特徴的ルール生成を用いた遺伝子発現解析

大河内 一弥* 牧 秀行* 森田 豊久* 富田 裕之**

(株)日立製作所 システム開発研究所* (株)日立製作所 ライフサイエンス推進事業部**

1. はじめに

近年、生命科学の分野では、遺伝子などに関する情報を大規模に取得し、そのデータの特徴を網羅的に解析しようとする研究が盛んに行われている。こうした研究の一分野として DNA チップ [1) [1]] を用いた遺伝子の発現解析があり、得られたデータに相関分析、主成分分析、クラスタリング(k-平均法、自己組織化マップ)など、様々な解析手法を適用した例が報告されている(具体的な研究については文献[2]などを参照)。

我々は、こうした試みの一つとして特徴的ルール生成 [3] を用いた遺伝子発現解析の実験を行った。特徴的ルール生成とはデータの特徴を IF～THEN 形式のルールとして抽出する分析手法である。本稿ではまず、遺伝子発現解析と特徴的ルール生成について説明する。次に、白血病患者の遺伝子発現データを用いて行った分析実験について述べる。その後、実験を通して判明した、DNA チップデータを分析する際の問題点について考察する。

2. DNA チップを用いた遺伝子発現解析

2.1 遺伝子発現解析

よく知られているように、遺伝子とは蛋白質の設計図であり、DNA 上では 4 種類の塩基(A、C、G、T)の配列として保持されている。これらの遺伝子は、その塩基配列がメッセンジャーRNA(mRNA)に転写され、mRNA が細胞内器官であるリボゾームで蛋白質に翻訳されることによって機能する。このように遺伝子が機能することを遺伝子の発現と呼ぶ。しかし、DNA 上の全ての遺伝子が常に発現するわけではなく、細胞の種類や状況によって発現する遺伝子は異なる。例えば、身体の部位によって発現する遺伝子は異なり、それによって、皮膚、骨、内臓など細胞の形態や機能に違いが生じる。

このような遺伝子の発現について、その規則性、しくみを調べるのが「遺伝子発現解析」である。

遺伝子発現解析では、まず、細胞中でどの遺伝子が発現しているかを観測する必要がある。遺伝子が発現する際には遺伝子の塩基配列を転写した mRNA が生成されるため、細胞中の mRNA の量を計測することで、どの遺伝子が、どの程度の強さで発現しているかがわかる。

2.2 DNA チップ

DNA チップは、スライドガラスなどの基盤上に、多数の DNA 断片や合成オリゴヌクレオチドのプローブを格子状に整列して貼り付けたものである。なお、DNA チップではプローブが配置された点のことを「スポット」と呼ぶ。

各プローブは、特定の mRNA と結合する性質を持ち、どのプローブがどのような塩基配列と結合するかは、DNA チップ製造時、基盤上にプローブを配置する段階でわかっている。計測対象となる細胞から抽出した、未知の mRNA を蛍光標識し、DNA チップと反応(これをハイブリッド化、ハイブリダイゼーションという)させた後に洗い流すと、プローブと結合した mRNA が DNA チップ上に残り、そのプローブ部位、すなわちスポットが蛍光を発する。この蛍光の強度を計測することにより、どの塩基配列を持つ mRNA が細胞中に存在していたかがわかる。すなわち、どの遺伝子が発現していたかがわかる。

これらの計測結果は、各スポットの蛍光強度を表す実数値として取得される。またこれらのデータは患者の疾患の有無など、サンプルの状態を表すフラグと共に提供される場合もある。DNA チップから取得されるデータの例を図 1 に示す。

Gene Expression Analysis Using Characteristic Rule Induction

* Kazuya OHKOUCHI, Hideyuki MAKI, Toyohisa MORITA, Hitachi, Ltd., Systems Development Laboratory.

