遺伝子発現量データ解析の基礎

富山大学和漢医薬学総合研究所　情報科学分野
奥　牧人

概要：漢方薬の複雑な作用機序を解明するためには，これまでにない新たなアプローチに積極的に取り組むと同時に，過去より受け継がれてきた基礎的な方法論も正しく理解し，必要に応じて最適な方法を選択出来ることが望ましい．本稿では，筆者が以前学生の教育用に作った資料を元に，DNA マイクロアレイにより計測された遺伝子発現量データに関する基礎的な解析法や可視化法について説明する．

1. はじめに

漢方医学は，しばしば中国の伝統医学と混同されるが，それが日本に渡り独自の発展を遂げてきたものである．名前の由来は，江戸時代にオランダから西洋医学が入ってきた際，それと区別するために元からあった医学体系を漢方と称するようになったと言われている．漢方薬は複数の生薬（植物，鉱物，動物の薬用部分）を組み合わせたものであり，多数の天然化合物を含んでいる．漢方薬は現在日本国内で広く使用されており，日本漢方生薬製剤協会による 2011 年の調査では，約 90 %の医師が漢方薬を使用していると回答している[1]．国としての承認や基準作りも進んでおり，医療用医薬品として 148 処方が薬価基準に，一般用医薬品として 294 処方が一般用漢方製剤製造販売承認基準にそれぞれ定められている．

漢方薬の複雑な作用機序を解明するため，これまでに多くの研究がなされてきたが，まだ十分には分かっていない．そこで，従来と異なる新たなアプローチとして，生態学や気候学といった他分野で開発が進んでいる統計的手法を生命科学へ移入し，漢方薬の複雑な作用機序の一端を明らかにしようというプロジェクトが立ち上がり，筆者も主要メンバーとして関わってきた．しかし，そこで筆者が強く感じたのは，新規手法を適用する前に通常の方法論が十分に試されていないのではないかという懸念である．その理由の一つとして，DNA マイクロアレイデータの基本的な解析によって何がどこまで分かるのか，多くの人にとってイメージしづらいことが挙げられる．

そこで本稿では，DNA マイクロアレイにより計測された遺伝子発現量データに関する基礎的な解析法や可視化法について説明する．対象は，自ら遺伝子発現量データ解析に挑戦したい学生や研究員に加え，自分で解析するつもりはないが，論文等で時折見かける生命情報学関係のグラフや表の意味をより深く知りたい，という方などを想定している．

2. 実験条件の確認と他の測定項目のプロット

遺伝子発現量データ解析で最初にすべきことは，実験条件および各サンプルの意味の確認である．意味不明の呪文だと思って解析を続行してしまうと，後でとんでもない取り違えを引き起こす危険がある．特に，何と発音したら良いか分からない単語があるとミスを引き起こしやすいので，最初に調べておく．

サンプルの意味の確認が済んだら，遺伝子発現量データ以外の測定項目を先にプロットする．このとき，エラーバーの種類に注意する．一般に，

標本標準偏差，標準誤差（標本標準偏差を$\sqrt{n}$で割ったもの），95 %信頼区間（標準誤差を約 2 倍したもの）の 3 種が使われている（図 1）．信頼区間は有意差について部分的な情報を持っている．具体的には，2 つの信頼区間が重なっている場合は有意差があるかどうか調べてみないと分からないが，重なりが無い場合は必ず有意差がある．

データのプロットは，従来折れ線グラフや棒グラフ，箱ひげ図などが用いられてきた．箱ひげ図の見方については統計学の教科書等に書いてあるはずなので説明は割愛する．一方，近年図 2 の右 2 つのような新しい描画法が出てきた．それぞれ蜂群図，バイオリン図と呼ぶ．

