Vol.2018-MPS-118 No.39
Vol.2018-BIO-54 No.39
2018/6/15

種固有の遺伝子共発現ネットワーク導出のためのトランスクリプトームスペースの形式化

大林武[†1]　青木裕一[†2]

概要：大量のトランスクリプトームデータに基づく遺伝子共発現情報（遺伝子発現プロファイルの類似性）は、遺伝子ネットワーク解析の基礎的な情報として、各生物種内の遺伝子ネットワーク解析のみならず、種間比較に基づく遺伝子ネットワークの進化に関しても研究されている。しかし、遺伝子共発現情報が所与の遺伝子発現プロファイルの要約という定義の元に計算される限り、トランスクリプトームデータの選択、データの補正、共発現関係の導出という複数の計算ステップの最適化を行うことはできず、種固有の共発現情報を定義することもできない。本報告ではトランスクリプトームデータにサンプリングバイアスが存在することを示し、共発現計算法において、このサンプリングバイアスを軽減するための手法を提案する。

キーワード：遺伝子ネットワーク，遺伝子共発現，トランスクリプトーム，進化

1. はじめに

細胞システムは多数の遺伝子の複雑な関係性によって構成されている。そのため、個々の遺伝子が担う分子機能に加えて、複数の遺伝子の関係性が作り出すパスウェイの理解が、細胞システム全体の鍵となる。遺伝子ネットワークはこのような遺伝子の関係性を可視化するものであり、遺伝子、パスウェイ、システム全体といった各レイヤーの構成と機能を考える上での強力なアプローチとなる。パスウェイを遺伝子群のなす機能単位と考えると、細胞はパスウェイ単位で発現調節していると想定できる。トランスクリプトームは、細胞機能の調整における最初の調節ポイントであり、またプロテオームなどオミクスデータよりも高精度にデータを測定することができるため、様々な研究対象における基礎的な情報として、精力的にデータの収集が行われている。機能の発現という観点からは、タンパク質の量、局在、活性など mRNA 量以外の情報も重要であるが、mRNA 量はタンパク質量の 94%を説明するという出芽酵母における報告[1]があるように、mRNA 量が細胞機能の発現に関して極めて多くの情報を有しているのは間違えない。

遺伝子発現の類似性に基づく遺伝子間関係は遺伝子共発現情報と呼ばれ、遺伝子発現情報という動的情報を、理解が容易な静的情報として表現していると解釈することができる。発現を共にする遺伝子は機能を共にするという機能連座制アプローチ（guilt-by-association）により、パスウェイレベルの遺伝子機能を、網羅的に精度良く推定する事が可能になる。我々のグループを含めて、多数のグループが遺伝子共発現データベースを構築、運用しており、個別研究から網羅的研究まで幅広く利用されている[2][3]。

このような遺伝子共発現ネットワークを比較解析することで、ネットワークの保存性と特異性を解析する研究も精力的に行われている。ある生物種における異なる状態の比較（組織の違い、正常組織とがん組織の違いなど）は細胞内の動的な性質を明らかにする[4]。さらに比較ゲノムと同様のアプローチにより、ネットワークの生物種間共通性や多様性を解析する研究も行われている[5]（図１）。

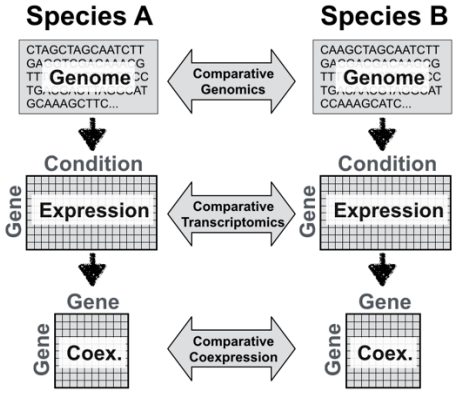

図１　比較トランスクリプトーム、比較共発現

この比較共発現という解析アプローチは、種固有のゲノムが種固有の遺伝子発現を決めるという観点から自然なものであろう。しかし、遺伝子という要素がハッキリした比較ゲノム解析とは異なり、遺伝子ネットワークの種間比較は様々な問題を抱えている。代表的な遺伝子ネットワークであるタンパク質間相互作用（PPI）ネットワークの比較解析においても、PPI を測定する手法の多様性、データの精度、網羅性、ダイナミクスの取り扱いなど、比較ゲノム解析にない難しさがある[6]。それでも PPI は実験による計測が可能であり、高精度の正解セットを構築できる物理的な関係である。これに対して、遺伝子共発現ネットワークは、物理的ではない関係性であるため何が真の共発現なのかという定義そのものが漠然としている。共発現ネットワークの種間比較は、種固有の共発現ネットワークがあり、それを比較することで、共発現ネットワークから見た種固有の

