様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １０ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ０６６２４

研究課題名（和文）リファレンストランスクリプトーム構築に基づく遺伝子共発現ネットワーク解析の高度化

研究課題名（英文）Improvement of gene coexpression network analysis with the construction of reference transcriptome data

研究代表者

大林　武（Obayashi, Takeshi）

東北大学・情報科学研究科・准教授

研究者番号：５０３９７０４８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：遺伝子共発現解析は発現プロファイルが似ている遺伝子群を特定することで、遺伝子の機能推定を行う手法である。公共のリポジトリに蓄積している大量の遺伝子発現データは多様な環境を反映しているが、サンプルの偏りをどのように扱うかが問題となる。本研究では、主成分分析を用いて遺伝子発現データを再構成し、アンサンブル計算と組み合わせることで、高精度の共発現情報を抽出できることを見出した。

研究成果の学術的意義や社会的意義
本研究では、データを収集する際のサンプリングバイアスを事後処理によって軽減する手法を提案し、公共のトランスクリプトームデータに基づく遺伝子ネットワーク解析を大きく改善することに成功した。サンプリングバイアスは大規模データを扱う上での大きな問題であり、それを解決する本手法は広い分野に適用できる可能性がある。

研究成果の概要（英文）：Gene coexpression analysis is a method for predicting gene function by identifying groups of genes with similar expression profiles. Although a large amount of publicly available gene expression data reflects diverse environments, sample bias in the data is a severe problem in coexpression analysis. We have developed a new methodology combining principal component analysis and ensemble calculation to retrieve highly accurate coexpression information.

研究分野：バイオインフォマティクス

キーワード：遺伝子共発現　トランスクリプトーム　データベース　ネットワーク解析　進化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1) ゲノム配列の決定コストが下がり続ける一方で、そこにコードされている遺伝子の機能解析は依然として容易ではない。遺伝子の機能を個々に解明していく個別研究では、着目する表現型を軸に関連する知見を組み合わせて実験をデザインし、解釈性の高い結果が得る。一方で、知見の乏しい遺伝子を解析するのは難しく、また、個別研究のアプローチを多数の遺伝子に展開するのも容易ではない。知見の乏しい遺伝子群を解析するには、ゲノムワイドな測定に基づくデータ駆動的アプローチであり、これにより研究されていない遺伝子と研究されている遺伝子を関連づけることができる。様々なゲノムワイドデータの中でも、トランスクリプトームは測定コストが比較的低く、また静的なゲノム配列から動的な機能発現を行う最初の制御点であることから、多くの研究で測定が行われている。トランスクリプトームは所与の細胞環境において必要な遺伝子群を示しており、多くの独立なサンプルにおいて遺伝子発現プロファイルの似ている遺伝子群は、類似の細胞機能を担っていると類推できる。このように遺伝子発現プロファイルの類似性が遺伝子機能の類似性を示しているという理屈で遺伝子発現データを解析するのが、遺伝子共発現解析である（Aoki et al. 2007、Usadel et al. 2009）。

(2) 遺伝子間の微かな機能の違いを正確に捉えるには、多くのサンプルにおける遺伝子発現プロファイルが必要になる。NCBI Gene Expression Omnibus（Barrett et al. 2013）などの公共のデータベースには大量の遺伝子発現データが蓄積しており、このような公共データに基づく遺伝子共発現解析が盛んに行われている。ここで、遺伝子発現制御は、発生段階や組織などの細胞内環境と、物理化学的刺激や細胞間伝達物質などの細胞外環境の両者に依存するため、細胞の異なる遺伝子発現状態を引き起こす環境のレパートリーは極めて多いことに注意しなければならない。公共のデータベースには大量のデータが蓄積しているとは言え、頻繁に測定されている環境と、全く測定されていない環境が不均一に存在していることが想定される。このことは遺伝子発現プロファイルを構築し、遺伝子間の類似性を導出する上で大きな問題となる。これまでにもサンプルの偏りを是正するための手法が提案されてきたが（Obayashi et al. 2008、Usadel et al. 2009）、依然として問題解決には至っていない。

２．研究の目的

サンプル分布に偏りのある公共の遺伝子発現データベースを再構成し、遺伝子共発現法の高精度化を達成する。また次の 3 つの観点を通じて個別研究を網羅的に支援する。

（１）遺伝子共発現法の性能を向上により、高精度な遺伝子機能の網羅的推定を実現する。

（２）サンプルの偏りを是正した、生物種本来の遺伝子共発現ネットワークを構築することで、生物種間の比較可能性を向上させる。進化的な洞察を可能し、モデル生物種から非モデル生物種への知見の転用を促進する。

