JOHNS HOPKINS
UNIVERSITY

Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs

410.734.81 and 410.734.82
Practical Introduction to Metagenomics

Topic: Metabolic reconstruction (military comic edition)

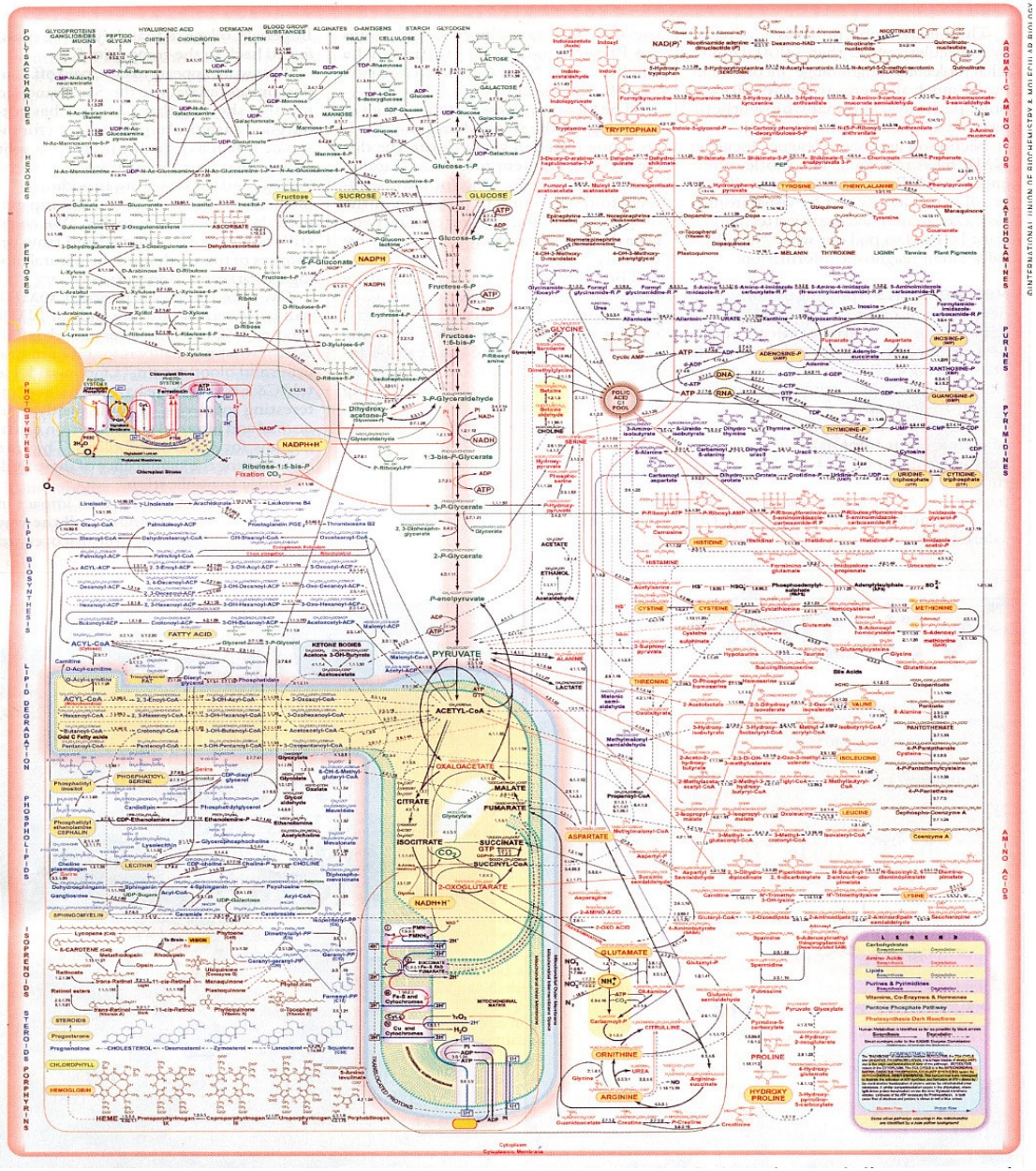
Instructor: Joshua Orvis

Introduction

What's there and what can it do?

