題目：遺伝子ネットワークを推(み)る

執筆者名（ふりがな）：中原　裕之（なかはら　ひろゆき）
勤務先（所属）：理論統合脳科学研究チーム
住所：351-0198　和光市広沢２－１
　　独立行政法人理化学研究所 脳科学総合研究センター
Email: hn@brain.riken.jp

１．序

脳機能解析にバイオインフォマティクスをどう上手く使うか、と良く聞かれる。この小論で考えてみたい。どう上手く使うのか？そもそもバイオインフォマティクスという用語が何を指すのか、ややはっきりしない。端的に言えば、この用語は、大規模な生物学的データをどのように扱い、解析するかという問題意識と共に使われる。ならば上手く使うための要諦は、簡単だ。それは、「個々の生物学の設問に応じつつ、その大規模データの特性を生かしたデータ解析をきちんと行えばよい」となる。そして、脳機能解析では、個々の脳機能解析の目的に即して、適切な大規模データを収集し、適切な解析を行えば、上手くいくハズである。上手くやるためにどの設問にどのデータを使うか、そしてそのデータにどの解析を用いるか、工夫の仕方は色々あろう。本小論では、その工夫の一つをお見せしたい。以下、第２節で、まず「数理的直観」の大切さに触れる。その上で、情報幾何を用いた高次相互作用を検出する解析手法[1,2]、具体的には、遺伝子マイクロアレイのデータにおける遺伝子カスケードの検出の解析例[2,3]を紹介する（第３節)。最後に、未来を簡潔に展望する（第４節)。以下、数学的に厳密な議論や数式の仔細を示すのは避け、やや粗くてもその心を伝えることを優先する。なるほどという気になることが大切だと思う。(数理的な仔細が気になる読者は、ぜひ巻末の原論文にあたってほしい。）ともかく始めよう。

２．なぜ「数理的直観」が大切か

バイオインフォマティクスでは、データが何らかの意味で大規模である。例えば、遺伝子マイクロアレイは多数の遺伝子が同時に計測されるという意味で大規模だし、functional MRI では同時に計測される脳活動のボクセルが多数という意味で大規模である。この大規模なデータというのは、数学的には高次元なデータということである。人間は３次元の世界に住む生き物である。(ちょっと物理が好きな人は、時間も含めれば４次元だというかもしれない。）１００次元の球と言われても不慣れな人は往生する。いずれにしても、高次元なデータというのは私たちの素のままの直観では騙されやすく、多数の相互作用を正しく把握するのが難しい。数理的に現象を把握・整理することで、正しい（数理的）直観が得られる。これがあれば、与えられたデータをもとに正しい解析手法を選択するのはそれほど難しくない。バイオインフォマティクスとは、この数理的直観を生かして、データを眺めやすくする、かつ正しく把握できるようにする、ということに他ならない。数理的直観の使い方にもバラエティがあり、その使い方の違いで、様々なバイオインフォマティクスの手法となっていくのである。数理的な議論というのは、この数理的直観を、数学的に分かりやすく書き下すことである（もちろん、それに慣れない人には、むしろ分かりづらい記述になってしまうのであるが)。

この小論で、大規模データ解析で特に重要な数理的直観を列挙し、各々をきちんと説明

するには、紙面が足りない。結構面白い話はあるのだが、これらの議論は涙を呑んですっとばす。後述する解析手法にも関係あるところで、少しだけ数理の話を述べておく。高次相関という視点と、条件付独立という概念とその「弱い定義」と「強い定義」について述べたい。「相関」とか「相互作用」という言葉を聞くと、通常、二つの「もの」（変数）の間にどれくらい関係（相関）があるかということが念頭に来るようだ。しかし、実はこの「相関」は、あくまで二者の相関、つまり「２次相関」に過ぎない。実は、例えば三つの「もの」（変数；X_1, X_2, X_3としよう）があるときには、そのうちの二つの間にある２次相関（例：X_1とX_2の間の相関）以外にも、三者間の相関「３次相関」が存在する。この３次相関は、例えば、３つの遺伝子が同時に発現する頻度を表すのに良い指標である。このように多数の変数が存在すれば、３次以上、４次、５次などなどの高次相関が存在する。この高次相関は、遺伝子データの場合、例えば、ある実験条件に特異的に発現している遺伝子集団や、あるいは遺伝子ネットワークや転写制御因子に関わる遺伝子群などを特定したいなどの課題には、有効に使われるべき情報である。後述する解析手法は、この高次相関を利用している。

