582605 Metabolic modeling (4cr)

- Lecturer: prof. Juho Rousu
- Course assistant: Markus Heinonen
- Lectures: Tuesdays and Fridays, 14.15-16, B119
- Exercises: 16.03.-24.04. Tuesdays 16.15-18, C221
- Course topics:
  - Reconstruction of metabolic networks (MN)
  - Structural analysis of MNs
  - Stoichiometric analysis of MNs
  - Metabolic flux analysis
  - Regulation of metabolism

Prerequisites

We will assume that you know at least something about the following

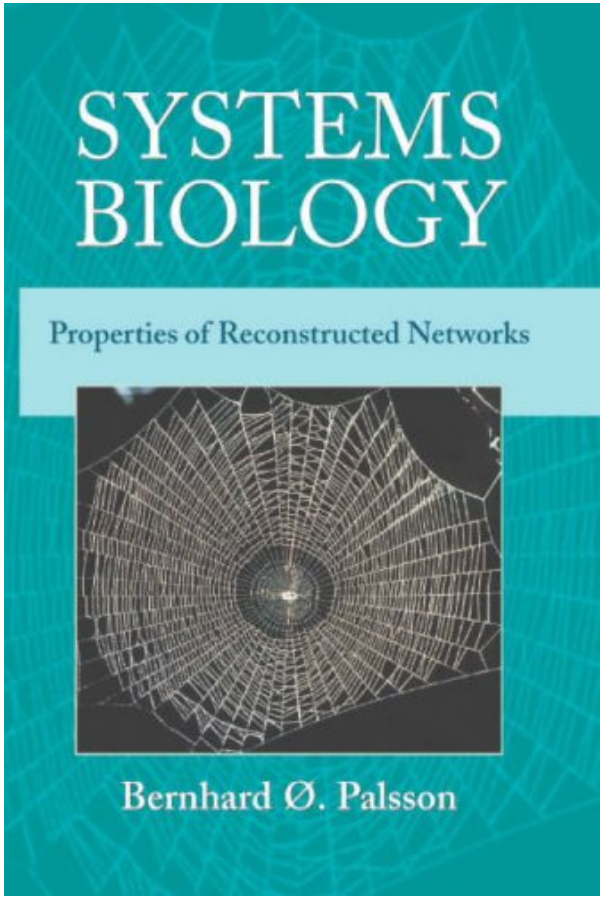
- Introduction to bioinformatics: protein, cell
- Data structures: graphs and networks
- Elementary probability calculus
- Basic linear algebra / Matrix computation

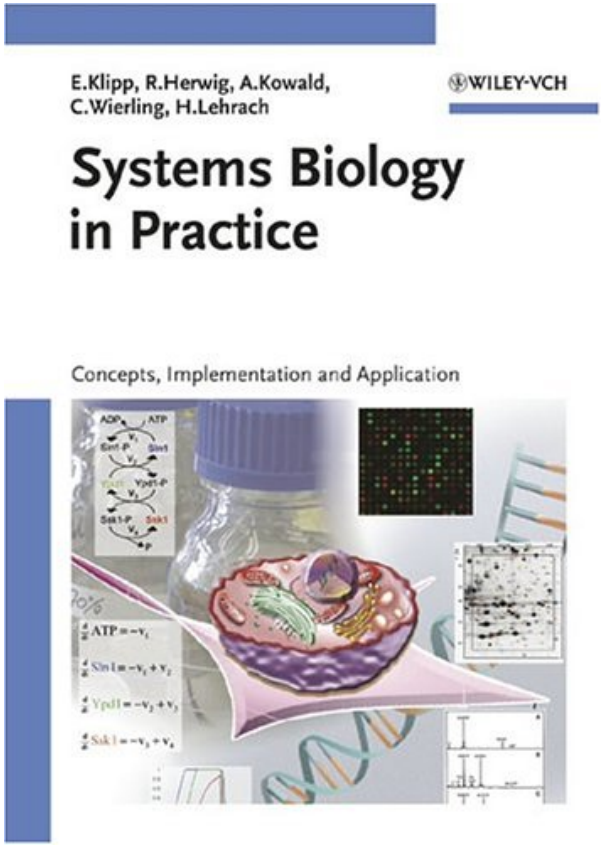
Passing the course

- Course exam (Wednesday 29.4.2009 9am-12pm, in A111): maximum 40 points
  - Examined contents: lecture slides and exercises
- Exercises: maximum 20 points, mix of different types:
  - Reading a paper, and presenting a summary
  - Assignments to be completed by pen and paper, mostly dealing with small metabolic systems
  - Computer assignments, calling for (a little a bit) of MATLAB or R programming
- Grading:
  - 30 points required for passing the course (grade 1/5),
  - 50 points gives maximum grade 5/5.

Additional reading

- For more broad coverage of the course topics, you may look at the following books
- The books are not required for passing the course





What is Metabolism?

Definitions (from the web):

- "Metabolism (from 'metabolismos' the Greek word for "change", or "overthrow" Etymonline), is the biochemical modification of chemical compounds in living organisms and cells...."
- "Enzymatic transformation of organic molecules. Synthesis corresponds to anabolism, and degradation to catabolism"
- "The sum of the processes by which a particular substance is handled (as by assimilation and incorporation, or by detoxification and excretion) in the living body."

What is not covered by metabolism?