** Hiroyuki TOMITA, Hitachi, Ltd., Life Science Group.

1) DNA チップにはいくつかの種類があり、厳密には「DNA チップ」や「DNA マイクロアレイ」などを区別して呼ぶ場合もあるが、本稿では特に区別せず、DNA チップと呼ぶ。

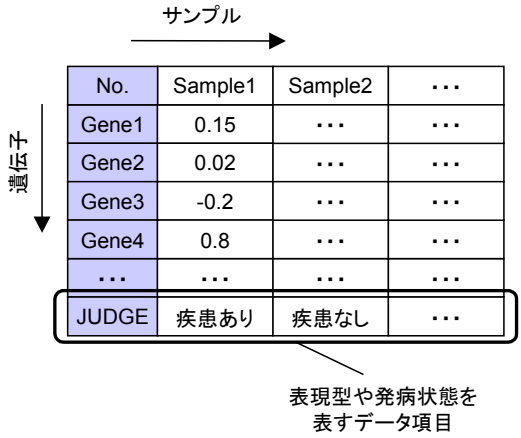

図 1 DNA チップから取得されるデータ

3. 特徴的ルール生成 [3]

本研究では上記の DNA チップから得られたデータを対象に、特徴的ルール生成の手法を用いて、その特徴を IF～THEN 形式のルール(図 2)として抽出した。

特徴的ルール生成の概念を図 3 に示す。特徴的ルール生成では、分析対象となるデータの数値項目をカテゴリに分割することで、数値解析の問題を部分空間の探索問題に変換する。結論部はユーザが指定し、これをうまく説明する条件部を探索する。

特徴的ルール生成では、ルールの評価値は以下のように定義される。「IF A THEN B」というルールにおいて、ルールの一般性(カバー率)、精度(ヒット率)を以下のように定義する。

カバー率： $P(A)$ ・・・・・・・・・・・・・・・・・・・・・・ (1)

ヒット率： $P(B \mid A)$ ・・・・・・・・・・・・・・・・・・・・ (2)

ここで $P(A)$、$P(B)$ はそれぞれ A、B の生起する確率である。ルールの評価値は以下の式で定義される。

$$\mu = P(A)^{\beta}\ P(B \mid A) \log\left\{\frac{P(B \mid A)}{P(B)}\right\} \cdots\cdots (3)$$

ここで、β は、カバー率とヒット率のどちらを重視してルールを評価するかを指定する定数である。

4. 白血病患者データの分析

4.1 データと分析方法

特徴的ルール生成の手法を実際の遺伝子発現データに適用する実験を行った。

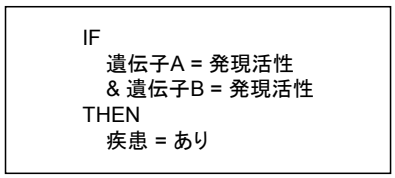

図 2 IF-THEN ルールの例

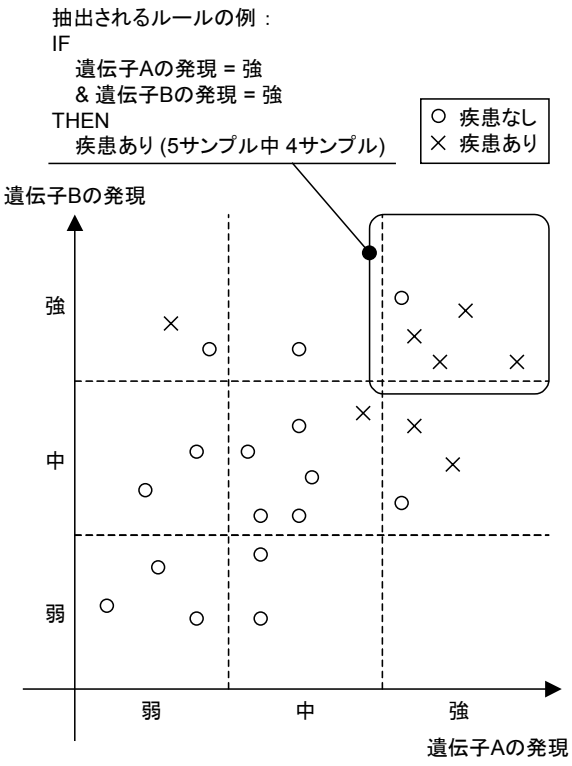

図 3 特徴的ルール生成の概念図

今回分析対象のデータとして用いたのは、米国Whitehead Instituteが公開している白血病患者の遺伝子発現データである [4]。このデータは、DNA チップデータ解析に関する会議である Critical Assessment of Microarray Data Analysis 2000 (CAMDA '00) [2] において、共通のデータセットとして使用されたものである。

