パスウェイデータベースの紹介

片山 俊明 <ktym@dbcls.jp>

http://jp.linkedin.com/in/toshiakikatayama

大学共同利用機関法人 情報・システム研究機構 *(ROIS)*
ライフサイエンス統合データベースセンター *(DBCLS)*

2015/8/4 @ AJACS#55, 米子

パスウェイデータベースとは

- タンパク質や化合物等の分子間相互作用ネットワークをDB化

- 目的
  - 網羅的な遺伝子のネットワークを見る（知識の集大成）
    - 注目する遺伝子や現象に関わる反応を調べる
  - ゲノムアノテーション、種間比較、進化解析
  - 遺伝子発現などのエンリッチメント解析
  - 数理モデル、シミュレーション、フラックス解析

Roche – Pathways

www.roche.com/sustainability/for_communities_and_environment/philanthropy/science_education/pathways.htm

Global Web Site | Roche in your country | Deutsch | Contact | Search

Roche

Home | About Roche | Products | **Sustainability** | Research & Development | Media | Investors | Careers

Sustainability > For communities and the environment > Philanthropy > **Biochemical Pathways**

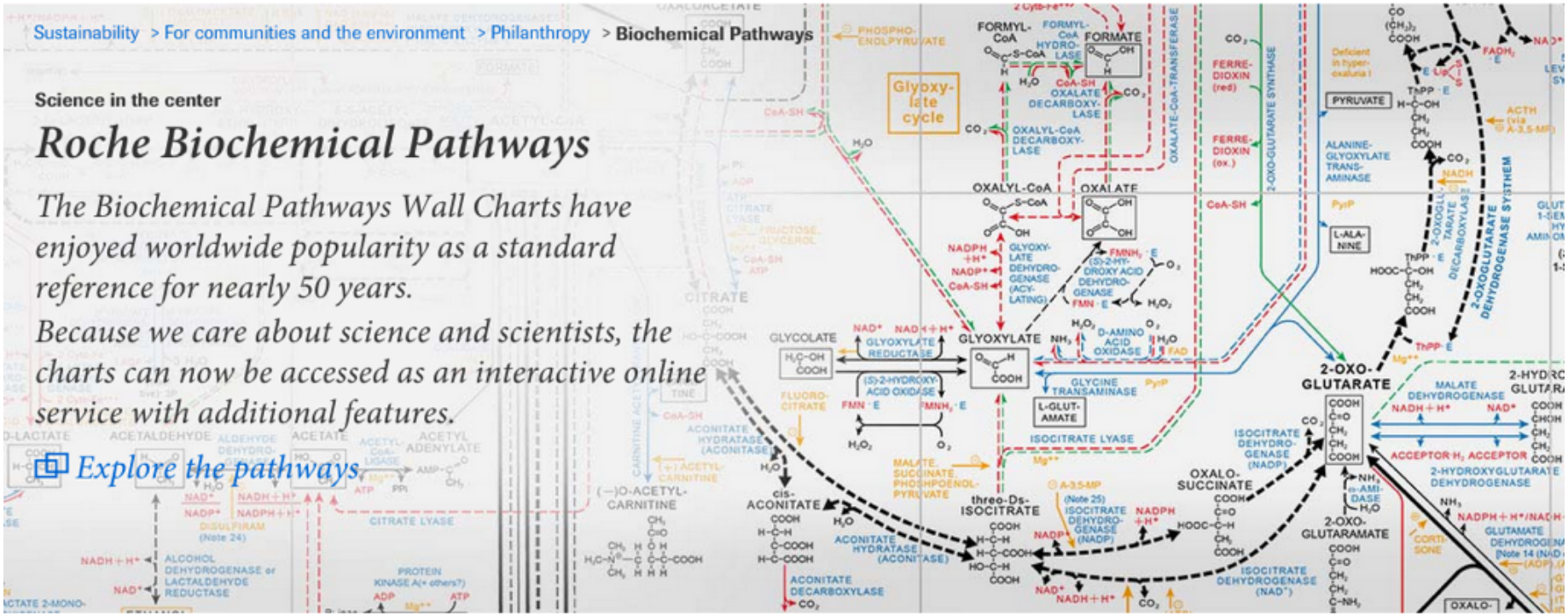
Science in the center

Roche Biochemical Pathways

The Biochemical Pathways Wall Charts have enjoyed worldwide popularity as a standard reference for nearly 50 years.

Because we care about science and scientists, the charts can now be accessed as an interactive online service with additional features.

Explore the pathways.



Share | Rate

Mapping the paths of life

Biochemical Pathways provide an overview of the chemical reactions of cells in various species and organs. Dr. Michal first compiled the Pathways Chart in 1965 and has been fine-tuning it ever since. Today, and with the collaboration of Roche, the two enormous posters can be found hanging in just about every research institute from Argentina to New Zealand.

> *You have to be someone with tenacity and patience. And love for science.*
>
> **Dr. Gerhard Michal**
> Editor of the Roche Biochemical Pathways

By the numbers

49 years

continuously fine-tuned by the editor himself.

パスウェイデータベースの歴史

- ベーリンガーマンハイム社（現ロシュ社）の代謝マップ
  - 1965年にポスターとして出版、1999年に書籍版
  - 1994年に ExPASy の電子版
    - http://web.expasy.org/pathways/
  - 2014年にロシュの電子版
    - http://www.roche.com/sustainability/for_communities_and_environment/philanthropy/science_education/pathways.htm
- 代謝反応データベース
  - 1996年 KEGG, BioCyc (EcoCyc) などが作られ始める
- パスウェイデータベース
  - ゲノムプロジェクトの進展で、遺伝子アノテーションを標準化する Gene Ontology の整備やヒトや病気のパスウェイなどに発展

Gene Ontology

- ウェブサイト
  - http://geneontology.org/
- 開発元
  - 主に米国 Gene Ontology Consortium
- 対象
  - 遺伝子機能のアノテーションに用いる用語と、パスウェイを含む概念階層の標準化を、Biological processes, Cellular components, Molecular functions の3カテゴリに分けて整備
- 論文
  - Nat Genet. 2000 May;25(1):25-9.
  - ：

Gene Ontology Consortium

geneontology.org

Gene Ontology Consortium

Home | Documentation | Downloads | User stories | Community | Tools | About | Contact us

Search GO data

terms and gene produc

Search

Enrichment analysis (beta)

Your gene IDs here...

biological process

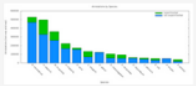
H. sapiens

Submit

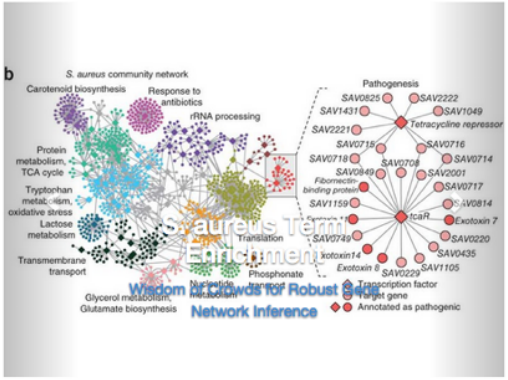
Advanced options

Powered by PANTHER

Statistics



Gene Ontology Consortium



What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology *Important*
- Gene Ontology tools
- Enrichment analysis
- Downloads

Recent news

Ten Quick Tips for Using the Gene Ontology

Important

Post date: 11/26/2013 - 08:22

Cardiovascular Gene Annotation Newsletter August 2015

Post date: 08/03/2015 - 02:40

BHF funded UCL-EBI collaboration leads to improved network analysis

Post date: 07/29/2015 - 06:26

Parkinson's Gene Annotation Newsletter July 2015

Post date: 07/01/2015 - 04:32

GO annotation of autism relevant proteins

Post date: 06/11/2015 - 06:21

more

Tags: navigation

User story: Everybody

Search

Highlighted GO term

Representi
biological p
The GOC h
a new term
(GO:00448
subclass of
This class r
period or st
biological p
more

Rand

- What
- What
- How
terms

View all FA

Tweets

jimh @jim
Getting rea
#geneonto
competitior
want to join
gowiki.tam

Gene

Compose

amigo.geneontology.org/amigo

AmiGO 2 | Home | Search | Tools & Resources | Help | Feedback | About | AmiGO 1.8

AmiGO 2

More information on quick search

Quick search

Search

Get Started with Grebe

Use the Grebe Search Wizard to **get started** in exploring the Gene Ontology data.

Go »

Advanced Search

Interactively **search** the Gene Ontology data for annotations, gene products, and terms using a powerful search syntax and filters.

Search

GOOSE

Use GOOSE to query a legacy GO database with **SQL** or edit one of the templates.

Go »

Term Enrichment Service

Your genes here...

biological process

H. sapiens | Submit

Powered by PANTHER

Advanced »

Statistics

View the most recent **statistics** about the Gene Ontology data on the main site.

Go »

And Much More...

Many **more tools** are available from the software list, such as alternate searching modes, Visualize, non-JavaScript pages.

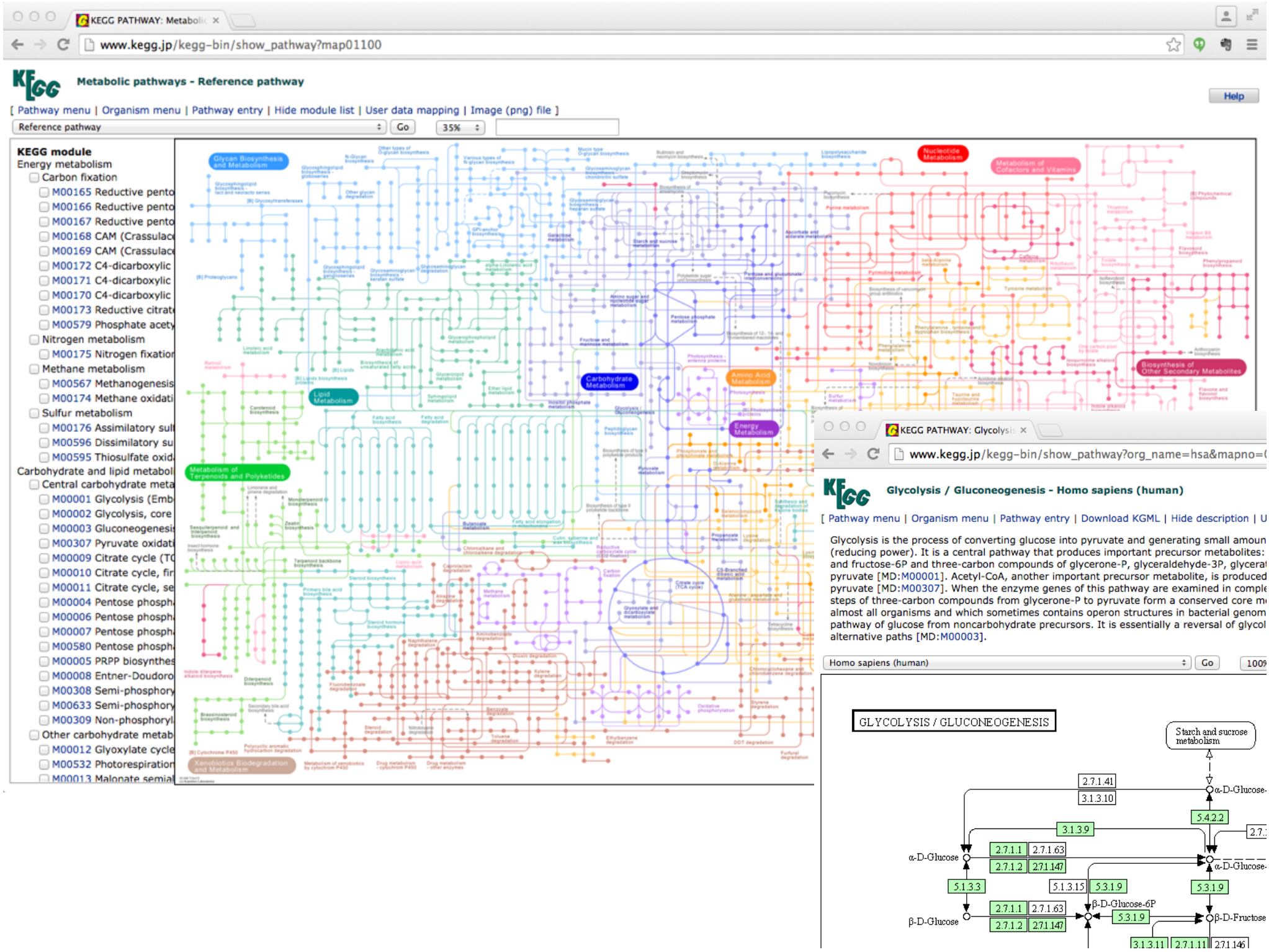
Go »

