖期の開始点が合うようにデータが補間されていることが分かる。

これらの結果は二段階目の対数増殖期の起点をうまく対応付けることができたことを示している。

		rpoS 欠損株の観測点												
		1	2	3	4	5	6	7	8	9	10	11	12	13
野生株の観測点	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	924	2144	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	3987	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	5196	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	6427	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	7790	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	9034	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	10147	0	0	0	0	0
	9	0	0	0	0	0	0	0	0	11602	0	0	0	0
	10	0	0	0	0	0	0	0	0	0	13265	0	0	0
	11	0	0	0	0	0	0	0	0	0	0	14822	0	0
	12	0	0	0	0	0	0	0	0	0	0	0	16136	0
	13	0	0	0	0	0	0	0	0	0	0	0	0	17894
	14	0	0	0	0	0	0	0	0	0	0	0	0	19627
	15	0	0	0	0	0	0	0	0	0	0	0	0	22217
	16	0	0	0	0	0	0	0	0	0	0	0	0	0
	17	0	0	0	0	0	0	0	0	0	0	0	0	0

図 9 提案手法から得られた最適経路
(相関係数: 0.52)

次に、図 10a, b にそれぞれ時系列 B の観測点 1 から 8 と時系列 A を対応付けた結果と時系列 B の観測点 9 から 17 と時系列 A を対応付けた結果を示す。

		rpoS 欠損株の観測点												
10a		1	2	3	4	5	6	7	8	9	10	11	12	13
野生株の観測点	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	924	2144	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	3987	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	5196	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	6427	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	7790	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	9034	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	10147	0	0	0	0	0

10a. 野生株の前半の観測点と *rpoS* 欠損株の比較 (相関係数: 0.63)

		rpoS 欠損株の観測点												
10b		1	2	3	4	5	6	7	8	9	10	11	12	13
野生株の観測点	9	0	0	0	0	0	0	0	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0
	11	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	917	0	0	0	0	0	0	0	0	0	0	0	0
	13	0	5012	0	0	0	0	0	0	0	0	0	0	0
	14	0	0	6897	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	8452	0	0	0	0	0	0	0	0	0
	16	0	0	0	0	10134	0	0	0	0	0	0	0	0
	17	0	0	0	0	11541	0	0	0	0	0	0	0	0

10b. 野生株の後半の観測点 *rpoS* 欠損株の比較 (相関係数: 0.58)

図 10 提案手法の応用

図 10a では図 9 と同じ経路となり、観測点の数に大きく差がある場合でも観測点の時間間隔が適切であれば誤差を考慮して対応付けを行えることを示している。

図 10b の結果を図 8a, b と照らし合わせると *rpoS* 遺伝子欠損株の 1 番ははじめの観測点と野生株の 9 から 11 番目の遺伝子が対応付けられていないが、これは *rpoS* 遺伝子欠損

株の観測点 1 と野生株の観測点 12 が最もよく類似しているためこの部分までで対応付けが終わったことを意味している。この結果を増殖曲線と比較すると *rpoS* 欠損株と野生株のどちらの点も対数増殖期の途中にあり、他の観測点から孤立していることが分かる。また、*rpoS* 遺伝子欠損株の 5 番目の観測点と野生株の 16, 17 番目の観測点に対応付けられているが、この結果は *rpoS* 遺伝子欠損株の二段階目の定常期と野生株のグルコース飢餓状態でかつラクトースが存在しているときの定常期とが類似した遺伝子発現パターンを持つことを示している。

これらの結果から、提案手法を利用すれば異なった観測点と異なった増殖曲線を持つデータの誤差を考慮に入れて精度よく対応付けることが可能になることが分かる。

5. 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、増殖曲線のデータを用いることなく遺伝子発現量のみでかつ、実験誤差を考慮して観測点の対応付けを行う手法を提案した。

現在、微生物の増殖の比較や生物の発生の研究では増殖曲線や AEL (After Egg Length) などのように時間をスケールとして比較実験が行われている。しかし、微生物の増殖や生物の発生の温度や栄養状態、pH などの条件が変わると同じ時間で観察したデータでも比較することが難しくなる。しかし、提案手法はマイクロアレイから得られた遺伝子の発現量と観測点の前後の情報のみからデータを比較することを目的としている。

提案手法では増殖曲線の情報を使わず遺伝子発現比率からデータの対応付けを行い、その結果を増殖曲線と照らし合わせて考察した。その結果、増殖曲線と提案手法での対応付けの結果が一致した。これらの結果から、提案手法を用いることで増殖曲線の形が異なることや観測点の場所や密度が異なることで今まで直接比較できなかったデータを比較す

ることが可能であると示すことができた。

本論文では提案手法がどの程度の精度で得られたデータまで利用可能であるのかは確認していない。今後、より多くのデータを用いて検証することで提案手法を適用するのに必要な実験条件を明記することができるように考えると考えられる。

参考文献

- 1) ArrayExpress at the ebi.
<http://www.ebi.ac.uk/arrayexpress/>.
(2006 年 11 月 現在).
- 2) OU BIOINFORMATICS CORE FACILITY at the University of Oklahoma.
<http://chase.ou.edu/oubcf/>.
(2006 年 11 月 現在)
- 3) Frederic R. Blattner, Guy Plunkett III, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, Ying Shao. The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, Vol. 277, pp. 1453-1462, 1997.
- 4) Axel Bernal, Uy Ear and Nikos Kyrpides. Genomes On Line Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research*, Vol. 29, No. 1, pp. 126-127, 2001.
- 5) Konstantinos Liolios, Nektarios Tavernarakis, Philip Hugenholtz and Nikos C. Kyrpides. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Research*, Vol. 34, Database issue, pp. D332-D334, 2006.
- 6) C.J Stoeckert Jr., H.C. Causton, and C.A. Ball. Microarray databases: standards and ontologies. *Nature Genetics*, Vol. 32, pp. 469-473, 2002.
- 7) Members of the Toxicogenomics Research Consortium. Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods*, Vol. 2, No. 5, pp. 351-356, 2005
- 8) Smith, T.F. and Waterman. M.S.. Identification of common molecular subsequences. *J. Mol. Biol.*, Vol. 147, pp. 195-197, 1981.
- 9) John Aach and George M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, Vol. 17, No. 6, pp. 495-508, 2001.
- 10) Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, Vol. 270, pp. 467-470, 1995.
- 11) Matthew F. Traxler, Dong-Eun Chang and Tyrrell Conway. Guanosine 3',5'-bisphosphate coordinates global gene expression during glucose-lactose diauxie in *Escherichia coli*. *PNAS*, Vol. 103, No. 7, pp. 2374-2379, 2006.
- 12) 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 21st Century*, 2003.
- 13) Quackenbush, J. Microarray data normalization and transformation. *Nature Genetics*, Vol. 32, pp. 496-501, 2002.
- 14) 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の比較のためのデータ補正方式. 情報処理学会 研究報告, 2005-BIO-2, pp. 25-32, 2005.
- 15) 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.