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(統合データ解析トライアル)

研究開発課題
「生化学反応ネットワーク統合解析環境の拡張」

研究開発終了報告書

研究開発期間:平成26年9月～平成27年2月

研究代表者:西田孝三
(理化学研究所 生命システム研究センター
テクニカルスタッフ)

§1 研究開発の概要

昨年度研究開発を行ったネットワーク解析環境への RDF 技術と代謝物プロファイル解析フレームワークの導入、そして生物学ネットワーク可視化プラットホーム Cytoscape(Smoot, 2011) が掲げる Cyberinfrastructure 上でのパイプライン化を試みた。ここでの Cyberinfrastructure とはデータの取得、統合、可視化を一つの機構上で行う環境を意味する。近年オミクス情報は測定技術の革新によって様々な分子レベルで網羅的に測定できるようになった。しかしながら代謝物プロファイルデータベースの統合に対応したソフトウェアは未だ発達していない。

今年度は代謝物プロファイル情報からデータ統合・解析・可視化まで一通りの解析が行える再現性の高い解析パイプラインを提供することをねらった。パイプラインの使用例にはシロイヌナズナの代謝物プロファイルデータベース AtMetExpress を用い、そのパイプラインは IPython Notebook(Perez, 2007)形式で公開した。RDF の活用までは至らなかったが、成果として代謝物プロファイルのパスウェイ統合の再現とその改変を利用者が容易に行うことが可能な統合解析環境の提供を実現した。

§2 研究開発のねらい

マルチオミクス情報は近年の測定技術革新(新型シーケンサによる RNA-seq など)によって遺伝子発現・代謝物といった様々な分子レベルで網羅的に測定できるようになった。また代表的パスウェイデータベースである KEGG(Kanehisa, 2014)は扱う情報を DRUG、DISEASE、Metagenome へと拡大しデータは増加の一途を辿っている。しかしながらトランスクリプトーム情報とパスウェイデータベースの統合を行うソフトウェアの進歩に対し、代謝物プロファイルをパスウェイ情報と統合・活用するソフトウェアは未だ発達していない。

昨年、研究代表者は KEGG データベースに含まれない DRUG・ゲノムアノテーション・トランスクリプトーム情報を統合し KEGG パスウェイを拡張するソフトウェア環境の構築を行った。本年度の開発では代謝物プロファイルデータベースの情報に基づき前年度成果の拡張と改善を行う。代謝物プロファイルデータベースを実用例として用いるために昨年度は大腸菌としていた解析対象をシロイヌナズナに移し、昨年度トランスクリプトームのみとしていた入力、解析対象を代謝物プロファイルに拡張する。この拡張による汎用性の向上により前年度の成果の改善とデータ統合の推進をねらう。

昨年度の成果において研究代表者は KEGG パスウェイに不足するデータ(代謝モデル、ドラッグターゲット)の統合を行うことで KEGG パスウェイの情報のみでは解釈できないオミクスデータに対し知見が得られることを示した。今年度の研究開発では代謝物プロファイルデータベースに対しても同様に KEGG パスウェイとの統合を行い前年度成果の拡張を行う。代謝物プロファイルデータベースとしては前年度の統合データ解析トライアルの成果の一つである AtMetExpress(http://atmetexpress.riken.jp/)があるが、その情報とパスウェイを統合するためのネットワーク解析環境は無い。

本研究では AtMetExpress 中のシロイヌナズナの代謝物プロファイルの実験条件間差異を KEGG パスウェイ上での可視化によって容易に確認することが可能な解析環境を構築する。KEGG パスウェイ上での可視化を主とした解析環境の先行研究例としては Pathview(Luo, 2013)があるが、代謝物プロファイルの統合解析とその可視化例の提示はまだ行えていない。

