

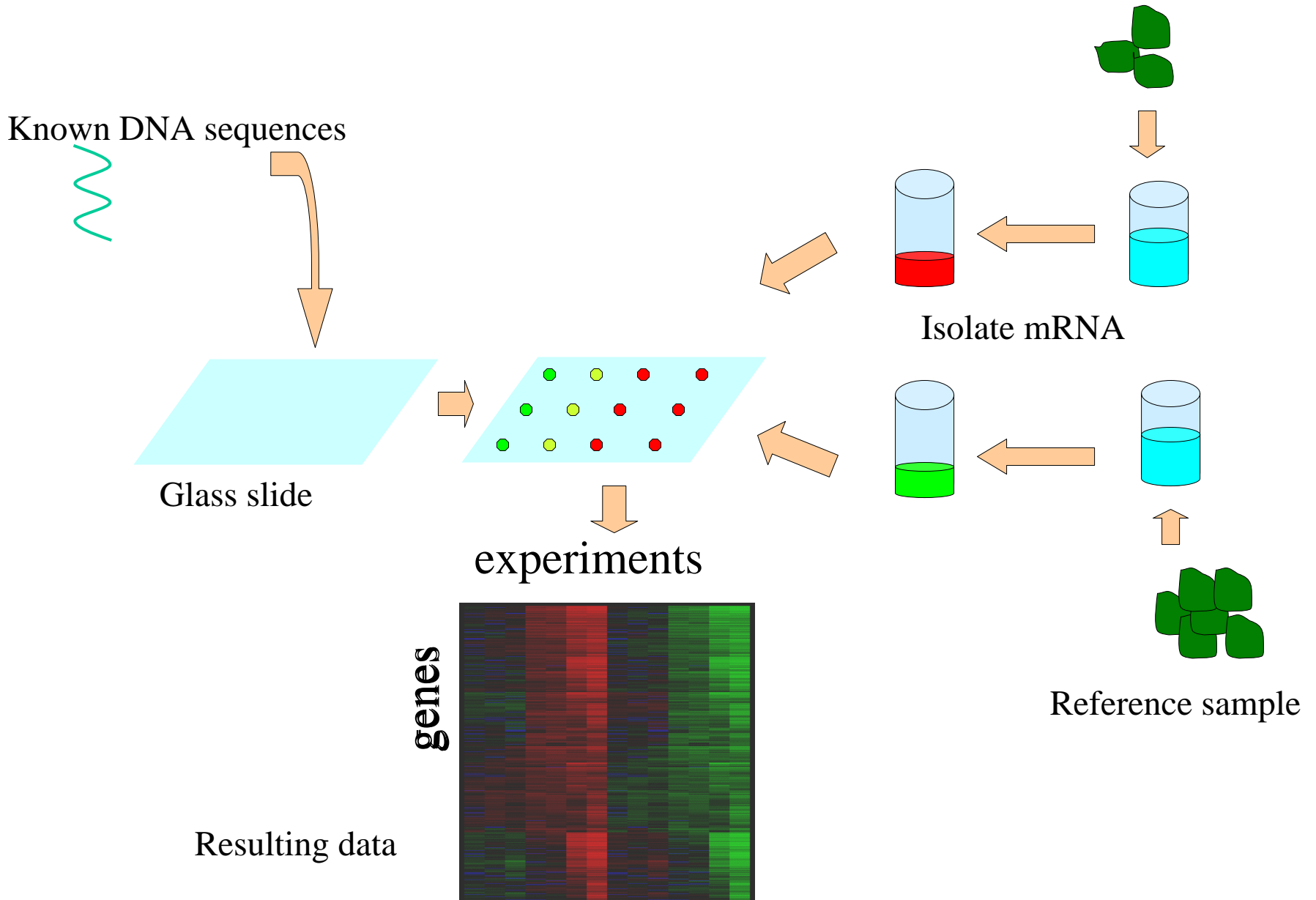
# Gene function prediction

Computational analysis of biological  
networks.

Olga Troyanskaya, PhD

# Available Data

# Coexpression - Microarrays



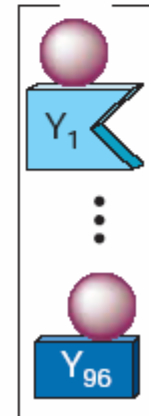
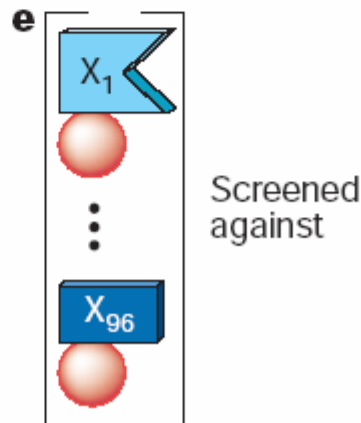
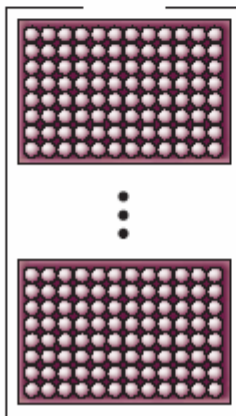
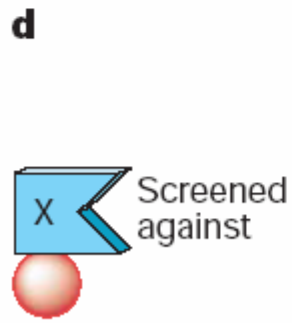
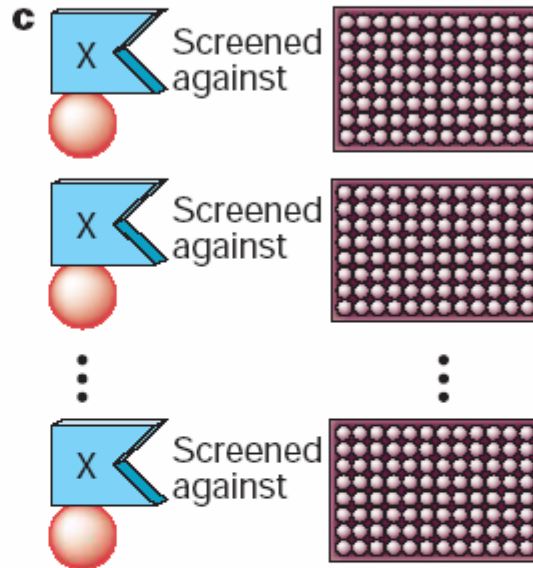
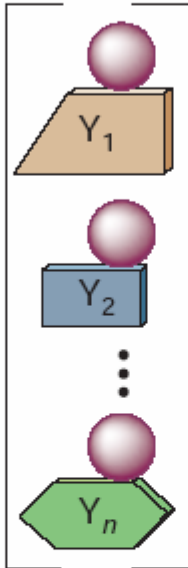
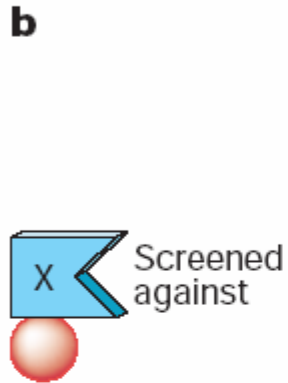
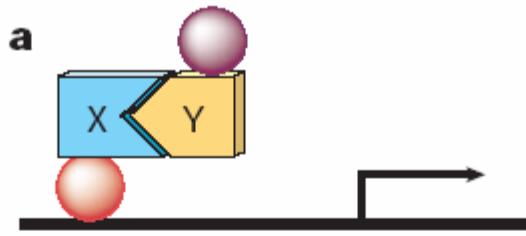
# Genetic interactions

- Synthetic lethality
- Synthetic interaction
  - Choice of phenotypes important

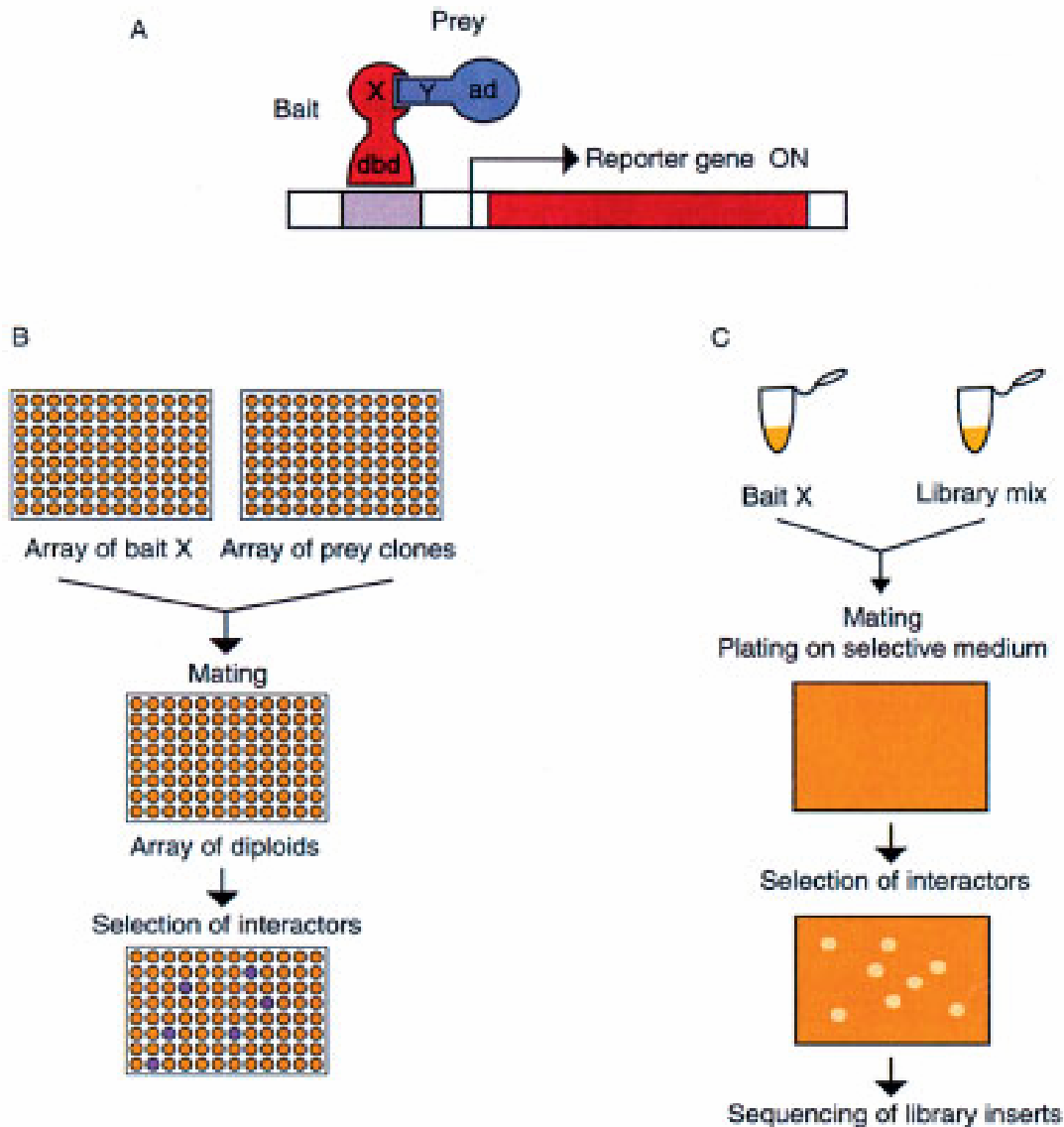
# Physical interactions

- Yeast two hybrid
- Co-IP precipitation
- FRET (doesn't have to be a true physical interaction, but has to be close)
- Protein arrays (can also test **molecular** function directly)

