生物学的知識を導入した遺伝子発現データの自動分類

生物学的知識を導入した遺伝子発現データの自動分類

生物学的知識を導入した遺伝子発現データの自動分類

岡田吉史[1)]、三林　光[2)]、長島知正[1,2)]
1）室蘭工業大学　サテライト・ベンチャー・ビジネス・ラボラトリー
2）室蘭工業大学　情報工学科

1．はじめに

DNAマイクロアレイの登場により，細胞内における数千～数万の遺伝子発現量（発現プロファイル）を同時に計測できるようになった．そのような極めて膨大な発現データから生物学的に有用な情報を抽出することは，人手による処理の限界を超えるため，情報技術の利用が必須となっている．

現在，マイクロアレイ実験によりもたらされた膨大な遺伝子発現データは，公共のゲノムデータベースに蓄積され，個々の遺伝子ごとに専門家による注釈付け（アノテーション）や関連知識の構造化が精力的に進められている．本研究のねらいは，このようなデータベースに蓄積される生物学的知識を既存のクラスタリング手法に導入することにより，計算機上で専門家の知識処理を導入した知的な遺伝子分類を実現することである．

以下，本稿では，公共ゲノムデータベースで定義される遺伝子のアノテーションを参照し，クラスタ間で遺伝子機能（の内訳）が互いに独立性を示すようなクラスタを自動生成する方法[1]を述べる．

2．遺伝子分類と階層的クラスタリング

特定環境下において，発現パターンが類似している遺伝子が存在するならば，それらは細胞内において同様の制御プロセス下にあり，生物学的機能も似ていると考えられている．したがって，発現パターンが類似した遺伝子同士に分類することで，特定の刺激に特徴的・特異的に応答する遺伝子群を抽出したり，機能既知の遺伝子発現パターンから未知遺伝子の機能を推定できるようになると考えられる．このように，遺伝子分類作業は，より高次のタスクへ拡張するための重要なステップとして位置づけられる．

一般に，コンピュータを用いて遺伝子分類を行う方法として，しばしば階層的クラスタリングが利用される[2]．

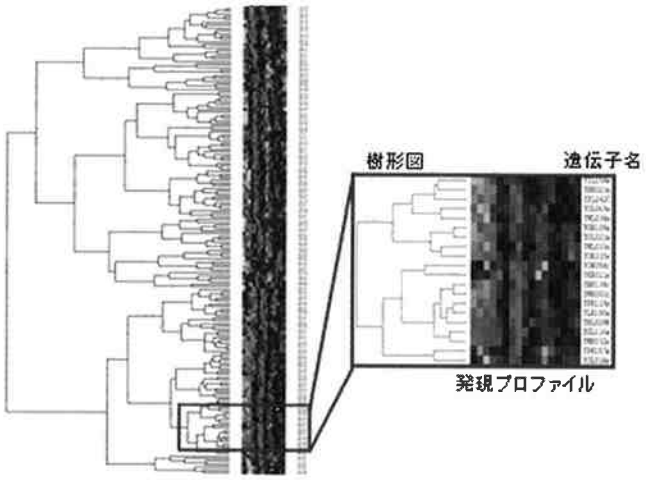

図１．遺伝子発現プロファイルの階層的クラスタリング結果

階層的クラスタリングは，遺伝子の発現パターンが似たものから順次融合してクラスタを作り，最終的なクラスタリング結果は樹形図の形式で与えられる（図1）．しかしながら，従来のクラスタリングはあくまでもデータの統計的な特徴を与えるに過ぎないため，樹形図から生物学的に意味のある遺伝子セットを抽出する作業は，解析者のトライアンドエラーによるところが大きかった．この作業は，データ数が膨大な場合には非常に困難な作業となるが，これまでのところ，生物学的知識に基づいて意味のあるクラスタを自動的に識別する合理的な手法は存在していない．我々の提案する方法は，遺伝子のアノテーション情報に基づき，クラスタ間で遺伝子機能分布が適切に分離し，個々のクラスタが機能的にまとまりをもつように，樹形図を利用して最適なクラスタ境界を自動的に推定する．本法により，上記に述べたようなクラスタリングを用いた遺伝子分類における手間を大幅に削減できると考えられる．

