SCIENCE CHINA
Life Sciences

SPECIAL TOPIC
• RESEARCH PAPERS •

January 2011 Vol.54 No.1: 60–64
doi: 10.1007/s11427-010-4125-y

TSdb: A database of transporter substrates linking metabolic pathways and transporter systems on a genome scale via their shared substrates

ZHAO Min[1,2], CHEN YanMing[3], QU DaCheng[3] & QU Hong[1*]

[1]*Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, China;*
[2]*Key Laboratory of Evolutionary Systematics of Vertebrates, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China;*
[3]*School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China*

Received November 8, 2010; accepted November 30, 2010

TSdb (http://tsdb.cbi.pku.edu.cn) is the first manually curated central repository that stores formatted information on the substrates of transporters. In total, 37608 transporters with 15075 substrates from 884 organisms were curated from UniProt functional annotation. A unique feature of TSdb is that all the substrates are mapped to identifiers from the KEGG Ligand compound database. Thus, TSdb links current metabolic pathway schema with compound transporter systems via the shared compounds in the pathways. Furthermore, all the transporter substrates in TSdb are classified according to their biochemical properties, biological roles and subcellular localizations. In addition to the functional annotation of transporters, extensive compound annotation that includes inhibitor information from the KEGG Ligand and BRENDA databases has been integrated, making TSdb a useful source for the discovery of potential inhibitory mechanisms linking transporter substrates and metabolic enzymes. User-friendly web interfaces are designed for easy access, query and download of the data. Text and BLAST searches against all transporters in the database are provided. We will regularly update the substrate data with evidence from new publications.

transporter substrate, biological database, compound classification, compound metabolic network

Citation: Zhao M, Chen Y M, Qu D C, *et al*. TSdb: A database of transporter substrates linking metabolic pathways and transporter systems on a genome scale via their shared substrates. Sci China Life Sci, 2011, 54: 60−64, doi: 10.1007/s11427-010-4125-y

Transporters are often membrane proteins that control the exchange of cellular metabolites, drugs, toxins, therapeutic drugs and environmental signals [1,2]. Besides their role in compound exchange and metabolic flux control, transporters, such as GLUT1, which was found to be associated with the Warburg effect and was shown to influence tumorigenic features, are regarded as therapeutic targets [2]. The indispensable roles of transporters in fundamental cellular processes rest with their substrates. Substrates bring nutrition and signals and also output metabolic waste to maintain metabolic states in cells [1]. However, studying an individual transporter system is not sufficient to discover the properties of global organization of transporting systems at the cellular level. Crucial questions at the systems level, such as how transporters coordinate substrate concentration, which types of compounds are more likely to be transported, and what the genome scale distribution patterns of the transporter substrates are, need to be addressed. Answers to these questions are beyond to identify the relationships between transporters and their substrates at the individual level.

Before the use of mass spectrometry, traditional studies of the relationship between transporters and their substrates

*Corresponding author (email: quh@mail.cbi.pku.edu.cn)

limit the number of samples that could be tested. High-throughput metabolomics based on mass spectrometry promises to become a tool that can be used for the comprehensive and quantitative analyses of metabolites on a genome scale [3]. However, genome level investigations of biological metabolites such as transporter substrates require a high quality reference dataset of metabolites. Thus the systematic collection, classification and annotation of transporters and their substrates is an essential first step towards the construction of transporter substrate metabolic networks and a high-throughput survey of their roles in metabolic flux control and the pharmacokinetics of drugs [2,4–8].

The existent transporter databases, TCDB (Transporter Classification Database), TransporterDB, ABCdb (ATP-binding cassette transporters database) and HMTD (Human Membrane Transporter Database) were all constructed for specific purposes like transporter classification or for comparative studies of transporters in different organisms [1,9–12]. Therefore, these databases seldom classify comprehensively the substrate information. In addition, the transporter substrate data in these databases are all in free text and not mapped to any compound database nor are they linked to any of the current pathway databases. Thus, it is still difficult to integrate transporters into any metabolic pathway schema. A comprehensive and well-annotated database of transporter substrates that maps all the substrates to compound identifiers (IDs) from the KEGG Ligand database could be used to link metabolic enzymes and transporters via their shared compounds. Here, we present the first curated transporter substrate database (TSdb) comprising 37608 transporters with 15075 substrates from 884 organisms, all of which are mapped to compound IDs from the KEGG Ligand database. TSdb is available for free at the following URL: http://tsdb.cbi.pku.edu.cn, thus providing a useful resource for biochemists and molecular biologists.

