ライフサイエンスデータベース統合推進事業
統合データ解析トライアル
研究開発課題
「マルチオミクスデータを用いたゲノム規模代謝モデリングのためのネットワーク解析システムの開発」

研究開発終了報告書

研究開発期間：　平成 25 年 9 月～平成 26 年 1 月

研究代表者：西田孝三
（(独)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 研究開発のねらい

ゲノム配列情報のみからタンパクや核酸分子の構造と機能を予測し、それらの間の相互作用ネットワークを定量的に記述し、細胞や個体の機能とを一対一に結び付ける「遺伝型からの表現型の予測」、この実現がゲノム解析の理想ではあるが現状では困難である。このため大腸菌のような比較的単純なモデル生物に対して「遺伝子に変異を加える」、「薬剤を投与する」といった何らかの攪乱条件と標準条件の間で定量マルチオミクスが与える網羅的な測定情報の比較を行い、比較結果をKEGG(金久ら、Nucleic Acids Res 2012)に代表されるパスウェイデータベースにマッピングすることによってそのネットワークを解析することが表現型の予測の現実解として行われている。しかしこの現実解でさえ、その解析は容易いことではない。その要因のひとつとして「特定の生物種に特化した詳細なメタデータを含む代謝モデルが KEGG のようなパスウェイマップと統合され」、「ユーザー独自のオミクスデータを容易にマッピング、可視化でき」、さらには「KEGG MEDICUS のような薬剤情報とパスウェイマップの関連性を辿ることができる」相互作用ネットワーク解析システムが無いことがある。そこで本研究ではこのようなネットワーク解析システムの構築を目的として「モデル生物としては最も単純なもののひとつである大腸菌(Escherichia coli; E. coli)K-12 MG1655 の全ゲノム規模の詳細な代謝モデル(カリフォルニア大学サンディエゴ校の Palsson らによる iJO1366(Orth ら、Mol Syst Biol 2011))と KEGG の統合」「統合情報にユーザーが独自のオミクスデータを GUI で KEGG のパスウェイにマッピング、可視化でき、」「可視化結果で着目したパスウェイについてはそのパスウェイをターゲットとする KEGG DRUG エントリの有無を画面遷移を伴うことなくユーザーに提示するシステムの開発」を行う。具体的な有効活用例としては本研究開発成果とマルチオミクスデータを用いたゲノム規模の代謝モデリング技術による薬剤応答、耐性機構予測の実現をねらう。

§2 研究成果

(1) 研究開発の経緯

近年いくつかのモデル生物で全ゲノム規模での代謝モデルの再構築結果が公開されているが、これらのモデルに「オミクス実験情報」、「KEGG に代表されるパスウェイ、そして KEGG MEDICUS のようなゆらぎ物質としての薬剤情報を集めたデータベース」を統合したネットワーク解析システムはまだ発展の途上にある。これらのモデルはオミクス実験を解釈する上で有用かつ KEGG が含まない情報を持っているが、その情報をマップし解釈するためのパスウェイネットワークとしては依然 KEGG パスウェイが有用であり、またそのパスウェイに関連付けられた薬剤情報データベース KEGG MEDICUS もオミクス実験に新たな解釈を与えるものとして有用である。しかしながらこれらの情報すべての統合をシステマチックに行うソフトウェア環境は確立されておらず本研究開発に至った。

(2) 研究開発の実施内容

本研究開発では前述の統合を行うネットワーク解析環境を実現した。この環境は

- Cytoscape 3.1 用 KEGG PATHWAY XML(KGML) インポート用プラグイン KEGGscape 0.5.1
- 分子間相互作用ネットワーク解析プラットホーム Cytoscape 3.1
- ドキュメントデータベース mongodb 2.4.9
- mongodb 操作のための Python 言語ライブラリ pymongo 2.6.3

