Unravelling the Function, Evolution and Interactions in Biochemical Pathways through Systems Biology

Dr Sophia Tsoka

Centre for Bioinformatics

School of Physical Sciences and Engineering

King's College London

Academic Background

- Chemistry (Aristotle University of Thessaloniki, Greece)
- MSc Food Biotechnology (University of Reading)
- PhD Biochemical Engineering (University College London)

- Research Associate
  Chemical Engineering, National Technical University of Athens
  (enzymatic systems for time-temperature integration)
- Postdoctoral Fellow
  European Bioinformatics Institute, Computational Genomics Group
  (representation and analysis of sequenced genomes, functional assignment, protein interactions)
- Staff Scientist – MRC Fellow
  European Bioinformatics Institute, Computational Genomics Group
  (metabolic reconstruction, stoichiometric modelling, network analysis, systems biology)
- Lecturer
  KCL Centre for Bioinformatics

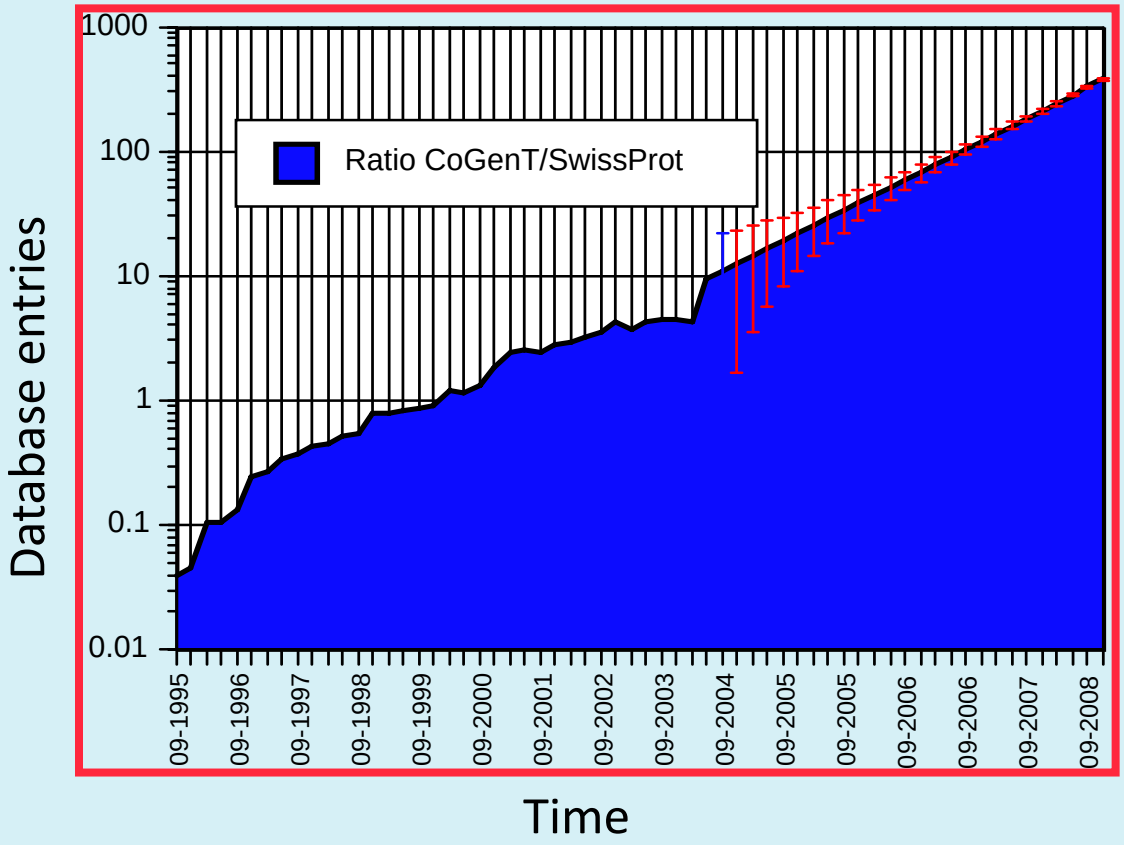
Presentation Outline

- General concepts in Bioinformatics
- Computational Genomics
  - Genome Analysis
    - sequence to function
  - Comparative Genomics
    - evolution of metabolic enzymes and pathways
  - Network Analysis
    - protein interactions
    - metabolic reconstruction
    - network robustness
- Conclusions

Bioinformatics and Computational Genomics

- *Bioinformatics:* development and implementation of computational methods for the analysis, storage and representation of biological information
- *Computational Genomics* involves the analysis of entire genomes
- *Systems Biology:* need to progress from individual assignments to detailed descriptions of cells as entire *systems*
- Advances in biological science and information technology force a new way of thinking about biological sciences and will ultimately lead to deeper understanding of nature and new experiments

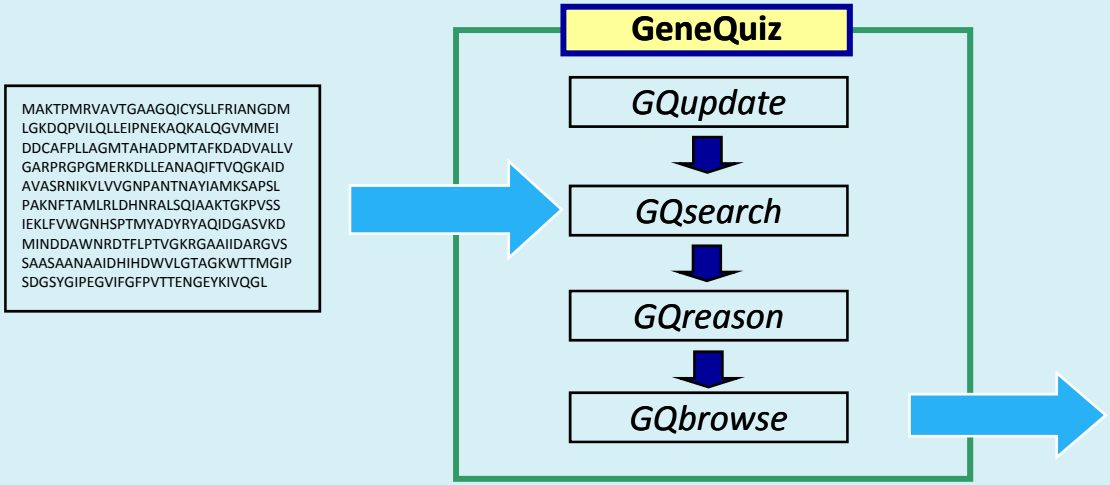
An Abundance of Data, but…



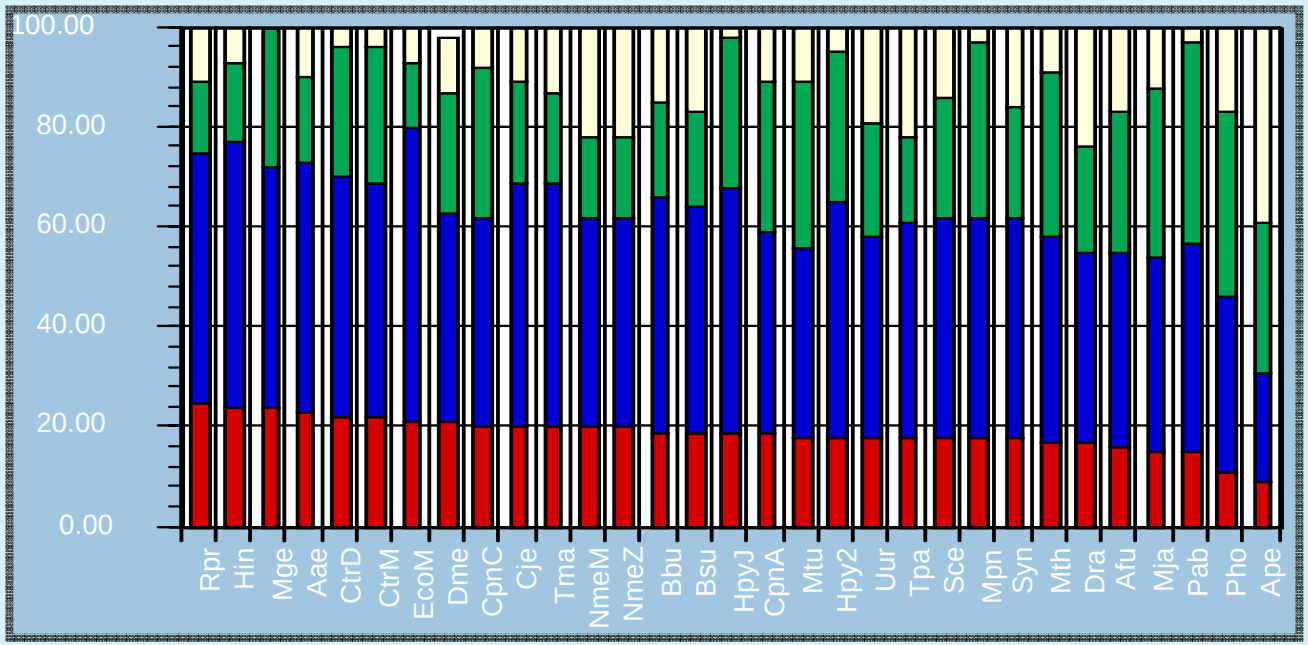
- **Bacteria: 658 complete, 1758 on-going**
- **Archaea: 53 complete, 90 on-going**
- **Eukaryotes: 86 complete, 934 on-going**
- **Metagenomes: 126 complete**