（３）再構成した遺伝子発現データならびにそれに基づく遺伝子共発現情報をデータベースとして公開し、さらなる網羅的解析の材料として広く提供する。

３．研究の方法

(1) シロイヌナズナは最も代表的なモデル植物として幅広いトランスクリプトーム解析が実施されており、またヒトやマウスよりも遺伝子制御ネットワークがシンプルであると推定される。そのため、シロイヌナズナを主な解析対象種とし、そこで開発した手法を他の生物種で検証する構成とする。

(2) トランスクリプトームデータとしては、近年 RNAseq の利用が増えてきたものの、マイクロアレイ技術も引き続き利用されている。今後 RNAseq の利用がさらに増えることを鑑みて、RNAseq データを手法開発の基本プラットフォームとし、その手法をマイクロアレイデータで検証することで、手法の普遍性を確認する。

(3) DDBJ DRA よりシロイヌナズナの Illumina RNAseq データをダウンロードし、MATATAKI（Okamura and Kinoshita 2018）で遺伝子のカウントに変換した。200 万カウント未満のサンプルを除き、平均 30 カウント未満の遺伝子を除いた。14741 サンプルに対する 18957 遺伝子の遺伝子発現データ（カウント）は、0.125 の疑似カウントを加えて対数値に変換した。また、各実験を単位としたバッチ補正は ComBat（Johnson et al. 2007）により行った。遺伝子の発現変動に着目するため、各遺伝子の平均発現レベルが 0 になるように遺伝子ごとに値の中心化を行った。

(4) 発現制御システムは、ある生物が遭遇する全ての環境において適切な遺伝子発現を実現できる。ここで、「ある生物が遭遇する全ての環境」は有限ではあるが、具体的に列挙することは難しい。また、実際にゲノム配列や制御システムに影響を与える個々の環境に起因する進化圧の強さは様々である。そのため、遺伝子発現への影響（発現制御システムへの進化圧）が同等となるような環境レパートリーは存在するはずであるが、それを陽に求めることはできない。そこで、理想的なレパートリーからの偏りを扱う代わりに、環境の独立な成分を求め、その分布を解析するところから始めた。

