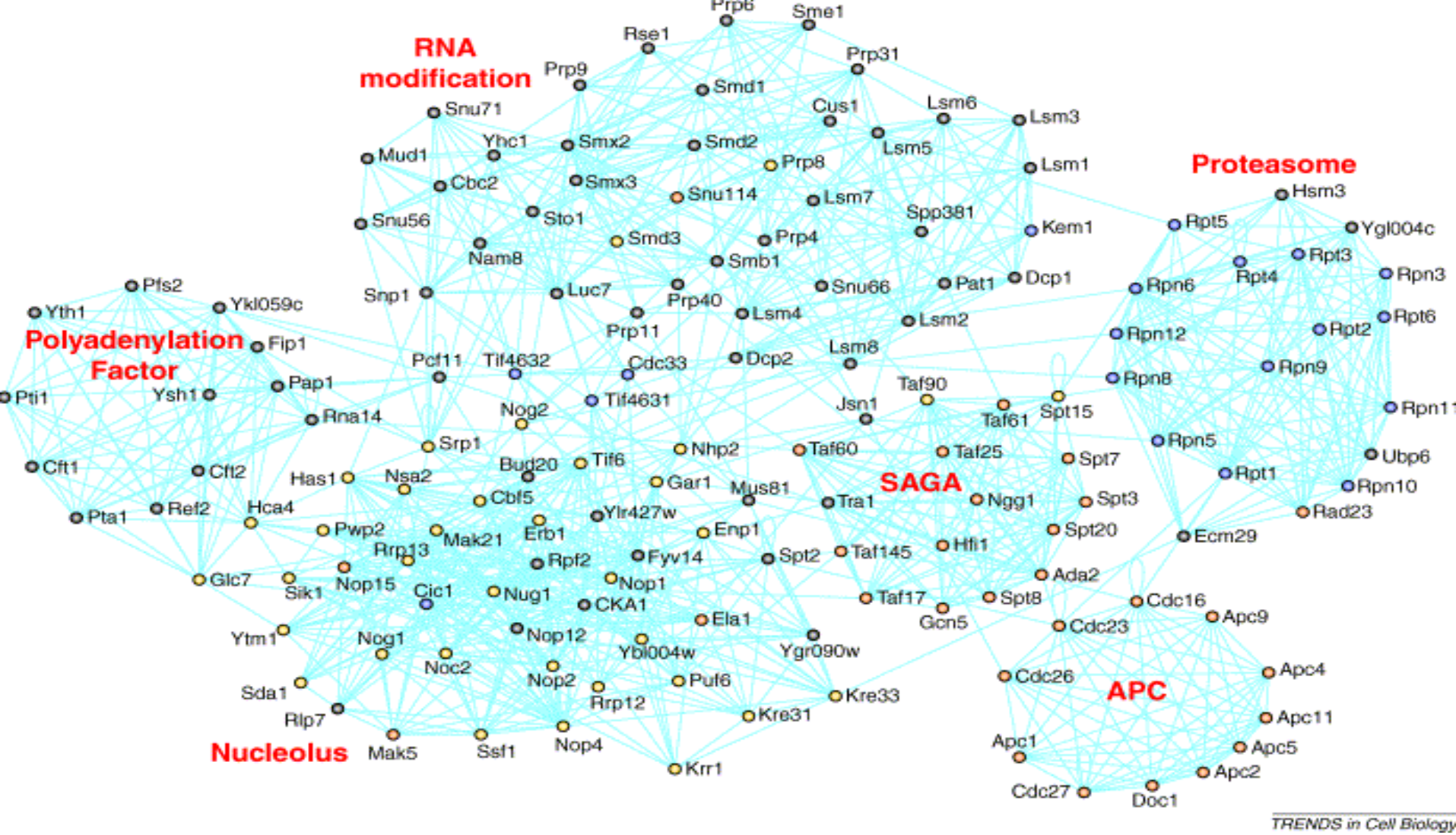
Methods for function prediction based on diverse large-scale data

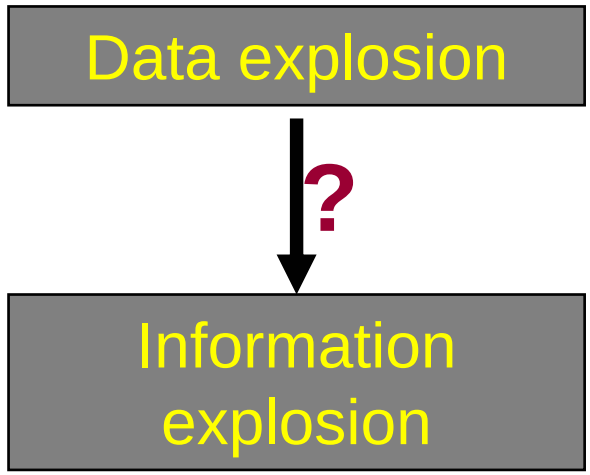
Some slides borrowed from Serafim Batzoglou, Stanford University



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.

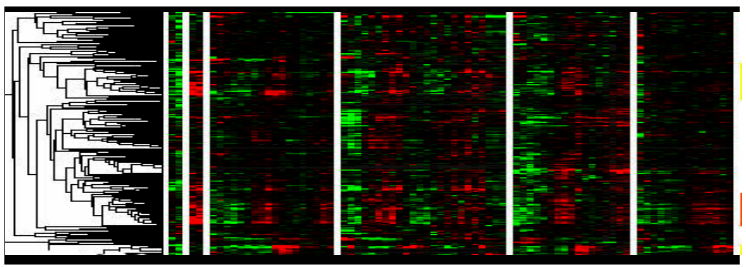
Abundance of high-throughput data

- 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

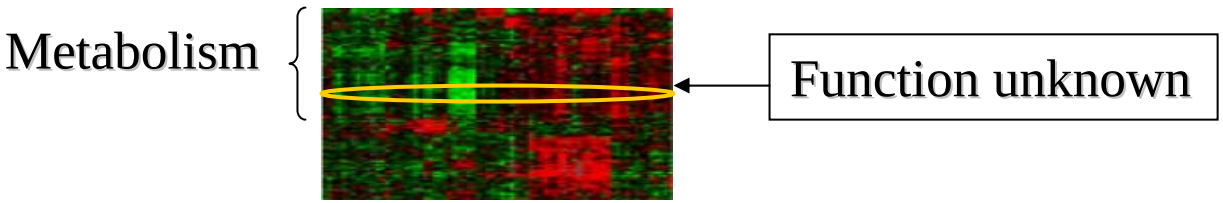
Gene clustering based on Microarray Analysis:



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

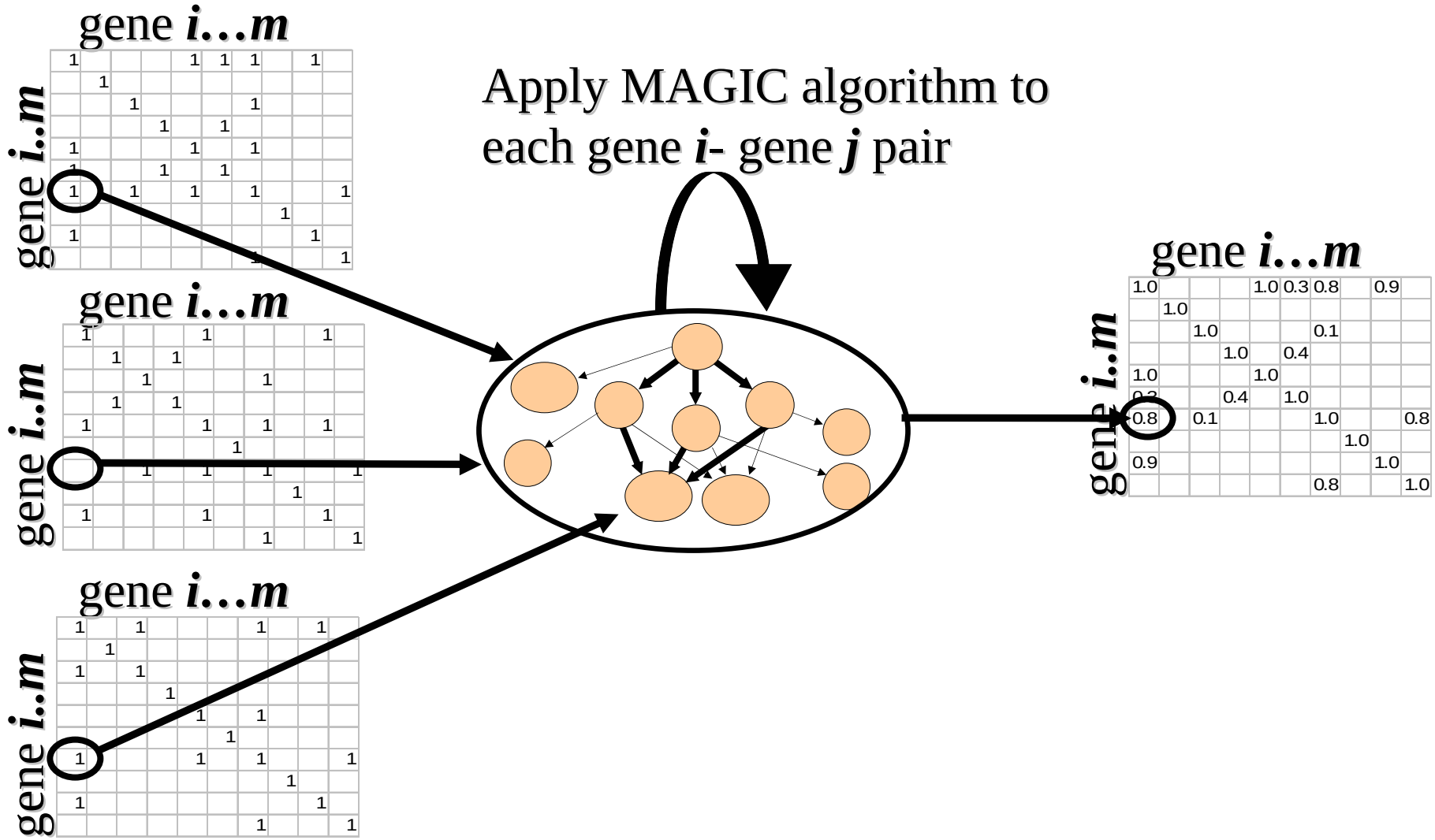
Gene A
Gene B
Gene C

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

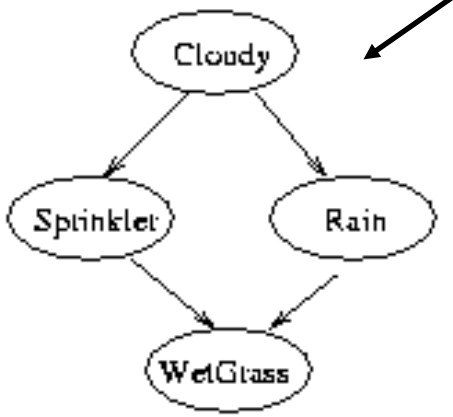


A brief intro to Bayesian networks

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

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

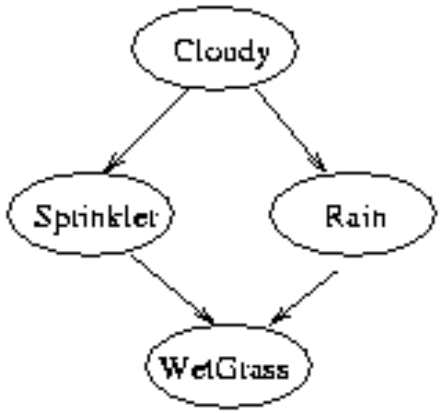
Conditional probability tables contain the priors

Murphy, K. “**A Brief Introduction to Graphical Models and Bayesian Networks**”

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

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

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

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

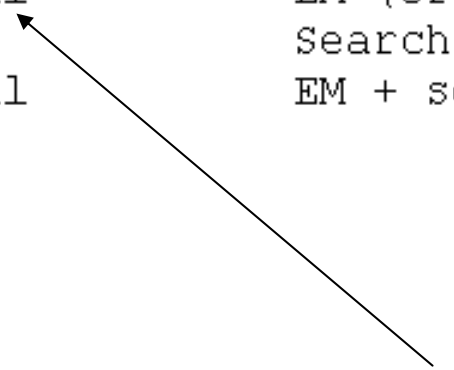
is a normalizing constant (probability (likelihood) of the data).