†1 東北大学大学院情報科学研究科
†2 東北大学東北メディカル・メガバンク機構

Vol.2018-MPS-118 No.39
Vol.2018-BIO-54 No.39
2018/6/15

機能や性質を明らかにすることが目的である。しかし、この種固有の共発現ネットワークはどのような性質を持ち、どのように導出するべきかについての議論はほとんどなされておらず、解析だけが先走りしている現状とも言える。

本報告では共発現ネットワークの進化解析の基盤となる、種固有の遺伝子共発現ネットワークに関して、トランスクリプトームスペースの観点から解析する。

2. 共発現行列の導出

2.1 一般的な導出手順

一般に遺伝子共発現情報は次の手順で導出される(図２)。

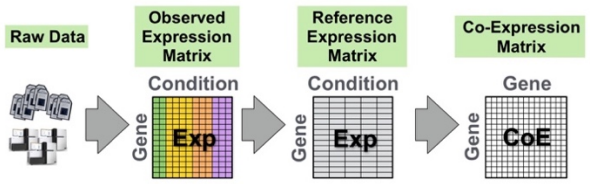

図２　共発現情報の導出手順

高精度な共発現情報を得るには遺伝子発現データは多い方が良い[7]。そのため GEO[8]や ArrayExpress[9]のような大規模なリポジトリから生データを取得し、一連の補正をして発現量行列を作成し、そこから遺伝子共発現行列（一般的にはピアソンの相関係数の行列）を計算するのが一般的である。ネットワークを描画するにあたっては、共発現行列からネットワークの隣接行列を作成する[10]が、可視化は焦点が異なるため、本報告では共発現行列そのものを議論の対象とする。

2.2 比較共発現の問題点

ネットワークの種間比較を行う場合に、比較するデータ間に存在するバイアスは大きな問題になる。同じ手法、同じ実験条件から導出した PPI ネットワークの種間比較[11]、同じ実験情報から導出した共発現ネットワークの種間比較[12]では、比較対象にバイアスがないため、正しい比較を行うことが可能である。しかし、同じ実験系で処理するサンプル数には限りがある。特定の条件における遺伝子ネットワークは、種固有のネットワークとは言えないだろう。さらに、遺伝子共発現関係ではサンプル数がネットワークの質に直接的に影響を与えるため[7]、推定される共発現関係の信頼性の問題も懸念される。

各生物種で利用できる全てのトランスクリプトームデータをもちいて導出した共発現ネットワークは、限られたデータのみから導出された共発現データよりも良い共発現である可能性が高いが、それは種固有の共発現関係の不偏推定になっているだろうか。いま、仮想的な例として、生物種 A と生物種 B の比較共発現解析を考える (図３)。生物種 A においても生物種 B においても、本来は強く共発現している遺伝子ペア（r=0.8, 遺伝子 X と遺伝子 Y）であったとしても、サンプリングに系統的な偏りがあった場合には、見かけ上の相関係数には違いが生じ得る。実際、研究者はランダムに研究対象を選択することはなく、生物種 A や生物種 B の生物学的ならびに経済学的な特徴を生かして、収集する組織や環境が偏ってくる。例えばイネであれば可食部である穀粒の成長に興味が集まる。アブラナ科特有の苦味成分（辛子油配糖体）を研究するならばシロイヌナズナを用いる。実際に収集される組織や環境が異なれば、本来同程度の共発現関係にある遺伝子ペアに対しても、図３のように異なる相関係数（r=0.7 vs 0.2）を示すことは十分に想定される。

