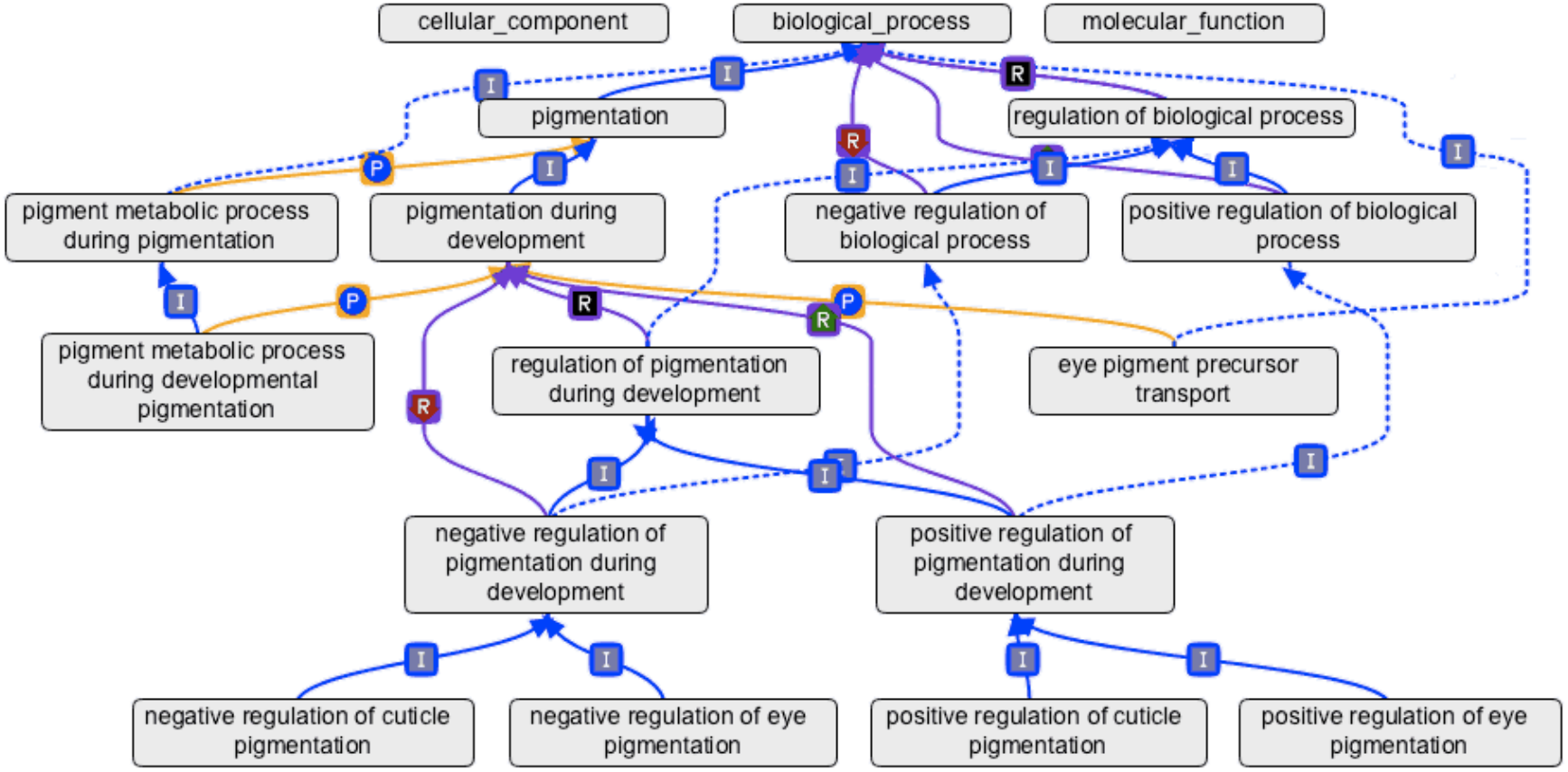
# Gene Ontology

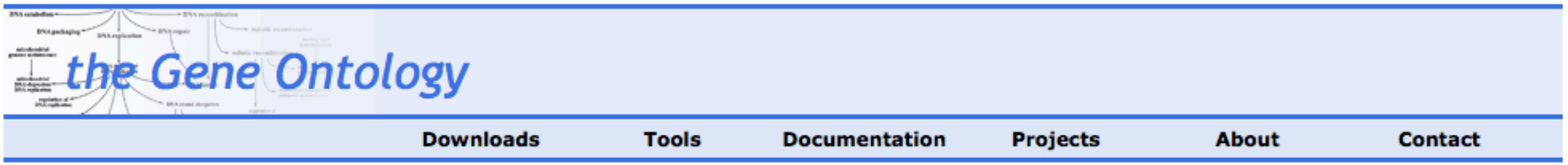
- "The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases."
- "The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner."
  - "A gene product might be associated with or located in one or more cellular components; it is active in one or more biological processes, during which it performs one or more molecular functions."

# GO is a directed acyclic graph



# GO to Gene association

- Gene can be associated to a GO term at any level.
  - The depth in the tree can be thought of as to resolution of what we know about the gene function
  - Since GO terms have a hierarchical structure, we can assume that the gene is also associated to the parents of the GO terms it is directly associated to.
- Gene can be associated to multiple GO terms.




## Gene Ontology site: search for BRCA1

http://geneontology.org/

# All results for BRCA1

Select all | Clear all | Perform an action with this page's selected gene products... | Go!

| rel ↓ | Symbol , full name | | Species |
|---|---|---|---|
| ☐ | **brca1**<br>Breast and ovarian cancer susceptibility protein | 1 association<br>BLAST | **protein** from *Xenopus laevis* |
| ☐ | **Brca1**<br>breast cancer 1 | 65 associations | **protein** from *Mus musculus* |
| ☐ | **Brca1**<br>breast cancer 1 | 66 associations<br>BLAST | **gene** from *Rattus norvegicus* |
| ☐ | **BRCA1**<br>Breast cancer 1, early onset | 21 associations<br>BLAST | **protein** from *Homo sapiens* |
| ☐ | **BRCA1**<br>Breast cancer 1, early onset | 22 associations<br>BLAST | **protein** from *Homo sapiens* |
| ☐ | **BRCA1**<br>Breast cancer 1, early onset | 22 associations<br>BLAST | **protein** from *Homo sapiens* |

# GO-terms associated to BRCA1

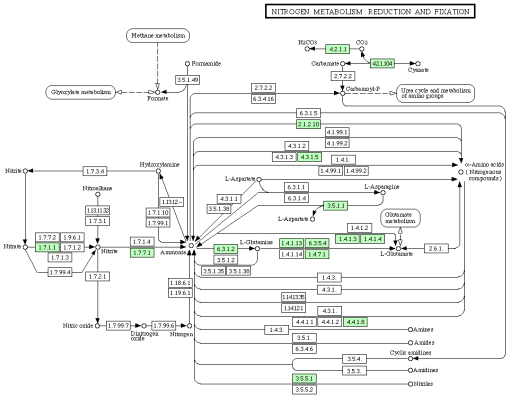
Select all | Clear all | Perform an action with this page's selected terms... | Go!

| Accession, Term | | Ontology |
|---|---|---|
| GO:0007420 : brain development | 3148 gene products<br>view in tree | biological process |
| GO:0007098 : centrosome cycle | 371 gene products<br>view in tree | biological process |
| GO:0043009 : chordate embryonic development | 3417 gene products<br>view in tree | biological process |
| GO:0006281 : DNA repair | 6229 gene products<br>view in tree | biological process |
| GO:0006260 : DNA replication | 5402 gene products<br>view in tree | biological |

# A tree view showing GO-term hierarchy

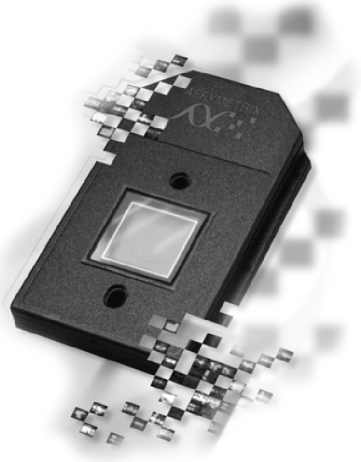
- all : all [556462 gene products]
  - GO:0008150 : biological_process [423988 gene products]
    - GO:0009987 : cellular process [251922 gene products]
      - GO:0044237 : cellular metabolic process [178145 gene products]
        - GO:0044249 : cellular biosynthetic process [97301 gene products]
          - GO:0034645 : cellular macromolecule biosynthetic process [66250 gene products]
            - GO:0006260 : DNA replication [5402 gene products]
        - GO:0044260 : cellular macromolecule metabolic process [112413 gene products]
          - GO:0034645 : cellular macromolecule biosynthetic process [66250 gene products]
            - GO:0006260 : DNA replication [5402 gene products]
          - GO:0006259 : DNA metabolic process [17561 gene products]
            - GO:0006260 : DNA replication [5402 gene products]
        - GO:0034641 : cellular nitrogen compound metabolic process [97222 gene products]
          - GO:0006139 : nucleobase-containing compound metabolic process [79473 gene products]
            - GO:0090304 : nucleic acid metabolic process [62644 gene products]
              - GO:0006259 : DNA metabolic process [17561 gene products]
                - GO:0006260 : DNA replication [5402 gene products]

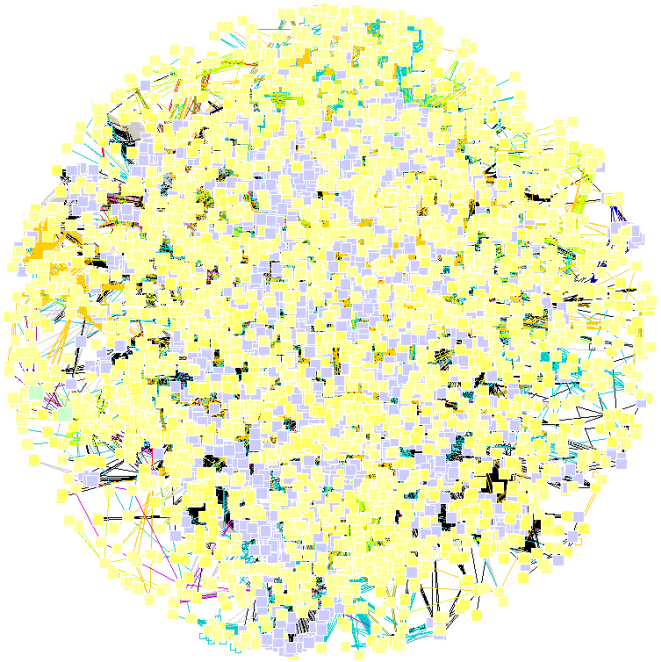
Metabolic pathways
(KEGG, ARACYC)




Regulatory
interactions
(AGRIS, Transfac)

Transcriptome
(AFFYMETRIX)





BIND
Homology based
protein-protein
interactions

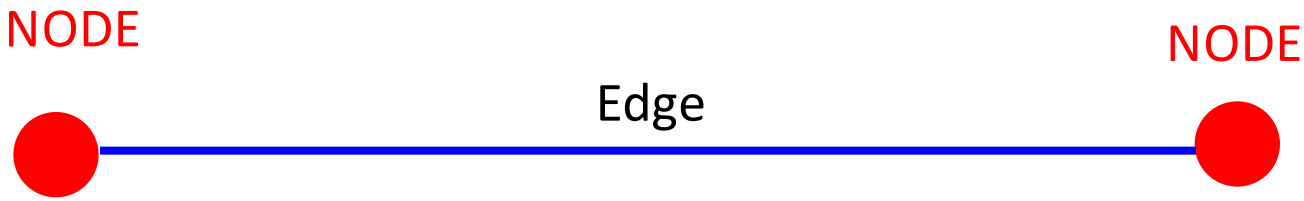
miRNA:RNA
interactions

Integrated Network

Literature
based
interactions

# Interactions between parts can be represented by Networks

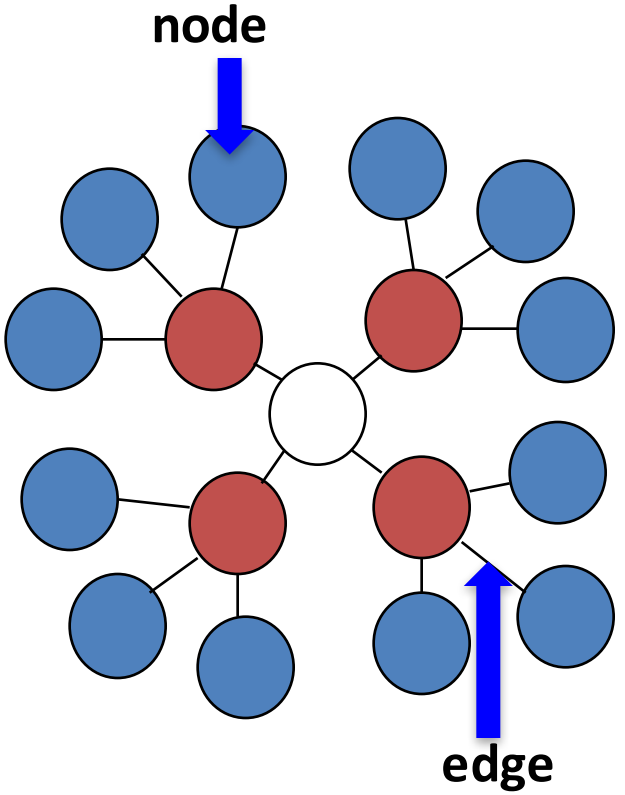
A network is made up of "nodes" and "edges" that connect them