Identification of organisms present in an environment and even comparing taxonomic profiles across environments is interesting, but others are more interested less in identifying individual members of a community than profiling what those members can **do**.

In this view, each organism can be thought of as a reservoir of a set of functional metabolic capability. A pool of organisms in an environment each contribute to that environment's overall metabolic ability.



Complete map of known human metabolic pathways

Enzyme commission

Just as we saw with ontologies in a previous lesson, if you wish to do systematic classification it's important to create a controlled reference set of curated data.

To further studies in metabolism, an Enzyme Commission was set up in 1955 to develop a nomenclature scheme, complete with short identifiers. Although completed before the computer age, these hierarchical identifiers are well-suited for easy parsing data modeling.

It's important to note that this is NOT a database of enzymes, but rather enzyme-catalyzed reactions. This means two different enzymes that catalyze the same reaction will have the same EC number.

Top-level EC #	Group	High-level reaction
1.x.x.x	Oxidoreductases	To catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another
2.x.x.x	Transferases	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group
3.x.x.x	Hydrolases	Formation of two products from a substrate by hydrolysis
4.x.x.x	Lyases	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved
5.x.x.x	Isomerases	Intramolecule rearrangement, i.e. isomerization changes within a single molecule
6.x.x.x	Ligases	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP

EC numbers (cont'd)

The hierarchical identifiers for EC are named using 4 digits separated by periods. Each level provides a more specific function. See the table below left for an example.

EC number	Reaction
4.x.x.x	Lyases
4.2.x.x	Carbon-oxygen lyases
4.2.1.x	Hydro-lyases
4.2.1.20	Tryptophan synthase

These levels provide a few benefits:

1. Enzymes can be assigned a level only as specific as the evidence suggests.
2. You can summarize functionality easily by clustering annotations at any given level.

Unfortunately, this is just the beginning for annotation. The most recent set was published in 1992 (though minor updates have been released yearly). Also, because it is only concerned with enzymatic activity, many other proteins cannot be assigned an EC number even if their function is known.

Also, the EC only provides the classification - there is no associated system included to actually query proteins and match them with an EC number. These are simply assigned by matching your query to any protein which already has one. This is less than ideal.

As you'll see, other new structured and organized datasets have risen to replace the EC classification, but it sticks around in part because of its early adoption into the evidence accepted by GenBank in their genome submissions.