http://www.genomesonline.org/ (19/05/2008)

… a Scarcity of Information



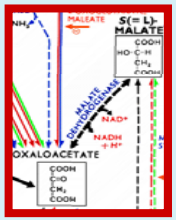
Database of results:
Protein function,
Homologous proteins,
3D structure

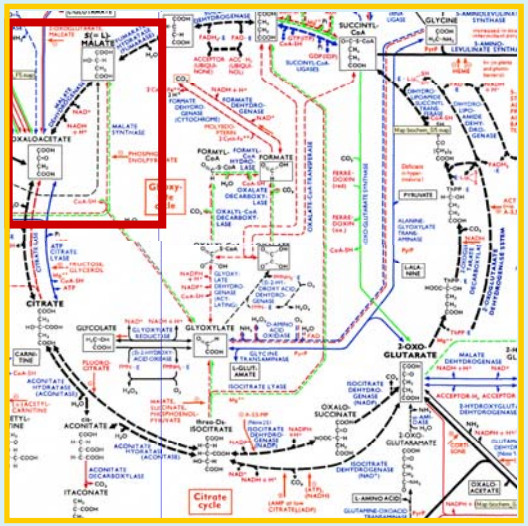


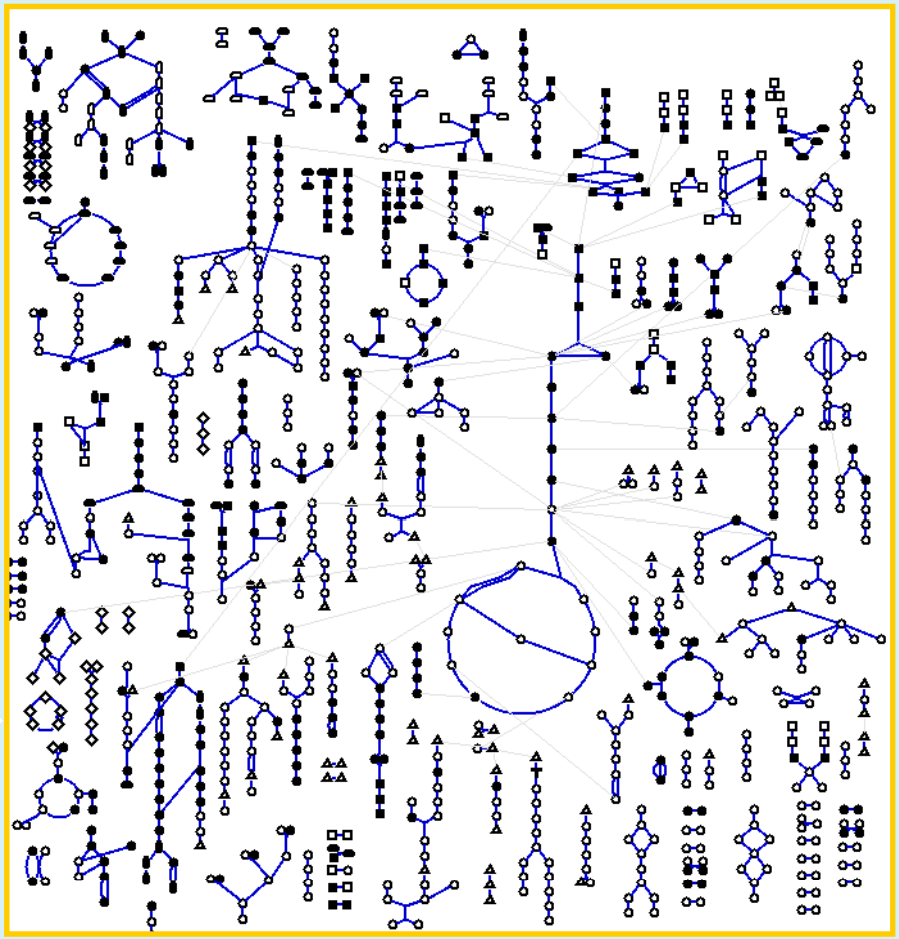
From individual proteins

to protein involvement in pathway

to networks



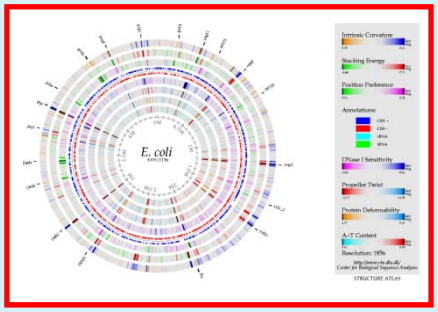




From
a Model Genome

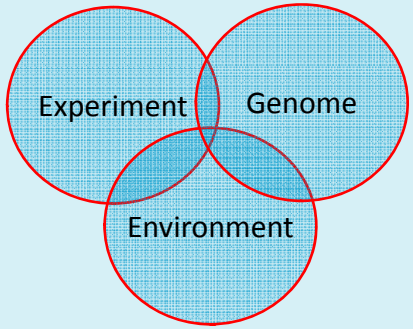
to
a Multitude of Genomes

to
Systems



http://www.cbs.dtu.dk/services/GenomeAtlas/

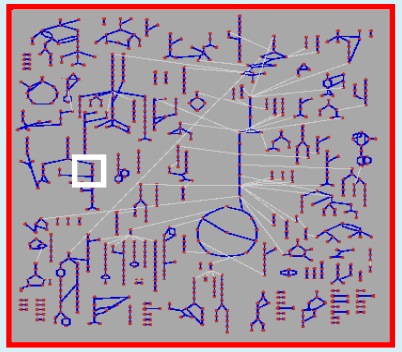
http://mkweb.bcgsc.ca/circos/

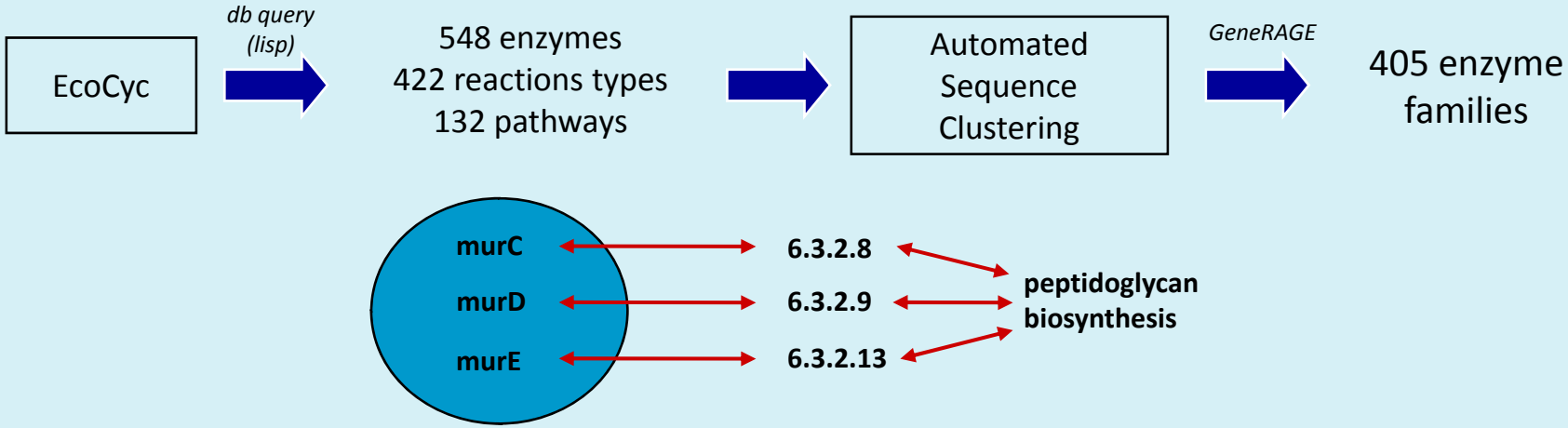


Computational Genomics ⟷ **Comparative Genomics** ⟷ **Systems Biology**

Understanding the Sequence – Function Relationships

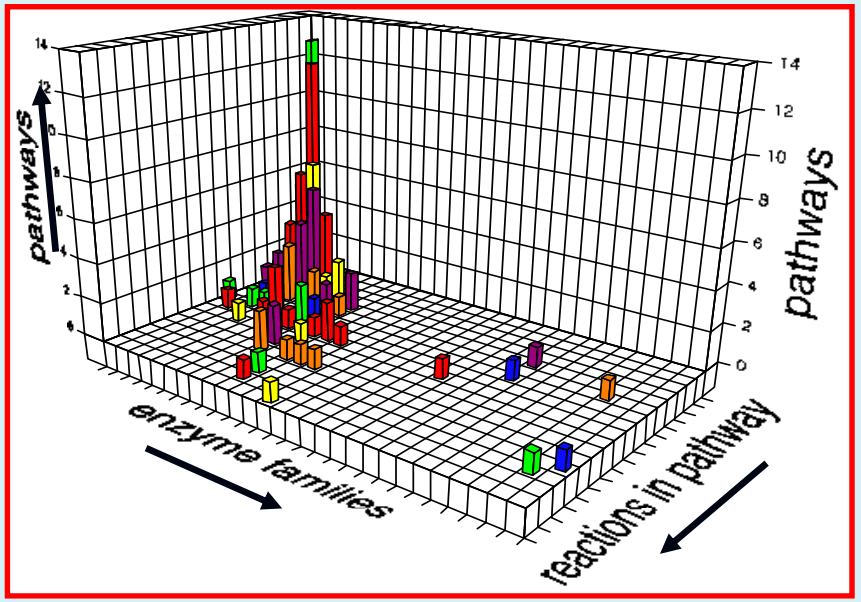
- Associate enzyme sequence to function for the entire *E. coli* metabolism
  - each family: functional versatility of family members
  - each reaction/pathway: molecular diversity of enzymes