1 Data source and contents

As shown in Figure 1, the comprehensive collection and mapping of transporter substrates to compound IDs from the KEGG Ligand database using the functional annotation from UniProt was achieved in four steps: (i) All UniProt entries with "transport" as a keyword were collected from the UniProt database [13–15]; (ii) functional annotations were extracted using the Swissknife module [16]; (iii) the annotations were grouped based on their content, allowing us to quickly and easily assess if and how the searched entries were related to transporters and providing the means to cross-check between different entries; and (iv) if the functional description from UniProt exactly matched a KEGG compound name, we assigned the KEGG compound to that description. The descriptions with mapped KEGG compounds were read manually to identify the substrates and to map them to the compound ID from KEGG Ligand [17,18]. The substrate dataset can be classified into several common types of compounds: carbohydrates, lipids, peptides, amino acid, nucleic acids, vitamins, and hormones. In total, 37608

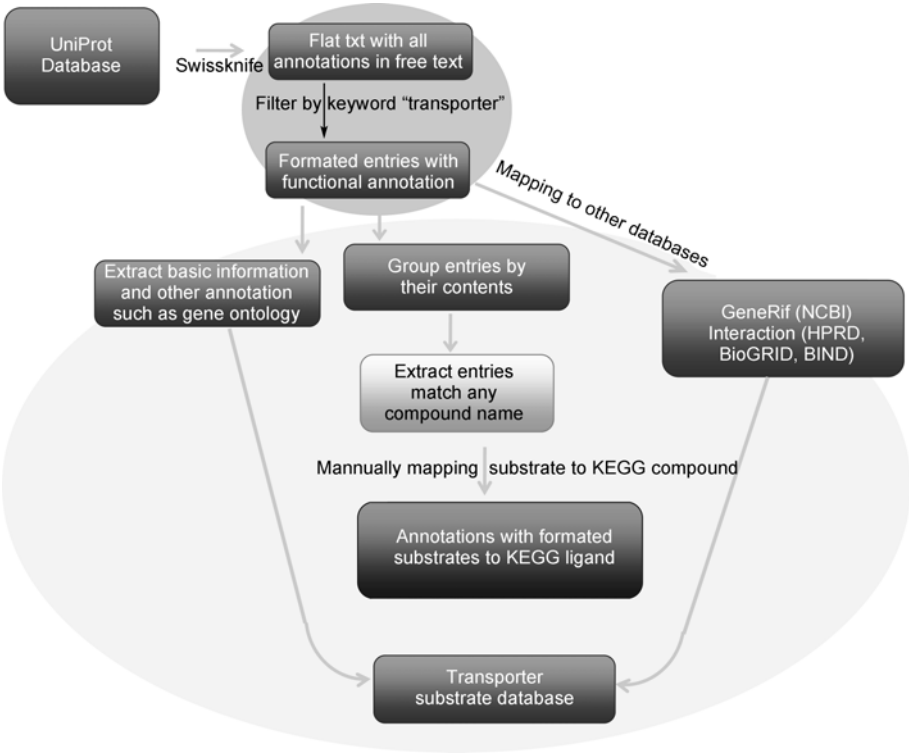

Figure 1 Pipeline for constructing transporter substrate database from UniProt database.

transporters with 15075 substrates from 884 organisms were collected. For accuracy, we only mapped substrates with clear descriptions to KEGG Ligand compound IDs. Transporters with ambiguous description such as "neutral amino acid" were included in the transporter dataset, but were not mapped to any KEGG compound. In addition to the formatted compound information in TSdb, we classified all the substrates according to their biochemical properties and biological functions; for example, enzyme effector, inhibitor, cofactor, GPCR ligand, ion ligand, P450 substrate or neurotransmitter.

Next, we retrieved and integrated extensive functional information about the transporters and their substrates from the public databases GeneRifs [19], UniProt [13,15], KEGG-Ligand 44.0 [18,20], BRENDA 7.1 [21], and HPRD [22]. For the transporters, protein functional annotation, such as subcellular localizations, protein family, gene ontology assignments, and known disease association, was collected from UniProt. For the transporters, 769 protein-protein interaction pairs from the BioGRID [23], HPRD [22], and BIND [24] databases and 12556 GeneRifs entries were also imported. For the substrates, all the enzymes and reactions involved in the anabolism and catabolism of the compounds were summarized from KEGG-Ligand database. All enzyme inhibitor information was extracted from the BRENDA database (version 7.1) and organism-specific inhibitors were recorded using the EC code from BRENDA. A semi-automatic method, similar to the method described in a previous study [25], was used to convert free text inhibitor information to KEGG compound IDs. For each enzyme, if the inhibitor description from BRENDA exactly matched a KEGG compound name, then the KEGG compound was assigned to that description. All assigned KEGG compounds were then grouped together by their KEGG compound IDs and all the mapping results were checked manually. Many man-made inhibitors such as EDTA are not produced *in vivo*. We selected the *in vivo* enzyme products of each organism from the organism-specific inhibitors dataset and found 351 transporter substrates that were also enzyme inhibitors. This information is useful to survey the inhibitory relations between transporter substrates and metabolic enzymes.