Poses from our childhood

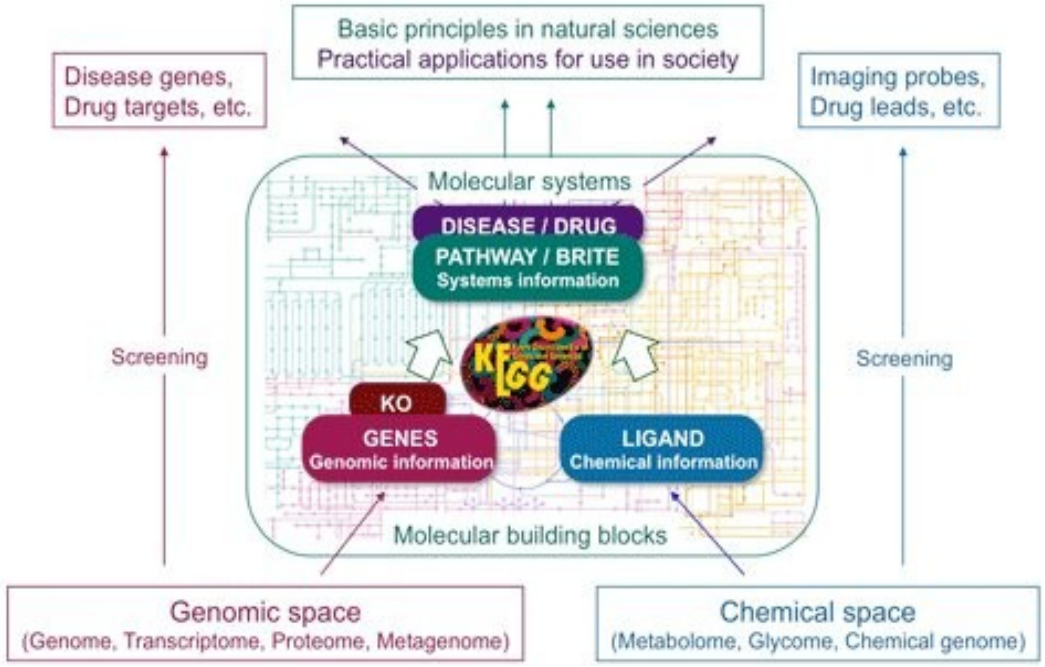
KEGG

In 1995 development began on the KEGG database, which initially used EC numbers to map proteins to metabolic pathways. In 2003 the KEGG Orthology (KO) became the central component of the system. Profoundly larger in scope than the EC effort, KEGG aims to be a complete representation of a biological system. From their site:

> KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information. It is a computer representation of the biological system, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information).

This is a very important resource to learn and use if you're interested in metabolism studies. KEGG is referenced in almost any publication related to the subject. It is also monolithic and can be very difficult to learn due to its history. Because the structure of it and resources have changed so much over its 17-year history, you should always use either the KEGG site itself or the most recent publications by the authors.

An overview of KEGG resources



KEGG (cont'd)

16 main KEGG databases

Category	Database	Content
Systems information	KEGG PATHWAY	KEGG pathway maps
	KEGG BRITE	BRITE functional hierarchies
	KEGG MODULE	KEGG modules of functional units
	KEGG DISEASE	Human diseases
	KEGG DRUG	Drugs
	KEGG ENVIRON	Crude drugs and health-related substances
Genomic information	KEGG ORTHOLOGY	KEGG Orthology (KO) groups
	KEGG GENOME	KEGG organisms with complete genomes
	KEGG GENES	Gene catalogs in complete genomes
	KEGG SSDB	Sequence similarity database for GENES
Chemical information	KEGG COMPOUND	Metabolites and other small molecules
	KEGG GLYCAN	Glycans
	KEGG REACTION	Biochemical reactions
	KEGG RPAIR	Reactant pair chemical transformations
	KEGG RCLASS	Reaction class defined by RPAIR
	KEGG ENZYME	Enzyme nomenclature

These databases are all interconnected via sets of well-organized identifiers. The website supplies visual navigation tools as well as an automated annotation server (KAAS).

The latest review published by the authors, which has descriptions of almost all of these, can be found here:

http://nar.oxfordjournals.org/content/40/D1/D109.full.pdf+html

A plea from KEGG [link]

KEGG has the common problem of being a heavily-used and under-funded. Also, funding agencies are always excited about the development of new tools. Keeping something going that has existed for almost two decades isn't very sexy. Due to many factors such as these, the database is no longer freely available to academics. As of last summer, access to the FTP site requires a license purchase.


Day 41
The trucks still do not realize that I am a tank.

PRIAM

I won't go into huge depth here about PRIAM, but it's important to at least know about it because many tools will either reference or us it.

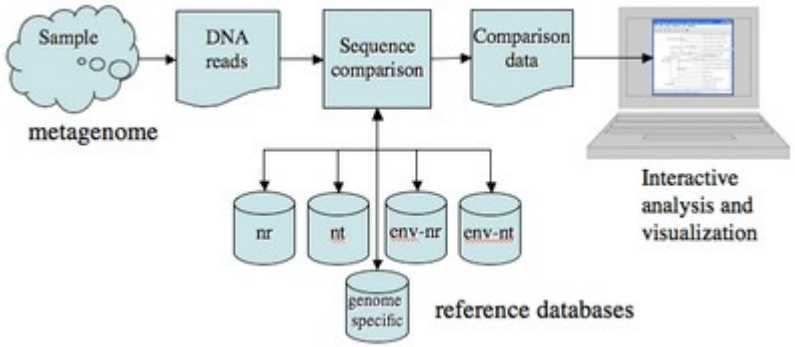
Initially published in 2003, this is a collection of enzyme-specific profiles from the ENZYME database. Updated approximately yearly, it can be used to predict metabolic pathways from complete genomes.

