

生化学反応ネットワーク 統合解析環境の拡張 (中間報告)

西田 孝三

理化学研究所・生命システム研究センター(QBiC)

Kozo Nishida @ JST-NBDC, Dec 26, 2014



©2014 西田孝三(理化学研究所) licensed under CC表示2.1日本

(ゲノムスケールの)生化学反応ネットワーク解析の流れ

- 定量生物学
- 分析化学

- 統計、機械学習

- データベース、オントロジー
- ゲノムアノテーション

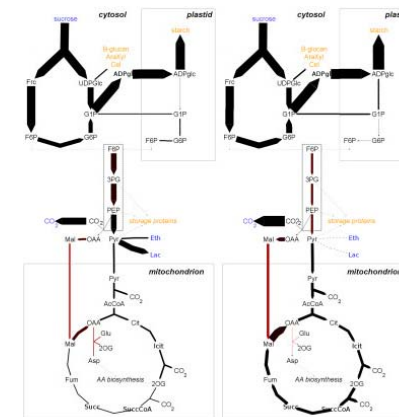
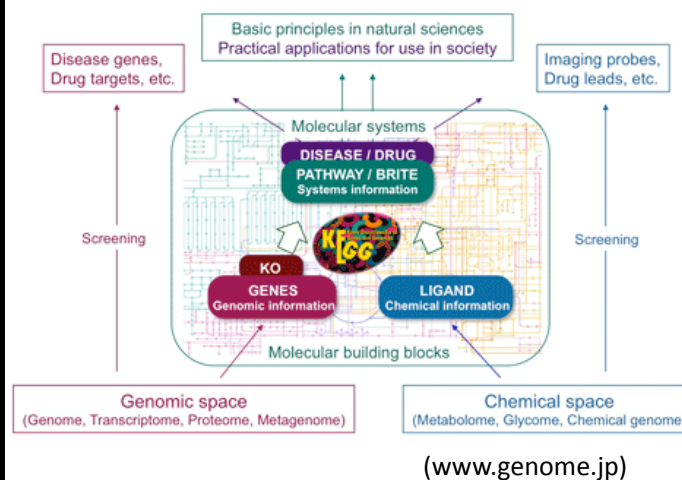
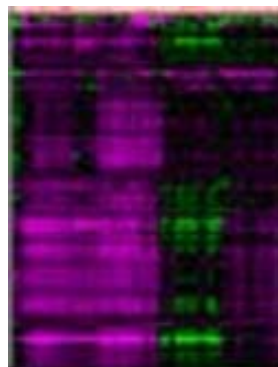
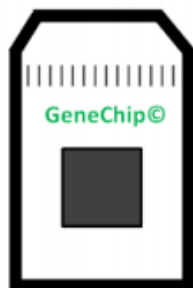
- ネットワーク
生物学

- トランスクリプトーム (GeneChip, RNA-seq)
- メタボローム (LC, GC-MS)

- 前処理(正規化)
- [統計]解析(有意差, 相関, 機械学習...)

反応ネットワーク(パスウェイ)への解析結果のマッピング(統合)・可視化

- パスウェイ比較・解析・可視化
- シミュレーションの制約条件への利用 (FBA...)

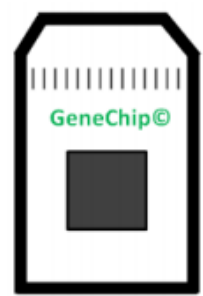


(Rohn, 2012)

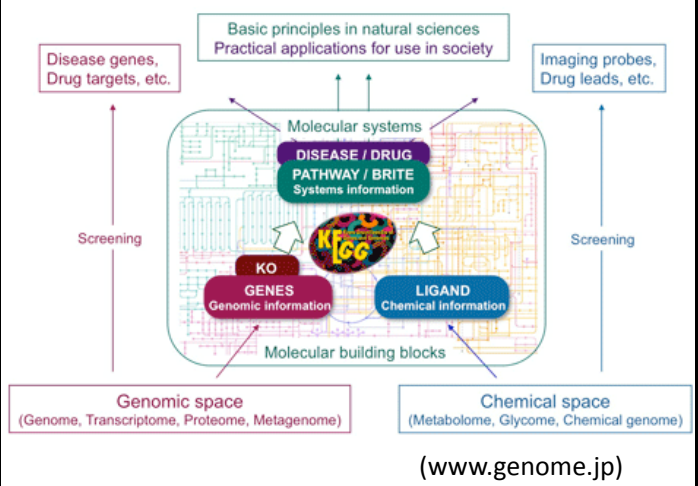
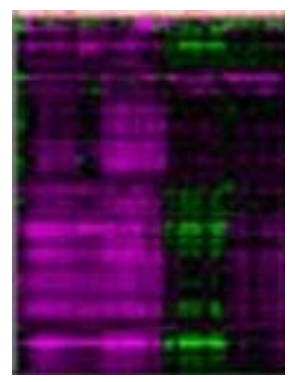
今年度目指すもの