# Yeast two hybrid

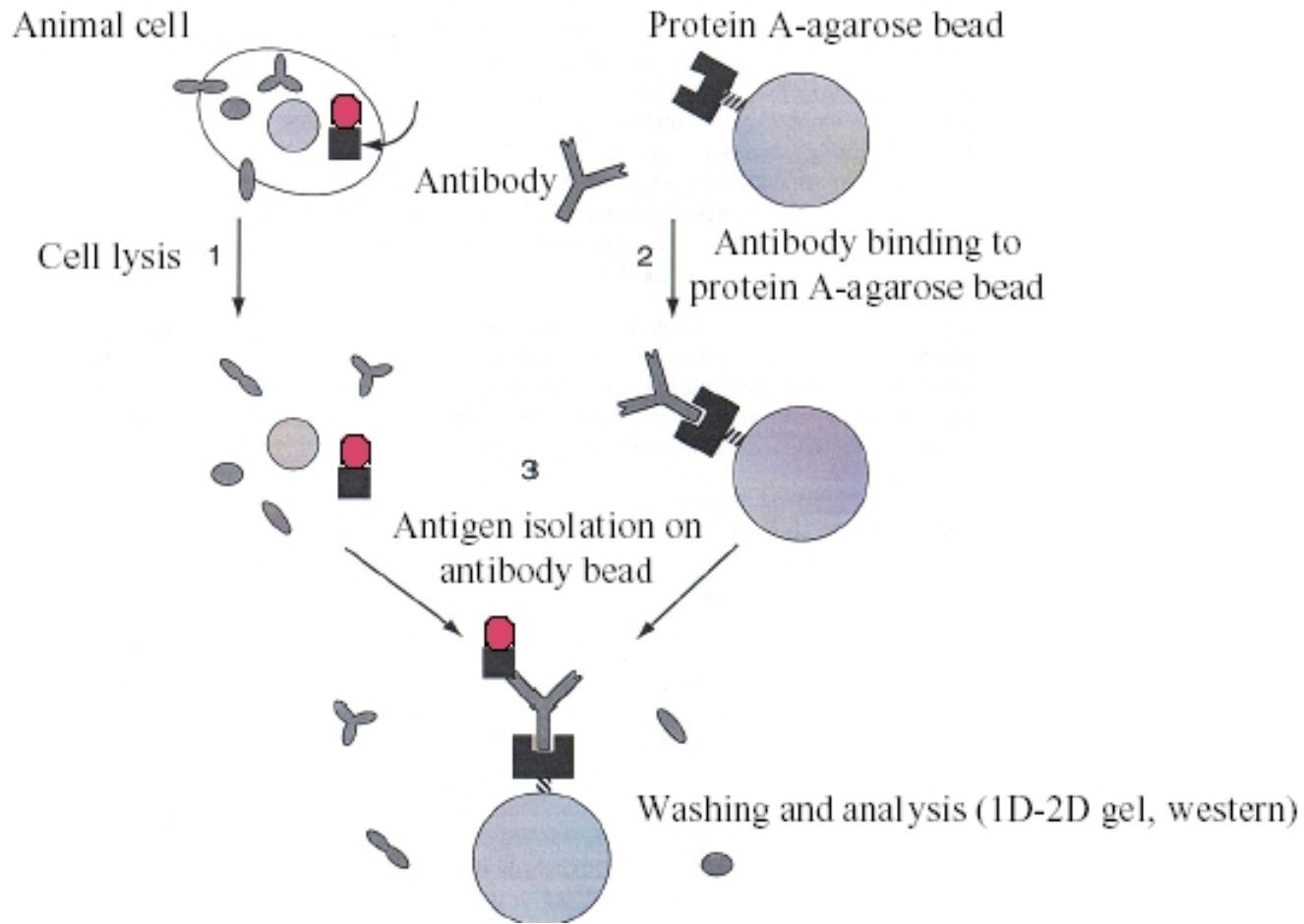


a) DNA-binding and activation domains (circles) are fused to proteins X and Y. The interaction of X and Y leads to reporter gene expression (arrow).

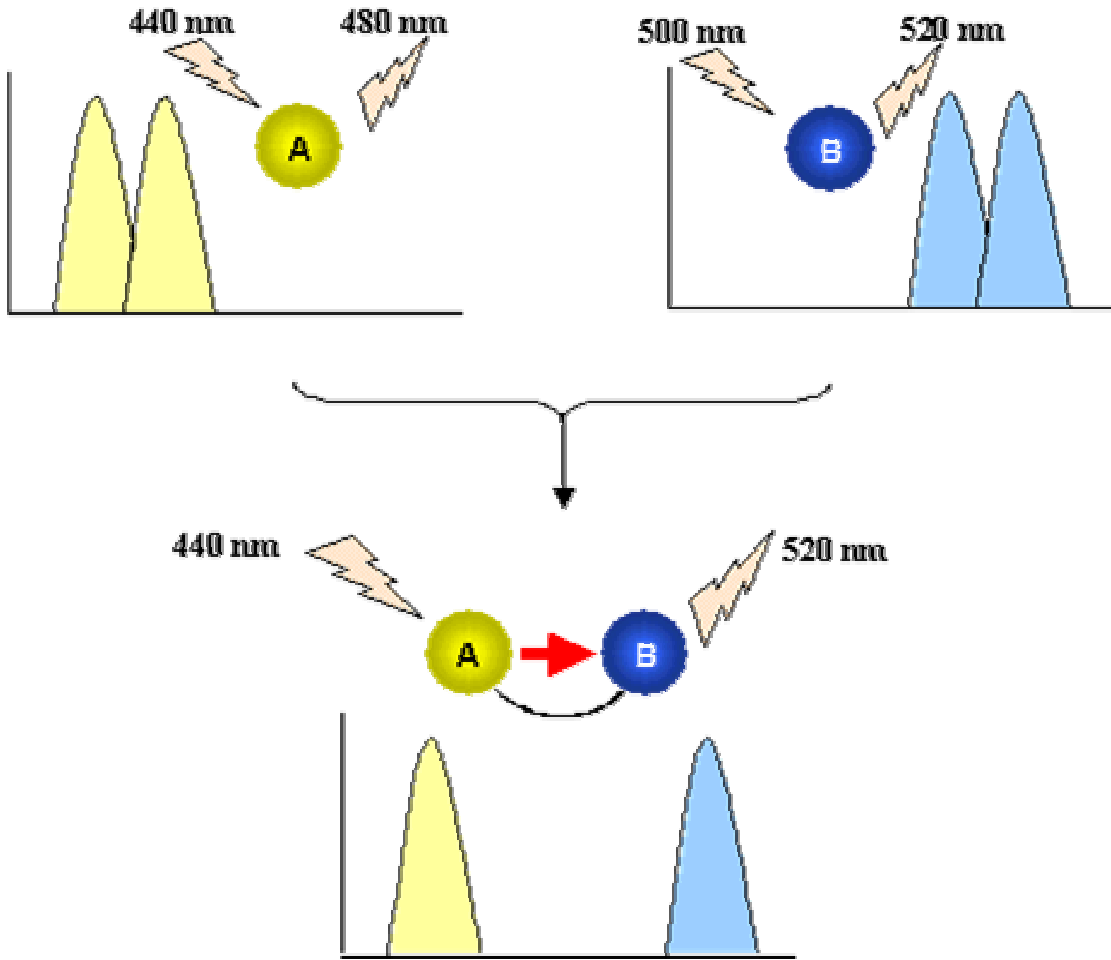


(A) The principle of the yeast two-hybrid system. Protein X is expressed as a fusion to the DBD and constitutes the bait. The DBD-X fusion protein is bound to the operator sites in the promoter region but does not activate transcription of the downstream reporter gene because it lacks an AD. The interaction of DBD-X with its partner Y fused to an AD recruits the AD-Y fusion protein to the promoter where it forms a functional transcriptional activator. Consequently, transcription of the reporter gene is switched on. (B) High-throughput yeast two-hybrid using the matrix approach. A matrix (or array) of prey clones is created by dispensing one yeast clone expressing a given AD-Y fusion protein into each well of a multiwell plate. Using a robot, the array of prey clones is then transferred to a multiwell plate containing yeast that express one DBD-X fusion and prey and bait clones are allowed to mate. Those diploids where DBD-X and a particular AD-Y interact are selected based on expression of a reporter gene, such as  $\beta$ -galactosidase (producing a blue color). (C) In the exhaustive library screening approach one DBD-fused bait X is screened against an entire library and positives are selected based on their ability to grow on selection plates. As opposed to the matrix approach, where each prey can be identified by its position in the array, diploids that have survived selection in the library screening need to be picked up and the library plasmids encoding the interacting prey have to be isolated and sequenced in order to identify the interacting protein. Libraries can be made either from random genomic or cDNA fragments or from full-length ORFs that are cloned separately and then pooled.

# Co-IP

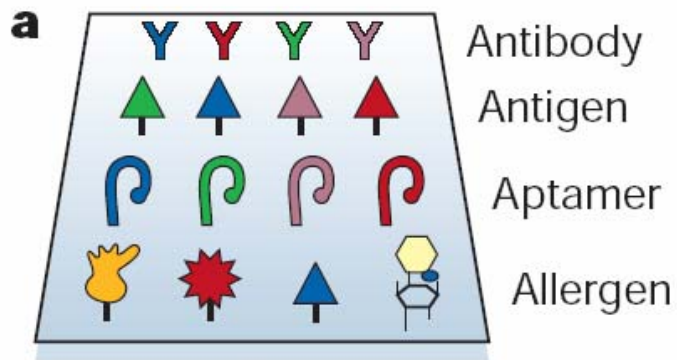


# FRET

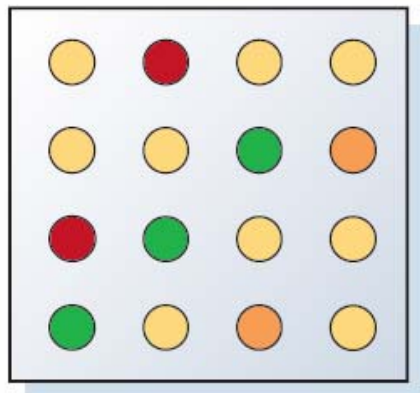


FRET (Fluorescence Resonance Energy Transfer) is a photophysical effect where energy that is absorbed by one fluorescent molecule (donor) is transferred non-radioactively to a second fluorescent molecule (acceptor).