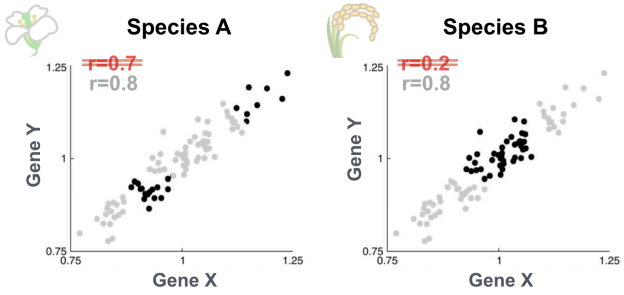

図３　遺伝子共発現の種間比較の問題

比較共発現解析では、相関係数の差の検定を行うが[4]、種間比較においては、母集団（トランスクリプトームスペース全体）からのランダムサンプリングではないため、検定の前提が成立しない。この問題に対して、我々は保存共発現と組み合わせることで系統特異的共発現を示す試みを行った[13]。しかし、実際にどの程度サンプリングバイアスが存在するのかは明らかではない。サンプリングバイアスが十分に小さければ、生物種間の関係性にとらわれることなく、各々の生物種に関して最適化を行った共発現ネットワークを種間比較すれば良い。逆に、似た系統の生物種は似たサンプリングバイアスを示すのであれば、種間比較から系統間比較にしても問題は解決しないことになる。

3. サンプリングバイアスの検出

種特異的な遺伝子共発現を考えるにあたり、共発現を計算する前の遺伝子発現量行列について考える必要がある（図２）。地球上に存在する物理化学的環境の中で、細胞が生存できる環境はその一部であり、その環境は生態学的ニッチを反映し生物種ごとに若干異なったものになると想像できる。この、細胞ひいては個体が生存できる環境がどの程度生物種間で重なっていて、どの程度異なっているのか。また重なっている環境条件については、生物種間でサンプリング密度が揃っているのかが問題となる。これらを見積もるために、複数の生物種のトランスクリプトームデータ

Vol.2018-MPS-118 No.39
Vol.2018-BIO-54 No.39
2018/6/15

を一度にクラスタリングすることで、生物種を超えたサンプル空間の全体像を描くことを試みた。

3.1 データセット

植物の共発現データベース ATTED-II[13]で公開されている共発現プラットフォームのうち、マイクロアレイと RNAseq の両方のプラットフォームによる遺伝子発現量のデータがある、シロイヌナズナ（Ath）、ダイズ（Gma）、ブドウ（Vvi）、イネ（Osa）、トウモロコシ（Zma）の 5 生物種を対象とした。利用可能なサンプル数は生物種ごと、プラットフォームごとに大きく異なる。サンプル数の違いの影響を軽減するため、各プラットフォームに関して、サンプル数が 200 に揃うようにランダムランプリングを行った（表 1）。

生物種	プラットフォーム	サンプル数	
		元データ	サンプリング後
シロイヌナズナ	マイクロアレイ	3229	200
	RNAseq	2121	200
ダイズ	マイクロアレイ	1115	200
	RNAseq	577	200
ブドウ	マイクロアレイ	245	200
	RNAseq	202	200
イネ	マイクロアレイ	2029	200
	RNAseq	264	200
トウモロコシ	マイクロアレイ	755	200
	RNAseq	684	200

表 1　種間比較用サンプルセット

生物種を繋ぐためのオーソログ関係は、OrthoFinder[14]で計算した。表 1 にある 10 のプラットフォームの全てで発現量を測定できたオーソロググループのみに注目する。各生物種の各オーソロググループ（メタ遺伝子）には複数の遺伝子が入るため、各オーソロググループに含まれる遺伝子発現量の真数スケールにおける総和を用いて、メタ遺伝子の発現量テーブルを作成した（図 4）。

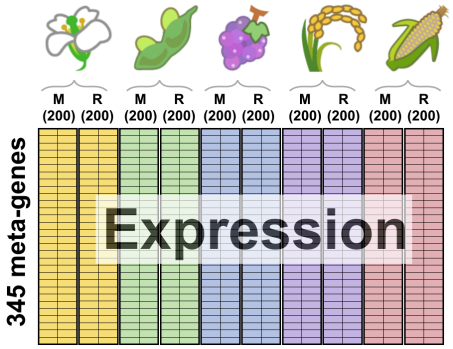

図 4：メタ遺伝子の発現量テーブル

サンプル収集のバイアス、ひいてはトランスクリプトームスペース全体の俯瞰することを目的に、この発現量テーブルのサンプル類似度について、以下の 2 種類のデータ解析を行った。