The Gene Ontology Consortium is supported by a P41 grant from the National Human Genome Research Institute (NHGRI) [grant 5U41HG002273-14]. The Gene Ontology Consortium would like to acknowledge the assistance of many more people than can be listed here. Please visit the acknowledgements page for the full list.

Last file loaded on 2015-08-01, see full details
AmiGO 2 version: 2.2.0 (amigo2a)

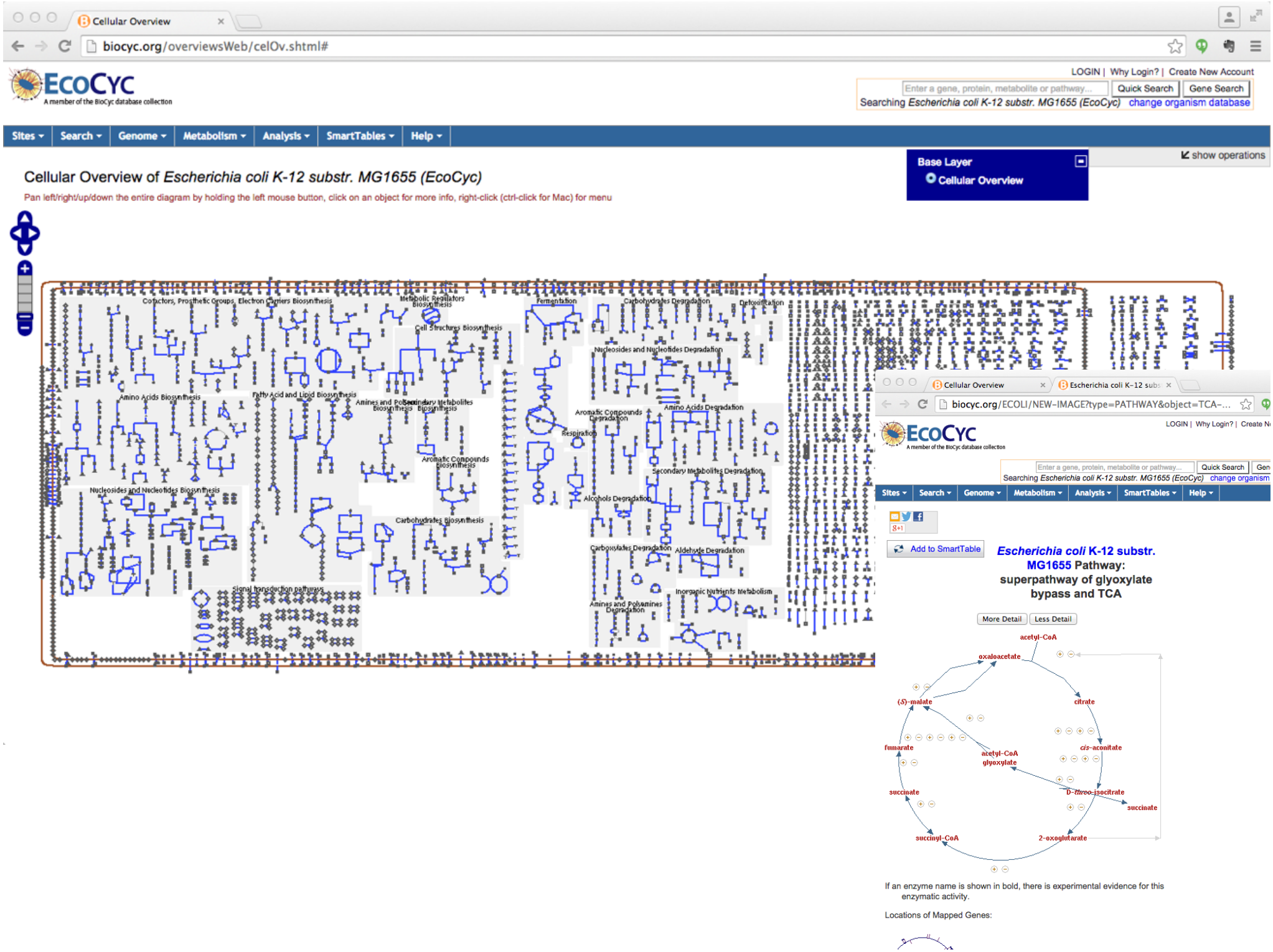
KEGG

- ウェブサイト
  - http://www.kegg.jp/
- 開発元
  - 京都大学 Kyoto Encyclopedia of Genes and Genomes
  - 金久實ら 今年20周年
- 対象
  - ゲノムの決まった全生物種（原核生物〜ヒトまで、手動・自動）
- 論文
  - Pac Symp Biocomput. 1997:175-86.
  - ：
  - Nucleic Acids Res. 2012 Jan;40(DB issue):D109-14.
  - Nucleic Acids Res. 2014 Jan;42(DB issue):D199-205.


KEGG PATHWAY: Metabolic
www.kegg.jp/kegg-bin/show_pathway?map01100
Metabolic pathways - Reference pathway
[Pathway menu | Organism menu | Pathway entry | Hide module list | User data mapping | Image (png) file]
Help
Reference pathway
Go
35%
KEGG module
Energy metabolism
Carbon fixation
M00165 Reductive pento
M00166 Reductive pento
M00167 Reductive pento
M00168 CAM (Crassulace
M00169 CAM (Crassulace
M00172 C4-dicarboxylic
M00171 C4-dicarboxylic
M00170 C4-dicarboxylic
M00173 Reductive citrate
M00579 Phosphate acety
Nitrogen metabolism
M00175 Nitrogen fixation
Methane metabolism
M00567 Methanogenesis
M00174 Methane oxidati
Sulfur metabolism
M00176 Assimilatory sul
M00596 Dissimilatory su
M00595 Thiosulfate oxid
Carbohydrate and lipid metabol
Central carbohydrate meta
M00001 Glycolysis (Emb
M00002 Glycolysis, core
M00003 Gluconeogenesis
M00307 Pyruvate oxidati
M00009 Citrate cycle (TC
M00010 Citrate cycle, fir
M00011 Citrate cycle, se
M00004 Pentose phospha
M00006 Pentose phospha
M00007 Pentose phospha
M00580 Pentose phospha
M00005 PRPP biosynthes
M00008 Entner-Doudoro
M00308 Semi-phosphory
M00633 Semi-phosphory
M00309 Non-phosphoryl
Other carbohydrate metab
M00012 Glyoxylate cycle
M00532 Photorespiration
M00013 Malonate semial
Glycan Biosynthesis and Metabolism
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism
KEGG PATHWAY: Glycolysis
www.kegg.jp/kegg-bin/show_pathway?org_name=hsa&mapno=
Glycolysis / Gluconeogenesis - Homo sapiens (human)
[Pathway menu | Organism menu | Pathway entry | Download KGML | Hide description | U
Glycolysis is the process of converting glucose into pyruvate and generating small amoun
(reducing power). It is a central pathway that produces important precursor metabolites:
and fructose-6P and three-carbon compounds of glycerone-P, glyceraldehyde-3P, glycerat
pyruvate [MD:M00001]. Acetyl-CoA, another important precursor metabolite, is produced
pyruvate [MD:M00307]. When the enzyme genes of this pathway are examined in comple
steps of three-carbon compounds from glycerone-P to pyruvate form a conserved core m
almost all organisms and which sometimes contains operon structures in bacterial genom
pathway of glucose from noncarbohydrate precursors. It is essentially a reversal of glycol
alternative paths [MD:M00003].
Homo sapiens (human)
Go
GLYCOLYSIS / GLUCONEOGENESIS
Starch and sucrose metabolism
2.7.1.41
3.1.3.10
α-D-Glucose-
5.4.2.2
3.1.3.9
2.7.1.1
2.7.1.63
α-D-Glucose
2.7.1.2
2.7.1.147
α-D-Glucose-
5.1.3.3
5.1.3.15
5.3.1.9
5.3.1.9
β-D-Glucose-6P
2.7.1.1
2.7.1.63
5.3.1.9
β-D-Glucose
2.7.1.2
2.7.1.147
β-D-Fructose
3.1.3.11
2.7.1.11
2.7.1.146