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

Learning Bayesian networks

- Two learning problems: structure + CPTs

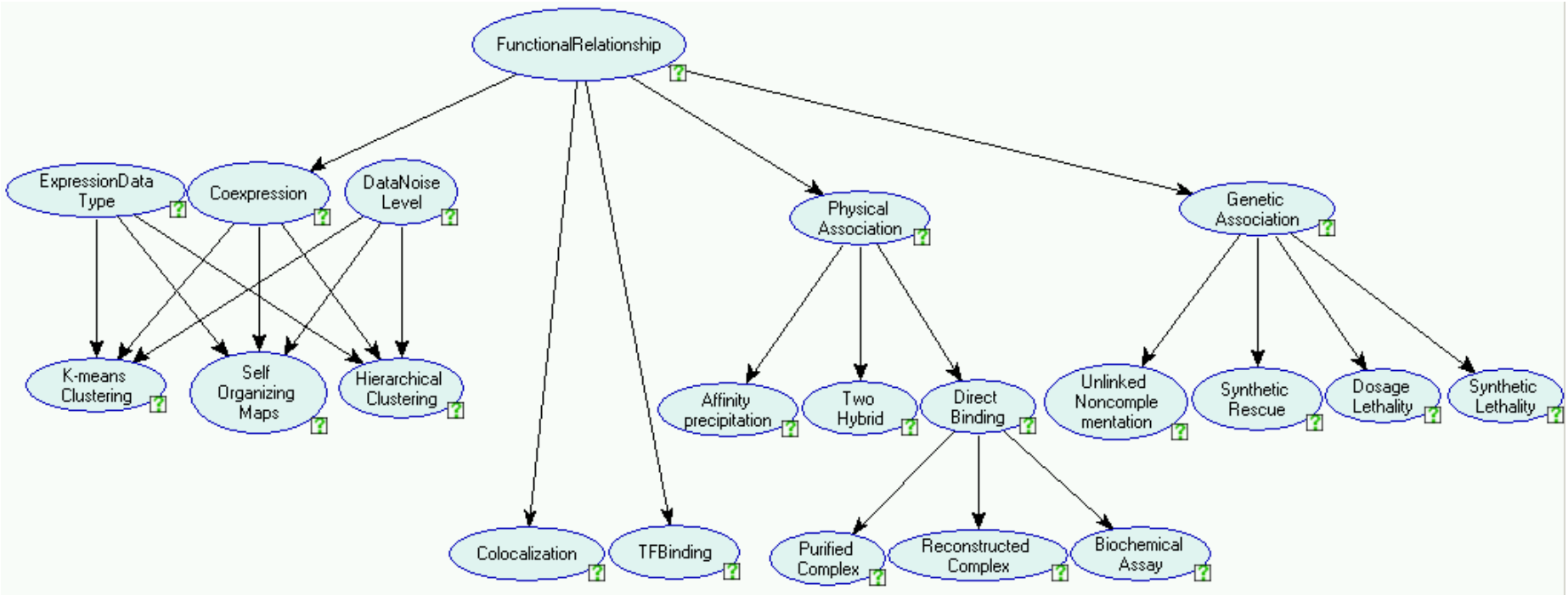
Structure	Observability	Method
Known	Full	Maximum Likelihood Estimation
Known	Partial	EM (or gradient ascent)
Unknown	Full	Search through model space
Unknown	Partial	EM + search through model space



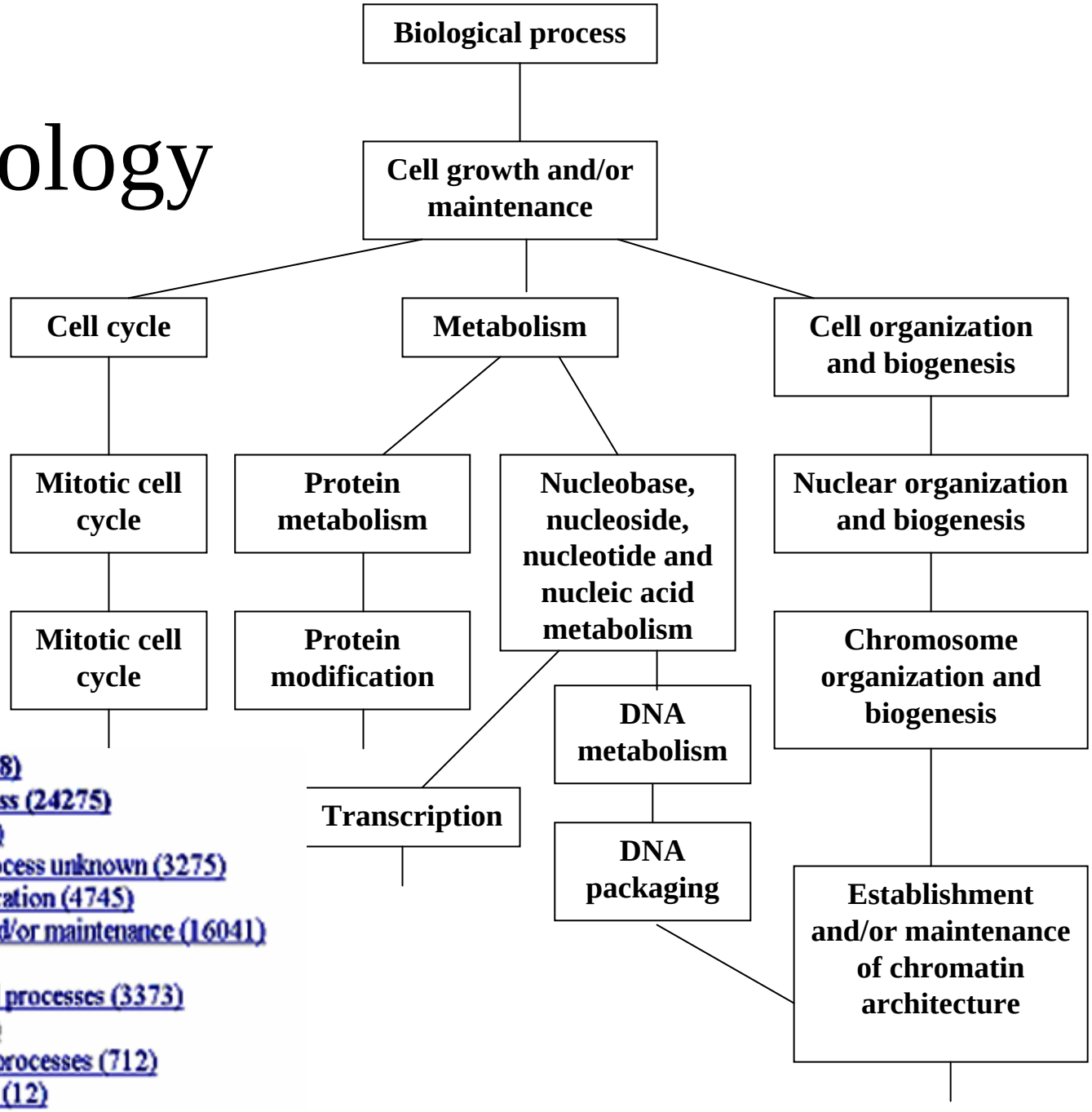
Due either to missing data or to hidden nodes

Back to gene function prediction with Bayes nets

MAGIC Bayesian network



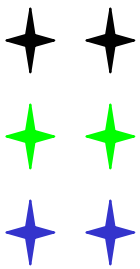
Gene Ontology

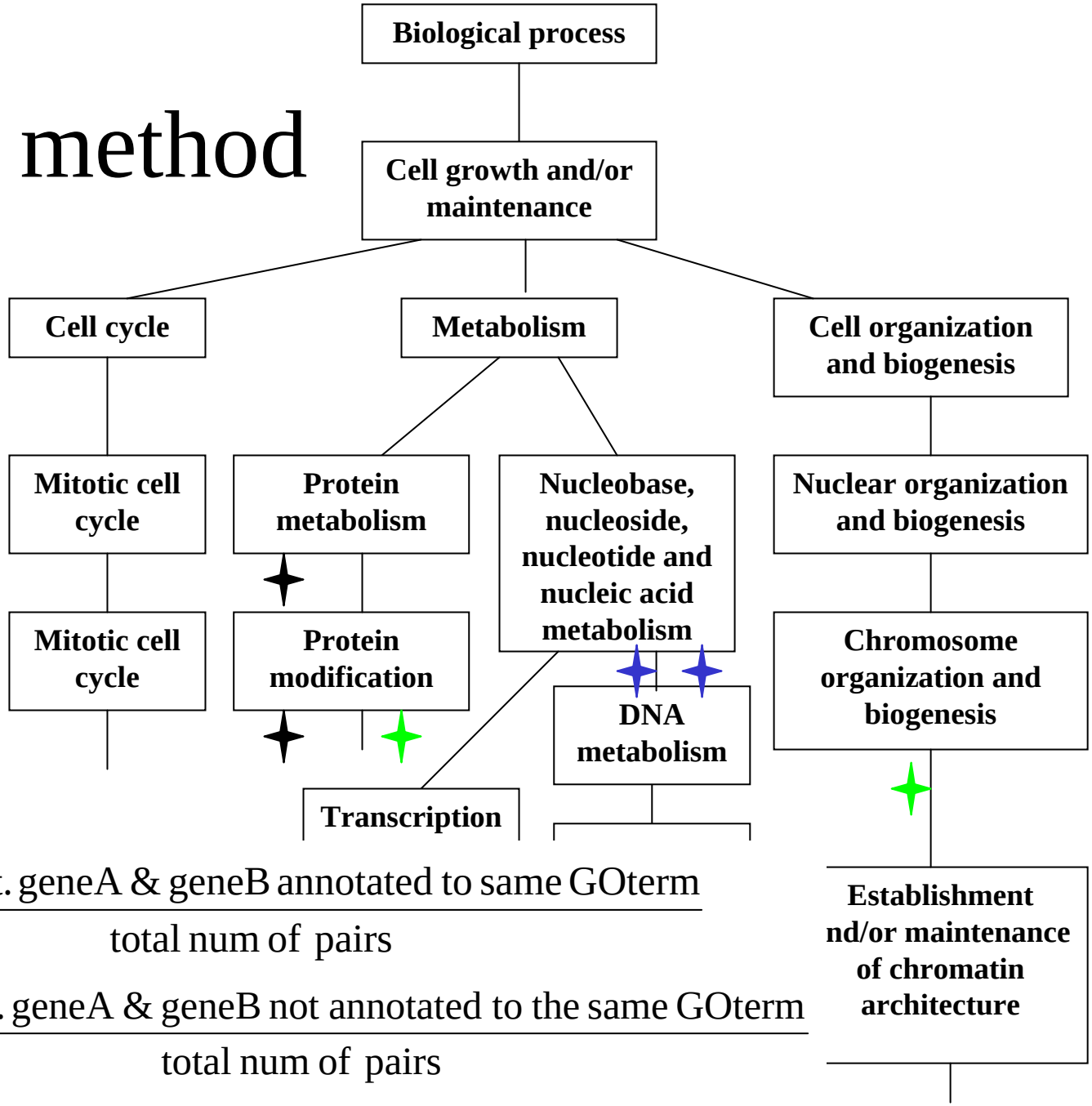


- GO:0003673 : Gene_Ontology (31688)
  - GO:0008150 : biological_process (24275)
    - GO:0007610 : behavior (216)
    - GO:0000004 : biological_process unknown (3275)
    - GO:0007154 : cell communication (4745)
    - GO:0008151 : cell growth and/or maintenance (16041)
    - GO:0016265 : death (385)
    - 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

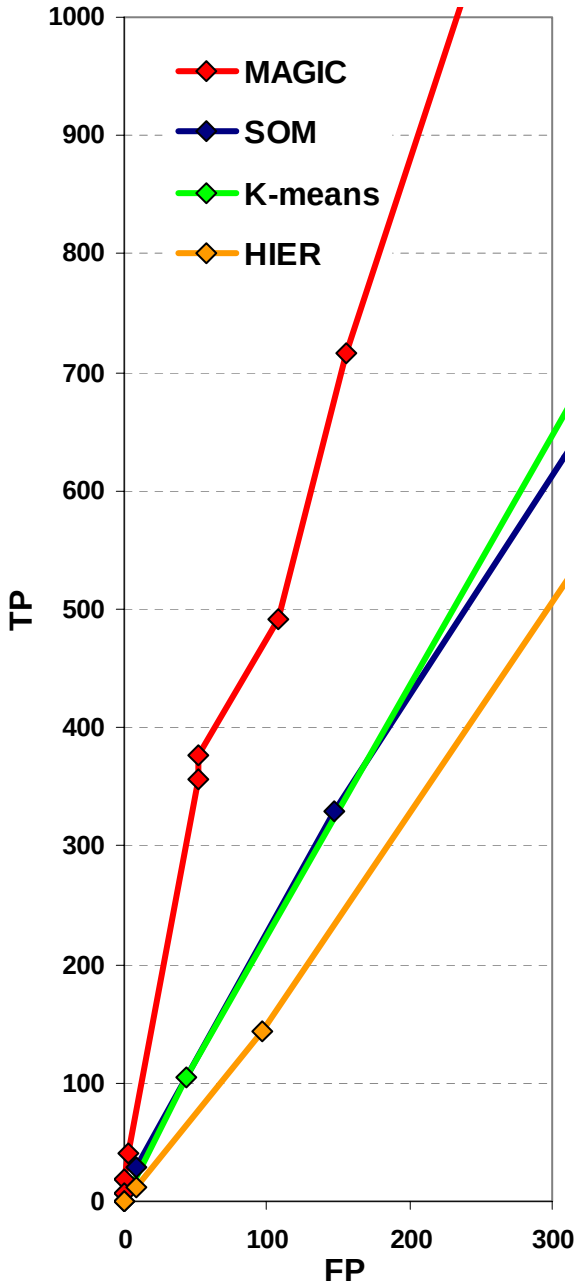


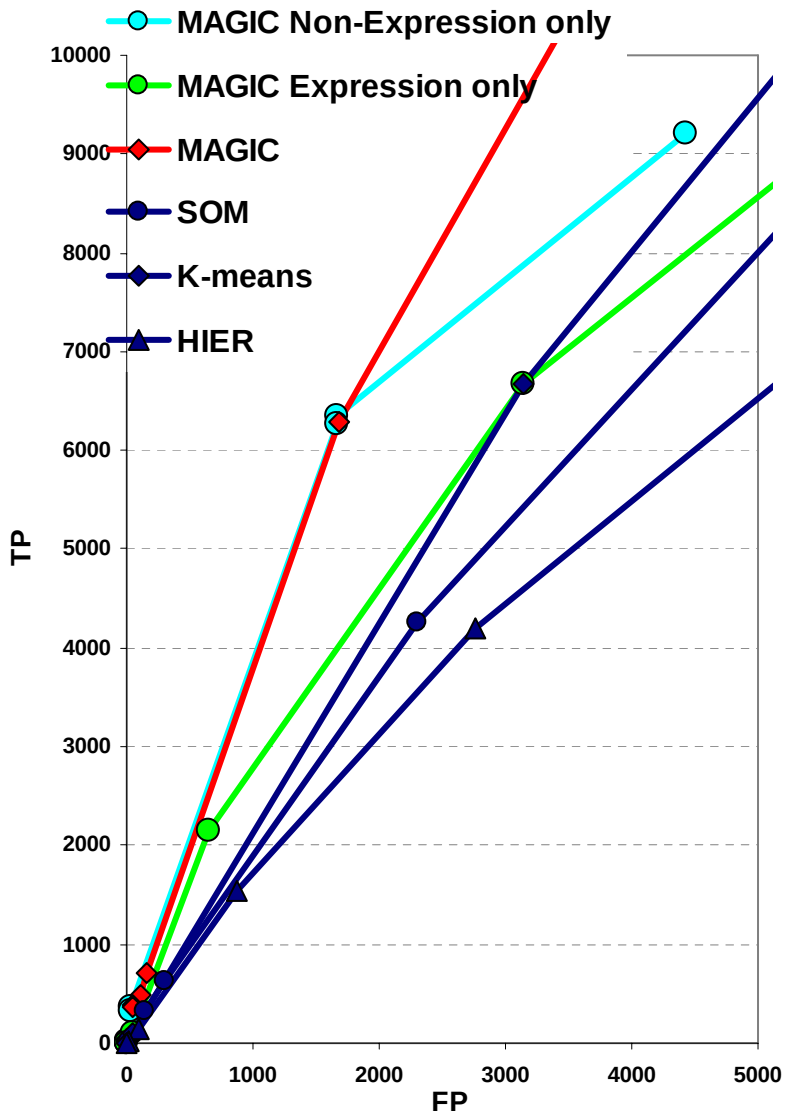


$$\text{proportionTP} = \frac{\text{num pairs s.t. geneA \& geneB annotated to same GOterm}}{\text{total num of pairs}}$$