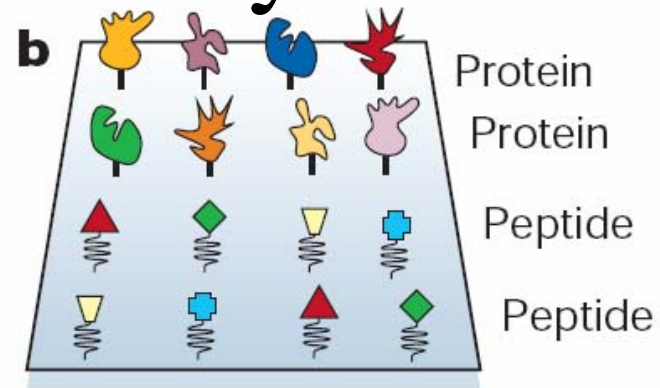
# Protein arrays



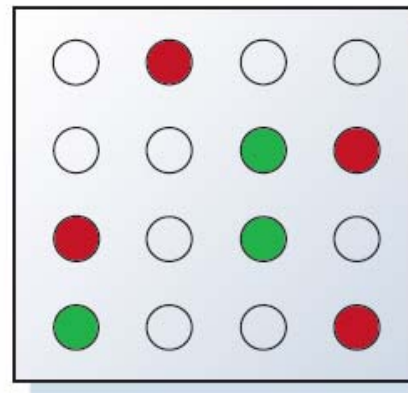
Serum probes  
Cell lysates  
Living cells



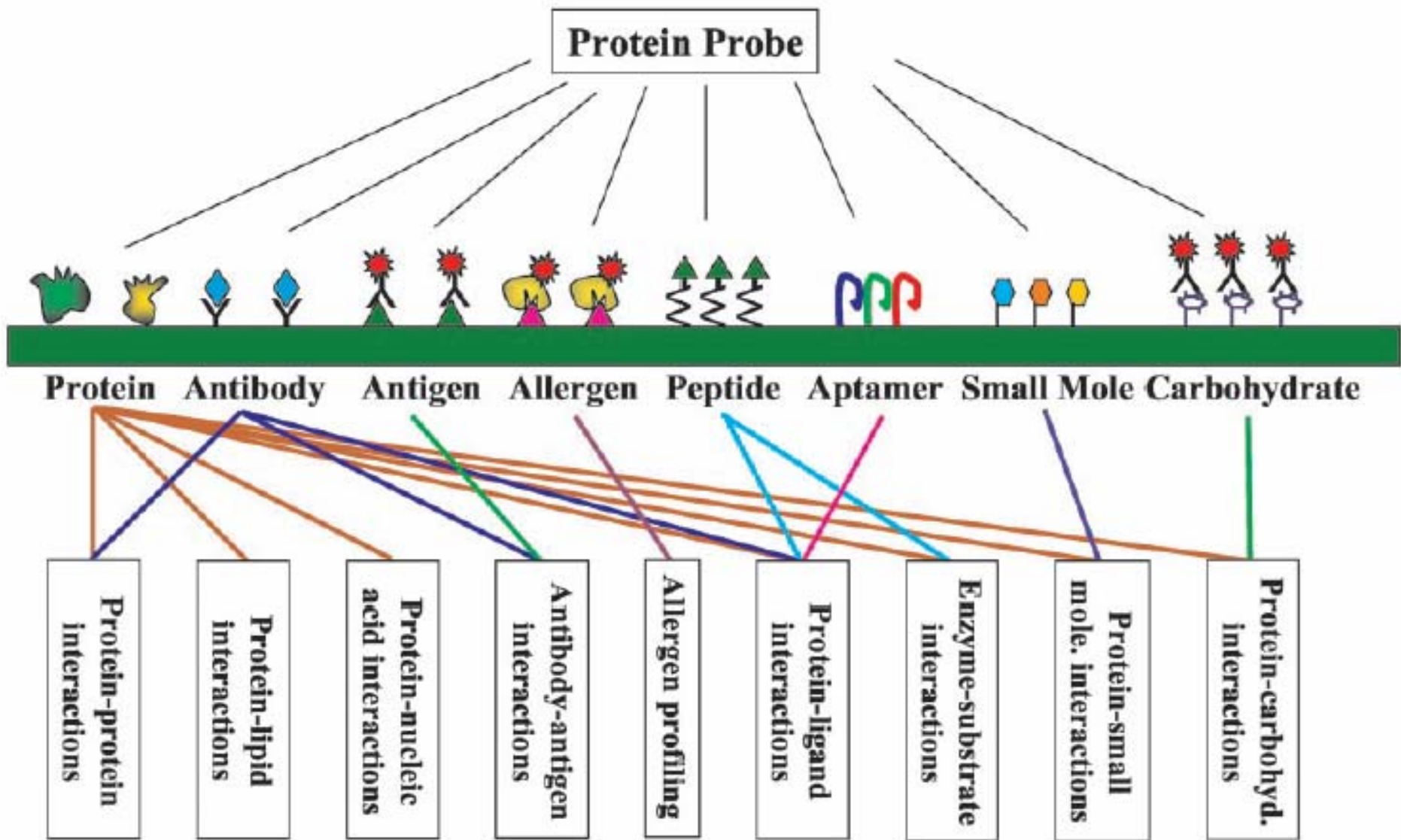
Protein expression level  
Protein profiling  
Diagnostics



Protein probes  
Nucleic acid probes  
Drug probes  
Enzymes



Protein binding properties  
Pathway building  
Drug discovery  
Post-translational modification

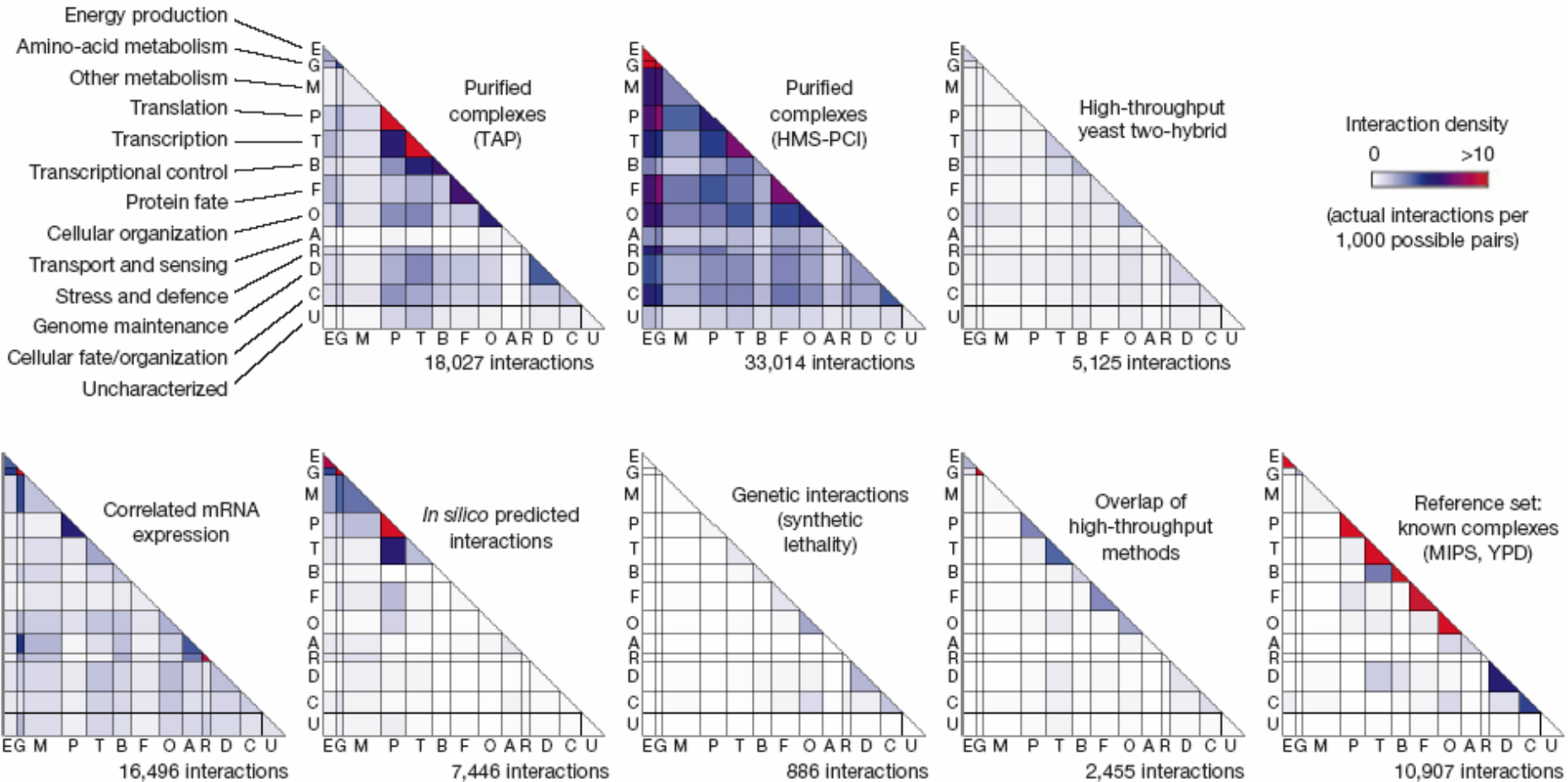


**Figure 3** Protein microarrays and their applications. Ligands, such as proteins, peptides, antibodies, antigens, allergens, and small molecules, are immobilized in high density on modified surfaces to form functional and analytical protein microarrays. These protein microarrays can also be used for various kinds of biochemical analysis.

**TABLE 1** Comparison of different technologies for interaction proteomics

<b>Approach</b>	<b>Application</b>	<b>Advantage</b>	<b>Disadvantage</b>
Yeast two-hybrid	Protein-protein interactions, protein-DNA interactions	High-throughput and systematic to reveal protein interactions	No control over interaction condition; interactions are usually in the nucleus
Affinity tagging/MS	Dissecting protein complexes	In vivo interactions that involve multiple partners	May miss transient or weak interactions, hard to identify false positives
Antibody array	Protein profiling, protein detection, clinical diagnostics	Very sensitive and low sample consumption, great potential in biomarker and drug development	Highly restricted by the quantity and quality of available antibodies; semiquantitative protein detection
Functional protein array	Diverse, e.g., protein-protein, protein-lipid, protein-small molecule, enzyme-substrate interactions as well as drug discovery and posttranslational modifications	Great potentials for analyzing biochemical activities of proteins and high-throughput drug and drug target screening	In vitro assays
Peptide array	Enzyme-substrate interaction and drug discovery	Sensitive and straightforward way to identify epitopes	Expensive to fabricate; in vitro assays
Carbohydrate array	Carbohydrate-mediated molecular recognition and anti-infection response	A new and sensitive way to study carbohydrate-mediated molecular events	In vitro arrays; tough to acquire carbohydrate molecules in pure forms
Small molecule array	Protein-small molecule interaction, drug discovery, enzyme specificity profiling	Minimum small molecule consumption and high sensitivity	In vitro assays; necessary to improve throughput to cover $10^6$ molecules in a normal combinatorial chemistry library