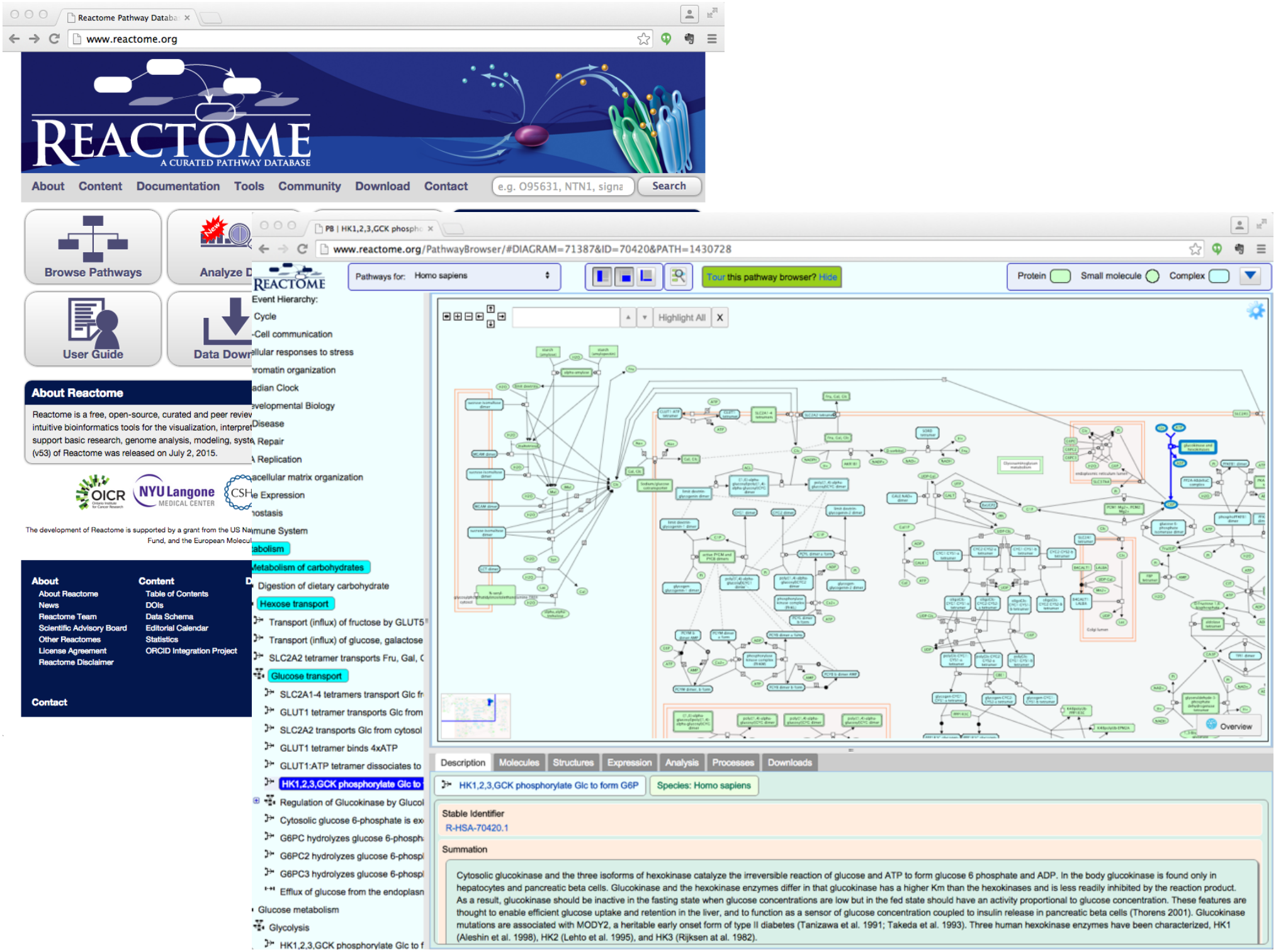
BioCyc

- ウェブサイト
  - http://biocyc.org/
- 開発元
  - 米国 SRI インターナショナル (Stanford Research Institute)
  - Peter Karp ら
- 対象
  - 大腸菌の代謝反応からヒトまで多生物種対応へ（手動・自動）
- 論文
  - Nucleic Acids Res. 1996 Jan 1;24(1):32-9. (EcoCyc)
  - Nucleic Acids Res. 2000 Jan 1;28(1):56-9. (MetaCyc)
  - ：
  - Nucleic Acids Res. 2014 Jan;42(DB issue):D459-71. (BioCyc)


Cellular Overview
biocyc.org/overviewsWeb/celOv.shtml#
EcoCyc
A member of the BioCyc database collection
LOGIN | Why Login? | Create New Account
Enter a gene, protein, metabolite or pathway...
Quick Search
Gene Search
Searching Escherichia coli K-12 substr. MG1655 (EcoCyc) change organism database
Sites
Search
Genome
Metabolism
Analysis
SmartTables
Help
show operations
Base Layer
Cellular Overview
Cellular Overview of Escherichia coli K-12 substr. MG1655 (EcoCyc)
Pan left/right/up/down the entire diagram by holding the left mouse button, click on an object for more info, right-click (ctrl-click for Mac) for menu
Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis
Metabolic Regulators Biosynthesis
Cell Structures Biosynthesis
Fermentation
Carbohydrates Degradation
Detoxification
Nucleosides and Nucleotides Degradation
Amino Acids Biosynthesis
Fatty Acid and Lipid Biosynthesis
Amines and Polyamines Biosynthesis
Secondary Metabolites Biosynthesis
Aromatic Compounds Degradation
Respiration
Amino Acids Degradation
Aromatic Compounds Biosynthesis
Secondary Metabolites Degradation
Nucleosides and Nucleotides Biosynthesis
Alcohols Degradation
Carbohydrates Biosynthesis
Carboxylates Degradation
Aldehyde Degradation
Signal transduction pathways
Inorganic Nutrients Metabolism
Amines and Polyamines Degradation
Cellular Overview
Escherichia coli K-12 subs
biocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=TCA-...
LOGIN | Why Login? | Create N
EcoCyc
A member of the BioCyc database collection
Enter a gene, protein, metabolite or pathway...
Quick Search
Gen
Searching Escherichia coli K-12 substr. MG1655 (EcoCyc) change organism
Sites
Search
Genome
Metabolism
Analysis
SmartTables
Help
Add to SmartTable
Escherichia coli K-12 substr. MG1655 Pathway: superpathway of glyoxylate bypass and TCA
More Detail
Less Detail
acetyl-CoA
oxaloacetate
citrate
(S)-malate
fumarate
acetyl-CoA
glyoxylate
cis-aconitate
succinate
D-threo-isocitrate
succinate
succinyl-CoA
2-oxoglutarate
If an enzyme name is shown in bold, there is experimental evidence for this enzymatic activity.
Locations of Mapped Genes:

Reactome

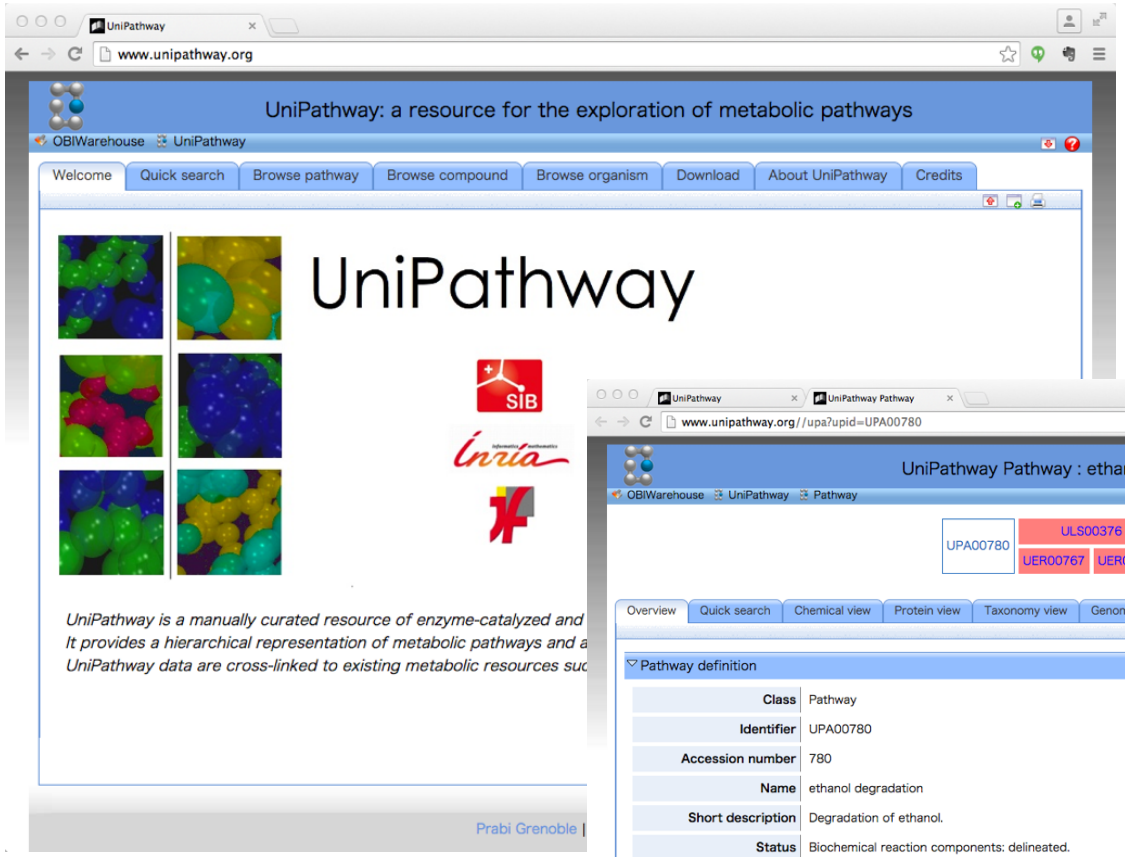
- ウェブサイト
  - http://www.reactome.org/
- 開発元
  - 英国 EMBL-EBI, 米国 Cold Spring Harbor Lab., カナダ OICR 等
- 対象
  - ヒトを中心に、脊椎動物、植物、酵母等の真核生物（手動）
- 論文
  - Nucleic Acids Res. 2005 Jan 1;33(DB issue):D428-32.