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

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



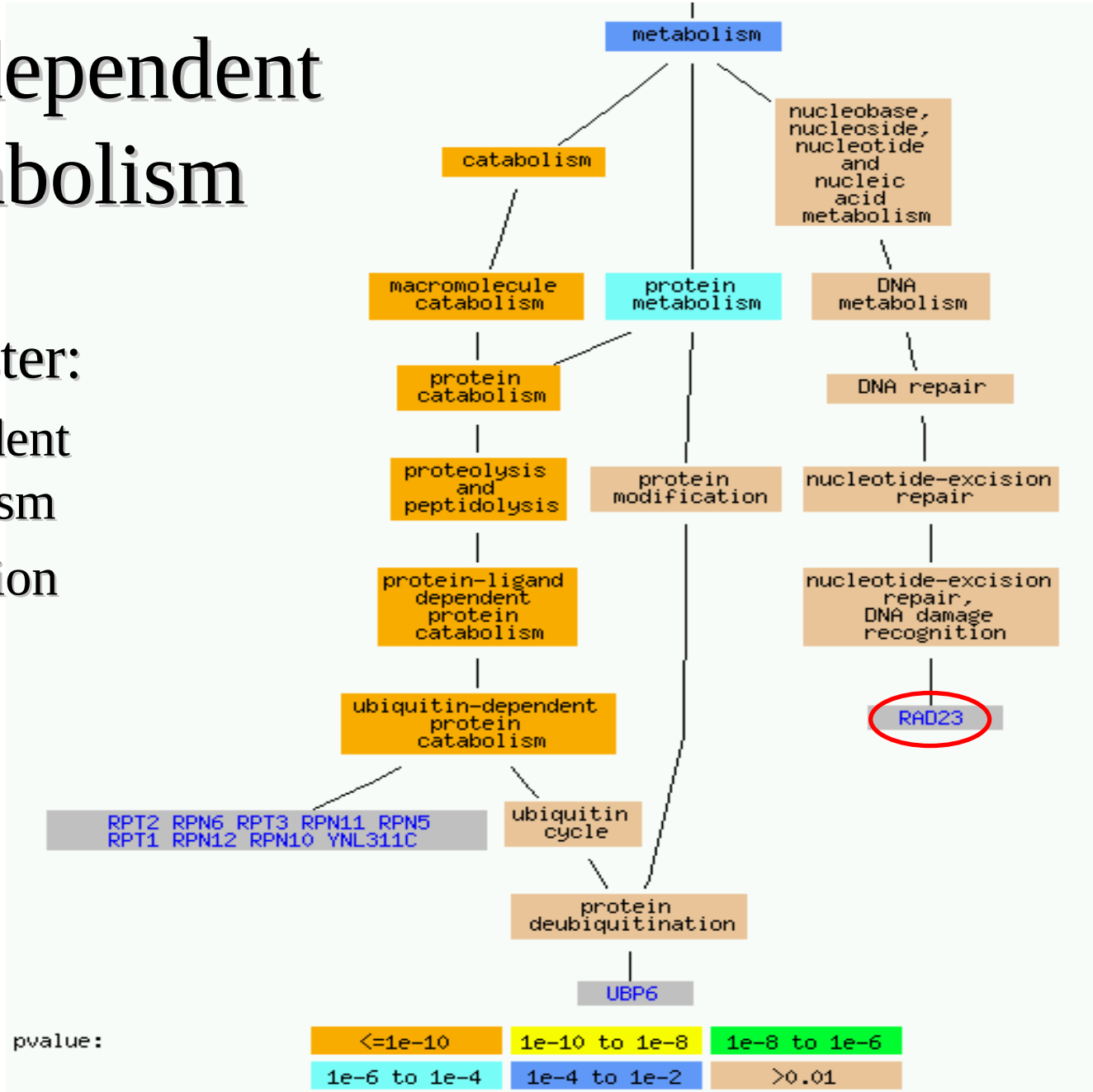


Ubiquitin-dependent protein catabolism

10 genes in cluster:

9 ubiquitin-dependent protein catabolism

1 nucleotide-excision repair

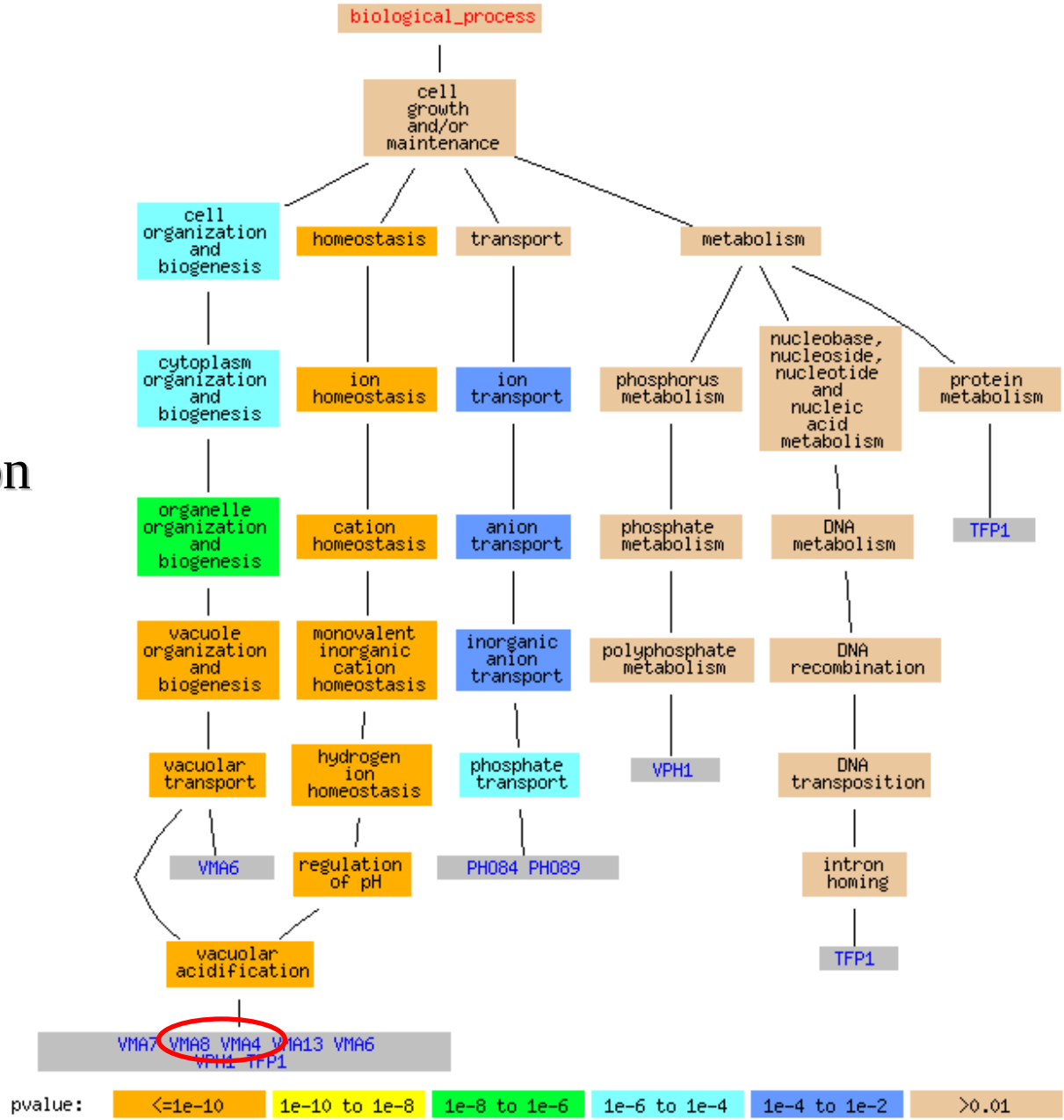


Vacuolar acidification cluster

9 genes in cluster:

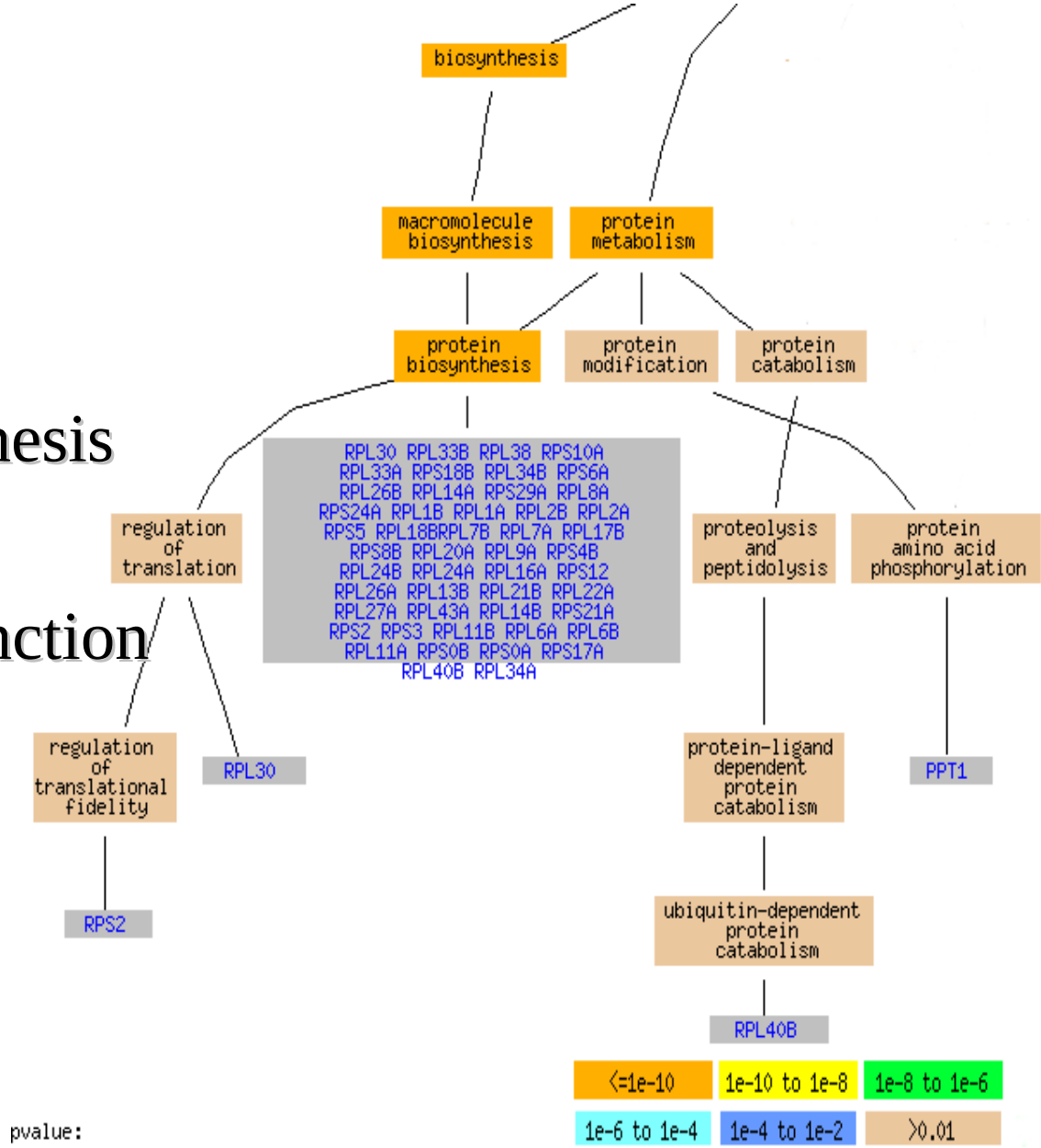
7 vacuolar acidification

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

Promoter finding & expression modules

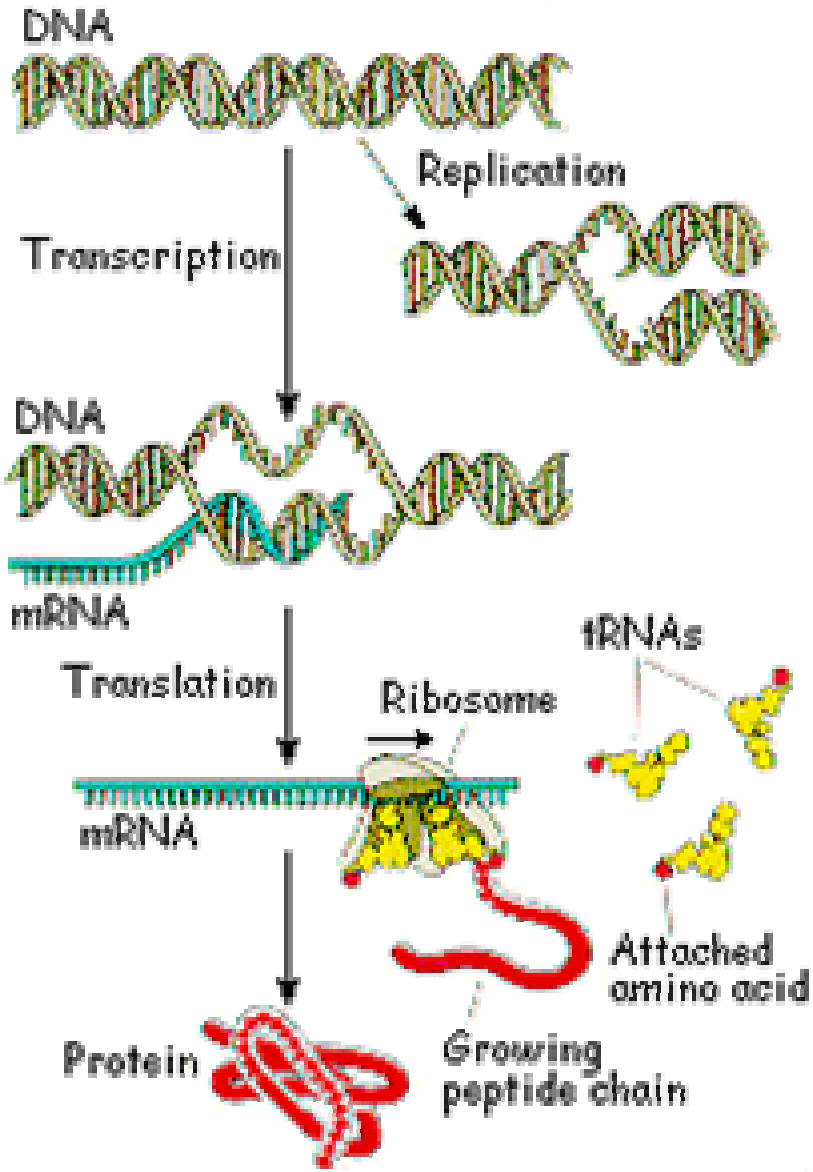
What are promoters?