# Interaction coverage - yeast



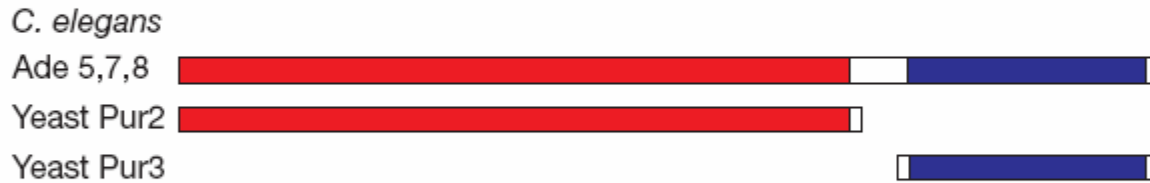
Methods for function prediction  
based on one type of data

# The Rosetta Stone method

General concept



Top sequence = fused domain that's homologous to two separate seqs from another species



Observed gene locations

Method of correlated gene neighbors



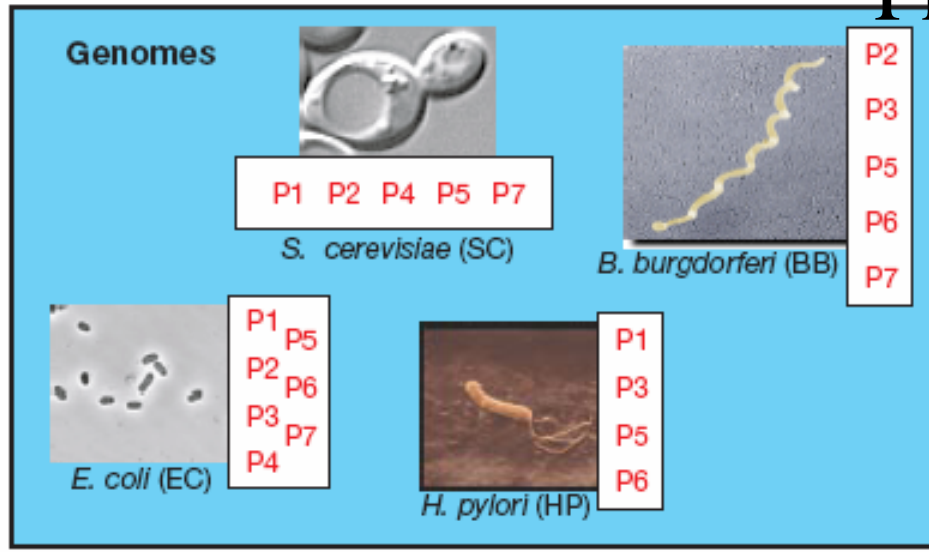
Inferred functional linkage



If two genes (blue and yellow in the figure) are found to be neighbours in several different genomes, a functional linkage may be inferred between the proteins they encode. The method is most robust for microbial genomes but may work to some extent even for human genes where operon-like clusters are observed (see, for example, ref. 26). The gene neighbour method correctly identifies functional links among eight enzymes in the biosynthetic pathway for arginine in *Mycobacterium tuberculosis*.

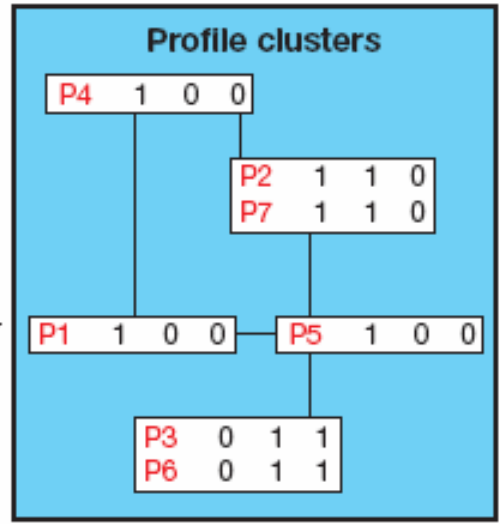
# Phylogenetic profiles method

Proteins are considered functionally linked if they share phylogenetic profiles (presence and absence in genomes). Proteins do not have to be homologous by sequence.



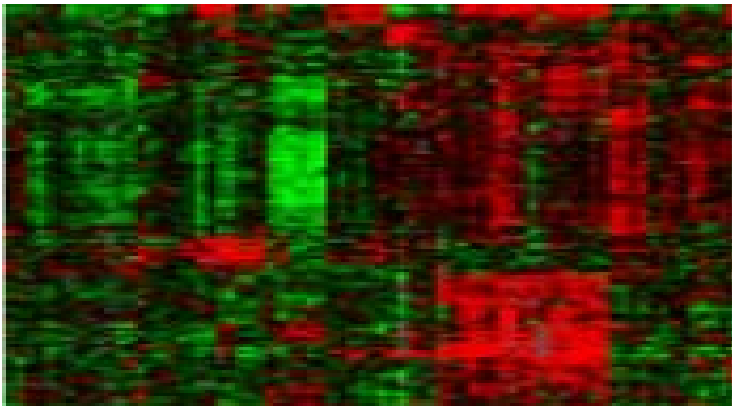
**Phylogenetic profile**

	EC	SC	BB	HP
P1	1	0	1	
P2	1	1	0	
P3	0	1	1	
P4	1	0	0	
P5	1	1	1	
P6	0	1	1	
P7	1	1	0	



**Conclusion** P2 and P7 are functionally linked, P3 and P6 are functionally linked

# Annotation assignment based on co-expression clusters



If enrichment for genes of a specific biological process, can claim unknowns are also involved in that process.

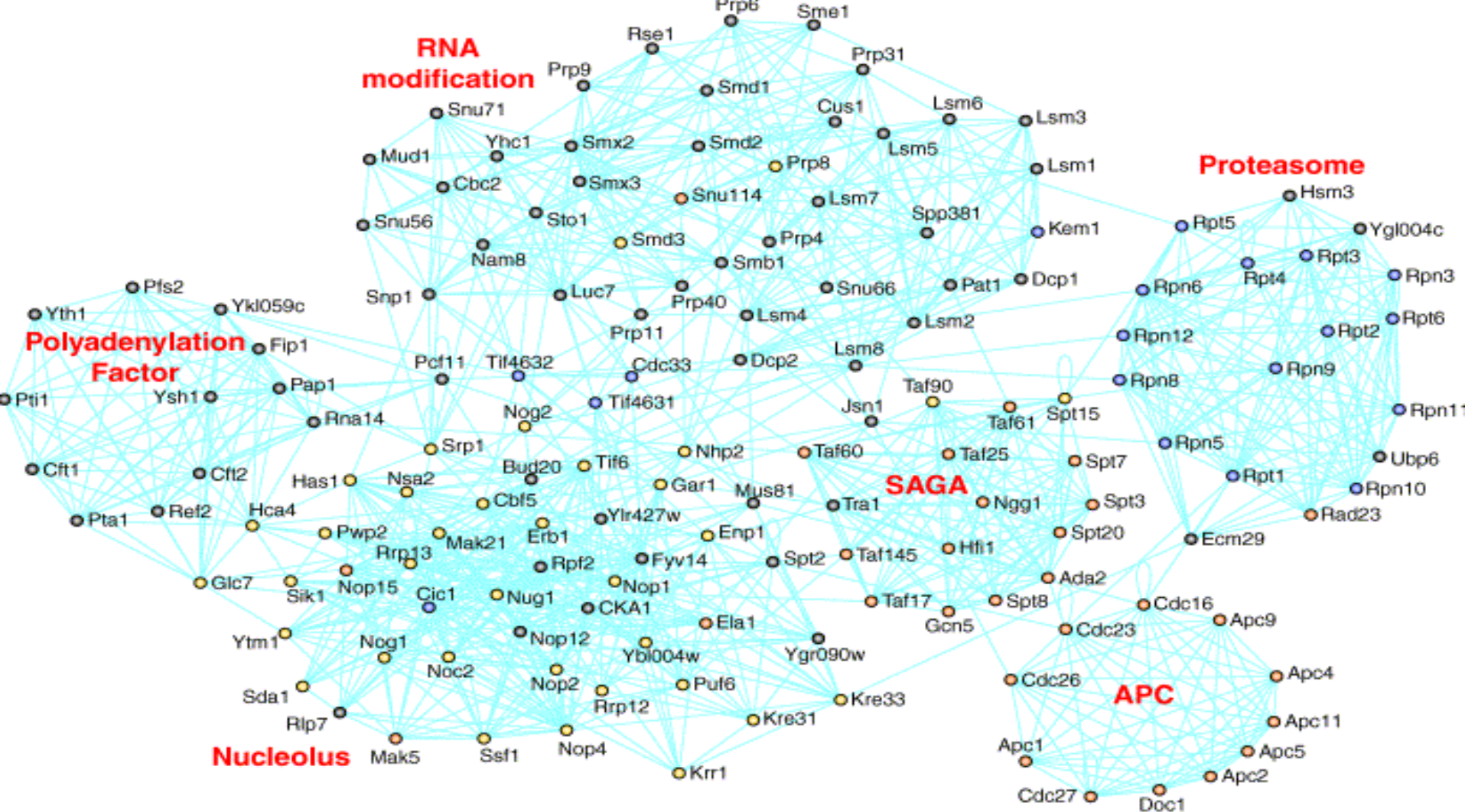
Prob of  $x$  out of  $n$  annotations assigned to the same GO term by chance

P( $x$  or more of  $n$  genes being annotated to a particular term)

$$\sum_{j=x}^n \left( \frac{n!}{j!(n-j)!} \right) \times p^j \times (1-p)^{(n-j)}$$

Num of permutations of  $x$  of  $n$  genes having the annotation

Methods for function prediction  
based on diverse large-scale data



The central densest region of interaction network (15000 interactions) from yeast. The interactions were collected from large-scale studies and the MIPS and BIND databases. Known molecular complexes can be seen clearly, as well as a large, previously unsuspected nucleolar complex.