A lot:

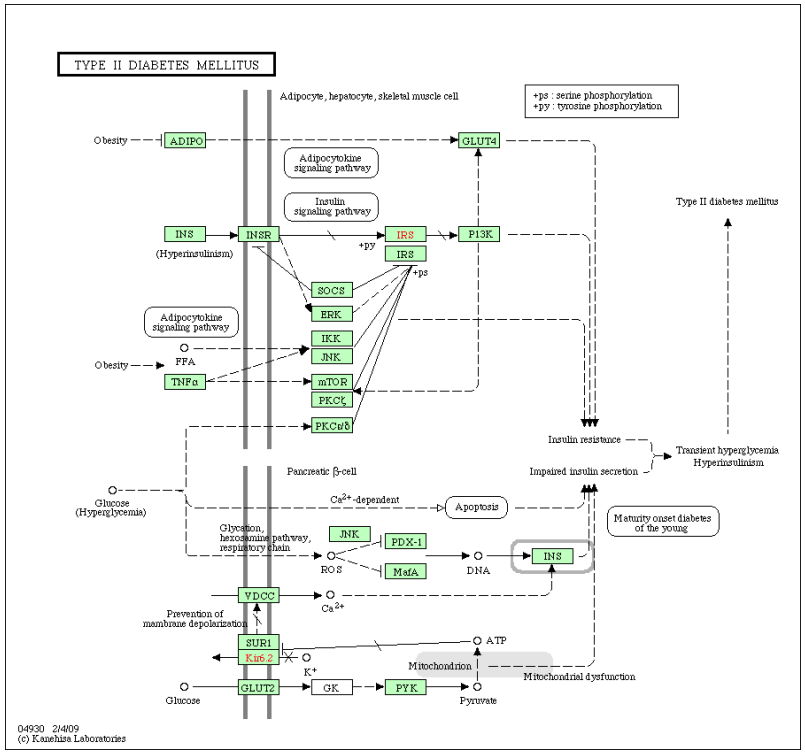
- Building of proteins: transcription, translation and protein folding: ready-made proteins are our building blocks
- Gene expression and protein expression (proteomics): we typically analyze situations where expression can be assumed to be constant
- Signaling between cells
- ...

Why metabolic modelling?

Why metabolic modelling?

Applications in medicine:

- Many diseases are linked to malfunction in metabolism (e.g. diabetes)
- These malfunctions are often properties of metabolic pathways, and cannot be pinned down to a single genetic defect in a single gene.
- Instead, a group of enzymes are working somehow incorrectly, putting the cellular system off-balance
- Restoring the balance (e.g. via a drug) might require modelling the whole pathway

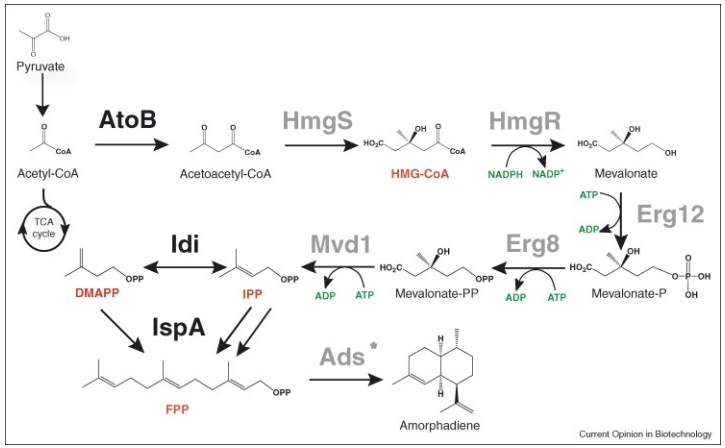


Pathways in type II diabetes, source: http://www.genome.jp/kegg/

Why metabolic modelling?

Applications in bioengineering:

- Suppose we want to engineer a microbe to produce biofuel (e.g. ethanol) from organic waste
- A significant problem is the yield: the microbes produce all kinds of products from the substrate, but the yield of the desired product might be too low for commerical use.
- Optimizing the yield typically requires modulating the activity of a set of enzymes (e.g. blocking some pathways, emphasizing others)



Aindrila Mukhopadhyay, Alyssa M Redding, Becky J Rutherford, Jay D Keasling. Current Opinion in Biotechnology 19, 3 (2008)

Outline of the course

Aim of the course: to learn techniques that are used to analyze metabolism

Particular techniques include

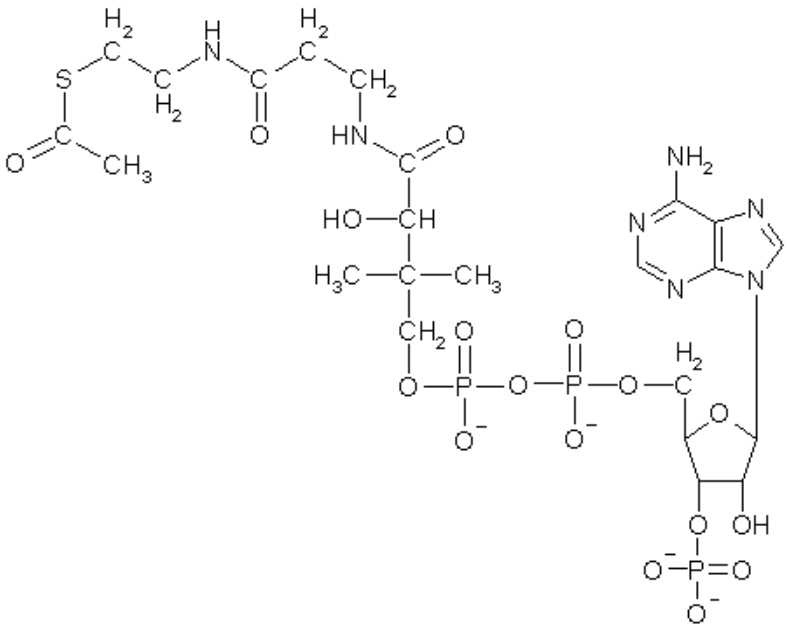
- Metabolic reconstruction: given a newly sequenced organism, how to estimate how the metabolism of the organism looks.
- Analysis of metabolic networks: what can we say about the organism just by looking at the metabolic production routes it has
- Flux estimation: given a metabolic network, estimate the activity of the different metabolic pathways
- Metabolic-level regulation: how does the cell react to sudden changes, when regulation of expression is too slow

Metabolism and metabolic networks

- Metabolism is the means by which cells acquire energy and building blocks for cellular material
- Metabolism is organized into sequences of biochemical reactions, metabolic pathways
- Pathways are interconnected in many ways, thus their total is a metabolic network, concisting of reactions and compounds (the metabolites).