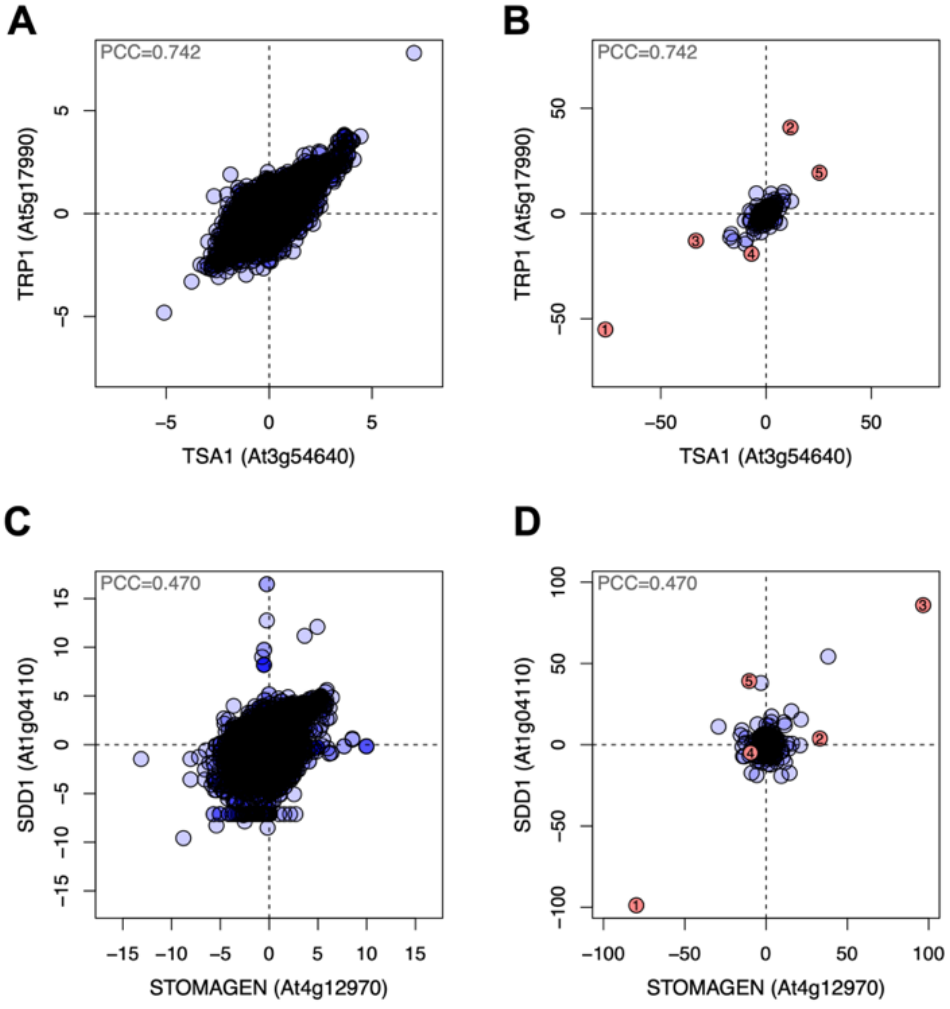
４．研究成果

(1) 主成分分析を用いて遺伝子発現データのサンプルを独立成分に再構成した。具体的には、各サンプルを軸とする 14741 次元の空間に、18957 遺伝子がプロットされているデータに対して、非中心化主成分分析による軸の回転を行った。遺伝子共発現解析の一つの技術的なポイントは、発現プロファイルの類似性をどのように計算するかである。ピアソンの相関係数が共発現指標として広く使用されており（Usadel et al. 2009）、我々もピアソンの相関係数の相互ランク基づく指標を用いている（Obayashi & Kinoshita 2009）。各遺伝子を中心化している条件では、ピアソンの相関係数はコサイン相関係数と同値であり、また軸の回転操作は任意の遺伝子ペアと原点を結ぶ角度を変化させないため、ピアソン相関係数に基づく遺伝子プロファイルの類似性は主成分分析の前後で変化しない。そのため、ここでの主成分分析は純粋にサンプル重複の解釈性を高める操作となる。ここで、各主成分はサンプル環境を構成する独立成分であると解釈できる。寄与率の低い主成分は主にランダムノイズにより構成されるため、そのような主成分を除くことで、遺伝子発現プロファイルの精度が向上することが期待できる。

(2) ノイズではない有益な主成分をどのように決定するかが問題になる。主成分の寄与率は傾き -1 の冪分布であった。このことは累積寄与率の大小によらず、寄与率の低い下位の主成分も遺伝子プロファイルを特徴づける重要な軸であることを示唆する（Geo et al. 2003）。仮に累積寄与率 90%点を指標とすれば上位 2350 主成分であるが、このことだけで有益な主成分を決定することはできない。

(3) 各軸の生物学的有用性を評価するために、各主成分と遺伝子機能の関連を調べた。KEGG パスウェイの濃縮検定の結果、おおよそ上位 200 主成分は遺伝子機能と関連しており、それ以降の主成分では関連性は有意とは言えなくなる。ただし、完全に関連性が消える訳ではない。

(4) 下位の主成分の共発現情報に対する影響を見積もるため、利用する主成分の数をパラメータとして、そこから導出される共発現情報の質を KEGG パスウェイアノテーションを用いて評価した。用いる遺伝子プロファイル類似性指標によって若干傾向の差があるものの、いずれのケースにおいて、全ての主成分を用いるよりも、大方上位 200 から 1000 主成分を用いた場合の方が共発現の性能が良いことを見出した。このことは、上位 1000 主成分以降の主成分は主にノイズから構成されていることを示唆する。上位 200 主成分から上位 1000 主成分については、共発現指標との組み合わせが示唆されたため、この上位 200 主成分と上位 1000 主成分の両者について引き続き検討を進めた。



(5) 各遺伝子の発現プロファイルは各種サンプルにおける遺伝子発現量のベクトルであるが、オリジナルの遺伝子発現量の代わりに、主成分スコアを用いて発現プロファイルを構成することができる。例として、TSA1 と TRP1 が共発現している様子を元データと主成分スコアで示す（図 A,B）。この 2 つの遺伝子はトリプトファン生合成系の遺伝子であり、協調的な制御を通じてトリプトファン合成量を調節している。図 A では 14741 サンプルが 2 変数正規分布のような分布で相関していることを示している。図 B は主成分スコアに基づく共発現であり、第 1 から第 5 主成分をハイライトしている。上位の主成分は全遺伝子で平均的に寄与の高い主成分であり、TSA1 と TRP1 においても大きな寄与を持っていることがわかる。一方、この 5 つの主成分の寄与は明確に大きいものの、これらを取り除いてもなおこの 2 つの遺伝子は共発現していることがわかる。そのため、TSA1 と TRP1 はどのような条件においても安定的に共発現している遺伝子ペアだと言える。このような恒常的関係は全ての共発現遺伝子ペアで見られるわけでない。図 C,D で示す STOMAGEN と SDD1 は共に孔辺細胞の数を制御している（Sugano et al. 2010）。オリジナルの遺伝子発現量を用いた共発現では全体的に弱く共発現しているように見える（図 C）。一方、主成分スコアを用いると、共発現に寄与しているのは主に第 1、第 3 主成分であり、他の主成分ではほとんど共発現をしないことがわかる（図 D）。すなわちこの 2 遺伝子の共発現は、条件特異的であると言える。STOMAGEN と SDD1 は共に制御に関わる遺伝子であり、もしこの 2 つの遺伝子が常に同じ発現変動をしていたならば、制御の複雑度としてはどちらか 1 つで十分ということになる。制御因子は各々異なる細胞環境を反映しており、それらが統合することで制御システムが実現していることを考えると、STOMAGEN と SDD1 の共発現が条件特異的であることは理にかなっている。