この高次相関を使う手法の例として、ベイジアン・ネットワーク（グラフィカルモデル）を用いた方法が既に試され、そこそこの成功を収めている[4]。この手法では、確率変数間の条件付独立性が重要な役割を果たす。実は、そこで用いられるのは「条件付独立性の“強い”定義」である。これに対して、我々の開発した手法は、「条件付独立性の“弱い”定義」を用いるのが工夫した点である（その上で、対数線形モデルで情報幾何座標系を利用する[1,2]）。この“弱い”と“強い”では、何が違うのか？一番簡単な例として、３つの変数X_1, X_2, X_3で、最初の二つの変数X_1とX_2が、最後の変数X_3に対して条件付独立な場合を考えよう。「強い定義」では、この場合は、

$$P(X_1, X_2 \mid X_3) = P(X_1 \mid X_3) P(X_2 \mid X_3)$$

と書ける。ここで、それぞれのPは確率分布を表している。等式の左側は、X_3に条件づけされているX_1とX_2の同時確率分布を表している。それが、等式の右側の二つの確率分布の積に等しいことが示されている。この二つの確率分布は、それぞれX_3に条件づけされたX_1の確率分布と、X_3に条件づけされたX_2の確率分布を表す。

このような関係を使ってベイジアン・ネットワーク（グラフィカルモデル）では、異なる変数（つまり遺伝子）間の関係を記述する。ここで、注意したいのは、上の等式は、X_3にあたる遺伝子がどのような発現をしても上の関係式が成立する、ことを主張していることである。遺伝子の発現が互いに影響を与える場面を考えると、この主張はやや強すぎる観がある。例えば、X_3の遺伝子が発現していないときと、いるときを、それぞれ、$X_3 = 0$、$X_3 = 1$と表してみよう。遺伝子制御、特に遺伝子転写制御、を考えると、私たちが知りたいのは、むしろ、この$X_3 = 0$、$X_3 = 1$の各々の場合で、他の遺伝子（例：X_1とX_2）がどのような

関係にあるかである。

実は、上で触れた「弱い定義」では、

$$P(X_1, X_2 \mid X_3 = 0) = P(X_1 \mid X_3 = 0)P(X_2 \mid X_3 = 0) \quad (1)$$

と、

$$P(X_1, X_2 \mid X_3 = 1) = P(X_1 \mid X_3 = 1)P(X_2 \mid X_3 = 1) \quad (2)$$

の二つを場合分けして、条件付独立を考える。この弱い定義でデータを解析したほうがよりきめの細かい解析が可能になる。例えば、(1) が成立する一方、(2) が成立しない場合は、$X_3 = 1$ (X_3 が発現しているとき) に、初めて二つの遺伝子の相互作用が生じる (正確には、正の相互作用と負の相互作用の両方がありえて、どちらが生じるかはデータ解析の過程で検出できる)。

このように高次相関 (相互作用) を検出し、かつ条件付独立の弱い定義を使うことで、遺伝子制御の推定を行う手法を開発したので、その例を次節に示す。

３．遺伝子ネットワーク、遺伝子発現調節の検出

難しい話は抜きにして、実際のデータ解析例を眺めよう (図１)。ここで使うデータは、白血病患者 327 人に対して骨髄から採取した細胞の遺伝子発現をマイクロアレイで検査した実験データである[5]。この雑誌の読者の多くはご存知だと思うが、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する。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制限に増殖してしまう病気である。したがって、一般に、分化にかかわる遺伝子の働きを調べることが重要となる。この問題意識を踏まえ、下記に述べる解析手法を試すためにこのデータを使ってみた。つまり、白血球 (あるいは白血病細胞) の遺伝子ネットワークを調べることを想定して、手法を適用している。

