2025 年度終了報告書

ライフサイエンスデータベース統合推進事業
統合化推進プログラム

【研究課題名】「非モデル植物のための遺伝子ネットワーク情報活用基盤」

研究代表者：
大林 武（東北大学 大学院情報科学研究科 教授）

研究分担者：
なし

1. 概要

(1) 研究および計画の概要

植物遺伝子の共発現データベース ATTED-II を基盤として、モデル植物の遺伝子共発現情報を充実し、その知見を非モデル植物へと橋渡しするための基盤を構築する。本データベースにより、モデル植物で得られた遺伝子共発現ネットワークの知見を非モデル植物の研究に活用可能とすることを目的とする。

(2) 成果の概要

本研究では、植物遺伝子共発現データベース ATTED-II を基盤として、モデル植物で得られた遺伝子ネットワーク情報を非モデル植物へと橋渡しするための解析基盤を構築した。非モデル植物では、トランスクリプトームデータの量的不足に加え、サンプル条件の偏りが共発現解析の精度と解釈を困難にするという課題がある。本課題では、主成分分析に基づくサンプル構造の抽出と可視化（CoexViewer、PC View）により、共発現が成立する条件の解釈を可能とした。

さらに、RNA-seq データに基づく共発現解析パイプラインの高精度化と自動化を進めるとともに、共発現マップ（CoexMap）を用いた種間比較機能を高度化し、複数植物種間で共発現構造を統合的に比較できる環境を整備した。特に、細胞内局在に基づくメゾスケールの比較手法を導入することで、機能モジュール単位での種間比較を可能とした。

最終的に、ATTED-II version 13.0 を公開し、共発現情報を収載する植物種を 19 種 20 系統に拡張した。また、主成分軸に対する KEGG 濃縮解析と大規模言語モデル（LLM）による要約生成を統合することで、共発現条件の解釈を支援する機能を実現した。この際、主成分ローディングに基づいて代表的な条件を選抜し、共通情報を除いたうえで LLM に入力することで、意味的統合に特化した解釈を可能とした。これにより、解析者が環境応答や進化的保存性の観点から遺伝子機能を理解できる統合的な解析基盤を確立した。

これらの成果により、非モデル植物における遺伝子機能解析および比較ゲノム研究の基盤が大幅に拡充された。ATTED-II は国際的な研究において継続的に活用されており、非モデル植物を含む解析基盤としての有効性と、研究コミュニティにおける波及効果が示された。

名称	概要
ATTED-II	植物の遺伝子共発現ネットワークを提供するデータベース。RNA-seq 共発現解析を基盤に、主成分解析によるサンプル条件可視化機能（CoexViewer、PC View）、共発現マップ（CoexMap）による種間比較機能、および LLM を用いた主成分要約機能を搭載し、19 種 20 系統の植物に対する共発現情報を提供する。

2. 目的・目標の達成状況

(1) 達成目標と達成状況

達成目標	達成状況
【項目 1】共発現提供生物種の情報拡充	
RNAseq 共発現の高精度化	定量法、主成分サバギング法、非標準種の扱いを統合した解析パイプラインを確立し、RNA-seq 共発現の精度と再現性を向上させた。
共発現モジュールが機能するサンプル条件の提示	主成分解析に基づく共発現条件の可視化機能（CoexViewer、PC View）を実装し、さらに LLM による主成分軸の要約機能を統合することで、共発現条件の解釈を可能とした。
連携機能の向上	Plant GARDEN との相互リンク、各生物種固有 ID 対応表の提供、Zenodo リポジトリによるデータの恒久保存および再利用基盤を整備した。
公開環境の整備	開発用・公開用サーバの物理分離、SSD 運用および非同期通信（Ajax）の導入により、高負荷アクセスに対応可能な堅牢な公開基盤を確立した。
遺伝子共発現情報の更新	RNA-seq 解析パイプラインの自動化と品質評価手法の整備により、ATTED-II version 13.0 を公開し、共発現情報を 19 種 20 系統に拡張した。
【項目 2】種間比較機能の高度化	
マクロスケールの種間比較システムの高度化	共発現マップ（CoexMap）の視認性を向上させるとともに、生物種の選択および複数種の並列表示機能を実装し、異なる植物種間での共発現構造の比較を可能とする基盤を確立した。
メゾスケールの種間比較システムの開発	細胞内局在に基づく共発現マップを導入し、機能モジュール単位での種間比較を可能とするメゾスケールの比較手法を確立した。