(6) この主成分スコアによる共発現の描像は、公共のデータベースに蓄積するサンプルの重複にどのように関係するだろうか。重複したサンプル環境は同じ主成分にまとまるため、主成分の寄与率はサンプル重複の影響を受けている。サンプルの重複を除くためには、主成分寄与率を無視して、全ての主成分を同じ寄与率に補正してから遺伝子間の類似性を計算するアプローチが考えられる（Hibbs et al. 2007）。しかし、実際の環境を構成している要素に、主要な要素とマイナーな要素があるという解釈は自然である。また、データに含まれるノイズは、必ずしも完全なホワイトノイズではなく、通常何かしらの構造があるため、SN 比が低くなる下位の主成分では上位の主成分よりもノイズの影響を強く受ける。そのため、全ての主成分を一律で同一に扱うのは情報量の観点からも得策ではない。

(7) 主成分スコアは冪分布であるため、共発現を計算する際に、上位の主成分が下位の主成分の影響をほぼ完全に隠蔽してしまうことが大きな問題である。上位だけでなく下位の主成分の値も考慮するために、2 つの類似性指標を検討した。一つ目は値ではなく順位を用いるスピアマン相関係数である。冪分布がもたらす大きな寄与率の違いに影響を受けることなく、遺伝子発現の類似性を計算できる。もう一つは少数の主成分をランダムサンプリングしてピアソン相関係数を計算するアンサンブル計算（Subsample aggregating: Subagging）である。

(8) この 2 つの類似性指標と他の要素の組み合わせを検討した。すなわち、(a)データを主成分分析するか否か。主成分分析を用いた場合、使用する主成分は上位 200 か 1000 か。(b)遺伝子の類似性はスピアマン相関係数かピアソン相関係数か。それらの類似性を相互ランクに変換するか。(c)アンサンブル計算を行うか。これらの組み合わせを検討したところ、上位 1000 主成分に対してピアソン相関係数の相互ランクをアンサンブルで計算するのが最も優れるという結果だった。また、この組み合わせには及ばないものの、上位 200 主成分に対してスピアマン相関係数を計算するだけの方法も良好な結果となった。

(9) シロイヌナズナの RNAseq データに対して得られたこの結果を、共発現データベース ATTED-II（Obayashi et al. 2018）で公開している 9 生物種、17 プラットフォームで検証をしたところ、シロイヌナズナ RNAseq と同様の結果が得られた。すなわち、今回の提案手法が普遍的に有効であると言える。ピアソン相関係数の相互ランク変換が良い共発現指標であることは、以前我々が報告したものであるが（Obayashi & Kinoshita 2009）、今回はこれに主成分分析とアンサンブル計算が加わったものになる。アンサンブル計算の各ステップでは、一部の主成分（すなわち環境要素）から共発現を導出するが、これは条件特異的な共発現であると解釈できる。すなわち、アンサンブル計算は、大量の条件特異的な共発現を導出し、その平均値として条件非特異的な共発現情報を導出していると解釈できる。

(10) 最後に主成分分析とアンサンブル計算による共発現導出のパスウェイ依存性について調べた。パスウェイごとにアンサンブル計算の有無を比較したところ、共発現解析が有効なパスウェイならば、パスウェイの性質によらず、平均的に遺伝子機能予測能力が向上していることが判明した。すなわち、特定のパスウェイの特性に影響された結果では、遺伝子共発現解析におけるサンプルの偏りという基盤的な問題を軽減することに成功したと言える。これらの成果をまとめ、Plant and Cell Physiology 誌に発表した（Obayashi et al. 2022）。

(11) また、この手法で構築した共発現データは植物の共発現データベース ATTED-II https://atted.jp 、動物の共発現データベース COXPRESdb https://coxpresdb.jp にて公開した。図 B,D で示す共発現スコアを用いた共発現の表示については現在ツールを開発中であり、次期バージョンにて公開できる見通しである。

引用文献

Aoki K., et al. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol.* 48: 381–390.

