デジタル信号処理に基づく遺伝子のクラスタリング

坪井宣洋† 松田秀雄† 橋本昭洋†

生物の遺伝子発現量の時系列データを利用して各生物が持つ遺伝子を分類する要求が高まっている．現在行われている分類は，いずれも発現データを直接ユークリッド距離などの尺度で比較していた．ところが，実験で得られるデータには実験誤差が含まれており，発現データをそのまま解析することは困難である．そこで，本研究では，離散フーリエ変換やウェーブレット変換などのデジタル信号処理の変換方式を適用し，時系列データの時間成分だけでなく，周波数成分も使った解析を行う．これにより，遺伝子発現量のような誤差を含む時系列データの解析が容易になるものと考えられる．本研究では，すでに全遺伝子の発現パターンが得られている出芽酵母を対象にして遺伝子のクラスタリングを行う．

Clustering of Genes Based on Digital Signal Processing

NOBUHIRO TSUBOI,† HIDEO MATSUDA† and AKIHIRO HASHIMOTO†

There is a growing need for a method of categorizing genes by analyzing large amounts of time series data obtained from gene expression experiments. Current methods of analyzing such data generally involve comparison based on measures such as Euclidean distance, but this direct comparison is difficult since experimentally derived data contains experimental error. In our work, we analyze not only the time component of the time series data but also the frequency component using digital signal processing methods, such as Fourier and wavelet transformations. These methods are considered helpful in analyzing gene expression data and other types of time sequence data that contain error. We apply these methods to the gene cluster analysis of budding yeast, for which expression data for all genes are available.

1. はじめに

これまでに 20 種類以上のゲノムの DNA 塩基配列が解読され，さらに多くのゲノムが解読されつつある．そして，生物が持つ静的な配列データのみならず，遺伝子相互の機能的関連により形成される動的なネットワーク構造を解明することが求められている．そのためには，従来のゲノム解析よりもさらに複雑かつ大量のデータを取り扱わなければならず，新たな情報科学的解析技術の開発が必要不可欠である．

このようなネットワーク構造を解明するための解析の 1 つとして，DNA マイクロアレイ[1]や DNA チップなどから得られる遺伝子発現データから，遺伝子制御ネットワークをブーリアンネットワーク[2]や微分方程式[3),4]などのモデルに基づいて構築する試みがなされている．しかし，そのためには大量の時系列データとして得られる発現データに適合するモデルを作成しなければならず，モデルパラメータの推定が容易ではなかった．

そこで，本研究では遺伝子発現データの解析の前段階として，発現パターン（遺伝子発現量の時間変化）相互の類似性を測る尺度を導入し，それをもとに類似した発現パターンを示す遺伝子を分類することを考えた．同様の発現パターンを示す遺伝子がクラスタにまとめられるので，ネットワーク構築に必要なパラメータ推定の量を減らすことができると考えられる．

発現パターンの類似性尺度としては，現在までにユークリッド距離[5)~7)]，相関係数[8),9)]などが提案されている．しかし，マイクロアレイなどから得られる遺伝子発現量のデータは 2 章で述べるようにかなりの実験誤差が存在するため，これらを直接，このような尺度で比較すると，データによっては正しい結果が得られない可能性があると考えられる．

そこで，本研究では，離散フーリエ変換やウェーブレット変換などのデジタル信号処理の変換方式を適用し，発現パターンの時間成分だけでなく，周波数成分

† 大阪大学大学院基礎工学研究科情報数理系専攻
Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University

も使った解析を行う．これにより，遺伝子発現パターンのような実験誤差を含む時系列データでは，そのデータから特徴量を抽出し，比較を行うことで，実験誤差の影響をある程度抑えることができると考えられる．

本手法の有効性を評価するため，ゲノムの配列が完全に決定され[10]，発現パターンのいくつかが公開されている[1),9)]出芽酵母を対象にしてクラスタリングを行い，その結果について考察する．

2. 遺伝子発現パターン

遺伝子発現パターンとは，ある条件下における遺伝子発現量の時系列データのことである．遺伝子発現量とは，ここでは遺伝子がコードされている領域のDNA塩基配列をもとに作られる転写産物（mRNA）の量を指す．