3.2 サンプル類似度の可視化

図 4 の 345 メタ遺伝子を 2 次元に圧縮し、2000 サンプル(200 サンプル、10 プラットフォーム)をプロットしたのが図 5 である。サンプル間の距離には、サンプル相関の 1−決定係数を用い、次元圧縮には tSNE [15]を用いた。10 のプラットフォームの色分けのため、生物種ごとにパネルを作成している。赤がマイクロアレイのサンプル(Xxx-m)、青が RNAseq のサンプル(Xxx-r)、灰色は他の 4 生物種のサンプルである。どの生物種においても赤い点と青い点は十分に混ざっており、マイクロアレイと RNAseq の技術の違いによる顕著な偏りは検出されなかった。一方で、生物種ごとには分布が大きく異なっており、サンプリング条件には生物種のバイアスが顕著に存在することがわかる。特にブドウにおいては、他のサンプルと交わっている領域が少なく、この 5 生物種のサンプル集合の中では、特異なサンプル集合だと言える。イネにも若干同様の傾向が見られる。ブドウの 2 つのプラットフォーム、ならびにイネの RNAseq は元のサンプル数が 300 未満（表 1）と少ないことが、サンプル条件の偏りをもたらしていると推定される。

3.3 サンプルのクラスタリング

同じデータを階層的クラスタリングした結果が図 6 になる。先ほどと同様に、サンプル間の距離は 1−決定係数であり、完全連結法にてクラスター数が 100 になるようにクラスタリングを行なった。マイクロアレイと RNAseq の偏りは小さいため、ここでは各生物種（マイクロアレイと RNAseq を合わせた 400 サンプル）ごとに、各クラスターに含まれるサンプル数をヒートマップで示す。図 6 は、列が 100 のサンプルクラスターであり、そのクラスターに含まれる生物種別のサンプル数を、黒から白の 7 段階のグレースケールで表している（0, 1, 2, 3~4, 5~8, 9~16, 17~32）。もしサンプルが均等に分散していれば、各生物種の各クラスターには 4 サンプルが属することになる。実際には、全く所属サンプルがないクラスター（黒）や所属サンプルの多いクラスター（白）が偏在しており、tSNE による可視化と同じ結論を読み取れる。

サンプル条件の網羅性を定量的に評価するため、各クラスターに含まれるサンプルの有無に注目した。100 のサンプルクラスターの中で、5 生物種全てについて 1 つ以上のサンプルを含むクラスターは 11 あり、4 生物種が含まれるクラスターは 19 ある（表 2）。これらの計 30 クラスターはどの生物も本来持つべき条件だと考えても良いだろう。