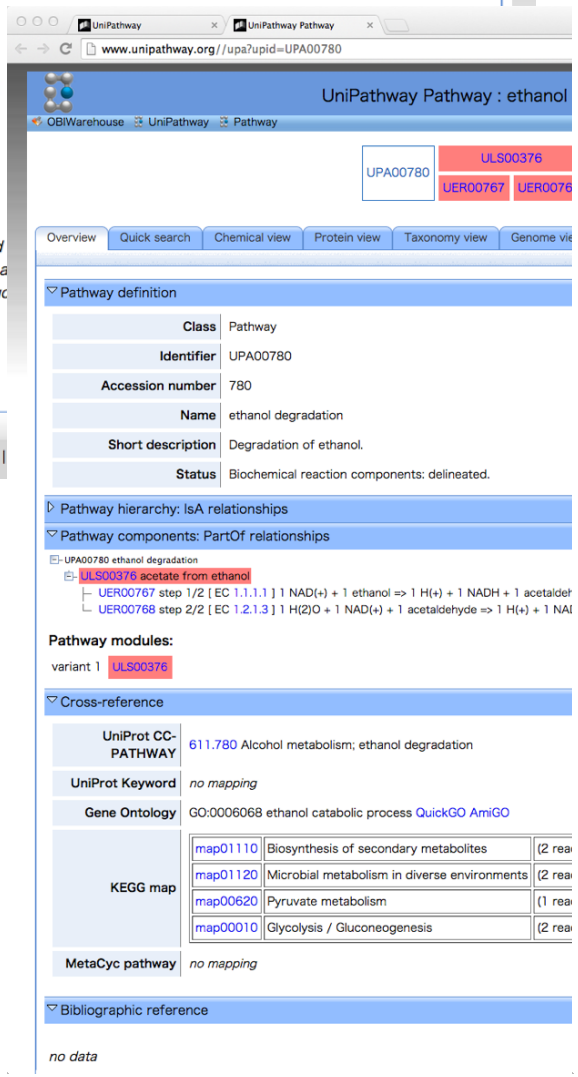

Reactome Pathway Database
www.reactome.org
REACTOME
A CURATED PATHWAY DATABASE
About Content Documentation Tools Community Download Contact
e.g. O95631, NTN1, signa
Search
Browse Pathways
User Guide
About Reactome
Reactome is a free, open-source, curated and peer reviev
intuitive bioinformatics tools for the visualization, interpret
support basic research, genome analysis, modeling, syste
(v53) of Reactome was released on July 2, 2015.
OICR
NYU Langone MEDICAL CENTER
The development of Reactome is supported by a grant from the US Na
Fund, and the European Molecul
About
About Reactome
News
Reactome Team
Scientific Advisory Board
Other Reactomes
License Agreement
Reactome Disclaimer
Contact
Content
Table of Contents
DOIs
Data Schema
Editorial Calendar
Statistics
ORCID Integration Project
PB | HK1,2,3,GCK phospho
www.reactome.org/PathwayBrowser/#DIAGRAM=71387&ID=70420&PATH=1430728
Pathways for: Homo sapiens
Tour this pathway browser? Hide
Protein
Small molecule
Complex
Event Hierarchy:
Cycle
Cell communication
Disease
Repair
Replication
Immune System
Metabolism
Metabolism of carbohydrates
Digestion of dietary carbohydrate
Hexose transport
Transport (influx) of fructose by GLUT5
Transport (influx) of glucose, galactose
SLC2A2 tetramer transports Fru, Gal,
Glucose transport
SLC2A1-4 tetramers transport Glc fr
GLUT1 tetramer transports Glc from
SLC2A2 transports Glc from cytosol
GLUT1 tetramer binds 4xATP
GLUT1:ATP tetramer dissociates to
HK1,2,3,GCK phosphorylate Glc to
Regulation of Glucokinase by Glucol
Cytosolic glucose 6-phosphate is ex
G6PC hydrolyzes glucose 6-phosph
G6PC2 hydrolyzes glucose 6-phospl
G6PC3 hydrolyzes glucose 6-phospl
Efflux of glucose from the endoplasn
Glucose metabolism
Glycolysis
HK1,2,3,GCK phosphorylate Glc to f
Highlight All
Overview
Description
Molecules
Structures
Expression
Analysis
Processes
Downloads
HK1,2,3,GCK phosphorylate Glc to form G6P
Species: Homo sapiens
Stable Identifier
R-HSA-70420.1
Summation
Cytosolic glucokinase and the three isoforms of hexokinase catalyze the irreversible reaction of glucose and ATP to form glucose 6 phosphate and ADP. In the body glucokinase is found only in hepatocytes and pancreatic beta cells. Glucokinase and the hexokinase enzymes differ in that glucokinase has a higher Km than the hexokinases and is less readily inhibited by the reaction product. As a result, glucokinase should be inactive in the fasting state when glucose concentrations are low but in the fed state should have an activity proportional to glucose concentration. These features are thought to enable efficient glucose uptake and retention in the liver, and to function as a sensor of glucose concentration coupled to insulin release in pancreatic beta cells (Thorens 2001). Glucokinase mutations are associated with MODY2, a heritable early onset form of type II diabetes (Tanizawa et al. 1991; Takeda et al. 1993). Three human hexokinase enzymes have been characterized, HK1 (Aleshin et al. 1998), HK2 (Lehto et al. 1995), and HK3 (Rijksen et al. 1982).

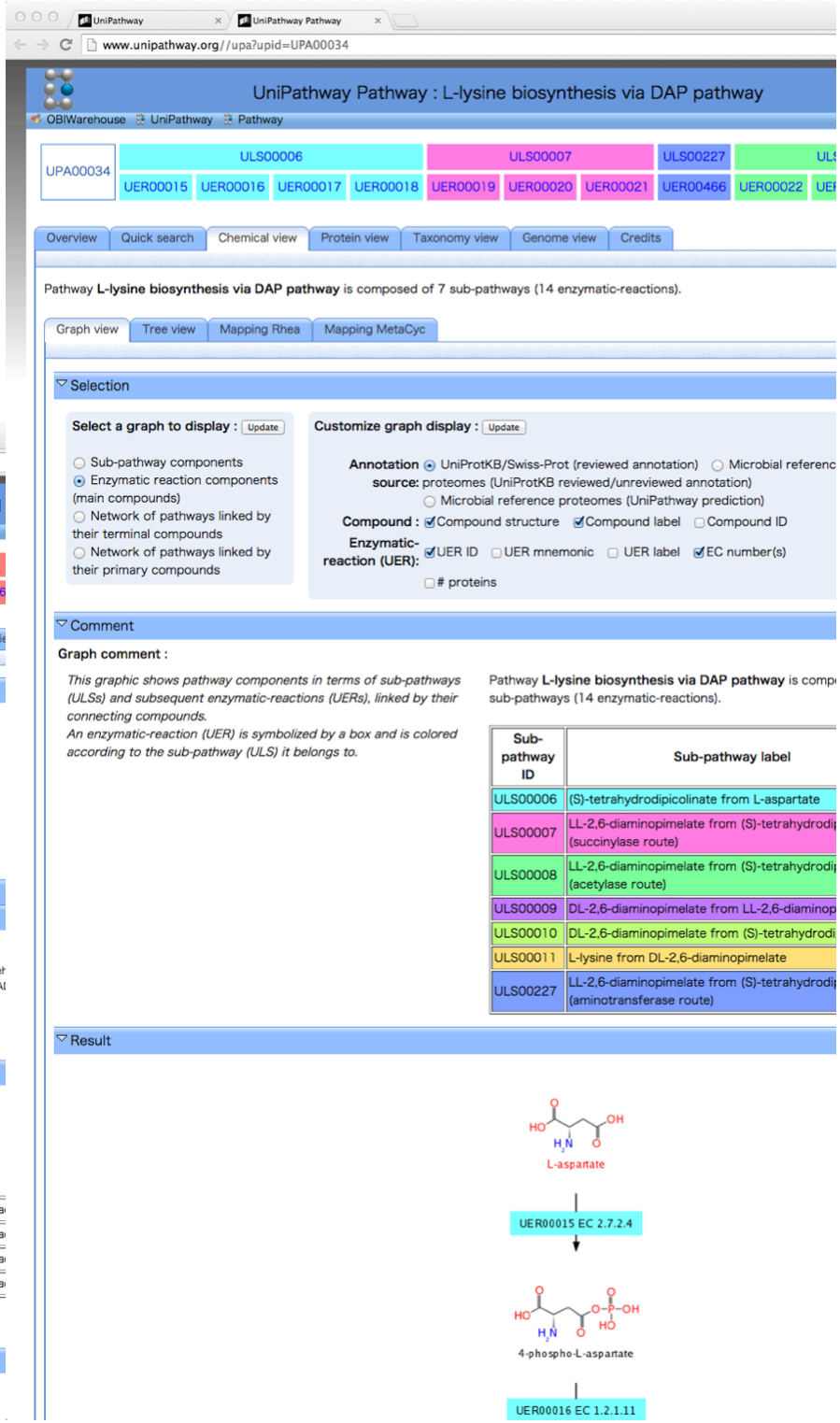
UniPathway

UniPathway

- ウェブサイト
  - http://www.unipathway.org/
- 開発元
  - スイス SIB, フランス INRIA, Joseph Fourier大学
- 対象
  - KEGG, MetaCyc, 特に UniProt のアノテーションにリンク
  - マニュアルキュレーション（手動）
- 論文
  - Nucleic Acids Res. 2012 Jan;40(DB issue):D761-9.


UniPathway
www.unipathway.org
UniPathway: a resource for the exploration of metabolic pathways
OBIWarehouse UniPathway
Welcome Quick search Browse pathway Browse compound Browse organism Download About UniPathway Credits
UniPathway
SIB
Inria
UniPathway is a manually curated resource of enzyme-catalyzed and
It provides a hierarchical representation of metabolic pathways and a
UniPathway data are cross-linked to existing metabolic resources suc
Prabi Grenoble