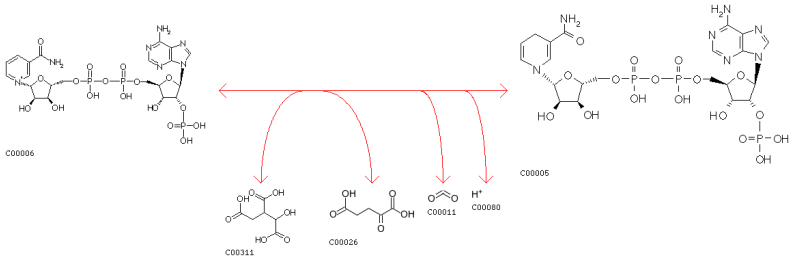
Metabolites

- Metabolites are small (typically < 50 atoms) organic compounds
- Acetyl-coenzyme-A (pictured) is among the largest metabolites in metabolism
- There are large number of metabolites, e.g. human metabolic network reconstruction by Duarte et al. (2007) contains 2766 metabolites



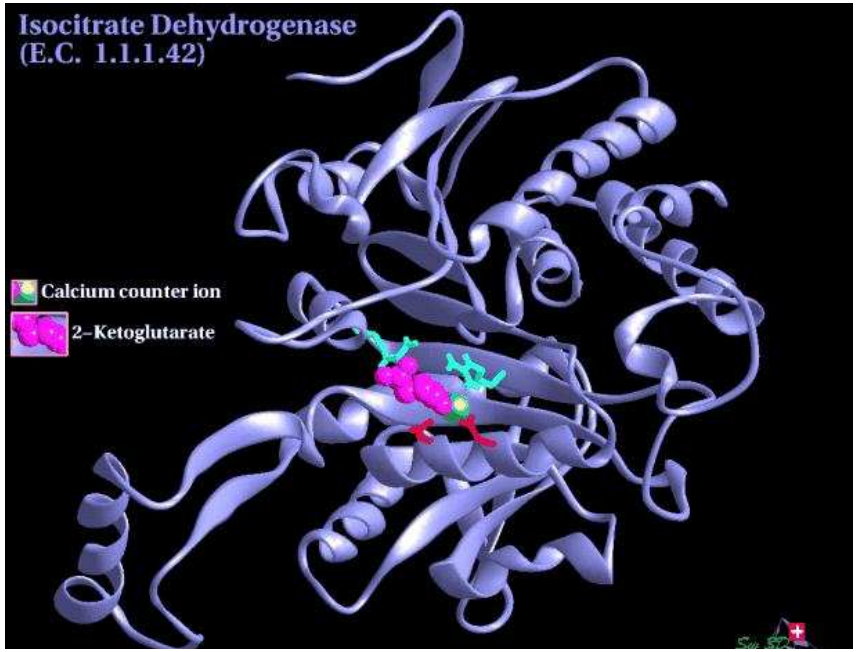
Reactions and enzymes

- The basic building block of metabolic networks is a (bio)chemical reaction.



- Most reactions that occur within a living cell are catalyzed by enzymes, a class of proteins.

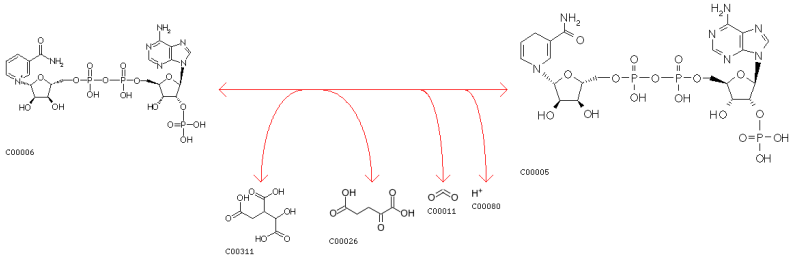
- Pictured is isocitrate dehydrogenase, an enzyme in the TCA cycle, together with the catalyzed reaction