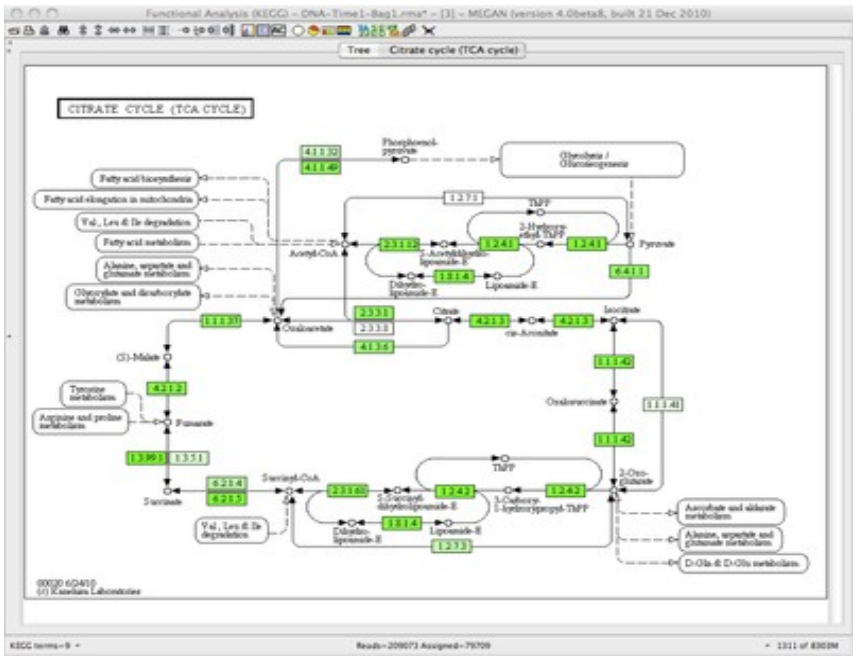
The profiles are stored as position-specific scoring matrices (PSSMs) and are painfully slow to search.

The original paper is perhaps the best source for additional information.

MEGAN 4

The MEtaGenome ANalyzer has grown from a simple taxonomy visualization tool in 2007 to, in the latest version 4, one that provides both taxonomic and functional analysis in an accessible user-interface.





The prerequisite step to running MEGAN is a BLAST search against a reference set like 'nr' or 'env-nr'. This is used as input to the system, which uses the BLAST results to taxonomically classify the reads.

For functional analysis, you can direct MEGAN to map each read to a SEED functional role or KEGG orthology (KO) accession number, the results of which are pictured on the lower left.

With different datasets, MEGAN also provides a comparison view that is based on a tree in which each node shows the number of reads assigned to it for each of the datasets. This can be done either as a pie chart, a bar chart or as a heat map.

MEGAN is available free to academics here.

Like Sirs

HUMAnN paper

Metabolic reconstruction for metagenomic data and its application to the human microbiome

Abubucker S, Huttenhower C, et al.
PLoS Computational Biology,
doi:10.1371/journal.pcbi.1002358, 2012.

"Here, we describe an alternative approach to infer the functional and metabolic potential of a microbial community metagenome. We determined the gene families and pathways present or absent within a community, as well as their relative abundances, directly from short sequence reads. We validated this methodology using a collection of synthetic metagenomes, recovering the presence and abundance both of large pathways and of small functional modules with high accuracy."

This work was used to classify all of the HMP samples. Their methodology was implemented in a software package called the HMP Unified Metabolic Analysis Network (HUMAnN), freely available online.

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome

Sahar Abubucker[1], Nicola Segata[2], Johannes Goll[3], Alyxandria M. Schubert[4], Jacques Izard[5,6], Brandi L. Cantarel[7], Beltran Rodriguez-Mueller[6], Jeremy Zucker[8], Mathangi Thiagarajan[3], Bernard Henrissat[9], Owen White[7], Scott T. Kelley[10], Barbara Methé[3], Patrick D. Schloss[4], Dirk Gevers[8], Makedonka Mitreva[1], Curtis Huttenhower[2,8]*

