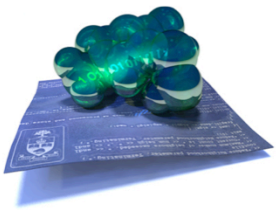
A
Bioinformatics
Course

Function Databases



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

EC (Enzyme Commission) numbers

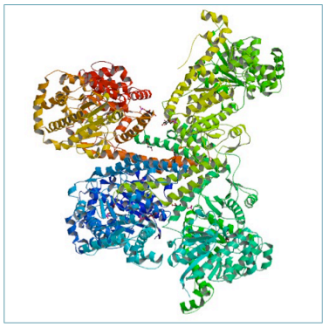
Enzymatic activity only.

Hierarchical classification, four levels.

Defines recommended names, need not be unique.

Multiple sites available to browse.

Annotated (by homology) for practically all enzymes in databases.



3NB0 – Glycogen Synthase as retrieved from the PDB with a search for the EC Number 2.4.1.11

```
1. -. -.-  Oxidoreductases.
2. -. -.-  Transferases.
3. -. -.-  Hydrolases.
4. -. -.-  Lyases.
5. -. -.-  Isomerases.
6. -. -.-  Ligases.
```

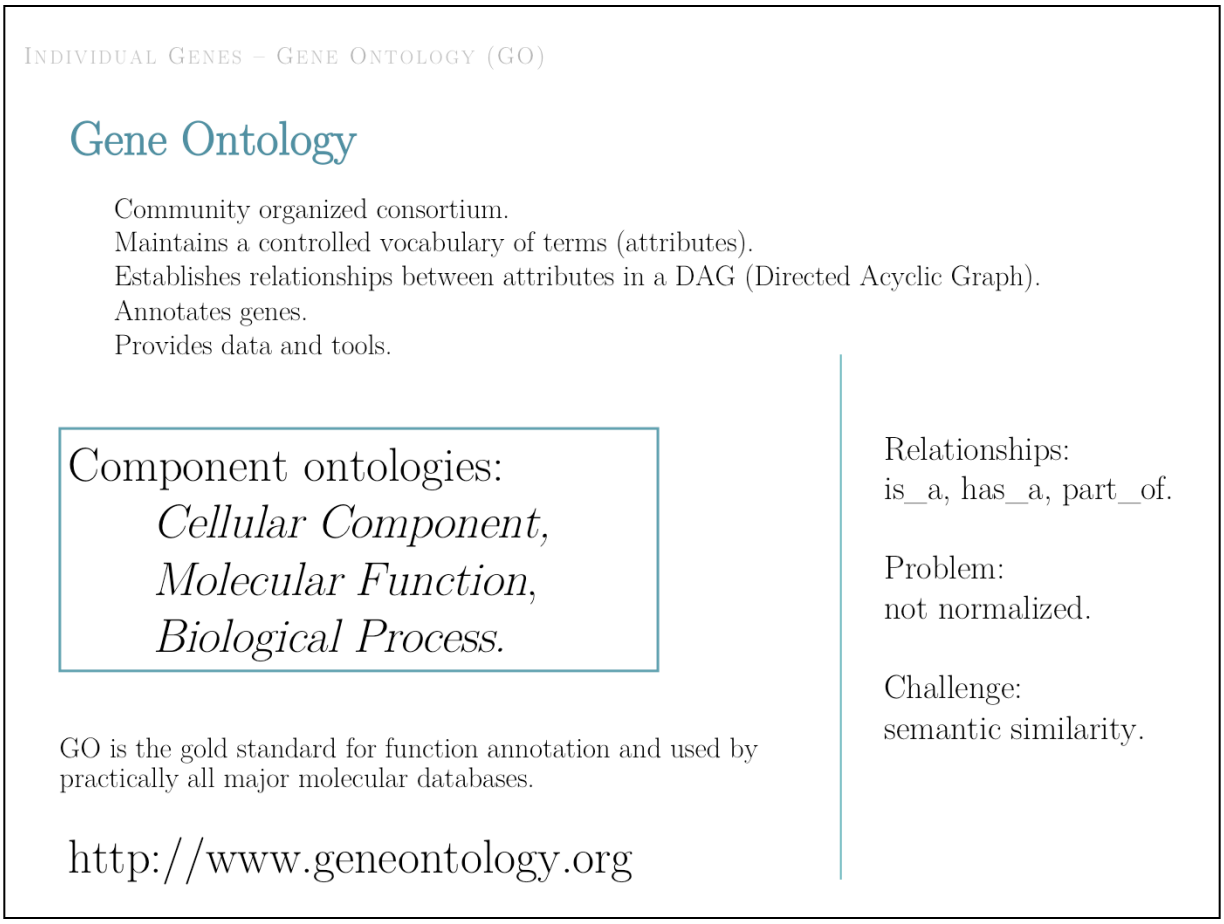
Example:

