プロジェクト・バイオ

多変量解析に一石を投じる―ConfeitoGUIplusの開発

萬年　一斗[1]・尾形　善之[2]・鈴木　秀幸[1]*
([1]（公財）かずさDNA研究所生体物質解析センター・[2]大阪府立大学大学院生命環境科学研究科）

近年，種々の分析機器のハイスループット化に伴い，いわゆるビッグデータと呼ばれるものが氾濫している．これらのデータを有効活用するためには，多変量解析は最早必須であると言っても過言ではない．今後，さらに多くのビッグデータが生み出されることを想定すると，複数の分析機器のデータ，特に異種データの統合解析が重要な役割を担う時代が遠からず訪れるであろう．たとえば，トランスクリプトームとメタボロームのデータを同時に処理し，挙動の類似性に着目すれば，未知の代謝経路で働く遺伝子と代謝産物がペアで予測できるようになるかもしれない．来るべき時代に備え，異種データの統合解析を自在に行うためのプラットフォームの整備は急務であると言える．

このようなプラットフォームとして，筆者らは多変量解析の一種である相関ネットワーク解析に光明を見いだし，独自の解析ソフトウェアConfeitoGUIplusの開発に取り組んできた．相関ネットワーク解析では，解析対象を点（ノード）で表し，相関のある点同士を線（エッジ）で結んで，ネットワーク状に関係性が表現される．ConfeitoGUIplusは，形成されたネットワークをさらに幾何学観点から精査することで，偽陽性・偽陰性を回避する工夫がなされている．これらの工夫により，先の例で言えば，遺伝子や代謝産物がノードとして同一ネットワーク中に表現されることになる．

アルゴリズムのプロトタイプはすでに発表していたものの[1]，今後起こり得るさまざまな用途を想定し，求められる3つの特徴（①入力データの自由度，②直感的なパラメーター設定，③許容幅のある関係性判定）を兼ね備えたソフトウェアにするため，アルゴリズムの練り直しを余儀なくされ，実装に長い年月を要した．本稿では，アルゴリズムの紹介とともに，上記3つの特徴をどのように実現させたのかについて述べたい．

入力ファイル形式とデータ構造

ConfeitoGUIplusは，入力ファイル形式としてタブ区切りテキストを採用している．極力プレーンな形式とすることで，各種分析機器メーカーの垣根を越えた入力データを容易に作成することを可能にした．ファイルの中身は，第1行にデータセット名（表1のA，Bなどに相当），第1列に要素名（表1のID0001，ID0002などに相当）を配した定量値テーブルであれば，機器分析結果，表現型の計測値，食味試験結果など，どういったものでも構わない．

表1．入力データ構造（Qは各々の定量値を表す）

	A	B	C	‥
ID0001	Q1	Q2	Q3	‥
ID0002	Q4	Q5	Q6	‥
ID0003	Q7	Q8	Q9	‥
⋮	⋮	⋮	⋮	

入力データを読み込むと，定量値の変動をもとに要素間の相関係数が算出され，相関係数の降順にソートされた相関係数リスト（図1）が全要素の数だけ自動生成される．

False-Positive Out（FPO）解析

続いて，リストの上位に位置する要素を仮のエッジで結び，モジュールの仮組みが行われる．この仮組みのモジュールから偽陽性（false-positive）の要素を取り除く（out）工程をFPO解析と呼んでいる．仮組みに用いられる要素数は，解析開始時にユーザーが設定する予想モジュールサイズ（詳細は後述）によって規定される．また，偽陽性の判定には，ノードとノードのつながり方に着目した2つの評価指標を用いる．

要素の評価指標　モジュールの構成要素iを評価する指標Vertex F-measure（$VF(i)$）は，Vertex Density（$VD(i)$）とVertex Specificity（$VS(i)$）を用いて，以下のように定義される．

$$VF(i)=\frac{1}{(1/VD(i)+1/VS(i))/2}$$

$$VD(i)=\frac{e(i)}{n-1},\quad VS(i)=\frac{e(i)}{d(i)}$$

（n：モジュールの要素数，e：エッジ数，d：次数）

著者連絡先　E-mail: hsuzuki@kazusa.or.jp　http://www.biosupport.kazusa.or.jp/sub_center3/