Barrett T., et al. (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41: D991–D995.

Geo J.B., et al. (2003) Principal component analysis of 1/f noise. *Phys. Lett. A* 314: 392–400.

Hibbs M.A., et al. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* 23: 2692–2699.

Johnson W.E., et al. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118–127.

Obayashi T, et al. (2008) COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res*. 36: D77-D82.

Obayashi T. and Kinoshita K. (2009) Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res*. 16: 249–260.

Obayashi T., et al. (2018) ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant Cell Physiol*. 59: e3

Obayashi T, et al. (2022) ATTED-II v11: a plant gene coexpression database using a sample balancing technique by subagging of principal components. *Plant Cell Physiology*, in printing.

Okamura Y. and Kinoshita K. (2018) Matataki: an ultrafast mRNA quantification method for large-scale reanalysis of RNA-Seq data. *BMC Bioinform*. 19: 266

Sugano SS, et al. (2010) Stomagen positively regulates stomatal density in Arabidopsis. *Nature*, 463: 241-244.

Usadel B, et al. (2009) Coexpression Tools for Plant Biology: Opportunities for Hypothesis Generation and Caveats. *Plant Cell and Environment*, 32: 1633-1651.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Obayashi Takeshi、Hibara Himiko、Kagaya Yuki、Aoki Yuichi、Kinoshita Kengo	４．巻 -
２．論文標題 ATTED-II v11: A Plant Gene Coexpression Database Using a Sample Balancing Technique by Subagging of Principal Components	５．発行年 2022年
３．雑誌名 Plant and Cell Physiology	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/pcp/pcac041	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Takeshi Obayashi, Yuichi Aoki, and Kengo Kinoshita
２．発表標題 全ゲノム遺伝子共発現情報の可視化手法の検討
３．学会等名 第9回生命医薬情報学連合大会
４．発表年 2020年

１．発表者名 Takeshi Obayashi, Yuichi Aoki, and Kengo Kinoshita
２．発表標題 Comparison of summarization methods of co-expressed gene list to predict gene function
３．学会等名 28th Conference on International Society for Computational Biology (ISMB2020)（国際学会）
４．発表年 2020年

１．発表者名 Yuichi Aoki, Shinichi Yamazaki, and Takeshi Obayashi
２．発表標題 Development of Coexpression-guided Generative Model for Plant Gene Expression Inference
３．学会等名 28th Conference on International Society for Computational Biology (ISMB2020)（国際学会）
４．発表年 2020年

１．発表者名
Takeshi Obayashi and Yuichi Aoki

２．発表標題
ATTED-II v10.2: a Plant Coexpression Database Providing Logit Score of Ensemble Mutual Rank as Coexpression Index to Enhance Usability for Genome-Wide Analyses.

３．学会等名
第62回日本植物生理学会年会

４．発表年
2021年

１．発表者名
Yuichi Aoki, Takeshi Obayashi

２．発表標題
Exploration of Gene Expression Latent Space in Higher Plants by using Generative Models.

３．学会等名
第62回日本植物生理学会年会

４．発表年
2021年

１．発表者名
Takeshi Obayashi, Yuichi Aoki, Kengo Kinoshita

２．発表標題
Integration and evaluation of scRNAseq-, bulk RNAseq and microarray-based gene coexpression data

３．学会等名
ISMB/ECCB 2019（国際学会）

４．発表年
2019年

１．発表者名
Takeshi Obayashi, Yuichi Aoki

２．発表標題
ATTED-II v10: a Plant Coexpression Database Providing Logit Score of Ensemble Mutual Rank as Coexpression Index to Enhance Usability for Genome-Wide Analyses.

３．学会等名
日本植物生理学会年会（大阪大会）

４．発表年
2020年

１．発表者名
Takeshi Obayashi, Aoki Yuichi

２．発表標題
Gene-to-gene Spearman correlation using the sample principal component scores is a simple sample-balancing methodology for gene coexpression calculation.

３．学会等名
日本植物生理学会年会（筑波大会）

４．発表年
2022年

１．発表者名
Takeshi Obayashi

２．発表標題
大規模遺伝子発現量データを構成する独立なサンプル数の極限値の見積もり

３．学会等名
第67回情報処理学会バイオ情報学研究会（SIGBIO）

４．発表年
2021年

１．発表者名
Takeshi Obayashi

２．発表標題
遺伝子共発現の評価における発現レベル依存性

３．学会等名
第66回情報処理学会バイオ情報学研究会（SIGBIO）

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関