2002 年情報論的学習理論ワークショップ
2002 Workshop on Information-Based Induction Sciences (IBIS2002)
Fujiyoshida, Japan, Sep 8 - 10, 2002.

DNA マイクロアレイからの遺伝子間高次相互作用の推定

Inference of high-order interactions of genes from DNA microarray data

西村　信一*†　井上　真郷*‡　堀　玄§　甘利　俊一*
Shin-ichi Nishimura　Masato Inoue　Gen Hori　Shun-ichi Amari
中原　裕之*
Hiroyuki Nakarhara

Abstract: DNA microarray is a tool for genomic-scale measurement of mRNAs, and will be of more importance in the future. In microarray data analysis, several methods such as Hierarchical Clustering are proven to be useful. However, these methods rely on second-order interaction, and detecting higher order interaction in microarray data is an important task. We propose the use of log-linear model to analyze higher-order interactions between the genes in a publicly available dataset, and discuss the biological meaning of our result.
Keywords: DNA microarray, log-linear model, DNA interaction, information geometry

1　はじめに

近年、ヒトや酵母・イネなどのゲノムが続々と解読され、今後はその情報をいかに活用すべきかが課題となってきた。それにつれて、Bioinformatics あるいは Functional genomics といった新しい研究の舞台が広がりつつある。

ゲノム (genome) は遺伝子 (gene) の集合である。ヒトの場合、ゲノム中に約 3 万個の遺伝子が含まれていると考えられている。細胞のおかれた状態に応じて遺伝子のうち必要な部分がその都度メッセンジャー RNA(mRNA) として転写され、これを鋳型としてタンパク質が作られる。したがって、mRNA を調べることで、ある時点で細胞（群）がどのような活動をしているかを推定することが可能となる。

分子生物学的な手法の進歩により、1980 年代後半に Polymerase Chain Reaction(PCR) 法による微量遺伝子の増幅が発明され、細胞内の微量 mRNA の増幅・検出が可能となった。1990 年代終わりに、様々な工学的テクノロジーの進歩と相まってマイクロアレイ技術が登場した。多数のプローブをチップ上に集積し、コンピュータによって制御されたスキャナーでチップ上の蛍光を読みとることにより、多数（数千から数万）の遺伝子の発現についての同時測定が可能となった。

マイクロアレイのデータ解析の手法については多数の報告があるが、一定の評価を受けて多用されているものとして、階層型クラスタリング [1] などが挙げられる。これらは、2 つの遺伝子の対についての発現比較を多くの遺伝子の組み合わせに行うことで、遺伝子をいくつかの機能的な群に分類するものである。

一方で、遺伝子は互いに高度な依存関係を持っており、遺伝子の相互関係を「遺伝子ネットワーク」の形で解明することは、分子生物学の一つの目標となっている。バイオインフォマティックスの分野においても、マイクロアレイデータの高次の相互作用から遺伝子ネットワークを推定するべく、Bayesian Networks などのグラフィカルモデルを適用した報告がされている [2, 3]。

我々はこれまでに高次相互作用を分解・分析するためのツールとして対数線形モデルを用いて、神経細胞群のスパイク発火における高次相互作用の解析 [4, 5] や遺伝子の相互作用の解析 [6, 7, 8] を行ってきた。紙面に限りがあるので本稿では、主として 3 次・4 次のモデル、即ち遺伝子 3・4 個の場合について詳述することにする。一般の場合については、他の文献を参照していただきた

*理化学研究所 脳科学総合研究センター 脳数理研究チーム, 〒 351-0198 埼玉県和光市広沢 2-1, Tel. 048-467-6845, e-mail {nishi@mns.,minoue@,hori@bsp.,amari@,hiro@,}brain.riken.go.jp
Lab. for Mathematical Neuroscience, RIKEN Brain Science Institute 2-1, Hirosawa,Wako-shi,Saitama-ken, 351-0198, Japan
†東京大学大学院医学系研究科 耳鼻咽喉科学専攻
Department of Otolaryngology, Graduate School of Medicine, University of Tokyo
‡京都大学大学院医学研究科 耳鼻咽喉科・頭頸部外科学
Department of Otolaryngology - Head and Neck Surgery, Graduate School of Medicine, Kyoto University
§理化学研究所 脳科学総合研究センター 脳信号処理研究チーム
Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute

い [5, 7, 9]。2 節では、まず対数線形モデルと情報幾何学的座標系、それを使った検定、条件付き独立などについて説明する。3 節で、マイクロアレイデータにその手法を適用した例を示す。最後に、4 節でまとめの議論をする。

2 対数線形モデルへの情報幾何学的アプローチ

ここで、2 値 (binary) の確率変数ベクトルに対する対数線形モデルについて簡潔に述べる [9]。

まず、n 次元の確率変数ベクトル $\boldsymbol{x} = (x_1, \cdots, x_n)$ を考え、それぞれの x_i が一つの遺伝子に相当し、mRNA が発現しているかどうかにより 0 ないし 1 の値をとるものとする。$\{p(\boldsymbol{x})\}$ がとりうる全ての確率分布は $(2^n - 1)$ 次元の多様体 $\boldsymbol{S}_n$ としてとらえることができる。

$\boldsymbol{S}_n$ を表す一つの座標系は、

$$\begin{aligned} \eta_i &= E[x_i], \ \eta_{ij} = E[x_i x_j], \ (i < j), ..., \\ \eta_{12\cdots n} &= E[x_i \cdots x_n] \end{aligned}$$

により表される $\eta-$ 座標系であり、これは $2^n - 1$ 個の次元を持つ。

一方、対数線形モデルを用いた $\theta-$ 座標系も同様に $2^n - 1$ 個の次元を持ち

$$\log p(\boldsymbol{x}) = \sum \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \sum_{i<j<k} \theta_{ijk} x_i x_j x_k + .. + \theta_{1\cdots n} x_1 .. x_n - \psi$$

と展開できる（ψ は $-\log p_{0\ldots0}$ に相当する正規化項である）。

この $\eta-$ 座標系と $\theta-$ 座標系は双対構造をなしており [10]、確率分布間の距離を情報幾何学的にとらえることで、変数間の相互作用を統一的かつ比較的容易に解析することができる [4, 5, 6, 9]。

2.1 3 次相関における有意水準の計算

前項では一般の場合について述べたが、ここでは 3 次の場合において相互作用の分解と有意水準の計算について説明する。この場合の対数線形モデルは

$$\begin{aligned} \log p(\boldsymbol{x}) &= \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \\ &+ \theta_{12} x_1 x_2 + \theta_{13} x_1 x_3 + \theta_{23} x_2 x_3 \\ &+ \theta_{123} x_1 x_2 x_3 - \psi \end{aligned}$$

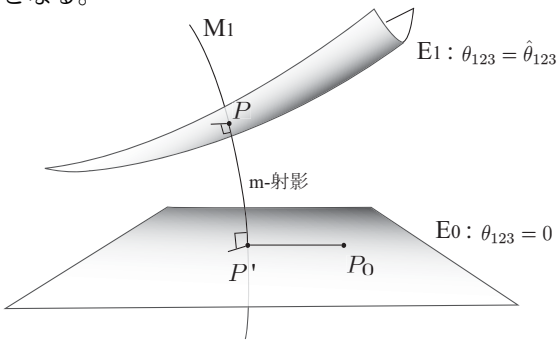
となる。ここで、データから得られる $p_{ijk} = Prob(x_1 = i, x_2 = j, x_3 = k)$ の推定値を $\hat{p}_{ijk}$ とすると最尤推定により

$$\begin{aligned} \hat{\theta}_1 &= \log \frac{\hat{p}_{100}}{\hat{p}_{000}}, \ \hat{\theta}_2 = \log \frac{\hat{p}_{010}}{\hat{p}_{000}}, \ \hat{\theta}_3 = \log \frac{\hat{p}_{001}}{\hat{p}_{000}}, \\ \hat{\theta}_{12} &= \log \frac{\hat{p}_{110}\hat{p}_{000}}{\hat{p}_{100}\hat{p}_{010}}, \ \hat{\theta}_{13} = \log \frac{\hat{p}_{101}\hat{p}_{000}}{\hat{p}_{100}\hat{p}_{001}}, \\ \hat{\theta}_{23} &= \log \frac{\hat{p}_{011}\hat{p}_{000}}{\hat{p}_{010}\hat{p}_{001}}, \ \hat{\theta}_{123} = \log \frac{\hat{p}_{111}\hat{p}_{100}\hat{p}_{010}\hat{p}_{001}}{\hat{p}_{110}\hat{p}_{101}\hat{p}_{011}\hat{p}_{000}} \end{aligned}$$

