

遺伝子発現データの高次元因子分析

東村 果穂

マイクロアレイや次世代シーケンサーの登場により、遺伝分野では、遺伝子の発現量や塩基配列の大規模なデータが入手できるようになった。2003 年にはヒトゲノムの全塩基配列が決定された。このような状況の中、遺伝子がどのような組み合わせで発現し、機能を持つかを同定することが大きな課題となっている。そこで、本研究では遺伝子の機能を因子とみなし、因子分析を行うことでその機能を推測することを目的とする。

遺伝子発現データは、個体数に比べて変数が非常に多いという特徴を持つため、統計ソフトで利用される最尤法や主因子法では計算出来ない。しかし、この問題は最尤法の 1 つである EM アルゴリズムを使うことで回避できる。そこで、本研究では、出芽酵母とヒトのマイクロアレイデータに EM アルゴリズムによる因子分析を適応した。

また、統計の分野では、より解釈のしやすい因子負荷量を得るために、様々な回転手法が開発されている。スパース因子分析は、分析の段階で解釈のしやすい因子負荷量を得る手法である。本研究では、遺伝子の機能は互いに影響を与えているという考えのもと、斜交 **geomin** 回転、**Simplimax** 回転、**Promax** 回転の 3 種類の斜交回転法とスパース因子分析によって得られた結果を比較した。

その結果、斜交 **geomin** 回転が最も簡便に解釈のしやすい因子負荷量を得る方法であることが分かった。因子負荷量の 0 成分の量を指定し、疎密状態を調整したい時には、スパース因子分析が最適な方法であった。**Simplimax** 回転は、0 成分の個数を決めるチューニングパラメーターの値に結果が左右されることが分かった。**Promax** 回転では、芳しい結果が得られなかった。

それぞれのデータに注目すると、出芽酵母のデータでは、G1 初期に「自己成長因子」、G1 後期に「修復因子」、S 期に「修復因子」と「出芽細胞成長因子」、G2 期に「出芽細胞成長因子」、M 期に「分裂因子」の負荷量が大きかった。ヒトのデータでは、正常な細胞標本とがん細胞標本では、「シグナル伝達因子」の負荷量の様相が全く異なっていた。「増殖因子」に関しては、がん細胞標本で大きな負荷量がみられた。「生存因子」は正常な細胞標本の方が、因子負荷量の正と負のコントラストが強かった。

そして、遺伝子の機能について得られた知見をまとめると、

- DNA 損傷の修復と細胞分裂が同時に起こらない（出芽酵母）
- 親細胞と出芽した細胞は交互に成長する（出芽酵母）
- がん細胞では、細胞の自殺や老化のような死ぬための機構が働かない（ヒト）
- がん細胞は、因子間相関が強いことから、自らの細胞の増殖が有利になるように遺伝子の発現を変化させている（ヒト）

ことが示唆された。

以上より、遺伝子の機能の推測に因子分析は有用な方法であることが示された。（行動統計科学）