2 Database access and web interface

TSdb consists of 10 tables that record the transporters, substrates, enzyme inhibitors, gene-gene interactions and other functional annotations such as GeneRif. It is designed specifically to do genome scale queries and comparisons. All the data in TSdb are stored in a MySQL relational database, which aids the building of custom queries using Structured Query Language (SQL). A user-friendly web interface was implemented in Perl-CGI, the most widely used server-side scripting language and module, running in an Apache environment. As shown in Figure 2A, text based searches include queries for transporters, transporter functional annotations and their substrate counts and contents. As shown in Figure 2B, the browser also allows users to browse all the substrates in TSdb by hierarchy according to their biochemical properties and biological roles. All the organisms and all the substrates in TSdb can also be found using our browser. Sequence-based BLAST searches are provided to identify similar sequences in TSdb (Figure 2D). All transporter-substrate pairs with UniProt Accession Numbers and KEGG Ligand IDs can be free downloaded as tab separated files. Transporter and substrate pages display the main database information. The transporter pages show basic information about the substrates and all the metabolic enzymes inhibited by the substrates (Figure 2C) and integrate extensive functional annotations such as GeneRif, SwissProt keywords and gene-gene interactions. Cross-referenced links to external database, including Entrez Gene, KEGG Ligand, and PubMed, are also provided. In the substrate pages, the enzymes are summarized and the reactions that produce and consume the substrates are listed, forming a useful link be-

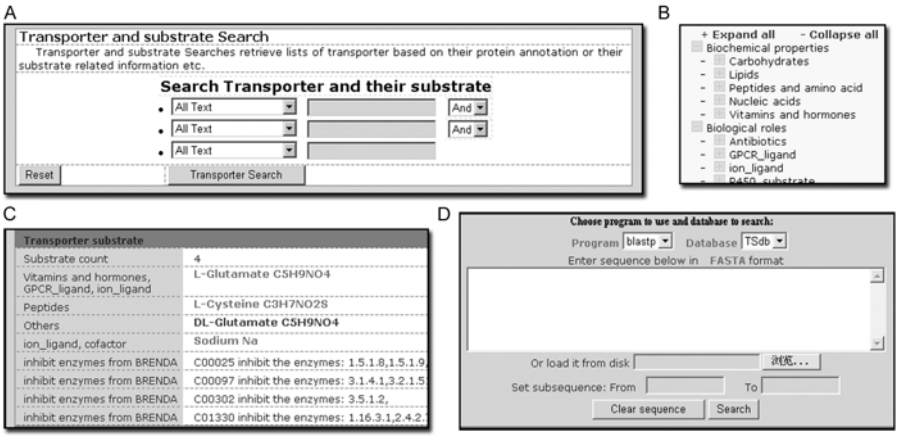

Figure 2 Web interface of TSdb. A, Query interface for text search. B, Data browser by biochemical properties and biological roles of transporter substrates. C, The mapped transporter substrate contents in each transporter page. D, Blast interface for sequence search against all transporters in TSdb.