また，遺伝子間には発現の制御関係がある．遺伝子Aの発現量の増加が遺伝子Bの発現量の増加をもたらすとき，AはBを活性化するといい，Aの発現量の増加がBの発現量の減少をもたらすとき，AはBを不活性化するという．そのような遺伝子の制御関係をグラフ構造で表現したものが，遺伝子制御ネットワークである．

遺伝子発現パターンの類似した遺伝子どうしは同じ制御関係に従う遺伝子（同じ遺伝子により活性化されるか，または同じ遺伝子により不活性化される遺伝子）であると推定され，機能的にも関連していると考えられる．そこで，機能未知の遺伝子Xの発現パターンが得られたとき，機能既知の遺伝子の発現パターンのデータベースに対し，発現パターンの類似性に基づく探索をすることにより，遺伝子Xの機能を予測することが可能となる．さらに，発現パターンの類似性に基づいた遺伝子の分類・整理を行っておけば，遺伝子制御ネットワークの推定に役立つと考えられる．

しかし遺伝子発現量の測定では，以下に述べるように実験誤差が問題となる．

まず，測定での検出感度を上げるためPCR反応による増幅を行うことが多い．しかし，同時に多数（96サンプル程度）のDNAを増幅するので反応条件がそれぞれの遺伝子ごとに最適とはならない．そのため，マイクロアレイに結合するプローブDNA量にばらつきが生じ，発現量に実験誤差が入る[11]．しかし，これについては発現量を測定したいサンプル以外に，発現量の基準となるリファレンスを加えて，それぞれ別の蛍光色素で染色し同じスポットでハイブリダイゼーションを行った後，リファレンスの発現量に対する相対比の形でサンプルの発現量を表すことで補正することができる．発現量の相対比（またはその対数）による表現は文献1), 9)など多くの実験で行われており，本論文でも相対比の対数により発現量を表すことにする．

これ以外の実験誤差の原因として，サンプルとプローブとを特異的にハイブリダイゼーションさせるときのランダムな変動（白色ノイズ）が指摘されている[12]．これについては，前述のリファレンスとの相対比をとることでは解消できない．しかし，ランダムな変動が原因であれば，発現パターンを周波数成分に分解し，高周波成分を取り除くことによって，実験誤差をある程度は補正できると考えられる．

そこで，本研究では，離散フーリエ変換やウェーブレット変換などのデジタル信号処理の変換方式を適用することにより後者の実験誤差を抑えることを目指している．

以下では，類似した遺伝子発現パターンを持つ遺伝子をまとめることにより分類する手法（クラスタリング）について説明する．

3. クラスタリング

3.1 クラスタリングの手法

クラスタリングの手法は，大きく次の2種類に分けられる[13]．

(1) 分割型

(2) 階層型

(1)の分割型は，与えられた集合中のデータをそれらの間の距離に基づいて，あらかじめ決められた個数のクラスタに分割する方法である（距離の開いたところで分割していく）．代表的な方法に k-means 法がある．

(2)の階層型は，やはり与えられた集合中のデータをそれらの間の距離に基づいて分類するが，分割型とは逆に距離の近いデータから順番に，各データを段階的にクラスタにまとめていく．クラスタリングの結果は1つの分類木で表現され，その分類木を適当な階層で切断することにより，部分木の集合が得られ，各部分木にまとめられた遺伝子集合がそれぞれクラスタとなる．

階層型は，分類木を構築していく段階で，すでにできた部分木に対して新たにデータを加えていくときに，どのデータを選択するかの基準の違いにより，さらに細かく単一連鎖（single linkage）法，平均連鎖（average linkage）法，完全連鎖（complete linkage）法などに分かれる．これらの方法では，距離の閾値を決めておいて，あるデータをクラスタにまとめるとき，

それぞれ，そのデータとクラスタ中のデータの間の距離の最小値が閾値以下（単一連鎖），距離の平均値が閾値以下（平均連鎖），距離の最大値が閾値以下（完全連鎖）のときにそのデータをクラスタにまとめるものである．閾値を段階的に大きくしていくことにより，データを階層的に分類木にまとめることができる．

