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）

2023 年度 研究開発実施報告

概要

研究開発課題名	（和文）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 （英文）Platform of gene network information for non-model plants
開発対象データベースの名称（URL）	ATTED-II (https://atted.jp/)
研究代表者氏名	大林　武（50397048）
所属・役職	東北大学 情報科学研究科　教授（2024 年 3 月時点）

目次

§1. 研究実施体制

グループ名	研究代表者または主たる共同研究者氏名	所属機関・役職名	研究題目
大林グループ	大林　武	東北大学・教授	遺伝子ネットワーク情報活用基盤 ATTED-IIの開発

§2. 研究開発対象とするデータベース・ツール等

(1) データベース一覧

【主なデータベース】

No.	名称	別称・略称	URL
1	植物の遺伝子共発現データベース ATTED-II	ATTED-II	https://atted.jp/

【その他のデータベース】

No.	名称	別称・略称	URL
1			

(2) ツール等一覧

No.	名称	別称・略称	URL
1			

§3. 実施内容

(1) 本年度の研究開発計画と達成目標

1) 共発現提供生物種の情報拡充のためのシステム開発

① RNAseq 共発現の高精度化

RNAseq は非モデル植物における代表的なトランスクリプトーム解析技術であるとともに、エコタイプの違いを扱いやすいなど、マイクロアレイにはない特性を持っており、非モデル植物の共発現解析の中心的なデータソースである。そのため、この RNAseq から高精度に共発現を導出する手法の開発を行う。2023 年度は非標準種を対象とする RNAseq データの影響を解析し、それに基づく高精度な共発現情報を抽出するパイプラインを構築する。特に多くの非標準種トランスクリプトームデータがすでに公共データベースに蓄積しているシロイヌナズナにおいて手法の開発と評価を行う。

② 共発現モジュールが機能するサンプル条件を提示する機能の開発

遺伝子発現プロファイルの類似性として計算される遺伝子共発現の値は、着目する遺伝子ペアがどの程度機能的に関連しているかを定量的に示す指標として利用できる。一方、この値は単純な値（スカラ）であるため、着目する共発現遺伝子ペアが、いつ、どのような細胞環境で連携して働くかはわからない。そこで、ある遺伝子ペアが共発現するサンプル条件を提示することで、共発現の解釈を向上させるためのシステムを構築する。共発現データを構築するためのサンプル数は、2021 月 2 月 23 日に公開したシロイヌナズナの RNAseq 共発現（Ath-r.c5-0）で 14,741 サンプルであり、今後も増えていくことが予想される。このサンプル数の規模感を踏まえて、関連するサンプル条件の提示は、「大まかな条件の傾向」と「詳細なサンプル条件の提示」の 2 段階でサンプル条件を提示する戦略とする。開発初年度である 2023 年度は、大まかな条件の傾向を得るため、主成分分析によって再構成したサンプル条件を提示するビューアを構築する。

③ データベース間の連携機能の高度化

ATTED-II で提供する共発現情報を、他のデータベースやツールで利用するための機能を開発する。ATTED-II は遺伝子の主キーとして NCBI Entrez Gene ID を用いており、他の ID への変換機能を組み込むことで、データベース横断的な利用を促進する。特に、非モデル植物研究の基盤となるゲノム情報との連携は不可欠であることから、統合化推進プログラム（2017~2021 年度）の支援で開発された植物ゲノムポータルの Plant GARDEN [https://plantgarden.jp/] と、遺伝子ごとに相互リンクを設置することで（主に標準種を対象とした）共発現情報とゲノム多型情報を相互参照できる仕組みを構築する。

また、共発現情報に基づき遺伝子機能アノテーションを自動で付与するシステムを構築し、共発現情報を遺伝子アノテーションとして簡便に利用できるようにする。2023 年度は自動付与の手法の比較を行う。

2) 種間比較機能の高度化

ATTED-II では遺伝子ペア単位で種間比較を行う仕組みがあり、共発現遺伝子ペアの進化的保存性を吟味することができる。一方で、遺伝子共発現ネットワークとしては種間の違いが大きく、容易に比較検討行うことができない。そのため、より広域な比較に着目することで、種間比較を行うツールを開発する。2023 年度はゲノムワイドな遺伝子共発現マップを種間比較するビューアを開発する。

3) 遺伝子共発現情報の構築と更新

コムギ、オオムギの共発現情報を新規に構築する。また、ATTED-II v11 の 9 生物種の共発現情報を更新する。共発現情報ならびにそれに付随する情報を ATTED-II v12 として提供する。