Finding regulatory regions from MSAs

Using coexpression in motif finding

Regulatory modules

What are promoters and why are they important?



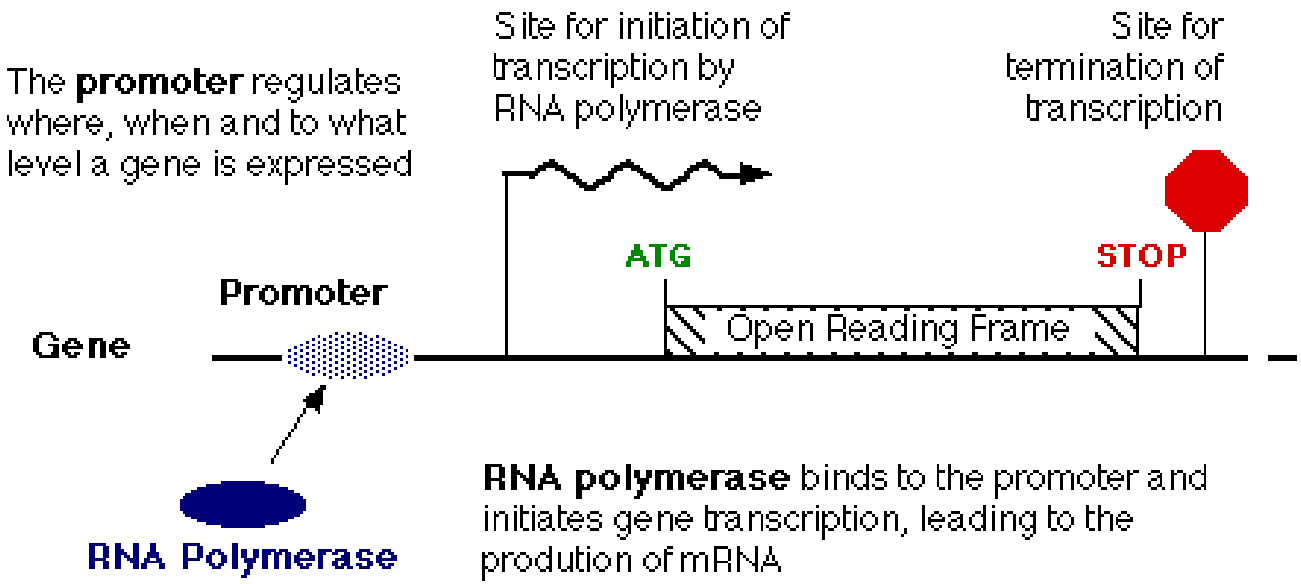
Opportunities for gene regulation

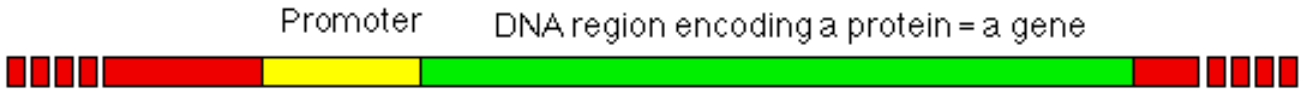
- Opening of DNA duplex
- Transcription
- mRNA stability
- Translation
- Protein stability
- Protein modification

Transcriptional regulation

- Thought to be the most used
- Does not waste intermediate products (mRNA, protein, etc)
- But transcriptional regulation is slow, and thus may not be used in cases when fast, transient regulation is necessarily

Promoters are important elements for gene expression

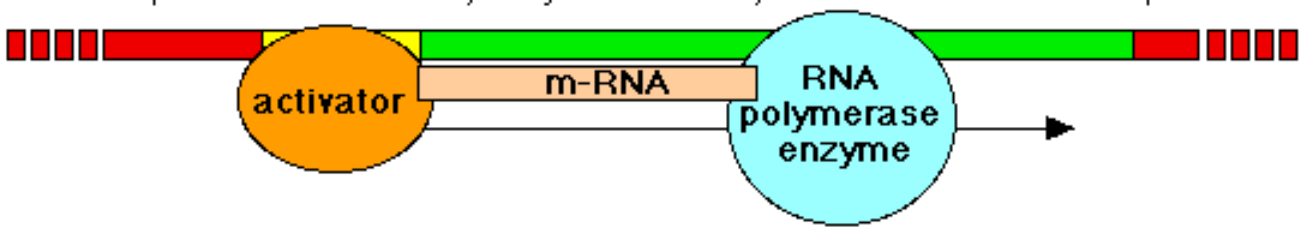




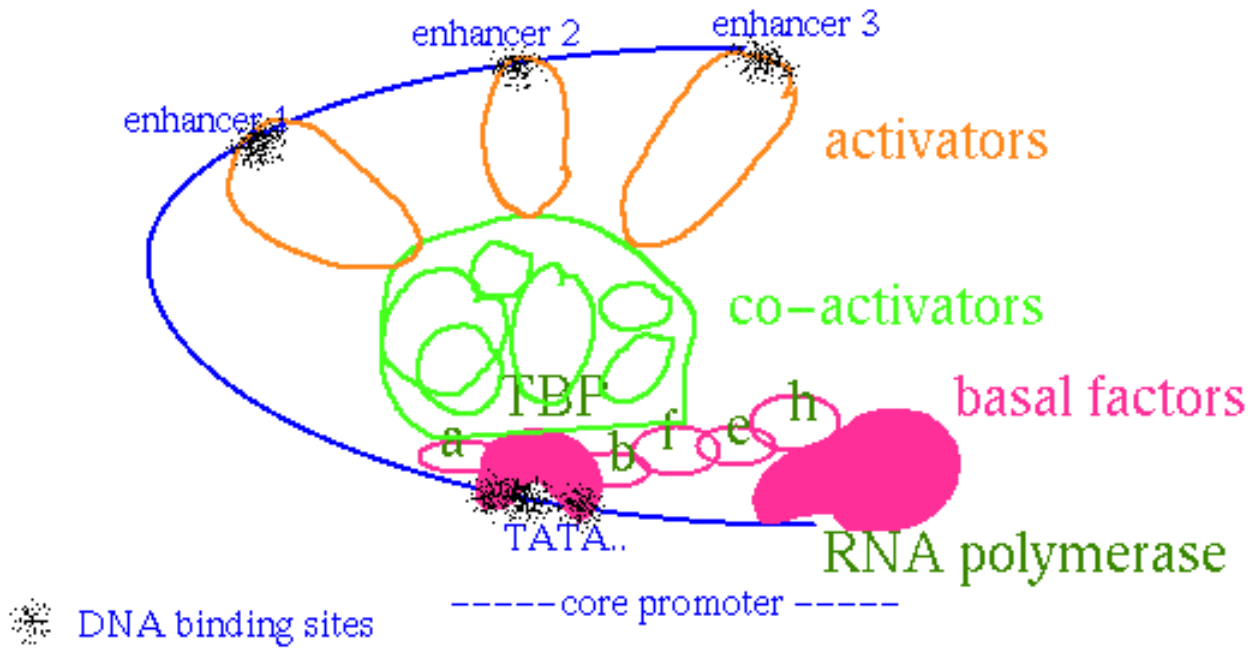
Will this gene be expressed? Depends on
whether specific activators bind to the promoter

State 1: no activation, enzyme can't bind, so no RNA is made. Gene is not expressed.

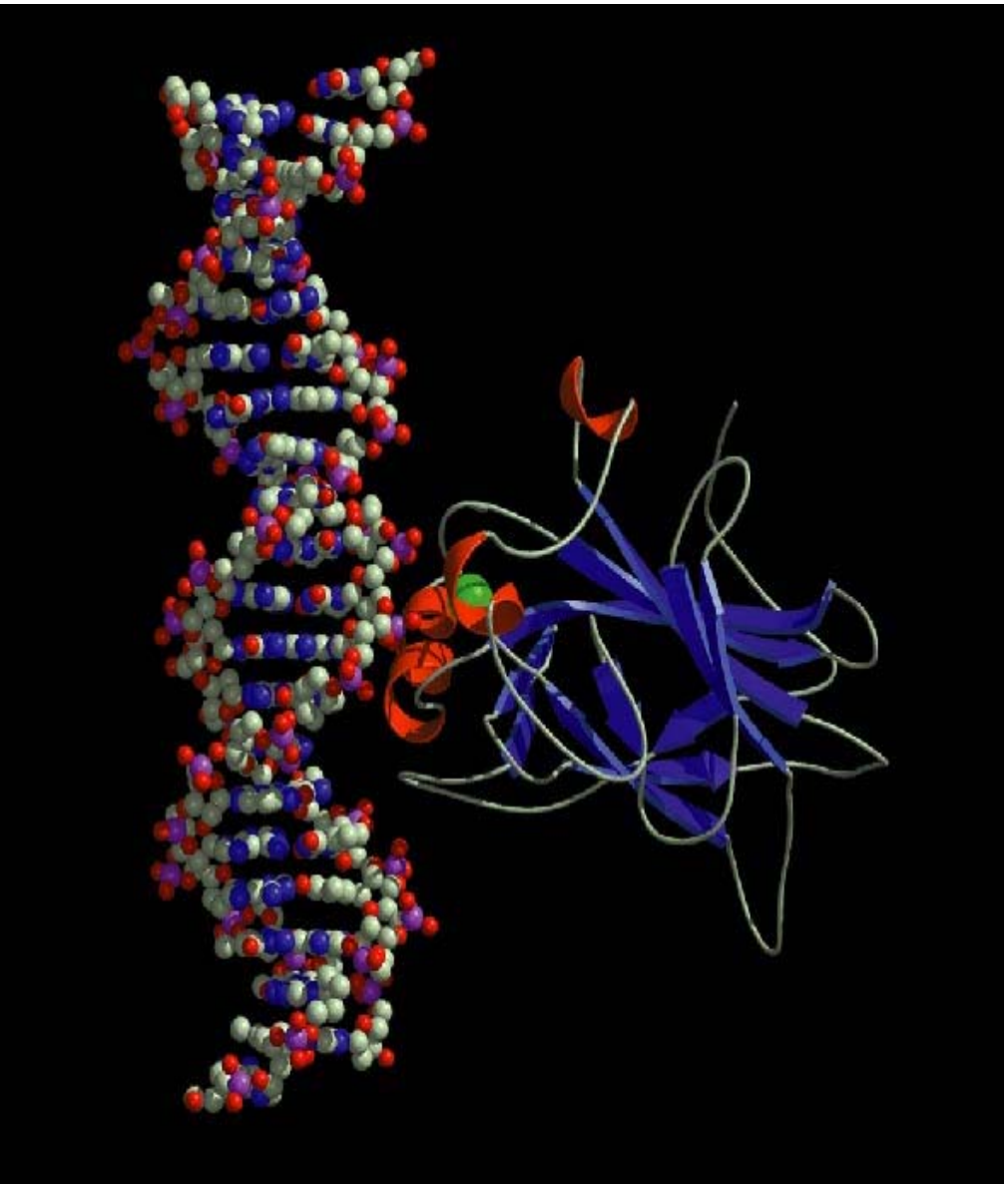
State 2: promoter is activated, enzyme cant bind, RNA is made. Gene is expressed.