本研究では，クラスタの数をあらかじめ決めることが困難と考え，階層型のクラスタリングを行うことにした．以下では，単一連鎖法，平均連鎖法，完全連鎖法を，それぞれ発現パターンによる遺伝子のクラスタリングに適用し，比較を行う．

3.2 クラスタの定義

前述のように，本研究では，階層型のクラスタリングを用いるため，すべての遺伝子が段階的に接続された分類木を，ある階層で切ることによってクラスタが得られる．

しかし，発現パターンの定量的な比較例が少ないため，発現パターンがどの程度類似していれば同じクラスタと考え，どれだけ異なっていればクラスタを分けるかについて，あらかじめ統一的に決めることは困難である．

本研究では，同一の遺伝子により発現調節されている一群の遺伝子集合（発現パターンは相互に類似する）の中で，直接発現調節されているものと，間接的に発現調節されているものとを分離することが重要と考え，直接発現調節されている遺伝子集合をクラスタと考えることにした．

具体的には，同一遺伝子により直接，活性化（または不活性化）されていることが知られている遺伝子の集合を与え，それらの遺伝子をすべて含む最小の部分木を調べ，その部分木の根の位置と同じ階層で分類木を切ることにより得られる各遺伝子集合をそれぞれクラスタと定義することにする．

ここで，クラスタとして最小の部分木をとると，実際には既知の遺伝子と同じ遺伝子により直接，活性化（または不活性化）されているが，それが知られていない遺伝子をクラスタから落としてしまう可能性がある．しかし，同一遺伝子からの発現調節が既知の遺伝子をできるだけ多く与えることにより，発現調節が既知でない遺伝子をできるだけ含むようにクラスタのサイズを調節できると考える．このクラスタの妥当性については後述する実験の 6 章で検討する．

前述の単一連鎖法，平均連鎖法，完全連鎖法のどれをとるかで，全体の分類木の中で包含関係にない部分木どうしの根の位置の上下関係が変わる可能性がある．そこで，クラスタリングの手法に依存しない，階層の上下関係の基準として，各部分木に含まれる遺伝子集合の中で遺伝子間距離の最大値をとることにした．この値が小さいほどその部分木の根は下位の階層に，大きいほど上位の階層に位置するものとする．

4. 遺伝子発現パターン間の類似性判定尺度

前章で述べたように，遺伝子発現パターンに基づくクラスタリングを行うには，遺伝子発現パターンの間の距離（類似性判定尺度）を定義する必要がある．

遺伝子発現パターンのような時系列データの類似性判定尺度としては，よく使われる方法に次のようなものがある．

(1) ユークリッド距離[5)~7)]

(2) 相関係数[8),9)]

(1) のユークリッド距離は，最も簡単な類似性判定尺度であり，2 つの時系列データを x_i，y_i とすると，

$$\sqrt{\sum_i (x_i - y_i)^2}$$

で定式化される．また，(2) の相関係数は，

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

で定式化され，1 に近いほど正の相関が強く，-1 に近いほど負の相関が強い．

ところが，遺伝子発現パターンには 1 章および 2 章で述べたように実験誤差が含まれているので，上記 (1)，(2) のように，x_i，y_i を直接用いた尺度は問題がある可能性がある．そこで，次章で述べるようにデジタル信号処理の変換方式を適用してデータの加工を行い，新たな類似性判定尺度を設けることにする．

5. デジタル信号処理に基づく類似性判定尺度

本研究では，次の 2 つのデジタル信号処理方式を考え，それらに基づく類似性判定尺度を定義した．

(1) 離散フーリエ変換[14)]

(2) Haar ウェーブレット変換[15)]

以下では，それらの尺度について説明する．

5.1 離散フーリエ変換（DFT）

元の時系列データに対し，次式で定義される離散フーリエ変換を施し，ある次数までの周波数成分を使って解析する．高次の周波数成分を取り除くことにより，元の時系列データの大域的な変動を調べることができる．

$$X_f = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \exp\left(\frac{-j2\pi ft}{n}\right)$$