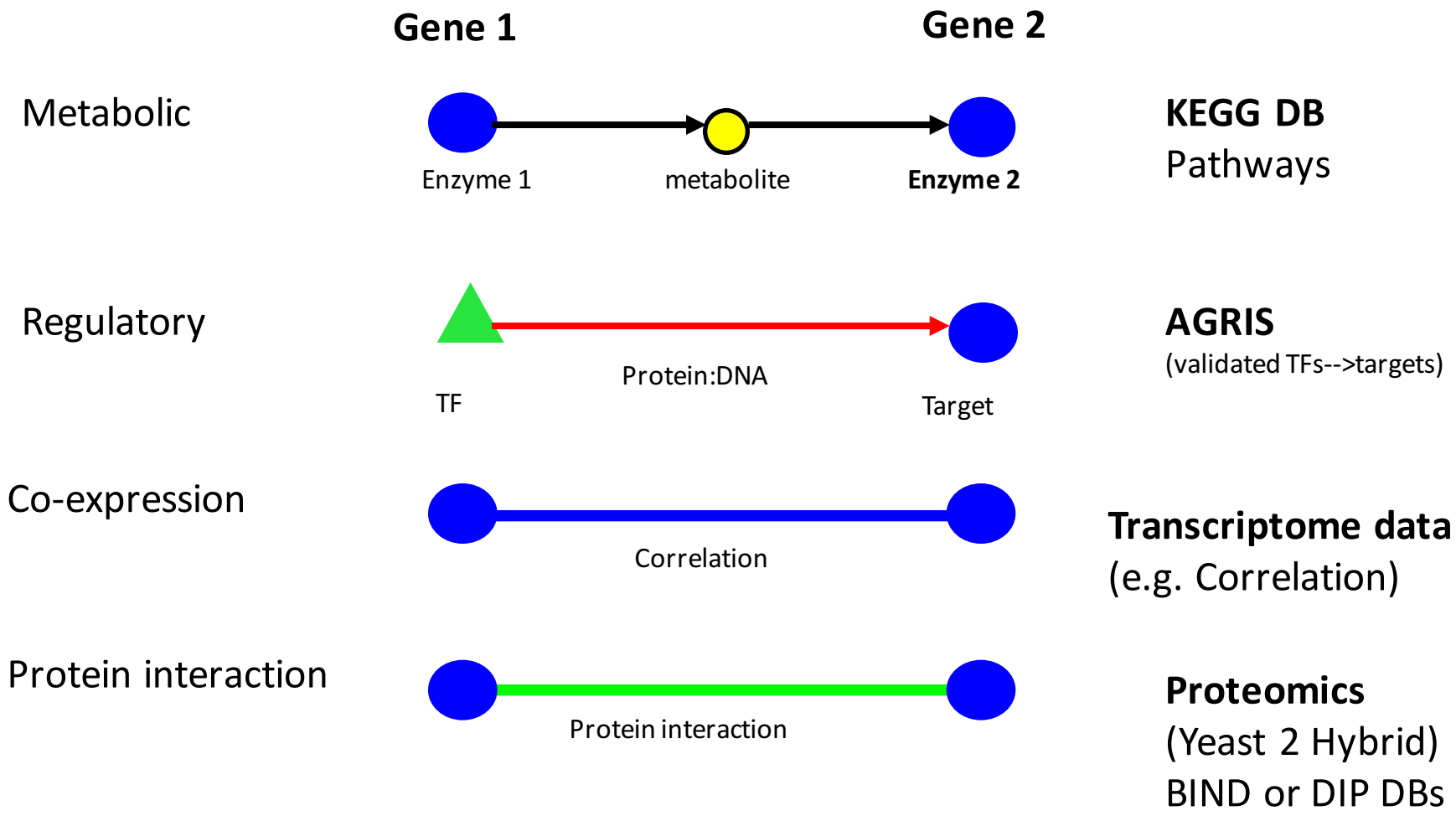

| Network | NODE | Edge | NODE |
|---|---|---|---|
| Social | You | Social interaction | Neighbor |
| Cities | New York | Highway connection | Chicago |
| **Gene** | **Tx Factor** | **Tx activation** | **Target Gene** |

# Network Lingo

- **A collection of nodes (vertices):**
  - Genes, proteins, metabolites
- **Connected by edges (links):**
  - Enzymatic reactions
  - Protein:DNA interactions
  - Protein:Protein Interactions
  - miRNA:RNA interactions
- **Displayed as a network graph**



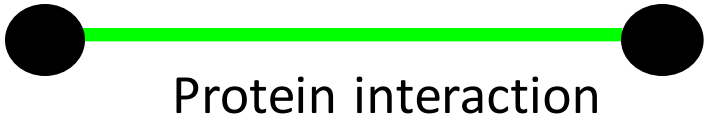
# Different types of "edges" can be used to connect genes in a Network



# Types of Edges (connections between nodes)

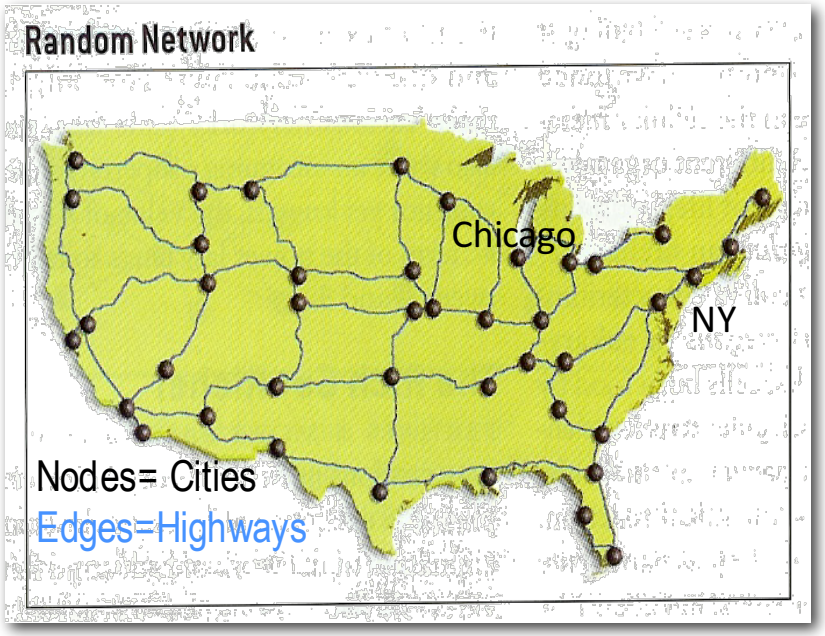
- **Directed**:
  - edges have a direction, only go one way
  - e.g. protein:DNA interactions, some enzymatic Rx

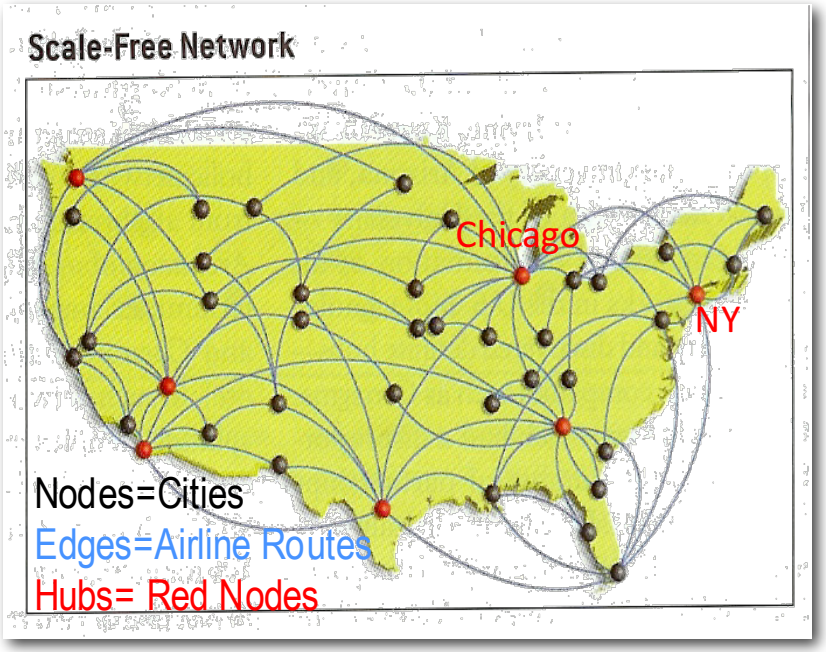

- **Undirected**:
  - no direction
  - e.g. protein:protein interactions

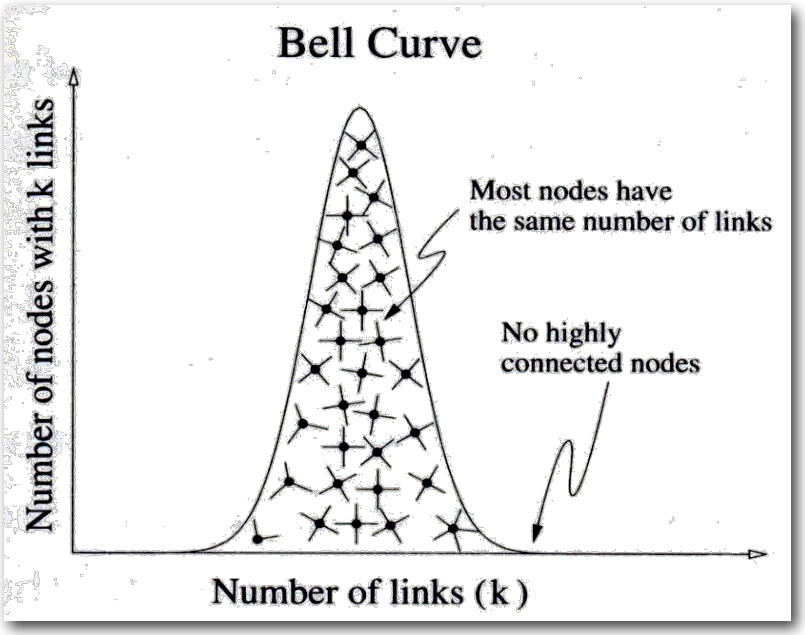


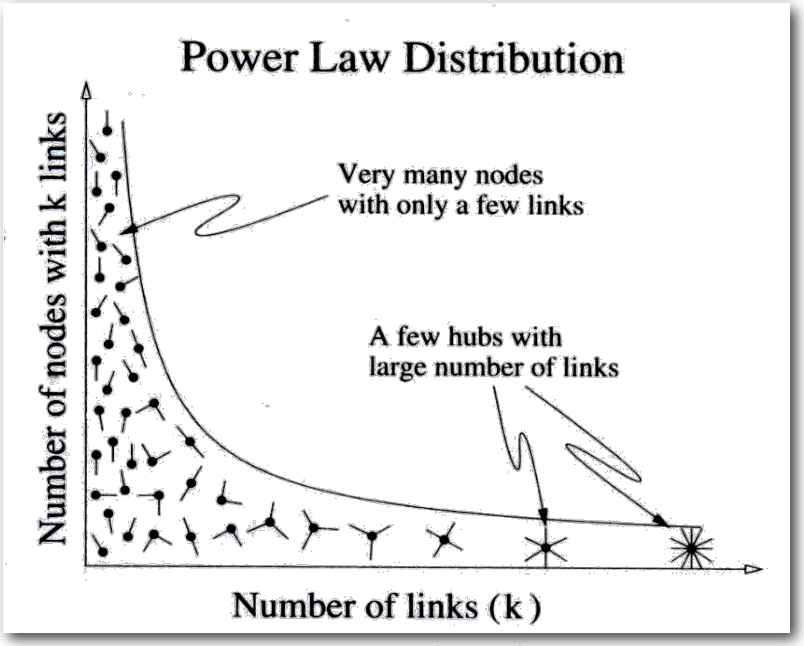

- **Weighted**
  - Not all edges are equal in value
  - e.g. weight the strength of edge connections
  - (e.g. based on correlation of gene expressio