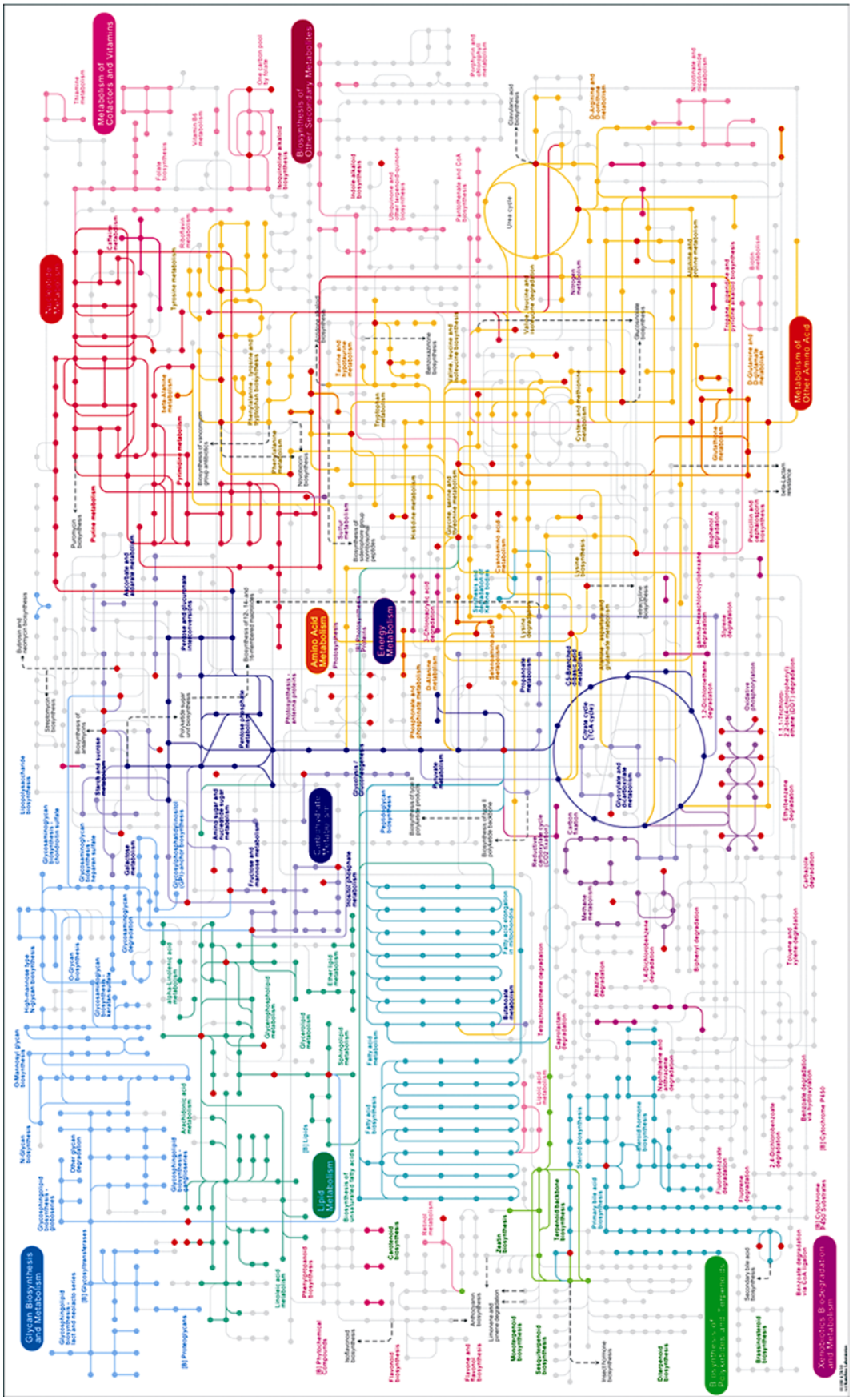

Figure 3 The marked KEGG metabolic pathway map by red points with all the transporter substrates in humans.

tween transporter systems and metabolic enzymes. In addition, cross-referenced links to KEGG Ligand, ChEBI and CAS databases are available.

3 Applications

A preliminary analysis of some of the data in TSdb has revealed interesting findings. For instance, 1082 of the transporter substrates are enzyme inhibitors; this may be a consequence of the central role of the substrates in the control of metabolic flux and metabolic regulation. In addition, we found a total of 761 gene-gene interaction pairs that are related to transporters in humans, thus linking compound metabolic networks and gene-gene interaction networks. Because all the substrates of our database are mapped to KEGG Ligand compound IDs, it is easy to map all the substrates in an organism to KEGG pathways and, via their shared substrates, to connect all the metabolic enzymes with transporter systems on a genome scale. All the transporter substrates in humans were mapped to KEGG pathways and all of them were connected to enzymes as shown in Figure 3.

4 Future perspectives

TSdb is the first high-quality resource of transporters and transporter substrates mapped to KEGG Ligand compound IDs. With summarized annotation and links that connect the transporters with metabolic enzymes via their shared compounds, the database will contribute significantly to the study of transporter systems. Our formatted database will also be useful for the comparison of transporter substrates in different organisms at the genome level. We are continuing to map and integrate substrate information from newly published literature into TSdb.

