



遺伝子発現差解析Webツールの開発

メタデータ	<p>言語: jpn</p> <p>出版者: 室蘭工業大学SVBL</p> <p>公開日: 2010-07-20</p> <p>キーワード (Ja):</p> <p>キーワード (En):</p> <p>作成者: 間澤, 直寛, 岡田, 吉史, 長島, 知正</p> <p>メールアドレス:</p> <p>所属:</p>
URL	http://hdl.handle.net/10258/514

遺伝子発現差解析Webツールの開発

著者	間澤 直寛, 岡田 吉史, 長島 知正
雑誌名	サテライト・ベンチャー・ビジネス・ラボラトリー 年報
巻	8
ページ	99-100
発行年	2009-03
URL	http://hdl.handle.net/10258/514

遺伝子発現差解析Webツールの開発

間澤直寛¹⁾, 岡田吉史¹⁾, 長島知正¹⁾

1) 室蘭工業大学情報工学科

1. 序論

近年、遺伝子発現データに基づいて病理診断を行うことを狙いとする研究が進められている。特に遺伝子選択法の開発は、病気識別に寄与する遺伝子群を抽出する上で重要なテーマとなっている。最近、我々はFSM(Forward variable (gene) selection method)と呼ばれる高精度な遺伝子選択法を開発した。FSMは従来の遺伝子選択法に比較して、より高精度に未知のサンプル(個々の患者)を適切なクラス(患者と健常者)に分類することが可能である。しかしながら、現時点では、本法はコマンドライン上での実行に限られており、一般への公開はなされていない。そこで、本研究では、医学・生物学、バイオインフォマティクス分野の研究者の利用を想定しFSMを用いた遺伝子選択に加え、抽出された遺伝子群を用いたクラス判別機能を実装したWebツールの開発を行う。

2. 手法

2.1. FSM

FSMはマハラノビス距離を用いた変数増加法に基づく遺伝選択法である。遺伝子間の関連情報を考慮することで、病理診断に有用な遺伝子をより高精度に抽出することができる。

以下に本研究で使用した計算のアルゴリズムを示す。

- (1) 全遺伝子について個々のF値を求める。F値は後述の式より $p=0$, $r=1$ としてそれぞれ計算し、最大値をとる遺伝子を第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 個の組に対するF値を算出する。
- (3) (2)を残しすべての遺伝子について繰り返し、最大のF値をとる値を第 k 位の遺伝子とする。
- (4) (2)と(3)を、すべての遺伝子順位が決定するまで繰り返す。

F値は以下の式で算出する。

$$F = \frac{(n^{[1]}+n^{[2]}-p-r-1)n^{[1]}n^{[2]}(D_{(p+r)}^2-D_p^2)}{r\{(n^{[1]}+n^{[2]}-2)(n^{[1]}+n^{[2]}+n^{[1]}n^{[2]}D_p^2)\}}$$

ここで、 n : 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。 D^2 の計算式は以下になる。

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

ここで、 \bar{u} は各群の平均ベクトル、 Σ は分散行列である。 D^2 は、あるベクトル因子の集団間の平均の差が大き

く、かつ、全体の分散が小さいと大きな値をとる。

2.2. 実装

遺伝子選択はFSMで行い、クラス判別にはマハラノビス距離を用いた非線形判別を行っている。また、クラス判別の際には遺伝子が1つ抽出されるごとにクラス判別を行い、より多く判別されたクラスを判別結果として扱い精度を高めている。ただし、多数決をするには奇数である必要があるため判別回数は奇数にする。また、抽出された遺伝子が8個以下の場合はすべての遺伝子で多数決による判別を行う。いくつかのデータでFSMが抽出した遺伝子による誤判別率を算出したところ、4つほど遺伝子が抽出されたあたりから良い判別率となり、最後の部分で再び誤判別率が増加する傾向が見られたため、抽出された遺伝子数が奇数の場合は上位3位と下位3位、偶数の場合は上位4位と下位3位を除いた部分で多数決をとることにした(図2)。

CGIプログラムはWebサーバソフトウェアApacheでWebサーバを立ち上げ、Perl言語を用いてブラウザ上に入力されたデータファイルを特定の名称と拡張子に変換し、FSMを用いて遺伝子選択及び、クラス判別を行う。抽出された遺伝子、判別されたクラス及び、エラーはそれぞれテキストに出力され、それが読み込まれてブラウザ上に表示される。エラーは出力画面下側に表示される。