# Significance of Scale-free networks and network hubs…



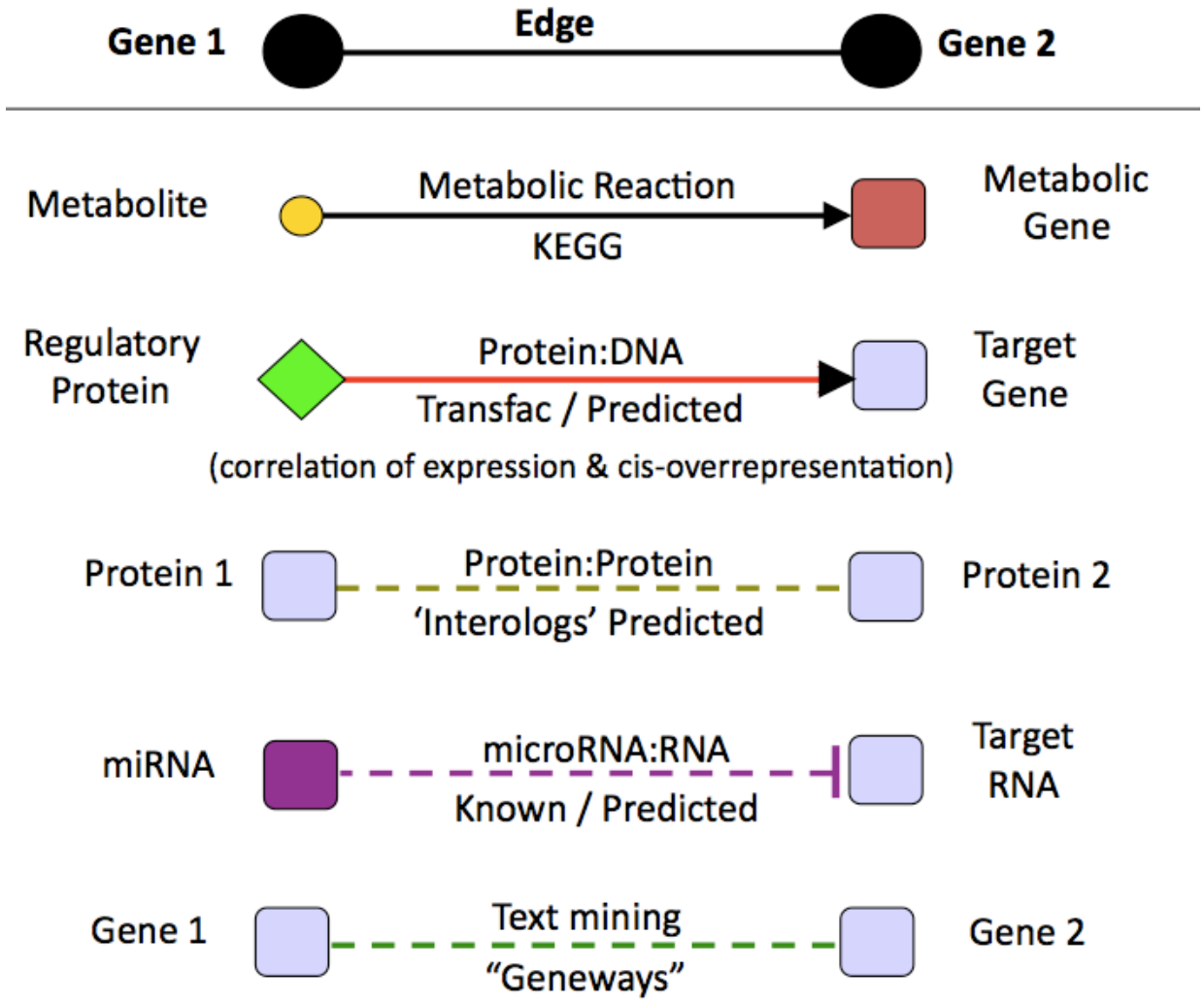




**Barabasi & Bonabeau (2003)**

# **Multinetwork**: Multiple types of EDGES link genes

**Gutiérrez et al. (2007)**
**Genome Biology**



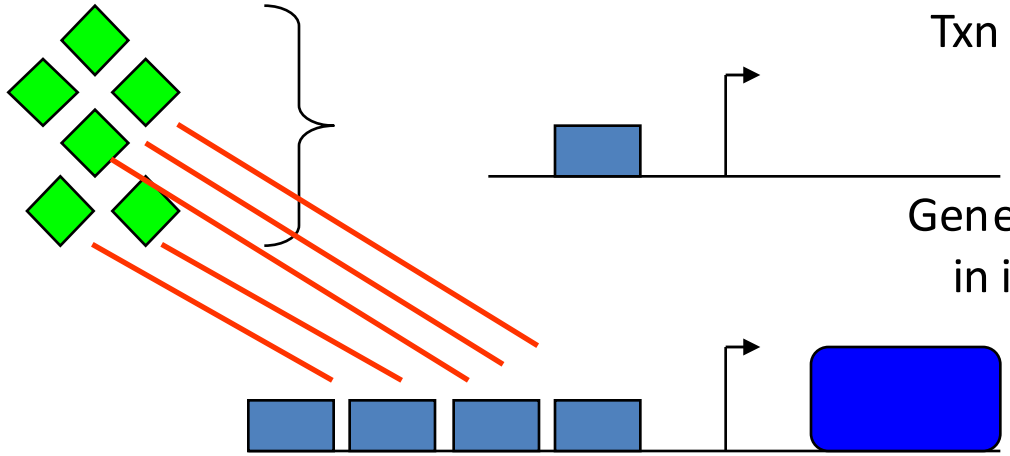

Green & Meyers
(U. Delaware)

A. Rhetszky
(U. Chicago.)

Empirically: Multiple layers of circumstantial evidence for genes that function together in vivo

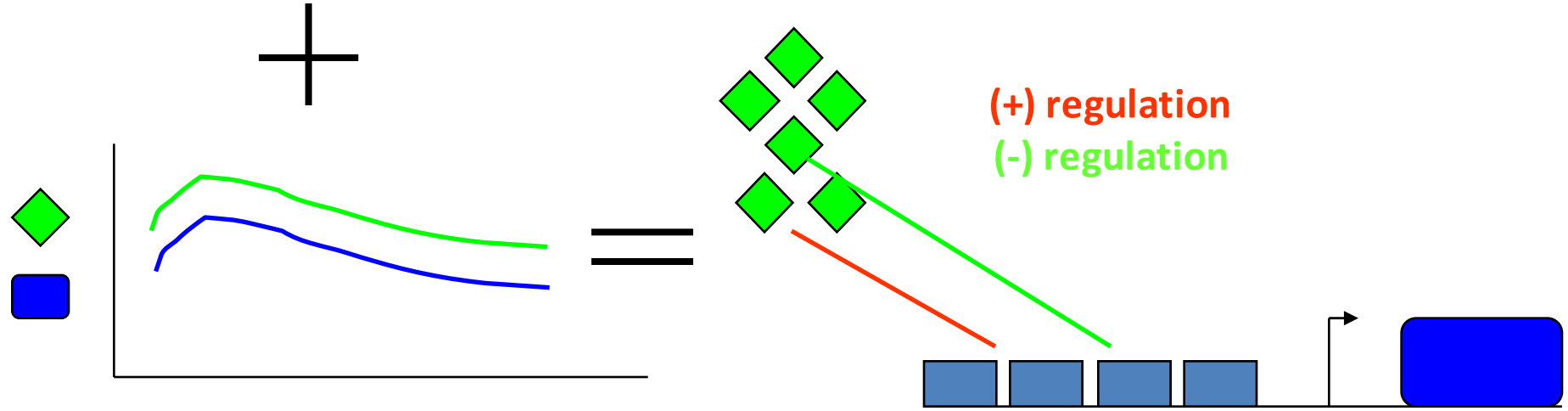
# Predicted Regulatory Interactions using Expression Data and functional cis-regulatory elments.

Damion Nero et al.



Txn Family associated to binding site
Data taken from AGRIS

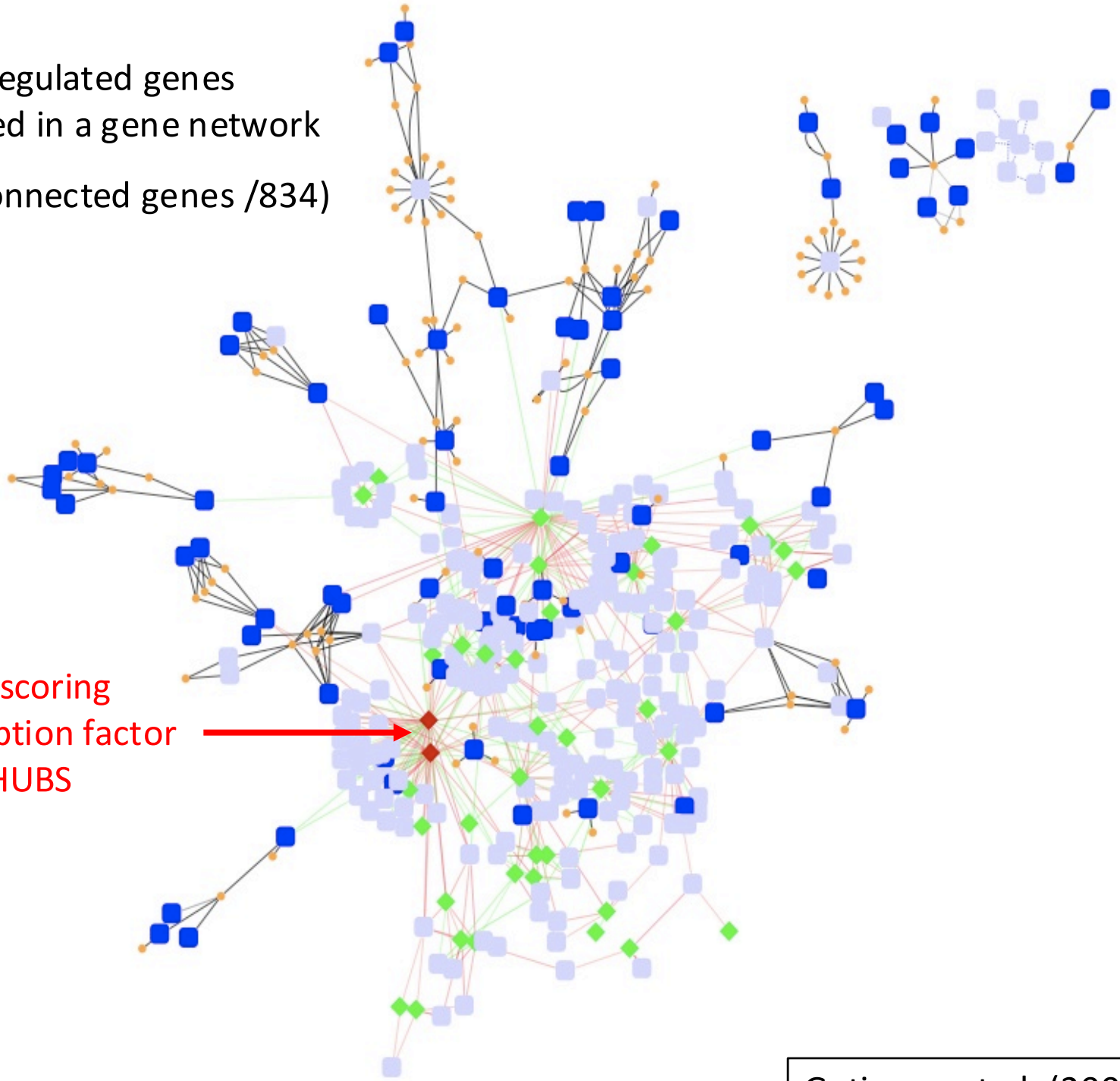
Gene with **over-representation** of motif in its promoter region is considered putative target




Consider only TxnF and target pair if they are **significantly correlated**.