Vol.2018-MPS-118 No.39
Vol.2018-BIO-54 No.39
2018/6/15

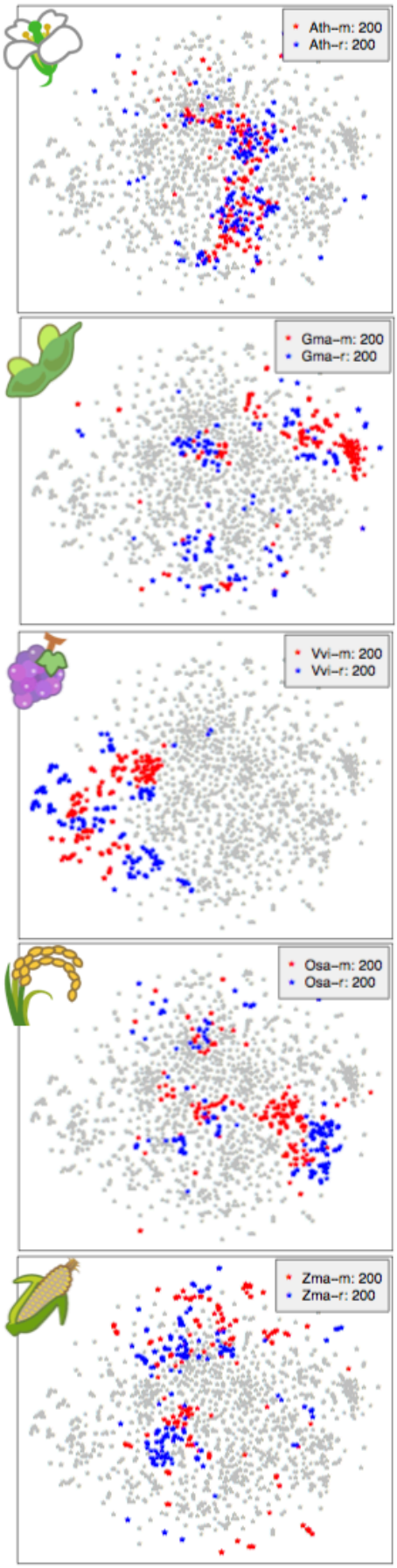

図５　tSNE によるサンプル分布の描画

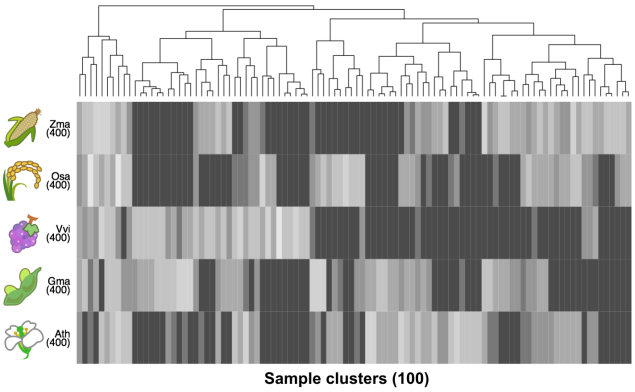

図６　サンプルの階層的クラスタリング

5種共通	11 クラスター
4種共通	19 クラスター
3種共通	24 クラスター
2種共通	32 クラスター
1種のみ	14 クラスター
計	100 クラスター

表２　サンプルクラスターに含まれる生物種数

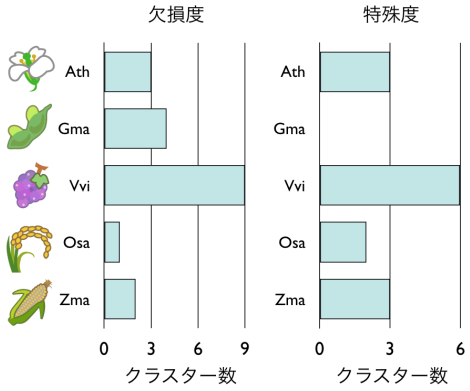

図７　サンプルの欠損度、特殊度

特定の生物種のみでデータが欠損している 19 のクラスターの内訳を図７に欠損度として示す。ダウンサンプリングの過程で解析データから落ちてしまった可能性もあるが、少なくともサンプルが薄いと解釈することができる。9 つのクラスターにおいて、ブドウのみでサンプルが欠損しており、ブドウのサンプル網羅度は低いことを示している。このことは、ダウンサンプリング前のサンプル数（表１）とも強く関係している。ブドウのサンプルはマイクロアレイと RNAseq を合わせて 450 程度と、他の生物種のサンプル数の半分以下であり、このことが十分な条件多様性を確保できていない理由だと思われる（ただし 450 というサンプル数は必ずしも少ないとは言えない）。そのほか全てではないにせよ各生物種固有の生存環境限界を反映している可

Vol.2018-MPS-118 No.39
Vol.2018-BIO-54 No.39
2018/6/15

能性もあり、その場合には努力しても決してサンプリングできない環境ということになる。

逆に 1 種のみで構成されているクラスターは特殊なサンプルだと判断できるので、それをサンプル特殊性と解釈して、図 7 に示した。先の欠損度と同様にブドウのサンプルの特殊性が際立っている。ブドウ特異的クラスターの多くはブドウの液果(berry)を対象としたサンプルであり、これは今回解析した他の生物種にはない組織のため特異的クラスターを形成している。また、ブドウの液果は可食ならびにワインの原料として重要であり、このことが液果に強く偏ったサンプリングになっていると推定できる。また、トウモロコシ特異的なサンプルクラスターは主に雄穂や胚乳サンプルであった。雄穂はトウモロコシのみの組織である。胚乳に関しては、トウモロコシだけでなくブドウやイネも有胚乳種子でありトウモロコシ特異的な組織という訳ではない。イネとブドウではサンプルが欠損しているか、性質が異なる組織である可能性が考えられる。

