Bioinformatics 3

V7 – Gene Regulation

Mon., Nov 16, 2015

Turn, Turn, Turn…

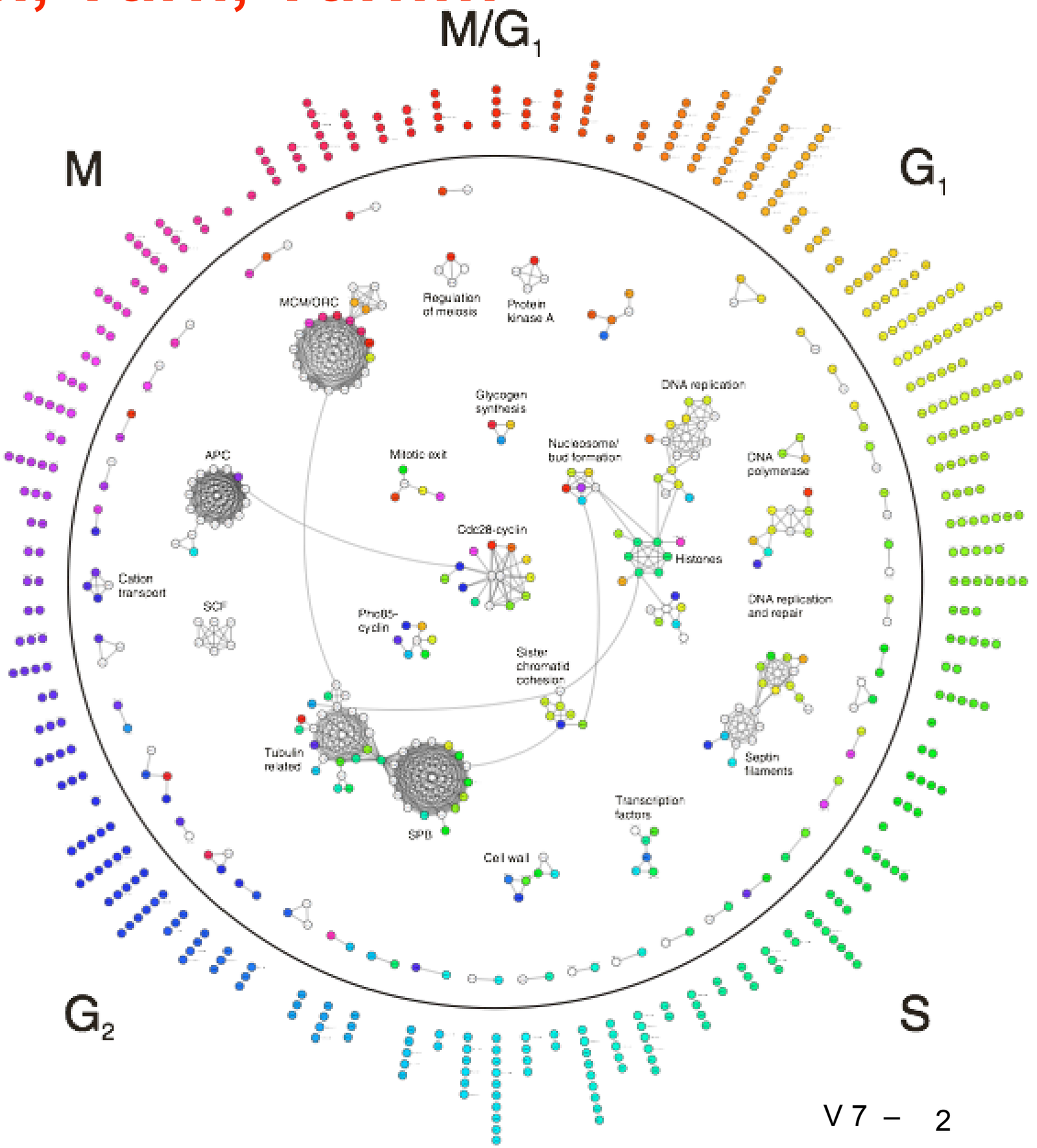
From Lichtenberg et al,
Science 307 (2005) 724:

The wheel represents the 4 stages of a cell cycle in yeast.

Colored proteins are components of protein complexes that are (only) expressed at certain stages.

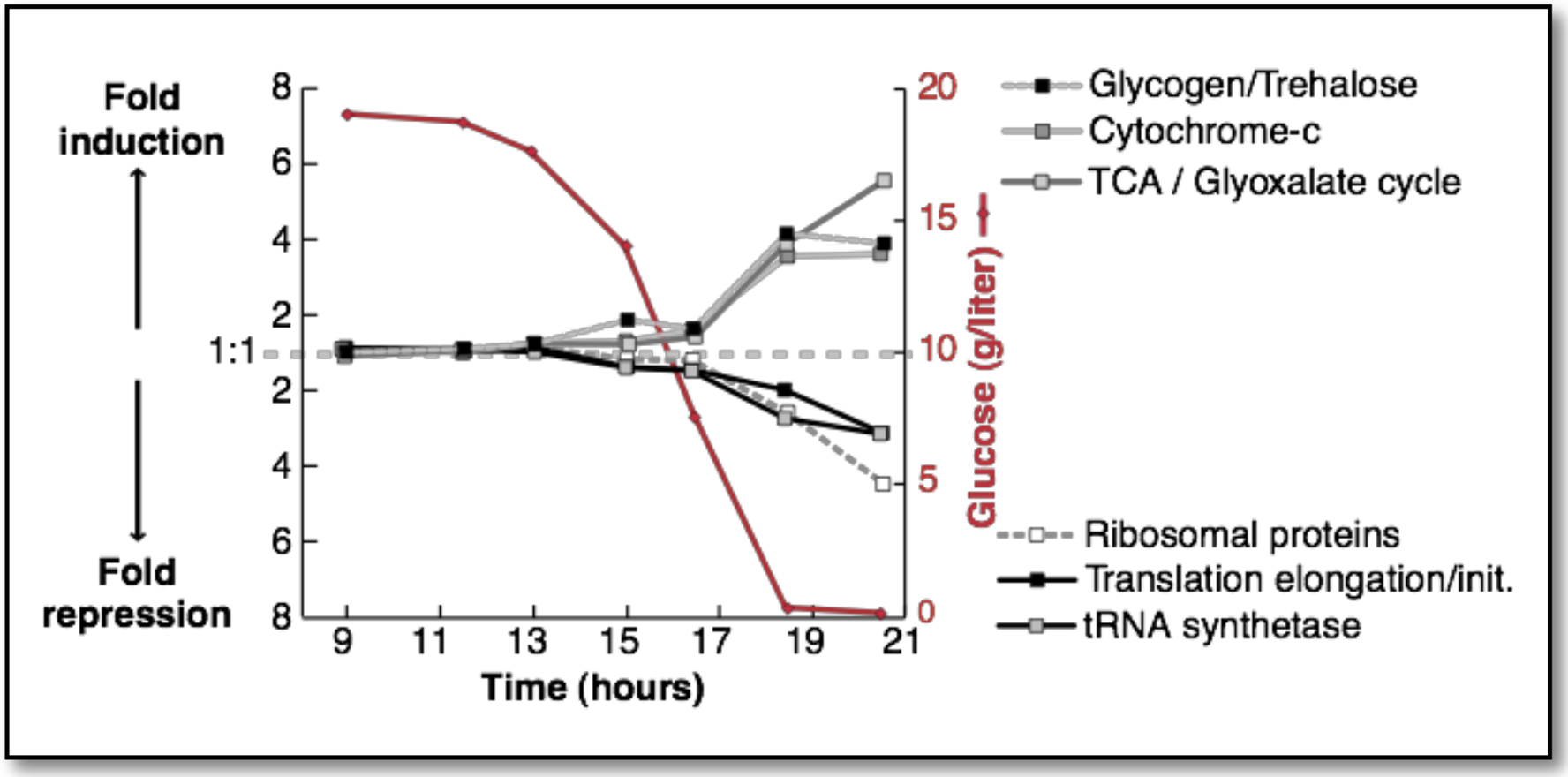
Other parts of these complexes have constant expression rates (white).

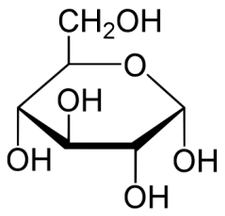
→ “assembly in time”

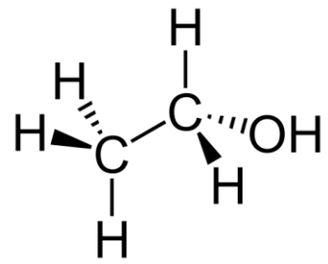


External Triggers affect transcriptome

Re-routing of metabolic fluxes during the diauxic shift in *S. cerevisiae*
→ changes in protein abundances (measured via mRNA levels)







anaerobic fermentation:
fast growth on glucose → ethanol

→ **Diauxic shift**

aerobic respiration:
ethanol as carbon source

DeRisi et al., *Science* **278** (1997) 680

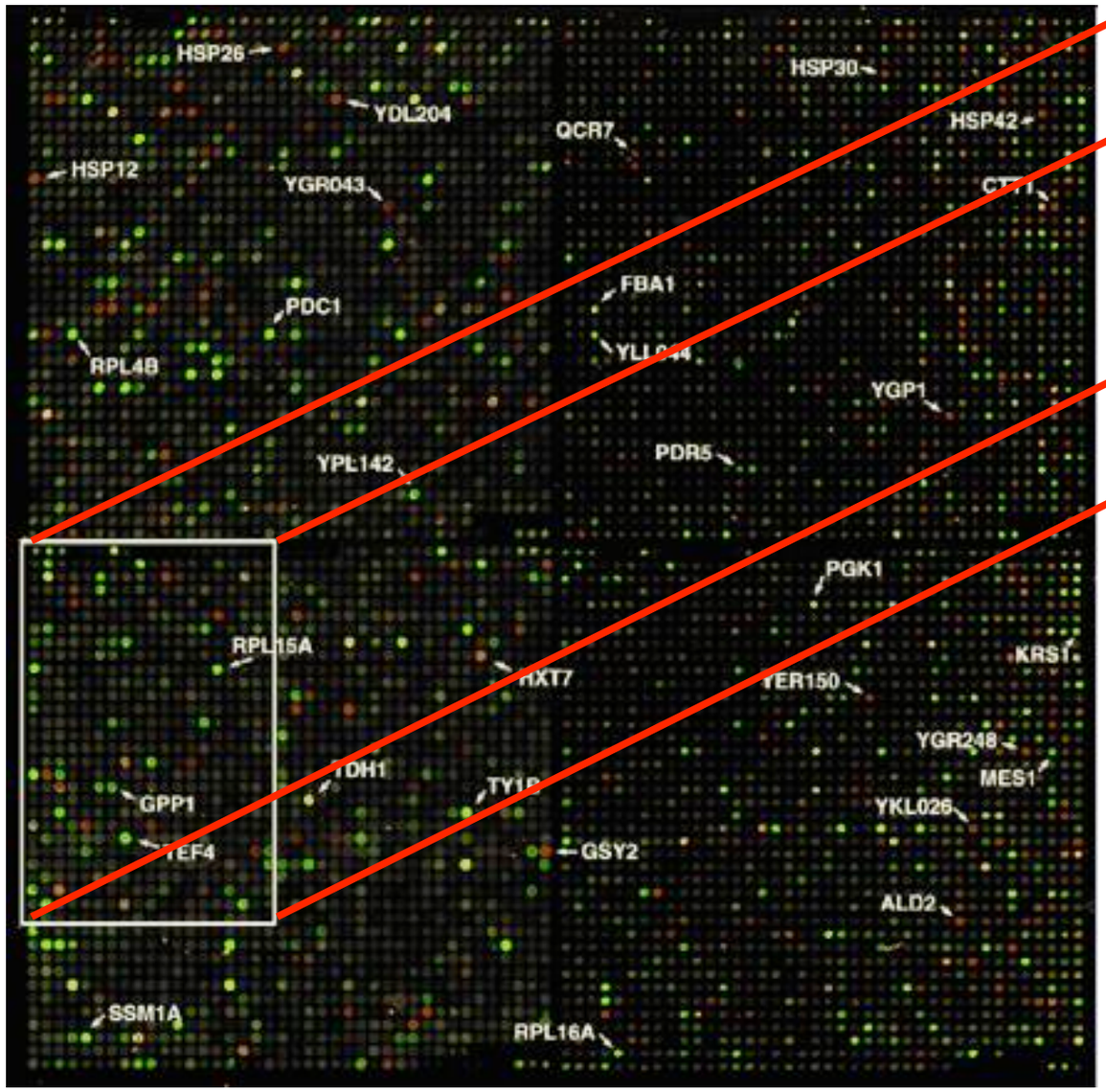
Diauxic shift affects hundreds of genes

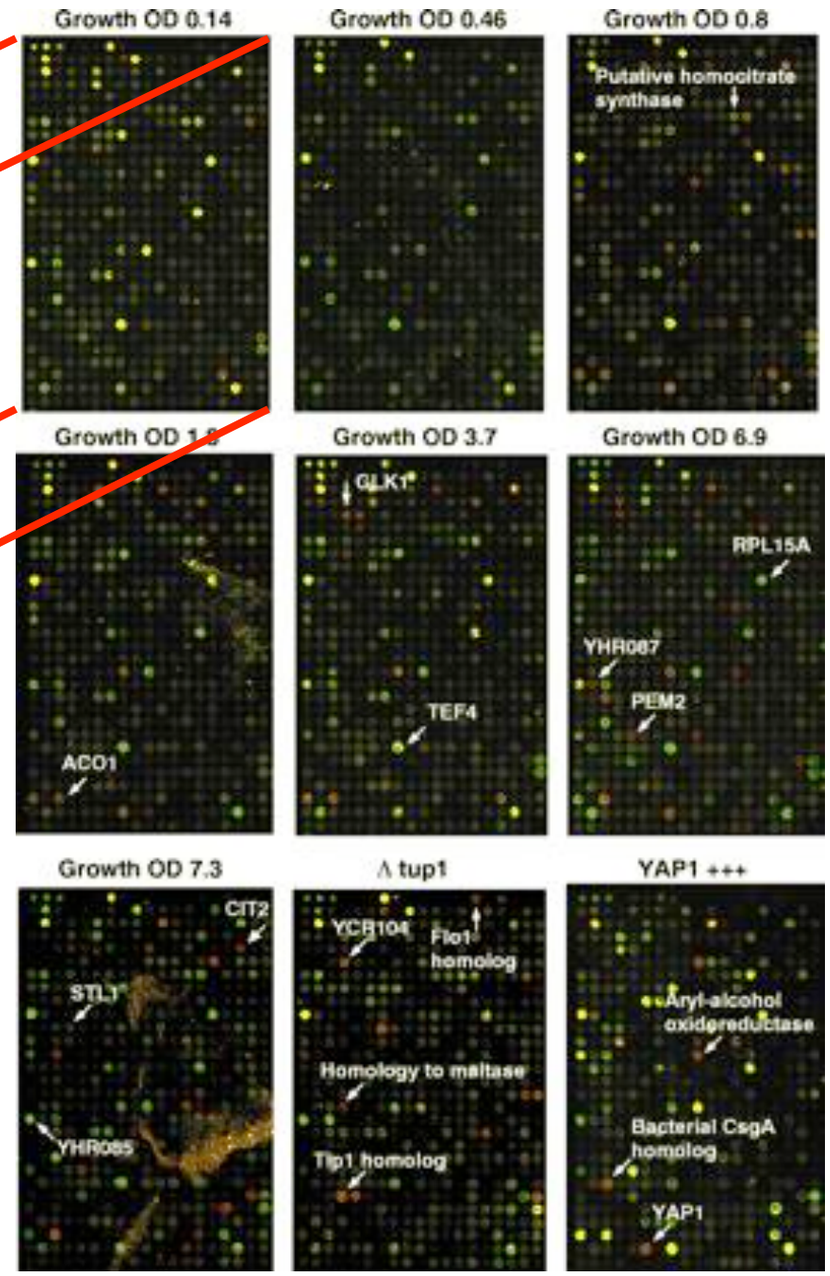
Cy3/Cy5 labels (these are 2 dye molecules for the 2-color microarray), comparison of 2 probes at 9.5 hours distance; w and w/o glucose

