# k平均法の初期条件依存度を用いた遺伝子の分類：DNAマイクロアレイデータの解析

# k平均法の初期条件依存度を用いた遺伝子の分類 : DNAマイクロアレイデータの解析

# k平均法の初期条件依存度を用いた遺伝子の分類
## —DNAマイクロアレイデータの解析

野村收作[1)], 広瀬臣吾[2)], 田中秀典[2)], 岡田吉史[1)], 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. 序

近年のゲノム科学におけるウェット実験系の急激な技術革新は、同時に情報技術の需要を飛躍的に向上させ、バイオインフォマティクスという1分野を開いた。そのバイオインフォマティクスも、今や単なる実験結果の解析技術から、より積極的にデータマイニング技術へと焦点が深化している。本研究ではそのようなバイオインフォマティクス研究の一端として、DNA マイクロアレイのデータを用い、ウェット実験に還元できるような DNA アノテーション情報（機能情報）のマイニングを試みた。特に DNA マイクロアレイデータのクラスター解析に焦点をあて、階層－非階層的クラスタリングの比較、非階層的クラスタリングにおける初期条件依存度を検討した結果、解析における諸境界条件に強く依存する遺伝子（群）、反対に殆ど依存しない遺伝子（群）の存在が示され、各遺伝子（群）に対する新規のクラスター化およびアノテーション付け方法が検討された。

### 2. DNA マイクロアレイとバイオインフォマティクス

近年、ゲノム科学において発達した強力な実験技術のひとつとしてDNA マイクロアレイ技術が挙げられる。DNA マイクロアレイはガラス基盤上に数千～数万個の遺伝子（対応する cDNA）をスポットッティングしたもので（図1）、これは例えばある生物個体内の全遺伝子の発現状況を同時かつ網羅的に調べ上げることができることを意味する。特に、DNA マイクロアレイを多数用いれば、生体内のあらゆる遺伝子の動態を経時的に追跡出来ることになり、複雑に絡み合った生体内の遺伝子発現ネットワークの解明に大いに寄与すると考えられる。他方、このような DNA マイクロアレイ技術によってウェット実験から得られるデータ量は膨大になり、それを整理・分類するような、或いは有用な情報・理論をマイニングするような情報技術（バイオインフォマティクス）の導入が不可欠となっている。具体的に前者としてはクラスター解析、主成分分析など、後者としてはネットワーク推定理論が挙げられる。しかしながら、あらゆる情報技術がそうであるように、各々の解析手法・測度の選択（広い意味で「境界条件」の選択とする）は解析に携わる研究者に委ねられている。従って、それら境界条件を生物学的な知識を活用して合理的に選択する必要がある。

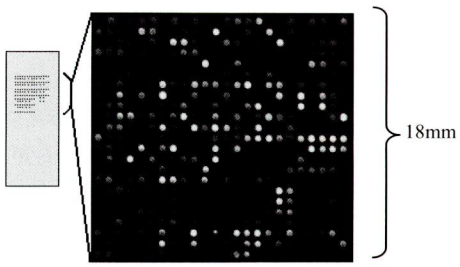


図1：DNA マイクロアレイと発現データ例

我々は既にクラスター解析における諸問題を検討し、クラスター解析を用いた未知遺伝子の機能推定方法として、既知遺伝子情報を導入した教師付きk平均クラスタリング法を提案し、模擬データを用いた計算機実験によりその性能を確かめた[1]。これに対し本研究では、実際のウェット実験のデータを用いてクラスター解析を行い、境界条件の問題を検討した。

### 3. クラスター解析

本研究では独立行政法人産業技術総合研究所北海道センターで行われた酵母（*S.cerevisiae*）の低温刺激による全遺伝子の DNA マイクロアレイ経時発現データ（遺伝子数約 6000 個×5 枚）を用いた。この実験の目的は主に、低温刺激によって発現（抑制）する遺伝子を整理・分類し、最終的に低温時における遺伝子発現ネットワークの解明（低温下における生体機能の解明）にある。従って、先ず各遺伝子をその動態によって分類しなければならない。しかしながら、ウェット実験が“低温刺激”時における各遺伝子の機能（発現の動態）をターゲットにしているため、先の研究のように既存の（MIPS[2]データベースなどにおける）既知遺伝子の情報を素朴に使うことはできない。換言すれば、“正解”が与えられない状態でよりよい境界条件を選択する必要がある。そこで本研究ではクラスター解析に関わる諸境界条件を検討した。

クラスター解析は主に階層的クラスタリングと非階層的クラスタリングに分類される。前者は互いに近い要素同士を一組づつ結んで行く方法で、結果は樹形図で示される。ただし、一般的に計算量が多くなることや、“似たもの（同じ