次元圧縮による可視化、ならびに階層的クラスタリングの全体的な結論としては、各生物種のサンプルには共通条件もあるものの、全般的には大きく偏っている。また、全くカバーされていない条件も存在するなど、このトランスクリプトームデータ、ひいてはこれに基づく遺伝子共発現データの種間比較には特別の注意が必要であると言える。

4. サンプリングバイアスの軽減

生物種ごとに重点的に収集しているサンプル条件が異なることがわかった。全く収集されていない条件については今後のデータ収集に期待するとして、すでに収集されている条件については、効率よく重複を取り除き、種固有のトランスクリプトームデータを構築したい。一方で、トランスクリプトームデータは特に低発現量遺伝子においては信頼性に欠けるため、単に代表サンプルを選択するだけでは良いデータにはならない。サンプルの次元圧縮は重複サンプルを扱う上で有力な方法であり、以下に主成分分析を用いたサンプルの次元圧縮の有効性について述べる。

4.1 データセット

共発現データの性能評価を行うため、遺伝子オントロジーがよく整備されているシロイヌナズナを対象とした。特に、共発現データベース ATTED-II にて公開されているシロイヌナズナの RNAseq のデータを用いた。サンプルのバッチ効果は ComBat[16]を用いて除き、さらに各遺伝子について平均が 0 になるように正規化した。

4.2 主成分空間における相関係数

2120 サンプルの主成分分析を行ったところ、80%の分散を保持するのに約 500 主成分、90%の分散を保持するのに約 1000 主成分必要なことがわかった（図８）。遺伝子 x の i 番目のサンプルにおける発現量を x_i、平均発現レベルを $\bar{x}$ とすると、相関係数 r は次の計算式になる。

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}\sqrt{\sum\left(y_i - \bar{y}\right)^2}}$$

回転前の遺伝子発現行列において、バッチ補正を行っており、その結果各遺伝子の平均発現量が 0 になっているため、上式の $\bar{x}, \bar{y}$ は 0 であり、ピアソンの相関係数はコサイン相関と等価になる。そのため、主成分分析による回転前後で遺伝子 x と遺伝子 y の相関係数 r は変化しない。主成分分析後の遺伝子ペアの相関関係の例を図 9 に示す。横軸の EGI:824629 はトリプトファン合成酵素アルファチェイン（TSA1）、縦軸の EGI:831666 はトリプトファン合成酵素 1（TRP1）であり、どちらもトリプトファン合成に必要な遺伝子として共発現していることがわかる。

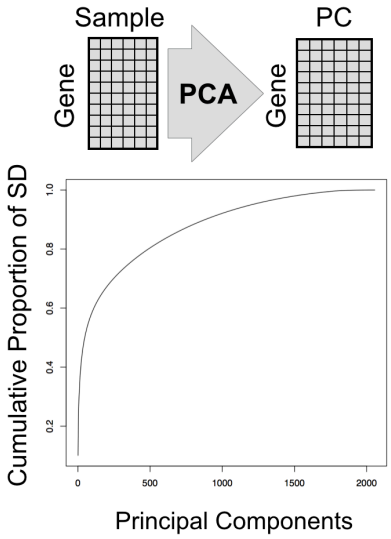

図８　サンプルの主成分分析

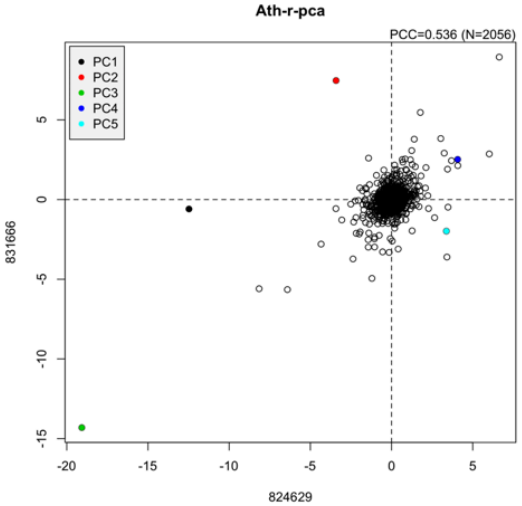

図９　主成分を用いた遺伝子ペアの相関

平均寄与率のもっと大きい５つ主成分（第１主成分から第５主成分）に注目すると、今回の遺伝子ペアについても変動度の大きい成分となっており、低次の主成分を省略しても相関係数にほとんど影響を与えないことがわかる。