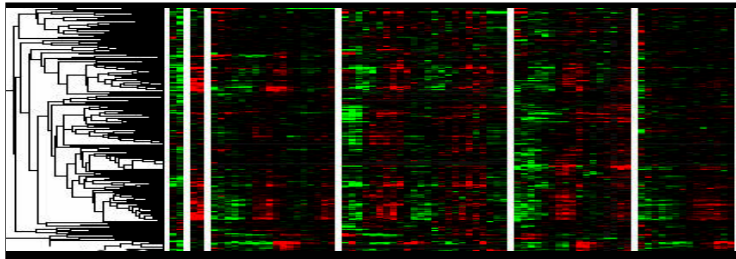
# Abundance of high-throughput data

- Sequence data
  - Transcription factor binding sites
  - Functional domains
- Location data
  - Colocalization
- Physical association data
  - Yeast two hybrid
  - Co-IP
- Genetic relationship data
  - Synthetic lethality
  - Synthetic rescue
  - Synthetic interaction



# Overcoming noise with data integration

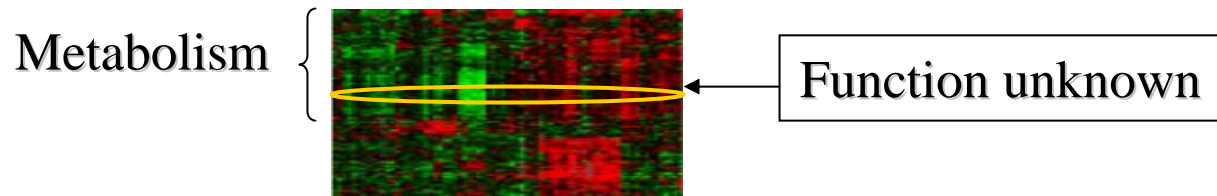
**Gene clustering based on Microarray Analysis:**



**Gene groupings based on Other Experimental Data:**

- TF binding sites
- Affinity precipitation
- Two Hybrid

**Gene grouping based on diverse high-throughput data**



# General representation for gene groupings

**Cluster 1**

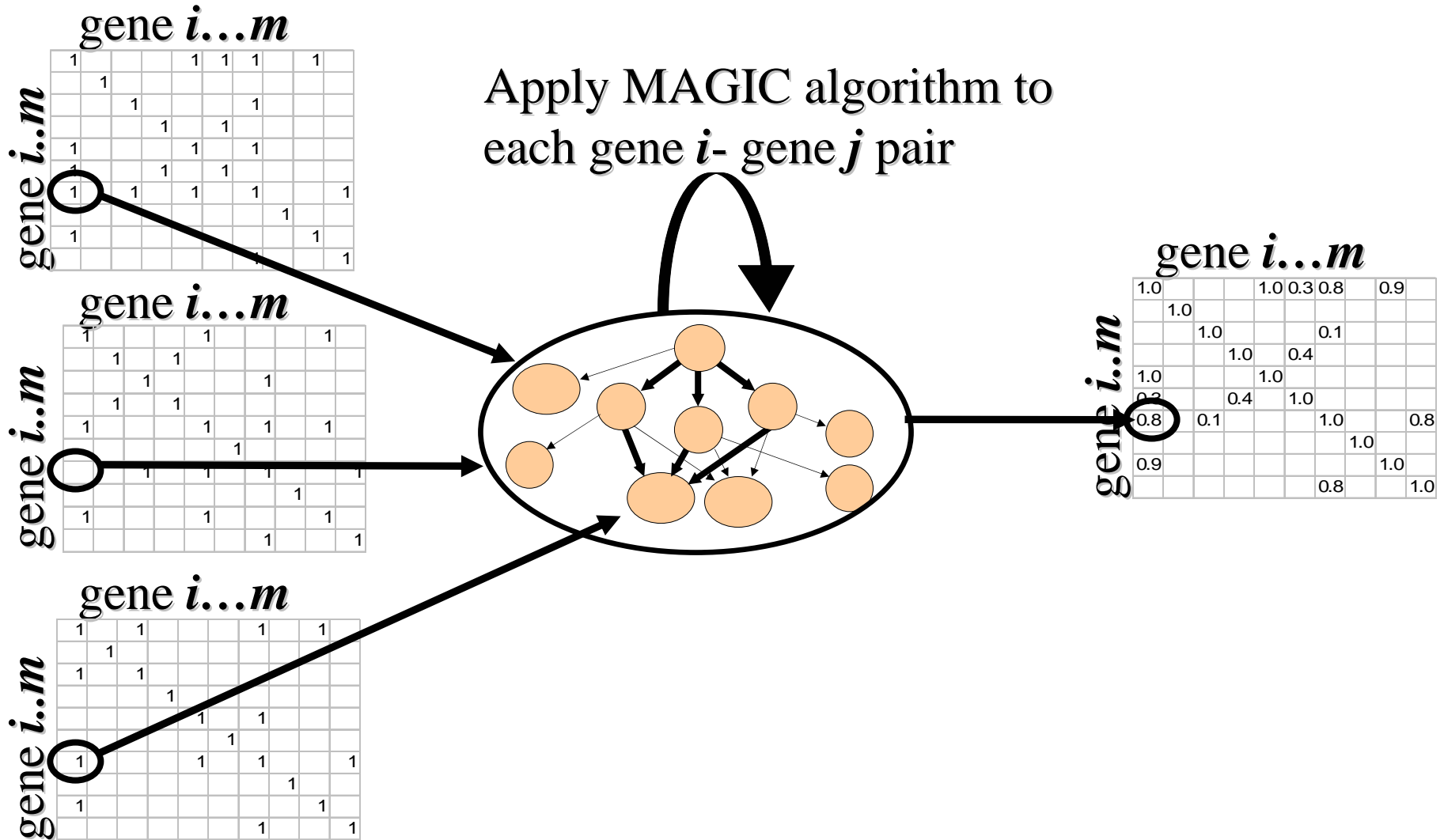
Gene A  
Gene B  
Gene C

**Cluster 2**

Gene A  
Gene D

	Gene A	Gene B	Gene C	Gene D
Gene A	1	1	1	1
Gene B	1	1	1	
Gene C	1	1	1	
Gene D	1			1

# MAGIC: Multi-source Association of Genes by Integration of Clusters



# **A brief intro to Bayesian networks**

- Bayes rule
- Bayesian networks – graphical probabilistic models
- Inference and learning

# Bayesian networks

*"The essence of the Bayesian approach is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise. The canonical example is to imagine that a precocious newborn observes his first sunset, and wonders whether the sun will rise again or not. He assigns equal prior probabilities to both possible outcomes, and represents this by placing one white and one black marble into a bag. The following day, when the sun rises, the child places another white marble in the bag. The probability that a marble plucked randomly from the bag will be white (ie, the child's degree of belief in future sunrises) has thus gone from a half to two-thirds. After sunrise the next day, the child adds another white marble, and the probability (and thus the degree of belief) goes from two-thirds to three-quarters. And so on. Gradually, the initial belief that the sun is just as likely as not to rise each morning is modified to become a near-certainty that the sun will always rise."*

From "The Economist"

# Bayes Rule

$$\text{posterior} = \frac{\text{conditional likelihood} * \text{prior}}{\text{likelihood}}$$

$$P(R=r | e) = \frac{P(e | R=r) P(R=r)}{P(e)} \quad \text{or} \quad P(B|A) = \frac{P(A \cap B)}{P(B)}$$

$P(R=r|e)$  is probability that random variable  $R$  has value  $r$  given evidence  $e$

Denominator is just a normalizing constant that ensures the posterior adds up to 1; it can be computed by summing up the numerator over all possible values of  $R$ , i.e.,  $P(e) = P(R=0, e) + P(R=1, e) + \dots = \sum_r P(e | R=r) P(R=r)$

# Example of Bayes rule use

- Suppose a patient tests positive for disease. What is the P she actually has the disease?
- Depends on:
  - accuracy and sensitivity of the test
  - the background (prior) probability of the disease
- $P(\text{Test}=\text{true} \mid \text{Disease}=\text{true}) = 0.95$  (FNR=5%)
- $P(\text{Test}=\text{true} \mid \text{Disease}=\text{false}) = 0.05$  (FPR=5%)
- $P(\text{Disease}=\text{true}) = 0.01$  (1% of total population)

# Example of Bayes rule use (cont)

$$P(R=r \mid e) = \frac{P(e \mid R=r) P(R=r)}{P(e)}$$

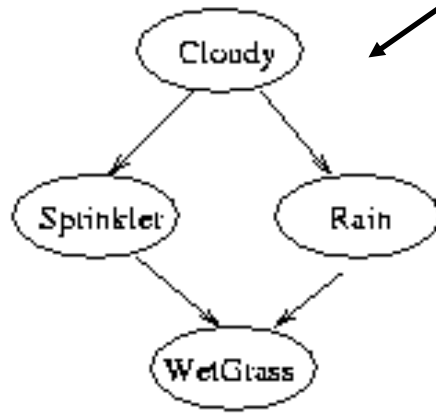
- D - Disease (R in the above equation)
- T - Test (e in the above equation)

$$\begin{aligned} P(D=\text{true} \mid T=\text{true}) &= \frac{P(T=\text{true} \mid D=\text{true}) * P(D=\text{true})}{P(T=\text{true} \mid D=\text{true}) * P(D=\text{true}) + P(T=\text{true} \mid D=\text{false}) * P(D=\text{false})} \\ &= \frac{0.95 * 0.01}{0.95 * 0.01 + 0.05 * 0.99} = \frac{0.0095}{0.0590} = 0.161 \end{aligned}$$

# The “famous” sprinkler Bayes net

Bayes nets are graphical probabilistic models

	$P(C=F)$	$P(C=T)$
	0.5	0.5



C	$P(S=F)$	$P(S=T)$
F	0.5	0.5
T	0.9	0.1

C	$P(R=F)$	$P(R=T)$
F	0.8	0.2
T	0.2	0.8

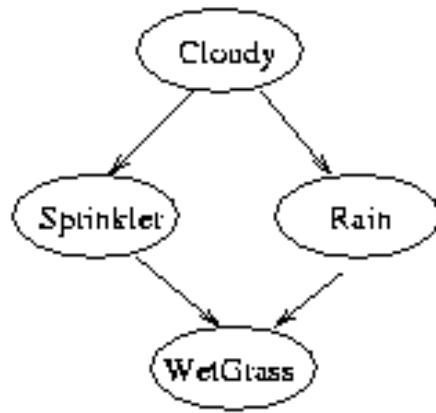
S	R	$P(W=F)$	$P(W=T)$
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Conditional probability tables contain the priors

# The “famous” sprinkler Bayes net

	$P(C=F)$	$P(C=T)$
	0.5	0.5

← Prior probability that it is cloudy



C	$P(S=F)$	$P(S=T)$
F	0.5	0.5
T	0.9	0.1

C	$P(R=F)$	$P(R=T)$
F	0.8	0.2
T	0.2	0.8

← Conditional probability that it rains when it's cloudy

S	R	$P(W=F)$	$P(W=T)$
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

← Probability that grass is wet when the sprinkler is off and it rains

Conditional probability that it rains when it's cloudy

Probability that grass is wet when the sprinkler is off and it rains

$$P(B|A) = \frac{P(A \cap B)}{P(B)}$$

# Inference

Observe: grass is wet

Two possible causes: either it is raining, or the sprinkler is on. Which is more likely? Use Bayes' rule to compute the posterior probability of each explanation.

$$\Pr(S = 1|W = 1) = \frac{\Pr(S = 1, W = 1)}{\Pr(W = 1)} = \frac{\sum_{c,r} \Pr(C = c, S = 1, R = r, W = 1)}{\Pr(W = 1)} = 0.2781/0.6471 = 0.430$$

$$\Pr(R = 1|W = 1) = \frac{\Pr(R = 1, W = 1)}{\Pr(W = 1)} = \frac{\sum_{c,s} \Pr(C = c, S = s, R = 1, W = 1)}{\Pr(W = 1)} = 0.4581/0.6471 = 0.708$$

where

$$\Pr(W = 1) = \sum_{c,r,s} \Pr(C = c, S = s, R = r, W = 1) = 0.6471$$

is a normalizing constant (probability (likelihood) of the data).

$\Rightarrow$  it is more likely that the grass is wet because it is raining: the likelihood ratio is  $0.7079/0.4298 = 1.647$ .