(2) 実施状況の詳細

【項目 1】共発現提供生物種の情報拡充

（項目 1A）RNAseq 共発現の高精度化

RNAseq に基づく共発現解析パイプラインの各工程を検討し、高精度化を行なった。サンプル選定では、ルールベースと手動チェックを組み合わせたフローを構築した。非標準種の影響については、シロイヌナズナにおいて、標準種のみの実験から構築した共発現データ（Ath-r）と、非標準種を含む実験から構築した共発現データ（Ath-e）を分離して提供する枠組みを導入し、データセットの性質に応じた解析を可能とした。発現量の定量では、ハッシュベースの手法からアライメントベースの手法に変更し、非標準種を含む場合においても安定した定量値を得られるようになった。さらに、ATTED-II version 11 で導入

した主成分分析に基づくサンプルバランシング手法を基盤として、サンプル条件のばらつきに起因するノイズの低減および低発現遺伝子における共発現推定の安定化を確認した。共発現導出および統合の段階では、これらの手法を統合した解析パイプラインを確立し、KEGG Pathway に基づく機能整合性スコアが安定的に向上することを確認した。これにより、非モデル植物を含む多様な植物種に対して比較可能な共発現データを提供する基盤が強化され、ATTED-II version 12.0 および version 13.0 の構築において中核的な役割を果たした。

（項目 1B）共発現モジュールが機能するサンプル条件の提示

遺伝子共発現関係は単一の数値として提示されるため利用が簡便であり、多数の遺伝子群の解析にも適しているが、その背後にあるサンプル条件は直接的には観察されない。共発現関係がどのようなサンプル条件下で成立するかは、遺伝子機能を生物学的に解釈するうえで重要であり、さらに仮説検証のための遺伝子ノックアウト実験において注目すべき環境条件を決定するためにも不可欠である。本課題の対象である非モデル植物では、利用可能な RNA-seq サンプルが特定の組織や条件に偏る傾向があり、同一の遺伝子ネットワークであってもサンプル条件の違いにより異なる共発現関係として観察される可能性がある。このため、共発現の種間比較を行ううえでも、サンプル条件の構造を適切に把握することが重要な課題となる。

これを踏まえ、本課題では、メタデータ整備、サンプル集合の理解、サンプルネットワーク構築の三要素を統合する枠組みとして、主成分分析に基づきサンプル条件を低次元空間で表現し、共発現の成立条件を可視化する手法を確立した。この枠組みは「CoexViewer」および「PC View」として実装されている。図 1 は特定の遺伝子ペアにおける発現量の散布図を例示しており、通常のサンプル空間ではなく主成分空間上での分布を可視化することで、共発現がどの主成分軸（すなわち細胞・組織・環境条件）に由来するのかを明確に示すことができる。RNA-seq サンプル数が数万件規模に及ぶ中で、共発現を特徴づける主要な十数個の主成分軸に着目することで、複雑なサンプル条件の構造を効率的に把握することが可能となる。また、サンプルのメタデータには表記揺れが多く含まれるが、各プロジェクト内での一貫性に着目し、プロジェクト単位で共通アノテーションを抽出し、主成分に応じて特徴的に変動するアノテーションを自動的に同定する手法を構築した（ATTED-II v12.0 で公開）。