3．クラスタ境界の自動決定

以下，我々が提案するクラスタ境界の自動決定アルゴリズムについて述べる．

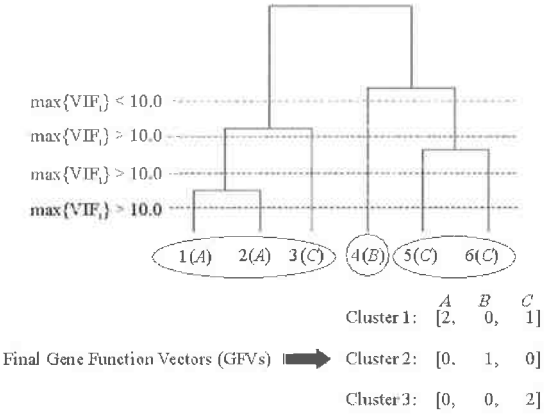

図 2. クラスタ境界の自動決定例

まず，発現プロファイルに対して階層的クラスタリングを行い，発現プロファイル間の類似性を階層的に表す樹形図を作成する（前頁の図1）．ただし，クラスタ間の類似度として余弦距離，クラスタの融合法には完全連結法を用いた．

次に，得られた樹形図の階層ごとに，クラスタ間の遺伝子機能分布の独立性を見積もることによりクラスタ数を自動決定する．本手法の基本的な考え方は，階層ごとにカットオフを設定して，得られるクラスタの遺伝子機能分布をベクトルとして表現し（MIPSデータベース[1]で定義される81種類の機能を参照），それらのベクトル間の独立性を評価することである．以下，このベクトルをGFV（Gene Function Vector）と呼ぶ．クラスタ間の GFV 間の独立性を見積る指標として，VIF（Variance Inflation Factor）を用いた．VIF は，重回帰分析において説明変数間の多重共線性を評価し，独立な説明変数を選択する際の指標として用いられる．VIF は，GFV 間の相関係数行列の逆行列の対角要素であり，通常は，共線性の有無の閾値として 10.0 が採用される．もし，あるクラスタの GFV に関する VIF 値が 10.0 以上ならば，そのクラスタは他のいずれかのクラスタと GFV について共線性があること示している．すなわち，遺伝子機能分布がよく似たクラスタが存在していることを意味している．もし，VIF 値が全て 10.0 より小さいならば，全てのクラスタの GFV が互いに独立である，すなわち，クラスタ間の遺伝子機能分布は互いに独立であると見なされる．本研究では，全ての VIF 値が 10.0 より小さくなる最初の階層を最適なクラスタ境界とした．

上記の処理を，図2に示す具体例で説明する．この図では，A, B, C のいずれかの機能を持つ6つの遺伝子の樹形図が示されている．最も低い階層では6つのクラスタが存在する．したがって，この階層でカットオフを設定すると，6つの GFV が作られる．ここで，遺伝子機能の種類数は A, B, C の3種類であるので，GFV は3次元ベクトルで表される．次に，GFV の相関係数行列の逆行列の6つの対角要素の最大値が 10.0 より小さいかどうか判定する．この条件を満たさない場合，6つのクラスタのいずれかが共線性を示すと判定され，次に高い階層である5つのクラスタについて判定を行う．この例では，クラスタ数が3のときに，上記の判定条件がはじめて満たされており，最適なクラスタ数は3と見積もられている．

4．酵母の細胞分裂周期データを用いた評価実験

本手法を，Cho らによって測定された酵母の細胞分裂周期に関するマイクロアレイデータに適用した．

4．1　細胞分裂周期データ

Cho らは，酵母の6000遺伝子に関する経時的な発現量を調べるため，10 分間隔・17 時間点においてマイクロアレイ実験を行った[3]．彼等は，得られた遺伝子発現プロファイルに基づき，発現ピークに応じて遺伝子を分類し，目視観察により5つの細胞分裂期（Early G1, Late G1, S, G2, M 期）に対応する 416 個の遺伝子を同定した．本稿では，これらの 416 遺伝子の発現プロファイルから欠損値を持つものを除去した 384 遺伝子の発現プロファイルに本法を適用し，評価実験を行う．

