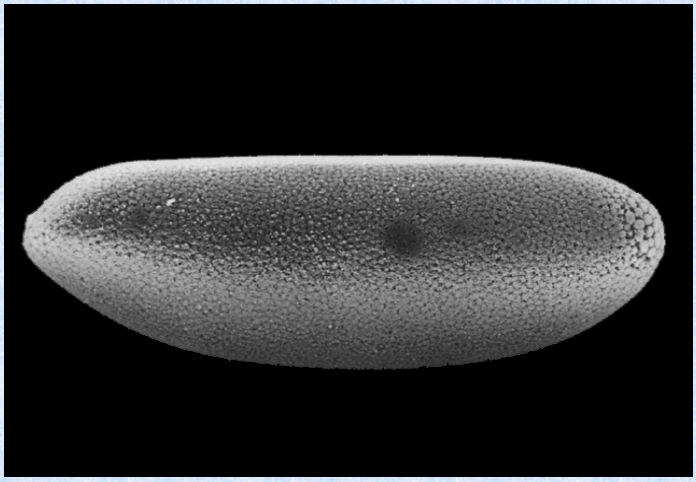
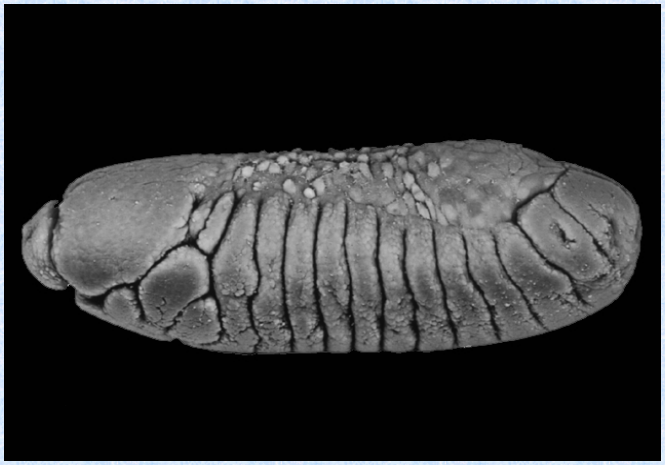
Regulatory modules: prediction and evolution

Massimo Vergassola
CNRS, Institut Pasteur
massimo@pasteur.fr

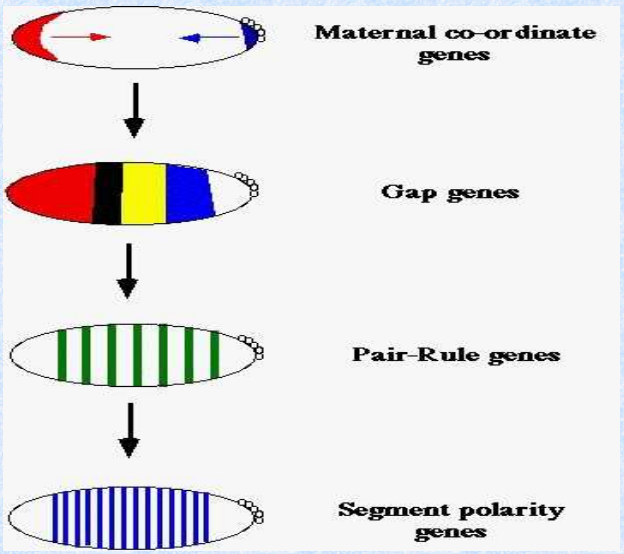
Syncytial blastoderm



Segmented embryo



Segmentation factors' cascade
(C. Nüsslein-Volhard, E. Wieschaus)