Red: genes induced by diauxic shift (710 genes 2-fold)

Green: genes repressed by diauxic shift (1030 genes change > 2-fold)

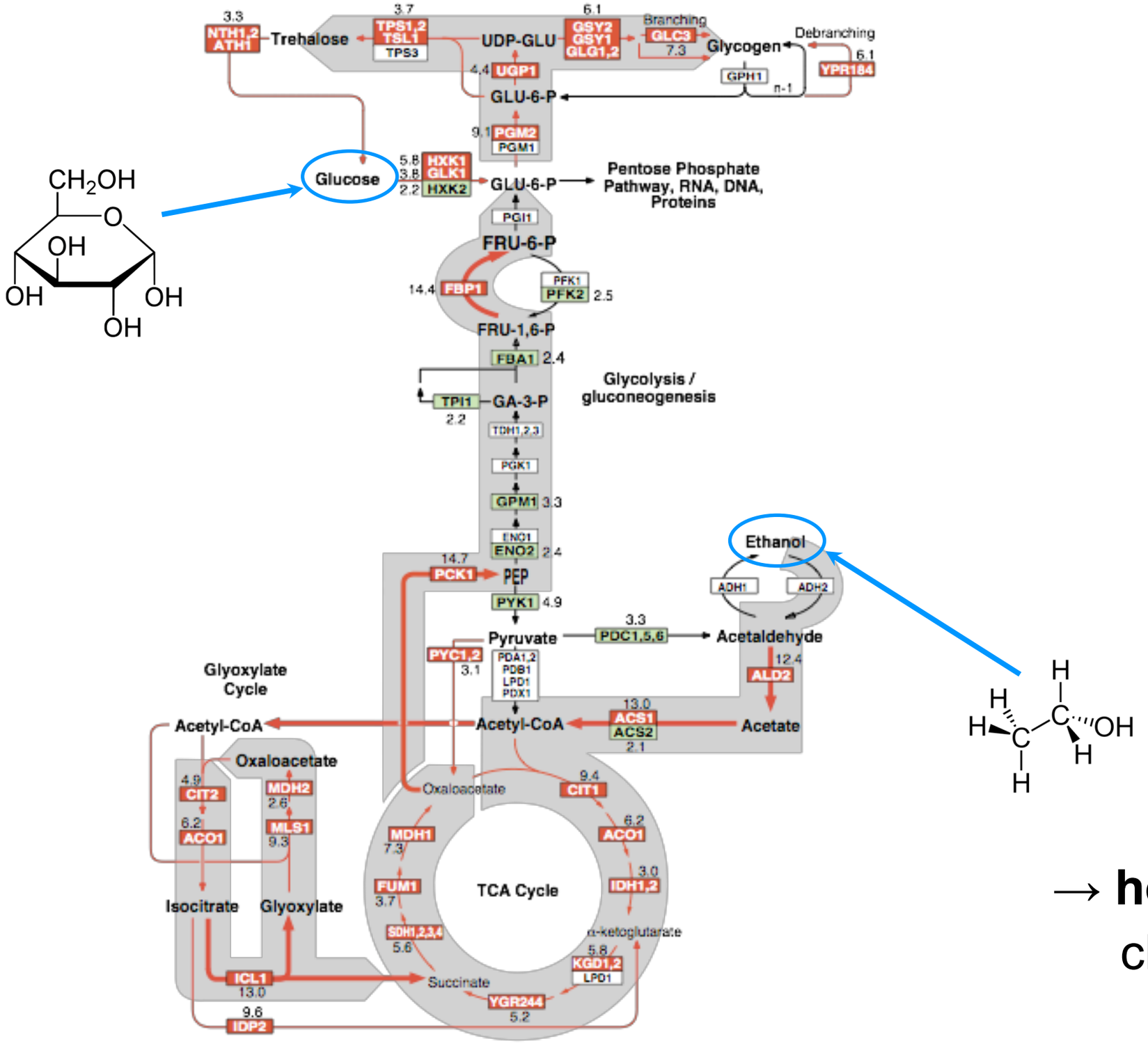
Optical density (OD) illustrates cell growth;





DeRisi et al., *Science* **278** (1997) 680

Flux Re-Routing during diauxic shift



fold change

expression increases

expression unchanged

expression diminishes

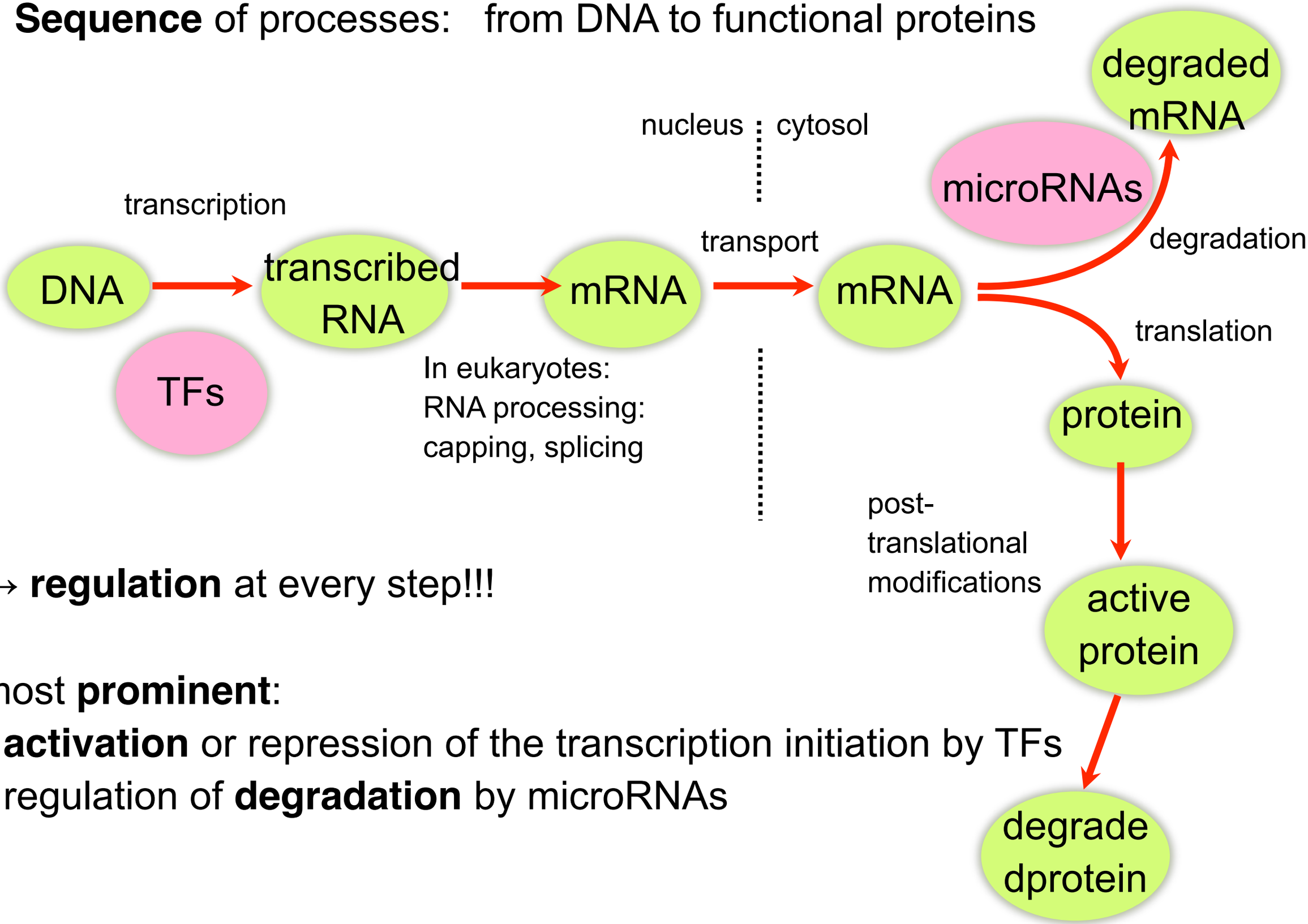
metabolic flux increases

→ **how** are these changes **coordinated?**

DeRisi et al., *Science* **278** (1997) 680

Gene Expression

Sequence of processes: from DNA to functional proteins



→ **regulation** at every step!!!

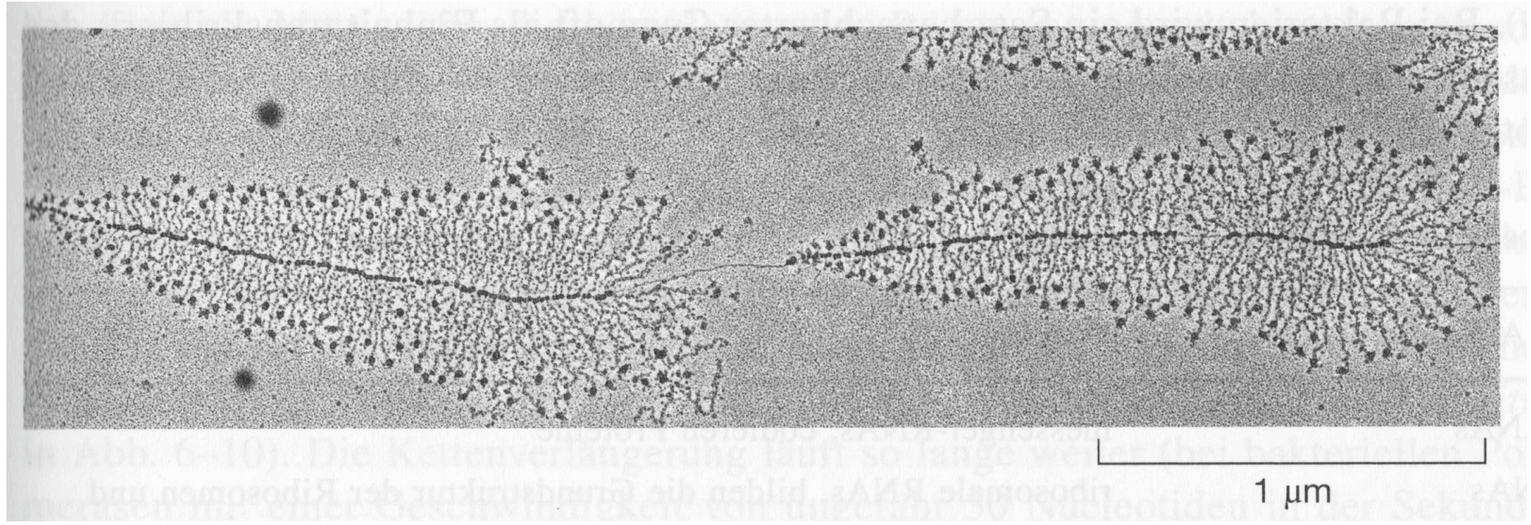
most **prominent**:

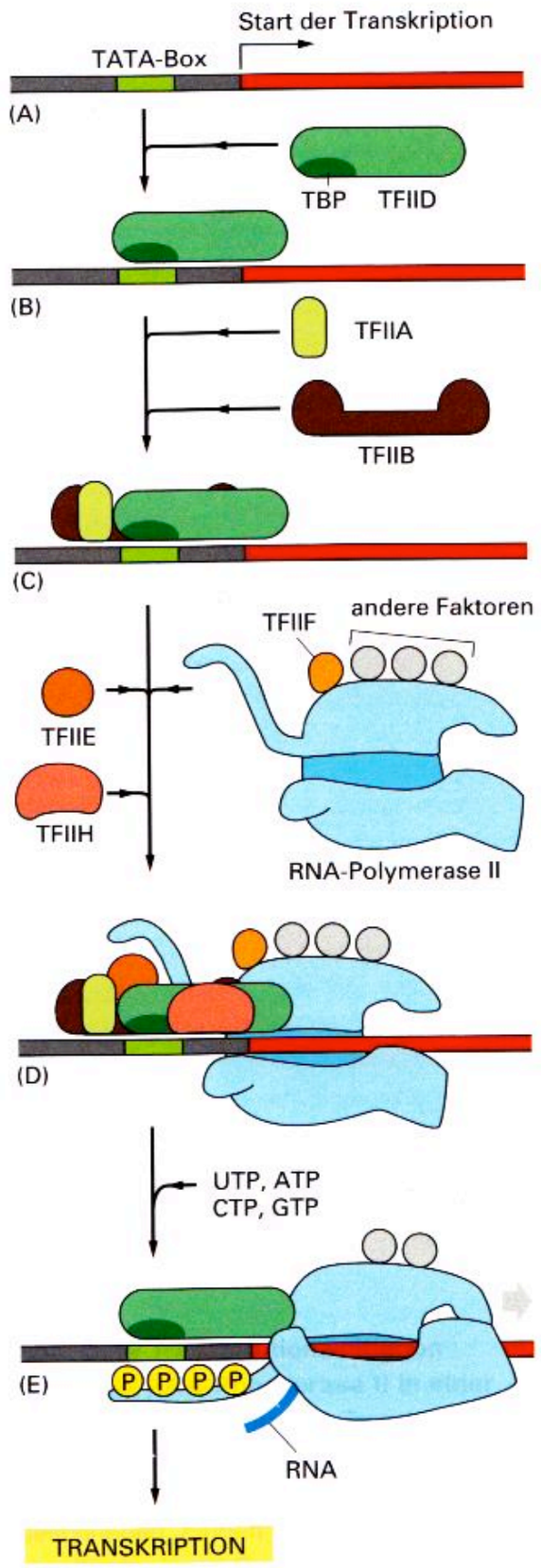
- **activation** or repression of the transcription initiation by TFs
- regulation of **degradation** by microRNAs

Transcription Initiation

In eukaryotes:

- several **general** transcription factors **have** to bind
- **specific** enhancers or repressors **may** bind
- then the RNA polymerase binds
- and starts transcription





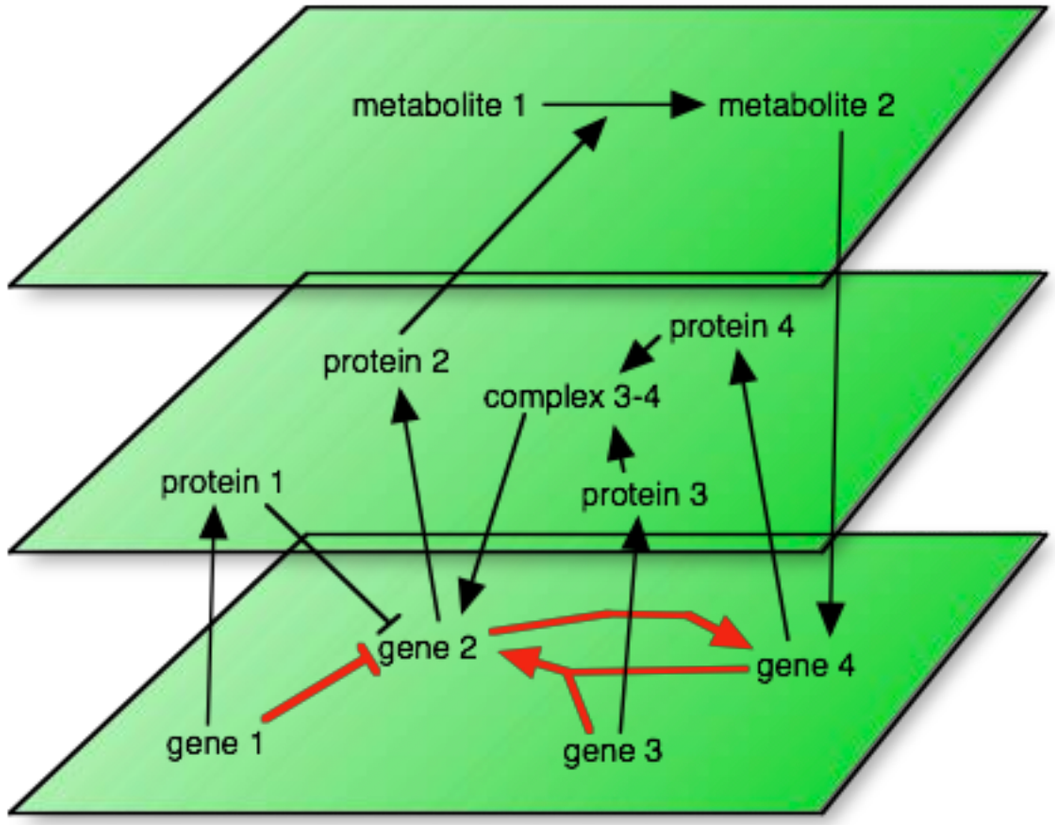
Alberts et al.
"Molekularbiologie der Zelle", 4. Aufl.