- Provide insight to:
  - function prediction based on sequence homology
    - quantify *convergent* and *divergent* evolution for an entire species
  - evolution of biochemical pathways



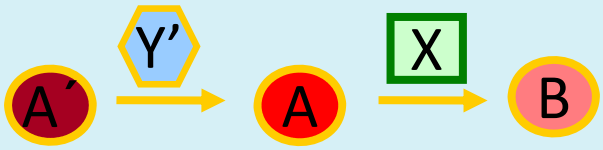


How Do Metabolic Networks Develop ?

- How many “building blocks” of enzyme families per metabolic pathway?

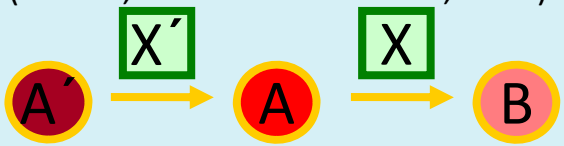


- Recruitment Mode



  - non-homologous enzymes within a pathway (Jensen, Ann. Rev. Microbiol., 1976)

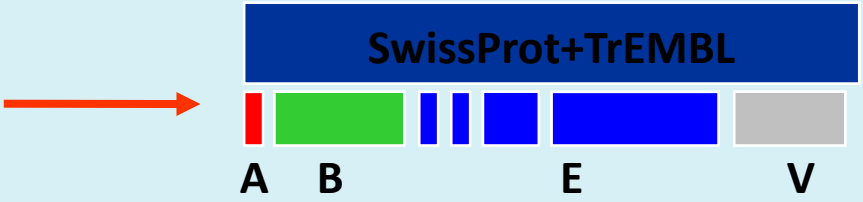
- Retrograde Mode



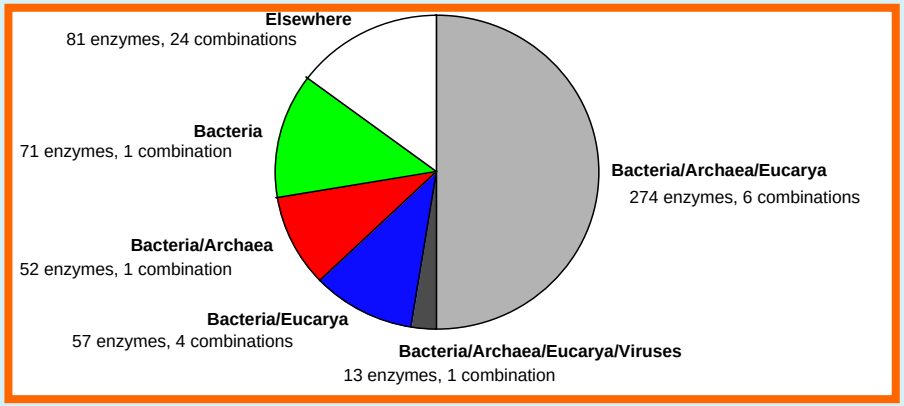
  - concentration of homologous enzymes within a pathway (Horowitz, PNAS, 1945)

How Do Metabolic Networks Evolve ?

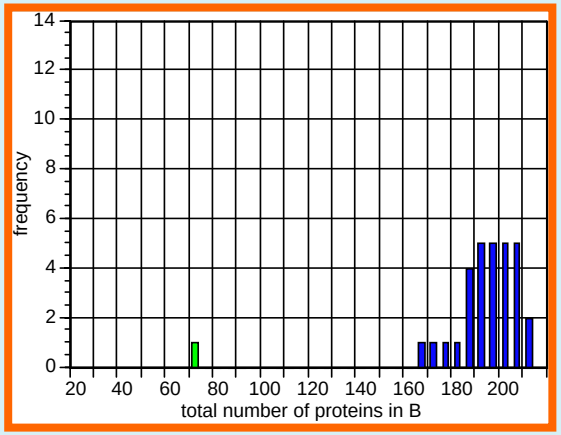
- E. coli enzymes (1 set of 548 proteins)
- Control samples (30 sets of 548 selected randomly)



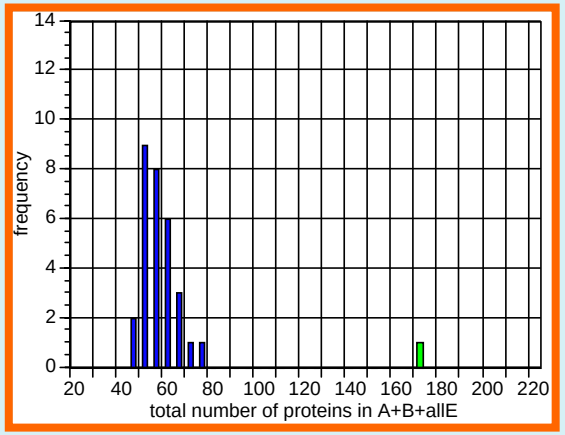
How Conserved is Small-Molecule Metabolism across Taxonomic Groups?



Are Metabolic Enzymes More/Less Conserved than Controls?