Transcriptional activation



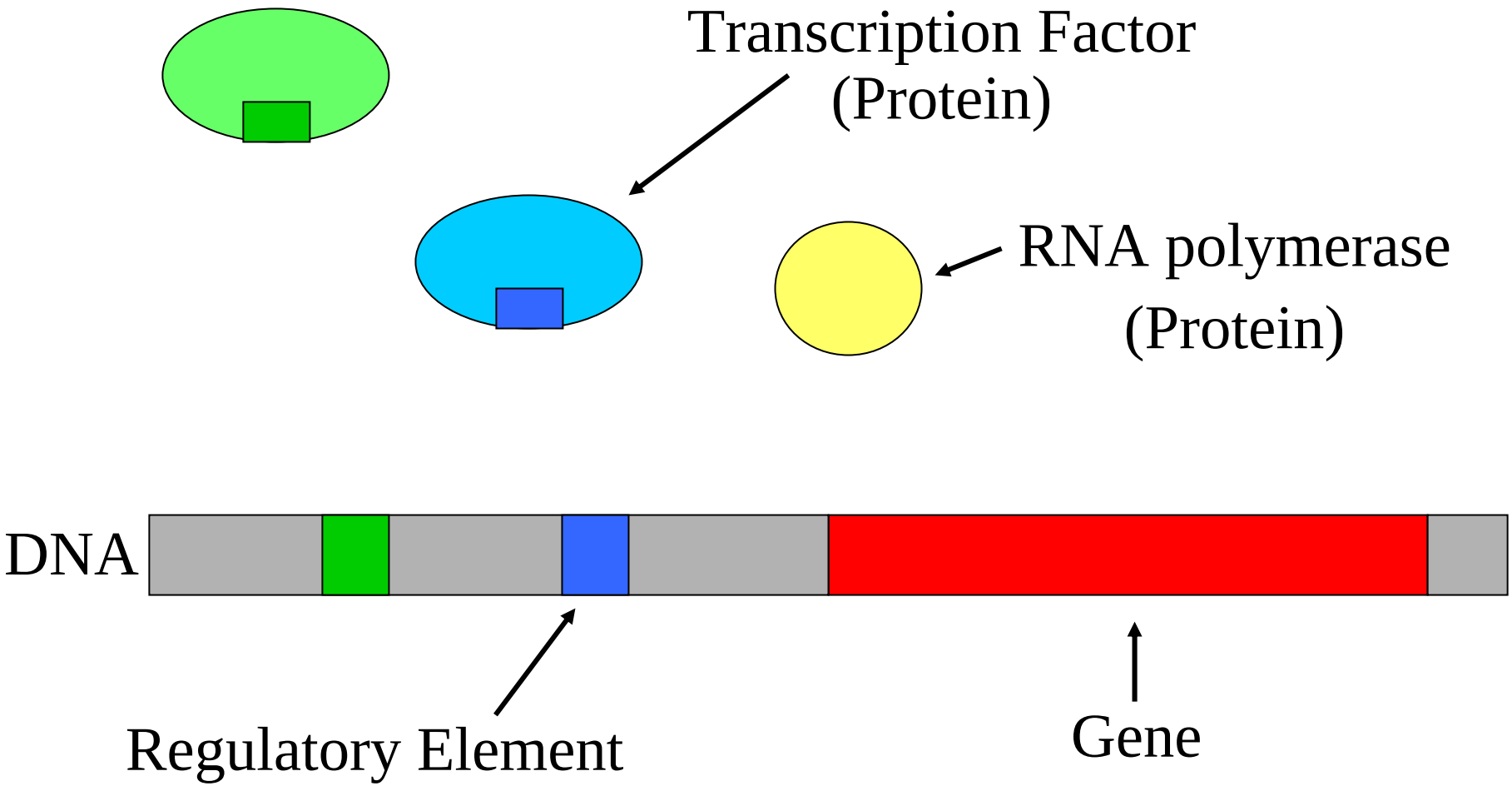
Transcription Factors Bind DNA



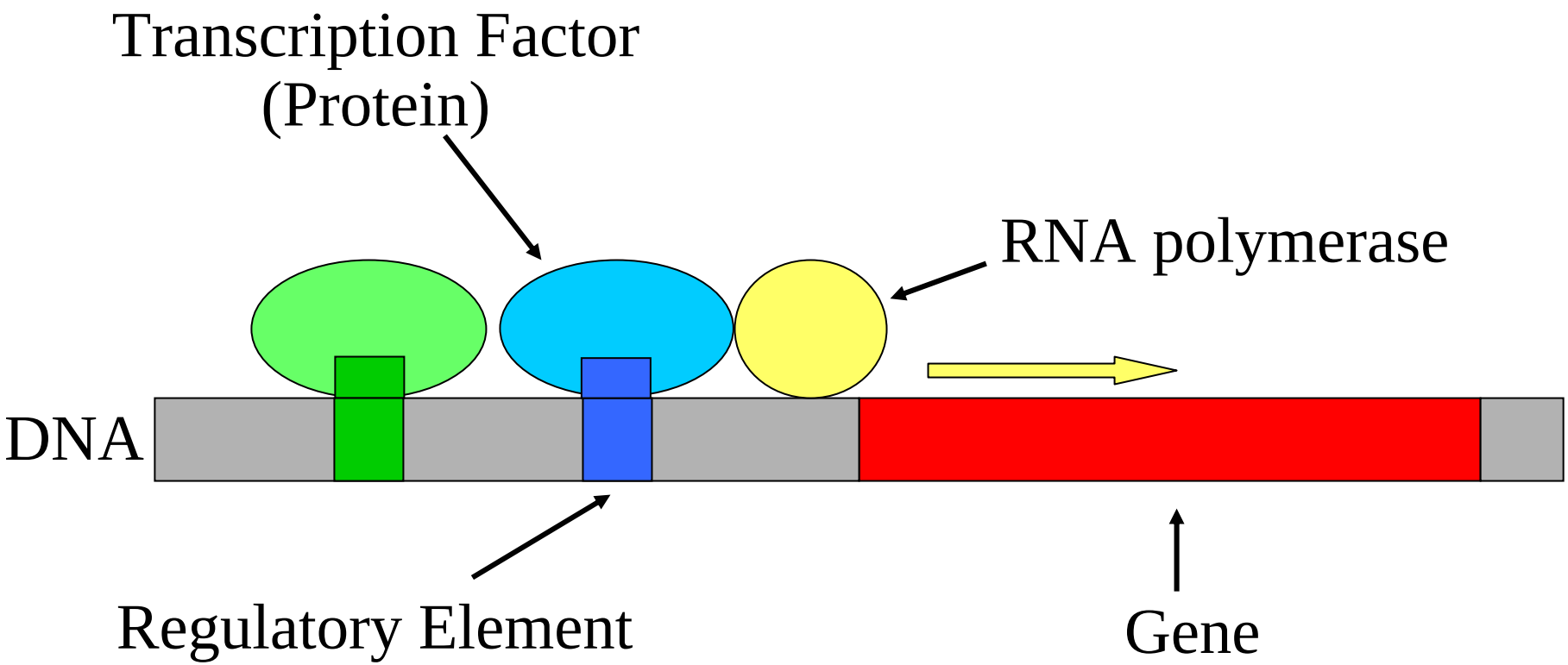
Transcription factors bind DNA in a specific manner.