2._._._	-	Transferase
2.4._._	-	Glycosyltransferase
2.4.1._	-	Hexosyltransferase
2.4.1.11	-	Glycogen synthase

Annotation of individual genes: enzymatic function

EC numbers are the oldest approach to categorize function. Strictly speaking, they apply to reactions, not gene products. And one must keep in mind that non-orthologous genes may catalze the exactly same reaction due to convergent evolution.

Many databases – such as the PDB for example – list E.C. numbers as a standard annotation for the enzymes in their records.

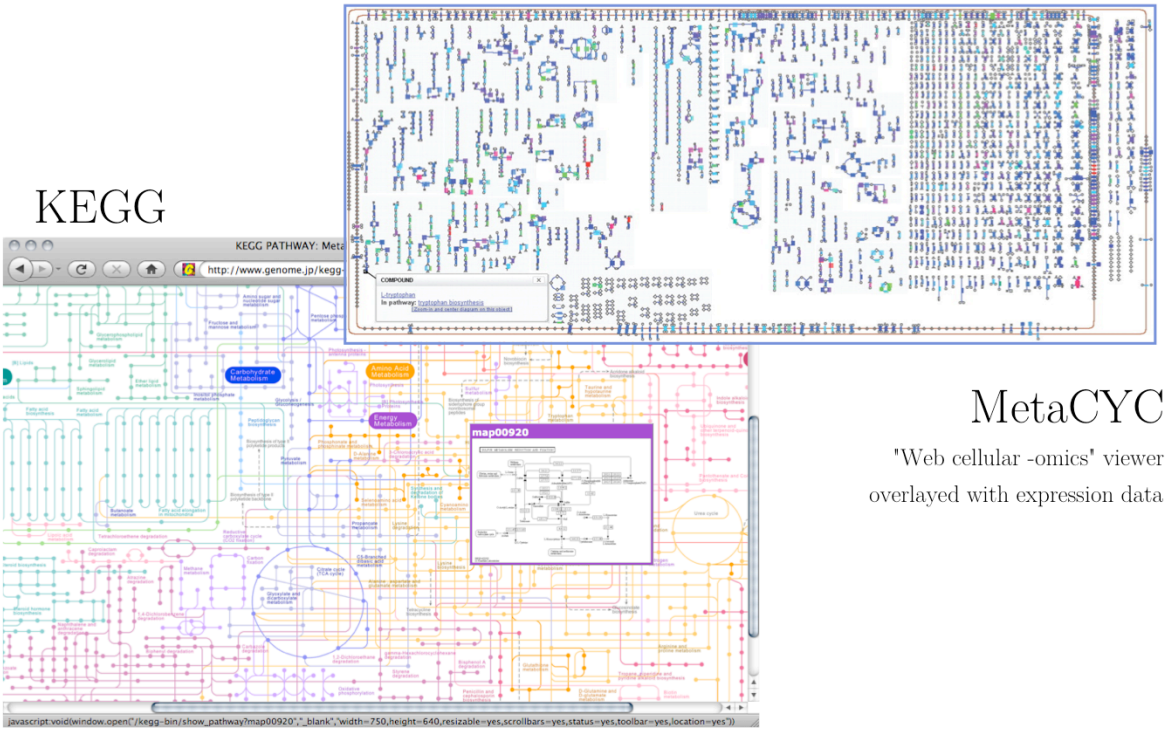


Annotation of individual genes: generalized function

To speak about function, we need a common language. A common language implies that we use the same terms to describe the same facts. To establish this for molecular biology is the goal of the Gene Ontology (GO) consortium.