Bacterial enzymes are less species-specific…

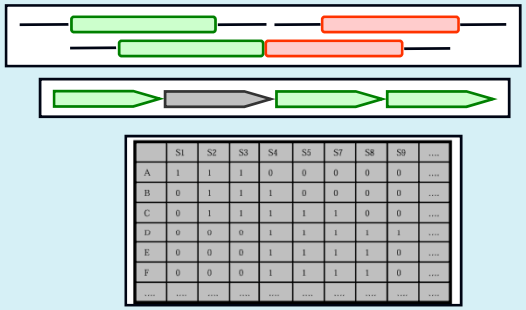


… and more phylogenetically diverse, compared to controls

Computational Prediction of Protein Interactions

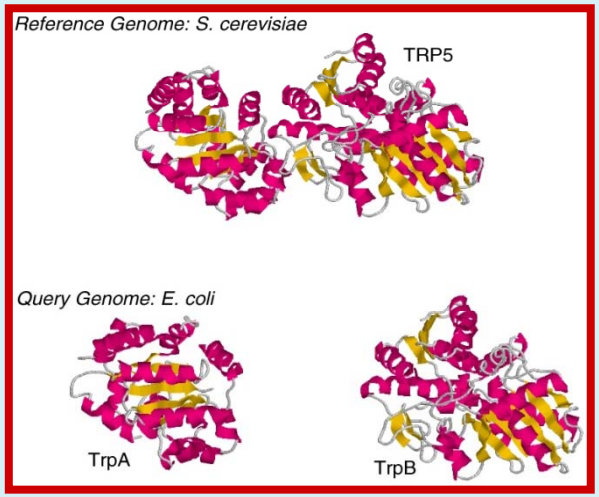
Detection of protein interactions based on genome features:

- gene fusion
- gene neighbourhood
- phylogenetic profiles

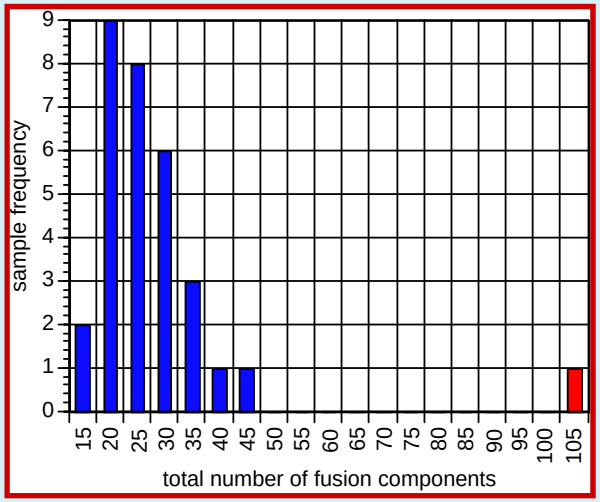


	S1	S2	S3	S4	S5	S7	S8	S9	....
A	1	1	1	0	0	0	0	0	....
B	0	1	1	1	0	0	0	0	....
C	0	1	1	1	1	1	0	0	....
D	0	0	0	1	1	1	1	1	....
E	0	0	0	1	1	1	1	0	....
F	0	0	0	1	1	1	1	0	....
....	....	....	....	....	....	....	....	....	....

A gene fusion event

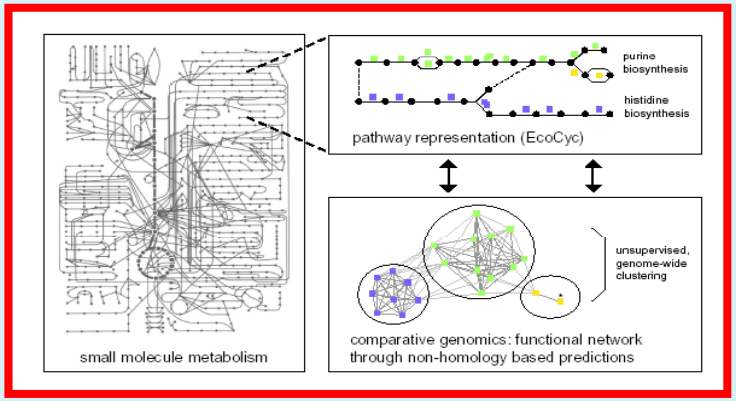


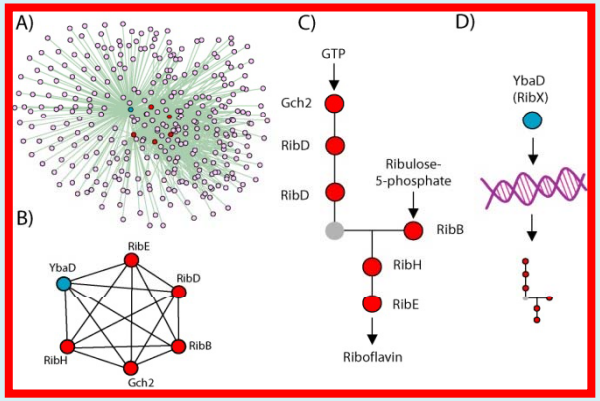
Metabolic enzymes frequently involved in gene fusion



Are Metabolic Interactions Described in Genome Structure ?

- integration of computational methods for detecting protein interactions
- protein network clustering
- benchmark against known metabolic pathways



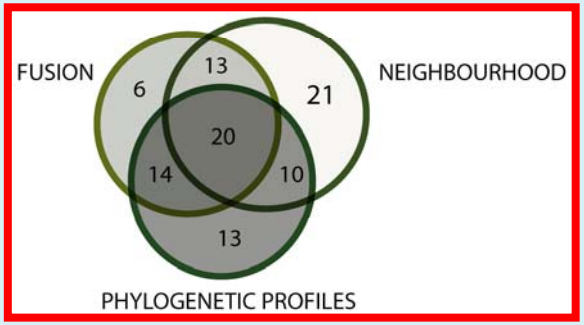


•metabolic enzymes cluster in modules (84% ave. pathway specificity)

•much of bacterial metabolism is encoded in genome features

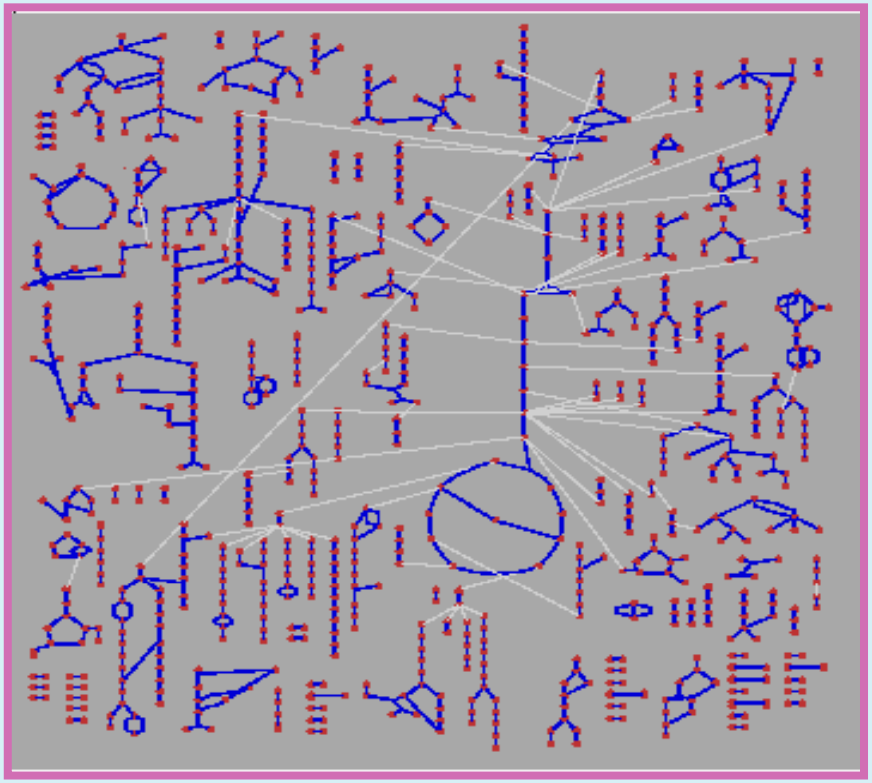
•some novel functions can be predicted (left)

the relative contribution of the methodologies used:



