WEB-BASED TOOLS AND REPOSITORIES

MICHAEL KOBAIVANOV
CONTENTS

• Tools:
  • Cytoscape - software platform for visualizing networks

• Repositories:
  • KEGG- Bioinformatics resource for deciphering the genome.
  • STRING- Search Tool for the Retrieval of Interacting Genes/Proteins
Yeast Protein–protein/Protein–DNA interaction network visualized by Cytoscape.
THE URGE FOR CYTOSCAPE

• Networks are pervasive in biology.

• Vast datasets are being gathered to describe networks of various types:
  • Networks of genetic information
  • Protein–protein interaction networks
  • Social networks
THE URGE FOR CYTOSCAPE

• Although various datasets can appear quite different in quality and quantity, they all are reflections of the same underlying biological system and its responses.

• Thus data integration, along with integrated visualization, is a key value for every researcher who deals with networks, and in particular biological networks.
INTRODUCTION

• Cytoscape, now in its 14th year of development, has become a standard tool for integrated analysis and visualization of biological networks.

• Its central organizing principle is a network graph, with biological entities (e.g. genes, proteins, cells, patients) represented as nodes and biological interactions represented as edges between nodes.
INTRODUCTION

• Data are integrated with the network using attributes, which map nodes or edges to specific data values such as gene expression levels or protein functions.

• Attribute values can be used to control visual aspects of nodes and edges (e.g. shape, color, size).

• Furthermore, it lets you perform complex network searches, filtering operations and other analysis.
HOW DOES IT LOOK?
SIGNIFICANT FEATURE - CUSTOM NODE GRAPHICS

• A key feature of Cytoscape is to allow diverse types of attribute data to be visualized on the nodes and edges of a biological network.

• Scalar data can be linked to simple visualization properties such as node color, shape or size as node and edge attributes.

Simple networks are shown with custom node images based on:
(A) pie chart displays
(B) line plots and bar charts
CUSTOM NODE GRAPHICS

• In the past, in order to represent multivariate data associated with a node through a custom graphical image, Cytoscape supplied a programming API.

• Version 2.8 introduces the ability for non-programmers to specify images through the Cytoscape GUI and to map these images to nodes using Cytoscope’s standard VizMapper interface.
BRIEF DEMONSTRATION
IN CONCLUSION-WHAT CAN YOU DO WITH CYTOSCAPE?

Cytoscape supports many use cases in molecular and systems biology, genomics, and proteomics:

• Load molecular and genetic interaction data sets.
• Visualize and analyze pathway datasets such as STRING, KEGG.
• Establish powerful visual mappings across these data.
• Perform advanced analysis and modeling (using Cytoscape Apps).
FINAL EXAMPLE
NICE TO KNOW

• Availability and implementation: Cytoscape is a desktop Java application released under the Library Gnu Public License (LGPL).

• Binary install bundles and source code for Cytoscape 2.8 are available for download from http://cytoscape.org

• Cytoscape also has a JavaScript-centric sister project named Sytoscape.js that can be used to analyze and visualize graphs in JavaScript environments, like a browser.
KEGG is a collection of databases dealing with:

- Genomes
- Biological pathways
- Diseases
- Drugs
- Chemical substances
• While the genome sequencing projects rapidly determine gene catalogs for an increasing number of organisms, functional annotation of individual genes is still largely incomplete.

• KEGG makes process of linking a set of genes in the genome with a network of interacting molecules in the cell, such as a pathway, representing a higher order biological function.
DATABASES

• KEGG consists of three main databases:
  • PATHWAY for representation of higher order functions in terms of the network of interacting molecules
  • Genomic information: GENOME for the collection of gene catalogs for all the completely sequenced genomes and GENES for genes and proteins in the complete genomes.
  • Chemical information: LIGAND for the collection of chemical compounds in the cell, enzyme molecules and enzymatic reactions. KEGG LIGAND consisted of three databases: KEGG COMPOUND for chemical compounds, KEGG REACTION for chemical reactions, and KEGG ENZYME for reactions in the enzyme nomenclature.
<table>
<thead>
<tr>
<th>Database</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>PATHWAY</td>
<td>2706 entries for pathway diagrams constructed from 143 manually drawn diagrams</td>
</tr>
<tr>
<td>GENES</td>
<td>110,018 entries in 24 complete genomes and 12 partial genomes</td>
</tr>
<tr>
<td>LIGAND</td>
<td>5645 entries in the COMPOUND section</td>
</tr>
<tr>
<td></td>
<td>3705 entries in the ENZYME section</td>
</tr>
<tr>
<td></td>
<td>5207 reactions in the REACTION section</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Auxiliary data</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ortholog group table</td>
<td>61 tables</td>
</tr>
<tr>
<td>Genome map</td>
<td>23 complete genomes and one partial genome</td>
</tr>
<tr>
<td>Comparative genome map</td>
<td>$23 \times 23$ complete genome comparisons</td>
</tr>
<tr>
<td>Expression map</td>
<td>Four sets of expression maps</td>
</tr>
<tr>
<td>Gene catalog</td>
<td>53 catalogs</td>
</tr>
<tr>
<td>Molecular catalog</td>
<td>Eight catalogs</td>
</tr>
<tr>
<td>Disease catalog</td>
<td>Three catalogs</td>
</tr>
</tbody>
</table>
### The Data Content of the KEGG Genes Entry