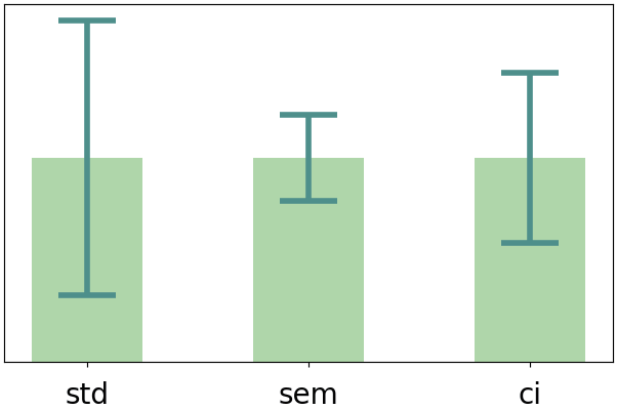

図 1．エラーバーの種類による違い．左から順に標本標準偏差，標準誤差，95 %信頼区間を同一のデータ (n=10) に対し表示している．

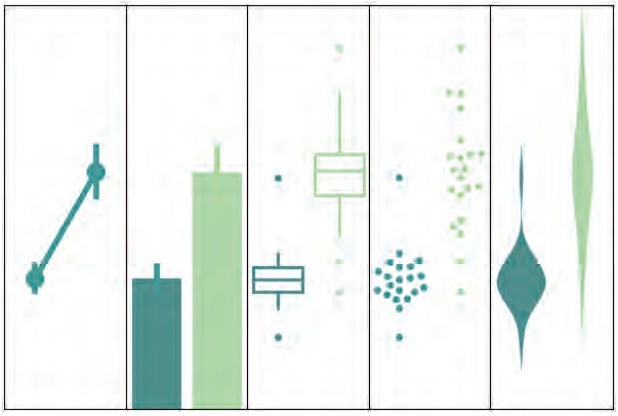

図 2. 同一のデータに対する異なる描画法の比較．左から順に折れ線グラフ，棒グラフ，箱ひげ図，蜂群図，バイオリン図を表す．

これらの描画法が登場した背景には，元データの分布が持つ情報の一部が従来のグラフでは失われているという問題意識がある．しかし，データが正規分布に従うと見なせる場合には折れ線グラフや棒グラフが簡潔で分かりやすく，そうでない場合でも，多数のデータを並べて比較する際は横線の明確な箱ひげ図が見やすい．従って，データと解析目的に応じて適切なものを選択するのが良いと考えられる．単にかっこいいからという理由だけで新しいものを選んではいけない．

3. データの前処理

データの前処理は最も手間のかかる工程である．何故なら，自動化や定型化が困難で人間が個別に判断する必要のある処理が多く含まれるからである．細かい注意点やノウハウを挙げ出したらページ数が足りないため，本稿では概説に留める．

DNA マイクロアレイデータの前処理は, 主に遺伝子の ID 変換，欠損値の処理，グローバル正規化，対数変換などから成る．これらの順番を入れ替えるとその後の結果が大きく変わるが，何が正しい順番かは筆者の知る限り決まっていない．

遺伝子の ID 変換では，表 1 に示す主な ID の種類を覚えておく必要がある．これらは遺伝子やそ

表 1．遺伝子や転写産物等を表す主な ID の種類．

ID の種類	例
遺伝子記号	*Tnf*
フルネーム	Tumor necrosis factor
Entrez	21926
Ensembl	ENSMUSG00000024401
RefSeq	NM_013693
UniProt	P06804
Affymetrix	1419607_at
Agilent	A_51_P385099

の転写産物，さらにそれを翻訳して作られるタンパク質，マイクロアレイのベンダーが設定したプローブ名などを含む．ID 変換をする際は，図 3 に示すような問題が生じるため，ID の意味を踏まえた上で対処法を個別に検討する必要がある．例えば，プローブ名から遺伝子記号への変換ではノンコーディングの部分の変換先が無い．それを解析から除外すべきか，それともプローブ名のまま残すべきかは解析の目的による．遺伝子から転写産物への変換ではスプライシングバリアントがあるため ID が増えるが，単に変換先の ID の昔の呼び名が後方互換のため併記されているだけの場合でも同様のことが起こる．従って，全てを残すか一つだけを残すかは状況による．プローブ名から転写産物への変換では，複数のプローブが単一の mRNA の異なる領域を担当している場合があり，その場合は最大値を取るのが適切だと一般に考えられている．しかし，それ以外のケースでは平均値にした方が良い場合もあるだろう．

データの前処理に関してもう一つ説明を要する用語がグローバル正規化である．これはサンプル間の分布のズレを補正する処理である（図 4）．通常，GEO 等のデータベースで公開されている遺伝子発現量データは，個々のサンプルに関しては既に推奨された統計的補正がかけてある．しかし，異なるサンプル間のズレの補正は済んでいる場合といない場合があるため，後者の場合は追加の

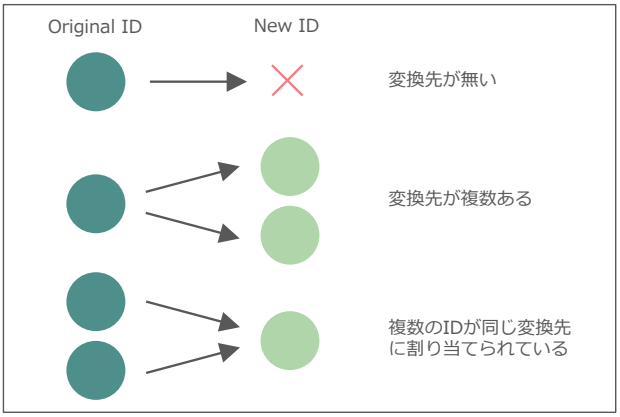

図 3．ID 変換時に生じる問題の例．

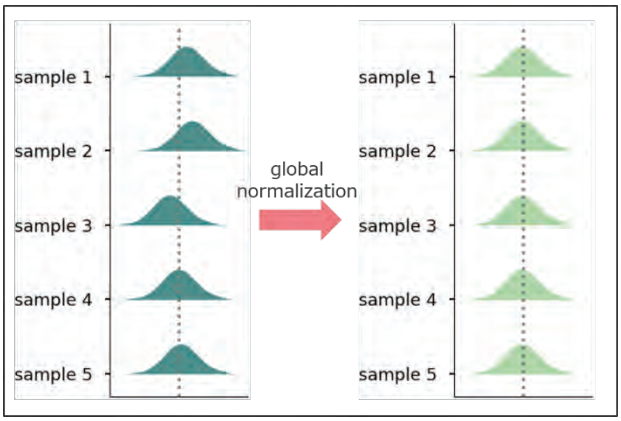

図 4．グローバル正規化の概念図．

正規化が必要となるのである．その背後にある考え方は，全ての遺伝子の発現量が一様に増加または減少するとは考えにくく，測定条件等による系統誤差と見なして問題ないだろうというものである．平均値または中央値を揃えるといった単純なやり方から，分布の形を全てのサンプルで完全に揃える quantile 正規化[2]などの複雑な手法まであり，データに合わせて適宜選択するのが良い．