# Bayes net encodes conditional independence relationships

A node is independent of its ancestors given its parents

By the chain rule of probability, the joint probability of all the nodes in the graph above is:

$$P(C, S, R, W) = P(C) * P(S|C) * P(R|C,S) * P(W|C,S,R)$$

Using conditional independence relationships, rewrite as:

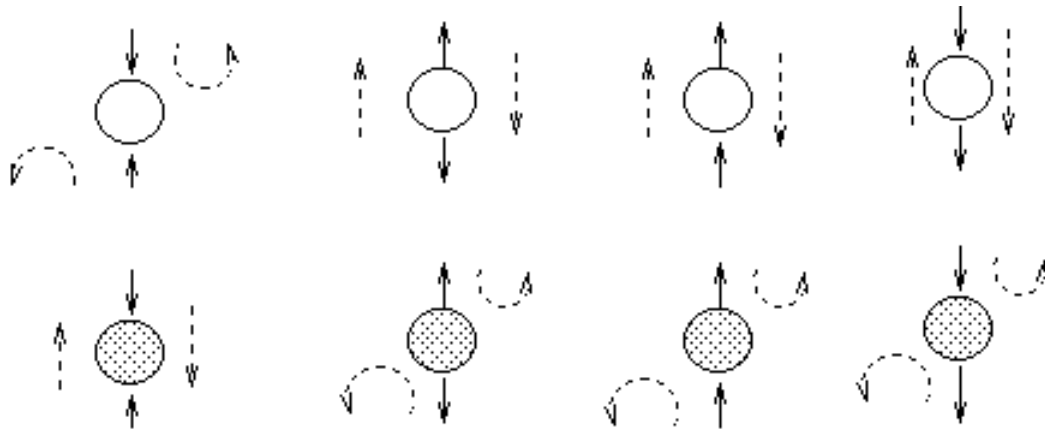
$$P(C, S, R, W) = P(C) * P(S|C) * P(R|C) * P(W|S,R)$$

simplified third term because R is independent of S given its parent C

simplified last term because W is independent of C given its parents S and R

# Conditional independence: “Bayes Ball”

Two (sets of) nodes A and B are conditionally independent (d-separated) given a set C if and only if there is no way for a ball to get from A to B in the graph. Hidden nodes are unshaded; observed nodes are shaded.



1<sup>st</sup> column: X is a "leaf" with two parents:

if X is hidden, its parents are marginally independent => the ball does not pass through

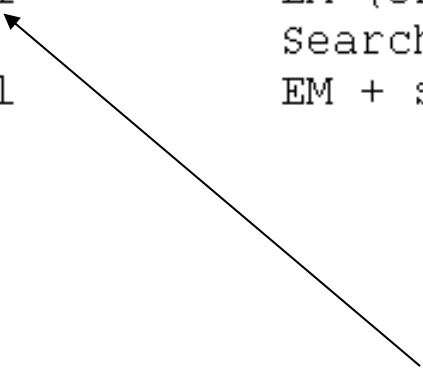
if X is observed, the parents become dependent, and the ball does pass through, because of the explaining away phenomenon.

# Learning Bayesian networks

- Two learning problems: structure + CPTs

Structure	Observability	Method
Known	Full	Maximum Likelihood Estimation
Known	Partial	EM (or gradient ascent)
Unknown	Full	Search through model space
Unknown	Partial	EM + search through model space

Due either to missing data  
or to hidden nodes



# Learning: ML estimation

Goal of learning is to find the values of the parameters of each CPD which maximizes the likelihood of the training data. The normalized log-likelihood of the training set  $D$  is:

$$L = \frac{1}{N} \sum_{i=1}^m \sum_{l=1}^S \log P(X_i | \text{Pa}(X_i), D_l)$$

What does this mean intuitively?

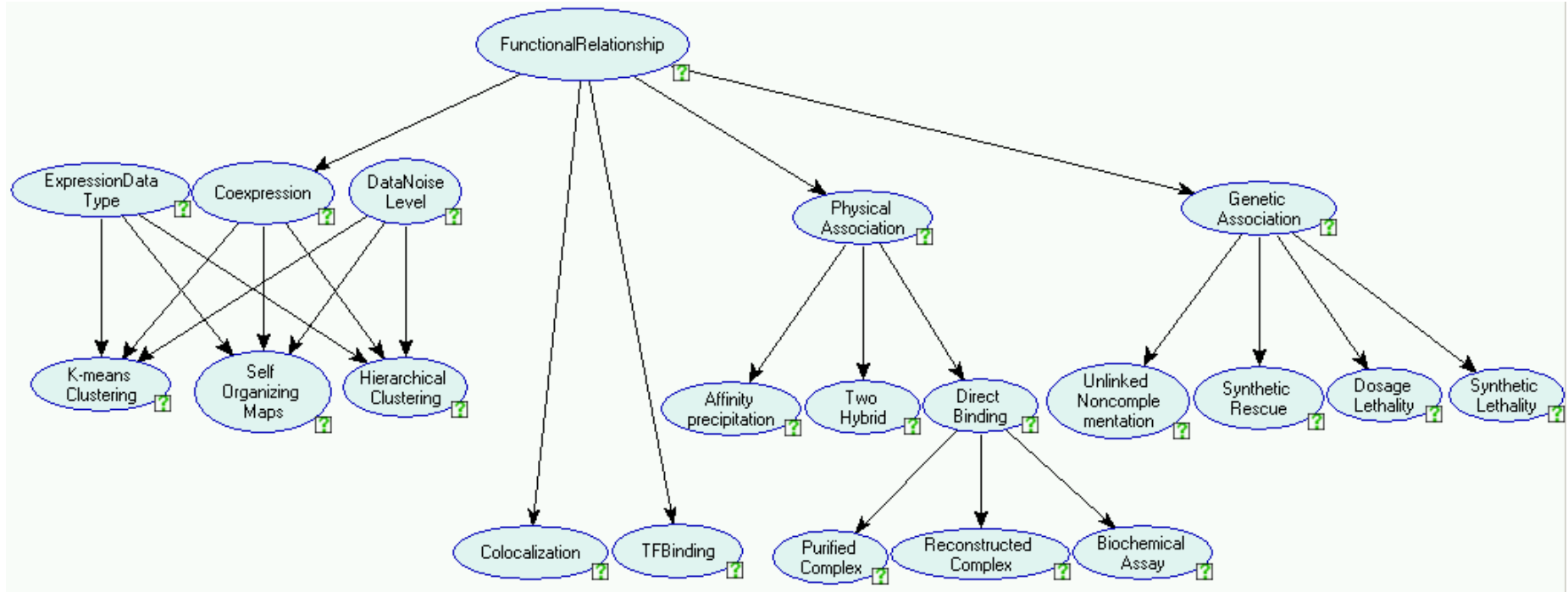
Just counting!

# Learning: EM

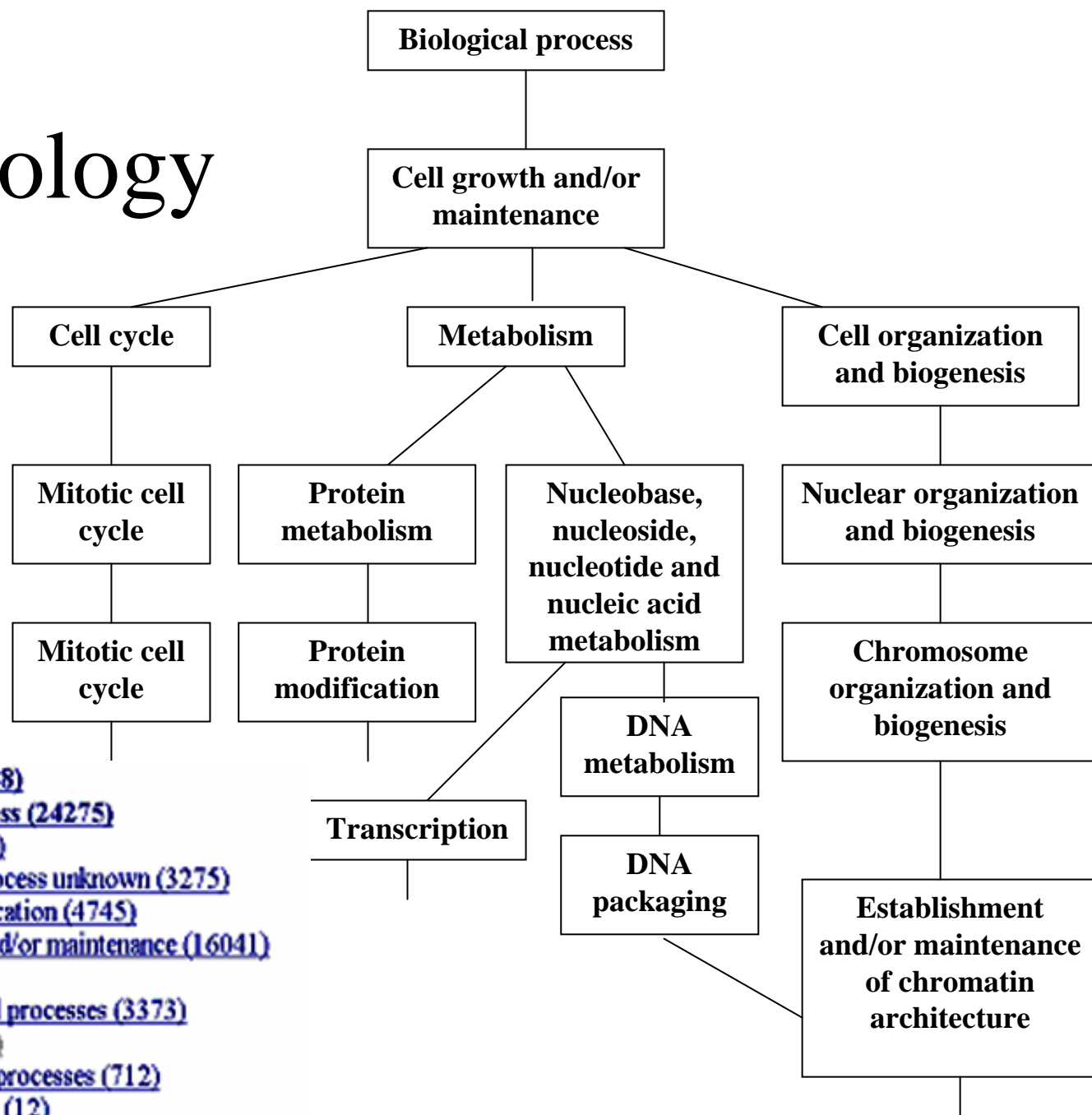
- EM algorithm finds a (locally) optimal Maximum Likelihood Estimate of the parameters.
- Idea behind EM: If we knew the values of all the nodes, learning (the M step) would be easy (ML).
  - E step: compute the expected values of all the nodes using an inference algorithm, then treat these expected values as observed.
  - Given the expected counts, maximize the parameters, then recompute the expected counts, etc.
- EM is guaranteed to converge to a local maximum of the likelihood surface.