1 The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, United States of America, **2** Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America, **3** J. Craig Venter Institute, Rockville, Maryland, United States of America, **4** Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, United States of America, **5** Department of Molecular Genetics, Forsyth Institute, Cambridge, Massachusetts, United States of America, **6** Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston, Massachusetts, United States of America, **7** Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, United States of America, **8** The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, **9** Architecture et Fonction des Macromolécules Biologiques, UMR 6098 CNRS, Université de la Méditerranée, Marseille, France, **10** Biology Department, San Diego State University, San Diego, California, United States of America

Abstract

Microbial communities carry out the majority of the biochemical activity on the planet, and they play integral roles in processes including metabolism and immune homeostasis in the human microbiome. Shotgun sequencing of such communities' metagenomes provides information complementary to organismal abundances from taxonomic markers, but the resulting data typically comprise short reads from hundreds of different organisms and are at best challenging to assemble comparably to single-organism genomes. Here, we describe an alternative approach to infer the functional and metabolic potential of a microbial community metagenome. We determined the gene families and pathways present or absent within a community, as well as their relative abundances, directly from short sequence reads. We validated this methodology using a collection of synthetic metagenomes, recovering the presence and abundance both of large pathways and of small functional modules with high accuracy. We subsequently applied this method, HUMAnN, to the microbial communities of 649 metagenomes drawn from seven primary body sites on 102 individuals as part of the Human Microbiome Project (HMP). This provided a means to compare functional diversity and organismal ecology in the human microbiome, and we determined a core of 24 ubiquitously present modules. Core pathways were often implemented by different enzyme families within different body sites, and 168 functional modules and 196 metabolic pathways varied in metagenomic abundance specifically to one or more niches within the microbiome. These included glycosaminoglycan degradation in the gut, as well as phosphate and amino acid transport linked to host phenotype (vaginal pH) in the posterior fornix. An implementation of our methodology is available at http://huttenhower.sph.harvard.edu/humann. This provides a means to accurately and efficiently characterize microbial metabolic pathways and functional modules directly from high-throughput sequencing reads, enabling the determination of community roles in the HMP cohort and in future metagenomic studies.

Citation: Abubucker S, Segata N, Goll J, Schubert AM, Izard J, et al. (2012) Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. PLoS Comput Biol 8(6): e1002358. doi:10.1371/journal.pcbi.1002358

Editor: Jonathan A. Eisen, University of California Davis, United States of America

Received August 6, 2011; **Accepted** December 7, 2011; **Published** June 13, 2012

Funding: This work was supported in part by grants NIH U54HG004968 (George Weinstock), University of Michigan Rackham Graduate Student Research Grant (AMS), NIH CA139193 and DE017106 (JI), NIH 5R01HG005975 (PDS), NIH U54HG004969 (DG), and NIH 1R01HG005969 (CH). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: chuttenh@hsph.harvard.edu

Introduction

Human-associated microbial communities interact directly with their hosts by means of metabolic products and immune modulation, and environmental communities are further responsible for a wide range of biochemical activities [1]. Metagenomic sequencing provides a culture-independent means of studying these diverse microbiota within different ecological niches, including sites in the human body that differ strikingly in microbial composition and subsequent impacts on health [2,3,4]. The gut microbiota in particular have been shown to play an important role in host metabolism [5,6] and immune response [4], and mechanisms of commensal microbial contribution to disease have been established e.g. in the vaginal [7] and skin [8] communities as well. These studies have demonstrated the importance of assaying microbial pathways, metabolism, and individual gene products by means of metagenomic sequencing to determine their roles in community-wide interactions and phenotypes. A functional interpretation of metagenomic sequences is thus key to connecting the metabolic and functional potential of a microbial community with its organismal population structure and with its influence on the surrounding environment or human host.

Reconstruction paper

Reconstruction of biochemical networks in microorganisms

Feist A, Palsson B, et al.
Nat Rev Microbiology. Vol 7, Feb 2009.

"This Review describes the detailed work flows that form the basis of the reconstruction process and provide key procedural information needed for the increasing number of researchers who are performing organism-specific reconstructions."

When automated passes aren't enough – this paper exhaustively reviews the methods involved in creating an organism-specific metabolic reconstruction on the order of detail expected for model organism databases.