では図１に戻ろう。紙面も限りがあるので、“生物学的”な意味を云々するのは極力避けて、ともかく図の上で“視覚的”に結果を見ることにする。このデータでは、約１万数千の遺伝子発現が同時に計測されている。(それが 327 サンプルある。) この約１万数千の遺伝子の中から、IGHM と XBP-1 の二つを取り上げた (この二つの遺伝子を取り上げたのは、主に解析手法のテストのためである。ただ興味深いのは、XBP-1 は、CREB に似た転写因子をコードする遺伝子であり、プラズマ細胞の分化に必要とされる。一方、IGHM はプラズマ細胞から分泌された免疫グロブリンのサブユニットをコードする遺伝子である。それゆえ、2 つの遺伝子の発現は何らかの形で関連すると予想できるのだが、一方で、XBP-1 による免疫グロブリンの直接の転写制御はありえそうもないという報告もある[6]。これより、この 2 つの遺伝子間相互作用に、寄与する他の遺伝子を見つけられないかという問いが思い浮かぶだろう。このような状況で、後述の解析手法が有効だろうと考えたので、この二

つの遺伝子を解析の題材に選んでいる。）この二つの遺伝子の 327 サンプルでの発現の計測値の散布図を図１A に示した。右下の COR の値は相関係数の値である。図１A からは、この二つの遺伝子の間に何か特有の関係があるようにはとても思えない。さて、我々の手法を使って、残りの約１万数千の遺伝子の中で、IGHM と XBP-1 の二つの遺伝子の間の“関係”、つまり相互作用、を最も変化させていそうな遺伝子を探した。見出されたのが ADPRT という遺伝子である。この遺伝子の発現の強弱に応じて、その発現が弱いときに限った IGHM と XBP-1 の散布図が図１B a)に示してある。一方、強いときに限った散布図が図１B b)である。図１B a)では、XBP-1 の発現が大きいときには、IGHM の発現も大きくなる傾向が強い。つまり正の相関がある。一方、図１B b)では、XBP-1 の発現が大きいときには、IGHM の発現が小さくなる傾向がある、つまり負の相関がある。このように、図１A ではほとんど関係がないように見えていた IGHM と XBP-1 の発現は、実は ADPRT の発現の大小を手がかりにすると、正の相関がある場合と負の相関がある場合にうまく区別されることが分かる。この ADPRT を見出すときに、もとのデータの中に隠されていた３次相関の情報が使われている。

同様の作業を繰り返すと、より詳細な関係が見えてくる。例えば、ADPRT の発現が小さい場合に着目しよう。そのような場合に、IGHM と XBP-1 の相互作用を最も変化させる遺伝子を、我々の手法で探すと、TM4SF2 が見つかった。先ほど同様にこの TM4SF2 の発現が小さい場合と大きい場合に分けて、IGHM と XBP-1 の散布図を表示したのが、図１C の a), b)である。一見して分かるように、図１B a)に見えた正の相関のほとんどは、この TM4SF2 の発現が大きい場合に由来することが分かる（図１C b)。この図１C の結果を見出すために使われた情報が４次相関である。図を眺めるコツを分かっていただけただろうか？さらに、図１C b)の場合について、５次の相関を手がかりに見出された遺伝子が AF1Q で、その結果が図１E に示してある。同様の手続は、興味のあるどの条件についても行うことができる。例えば、図１B、つまり ADPRT の発現が大きい場合について、同様に調べた結果が図１D,F に示してある。

以上の解析結果をもとに想定しうる遺伝子の互いの制御の関係を図示したのが図２A である。念のため断っておくが、この図はあくまで上述のデータ解析から推測される遺伝子間の相互作用を表した模式図である。これで生物学的知見が確定されたと思うのは早計である。この図は、むしろ詳細な生物学的実験を更に行うための案内図と理解すべきである。参考までに、遺伝子マイクロアレイのデータ解析に頻繁に利用されるヒエラルキカル・クラスタリングを行って遺伝子群の順序を並べ替えたが図２B である。そこに、図２A の遺伝子がどこに位置するかを示してある。これらの遺伝子群が、ヒエラルキカル・クラスタリングで順番づけの上では、ずいぶん離れた場所に位置することに注目してほしい。ヒエラルキカル・クラスタリングが２次相関のみを利用する手法であるのに対して、上述の解析手法は２次・３次そしてそれ以上の高次相関を利用している。それゆえ、ヒエラルキカル・クラスタリングでは見ることができない相互作用が、図１の手法では検出できている。

