混合分布モデルを用いたマイクロアレイの遺伝子発現量推定

竹谷 勝† 松田 岳博‡ 津村 徳道‡ 岩本 政雄† 三宅 洋一‡
独立行政法人 農業生物資源研究所† 千葉大学自然科学研究科‡

1. はじめに

近年，多くの相補的遺伝子(cDNA)をスライドガラスに付着させたマイクロアレイが開発され，大規模な遺伝子発現解析が 1 回の実験で行えるようになった．しかし，マイクロアレイは生物的・物理的特性によるノイズの影響を受けやすく観測データが不安定であるという問題点も抱えている．これまで，ノイズ評価用のマイクロアレイからノイズの特性をモデル化して，ベイズ推定により遺伝子発現量を推定するという研究が報告されている[1]が，ノイズのモデル化のために多くのマイクロアレイ実験を行うことは多大な労力と時間を要する．

今回，我々はマイクロアレイの複製データを用いて，ノイズによる真値と観測値の差異を混合分布としてモデル化し，観測値から遺伝子発現量を確率分布として推定した．

2. マイクロアレイデータ

本研究で用いたマイクロアレイは同一遺伝子を左右ブロックに複製して配置している．図 1 に概念図を示す．生化学反応後のマイクロアレイを専用イメージスキャナーで読み取り，遺伝子のスポットごとに画素強度の総和の対数を観測値とした．

図 2 にマイクロアレイの左ブロックの観測値を x_1 軸，対応する右ブロックの観測値を x_2 軸としたマイクロアレイ上の全遺伝子データの散布図を示す．複製データに関わらず，ノイズにより二次元的に分布していることが分かる．

3. 観測データの混合分布モデル

左右ブロックから得られる観測値の二次元データセットを主成分分析し，第一主成分軸上に真値である遺伝子発現量が存在すると仮定した．

観測値はすべての真値からノイズにより生じる可能性がある．そこで，第一主成分軸上に任意の間隔で真値を仮定し，各真値のノイズによる発生分布を二次元正規分布とみなして，全遺伝子の観測データを複数の二次元正規分布で構成される混合分布によりモデル化した．図 3 に混合分布モデルの概念図を示す．

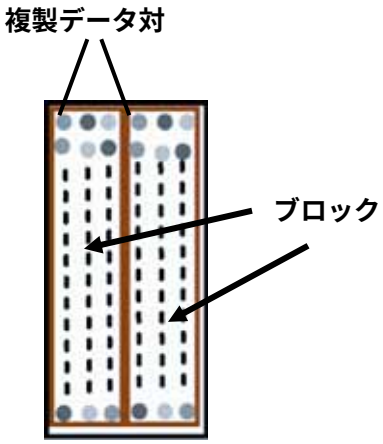

図 1 マイクロアレイデータの概念図

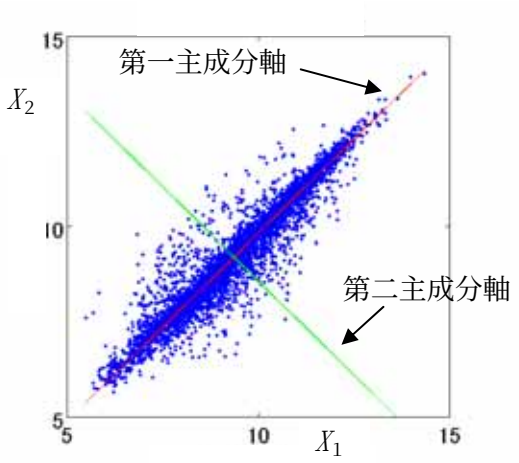

図 2 観測データの散布図

Estimation of gene expression on microarray by using mixture distribution modeling
† M. Takeya, M. Iwamoto, National Institute of Agrobiological Sciences
‡ T. Matsuda, N. Tsumura, Y. Miyake, Graduate School of Science and Technology, Chiba University