$f = 0, 1, \ldots, n-1$

2つの時系列データ p, q の距離は，次式のように離散フーリエ変換で得られる係数列をベクトルとしてみたときのユークリッド距離で表す．

ここで係数を何次まで使うかが問題となるが，これについては後述する．

$$Distance(p,q) = \sqrt{\sum_f (X_f(p) - X_f(q))^2}$$

5.2 Haar ウェーブレット変換（Haar WT）

次の Haar ウェーブレット ψ を用いて，元の時系列データを係数 $c_{i,j}$ の系列に変換する．絶対的な値ではなく，値の変動量を調べることができる（**図 1** 参照）．

$$\psi(x) = \begin{cases} 1 & \text{for } 0 \le x < \frac{1}{2} \\ -1 & \text{for } \frac{1}{2} \le x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_{m,n}(x) = \psi(2^{-m}x - n), \quad m > 0, n = 0 \ldots 2^m.$$

$$f = f^0 + \sum_{m=0}^{N} \sum_{l=0}^{2^m} c_{m,l}\psi_{m,l},$$

ただし，f は元の時系列データ，f^0 は f の平均値である．ここで求まる係数 $c_{i,j}$ の列 $Haar(f)$ は Haar 表現と呼ばれる．

$$Haar(f) = \{c_{i,j} : i = s_{max} \ldots s_{min}, j = 1 \ldots 2^i\}$$

s_{max} と s_{min} は，どれだけの時間での変動を調べるかを表している．s_{max} が最も荒い解像度のレベルに対応し，s_{min} が最も細かい解像度のレベルに対応する．

2つの時系列データ f, g の Haar 表現が与えられると，その間の距離は離散フーリエ変換と同様に，次式のように係数の列をベクトルとしてみたときのユークリッド距離で表すことができる．

$$Distance(f,g) = \sqrt{\sum_{i,j}^{m,n} (c_f^{i,j} - c_g^{i,j})^2}$$

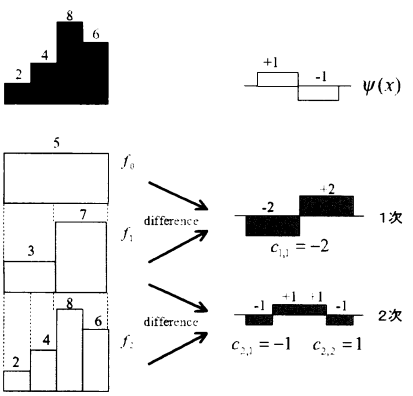

図 1 Haar ウェーブレット変換
Fig. 1 The Haar wavelet transform.

6. 実験と考察

全遺伝子の発現パターンが得られている出芽酵母（*S. cerevisiae*）を対象に実際にクラスタリングを行ったので，その結果について述べる．

6.1 実験 1 diauxic shift

文献 1) で示されている発現パターンを使ってクラスタリングを行った．3.2 節のクラスタの定義で述べたように，同じ遺伝子により直接，活性化されていることが知られている 4 つの遺伝子（HSP12, HSP26, HSP42, CTT1）を与え，それらの遺伝子をもとにクラスタを求めた．

この発現パターンは 7 点の時刻でとった時系列データとなっている．この発現パターンの例として，これら 4 遺伝子の発現パターンを**図 2** に示す．

この 4 遺伝子は同一遺伝子により直接活性化されているため，クラスタリングが 3.2 節で意図したようにうまく行えれば，同じ遺伝子で直接発現調節されているが，それがまだ知られていない遺伝子がクラスタに入ることを除けば，他の遺伝子がクラスタに入ってくることは考えられない．

そこで，これら 4 遺伝子が分類されたクラスタ内の遺伝子数と，そのクラスタの階層上の位置（クラスタに含まれるすべての遺伝子間距離の最大値）を，クラスタリング結果の評価基準とした（どちらも小さい方がよい）．