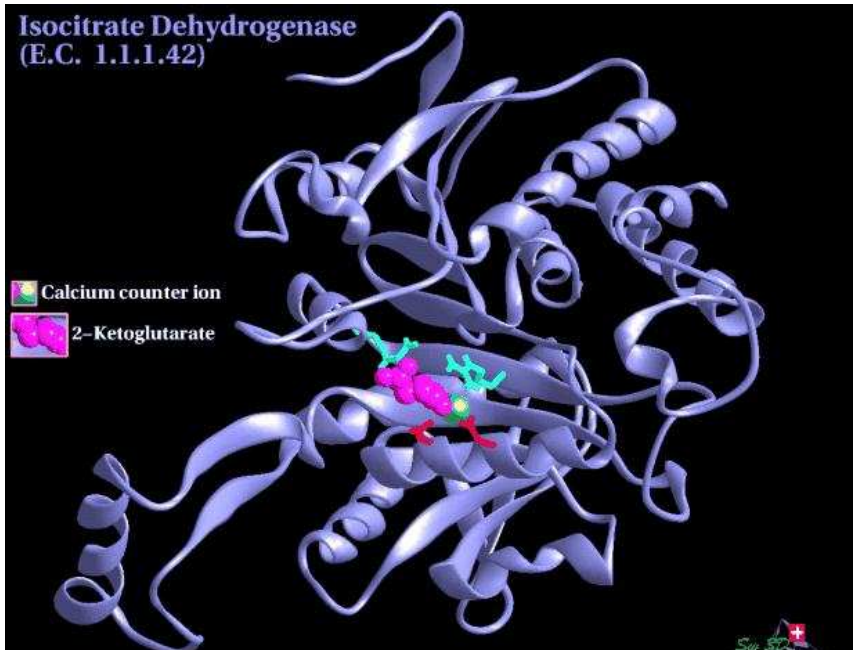


Picture from SWISS-3D Database, http://www.expasy.ch/sw3d/

Reactions and enzymes

- Enzymes are highly specific, a single enzyme can catalyze only one (or at most a couple) kind of a reaction.
- This enables the cell to control the production of certain metabolites without altering everything else at the same time.
- For example, isocitrate dehydrogenase is not known to catalyze any other biochemical reaction than the one pictured

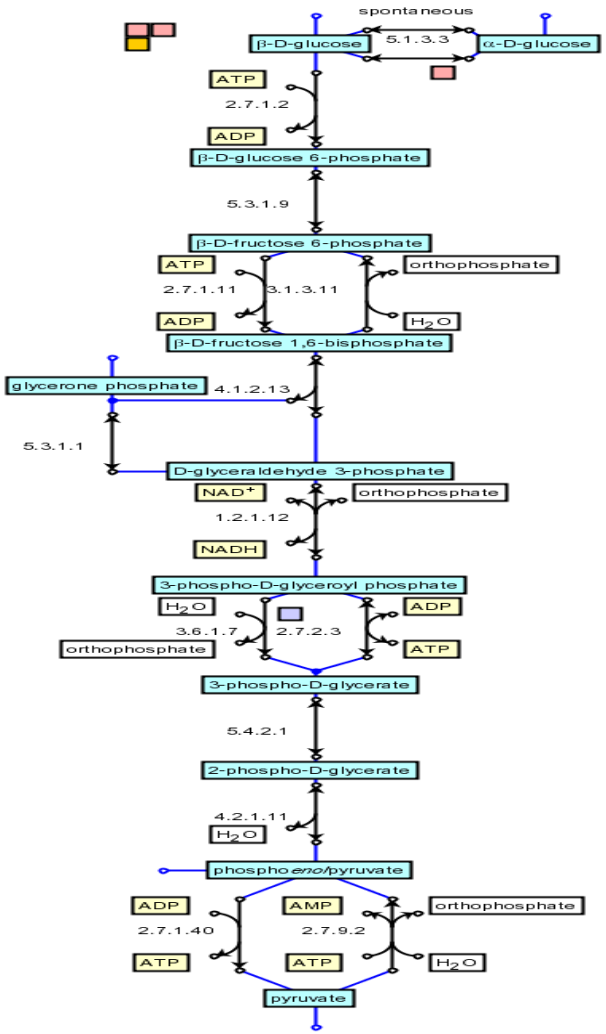




Picture from SWISS-3D Database, http://www.expasy.ch/sw3d/

Metabolic networks

- The individual enzymatic reactions are organized into pathways, sequences of reactions.
- The pathways are interconnected in many ways, which makes the metabolism a directed network.
- The network contains both cycles and biconnected components, i.e. alternative routes from one compound to another



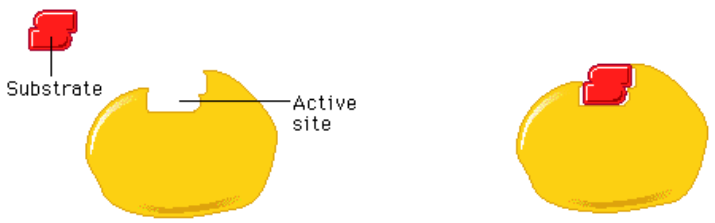
Picture:E.Coli glycolysis, EMP database, www.empproject.com/

Types of reactions

- *Fueling reactions* produce the precursor molecules needed for biosynthesis. In addition they generate energy, in the form of ATP, which is used by biosynthesis, polymerization and assembly reactions.
- *Biosynthetic reactions* produce building blocks used by the polymerization reactions. Biosynthetic reactions are organized into biosynthetic pathways, reation sequences of one to a dozen reactions. All biosynthetic pathways begin with one of 12 precursor molecules.
- *Polymerization reactions* link molecules into long polymeric chains.
- *Assembly reactions* carry out modifications of macromolecules, their transport to prespecified locations in the cell and their association to form cellular structure such as cell wall, membranes, nucleus, etc.

How does an enzyme work?

An enzyme works by binding the substrate molecules into the so called active site. In the active site, the substrates end up in such a mutual geometric conformation that the reaction occurs effectively.



The occurence of the reaction causes the enzyme to change its conformation, which releases the products. After that, the enzyme is ready to bind another set of substrates. The enzyme itself stays unchanged in the reaction.

Enzyme activity

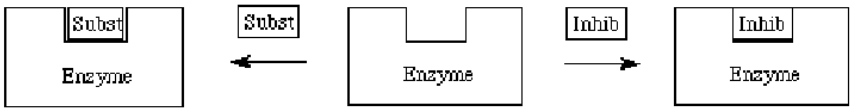
The rate of certain enzyme-catalyzed reaction depends on the concentration (amount) of the enzyme and the specific activity of the enzyme (how fast a single enzyme molecule works).
The specific activity of the enzyme depends on

- pH and temperature
- positively on the concentration of the substrates
- negatively on the concentration of the end-product of the pathway (inhibition).