4．データ全体の傾向把握

遺伝子発現量データの全体の傾向をつかむため，主成分分析（PCA）などの次元圧縮法がよく用いられる．これにより，多次元の膨大なデータの縮図が得られ，そこから様々なことを読み取ることが出来る．例えば，各サンプルが条件毎に分かれているかどうか，外れ値はないかなどが分かる．もしも直感に反するプロットが得られた場合は，それ以前の作業工程で何かミスは無かったか，用いた前処理法は本当にそのデータに対して適切なものであったか，戻って再検討すべきである．

図 5 に代表的な次元圧縮法のプロットを示す．始めは最も単純な主成分分析を使い，それでうまく傾向が捉えられない場合に限って，より複雑な多次元尺度法（MDS）や t-SNE 法[3]などを順次試すのが良いと考えらえる．何故なら，複雑な方法

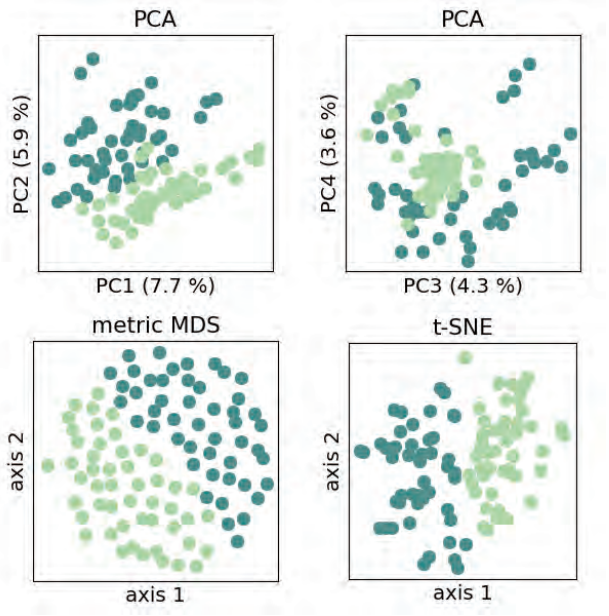

図 5. 同一のデータに対する異なる次元圧縮法の比較．上段は主成分分析，左下は計量多次元尺度法，右下は t-SNE 法の結果をそれぞれ表す．各点はサンプルを，色は実験条件を表す．

は決して上位互換ではなく，疑似乱数の出方次第で結果が異なる上，解析者が任意に調節可能な多数のパラメータを持っているため，どうしても結果の恣意性の問題が出てきてしまうからである．これを防ぐためには，疑似乱数の種を変えたり他の方法やパラメータを使ったりしても同様の結果が得られていることを確認する必要がある．

主成分分析では，各主成分が重要な順に出力されるため，それらの寄与率（データ全体の分散のうち，各主成分がどれだけの割合を説明するか）が通常重要視される．一方，多次元尺度法や t-SNE 法では各軸の重要性に優劣は無い．何故かと言うと，これらの手法は高次元のデータを互いの距離関係をなるべく保ったまま 2 次元平面に写像することを目的としており，どの点とどの点が近いか，どれとどれが離れているかという情報のみが考慮されているからである．従って，たとえプロット全体を時計回りに回転したとしても，点同士の距離関係は変わらないので問題ない．

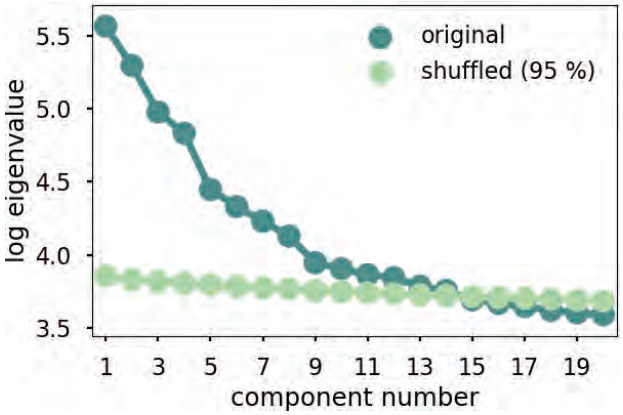

図 6. 主成分分析の有効成分数の推定．元データとそれをシャッフルしたものを比較している．

主成分分析に話を戻すと，何番目までの主成分が重要かをデータから推定する方法があるので，確かめておくと安心である．ここでの目的はあくまでもデータ全体の傾向を捉えることなので，必ずしも重要と判定された主成分全てを調べる必要はないが，一つの目安にはなる．注意として，古典的な方法の中には使うべきでないとされるものも多い[4]（1 より大きい固有値の数や，寄与率の総和が一定値を超えるまで，目視による scree プロットの変化点など）．信頼のおける方法の一つは，元データと同じサイズのランダムデータを複数用意し，それらを PCA にかけた結果と元データの結果とを比較する平行分析である（図 6）．元データの固有値がランダムデータのものより有意に上側にある主成分を重要なものとする．ランダムデータは，元データをサンプル毎にシャッフルしたものなどを用いる．狭義の平行分析では標準正規分布に従う独立な疑似乱数を使用するが，ここでは広義の意味で呼んでいる．