Back to gene function prediction  
with Bayes nets

# MAGIC Bayesian network



# Gene Ontology



GO:0003673 : Gene\_Ontology (31688)

GO:0008150 : biological\_process (24275)

GO:0007610 : behavior (216)

GO:0000004 : biological\_process\_unknown (3275)

GO:0007154 : cell\_communication (4745)

GO:0008151 : cell\_growth\_and/or\_maintenance (16041)

GO:0016265 : death (385)

GO:0007275 : developmental\_processes (3373)

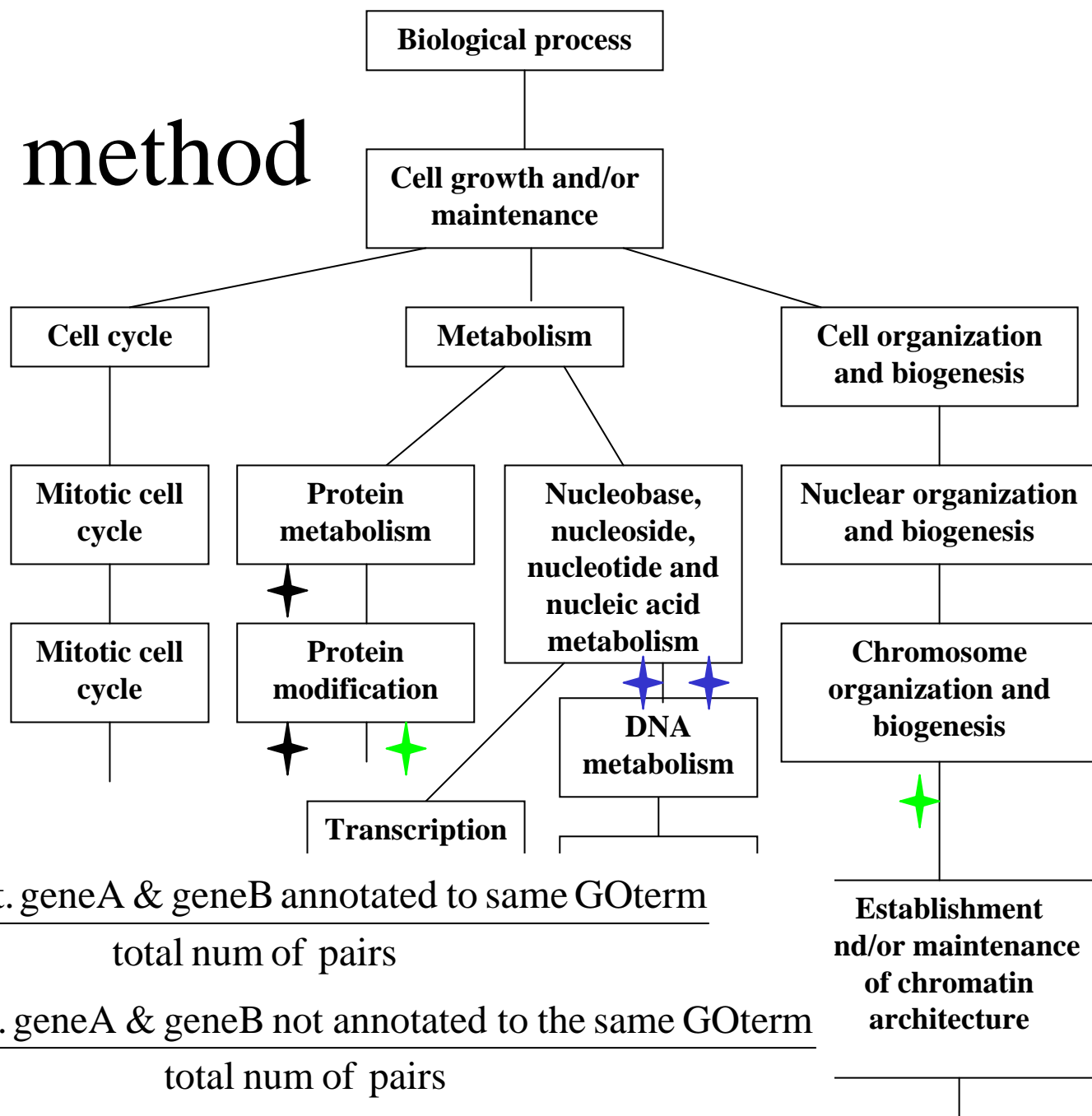
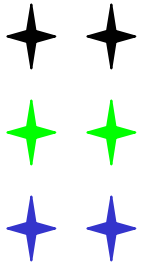
GO:0008371 : obsolete (761)

GO:0007582 : physiological\_processes (712)

GO:0016032 : viral\_life\_cycle (12)

# Evaluation method

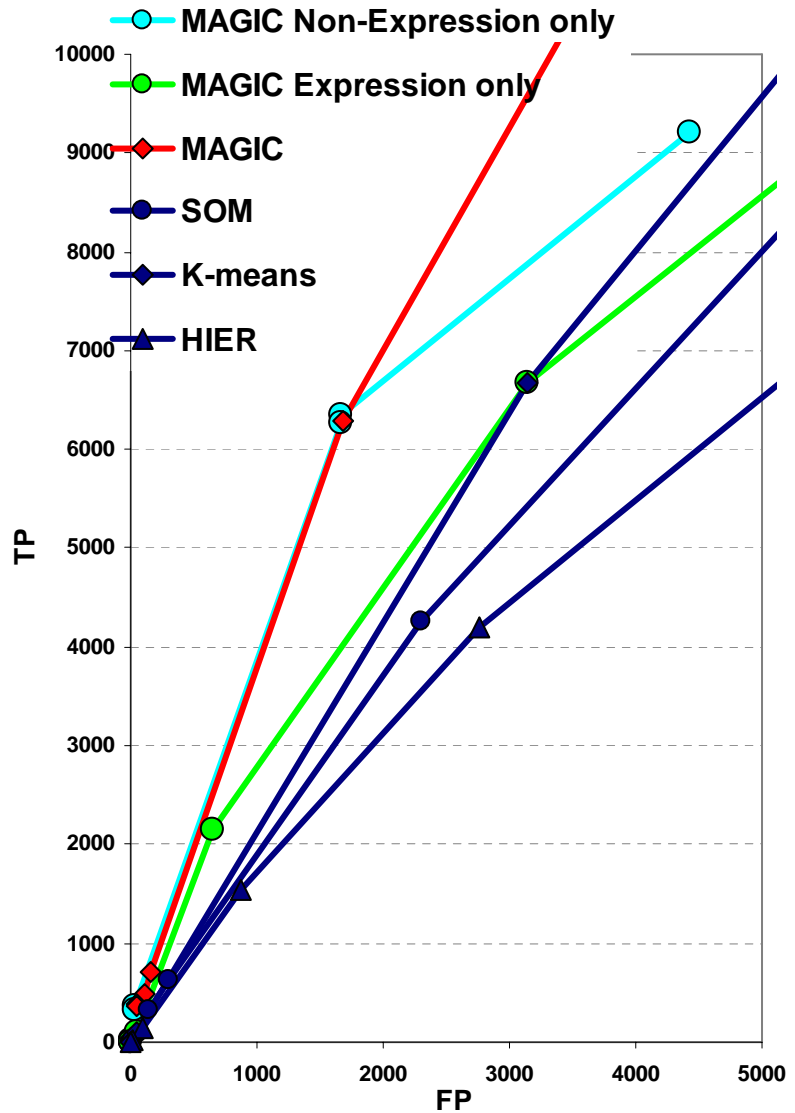
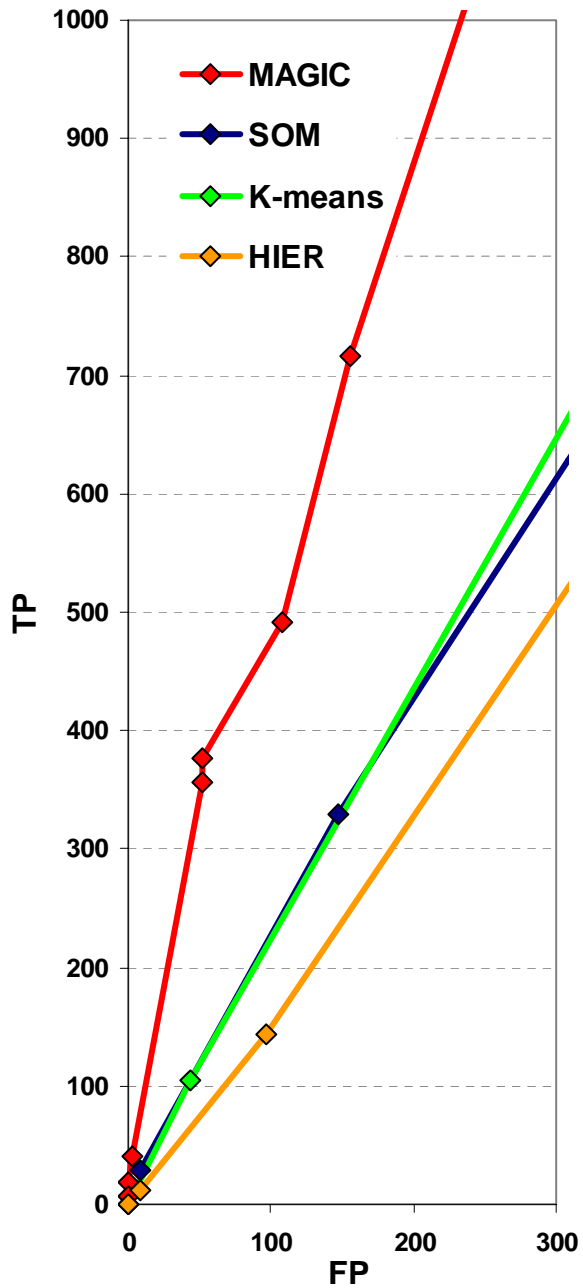
## Predicted Gene Pairs



$$\text{proportionTP} = \frac{\text{num pairs s.t. geneA \& geneB annotated to same GOterm}}{\text{total num of pairs}}$$

