Biological Pathways & Networks

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB
Main topics

- Biological pathways
  - KEGG & SEED & MetaCyc databases
  - Reactome
  - Pathway reconstruction
- Biological networks
  - PPI networks
  - Network analysis
- Biological network inference
  - Computational inference methods
Pathways versus networks

- “Many pathways have no real boundaries, and they often work together to accomplish tasks. When multiple biological pathways interact with each other, it is called a biological network.” (from http://www.genome.gov/27530687#al-3)
Biological pathways are essential to the understanding of biological functions.
Smaller units (e.g., KEGG pathways) are extremely important for the understanding of biological functions.
Pathways are often used to study the functionality encoded by a genome

Genome of an endosymbiont coupling N2 fixation to cellulolysis within protist cells in termite gut

Image from: http://www.sciencemag.org/cgi/content/full/322/5904/1108/DC1
KEGG pathway

- A collection of manually drawn pathway maps representing current knowledge on the molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes, and human disease.
- Functions represented by K numbers
- Mapping between K numbers and pathways
- Complete genomes: Eukaryotes: 299  Bacteria: 3010  Archaea: 175
- KEGG Metagenomes
  - Environmental: 84  (Ocean: 82, Microbial mat: 2)
KEGG Hierarchy

- Metabolism
- Genetic Information Processing
  - Transcription; translation; …
- Environmental Information Processing
  - Membrane transport; signal transduction; …
- Cellular Processes
  - Transport and catabolism; cell motility; cell growth and death..
- Organismal Systems
  - Immune system; endocrine system; …
- Human Diseases
  - Cancers; microRNAs in cancer; viral carcinogenesis…
- Drug development
Metabolism

1. Metabolism
1.1 Carbohydrate Metabolism
   Glycolysis / Gluconeogenesis
   Citrate cycle (TCA cycle)
   Pentose phosphate pathway
   Pentose and glucuronate interconversions
   Fructose and mannose metabolism

...
Main types of pathways

- Metabolic pathways
  - Metabolic pathways make possible the chemical reactions that occur in our bodies
- Gene regulation pathways
  - Gene regulation pathways turn genes on and off
- Signal transduction pathways
  - Signal transduction pathways move a signal from a cell's exterior to its interior
SEED subsystem

- A subsystem is a group of related functional roles jointly involved in a specific aspect of the cellular machinery.
- A subsystem includes annotations for “many” organisms
  - comparative analysis of genomes
- A subsystem is the sum of the pathways of all organisms under study

- [http://theseed.uchicago.edu/FIG/](http://theseed.uchicago.edu/FIG/) (58 archaeal, 868 bacterial and 29 eukaryal genomes are more-or-less complete)
How does subsystem work in SEED

1) A list of functional roles
2) Annotations in various species

Individual organisms

Subsystem
MetaCyc

- Database of nonredundant, experimentally elucidated metabolic pathways. MetaCyc contains more than 2260 pathways from more than 2600 different organisms (Dec 2014)
- Curated from the scientific experimental literature.
- Pathways involved in both primary and secondary metabolism

http://metacyc.org/

Snapshot of MetaCyc pathway ontology as of Dec, 2014
Reactome—a curated knowledgebase of biological pathways

- **Key data classes**
  - **PhysicalEntity** (individual molecules, multi-molecular complexes, and sets of molecules or complexes grouped together on the basis of shared characteristics)
  - **CatalystActivity** (molecular functions taken from the Gene Ontology molecular function controlled vocabulary to describe instances of biological catalysis.)
  - **Events** (the conversion of input entities to output entities in one or more steps, the building blocks used in Reactome to represent all biological processes)
Reactome: apoptosis

Apoptosis and disease: a life or death decision.

http://www.reactome.org/cgi-bin/eventbrowser?DB=gk_current&FOCUS_SPECIES=Homo%20sapiens&ID=109607&
Pathway reconstruction

- We have pathway annotation for reference genomes (which are not necessarily perfect)
- When a new genome arrives, we first annotate the functions of the encoded genes
- Then try to figure out what are the possible pathways encoded by the genome
A simple pathway reconstruction approach

List of functions

$f_1$ $f_2$ $f_3$ $f_4$ $f$ $f_5$ $f_6$

mapping

$p_1$ $p_2$ $p_3$ $p_4$

List of pathways
Protein-protein interaction (PPI)

Nodes: proteins
Links: physical interactions

(Jeong et al., 2001)
Experimental methods for PPI detection

- Yeast two-hybrid
- Proteome chips
- Tagged Fusion Proteins
- Coimmunoprecipitation
- X-ray Diffraction
- …
PPI databases