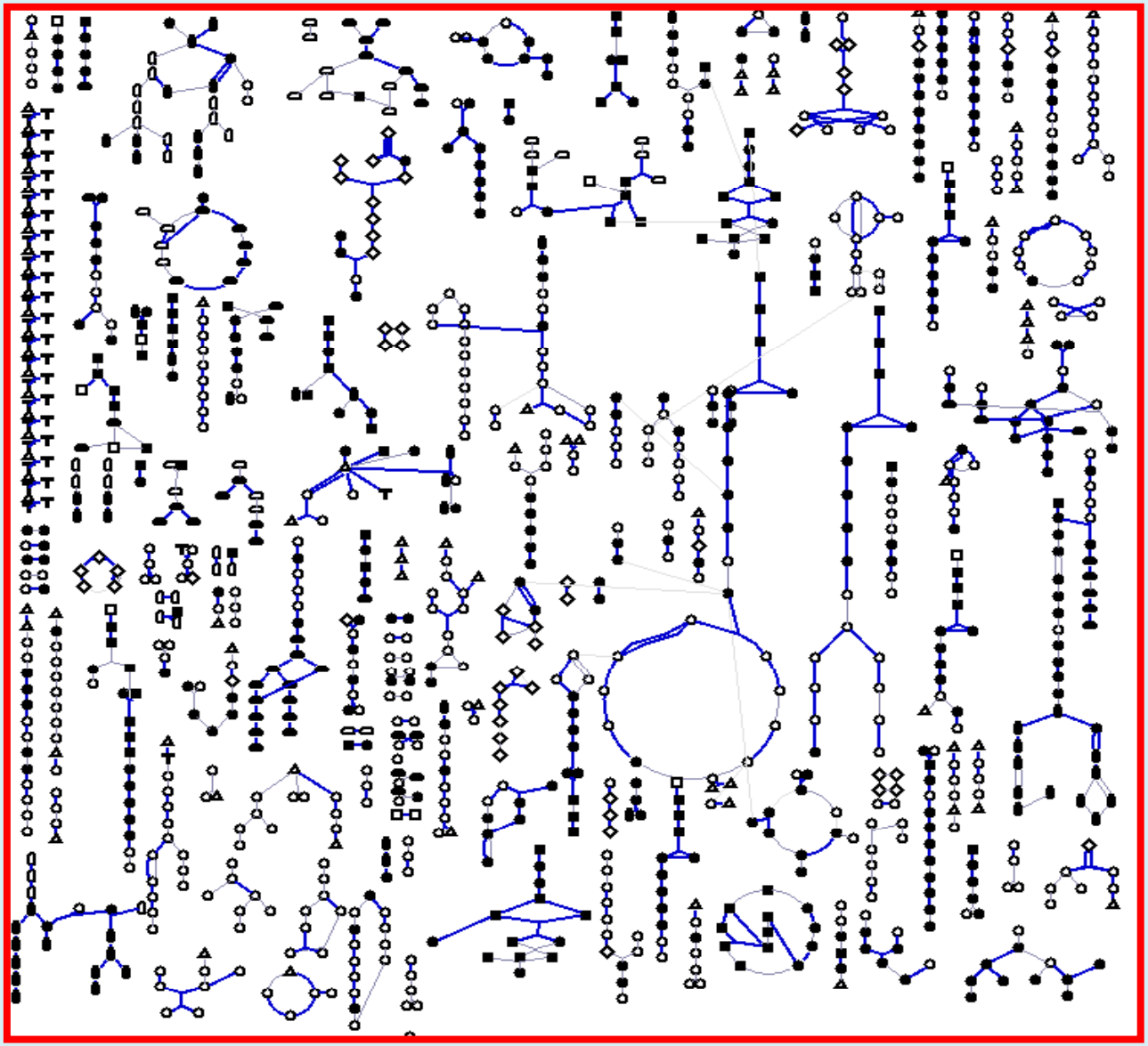
Metabolic Databases: Design and Implementations

- EcoCyc: experimental information for the entire known metabolic complement of the bacterium *Escherichia coli*
- Object-oriented architecture, ontology permit large-scale data mining
- Representation of metabolic pathways, includes:
  - Reactions
  - Compounds
  - Stoichiometry
  - Inhibitors
  - … etc
- Powerful query capabilities using lisp, java



http://www.ecocyc.com/

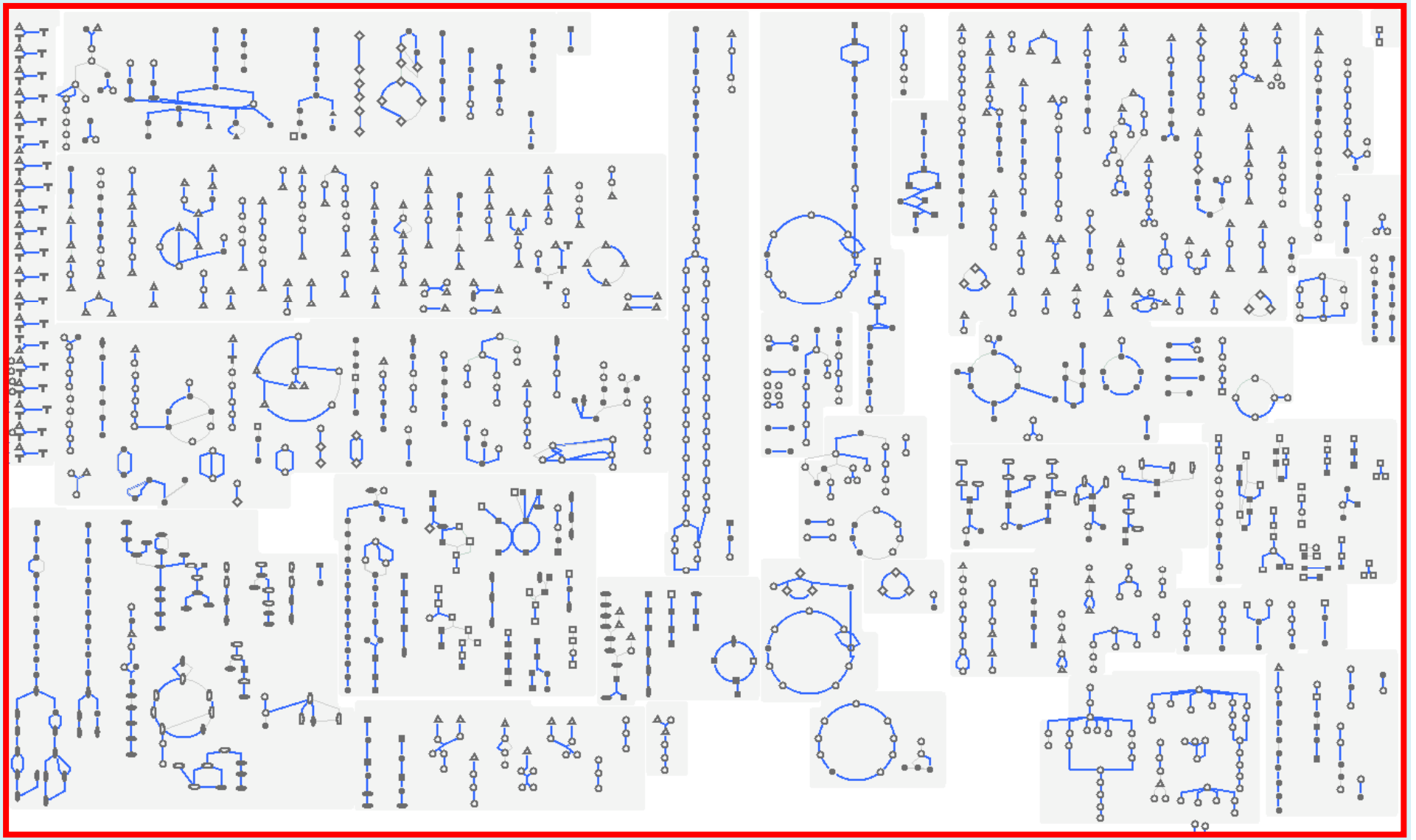
The *Plasmodium falciparum* Metabolic Network



