数理モデル化と問題解決 27－2
（1999．11．25）

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等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，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，既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．

このようなネットワーク構造を解明するための解析の一つとして，DNAマイクロアレイ[1]等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，今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，マイクロアレイ等から得られる遺伝子発現データは実験誤差がかなり大きいと見積もられており，発現パターンをそのままこのような尺度で比較すると，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，周波数成分も使った解析を行う．これにより，実験誤差を含む時系列データから，特徴量を抽出し，比較を行うことで，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[11)]，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．

† 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

2. 遺伝子発現パターン

遺伝子発現パターンとは，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．

遺伝子発現量の測定では，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，同時に多数のDNAを増幅するので，生成物の量が遺伝子の長さ等によってばらつきが生じる[12]ため，異なる遺伝子間の発現量の直接の比較は不可能である．そのため，遺伝子発現パターンは，ある時刻の発現量と，発現量測定の基準となる状態の発現量との相対比もしくは相対比の対数値で表すことが多い．

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(減少)をもたらすとき，AはBを活性化(不活性化)すると言う．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は同じ制御関係に従うと推定されるため，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思われる．

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．代表的な方法にk-means法がある．

(2)の階層型は，各データをクラスタにまとめる基準に基づいて階層的にクラスタを構成する方法である．代表的な方法に単一連鎖(single linkage)法，平均連鎖(average linkage)法，完全(complete linkage)法などがある．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で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は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，データを階層的にクラスタ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．

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．

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，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，二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，

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

で定式化され，1(−1)に近いほど正(負)の相関が強い．

ところが，遺伝子発現パターンには1章で述べたように実験誤差が含まれているので，上記のような，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. デジタル信号処理を用いた類似性判定尺度

5.1 離散フーリエ変換[14)]

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．

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

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時のユークリッド距離で表す．

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．

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

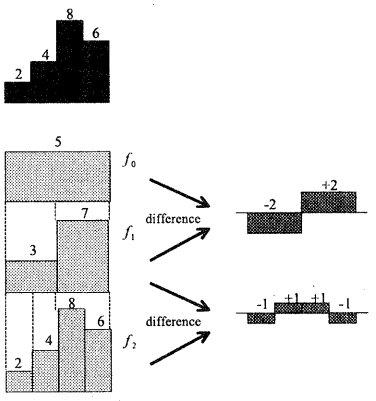

図1　Haar ウェーブレット変換
Fig. 1　The Haar wavelet transform.

5.2 Haar ウェーブレット変換[15)]

次の Haar ウェーブレットψを用いて，元の時系列データを分解し，係数$c_{i,j}$を使って解析する．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 $Haar(f)$ が，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，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．

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

6. 実　　験

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い，同じ遺伝子により制御されていることが知られている 4 つの遺伝子 (HSP12,HSP26, HSP42, CTT1) 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 7 点の時刻でとった時系列データである．

表1　diauxic shift（正規化された距離）
Table 1　diauxic shift (normalized distance)

	平均連鎖	完全連鎖	単一連鎖
ユークリッド距離	1.03641	2.05927	0.47994
相関係数	0.56501	4.17978	0.05184
フーリエ変換 (0〜1 次)	0.90105	1.66466	0.40625
フーリエ変換 (0〜2 次)	0.97211	1.56275	0.44609
ウェーブレット変換 (1 次)	1.2001	2.17709	0.64291
ウェーブレット変換 (1〜2 次)	1.31633	2.93600	0.40269

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 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	365
フーリエ変換 (0〜1 次)	52	59	631
フーリエ変換 (0〜2 次)	52	52	631
ウェーブレット変換 (1 次)	72	52	635
ウェーブレット変換 (1〜2 次)	102	102	620

表 1 はこれら 4 つの遺伝子が属するクラスタ内の各要素間の距離の最大値を，すべての遺伝子間の距離の平均値で割ったもの（正規化された距離）である．この値が小さい程，これら 4 つの遺伝子が属するクラスタ内の遺伝子間距離が小さいことを示す．

また，表 2 はこれら 4 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．表 1 と同様，この値も小さい方がよい．

正規化された距離 (表 1) での比較では相関係数を距離として，単一連鎖法によりクラスタリングしたものがクラスタ内の遺伝子間距離が最小となった．しかし，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．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，時系列データの観測点が 7 点と少なくフーリエ変換やウェーブレット変換で周波数成分を特徴量として取り出した効果が出にくかったことが原因と考えられる．

6.2 実験 2 Cell Cycle

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，同じ遺伝子により制御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である．

表 1，表 2 と同様，表 3，表 4 は，正規化された距離，同一クラスタ内の遺伝子数を表す．

表3　Cell Cycle（正規化された距離）
Table 3　Cell Cycle (normalized distance)

	平均連鎖	完全連鎖	単一連鎖
ユークリッド距離	2.21015	4.72438	0.45873
相関係数	1.03884	1.00456	0.13733
フーリエ変換 (0～1 次)	1.29050	3.68037	0.12293
フーリエ変換 (0～2 次)	1.39875	2.41628	0.20218
ウェーブレット変換 (1 次)	2.69609	1.46468	0.72453
ウェーブレット変換 (1～2 次)	1.56902	2.02722	0.18461
ウェーブレット変換 (1～3 次)	0.56775	3.10212	0.29341

表4　Cell Cycle（同一クラスタ内の遺伝子数）
Table 4　Cell Cycle (the number of genes in 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	350
フーリエ変換 (0～1 次)	556	595	351
フーリエ変換 (0～2 次)	427	515	401
ウェーブレット変換 (1 次)	614	498	412
ウェーブレット変換 (1～2 次)	505	342	482
ウェーブレット変換 (1～3 次)	27	593	482

正規化された距離 (表 3) での比較ではフーリエ変換で 0 次と 1 次の係数から距離を求め，単一連鎖法によりクラスタリングしたものがクラスタ内の遺伝子間距離が最小となった．また，クラスタ内遺伝子数 (表 4) での比較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ものが遺伝子数最小となった．この実験では，時系列データの観測点が実験 1 と比べて 16 点と多いため，フーリエ変換やウェーブレット変換の効果が現れたと考えられる．

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，よりよ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取るかについて検討することが考えられる．

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) 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).
11) Goffeau, A., et al.: The Yeast Genome Directory, *Nature*, Vol. 387, suppl. (1997).
12) Richmond, C. R., Glasner, J. D., Mau, R., Jin, H. and Blattner, F. R.: Genome-wide Expression Profiling in *Escherichia coli* K-12, *Nucleic Acids Research*, Vol. 27, No. 9, pp. 3821–3835 (1999).
13) J. A. Hartigan, *Clustering Algorithms* John Wiley & Sons, Inc., New York (1975).
14) Agrawal, R., Faloutsos, C., Swami, A.: Efficient Similarity Search In Sequence Databases, *Proc. FODO '93*, LNCS No. 730) (1993).
15) Struzik, Z. R., Siebes, A.: The Haar Wavelet Transform in the Time Series Similarity Paradigm, *Proc. PKDD '99*, LNAI No. 1704, pp. 12-22 (1999).