汎用的かつ各解析を最終結果まで一気通貫で行える
統合解析環境(Cyberinfrastructure)の実現

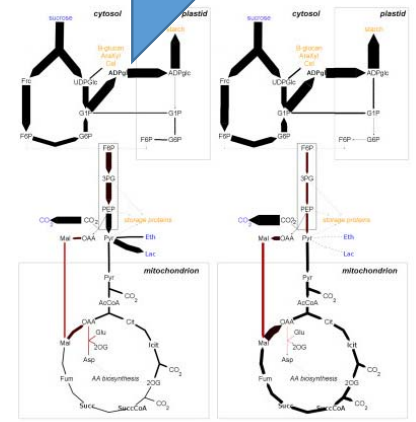
- メタボローム(LC, GC-MS)



機械学習...

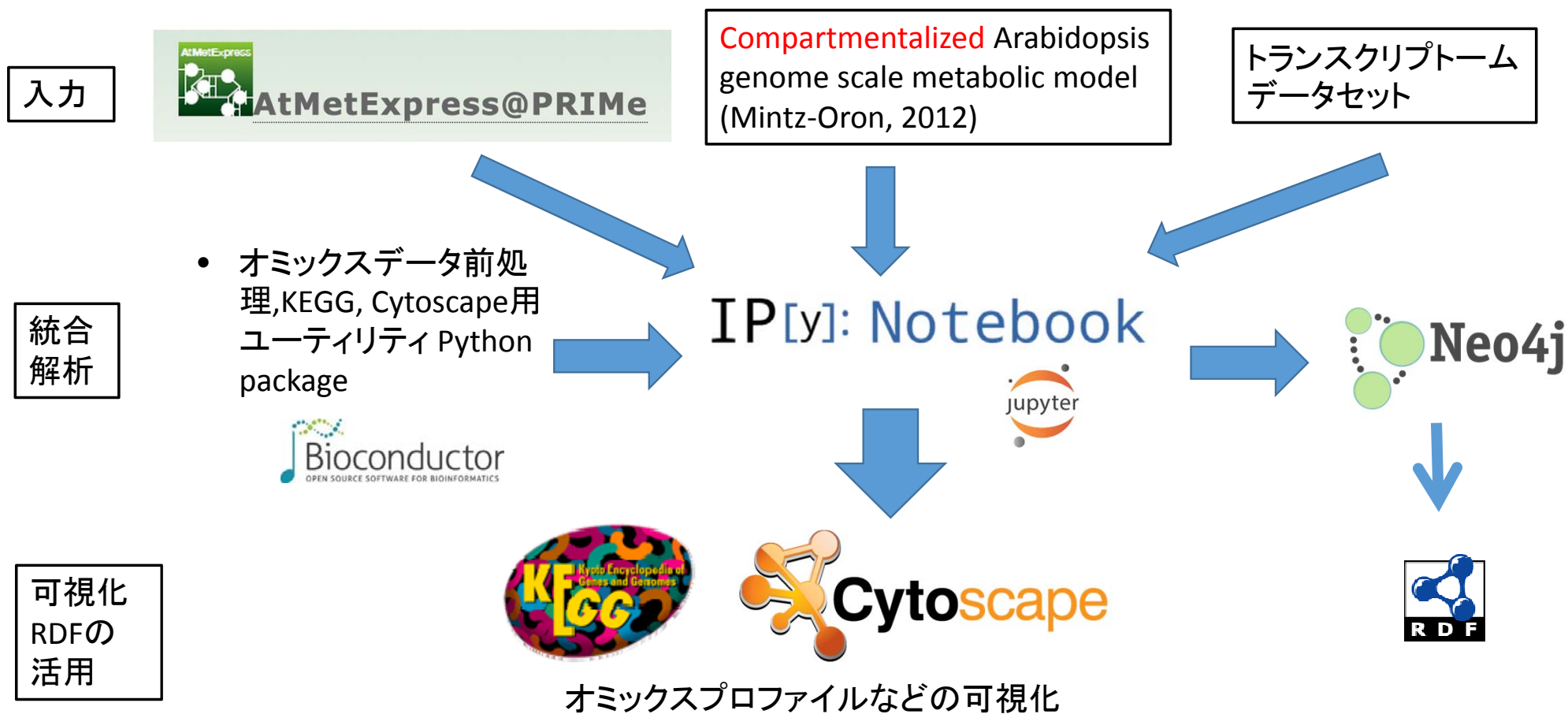


約...の制



(Rohn, 2012)