もの）同士”が多い環境だとクラスタリング結果は測度の選び方やクラスタリング順序によって大きく異なる。後者（非階層）は予め核となる点をいくらか用意し、各要素はそれぞれ一番近い核のあるグループへ編入される。その後、各グループで重心等を計算して新たな核を生成し再び各要素との距離を測りグループを再構成する。この様な手続きが収束するまでグループを再構成してゆく。この場合階層的クラスタリングのような問題は生じないが、その代わり非階層の場合は初めの核の選び方によってクラスタリング結果が大きく異なるという問題が生じる（特に要素数が少ない場合）。また、核の個数（クラスター数）を初めに決めなければならないという問題点もある（階層的クラスタリングではグループ数を事後的に解析者が選択できるが、原理的に自動決定出来ない点では同じである）。

## 4. 階層－非階層クラスタリングの比較

前項で述べたような“正解”が無い状況では、階層・非階層双方ともにどちらが良いという議論は出来ない。しかし、非階層クラスタリングの初期状態依存問題は核をランダムに変更して計算を繰り返すことにより間接的に回避されると思われる。また、今回取り扱う生データは約6000個の 5 次元ベクトルあり、例えば各要素を＋（発現）、－（抑制）で考えた場合、大まかなベクトルのパターンは $2^4 = 16$ 通りである。従って、類似のデータが多数存在すると思われるので非階層的クラスタリングのほうが良いと思われる。

そこで、非階層的クラスタリングの結果から有用な情報を引き出すことを考え、先ず階層クラスタリングと非階層クラスタリングの結果を比較した。生データとしては比較を容易にするため、MIPS データベースにおいて“energy”と“metabolism”に分類されている340個の遺伝子を使用した。非階層クラスタリングとしてk平均法を用いた。グループ数は階層クラスタリングの結果から事後的に選んだグループ数である8を、さらに初期条件（核）をランダムに変え 1000 回計算実験を行った。距離測度としては階層・非階層ともに

$$\text{Standard Correlation} = \boldsymbol{a}\cdot\boldsymbol{b}/|\boldsymbol{a}||\boldsymbol{b}|$$

を用いた。但し $\boldsymbol{a},\boldsymbol{b}$ はある遺伝子の発現データベクトル。

表1：階層－非階層クラスタリングの比較（一致率）及び非階層クラスタリングの収束回数

| | 平均 | 最大 | 最小 |
|---|---|---|---|
| 一致率 | 79.7% | 81.6% | 70.9% |
| 収束回数 | 13.5 回 | 38 回 | 4 回 |

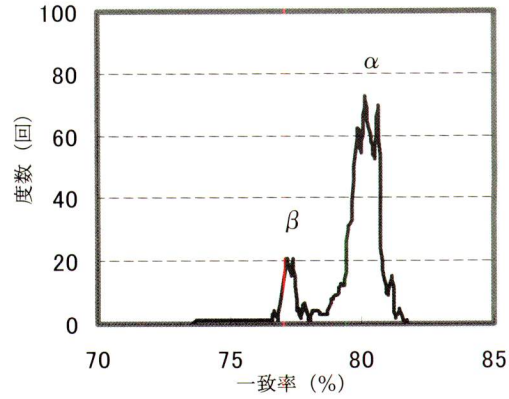


図2：階層－非階層クラスタリングの一致率の分布

表1は階層－非階層クラスタリングの“一致率”と非階層クラスタリングにおける収束までの回数をそれぞれ示している。“一致率”とは、ある 2 つの要素だけに注目した時にそれらが階層－非階層とも同様に分類（同じグループか否か）されている確率を表す。表 1 によると、階層－非階層の比較においては平均約 80%の一致率となっており、各々の遺伝子の分類は良く一致している。ただし、今は“正解”無しの状況なのでただ「良く一致している」以上でも以下でもない。また、階層クラスタリングの結果は 1 種類のクラスタリングパターンのみであることを考えれば、むしろ非階層クラスタリングにおける初期条件由来のパターンのバリエーションに注目したほうが有用であると思われる。

図2に 1000 回の計算実験における一致率の度数分布を表す。初期条件をランダムに設定したにも関わらず、一致率は2つのピークをもつ非連続的な分布を示すことがわかる。非階層クラスタリングにおけるパターンのバリエーションは表 1 における一致率のばらつきであると仮定すれば、これは非階層クラスタリングにおける代表的な2つのパターンの存在を示唆すると思われる。そこで図2の各ピーク（図2中のα、β）におけるクラスター構成（それぞれパターンα、βとする）を詳しく比較してみたところ、パターンαにおけるクラスター内のある遺伝子群がそっくりパターンβの別のクラスターと融合していたり、或いは他のクラスター内の遺伝子群と融合して新しいクラスターを形成していたりすることが分かった。図3は a)パターンαの例と b)クラスター内の遺伝子 ID である（A,B,C クラスターのみ）。遺伝子 ID の内、赤色と青色のものは群としてパターンβでも現れることを意味している。特に赤色のものはパターンβでもクラスター内の過半数を占めており、非階層クラスタリングの初期状態変化に対して“安定”なグループと受け止められる。反対に黒色の遺伝子は単独で他のグループに現れ、初期状態に強く依存していると考えられる。

