

Protein annotation based on protein-protein interactions and machine learning

Xingming Zhao

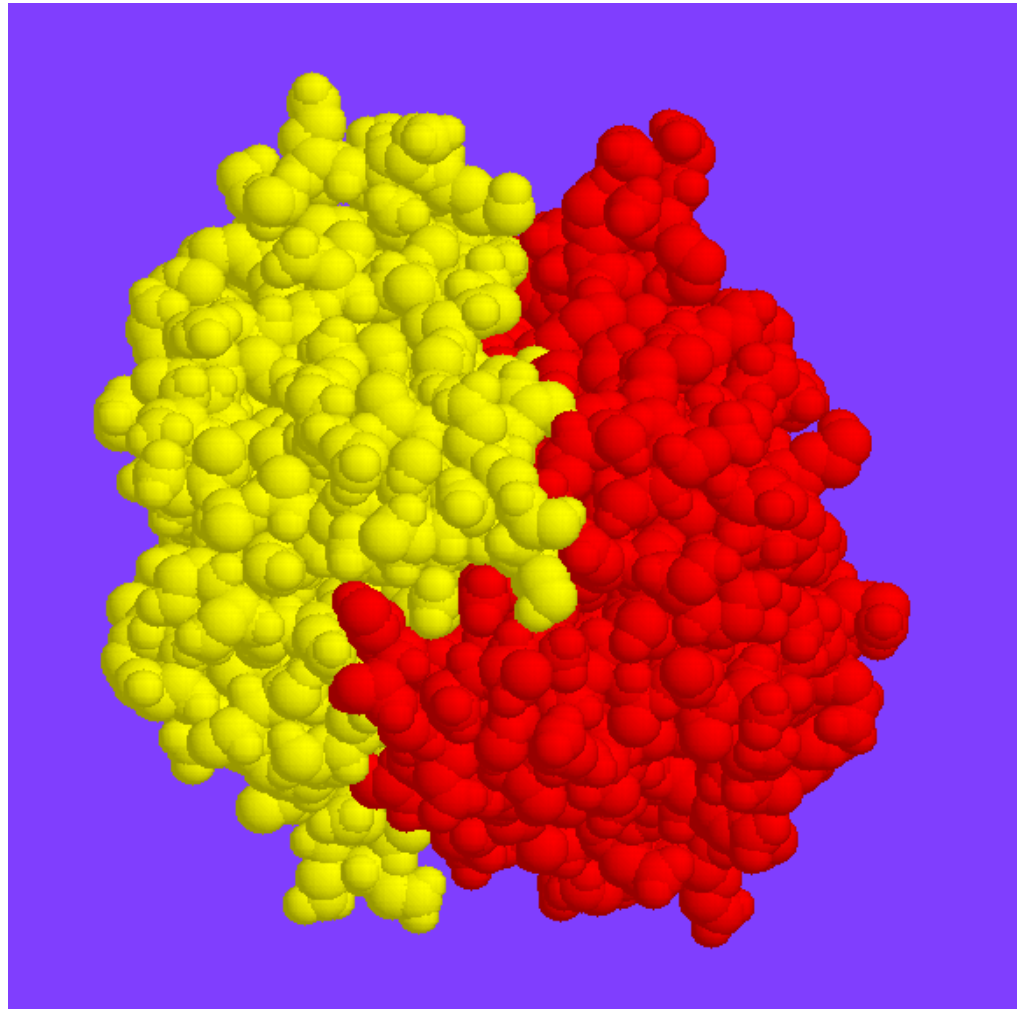
Institute of system biology, Shanghai university

Outline

- Protein-protein interaction (PPI)
- Protein annotation
- Protein annotation based on PPI and machine learning
- Conclusions