代謝物プロファイルの統合において活用可能な化合物やパスウェイ情報の RDF とそのエンドポイントには PIERO(Kotera, 2014), ChEBI(Hastings, 2013), Reactome(Croft, 2014), MetaCyc(Capsi, 2014)があるが、これらと代謝物プロファイルをリンクさせ解析パイプラインを提供している解析環境は無い。RDF の活用により、 オントロジーの容易な拡張やスキーマに囚われない横断検索の実現、自動処理時におけるデータのパースや API の設計が不要になるといった利点が得られる。しかしながら代謝物のプロファイルの解析パイプラインと RDF を基盤としたソフトウェア環境は未だ発達して

いない。
パイプラインは「KEGG と Cytoscape をつなぐ拡張機能 KEGGscape(Nishida, 2014)」、「Cytoscape とパイプライン実行用途としての Python 言語(https://www.python.org/)をつなぐ拡張機能 cyREST(https://github.com/idekerlab/cyREST)」、「再現性が高く処理の変更も容易な Python 言語実行環境 IPython Notebook」の三つを組みあわせ構築を行う。 バイオインフォマティクス汎用パイプライン実行環境の先行研究例としては Galaxy(Goecks, 2010)があるが、Galaxy と比較し IPython Notebook は利用者が Python 言語を用いる必要があるもののパイプラインの変更が容易という利点がある。

§3 研究開発計画

(1)当初の研究開発計画
下記の三項目の順に研究開発を進めることを計画していた。

(I) KEGG とシロイヌナズナのパスウェイ情報の RDF を利用した統合
ChEBI の owl と代謝物プロファイルの ID をリンクさせた RDF データを利用し ChEBI のオントロジーを用いた AtMetExpress 中の代謝物のエンリッチメント解析に利用する。

(II) シロイヌナズナのオミクスプロファイルの Bioconductor を用いた解析とネットワーク解析プラットフォーム Cytoscape 上でのパスウェイ情報可視化のパイプライン化
R 言語と Python 言語から Cytoscape を操作する RESTFUL API を利用し、Bioconductor と Cytoscape をシームレスにプログラム操作するための開発を行う。既存研究においては app とよばれる拡張機能を Java 言語で記述し Cytoscape 内で解析を完結させるものが大半を占める。一方本研究は Pathview のようにその操作のほとんどを外部の非 Java 言語からプログラマブルに行えるように設計を行うことで、パイプラインの変更を容易にするとともに解析の再現性を高める。

(III)パイプライン処理実行用ウェブブラウザーUI の実装
ウェブブラウザーから 1) 2)で挙げた処理をパイプラインとして実行するためのプログラムを IPython notebook を用い実装する。IPython notebook を UI として用いることで、本研究開発の成果(データ取得統合・解析の実行・可視化)はインタラクティブかつプログラマブルに再現でき、またその情報共有の容易さからコミュニティによる発展が期待できる。

(2)新たに追加・修正など変更した研究開発計画
(I) に関わる変更
後述の理由により今年度の RDF の利用を見送った。まず本研究の目的と ChEBI の化合物のオントロジーが有用な領域が異なっていた。本研究が求める化合物のオントロジーが原子の由来関係に基づいた代謝産物の生合成経路に関するものであるのに対し、ChEBI のオントロジーがターゲットとするものは構造情報が意味を持つタンパク質間相互作用や薬物活性のような領域であった。
化合物だけでなく反応も含んだシロイヌナズナの代謝物プロファイル解釈のためオントロジーは PIERO オントロジーや KCF-S(Kotera, 2013)とも異なっていた。特にフラボノイドなど AtMetExpress で確認される二次代謝物の理解に重要な化合物、反応情報については KEGG の ID と一対一の対応を取ることができず、新しく反応ネットワークのオントロジーの設計を行う必要があることがわかった。これは KEGG だけでなく AraCyc(Zhang, 2005)など他のパスウェイデータベースに対しても同様であった。
また代謝物プロファイルにおいては発現プロファイルと比較しアノテーションシステムが確立されておらず RDF 設計を根本的に見直す必要があった。当初の研究開発計画においては同定