具体的なユースケース



具体的なユースケース

- KEGGが含まないメタデータとオミックスデータの統合(メタボロームデータを追加)
- IPython notebook, Cytoscapeを用いた再現性の高い統合解析、可視化環境の構築
- RDFを利用した反応ネットワークの解析

可視化
RDFの
活用



Omicsプロファイル、メタデータ(ゲノムアノテーション、
器官)差異の可視化

達成事項と未達成事項

- 達成事項

- AtMetExpressのメタボロームデータセットに合わせた解析パイプラインの構築

- 未達成事項

- オミックスデータ前処理, KEGG, Cytoscape用Utility Python package作成, Bioconductor(R) の解析パイプラインの統合
- AtMetExpressとArabidopsis thalianaのgenome scale metabolic model(Mintz-Oron, 2012)のdata統合
- トランスクリプトームデータセットの統合
- Neo4jを利用したデータ統合とRDFを利用した解析

デモ

メタボロームデータセットの詳細と パイプラインについて

- Metabolomics dataの前処理
 - 同一compoundと判定されるが蓄積量profileが複数種存在するよう扱わざるを得なかったpeakのgroup化
 - Phenotype, 実験条件(Tissue, 光周期など)毎の複数の蓄積量profileのgroup化
- KEGG用ユーティリティ (KEGG APIの処理を組合せたもの)
- Cytoscape用ユーティリティ(データ統合, 可視化をプログラマブルに行うもの)

達成事項と未達成事項

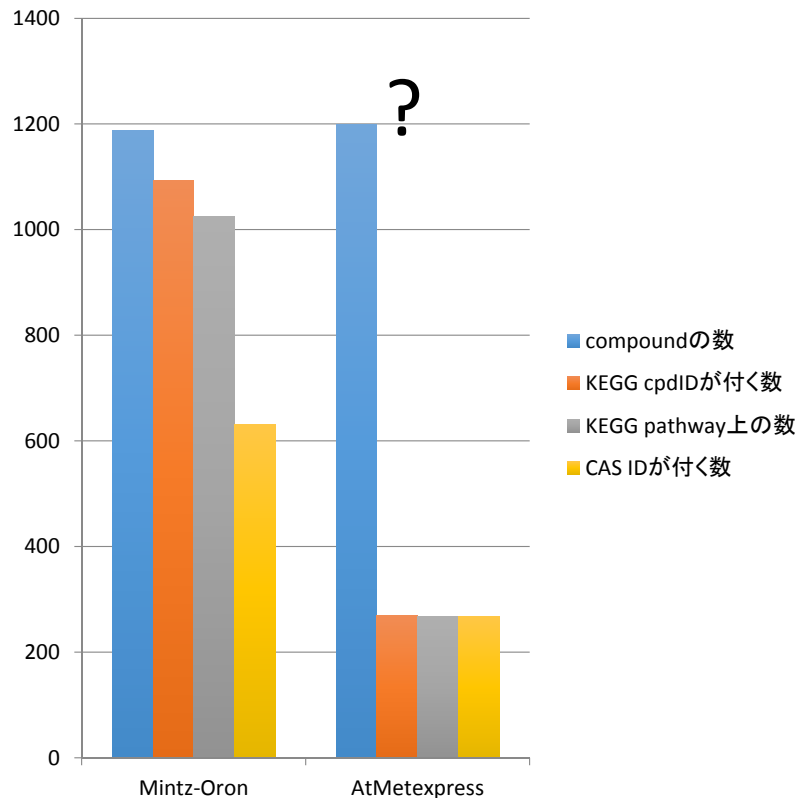
- 達成事項

- AtMetExpressのメタボロームデータセットに合わせた解析パイプラインの構築

- 未達成事項

- オミックスデータ前処理, KEGG, Cytoscape用ユーティリティPython package作成, Bioconductor(R) の解析パイプラインの統合
- AtMetExpressとArabidopsis thalianaのgenome scale metabolic model(Mintz-Oron, 2012)のdata統合
- トランスクリプトームデータセットの統合
- Neo4jを利用したデータ統合とRDFを利用した解析