Layers upon Layers

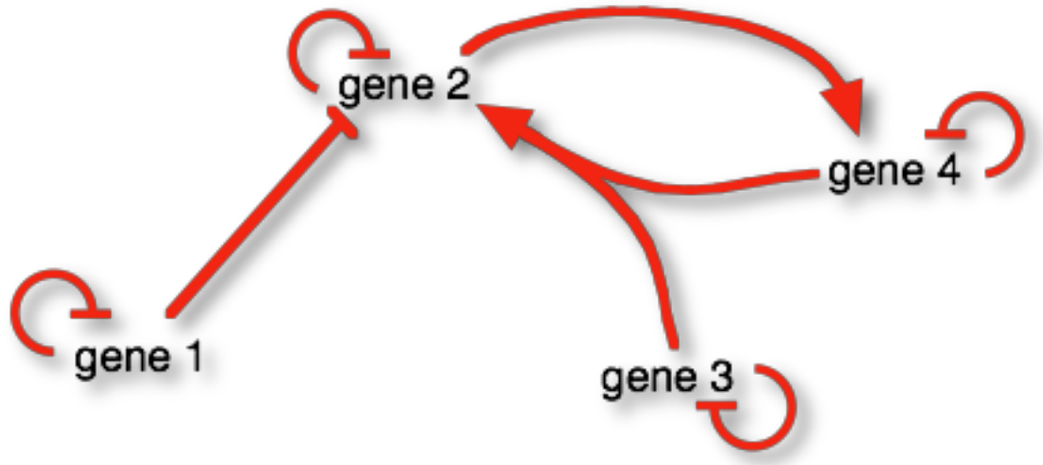
Biological regulation
via proteins and metabolites

<=>

Projected regulatory network



<=>



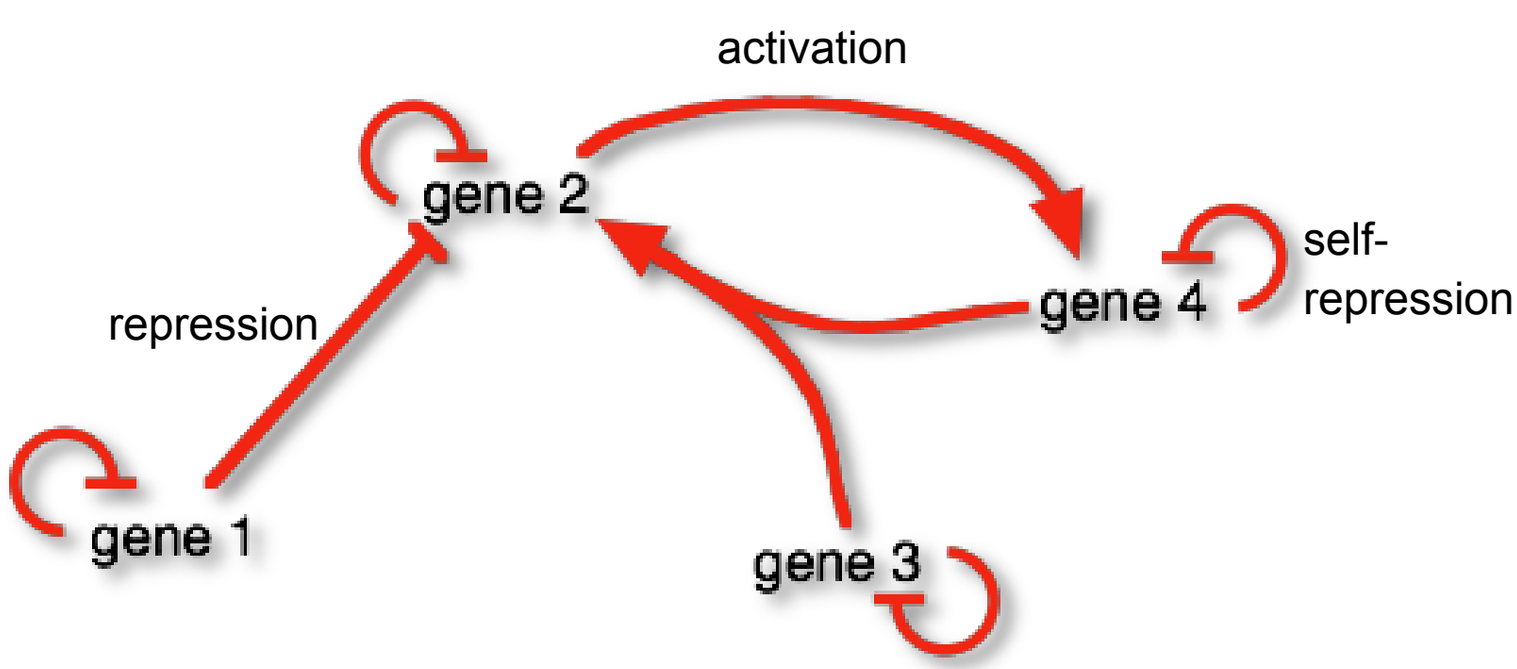
Remember:
genes do not interact directly

Conventions for GRN Graphs

Nodes: genes that code for proteins which catalyze products …
→ everything is projected onto respective gene

Gene regulation networks have "cause and action"
→ **directed** networks

A gene can enhance or suppress the expression of another gene
→ **two types** of arrows



What is a GRN?

Gene regulatory networks (GRN) are model representations of how genes regulate the expression levels of each other.

In transcriptional regulation, proteins called **transcription factors (TFs)** regulate the transcription of their **target genes** to produce messenger RNA (mRNA),
whereas in **post-transcriptional regulation microRNAs** (miRNAs) cause degradation and repression of target mRNAs.

These interactions are represented in a GRN by adding edges linking TF or miRNA genes to their target mRNAs.

Narang et al. (2015). PLoS Comput Biol 11(9): e1004504

What is a GRN?

Since these physical interactions are fixed,
we can represent a GRN as a **static network**
even though regulatory interactions occur
dynamically in space and time.

A GRN provides a systemic view of gene regulation
by coordinated activity of multiple TFs and miRNAs
and thus serves as a medium for
understanding the mechanism of gene regulation.

Narang et al. (2015). PLoS Comput Biol 11(9): e1004504

Which TF binds where?

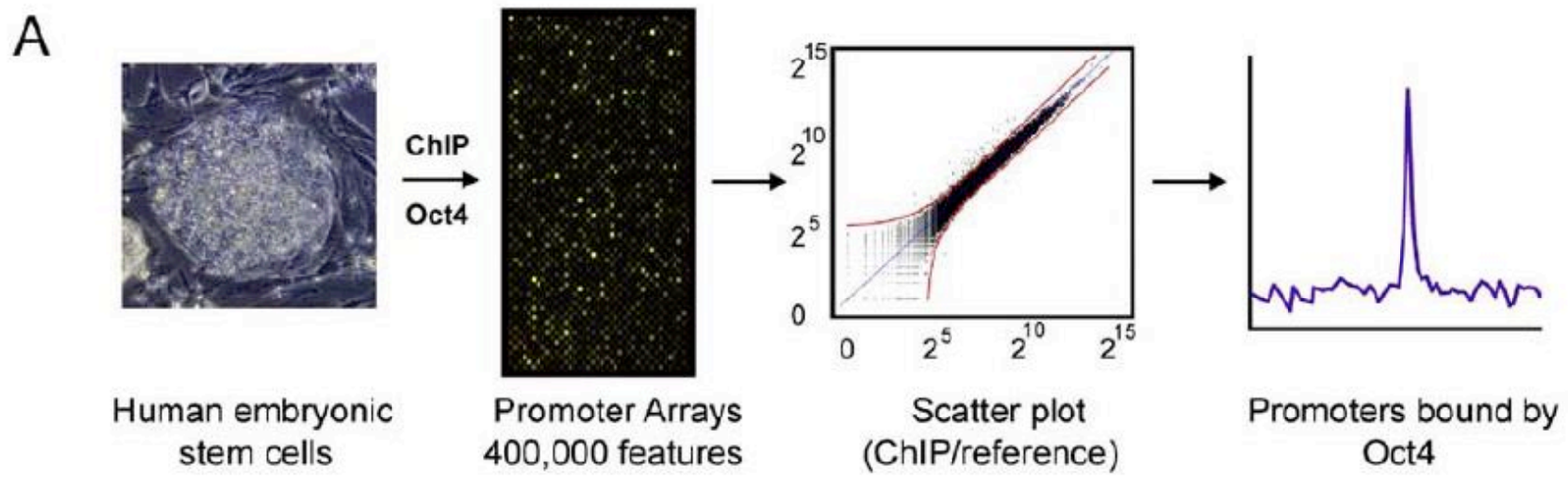
Chromatin immuno precipitation: use e.g. antibody against Oct4

➔ ”fish“ all DNA fragments that bind Oct4

➔ sequence DNA fragments bound to Oct4

➔ align them + extract characteristic sequence features

➔ Oct4 binding motif



Boyer et al. Cell 122, 947 (2005)

Sequence logos represent binding motifs

A **logo** represents each column of the alignment by a stack of letters, with the height of each letter proportional to the **observed frequency** of the corresponding amino acid or nucleotide, and the overall height of each stack proportional to the sequence conservation, measured in bits, at that position.

Sequence conservation is defined as difference between the maximum possible entropy and the entropy of the observed symbol distribution:

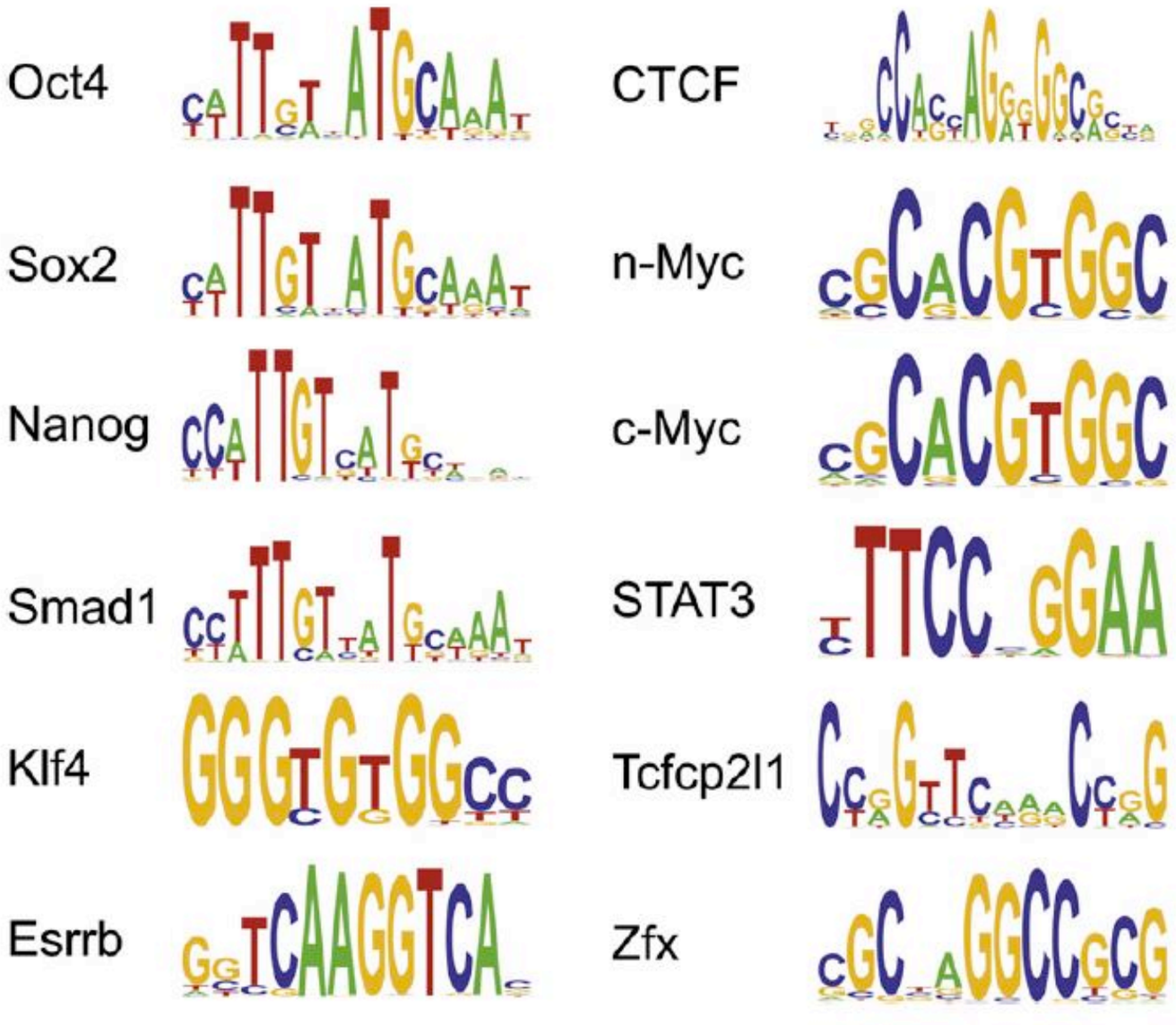
$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

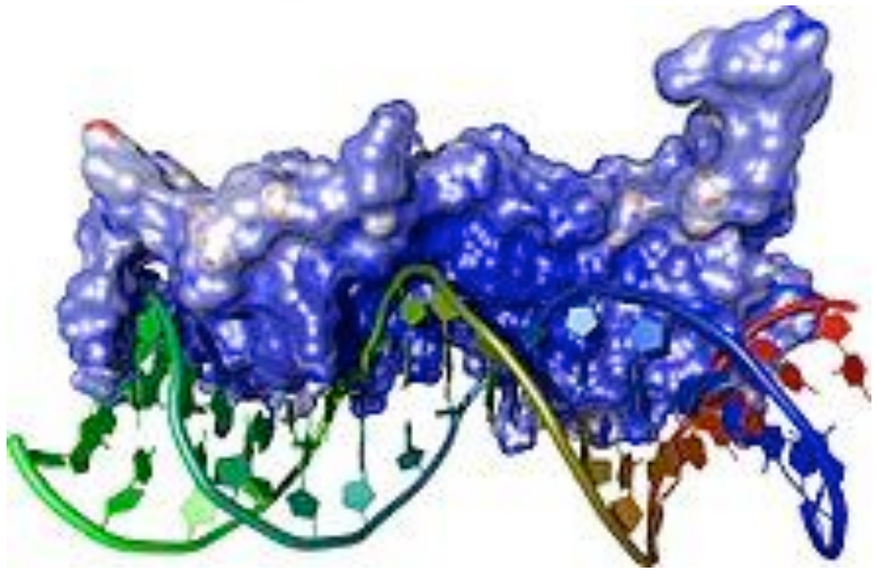
p_n : observed frequency of symbol *n* at a particular sequence position

N : number of distinct symbols for the given sequence type, either 4 for DNA/RNA or 20 for protein.

Crooks et al., Genome Research
14:1188–1190 (2004)

Construct preferred binding motifs





DNA-binding domain of a glucocorticoid - receptor from *Rattus norvegicus* with the matching DNA fragment ; www.wikipedia.de

Chen et al., Cell 133, 1106-1117 (2008)

Position specific weight matrix

Build list of genes that share a TF binding motif.

Generate multiple sequence alignment of their sequences.