Protein-protein interaction

- **Physic PPI**



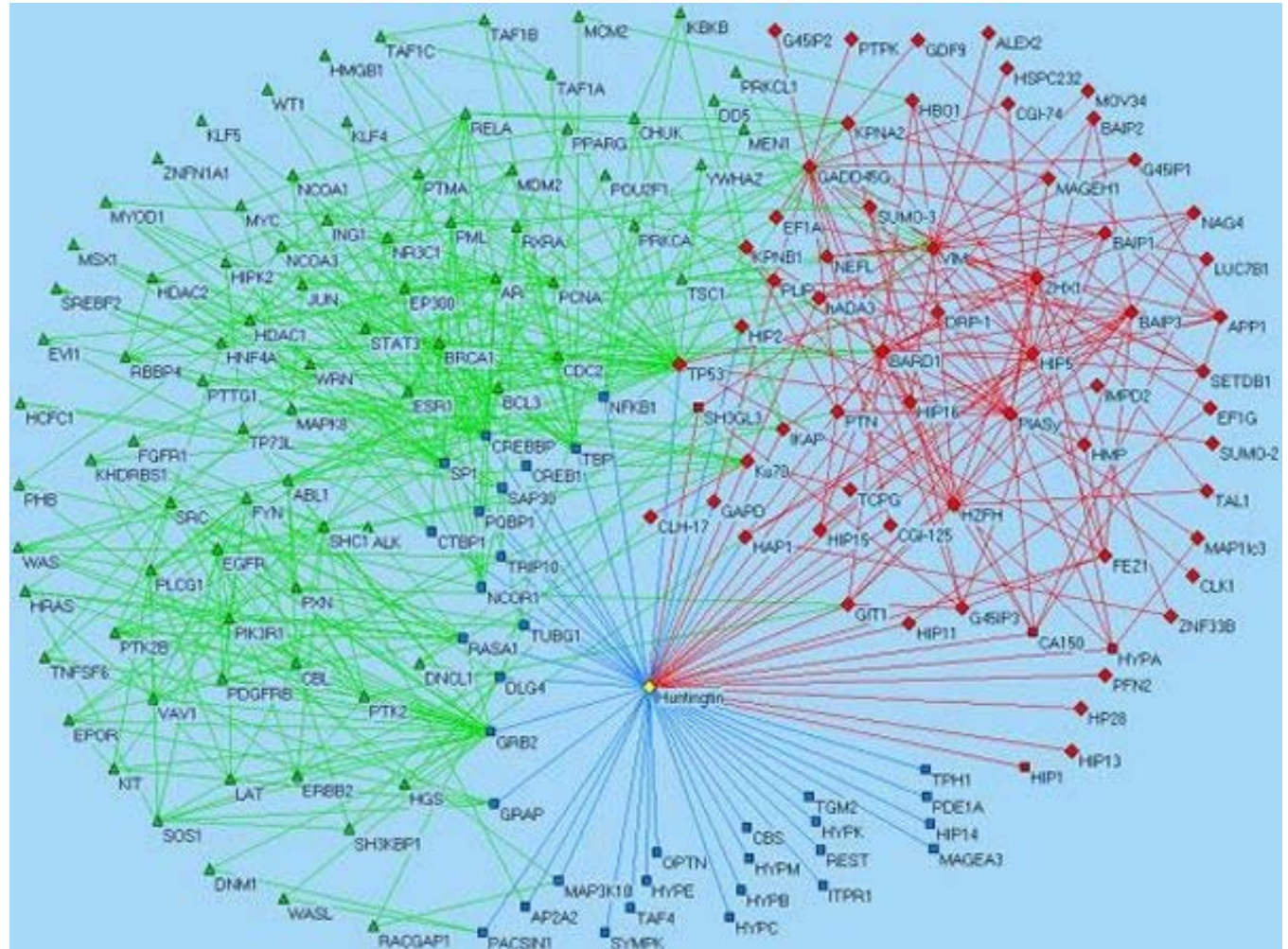
• PPI network

Undirected Graph

$G(V,E)$

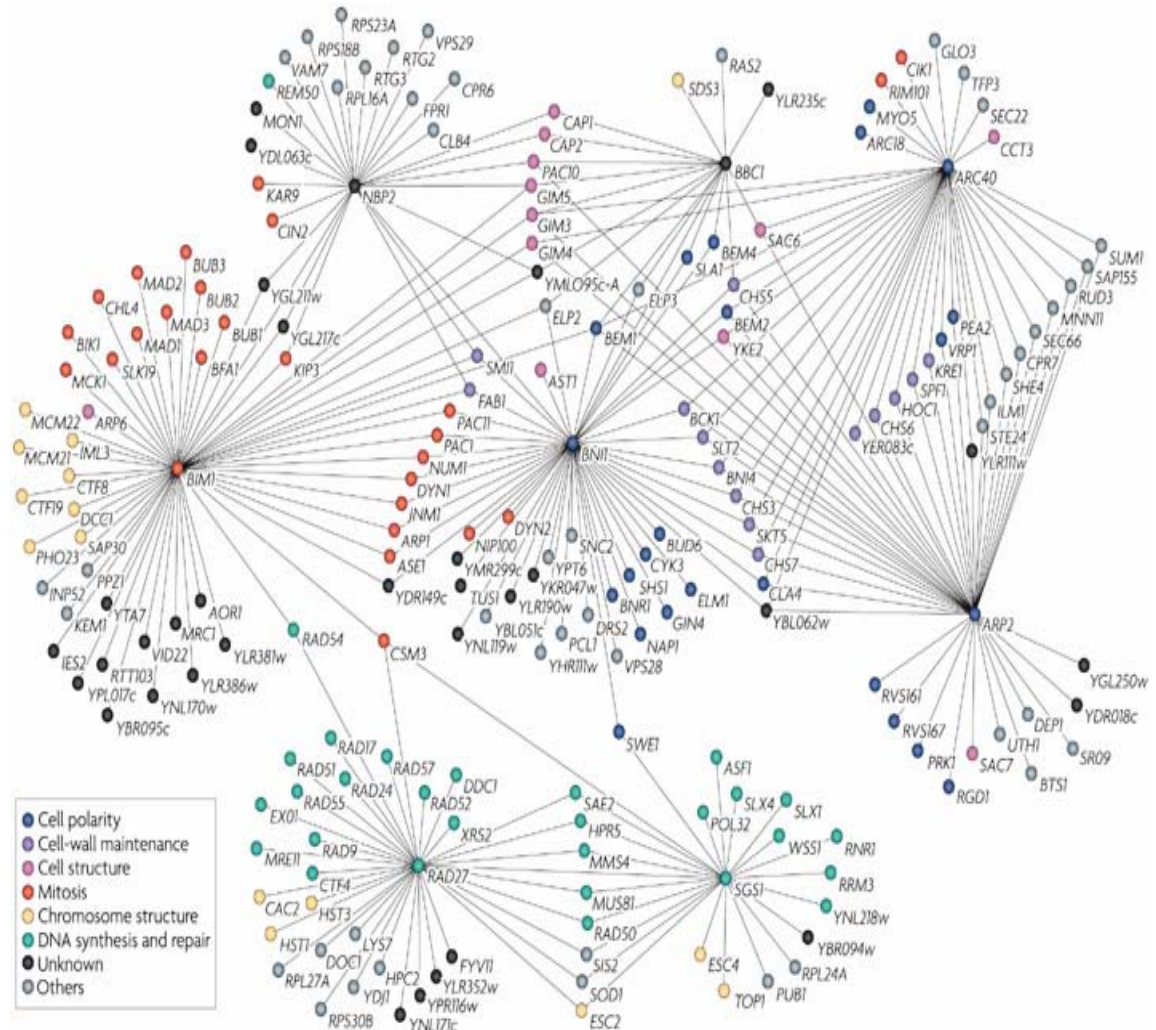
Node - proteins