Mintz-OronのmodelとAtMetExpressについて



- Mintz-OronのModelはAraCycとKEGGに含まれる pathway 情報(ID)で構成される
- Mintz-Oronのgenome scale metabolic modelの reaction networkはKEGG pathway上でそのほとんどが可視化可能
- AtMetexpressのcompoundはCASがDB照合用の primary IDだったためこれをkeyとし照合
- Modelの 1/5 程度 をcover
- AtMetExpressのannotationの見直し

達成事項と未達成事項

- 達成事項

- AtMetExpressのメタボロームデータセットに合わせた解析パイプラインの構築

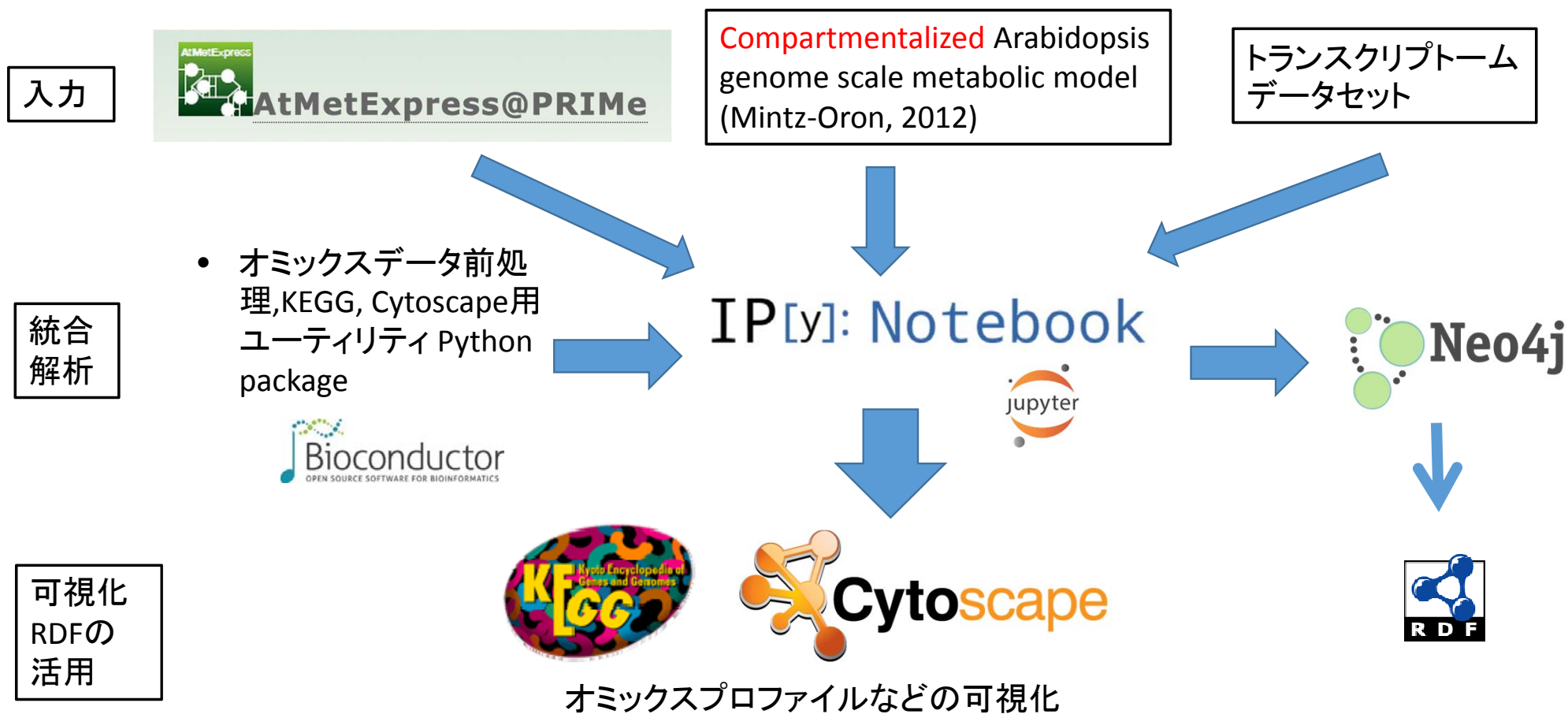
- 未達成事項

- オミックスデータ前処理, KEGG, Cytoscape用ユーティリティPython package作成, Bioconductor(R) の解析パイプラインの統合
- AtMetExpressとArabidopsis thalianaのgenome scale metabolic model(Mintz-Oron, 2012)のdata統合
- Transcriptome datasetの統合
- Neo4jを利用したdata統合とRDFを利用した解析