http://plasmocyc.stanford.edu/

Yeh, Hanekamp, Tsoka, Karp, Altman, Genome Research, 14, 917, 2004

The Human Metabolic Network

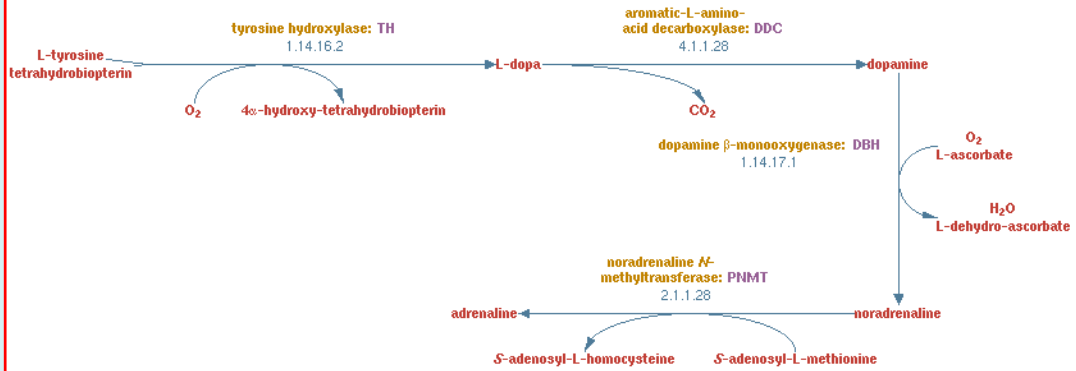


http://humancyc.org/

H. sapiens Pathway: catecholamine biosynthesis

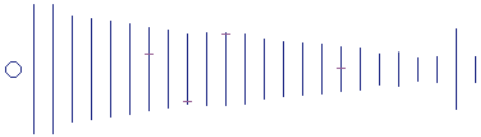
Login (Optional): Why Login?
Create New Account | Help

Customize Diagram | More Detail | Less Detail | Cross-Species Comparison | Download Genes | BioPAX format



If an enzyme name is shown in bold, there is experimental evidence for this enzymatic activity.

Locations of Mapped Genes:



Synonyms: dopamine biosynthesis, noradrenaline biosynthesis, adrenaline biosynthesis

Superclasses: Biosynthesis -> Hormones Biosynthesis

Pathway Summary from MetaCyc:

The catecholamines (norepinephrine, epinephrine and dopamine) are synthesized in the central nervous system (CNS), sympathetic nerves and in the chromaffin cells of the adrenal medulla. Nonneuronal cells in the gastrointestinal tract and the kidneys are among other tissues capable of producing catecholamines.

In the CNS, dopamine and norepinephrine are widely distributed, whereas epinephrine is found in the mammalian brain in relatively low concentrations. In the periphery, norepinephrine is the transmitter of the postganglionic sympathetic nervous system and dopamine is involved in the regulation of renal and gastrointestinal function. The main source of epinephrine outside of the CNS are the chromaffin cells of the adrenal medulla.

Dopamine, norepinephrine and epinephrine are synthesized by a series of enzymes with cytoplasmic and vesicular locations. This biosynthetic pathway was first postulated in 1939 [Blaschko59], and the rate-limiting step was experimentally confirmed in1964 [Nagatsu64 , Nagatsu64a] Synthesis of dopamine, norepinephrine and epinephrine ends at different points in the pathway depending on the availability of the various biosynthetic enzymes involved in each step.

Pathway Evidence Glyph:

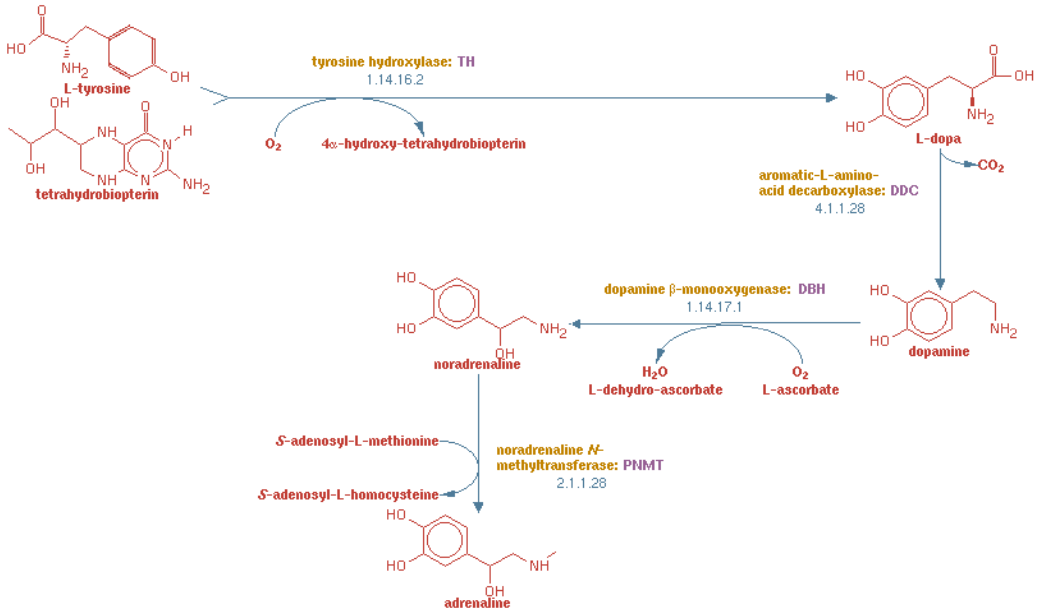
Key to pathway glyph edge colors:

An enzyme catalyzing this reaction is present in this organism
The reaction and any enzyme that catalyzes it (if one has been identified) is unique to this pathway

H. sapiens Pathway: catecholamine biosynthesis

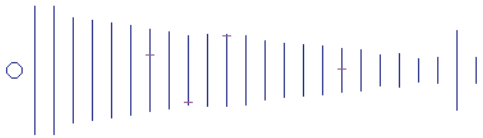
Login (Optional): Why Login?
Create New Account | Help

Customize Diagram | More Detail | Less Detail | Cross-Species Comparison | Download Genes | BioPAX format



If an enzyme name is shown in bold, there is experimental evidence for this enzymatic activity.

Locations of Mapped Genes:



Synonyms: dopamine biosynthesis, noradrenaline biosynthesis, adrenaline biosynthesis

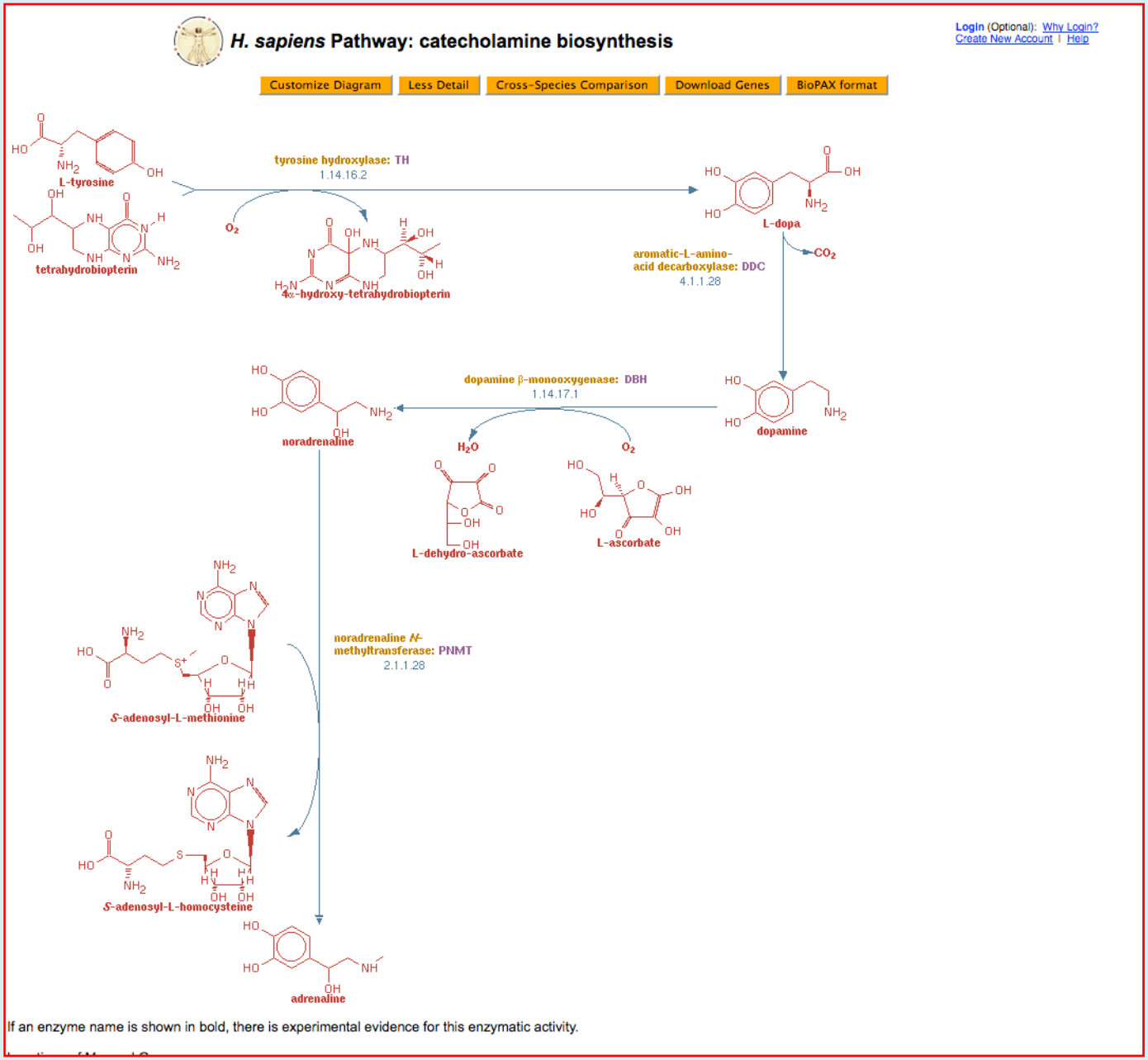
Superclasses: Biosynthesis -> Hormones Biosynthesis