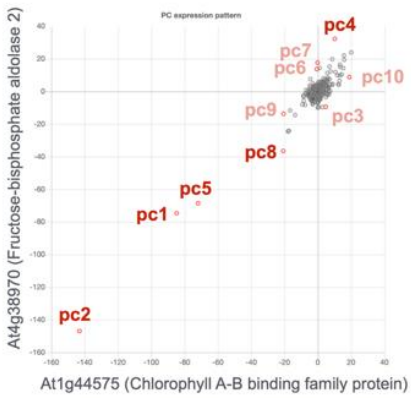

図 1　サンプル主成分を用いた遺伝子共発現の導出と理解

さらに最終年度には、主成分ごとの生物学的特徴を関連遺伝子の KEGG 濃縮検定により定量的に要約するとともに、大規模言語モデル（LLM）を用いて各主成分軸に対するラベルおよび説明文を自動生成する機能を実装した（図 2）。この際、LLM への入力としては、主成分スコアに基づいて選抜した遺伝子群の KEGG 濃縮解析結果と、主成分ローディングに基づいて選抜した代表的なサンプルのアノテーション情報を用いた。具体的には、主成分ローディングの大きいサンプルから代表的な条件を抽出し、共通情報を除いて LLM に入力している。このように、統計的手法により重要なメタデータを選抜したうえで LLM に要約を担わせることで、情報抽出ではなく意味的統合に特化した解釈支援を実現した。ATTED-II version 13.0 では、これらの要約情報をラベル・短い要約・詳細説明の三段階で提供するとともに、CoexViewer 上に統合表示することで、共発現条件の直感的理解を支援する仕組みへと発展させた。これにより、膨大な RNA-seq サンプルの条件構造を統合的に理解し、遺伝子群の共発現関係を生物学的文脈で解釈可能とする解析環境を実現した。

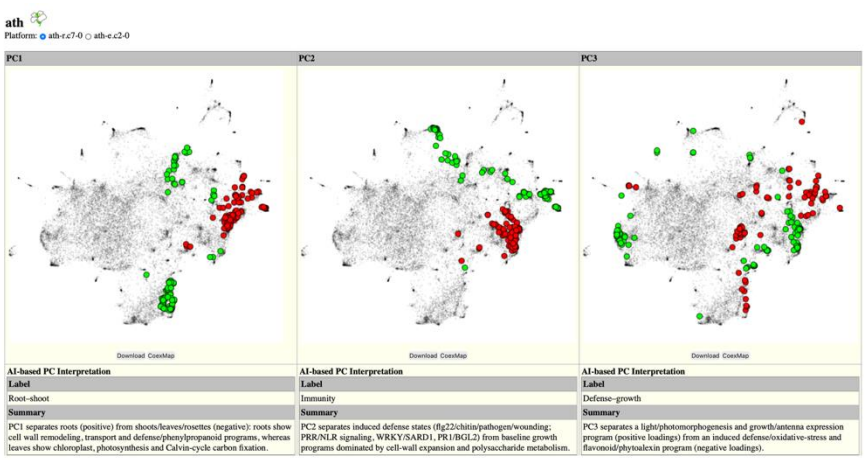

図２　主成分軸の生物学的解釈の自動化と統合表示（ATTED-II v13）

（項目 1C）連携機能の向上

本課題では、異なるデータベースやツールにおいて ATTED-II の共発現情報を二次利用可能とするため、当初計画で掲げた(1)遺伝子 ID 体系の統合と(2)共発現情報の要約・再利用の促進に取り組んだ。まず、Plant GARDEN（2017-2022 年度統合化推進プログラム採択）との相互参照性を強化し、ATTED-II の遺伝子ページから Plant GARDEN および同データベースが提供する JBrowse への直リンクを実装した。これにより、共発現情報から遺伝子多型情報やゲノム構造情報へ直接アクセスできるようになり、遺伝子探索や品種改良研究における利便性を向上させた。

また、非モデル植物を含む各生物種固有の遺伝子 ID と Entrez Gene ID の対応表を整備し、ATTED-II のダウンロードページ（図 3）から入手できるようにした。これにより、各種ゲノムアノテーションや他データベースで一般的に用いられる遺伝子 ID 体系との互換性が向上し、ユーザーが自身の研究環境で共発現情報を容易に統合できるようになった。さらに、共発現データを長期的に保存・共有するため、Zenodo リポジトリを恒久リポジトリとして