となる。ここで、θ_{123} は、3 次相関を表している。今、推定された確率分布を P とおく。この P の $\hat{\theta}_{123}$ の帰無仮説 $\theta_{123} = 0$ に対する検定を考えよう。帰無仮説の確率分布を P' とおく。このとき、$\theta_{123} = 0$ である任意の確率分布を P'' とすると、Kullback-Leibler Divergence を利用して、

$$P' = \arg \min_{P''(x) \in E_0; \theta_{123}=0} D[P(x); P''(x)]$$

で与えられる。即ち、P から E_0 におろした $m-$ 射影が P' となる。



混合座標系で表すと

$$\begin{aligned} & P(\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3, \hat{\eta}_{12}, \hat{\eta}_{13}, \hat{\eta}_{23}, \hat{\theta}_{123}) \\ & \stackrel{m-\text{射影}}{\longrightarrow} P'(\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3, \hat{\eta}_{12}, \hat{\eta}_{13}, \hat{\eta}_{23}, 0) \end{aligned}$$

となる。Fisher information matrix を $G = (g_{ij})$、標本数を N とすると、尤度比検定により、検定統計量 λ は、

$$\begin{aligned} \lambda &= 2 \log \frac{\prod p(\boldsymbol{x}_i; \hat{\eta}, \hat{\theta}_{123})}{\prod p(\boldsymbol{x}_i; \hat{\eta}, 0)} \\ &= 2 \sum_{i=1}^{N} \log \frac{p(\boldsymbol{x}_i; \hat{\eta}, \hat{\theta}_{123})}{p(\boldsymbol{x}_i; \hat{\eta}, 0)} \\ &\simeq 2NE \left[\log \frac{p(\boldsymbol{x}; \hat{\eta}, \hat{\theta}_{123})}{p(\boldsymbol{x}; \hat{\eta}, 0)} \right] \\ &\simeq 2ND \left[p(\boldsymbol{x}; \hat{\eta}, \hat{\theta}_{123}); p(\boldsymbol{x}; \hat{\eta}, 0) \right] \\ &\simeq N g_{77} {\hat{\theta}_{123}}^2 \\ &\sim \chi(1) \end{aligned}$$

となり、有意水準を計算できる。

先ほどの式の最後で、Kullback-Leibler Divergence を計算するところについて補足する。情報幾何学において、

Fisher information matrix は Riemannian metric tensor として働き

$$2D\left[p(x,\boldsymbol{\xi});p(x,\boldsymbol{\xi}+d\boldsymbol{\xi})\right]=\sum g_{ij}d\boldsymbol{\xi}_i d\boldsymbol{\xi}_j$$

となることが知られている。先ほどのように混合座標系を使った場合、η と θ は直交しているため、$d\boldsymbol{\xi}$ が θ_{123} のみを変える方向に動いた場合

$$\sum g_{ij}d\boldsymbol{\xi}_i d\boldsymbol{\xi}_j = g_{77}{d\boldsymbol{\xi}_7}^2$$

となる。

ここでは θ_{123} について紹介したが、ここでの議論は他のパラメータについても、あるいは次元が高くなった場合にも成り立つ。

2.2 条件付き独立の強い定義・弱い定義

グラフィカルモデルにおいては条件付き独立の概念が非常に重要となる。一般的に、「X と Y は Z を与えられた場合に独立である」という関係がある場合、$X-Z-Y$ というグラフが得られる。この場合、

$$P(X,Y|Z)=P(X|Z)P(Y|Z)$$

という関係式が成り立つ [11]。ここでは、条件付き独立の「強い定義」が用いられている。一方、例えば、2 値の変数で考えた場合、$Z=0$ と $Z=1$ の 2 通りで、上の関係が異なっている場合があり得る。

$$\begin{aligned}P(X,Y|Z=0) &= P(X|Z=0)P(Y|Z=0)\\ P(X,Y|Z=1) &\neq P(X|Z=1)P(Y|Z=1)\end{aligned}$$