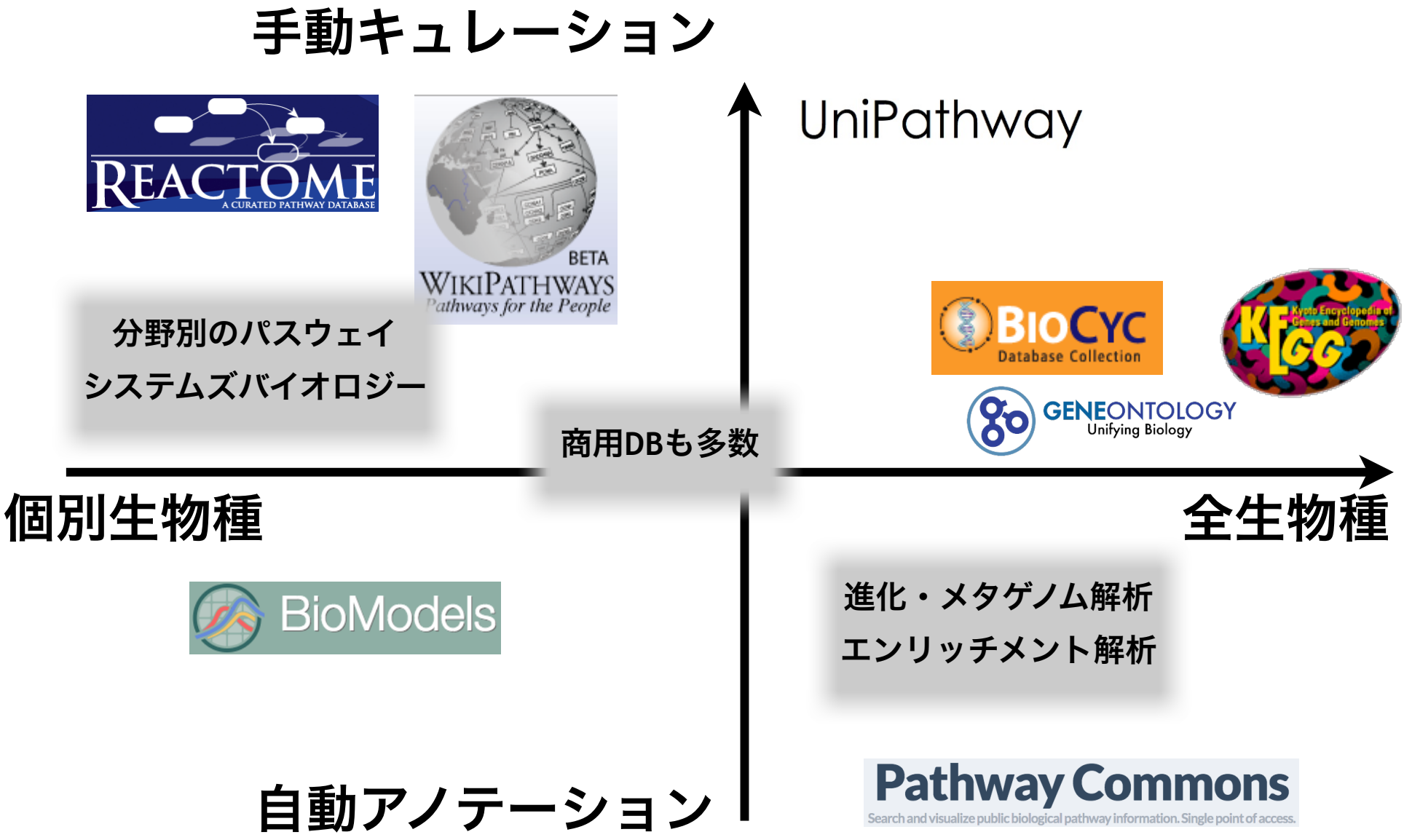

UniPathway
UniPathway Pathway
www.unipathway.org//upa?upid=UPA00780
UniPathway Pathway : ethanol
OBIWarehouse UniPathway Pathway
UPA00780
ULS00376
UER00767
Overview Quick search Chemical view Protein view Taxonomy view
Pathway definition
Class Pathway
Identifier UPA00780
Accession number 780
Name ethanol degradation
Short description Degradation of ethanol.
Status Biochemical reaction components: delineated.
Pathway hierarchy: IsA relationships
Pathway components: PartOf relationships
UPA00780 ethanol degradation
ULS00376 acetate from ethanol
UER00767 step 1/2 [EC 1.1.1.1] 1 NAD(+) + 1 ethanol => 1 H(+) + 1 NADH + 1 acetaldeh
UER00768 step 2/2 [EC 1.2.1.3] 1 H(2)O + 1 NAD(+) + 1 acetaldehyde => 1 H(+) + 1 NAI
Pathway modules:
variant 1 ULS00376
Cross-reference
UniProt CC-PATHWAY 611.780 Alcohol metabolism; ethanol degradation
UniProt Keyword no mapping
Gene Ontology GO:0006068 ethanol catabolic process QuickGO AmiGO
KEGG map
map01110 Biosynthesis of secondary metabolites
map01120 Microbial metabolism in diverse environments
map00620 Pyruvate metabolism
map00010 Glycolysis / Gluconeogenesis
MetaCyc pathway no mapping
Bibliographic reference
no data


UniPathway
UniPathway Pathway
www.unipathway.org//upa?upid=UPA00034
UniPathway Pathway : L-lysine biosynthesis via DAP pathway
OBIWarehouse UniPathway Pathway
UPA00034
ULS00006 ULS00007 ULS00227
UER00015 UER00016 UER00017 UER00018 UER00019 UER00020 UER00021 UER00466 UER00022
Overview Quick search Chemical view Protein view Taxonomy view Genome view Credits
Pathway L-lysine biosynthesis via DAP pathway is composed of 7 sub-pathways (14 enzymatic-reactions).
Graph view Tree view Mapping Rhea Mapping MetaCyc
Selection
Select a graph to display : Update
Sub-pathway components
Enzymatic reaction components (main compounds)
Network of pathways linked by their terminal compounds
Network of pathways linked by their primary compounds
Customize graph display : Update
Annotation source: UniProtKB/Swiss-Prot (reviewed annotation) Microbial referenc
proteomes (UniProtKB reviewed/unreviewed annotation)
Microbial reference proteomes (UniPathway prediction)
Compound : Compound structure Compound label Compound ID
Enzymatic-reaction (UER): UER ID UER mnemonic UER label EC number(s)
proteins
Comment
Graph comment :
This graphic shows pathway components in terms of sub-pathways (ULSs) and subsequent enzymatic-reactions (UERs), linked by their connecting compounds.
An enzymatic-reaction (UER) is symbolized by a box and is colored according to the sub-pathway (ULS) it belongs to.
Pathway L-lysine biosynthesis via DAP pathway is compo
sub-pathways (14 enzymatic-reactions).
Sub-pathway ID
Sub-pathway label
ULS00006 (S)-tetrahydrodipicolinate from L-aspartate
ULS00007 LL-2,6-diaminopimelate from (S)-tetrahydrodi (succinylase route)
ULS00008 LL-2,6-diaminopimelate from (S)-tetrahydrodi (acetylase route)
ULS00009 DL-2,6-diaminopimelate from LL-2,6-diaminop
ULS00010 DL-2,6-diaminopimelate from (S)-tetrahydrodi
ULS00011 L-lysine from DL-2,6-diaminopimelate
ULS00227 LL-2,6-diaminopimelate from (S)-tetrahydrodi (aminotransferase route)
Result
L-aspartate
UER00015 EC 2.7.2.4
4-phospho-L-aspartate
UER00016 EC 1.2.1.11

パスウェイデータベースの利用

- パスウェイデータベースの進展
  - 生化学中心のリファレンス整備: KEGG, BioCyc, UniPathway, ...
  - ヒト中心のリファレンス整備: Reactome, PID, AlzPathway, ...
  - 統合化: Gene Ontology, Pathway Commons, WikiPathways, ...
- パスウェイデータベースの形式
  - KGML, SBML, CSML, BioPAX, ...
- パスウェイデータベースのツール
  - エンリッチメント解析: R, g:Profiler, Ingenuity (IPA), ...
  - シミュレーション: CellDesigner, Cell Illustrator, Gepasi, ...
  - 可視化: Cytoscape, ...

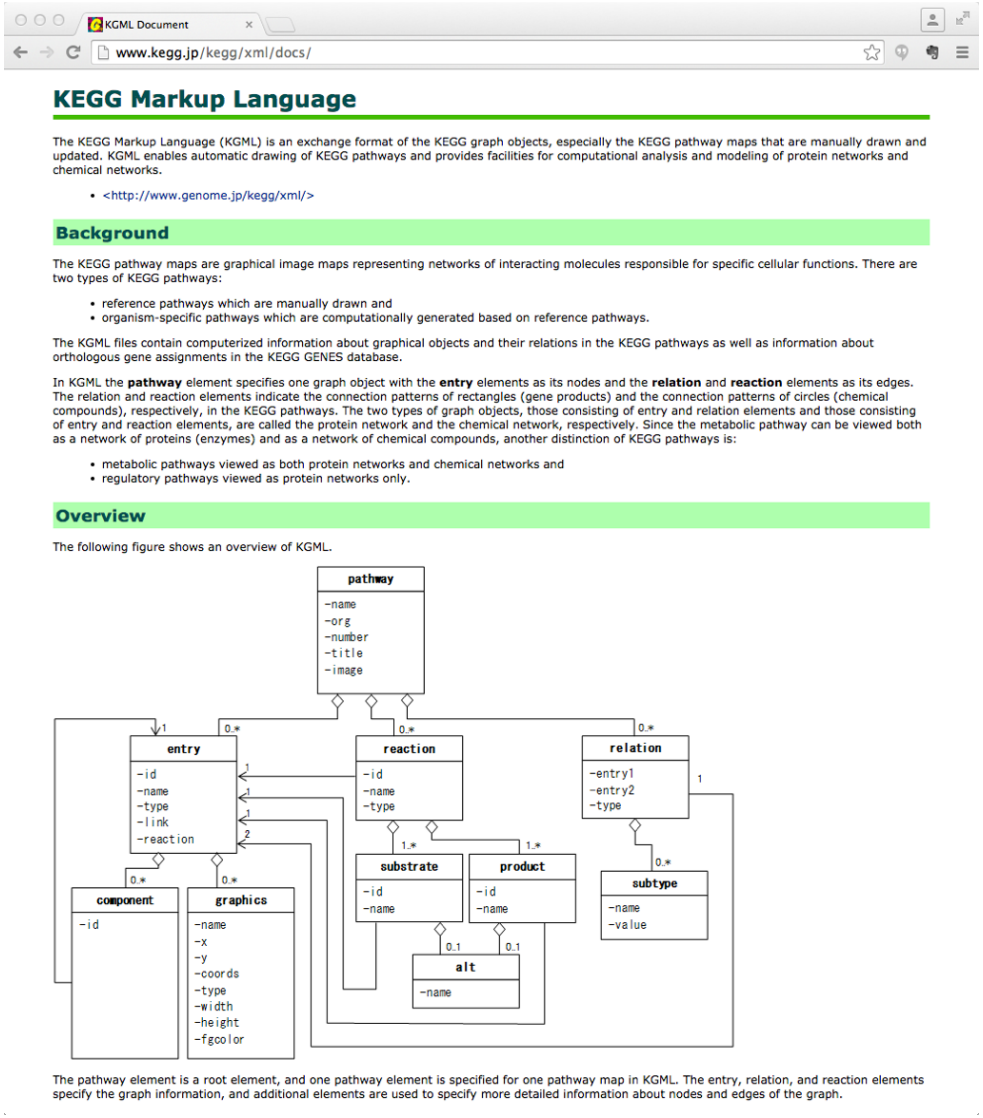
パスウェイデータベースの分類



パスウェイデータベースの形式

- KEGG
  - KGML: KEGG 独自の XML 形式
    - 生体分子とレイアウトを含む、パスウェイはスタティック
- システムバイオロジー
  - SBML: BioModels, Cell Designer 等で利用される XML 形式
    - シミュレーションのため、分子の変化や制御をモデル化
- 標準化
  - BioPAX: Pathway Commons 等で利用される RDF/OWL 形式
    - さまざまなパスウェイ表現の標準化を目指す
- 現状
  - パスウェイDBの一覧と対応形式 http://www.pathguide.org/

KGML: KEGG Markup Language



KEGGパスウェイの酵素・基質・生成物などの関係と、図のレイアウトをコンピュータで取り扱えるようにするための記述言語

京都大学の金久實らが開発

http://www.kegg.jp/kegg/xml/

SBML：System Biology Markup Language

パスウェイにおける生化学反応をシミュレーション等、コンピュータで取り扱えるようにするための記述言語

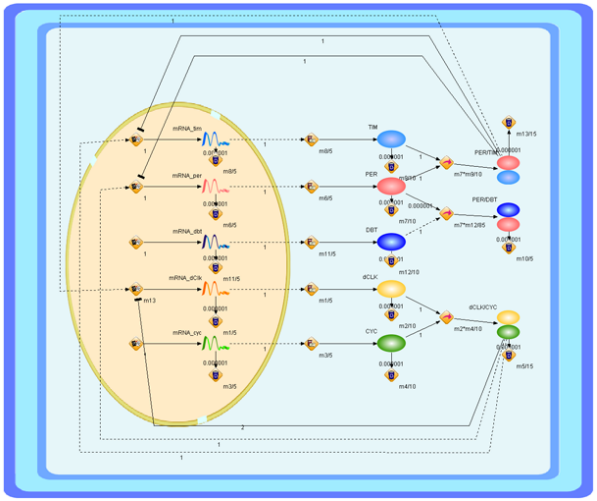
ERATO/ソニーの北野宏明らが開発

http://sbml.org/

CSML：Cell System Markup Language

パスウェイのモデリング・可視化・シミュレーションのための記述言語

東大医科研の宮野悟らが開発



http://csml.org/

BioPAX: Biological Pathway Exchange

数百を超えるパスウェイデータベースの、データ交換フォーマットを定義する国際コンソーシアム

http://www.biopax.org/

PathwayCommons

- ウェブサイト
  - http://www.unipathway.org/
- 開発元
  - 米国 MSKCC、カナダ トロント大
- 対象
  - 様々なパスウェイデータベースからのコレクション
  - BioPAX 形式に標準化
- 論文
  - Nucleic Acids Res. 2011 Jan;39(DB issue):D685-90

Pathway Commons: A Res

www.pathwaycommons.org/about/

Pathway Commons | Download | F.A.Q. | Publications | Contact

March 4, 2015 - We have released a new version of Pathway Commons 2 (v7), which serves 31,698 pathways and 1,151,476 interactions from 18 data sources.