"Although the automated extraction of metabolic reactions from databases provides an initial set of candidate biochemical reactions encoded on a genome, it cannot establish certain organism-specific features, such as substrate or cofactor specificity and subcellular localization. Such information requires domain-specific knowledge of the organism."

The importance of lab work remains.

SYSTEMS BIOLOGY

Reconstruction of biochemical networks in microorganisms

Adam M. Feist, Markus J. Herrgård*‡, Ines Thiele*, Jennie L. Reed§ and Bernhard Ø. Palsson*||*

Abstract | Systems analysis of metabolic and growth functions in microbial organisms is rapidly developing and maturing. Such studies are enabled by reconstruction, at the genomic scale, of the biochemical reaction networks that underlie cellular processes. The network reconstruction process is organism specific and is based on an annotated genome sequence, high-throughput network-wide data sets and bibliomic data on the detailed properties of individual network components. Here we describe the process that is currently used to achieve comprehensive network reconstructions and discuss how these reconstructions are curated and validated. This Review should aid the growing number of researchers who are carrying out reconstructions for particular target organisms.

Reconstructed networks of biochemical reactions are at the core of systems analyses of cellular processes. Such networks form a common denominator for both experimental data analysis and computational studies in systems biology. The conceptual basis for the reconstruction process has been outlined[1], and computational methods and tools used to characterize them have been reviewed[2,3]. Furthermore, the number of available, well-curated organism-specific network reconstructions is increasing (Supplementary information S1 (table)) and the spectrum of their uses is broadening[4].

This Review describes the detailed work flows that form the basis of the reconstruction process and provide key procedural information needed for the increasing number of researchers who are performing organism-specific reconstructions. We describe the procedures in which various experimental data types are integrated to reconstruct biochemical networks, the current status of network reconstructions and how network reconstructions can be used in a prospective manner to discover new interactions and pathways. We will focus on the networks that underlie three key cellular processes: metabolism, transcription and translation, and transcriptional regulation. The reconstruction process for genome-scale metabolic networks is well developed, whereas the process for the reconstruction of transcriptional regulation and for transcriptional and translational processes at the genome-scale is only now developing. In addition, we will briefly discuss the impact of network content on modelling and integration of these types of networks, as well as the prospects of reconstructing other types of networks, such as signalling and small RNA (sRNA) pathways.

*Department of Bioengineering, University of California, San Diego, La Jolla, California 92093, USA.
‡Synthetic Genomics, 11149 N. Torrey Pines Road, La Jolla, California 92037, USA.
§Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.
||Centre for Systems Biology, University of Iceland, Vatnsmyrarvegi 16, IS-101 Reykjavik, Iceland.
Correspondence to B.Ø.P.
e-mail: palsson@ucsd.edu
doi:10.1038/nrmicro1949
Published online
31 December 2008

Metabolic networks

Before annotated genomic sequences were available, primary literature and biochemical characterization of enzymes provided the main sources of information for reconstructing metabolic networks in a select number of organisms. Accordingly, some of the earliest metabolic reconstructions that were subsequently used in modelling applications were for *Clostridium acetobutylicum*[5], *Bacillus subtilis*[6] and *Escherichia coli*[7–10].

Today, with the ability to sequence and annotate whole genomes, we can generate metabolic network reconstructions at a genome scale, even for organisms for which little direct biochemical information is available in the published literature. To implement the metabolic reconstruction process, we need to answer the following questions for each of the enzymes in a metabolic network: what substrates and products does an enzyme act on; what are the stoichiometric coefficients for each metabolite that participates in the reaction (or reactions) catalysed by an enzyme; are the outlined reactions reversible; and where does the reaction occur in the cell (for example, the cytoplasm or periplasm)? These data come from a range of sources. Establishing a set of the chemical reactions that constitute a reaction network culminates in a database of chemical equations. Each reaction also

Further reading

KEGG primer: An introduction to pathway analysis using KEGG
http://pid.nci.nih.gov/PID/2007/071009/full/pid.2007.2.shtml

Enzyme-specific profiles for genome annotation: PRIAM
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC275543/

MEGAN 4: MEtaGenome ANalyzer
http://ab.inf.uni-tuebingen.de/software/megan/