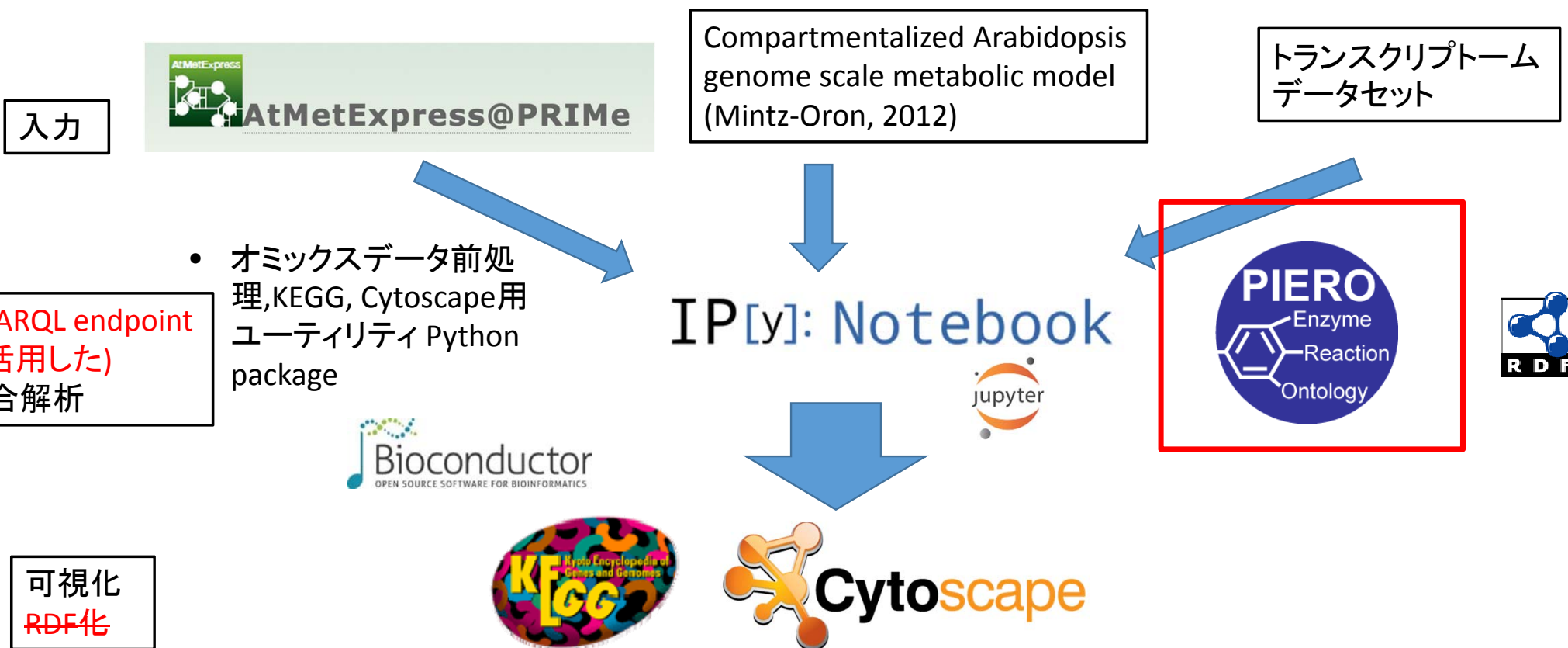
PIERO Enzyme Reaction Ontology

- KEGGの反応ネットワーク情報をRDF化。10月7日に published(Kotera, Nishimura 2014)
- KEGGのreaction databaseのSPARQL endpointを提供
 - 提案当初に予定していたKEGGのreaction情報のRDF化は不要に
- Reaction network 解析に有用なontology

具体的なユースケース



具体的なユースケース(変更案)



まとめ

- メタボロームデータセットの反応ネットワーク上での可視化を一気通貫で行うためのパイプライン処理の基盤の完成を示した

今後の予定

- パイプライン処理を容易に行うための汎用ライブラリ作成
- AtMetExpressのアノテーションのチェック
- メタボローム実験条件に合わせたトランスクリプトーム情報の可視化
- KEGGのSPARQL endpoint, PIEROの活用