Pathway Commons

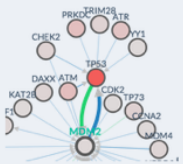
Search and visualize public biological pathway information. Single point of access.

BRCA1, BRCA2, MDM2 | Start exploring »

Pathway Commons is a network biology resource and acts as a convenient point of access to biological pathway information collected from public pathway databases, which you can search, visualize and download. All data is freely available, under the license terms of each contributing database.

For biologists

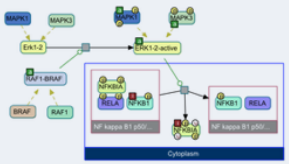
Search, visualize and download Pathway Commons pathways as part of an integrated network analysis (more)



Simple

See genes in pathway context

PCViz



Advanced

See detailed processes

ChiBE

Analyze

Search and analyze pathway relationships

CyPath2

For computational biologists and software developers

Download all pathways in BioPAX, SIF and other formats for pathway and network analysis. Build software on top of Pathway Commons using our web service API (more)

PCViz: Pathway Commons

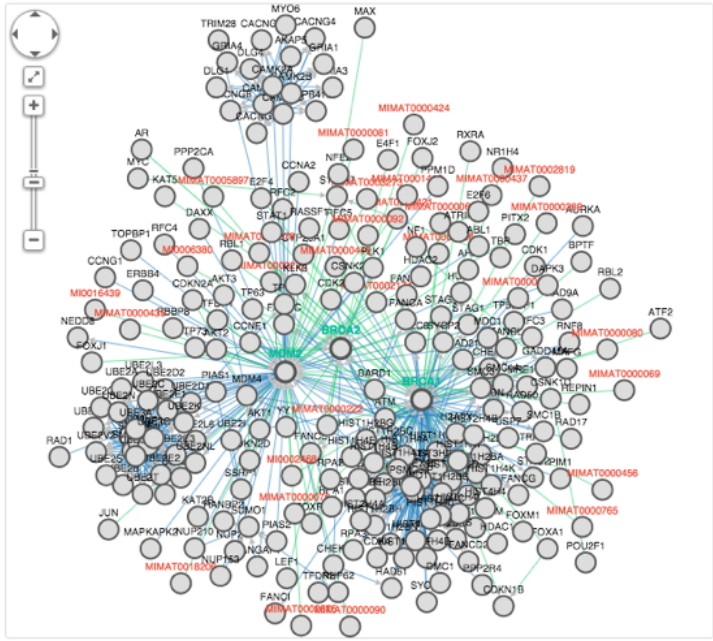
www.pathwaycommons.org/pcviz/#neighborhood/BRCA1, BRCA2, MDM2

PCViz Pathway Commons Network Visualizer

Genes of interest

BRCA1 | BRCA2 | MDM2 | +

Details



Click on genes in details...

Download | Embed | Reset | Full screen

About

PCViz is an open-source web-based network visualization tool that helps users query Pathway Commons and obtain details about genes and their interactions extracted from multiple pathway data resources.

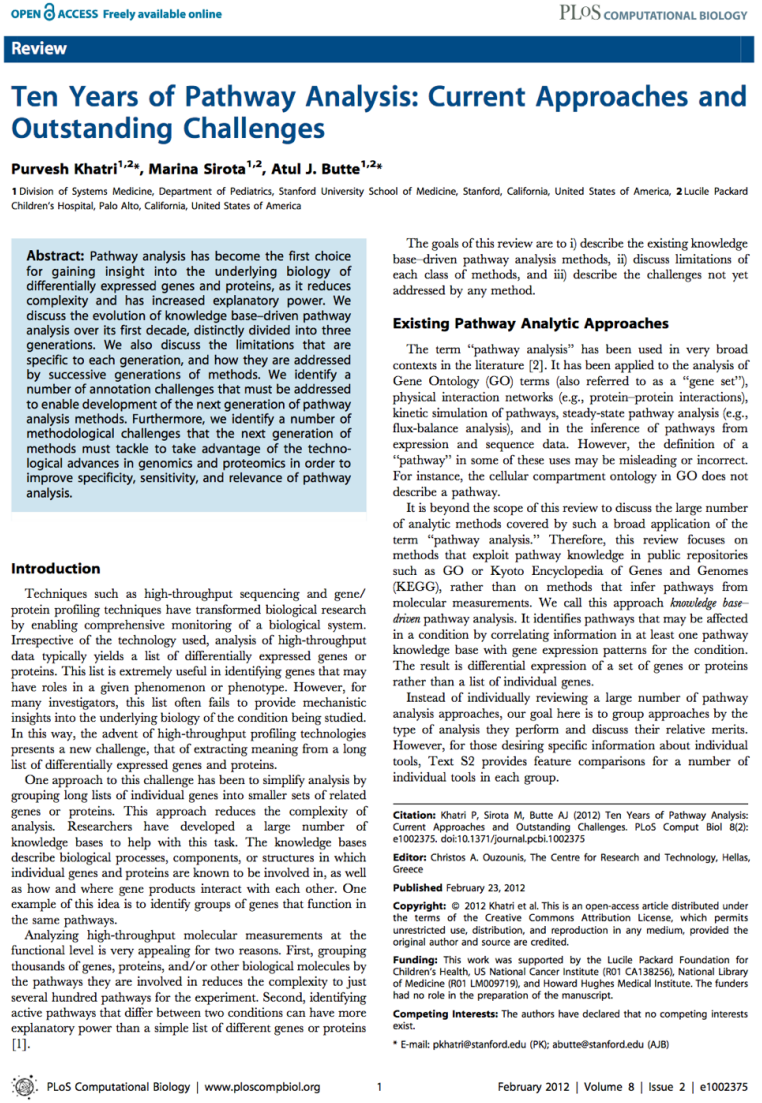
実習

パスウェイデータベースの利用

- 現状 Gene Ontology や KEGG など網羅的なパスウェイ DB を用いた解析が多い

- 手法は大きく分けて３世代

1. Over-Representation Analysis (ORA)
2. Functional Class Scoring analysis (FCS)
3. Pathway Topology-based analysis (PT)



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Review

Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges

Purvesh Khatri[1,2]*, Marina Sirota[1,2], Atul J. Butte[1,2]*

1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States of America, **2** Lucile Packard Children's Hospital, Palo Alto, California, United States of America

Abstract: Pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins, as it reduces complexity and has increased explanatory power. We discuss the evolution of knowledge base–driven pathway analysis over its first decade, distinctly divided into three generations. We also discuss the limitations that are specific to each generation, and how they are addressed by successive generations of methods. We identify a number of annotation challenges that must be addressed to enable development of the next generation of pathway analysis methods. Furthermore, we identify a number of methodological challenges that the next generation of methods must tackle to take advantage of the technological advances in genomics and proteomics in order to improve specificity, sensitivity, and relevance of pathway analysis.

Introduction

Techniques such as high-throughput sequencing and gene/protein profiling techniques have transformed biological research by enabling comprehensive monitoring of a biological system. Irrespective of the technology used, analysis of high-throughput data typically yields a list of differentially expressed genes or proteins. This list is extremely useful in identifying genes that may have roles in a given phenomenon or phenotype. However, for many investigators, this list often fails to provide mechanistic insights into the underlying biology of the condition being studied. In this way, the advent of high-throughput profiling technologies presents a new challenge, that of extracting meaning from a long list of differentially expressed genes and proteins.

One approach to this challenge has been to simplify analysis by grouping long lists of individual genes into smaller sets of related genes or proteins. This approach reduces the complexity of analysis. Researchers have developed a large number of knowledge bases to help with this task. The knowledge bases describe biological processes, components, or structures in which individual genes and proteins are known to be involved in, as well as how and where gene products interact with each other. One example of this idea is to identify groups of genes that function in the same pathways.

Analyzing high-throughput molecular measurements at the functional level is very appealing for two reasons. First, grouping thousands of genes, proteins, and/or other biological molecules by the pathways they are involved in reduces the complexity to just several hundred pathways for the experiment. Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of different genes or proteins [1].

The goals of this review are to i) describe the existing knowledge base–driven pathway analysis methods, ii) discuss limitations of each class of methods, and iii) describe the challenges not yet addressed by any method.

Existing Pathway Analytic Approaches

The term "pathway analysis" has been used in very broad contexts in the literature [2]. It has been applied to the analysis of Gene Ontology (GO) terms (also referred to as a "gene set"), physical interaction networks (e.g., protein–protein interactions), kinetic simulation of pathways, steady-state pathway analysis (e.g., flux-balance analysis), and in the inference of pathways from expression and sequence data. However, the definition of a "pathway" in some of these uses may be misleading or incorrect. For instance, the cellular compartment ontology in GO does not describe a pathway.

It is beyond the scope of this review to discuss the large number of analytic methods covered by such a broad application of the term "pathway analysis." Therefore, this review focuses on methods that exploit pathway knowledge in public repositories such as GO or Kyoto Encyclopedia of Genes and Genomes (KEGG), rather than on methods that infer pathways from molecular measurements. We call this approach *knowledge base–driven* pathway analysis. It identifies pathways that may be affected in a condition by correlating information in at least one pathway knowledge base with gene expression patterns for the condition. The result is differential expression of a set of genes or proteins rather than a list of individual genes.

Instead of individually reviewing a large number of pathway analysis approaches, our goal here is to group approaches by the type of analysis they perform and discuss their relative merits. However, for those desiring specific information about individual tools, Text S2 provides feature comparisons for a number of individual tools in each group.

Citation: Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol 8(2): e1002375. doi:10.1371/journal.pcbi.1002375

Editor: Christos A. Ouzounis, The Centre for Research and Technology, Hellas, Greece

Published February 23, 2012

Funding: This work was supported by the Lucile Packard Foundation for Children's Health, US National Cancer Institute (R01 CA138256), National Library of Medicine (R01 LM009719), and Howard Hughes Medical Institute. The funders had no role in the preparation of the manuscript.