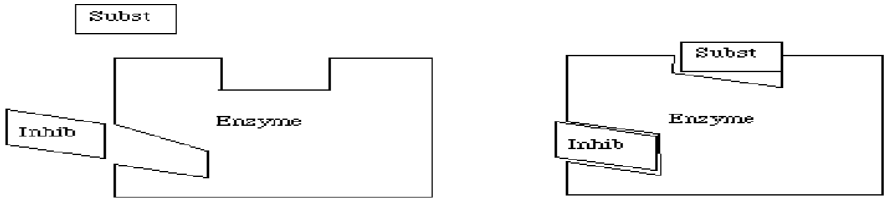
Note that transcription level gene regulation **directly** affects only the concentration of the enzyme.

Inhibition of Enzymes & Metabolic-level regulation

- The activity of enzymes is regulated in the metabolic level by inhibition: certain metabolites bind to the enzyme hampering its ability of catalysing reactions.
- In competitive inhibition, the inhibitor allocates the active site of the enzyme, thus stopping the substrate from entering the active site.



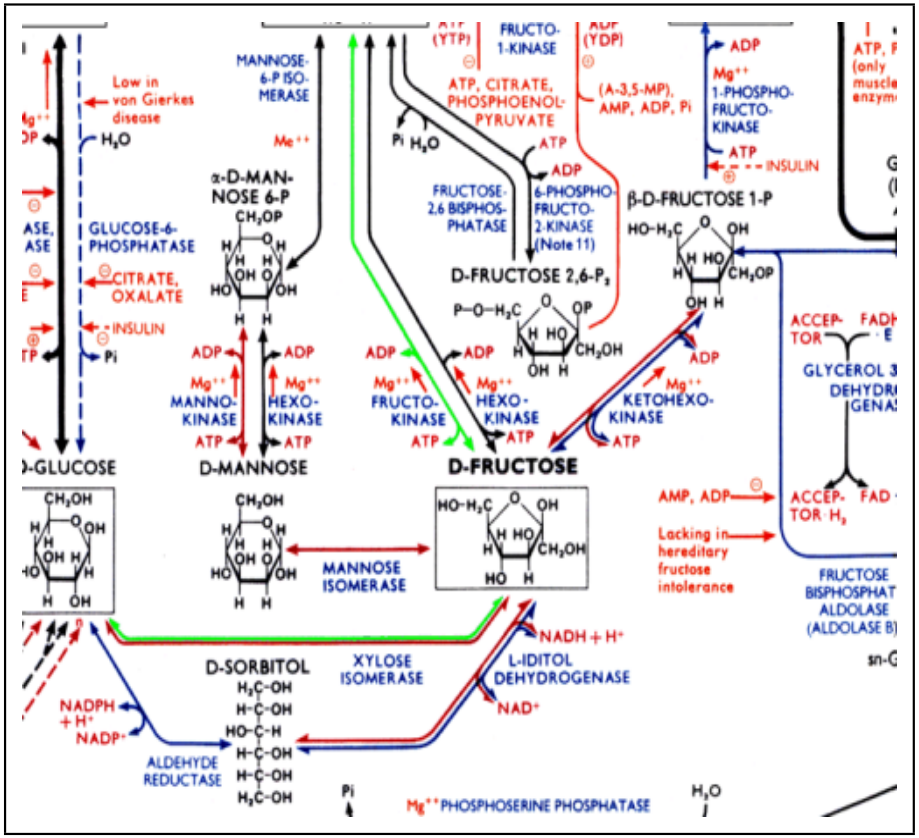
- In non-competitive inhibition, the inhibitor molecule binds to the enzyme outside the active site, causing the active site to change conformation and making the catalysis less efficient.



Metabolic reconstruction problem

From the sequenced genome, we want to infer the encoded metabolic network.

```
atagtgttgc attcctctct gccttcccat caccacaaaa
agtgtaataa atgctggtat gtccagctga agccagttcc
cttgctcgtg gccagctggg gccatacaca gccctgggga
cttgtgtctg agggtggtga cagctgtttt ctgcctcagg
ttggaggaac ttcctacaat gatgcagcac ttctcacagt
tttgttggag acaaggtaat gggggcatgt gatgaggaca
ctatgttaca gagattccag cccacacatt cttggccttc
ttcctcgcct atgatgtcct tgacctccac cgtatatttg
tttccaaatc tgaaggactt catctcccgc tttgaggtga
tttgatgccc cttgttccgt tacctccttt cagatgcttt
aagaataact tgcatttatt gagtgctggc ttcatgccag
tacctatcgt gtggaatttg aaatttccaa cattcctaca
ccagtggagg ctgtgctggg ctccctgtga gcatctggat
ctatgggtgg cagtcagggc tctccctttt gtgacaaaag
aaagaagcct caggcctcat ccagcctgga tttcacagcc
cagggcactt tggaagaggc agagaacttt aggagcatgg
atgcagctgg caatagtagg actgacacac ggtggcattg
acgtcgagta cgaaacccac aggcagtatt catagctact
cccagaagct ttgcacgatc agacccccac gtggggaatc
```



Data sources for Metabolic Reconstruction

The principal kinds of data for reconstruction (roughly in the order of reliability) are:

- Biochemistry: an enzyme has been isolated from an organism, and its function has been demonstrated (experimentally in test tube, or uncovering its 3D structure and simulating its behaviour in a computer).
- Genomics. Functional assigment to open reading frames (ORFs) based on DNA sequence homology. These annotations are often subject to revision and updates.

Data sources for Metabolic Reconstruction

- Physiology and indirect information. Physiological ability of the cell (e.g. capability to produce certain metabolite) may lead us to "fill in the pathway" so that the resulting network has this ability
- Modeling and simulation studies. The network needs to be able to simulate cell behaviour in silico (e.g. it needs to be able to produce all necessary components of biomass)

Resources in the web

There are numerous online resources that can be used to aid metabolic reconstruction. Rougly, they can be divided into the following categories.