後の化合物 ID が代謝物プロファイルの RDF の URI になるものと想定していたが代謝物プロファイルにおける ID はスペクトル情報であった。マイクロアレイを用いた発現プロファイルにおける ID はプローブ配列でありパスウェイとの対応付けが公開、共有されているが代謝物プロファイルにおける ID(スペクトル情報)に対してはそのような情報基盤が全く確立されていないことが研究開発後にわかった。メタボローム実験においては各研究者・研究機関が保有している標準品が異なること、さらには標準品が無い場合にはスペクトルライブラリとの比較によりスペクトルの化合物を推定せざるを得ない、といった事情が裏にあり、アノテーションのレベルの差異が大きかった。このアノテーションの問題は解析パイプラインの入力定義に関わるだけでなく RDF の利点(意味定義の明確化、L 体 D 体の推論、組織を超えたデータ利用など)を最も生かすことできる重要な箇所だが、計画の規模が大きすぎたため今年度での解析トライアルでの着手を断念した。

§４ 研究開発成果

AtMetExpress のデータの内 GC-MS により測定を行ったシロイヌナズナの葉と根の時系列代謝物プロファイルの組織間差異の可視化を行う解析パイプラインの公開を Github(https://github.com/kozo2/togotrial2014)にて行った。この解析パイプラインを構成する各機能モジュールは KEGGscape(https://github.com/idekerlab/KEGGscape) 、keggutil(https://github.com/kozo2/keggutil)であり同様に Github にて公開を行っている。本解析パイプラインを用いることにより、任意の KEGG パスウェイ上で組織間の代謝物プロファイルの差異が確認できる(別紙参照)。KEGG パスウェイは代謝物プロファイルをマップする用途に活用しており、AtMetExpress の Col0Timecourse データセットの葉、根の時系列プロファイルが KEGG の Biosynthesis of amino acids パスウェイ(rn01230)上の Leucine 分岐経路などにおいて組織間で異なっていることが発見可能。またプロファイルのマッピングに利用する KEGG パスウェイは IPython Notebook で容易に任意のものに変更することが可能。

§５ 研究開発計画に対する達成状況と将来展望

(1)達成状況

入力となる代謝物プロファイルのアノテーションの質や、代謝物とデータベースとの対応付けのための情報基盤が発現プロファイルと比して非常に困難であることが研究開発計画時に把握できておらず、当初掲げていた RDF の活用やゲノムスケールの代謝モデルのデータ統合が達成できなかった。この入力データの前処理の困難が後の解析パイプラインの内容にも影響し、当初計画していた AtMetExpress のすべてのデータセットに対応する可視化パイプラインの構築も達成できずその内の主要データセットへの対応に留まった。しかしながら代謝物のアノテーション環境が現状未整備であることはこれから行うべきスペクトルをトリプルの起点とした RDF、スペクトルデータベースとのリンク、そしてアノテーションに必要なオントロジーの作成の方向性を明らかにした。同定後の化合物 ID ではなくスペクトル情報とアノテーションを結ぶ用途にこそ RDF が適していること、また既存のデータベースを統合するだけでは代謝物プロファイルのすべてのアノテーションは現状行う事ができないため、オントロジー拡張用途にも RDF を用いたデータ整備が本研究の完成に必要であることがわかった。

(2)ツール等の将来展望

非公開

§6 研究参加者

氏名	所属	役職	研究開発項目	参加時期
○西田孝三	理化学研究所	テクニカルスタッフ	開発全般	H26.9-H27.2

§7 成果発表等

(1)原著論文発表 (国内(和文)誌 件、国際 (欧文) 誌 件)

(2)その他の著作物(総説、書籍など)

1. Fukushima, A., Kanaya, S., & Nishida, K. Integrated network analysis and effective tools in plant systems biology. Frontiers in plant science, 5. (2014).