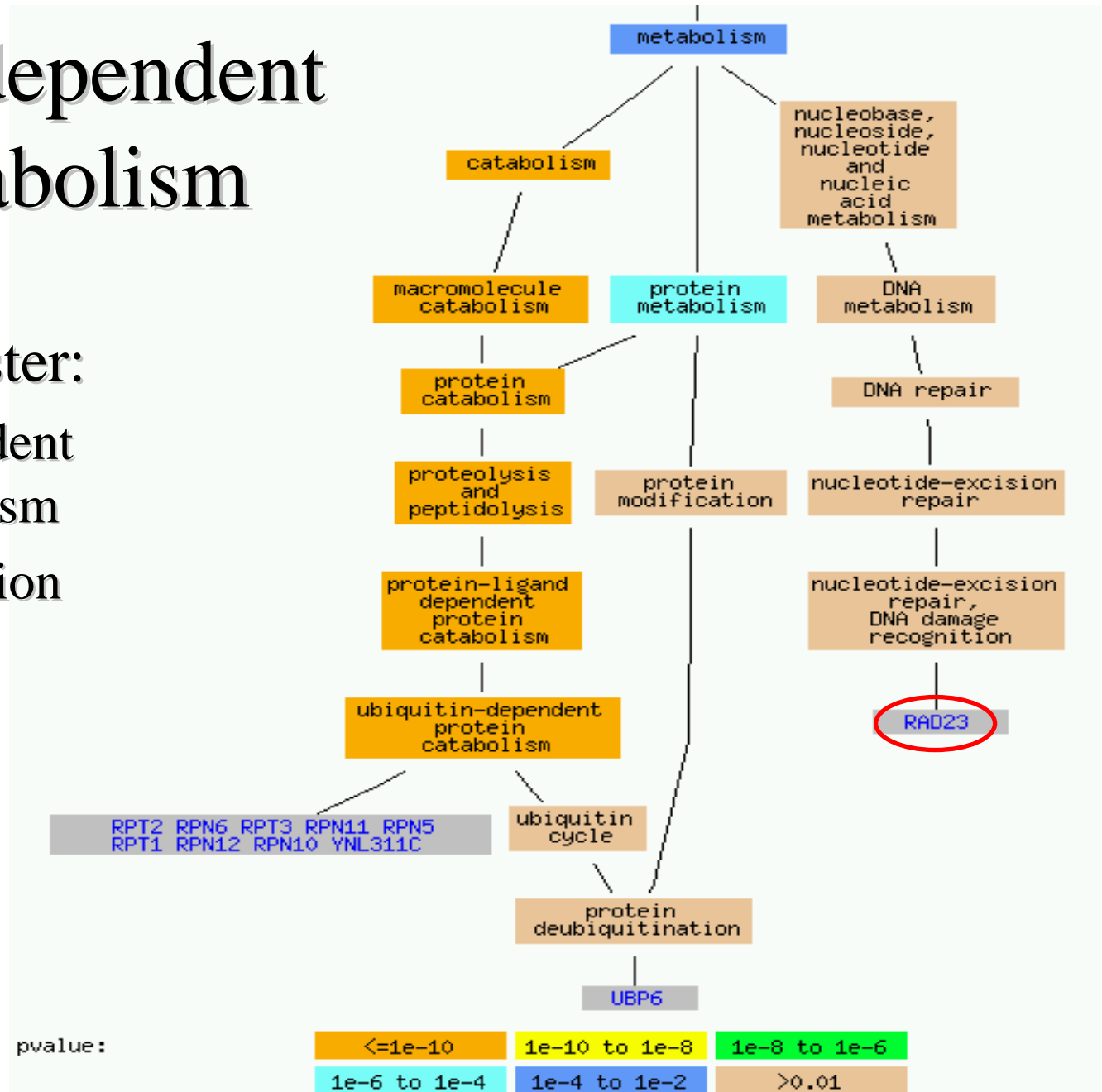
$$\text{proportionFP} = \frac{\text{num pairs s.t. geneA \& geneB not annotated to the same GOterm}}{\text{total num of pairs}}$$

MAGIC performs better than input methods over a range of FP levels



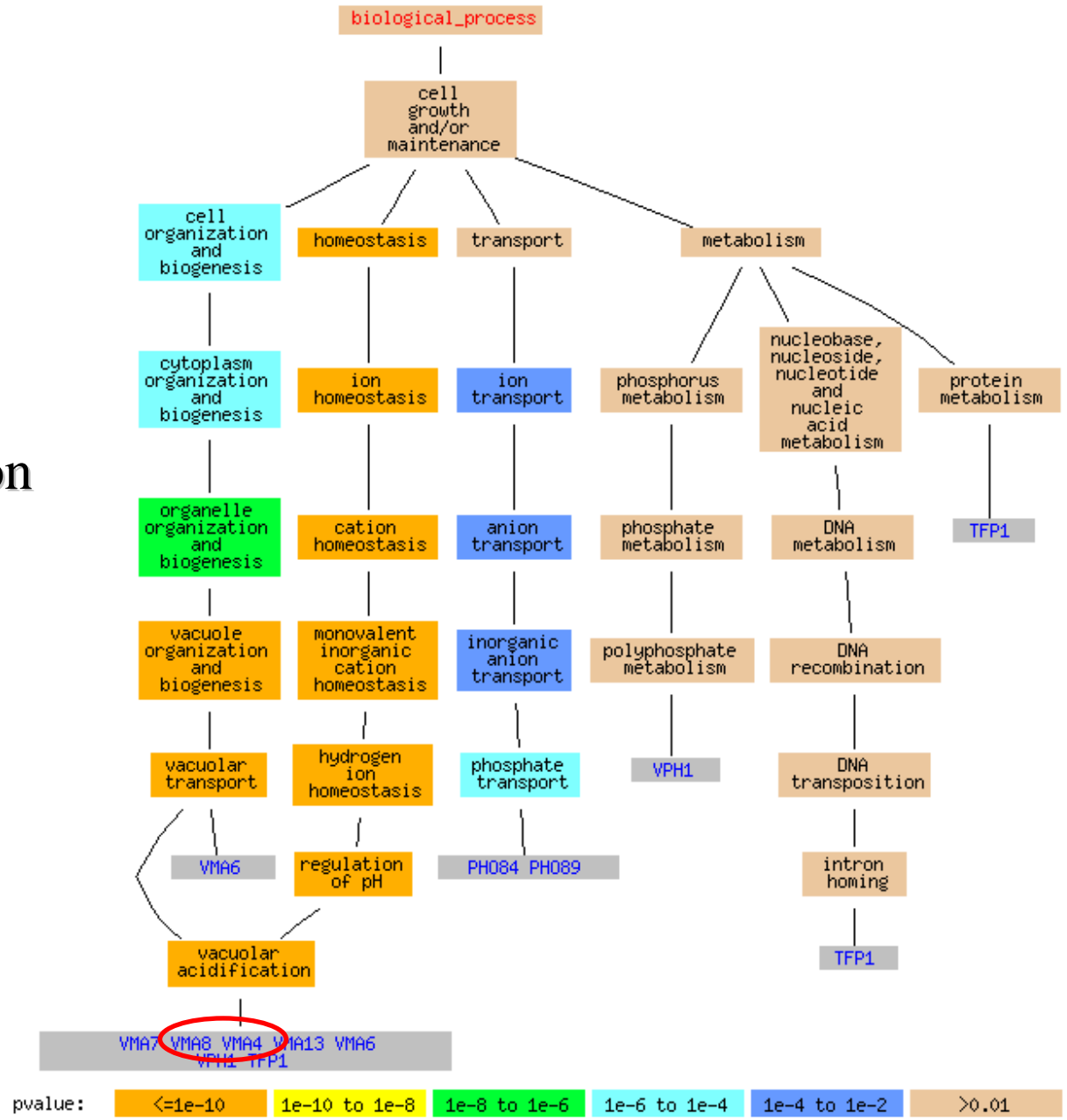
# Ubiquitin-dependent protein catabolism

10 genes in cluster:  
 9 ubiquitin-dependent protein catabolism  
 1 nucleotide-excision repair



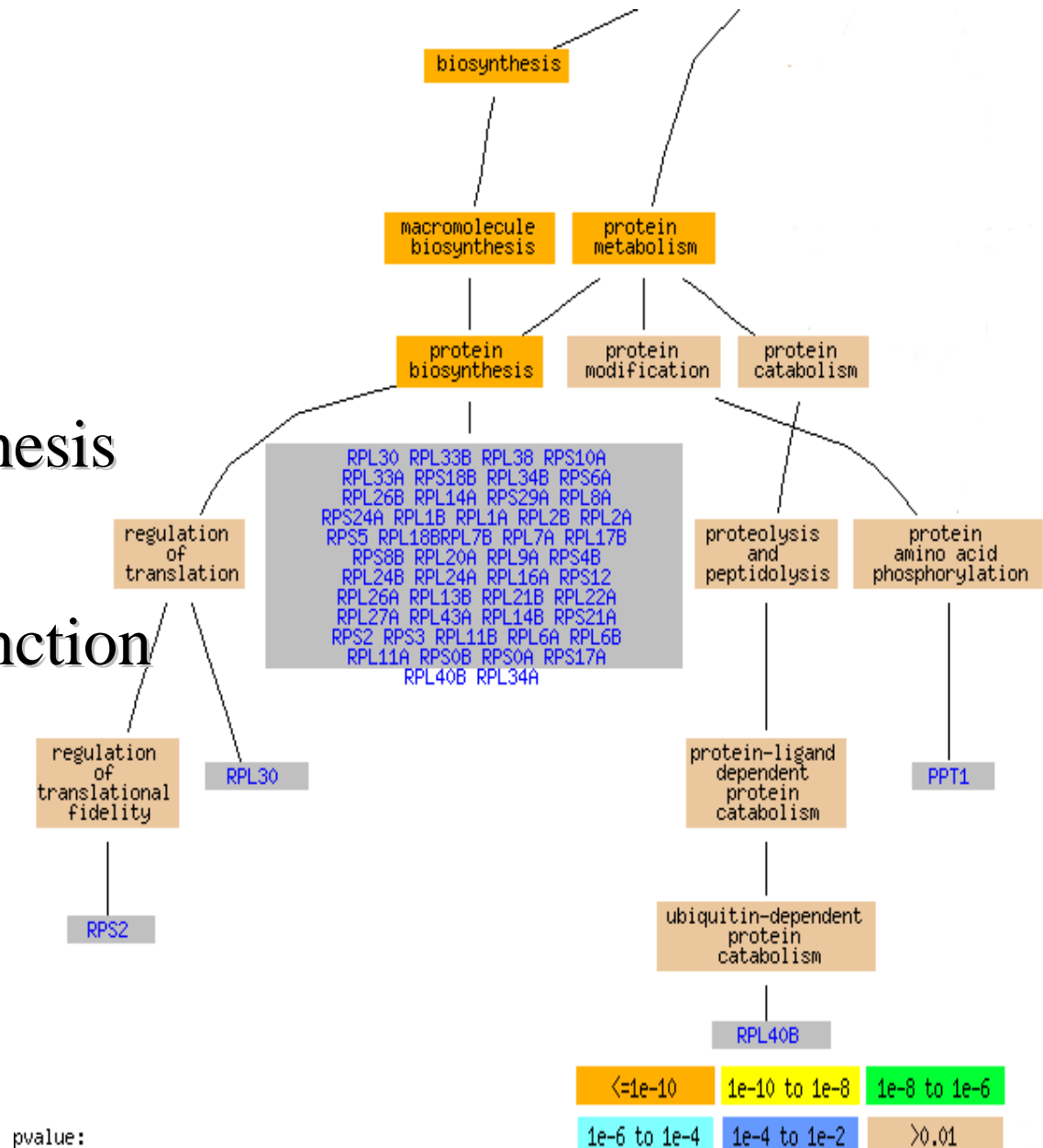
# Vacuolar acidification cluster

9 genes in cluster:  
 7 vacuolar acidification  
 2 phosphate transport



# Protein biosynthesis cluster

- 49 protein biosynthesis
- 9 other functions
- 10 unknown ← function prediction



# Next Steps

- Experiment with automatically learning the weights (in progress)
- MAGIC for specific processes (in progress)
- Test best predictions experimentally
- Integrate system with the *Saccharomyces* Genome Database for public use