N-regulated genes
connected in a gene network

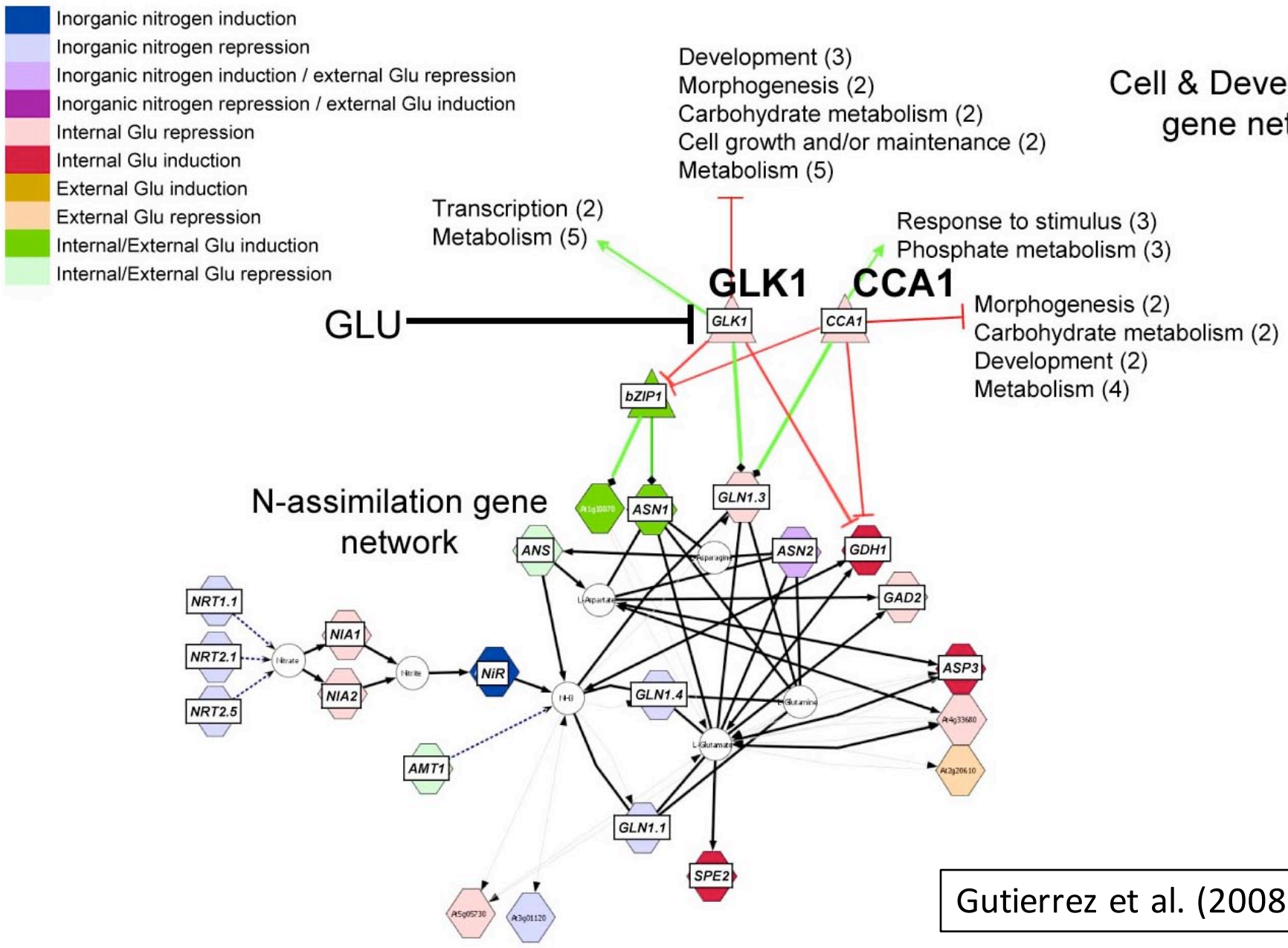
(369 connected genes /834)



Top scoring
Transcription factor
HUBS

Gutierrez et al. (2008) PNAS

# N-regulation of CCA1 controls a nitrogen responsive gene regulatory network




Gutierrez et al. (2008) PNAS

# Cluster Coefficient

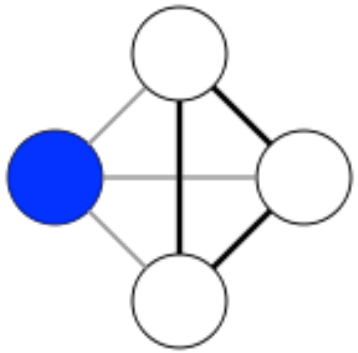
- How close are the neighbors ?

- 2 times the number of triangles (connected neighbors) divided by degree times degree – 1.

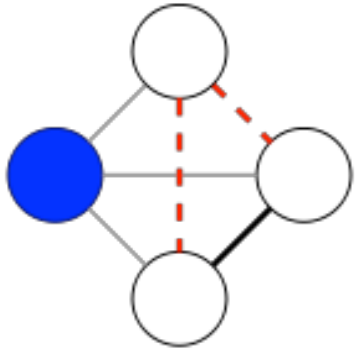
$$c_v = \frac{2T(v)}{deg(v)(deg(v)-1)}$$

http://networkx.lanl.gov/

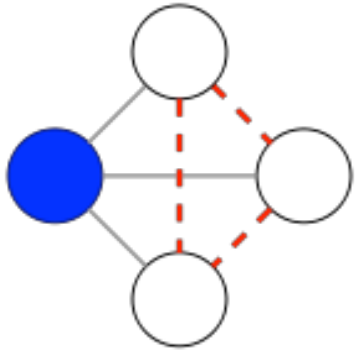
http://en.wikipedia.org/wiki/Clustering_coefficient



c = 1



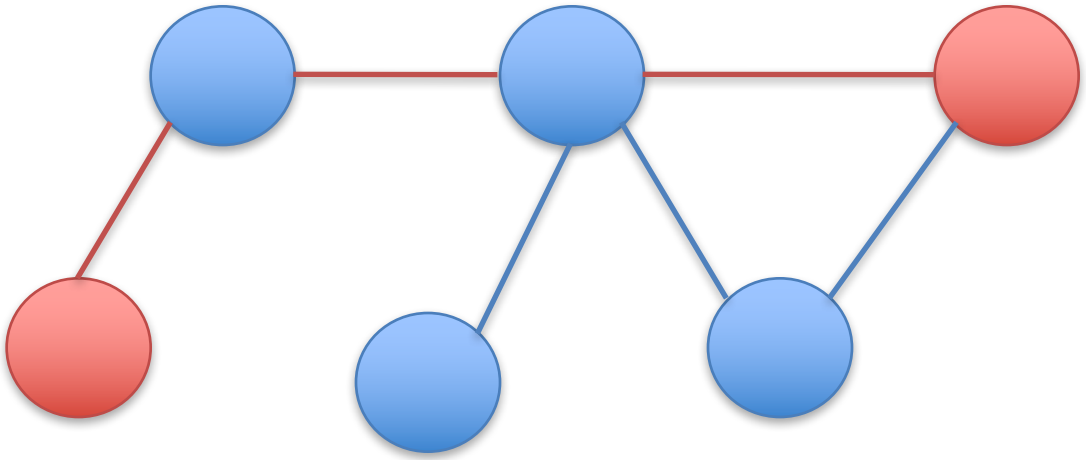
c = 1/3



c = 0

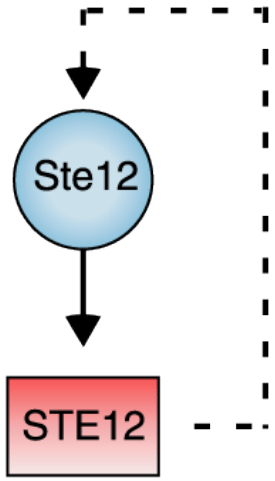
# Shortest Path

- Quickest way to get from one node to another.
- Diameter of a graph is the longest shortest path between any two nodes in the graph

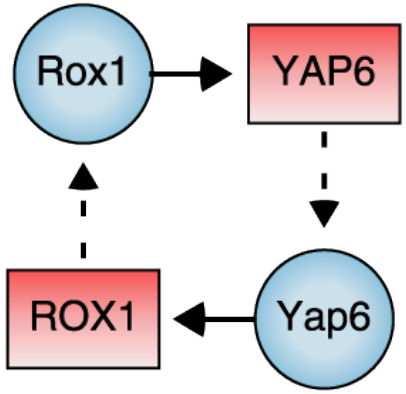


# Examples of network motifs

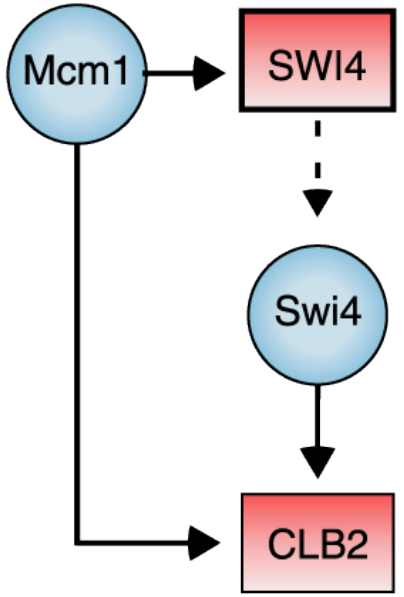
Autoregulation




Multicomponent Loop

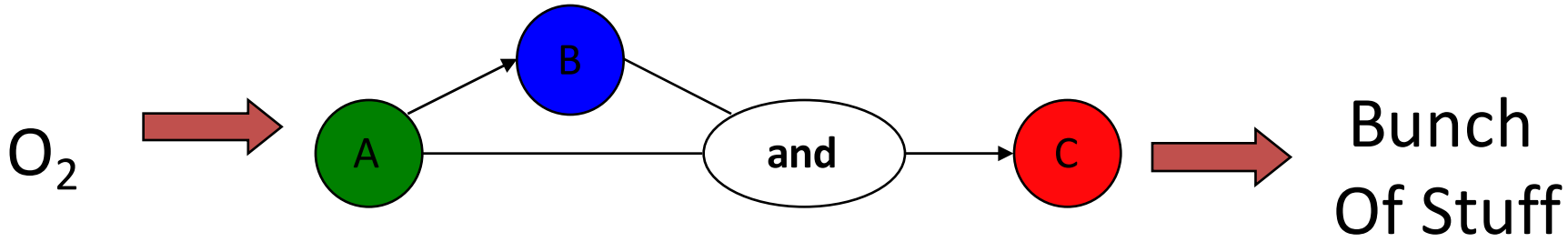



Feedforward Loop

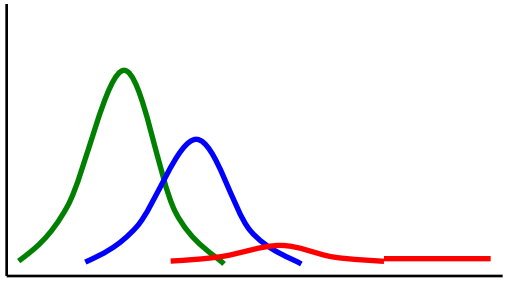


# Feed-forward loop: example

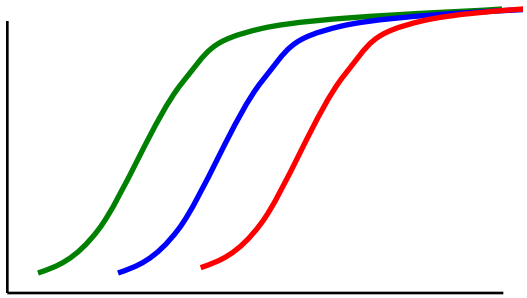
Switch from anaerobic to aerobic metabolism => requires ability to detect ***sustained*** $O_2$ availability (see paper on local motifs in *E. coli* by Uri Alon et al.)