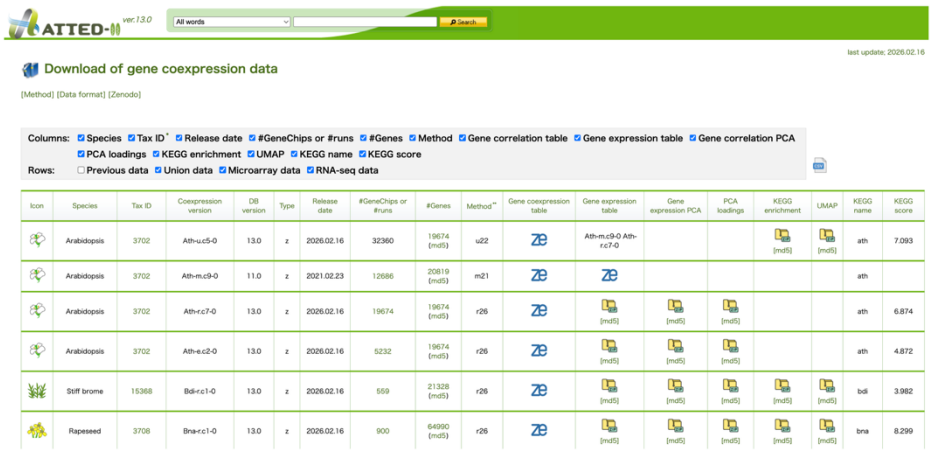

図３　ATTED-II v13におけるデータダウンロードページ

活用した。Zenodo の DOI 付与機能によるデータの引用可能性を確保するとともに、ATTED-II サーバのトラブルの際にもデータの参照が保証される。Zenodo へのデータの登録が完了するまでの期間はローカルサーバからバルクダウンロードファイルを提供し、登録後は半自動的に Zenodo への直リンクに切り替わるようにダウンロードページを整備した。図 3 中の「ZE」は Zenodo にアップロードされたファイルの直リンクを示し、それ以外のリンクはローカルサーバ上のファイルを示している。これにより、共発現データ、主成分解析結果、KEGG Pathway 濃縮解析結果に加え、主成分軸に対する要約情報などを含む ATTED-II version 13.0 の大規模データを、恒久的かつ再利用可能な形で提供できるようになった。API および RDF による外部データベースや解析ツールからの ATTED-II データ利用の仕組みとあわせて、共発現情報の透明性と再現性が向上し、ATTED-II が植物研究コミュニティにおけるデータ統合基盤として機能する体制が強化された。

（項目 1D）公開環境の整備

当初計画には含まれていなかったが、物理サーバ資源の制約とアクセス数の増加により、サービス運用と開発の両立が困難になったことから、2023 年度および 2025 年度に追加実施項目として公開環境の整備を行った。2023 年度の追加実施では、他研究室との共同利用サーバ上でのリソース競合や応答遅延、ネットワーク障害が頻発していた状況を改善するため、ATTED-II 専用の物理サーバを新規導入した。これにより、仮想マシン環境を維持したまま、開発・公開・アーカイブを安定的に運用できる体制を構築した。共発現データを SSD 上に配置し、I/O ボトルネックを解消したことで、共発現リストの表示時間を従来の平均 5〜15 秒から数秒程度に短縮し、100 遺伝子に制限されていたネットワーク描画機能の拡張（最大 500 遺伝子規模）にも対応可能となった。さらに 2025 年度には、アクセス増加と AI クローラを含む大量アクセスへの対応が課題となり、サービス応答性の維持と安定的な開発作業の両立を図るため、開発系と公開系の仮想環境を物理的に分離する構成を整備した。新たに通信帯域・CPU・I/O 性能を強化した公開専用サーバを導入し、既存サーバを開発専用に再構成することで、リソース競合を根本的に解消した。この 2 台構成により、公開環境と開発環境を相互にバックアップできる冗長性を確保するとともに、DDoS 攻撃や AI クローラによる大量アクセス時にもサービスを継続できる耐障害性を備えた運用基盤を実現した。