Binding recognizes DNA substrings called **regulatory motifs**

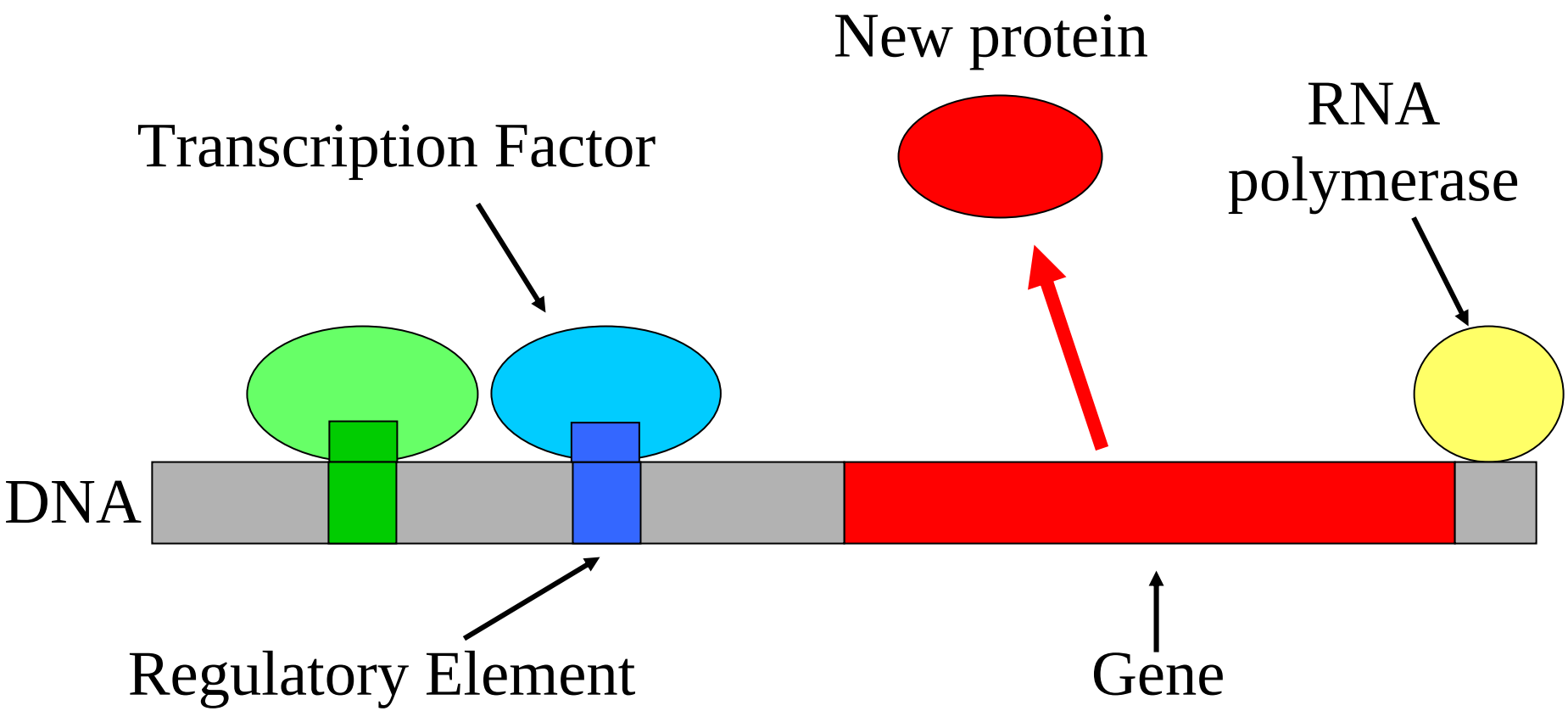
Regulation of Genes



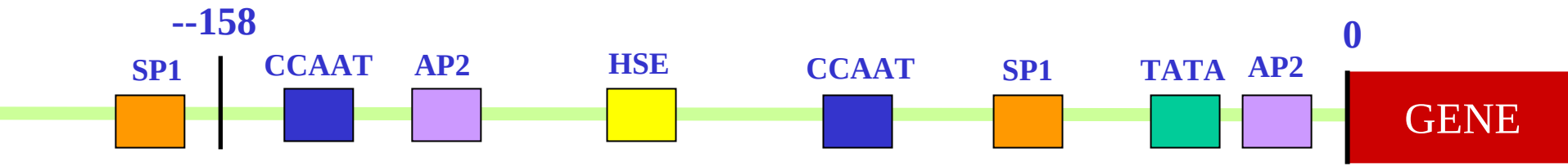
Regulation of Genes



Regulation of Genes



Example: A Human heat shock protein

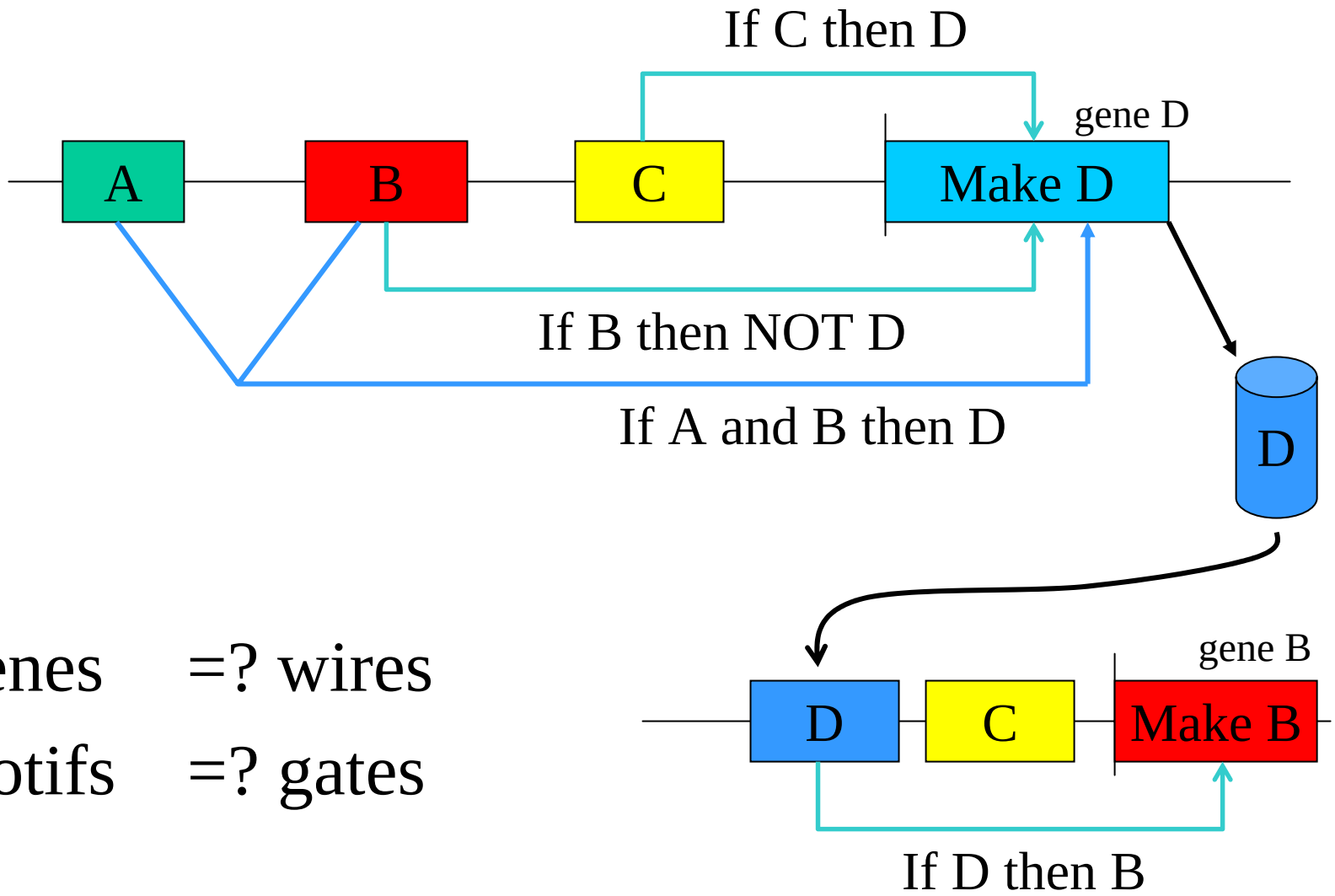


promoter of heat shock hsp70

- TATA box: positioning transcription start

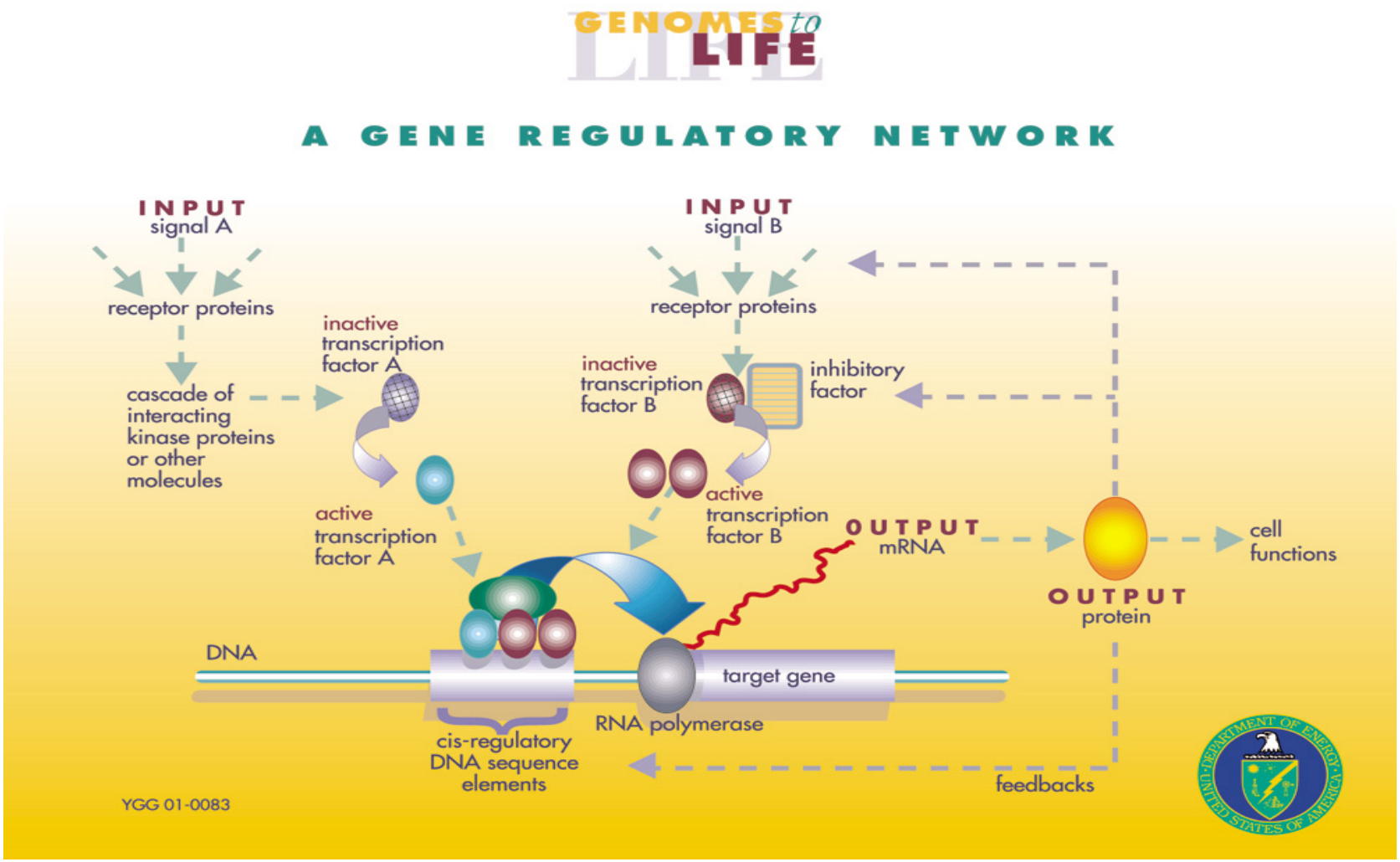
- TATA, CCAAT: constitutive transcription
- GRE: glucocorticoid response
- MRE: metal response
- HSE: heat shock element

Gene Regulatory Network

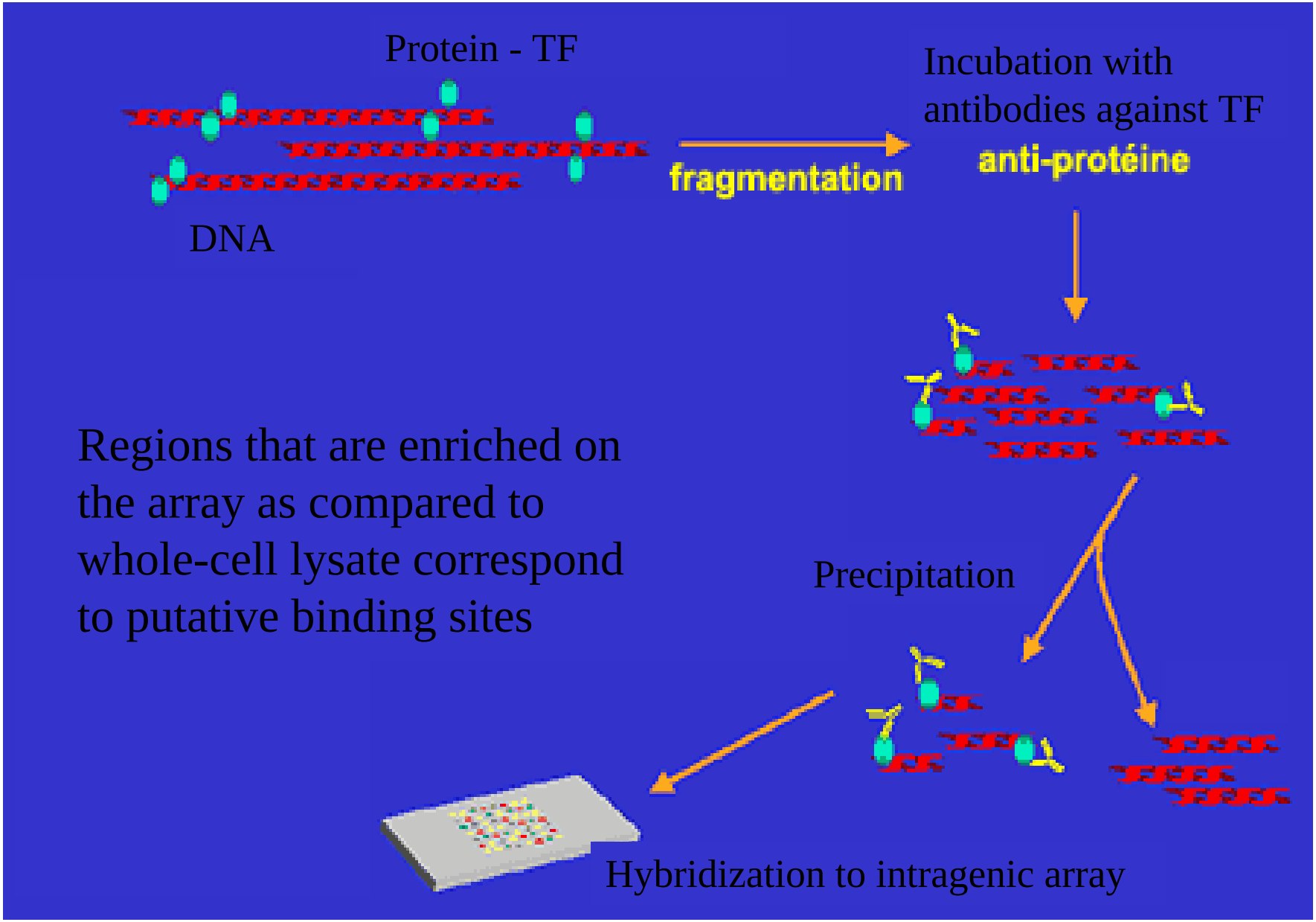


- Genes =? wires
- Motifs =? gates

Regulatory Networks

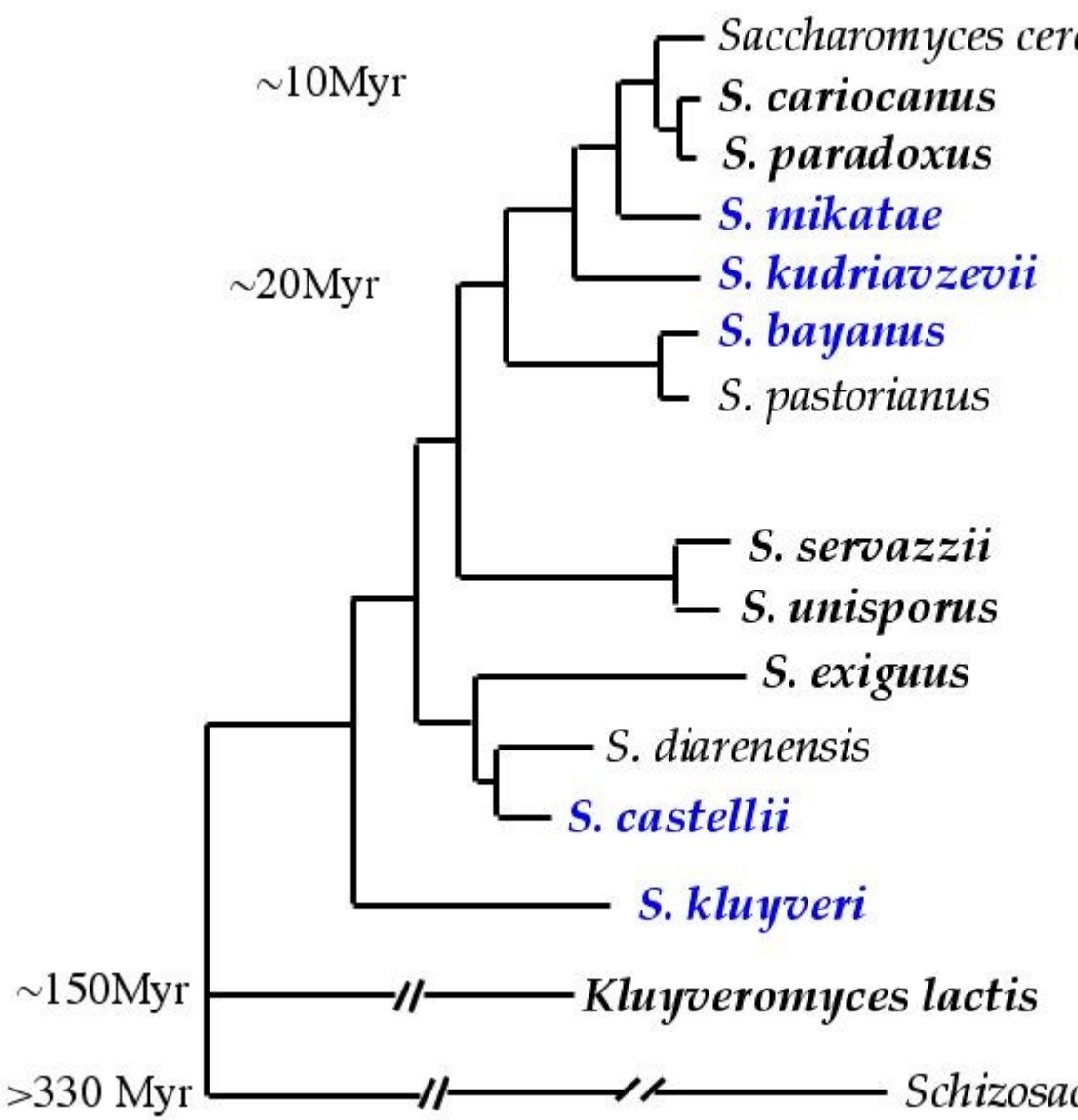


Identifying TF factor binding sites directly – “ChIP on Chip”


Protein - TF
DNA
fragmentation
Incubation with
antibodies against TF
anti-protéine
Regions that are enriched on
the array as compared to
whole-cell lysate correspond
to putative binding sites
Precipitation
Hybridization to intragenic array

Promoter finding from genome sequences

Multiple sequence alignment



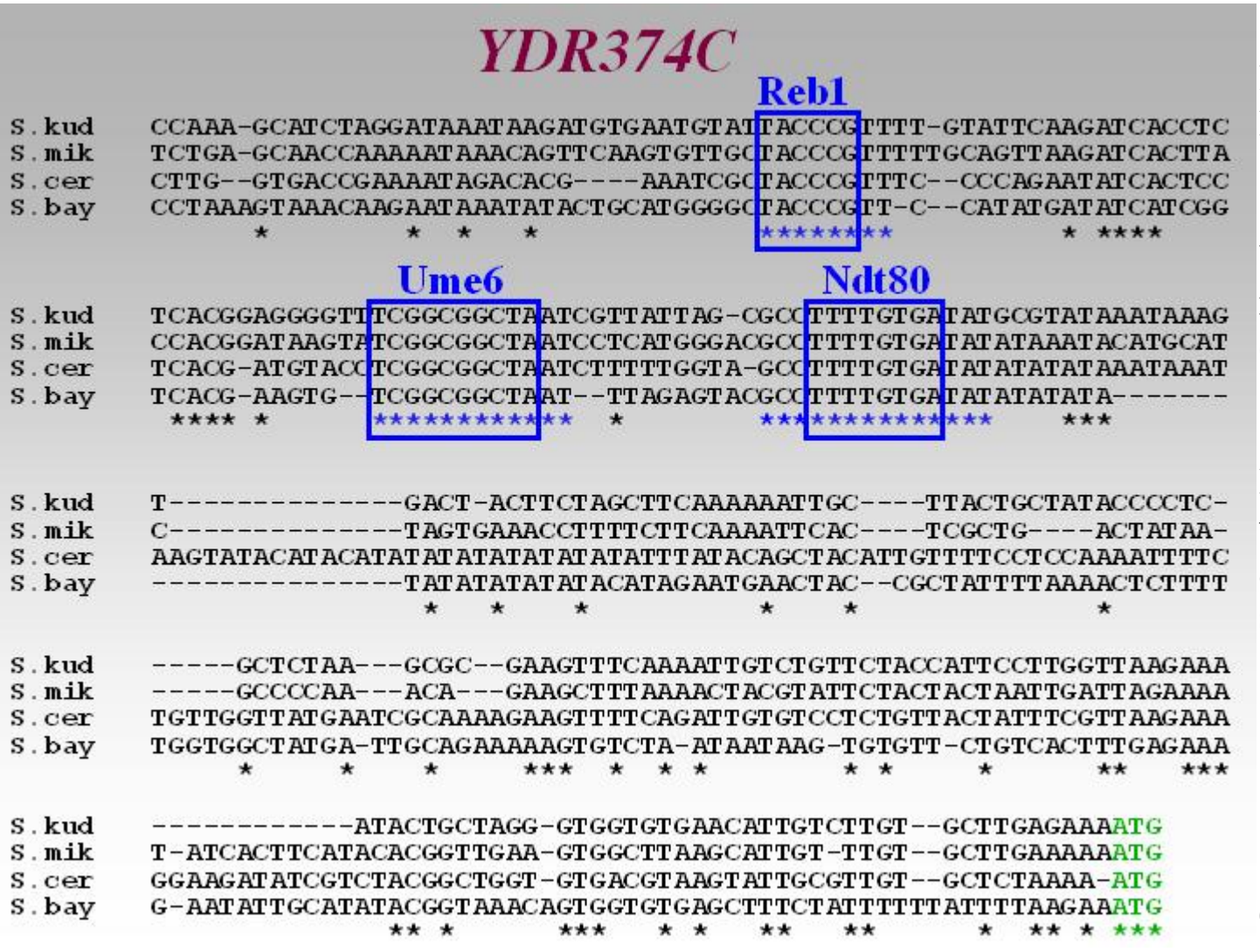
Multiple sequence alignment of *Saccharomyces species*.