ただし，遺伝子間距離の最大値はそのままでは，ユークリッド距離や相関係数などの距離の与え方によって

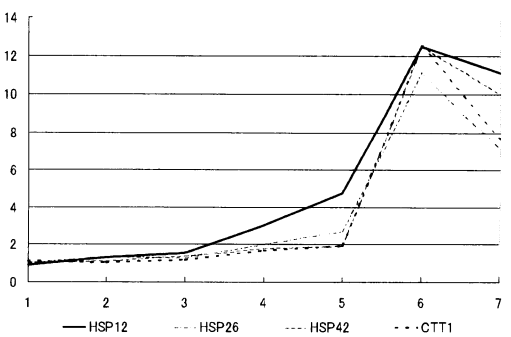

図 2 発現パターンの例（diauxic shift）
Fig. 2 Examples of expression patterns (diauxic shift).

表 1 diauxic shift（Z スコア）
Table 1 diauxic shift (Z score).

	平均連鎖	完全連鎖	単一連鎖
ユークリッド距離	0.769	2.209	5.282
相関係数	0.566	4.073	2.302
DFT (0〜1 次)	0.928	1.078	5.216
DFT (0〜2 次)	1.005	1.005	5.245
WT (1 次)	1.768	2.324	5.364
WT (1〜2 次)	2.470	3.553	4.418
WT (1〜3 次)	5.944	5.944	5.944

表 2 diauxic shift（同一クラスタ内の遺伝子数）
Table 2 diauxic shift (the number of genes in the same cluster).

	平均連鎖	完全連鎖	単一連鎖
ユークリッド距離	31	46	622
相関係数	164	479	361
DFT (0〜1 次)	52	59	631
DFT (0〜2 次)	52	52	631
WT (1 次)	72	128	634
WT (1〜2 次)	102	102	620
WT (1〜3 次)	620	635	632

その値やその分布が大きく異なるため，分布を考慮した基準として次式で表される Z スコアを用いた．

$$Z = \frac{x - m}{\sigma}$$

ここで，x は 4 遺伝子が分類されたクラスタ内の遺伝子間距離の最大値，m はすべての遺伝子間の距離の平均値，σ はすべての遺伝子間の距離の標準偏差である．

表 1，**表 2** はそれぞれ，前述の 4 遺伝子が属するクラスタの Z スコア，およびクラスタ内の遺伝子数を表す．

Z スコア（表 1）での比較では，相関係数を距離として平均連鎖法によりクラスタリングしたものが最小となった．また，クラスタ内遺伝子数（表 2）での比較ではユークリッド距離で平均連鎖法によりクラスタリングしたものが遺伝子数最小となった．

3.2 節で述べたように，前述の 4 遺伝子の他に別の機能未知の遺伝子が，この 4 つの遺伝子と同じ遺伝子によって直接，活性化されている可能性もあるので，クラスタ内遺伝子数や Z スコアが最小であることが必ずしもクラスタリング結果の優劣に結び付くとは限らない．しかし，表 2 の結果では，特にウェーブレット変換に基づく距離を求めたときのクラスタ内遺伝子数がかなり大きく，これだけ多数の遺伝子が同じ遺伝子から直接，発現調節されているとは考えにくい．

以上のことから，このデータのクラスタリングに関しては，ウェーブレット変換の効果があまり出ていないものと考えられる．

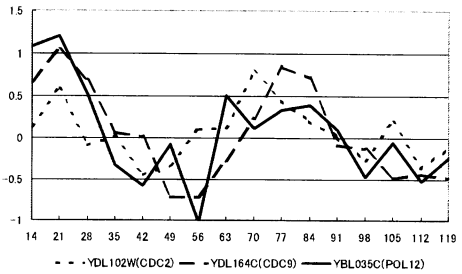

図 3 発現パターンの例（細胞周期）
Fig. 3 Examples of expression patterns (cell cycle).

表 3 細胞周期（Z スコア）
Table 3 Cell Cycle (Z score).