合わせてウェブサーバの移行とウェブデザインの見直しを行い、ユーザビリティを向上させる。まず、現行の ATTED-II（v11.1）の応答速度の向上させるために、仮想マシンをより高速なウェブサーバにて実行する。さらに、データベース間連携を促進するためのユーザインターフェイスの改良を行う。具体的には、共発現遺伝子リストを提示するページ［ https://atted.jp/coex/(gene_id) ］を、遺伝子ネットワークを表示する遺伝子のページ［https://atted.jp/locus/(gene_id)］に統合化し、外部データベース（あるいはツール）から ATTED-II 共発現情報に遺伝子単位でリンクする際のリンク先を一本化する。

4) 公開環境の整備（2023 年 7 月 24 日に追加）【追加支援】

ATTED-II の応答遅延やトラブルを解消し、ユーザビリティを大幅に向上させるために、専用の物理計算サーバーを新たに調達し、運用する。新サーバーにより連続利用可能性と遺伝子数制限の緩和が期待され、安定的なシステム運用を実現する。

(2) 進捗状況

1) 共発現提供生物種の情報拡充のためのシステム開発

① RNAseq 共発現の高精度化

まずシロイヌナズナにおいて非標準種（すなわち Col-1 以外）の RNAseq を、標準種（Col-1）の RNAseq データから分離し、各々遺伝子共発現データを構築し、KEGG アノテーションとの一致度である、KEGG スコアを用いて評価した。

- 標準種（Col-1）のみのシロイヌナズナ RNAseq 共発現データセット（Ath-r）：7.495 (22,408 サンプル)
- 標準種（Col-1）を含まない非標準種のみのシロイヌナズナ RNAseq 共発現データセット（Ath-e）：5.279 (5,256 サンプル)
- 参考：v11.1 の Ath-r：7.275 (14,741 サンプル)

非標準種を含む RNAseq 共発現が、サンプル数から期待する KEGG スコアよりも低い結果となり、マッピング方法やバッチ単位について検討する必要があると考えられた。そこで、エコタイプをバッチ補正の共変量としてバッチ補正を行ったが、明確な精度向上には繋がらなかった。一方で、マッピング方法をアライメントフリーの MATATAKI からアライメントベースの bowtie2 に変更すると若干の改善が確認できた。これは非標準種の利用ではある程度の配列多様性が生じてしまうことから、それを前提とした RNAseq データの定量化が必要であると解釈できる。なお、遺伝子発現量のデータ解析ではバッチ効果の影響がその後のデータ解析に甚大な影響を与えることから、ATTED-II におけるデータ処理では一貫して実験単位で操作を行っている。2023 年度に実施した標準種と非標準種の分離においても実験単位で分離しているため、非標準種を含む実験セット（Ath-e）には、コントロール的な位置付けとして標準種も含まれていることがある。

非標準種の共発現データの解析と並行して、単一生物種のサンプルを亜種などで分類して共発現データを構築するためのパイプラインを構築し、開発サーバ上で動作することを確認した。

RNAseq の定量化方法の検討は、異質倍数体であるコムギの RNAseq 共発現の構築でも重要である。コムギの RNAseq に対して、MATATAKI と bowtie2 による RNAseq の定量化ならびに共発現の導出を行ったので、次年度も継続して検討を進める。

② 共発現モジュールが機能するサンプル条件を提示する機能の開発

遺伝子発現プロファイルの類似性として計算される遺伝子共発現の値は、着目する遺伝子ペアがどの程度機能的に関連しているかを定量的に示す指標として利用できる。一方、この値は単純な値（スカラ）であるため、着目する共発現遺伝子ペアが、いつ、どのような細胞環境で連携して働くかはわからない。そこで、ある遺伝子ペアが共発現するサンプル条件を提示することで、共発現の解釈を向上させるためのシステムを構築する。関連するサンプル条件の提示は、「大まかな条件の傾向」と「詳細なサンプル条件の提示」の 2 段階でサンプル条件を提示する戦略とする。開発初年度である 2023 年度は、大まかな条件の傾向を得るため、主成分分析によって再構成したサンプル条件を提示するビューアのプロトタイプを構築し、開発サーバにて設置した。次年度、引き続き検証しながら機能の高度化を進めていく。

③ データベース間の連携機能の高度化

これまでの ATTED-II では、遺伝子アノテーション、遺伝子共発現、オーソログを別々のページで提供していたが、これらは互いに強く関連するため、統一的に扱うことができるユーザインターフェースの構築を進めた。このことは外部のデータベースとの連携において、ATTED-II の利用を促進するために重要あり、ページデザインが確定し次第、Plant GARDEN とのリンクを行う。合わせて、ATTED-II の API の整備を実施し、データ提供機能の向上を行った。遺伝子機能アノテーションの付与方法については、広く利用されている Gene Set Enrichment Analysis などの統計的手法よりも、遺伝子共発現の値を直接用いて推定する方法が良いことを見出した。