Pathway Summary from MetaCyc:

The catecholamines (norepinephrine, epinephrine and dopamine) are synthesized in the central nervous system (CNS), sympathetic nerves and in the chromaffin cells of the adrenal medulla. Nonneuronal cells in the gastrointestinal tract and the kidneys are among other tissues capable of producing catecholamines.

In the CNS, dopamine and norepinephrine are widely distributed, whereas epinephrine is found in the mammalian brain in relatively low concentrations. In the periphery, norepinephrine is the transmitter of the postganglionic sympathetic nervous system and dopamine is involved in the regulation of renal and gastrointestinal function. The main source of epinephrine outside of the CNS are the chromaffin cells of the adrenal medulla.

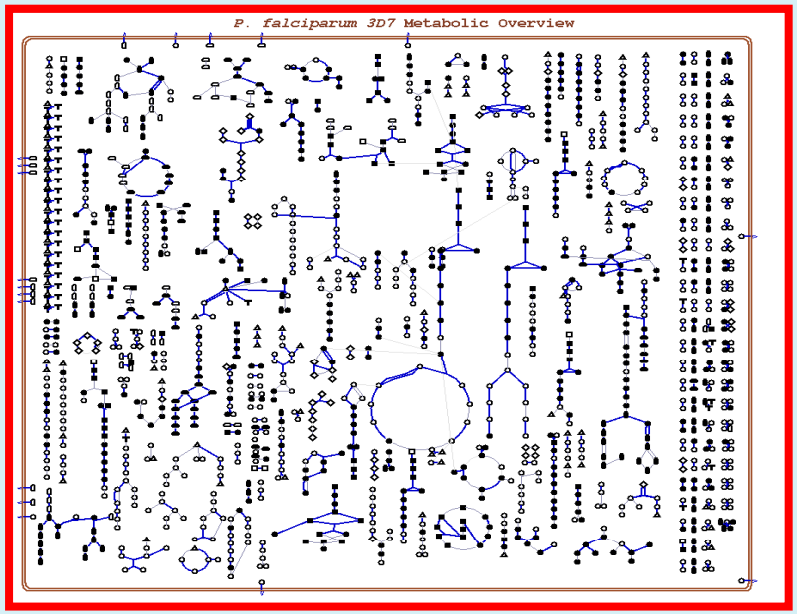
Dopamine, norepinephrine and epinephrine are synthesized by a series of enzymes with cytoplasmic and vesicular locations. This biosynthetic pathway was first postulated in 1939 [Blaschko59], and the rate-limiting step was experimentally confirmed in1964 [Nagatsu64 , Nagatsu64a] Synthesis of dopamine, norepinephrine and epinephrine ends at different


H. sapiens Pathway: catecholamine biosynthesis
Login (Optional): Why Login?
Create New Account | Help
Customize Diagram
Less Detail
Cross-Species Comparison
Download Genes
BioPAX format
L-tyrosine
tetrahydrobiopterin
tyrosine hydroxylase: TH
1.14.16.2
O2
4α-hydroxy-tetrahydrobiopterin
L-dopa
CO2
aromatic-L-amino-acid decarboxylase: DDC
4.1.1.28
dopamine
dopamine β-monooxygenase: DBH
1.14.17.1
H2O
O2
noradrenaline
L-dehydro-ascorbate
L-ascorbate
noradrenaline N-methyltransferase: PNMT
2.1.1.28
S-adenosyl-L-methionine
S-adenosyl-L-homocysteine
adrenaline
If an enzyme name is shown in bold, there is experimental evidence for this enzymatic activity.

Metabolic Reconstruction for Drug Target Discovery

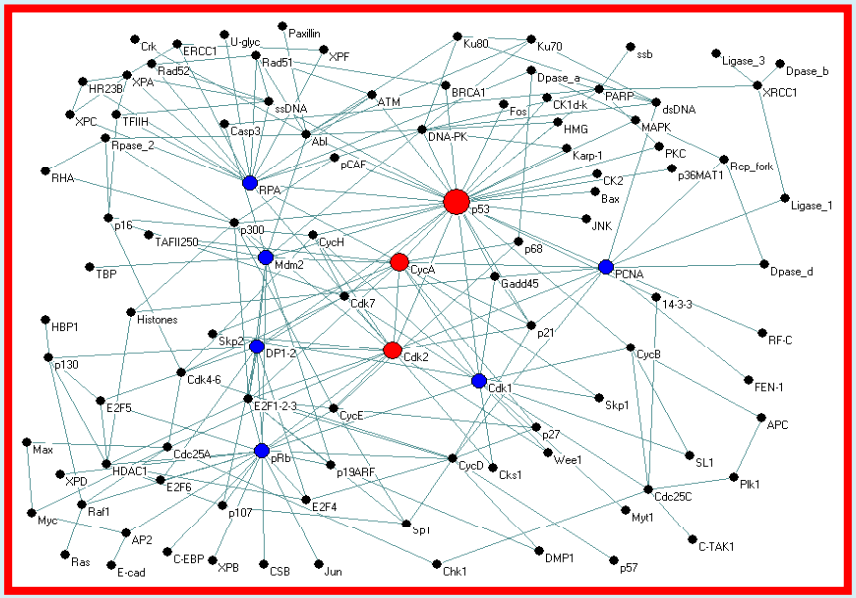
Plasmodium falciparum

- 5366 proteins
- 122 pathways
- 697 reactions
- 861 enzymatic reaction
- 525 compounds
- 216 chokepoint reactions as drug targets



http://plasmocyc.stanford.edu/

Network Robustness



analysis of the p53 protein interaction network

104 nodes, 226 interactions

network robustness through change in network diameter after node deletion

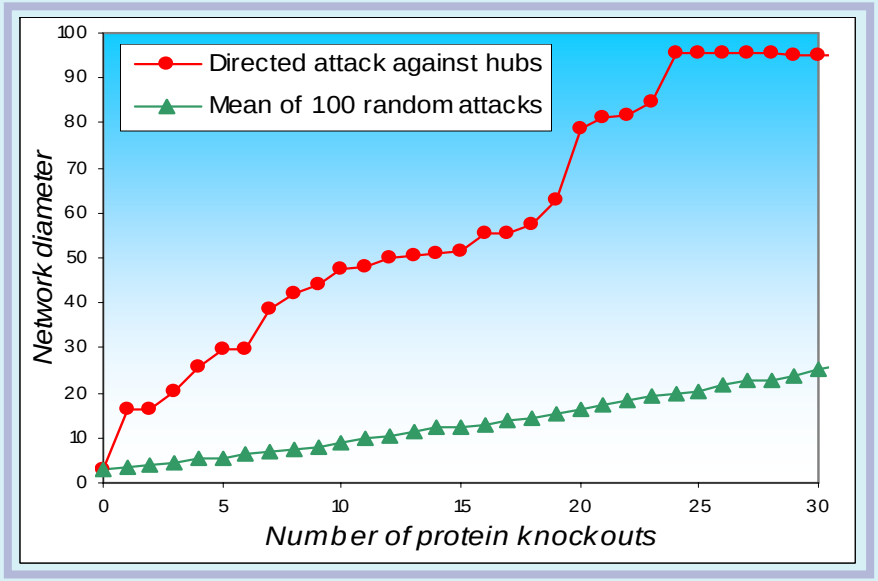
Average path length (APL) is the mean of the shortest paths to all other nodes, a measure of node centrality in the network