Species relatively closely related to S. cerevisiae are likely to be most useful for identifying conserved non-coding sequences.

Alignments of S. cerevisiae protein sequences to their orthologs in species that are more distantly related to S. cerevisiae are likely to be more useful for identifying possible functional domains in protein sequence

Cliften et al. Science 301 2003

Promoters can be identified from multiple sequence alignment



Reb1 - RNA polymerase I Enhancer Binding protein

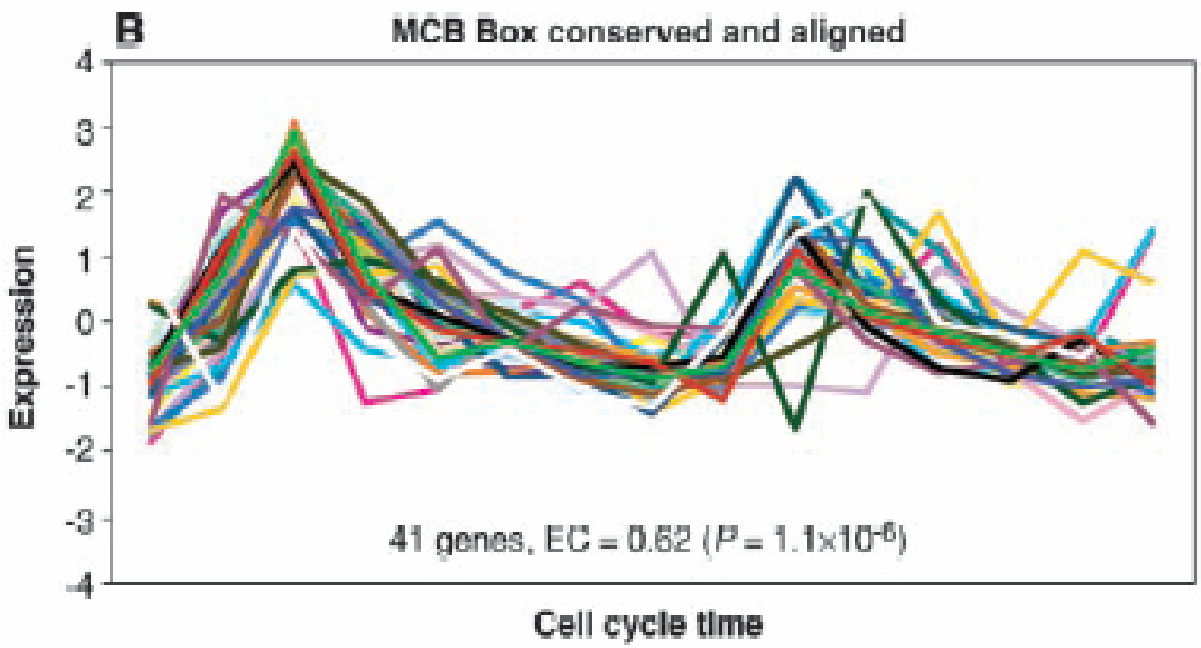
Ume6 - Regulator of both repression and induction of early meiotic genes.

Ndt80 - DNA binding transcription factor that activates middle sporulation genes

Cliften et al. Science 301 2003

Promoter finding based on transcriptional profiles

Co-regulated genes are co-expressed



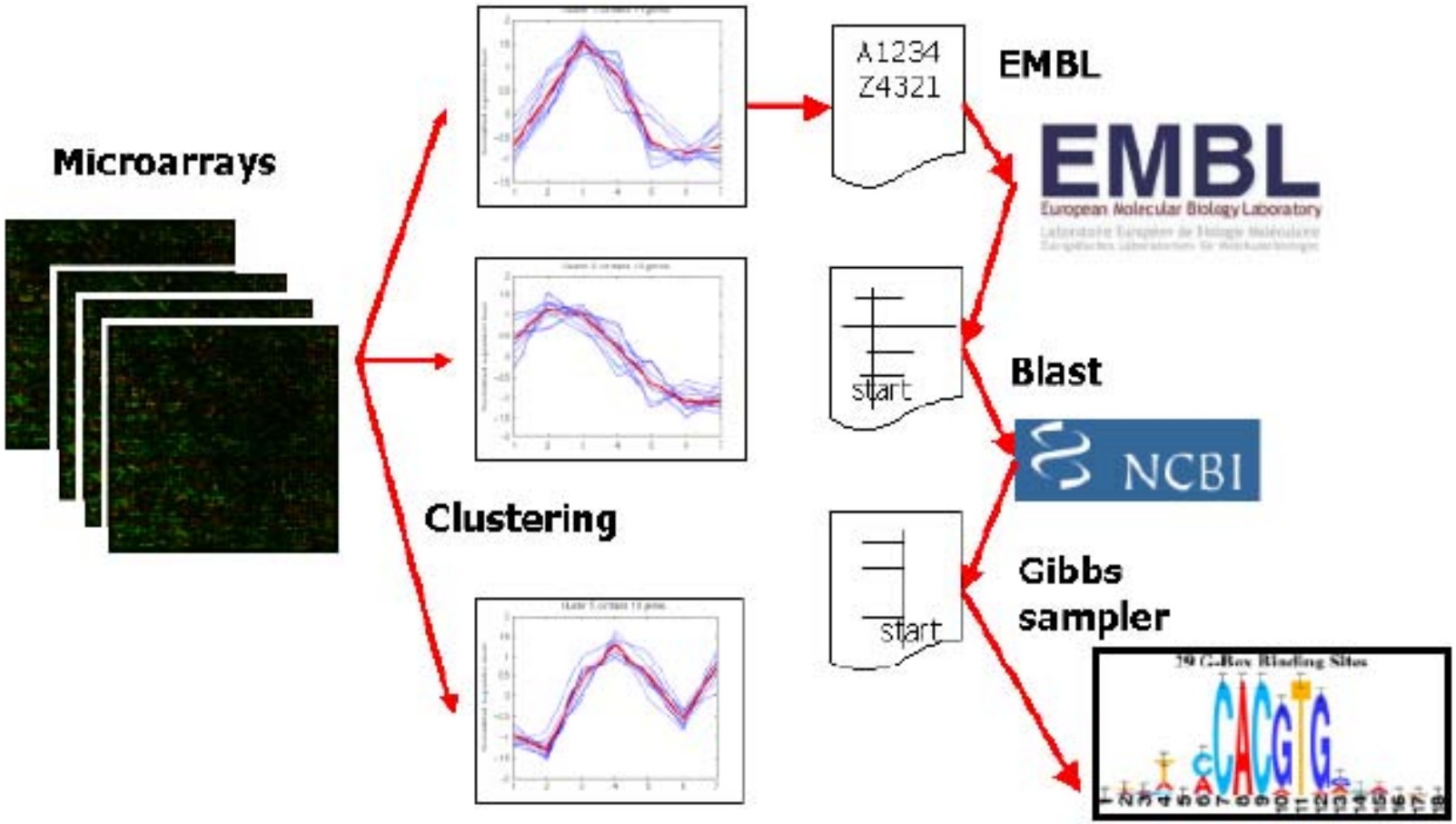
Expression profiles of 53 genes in *S. cerevisiae* genome that contain the exact match to an MCB box in their promoters (profiles normalized by mean & variance).

Cliften et al. Science 301 2003

Gibbs Sampling in Motif Finding

Idea

- Identify sets of co-regulated genes from microarrays
  - Unsupervised analysis - clustering
  - Supervised analysis
- Identify common motifs in regulatory regions of co-regulated genes
  - Combinatorial methods (enumeration with tricks)
  - Probabilistic methods (EM, Gibbs Sampling – a special case of MCMC)


Microarrays
Clustering
A1234
Z4321
EMBL
EMBL
European Molecular Biology Laboratory
start
Blast
NCBI
start
Gibbs
sampler

Gibbs Sampling

- Given:
  - $x^1, \ldots, x^N$ sequences
  - motif length K,
  - background B,
- Find:
  - Model M
  - Locations $a_1, \ldots, a_N$ in $x^1, \ldots, x^N$

Maximizing log-odds likelihood ratio:

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \log \frac{M(k, x^i_{a_i+k})}{B(x^i_{a_i+k})}$$

Gibbs Sampling (2)

- AlignACE: first statistical motif finder
- BioProspector: more recent, faster algorithm with higher accuracy

Algorithm (sketch):

1. Initialization:
   a. Select random locations in sequences x^1, …, x^N
   b. Compute an initial model M from these locations

2. Sampling Iterations:
   a. Remove one sequence x^i
   b. Recalculate model
   c. Pick a new location of motif in x^i according to probability the location is a motif occurrence

Gibbs Sampling (3)

Initialization:

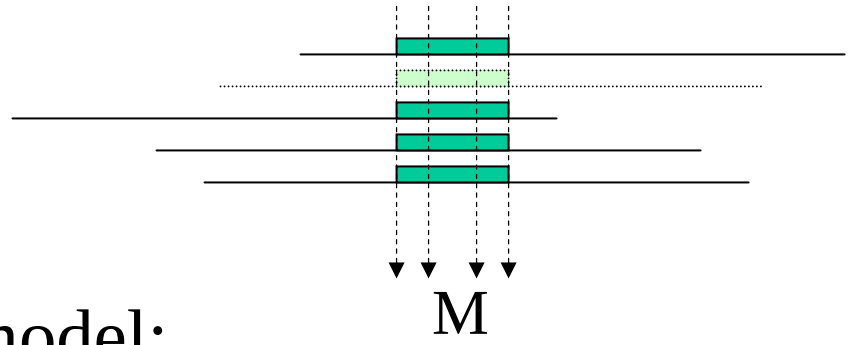
- Select random locations $a_1, \ldots, a_N$ in $x^1, \ldots, x^N$
- For these locations, compute M:

$$M_{kj} = \frac{1}{N} \sum_{i=1}^{N} (x_{a_i + k} = j)$$

- That is, M_{kj} is the number of occurrences of letter j in motif position k, over the total

Gibbs Sampling (4)

Predictive Update:



- Select a sequence $x = x^i$
- Remove x^i, recompute model:

Again, M_{kj} is the proportion of occurrences of letter j in motif position k

$$M_{kj} = \frac{1}{(N-1)+B}(\beta_j + \sum_{s=1, s\neq i}^{N}(x_{a_s+k} = j))$$

where β_j are pseudocounts to avoid 0s,

and B = $\Sigma_j \beta_j$

Gibbs Sampling (5)

Sampling:

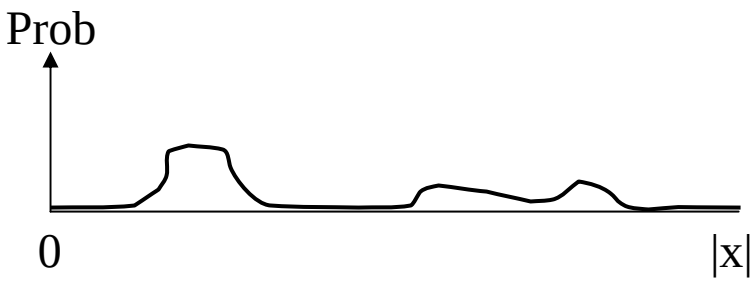
For every K-long word $x_j,\ldots,x_{j+k-1}$ in x:

$$Q_j = P(\text{ word } | \text{ motif }) = M(1,x_j)\times\ldots\times M(k,x_{j+k-1})$$
$$P_i = P(\text{ word } | \text{ background }) = B(x_j)\times\ldots\times B(x_{j+k-1})$$

How "overrepresented" this word is in motif vs. background

Let $A_j = \dfrac{Q_j / P_j}{\sum_{j=1}^{|x|-k+1} Q_j / P_j}$

Represents weights for sampling (words more different from background get higher weight)



Sample a random new position a_i according to the probabilities $A_1,\ldots, A_{|x|-k+1}$ (new location for the motif)

Gibbs Sampling (6)

Running Gibbs Sampling:

1. Initialize

2. Run until convergence

3. Repeat 1,2 several times, report common motifs

Advantages / Disadvantages

- Very similar to EM (essentially EM's stochastic analog)

Advantages:

- Easier to implement
- Less dependent on initial parameters
- Less likely to converge to local minima than EM
- More versatile, easier to enhance with heuristics

Disadvantages:

- More dependent on all sequences to exhibit the motif
- Less systematic search of initial parameter space (doesn't converge to point estimate like EM)

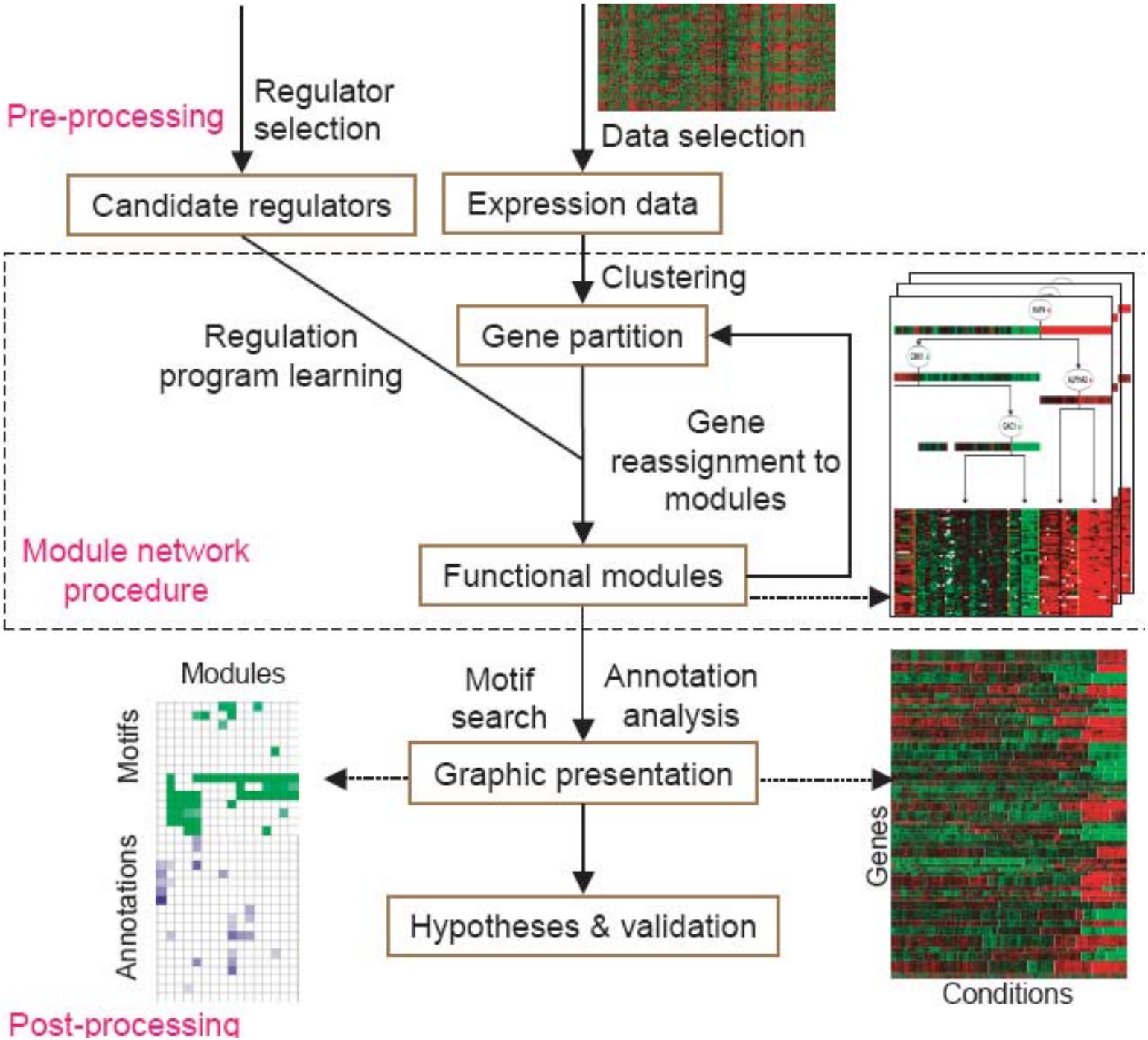
Regulatory Modules

Segal et.al. Nature Genetics 32(2):166 2004

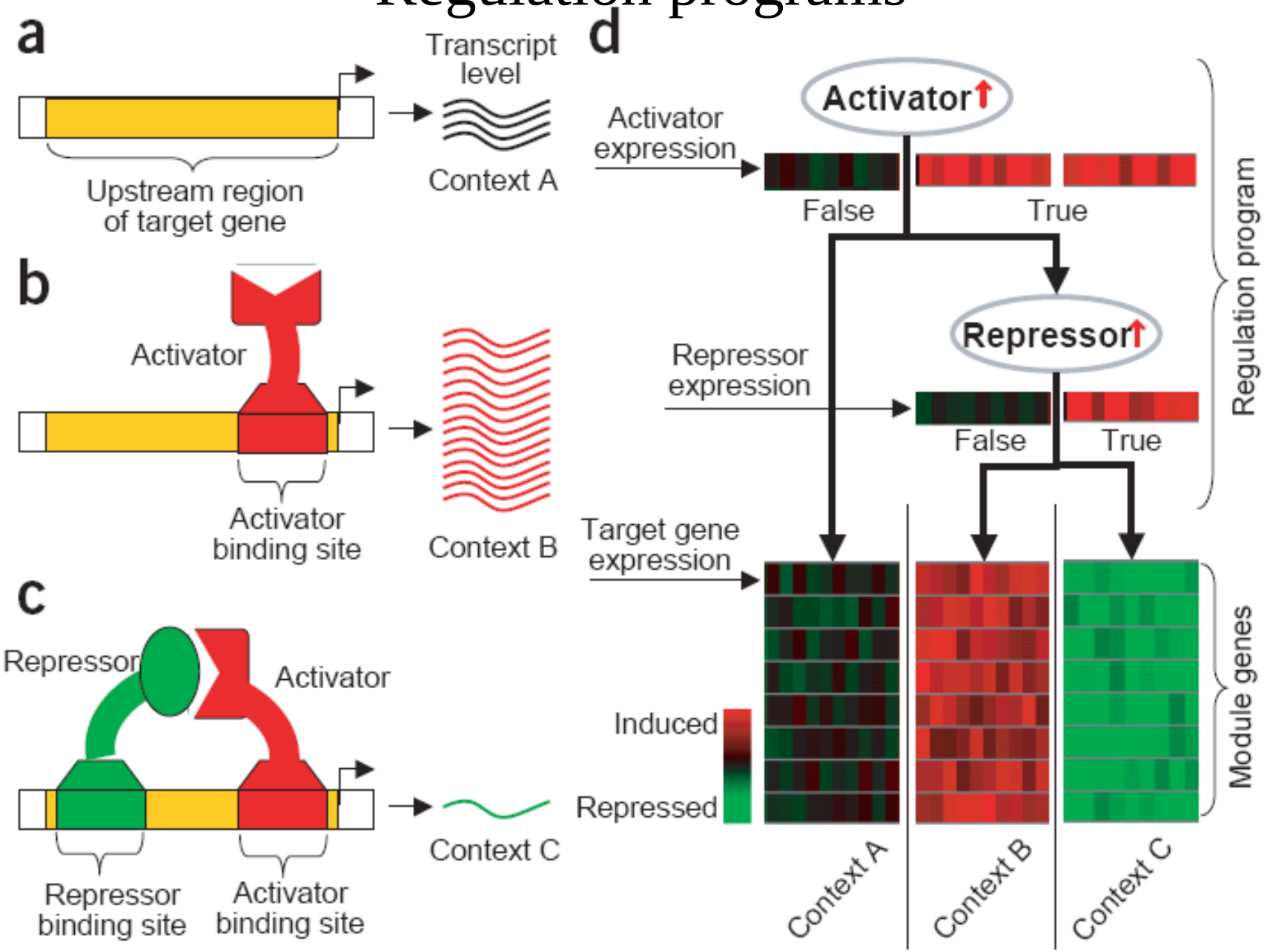
Regulatory modules

- Regulatory Module – set of genes that are co-regulated by a shared regulation program
- Regulation program – set of regulatory genes and expression profile of genes in the module as a function of expression of regulatory genes
- Assumption – regulators are themselves transcriptionally regulated

- Input: gene expression data set, list of putative regulator genes
- Output: partition of genes into modules and a regulation program for each module


Pre-processing
Regulator selection
Data selection
Candidate regulators
Expression data
Clustering
Gene partition
Regulation program learning
Gene reassignment to modules
Module network procedure
Functional modules
Modules
Motifs
Annotations
Motif search
Annotation analysis
Graphic presentation
Genes
Conditions
Hypotheses & validation
Post-processing

Regulation programs



Defining regulation programs

- Regulation program specifies:
  - Set of contexts (rules describing behavior of genes in modules e.g. upregulation)
  - Response of modules in each context
- Contexts organized in regression tree
  - Decision nodes are regulators
  - Each path to a leaf defines a context using texts on the path
  - Contexts effectively specify sets of arrays
  - Context model: normal distribution over the expression of the module's genes in these arrays (mean, variance stored in corresponding leaf)
  - Small variance => tight regulation

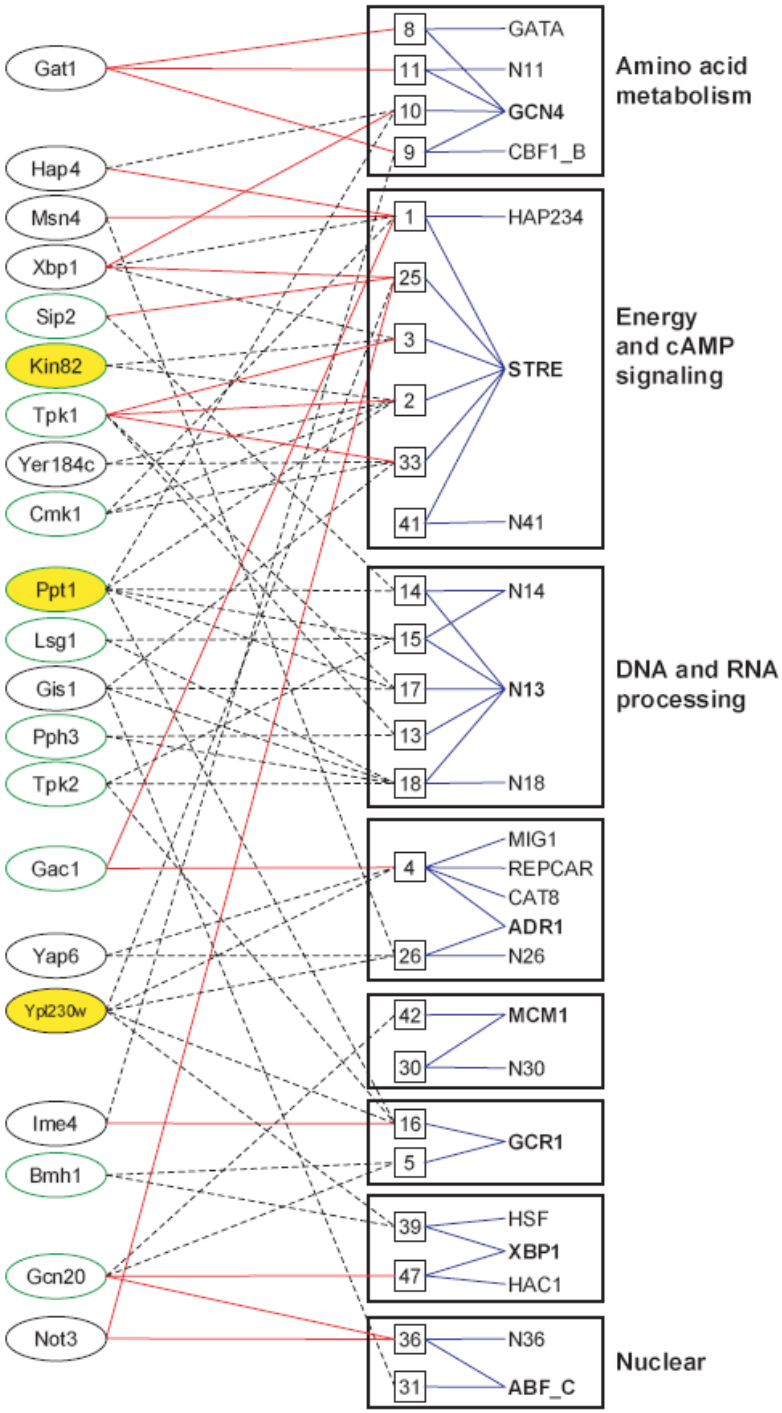
Learning Module Networks with EM

- E-step - Given g's inferred regulation program, find module that best predicts g's behavior
  - For each gene g:
    - Calculate:
      $P(g \mid regulatory\_program) = \prod_{arrays} p(g_j \mid array_j context)$
      context c of array j is defined as $N(\mu_c, \sigma_c)$
    - Reassign gene g to the program that gives highest
      $P(g \mid regulatory\_program)$
- M-step - Given partition of genes into modules, learn best regulation program (tree) through combinatorial search of trees
  - Tree grown from root to leaves, for each regulatory node:
    - Choose query that best partitions gene expression into two distinct distributions
    - Stop when no such split exists

EM learning details

- Initialize with clusters
- Converges after 23 iterations to the 50 modules (initial assignments changed for 49% of genes)

Predicted Modules



- Regulators (in ovals) are connected to modules (numbered squares)
  - Red line – regulation supported in literature, dashed line - inferred
- Module groups (boxes) share common motifs and sometimes common function
- Yellow regulators tested experimentally

Limitations

- Requires a set of putative transcription factors as input and a predefined number of modules
- Cannot find regulators whose expression does not change sufficiently for detection
- Cannot identify multiple regulators that participate in a regulatory even, will only identify one of them
- Can mistakenly identify a gene as regulator because it is highly predictive of a module either b/c it's a member of the module or by chance (gene has to be a member of the putative regulator set
- Will not identify regulatory events specific to regulator and its target
- Can only handle non-overlapping modules – a gene can belong to only one module (other methods address this problem)