2) 種間比較機能の高度化

ゲノムワイドな遺伝子共発現マップを種間比較するためのビューアの開発において、現在用いている UMAP 法による次元圧縮では次元圧縮後の全体形状についての制約がなく、極度の外れ値が生じることがある。このことは単一生物種での共発現マップの利用だけでなく、種間のアラインメントでも問題となる。そのため、この外れ値に対して適切な位置に配置し直すアルゴリズムを実装し、共発現マップの視認性と整列可能性を向上させた。今後、複数の共発現マップを比較するための評価法、ならびに解析ツールの開発を行っていく。

3) 遺伝子共発現情報の構築と更新

コムギ、オオムギの共発現情報を新規に構築するとともに、ATTED-II v11 の 9 生物種の共発現情報を更新し、開発サーバで閲覧できるようにした。ATTED-II v11.1 と新しく構築した RNAseq 共発現データについ

て、以下に各生物種の RNAseq run の数と KEGG スコア（大きい方が KEGG アノテーションと一致する）を示す。なおオオムギは KEGG アノテーションが利用できないため、評価値はない。

	#Runs	KEGG スコア
Arabidopsis:	14741→22408	7.275 → 7.495
Field mustard:	212→388	6.944→6.410
Soybean:	1082→2141	5.128→5.474
Barley:	294	(評価値なし)
Medicago:	310→730	5.404→5.553
Rice:	778→1149	4.719→4.672
Poplar:	648→868	4.211→4.330
Tomato:	749→1474	6.023→6.767
Wheat:	1667	10.314
Grape:	1235→1436	4.166→4.282
Maize:	4411→5485	6.428→6.428

RNAseq run の数量が上昇し、軒並み KEGG スコアが向上する結果となった。ただし、RNAseq の定量法に関して検討の余地が残っており、ATTED-II v12 としての年度内公開は見送った。マッピング効率などの検討を行い、2024 年度上半期の公開を予定している。

また、速度向上のため仮想マシンをホストするホストサーバの移行を行った。合わせてユーザビリティの向上のため、改良したユーザインターフェイスのプロトタイプを作成した。このインターフェイスは改善の余地があり、引き続き次年度に検討を続け、公開版に反映させる。

4) 公開環境の整備【追加実施】

ATTED-II の提供用に新しい物理サーバーを導入し、提供中の ATTED-II v11.1 が動作することを確認した。セキュリティ設定などの確認ののち 2024 年上半期から実運用を行う。

§4. 成果発表等

(1) 原著論文発表

① 論文数概要

種別	国内外	件数
発行済論文	国内(和文)	0 件
	国際(欧文)	0 件
未発行論文 (accepted, in press 等)	国内(和文)	0 件
	国際(欧文)	0 件

② 論文詳細情報

該当なし

(2) その他の著作物(総説、書籍など)

該当なし

(3) 国際学会および国内学会発表

① 概要

種別	国内外	件数
招待講演	国内	3 件
	国際	0 件
口頭発表	国内	1 件
	国際	1 件
ポスター発表	国内	2 件
	国際	0 件

② 招待講演

〈国内〉

1. 大林武、Platform of Gene Network Information for Non-model Plants、統合化推進プログラム(DICP)キックオフ・ミーティング、オンライン、2023 年 5 月 24 日
2. 大林武、遺伝子共発現データベース ATTED-II における種固有の共発現情報の導出と利用の展望、第 1 回北海道バイオ"Mix up"、北海道大学、2023 年 8 月 8 日
3. 大林武、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、トーゴーの日シンポジウム、日本科学未来館、2023 年 10 月 5 日

〈国際〉

該当なし

③ 口頭講演

〈国内〉

1. Obayashi T、Enhancing ATTED-II Database for Diverse Plant Species Research、第 65 回日本植物生理学会年会、神戸国際会議場、2024 年 3 月 17 日

〈国際〉

1. Obayashi T. Subagging of Principal Components for Sample Balancing: Building a Condition-Independent Gene Coexpression Resource from Public Transcriptome Data, Function COSI, ISMB/ECCB 2023 (Intelligent Systems For Molecular Biology / European Conference On Conputational Biology) 26 Jul 2023, Lyon, France.