システム面では、OS を CentOS7 から Rocky Linux 9 へ、Web サーバを Apache から Nginx へ移行し、MariaDB のデータ構造最適化と仮想マシン構成の見直しを行った。また、主要ページ（共発現、NetworkDrawer、CoexViewer など）に非同期通信処理（Ajax）を導入し、応答速度を平均 40%以上短縮した。2025 年 1 月には AI クローラによるアクセス急増（約 320 万アクセス/月）により一時的な不安定化が発生したが、アクセス制限、キャッシュ最適化およびサーバ増強により速やかに安定稼働を回復した。これらの改善により、ATTED-II version 13.0 における生物種拡充および機能強化に対応可能な公開基盤が整備された。

以上の取り組みにより、ATTED-II の公開環境は高い拡張性と冗長性を備え、今後の生物種拡充や機能強化にも耐えうる堅牢な運用基盤が確立された。結果として、国内外からの利用増加に対応しつつ、安定したサービス提供と継続的な開発を両立できる体制を実現した。

遺伝子共発現情報の更新

当初は毎年度の共発現データ更新を計画していたが、対象生物種の拡大に伴う課題として、更新頻度よりも精度検証と自動化を重視した体制へと方針を転換した。これにより、共発現導出から品質評価までを一貫して処理できる RNA-seq 解析パイプラインの再構築を進めた。その成果として、2024 年度には主要作物であるコムギおよびオオムギを新たに追加し、対象生物種を 9 種から 11 種に拡充した ATTED-II version 12.0 を 2024 年 9 月に公開した。さらに 2025 年度には、最終年度目標である 19 種の共発現データ公開を見据え、共発現導出結果を自動評価するパイプラインを整備するとともに、KEGG Pathway による品質評価の際にパラログを考慮する手法を開発した。これらの取り組みの成果として、2026 年 2 月には ATTED-II version 13.0 を公開し、共発現情報を収載する植物種を 19 種 20 系統に拡張した。これにより、モデル植物から主要作物までを含む幅広い植物群に対して、統一的な手法に基づく共発現データを提供することが可能となった。また、構築した自動化パイプラインにより、多様な生物種に対して高信頼かつ比較可能な共発現情報を継続的に更新可能な体制が確立された。

【項目 2】種間比較機能の高度化

（項目 2.1）マクロスケールの種間比較システムの高度化

ATTED-II では、version 11 より全遺伝子を対象とした共発現マップ（CoexMap：UMAP による遺伝子の 2 次元表示）を提供し、種間での共発現構造の比較を可能にしてきた。しかし、従来は表示スケールや座標整列に起因する視認性の制限があり、遠縁種の比較では構造の対応が不十分であった。そこで、UMAP 法を用いた共発現マップに外れ値補正アルゴリズムを導入し、散布図全体のスケーリングを調整することで、マップをできるだけ正方形領域に整形して視認性を向上させた。これにより、複数植物種の共発現マップを比較する際に、構造的対応関係をより直感的に把握できるようになった。また、異なる植物種間で共発現構造を整列させ、進化的に保存された遺伝子モジュールを俯瞰的に評価できる比較基盤を確立した。

さらに、比較対象種の増加に伴い横長の表示が問題となったため、表示する生物種を選択して比較できるインターフェイスを実装した（図 4）。Ajax による非同期通信を導入し、複数種のマップを動的に切り替えながら比較できるようにしたことで、操作応答性も大幅に改善した。ATTED-II version 13.0 では、これらの改良により、多数の植物種に対するマクロスケールの共発現構造を統一的に比較可能な環境を実現した。また、マクロスケール（全遺伝子）に加え、メゾスケール（機能モジュール単位）およびミクロスケール（個別遺伝子ペア）を関連づけて参照できる構成を整備し、スケール横断的な種間比較を支援する環境を実現した。