さて、前節において、サンプルの偏りが問題になっている。仮に N 個の同一のサンプルが発現行列データに含まれていた場合、その共分散はルート N 倍となる。主成分分析を行うことで偏りの軽減にはなっているが、完全には排除できていない。言い換えれば、各主成分の寄与率はサンプルの偏りの結果でもあり、寄与率の大小について何らかの補正が必要になる。相関係数の計算においては、大きな寄与のサンプル（本節では主成分）が小さな寄与のサンプルの効果を隠蔽してしまう。そこで、ブートストラップを行い、確率的に上位の主成分を取り除いた共発現を計算し、ブートストラップ共発現の平均値を最終的な共発現とする手法を提案する（バギング共発現）。

共発現評価の指標として遺伝子オントロジーとの一致率がよく用いられる[13]。この手法で提案手法による共発現の質を評価したところ、共発現指標としてピアソンの相関係数を用いた場合にも（5.5 から 5.9 へ）、相互ランク[17]を用いた場合にも（6.2 から 6.5 へ）上昇しており、サンプル重複の問題を軽減することで共発現の精度を向上することに成功した。

5. まとめ

本報告では、共発現情報の種間比較において、サンプル条件の種間バイアスを指摘し、バイアスの効果を軽減する手法の提案を行った。提案手法は生物種単独の共発現スコアを向上させたため、サンプルバイアスの軽減に有効であると考えられる。今後、種固有のトランスクリプトームスペースをより正確に構築するに当たっては、現在のトランスクリプトームデータが必ずしも単一組織に由来するものではないため、サンプルを組織に分解することが大きな課題となる。

参考文献

[1] Li, J.J., et al.. Quantitating translational control: mRNA abundance-dependent and independent contributions and the mRNA sequences that specify them. Nucleic Acids Res. 2017, vol. 16; p. 11821-11836.
[2] Usadel, B. et al.. Coexpression Tools for Plant Biology: Opportunities for Hypothesis Generation and Caveats. Plant Cell and Environment. 2009, vol. 32, p. 1633-1651.
[3] Obayashi, T. and Kinoshita, K. Coexpression landscape in ATTED-II: usage of gene list and gene network for various types of pathways. J. Plant Research. 2010, vol. 123, p. 311-319.
[4] De la Fuente. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet. 2010, vol. 26, p. 326-33
[5] Ruprecht, C. et al.. Beyond Genomics: Studying Evolution with Gene Coexpression Networks. Trends Plant Sci. 2017, vol. 22, p. 298-307.
[6] Kiemer, L. and Cesareni, G. Comparative interactomics: comparing apples and pears? Trends Biotechnol. 2007, vol. 25, p. 448-454.
[7] Huang, J. et al.. Construction and Optimization of a Large Gene Coexpression Network in Maize Using RNA-Seq Data. Plant Physiol. 2017, vol. 175, p. 568-583.
[8] Barrett, T. et al.. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013, vol. 41, p. D991-995.
[9] Kolesnikov, N. et al.. ArrayExpress update--simplifying data submissions. Nucleic Acids Res. 2015, vol. 43, p. D1113-1116.
[10] Zhang, B. and Horvath, S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005, vol. 4, Article 17.
[11] Wan, C. et al.. Panorama of ancient metazoan macromolecular complexes. Nature. 2015, vol. 525, p. 339-344. Rolland, T. et al.. A proteome-scale map of the human interactome network. Cell. 2014, vol 159, p.1212-1226.
[12] Hu, G. et al.. Evolutionary Conservation and Divergence of Gene Coexpression Networks in Gossypium (Cotton) Seeds. Genome Biol Evol. 2016, vol. 8, p. 3765-3783.
[13] Obayashi, T. et al.. ATTED-II in 2018: A Plant Coexpression Database based on Investigation of Statistical Property of the Mutual Rank Index. Plant Cell Physiology. 2018, vol. 59, p. e3.
[14] Emms, D.M. and Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. 2015, vol. 16, p. 157.
[15] van der Maaten, L.J.P. and Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research. 2008, vol. 9, p. 2579-2605.
[16] Johnson, W.E. et al.. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007, vol. 8, p. 118-127.
[17] Obayashi, T. and Kinoshita K. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Research. 2009, vol. 16, p. 249-260.