このように、高次の遺伝子相互作用を調べることによって、遺伝子発現調節の「スイッチ」となっている遺伝子を探し出せる可能性がある。このような用途に利用できる解析手法は、分子生物学の研究現場では強く求められていると思う。なお上記手法は、あくまで解析手法の適用を示す一例である。提案した解析手法は、情報幾何という数理的な一般的枠組の上に構想されている。それゆえ、実際の実験現場の設問に応じて、手法を調整することが可能である。(今回紹介した手法のみならず、一般に、実験の設問に応じて手法は調整し、ベストのアプローチをしたほうが良い。) 例えば、生物学的知見から（二つだけでなく）もっと多い複数の遺伝子群を最初に想定して、その群全体の発現に変化を与える遺伝子を検出したりすることもできる。そして、図２Ａのような図を作成し、これを案内図として、より詳細な生物学的実験を行うことで、生物学的知見を確定することが望まれる。

4. 結語

本小論では、バイオインフォマティクスの解析の一例を駆け足で見た。つい忘れられがちだが、数理的直観が大切なことも述べた。実は、上述した解析手法を開発するときに用いた数理的議論、それを支える数理的直観の基礎は、情報幾何と呼ばれる数理情報科学の研究である[7]。紙面の都合と読者層を考え、数理的議論は差し控えたが、数理の基礎研究が実は上述の実用を意識したデータ解析手法の研究にとても役立っていることは強調しておきたい。また、本小論では、遺伝子解析への適用として例を示したが、実はこの手法は、他のバイオインフォマティクスのデータにも適用が可能である。例えば、神経細胞の活動記録などでも高次相関が重要である。実際、上述の解析手法の研究は、最初は、多数の神経細胞の活動を同時記録したデータの解析手法の研究として始まった。この解析では、例えば、多数の神経細胞の間の情報のやりとりを、多電極細胞記録の計測結果から推定する。その研究も現在進行中である。私の研究室では、脳全体あるいは神経細胞集団レベルでの回路、つまり脳ネットワークの研究と、遺伝子レベルでの回路、つまり遺伝子ネットワークの研究を繋げていきたいと考えている（ラボの URL: www.itn.brain.riken.jp；現在研究員等を募集中)。近年、脳研究に役立つ様々な計測技術が急速に発達している。解析手法の発展が、これからますます重要になってくる。今回紹介した手法は、広い意味で、計測技術の発展に伴った解析技術の発展と捉えることができる。

脳機能解析にバイオインフォマティクスは有効である。上手く使えばよいのだ。どうだろうか？読者がそう確信するのに本小論が少しでも役立てば、幸いである。

1) Nakahara, H. & Amari, S. Information geometric measure for neural spikes. *Neural Computation* 14(10), 2269-2316 (2002).

2) Nakahara, H., Nishimura, S., Inoue, M., Hori, G. & Amari, S. Gene interaction in DNA microarray data is decomposed by information geometric measure. *Bioinformatics* 19(9), 1124-1131

(2003).
3) 中原 裕之, 堀 玄, 井上 真郷, 西村 信一. (2002). 遺伝子解析と脳の数理,『脳情報数理科学の発展(数理科学･別冊)』, pp133-143.
4) Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17 Suppl 1, S215-S224 (2001).
5) Yeoh, E.-J. et al. Classification, subtype discovery, and prediction of outcome in pediatric acute important leukemia by gene expression profiling. *Cancer Cell* 1(2), 133-143 (2002).
6） Reimold,A.M. *et al.* (2001) Plasma cell differentiation requires the transcription factor xbp-1. *Nature*, **412**, 300–307.
7) Amari, S. & Nagaoka, H. Methods of information geometry. (Oxford university press, 2000).