	平均連鎖	完全連鎖	単一連鎖
ユークリッド距離	5.715	8.021	4.122
相関係数	2.251	0.010	2.358
DFT (0〜1 次)	2.223	3.521	0.385
DFT (0〜2 次)	2.503	2.572	1.370
WT (1 次)	0.628	4.398	0.409
WT (1〜2 次)	3.196	1.605	1.350
WT (1〜3 次)	−0.222	4.112	2.300

表 4 細胞周期（同一クラスタ内の遺伝子数）
Table 4 Cell Cycle (the number of genes in the same cluster).

	平均連鎖	完全連鎖	単一連鎖
ユークリッド距離	630	656	523
相関係数	410	113	348
DFT (0〜1 次)	556	595	351
DFT (0〜2 次)	427	515	388
WT (1 次)	535	637	505
WT (1〜2 次)	505	342	415
WT (1〜3 次)	27	593	482

6.2 実験 2 細胞周期

文献 9) で示されている発現パターンを使ってクラスタリングを行い，同じ遺伝子により直接，活性化されていることが知られている 3 つの遺伝子（CDC2, CDC9, POL12）がどの程度まとまっているかを調べた．この発現パターンは 16 点の時刻でとった時系列データとなっている．この発現パターンの例として，前述の遺伝子の発現パターンを**図 3** に示す．

表 1，表 2 と同様に，**表 3**，**表 4** はそれぞれ，Z スコアと，これら 3 つの遺伝子が属するクラスタにまとめられた遺伝子の数を表す．

Z スコア（表 3）での比較と，クラスタ内遺伝子数（表 4）での比較の両者とも，ウェーブレット変換の 1 次から 3 次の係数から距離を求め，平均連鎖法によりクラスタリングした場合の値が最小となった．

実験 1 と同様，この 3 つの遺伝子以外の機能未知の

遺伝子が，この3つの遺伝子と同じ遺伝子により直接，発現調節されている可能性も考えられるが，ウェーブレット変換の1次〜3次と平均連鎖法以外の組合せでは，いずれもクラスタ内遺伝子数が100以上と非常に多い．したがって，少なくともこの実験に関しては，ウェーブレット変換の1次〜3次と平均連鎖法との組合せが良い結果を出しているものと考える．

6.3 観測点の数による結果の違い

実験1と実験2との結果を比較すると，実験1ではフーリエ変換やウェーブレット変換が良い結果を出していないが，実験2ではウェーブレット変換の1次から3次の係数をとったものが最も良い結果を出している．実験1でウェーブレット変換で良い結果が出なかった理由としては，1つには時系列の観測点数が7点と少ないこと，2つめは図2と図3の発現パターン例から分かるように，実験1では発現パターンが単調増加に近く周波数成分での特徴が得にくかったことが考えられる．

そこで，実験2の16点の時刻で観測した時系列データの観測点を8点に減らし Z スコアと同一クラスタ内遺伝子数による評価を行うことにより，観測点の減少がクラスタリング結果に及ぼす影響を調べた．観測点は，16点のうちの奇数番目をとったもので実験を行った（**表5**，**表6** 参照）．

表5と表6から，観測点の減少により結果は悪くなるものの，Z スコアではウェーブレット変換の1次の係数から距離を求め，完全連鎖法または単一連鎖法によりクラスタリングした場合が，クラスタ内遺伝子数ではウェーブレット変換の1次から3次の係数から距離を求め，平均連鎖法によりクラスタリングした場合の値が最小となった．

表5 観測点を減らしたクラスタリング結果（Zスコア）
Table 5 Clustering result by reducing observed points (Z score).

	平均連鎖	完全連鎖	単一連鎖
ユークリッド距離	4.674	7.566	2.423
相関係数	2.000	2.047	1.560
DFT (0〜1次)	3.660	5.735	2.087
DFT (0〜2次)	4.359	5.701	2.659
WT (1次)	5.776	−0.006	−0.006
WT (1〜2次)	6.647	6.647	3.231
WT (1〜3次)	4.327	5.177	2.218

表6 観測点を減らしたクラスタリング結果（同一クラスタ内の遺伝子数）
Table 6 Clustering result by reducing observed points (the number of genes in the same cluster).

