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)
From individual proteins to protein involvement in pathway to networks
From a Model Genome to a Multitude of Genomes to Systems

Computational Genomics  ↔  Comparative Genomics  ↔  Systems Biology

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

http://mkweb.bcgsc.ca/cirros/
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

---

Tsoka and Ouzounis, Genome Research, 11, 1503, 2001

www.kcl.ac.uk
How Do Metabolic Networks Develop?

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

- **Recruitment Mode**

- **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

Peregrin-Alvarez, Tsoka, Ouzounis, Genome Research, 13, 422, 2003
Computational Prediction of Protein Interactions

Detection of protein interactions based on genome features:
- gene fusion
- gene neighbourhood
- phylogenetic profiles

A gene fusion event

Metabolic enzymes frequently involved in gene fusion

Tsoka, Ouzounis, Nature Genetics, 26, 141, 2000
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:

von Mering, Zdobnov, Tsoka, Ciccarelli, Pereira-Leal, Ouzounis, Bork, PNAS, 100, 15428, 2003
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/

P.D. Karp, Nucleic. Acids Res., 33, 6083-6089, 2005
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

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

Locations of Mapped Genes:

Synonym: dopamine biosynthesis, noradrenaline biosynthesis, adrenaline biosynthesis
Superclass: Biochemistry > 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 [Bachofen], and the rate-limiting step was experimentally confirmed in 1954 [Nagatsu, Nagatsu, Ikeda] 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

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 [Blaichko59] and the rate-limiting step was experimentally confirmed in 1964 [Nagatsu54, Nagatsu54a]. Synthesis of dopamine, norepinephrine and epinephrine ends at different
**H. sapiens Pathway: catecholamine biosynthesis**

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/

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

\[
D = \frac{\sum_i APL_i}{N}
\]

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

<table>
<thead>
<tr>
<th>TIV</th>
<th>Cellular proteins targeted</th>
<th>Network diameter after knockout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenovirus</td>
<td>p53, pRb</td>
<td>24.98</td>
</tr>
<tr>
<td>Coxsackie</td>
<td>Cyclin D1</td>
<td>5.00</td>
</tr>
<tr>
<td>HCMV</td>
<td>pRb, p107, p130</td>
<td>14.37</td>
</tr>
<tr>
<td>HPV 16/18</td>
<td>p53, pRb, p107, p130</td>
<td>27.07</td>
</tr>
<tr>
<td>SV40</td>
<td>p53, pRb</td>
<td>24.98</td>
</tr>
</tbody>
</table>

Tumour-inducing viruses behave like biological hackers against this vulnerability

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

Dartnell, Simeonidis, Hubank, Tsoka, Papageorgiou, FEBS Letters, 579, 3037, 2005

www.kcl.ac.uk
Network Modularity

- measures the quality of partitioning a network into communities
- larger \textit{modularity} value gives better quality of network partition into \textit{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, *Nature Genetics*, 26, 141, 2000
- von Mering, Zdobnov, Tsoka, Ciccarelli, Pereira-Leal, Ouzounis, Bork, *PNAS*, 100, 15428, 2003
- Yeh, Hanekamp, Tsoka, Karp, Altman, *Genome Research*, 14, 917, 2004

Funding

Medical Research Council
EU Training and Mobility

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