Network diameter is the average of all APLs and a measure of *communication* in the network

$$D = \frac{\sum_i APL_i}{N}$$

APL	Protein nodes
1.9	p53
2.1	Cdk2
2.2	CycA
2.3	Cdk1, Mdm2, DP1-2, pRb
2.4	PCNA, RPA
2.5	DNA-PK, p21, p300, E2F1-2-3, Cdk7, CycH
2.6	Abl, Gadd45
2.7	CycB, CycD, CycE, PARP, ATM
2.8	ssDNA, Cdc25A, 14-3-3, pCAF, PKC
2.9	HMG, Karp-1, BRCA1

How Robust is a Cell Signalling Network ?



Robustness of network studied through change in network diameter after node deletion

Degeneration of the p53 network diameter under simulated node attack:

- network robust to random protein knockouts,
- the network rapidly fragments under an attack directed against the hubs

Tumour-inducing viruses behave like biological hackers against this vulnerability

The TIV directed strikes are effective at disrupting communication within the p53 network

TIV	Cellular proteins targeted	Network diameter after knockout
Adenovirus	p53, pRb	24.98
Coxsackie	Cyclin D1	5.00
HCMV	pRb, p107, p130	14.37
HPV 16/18	p53, pRb, p107, p130	27.07
SV40	p53, pRb	24.98

Network Modularity

- measures the quality of partitioning a network into communities
- larger *modularity* value gives better quality of network partition into *modules*

Given:

An undirected network consisting of N nodes and L links

Determine:

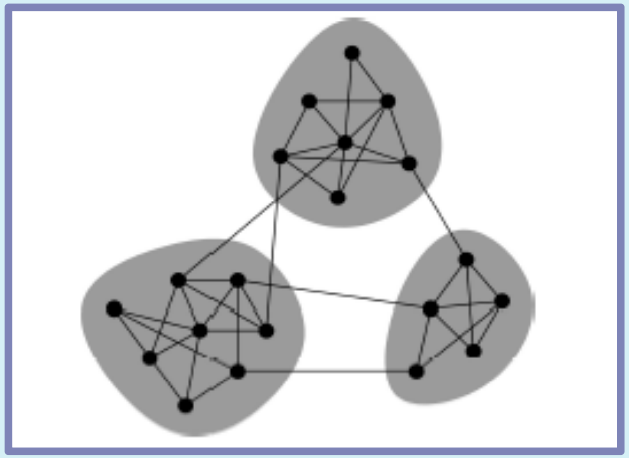
The optimal number of communities

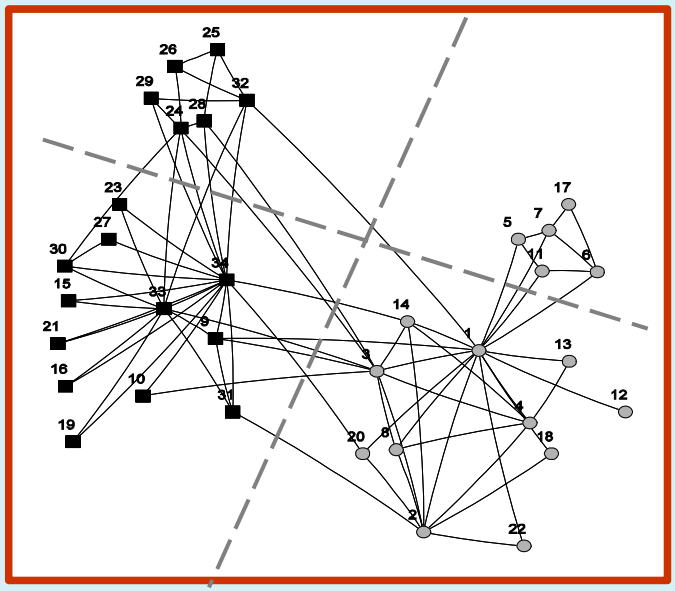
Node-module allocation

So as to:

Maximise the network modularity metric

$$Q = \sum_m \left[\frac{L_m}{L} - \left(\frac{D_m}{2L} \right)^2 \right]$$





Applied on social and biological networks

Optimal community structures with maximum modularity measure achieved

Conclusions

- Automated function prediction and classification
- Evolution of metabolic pathways
- Prediction of protein interactions
- Network reconstruction
- Network analysis

- Improved *understanding* of the rules that govern biological systems

- Improved *design* of biocatalytical processes using Systems Biology

Collaborations

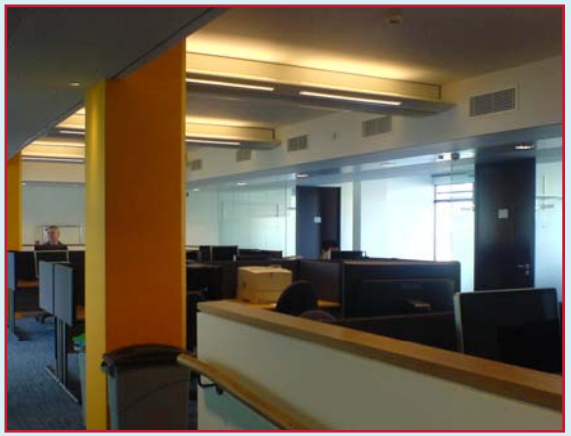
Prof Russ Altman	Stanford University, USA
Dr Benjamin Audit	ENS, Lyon, France
Dr Peer Bork	EMBL, Heidelberg, Germany
Prof David Bogle	UCL, London, UK
Dr Richard Coulson	EBI, Cambridge, UK
Dr Anton Enright	EBI, Cambridge, UK
Dr Wally Gilks	Univ. of Leeds, UK
Dr Paul Janssen	SCKCEN, Belgium
Dr Peter Karp	SRI Menlo Park, USA
Prof Christos Ouzounis	KCL Centre for Bioinformatics
Dr Lazaros Papageorgiou	UCL, London, UK
Dr Jose M Peregrin-Alvarez	Hospital for Sick Children, Toronto, Canada
Dr Jose B Pereira-Leal	Instituto Gulbenkian de Ciencia, Lisbon, Portugal
Dr Chris Sander	Memorial Sloan Kettering, New York, USA
Prof Anna Tramontano	University of Rome, Italy
Prof Alfonso Valencia	CNIO, Madrid, Spain

Associated publications

- Tsoka, Ouzounis, *Genome Research*, 11, 1503, 2001
- Peregrin-Alvarez, Tsoka, Ouzounis, *Genome Research*, 13, 422, 2003
- Tsoka, Ouzounis, *Nature Genetics*, 26, 141, 2000
- von Mering, Zdobnov, Tsoka, Ciccarelli, Pereira-Leal, Ouzounis, Bork, *PNAS*, 100, 15428, 2003
- Tsoka, Simon, Ouzounis, *Archaea*, 1(4), 223, 2004
- Yeh, Hanekamp, Tsoka, Karp, Altman, *Genome Research*, 14, 917, 2004
- Dartnell, Simeonidis, Hubank, Tsoka, Papageorgiou, *FEBS Letters*, 579, 3037, 2005
- Xu, Tsoka & Papageorgiou, *Eur. Phys. J. B*, 60, 231-239, 2007

Funding

Medical Research Council
EU Training and Mobility



The King's Centre for Bioinformatics
http://www.kcl.ac.uk/schools/pse/bioinform/