	平均連鎖	完全連鎖	単一連鎖
ユークリッド距離	648	656	461
相関係数	369	656	183
DFT (0〜1次)	595	648	538
DFT (0〜2次)	623	631	496
WT (1次)	651	378	378
WT (1〜2次)	628	656	529
WT (1〜3次)	155	633	455

このことから，時系列の観測点数はクラスタリング結果に大きく影響するが，観測点数が少ない場合でも発現量が細かく変化するようなデータに対しては，ウェーブレット変換の効果がある程度得られると考えられる．

6.4 既知の遺伝子発現の制御関係との関連

実験2で使用した16点の時系列データでのクラスタリング結果で最も良い結果が得られたウェーブレット変換の1次から3次の係数から距離を求め，平均連鎖法により得た合計50個のクラスタについて，既知の遺伝子発現の制御関係との関連を調べた．

また，文献8)で使われたクラスタリングプログラム（相関係数を類似度として，平均連鎖法によりクラスタリングを行う）が著者らのWWWサイト☆で公開されていたので，これを使って同様の処理を行った．

文献8)のプログラムにより得られた分類木は，実験2で示した相関係数から距離を求めて平均連鎖法によりクラスタリングにより得られた分類木とほぼ一致した（表4と同様の基準でクラスタ内遺伝子数は152個となった）．実験2での結果と一部違いが出たのは，我々の手法では -1 から1の値をとる相関係数の値を1から引くことにより，相関係数1のとき距離が0で相関係数 -1 のとき距離が2となるよう変換処理をしてから距離の小さい順にクラスタにまとめる平均連鎖法を行っているのに対して，文献8)では相関係数を直接使ってその値の大きいものから順にクラスタにまとめているという違いによるためだと考えられる．

前述の実験2での3個の遺伝子（グループAと呼ぶ）と同様，同じ遺伝子により，直接，活性化されていることが知られている遺伝子の集合2組 {HHF1, HHF2, HHT1, HTA1, HTB1, HTB2}（グループBと呼ぶ）と {EGT2, FUS1, PCL2, PCL9, RME1, SIC1}（グループCと呼ぶ）をYPD（Yeast Protein Database）[16]により調べ，これらのグループがどの程度クラスタと関連しているかを調べた．

その結果，我々の手法ではこれらのグループは別々のクラスタにきれいに分かれ混在することはなかった．また，グループBは1個だけ別のクラスタに分かれた

☆ http://rana.stanford.edu/software/

が残りの5個は1つのクラスタに分類できた．グループCも半数の3個が別々のクラスタに分散したが，残りの3個は1つのクラスタに分類できた．

しかし，文献8)のプログラムの結果では，これらのグループが完全には分離されず，たとえばグループAのクラスタの中にグループCのRME1が入るなどの混在が見られた．

以上のことから，本手法によるクラスタリングは従来手法よりも，同じ遺伝子により直接，活性化されている遺伝子集合を推定するうえで良い結果を出しているものと考えられる．

7. おわりに

デジタル信号処理を利用した，遺伝子発現パターンに基づく遺伝子のクラスタリングの手法を提案した．時系列データにデジタル信号処理の変換方式を適用することにより，観測点が多い時系列データに対してはデータの持つ特徴をうまくとらえ，より良いクラスタリング結果が得られることが確認できた．

今後は，データの性質に応じて，最適な変換方法や変換後の係数をどの次数までとるかなどについて検討することが考えられる．

謝辞 本研究は一部，文部省科学研究費補助金特定領域研究「ゲノムサイエンス」(課題番号08283103)，科学技術振興事業団戦略基礎研究推進事業および計算科学技術活用型特定研究開発推進事業によっている．