(3)国際学会発表及び主要な国内学会発表

① 招待講演 (国内会議 0 件、国際会議 0 件)
② 口頭発表 (国内会議 0 件、国際会議 1 件)

1. Kozo Nishida (Riken)、Integrated omics analysis pipeline for model organism with Cytoscape、RECOMB 2014、San Diego、11 月 11 日

③ ポスター発表 (国内会議 0 件、国際会議 1 件)

1. Kozo Nishida (Riken)、KEGGscape: Cytoscape plugin for mapping metadata on KEGG PATHWAY、The International Conference on Systems Biology 2014、Melbourne、9 月 14-18 日

(4)知財出願

①国内出願 (0 件)
②海外出願 (0 件)
③その他の知的財産権
なし

(5)受賞・報道等

なし

§8 自己評価

公開を行ったソフトウェアの IPython Notebook と Cytoscape からなる構成や可視化結果については Pathview においてもサポートされていない詳細な可視化が実現できたと考える。しかしパイプラインへの入力となる代謝物プロファイルのデータ構造とデータ処理の設計に関しては全く整理を行うことができなかった。ただしその解決策の将来展望が前述第 5 節のように見えた点で本トライアルは非常に有意義であったと考える。

ツールの機能、有用性

- 任意のKEGG PATHWAY上でのAtMetExpress代謝物プロファイル中のCol0_TimeCourseデータセット等の実験条件間の差異の可視化を行う

- IPython NotebookとCytoscapeを連携させ、プロファイルから可視化までの一連の解析パイプラインを提供

- Cytoscapeの豊富な機能と、再現性が高くプログラマブルなパイプラインの双方を統合した環境が利用可能

使用方法

- IPython Notebook, Cytoscape, KEGGscape, cyREST, keggutilのインストール後、https://github.com/kozo2/togotrial2014 をクローンし leaf-root.ipynb をウェブブラウザー上から実行する

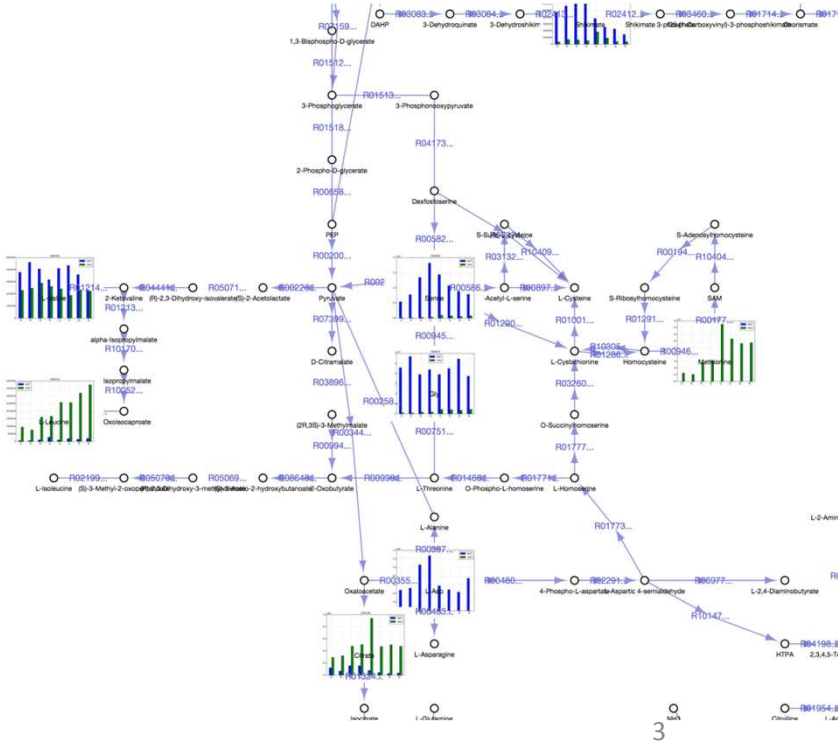
- Programのカスタマイズ可能箇所は*.ipynb中に指示があり、これらを変更することで処理の変更が可能

入力と出力

AtMetExpressの
代謝物プロファイル中の
データセット

IPython Notebookを用いた解析パイプライン

KEGG と Cytoscapeを用いた
代謝物プロファイルの統合
と可視化



実験条件間差異の例

KEGG Biosynthesis of amino acidsパスウェイ
葉(青)、根(緑)間の時系列代謝物プロファイルの
差異がL-Valine、L-Leucine間等で視認可能