4．2　最適クラスタ数の決定

本手法に基づいて，384 遺伝子発現プロファイルの最適クラスタ数を決定した．図3に，VIF 値が 10 以上となるクラスタの割合を示した．クラスタ数が 57 以上の場合には，GFV の相関係数行列において従属性の強い GFV が存在し，ランク落ちが発生するため逆行列の算出が不能であった．したがって，図ではクラスタ数が 56 以下の場合について示されている．この図からわかるように，クラスタ数が少なくなるにしたがい，共線性を示す

[1] MIPS: http://mips.gsf.de/genre/proj/yeast/index.jsp

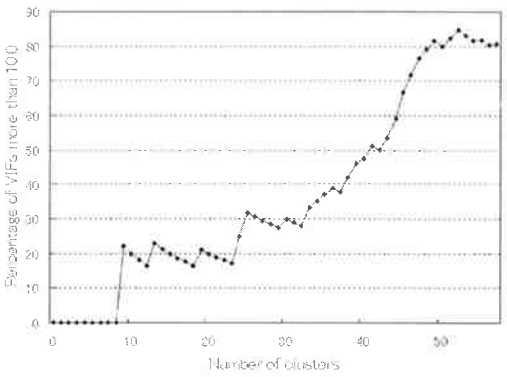

図 3. クラスタ間の独立性

横軸は樹形図の各階層でカットオフを設定して得られるクラスタ数，縦軸はすべてのVIFのうち 10.0 以上を示すVIFの割合である。

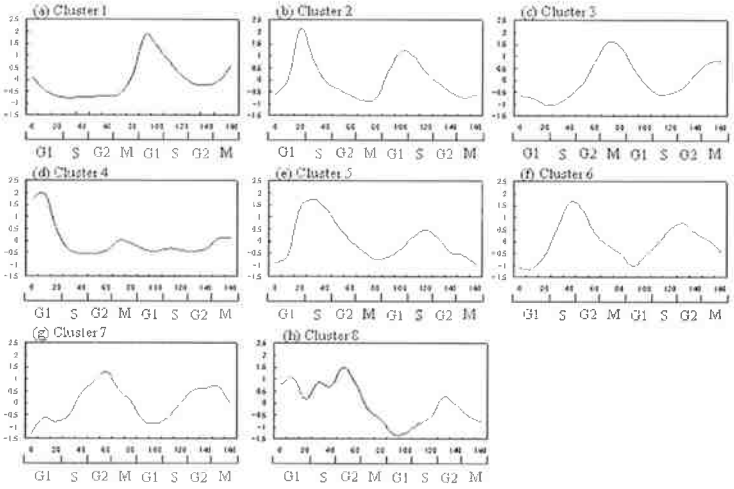

図 4. 各クラスタの遺伝子平均プロファイル

表 1. 各クラスタにおける細胞分裂期の遺伝子内訳

Cluster	Early G1	Late G1	S	G2	M
1 (71)	48	17	5	0	1
2 (141)	5	112	24	0	0
3 (66)	12	0	0	5	49
4 (6)	2	1	1	1	1
5 (20)	0	5	14	1	0
6 (45)	0	0	25	19	1
7 (25)	0	0	3	19	3
8 (10)	0	0	3	7	0

GFV の割合は徐々に減少し，クラスタ数が 8 個でその割合は0%になっている．つまり，クラスタ数の減少に伴い遺伝子の機能分布の類似したクラスタは順次融合され，8クラスタにおいてはじめて全てのクラスタ間の遺伝子機能分布が独立性を示したといえる．したがって，本手法により，384 遺伝子の発現プロファイルに関する最適なクラスタ数は8個と判定された．

4．3　Cho らの分類結果との比較評価

本手法で得られた結果の妥当性を調べるため，上記で生成された8クラスタに含まれる遺伝子分布と，Choらの手作業による分類結果(Early G1, Late G1, S, G2, M 期の5グループ)との比較を行った．表1は，各クラスタにおける5つの細胞分裂期に属する遺伝子の内訳である．クラスタ1,2,3,5,7 は，それぞれEarly G1, Late G1, M, S, G2 期の遺伝子を特に多く含んでいる．また，図4は，生成された8クラスタに含まれる遺伝子の平均発現プロファイルである．この図からも，クラスタ1,2,3,5,7 のプロファイルは，それぞれ Early G1, Late G1, M, S, G2 期において発現レベルのピークをとっていることがわかる．このように，本手法は Cho らによって分類された5つの細胞分裂期をよく特徴付けるクラスタを適切に生成できていることがわかる．