Transient $O_2$ exposure

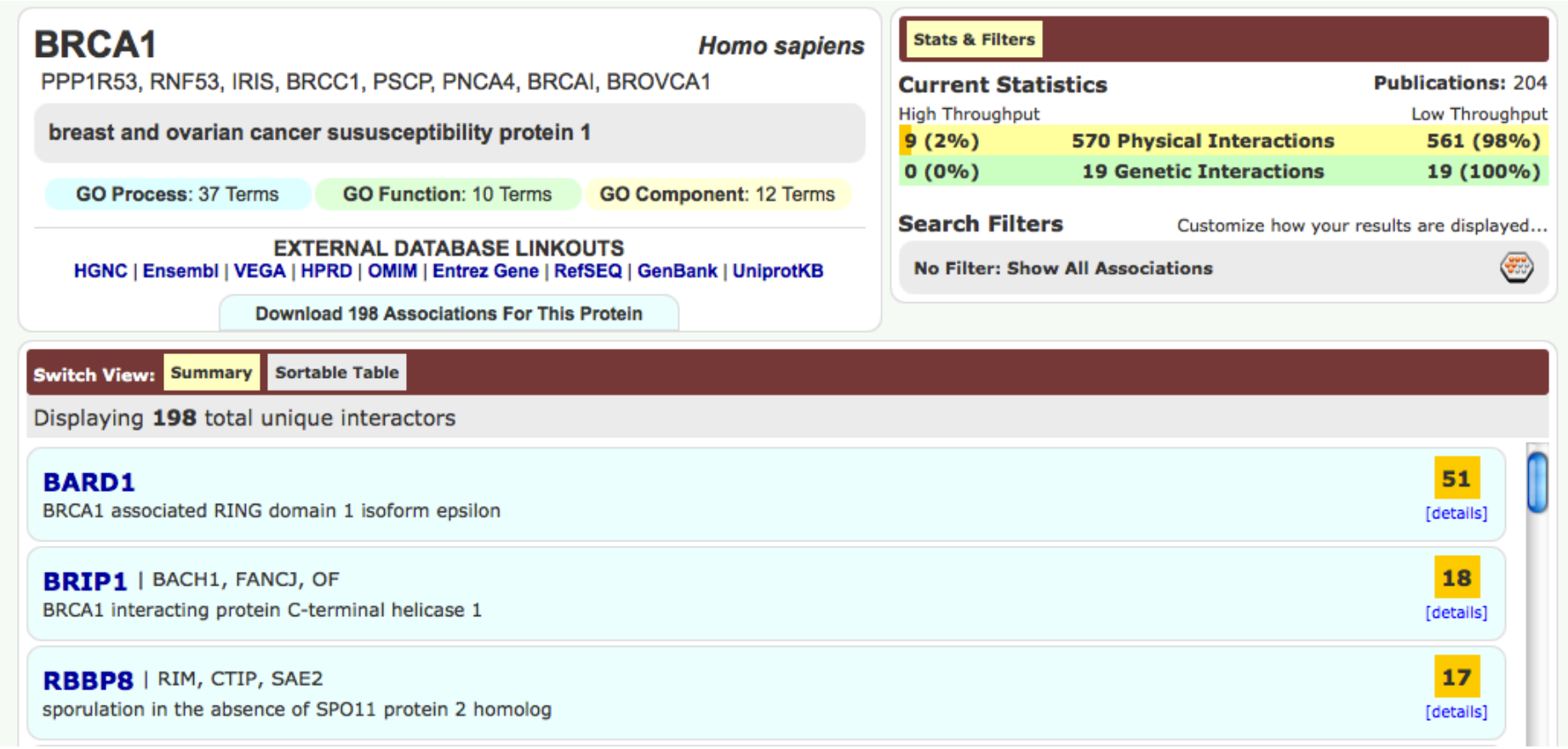


Sustained $O_2$ exposure

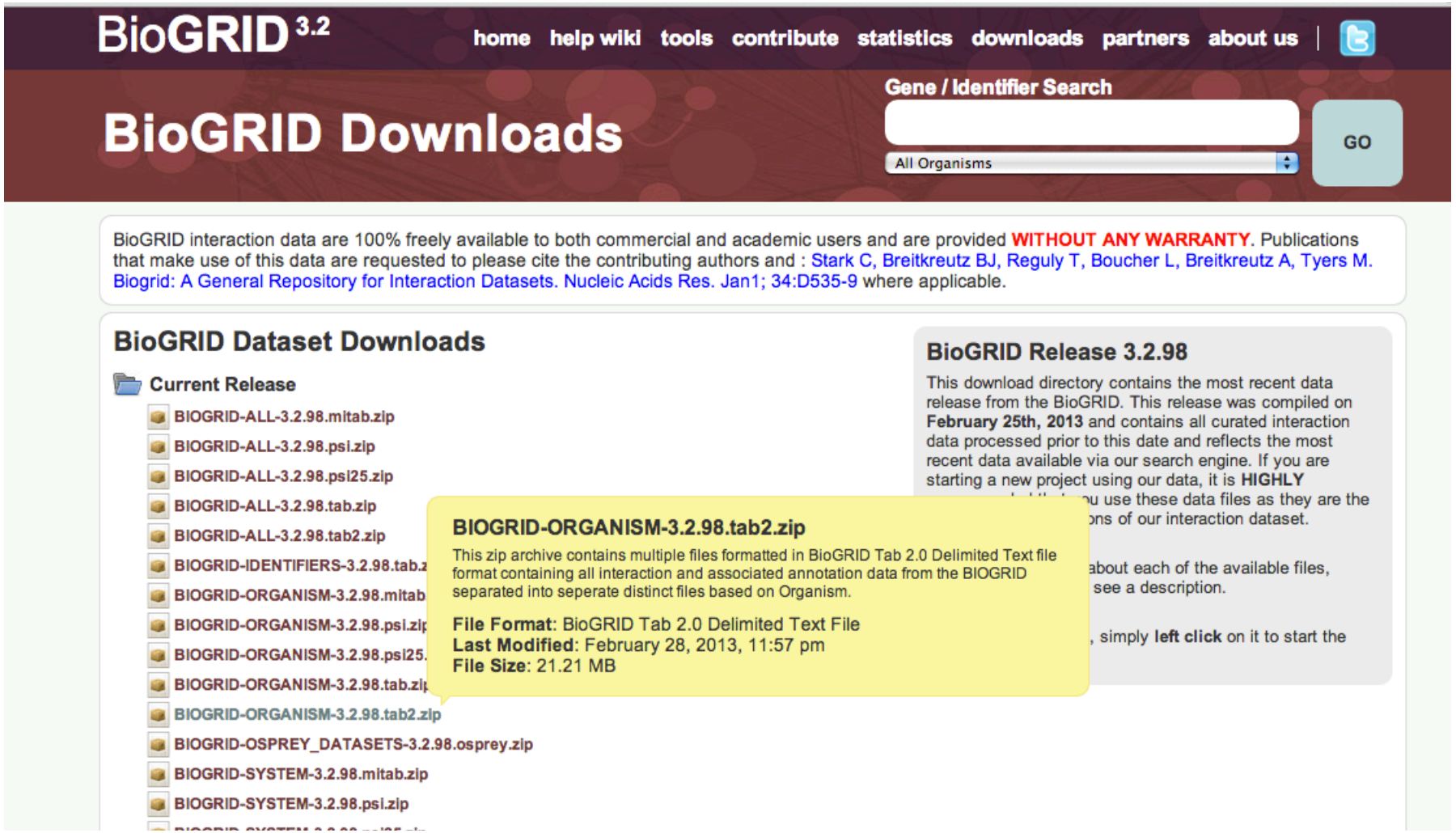


# BIOGRID: searching for BRCA1 interactions

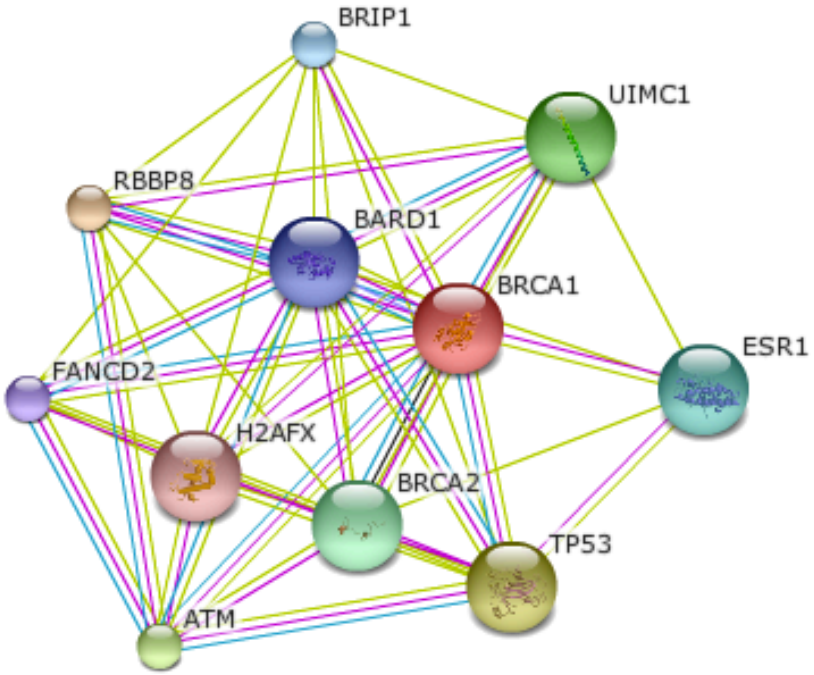
- http://thebiogrid.org



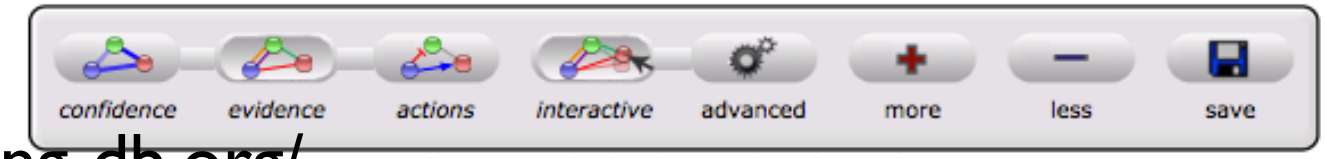
# Download interactions



# String: searching for BRCA1 interactions




This is the **evidence view**. Different line colors represent the types of evidence for the association.




(requires Flash player 10 or better)