- Many databases
- DIP
  - Established in 1999 in UCLA
  - extract and integrate protein-protein info and build a user-friendly environment
- BIND
STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database covers 5'214'234 proteins from 1133 organisms (the numbers were 2,590,259 proteins from 630 organisms back in Nov 2009).
Graph theory

- Modeling real-world phenomena, e.g. World Wide Web, electronic circuits, collaborations between scientists, co-citations, biological networks, etc.
- Global properties: e.g. diameter, clustering, degree distribution
- Local properties: vertex density, motif and graphlet
Topological analysis

- **Definitions**
  - **Graph**
    
    \[ G(V, E) \]
    
    *V*: vertex set  
    *E*: edge set
  
  \(|V|, |E|: sizes\)

  - **Vertex (or Node)**
    
    Degree: number of edges connected to the vertex.

  - **Edge**

- e.g.
  
  \(|V| = 4\)  
  \(|E| = 6\)
Topological analysis

- **Degree distribution** $P(k)$
  - the probability of a vertex has degree of $k$.
  - power law:

$$P(k) \sim k^{-\gamma}$$

- **Diameter (length)**
  - the shortest path from one vertex to another
Topological analysis

- **Clustering coefficient** \((C)\)
  \[ C_i = \frac{2e_i}{k_i^\ast(k_i - 1)} \]
  - \(e_i\): # of edges between neighbors of vertex \(i\)
  - \(k_i\): # of neighboring vertices of \(i\)
  - \(i\) not included in both

- **Vertex density** \((D)\)
  - Same as \(C\) but includes \(i\)
Analysis of biological networks (what can networks tell us?)

- **Scale-free**
  - Degree distribution follows a power law of the form $P(k) \sim k^{-\gamma}$.
  - Robustness and fragility (Hub proteins)

- **Small-world networks**
  - Small world network lies between two extremes of graph, completely regular and completely random graph.
  - Regular networks have long path lengths, and are clustered, while random graphs have short path length but show little clustering
  - Small-world networks have short path lengths but highly clustered.
Identify modules from biological networks

- Modules: highly connected clusters
- A “module” in a biological system is a discrete unit whose function is separable from those of other modules
- Identifying functional modules and their relationship from biological networks will help to the understanding of the organization, evolution and interaction of the cellular systems they represent
Biological network inference

- A network is a set of nodes and a set of directed or undirected edges between the nodes

- Transcriptional regulatory networks.
  - Genes are the nodes and the edges are directed
  - Primary input: gene expression data (e.g., microarray data, and now RNA-seq)

- Signal transduction network
  - Proteins are the nodes and the edges are directed
  - Primary input: experiments measuring protein activation / inactivation

- Metabolite network
  - Metabolites are the nodes and the edges are directed.
  - Primary input: measurements of metabolite levels
How to infer gene/protein connectivity

- **Clustering approaches**
  - Cluster analysis and display of genome-wide expression patterns, PNAS, 98
  - Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, PNAS, 99
  - Genetic network inference: from co-expression clustering to reverse engineering, Bioinformatics, 2000

- **Information theory methods**
  - Reverse engineering of regulatory networks in human B cells, Nature Genetics, 2005

- **Bayesian methods**
  - Advances to bayesian network inference for generating causal networks from observational biological data, Bioinformatics, 2004
Protein–protein interaction networks: how can a hub protein bind so many different partners?

- Multiple binding sites
- Flexibility
- Disorder proteins
- Big size (larger proteins)
- Incorporation of time into the networks (‘date’ and ‘party’ hub proteins)
- ...
- Still limited
- Tsai et al said this problem actually does not even exist (Trends in Biochemical Sciences, 2009)
p53 is one of the most connected nodes in either the protein–protein interaction network or the gene regulation network; protein products derived from a single gene may involve many interactions!
Network visualization (and analysis)

http://www.cytoscape.org/
Cytoscape network analysis workflow

Image source: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3649846/figure/F2/
Systems biology does not investigate individual genes or proteins one at a time, as has been the highly successful mode of biology for the past 30 years. Rather, it investigates the behavior and relationships of all of the elements in a particular biological system while it is functioning. These data can then be integrated, graphically displayed, and ultimately modeled computationally.

Image source: http://idekerlab.ucsd.edu/research/Pages/systems-biology.aspx
A mathematical model of the unfolded protein stress response reveals the decision mechanism for recovery, adaptation and apoptosis (three states).