The term "ontology" comes from the domain of *knowledge engineering*. An ontology collects terms that describe facts, their definitions, and their relationships.

Many more information resources have grown around GO, among them GOA – a collection of GO term annotations to genes in organisms, and GO-Slims – high levels of subsets of GO terms that give a broad overview of functional categories.

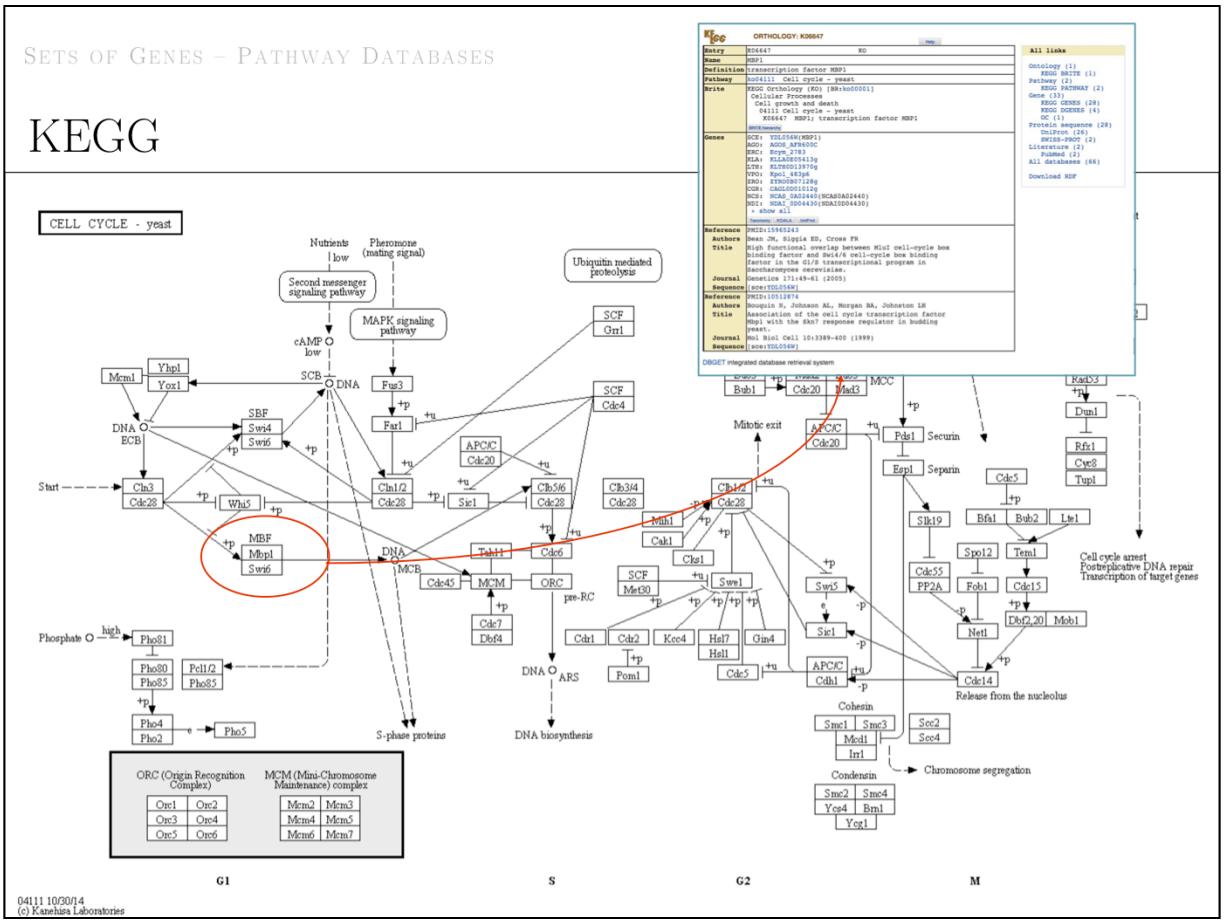


Annotation of gene sets: pathways

The notion of biochemical pathways also stems from the 1950s, when it was discovered that complex reactions via variety of intermediates could be catalyzed through a series of collaborating enzymes – even if the free energy of intermediates was not favourable for the reaction to progress.