5. 発現変動遺伝子の取得

発現変動遺伝子（differentially expressed gene, DEG）とは，異なる条件やグループ間の比較において発現量が大きく上昇または減少した遺伝子のこと

である．発現変動遺伝子の同定には大きく 2 通りの方法がある．一つは倍率変化による選別であり，もう一つは仮説検定に基づく判定である．

倍率変化による選別では，一方の発現量の値が他方と比べて 2 倍より大きく増えたか, 或いは 1/2 倍より小さくなった場合に, DEG と判定することが多い．この慣習的に用いられる閾値に何ら根拠は無く，4 倍と 1/4 倍など，他の値に設定しても構わない．比較したい 2 つのグループのそれぞれに複数のサンプルがある場合は，発現量の対数を取ったものの平均値を用いる．先に算術平均をとってから対数変換するのではない．

仮説検定の方は, DNA マイクロアレイのデータに限った話だが，対数変換するとほぼ正規分布になるため，一般的に t 検定が用いられる．一口に t 検定と言っても様々な種類があるが，通常は Welch の t 検定を両側検定で使うことが多いだろう（Welch の, と言った時点で, 2 群比較, 独立性, 異分散性を仮定したことを意味する）．

ヒトやマウスの遺伝子は 2 万個以上あるので，各遺伝子について t 検定を繰り返せば，当然多重比較のための特別な処置が必要となってくる．通常の実験研究では Bonferroni 補正や Tukey-Kramer 法などが使われるが，遺伝子発現量データ解析では Benjamini-Hochberg 法（BH 法）による FDR 制御がよく用いられる．これについて以下で少し詳しく説明する．

まず，基本用語の意味についておさらいする．統計量とは，予め定められた手順によって観測データから算出される数値である．例えば t 値などである．帰無仮説が真のときに統計量の値が従う分布を考える．このとき，ある観測データに基づく統計量の値に対して，その値またはそれより極端な値が出る確率の総和を p 値と呼ぶ（図 7）．そして，p 値が予め設定しておいた有意水準 α 以下だった場合に帰無仮説が棄却される．これは t 検定では有意差があることを意味する．差があることを陽性（P），無いことを陰性（N）と表すと，表 2 に示す混同行列が書ける．

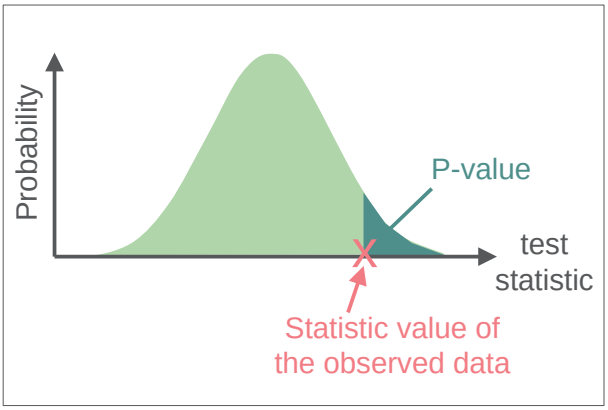

図 7．p 値の概念図（片側検定の場合）．×印が統計量の値，濃い色の部分の面積が p 値を表す．

表 2．混同行列．真陽性（TP），偽陰性（FN），偽陽性（FP），真陰性（TN）の 4 つの場合がある．

	P (predicted)	N (predicted)
P (actual)	TP	FN
N (actual)	FP	TN

多重比較における一番の問題は, FP が過度に増えることである．例えば，1 万個の遺伝子があって，そのうち真に発現変動を起こしていると予想される遺伝子が 500 個程度だったとする（図 8）．残りは全て陰性のはずだが，通常の有意水準 0.05 で判定すると，そのうちの 5 %は陽性として出てくる．その数は約 500 個にものぼる．本当に差がある 500 個が仮に全て正しく陽性と判定出来たとしても, 合計で 1000 個ほど出てくる DEG のうち，約半分は偽物ということになる．

Bonferroni 法は，有意水準 α を多重比較数 n で割ることで，FP の増加を抑える．しかし，この操作は同時に，真に陽性である遺伝子を検出することも困難にし, FN の増加を引き起こしてしまう．特に n が大きいほどこのデメリットは大きい．そこで，FP と FN をバランスよく抑えるための新た

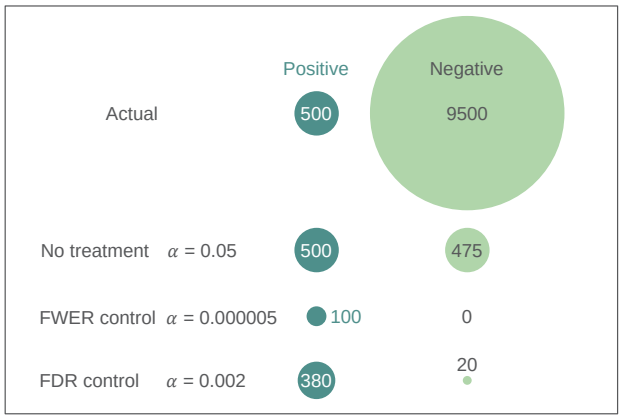

図 8. 多重比較における問題とその対処法の概念図．1 行目が真の内訳，下 3 行が各枠組みの中で陽性と判定された遺伝子の内訳を表す．

な枠組みとして登場したのが FDR 制御という考え方であり，それを実現するための手順の一つが BH 法である．FDR 制御に対し，元からあった考え方を FWER 制御と呼ぶ．Bonferroni 法は FWER 制御を実現する手順の一つである．

両者の違いを図 8 に示す．FWER 制御では DEG と判定された遺伝子の中に一つ以上の FP が混入する確率を 5 %以下に抑える．言い換えると 95 %の確率で DEG は TP のみから構成される．一方，FDR 制御では DEG の中に含まれる FP の割合を 5 %以下に抑えることを目標とする．つまり，毎回ある程度は FP が混じっていることになる．

BH 法の詳細な説明は割愛するが，その解釈について注意点を述べる．まず，補正するのは p 値ではなく有意水準 α の方である．また，個々の比較毎に異なる補正係数を乗じた p 値や有意水準を用いるのではなく，実質的に全ての比較で同じ値の α を使用していることに相当する．

つまり，まず初めに許容される FP 混入率（通常 5 %）を解析者が指定すると，それに応じて最適な有意水準の値がデータから算出され，p 値がそれ以下の遺伝子が陽性と判定され出力される．その有意水準の値とは，陽性と判定された遺伝子の p 値の中で最大のものに他ならない．

6. 発現変動遺伝子リスト間の重複度合の確認

複数の発現変動遺伝子リストが得られている時はベン図を描くことが多い（図 9）．これにより，異なるリスト間で共通する遺伝子や，片方のみに含まれる遺伝子の個数などが把握しやすくなる．しかし，ベン図が有効な集合の個数はせいぜい 4 個が限界である．何故なら，ベン図中の領域の数は集合の個数に対して 2 のべき乗で増えていくため，集合の数が多くなるとベン図から意味のある情報を読み取れなくなるからである．

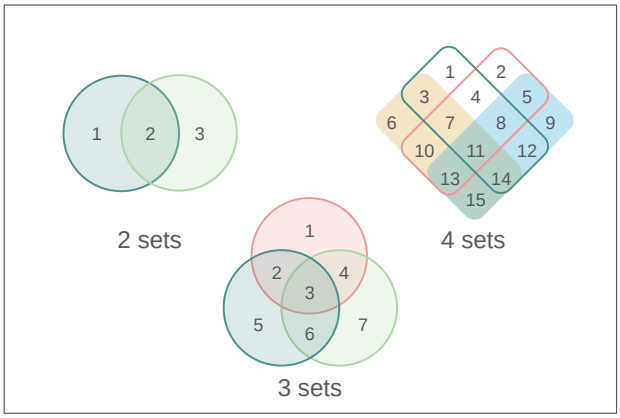

図 9. ベン図の例．実際に使う際は，各領域に含まれる遺伝子数を調べて記入する．

7. クラスタリング

クラスタリングとは，似たようなデータ点をグループ別にまとめる手法である．遺伝子発現量データ解析で頻繁に使用される．例えば時系列データの場合，時間経過とともに発現量が上昇する遺伝子グループ，徐々に減少するグループ，ほぼ一定値のグループなどに切り分けることが出来る．

クラスタリングには様々な種類があるが，遺伝子発現量データ解析では一般に凝集型の階層的クラスタリングがよく用いられている．凝集型でない階層的クラスタリング（分割型と呼ばれる）はほとんど用いられないため，以降では単に階層

的クラスタリングと呼ぶ.

階層的クラスタリングの前処理として，遺伝子毎に z 変換をかける場合が多い．その理由は，そうしない場合に，平均的な発現量の高い遺伝子同士，低い遺伝子同士がそれぞれクラスタを作り，それ以外の観点による分類（例えば，投薬により発現量が徐々に上昇するグループと下降するグループなど）が出来なくなってしまうからである.

階層的クラスタリングには 3 つの設定項目がある. 1 つ目の設定項目は類似度・非類似度である. 類似度とは似ているほど値が大きくなる指標，非類似度とは似ていないほど値が大きくなる指標のことである．類似度は非類似度に変換して使用する．非類似度のうち，三角不等式など幾つかの条件を満たすものを特に距離と呼ぶ．表 3 に代表的な類似度と非類似度をまとめるが，筆者の経験上，ユークリッド距離か相関係数のいずれかを使っておけば大体問題ない.

2 つ目の設定項目は連結法である. 連結法とは，クラスタ間の非類似度をどう決めるかを表す．図 10 に代表的な 3 つの連結法を示すが，遺伝子発現量データ解析において単連結法は使うべきでない．何故なら，chaining という現象[5]が発生し，データが本来有するクラスタ構造が得られないからである（図 11）.

3 つ目の設定項目は樹形図の分割基準である. 階層的クラスタリングでは，樹形図をある閾値で切断することにより最終的なクラスタ分けを行う（図 12）．閾値が高いと少数の大きなクラスタが得られ，低いと細かなクラスタが多数出てくる. 閾値を直接指定する場合もあれば，指定したクラスタ数となるように分割する場合もある．何をもって最適な分割と呼べるかはデータや解析目的により異なるため，人間が樹形図やヒートマップを目視で確認してその都度分割基準を決めるのが良いだろう.

階層的クラスタリングの結果をヒートマップで表すとき，赤から緑へ変化するカラースケール

表 3. 主な類似度と非類似度.

名前	種別
ユークリッド距離，L2 ノルム	距離
マンハッタン距離，L1 ノルム	距離
チェビシェフ距離，最大値ノルム	距離
マハラノビス距離	距離
相関係数	類似度
コサイン類似度	類似度

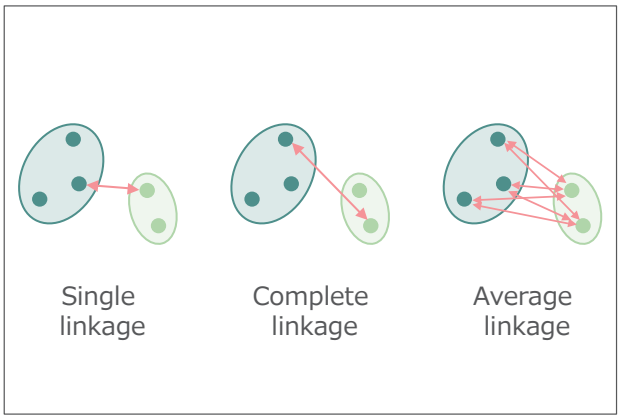

図 10. 代表的な連結法の概念図. 左から順に単連結法，完全連結法，平均連結法を表す．点同士の非類似度の最小，最大，平均値をそれぞれ用いる.

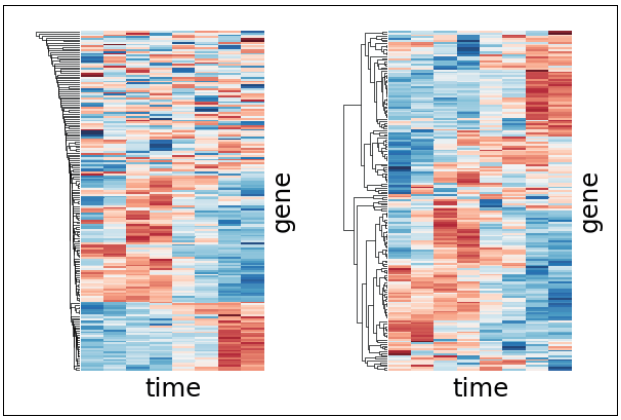

図 11.（左）単連結法による失敗例. Chaining 現象により，上側 3 分の 1 程度が構造化出来ていない.（右）同じデータに平均連結法を適用したもの.

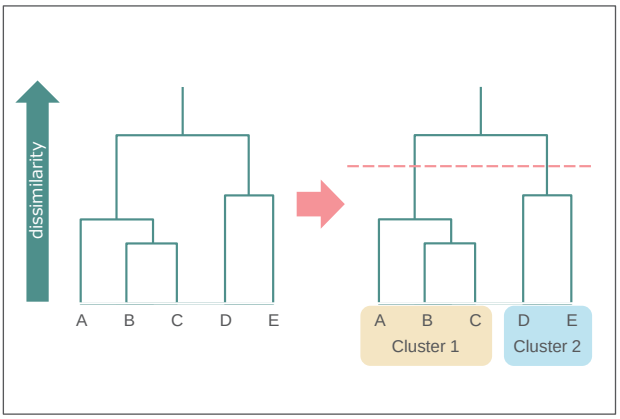

図 12. クラスタ分割の概念図.

を使ってはいけない．何故なら，色覚異常の人が区別できないからである．昔の論文では広く使われていたが，それらを真似しようとすべきでない．代わりに，赤と青，マゼンタと緑などの組み合わせを使う．

8. エンリッチメント解析

エンリッチメント解析とは，ある遺伝子リストがあったとき，その中にどのようなタイプの遺伝子が多く含まれているかを調べるものである．例えば，100 個の発現変動遺伝子をこの解析にかけると，その中に炎症反応に関わる遺伝子が 30 個含まれていたとか，インスリンのシグナル伝達経路に関わる遺伝子が 10 個いたとか，などの情報がまとめて列挙される．

個々の遺伝子には，それが生体内でどのような機能や役割を果たしているかを表すためのタグが複数付けられている．英語では annotation と呼ぶため注釈と訳すこともある．表記ゆれを防ぐため，タグとして使って良いものは予め決まっている．遺伝子発現量データ解析でよく使うのは，GO（gene ontology）タグと，KEGG パスウェイのタグである（表 4）.

GO タグは大きく 3 つのグループに分類されているが，通常は BP（biological process）グループ

表 4. 遺伝子に付いているタグの例.

タグ	種別
inflammatory response	GO (BP)
macrophage cytokine production	GO (BP)
insulin signaling pathway	KEGG
cell cycle	KEGG

に属するタグだけ調べれば十分だろう．何故なら，残りの 2 グループは結果の解釈が難しいからである．GO タグには大まかな内容のタグから詳細なタグまで様々なものがある．あまりにも内容が漠然としたものや逆に細か過ぎるタグはエンリッチメント解析の役に立たないので，有用な GO タグだけを集めたサブセットも幾つか考案されている．有名なものでは，GO slim や GO FAT などがある.

KEGG のタグの方は，各遺伝子がどの経路に関わっているかを表す．GO タグと内容的に重複する部分も大きいが，多くの遺伝子が関わる経路が見つかったら，それらが経路上のどこに位置するかを図示出来るという特徴がある（図 13）.

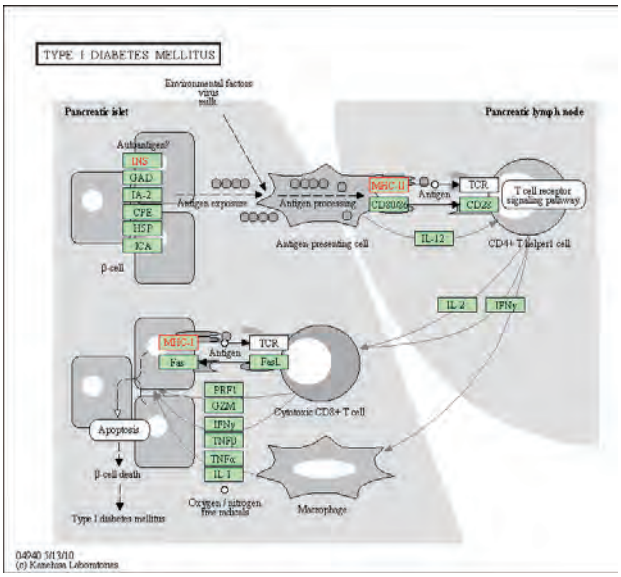

図 13. KEGG パスウェイの例．注目している遺伝子を図中でハイライト表示することが出来る.

エンリッチメント解析では，解析対象の遺伝子集合と，あるタグを持つ遺伝子集合とを比較し，有意な重複があるかどうかを調べる（図 14）．これを全てのタグに対して繰り返し行い，有意だったタグの一覧を出力する．有意性の判定にはフィッシャーの正確検定を使う（図 15）．これは，もしも 2 つの集合が互いに無関係だった場合に偶然起こり得る重複度合に対して，実際の重複度合が有意に大きいかどうかを調べるものである．

エンリッチメント解析の結果を図示する際，慣習的に p 値の対数を取って符号を反転したものを棒グラフで表すことが多い．しかし，そもそも p 値の大小は比較すべきものではないので，そのような慣行の蔓延に筆者は強い懸念を抱いている．

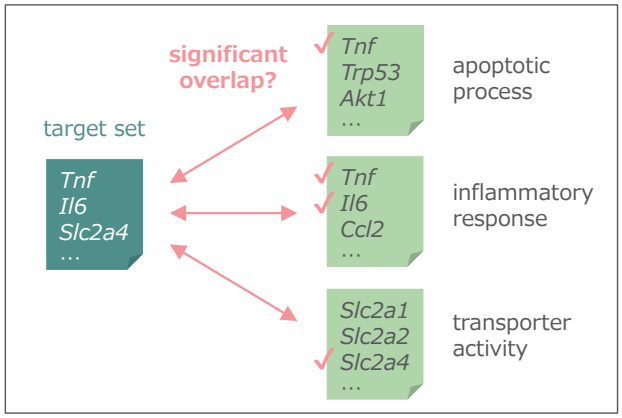

図 14. エンリッチメント解析の概念図.

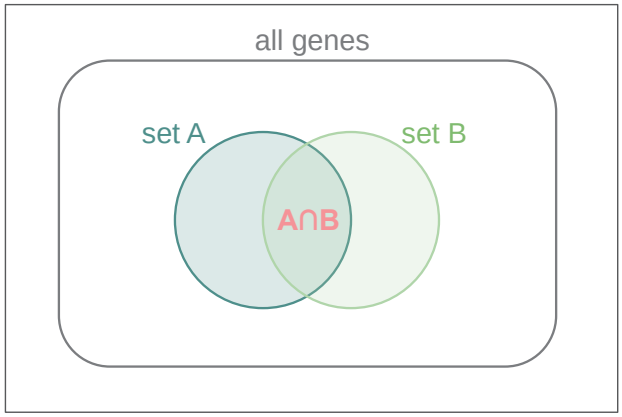

図 15. フィッシャーの正確検定の概念図．比較する 2 つの集合のそれぞれのサイズ, 重複の大きさ, 全遺伝子数の計 4 つの数値を用いる.