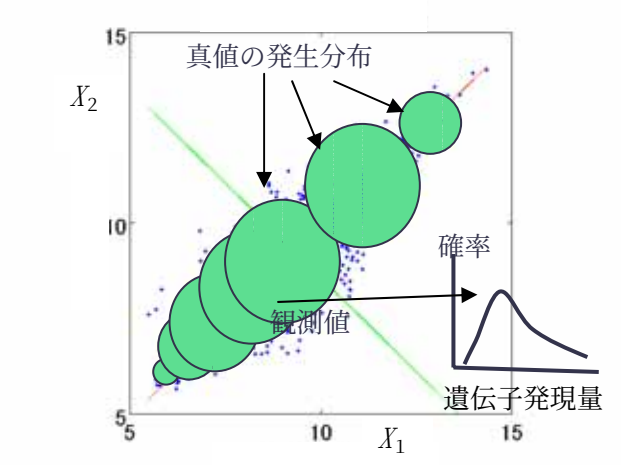

図 3　混合分布モデルの概念図

4.　EM アルゴリズムによる解法

遺伝子発現量の真値が g 個，マイクロアレイ上の遺伝子を n 個とする．ここで，各真値の発生確率を $\pi_1,\cdots,\pi_g$，各真値における観測値の確率密度関数を $f_i(w,\theta_i)$とする．観測値 w の確率密度関数は g 個の成分の混合として，以下のように表される．

$$f(w;\Psi)=\sum_{i=1}^{g}\pi_i f_i(w;\theta_i) \qquad (1)$$

ここで，$\Psi=(\pi_1,\cdots,\pi_{g-1},\ \theta_1,\cdots,\ \theta_g)$である．Ψに対する対数尤度関数は，

$$\begin{aligned}\log L(\Psi)&=\sum_{j=1}^{n}\log f(w_j;\Psi)\\&=\sum_{j=1}^{n}\log\{\sum_{i=1}^{g}\pi_i f_i(w_j;\theta_i)\}\end{aligned} \qquad (2)$$

となる．いま，未観測データとして以下を導入する．

$$z=(z_1^T,...,z_n^T)^T \qquad (3)$$

ここで，z_j は g 次元のベクトルであり，その要素 z_{ij} は，j 番目の観測値が i 番目の真値から発生したものなら 1，そうでなければ 0 を与える．z を用いてΨに対する完全データの対数尤度は以下の式で与えられる．

$$\begin{aligned}\log L_c(\Psi)&=\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}\log\pi_i\\&+\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}\log f_i(w_j;\theta_i)\end{aligned} \qquad (4)$$

次に，観測値 w_j の第一主成分を w_{1j}，第二主成分を w_{2j} とする．また，i 番目の真値の第一主成分を μ_{1i}，第二主成分を μ_{2i} とし．これら真値の位置は固定値である．i 番目の真値からのノイズによる発生分布において，第一・第二主成分の独立性を仮定すると，共分散は 0 であり，第一主成分方向の分散は σ_{1i}^2，第二主成分方向は σ_{2i}^2 とする．ここで，EM アルゴリズム[3]により式(4)を解いて各発生分布の分散を算出すると，

$$\sigma_{li}^{2(k+1)}=\frac{\sum_{j=1}^{n}z_{ij}^{(k)}(w_{lj}-\mu_{li})^2}{\sum_{j=1}^{n}z_{ij}^{(k)}} \qquad (5)$$

となる．l=1,2，k は繰り返しの回数を表し，閾値を満たすまで式(5)の計算が行われる．

5.　クラスタリング

イネの葉から採取した遺伝子 4475 個のマイクロアレイ 6 枚に対して提案法を適用し，遺伝子発現量を推定した．さらにマイクロアレイ間での発現量変化に基づき遺伝子機能を予測するため，発現量の確率分布を利用して遺伝子のクラスタリングを行った．中心点に集まったサンプルとそれ以外の遠方サンプルとの境界を探索する Adaptive quality-based clustering[3]の結果を利用して，クラスタリングの構成データではないが発現量の確率分布がクラスタリングに含まる遺伝子を割合に応じてリストアップした．その結果，今まではクラスタリングから欠落していたデータを候補として扱えるようになった．

6.　まとめ

マイクロアレイ解析を行う際に通常作成される複製データを有効に活用して，遺伝子発現量を確率分布として推定する手法を開発し，実際のマイクロアレイデータに適用して有効性を確認した．今後は，発現量の確率分布をより効率的に利用できるクラスタリング法を開発する．

参考文献

[1] Ron O. Dror et al., Bayesian Estimation of Transcript Levels Using a General Model of Array Measurement Noise, Journal of Computational Biology, 10, pp.433-452, 2003
[2] Geoffrey J. McLachlan and Thriyambakam Krishnan, The EM Algorithm and Extensions, John Wiley & Sons, Inc, 1997
[3] Frank D. Smet et al., Adaptive quality-based clustering of gene expression profiles, Bioinformatics, 18, pp.735-746, 2002