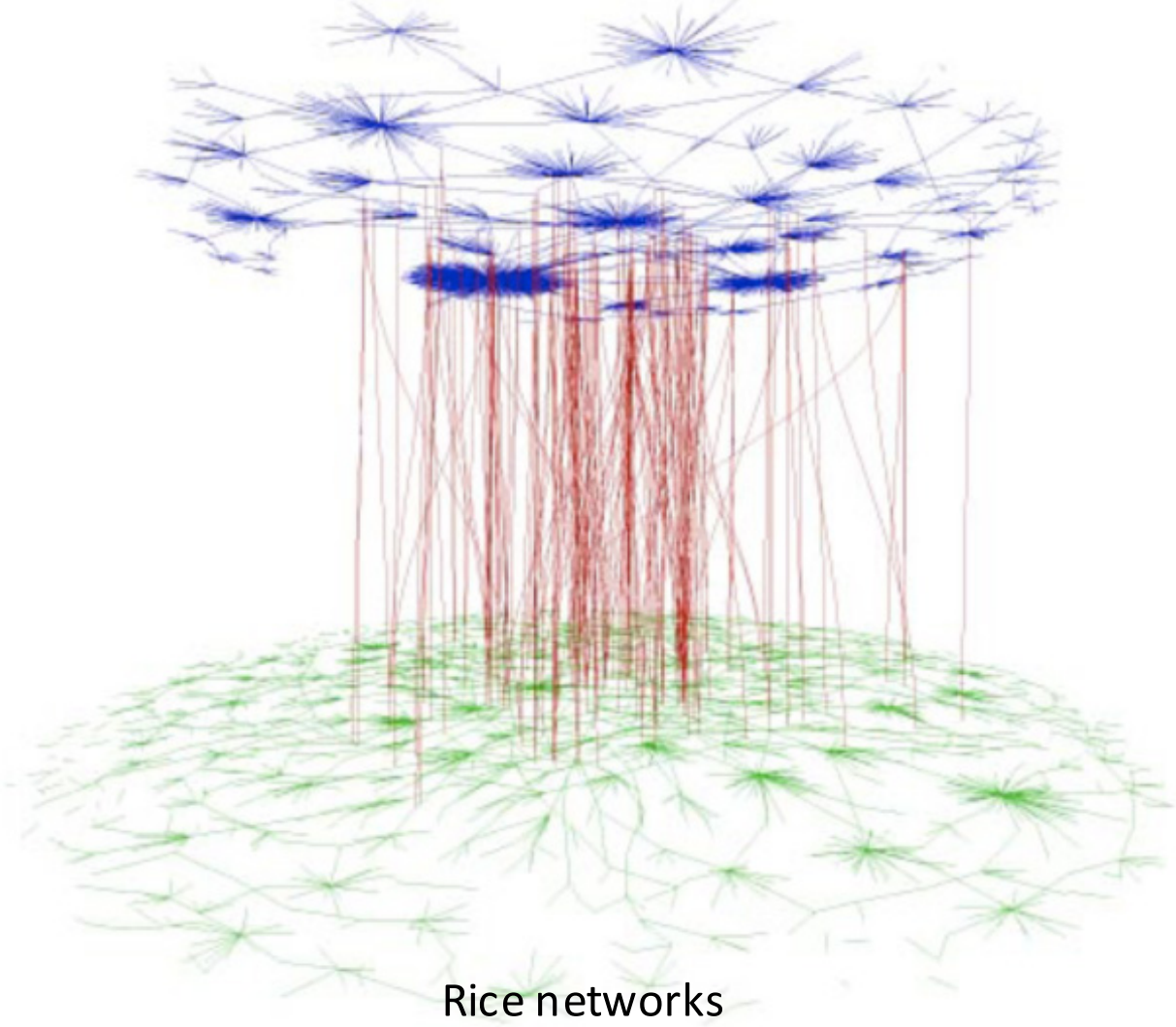
- http://string-db.org/

# Comparative Network Databases

Manpreet S. Katari

# Next frontier:
# Orthologous networks

Arabidopsis networks



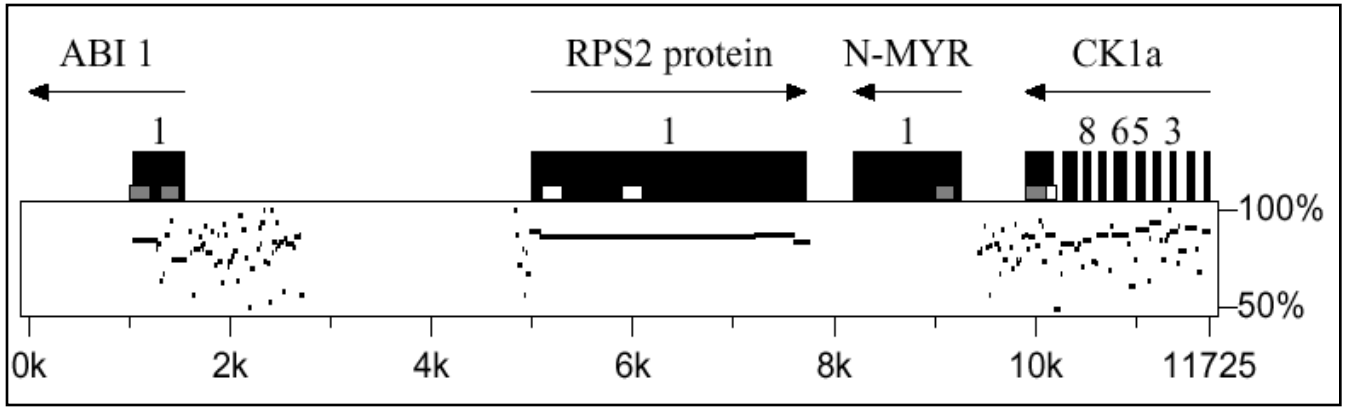
Rice networks

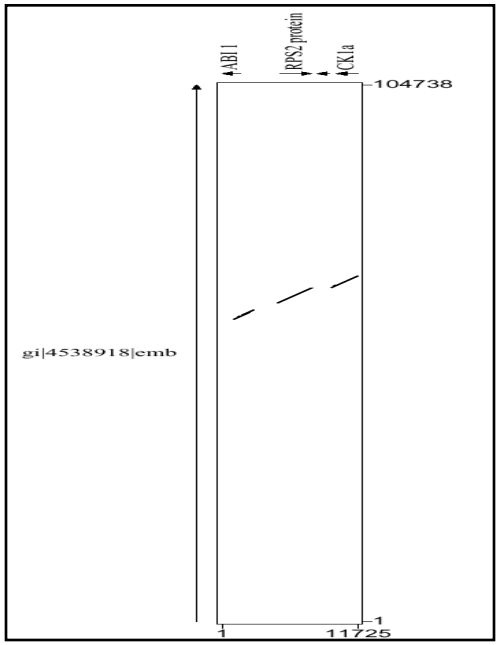
Adapted from Imperial College of London 2008

# Why compare genomes?

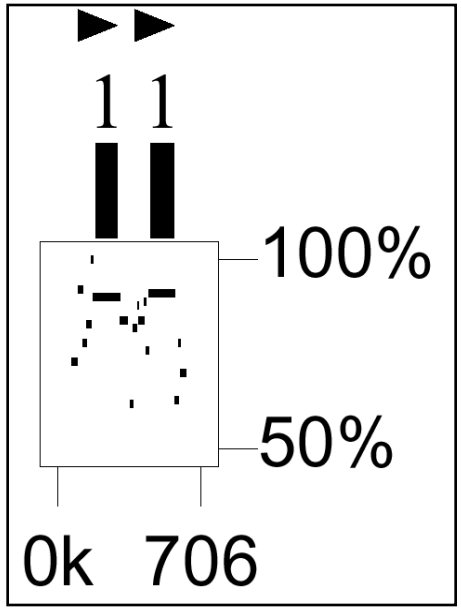
- Improve genome annotation
  - Purifying selection – Conserved sequences between species are biologically significant.
    - Genes (Coding/Non-Coding) Regions
    - Regulatory Region
- Predict gene function
  - Conserved protein motifs may suggest conserved function.
- Depending on the biological question it may be favorable to compare distant species rather then close species and vice versa.

# Conservation of Coding Regions Identified in PiP Plots



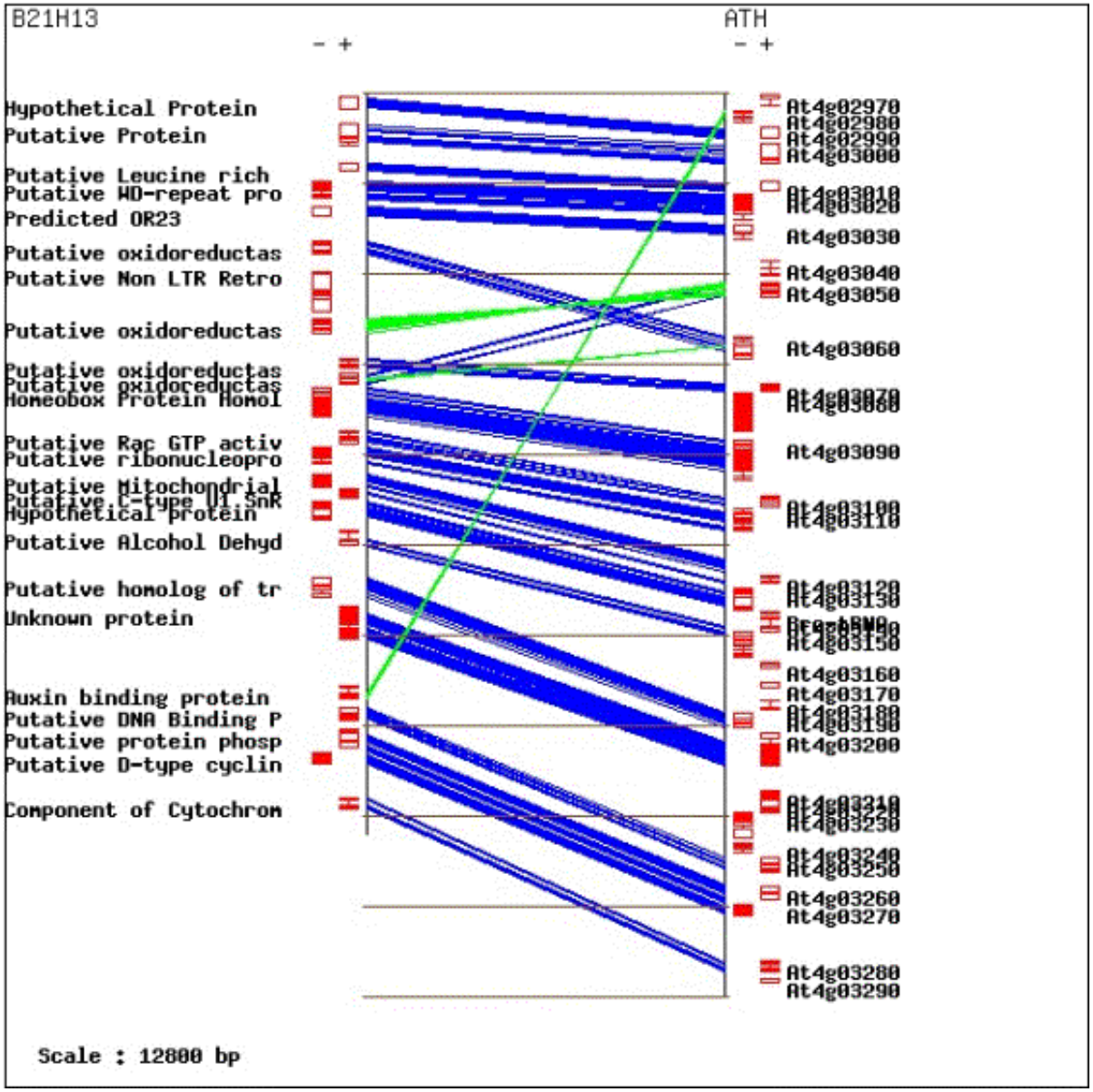




The Brassica sequence is from Quiros et al. Genetics, 2001. PiPMaker (Schwartz et al. Genome Res. 2000) was used to create the images. *Brassica* is on the X-axis and it is compared to its putative homologous region in *Arabidopsis*. N-MYR is not present in *Arabidopsis*.




A *Brassica* read aligns to a hypothetical protein in *Arabidopsis*. The rectangles represent annotated exons, which correspond well with alignment.

# Conserved Gene order between Arabidopsis and Brassica

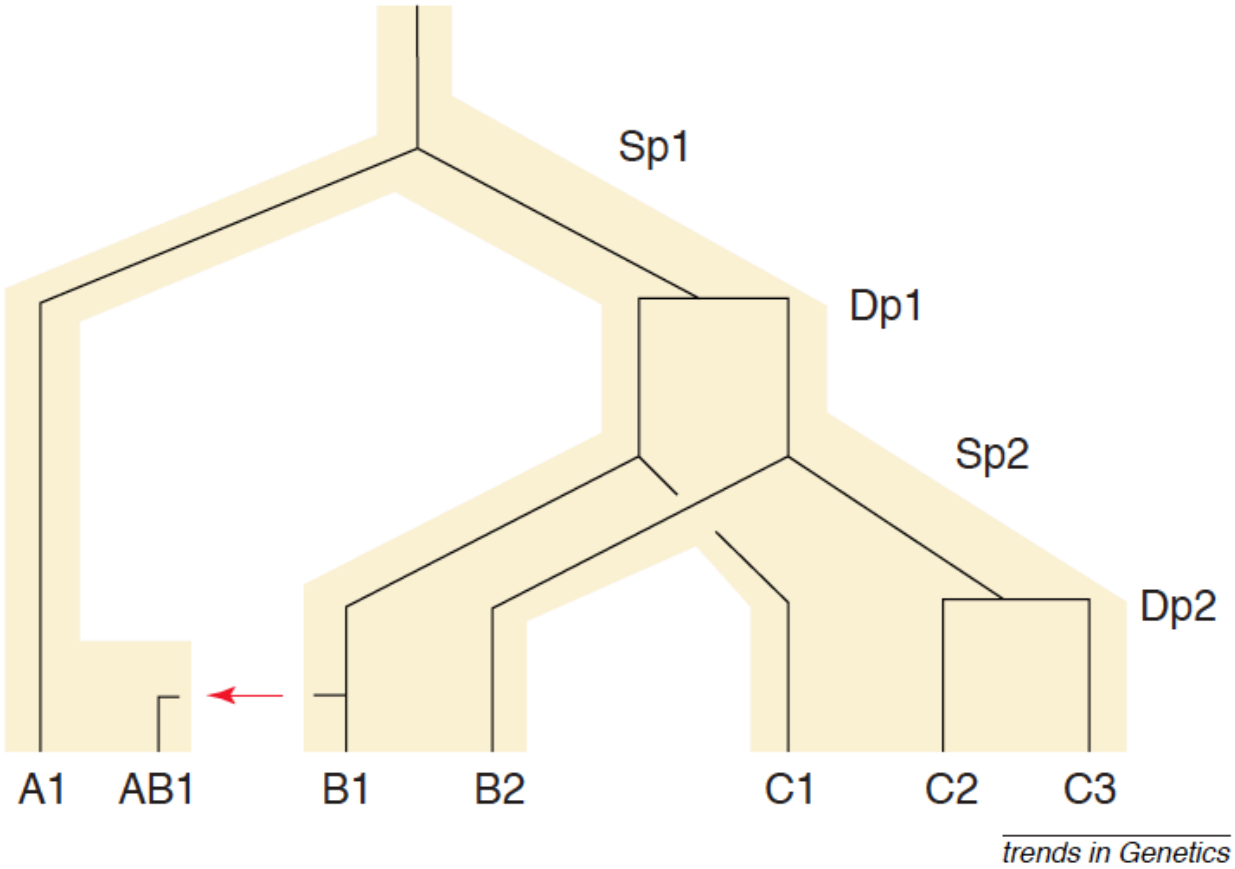


# Some Useful Definitions

- **Homology -** The relationship of any two characters that have descended, usually with divergence, from a common ancestral character.
- **Characters -** Any genic, structural or behavioral feature of an organism having at least two forms of the feature called character states.
- **Orthology -** The relationship of any two homologous characters whose common ancestor lies in the cenancestor of the taxa from which the two sequences were obtained.
- **Paralogy -** The relationship of any two homologous characters arising from a duplication of the gene for that character.
- **Xenology -** The relationship of any two homologous characters whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material for at least one of those characters.