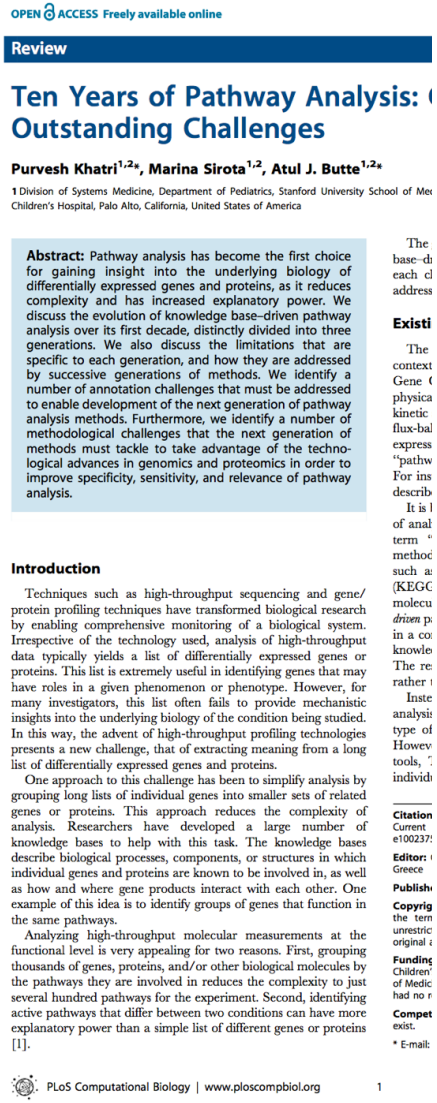
Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pkhatri@stanford.edu (PK); abutte@stanford.edu (AJB)

パスウェイデータベースの利用

1. Over-Representation Analysis (ORA)

- 遺伝子群がどのパスウェイに主に関係しているかを推定
  - 与える遺伝子リストを発現変動の閾値などで指定
  - 各パスウェイに出現する遺伝子頻度から有意性を検定
- ツール
  - GenMAPP や GOstat など多数
- 課題
  - 発現強度などが閾値以下の遺伝子は解析の対象外に
  - 検定では遺伝子数しか使わず、発現強度などは無視
  - 個々の遺伝子やパスウェイが相互に関係ないことを仮定



OPEN ACCESS Freely available online

Review

Ten Years of Pathway Analysis: Outstanding Challenges

Purvesh Khatri[1,2]*, Marina Sirota[1,2], Atul J. Butte[1,2]*

1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Me
Children's Hospital, Palo Alto, California, United States of America

Abstract: Pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins, as it reduces complexity and has increased explanatory power. We discuss the evolution of knowledge base–driven pathway analysis over its first decade, distinctly divided into three generations. We also discuss the limitations that are specific to each generation, and how they are addressed by successive generations of methods. We identify a number of annotation challenges that must be addressed to enable development of the next generation of pathway analysis methods. Furthermore, we identify a number of methodological challenges that the next generation of methods must tackle to take advantage of the technological advances in genomics and proteomics in order to improve specificity, sensitivity, and relevance of pathway analysis.

Introduction

Techniques such as high-throughput sequencing and gene/protein profiling techniques have transformed biological research by enabling comprehensive monitoring of a biological system. Irrespective of the technology used, analysis of high-throughput data typically yields a list of differentially expressed genes or proteins. This list is extremely useful in identifying genes that may have roles in a given phenomenon or phenotype. However, for many investigators, this list often fails to provide mechanistic insights into the underlying biology of the condition being studied. In this way, the advent of high-throughput profiling technologies presents a new challenge, that of extracting meaning from a long list of differentially expressed genes and proteins.

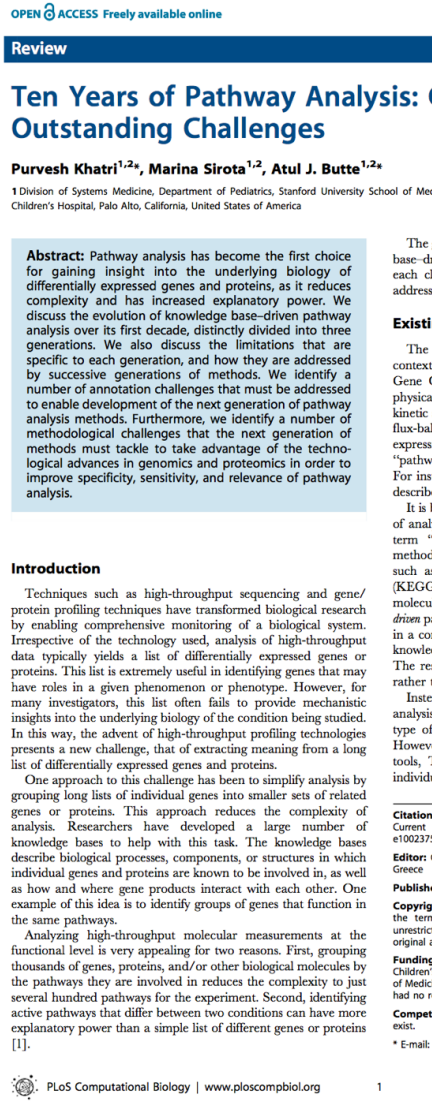
One approach to this challenge has been to simplify analysis by grouping long lists of individual genes into smaller sets of related genes or proteins. This approach reduces the complexity of analysis. Researchers have developed a large number of knowledge bases to help with this task. The knowledge bases describe biological processes, components, or structures in which individual genes and proteins are known to be involved in, as well as how and where gene products interact with each other. One example of this idea is to identify groups of genes that function in the same pathways.

Analyzing high-throughput molecular measurements at the functional level is very appealing for two reasons. First, grouping thousands of genes, proteins, and/or other biological molecules by the pathways they are involved in reduces the complexity to just several hundred pathways for the experiment. Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of different genes or proteins [1].

パスウェイデータベースの利用

2. Functional Class Scoring analysis (FCS)

- 変動の弱い遺伝子もパスウェイに影響がある可能性を考慮
  - 各遺伝子の発現レベル変動など統計値を計算
  - パスウェイ中の全遺伝子の統計値を集計
  - パスウェイに含まれる遺伝子の影響の有意さを検定
- ツール
  - GSEA や R の BioConductor パッケージなど多数
- 課題
  - ORA と同様 FCA も各パスウェイを独立に解析
    - 実際には複数のパスウェイに関係する遺伝子も多い
  - 遺伝子変動からランク付けしたあとは変動量は無視されがち



OPEN ACCESS Freely available online

Review

Ten Years of Pathway Analysis: (
Outstanding Challenges

Purvesh Khatri[1,2]*, Marina Sirota[1,2], Atul J. Butte[1,2]*

1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Me
Children's Hospital, Palo Alto, California, United States of America

Abstract: Pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins, as it reduces complexity and has increased explanatory power. We discuss the evolution of knowledge base–driven pathway analysis over its first decade, distinctly divided into three generations. We also discuss the limitations that are specific to each generation, and how they are addressed by successive generations of methods. We identify a number of annotation challenges that must be addressed to enable development of the next generation of pathway analysis methods. Furthermore, we identify a number of methodological challenges that the next generation of methods must tackle to take advantage of the technological advances in genomics and proteomics in order to improve specificity, sensitivity, and relevance of pathway analysis.

Introduction

Techniques such as high-throughput sequencing and gene/protein profiling techniques have transformed biological research by enabling comprehensive monitoring of a biological system. Irrespective of the technology used, analysis of high-throughput data typically yields a list of differentially expressed genes or proteins. This list is extremely useful in identifying genes that may have roles in a given phenomenon or phenotype. However, for many investigators, this list often fails to provide mechanistic insights into the underlying biology of the condition being studied. In this way, the advent of high-throughput profiling technologies presents a new challenge, that of extracting meaning from a long list of differentially expressed genes and proteins.

One approach to this challenge has been to simplify analysis by grouping long lists of individual genes into smaller sets of related genes or proteins. This approach reduces the complexity of analysis. Researchers have developed a large number of knowledge bases to help with this task. The knowledge bases describe biological processes, components, or structures in which individual genes and proteins are known to be involved in, as well as how and where gene products interact with each other. One example of this idea is to identify groups of genes that function in the same pathways.

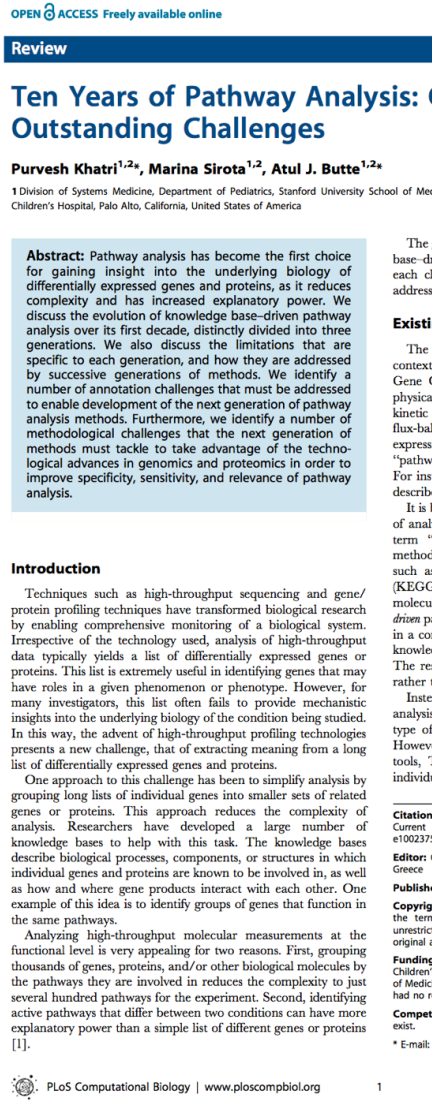
Analyzing high-throughput molecular measurements at the functional level is very appealing for two reasons. First, grouping thousands of genes, proteins, and/or other biological molecules by the pathways they are involved in reduces the complexity to just several hundred pathways for the experiment. Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of different genes or proteins [1].

PLoS Computational Biology | www.ploscompbiol.org 1

パスウェイデータベースの利用

3. Pathway Topology-based analysis (PT)

- 相互作用する遺伝子間の関係も考慮
  - GO, MsigDB: パスウェイに対応する遺伝子のリストのみ
  - KEGG, MetaCyc, Reactome, RegulonDB, STKE, BioCarta, PantherDB: 遺伝子産物がどこでどのように相互作用するか
  - 基本的には FCS と同じ手法でパスウェイのトポロジーを考慮
    - パスウェイ中の遺伝子間の類似性を反応経路上の近さで定義
- ツール
  - ScorePAGE や NetGSA などいくつか
- 課題
  - 実際のパスウェイは細胞ごとに違う可能性
  - パスウェイの動的な状態は表現できない
  - パスウェイ間の関係はを考慮できていない



OPEN ACCESS Freely available online

Review

Ten Years of Pathway Analysis: Outstanding Challenges

Purvesh Khatri[1,2]*, Marina Sirota[1,2], Atul J. Butte[1,2]*

1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Me
Children's Hospital, Palo Alto, California, United States of America

Abstract: Pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins, as it reduces complexity and has increased explanatory power. We discuss the evolution of knowledge base–driven pathway analysis over its first decade, distinctly divided into three generations. We also discuss the limitations that are specific to each generation, and how they are addressed by successive generations of methods. We identify a number of annotation challenges that must be addressed to enable development of the next generation of pathway analysis methods. Furthermore, we identify a number of methodological challenges that the next generation of methods must tackle to take advantage of the technological advances in genomics and proteomics in order to improve specificity, sensitivity, and relevance of pathway analysis.