## 5. 議論と結語

前項の結果は、非階層クラスタリングで階層化されたクラスター内にも階層構造を見出せることを示している。換言すれば、初期グループ数の設定に関わらず非階層クラスタリングにおける現実的な（最適ではないかもしれないが）クラスター数の決定が可能であることを示している。また、黒色にコードされた遺伝子は特にはずれ値のような値ではないことから（その様なデータはクラスタリングの前処理で既に除かれている）、他の遺伝子の発現パターンとの関係で常にクラスターの境界上にある遺伝子であると考えられ、機能推定の見地から注目に値すると思われる。また、パターンα、βにおいてほぼ同一と考えられるクラスターは8クラスター中1つしかなかった。他は赤、青にコードされた遺伝子群が互いに組み合わさって全く別々のクラスターを構成していた。このことはクラスター単位でアノテーション付けを行っていた場合、その内容が大きく異なることを意味する。

本研究では、手続きを容易にする為、非階層クラスタリングに孕む初期条件依存問題を階層クラスタリングと比べることで示したが、本来は（自動化するのは非常に難しいが）、非階層クラスタリングのあらゆるパターン間で比較し、個々の遺伝子の初期条件依存性を調べる必要がある。その上で初期条件依存性の無い遺伝子グループにおけるアノテーション付け、及び、依存性の高い遺伝子に対する個別的なリサーチが必要と考えられる。

表2："energy"と"metabolism"に分類された遺伝子340個中で黒色（単独でクラスター間を移る）にコードされたもの

| ORF | Gene name | Classification |
|---|---|---|
| YCL025c | AGP1 | known protein |
| YDL085w | NDH2 | known protein |
| YDL247w | - | strong similarity to known protein |
| YDR111c | - | strong similarity to known protein |
| YFL030w | - | similarity to known protein |
| YFR015c | GSY1 | known protein |
| YKL087c | CYT2 | known protein |
| YMR013c | SEC59 | known protein |
| YPL171c | OYE3 | known protein |

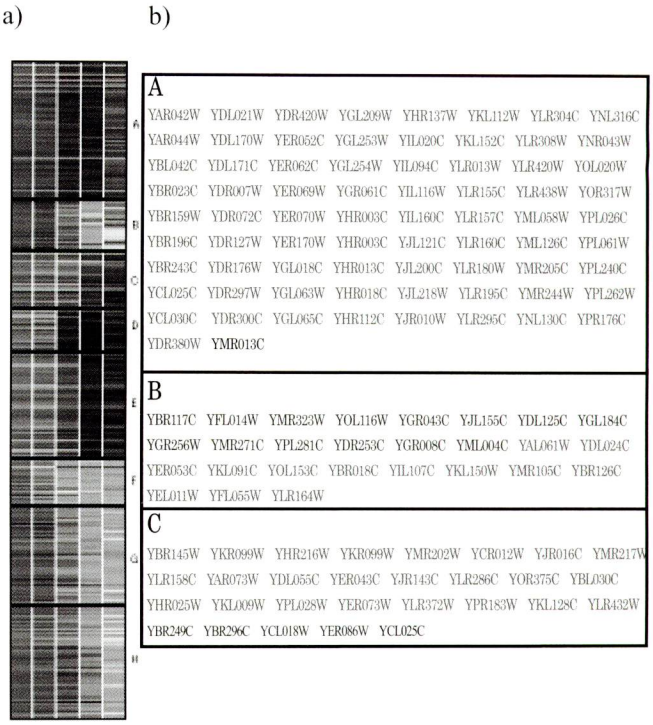


図3：a)非階層クラスタリング結果のうちパターンαの結果および、b)パターンβにおける分類（色分けは本文参照）

**参照資料**

[1]野村收作　他：「DNA マイクロアレイ発現データにおける既知遺伝子情報の導入－教師付 k 平均クラスタリング法を用いた解析」、室蘭工業大学 S.V.B.L 年報 Vol.3,pp30-32,2002

[2]Munich Information center for Protein sequences(MIPS) http://mips.gsf.de

**参考資料**

1)北野宏明：「システムバイオロジー」、秀潤社、2001

2)松原健一、榊佳之：「ゲノム機能　発現プロファイルとトランスクリプトーム」、中山書店、2000

3)林崎良英：「DNA マイクロアレイ実践マニュアル」、羊土社、2000