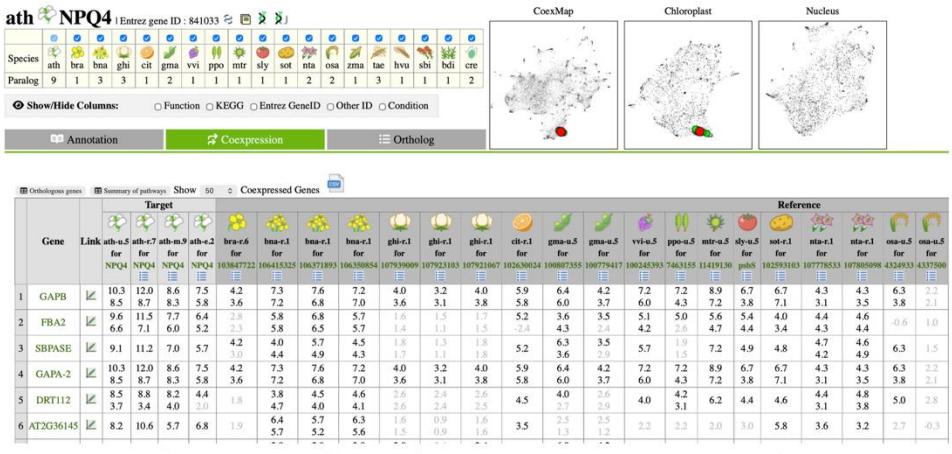

図４　共発現解析における多機能ビューの統合インターフェース（ATTED-II v13）

（項目 2.2）メゾスケールの種間比較システムの開発

マクロスケールの比較では全体構造を俯瞰できる一方で、局所的な共発現構造の差異を十分に捉えにくい課題があった。図 5 上段は全遺伝子（約 25,000 遺伝子）を対象とした共発現マップであるが、生物種間でマップ構造が必ずしも相同とはならず、遠縁種どうしの比較が困難であることが分かる。一方で、図 5 下段に示す葉緑体局在遺伝子（約 4,000 遺伝子）のみを対象とした共発現マップでは構造の相同性が高く、遠縁の植物種間でも対応関係を捉えられることを見出した。同様の傾向は細胞核に局在する遺伝子群においても確認された。

細胞内局在は生物に普遍的な細胞構造に基づく分類であり、この情報に基づいて遺伝子集合を制限することで、全遺伝子を単一の二次元空間に投影した際に失われる局所的な共発現パターンを補完し、機能モジュール単位での種間比較を可能とすることができる。この考えに基づき、細胞内局在別に共発現マップを構築するメゾスケールの比較手法を確立した。

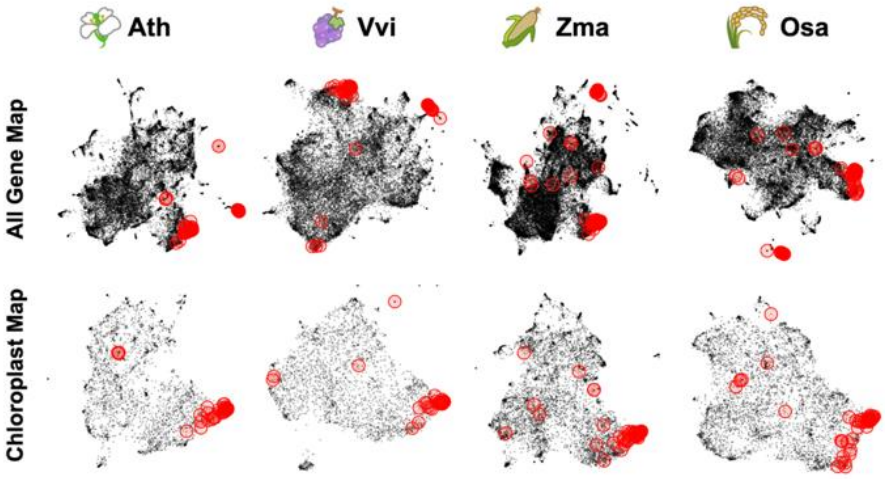

図５　葉緑体局在遺伝子のメソスケール共発現種間比較

ATTED-II version 13.0 では、このメゾスケールの比較概念を実装し、細胞内局在に基づく共発現マップを複数植物種間で比較可能な形で提供することで、マクロスケールの比較では捉えにくい機能モジュールの保存性および差異を評価できる環境を実現した。