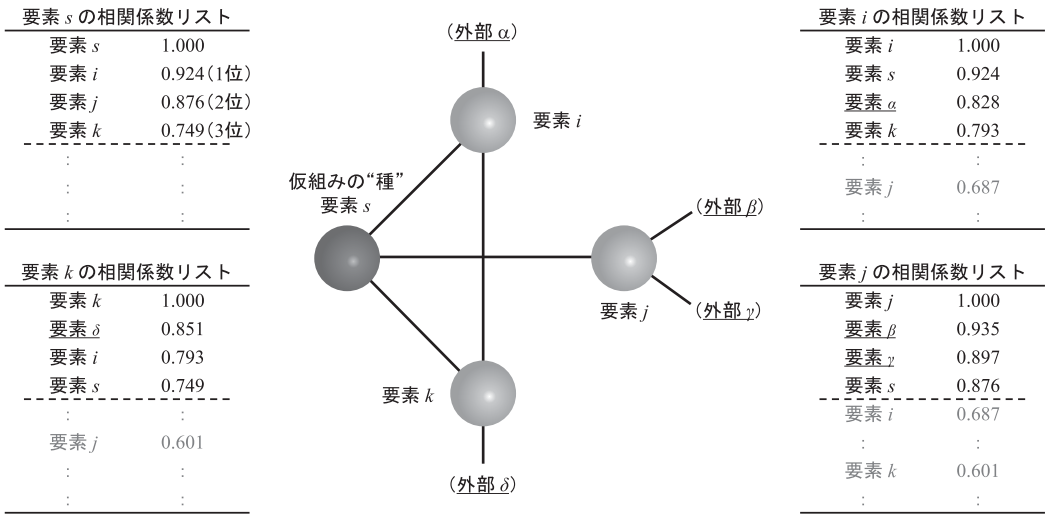

図1．相関係数リストとモジュールの内部・外部の概念

数式とにらみ合いをするよりも意味合いを捉えた方がわかりやすいと思う．解説の都合，順番を前後させて*VD*(*i*)から記述する．なお，【　】内には図1に示したモジュールにおける具体的な数値を併記する．

VD(*i*)は，エッジの密度を表し，要素*i*から伸びるエッジ数【2】を理論最大値【要素*i*以外のすべてのノードにエッジが伸びたと考えて3】で除して求められる．*VD*(*i*)の値が大きい要素ほど，他の要素との結びつきが強いと言える．

一方，*VS*(*i*)は，エッジの指向性を表現しており，モジュールの外部についても考慮に入れる必要がある．ここでの外部とは，*VS*(*i*)演算の対象とする要素の相関係数リスト内で上位に含まれる要素のうち，モジュールの構成要素以外の要素群を意味する．たとえば，図1の要素*i*は，外部にもエッジを1本伸ばしているので，内部のエッジ数【2】を内外のエッジの和【2 + 1】で除することで，*VS*(*i*)を求めることができる．この値の大小が，モジュール内外のどちらに要素を配置するのが適切かを決める判断材料となる．

VF(*i*)は，*VD*(*i*)と*VS*(*i*)の調和平均であり，両者の特性を併せ持つ指標となっている．すなわち，高い*VF*(*i*)を示す要素は，モジュール内部に強固なつながりを持ち，そのモジュールの構成要素として相応しい要素と言える．

モジュールの評価指標　　モジュール全体を評価する指標Network *F*-measure（*NF*）は，Network Density（*ND*）とNetwork Specificity（*NS*）を用いて，以下のように定義される．

$$NF = \frac{1}{(1/ND + 1/NS)/2}$$

$$ND = \frac{\sum e(i)}{n \cdot (n-1)}, \quad NS = \frac{\sum e(i)}{\sum d(i)}$$

（*n*：モジュールの要素数，*e*：エッジ数，*d*：次数）

*NF*は，*VF*(*i*)と同様，*ND*と*NS*を一括りに扱う指標であり，この値が大きいほど，偽陽性が少ない状態であることを示す．では，*ND*と*NS*はそれぞれ何を意味しているのだろうか．これらは，*VD*(*i*)と*VS*(*i*)の考え方をモジュール全体に拡張したものと捉えることができる．

つまり，モジュール全体として，結びつきが強いか弱いか（*ND*），外部とのつながりが多いか少ないか（*NS*）をそれぞれ表している．*ND*は，モジュール内部だけを見て，各ノードから伸びるエッジの総和【重複を除いて数えて4】を理論最大値【4つのノードを頂点とする図形の辺と対角線の和6】で除して求められる．一方，内外のエッジを加味し，内部の総和と内外の総和の比【分子･分母とも要素*s*から時計回りに見て(3 + 2 + 1 + 2)/{(3 + 0) + (2 + 1) + (1 + 2) + (2 + 1)}】をとると，*NS*を求めることができる．

仮組みモジュールの構築　　評価指標の紹介が終わったところで，改めてモジュールの仮組みの手順に触れておきたい．たとえば，予想モジュールサイズが5から50であった場合について考える．まず，仮組みの“種”にする要素の相関係数リストの5位までを内部として認識し，仮のエッジで結ぶ．次いで，ランクインした要素間に仮のエッジを結ぶか否かを判定するため，各要素の相関係数リストを参照し，“種”と5位の要素の相関係数の値よりも大きい相関係数を示す要素を上位として仮の

エッジで結んでいき，モジュールの仮組みを行う．

こうして仮組みされたモジュールに対して，*NF*と各構成要素の*VF*(*i*)を計算し，*VF*(*i*)が最小となる要素を偽陽性と判定，これを取り除いた残りに対して，再び*NF*と各構成要素の*VF*(*i*)を計算する．この最適化処理を予想モジュールサイズの下限+1になるまで繰り返し，*NF*が最大であった時点を過不足なく偽陽性が除かれた状態と判定し，その時点の仮組みモジュールを一時記憶する．