Edge - interaction

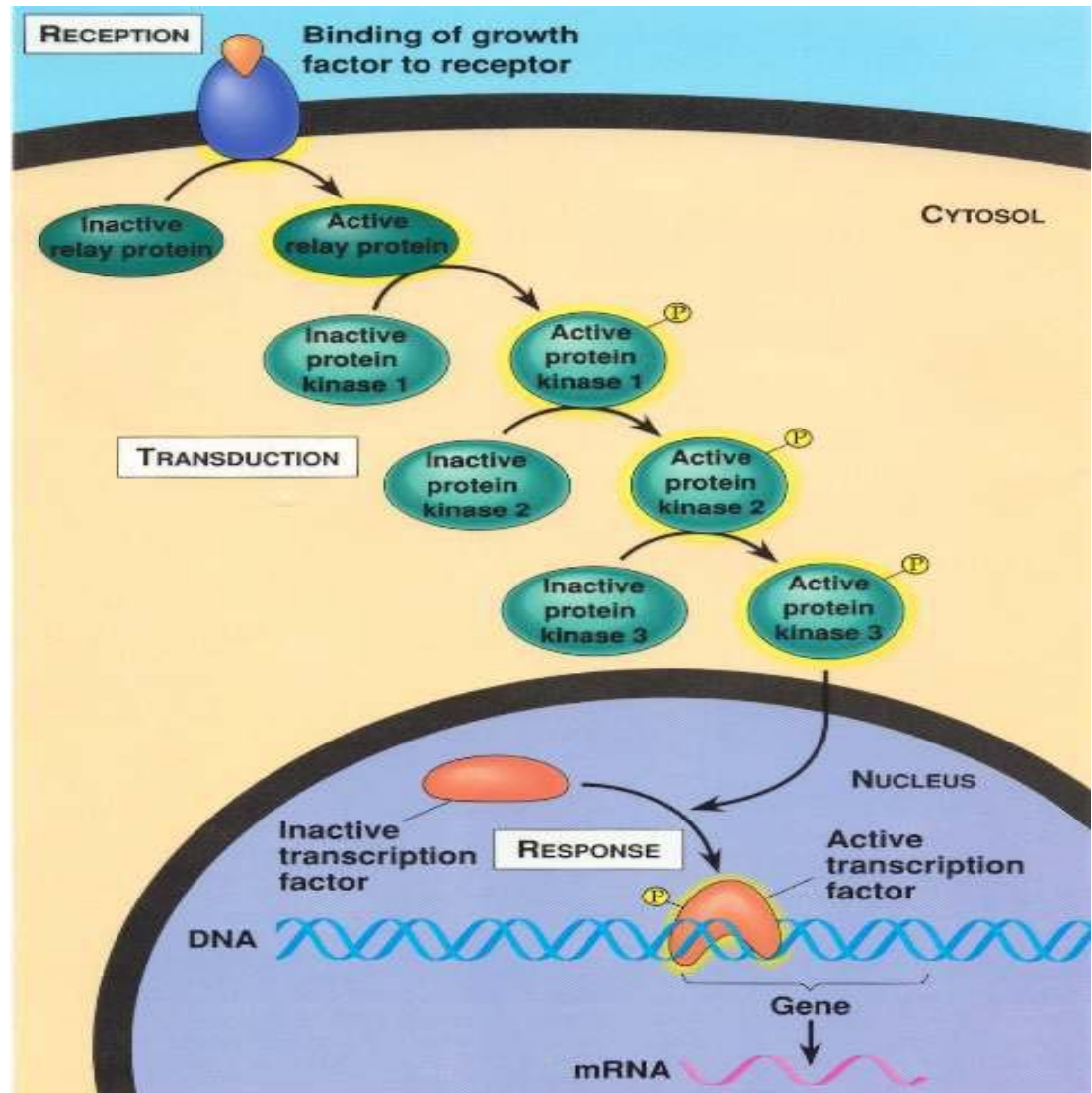


Genetic interaction

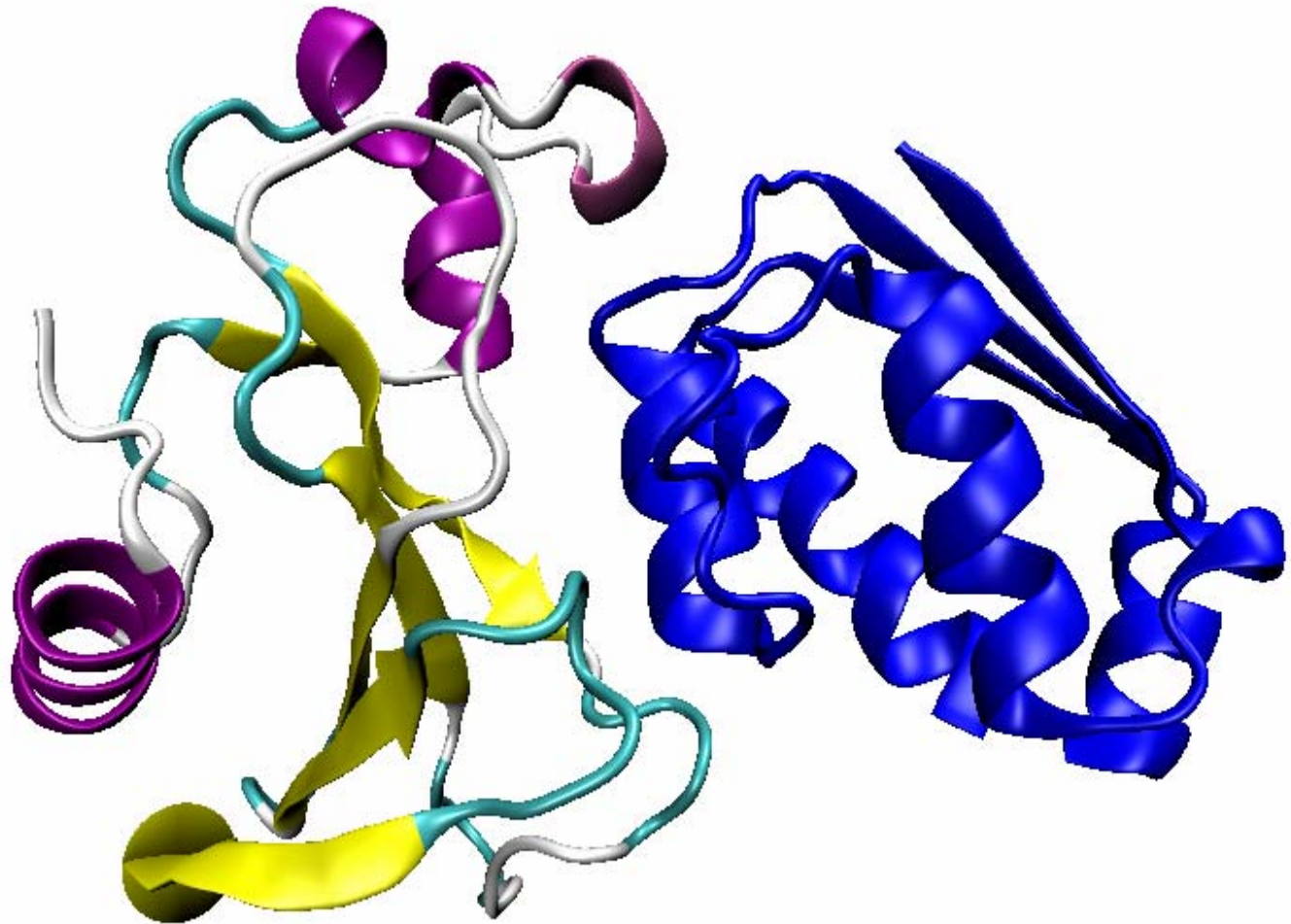
two mutations
have a combined
effect not exhibited
by either mutation
alone,