Our current understanding of biochemical pathways is reasonably sophisticated, with active research focussing on the automated inference of pathways from individual activities, and on attempts to infer the metabolic repertoire of newly sequenced genomes through sequence homology.

Nevertheless, the original view of discrete pathways has been superseded in many aspects by reaction networks that allow parallel pathways and side-reactions, globally driven by sources and sinks on the level of the entire metabolome. “Flux balance analysis” is one computational approach that aims to elucidate the flow of metabolites through entire cells.



Annotation of gene sets: generalized pathways

Pathways are not just metabolic however: we have defined developmental pathways for tissues and organisms, and a large number of gene regulatory pathways for all aspects of cellular biology.

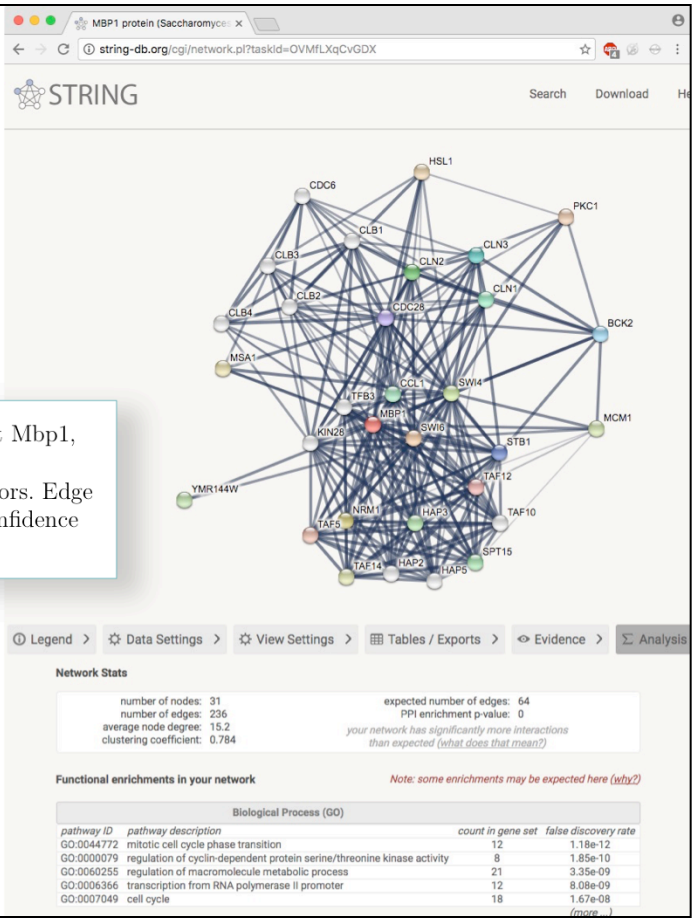
The "Kyoto Encyclopedia of Genes and Genomes" (KEGG), pioneered by Minoru Kanehisa is a carefully curated data resource that defines the relationships of gene products in a multitude of pathways and visualizes them in network diagrams that summarize key functional aspects.

NETWORKS – STRING

Superbly integrated tool for discovery of functional relationships from experimental data (yeast two-hybrid, TAP-tag MSMS, co-purification, genetic interactions), curated databases (e.g. KEGG pathways), coexpression, text-mining, gene-fusion, co-occurrence across species, and gene neighborhood.

Example search: yeast Mbp1, interactors and some interactors of interactors. Edge thickness indicates confidence level for interaction.

Lists functionally interacting proteins sorted by confidence score, with database cross-references and annotations. Calculates GO term enrichment, lists pathways that include the recovered proteins, calculates domain and feature enrichment etc. Results can be downloaded for further analysis.



Annotation of gene sets: genome scale functional networks

The STRING database makes it easy to build networks of potentially collaborating genes for function annotation.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA