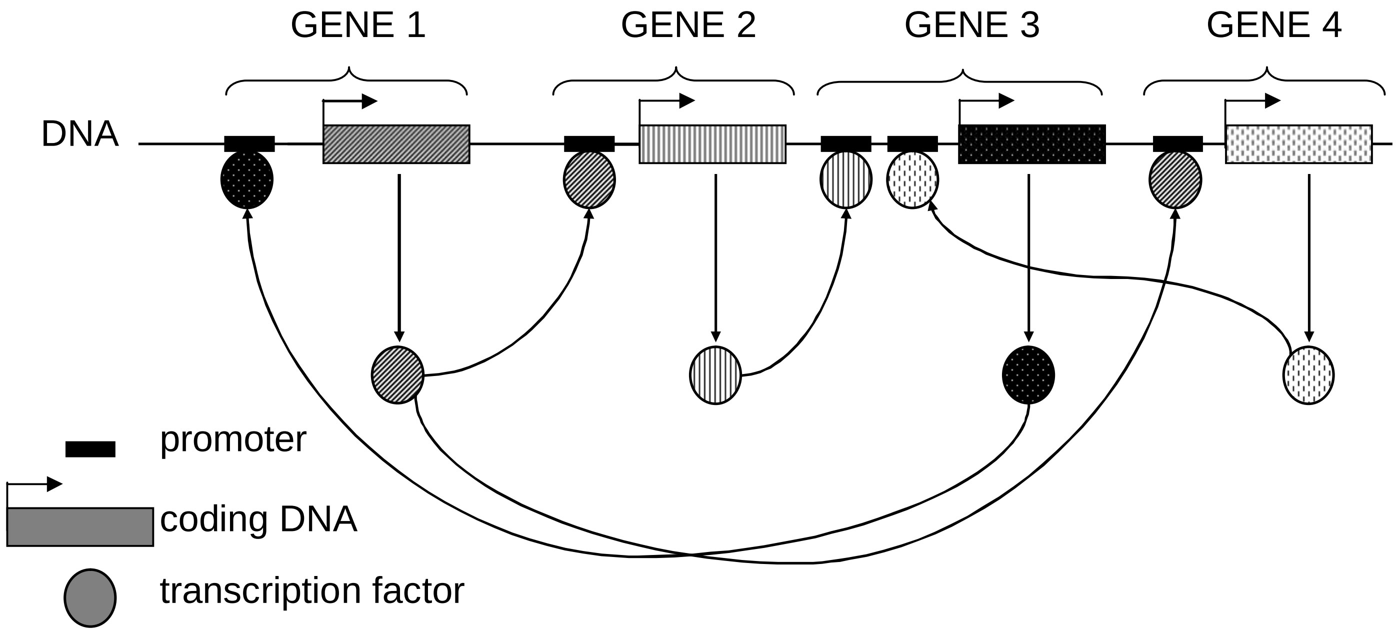
Introduction to Microarray Data Analysis and Gene Networks lecture 7

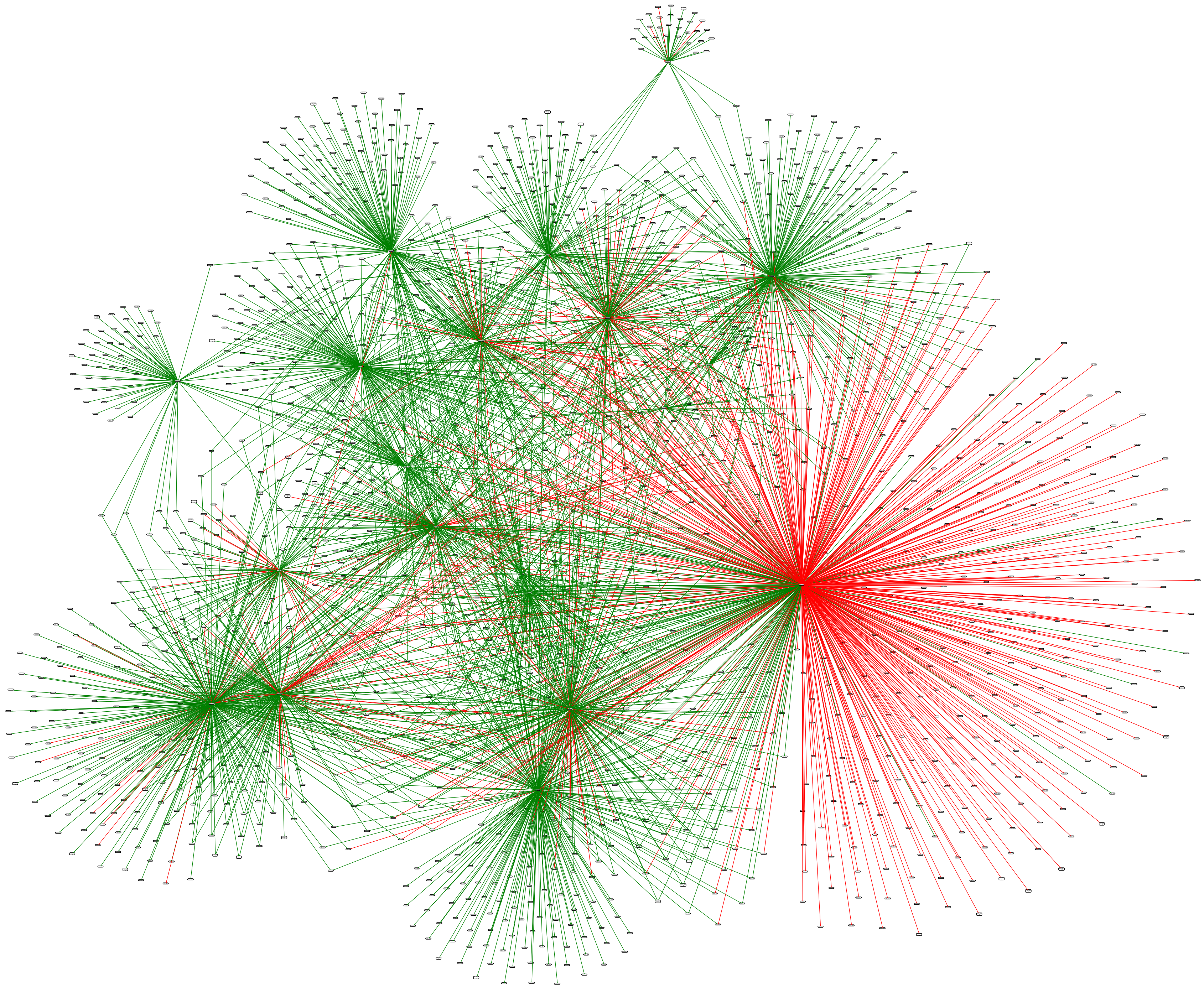
Alvis Brazma

European Bioinformatics Institute

Lecture 7

- Gene networks – part 1


GENE 1
GENE 2
GENE 3
GENE 4
DNA
promoter
coding DNA
transcription factor



Questions I want to ask

- What does it mean to understand a network of thousands of genes and connections?
- How does a simple cell (with < 6000 genes) work?
- What does it mean to understand ‘how does a cell work’?
- Can a descriptive approach to biology ever provide the answer?

Modelling approach

- Develop a model (a formal language) describing gene networks
- Study the properties of the model instead of the real world gene networks directly
- Make predictions about real world gene networks based on the properties of the model
- Test the predictions in the real world
- If the predictions are correct – the model is correct

All models are wrong, but some are useful

- George E. P. Box

Simulation and reverse engineering of gene networks

- Simulation - given a model, observe its behaviour and compare to gene expression data from real networks
- Reverse engineering - given gene expression data construct a particular model (in the given model class) that is consistent with the data

Approach to gene network modelling - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …
- **Topology** – a graph describing the connections between the parts
- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)
- **Dynamics** – how does it all work in real time

Each level models different network properties, each next level includes more detail

Gene Networks - four levels of hierarchical description

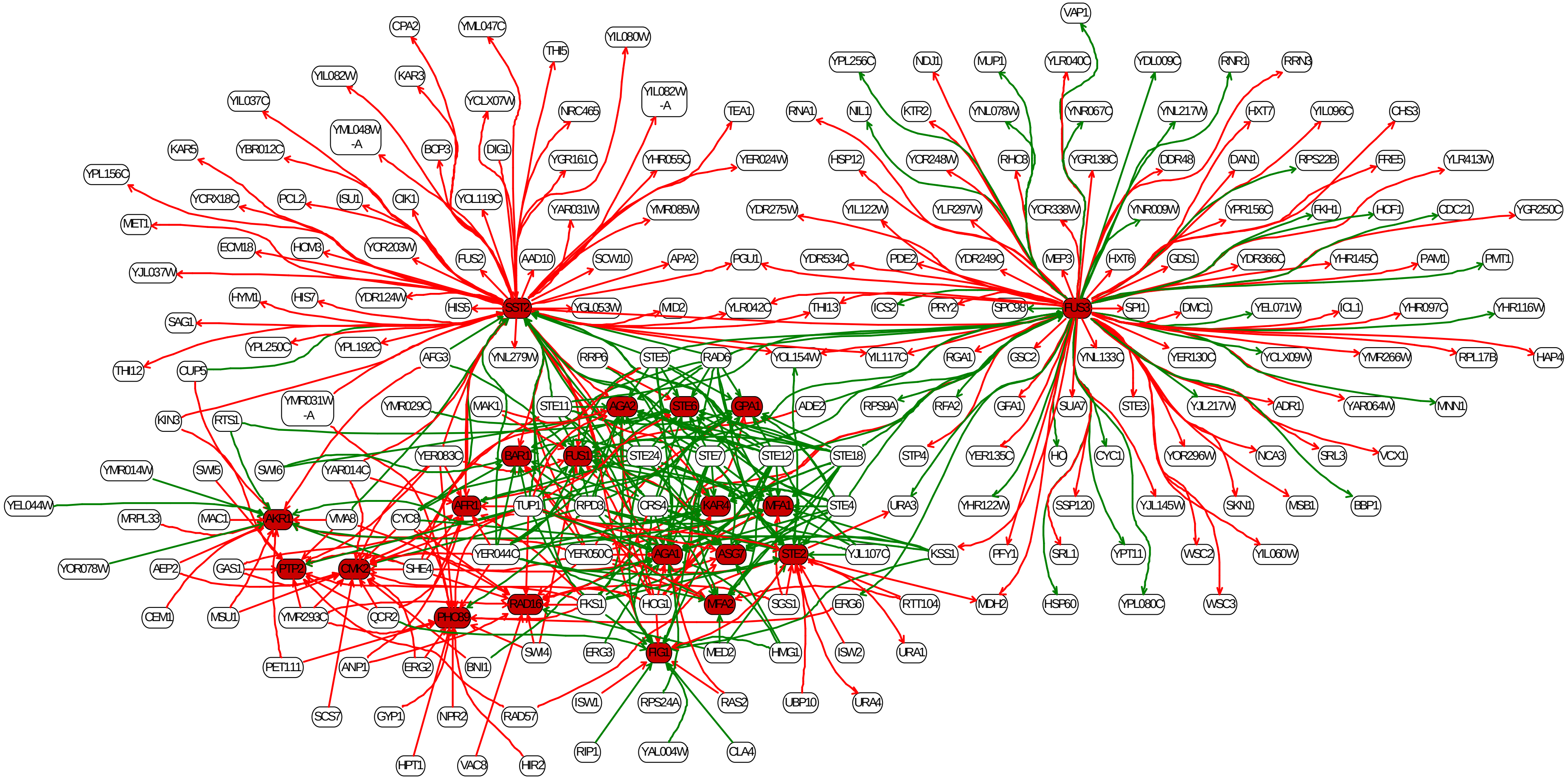
- **Parts list** – genes, transcription factors, promoters, binding sites, …
- **Topology** – a graph describing the connections between the parts
- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)
- **Dynamics** – how does it all work in real time

Genes and gene products, proteins

Organism	The number of predicted genes	Part of the genome that encodes proteins (exons)
E.Coli (bacteria)	5000	90%
Yeast	6000	70%
Worm	18,000	27%
Fly	14,000	20%
Weed	25,500	20%
Human	25,000	< 5%

Gene Networks - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …
- **Topology** – a graph describing the connections between the parts
- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)
- **Dynamics** – how does it all work in real time

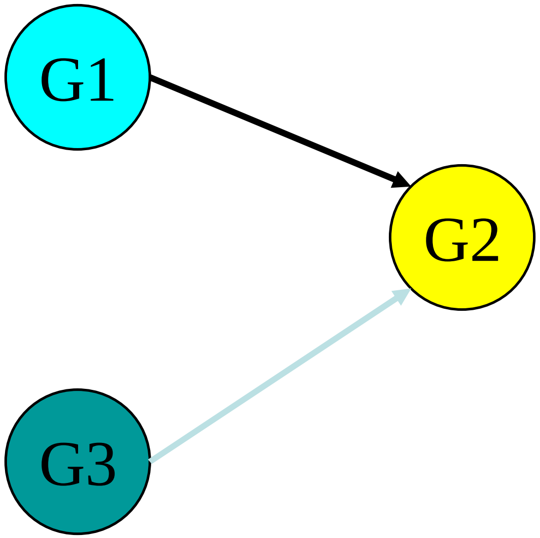


VAP1
CPA2
YML047C
YIL080W
THI5
YPL256C
NDJ1
MUP1
YLR040C
YDL009C
RNR1
RRN3
YIL082W
KAR3
YIL082W -A
YIL037C
YCLX07W
NRC465
TEA1
RNA1
NIL1
KTR2
YNL078W
YNR067C
YNL217W
HXT7
YIL096C
CHS3
YML048W -A
KAR5
YBR012C
BOP3
DIG1
YGR161C
YHR055C
YER024W
HSP12
YOR248W
RHO3
YGR138C
DDR48
DAN1
RPS22B
FRE5
YLR413W
YPL156C
YCRX18C
PCL2
ISU1
CIK1
YOL119C
YAR031W
YMR085W
YDR275W
YIL122W
YLR297W
YOR338W
YNR009W
YPR156C
FKH1
HOF1
CDC21
YGR250C
MET1
ECM18
HOM3
YOR203W
FUS2
AAD10
SCW10
APA2
PGU1
YDR534C
PDE2
YDR249C
MEP3
HXT6
GDS1
YDR366C
YHR145C
PAM1
PMT1
YJL037W
HYM1
HIS7
YDR124W
HIS5
SST2
YGL053W
MID2
YLR042C
THI13
ICS2
PRY2
SPC98
FUS3
SPI1
DMC1
YEL071W
ICL1
YHR097C
YHR116W
SAG1
YPL250C
YPL192C
AFG3
YNL279W
RRP6
STE5
RAD6
YOL154W
YIL117C
RGA1
GSC2
YNL133C
YER130C
YCLX09W
YMR266W
RPL17B
HAP4
THI12
CUP5
YMR031W -A
YMR029C
MAK1
STE11
AGA2
STE6
GPA1
ADE2
RPS9A
RFA2
GFA1
SUA7
STE3
YJL217W
ADR1
YAR064W
MNN1
KIN3
RTS1
YER083C
BAR1
FUS1
STE24
STE7
STE12
STE18
STP4
YER135C
HO
CYC1
YOR296W
NCA3
SRL3
VCX1
YMR014W
SWI5
SWI6
YAR014C
AFR1
TUP1
RPD3
CRS4
KAR4
MFA1
STE4
URA3
YHR122W
SSP120
YJL145W
SKN1
MSB1
BBP1
YEL044W
MRPL33
MAC1
AKR1
VMA8
CYC8
YER044C
YER050C
AGA1
ASG7
STE2
YJL107C
KSS1
PFY1
SRL1
YPT11
WSC2
YIL060W
YOR078W
AEP2
GAS1
PTP2
CMK2
SHE4
RAD16
FKS1
HOG1
MFA2
SGS1
ERG6
RTT104
MDH2
HSP60
YPL080C
WSC3
CEM1
MSU1
YMR293C
QCR2
PHO89
SWI4
ERG3
FIG1
MED2
HMG1
ISW2
URA1
PET111
ANP1
ERG2
BNI1
ISW1
RPS24A
RAS2
UBP10
URA4
SCS7
GYP1
NPR2
RAD57
RIP1
YAL004W
CLA4
HPT1
VAC8
HIR2

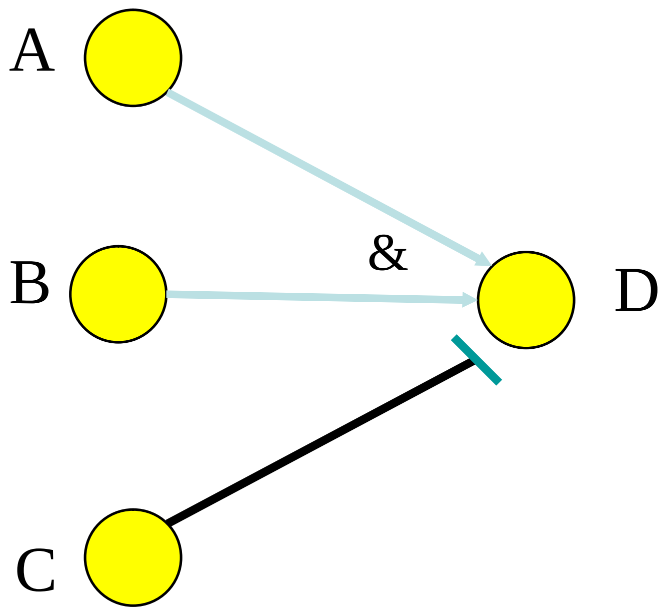
Gene Networks - four levels of hierarchical description

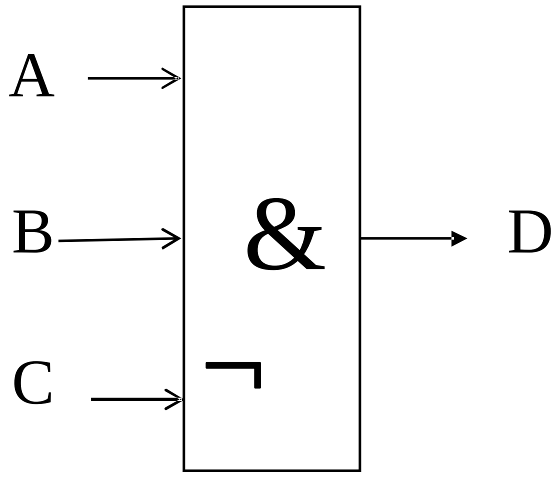
- **Parts list** – genes, transcription factors, promoters, binding sites, …
- **Topology** – a graph describing the connections between the parts
- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)
- **Dynamics** – how does it all work in real time

Gene activation, repression, more complex combinatorial effects

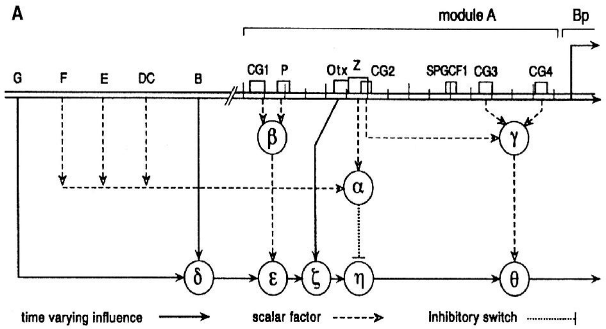


Logics





$$D = A \,\&\, B \,\&\, \neg C$$



B

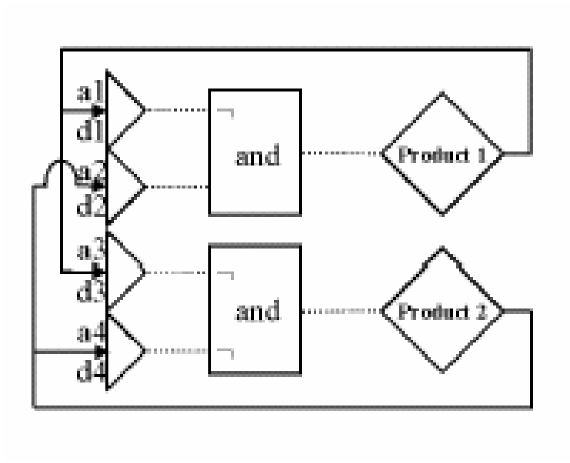
if (F = 1 or E = 1 or CD = 1) and (Z = 1) $\alpha = 1$ **else** $\alpha = 0$	Repression functions of modules F, E, and DC mediated by Z site
if (P = 1 and CG_1 = 1) $\beta = 2$ **else** $\beta = 0$	Both P and CG_1 needed for synergistic link with module B
if (CG_2 = 1 and CG_3 = 1 and CG_4 = 1) $\gamma = 2$ **else** $\gamma = 1$	Final step up of system output
$\delta(t) = B(t) + G(t)$	Positive input from modules B and G
$\varepsilon(t) = \beta^{*}\delta(t)$	Synergistic amplification of module B output by CG_1-P subsystem
if ($\varepsilon(t) = 0$) $\xi(t) = Otx(t)$ **else** $\xi(t) = \varepsilon(t)$	Switch determining whether Otx site in module A, or upstream modules (i.e., mainly module B), will control level of activity
if ($\alpha = 1$) $\eta(t) = 0$ **else** $\eta(t) = \xi(t)$	Repression function inoperative in endoderm but blocks activity elsewhere
$\Theta(t) = \gamma^{*}\eta(t)$	Final output communicated to BTA

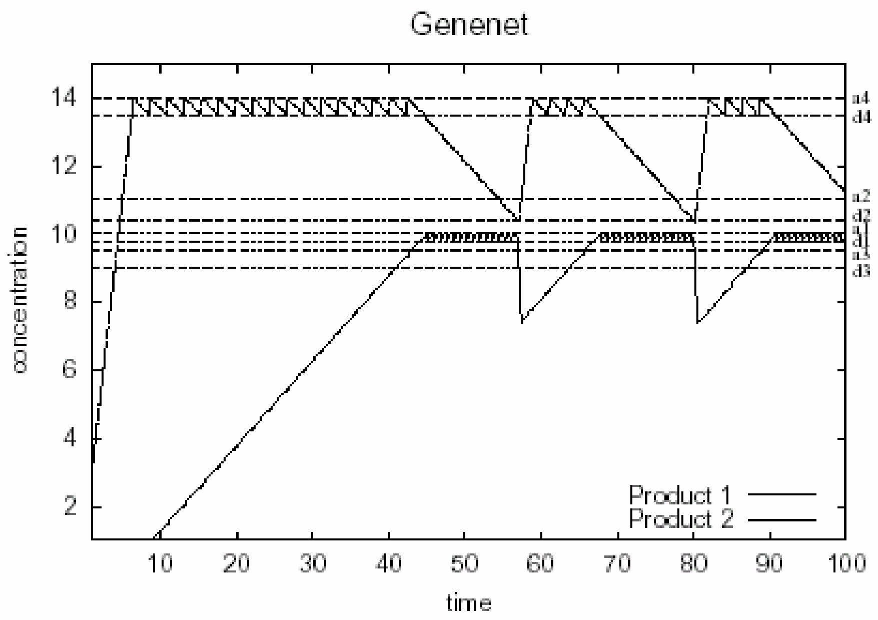
Yuh, C.H., Bolouri, H. and Davidson, E.H. (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279, 1896-902

Gene Networks - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …
- **Topology** – a graph describing the connections between the parts
- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)
- **Dynamics** – how does it all work in real time

Simulations on a FSLM





Gene Networks - four levels of hierarchical description

- **Parts list** – genome scale
- **Topology** – genome scale for smaller genomes (Yeasts, E. Coli)
- **Control logics** – tens of genes
- **Dynamics** – typically a couple of genes

Gene Networks - four levels of hierarchical description

- **Parts list** – genome scale
- **Topology** – genome scale for smaller genomes (Yeasts, E. Coli)
- **Control logics** – tens of genes
- **Dynamics** – typically a couple of genes

Parts list

- Classification of all genes and their products (transcripts, proteins) – *an ontology or a database*
  - Some particularly important for regulation classes of genes – transcription factors, signalling proteins
  - Transcription factor binding sites, promoters - e.g, TRANSFAC database

Gene Ontology (GO)

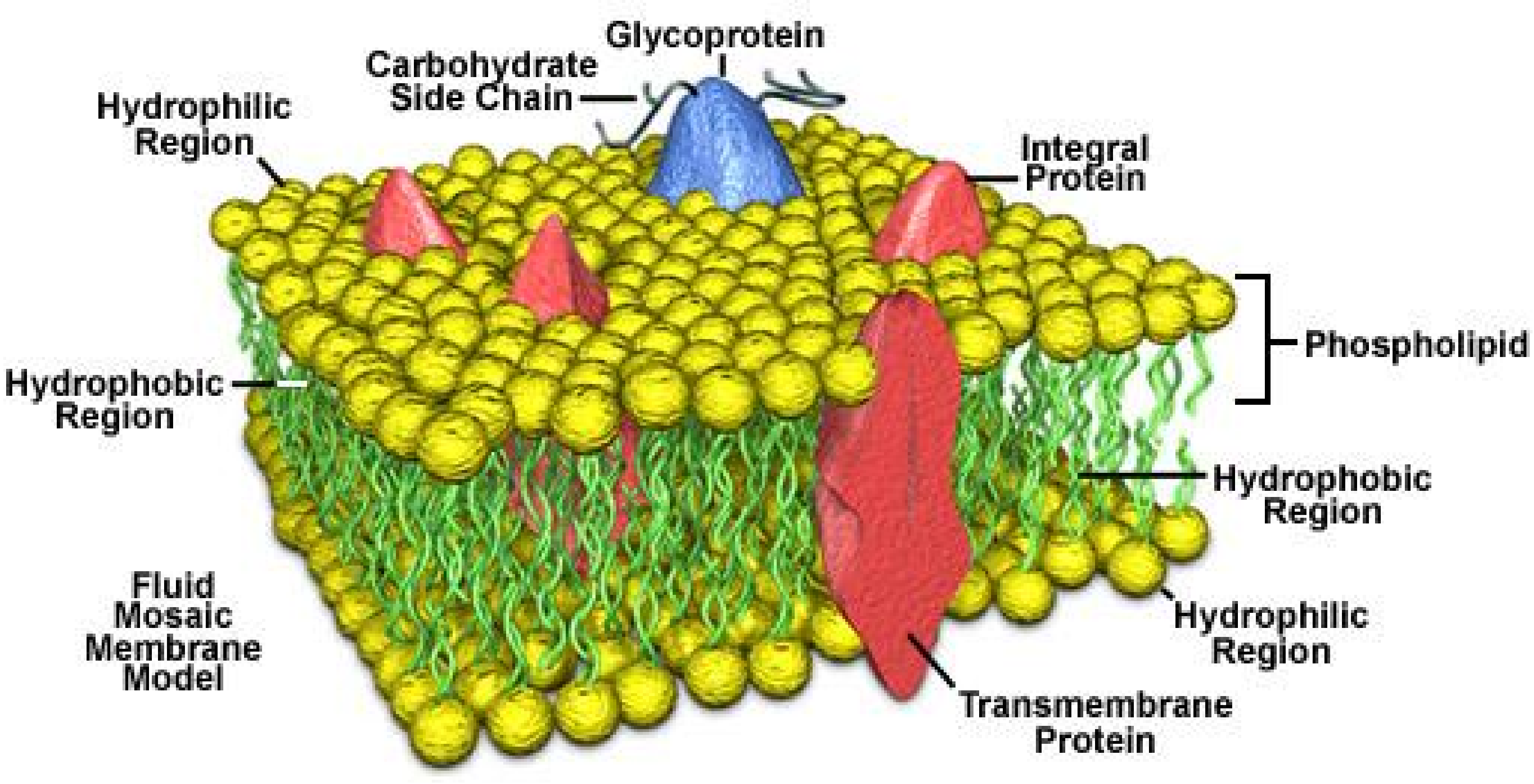
What information might we want to capture about a gene product?

- What does the gene product do?
- Where and when does it act?
- Why does it perform these activities?

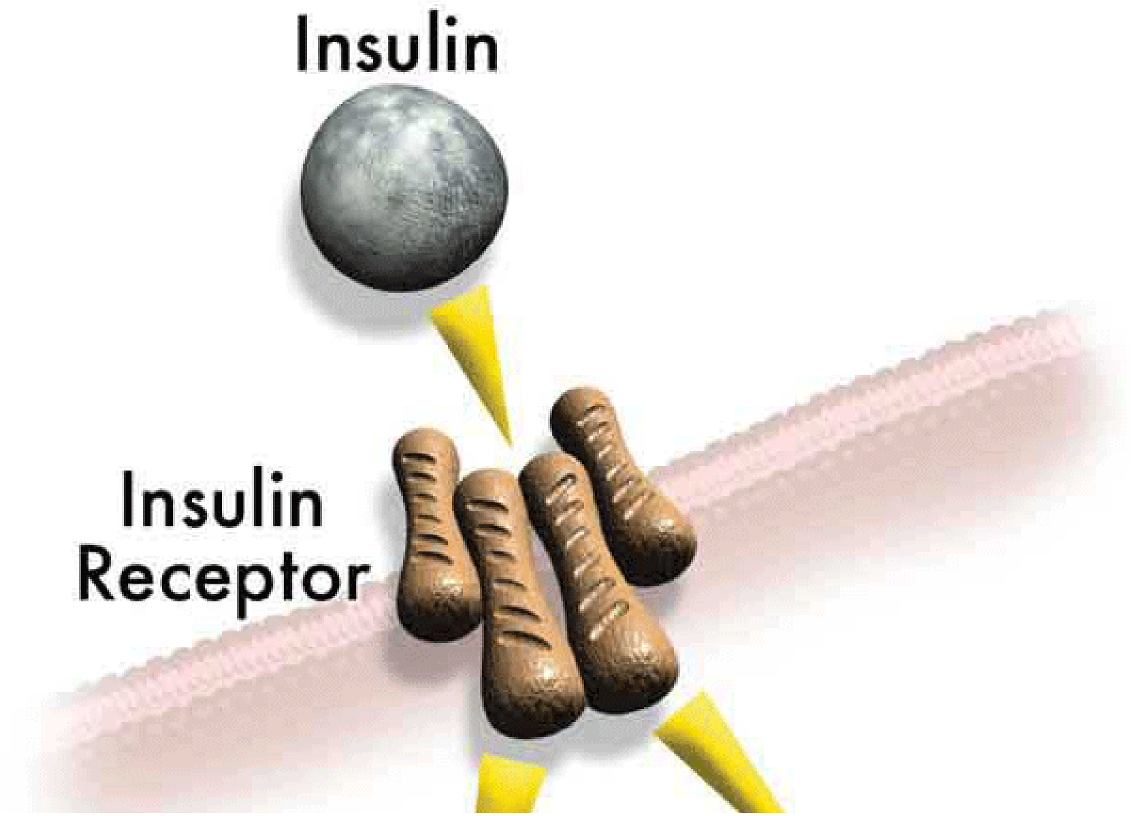
GO structure

- GO terms divided into three parts:
  - cellular component
  - molecular function
  - biological process

Cellular Component



Molecular Function



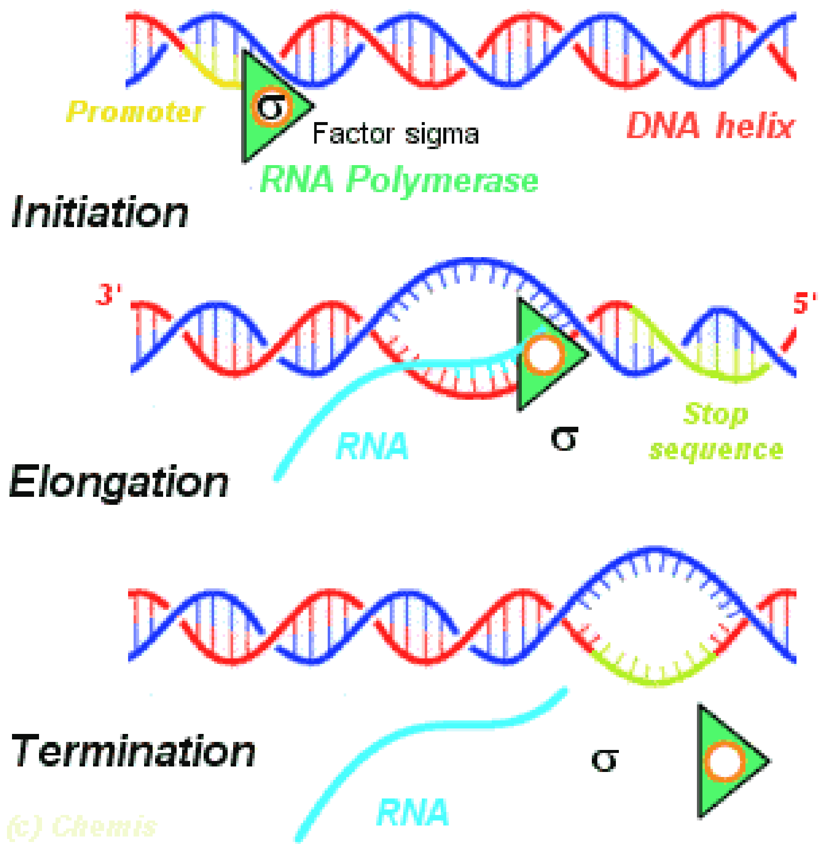
insulin binding

insulin receptor activity

Molecular Function

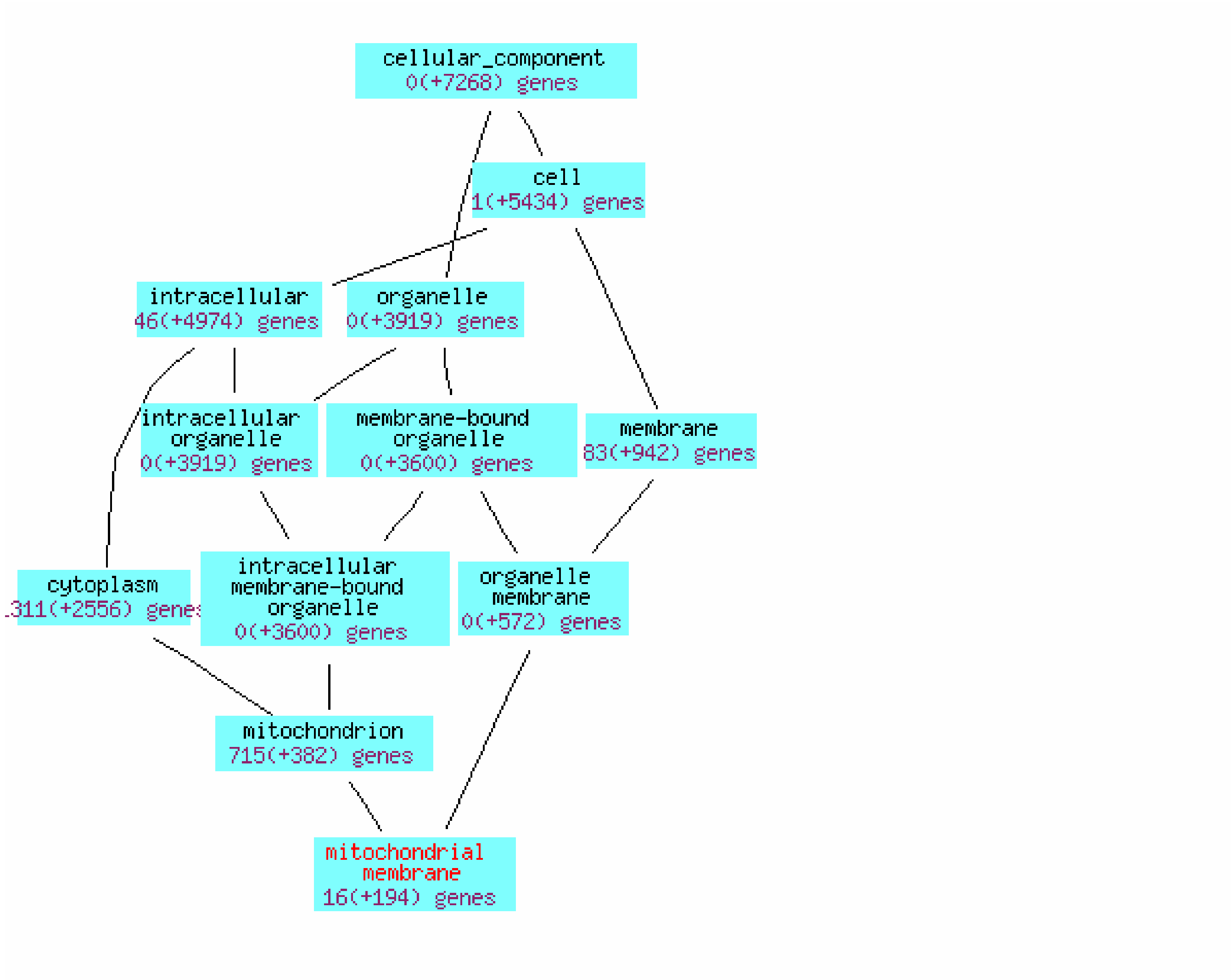
- A gene product may have several functions; a function term refers to a single reaction or activity, not a gene product.
- Sets of functions make up a biological process.

Biological Process

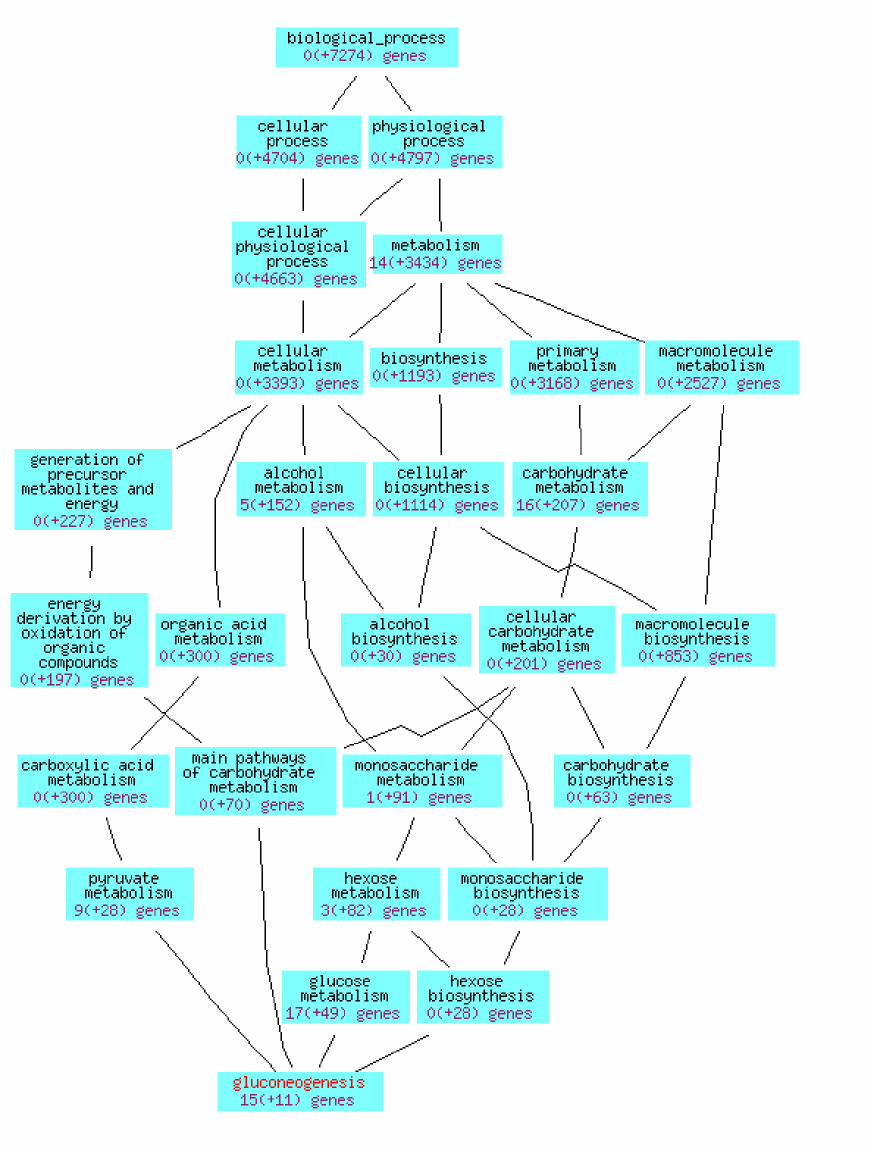


transcription

DAG structure - Mitochondrial membrane



Gluconeogenesis

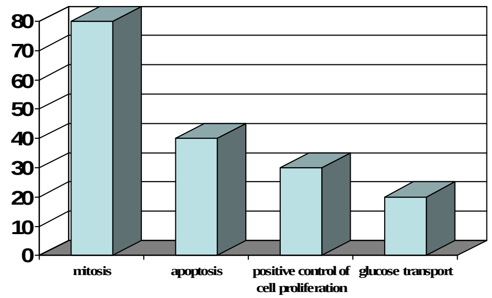


GO for microarray analysis

- Annotations give ‘function’ label to genes
- Ask meaningful questions of microarray data e.g.
  - genes involved in the same process, same/different expression patterns?

Using GO in practice

- statistical measure
  - how likely your differentially regulated genes fall into that category by chance



microarray

1000 genes

→ experiment

→ 100 genes differentially regualted

→

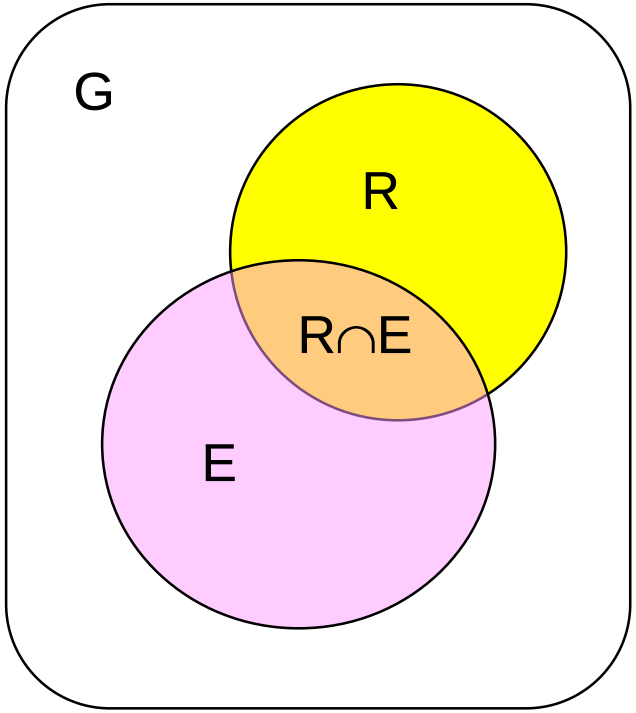
mitosis – 80/100
apoptosis – 40/100
p. ctrl. cell prol. – 30/100
glucose transp. – 20/100

Using GO in practice

- However, when you look at the distribution of all genes on the microarray:

Process	Genes on array	# genes expected in 100 random genes	occurred
mitosis	800/1000	80	80
apoptosis	400/1000	40	40
p. ctrl. cell prol.	100/1000	10	30
glucose transp.	50/1000	5	20

How to estimate that the overlap is more than expected by random?

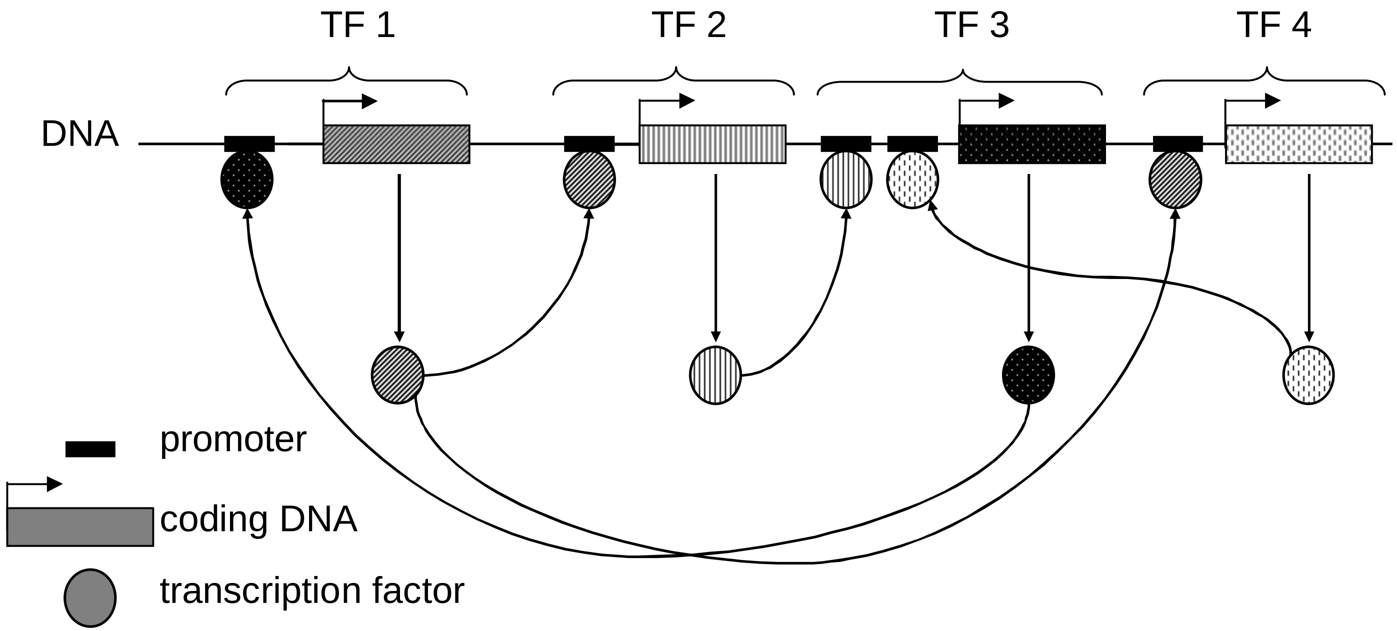


We assume that the elements of the set E are marked, and pick the set of size |R| at random. Then the size x=|R∩E| of the intersection are distributed according to *hypergeometric* distribution.
The probability of observing an intersection of size k or larger can be computed according to formula:

$$P(x \geq k) = 1 - \sum_{i=0}^{k} \frac{\binom{|E|}{i}\binom{|G|-|E|}{|R|-i}}{\binom{|G|}{|R|}}$$

Parts list - number of transcription factors

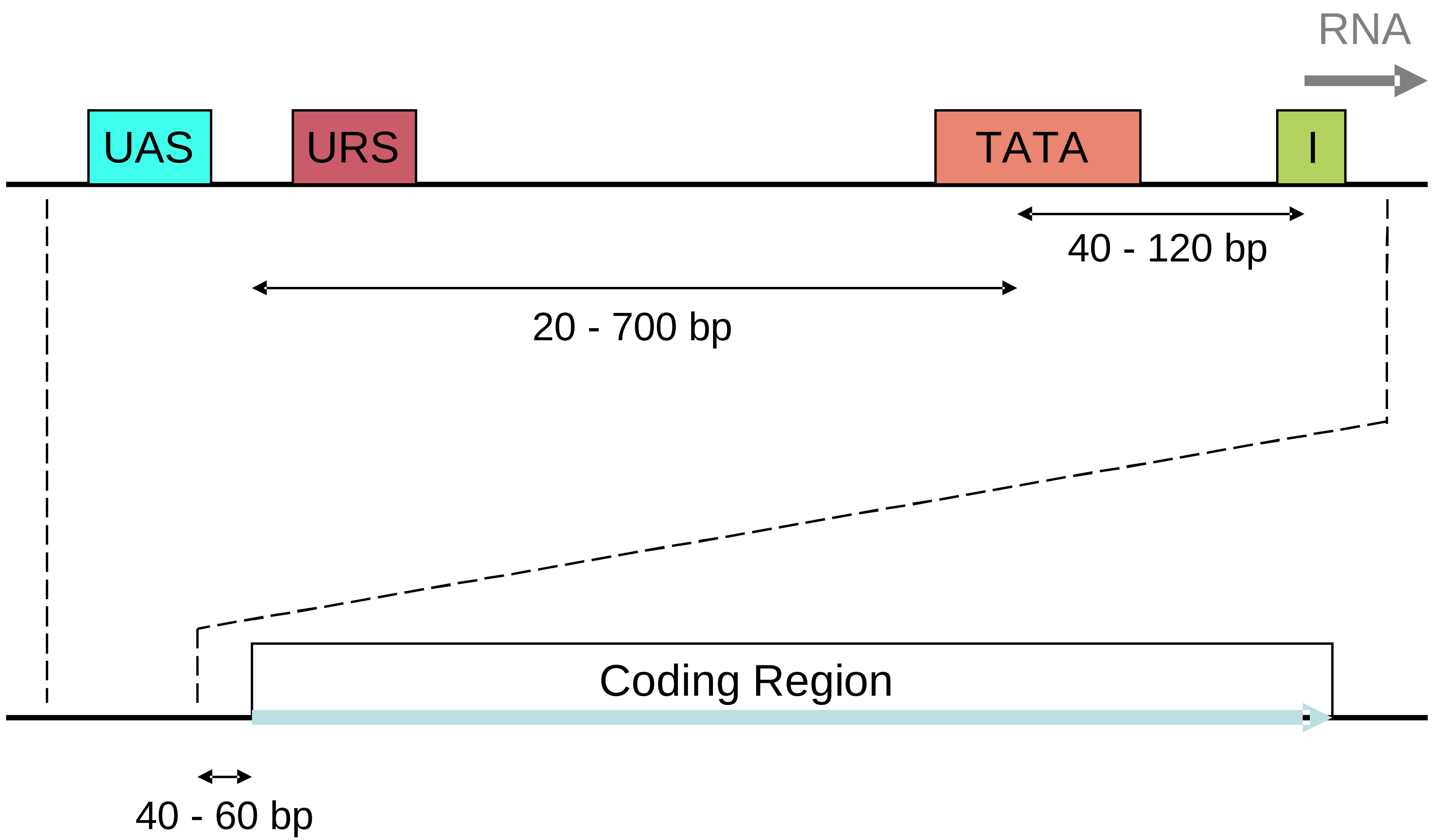
Organism	Number of genes	Number of transcription regulators (GO:0030528)
Yeast	6682	312 (4.7%)
Fly	13525	492 (3.6%)
Human	22287	1034 (4.6%)


TF 1
TF 2
TF 3
TF 4
DNA
promoter
coding DNA
transcription factor

Transcription factor binding sites and promoters

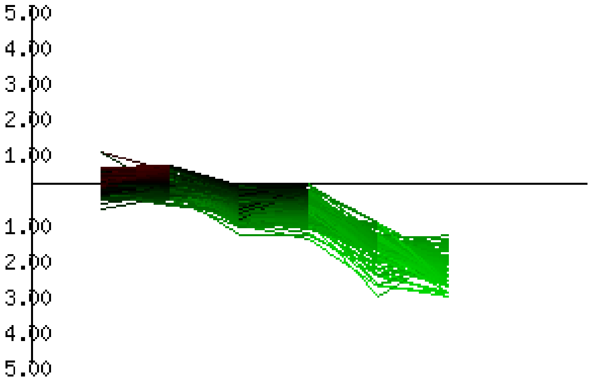
- Binding site identification is much more elusive than gene identification

Organization of a typical yeast promoter

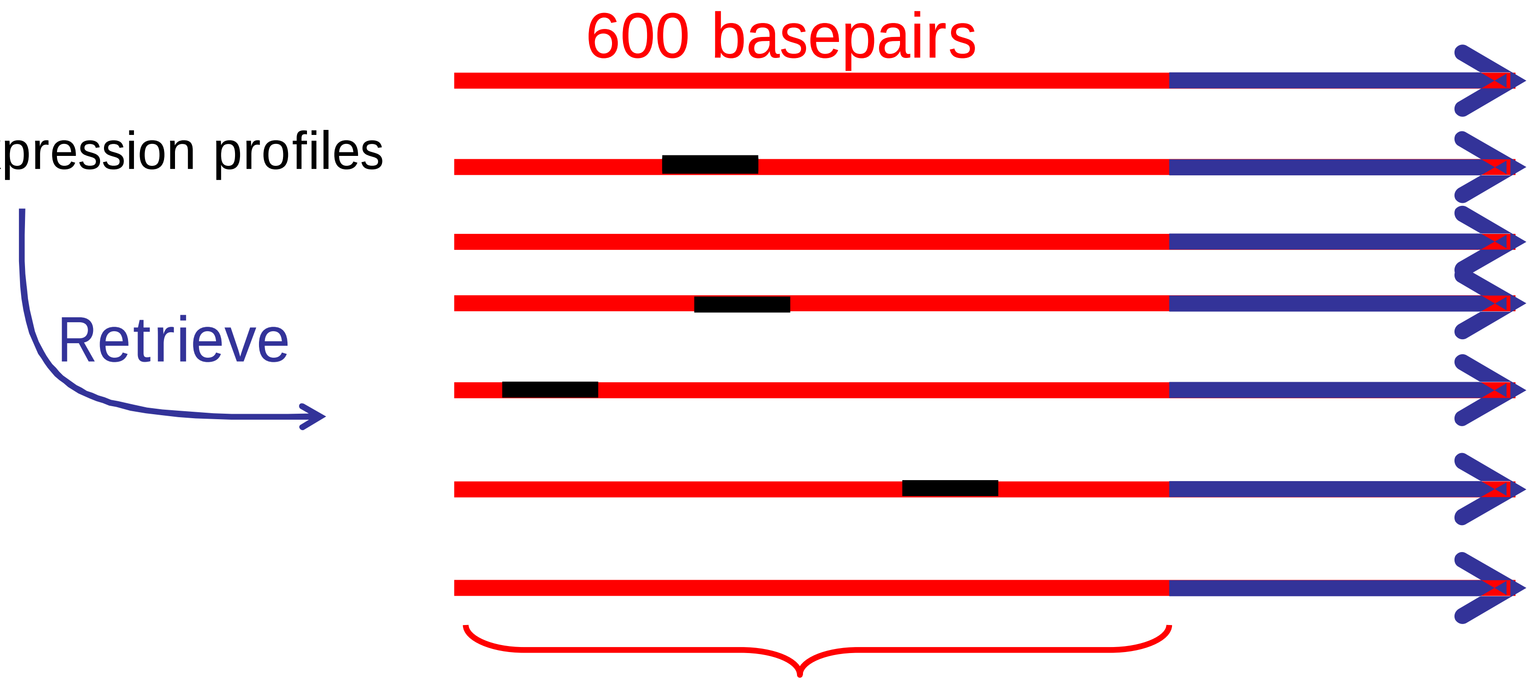


Computational identification of TF binding sites

Search for over-represented patterns in clusters of putative regulatory regions



Expression profiles

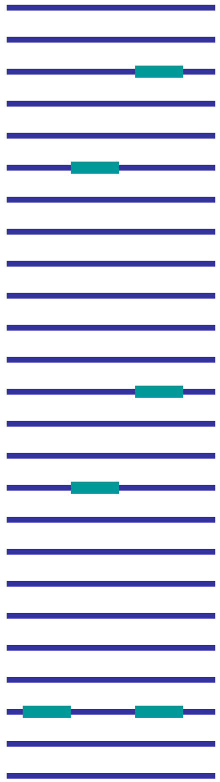


Upstream regions

▬ Pattern over-represented in cluster

Pattern selection criteria
Binomial distribution

Background - ALL upstream sequences



Cluster:

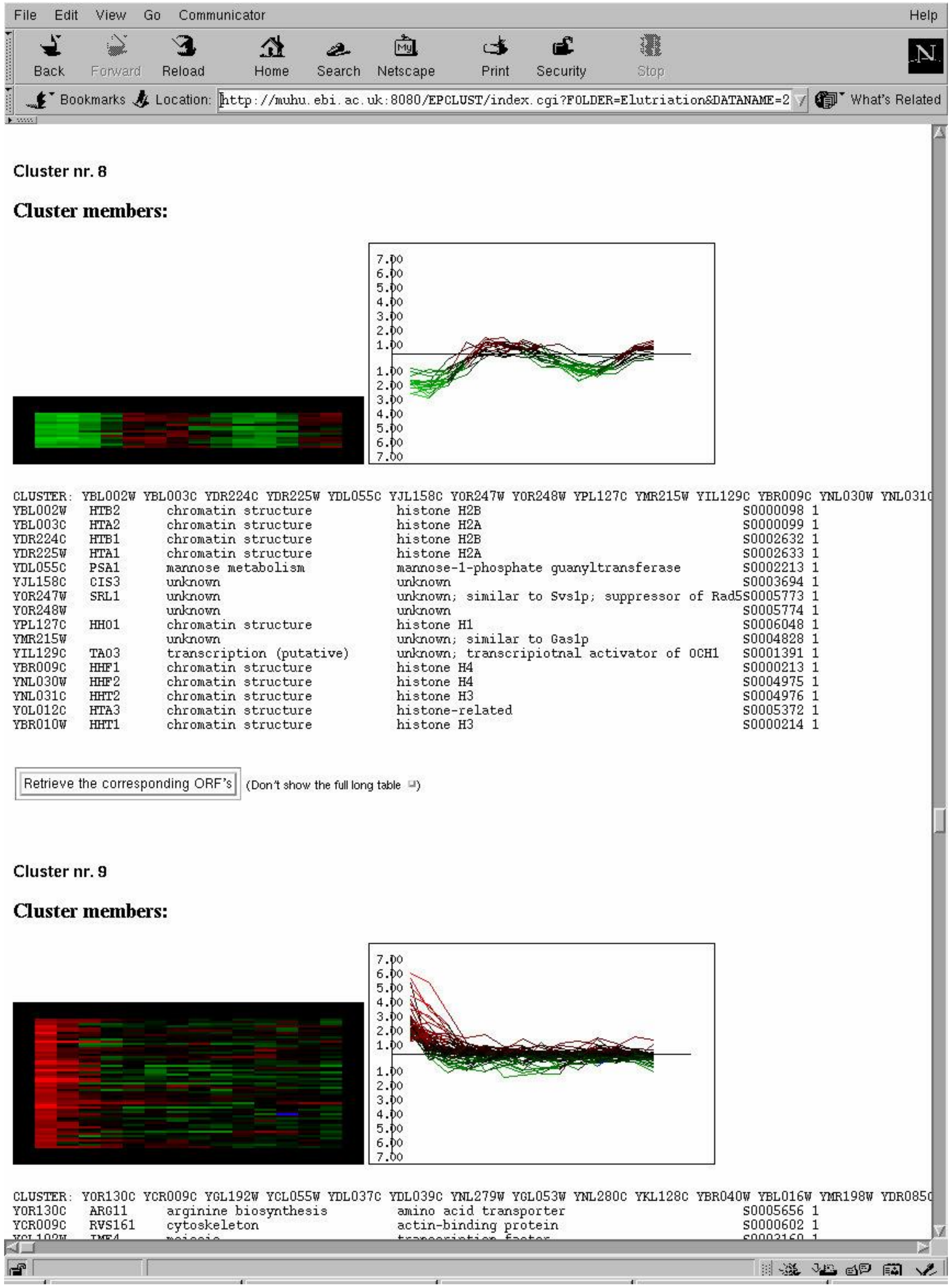
π occurs 3 times

P(π,6) is probability of having 3 or more matches in 6 sequences

5 out of 25, ***p* = 0.2**

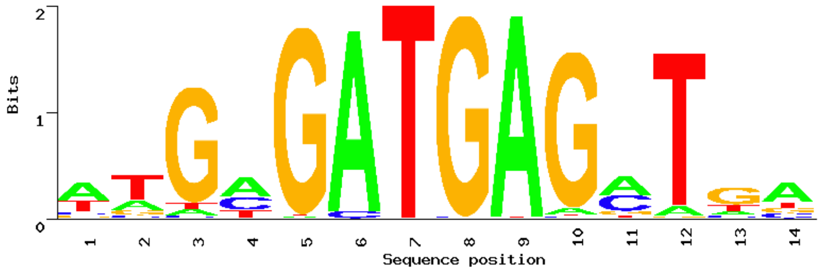
P(π,6) =0.0989

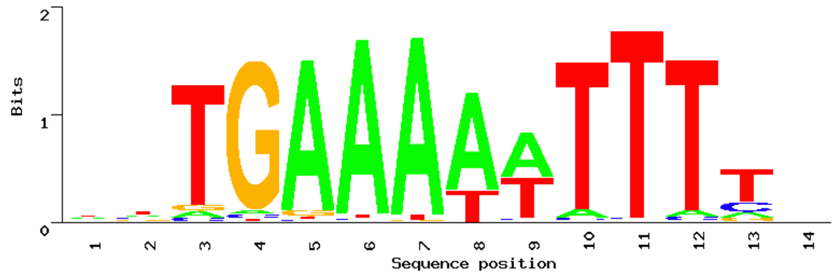
Clustering of gene expression data (K-means clustering)



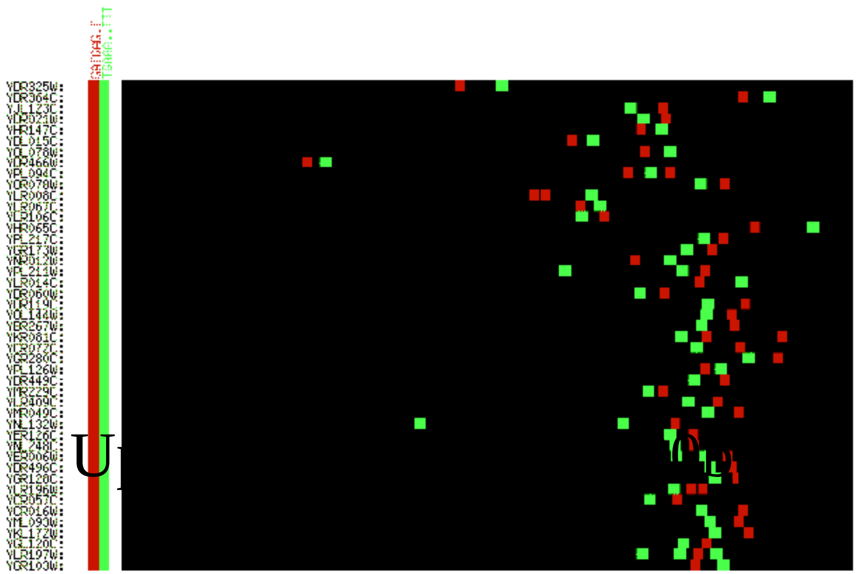
1 mismatch

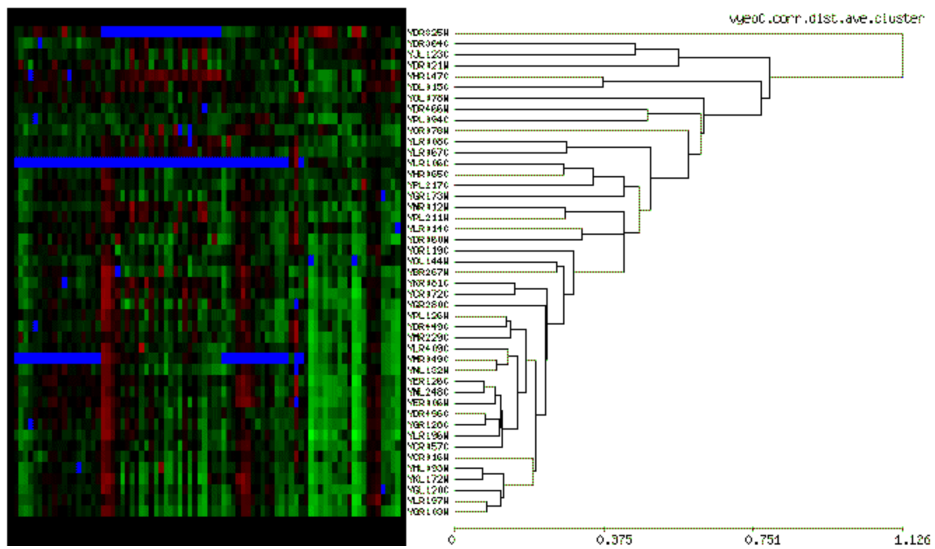
GATGAG.T
TGAAA..TTT



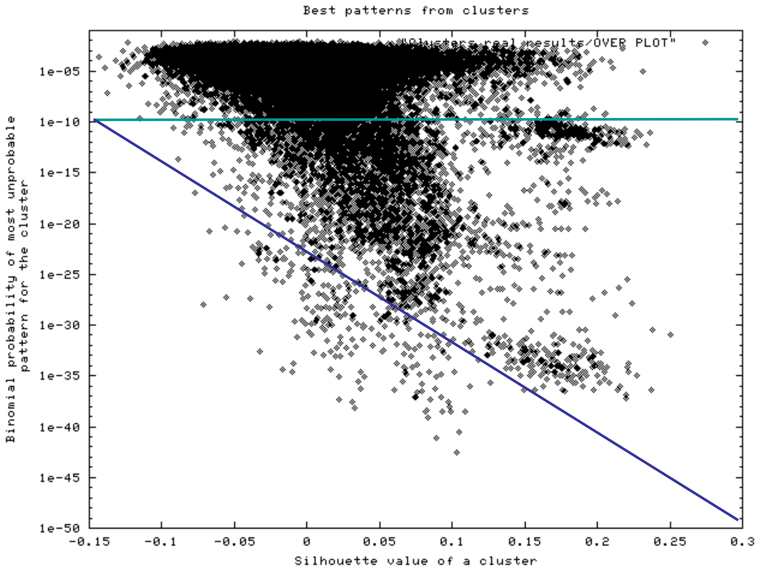


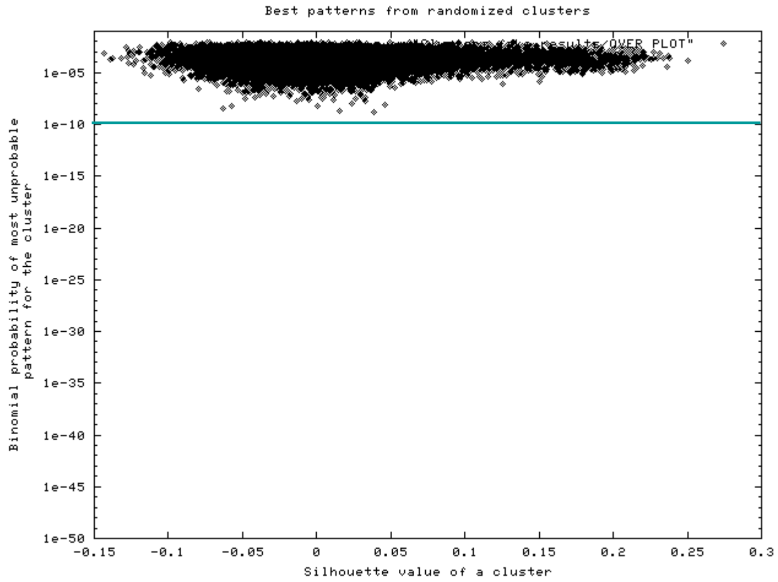
GATGAG.T W/30
TGAAA..TTT





Pattern vs cluster "strength"

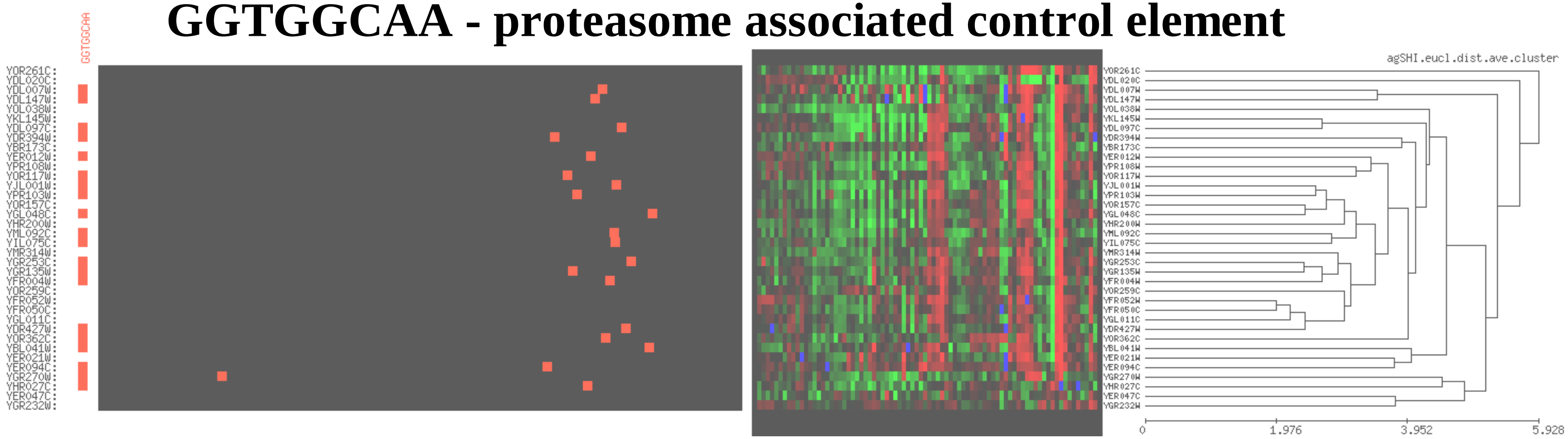




The pattern probability vs. the average silhouette for the cluster

The same for randomised clusters

GGTGGCAA - proteasome associated control element



YOR261C	YOR261C	RPN8	protein degradation	26S proteasome regulatory subunit	S0005787	1
YDL020C	YDL020C	RPN4	protein degradation, ubiquitin	26S proteasome subunit	S0002178	1
YDL007W	YDL007W	RPT2	protein degradation	26S proteasome subunit	S0002165	1
YDL147W	YDL147W	RPN5	protein degradation	26S proteasome subunit	S0002306	1
YOL038W	YOL038W	PRE6	protein degradation	20S proteasome subunit (alpha4)	S0005398	1
YKL145W	YKL145W	RPT1	protein degradation, ubiquitin	26S proteasome subunit	S0001628	1
YDL097C	YDL097C	RPN6	protein degradation	26S proteasome regulatory subunit	S0002255	1
YDR394W	YDR394W	RPT3	protein degradation	26S proteasome subunit	S0002802	1
YBR173C	YBR173C	UMP1	protein degradation, ubiquitin	20S proteasome maturation factor	S0000377	1
YER012W	YER012W	PRE1	protein degradation	20S proteasome subunit C11(beta4)	S0000814	1
YPR108W	YPR108W	RPN7	protein degradation	26S proteasome regulatory subunit	S0006312	1
YOR117W	YOR117W	RPT5	protein degradation	26S proteasome regulatory subunit	S0005643	1
YJL001W	YJL001W	PRE3	protein degradation	20S proteasome subunit (beta1)	S0003538	1
YPR103W	YPR103W	PRE2	protein degradation	20S proteasome subunit (beta5)	S0006307	1
YOR157C	YOR157C	PUP1	protein degradation	20S proteasome subunit (beta2)	S0005683	1
YGL048C	YGL048C	RPT6	protein degradation	26S proteasome regulatory subunit	S0003016	1
YHR200W	YHR200W	RPN10	protein degradation	26S proteasome subunit	S0001243	1
YML092C	YML092C	PRE8	protein degradation	20S proteasome subunit Y7 (alpha2	S0004557	1
YIL075C	YIL075C	RPN2	tRNA processing	26S proteasome subunit)	S0001337	1
YMR314W	YMR314W	PRE5	protein degradation	20S proteasome subunit(alpha6)	S0004931	1
YGR253C	YGR253C	PUP2	protein degradation	20S proteasome subunit(alpha5)	S0003485	1
YGR135W	YGR135W	PRE9	protein degradation	20S proteasome subunit Y13 (alpha3)	S0003367	1
YFR004W	YFR004W	RPN11	transcription	putative global regulator	S0001900	1
YOR259C	YOR259C	RPT4	protein degradation	26S proteasome regulatory subunit	S0005785	1
YFR052W	YFR052W	RPN12	protein degradation	26S proteasome regulatory subunit	S0001948	1
YFR050C	YFR050C	PRE4	protein degradation	proteasome subunit, B type	S0001946	1
YGL011C	YGL011C	SCL1	protein degradation	20S proteasome subunit YC7ALPHA/Y8	S0002979	1
YDR427W	YDR427W	RPN9	protein degradation	26S proteasome regulatory subunit	S0002835	1
YOR362C	YOR362C	PRE10	protein degradation	20S proteasome subunit C1 (alpha7)	S0005889	1
YBL041W	YBL041W	PRE7	protein degradation	20S proteasome subunit	S0000137	1
YER021W	YER021W	RPN3	protein degradation	26S proteasome regulatory subunit	S0000823	1
YER094C	YER094C	PUP3	protein degradation	20S proteasome subunit (beta3	S0000896	1
YGR270W	YGR270W	YTA7	protein degradation	26S proteasome subunit; ATPase	S0003502	1
YHR027C	YHR027C	RPN1	protein degradation	26S proteasome regulatory subunit	S0001069	1
YER047C	YER047C	SAP1	mating type switching	AAA family protein	S0000849	1
YGR232W	YGR232W		unknown	unknown	S0003464	1

GGTGGCAA is a binding site for RPN4

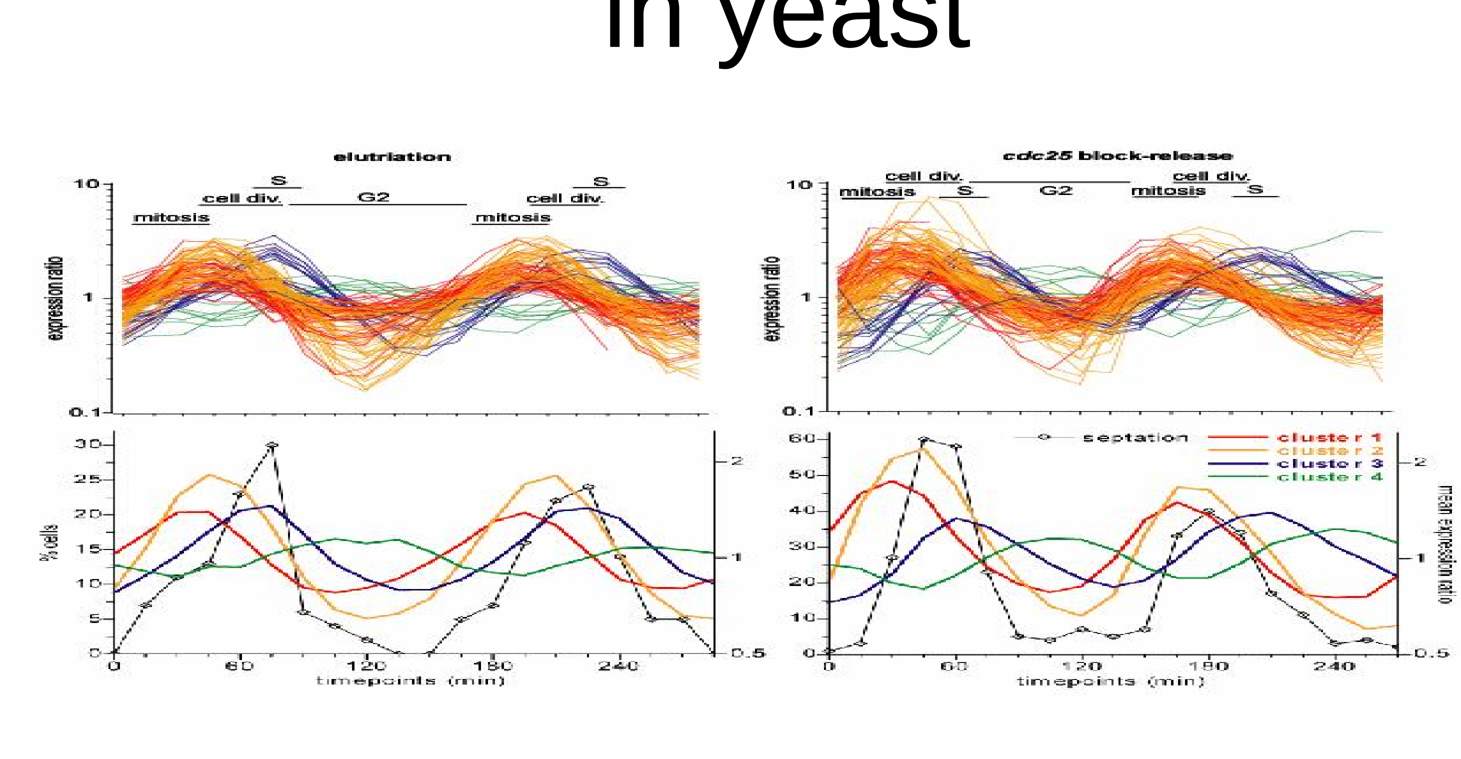
FEBS Lett **1999 Apr 30;450(1-2):27-34**

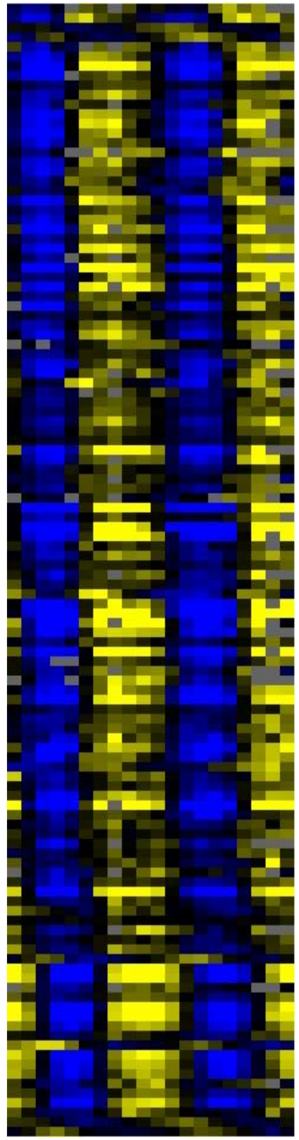
Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast.

Mannhaupt G, Schnall R, Karpov V, Vetter I, Feldmann H
Adolf-Butenandt-Institut der Ludwig-Maximilians-Universitat Munchen, Germany.

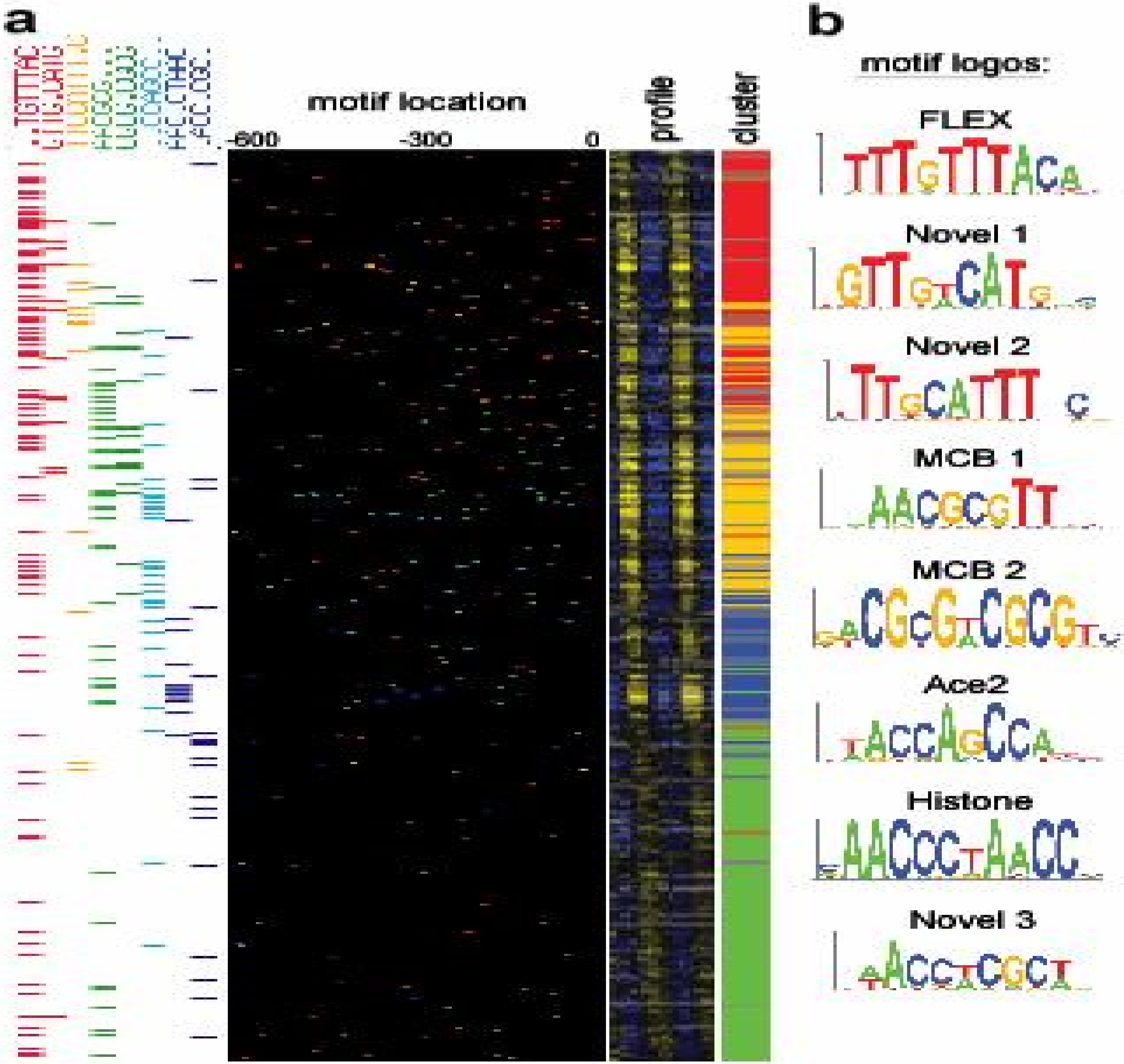
We identified a new, unique upstream activating sequence (5'-**GGTGGCAAA**-3') in the promoters of 26 out of the 32 proteasomal yeast genes characterized to date, which we propose to call proteasome-associated control element. By using the one-hybrid method, we show that the factor binding to the proteasome-associated control element is Rpn4p, a protein containing a C2H2-type finger motif and two acidic domains. ...

Microarray expression measurements in cell cycle for over 400 periodic genes in yeast





Rustici et al, Nature Genetics, 2004



Kivinen at al

Computational TF binding site identification

- Works OK for yeasts, but even for S. Pombe difficult
- For human this type of strategy does not work – the putative promoter regions are too long
- The best that has been achieved for human is to use known binding site patterns, to try to find where they are in the genome

Known binding sites for *S.pombe*

CAGTCACA	HomolD	(translation)
ACCCTACCCT	HomolE	(translation)
ACGCGT	MCB	(cell cycle)
TTCTTTGTTY	TR-box	(mating)
ACAAT	M-box	(mating)
TTTGTTTAC	FLEX	(meiosis)
GAAnnTTC	HSE	(stress)
TGACGTCA	CRE	(stress)

Alternative methods for high throughput binding site identification

- ChIP-on-chip – identify intragenic sequences of a few hundred base-pairs binding a particular transcription factor, then look for an overrepresented sequence elements
- Protein binding arrays – hybridise the transcription factor directly on the array
- Phylogenetic foot printing or shadowing

ChIP-chip (Chromatin Immuno Precipitation on chip) experiments to identify TF binding sites

- The method
  - TF are cross-linked to genomic DNA with Chromatin IP
  - The DNA is fragmented and nonprotein binding bits washed away
  - The remaining DNA is labelled and hybridised on a microarray containing intragenic regions
  - The spot brightness now tells where TF were bound
- Problems
  - Binding is still condition specific
  - Are the binding functional?

Problems in binding site identification

- They are all based on the assumption that statistically overrepresented sequence elements are functional
- They are all based the assumption that the binding sites can be described by regular expressions or position weight matrices
- They work on yeasts around ~50% OK, but so far they have failed in higher organisms
- On the order of 4000 TF BS location for S. Cerevisiae

Parts list - conclusions

- Gene identification - OK
- Gene function – only 1/3 of the genes have known function
- Transcription Factors – 1/3 – 2/3
- Transcription factor binding site identification – More or less OK for yeast, rather poor for higher organism

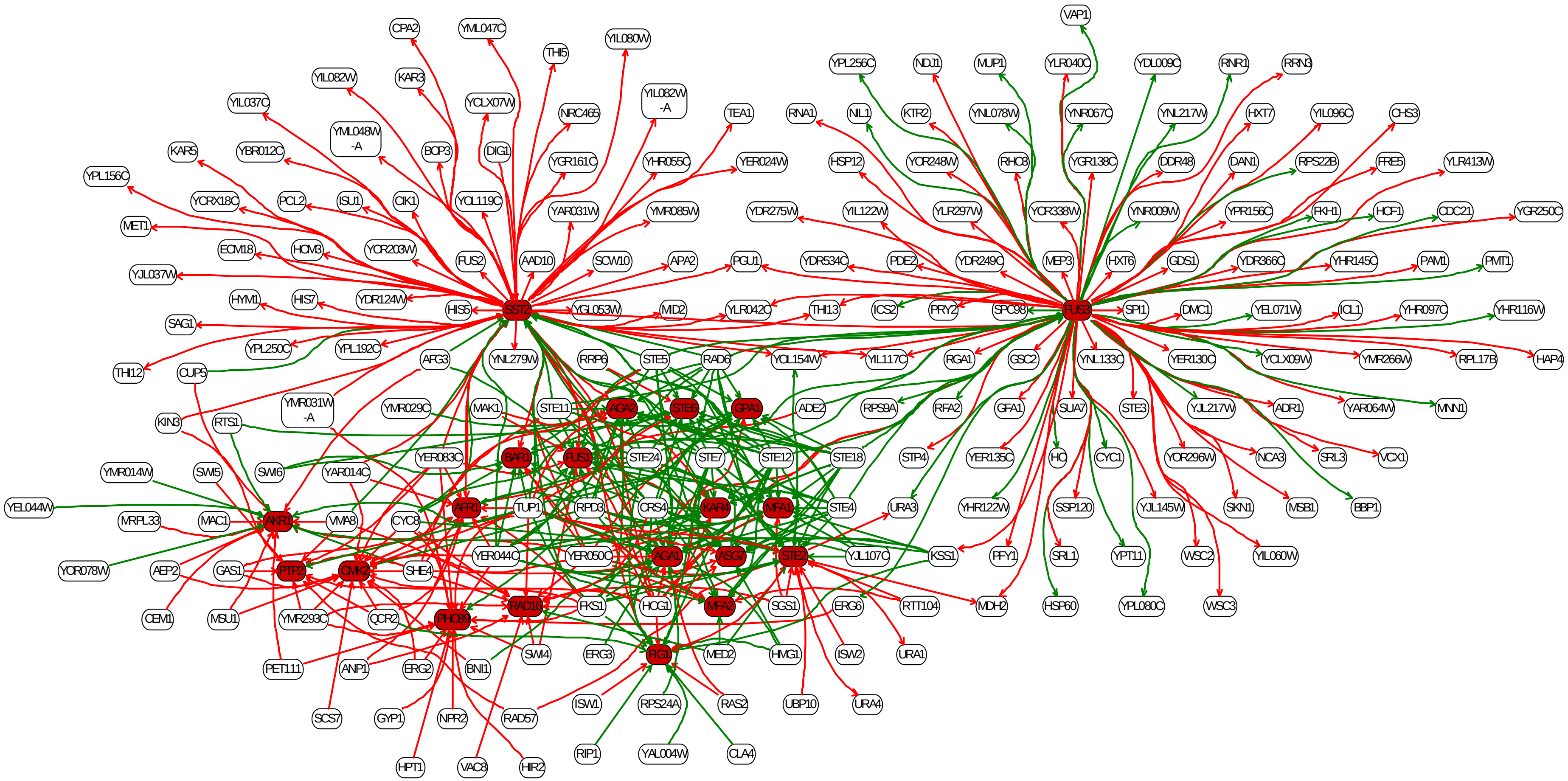
Gene Networks - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …
- **Topology** – a graph describing the connections between the parts
- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)
- **Dynamics** – how does it all work in real time

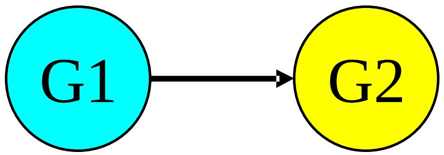
Gene Networks - four levels of hierarchical description

- **Parts list** – genes, transcription factors, promoters, binding sites, …
- **Topology** – a graph describing the connections between the parts
- **Control logics** – how combinations of regulatory signals interact (e.g., promoter logics)
- **Dynamics** – how does it all work in real time

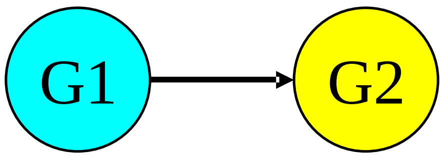
Topology – a graph where nodes represent genes and edges (arcs) represent relationships between genes



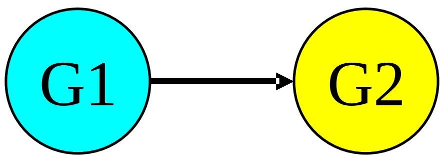
The arcs can have different meaning



- The product of gene G1 is a transcription factor, which binds to the promoter of gene G2 (in Chip-chip experiment) – physical interaction network (direct network)



- Gene G2 contains a binding site for gene G1 (in silico BS identification)



- The disruption of gene G1 changes the expression level of gene G2 – data interpretation network (indirect network)

What kind of things we can study on this level?

- Ideally this graph should tell us which gene can potentially regulate which others and which are independent
- How complex is this graph? What are the connectivity properties? Can we find modules?

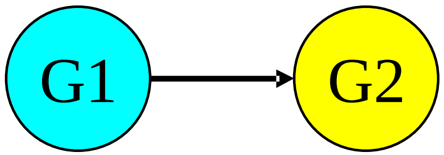
What kind of things we can study on this level

- *Source genes* and *target genes* – nodes with outgoing arcs and nodes with incoming arcs respectively
- For every source gene we can define the set of target genes (target set of a gene)
- How graphs with different edges relate? How do target sets of the same gene compare in different networks?

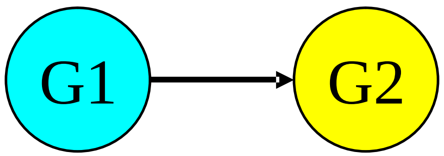
Data for *S. cerevisiae*

- ChIP network (Young lab, Science 2002, Nature 2004) – binding locations for 177 transcription factors (about 4500 locations in the genome) – direct network
- *Mutation* network (Hughes et al, Cell 2000) – 228 yeast mutant expression data for all genes – indirect network

The arcs can have different meaning

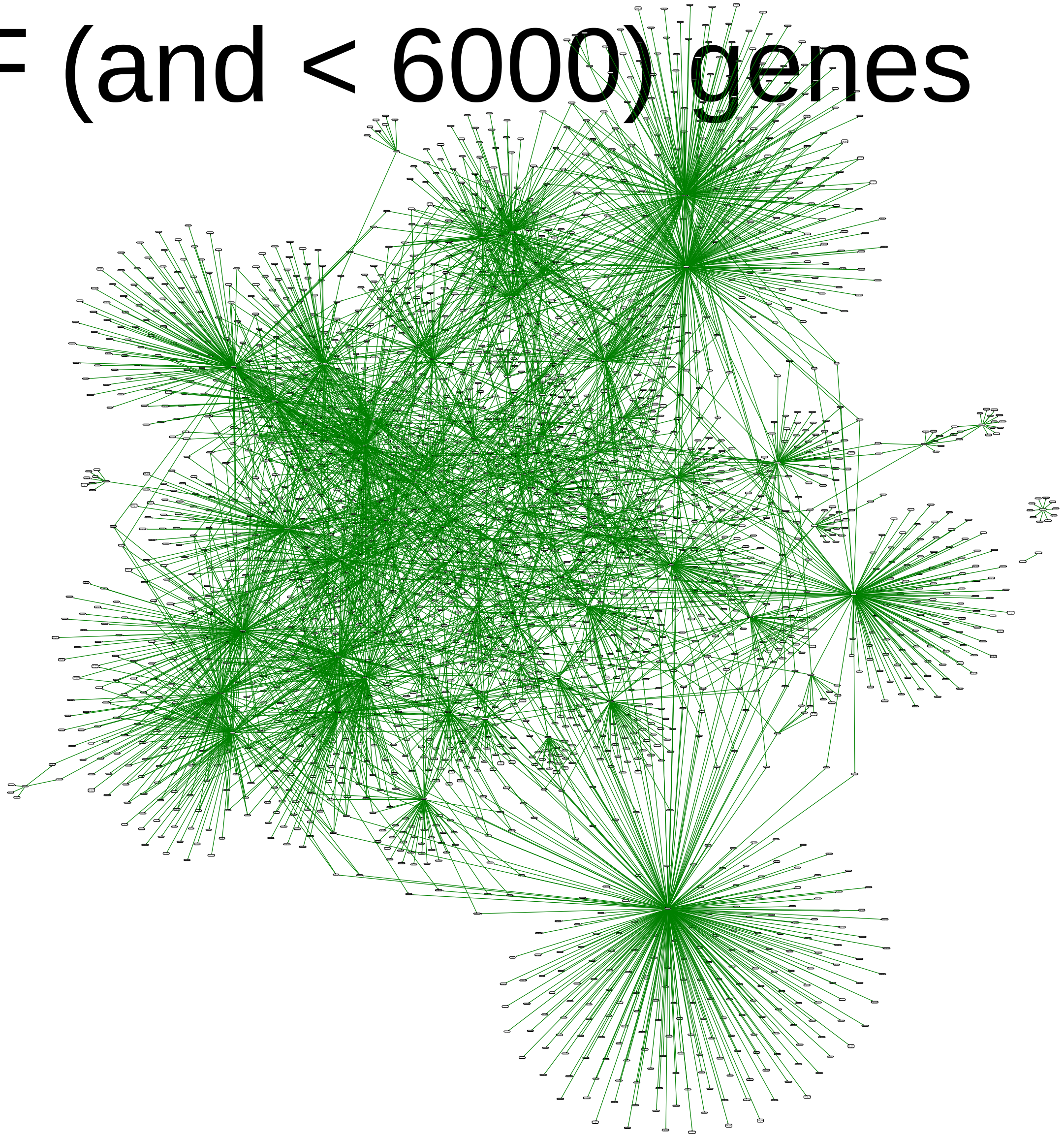


- The product of gene G1 is a transcription factor, which binds to the promoter of gene G2 (in Chip-chip experiment) – physical interaction network (direct network)



- Gene G2 contains a binding site for gene G1 (in silico BS identification)

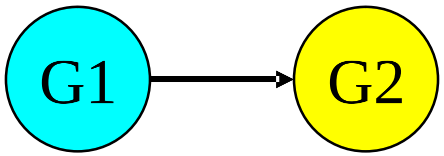
Transcription factor binding network in *S. cerevisiae* for ~100 TF (and < 6000) genes



Data for *S. cerevisiae*

- ChIP network (Young lab, Science 2002, Nature 2004) – binding locations for 177 transcription factors (about 4500 locations in the genome) – direct network
- The presence of derived transcription factor binding site data used to improve the networks

The arcs can have different meaning

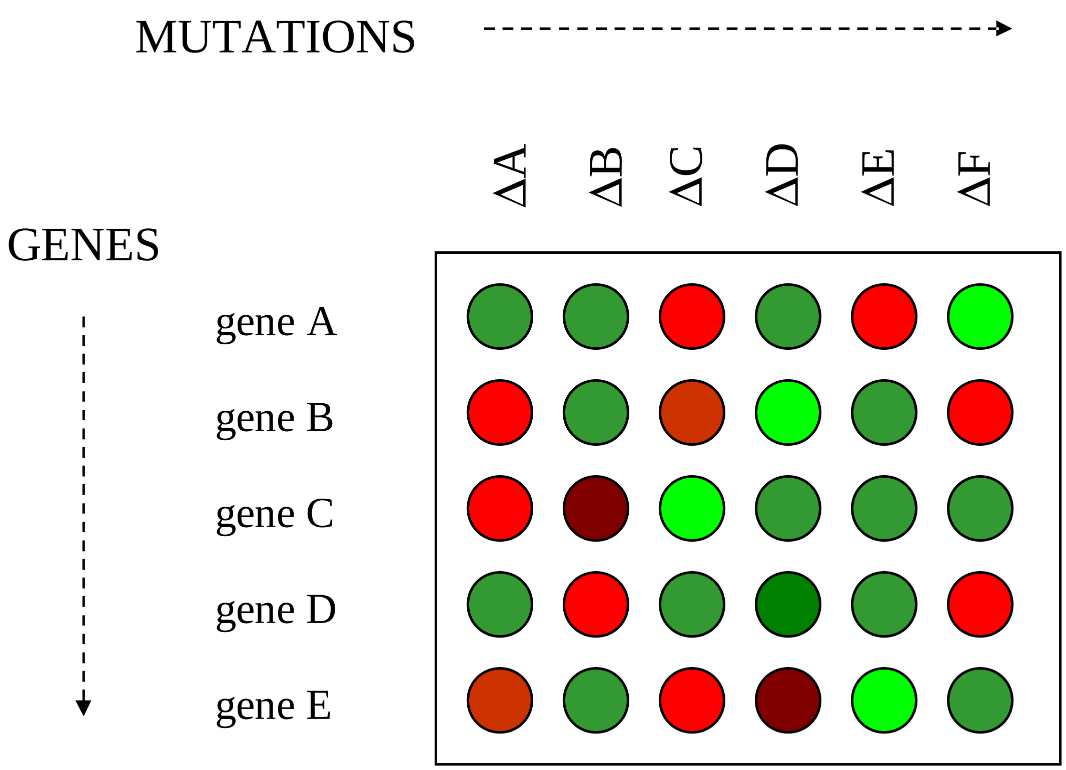


- The disruption of gene G1 changes the expression level of gene G2 – data interpretation network (indirect network)

Data for *S. cerevisiae*

- *Mutation* network (Hughes et al, Cell 2000) – 228 yeast mutant expression data for all genes – indirect network

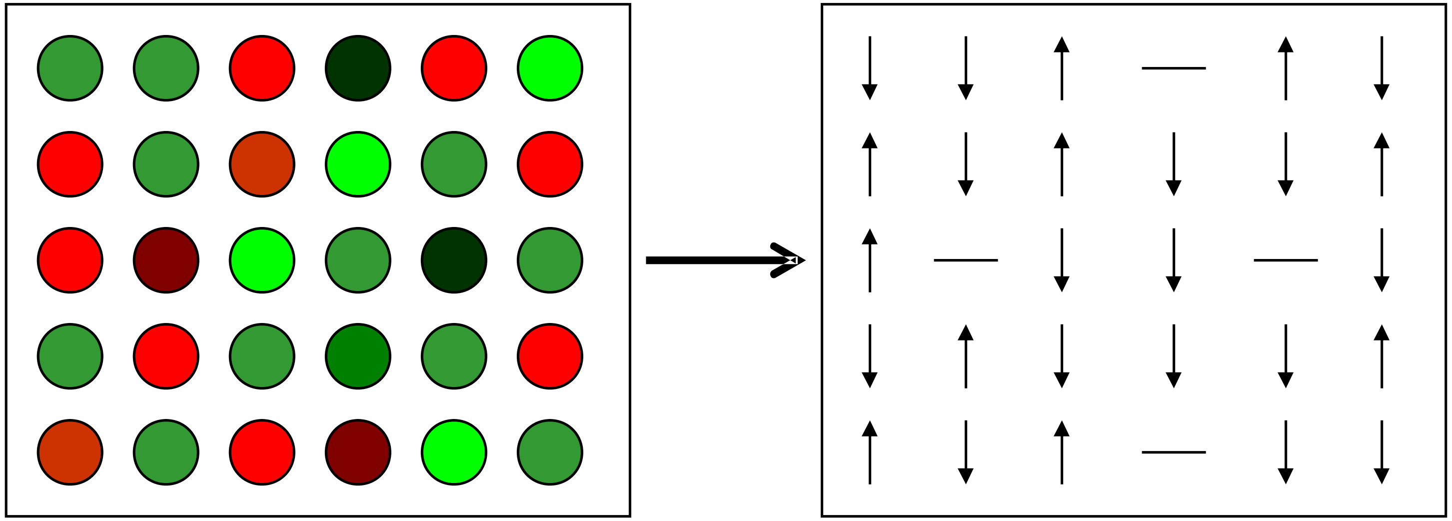
The mutation microarray data matrix



Discretization of the data

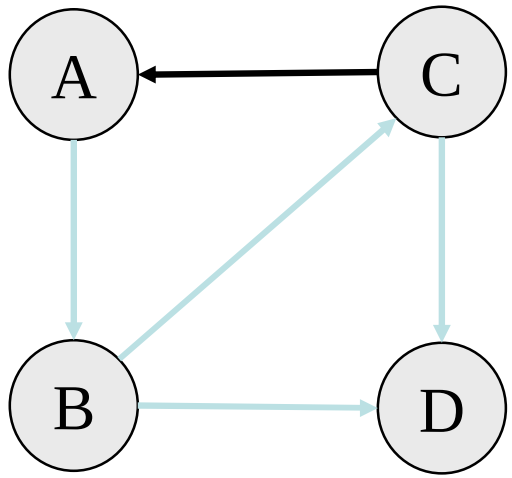
The normalized expression log(ratios) are discretized using two thresholds:

$X \leq C \Rightarrow X' = -1$ $\quad C \leq X \leq C \Rightarrow X' = 0$ $\quad X \geq C \Rightarrow X' = 1$



Gene disruption networks

	ΔA	ΔB	ΔC
gene A	-1	0	-1
gene B	1	-1	0
gene C	0	1	-1
gene D	0	1	1



Yeast mutation data

- Gene expression data for all ~6000 genes for ~300 systematic mutation experiments in yeast published by Rosetta (Gene Expression Compendium, Cell, 2000)
- Additionally ~60 replicates for the wild-type, and an error model, together allowing to discretize the data

Dataset

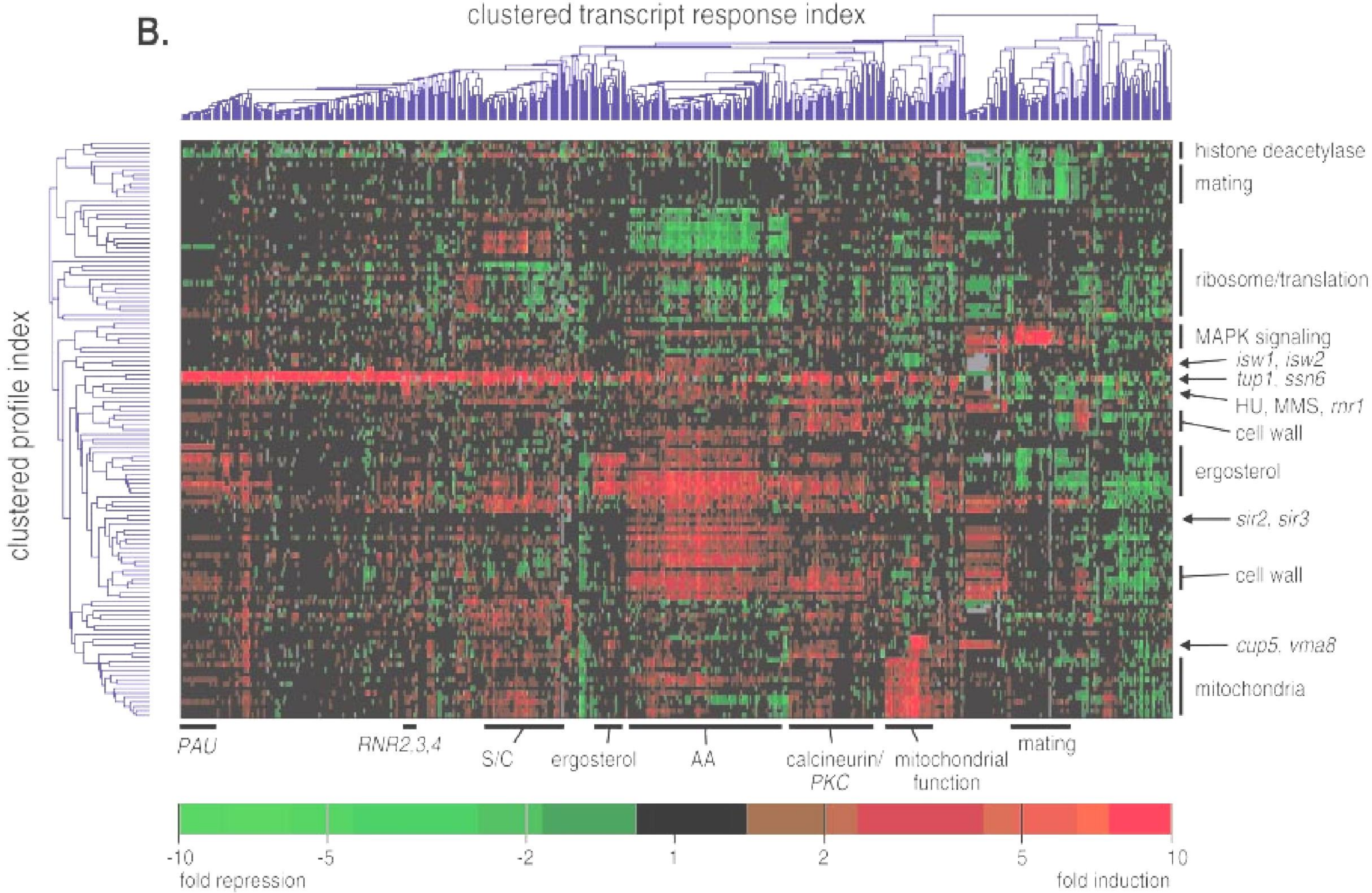
The dataset used is coming from Hughes *et al.*:
"Functional discovery via a compendium of expression profiles", Cell 102, 109-126 (2000)

- Yeast data, 6316 gene expression profiles over 300 experiments
- 276 deletion mutants (274 single, 2 double)
- 11 tet-promotor mutants
- 13 compound treatments

We have selected a subset of 207 experiments:

- Single deletion mutants
- Diploid cells only
- All chromosomes present

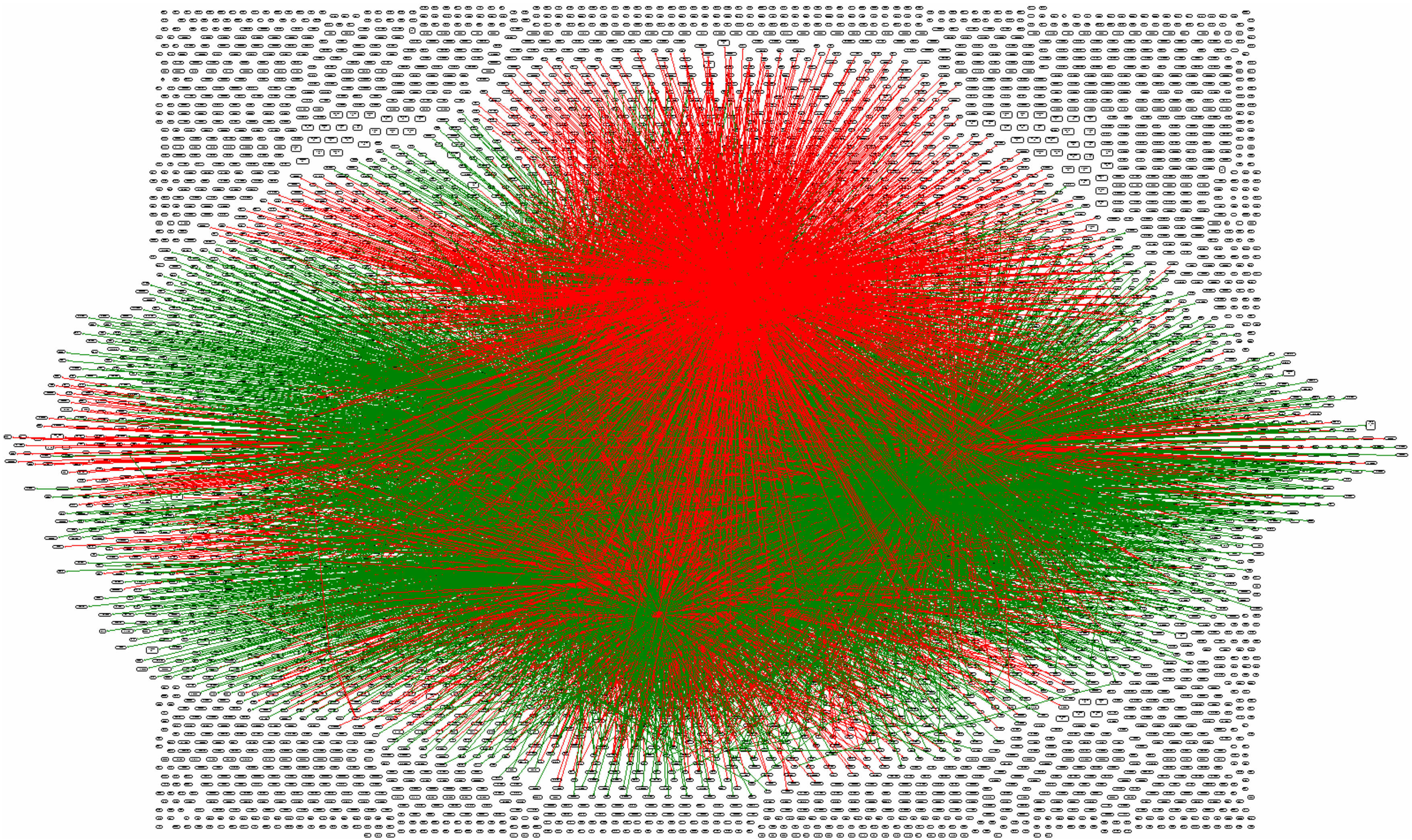
B.


clustered transcript response index
clustered profile index
histone deacetylase
mating
ribosome/translation
MAPK signaling
isw1, isw2
tup1, ssn6
HU, MMS, rnr1
cell wall
ergosterol
sir2, sir3
cell wall
cup5, vma8
mitochondria
PAU
RNR2,3,4
S/C
ergosterol
AA
calcineurin/
PKC
mitochondrial
function
mating
-10
-5
-2
1
2
5
10
fold repression
fold induction

Gene specific thresholds

- ~ 80 experiments on ‘wild-type’ yeast were performed, revealing wide variation in gene expression dependent on particular gene
- Gene expression variation for each gene can be assumed to have normal distribution
- Standard deviation for each gene can be used therefore for assessing the threshold on gene by gene basis

Mutation network for S. Cerevisiae



Why topology is important?

- Reduce hypothesis space when analysing next layers of model complexity – instead of default – all genes depend on all, topology tells us which genes are independent
- What is the complexity of gene regulation
  - Given a transcription factor T – how many genes does T regulate?
  - Given a gene A, how many transcription factors regulate A?
- Are networks modular?