④ ポスター発表

〈国内〉

1. 大林武、共発現データベースATTED-IIにおける種間比較の取り組み、統合化推進プログラム(DICP)研究交流会、日本科学未来館、2023年10月5日

2. 火原日美子. 共発現データベースATTED-IIにおけるバージョン管理, 統合化推進プログラム(DICP)研究交流会、日本科学未来館、2023年10月5日

(4) 知的財産権の出願 (国内の出願件数のみ公開)

出願件数

種別		件数
特許出願	国内	0件

(5) 受賞・報道等

① 受賞

該当なし

② メディア報道

該当なし

③ その他の成果発表

該当なし

§5. 主要なデータベースの利活用状況

(1) アクセス数

1) 実績

表 1 研究開発対象の主要なデータベースの利用状況(月平均)

DB 名	種別	2023 年度
ATTED-II	訪問者数	1,067
	訪問数	2,244
	閲覧ページ数	14,588

※ Google Analytics の結果

2) 分析

AWStats と Google Analytics によるアクセス統計が大きく異なるため数値の妥当性の判断が難しい。上記は Google Analytics の結果を示す(Analytics のタグを設置していない API を含んでいない)。この数値は、過去 5 年間と比較して数割減少している。Google Analytics のバージョンが GA4 に上がったことが関連しているのかもしれない。

AWStats は年度の前半後半で、集計手法を変更した。年度前半では、ホスト OS の nginx のログ、年度後半は仮想マシンの nginx のログを用いている。前半の数値としては、月間で Unique User: 3,000, TotalVisits 8,000, Page View 66,593 と、Google Analytics の数倍になっている。

一方で、年度後半の AWStats の統計では、月間 Page View が 4 万から 85 万と大きくばらついている。ホスト OS からのアクセスの転送のため、ユーザの IP アドレスの取得ができておらず、クローラの判定精度が下がっているかもしれない。アクセスログはユーザ同行を把握するための重要な情報源であるため、今後、AWStats をホスト OS にも設置し、アクセスログ解析の精度向上を目指す。

(2) データベースの利用状況を示すアクセス数以外の指標

Google Scholar によると 2023 年に出版された ATTED-II を引用した論文は 88 件だった。これは 2018 年から 2023 年の各年の被引用数が 111 件, 109 件, 108 件, 101 件, 108 件であったことと比べて、若干減少している可能性がある。

(3) データベースの利活用により得られた研究成果(生命科学研究への波及効果)

以下の 10 件は全て研究参加者と直接関係がない研究グループによる研究成果。1 件目の論文(Apodiakou and Hoefgen, 2023)については、メール、オンライン打合せ、対面打合せにより ATTED-II 利用のコンサルティングを行なった。

1. O-アセチルセリン(OSA)の蓄積とともに誘導されるシロイヌナズナ遺伝子である OSA 遺伝子群は、OSA

が蓄積する硫黄欠乏条件、暗黒条件、活性酸素種蓄積条件などで誘導される。本総説論文では、ATTED-II 遺伝子共発現ネットワークを基盤情報として、OSA 制御経路の全体像を議論するとともに、既存の Dap-seq データなどの遺伝子制御データと組み合わせることで遺伝子制御ネットワークの全体像を議論し、特に硫黄欠乏関連転写因子である EIL3/SLIM1 が重要な機能を担うと結論づけた。
Apodiakou A, Hoefgen R. New insights into the regulation of plant metabolism by O-acetylserine: sulfate and beyond. J Exp Bot. 2023;74: 3361–3378.

(4) データベースの利活用によりもたらされた産業への波及効果や科学技術のイノベーション（産業や科学技術への波及効果）

ゲノム配列決定が容易になった現在、遺伝子機能解析を行うために遺伝子共発現情報を利用するニーズはこれまでになく高いと想像している。Google Patents を ATTED-II で検索すると、国内外の特許が 19 件ヒットする。これらを個々に精査し、実際に ATTED-II を利用した結果であることを確認した。企業の研究者からは、ATTED-II の生物種の拡張について議論することもあり、本プログラムにより、産業への波及効果が大きく増すことを期待している。

§6. 研究開発期間中に主催した活動(ワークショップ等)

(1) 進捗ミーティング

年月日	名称	場所	参加人数	目的・概要
2023 年 9 月 2 日	チーム外ミーティング(非公開)	大阪	4 人	硫黄代謝経路解析における遺伝子共発現情報の利用のコンサルティング
2023 年 10 月 3 日	チーム内ミーティング(非公開)	東北大学	4 人	研究進捗報告のためのミーティング
2023 年 12 月 15 日	チーム内ミーティング(非公開)	東北大学	4 人	同上

(2) 主催したワークショップ、シンポジウム、アウトリーチ活動等

年月日	名称	場所	参加人数	目的・概要

以上

別紙1　既公開のデータベース・ウェブツール等

No.	正式名称	別称・略称	概要	URL	公開日	状態	分類	関連論文
1	植物の遺伝子共発現データベースATTED-II	ATTED-II	代表的な植物の遺伝子共発現情報を提供するデータベース	https://atted.jp	2004/11/14	維持・発展	データベース等	Obayashi T, Hibara H, Kagaya Y, Aoki Y, Kinoshita K. (2022) ATTED-II v11: A Plant Gene Coexpression Database Using a Sample Balancing Technique by Subagging of Principal Components. Plant Cell Physiol. 15;63(6):869-881. doi: 10.1093/pcp/pcac041.