Signal transduction
via PPI



Protein complex
formation via PPI



- PPI database:

- HPRD

- <http://www.hprd.org/>

- BioGRID (physic and genetic interaction)

- <http://www.thebiogrid.org/>

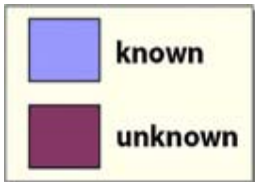
- DIP (experimentally determined)

- <http://dip.doe-mbi.ucla.edu/dip/Main.cgi>

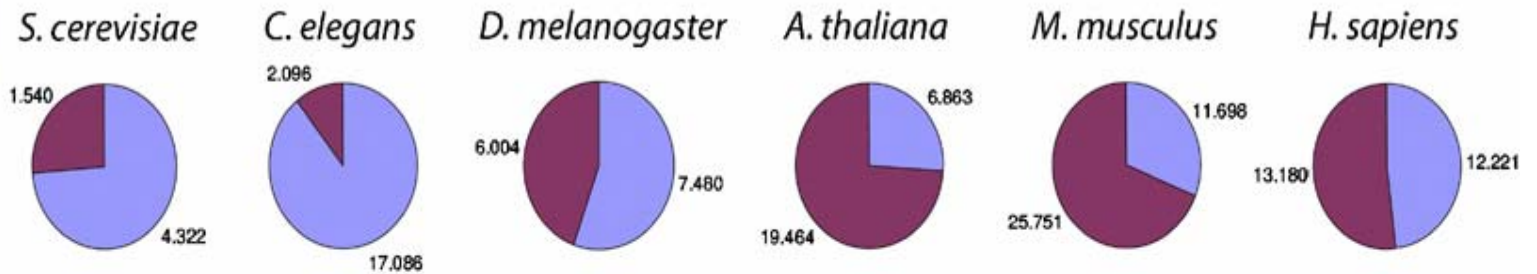
PPI network provide:

- Global view of functional relationship among proteins;
- Differences between normal and disease states;
- Local view of protein complexes;
- Details of signal pathways

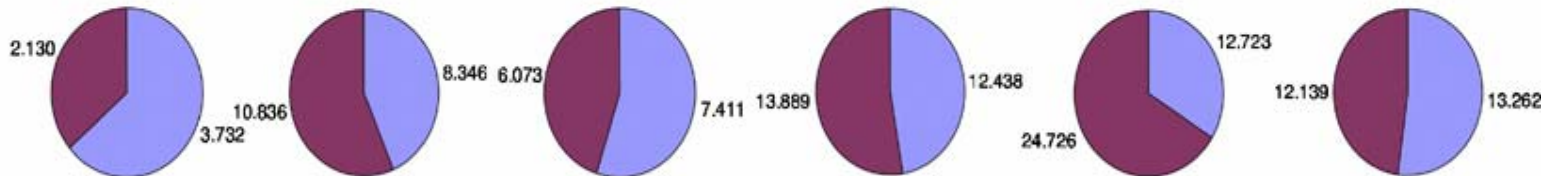
Protein annotation



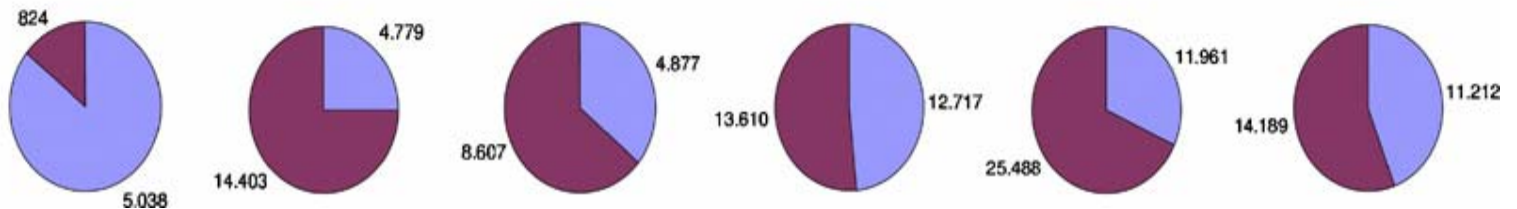
Biological process



Molecular function



Cellular component



Information

Name(s) Protein serine/threonine kinase essential for cell wall remodeling during growth
Type gene
Species [Saccharomyces cerevisiae](#)
Synonyms CLY15
 HPO2
 STT1
 YBL105C
Database SGD, [SGD:S000000201](#)
Sequence [View sequence](#); [use as BLAST query sequence](#)

[Back to top](#)

Term Associations

Filter associations displayed ?