Alignment matrix: how often does each letter occur

at each position in the alignment?

a) **Alignment Matrix**

```
A A T T G A
A G G T C C
A G G A T G
A G G C G T
```

	1	2	3	4	5	6
A	4	1	0	1	0	1
C	0	0	0	1	1	1
G	0	3	3	0	2	1
T	0	0	1	2	1	1

consensus: A G G T G N

$$\ln \frac{(n_{i,j} + p_i) / (N + 1)}{p_i} \approx \ln \frac{f_{i,j}}{p_i}$$

b) **Weight Matrix**

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	.96	.96	-1.6	.59	0
T	-1.6	-1.6	0	.59	0	0

test sequence: A G G T G C

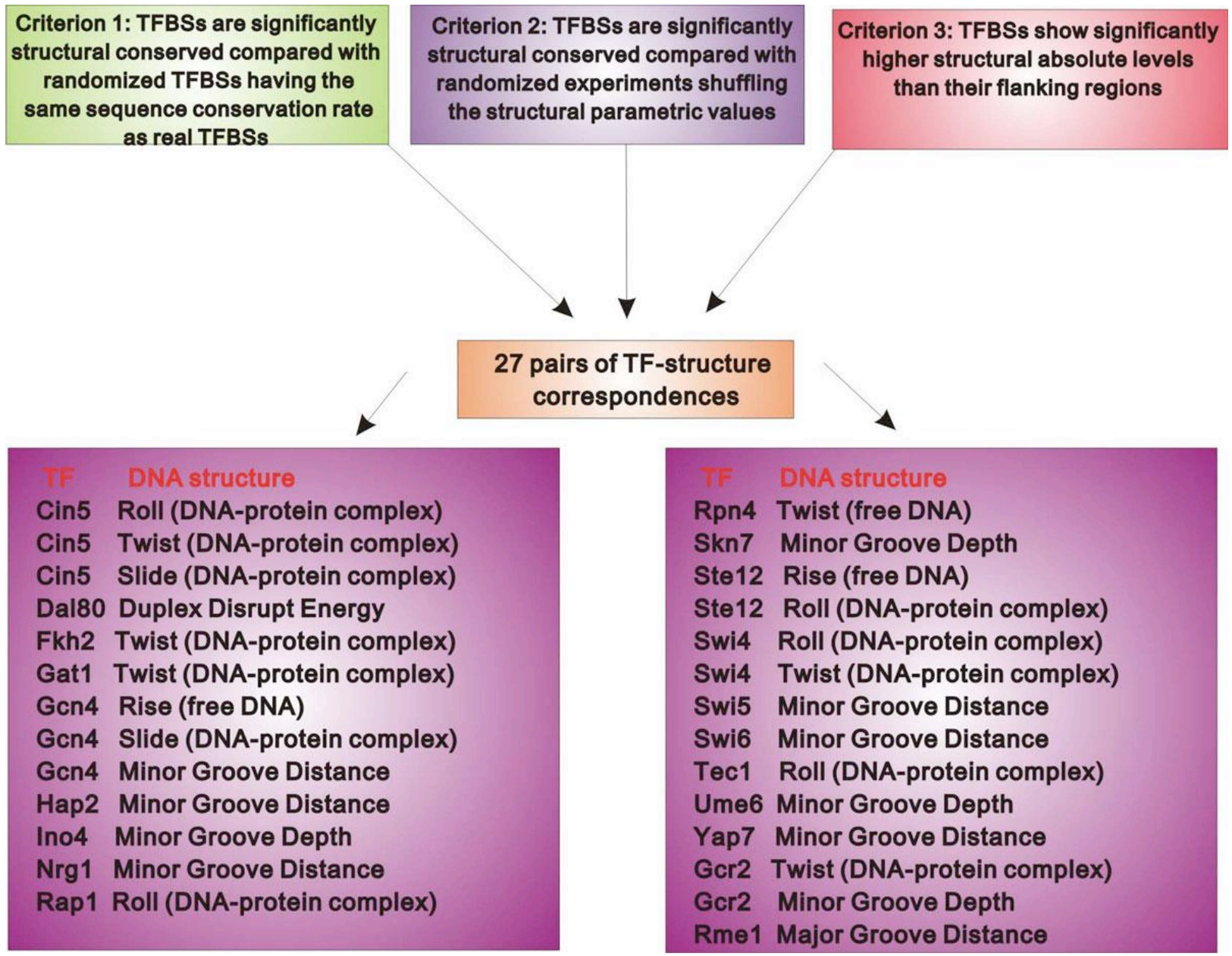
Fig. 1. Examples of the simple matrix model for summarizing a DNA alignment. (**a**) An alignment matrix describing the alignment of the four 6-mers on top. The matrix contains the number of times, $n_{i,j}$, that letter i is observed at position j of this alignment. Below the matrix is the consensus sequence corresponding to the alignment (N indicates that there is no nucleotide preference). (**b**) A weight matrix derived from the alignment in (a). The formula used for transforming the alignment matrix to a weight matrix is shown above the arrow. In this formula, N is the total number of sequences (four in this example), p_i is the *a priori* probability of letter i (0.25 for all the bases in this example) and $f_{i,j} = n_{i,j}/N$ is the frequency of letter i at position j. The numbers enclosed in blocks are summed to give the overall score of the test sequence. The overall score is 4.3, which is also the maximum possible score with this weight matrix.

Hertz, Stormo (1999) Bioinformatics 15, 563

What do TFs recognize?

(1) Amino acids of the TFs make specific contacts (e.g. hydrogen bonds) with DNA base pairs

(2) DNA conformation depends on its sequence
→ Some TFs „measure“ different aspects of the DNA conformation



Dai et al. *BMC Genomics* 2015, **16**(Suppl 3):S8

E. coli Regulatory Network

BMC Bioinformatics

BioMed Central

Research article

Open Access

Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach

Hong-Wu Ma[1], Jan Buer[2,3] and An-Ping Zeng*[1]

Address: [1]Department of Genome Analysis, GBF – German Research Center for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany, [2]Department of Mucosal Immunity, GBF – German Research Center for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany and [3]Medical Microbiology and Hospital Hygiene, Medical School Hannover, Carl-Neuberg-Str. 1, 30625 Hannover, Germany

Email: Hong-Wu Ma - hwm@gbf.de; Jan Buer - jab@gbf.de; An-Ping Zeng* - aze@gbf.de

* Corresponding author

Published: 16 December 2004

BMC Bioinformatics 2004, **5**:199 doi:10.1186/1471-2105-5-199

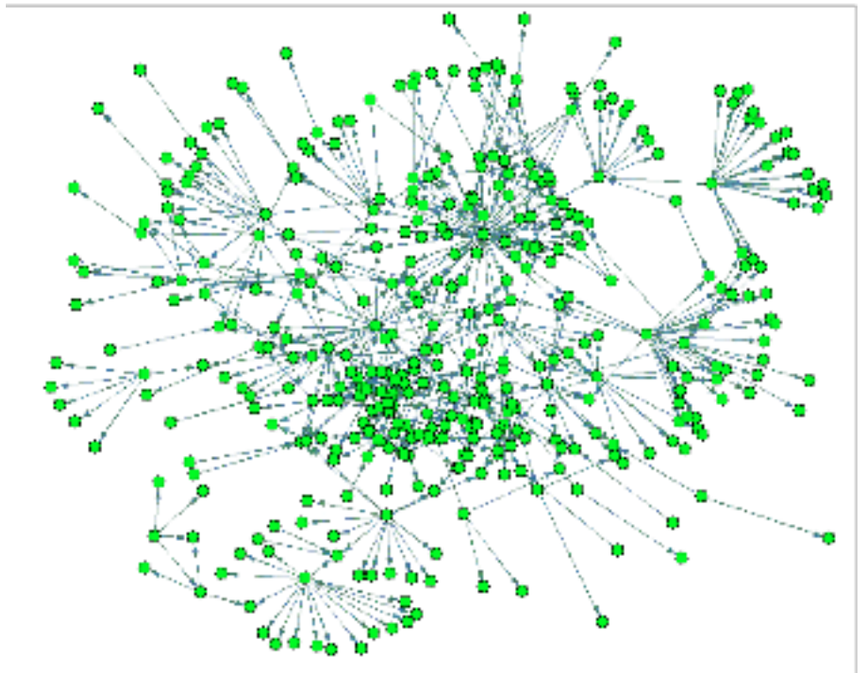
Received: 28 July 2004
Accepted: 16 December 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/199

BMC Bioinformatics **5** (2004) 199

Simple organisms have hierarchical GRNs

Largest weakly connected component
(ignore directions of regulation)
: 325 operons
(3/4 of the complete network)

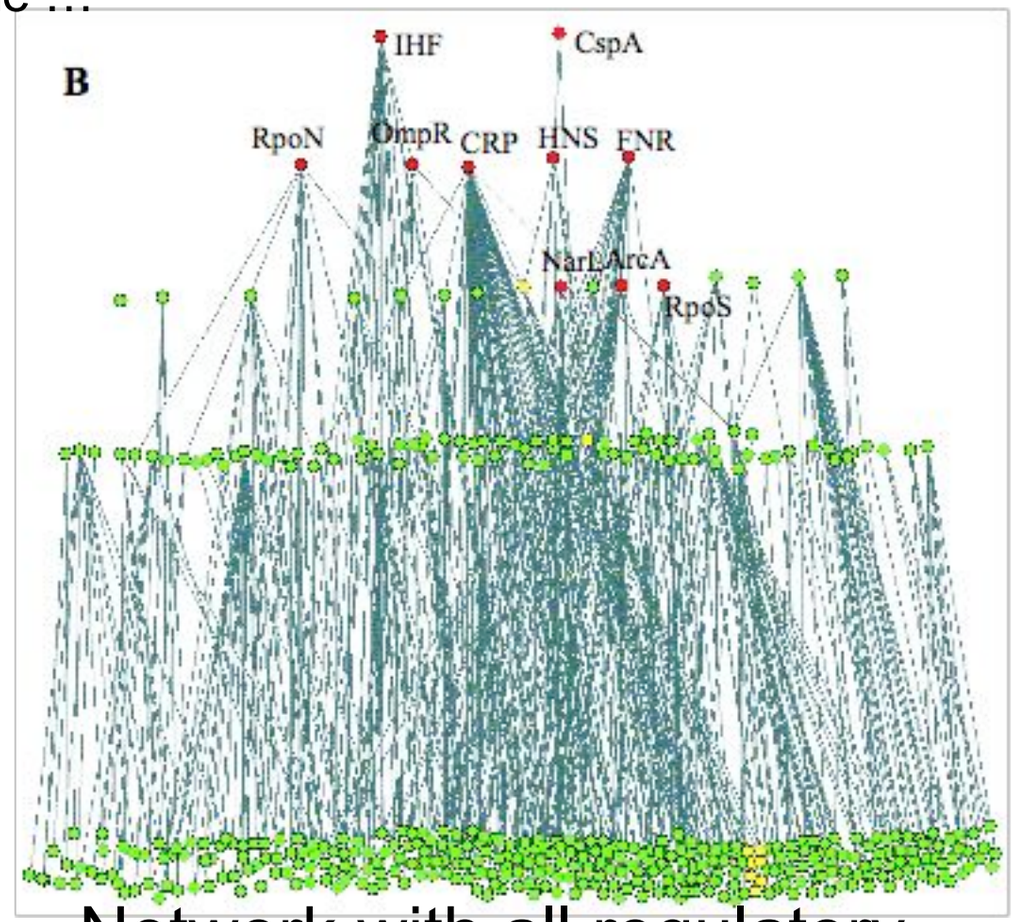


Network from standard layout algorithm

→

Lowest level: operons that code for TFs with only auto-regulation, or no TFs

Next layer: delete nodes of lower layer, identify TFs that do not regulate other operons in this layer (only lower layers)
Continue …



Network with all regulatory edges pointing downwards

→ a few global regulators (•) control all the details

Ma et al., *BMC Bioinformatics* **5** (2004) 199

Global Regulators in *E. coli*

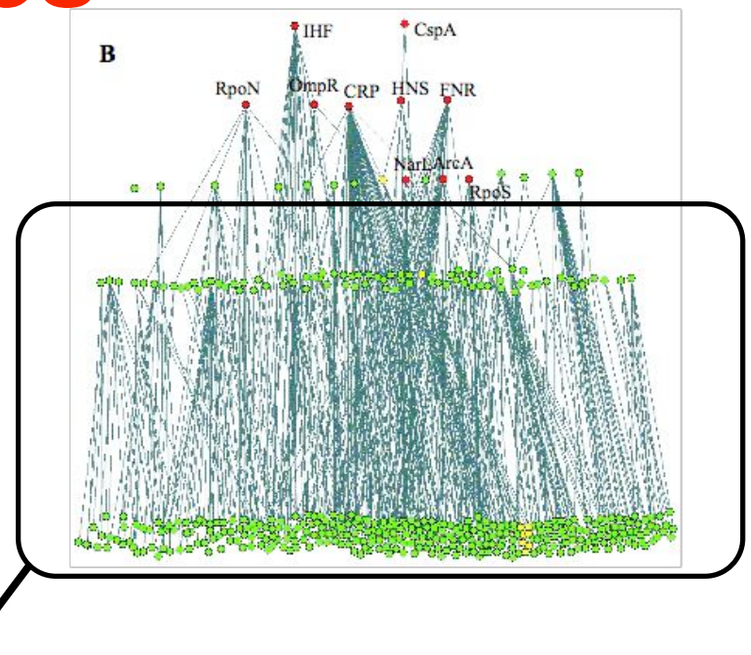
Table 1: Global regulators and their regulated operons and functions in the regulatory network of *E. coli*.

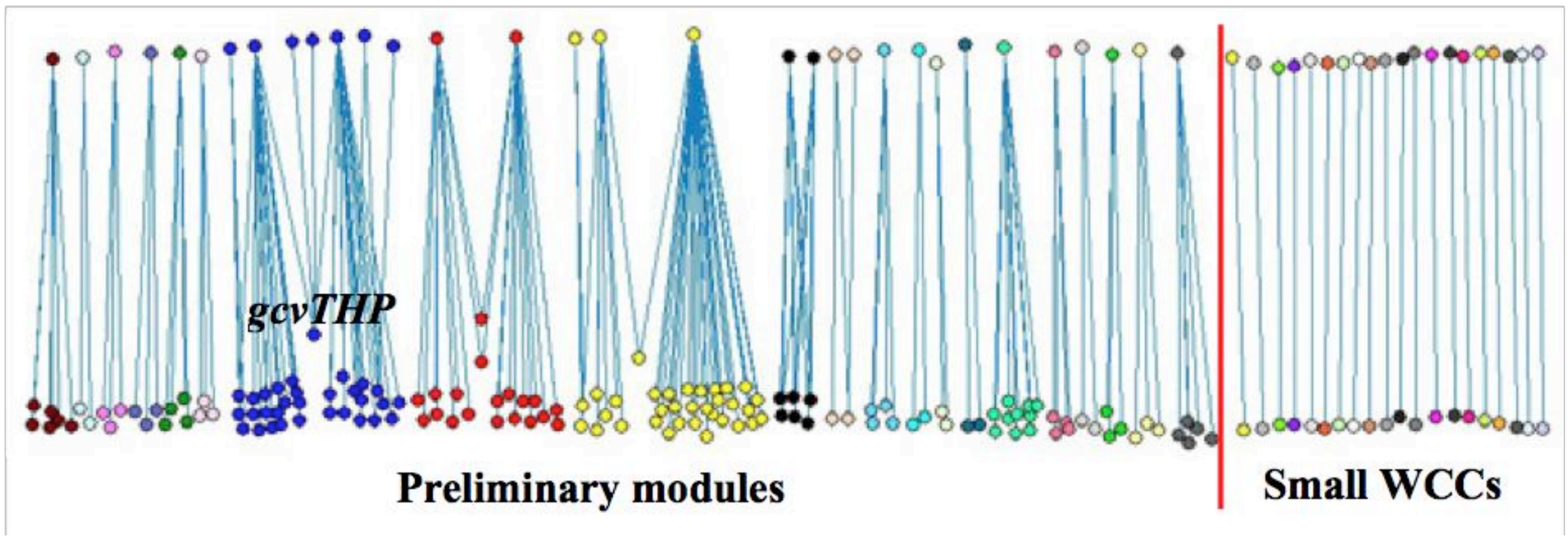
Global regulator	directly regulated Operons	Total regulated operons	Modules regulated	Function
IHF	21	39	15	integration host factor
CspA	2	24	5	Cold shock protein
CRP	72	112	21	cAMP receptor protein
FNR	22	38	16	anaerobic regulator, regulatory gene for nitrite and nitrate reductases, fumarate reductase
HNS	7	22	5	DNA-binding global regulator; involved in chromosome organization; preferentially binds bent DNA
OmpR	6	20	3	Response regulator for osmoregulation; regulates production of membrane proteins
RpoN	12	17	4	RNA polymerase sigma 54 subunit
RpoS	14	24	8	stationary phase sigma factor
ArcA	20	21	6	Response regulator protein represses aerobic genes under anaerobic growth conditions and activates some anaerobic genes
NarL	13	15	5	Two-component regulator protein for nitrate/nitrite response

Ma et al., *BMC Bioinformatics* **5** (2004) 199

E.coli GRN modules

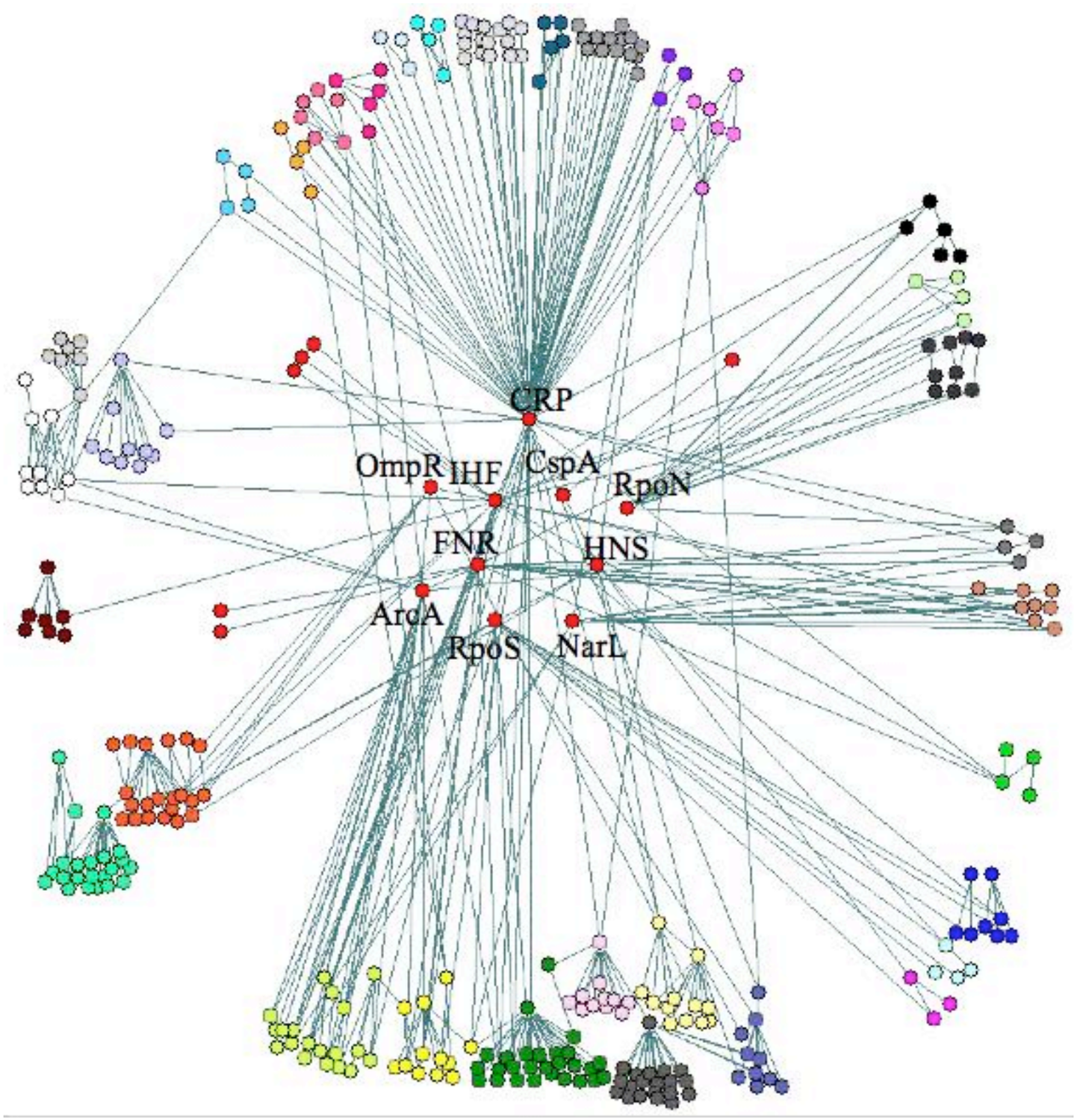
Remove top 3 layers and determine WCCs
→ just a few modules





Ma et al., *BMC Bioinformatics* **5** (2004) 199

Putting it back together



The 10 global regulators are at the core of the network,

some hierarchies exist between the modules

Ma et al., *BMC Bioinformatics* **5** (2004) 199

Modules have specific functions

Table 2: Functional investigation of modules identified.

index	Operons included	Biological function description
1	*aceBAK, acs, adhE, fruBKA, fruR, icdA, iclMR, mlc, ppsA, ptsG, ptsHI_crr, pykF*	Hexose PTS transport system, PEP generation, Acetate usage, glyoxylate shunt
2	*acnA, fpr, fumC, marRAB, nfo, sodA, soxR, soxS, zwf*	Oxidative stress response
3	*ada_alkB, aidB, alkA, ahpCF, dps, gorA, katG, oxyR*	Oxidative stress response, Alkylation
4	*alaWX, aldB, argU, argW, argX_hisR_leuT_proM, aspV, dnaA, leuQPV, leuX, lysT_valT_lysW, metT_leuW_glnUW_metU_glnVX, metY_yhbC_nusA_infB, nrdAB, pdhR_aceEF_lpdA, pheU, pheV, proK, proL, proP, sdhCDAB_b0725_sucABCD, serT, serX, thrU_tyrU_glyT_thrT, thrW, tyrTV, valUXY_lysV, yhdG_fis*	rRNA, tRNA genes, DNA synthesis system, pyruvate dehydrogenase and ketoglutarate dehydrogenase system
5	*araBAD, araC, araE, araFGH, araJ*	Arabinose uptake and usage
6	*argCBH, argD, argE, argF, argI, argR, carAB*	Arginine usage, urea cycle
7	*caiF, caiTABCDE, fixABCX*	Carnitine usage
8	*clpP, dnaKJ, grpE, hflB, htpG, htpY, ibpAB, lon, mopA, mopB, rpoH*	Heat shock response
9	*codBA, cvpA_purF_ubiX, glnB, glyA, guaBA, metA, metH, metR, prsA, purC, purEK, purHD, purL, purMN, purR, pyrC, pyrD, speA, ycfC_purB, metC, metF, metJ*	Purine synthesis, purine and pyrimidine salvage pathway, methionine synthesis
10	*cpxAR, cpxP, dsbA, ecfI, htrA, motABcheAW, ppiA, skp_lpxDA_fabZ, tsr, xprB_dsbC_recJ*	Stress response, Conjugative plasmid expression, cell motility and Chemotaxis
11	*dctA, dcuB_fumB, frdABCD, yjdHG*	C4 dicarboxylate uptake
12	*edd_eda, gntKU, gntR, gntT*	Gluconate usage, ED pathway
13	*csgBA, csgDEFG, envY_ompT, evgA, gcvA, gcvR, gcvTHP, gltBDF, ilvIH, kbl_tdh, livJ, livKHMGF, lrp, lysU, ompC, ompF, oppABCDF, osmC, sdaA, serA, stpA*	Amino acid uptake and usage
14	*fdhF, fhlA, hycABCDEFGH, hypABCDE*	Formate hydrogenlyase system
15	*flgAMN, flgBCDEFGHIJ, flgKL, flgMN, flhBAE, flhDC, fliAZY, fliC, fliDST, fliE, fliFGHIJK, fliLMNOPQR, tarTapcheRBYZ*	Flagella motility system
16	*ftsQAZ, rcsAB, wza_wzb_b2060_wcaA_wcaB*	Capsule synthesis, cell division
17	*gdhA, glnALG, glnHPQ, nac, putAP*	Glutamine and proline utilization
18	*glmUS, manXYZ, nagBACD, nagE*	Glucosamine, mannose utilization
19	*glpACB, glpD, glpFK, glpR, glpTQ*	Glycerol phosphate utilization
20	*lysA, lysR, tdcABCDEFG, tdcR*	Serine, threonine usage
21	*[illegible] malK_lamB_malM_malPQ_malS_malT, malZ*	Maltose utilization

Ma et al., *BMC Bioinformatics* **5** (2004) 199

Transcription factors in yeast *S. cereviseae*

Q: How can one define transcription factors?

Hughes & de Boer consider as TFs proteins that
(a) bind DNA directly and in a sequence-specific manner and
(b) function to regulate transcription nearby sequences they bind

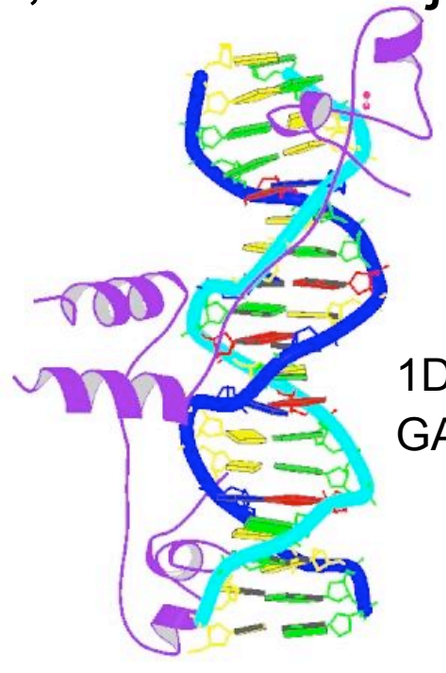
Q: Is this a good definition?

E.g. only 8 of 545 human proteins that bind specific DNA sequences and regulate transcription lack a known DNA-binding domain (DBD).

Hughes, de Boer (2013) Genetics 195, 9-36

Transcription factors in yeast

Hughes and de Boer list 209 known and putative yeast TFs, the vast majority of which contain a canonical DNA-binding domain.



1D66.pdb
GAL4 family

<u>Most abundant:</u>

- GAL4/zinc cluster domain (57 proteins),
  largely specific to fungi (e.g. yeast)

- zinc finger C2H2 domain (41 proteins),
  most common among all eukaryotes.

<u>Other classes :</u>

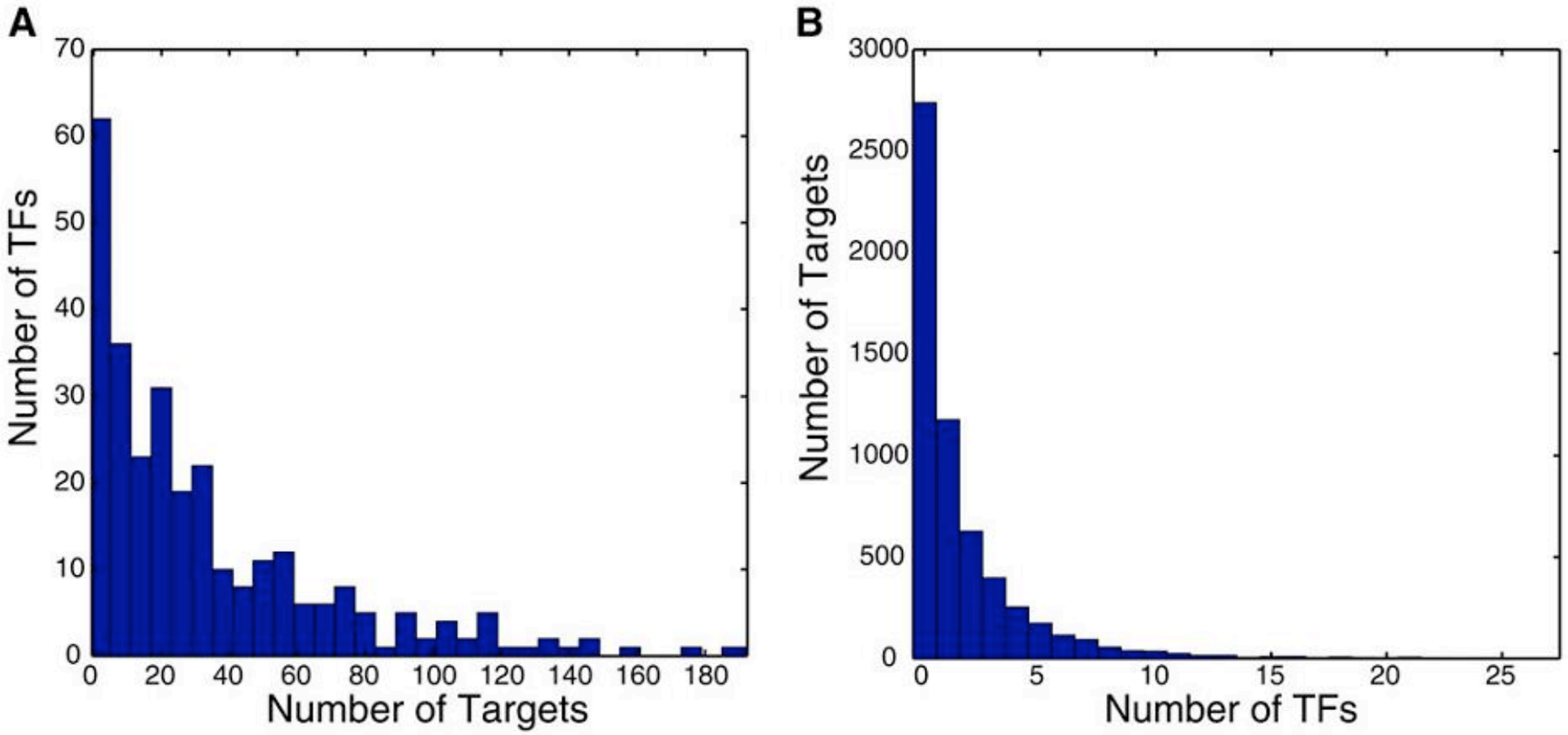
- bZIP (15),
- Homeodomain (12),
- GATA (10), and
- basic helix-loop-helix (bHLH) (8).

Hughes, de Boer (2013) Genetics 195, 9-36

TFs of *S. cereviseae*

(A) Most TFs tend to bind relatively few targets.
57 out of 155 unique proteins bind to ≤ 5 promoters in at least one condition.
17 did not significantly bind to any promoters under any condition tested.

In contrast, several TFs have hundreds of promoter targets. These TFs include the general regulatory factors (GRFs), which play a global role in transcription under diverse conditions.



Hughes, de Boer (2013) Genetics 195, 9-36

Co-expression of TFs and target genes?

Overexpression of a TF often leads to induction or repression of target genes.

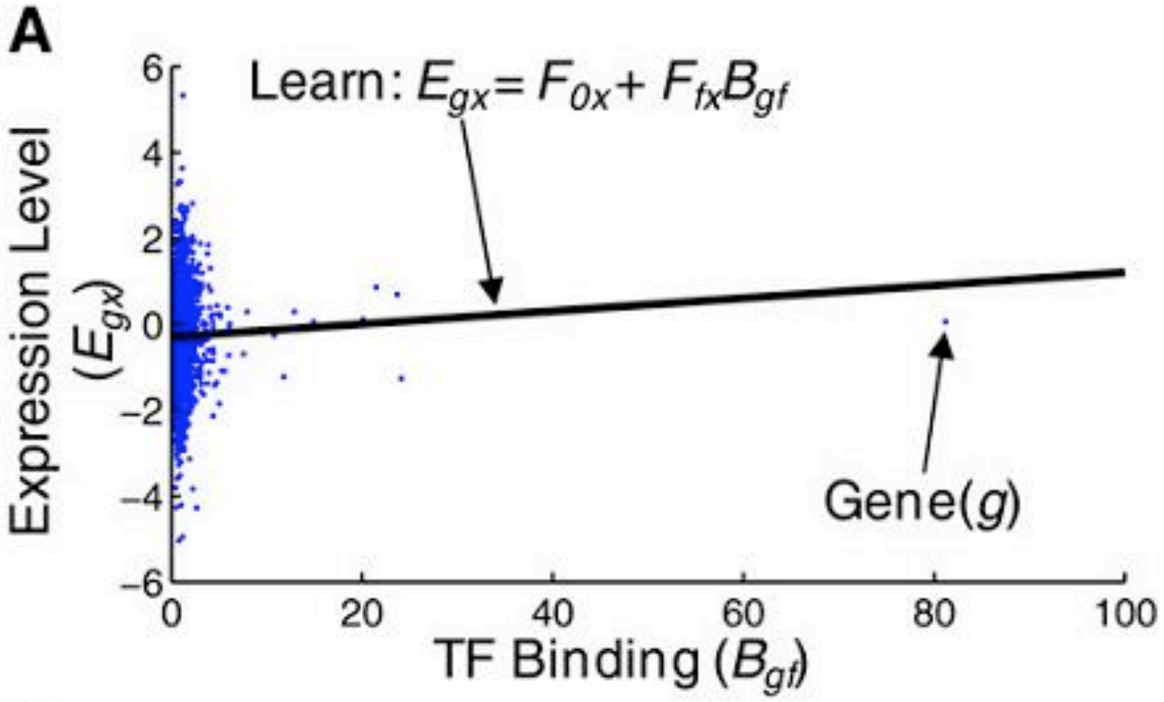
This suggests that many TFs can be regulated simply by the abundance (expression levels) of the TF.

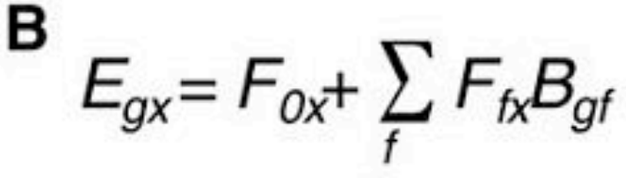
However, across 1000 microarray expression experiments for yeast, the **correlation** between a TF's expression and that of its ChIP-based targets was typically **very low** (only between 0 and 0.25).

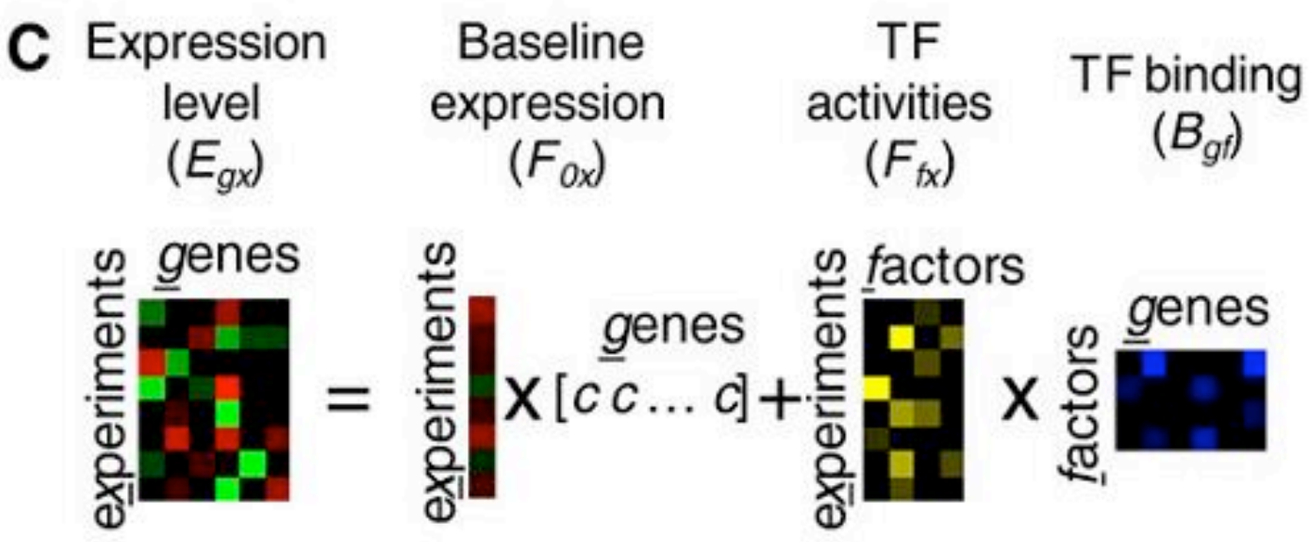
Considering that at least some of this correlation can be accounted for by the fact that a subset of TFs autoregulate, this finding supports the notion that TF expression accounts for only a minority of the regulation of TF activity in yeast.

Hughes, de Boer (2013) Genetics 195, 9-36

Using regression to predict gene expression







(A) Example where the relationship between expression level (E_{gx}) and TF binding to promoters (B_{gf}) is found for a single experiment (x) and a single TF (f). Here, the model learns 2 parameters: the background expression level for all genes in the experiment (F_{0x}) and the activity of the transcription factor in the given experiment (F_{fx}).

(B) The generalized equation for multiple factors and multiple experiments.

(C) Matrix representation of the generalized equation.

Baseline expression is the same for all genes and so is represented as a single vector multiplied by a row vector of constants where c = 1/(no. genes).

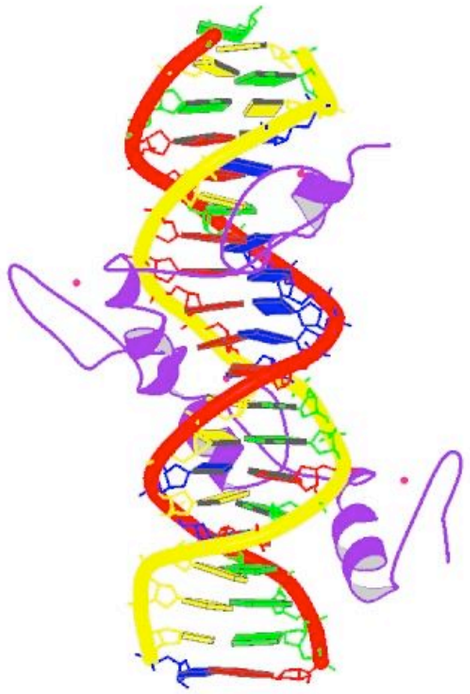
Hughes, de Boer (2013) Genetics 195, 9-36

Transcription factors in human: ENCODE

Some TFs can activate and express target genes.

YY1 shows largest mixed group of target genes.

TF	Ubiquitously activated	Ubiquitously repressed
YY1	COQ5[cd]	AC091153.1
	CPNE1	ATP5O
	CPSF2 [cd]	BIRC6[d]
	CR613718	CAPZA2
	IP6K2[a]	CXorf26
	NARS[ac]	DKFZp434H247
	PAK4[d]	EFHA1
	PSMB4[ac]	MRPS10[c]
	UBR5	MRPS18B[acd]
		NUP160
		OXCT1
		PSMD8[ac]
		SNX27
		SNX3[ad]
		SRP68[ad]
		TNKS



1UBD.pdb
human YY1

Whitfield et al. Genome Biology 2012, 13:R50

YY1 binding motifs

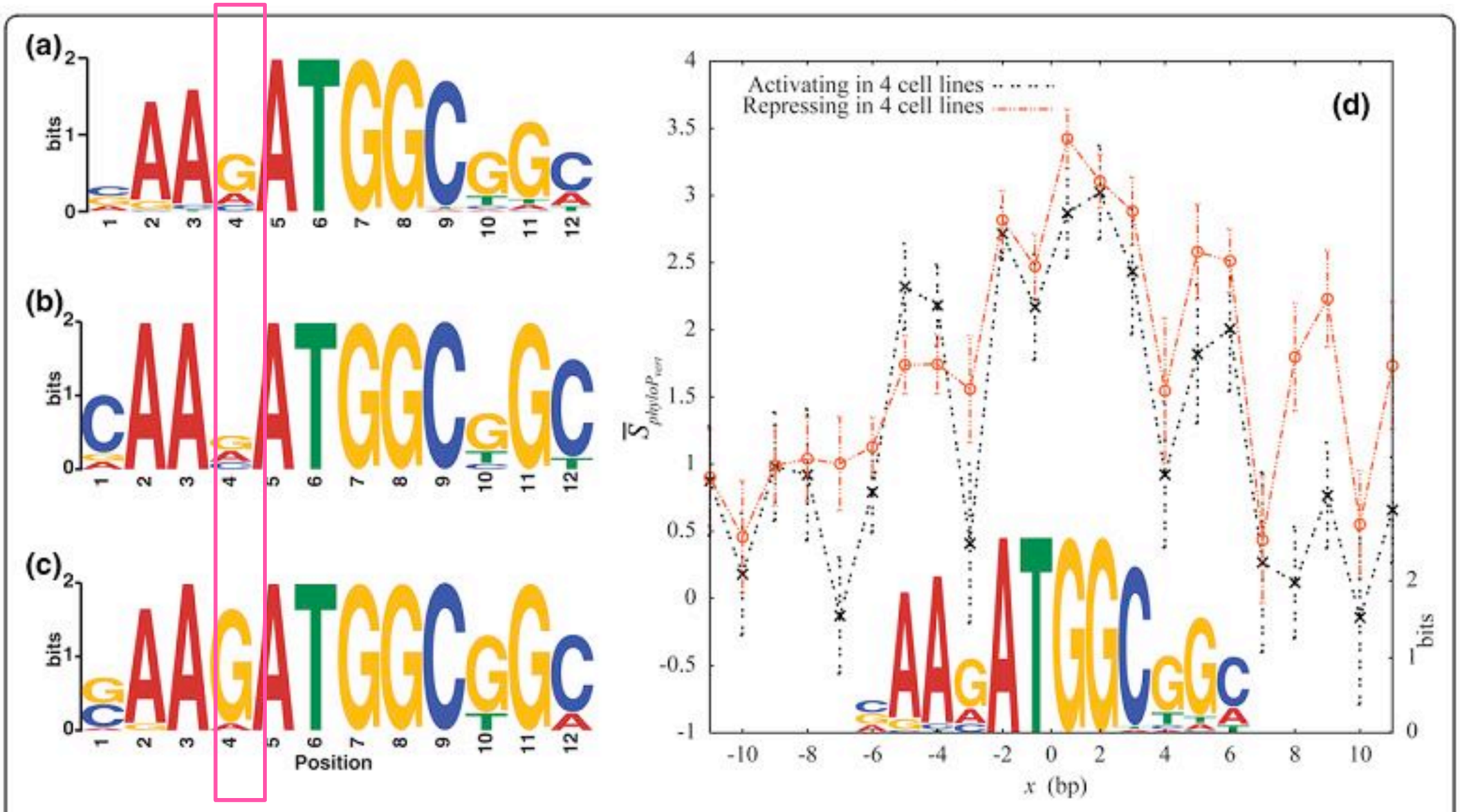
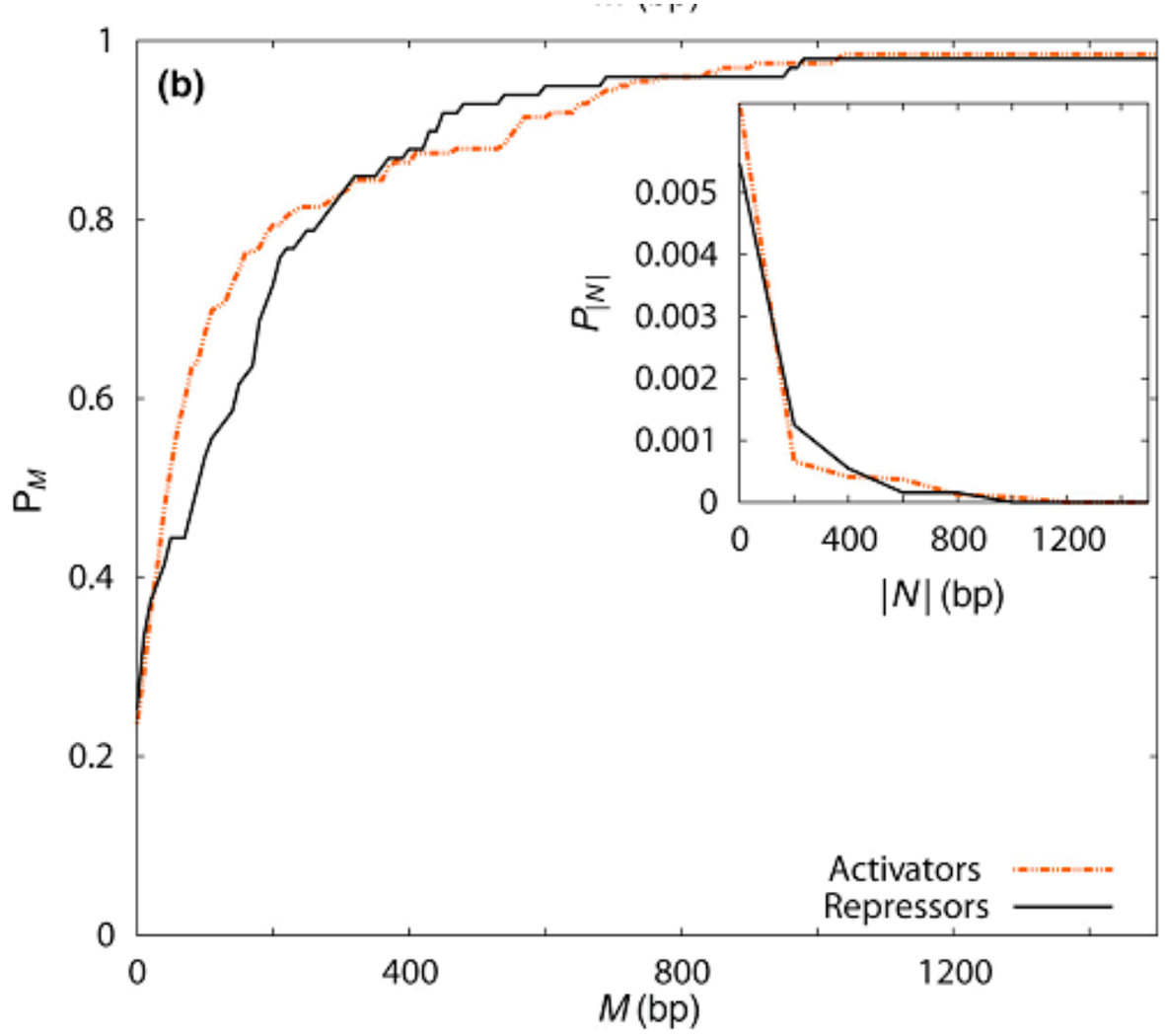


Figure 2 Characterization of functional YY1 binding sites. Sequence logo [102] for YY1 binding sites from **(a)** PWM and sites that are functionally **(b)** ubiquitously activating (9 BS) or **(c)** ubiquitously repressive (16 BS) in four human cell lines. In **(d)**, we plot the mean vertebrate phyloP conservation score [90] around functional YY1 binding sites. The mean score, $\bar{S}_{phyloP_{vert}}$, was computed at each base for sites where the binding event ubiquitously activated (black line) or repressed (red line) transcription in all four cell lines. The position weight matrix that was used to predict YY1 binding sites is shown (scale on the right axis).

Whitfield et al. Genome Biology 2012, 13:R50

Where are TF binding sites wrt TSS?



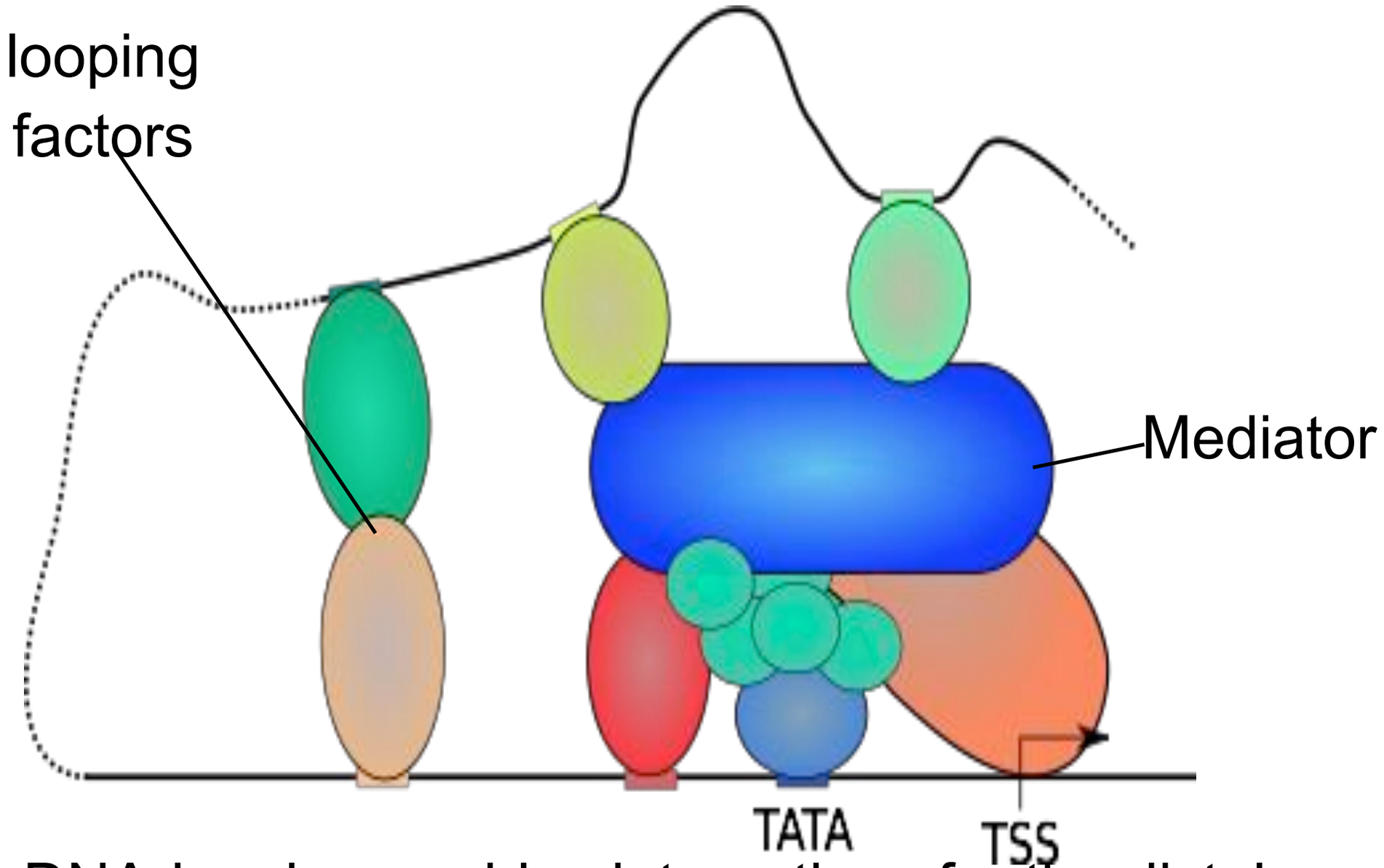
Inset: probability to find binding site at position N from transcriptional start site (TSS)

Main plot: cumulative distribution.

activating TF binding sites are significantly closer to the TSS than repressing TF binding sites ($p = 4.7 \times 10^{-2}$).

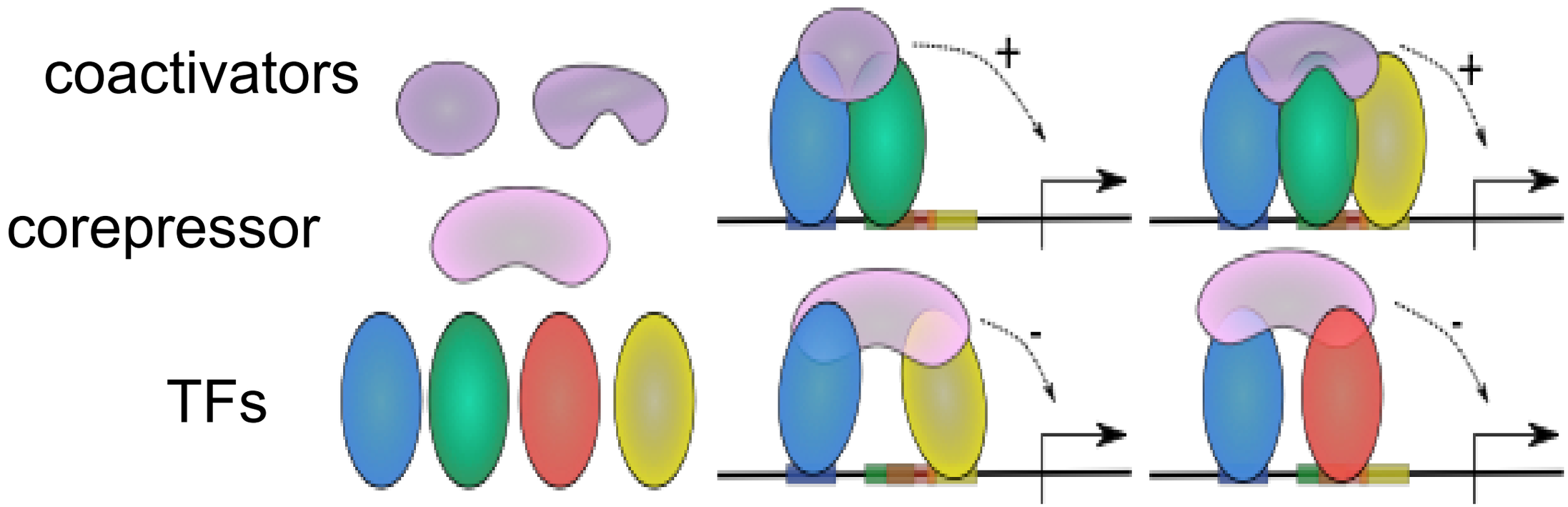
Whitfield et al. Genome Biology 2012, 13:R50

Cooperative transcriptional activation



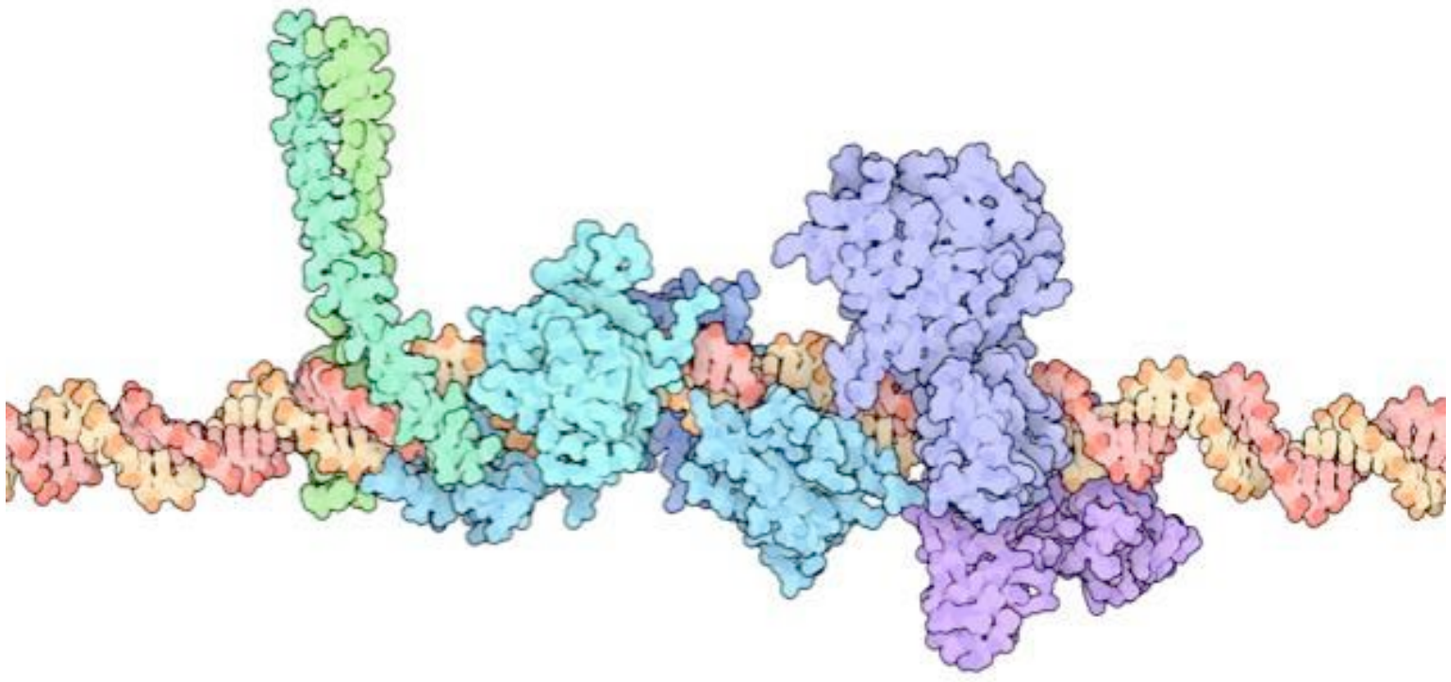
DNA-looping enables interactions for the distal promotor regions,
Mediator cofactor-complex serves as a huge linker

cis-regulatory modules



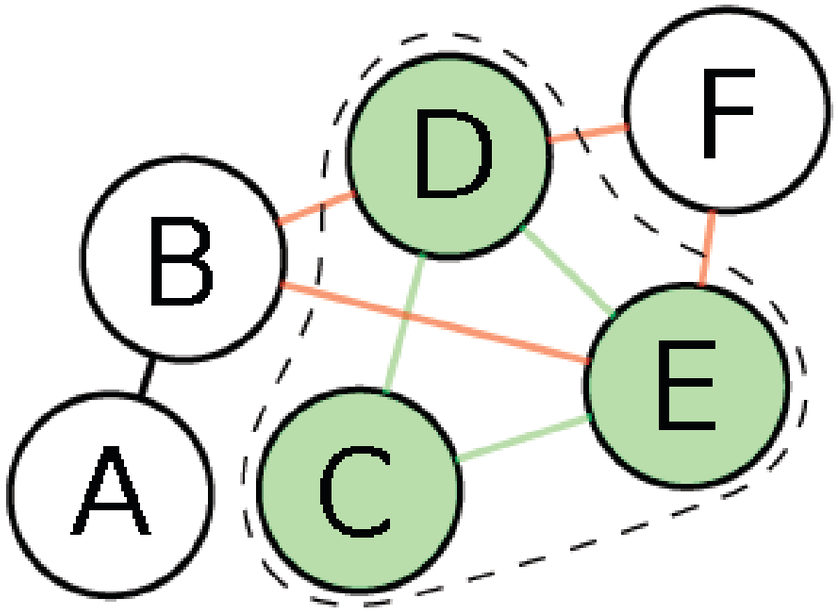
TFs are not dedicated activators or respressors!

It‘s the assembly that is crucial.



IFN-enhanceosome from RCSB Protein Data Bank, 2010

Protein complexes involving multiple transcription factors



Borrow idea from ClusterOne method:

Identify candidates of TF complexes

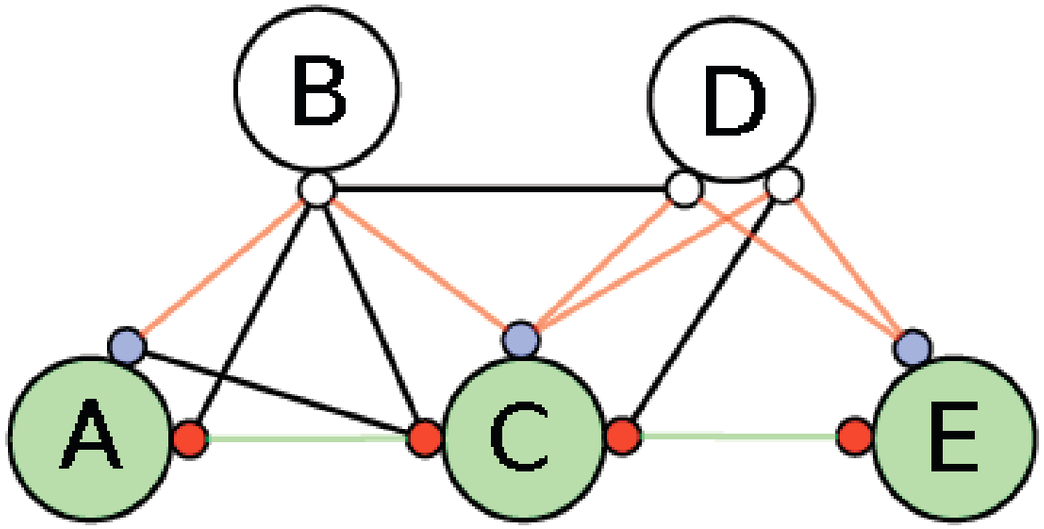
in protein-protein interaction graph

by **optimizing the cohesiveness**

$$f(V) = \frac{w^{in}(V)}{w^{in}(V) + w^{bound}(V)}$$

underlying domain-domain representation of PPIs

Assumption: every domain supports only one interaction.

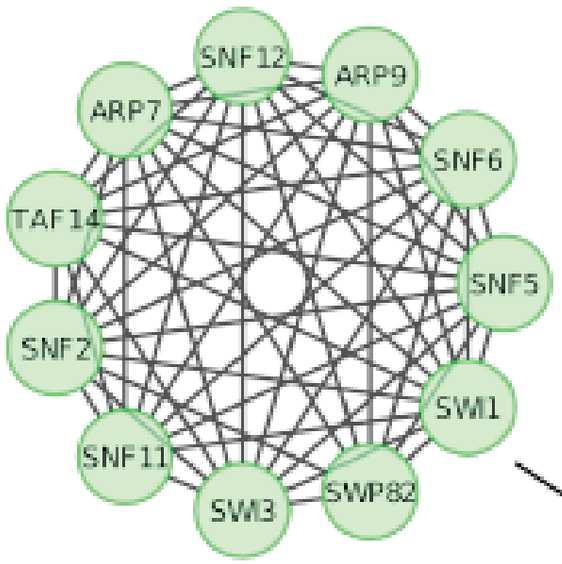


Green proteins A, C, E form actual complex.

Their red domains are connected by the two green edges.

B and D are incident proteins. They could form new interactions (red edges) with unused domains (blue) of A, C, E

data source used: Yeast Promoter Atlas, PPI and DDI



weighted protein-protein interaction data, such as

PrePPI

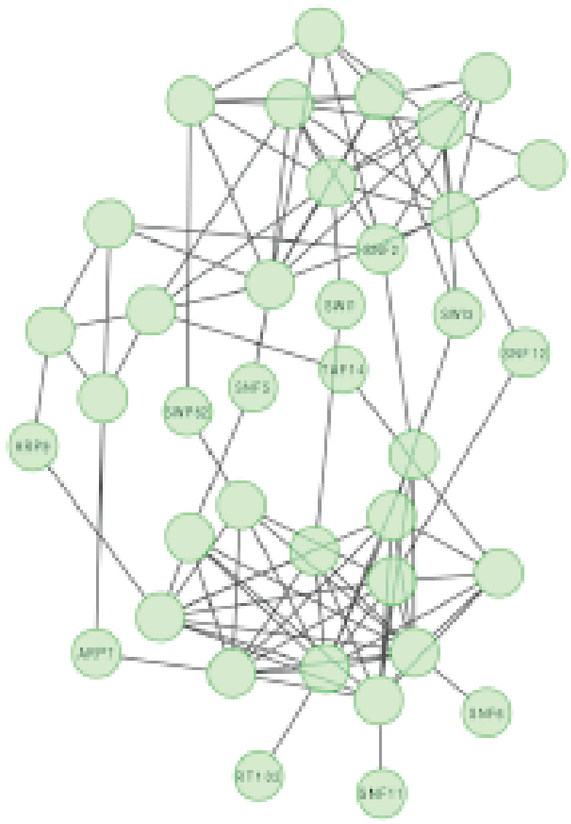
SIF format with UniProt IDs

seed proteins,
threshold for pairs,
max. depth of search

domain-aware cohesiveness optimization algorithm

transcription factor complex candidates

UniProt EMBL-EBI



domain-domain interaction network

data retrieval and buildup automatically

Pfam InterPro DOMINE IDDI

Will, T. and Helms, V. (2014)
Bioinformatics, 30, i415-i421

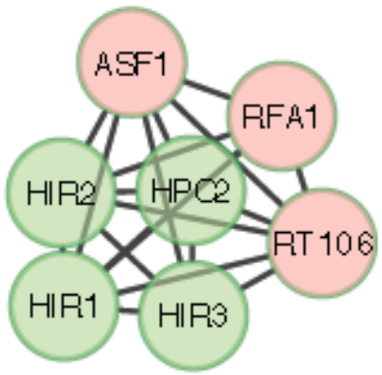
Daco identifies far more TF complexes than other methods

	DACO	Cl1ps	Cl1s	Cl1	MCD	MCL
TF complexes	1375	175/176	61/63	106/106	16/38	75/79
TF variants	412	134/138	59/61	80/80	16/38	75/79

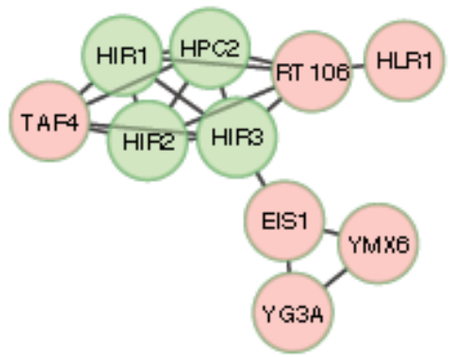
ClusterOne (Cl1), MCD and MCL are other methods to generate protein complexes from PP interaction data.

Listed here are the number of disjoint protein complexes generated by these methods that involve at least 2 TFs.

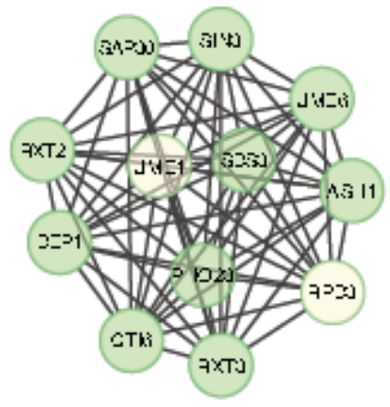
Examples of TF complexes – comparison with ClusterONE



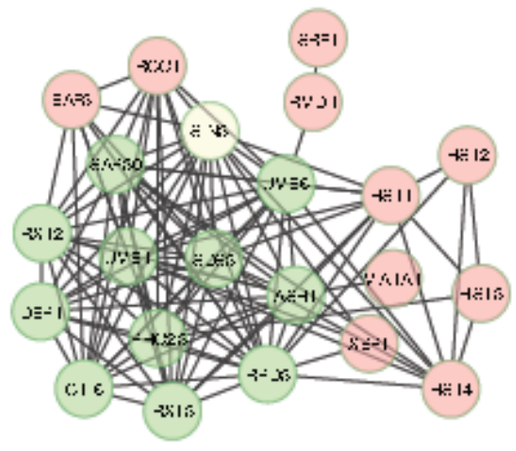
(a) HIR(SGD) / DACO



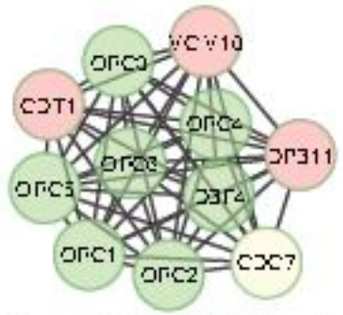
(b) HIR(SGD) / ClusterONE



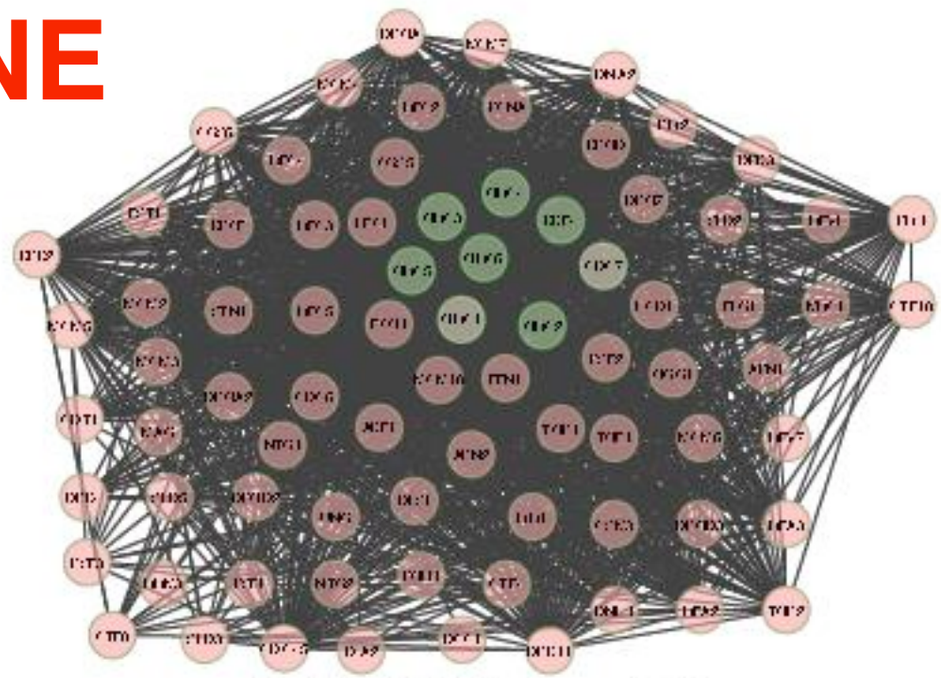
(c) RPD3L(CYC2008) / DACO



(d) RPD3L(CYC2008) / ClusterONE



(e) ORC(MIPS) / DACO



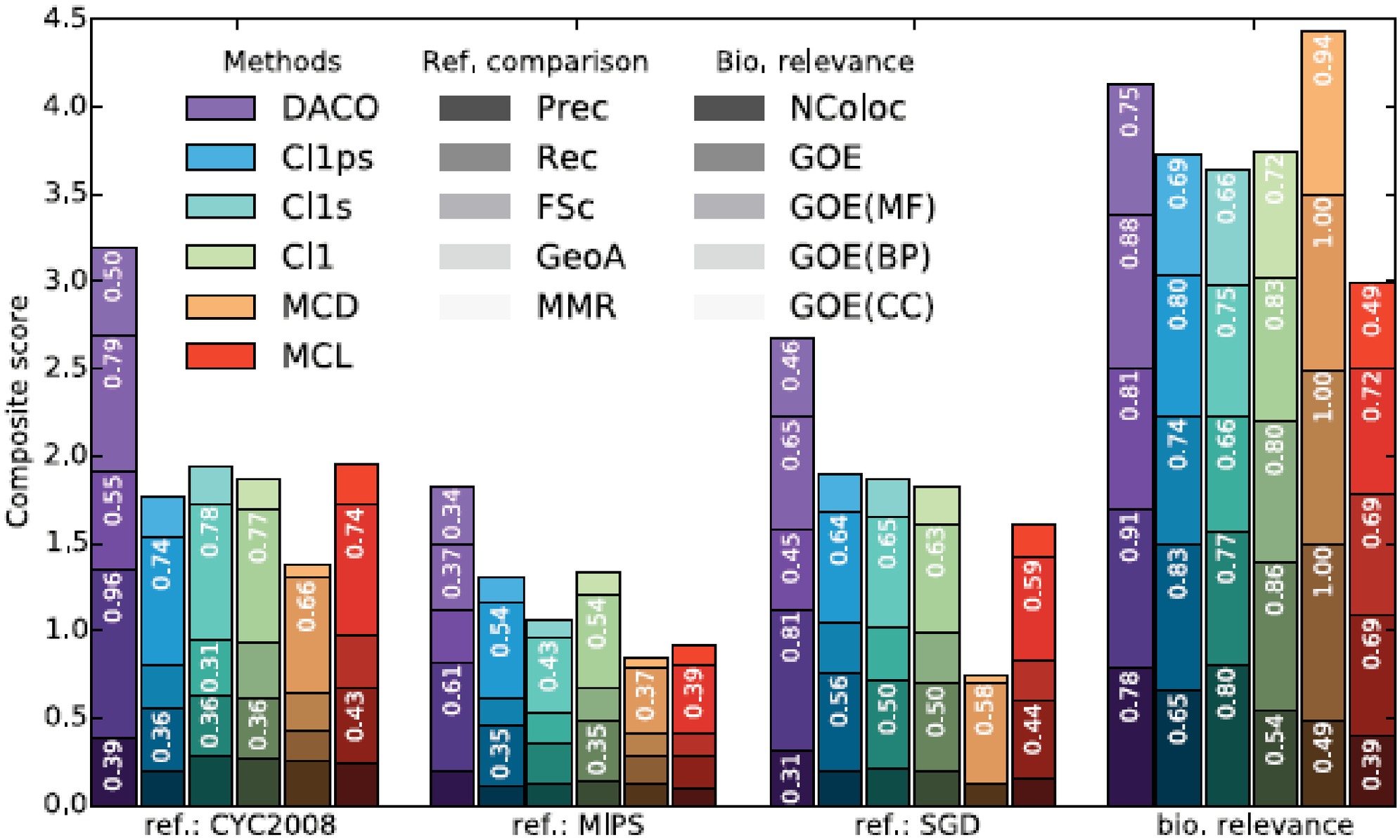
(f) ORC(MIPS) / ClusterONE

Green nodes: proteins in the reference that were matched by the prediction

red nodes: proteins that are in the predicted complex, but not part of the reference.

→ DACO complexes are more compact than ClusterONE complexes

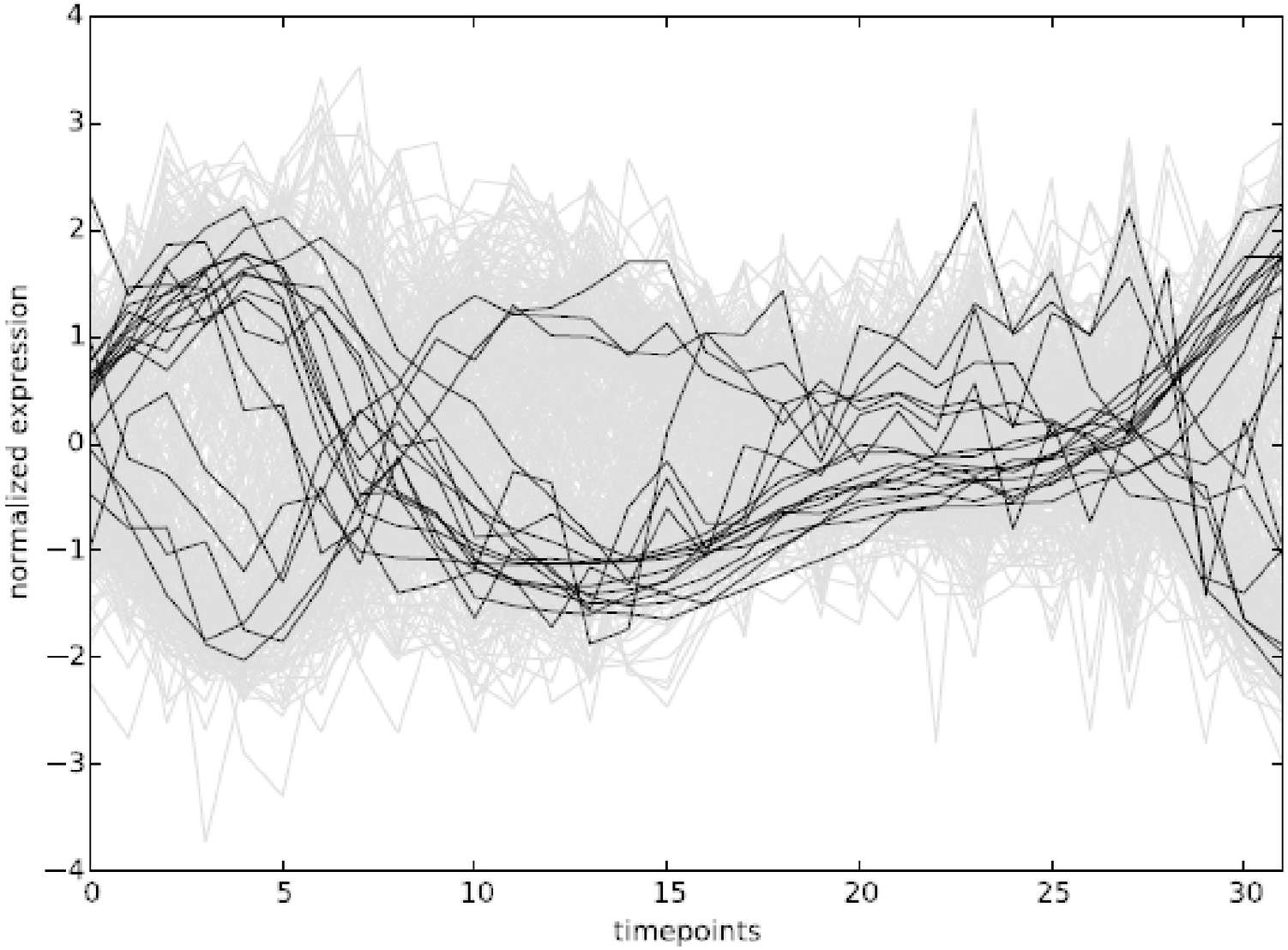
Performance evaluation



Columns 1-3: overlap of predicted complexes with gold-standard sets

Column 4: functional homogeneity (GO terms) of complex components

Co-expressed target genes of MET4/MET32 TF complex during yeast cell cycle



X-axis: 32 time points during yeast cell cycle

Y-axis: normalized expression of target genes of TFs MET4 and MET32

Grey: target genes of either MET4 **or** MET32 show scattered expression

Black: target genes of MET4 **and** MET32 show 2 expression modes

Functional role of TF complexes

TFs	P_{dECS}	bind. mode	targets	reg. influence	GO process enrichment ($P < 0.05$) in targets
MET4/MET32	0.0010	coloc.	19	+	methionine metabolic process
TBP/HAP5	0.0335	med.	47	+	/
GLN3/DAL80	0.0009	med.	28	/	allantoin catabolic process
DIG1/STE12/SWI6	0.0369	all	15	/	fungal-type cell wall organization
FHL1/RAP1	0.0001	coloc.	116	+	rRNA transport
RPH1/GIS1	0.0001	med.	100	-	hexose catabolic process
CBF1/MET32	0.0002	coloc.	33	o	sulfate assimilation
DIG1/STE12	0.0003	med.	34	-	response to pheromone
GCN4/RAP1	0.033	med.	62	+	/
MSN4/MSN2	0.0021	med.	105	+	oligosaccharide biosynthetic process
DAL80/GZF3	0.0044	med.	20	-	purine nucleobase metabolic process
SWI6/SWI4	0.0039	med.	53	+	regulation of cyclin-dependent protein serine/threonine kinase activity
STB1/SWI6	0.0275	all	47	+	/
TBP/SWI6	0.0159	med.	14	+	/
GLN3/GZF3	0.0120	adj.	31	/	allantoin catabolic process
MBP1/SWI6/SWI4	0.0307	med.	18	+	regulation of cyclin-dependent protein serine/threonine kinase activity
MBP1/SWI6	0.0124	adj.	25	/	cell cycle process