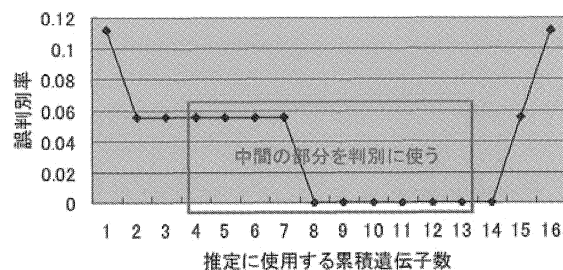


図1: FSMの誤判別率

2.3. インターフェース

図2は、本ツールの実行画面である。ブラウザ上のDataSetに遺伝子抽出を行いたいデータファイルを指定し、実行ボタンを押すことでFSMを用いて二群からの遺伝子抽出を行う。UnknownのデータファイルのサンプルがDataSetで抽出された遺伝子に基づいてどちらのクラスに分類されるかを判別する。未知サンプルの識別の必要がなければUnknownの入力を省略することができる。

DataSetのフォーマットはデータの列をサンプル、行を遺伝子とする。また、クラスラベルを一行目に指定(1群、2群をそれぞれ“1”、“2”で表す)する。



図2：作成した Web ツールとその使用例

Unknown のフォーマットはデータの列をサンプル、行を遺伝子とする。

3. 結果

図2は、本ツールに Golub の白血病のデータを入力する場合の例を示している。2.3.に従ってデータを指定し、実行すると、図2のように抽出された遺伝子が表示される。

未知のサンプルデータを識別するためには図2に示されるように、そのサンプルの発現値を記述したファイルを指定する。この際に複数のサンプルを指定することも可能である(図2では3つのサンプルを指定)。判別結果は、ウィンドウの右側のパネルに表示される(図3)。例えば、図2で指定した一列目のサンプル(sample1)は、クラス1(class1)に分類されていることを示している

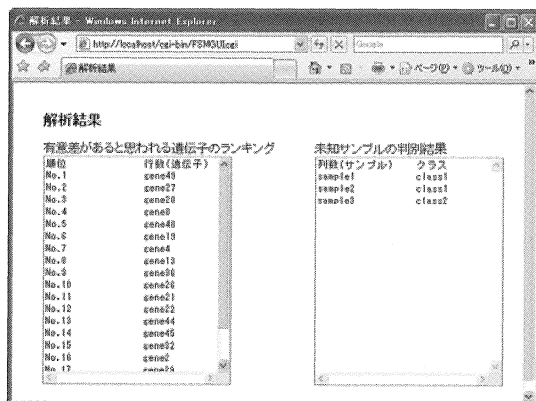


図3：実行結果

4. まとめ

本研究で、コマンドライン上での実行に限られていた FSM による遺伝子選択機能とクラス判別機能が実装された Web ツールを開発した。本ツールを使用することで容易に判別に寄与する遺伝子を抽出し、クラス判別を行うことができ、他の研究者や医療機関等で役立つことが期待される。

しかし、FSM にはいくつかの問題点がある。まず、抽出できる遺伝子数がサンプル数に依存することである。F 値の性質から(サンプル数-2)で F 値が 0 になるためそれ以降は計算できなくなる。次に、マハラノビス距離は行列を乗算していく式であるため、単純に行列が大きくなるにつれて値が大きくなる。このため、マハラノビス距離がオーバーフローしてしまうことがある。また、逆行列の発散によりマハラノビス距離がオーバーフローすることもある。

実際に使用するデータのほとんどは遺伝子数が数千〜数万にも及ぶ。そこで問題となるのは、メモリの確保と計算時間である。現在使用している PC の性能ではサンプル数 20、遺伝子数 1 万 3 千程度までのメモリ確保ができる。また、遺伝子数 50 (2.1.における k) の場合では 0.5 秒以内に計算できるが、遺伝子数 1 万の場合では数分から 10 分程度かかる (Intel (R) core (TM) 2 CPU 6600 2.4GHz, 2GB RAM)。今後は、PC クラスタを用いた並列処理によりこれらの問題の解決を検討している。また、本ツールの Web サイトを英語化し、遺伝子数から計算予想時間を表示するプログラムの追加も行っていく。

参考文献

- [1] 永田靖, 棟近雅彦: 多変量解析法入門, サイエンス社, pp.99-118, 2007.
- [2] 石村貞夫: すぐわかる多変量解析, 東京図書 pp.116-163, 1997.
- [3] H.Mitsubayashi et al.: Accurate and Robust Gene Selection for Disease Classification Using a Simple Statistics, Bioinformation 3(1), 68-71 (2008)
- [4] 実験医学 バイオキーワード集
<http://www.yodosha.co.jp/jikkenigaku/keyword/index.html>
- [5] T.R. Golub et al.: Science, (1999) 286:531 [PMID:10521349]