から成る。この内 KEGGscape は研究開発提案者の開発によるものであり、残りの構成要素と組み合わせ、

- 薬剤情報データベース Drugbank 中の大腸菌のタンパクをターゲットとする薬剤の KEGG パスウェイへのマッピング
- 大腸菌 K-12 MG1655 の全ゲノム規模の詳細な代謝モデル iJO1366 の 1 世代前の代謝モデル iAF1260 の KEGG パスウェイへのマッピング
- 変異株 2 種(Δpta/Δadhe/ΔpfkA/Δglk と Δpta/ΔpfkA)の公開マイクロアレイデータの比較解析結果の KEGG パスウェイへのマッピング

を行い、これを再現するドキュメントと合わせ公開した。

(3) 本研究開発の成果として作成したプログラムなど

- KEGGscape
  - 公開 URL http://apps.cytoscape.org/apps/keggscape
  - ソースコード公開 URL https://github.com/idekerlab/KEGGscape
  - ドキュメント URL http://keggscape.readthedocs.org/en/latest/

(4) 研究開発の成果の汎用性、新たな活用法

成果であるネットワーク解析環境がユーザーに求めるものは代謝モデル情報やオミクス実験に限らない。mongodb が扱うことが可能なファイル形式(json, csv, tsv)に KGML 中の KEGG ID を含むものであればすべて KEGG パスウェイへのマッピングが可能である。iAF1260 に限定されず、KEGG の反応 ID との対応がとれるものであればすべて KEGG パスウェイとの情報統合が可能なため、大腸菌以外の生物種に対しても同様の解析が可能である。iAF1260 を作成した Palsson のグループだけでも大腸菌以外に Bacillus subtilis、Haemophilus influenzae、Helicobacter pylori、

Mycobacterium tuberculosis、Staphylococcus aureus といった代謝モデル情報があり「代謝モデル、マルチオミクスデータ、KEGG パスウェイ、薬剤」を統合したネットワーク解析に活用が可能である。

(5)研究開発の成果からの有用な知見の発見

KEGG orthology(KO)でのアノテーションの漏れが可視化されることで、KEGG の情報のみでは解釈できなかったマルチオミクスデータの解釈が可能となった(具体例は別紙を参照)。また本研究によりヒト以外の生物種の代謝に対するゆらぎ物質としての薬剤情報は KEGG MEDICUS には不足していることがわかった(§3(1)b)薬剤情報の統合について、を参照)。最後に知見の発見ではないが薬剤ターゲット情報をもとにパスウェイの着目点をユーザーに提供することが可能となった。

§3 研究開発計画および計画に対する達成状況

(1)達成状況、方針変更など

a) 薬剤情報の統合について

KEGG DRUG 中のターゲット情報はヒト以外の生物種については KO を対象としている。これら KO に大腸菌(KEGG の生物種 ID eco)が含まれる薬剤を調べあげたがペプチドグリカン合成系、葉酸合成系、タンパク合成系といった抗生物質としての薬剤しか存在しなかった。これらは代謝に対するゆらぎ物質としての薬剤には適していないと考え、薬剤情報の統合対象は当初想定していた KEGG MEDICUS から Drugbank に変更を行った。Drugbank の薬剤ターゲットについてはすべて KEGG エントリとの照合が行えるようになっており、薬剤情報と KEGG パスウェイの統合が達成できている。

b) 代謝モデル情報の統合について

当初想定していた最新の大腸菌代謝モデル iJO1366 から iAF1260 に統合対象を変更した。iAF1260 が KEGG の反応 ID との対応を有していたのに対し iJO1366 はこれを有していなかったため。iAF1260 については統合のしくみが実現できている。iAF1260 と iJO1366 の差分について統合を行うしくみはまだ実現できていない

c) マイクロアレイデータの統合について

当初は Palsson らが公開しているマイクロアレイデータ(http://systemsbiology.ucsd.edu/InSilicoOrganisms/Ecoli/EcoliExpression2)を KEGG パスウェイにマップすることのみを宣言しており、具体的にどの実験条件のマイクロアレイデータを比較解析を行うかまでは予定に含めていなかったが、変異株 2 種(Δpta/Δadhe/ΔpfkA/Δglk と Δpta/ΔpfkA)の各アレイデータを KEGG の Galactose metabolism パスウェイにマップし前項 a) b)と合わせ可視化を行った。これら変異株のアレイは遺伝子 glk の産物が a)の薬剤ターゲットに含まれること、b)の代謝モデルにのみ含まれる反応が薬剤ターゲットが機能する Galactose metabolism パスウェイに存在したことから選定した。発現を KEGG パスウェイにマップするしくみは実現できている。

d) その他達成できなかったもの

当初は a) b) c)での統合結果をグラフデータベースとして適した形式で Web 上で公開すると宣言していたが a) b) c)を確実にユーザーの手元に再現するソフトウェアとそのドキュメントを提供するに留まった。