Filter Associations

Evidence Code	Ontology
All Curator Approved	All
IC	Biological Process
IDA	Cellular Component
IEP	Molecular Function

Qualifier	Term	Ontology	Evidence	Reference	Assigned by
	cytoplasm [view associations]	cellular component	IDA	PMID:11545731	SGD
	cytoskeleton [view associations]	cellular component	IDA	PMID:15910746	SGD
	nucleus [view associations]	cellular component	IDA	PMID:11545731	SGD
	site of polarized growth [view associations]	cellular component	IDA	PMID:10893184	SGD
	protein kinase C activity [view associations]	molecular function	IDA	PMID:8207005	SGD
	actin filament organization [view associations]	biological process	IGI	PMID:12810699	SGD
	cell wall organization and biogenesis [view associations]	biological process	IMP	PMID:7874200	SGD
	protein amino acid phosphorylation [view associations]	biological process	IDA	PMID:8207005	SGD

Done

Annotation database:

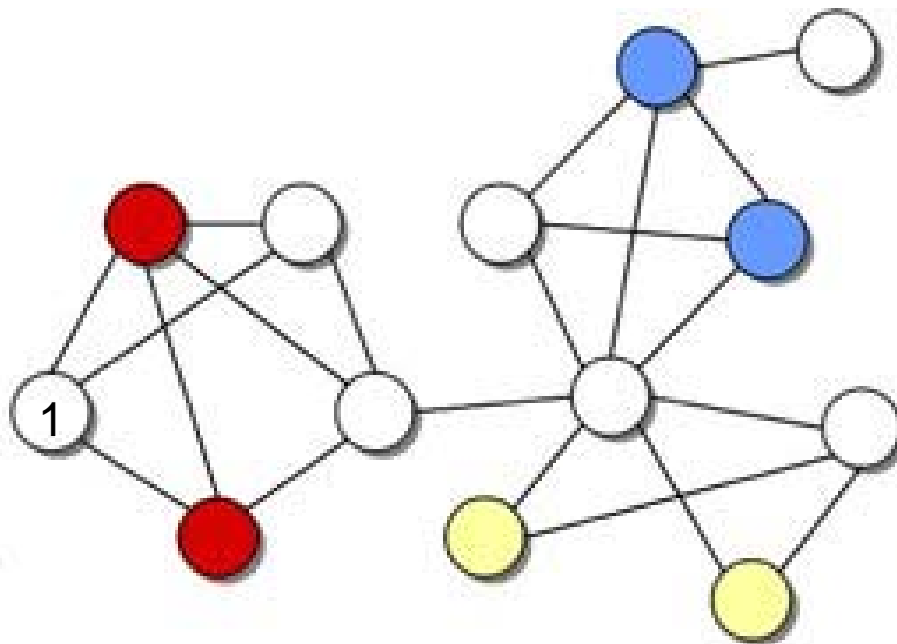
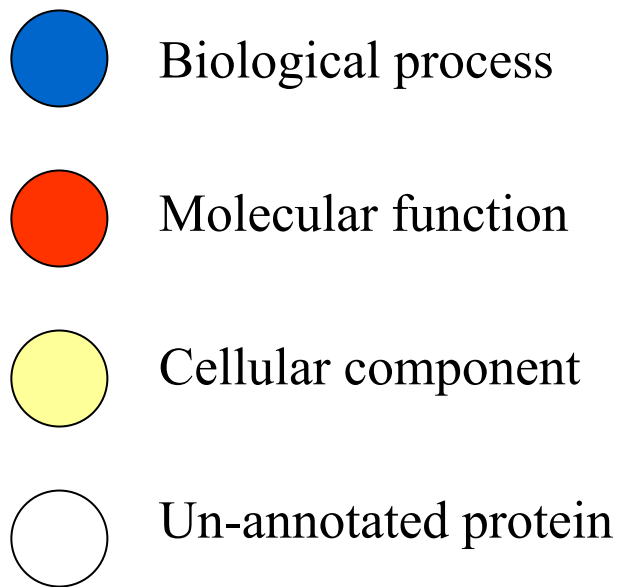
GO: directed, acyclic graph;