(3) 主な成果論文等

1. Kanako Bessho-Uehara, Takeshi Obayashi. Evolutionary approaches for narrowing down candidate genes from an unannotated gene list, Plant and Cell Physiology, 2025,66,287-290 (DOI: 10.1093/pcp/pcaf003).
   [概要] 非モデル植物を含む多様な植物種における遺伝子機能解析の進化学的アプローチを概説し、共発現ネットワークとオーソログ情報を活用した知識移転の重要性を示した。本課題が目指す「種間比較による遺伝子機能推定基盤」の方向性を学術的に位置づけるものである。
2. Takeshi Obayashi. Subagging of Principal Components for Sample Balancing: Building a Condition-Independent Gene Coexpression Resource from Public Transcriptome Data, *Function COSI, ISMB/ECCB 2023* (口頭発表, 2023 年 7 月 26 日).
   ［概要］ RNA-seq 共発現解析における主成分サバギング法を提案し、サンプルバランス補正による高精度な共発現導出法が ISMB/ECCB にて口頭発表採択となった。深層学習を活用した報告が大多数を占め、過学習の問題が議論される中で、記述統計的アプローチの高度化により共発現精度を高める本研究は意外性をもって受け止められ、活発な議論を呼んだ。

3. 受賞: Takeshi Obayashi et al. ATTED-II v11: a plant gene coexpression database using a sample balancing technique by subagging of principal components, Plant and Cell Physiology, 2022, 63, 869-881.
[概要] 本課題の基礎となる主成分分析による遺伝子共発現法を報告した ATTED-II の論文であり、2026 年度 PCP Top Cited Paper Award を受賞した（授賞式 2026 年 3 月 14 日）。

(4) 主要なデータベースの利活用状況

Google Analytics による月平均アクセスは、2023 年度約 2,200 件、2024 年度約 2,100 件、2025 年度（9 月末時点）約 3,300 件と、着実な増加傾向を示した。一方、AWStats によるサーバログ解析では、2024 年後半以降、外部サービスや自動プログラムによる API アクセスが急増し、2025 年 1 月には月間 320 万アクセスを超えるピークを記録した。これらのアクセスの多くは、Araport など海外研究者によるシロイヌナズナ遺伝子情報参照や、他データベースとのリンクチェックなどの自動取得に加え、近年の AI 開発の基盤としての利用を目的としたクローラによるものと考えられる。320 万アクセスを記録した 2025 年 1 月にはサービス提供が断続的に停止するなどの問題が生じたが、その後、アクセス制限、物理サーバの増強およびシステムの高速化により、現在は安定したサービス提供を実現している。これらの結果から、ATTED-II は国内外の研究者による継続的な参照に加え、API を通じた機械的データ取得および外部データベースとの連携において重要な役割を果たしており、ATTED-II version 13.0 の公開後は、種数拡充および機能強化に伴い、その利用基盤としての重要性がさらに高まっていると判断される。

(5) データベースを利用して得られた研究成果・産業応用の例

ATTED-II は多くの学術研究で活用されているだけでなく、特許出願や企業研究においても共発現情報が参照されている。これらは、同データベースが代謝改変や機能性成分生産などの応用領域において、分子設計や候補遺伝子抽出を支える基盤ツールとして機能していることを示している。