- Databases with annotated genomes and annotation software
- Enzyme databases
- Pathway databases
- Automatic reconstruction tools

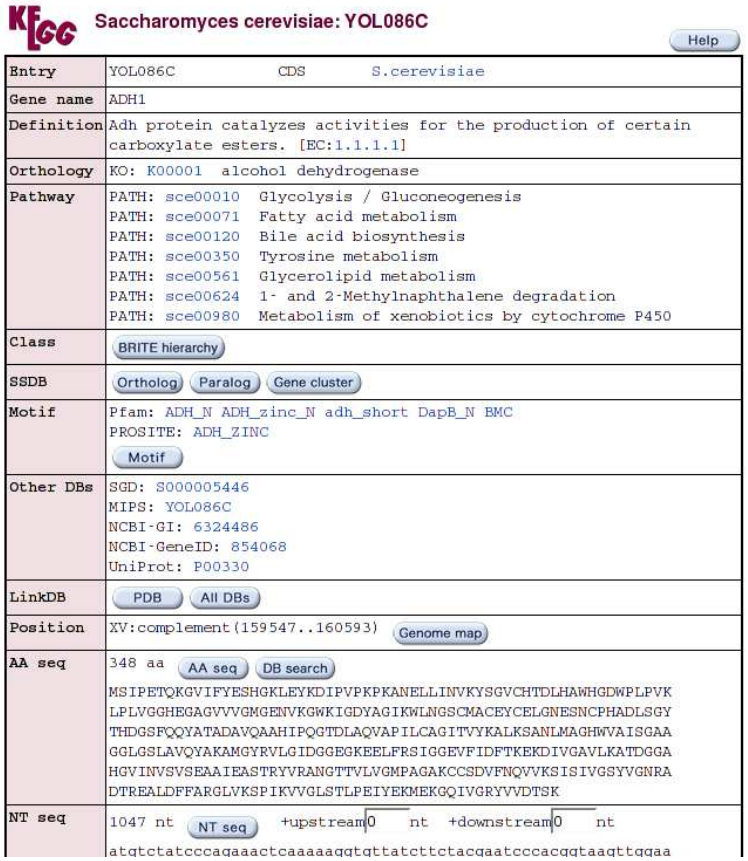
Most services in the web provide some mixture of these tools

KEGG - Kyoto Encyclopedia of Genes and Genomes (http://kegg.com)

- Knowledge base aiming to integrate genetic and higher-level information
- Project initiated in 1995 under the Human Genome Project.
- Genetic information contained in GENES database
- Higher-order functional information in PATHWAY database
- LIGAND databse contains information about chemical compounds, enzyme molecules and enzymatic reactions.
- Downloadable for academic users via ftp://ftp.genome.ad.jp/pub/kegg/.

GENES database

- Data from ca. 1000 genomes, majority completely sequenced
- $>$ 4,000,000 entries
- For each gene
  - Identification
  - Classification according to KEGG/PATHWAYS
  - Known sequence motifs
  - Chromosomal position
  - Amino acid and nucleotide sequences
  - Links to other databases (Genbank, SWISS-PROT)



KEGG Saccharomyces cerevisiae: YOL086C

Help

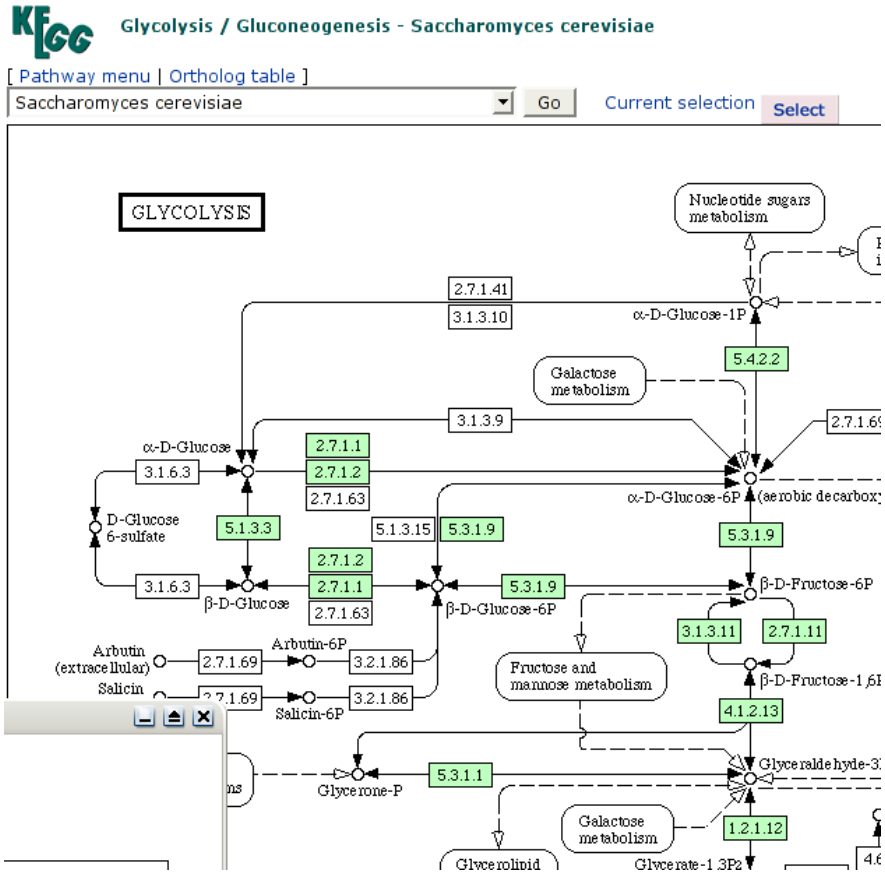
Entry	YOL086C CDS S.cerevisiae
Gene name	ADH1
Definition	Adh protein catalyzes activities for the production of certain carboxylate esters. [EC:1.1.1.1]
Orthology	KO: K00001 alcohol dehydrogenase
Pathway	PATH: sce00010 Glycolysis / Gluconeogenesis PATH: sce00071 Fatty acid metabolism PATH: sce00120 Bile acid biosynthesis PATH: sce00350 Tyrosine metabolism PATH: sce00561 Glycerolipid metabolism PATH: sce00624 1- and 2-Methylnaphthalene degradation PATH: sce00980 Metabolism of xenobiotics by cytochrome P450
Class	BRITE hierarchy
SSDB	Ortholog Paralog Gene cluster
Motif	Pfam: ADH_N ADH_zinc_N adh_short DapB_N BMC PROSITE: ADH_ZINC Motif
Other DBs	SGD: S000005446 MIPS: YOL086C NCBI-GI: 6324486 NCBI-GeneID: 854068 UniProt: P00330
LinkDB	PDB All DBs
Position	XV:complement(159547..160593) Genome map
AA seq	348 aa AA seq DB search MSIPETQKGVIFYESHGKLEYKDIPVPKPKANELLINVKYSGVCHTDLHAWHGDWPLPVK LPLVGGHEGAGVVVGMGENVKGWKIGDYAGIKWLNGSCMACEYCELGNESNCPHADLSGY THDGSFQQYATADAVQAAHIPQGTDLAQVAPILCAGITVYKALKSANLMAGHWVAISGAA GGLGSLAVQYAKAMGYRVLGIDGGEGKEELFRSIGGEVFIDFTKEKDIVGAVLKATDGGA HGVINVSVSEAAIEASTRYVRANGTTVLVGMPAGAKCCSDVFNQVVKSISIVGSYVGNRA DTREALDFFARGLVKSPIKVVGLSTLPEIYEKMEKGQIVGRYVVDTSK
NT seq	1047 nt NT seq +upstream 0 nt +downstream 0 nt atgtctatcccagaaactcaaaaagtgttatcttctacgaatcccacggtaagttggaa

