統計的変数選択法の嚢胞性線維症マイクロアレイデータへの応用

統計的変数選択法の嚢胞性線維症マイクロアレイデータへの応用

統計的変数選択法の嚢胞性線維症マイクロアレイデータへの応用

藤井祐輔[1)]，岡田吉史[1)]，長島知正[1)]
1）室蘭工業大学情報工学科

1. はじめに

近年、バイオインフォマティクス技術の医療の分野への応用が進み、遺伝子医療への可能性が開かれつつある。そこで、遺伝子発現データに基づいた病理診断が重要になる．現在、遺伝子診断に有用な遺伝子を抽出するために、DNAマイクロアレイから得られた情報を基に解析を行う遺伝子発現差解析の研究が進んできている。最近、我々は単純な統計値であるF値に基づいて、異なるクラス(患者と健常者など)の識別に寄与する遺伝子群を組み合わせ的に探索するForward variable (gene) Selection Method(FSM)を開発した。

本研究では、DNAマイクロアレイから得られた嚢胞性線維症のデータにFSMを適用し、嚢胞性線維症(Cystic Fibrosis, CF)患者と健常者の識別に有効な遺伝子群を抽出する。

2. 方法

2.1. マイクロアレイデータセット

本研究で使用するデータセットは、米国デンバー州のある病院の子供の患者9人と健常者9人のサンプルにより構成されている。サンプル一人分につき、12,625個の遺伝子発現量のデータがある。以下にそれぞれのサンプルの遺伝子発現量を数値化したデータセット例を示す(図1)。このデータセットにおいて、行は遺伝子であり、列はサンプルを表している。

Sample	CF1		CF9	N1		N9
遺伝子1	4	…	5	1	…	1
遺伝子2	8	…	7	2	…	4
…						
遺伝子*N*	6	…	9	4	…	1

図1: データセット例

2.2. FSMによる遺伝子選択

本研究で実行した計算のアルゴリズムを以下に示す。
(1)全遺伝子について個々にF値を求める。F値は式(1)により、p=0、r=1としてそれぞれ計算し、最大値の遺伝子を第1位の遺伝子とする
(2)第1位から第(k-1)位までの遺伝子にk番目の遺伝子を加え、k個の遺伝子の組に対するF値を算出する
(3)(2)を残り全ての遺伝子について繰り返し、一番F値の大きい遺伝子を第k位の遺伝子とする
(4)(2)と(3)を、全ての遺伝子順位が決定するまで繰り返す

FSMのイメージ図(図2)、F値の計算式を以下に示す。

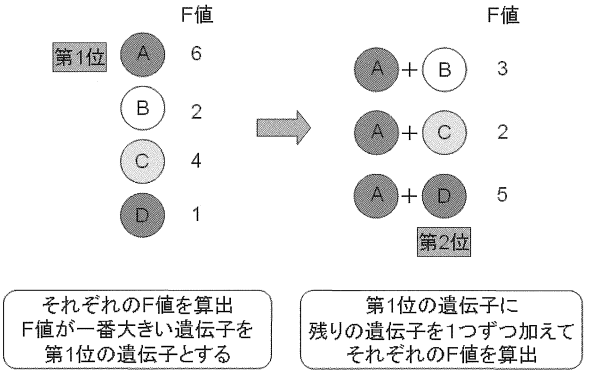

図2: FSMのイメージ図

$$F = \frac{(n^{[1]}+n^{[2]}-p-r-1)n^{[1]}n^{[2]}(D^2_{(p+r)}-D^2_{(p)})}{r\{(n^{[1]}+n^{[2]}-2)(n^{[1]}+n^{[2]})+n^{[1]}n^{[2]}D^2_{(p)}\}} \quad (1)$$

ここで、$n^{[c]}$:クラスcのサンプル数、p:変数を追加する前の変数の個数、r:追加する変数の個数、D^2:判別効率（２つの母平均のマハラノビス距離の二乗）である。本研究では、r=1とした。D^2の計算式は以下になる。

$$D^2 = (\vec{u}^{[1]} - \vec{u}^{[2]})' \Sigma^{-1} (\vec{u}^{[1]} - \vec{u}^{[2]})$$

ここで、$\vec{u}$は各群の平均ベクトル、Σは分散共分散行列である。

3. 判別の評価方法

3.1. 評価方法

クラス推定にはマハラノビス距離を用いた判別分析を用い、LOOCV(Leave-one-out cross validation)により誤判別率を算出した。LOOCVとは、標本群から１つの事例だけを抜き出してテスト事例とし、残りを訓練事例とする手法である。本研究では、テストサンプル(各遺伝子の