3. 今後の計画および展望

（1）ATTED-II の開発計画

本研究開発により、ATTED-II version 13.0 を公開し、共発現情報を収載する植物種の拡張と、主成分解析に基づくサンプル条件の解釈機能および種間比較機能の高度化を実現した。今後は、これらの成果を基盤として、ATTED-II version 13.0 の論文化を進めるとともに、対象生物種のさらなる拡充および比較解析機能の高度化を継続する予定である。特に、遺伝子共発現と配列類似性を統合した解析手法や、系統群単位での同時クラスタリングの導入により、進化的関係と共発現構造を統合的に理解する枠組みの構築を目指す。また、主成分軸の解釈支援についても、LLM による要約の精緻化および外部知識との統合を進めることで、より高度な生物学的解釈を可能とする方向へ発展させる。これらの開発を通じて、共発現情報をゲノム配列と機能モジュールを結ぶ中間レイヤーとして体系化し、非モデル植物を含む多様な生物種における遺伝子機能解析の基盤としての役割をさらに強化することを目指す。

（2）他分野への展開と中長期的展望

本課題で確立した遺伝子共発現に基づく機能推定および条件解釈の枠組みは、植物に限らず、環境応答を反映するオミクスデータ全般に適用可能である。この特性を踏まえ、既存の動物共発現データベース COXPRESdb においては、本課題で開発した手法を比較的直接的に適用することが可能であり、ヒトと実験動物の種間比較に加え、海洋性動物など非モデル生物種への応用を進めることで、環境科学や食の安全といった分野への展開が期待される。

一方で、海洋プランクトンを対象とするエピゲノミクス研究では、ゲノム情報や遺伝子アノテーションが未整備な生物種を扱う必要があるため、より挑戦的な課題となる。現在開発中のメタエピゲノミクスプラットフォーム PlanDyO において、本課題で開発した共発現モジュール推定および機能アノテーション技術の適用可能性の検証を進めている。エピゲノミクスはトランスクリプトームと同様に環境応答を反映することから、共発現モジュールの概念を導入することで、機能アノテーションやパスウェイ推定、細胞状態解析への新たな展開が期待される。

このように、共発現データベースは生物機能の理解に加えて生態系の理解を支える基盤情報として重要性を増しており、本課題で構築した基盤をエピゲノミクスや環境ゲノミクスへ拡張することで、将来的にはアカデミアのみならず産業界や環境政策分野にも貢献する汎用的なデータ基盤の構築を目指す。

4. 計画・実施体制等の妥当性

(1) 各グループの担当項目

(1)-1. 大林グループ（東北大学）

大林グループでは、植物遺伝子共発現データベース ATTED-II の開発・運用を中心に、非モデル植物への展開を目的とした技術基盤の整備を担当した。具体的には、RNA-seq データに基づく共発現解析パイプラインの高精度化、主成分分析によるサンプル条件可視化機能（PC View）の実装、種間比較機能の高度化、公開環境の刷新と安定化を主導した。また、Arabidopsis の共発現データと Plant GARDEN のゲノム多型情報との連携を実現し、外部データベースとの相互運用性を高めた。

研究代表者が全体統括を担い、開発計画の策定から評価・公開までを一貫して管理し、研究メンバー間の役割を明確に分担することで、効率的な開発体制を維持した。単一チーム構成ながら、DICP 研究交流会やトーゴーの日シンポジウムなどを通じて他チームおよび DBCLS との情報交換を行い、データ共有や技術連携の観点からオープンな開発体制を確立した。特に、DICP 研究交流会での議論を契機として、大規模言語モデル（LLM）を用いた主成分サンプル要約のアイデアが生まれ、PC View の自動説明生成機能として発展した。

また、若手研究者がデータ解析、可視化機能開発、アクセスログ解析などを担当し、プロジェクトを通じてソフトウェア開発やデータベース運用に関する実践的スキルを獲得した。これらの体制により、ATTED-II version 13.0 における機能拡張および生物種拡充を実現し、モデル植物で確立された共発現解析を非モデル植物へ拡張するための基盤を整備するとともに、安定的かつ国際的に信頼されるデータ提供体制を確立した。