Introduction

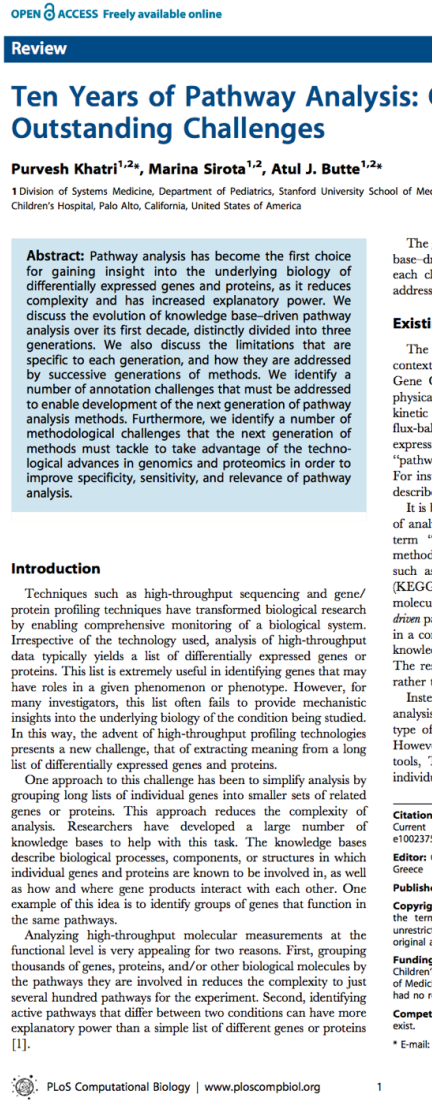
Techniques such as high-throughput sequencing and gene/protein profiling techniques have transformed biological research by enabling comprehensive monitoring of a biological system. Irrespective of the technology used, analysis of high-throughput data typically yields a list of differentially expressed genes or proteins. This list is extremely useful in identifying genes that may have roles in a given phenomenon or phenotype. However, for many investigators, this list often fails to provide mechanistic insights into the underlying biology of the condition being studied. In this way, the advent of high-throughput profiling technologies presents a new challenge, that of extracting meaning from a long list of differentially expressed genes and proteins.

One approach to this challenge has been to simplify analysis by grouping long lists of individual genes into smaller sets of related genes or proteins. This approach reduces the complexity of analysis. Researchers have developed a large number of knowledge bases to help with this task. The knowledge bases describe biological processes, components, or structures in which individual genes and proteins are known to be involved in, as well as how and where gene products interact with each other. One example of this idea is to identify groups of genes that function in the same pathways.

Analyzing high-throughput molecular measurements at the functional level is very appealing for two reasons. First, grouping thousands of genes, proteins, and/or other biological molecules by the pathways they are involved in reduces the complexity to just several hundred pathways for the experiment. Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of different genes or proteins [1].

パスウェイデータベースの課題

- アノテーションの課題
  - パスウェイDBは知識ベースとして解像度が不足
    - 遺伝子のバリアントごとに機能が違う可能性
    - SNP がどのようにパスウェイに影響を及ぼすか
    - → トランスクリプトや SNP レベルのアノテーションが必要
  - 不正確なアノテーション
    - GO ではヒト 18587 遺伝子がアノテーションされている
    - NCBI Gene では 45283 遺伝子あり、内 14162 は偽遺伝子
    - → 偽遺伝子由来の siRNAs も遺伝子発現に関与
    - 多くの遺伝子アノテーションは IEA (Inferred from Electronic Annotations)
  - 実験条件や細胞種特異的な情報の記載が欠けている
    - パスウェイの知識の多くはこれらの異なる実験由来



OPEN ACCESS Freely available online

Review

Ten Years of Pathway Analysis: Outstanding Challenges

Purvesh Khatri[1,2*], Marina Sirota[1,2], Atul J. Butte[1,2*]

1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Me
Children's Hospital, Palo Alto, California, United States of America

Abstract: Pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins, as it reduces complexity and has increased explanatory power. We discuss the evolution of knowledge base–driven pathway analysis over its first decade, distinctly divided into three generations. We also discuss the limitations that are specific to each generation, and how they are addressed by successive generations of methods. We identify a number of annotation challenges that must be addressed to enable development of the next generation of pathway analysis methods. Furthermore, we identify a number of methodological challenges that the next generation of methods must tackle to take advantage of the technological advances in genomics and proteomics in order to improve specificity, sensitivity, and relevance of pathway analysis.

Introduction

Techniques such as high-throughput sequencing and gene/protein profiling techniques have transformed biological research by enabling comprehensive monitoring of a biological system. Irrespective of the technology used, analysis of high-throughput data typically yields a list of differentially expressed genes or proteins. This list is extremely useful in identifying genes that may have roles in a given phenomenon or phenotype. However, for many investigators, this list often fails to provide mechanistic insights into the underlying biology of the condition being studied. In this way, the advent of high-throughput profiling technologies presents a new challenge, that of extracting meaning from a long list of differentially expressed genes and proteins.

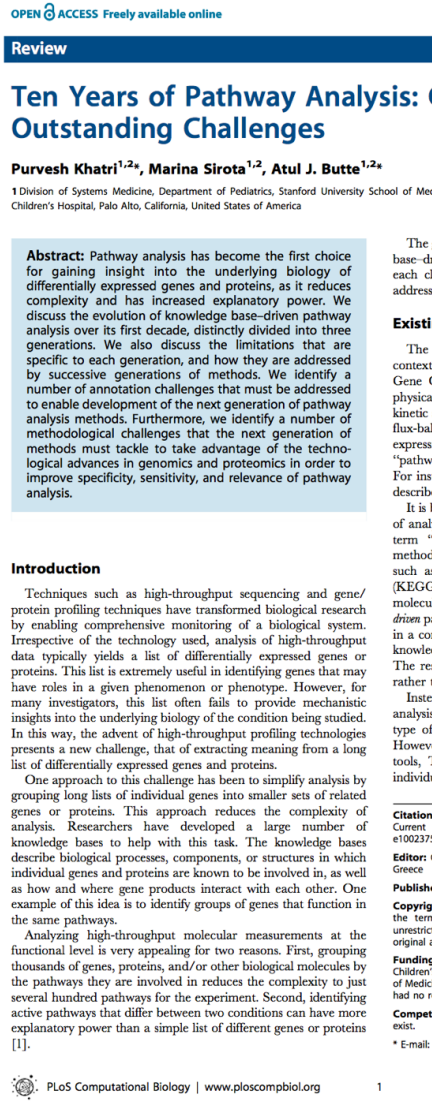
One approach to this challenge has been to simplify analysis by grouping long lists of individual genes into smaller sets of related genes or proteins. This approach reduces the complexity of analysis. Researchers have developed a large number of knowledge bases to help with this task. The knowledge bases describe biological processes, components, or structures in which individual genes and proteins are known to be involved in, as well as how and where gene products interact with each other. One example of this idea is to identify groups of genes that function in the same pathways.

Analyzing high-throughput molecular measurements at the functional level is very appealing for two reasons. First, grouping thousands of genes, proteins, and/or other biological molecules by the pathways they are involved in reduces the complexity to just several hundred pathways for the experiment. Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of different genes or proteins [1].

PLoS Computational Biology | www.ploscompbiol.org 1

パスウェイデータベースの課題

- 手法の課題
  - どの手法が何に優れているのかを比較することが困難
    - 比較するためのベンチマークとなるデータセットが不足
    - DBごとにどの粒度でパスウェイを分けるかも異なる
    - → KEGG では１つのパスウェイでも GO では複数など
  - 動的な応答をモデル化し解析することができない
    - 各パスウェイは相互に独立していると仮定しがち
    - → パスウェイ間全体が１つのシステムとして扱えるとよい
    - 測定された発現情報は変化しないことを仮定しがち
    - → ポジティブ・ネガティブループなど経時的な変化も
  - 外部からの影響をモデル化できていない
    - イオン濃度で影響を受ける転写因子が異なるなど



OPEN ACCESS Freely available online

Review

Ten Years of Pathway Analysis: (
Outstanding Challenges

Purvesh Khatri[1,2]*, Marina Sirota[1,2], Atul J. Butte[1,2]*

1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Me
Children's Hospital, Palo Alto, California, United States of America

Abstract: Pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins, as it reduces complexity and has increased explanatory power. We discuss the evolution of knowledge base–driven pathway analysis over its first decade, distinctly divided into three generations. We also discuss the limitations that are specific to each generation, and how they are addressed by successive generations of methods. We identify a number of annotation challenges that must be addressed to enable development of the next generation of pathway analysis methods. Furthermore, we identify a number of methodological challenges that the next generation of methods must tackle to take advantage of the technological advances in genomics and proteomics in order to improve specificity, sensitivity, and relevance of pathway analysis.

The
base–d
each cl
address

Existi

The
context
Gene (
physica
kinetic
flux-bal
express
"pathw
For ins
describ

It is
of anal
term "
method
such a
(KEGG
molecu
driven p
in a co
knowle
The re
rather

Inste
analysis
type of
Howev
tools, "
individ

Introduction

Techniques such as high-throughput sequencing and gene/protein profiling techniques have transformed biological research by enabling comprehensive monitoring of a biological system. Irrespective of the technology used, analysis of high-throughput data typically yields a list of differentially expressed genes or proteins. This list is extremely useful in identifying genes that may have roles in a given phenomenon or phenotype. However, for many investigators, this list often fails to provide mechanistic insights into the underlying biology of the condition being studied. In this way, the advent of high-throughput profiling technologies presents a new challenge, that of extracting meaning from a long list of differentially expressed genes and proteins.

One approach to this challenge has been to simplify analysis by grouping long lists of individual genes into smaller sets of related genes or proteins. This approach reduces the complexity of analysis. Researchers have developed a large number of knowledge bases to help with this task. The knowledge bases describe biological processes, components, or structures in which individual genes and proteins are known to be involved in, as well as how and where gene products interact with each other. One example of this idea is to identify groups of genes that function in the same pathways.

Analyzing high-throughput molecular measurements at the functional level is very appealing for two reasons. First, grouping thousands of genes, proteins, and/or other biological molecules by the pathways they are involved in reduces the complexity to just several hundred pathways for the experiment. Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of different genes or proteins [1].

Citation
Current
e100237

Editor:
Greece

Publish

Copyrig
the ter
unrestric
original

Funding
Children'
of Medic
had no

Compet
exist.

* E-mail:

PLoS Computational Biology | www.ploscompbiol.org 1

パスウェイデータベースの問題

- 網羅性
  - まだ全ての遺伝子がパスウェイに記述されているわけではない
  - アノテーションの不十分な生物種とオーソログ対応を取ると再構築が不完全に
  - 臓器・細胞種別・遺伝子のバリアント等を表現する精度がない
  - → 研究コミュニティによるアノテーションの継続が必須
- 酵素反応の方向性
  - 矢印で描かれた酵素反応の方向性は、ほとんどが実験的に確かめられていない
  - → 元々 kinetics の情報が集積されてないため。図を鵜呑みにしてはいけない
- 細胞内局在
  - ミトコンドリアやアピコプラストなど、細胞内小器官によって空間的に隔てられている場合も、パスウェイとして分子だけ見ると繋がって見えるので解釈に注意
- 解析に必要な平衡定数や時系列の制御はまだ表現できていない
  - → より高精細なデータを DB 化しシミュレーションへ