このように、Z の実現値を区別して独立性を考えるのは、条件付き独立の「弱い定義」と呼ばれる。一般的なグラフィカルモデルの枠組みでは、この「弱い定義」の関係を表現することはできない。しかし、対数線形モデルにおいては

$$\begin{cases}P(X_1,X_2|X_3=0)=P(X_1|X_3=0)P(X_2|X_3=0)\\ P(X_1,X_2|X_3=1)\neq P(X_1|X_3=1)P(X_2|X_3=1)\end{cases}$$

$$\Rightarrow \theta_{12}=0, \theta_{12}+\theta_{123}\neq 0$$

のように表すことが可能である。

さらに高次元になった場合にも同様であり、例えば4次元では

$$\begin{cases}P(X_1,X_2|X_3=0,X_4=0)\\ \quad =P(X_1|X_3=0,X_4=0)P(X_2|X_3=0,X_4=0)\\ P(X_1,X_2|X_3=1,X_4=1)\\ \quad \neq P(X_1|X_3=1,X_4=1)P(X_2|X_3=1,X_4=1)\end{cases}$$

$$\Rightarrow \theta_{12}=0, \theta_{12}+\theta_{123}+\theta_{123}+\theta_{124}+\theta_{1234}\neq 0$$

と、$\theta-$ 座標系から、変数間の相互作用について（一般的なグラフィカルモデルよりもさらに）細かな解析が可能である。

生物学的な立場からみると、遺伝子が発現していないときを0，発現している時を1と考え、遺伝子 X_3 が遺伝子 X_1 と X_2 両方の発現に関与する場合、

- 遺伝子 $X_3=0$ の場合：X_1 と X_2 は独立
- 遺伝子 $X_3=1$ の場合：X_1 と X_2 の間に相関が見られる

という関係が成立する可能性は大いに高い。X_3 が抑制的に働く場合には、逆の関係になることもあり得る。

4次の場合にはさらに、

- 遺伝子 $X_3=1, X_4=1$ の場合：X_1 と X_2 の間に相関が見られる
- 上記以外の場合：X_1 と X_2 は独立

という関係がみられた場合、X_3 と X_4 という遺伝子から作られるタンパクが2つまとまって効果を現す、というタンパク間の相互作用にまで踏み込んでいける可能性がある。

2.3 対数線形モデルによる遺伝子発現の解析

遺伝子ネットワークの解明は、バイオインフォマティックスにおいて一つの重要な分野であり、Bayesian Network を適用したものなど様々な手法が提案されている。これらの場合にも、データを「量子化」して扱うべきか、そのままのデータを用いた線形モデルを使うべきか、議論がなされているところである。前者の場合、閾値を設定してデータを量子化するのであるが、何段階に量子化するか、あるいは閾値をどこにおくか、により結果が大きく変わりうることが指摘されている。

マイクロアレイのデータ解析においては、実験回数が少ないというのが最大の制約である。一方で、データを量子化してしまうと、データの情報の一部がある意味で捨てられてしまうわけで効率が余りよいとはいえない。

従って、我々はデータが $[0,1]$ に収まるように変形し、それを「データが1である確率」として取り扱う。これにより、例えば、

$$\begin{aligned}\eta_i &= E\left[x_i\right],\ \eta_{ij}=E\left[x_i x_j\right],\ (i<j),\ldots,\\ \eta_{12\cdots n} &= E\left[x_i\cdots x_n\right]\end{aligned}$$

と η は計算でき、また θ も 2 値の場合と同様に計算できる [6, 7]。これにより、量子化によるデータの損失を抑えることが出来た。

問題は、いかに $[0,1]$ に変形するかということになるが、

- 最小値が 0 に、最大値が 1 になるように線形変換
- 何らかの非線形関数、例えばシグモイド関数などで変換
- 小さい順にソートし、最小のものを 0，最大のものが 1 になるようにランクをつける

などが考えられ、本研究においては順位を用いた変換を行った。なお、順位を用いたものは、２次においては Spearmann の順位相関に相当するが、今回提案する手法ではより高次の相互作用をみることができる。

3 実験

Yeoh らは、小児の白血病の予後 ($\simeq$ 生存期間) と白血病細胞の遺伝子発現について報告をおこない [12]、その実験に用いたマイクロアレイのほぼ生のデータを

http://www.stjuderesearch.org/data/ALL1

において公表した。この dataset は 327 人の急性リンパ性白血病患者に対し 12,558 種の遺伝子発現を検査したものである[1]。

まず、データに対する前処理として、実験全体にわたっての発現量の変化が小さいものを除外し（その結果遺伝子の総数は 11,826 になった)、前項で述べた順位統計量を用いて計算を行った。

3.1 実験結果

今回我々が提案する手法が、２次の相関のみを取り扱う Hierarchical Clustering などと比べてどのように違った情報を得ることができるのか、それを端的に示すために単純に純３次相互作用（これは 2 次相関の計算では導き得ない）のみに注目し、それが大きい遺伝子の組を選び出した。

遺伝子の組 (X_1, X_2, X_3) について高い３次相関を示したものと、その $p-$ 値を表 1 に示す。

転写因子・転写制御因子が多く登場すること、代謝に関わるものが少ないことが見て取れる。

このうち、POU2AF1 という遺伝子は、転写調節因子であり、白血球の分化にかかわること、HLA-DRA,HLA-DMA 遺伝子の転写制御をおこなうことがわかっている。

[1] 正常な体内では、骨髄に存在する造血幹細胞が、様々な刺激を受けることにより赤血球や白血球、血小板などに様々な段階を経ながら分化していく。それに対して、白血病とは、白血球のがんであり、遺伝子の何らかの異常によりこの分化の途中のいずれかの段階で細胞が腫瘍化して無軌道に増殖してしまう病気である。

表 1　高い 3 次相関を示した遺伝子の組

太字は転写制御因子、斜体はシグナル伝達に関する遺伝子

X1	X2	X3	p-value
POU2AF1	PDLIM1	HLA-DRA	0.0166
POU2AF1	CD9	HLA-DMA	0.0193
PDLIM1	HLA-DPA1	**FOX01A**	0.0215
ZAP70	*ICAM3*	CD24	0.0231
FBP17	HELO1	**ADPRT**	0.0238
IL1B	HELO1	**ADPRT**	0.0243
PSAP	*PTPRC*	*FLT3*	0.0253
STK38	GPX1	**ADPRT**	0.0256
FBP17	GLUL	CD24	0.0282
SLC9A3R1	HELO1	TM4SF2	0.0315
IGLL2	*BTK*	*FLT3*	0.0319

この遺伝子についてさらに高次相関関係を調べてみることとする。

X_3 を a) POU2AF1 の場合、b) ハウスキーピング遺伝子[2]としてよく使われる G6PD の場合、c) 転写因子である XBP-1 の場合、に、固定して考える。X_1, X_2 にあてはめうる遺伝子の組 69,909,400 通りすべてについて $(\theta_{12}, \theta_{12}+\theta_{123})$ をプロットした結果を図 1 に示す[3]。

a) の POU2AF1 の場合は、b) の G6PD と比較して左上方向・右下方向への点の分散が大きいことが見て取れるが、これは「弱い定義の」条件付き独立が大きくなる方向である。(強い定義の条件付き独立では $\theta_{12}=\theta_{123}=0$ となりこの図では原点にあたる）$\theta_{12}=0$ の場合 $X_3=1$ で独立、$\theta_{12}+\theta_{123}=0$ の場合 $X_3=0$ で独立、に相当する。

生物学的に見ると、POU2AF1 は、白血球の分化にかかわる遺伝子であり、これが欠損している場合白血球の分化が途中で止まってしまうことが報告されている。一方、G6PD はほぼすべての細胞に普遍的に存在する遺伝子で、ほかの遺伝子の発現に影響を与えることはないと考えられている。白血球の分化にかかわる転写因子として知られている XBP-1 という遺伝子について同様のプロットを行った場合も、POU2AF1 と似た結果となり、分化のための「スイッチ」として働くような遺伝子は、他の遺伝子に対して「弱い」条件付き独立性を与える可能性がある。

次に、どのような遺伝子が POU2AF1 によって制御されているのかを調べてみる。図 1 において、$\theta_{12}<0$ で $\theta_{12}+\theta_{123}\simeq 0$ のところの点に注目し、$\theta_{12}<0$ という条件で帰無仮説 $\theta_{12}=0$ に関する $p-$ 値を求めた。

[2] 細胞の定常機能の維持のため、常に一定の発現をしていると一般的に認められている遺伝子。発現量を比較する際に、対照として用いられることが多い。

[3] ここで、図に示すにあたって $\theta_{12}, \theta_{12}+\theta_{123}$ について計算結果をそのまま使用したが、実際に遺伝子を選択する際には情報幾何学的尺度で $p-$ 値を計算して行うべきである。

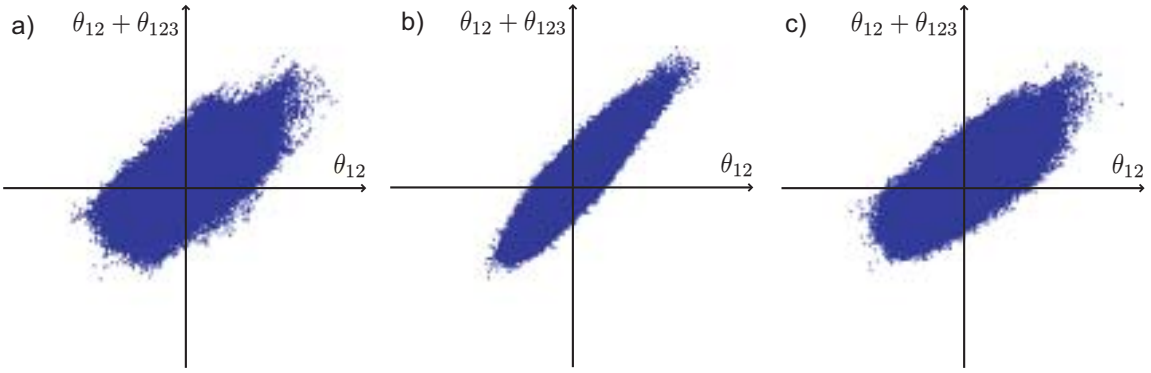


図 1　X_3 = a)POU2AF1, b)G6PD,c)XBP1 を固定した条件での $(\theta_{12}, \theta_{12}+\theta_{123})$ の分布

表 2　X_3 =POU2AF1 で θ_{12} が小さい遺伝子の組

X1	X2	p-value
GPX1	LCK	0.000219
FLT3	ZAP70	0.000268
GPX1	ZAP70	0.000275
CDKN1A	ZAP70	0.000331
FLT3	LCK	0.000351
GPX1	CD3Z	0.000541
CDKN1A	LCK	0.000544

表 2 の遺伝子のうち、LCK,FLT3,ZAP70 といった遺伝子はタンパク質キナーゼ、すなわち細胞内でほかのタンパクをリン酸化することで情報伝達を行う情報伝達物質である。また CD3Z も細胞膜から細胞内に情報を伝達するのに必要であることがわかっている。これらの遺伝子も、遺伝子発現には直接かかわらないにしても白血球の分化に関係していることがわかっている。

これらの遺伝子は POU2AF1=1 のとき互いに独立で、POU2AF1=0 のときに相関がある、ということになっている。ここで、(X_1, X_2, X_3) = (GPX1,LCK,POU2AF1) にさらに新たに X_4 という遺伝子を付け加えて、

