G-010

遺伝子発現データを用いた病理診断における遺伝子選択手法の提案
Selection of Gene Relevant to Particular from DNA-microarray

吉田 育未† Goutam Chakraborty† 馬淵 浩司† 松原 雅文† 山下 和彦‡
Ikumi Yoshida Goutam Chakraborty Hiroshi Mabuch Masafumi Matsuhara Kazuhiko Yamashita

1. はじめに

DNA マイクロアレイ技術の向上により，多数の遺伝子と，多数の mRNA を同時にかつ高速に測定することが可能となった [1]．このような観点から，DNA マイクロアレイは，複数の遺伝子が複雑に関与している癌の診断ツールのひとつとして期待されており，実際の医療現場への応用研究が行われている．これまでに，DNA マイクロアレイの網羅的な一次スクリーニングから得られた大量な遺伝子発現情報に対して，様々な統計的手法による疾病と関係性の高い遺伝子座の解析が行われてきた．本研究では，疾病と関係性の高い遺伝子座選択の新たな統計的手法を提案し，既存手法との評価実験を行う．そして，得られた結果より本手法の有効性を検証する．

2. 既存手法

DNA マイクロアレイで得られた遺伝子発現データの中には，健常者と癌患者が中間的な発現値を持つ遺伝子座が存在する．統計的手法おいて，一方のクラスタの発現値が一様に高く，もう一方のクラスタは低いような差異が現れている遺伝子座の選択し，診断に有効な遺伝子座とそうではない遺伝子座のふるい分けを行うことを目的とする．その統計的方法代表して，近傍解析と Mann Whitney 検定があげられる．以下にその統計的手法の説明を行う．

2.1 近傍解析

癌患者をクラス 1，健常者をクラス 2 とし，2 クラスにおける遺伝子 g の発現量の平均 $\mu_1(g)$ と $\mu_2(g)$ と標準偏差 $\sigma_1(g)$ と$\sigma_2(g)$ を求める．平均は，データ全体の性質を示す代表値であり，2 クラスの大小関係に関する位置の尺度である．一方のクラスの発現値が高く，もう一方のクラスの発現値が小さければ，データの相対的な位置関係に差がある．しかし，平均に大きな差があったとしても，データの広がりの程度によって，2 クラスの発現値が重なる可能性がある．よって，2 クラスの平均が大きく分かれていて，かつ，標準偏差が小さい遺伝子座の選択を行う．以下の式により，

$$P(g) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)} \qquad (1)$$

遺伝子座 (P) の 2 クラス間の遺伝子発現の差異を評価する．$P(g)$ の値が大きいほど，診断に有効な遺伝子座である [3]．

2.2 Mann Whitney 検定

Mann Whitney 検定はノンパラメトリック法による独立の 2 群の差の検定である [2]．遺伝子座 g の癌患者と健常者の発現値を一括して昇順にソートし，上から順に発現値に順位をつける．癌患者の発現値が高く，健常者の発現値が小さい場合，順位和は，癌患者のほうが大きく健常者が小さくなる．逆に癌患者の発現値が小さく，健常者の発現値が大きい場合，順位和は，癌患者のほう小さく健常者が大きくなる．このように，2 群の点の配置に差がある遺伝子座を選択する．

3. 提案手法

統計的手法の多くは，扱うデータの制約があり，使用するデータによっては，適応できない場合がある．本手法では，扱うデータの形式にとらわれない統計的手法を提案する．

3.1 クラスタの中心角度による遺伝子座の選択

癌患者の発現値を X，健常者の値を Y とし，癌患者と健常者の発現値のデータ群の中心位置を K-Means 法により求める．癌と関係性の高い遺伝子座は，癌患者の発現値は高く，健常者の発現値は低い，また，癌患者の発現値は低く，健常者の発現値は高い場合も考えられる．よって，データ群のクラスタの中心は，グラフ上で，右下か左上の位置ある．癌患者と健常者の発現値が同じであれば，クラスタの中心位置は，45 度線付近にくるので，クラスタの中心位置と 45 度線の角度が大きく離れている遺伝子座の選択を行う図 1.

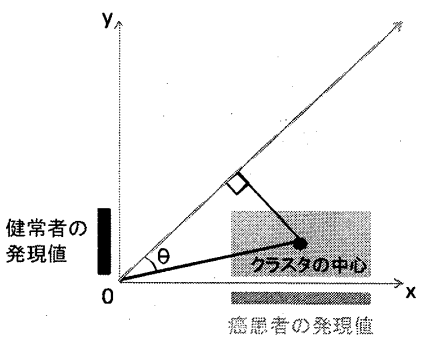

図 1: 角度による遺伝子座の選択

4. 評価実験

既存手法と提案手法を用いて，癌と関係性の高い遺伝子座の選択を行い，4，026 個の遺伝子座の中から，評価の優れた 500 個の遺伝子座を選択し，共通して選ばれた遺伝子座の個数，または，共通に選択されなかった遺伝子座選の違いを比較し，本手法の有効性を示す．実験には，癌患者 67 名と健常者 29 名の 4，026 個の遺伝子発現データを用いる．遺伝子発現データは，実数で与えら