KEGG LIGAND database

- `http://www.genome.ad.jp/dbget/ligand.html`
- A database of enzymatic reactions
- $\approx$ 5000 enzymes, 15000 compounds and 8000 reactions
- Supports similarity searches between compounds, and reaction prediction between compunds
- Pathway computation capability, i.e. queries returning all possible pathways between two compounds.

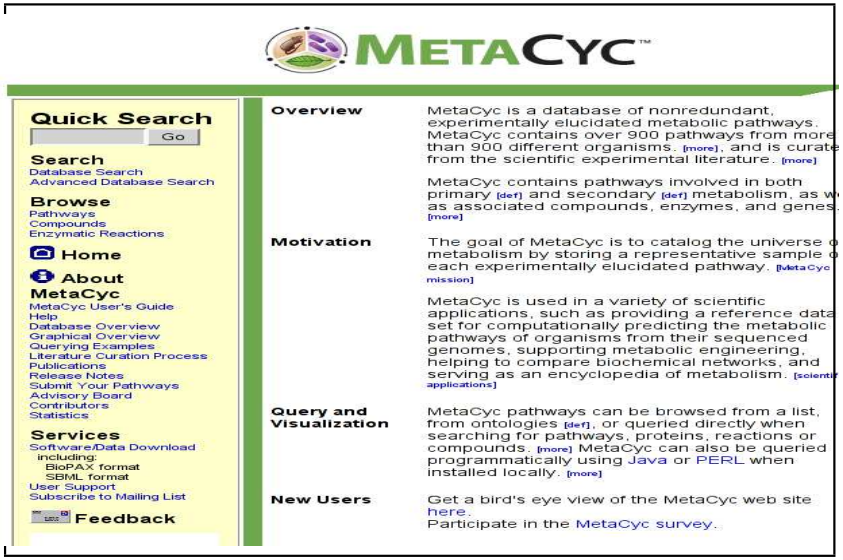
KEGG PATHWAY database

- PATHWAY database contains maps of metabolic pathways of many organimsm.
- The enzymes and compounds are clickable in the map and lead to the LIGAND and GENES database entries.
- Kegg PATHWAY maps are frequently used by biologists in their presentations



BioCyc (http://www.biocyc.org/)

- BioCyc is a collection of over 400 pathway/genome databases, mostly containing whole genome databases dedicated to certain organisms.
- One organism specific database, EcoCyc, is a highly detailed bioinformatics database on the genome and metabolic reconstruction of Escherichia Coli
- MetaCyc, an encyclopedia of metabolic pathways, contains information on metabolic reactions derived from over 1500 different organisms.



Taxonomy of enzyme function: EC classification

- The Enzyme Commission (EC) classification scheme divides enzymes classes based on their function.
- The scheme has four levels, the three first level specifying the general kind of the reaction (oxidation, hydrolysis, which kind of bonds are acted on, which co-factors are used and so on.The fourth level contains individual enzymes.
- The EC scheme is the current standard for denoting enzyme function

Enzyme EC numbers

EC (Enzyme Commission) numbers assigned by IUPAC–IUBMB

Pathway Search by [EC | Cpd | Gene | Seq]
[1st Level | 2nd Level | 3rd Level | 4th Level | Text Search]