<http://www.geneontology.org/>

MIPS: tree-like structure

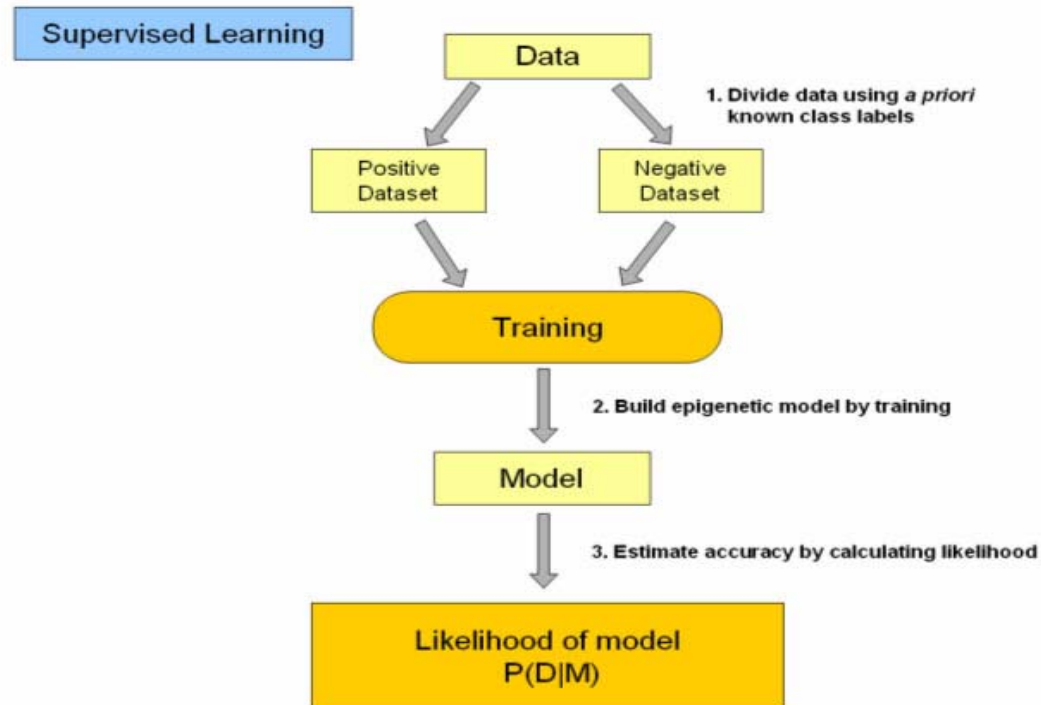
<http://mips.gsf.de/projects/funcat>

PPI & protein annotation



Protein annotation based on PPI and machine learning

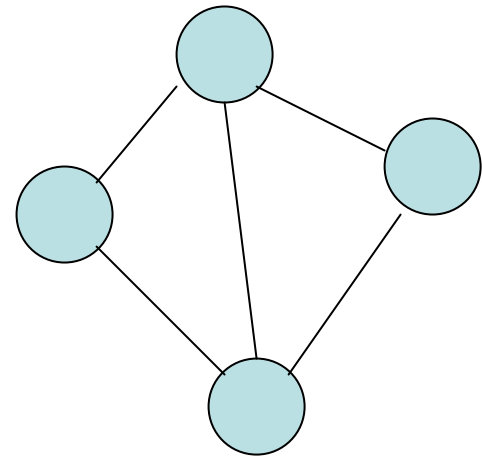
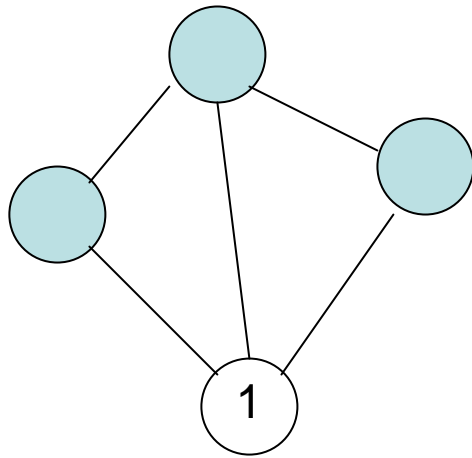
Supervised learning



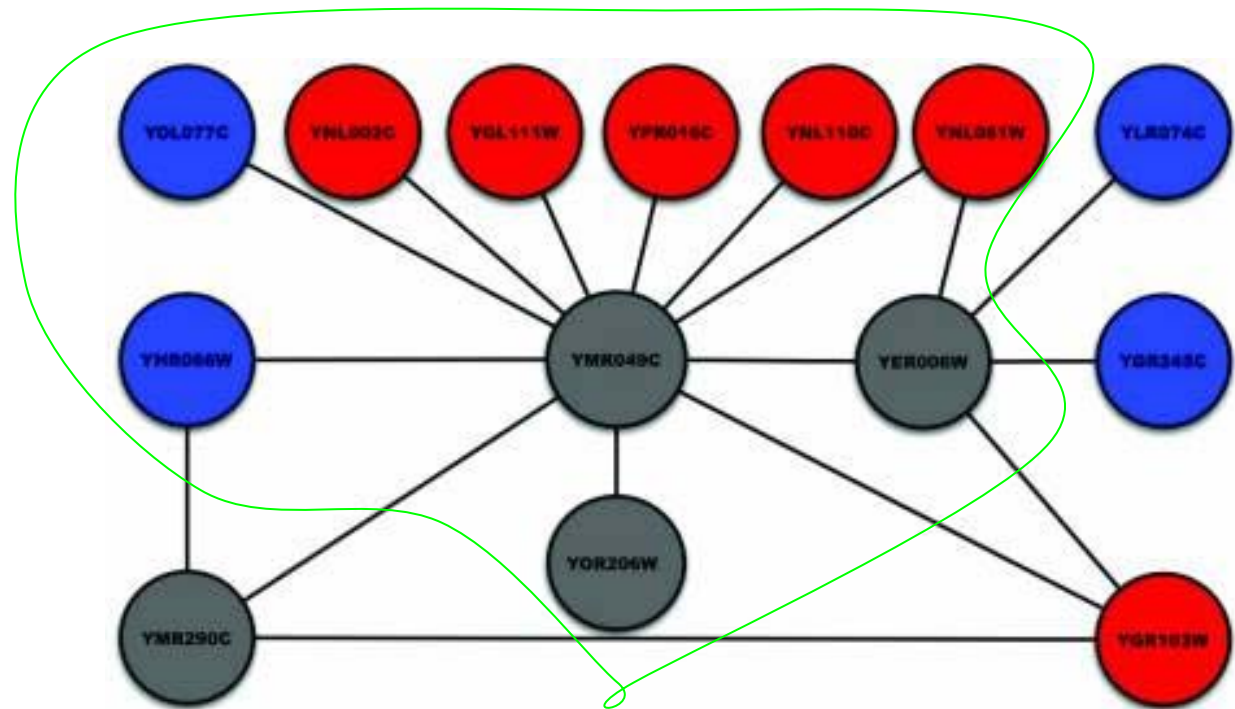
An example of a supervised learning process. Training is guided by *a priori* knowledge obtained experimentally. Accuracy is improved using positive and negative data rather than merely using positive data and random background information. The accuracy of the model is then estimated by calculating its likelihood, $P(D|M)$ where P is the probability of the dependency model M using data D .

Supervised learning

Association method



Neighborhood



Markov random field (MRF)

The conditional probability on the functional labeling is proportional to $\exp(-U(x))$, where x is the value of X , and

