3V-05　クラスタ分析によるＤＮＡチップ遺伝子発現データの分類

中重 亮　野崎康行　渡辺恒彦　田村卓郎
日立ソフトウェアエンジニアリング（株）

１．はじめに

最新遺伝子解析技術としてＤＮＡチップが登場し、細胞内での複数遺伝子の働きを一挙に観測して、データ化することが可能となった。

一枚のＤＮＡチップからは数千から１万程度の遺伝子について、ある細胞内におけるそれぞれの働き具合を表す発現強度データが一度に得られる。したがって、バイオ研究者がこのデータから生物学的に新しい意味情報を抽出する際には、研究対象としてどの遺伝子に着目すべきかを選択するために、まず発現データに応じて分類し、遺伝子をグループに分ける処理が重要である。そこで、我々は統計分析の分野で広く利用されているクラスタ分析手法に着目し、遺伝子発現データに対する適用を試みた。

２．ＤＮＡチップシステムの概要

図１はＤＮＡチップを利用した生物実験の例を示している。ここではマウスの脳の中で特別な働きをする遺伝子を調べるための実験を想定している。まず、マウスの全遺伝子(約１０万個)の中で、脳内で特別な働きをすると思われる遺伝子約１万個を選別しておき、これらの遺伝子に対して検出用の相補的なＤＮＡをチップ化しておく。このマウス用のＤＮＡチップに対して、マウスの脳における各部位の細胞から採取した試料（発現遺伝子に蛍光ラベルを施したもの）を反応させると、細胞内で活動中のそれぞれの遺伝子に対応するところで蛍光ラベルを捕獲させることができ、蛍光スキャナ装置によりこれを観測してデータ化することができる。実際にシステムから出力される発現データは、特定の成長時点や個体内の部位など、それぞれの実験における各遺伝子の発現強度の計測値である。その変化も含めて捉えたデータを遺伝子の発現パターンと呼んでいる。

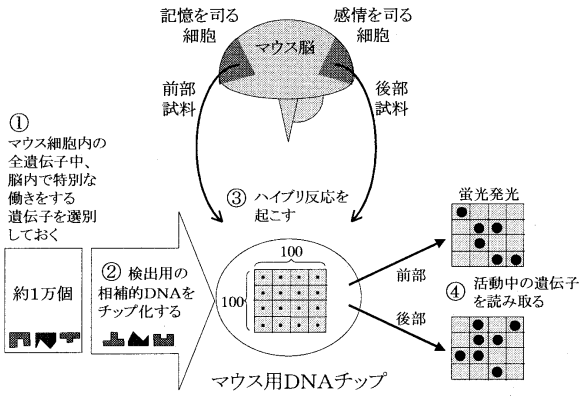

図１ ＤＮＡチップを利用した生物実験例

３．マウスの実験データへの適用

クラスタ分析プログラムを試作し、マウスの実験データに適用して以下の結果を得た。クラスタ分析には非類似度として最も代表的なユークリッド平方距離を採用した。また、併合アルゴリズムとしては最短距離法とウォード法の二つを実装してその分類結果を比較し、それぞれの効果を確認した。

データに関してはＤＮＡチップによる発現データの入手が困難だったため、チップから得られるデータと等価な従来手法によるマウス遺伝子の時系列発現データ（遺伝子数：５，９０６、誕生からの成長：３時点）を入手し、クラスタ分析プログラムを適用した。この分析結果を樹状図で表示し、遺伝子定義情報から抽出した生物学上のキーワードに着目して分類状況を考察した。

４．クラスタ分析による分類結果

図２に示すように、最短距離法の樹状図からは、ほとんどのクラスタ併合において要素を一つづつ吸収する処理になっており、たくさんの

Gouping of DNA Chip Gene Expression Data Using Cluster Analysis
Ryo Nakashige, Yasuyuki Nozaki, Tsunehiko Watanabe, and Takuro Tamura
Hitachi Software Engineering Co., Ltd.

要素を含む大きなクラスタが１個だけ成長していった様子が確認できた。それ以外では要素数５以下の孤立したクラスタが３６個あった。これらは発現パターンが他と非常に異なるものから構成されている。一方、ウォード法の樹状図からは各クラスタ内要素数が比較的平均化した分類が得られており、高さ１％以内の部分木が全部で２４個、そのうち約６割の１４個が５以上１００以下の要素を含んでいた。

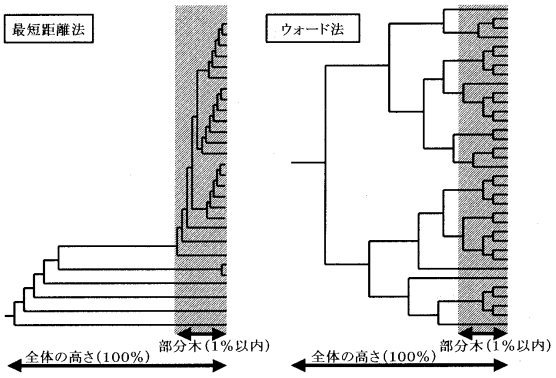

図２　クラスタ分析結果（樹状図）の定性的差異

キーワードに着目した分類状況については、代表的な３種類のキーワード（リボソーム・ＡＴＰ・ミトコンドリア）に関連する遺伝子が、いくつかの部分木の中に発現パターンの似かよったもの同士で集まっていることが確認できた。特に、併合アルゴリズムにウォード法を採用した場合、リボソームに関連する５５個の遺伝子の中で、２個以上の遺伝子が集まったグループを１１個確認できた。これらをまとめたものが表１である。ここでは部分木に対する閾値として樹状図全体の高さの１％以下を採用し、これを満たす部分木内に集まった２個以上の遺伝子は同じグループに属すると考えた。

今後は、まず実際のチップデータで同様の分析結果が行えるかどうか、ＤＮＡチップデータに特化した調整が必要かどうかを検証する必要がある。また、図２および表１の結果から遺伝子グループは併合アルゴリズムの選択で大きく変わることが分かる。最短距離法では大きなグループが１個と、孤立した小さなグループが多数できたが、ウォード法ではそれぞれのグループの大きさに関してバランスの取れた分類が得られている。このため、それぞれのアルゴリズムを選択した場合の分類が、遺伝子機能の観点からどのような特徴を持つか、相互に結果を比較して詳細を考察する必要がある。

表１　クラスタ分析による遺伝子グループ

#	特徴量		最短距離法	ウォード法
1	リボソーム	関連遺伝子の総数	５５	
		グループ数	２	１１
		１グループ内の遺伝子数	４６，２	１０，７，６，５，５，４，４，３，３，３，２
2	ＡＴＰ	関連遺伝子の総数	１４	
		グループ数	１	３
		１グループ内の遺伝子数	１１	４，２，２
3	ミトコンドリア	関連遺伝子の総数	１２	
		グループ数	１	３
		１グループ内の遺伝子数	１１	３，３，２

５．まとめ

ＤＮＡチップから得られる大量の遺伝子発現データから生物学的に意味のある情報を取り出すデータマイニングを行うために、クラスタ分析手法の適用を検討した。今後はＤＮＡチップデータに対する適用と各種距離・併合アルゴリズムの効果の比較を行う。

参考文献

[1] 松原謙一，中村桂子：ゲノムを読む：紀伊國屋書店（1996）

[2] 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)

[3] 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（特集／ゲノム解析），Vol.37，No.6，pp.33-39（1999）