1. Oxidoreductases

2. Transferases
  2.1 Transferring one-carbon groups
    2.1.1 Methyltransferases
    2.1.2 Hydroxymethyl-, formyl- and related transferases
    2.1.3 Carboxyl- and carbamoyltransferases
    2.1.4 Amidinotransferases
      2.1.4.1 Glycine amidinotransferase
      2.1.4.2 Inosamine-phosphate amidinotransferase
  2.2 Transferring aldehyde or ketone residues
  2.3 Acyltransferases
  2.4 Glycosyltransferases
  2.5 Transferring alkyl or aryl groups, other than methyl groups
  2.6 Transferring nitrogenous groups
  2.7 Transferring phosphorus-containing groups
  2.8 Transferring sulfur-containing groups
  2.9 Transferring other groups

3. Hydrolases

4. Lyases

5. Isomerases

6. Ligases

[KEGG Home Page | GenomeNet Home Page | DBGET Links Diagram]

Metabolic reconstruction workflow

- Start from a sequenced genome of an organism
- Obtain annotations for ORFs via sequence homology and pick those with annotated enzymatic reaction (EC class)
- Pick reactions that have multiple polypeptides (or ORFs) associated and decide if they correspond to protein complexes or isozymes. (If available protein-protein interaction data could be used here)
- Fill in gaps in the metabolism: metabolites that cannot be produced by the reactions although they are empirically observed. Here sources other than sequence homology data are useful (phylogenetic profiling, metabolite concentrations, literature)

Constructing whole-genome metabolic reconstructions is a non-trivial exercise: each such reconstruction is typically worth a publication.

Genome annotation

Since few organism have extensive biochemical information available, reconstruction relies heavily on an annotated genome sequence.

Traditional techniques for annotation include

- Experimental methods: gene cloning or knockout and observation of changes in the phenotype
- Sequence homology: comparing the sequence to genes with known function in other organisms

Genome annotation

More recent techniques include:

- Protein-protein interaction data: if two enzymes are known to form a complex, it is likely that they together catalyze the same or adjacent reactions in the metabolic network
- Correlated mRNA expression: an enzyme that has similar expression profile (over a set of conditions) might have a similar function
- Phylogenetic profiling: based on the assumption that proteins that function together in a pathway or structural complex are likely to evolve in a correlated fashion. Functionally linked proteins tend to same similar occurrence profiles accross species.

Finding similar sequences

- Alignment: Use the BLAST or FASTA family of methods to align ORFs with the sequences of known enzymes function contained in enzyme databases such as IntEnz (www.ebi.ac.uk/intenz) or Uni-Prot (www.expasy.ch).

Finding similar sequences

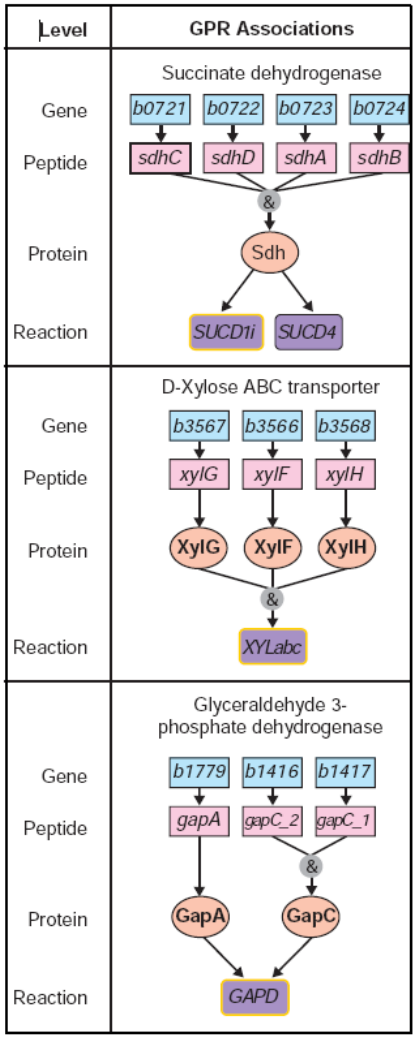
- Alignment: Use the BLAST or FASTA family of methods to align ORFs with the sequences of known enzymes function contained in enzyme databases such as IntEnz (www.ebi.ac.uk/intenz) or Uni-Prot (www.expasy.ch).
  - Function can be reliably assigned for sequences that are evolutionarily close but it is not reliable for distant homologs.

Finding similar sequences

- Alignment: Use the BLAST or FASTA family of methods to align ORFs with the sequences of known enzymes function contained in enzyme databases such as IntEnz (www.ebi.ac.uk/intenz) or Uni-Prot (www.expasy.ch).
  - Function can be reliably assigned for sequences that are evolutionarily close but it is not reliable for distant homologs.
- Conserved motifs: find groups of conserved amino acids, 'motifs' that are stored in a database such as PROSITE (www.expasy.ch/prosite/).
  - The idea is to define certain conserved amino acid patterns that are related to function, e.g. they are residues close to the active site.
  - These methods are more sensitive for function determination than alignment techniques.

Gene-protein-reaction interactions

- Peptides from several genes may be used to encode single protein which may catalyze several reactions (top picture)
- Several proteins may form a complex to catalyze a single reaction (middle picture)
- Different genes may encode isozymes (proteins with identical function) that catalyze the same reaction (bottom picture)



(picture from Reed et al. Genome Biology 4, 2003)

Pathway Tools (http://bioinformatics.ai.sri.com/ptools/)

- One of the few software packages that assists in the construction of pathway/genome databases such as EcoCyc.
- PathoLogic tool takes an annotated genome for an organism and infers probable metabolic pathways to produce a new pathway/genome database.
- This can be followed by application of the Pathway Hole Filler, which predicts likely genes to fill "holes" (missing steps) in predicted pathways.
- In addition there are Navigation and editing tools by which the user can visualize, analyze, access and update the database.
- The rationale: Pathway Tools give a rapid first blueprint of the metabolic network that can be iteratively refined.