$$U(x) = -\alpha N_1 - \beta_1 N_{10} - \gamma_1 N_{11} - \kappa_1 N_{00}$$

$$N_1 = \sum_{i=1}^N x_i$$

$$N_{11} = \sum_{(i,j) \in S} x_i x_j,$$

$$N_{10} = \sum_{(i,j) \in S} (1 - x_i) x_j + (1 - x_j) x_i,$$

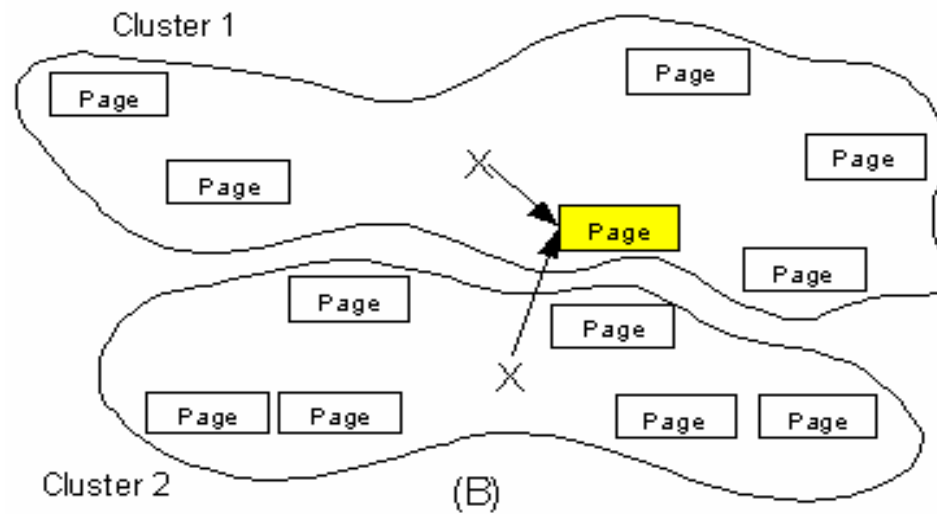
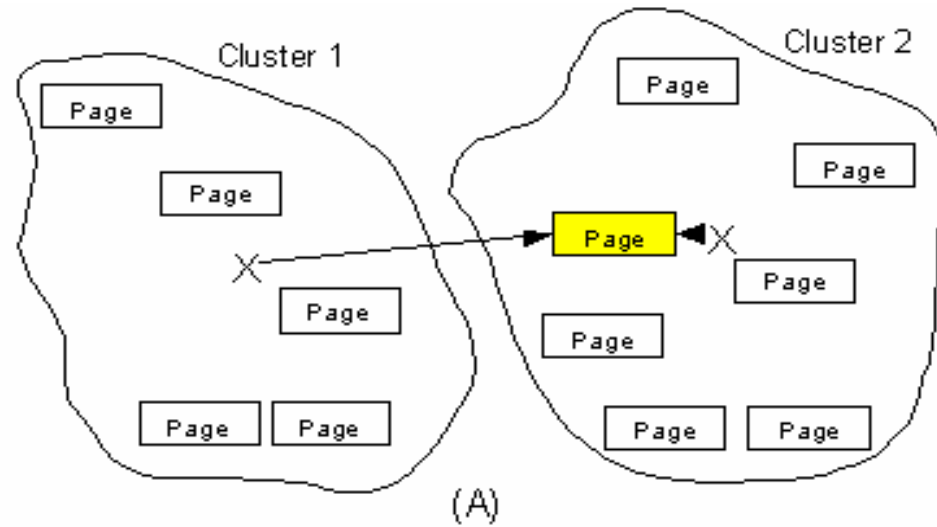
$$N_{00} = \sum_{(i,j) \in S} (1 - x_i)(1 - x_j).$$

Global optimization

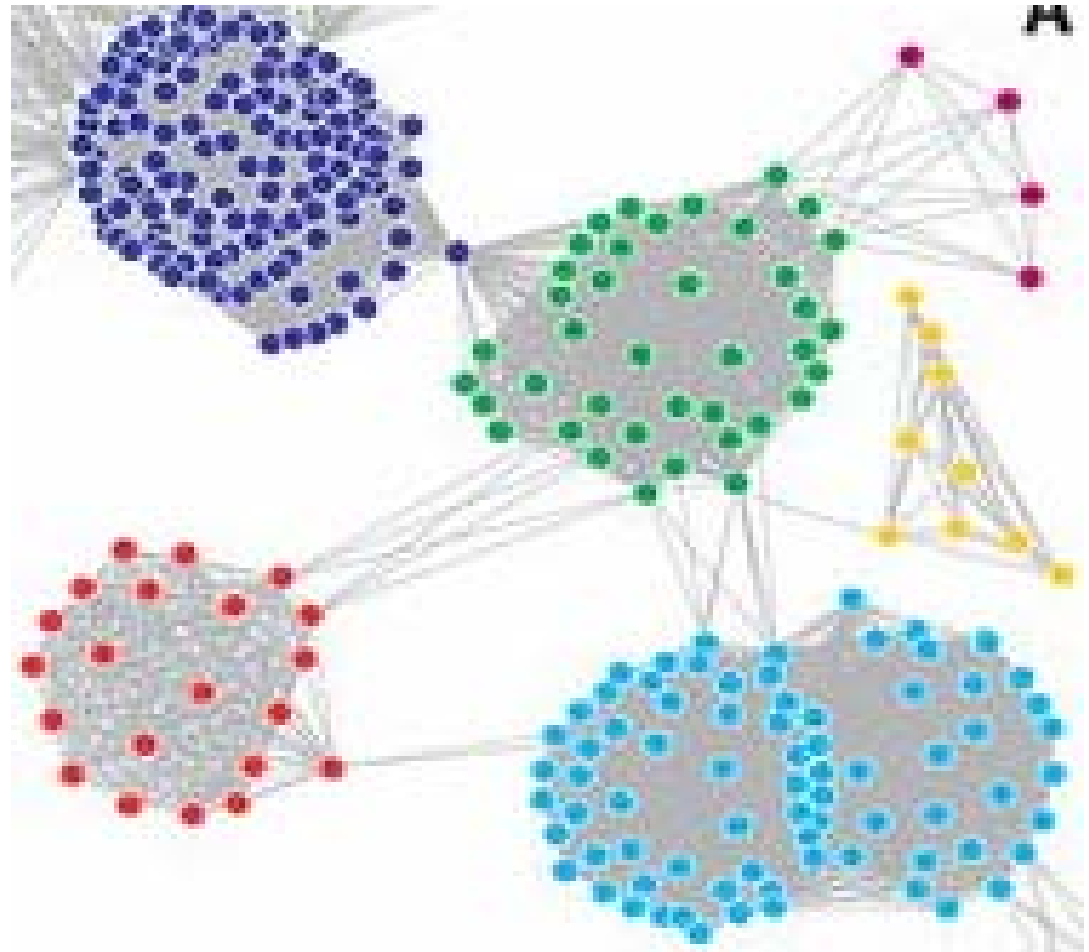
$$E = - \sum_{i,j} J_{ij} \delta(\sigma_i, \sigma_j) - \sum_i h_i(\sigma_i)$$

where J_{ij} is the element of the adjacency matrix of the interaction network, $\delta(i, j)$ is the discrete δ function and $h_i(\sigma_i)$ is the number of partners of protein i annotated with function σ_i . The simulated annealing is employed to minimize