Fitch WM (May 2000). “Homology a personal view on some of the problems”. Trends Genet. 16 (5): 227–31.

**FIGURE 1. Orthology, paralogy and xenology**




There are two speciation events (Sp1 and Sp2) and two gene-duplication events (Dp1 and Dp2)
Two genes whose common ancestor resides at speciation are orthologous.
Two genes whose common ancestor resides at gene duplications are paralogous.
The red arrow denotes the transfer of the B1 gene from species B to species A.
As a result, the AB1 gene is xenologous to all six other genes.

Fitch WM (May 2000). "Homology a personal view on some of the problems". Trends Genet. 16 (5): 227–31.

# Some common methods and databases to define homology

- Top Reciprocal Blast Hit
  - Align all proteins of Species A to Species B.
  - For protein PA in Species A, see the top hit in species B (PB). Then check if PA is the top hit of PB. If so they are putative orthologs.
- COG (Cluster of Orthologous Genes)
  - Take an exhaustive Top Reciprocal Blast Hit approach on as many species as you care to.
  - Cluster together proteins that are top hits of each other across several species.

# Dealing with Paralogy

- Inparanoid
  - Inparalogs = paralogs that arose through a gene duplication event after speciation
  - Outparalogs = paralogs that arose following a gene duplication preceding speciation (can never be orthologous)
  - Clustering of inparalogs can help define one to one and one to many orthologs.
  - Perform Blast comparing the different species but also blast within the species.
- orthoMCL
  - Inparanoid is done pairwise (2 species at a time), which is not very scalable.
  - orthoMCL uses a markov cluster algorithm to make it scalable.

# OrthoMCL pipeline

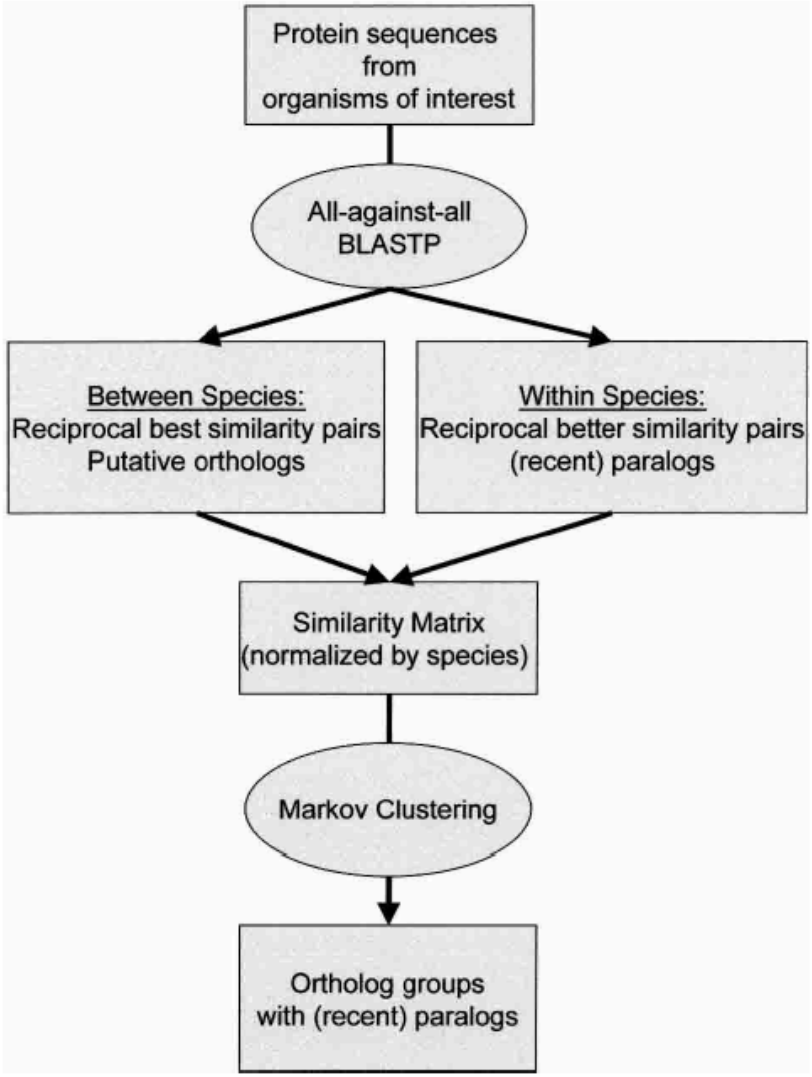
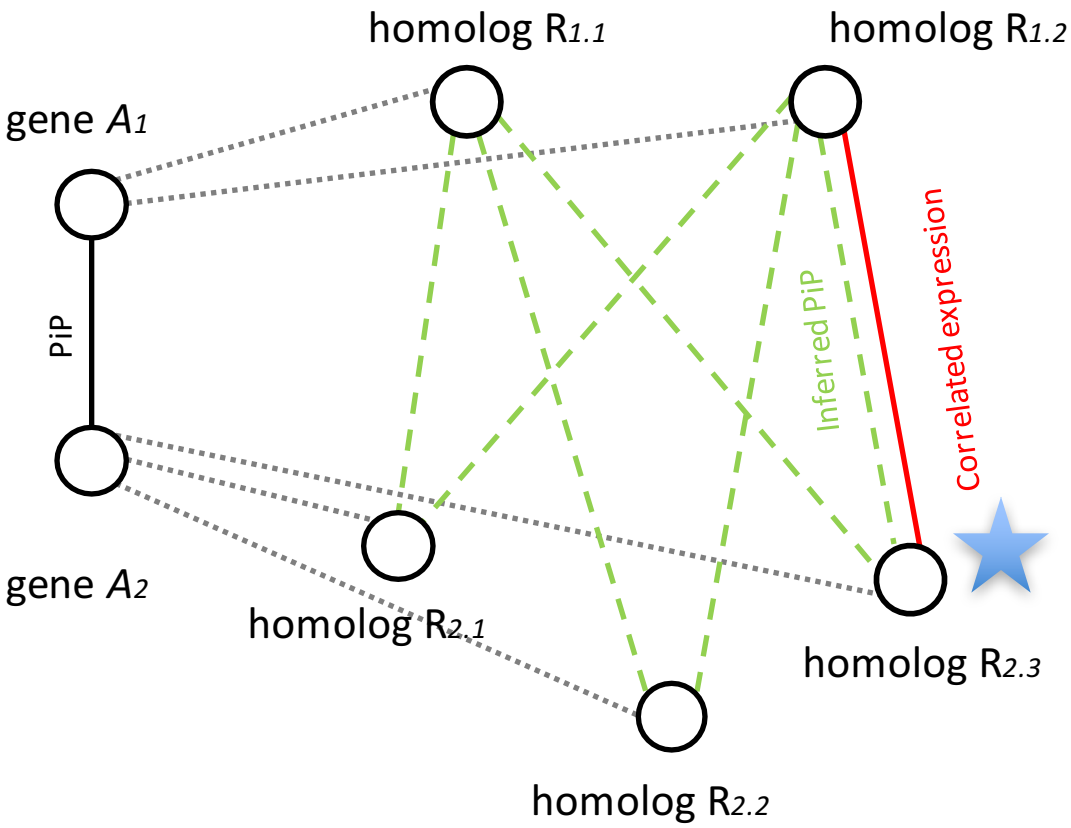



**Figure 1** Flow chart of the OrthoMCL algorithm for clustering orthologous proteins.

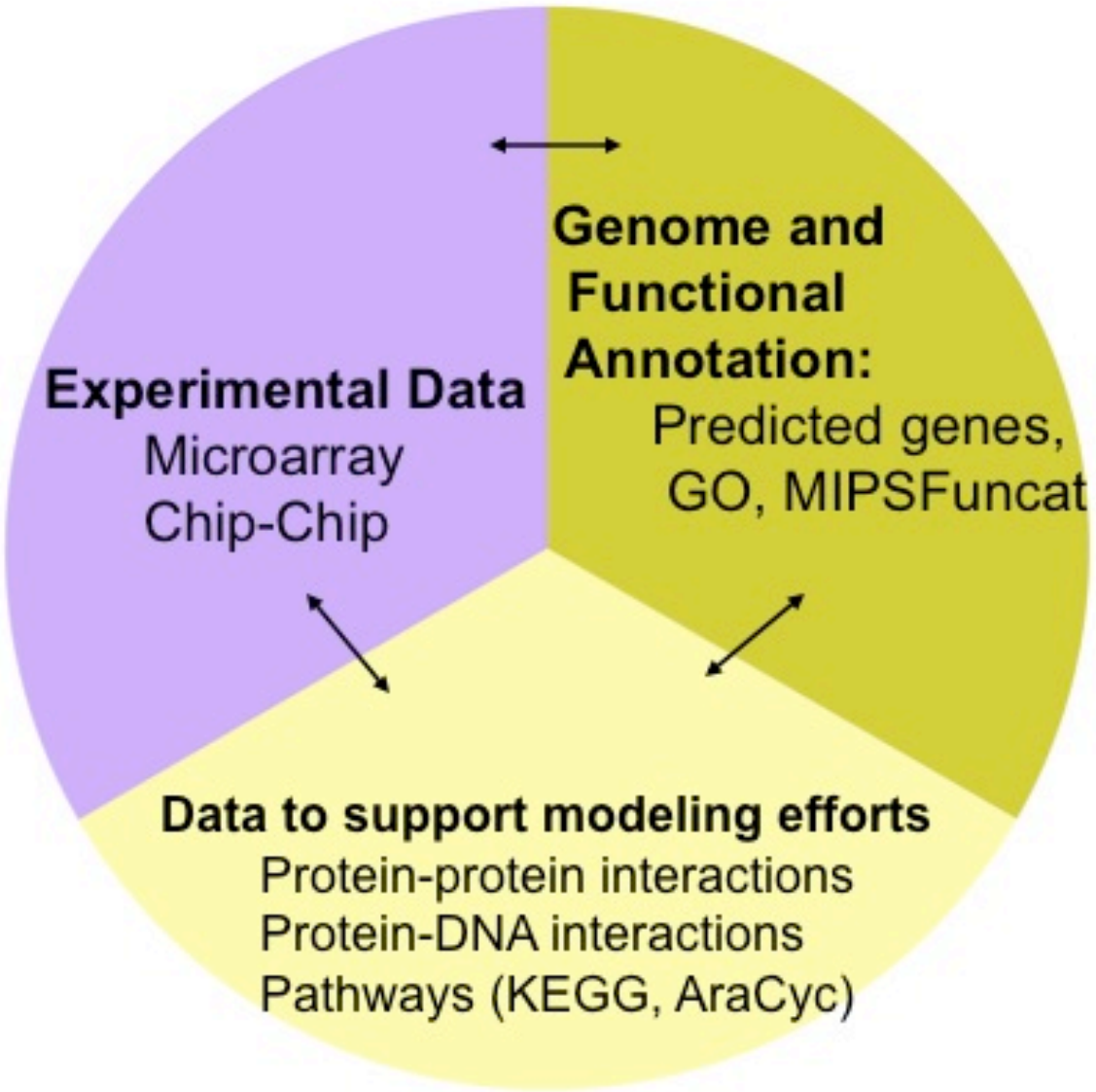
# Network Inference using correlation



How would you design the database schema for Homology Data ?

How would you design the database schema for Interaction Data ?