The comprehensive and formatted information on transporter substrates in TSdb will be useful for understanding the roles of transporters and their substrates in broader molecular networks. In addition to the on-going data curation from the literature and the integration of other transporter-related databases, we are also planning the construction of a metabolic network of transporter substrates in mammals which will emphasize their regulatory roles in nutrition and enzyme inhibition.

This work was supported by the National High Technology Research and Development Program of China (*Grant Nos. 2006AA02Z334, 2006AA02Z314, 2006AA02A312 and 2007AA02Z165*), *and the National Basic Research Program of China* (*Grant Nos. 2006CB910404 and 2007CB946904*). *We gratefully acknowledge the support of the K. C. Wong Education Foundation, Hong Kong.*

1 Nelson D L, Cox M M. Lehninger's principles of biochemistry. 3rd ed. New York: Worth Publishers, 2000
2 Ren Q H, Chen K X, Paulsen I T. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res, 2007, 35: D274–D279
3 Amann T, Hellerbrand C. GLUT1 as a therapeutic target in hepatocellular carcinoma. Expert Opin Ther Targets, 2009, 13: 1411–1427
4 Dettmer K, Aronov P A, Hammock B D. Mass spectrometry-based metabolomics. Mass Spectrom Rev, 2007, 26: 51–78
5 Delgado-Lista J, Perez-Martinez P, Perez-Jimenez F, *et al*. ABCA1 gene variants regulate postprandial lipid metabolism in healthy men. Arterioscler Thromb Vasc Biol, 2010, 30: 1051–1057
6 Kim I W, Booth-Genthe C, Ambudkar S V. Relationship between drugs and functional activity of various mammalian P-glycoproteins (ABCB1). Mini Rev Med Chem, 2008, 8: 193–200
7 Gradhand U, Kim R B. Pharmacogenomics of MRP transporters (ABCC1-5) and BCRP (ABCG2). Drug Metab Rev, 2008, 40: 317–354
8 Solinas M, Yasar S, Goldberg S R. Endocannabinoid system involvement in brain reward processes related to drug abuse. Pharmacol Res, 2007, 56: 393–405
9 Higgins C F. Multiple molecular mechanisms for multidrug resistance transporters. Nature, 2007, 446: 749–757
10 Fichant G, Basse M J, Quentin Y. ABCdb: an online resource for ABC transporter repertories from sequenced archaeal and bacterial genomes. FEMS Microbiol Lett, 2006, 256: 333–339
11 Ren Q, Kang K H, Paulsen I T. TransportDB: a relational database of cellular membrane transport systems. Nucleic Acids Res, 2004, 32: D284–D288
12 Saier M H, Tran C V Jr., Barabote R D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res, 2006, 34: D181–D186
13 Yan Q, Sadee W. Human membrane transporter database: a web-accessible relational database for drug transport studies and pharmacogenomics. AAPS PharmSci, 2000, 2: E20
14 UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res, 2010, 38: D142–D148
15 Jain E, Bairoch A, Duvaud S, *et al*. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics, 2009, 10: 136
16 Bairoch A, Apweiler R, Wu C H, *et al*. The Universal Protein Resource (UniProt). Nucleic Acids Res, 2005, 33: D154–D159
17 Hermjakob H, Fleischmann W, Apweiler R. Swissknife—'lazy parsing' of SWISS-PROT entries. Bioinformatics, 1999, 15: 771–772
18 Kanehisa M, Araki M, Goto S, *et al*. KEGG for linking genomes to life and the environment. Nucleic Acids Res, 2008, 36: D480–D484
19 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000, 28: 27–30
20 Lu Z, Cohen K B, Hunter L. GeneRIF quality assurance as summary revision. Pac Symp Biocomput, 2007, 269–280
21 Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res, 2006, 34: D354–D357
22 Chang A, Scheer M, Grote A, *et al*. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res, 2009, 37: D588–D592
23 Prasad T S K, Goel R, Kandasamy K, *et al.* Human Protein Reference Database—2009 update. Nucleic Acids Res, 2009, 37: D767–772
24 Breitkreutz B J, Stark C, Reguly T, *et al*. The BioGRID Interaction Database: 2008 update. Nucleic Acids Res, 2008, 36: D637–D640
25 Willis R C, Hogue C W. Searching, viewing, and visualizing data in the Biomolecular Interaction Network Database (BIND). Curr Protoc Bioinformatics online publication 1 January 2006; doi: 10.1002/0471250953.bi0809s12
26 Gutteridge A, Kanehisa M, Goto S. Regulation of metabolic networks by small molecule metabolites. BMC Bioinformatics, 2007, 8: 88