発現量からなるベクトル)は、各群の平均ベクトルにマハラノビス距離がより近い群に判別される。判別の精度は、上記の操作を全てのテストサンプルに対して行った後、判別を誤ったサンプル数を総サンプル数で割った誤判別率により評価される。

3.2. 結果

本研究では、FSMで抽出される上位n個(n=1~16個)の遺伝子群に対してLOOCVを行った。図2に、前述の方法により抽出した遺伝子によるクラス推定の結果を示す。横軸はテストサンプルのクラス推定に使用する累積遺伝子数であり、縦軸は誤判別率を示す。誤判別率は推定に使用する遺伝子数が上位8個までの間に低下していく。そして、遺伝子数が8個から14個では誤判別率は0であり、良い結果を示している。その後15,16個では誤判別率はやや上昇する。これは、判別に全く寄与しないノイズを含んだ遺伝子が増加するためであると考えられる。

ただし、この遺伝子の抽出においては逆行列とマハラノビス距離に条件を付けている。遺伝子選択の際に計算される逆行列の対角成分に0.01以下の要素が存在する遺伝子、遺伝子追加後のマハラノビス距離が遺伝子追加前のマハラノビス距離の5倍以上に発散してしまった遺伝子は遺伝子選択から除外している。

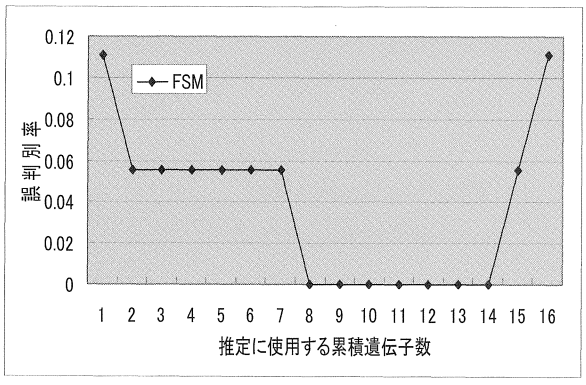

図2: 累積遺伝子数と誤判別率の関係

4. 考察

図2の結果は、抽出された遺伝子の誤判別率が最大でも約0.12であり良い結果が得られている。一方、データセットのサンプルの遺伝子数は10000個以上あるのに対して、本研究で抽出できた遺伝子数はわずか16個であった。この理由の1つとしては、サンプル数が全18検体しかなかったことが挙げられる。

F値の式の分子$n^{[1]}+n^{[2]}-p-r-1$の部分において$n^{[1]}+n^{[2]}$は2つのクラスのサンプル数の和であるので、本研究では18である。また追加する変数の個数rは1であるから、上記の式は16-pとなる。ここでFSMにおけるランキング17位の遺伝子の抽出を考えると、そのときのpの値は追加する前の遺伝子の個数なので16になる。このとき上記の式は0になるからF値も0になり、F値によるランキングができなくなるということが起こる。以上によりFSMで抽出できる遺伝子数はサンプル数によって決定されているということがわかる。

また、本研究で抽出された遺伝子のアノテーションについて調べたが直接CFに関連があるCFTR遺伝子は抽出されていなかった。しかし、CFTR遺伝子が間接的にある遺伝子の機能異常を引き起こしている可能性がある。本研究で抽出された遺伝子が、CFTR遺伝子によって悪影響を与えられている遺伝子である可能性は十分に考えられる。

5. まとめ

本研究では、嚢胞性線維症患者と健常者の識別に有効な遺伝子群を抽出するために、FSMを用いて遺伝子発現差解析を行った。その結果、2つのクラスを完全に識別可能な遺伝子群を抽出できた。しかし、直接CFに関連がある遺伝子は抽出されていなかった。実験で抽出された遺伝子が間接的にCFに関連があるかどうかの生物学的意味の解明は、今後検討すべき課題であると言える。

さらに、抽出できる遺伝子数を増やすためにFSMのアルゴリズムの改良を行う必要がある。また、抽出遺伝子の信頼性の確保のために、本研究で扱った疾患と同じ疾患の別のサンプルで発現差解析を行う必要があると考えられる。

参考文献

[1] 永田靖, 棟近雅彦：多変量解析入門, サイエンス社 pp.99-118, 2007.

[2] H.Mitsubayashi et al.:Accurate and Robust Gene Selection for Disease Classification Using a Simple Statisics,Bioinformation 3(1),68-71(2008)

[3] 小野修司:Cluster Validity Indexを利用した遺伝子発現差解析手法の評価.Master's thesis,室蘭工業大学大学院研究科,1 2006.

[4] 実験医学　バイオキーワード集 http://www.yodosha.co.jp/jikkenigaku/keyword/index.html

[5] 石村貞夫:すぐわかる多変量解析,東京図書株式会社, 1992.