†岩手県立大学ソフトウェア情報学部
‡岩手県立大学大学院ソフトウェア情報学研究科

れている．ex は遺伝子発現値であり，ex^1 の上付きの添

図 2: 実験データ

	癌間患者 67 名	健常者 29 名
$gene1$	$ex_1^1 \cdots ex_{67}^1$	$ex_{68}^1 \cdots ex_{96}^1$
$gene2$	$ex_1^2 \cdots ex_{67}^2$	$ex_{68}^2 \cdots ex_{96}^2$
$\vdots$	$\vdots$	$\vdots$
$genei$	$ex_1^i \cdots ex_{67}^i$	$ex_{68}^i \cdots ex_{96}^i$
$\vdots$	$\vdots$	$\vdots$
$gene4026$	$ex_1^{4026} \cdots ex_{67}^{4026}$	$ex_{68}^{4026} \cdots ex_{96}^{4026}$

え字は遺伝子座の番号を示し，ex_1 の下付きの添え字は癌患者と健常者の番号を表している．

4.1 結果

3 つの手法により，選択した遺伝子座と 3 つの手法で選択された遺伝子座の割合を以下の，図 3，図 4，図 5，図 6 に示す．

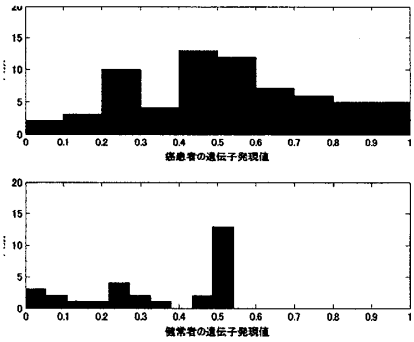

図 3: 提案手法

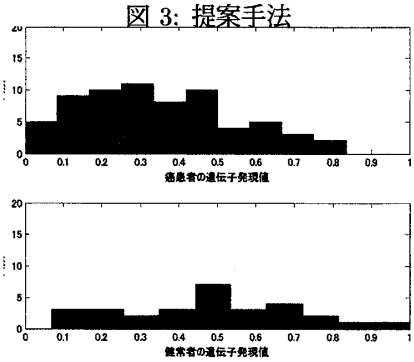

図 4: 近傍解析

図 6 から 3 つの統計的手法で選択した 500 個の遺伝子座の内，289 個の遺伝子座を共通して選択されていることが読み取れる．3 つ手法による遺伝子座選択おいてある程度同等の精度が得られることが分かる．

4.2 考察

2 クラスの平均差と標準偏差基づいた近傍解析において，遺伝子発現データは，正規分布であることが望ましい．図 4 は，ベルカーブを描いていない．平均は，左右対称な分布の中心の位置を示す尺度であり，標準偏差は，平均からの左右の散らばりの程度を示す尺度である．偏りのある遺伝子発現データでは，評価値に誤差が見ら

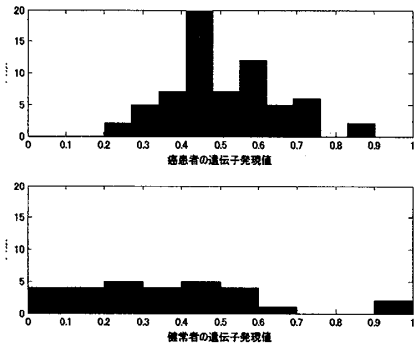

図 5: Mann Whiteny 検定

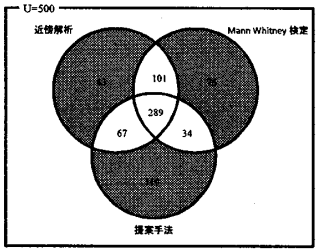

図 6: 3 つの手法による遺伝子座の選択個数の割合

れる．Mann Whitney 検定は，遺伝子発現データのに基づいた独立の 2 群の差異を検定する方法であり，直接的に遺伝子発現データを評価していない．また，Mann Whitney 検定は，2 標本間の散布度が等質であることを前提とするが，図 5 より，2 標本間の散布度が等質性があるとは言い難い．提案手法では，データの形式の影響を受けないため，既存手法より有効性がある．

5. おわり

平均と標準偏差，Mann Whitney 検定，提案手法の 3 つの統計的手法により遺伝子座の選択を行った．遺伝子座の選択を行い，結果から本手法の有効性を示した．今回，使用した癌のデータは 3 種類の悪性リンパ腫であり，癌同士の発現値が重なり合っている部分が多かった．今後の予定としては，異なる部位の癌から得られた遺伝子発現データを用いて，癌と関係の高い遺伝子座の選択を行い既存手法と提案手法の評価実験を行う．そして，提案手法によりデータの形式にとらわれない遺伝子座選択の可能性を実証する．

参考文献

[1] 五條堀 孝，生命情報学，シュプリンガー・フェアラーク東京株式会社，2003 年．

[2] 市原 清市，バイオサイエンスの統計学・正しく活用するための実践理論，株式会社南江堂，19990 年，2001 年．

[3] T. R. Goulb, et al., "Molecluar classification of cancer: class discovery and class prediction by gene expression monitoring", Sicence, Vol. 286, pp. 531-537, 1999.