参考文献

1) DeRisi, J.L., Iyer, V.R. and Brown, P.O.: Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale, *Science*, Vol.278, pp.680–686 (1997).
2) Akutsu, T., Kuhara, S., Maruyama, O. and Miyano, S.: Identification of Gene Regulatory Networks by Strategic Gene Disruptions and Gene Overexpressions, *Proc. 9th ACM-SIAM Symp. Discrete Algorithms*, pp.695–702 (1998).
3) Chen, T., He, H.L. and Church, G.M.: Modeling Gene Expression with Differential Equations, *Proc. Pacific Symp. Biocomputing '99*, pp.29–40 (1999).
4) D'haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R.: Linear Modeling of mRNA Expression Levels during CNS Development and Injury, *Proc. Pacific Symp. Biocomputing '99*, pp.41–52 (1999).
5) Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X. and Somogyi, R.: Cluster Analysis and Data Visualization of Large-scale Gene Expression Data, *Proc. Pacific Symp. Biocomputing '98*, pp.42–53 (1998).
6) Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. and Somogyi, R.: Large-scale Temporal Gene Expression Mapping of Central Nervous System Development, *Proc. Natl. Acad. Sci. USA*, Vol.95, pp.334–339 (1998).
7) Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M.: Systematic Determination of Genetic Network Architecture, *Nature Genetics*, Vol.22, pp.281–285 (1999).
8) Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein D.: Cluster Analysis and Display of Genome-wide Expression Patterns, *Proc. Natl. Acad. Sci. USA*, Vol.95, pp.14863–14868 (1998).
9) Spellman, P.T., 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).
10) Goffeau, A., et al.: The Yeast Genome Directory, *Nature*, Vol.387, suppl. (1997).
11) 久原 哲，田代康介，牟田 滋：DNAチップの情報科学的取り扱い，数理科学，No.432, pp.33–39 (1999).
12) 大久保公策：ゲノムワイドな遺伝子発現情報の収集と解析—ゲノムプログラムの実行情報からの解読，実験医学，Vol.17, No.19, pp.2537–2543 (1999).
13) Hartigan, J.A.: *Clustering Algorithms*, John Wiley & Sons, New York (1975).
14) Agrawal, R., Faloutsos, C. and Swami, A.: Efficient Similarity Search In Sequence Databases, *Proc. FODO '93*, LNCS No.730, pp.69–84 (1993).
15) Struzik, Z.R. and Siebes, A.: The Haar Wavelet Transform in the Time Series Similarity Paradigm, *Proc. PKDD '99*, LNAI No.1704, pp.12–22 (1999).
16) Hodges, P.E., McKee, A.H.Z., Davis, B.P., Payne, W.E. and Garrels, J.I.: Yeast Proteome Database (YPD): a Model for the Organization and Presentation of Genome-wide Functional Data, *Nucleic Acids Research*, Vol.27, No.1, pp.69–73 (1999).

(平成11年10月29日受付)
(平成11年12月17日再受付)
(平成12年12月28日採録)

坪井　宣洋（学生会員）

昭和51年生．平成10年大阪大学基礎工学部情報工学科中退．現在大阪大学大学院基礎工学研究科情報数理系専攻（修士課程）在学中．

松田　秀雄（正会員）

昭和34年生．昭和57年神戸大学理学部物理学科卒業．昭和59年同大学院工学研究科システム工学専攻（修士課程）修了．昭和62年同大学院自然科学研究科（博士課程）修了．同年同大学工学部助手となり，同大学講師，助教授を経て，平成6年10月より大阪大学基礎工学部情報工学科助教授，現在に至る．この間，平成3年4月より10カ月間米国アルゴンヌ国立研究所客員研究員．学術博士．論理型言語による並列処理，遺伝子情報処理の研究に従事．電子情報通信学会，IEEE CS，ACM各会員．

橋本　昭洋（正会員）

昭和36年大阪大学工学部通信工学科卒業．昭和41年同大学院工学研究科博士課程修了．工学博士．同年NTT電気通信研究所に勤務．昭和44～46年イリノイ大学計算機科学科客員助教授．昭和60年NTTデータ処理研究部長，昭和62年情報科学研究部長．この間計算機の故障診断，自動設計，大型計算機DIPSの開発等に従事．平成元年大阪大学基礎工学部情報工学科教授，平成6年同大学情報処理教育センター長を併任．現在に至る．最近は分子生物学関連の情報処理技術の研究に従事．著書：計算機アーキテクチャ（平成7年，昭晃堂）．電子情報通信学会，IEEE，ACM各会員．