このデータに含まれる患者の白血病のタイプには急性リンパ性白血病(ALL) と急性骨髄性白血病(AML) の2 種類がある。これらのタイプは、すべての患者についてデータ中に記述されている。遺伝子の発現データとしては、各患者の血液中の遺伝子の発現量をDNAチップを用いて計測した、7129 項目の測定値(実数値) が得られている。また、このデータに含まれる患者数は72 名である。つまり、分析対象は、72 サンプル×

表 1　生成されたルール一覧 (評価値が上位のもの)

結論部: 白血病のタイプ = ALL型

#	評価値	ヒット率	カバー率	ヒット数	カバー数	条件部	Unigene 2)
1	0.148	100%	34.7%	25	25	U07139_at = 大	CACNB3
1	0.148	100%	34.7%	25	25	M94633_at = 大	RAG2
3	0.142	100%	33.3%	24	24	U29175_at = 大	SMARCA4
3	0.142	100%	33.3%	24	24	M28170_at = 大	CD19
3	0.142	100%	33.3%	24	24	U27460_at = 大	UGP2
3	0.142	100%	33.3%	24	24	M31523_at = 大	TCF3
3	0.142	100%	33.3%	24	24	X97267_rna1_s_at = 大	PTPRCAP
3	0.142	100%	33.3%	24	24	X85116_rna1_s_at = 大	EPB72
3	0.142	100%	33.3%	24	24	M84371_rna1_s_at = 大	CD19
3	0.142	100%	33.3%	24	24	M12959_s_at = 大	TRA@
3	0.142	100%	33.3%	24	24	L09209_s_at = 小	APLP2
3	0.142	100%	33.3%	24	24	U22376_cds2_s_at = 大	MYB
3	0.142	100%	33.3%	24	24	D26156_s_at = 大	SMARCA4
3	0.142	100%	33.3%	24	24	Z49194_at = 大	POU2AF1
3	0.142	100%	33.3%	24	24	Y08612_at = 大	NUP88
3	0.142	100%	33.3%	24	24	X99920_at = 大	S100A13
3	0.142	100%	33.3%	24	24	X95735_at = 小	ZYX
3	0.142	100%	33.3%	24	24	X93512_at = 大	TERF2
3	0.142	100%	33.3%	24	24	X82240_rna1_at = 大	TCL1A
3	0.142	100%	33.3%	24	24	X68560_at = 大	SP3
3	ほか、評価値の同じルールが多数生成された。						

7130項目(遺伝子の発現量7129項目＋白血病のタイプ1項目) の表形式データになっている。

このデータを対象に、特徴的ルール生成の分析実験を行った。ルール生成のツールとしてはDATAFRONT [5] を用いた。

この時の実験条件を以下に述べる。IF以下の条件項目としては7129遺伝子の発現データを用いた。THEN以下の結論項目は「白血病のタイプ = ALL型」とした。全72サンプル中、47サンプルがALL型なので、事前確率は65%である。遺伝子の発現量を表す各項目は、等数分割によって大・中・小の3つのカテゴリに分割した。等数分割とは、各カテゴリができるだけ同数のサンプルを含むようにしきい値を定める分割方法である。また、カバー率とヒット率の重みを決めるパラメータ(式(3)の β)は β = 1としてカバー率重視でルールを評価した。また、今回は最大条件節数を1 としてルール生成を行った。

4.2　結果検討

上記で述べた条件のもとで分析を行った結果を表1に示す。出力されたルールは評価値によってソートし、上位のものを掲載した。この評価値は前述した式(3)より算出される値である。なお評価値で順位が3位のものは同評価値のルールが多数出力されたため、その一部を挙げるにとどめた。表1は、1 行がそれぞれ1 つの独立したルールになっており、評価値の降順に並んでいる。評価値が同じルールが複数ある場合、それらの中の並び順には特に意味はない。

例えば、表1において1 行目のルールは「遺伝子U07139_at の発現量が大ならば、白血病のタイプがALLである傾向がある」を意味し、ヒット数25、カバー数25 は「遺伝子U07139_at の発現量が大」の条件を満たすサンプル数が25(カバー数) で、そのうち、さらに「白血病のタイプがALLである」が成り立つサンプル数が25(ヒット数) であることを意味する。カバー率は全サンプル中に占めるそのルールのカバー数の割合

2) Unigene は米 NCBI (National Center for Biotechnology Information) によって提供されている遺伝子に関するデータベースである。このカラムには Unigene で遺伝子の識別に用いられている名称を記した。

(25/72 = 34.7%)、ヒット率は、カバー数に占めるヒット数の割合である(25/25 = 100%)。

このルール生成結果を見ると、ここに取り上げたルールでは、ヒット率が 100% で、各ルールにおいて、そのルールがカバーしているサンプルについては正確に特徴を言い表している(すなわち、各遺伝子と白血病のタイプの強い関係を発見している)事がわかる。これらのルールのカバー事例数は 25～24 で、白血病のタイプが ALL であるサンプルの数は、もともと 47 なので、各ルールはこれらの約半分を説明していることになる。

次に、これらのルールに出現する遺伝子についてアノテーション情報を参照し、その意味について考察した。アノテーションとは遺伝子に関する注釈のことである。分析結果には、従来ガンと関係の深いことが知られている、細胞増殖、転写制御、細胞内情報伝達の遺伝子がルール上位にリストアップされており、この手法によって生物学的に妥当と思われる結果を得られたことが確認できた。表 1 に現れる遺伝子では、CACNB3 、 RAG2 、 SMARCA4 、 TCF3 、 PTPRCAP 、 APLP2 、 MYB 、 POU2AF1 、 S100A13、ZYX、TERF2、TCL1A、SP3 が上記の遺伝子にあたる。

5. 考察

一般に 1 枚のDNA チップからは数千～1万以上の遺伝子発現データが得られるが、検体を集めて来て、DNA チップの計測にかけるにはかなりの手間がかかるため、DNA チップデータは、属性数(遺伝子数)に比べてサンプル数が非常に少ないデータとなる傾向にある。通常の統計解析や機械学習のアルゴリズムを用いてこのようなデータを扱うケースでは、サンプル数に対して多数の属性を用いてモデルを構築するのが困難な場合がある。

一方で、特徴的ルール生成のアルゴリズムでは、あらかじめ設定した最大条件節数を上限とした属性の組み合わせを探索するため、属性が非常に多いデータにも対応できる。また、今回の実験では 1 条件節を持つルールの抽出を行ったが、このアルゴリズムを用いて複数の条件節を持つルールを抽出することも可能である。遺伝子ネットワークに代表されるように、遺伝子間の関連性は重要な研究テーマであるので、複数の遺伝子の関係を特徴として抽出できるこのような手法は有用である。

特徴的ルール生成の問題点としては、ルールの最大条件節数を大きくしたとき、探索のための計算量が非常に大きくなるという点が挙げられる。アルゴリズムの改良、グリッドなどの並列処理による高速計算などを検討し、複数の遺伝子を含むルールを実用的な計算時間で提示することが今後の課題である。

6. まとめ

本研究では遺伝子発現解析について、遺伝子の発現状態を一括して測定する DNA チップに着目し、特徴的ルール生成の手法を用いた計測データの分析について検討した。

また、実データへの適用例として、白血病患者の遺伝子発現データを用いた分析実験を行った。白血病のタイプをルールの結論部に設定して上記手法を適用した結果「遺伝子 U07139_at の発現量が大ならば、白血病のタイプは ALL である傾向がある」のようなルールを得た。これらのルールを既知のアノテーション情報と照らし合わせた結果、細胞増殖関連の遺伝子など、ガンとの関係が指摘されているいくつかの遺伝子が抽出された事を確認できた。

参考文献

[1] 角田,「マイクロアレイデータの解析」, ゲノム機能 ― 発現プロファイルとトランスクリプトーム pp.118-133, 中山書店, 2000

[2] Methods of Microarray Data Analysis - Papers from CAMDA 2000, Kluwer Academic Publishers, 2001

[3] 芦田・前田・高橋, 「データマイニングにおける特徴的ルール生成方式」, 情報処理学会第 50 回全国大会, 1995

[4] T. R. Golub et al., “Molecular Classification of Canser: Class Discovery and Class Prediction by Gene Expression Monitoring”, Science Vol.286 pp.531-537, 1999

[5] DATAFRONT ホームページ：
http://www.hitachi.co.jp/Prod/comp/soft1/datafront/datafront_home.htm