4．4　クラスタ妥当性指標を用いた評価

本手法は，クラスタ間の遺伝子機能の分布が互いに独立性をなすようなクラスタ，すなわち個々のクラスタが機能的にまとまりを示すようなクラスタを構成することを狙いとしている．そこで，クラスタ間での機能の分離具合および各クラスタ内部での機能のまとまり具合を調べるため，一般的なクラスタ妥当性指標を用いた評価を行った．ここでは，機能の分離具合を表す Separation，機能のまとまり具合を表す Entropy に関する結果を示す．これらの妥当性指標を用いたクラスタ評価の詳細については，文献[1]または[4]を参照されたい．評価は，本研究と同様の遺伝子分類の目的で使用される3つのクラスタリング法(k-means[5], SOM[6], Autoclass[7])に関するクラスタ妥当性との比較により行われた．表2は，本手法(Hierarchical)を含む4つのクラスタリング法によって生成されたクラスタの Separation と Entropy を示している．これらの値は，小さいほど良いクラスタであると見なされる．1行目の括弧内は，各手法で生成される最適クラスタ数である．ここで，k-means と SOM に関しては，本手法と同様に VIF に基づいてクラスタ数を決定した．

表 2. クラスタ妥当性指標を用いた他クラスタリング法との比較

Algorithm	Cluster validity index	
	Separation	Entropy
Hierarchical (8)	0.843	0.721
K-means (7)	0.928	0.707
SOM (6)	0.941	0.946
AutoClass (7)	0.950	0.722

AutoClass はベイジアンネットに基づいて最適クラスタ数を自動決定するため，生成されたクラスタ数をそのまま用いた．Separation に関しては，本手法が最小値を示しており，クラスタ間の遺伝子機能の分布が最もよく分離されていることを示している．Entropy については，本手法は4つの手法の中で2番目に小さく，個々のクラスタが特定の遺伝子機能によって特徴付けられ，機能的なまとまりを示しているといえる．以上の結果から，本手法が生成するクラスタは4つの手法の中で遺伝子機能の分離・まとまり具合の観点から，最も良好なクラスタを構成しているといえる．

5．まとめ

本研究では，公共ゲノムデータベースで定義される遺伝子のアノテーション情報（既知遺伝子の機能情報）を参照し，クラスタ間で遺伝子機能内訳が互いに独立性をなすようなクラスタを自動的に生成するアルゴリズムを開発した．本稿では，評価実験として，細胞周期におけるG1期，S期，G2期，M期のような生物学的に異なるプロセスで機能する遺伝子群を効率的に分離できることを示し，さらに，他のクラスタリング手法（k-menas，SOM，AutoClass）とのクラスタ妥当性の比較評価をとおして，提案手法が生物学的に意味解釈可能な優れたクラスタを構成することを示した．本研究により，ウェット実験を行う生物学研究者の遺伝子分類作業における手間を大幅に削減できることが期待される．

今後は，生成されたクラスタから，系統的に未知遺伝子の機能や遺伝子間（クラスタ間）のネットワークを自動推定する技術へと発展させていきたい．

参考文献

[1] Y. Okada, T. Shahara, H. Mitsubayashi, S. Ohgiya, T. Nagashima, Knowledge-assisted recognition of cluster boundaries in gene expression data, Artificial Intelligence in Medicine, 2005 (accepted).

[2] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA. 95*, 14863-14868, 1998.

[3] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E Gabrielian, D. Landsman and D.J. Lockhart, Genome-Wide transcriptional analysis of the mitotic cell cycle, *Mol. Cell. 2*, 65-73, 1998.

[4] J. He, A. H. Tan and C.L. Tan, Modified ART 2A growing network capable of generating a fixed number of nodes, *IEEE Trans. on Neural Networks. 15* , 728-737, 2004.

[5] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church, Systematic determination of genetic network architecture, *Nat. Genet. 22*, 281-285, 1999.

[6] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander and T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci. USA 96*, 2907-2912, 1999.

[7] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor and D. Freeman, Autoclass: a Bayesian classification system, *Proc. 4th International conference on Machine Learning*, 54-64, 1988.