（２）ツールの将来性への展望

本研究開発ではオミクスデータの標準化(マイクロアレイデータの GCRMA)は終わった後のデータを用いた。オミクスデータの比較解析のための標準化は特にメタボロミクスにおいて重要であるがそれだけで多大な労力が伴うためデータベースの構築とそのビューアとなる Web アプリケーションの構築に手一杯というのが現状である。こういった比較オミクスのためのデータベースに本件のようなさまざまなメタデータを統合するネットワーク解析システムをクライアントとして提供しさらなる統合を行うことでマルチオミクス解析結果をベースにした生命システムの理解が期待される。また本研究においてはメタデータ統合はドキュメントデータベース mongodb によりデータの取り扱いはローカルテキストファイルもしくは関係データベースと比較し簡潔にはなっているもののその表形式のドキュメントの内容については依然熟知する必要があり、スクリプト言語を用いた文字列のパース、また ID 照合をユーザーに強いている。この点については Drugbank, KEGG, 代謝モデルさらにはマルチオミクス情報これらすべての linked open data 化を行うことが望ましい。この linked open data 化が実現されれば、本研究で提案したような Python スクリプトの作成をユーザーが行う必要は無くなる。 しかしながら取り上げた全てのデータベース情報の迅速な linked open data 化は困難である。今後は統合を行った後のデータをドキュメント指向と linked open data 指向の中間に位置するグラフデータベースに格納、公開し、パスウェイネットワーク解析の利便化を行う予定である。

§４ 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	すべて	H25.10-H26.1

§５ 成果発表等

(1)原著論文発表 （国内(和文)誌 0 件、国際（欧文）誌 0 件）

(2)その他の著作物(総説、書籍など)

(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 0 件）

③ ポスター発表 （国内会議 0 件、国際会議 1 件）

１．Kozo Nishida（Riken）、Importing KEGG signaling pathways with KEGGscape、Cytoscape retreat2013、Institut Pasteur Paris, France、2013 年 10 月 9.10 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