**Table 2.** The data content of the GENES database entry

<table>
<thead>
<tr>
<th>Field</th>
<th>Content</th>
<th>Links</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTRY</td>
<td>Entry identifier (gene accession number)</td>
<td>LinkDB database</td>
<td>GenBank or original database</td>
</tr>
<tr>
<td>NAME</td>
<td>Gene names and alternative names</td>
<td></td>
<td>GenBank or original database</td>
</tr>
<tr>
<td>DEFINITION</td>
<td>Annotation of gene function</td>
<td>LIGAND/ENZYME database, SWISS-PROT database and PubMed database</td>
<td>GenBank, original database, SWISS-PROT and KEGG</td>
</tr>
<tr>
<td>CLASS</td>
<td>Classification of genes according to the KEGG pathways</td>
<td>KEGG/PATHWAY database</td>
<td>KEGG</td>
</tr>
<tr>
<td>POSITION</td>
<td>Chromosomal position</td>
<td>KEGG/GENOME map</td>
<td>GenBank</td>
</tr>
<tr>
<td>DBLINKS</td>
<td>Outside links</td>
<td>Original databases and NCBI Entrez database</td>
<td></td>
</tr>
<tr>
<td>CODON_USAGE</td>
<td>Codon usage</td>
<td></td>
<td>Computed</td>
</tr>
<tr>
<td>AASEQ</td>
<td>Amino acid sequence</td>
<td>see footnote(^a)</td>
<td>GenBank or original database</td>
</tr>
<tr>
<td>NTSEQ</td>
<td>Nucleotide sequence</td>
<td></td>
<td>GenBank or original database</td>
</tr>
</tbody>
</table>

\(^a\)Computational links are available including sequence similarity searches (FASTA and BLAST), motif search (MOTIF), membrane protein predictions (SOSUI and TSEG), and cellular localization site prediction (PSORT).
EXAMPLE FROM KEGG PATHWAYS - GENETIC INFORMATION PROCESSING

- Transcription
- RNA POLYMERASE
**EXAMPLE FROM KEGG ENZYME - PEPSIN ENZYME**

Name: pepsin A  
Class: Hydrolases  
Comment: The predominant endopeptidase in the gastric juice of vertebrates, formed from pepsinogen A by limited proteolysis. Human pepsin A occurs in five molecular forms.

**Other databases:** ExplorEnz - The Enzyme Database; IUBMB Enzyme Nomenclature; ExPASy - ENZYME nomenclature database; BREND, the Enzyme Database.
Welcome to STRING

Protein-Protein Interaction Networks

ORGANISMS | PROTEINS | INTERACTIONS
---|---|---
2031 | 9.6 mio | 184 mio

SEARCH
INTRODUCTION

• STRING is a biological database and web resource of known and predicted protein – protein interactions.

• The STRING database contains information from numerous sources, including experimental data, computational prediction methods and public text collections. It is freely accessible and it is regularly updated.

• The latest version 10.0 contains information on about 9.6 millions proteins from more than 2000 organisms.
• Protein–protein interaction networks are an important ingredient for the system-level understanding of cellular processes.
• Such networks can be used for filtering and assessing functional genomics data and for providing an intuitive platform for annotating structural, functional and evolutionary properties of proteins.
• Exploring the predicted interaction networks can suggest new directions for future experimental research and provide cross-species predictions for efficient interaction mapping.
FEATURES

• A web interface is available to access the data and to give a fast overview of the proteins and their interactions.

• Predicted interactions are propagated to proteins in other organisms which share an ancestry.

• A plug-in for Sytoscape to use STRING data is available. Another possibility to access data STRING is to use API) by constructing a URL that contain the request.
• Asked information about trpB protein
STRING EXAMPLE
BIBLIOGRAPHY

• Cytoscape 2.8: new features for data integration and network visualization
  • Michael E. Smoot1,2,*, Keiichiro Ono1,2, Johannes Ruscheinski1,2, Peng-Liang Wang1,2 and Trey Ideker1,2
  • https://en.wikipedia.org/wiki/Cytoscape

• KEGG: Kyoto Encyclopedia of Genes and Genome
  • Minoru Kanehisa* and Susumu Goto
  • https://en.wikipedia.org/wiki/KEGG

• STRING 8—a global view on proteins and their functional interactions in 630 organisms
  • Lars J. Jensen1,2, Michael Kuhn1, Manuel Stark3, Samuel Chaffron3, Chris Creevey1, Jean Muller1, Tobias Doerks1, Philippe Julien4, Alexander Roth3, Milan Simonovic3, Peer Bork1,5,* and Christian von Mering3
  • https://en.wikipedia.org/wiki/STRING
QUESTIONS ?