- 遺伝子 $X_3 = 0, X_4 = 0$ の場合：X_1 と X_2 の間に相関が見られる
- 上記以外の場合：X_1 と X_2 は独立

という条件で当てはまる遺伝子を探したところ、$p-$ 値は 0.1045 で IGLL2 という遺伝子が最も当てはまるという結果になった（図 2 ）。

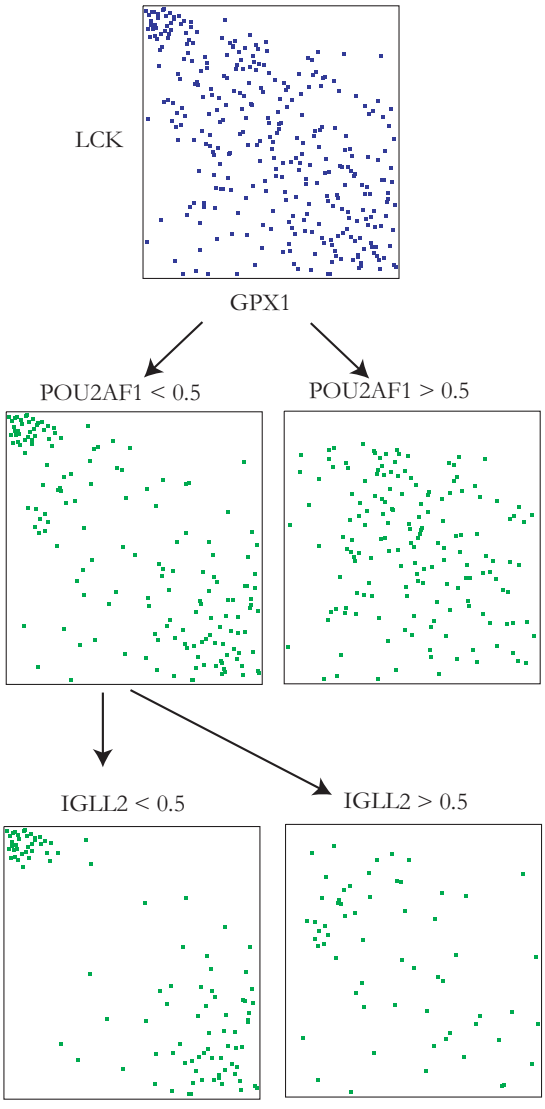


図 2　GPX1 と LCK の相互作用を POU2AF1,IGLL2 で分解

4 まとめ

対数線形モデルを DNA マイクロアレイにより得られた遺伝子発現データに適用し、高次相互作用を解析した。実験に用いたデータは白血病細胞のものであったが、我々の手法により、白血球の分化に関する遺伝子が多数ピックアップされたのは、我々の手法の生物学的な有効性を示しているものと考えた。とくに、POU2AF1 や、表２に含まれる細胞内シグナル伝達に関わる遺伝子など、細胞の機能を大きく切り替える「スイッチ」のような役割を果たすものに強い３次相互作用がみられるというのは、生物学的にも興味深い結果である。

また、対数線形モデルによりある遺伝子の影響を受ける遺伝子の候補を選び出すことができた。このように、対数線形モデルは遺伝子相互作用の解明に向けての有力なツールになりうると考えられる。

参考文献

[1] Eisen, M. B. et al. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS USA*. 95(25): 14863-8.

[2] Akutsu, T. et al. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Proc. Pacific Symp. on Biocomputing* 4: 17-28.

[3] Pe'er, D. et al. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17 Suppl 1: S215-24.

[4] Nakahara, H. and S. Amari (2002). Information-geometric decomposition in spike analysis. *Advances in Neural Information Processing Systems* 14.

[5] Nakahara, H. and S. Amari (2002). Information geometric measure for neural spikes. *To appear in Neural Computation.*

[6] Nakahara, H. et al. (2002) Log linear model and gene interaction in DNA microarray data. *as RIKEN BSI BSIS Tech Report No02-2*

[7] Nakahara, H. et al. (2002) *submitted.* Gene interaction in DNA microarray data is decomposed by information geometric measure.

[8] 中原裕之ほか (2002) 遺伝子解析と脳の数理 **数理科学** 10 月号別冊 (予定)

[9] Amari, S. (2001). Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory* 47(5): 1701-11.

[10] Amari, S. and H. Nagaoka (2000). Methods of Information Geometry. *AMS and Oxford University Press.*

[11] Edwards, D. (2000) Introduction to Graphical Modelling, Second Edition. *Springer,New York*

[12] Yeoh, E. J. et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 1(2):133-43