(5)受賞・報道等

① 受賞

② マスコミ(新聞・TV等)報道

④ その他

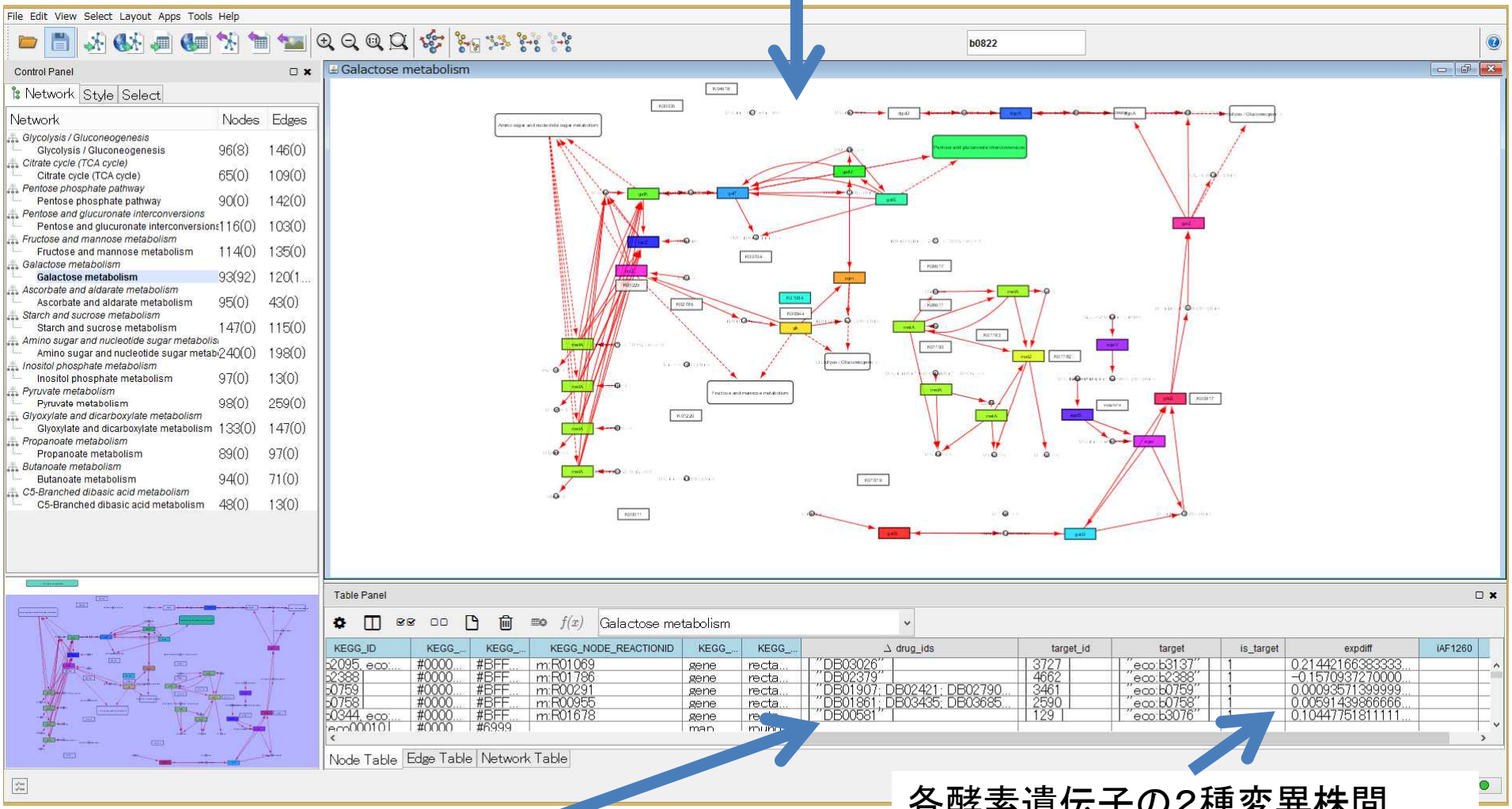
§6 自己評価

研究開発のねらいを実現するために立てた方針、データベースと各要素を組み合わせて作ったソフトウェアシステムには自信があり、ほぼ§1を満たせたと思うが、申請前での KEGG MEDICUS の調査が甘く、薬剤情報に KEGG MEDICUS のデータを用いると言っておきながら直接使うことはなかったことについては咎められても仕方がない。このため Drugbank の薬剤情報の統合に多くの時間をかけることとなり、マッピング結果を考察する、もしくは2次データの公開に関する案を進める、といったことができなかった。また各ソフトウエアコンポーネントの作り込み（例外処理、可視化の見栄え、汎用化など）が甘く発展の余地がある。