# Contents of an Integrated Database



# Current data integration strategy

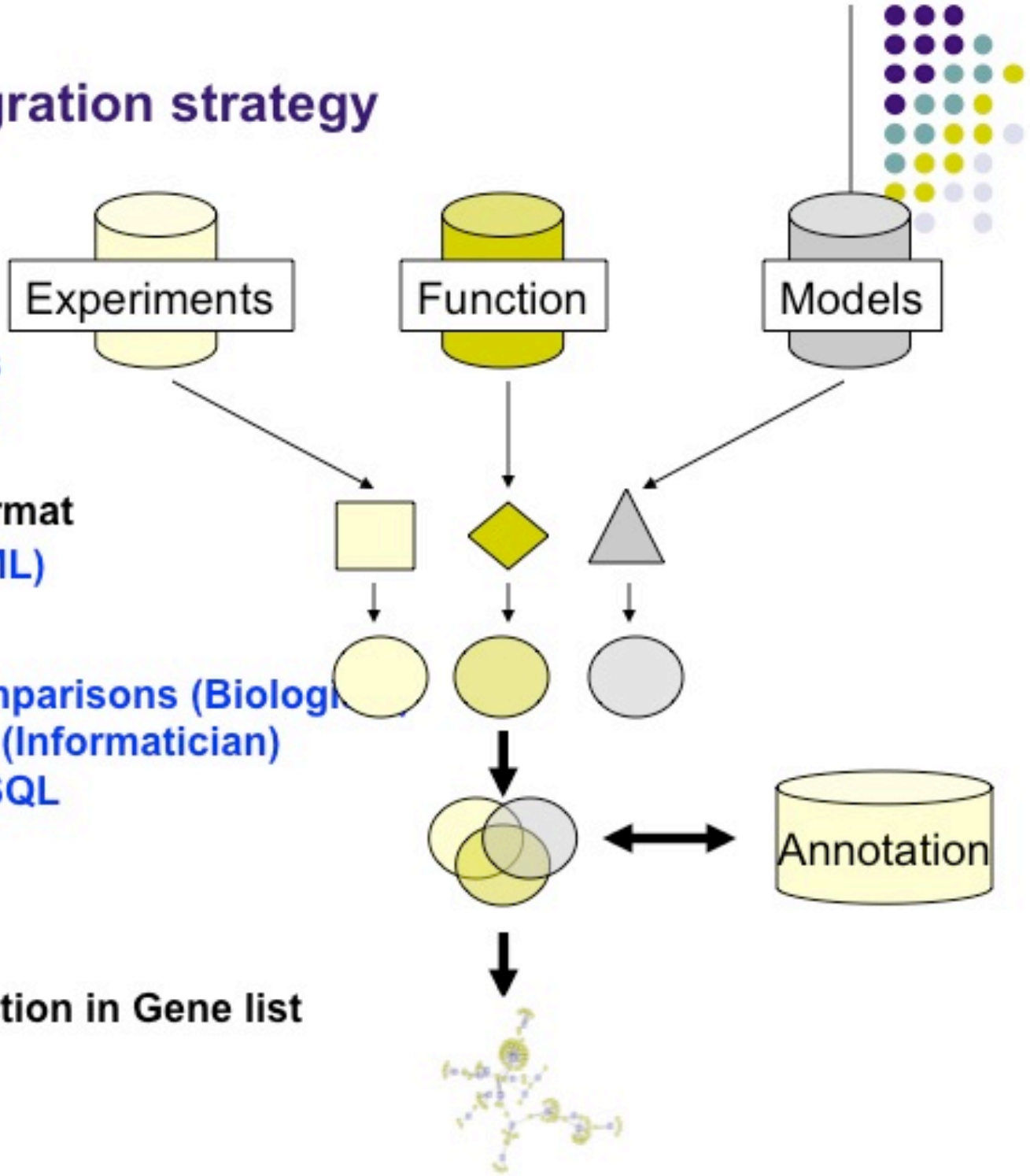
1) Find the data.
   Decentralized databases
   Data in different formats

2) Convert to a common format
   XML is a good idea (SBML)

3) Data integration.
   Manual: Excel sheet comparisons (Biologi
   Automated: Perl Scripts (Informatician)
   Database: Queries e.g. SQL
   (High-production labs)

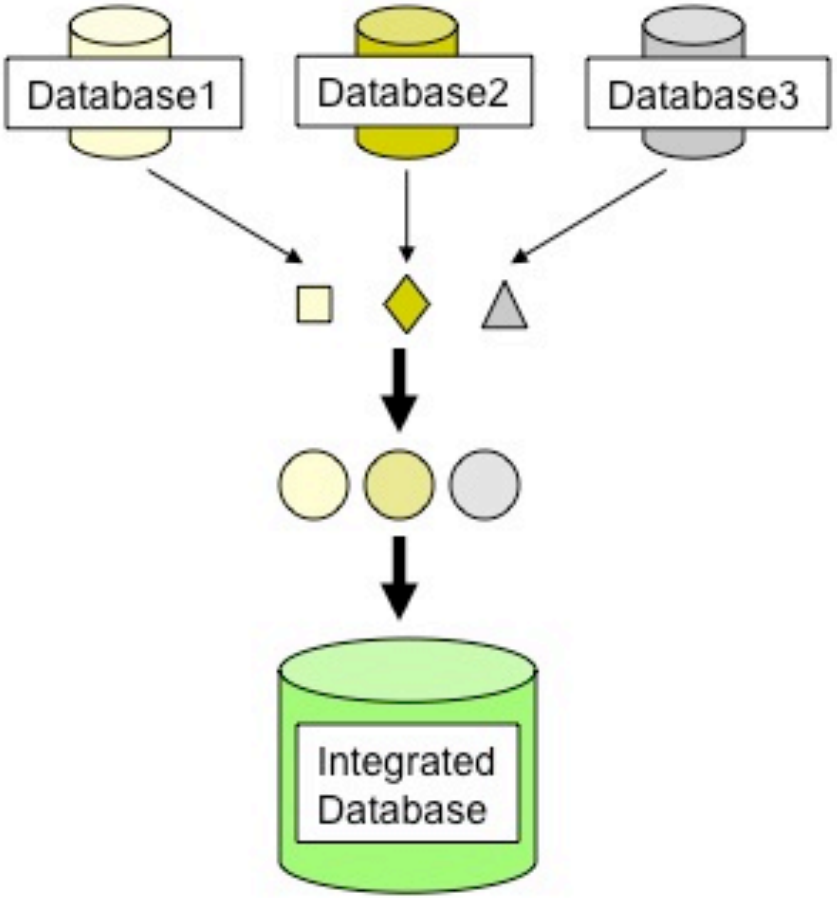
4) Gene list intersect.

5) Modeling Biological function in Gene list
   Need visualization and
   network modeling tools

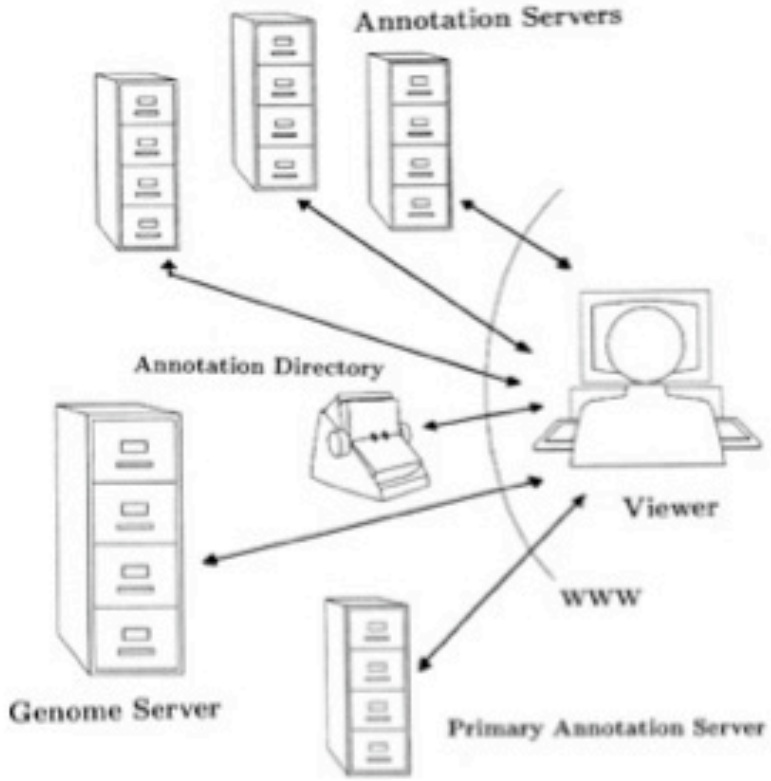


# Suggested Solutions for Data Integration:

1)One database stores all information.

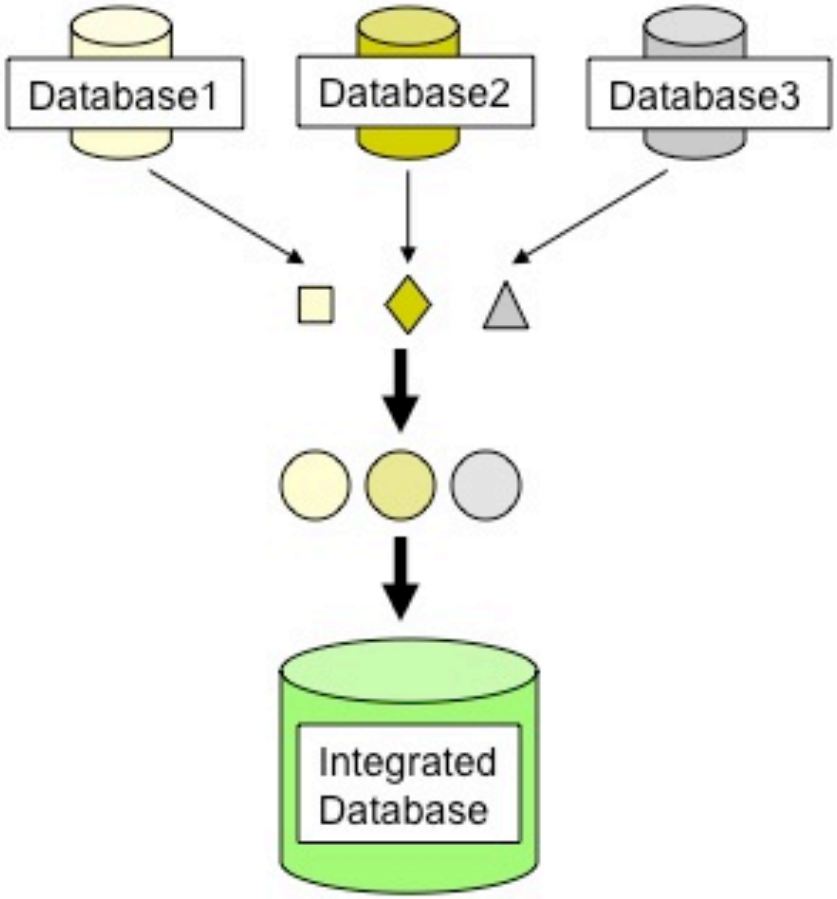



2)Remote Databases in standardized format linked to client. (DAS, BioMOBY, ISYS, etc.)



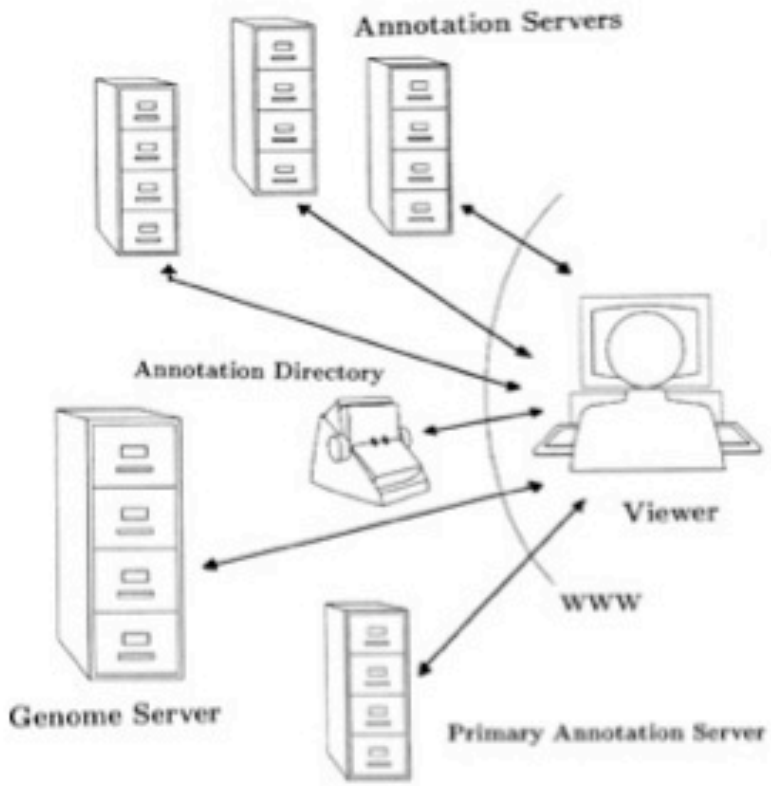

Robin Dowell, Wash. Univ.
http://selab.wustl.edu/people/robin/das/bosc00/sld001.htm

# One database stores all information.



# Remote Databases in standardized format linked to client. (DAS, BioMOBY, ISYS, etc.)

2)



# Examples of Remote Database servers

- http://intermine.org/
- http://www.biomart.org/