図のキャプション

図１　情報幾何座標系を利用した遺伝子相互作用の解析例。文献の図を改変[2]。

図２　A.　図１の解析から推定される相互作用（文献[2]より)。　B.　ヒエラルキカル・クラスタリングの図（文献[2]より)。

追記：この小論をもとに「脳２１」掲載の文章が書かれた。

図1

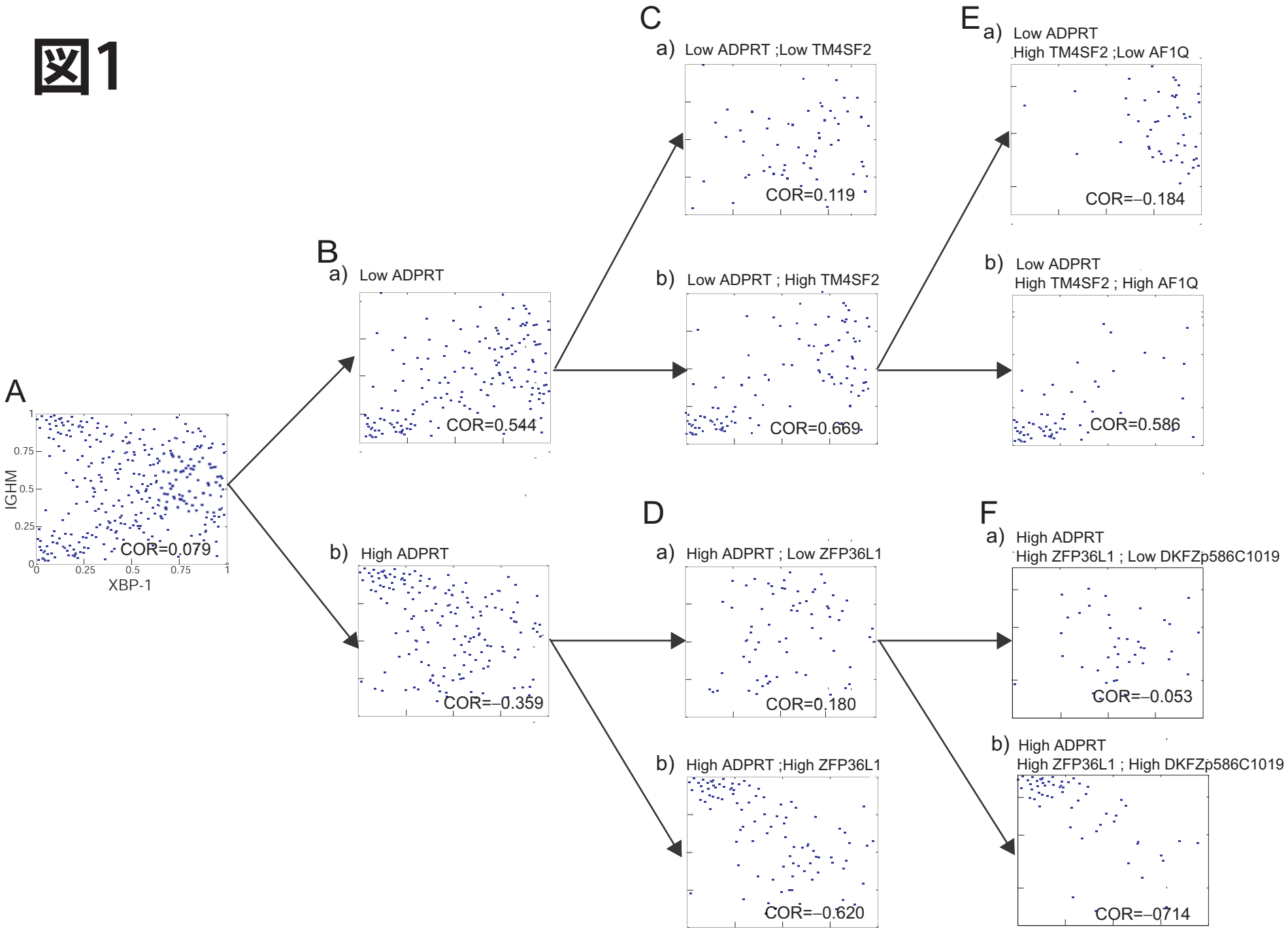
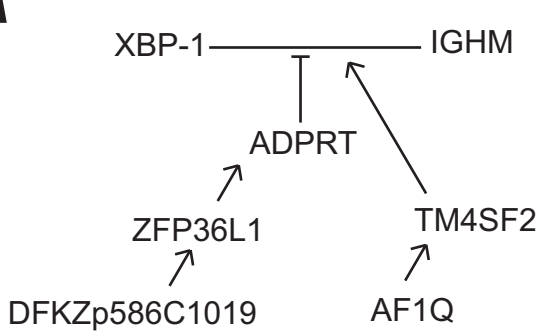
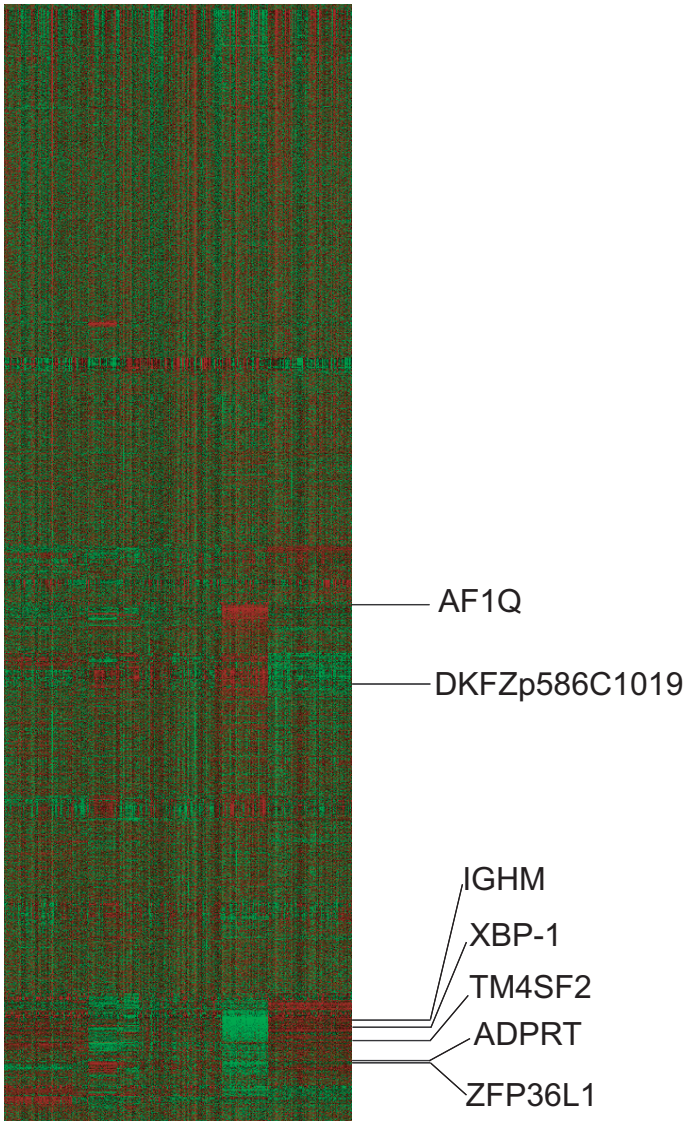

A
IGHM
XBP-1
COR=0.079
B
a) Low ADPRT
COR=0.544
b) High ADPRT
COR=–0.359
C
a) Low ADPRT ;Low TM4SF2
COR=0.119
b) Low ADPRT ; High TM4SF2
COR=0.669
D
a) High ADPRT ; Low ZFP36L1
COR=0.180
b) High ADPRT ;High ZFP36L1
COR=–0.620
E
a) Low ADPRT
High TM4SF2 ;Low AF1Q
COR=–0.184
b) Low ADPRT
High TM4SF2 ; High AF1Q
COR=0.586
F
a) High ADPRT
High ZFP36L1 ; Low DKFZp586C1019
COR=–0.053
b) High ADPRT
High ZFP36L1 ; High DKFZp586C1019
COR=–0714

図2

A


XBP-1
IGHM
ADPRT
ZFP36L1
TM4SF2
DFKZp586C1019
AF1Q

B


AF1Q
DKFZp586C1019
IGHM
XBP-1
TM4SF2
ADPRT
ZFP36L1