以上

ツールの機能

KEGGのパスウェイレイアウトに基づいたネットワークの
Cytoscapeへのインポート機能

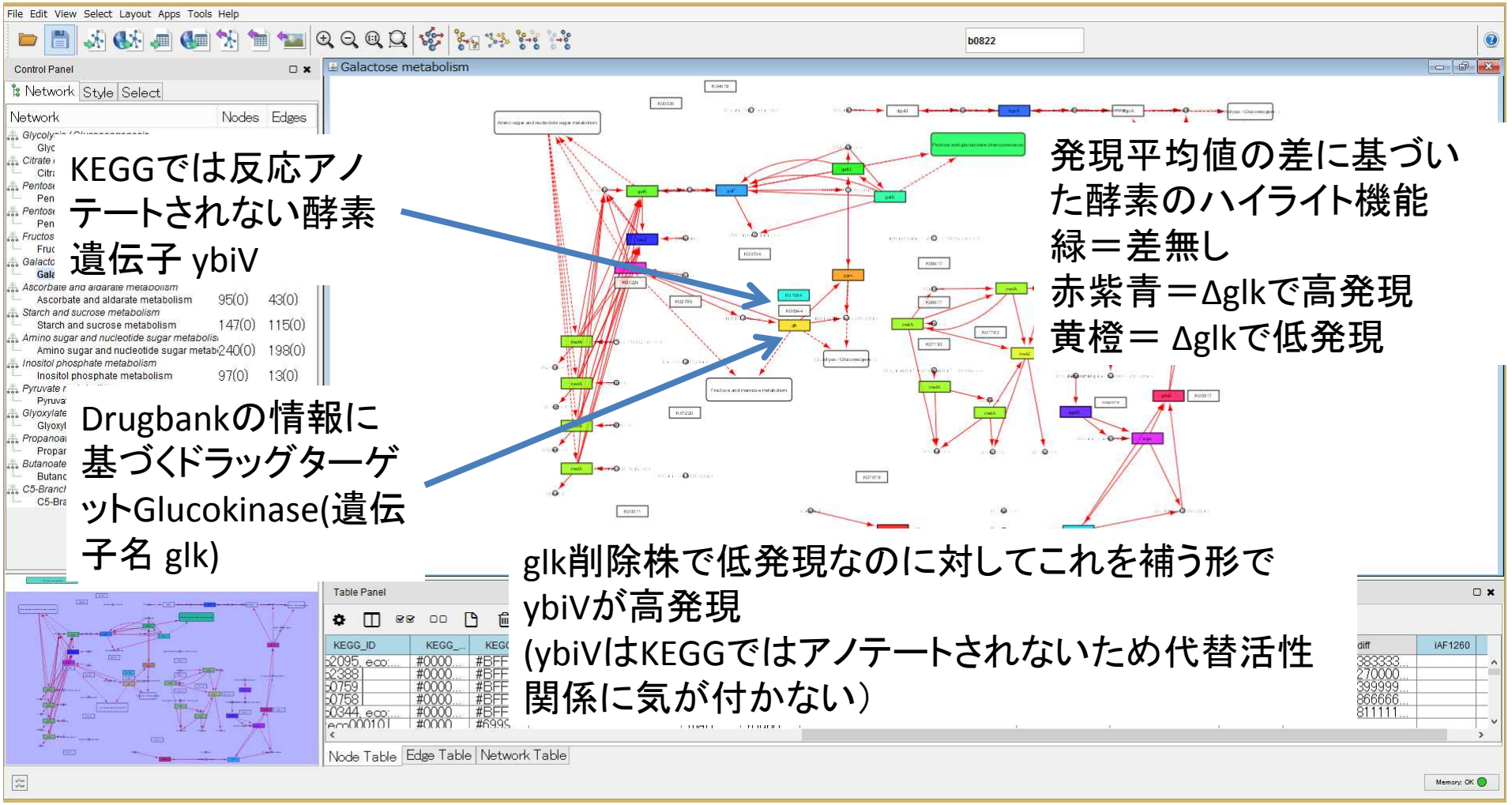


ネットワーク中の酵素にターゲットとする
Drugbankの薬剤IDが関連付けられる

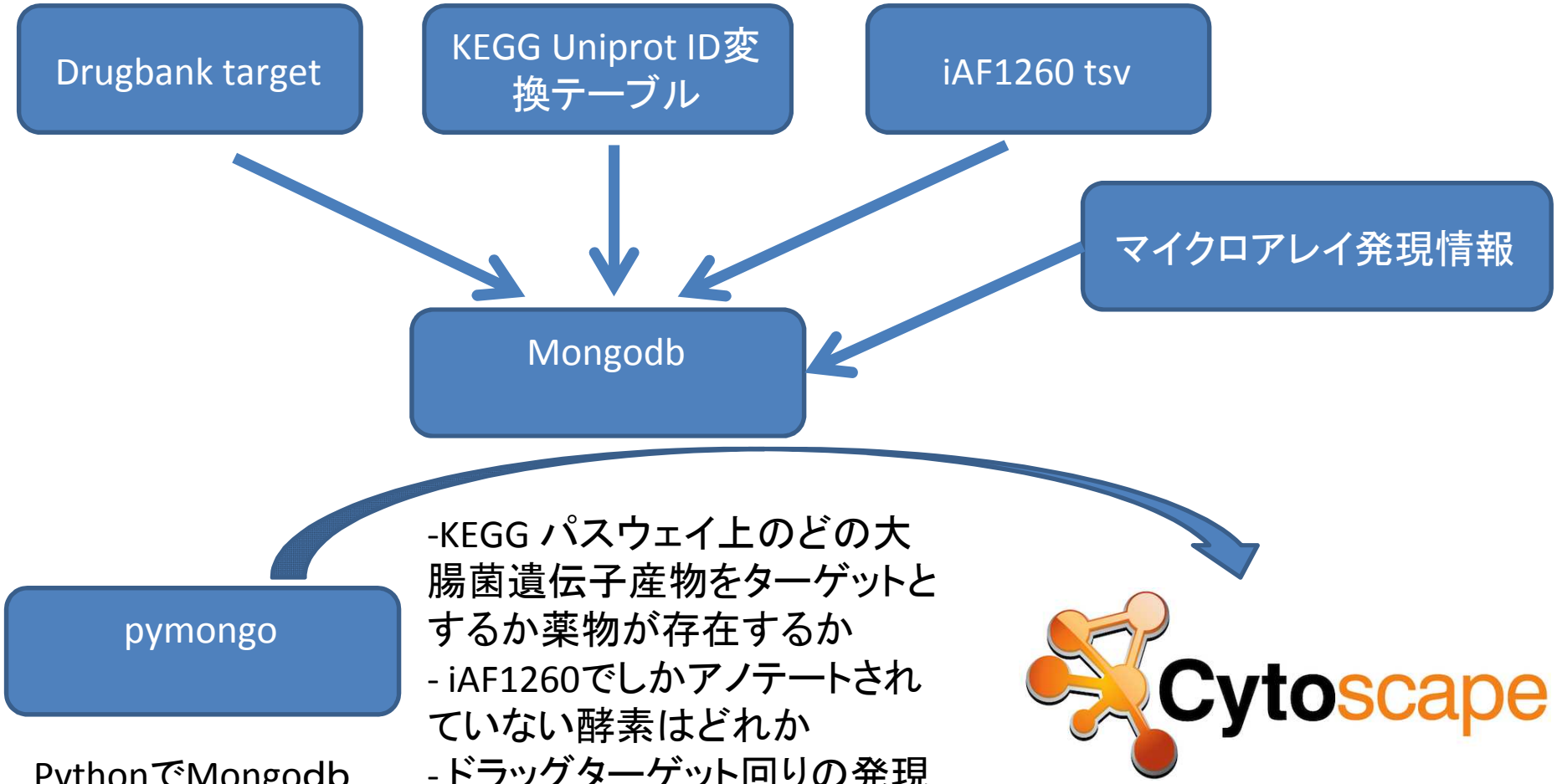
各酵素遺伝子の2種変異株間
(Δpta/Δadhe/ΔpfkA/Δglkと
Δpta/ΔpfkA) の発現平均値の差

ツールの有用性

薬剤＋オミクス情報（発現）＋代謝モデルの全情報を統合した生命システムの考察を可能に



入出力の関係


Drugbank target
KEGG Uniprot ID変換テーブル
iAF1260 tsv
マイクロアレイ発現情報
Mongodb
pymongo
-KEGG パスウェイ上のどの大腸菌遺伝子産物をターゲットとするか薬物が存在するか
- iAF1260でしかアノテートされていない酵素はどれか
- ドラッグターゲット回りの発現差異
Cytoscape
PythonでMongodb中のデータをアグリゲートしcsvで出力
csvを受け取りパスウェイネットワークと関連付けて可視化

現時点での使用方法

- アグリゲート後のデータの配布を行うのは問題があるためすべての手続きをドキュメント化
  http://keggscape.readthedocs.org/en/latest/
- ユーザ自身で各データのダウンロード、スクリプトの実行を行ってもらう