内部・上位の定義を変更し，“種”の相関係数リストの6位まで，7位まで……50位までといった具合で全46パターンの仮組みモジュールを想定し，同様な処理を実行する．最適化された状態をすべてのパターンで比較し，その中で最大の*NF*を示す仮組みモジュールを採用する．

仮組みモジュールの統合　ここまでの処理は，仮組みの“種”となり得る相関係数リストごとにパラレルに実行される．つまり，全要素の数だけ仮組みモジュールが構築され，モジュール間には要素の重複が生じていることになる．どんなに仮組みモジュールを最適化しても，要素が含まれるモジュールが一意に定まらないのでは，大きな偽陽性を見過ごしている状態にあると言える．

この問題を解消するため，モジュール化と呼ばれる工程が実行される．この工程においても，*VF*(*i*)が重要な指標となる．まず，仮に結んでいたエッジを外し，*VF*(*i*) ≥ 0.50の要素同士を改めてエッジで結ぶ．このとき，モジュール間に要素の重複があれば，その要素を介してモジュール同士の統合も行う（図2，黒・白のノードは要素の重複を表す）．同様な処理を0.50 ≤ *VF*(*i*) ≤ 0.99の範囲で閾値を0.01刻みで変化させて実行し，予想モジュールサイズの範囲に収まるモジュールの数が最大となるモジュール群を採用する．こうして重複を除いたモジュール群のエッジを相関係数リストに則ったルールで再描画し，最終的なFPO解析結果とする．

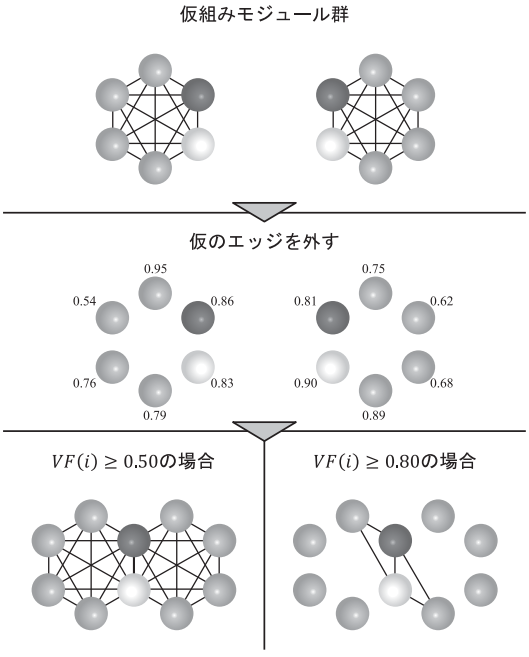

図2．*VF*(*i*)の閾値の違いによるモジュールの変化

このような遠回りに見える処理を行うことで，モジュールサイズの予想が外れた場合であっても，サイズ指定によるバイアスを最小限にした解析結果を得ることができる．これに加え，ユーザーは予想モジュールサイズ，言い換えれば，何種類の要素が協調的な変動をしているのかという直感的なパラメーターを入力するだけで解析が可能であり，利便性の高いインターフェースの恩恵を享受することができる．

False-Negative In（FNI）解析

前段のFPO解析で得られたモジュール群のうち，ユーザーが指定した1つのモジュールに対して，偽陰性（false-negative）の要素を取り込む（in）工程をFNI解析と呼んでいる．対象モジュールの構成要素を除く，すべての要素が取り込みの候補となり，判定には下記に示すMax *VS*(*i*)が用いられる．各要素の相関係数リストを上から順に見て，外部のエッジが出現する度に*VS*(*i*)を計算し直し，その中で最大となるものをMax *VS*(*i*)とする．解析開始時にユーザーが設定するMax *VS*(*i*)の閾値を超える要素を偽陰性と判定し，破線のエッジで対象モジュールに紐付けていく．これにより得られる実線・破線混じりのモジュールがFNI解析結果となる．

偽陰性判定には*VS*(*i*)のみが考慮されているため，比較的緩やかな条件と言えるが，*VF*(*i*)と*NF*で厳密な判定を行うFPO解析と組み合わせることで，関係性抽出に許容幅をもたらしている．こうした許容幅の導入が，FPO解析で退け合う可能性がある異種データ同士を関連付けるために重要な役割を担っている．

おわりに

紙面の都合上，解析実績の紹介はここでは割愛するが，その概要については2016年大会の講演要旨集[2]を参考にしていただきたい．当該大会の反響は大きく，多くの方から利用検討の申請をいただいている．解析実績数を伸ばし，いずれはConfeitoGUIplusを異種データ統合解析のデファクトスタンダードへと成長させていきたい．

文　献

1) Ogata, Y. *et al.*: *Genome Inform.*, **23**, 117 (2009).
2) 萬年一斗ら：日本生物工学会大会講演要旨集，p. 145 (2016).