9. ネットワーク図の描画

ネットワーク図は見た目が派手なので，読者に強い印象を与えることを第一の目的とした作成の依頼を受けることも少なくない．しかし，研究における全てのグラフについて言えることだが，その図から何が読み取れるのか，読者に伝えたいメッセージとは何かが明確でない図に研究上の価値は無い．ネットワーク可視化の分野では，各頂点のラベルを表示しない場合，頂点数が約 1000 個を超えるものは一枚絵として描画しても仕方がないと一般的に考えられている．何故なら，頂点や枝が過度に折り重なり，何も構造が見て取れなくなるからである．「巨大な毛玉」と揶揄されることもある．一方，各頂点のラベルを表示する場合，描画可能なネットワークのサイズはさらに小さくなる．何故なら，ラベルの文字が判読可能である必要があるからである．筆者の経験上，せいぜい数十頂点が限界だろう．そのサイズの図を出して一体何を伝えたいのか，必ずはっきりさせるべきである．

ネットワークの取得法には大きく 2 通りある．一つは遺伝子発現量データから各遺伝子間の相関係数を計算し，その絶対値がある閾値以上の場合に枝を張るというものである．もう一つはデータベースで検索を行い，相互作用があることが知られている遺伝子間や，相互作用があると予測されている遺伝子間に枝を張る方法である（図 16）．

ネットワーク内に複数のサブクラスタが形成されている場合は，コミュニティ検出が有効である．コミュニティとはネットワーク解析の用語で，比較的密に結合している部分グラフのことを指す．クラスタと読み替えても問題ない．検出には Louvain 法や infomap 法などがよく用いられる．

ネットワークの可視化では，各頂点の大きさ，形，色に何を対応させるのか，各枝の太さや色に何を割り当てるのかを決める必要がある（図 16）．

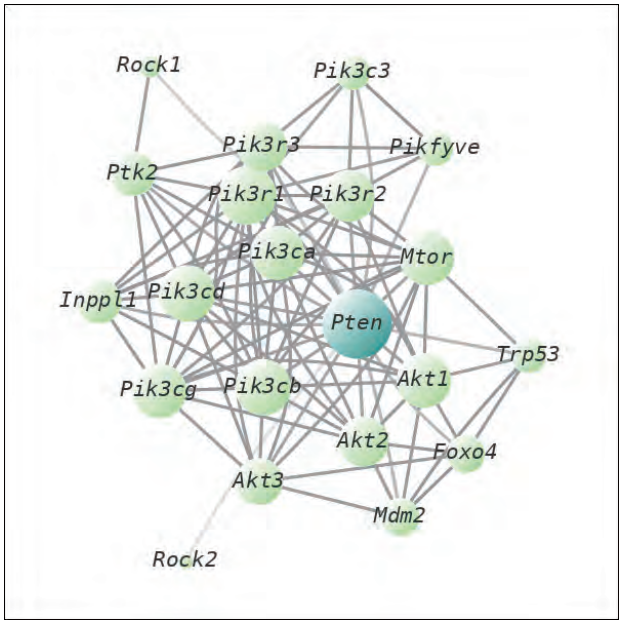

図 16. データベースを検索し *Pten* と関連性の強い上位 20 個の遺伝子のネットワークを取得した結果．頂点の大きさには次数を，枝の輝度には媒介中心性をそれぞれ割り当てている．

また，頂点の配置をグラフ構造から自動で計算する手法も幾つかあるので，それらの中から適宜データと目的に合ったものを選ぶ．

10. まとめ

本稿では DNA マイクロアレイデータ解析の基礎について説明した．全体を通して言えることは，多くの作業工程において正解と呼べる手順はいまだ確立しておらず，解析者は日々難しい選択を迫られているということである．本稿が判断の一助となれば幸いである．また，本稿で紹介した基礎的な手法が漢方薬の複雑な作用機序の解明に役立つことを願う．

本稿で記述した内容の少なくとも一部は，同じトランスクリプトームを対象とした次世代シーケンサーの RNA-seq データにも適用可能と考えられる．また，基本的な考え方はプロテオーム，メタボローム，マイクロバイオームなど他のオミクスデータにもある程度通じる部分があると期待される．

図 5，6，11 では公開データ GSE2565 を使用した．図 13 では KEGG データベースを，図 16 では STRING データベースをそれぞれ利用した．残りは人工的に作成したデータを用いた．

謝辞

本研究は富山大学のプロジェクト「医薬学と複雑系数理学からの挑戦　～「未病」の解明、そして新たな医療体系の構築と地域との連携による健康人口の増加～」（通称，未病プロジェクト）関係者の先生方からの多大なご支援を頂きながら進めております．附属病院長でプロジェクトリーダーをされている齋藤滋先生を始め，プロジェクトの中心的メンバーとしてご活躍されている和漢医薬学総合研究所の門脇真先生，小泉桂一先生，林周作先生，都市デザイン学部の春木孝之先生，人間発達科学部の成行泰裕先生に深く感謝致します．また，本プロジェクトの発足以前より継続して指導頂いている東京大学生産技術研究所の合原一幸先生に厚く御礼申し上げます．

参考文献

[1] 日本漢方生薬製剤協会，漢方薬処方実態調査 2011, http://www.nikkankyo.org/serv/serv1.htm

[2] BM Bolstad, et al., *Bioinform.*, 19(2):185-193 (2003).

[3] L van der Maaten and G Hinton, *JMLR*, 9:2579-2605 (2008).

[4] DA Jackson, *Ecology*, 74(8):2204-2214 (1993).

[5] 神嶌，人工知能学会誌, 18(1):59-65 (2003).

（2018 年 4 月，和漢医薬学総合研究所年報の総説として執筆）