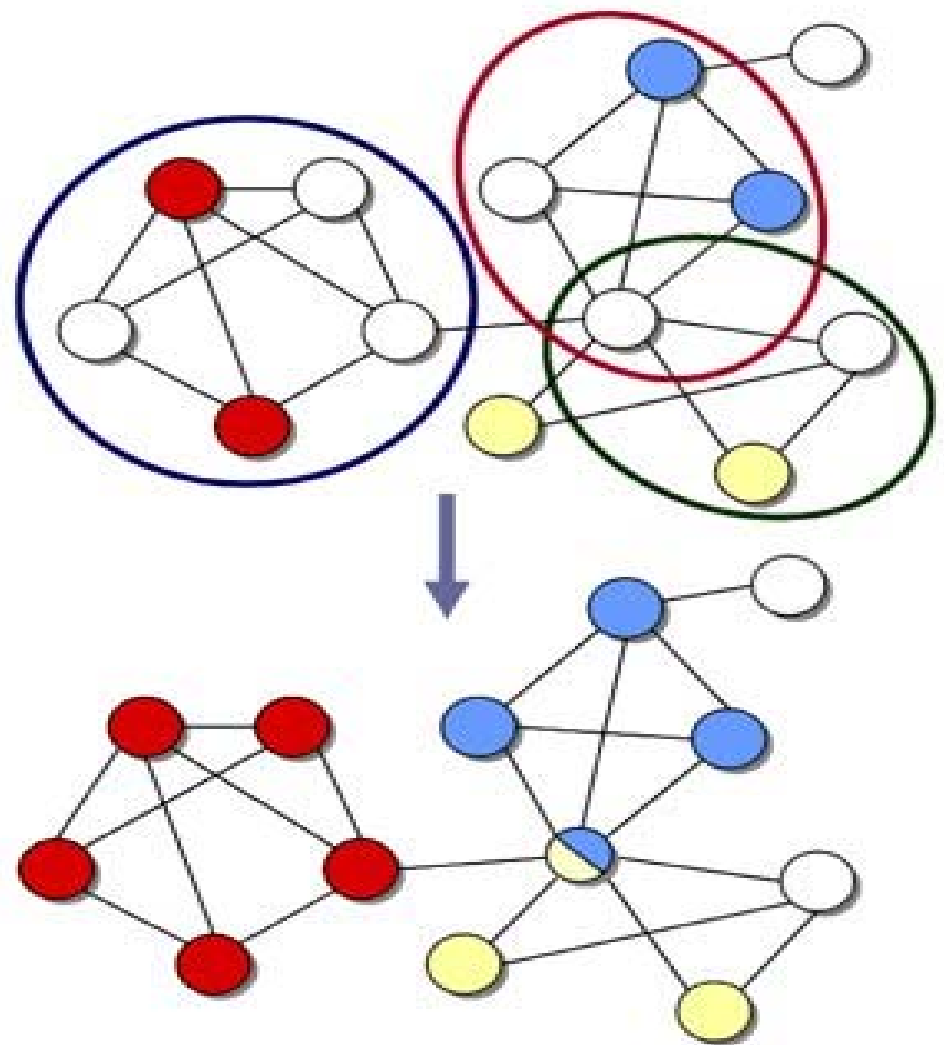
Unsupervised learning



Modular structure of PPI network



Clustering



- Similarity based clustering

- P-value

$$P(N, N_A, N_B, m) = \frac{\binom{N}{m} \binom{N-m}{N_A-m} \binom{N-N_B}{N_B-m}}{\binom{N}{N_A} \binom{N}{N_B}}$$

where N is the total number of proteins in the PPI network, N_A and N_B are respectively the number of interaction partners of A and B, and m is the number of proteins in common between N_A and N_B .

- shortest path distance

For weighted PPI;

- Neighborhood similarity

$$D(g1, g2) = \frac{|N_{g1} \Delta N_{g2}|}{|N_{g1} \cup N_{g2}| + |N_{g1} \cap N_{g2}|}$$

MCODE

1) vertex weighting

$$C_i = 2n/k_i(k_i-1)$$

where k_i is the vertex size of the neighborhood of vertex i and n is the number of edges in the neighborhood.

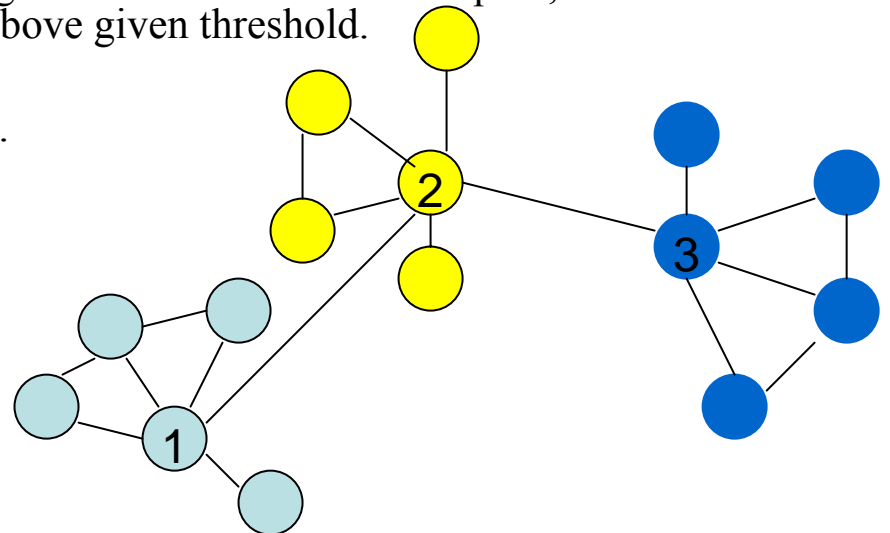
2) Finding complex

Given highest weighted vertex s , add the neighbors of s to the same complex, where the weights of its corresponding neighbors are above given threshold.

Proceed until no more vertices can be added.

Remove found complex;

Repeat above procedure.



Conclusions

1. PPI network provide insights into protein functions;
2. Machine learning techniques are powerful in predicting protein functions.

Thank you for your attention!