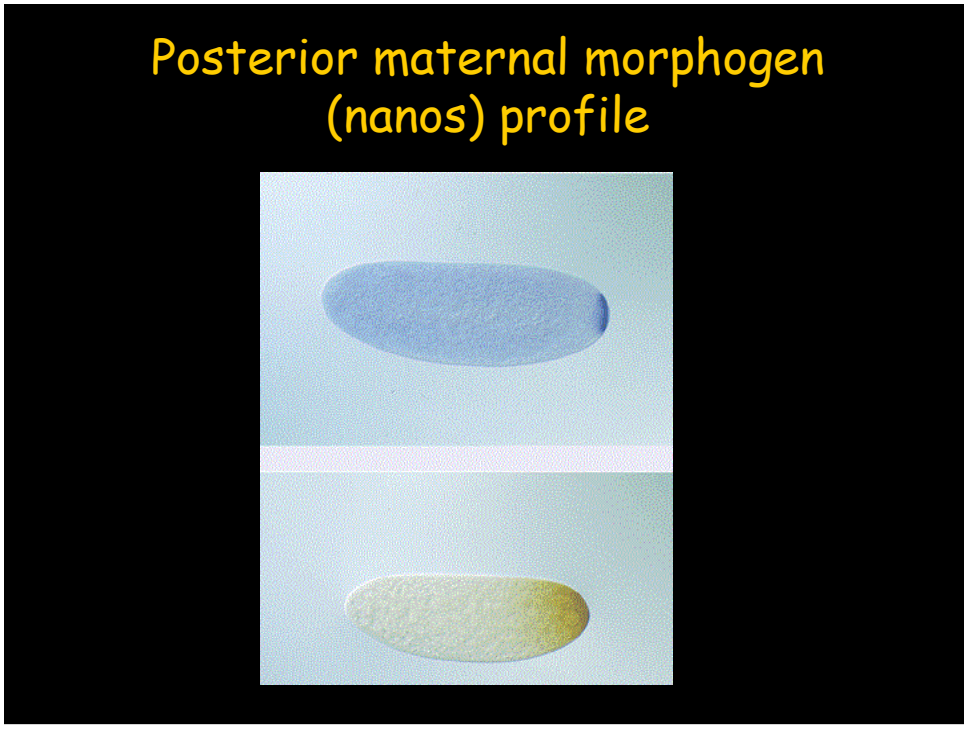


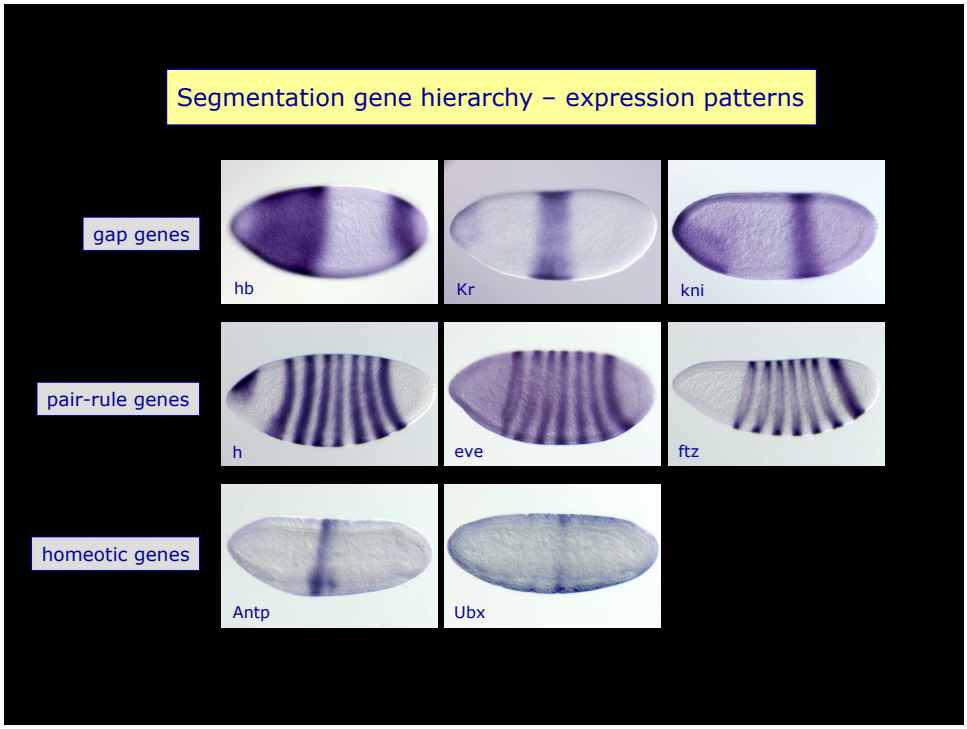
Bcd,nos,tor, dl,cad,…

Hb,Kr,tll,...

Eve,ftz,h,...

En,wg,dsh,...


Posterior maternal morphogen
(nanos) profile

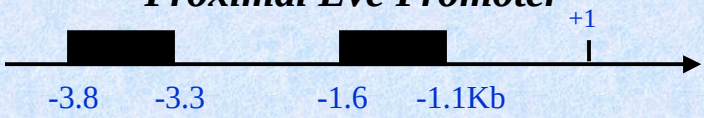

Segmentation gene hierarchy – expression patterns
gap genes
pair-rule genes
homeotic genes
hb
Kr
kni
h
eve
ftz
Antp
Ubx

Cis-regulatory modules
(review: E. Davidson's book, 2001)

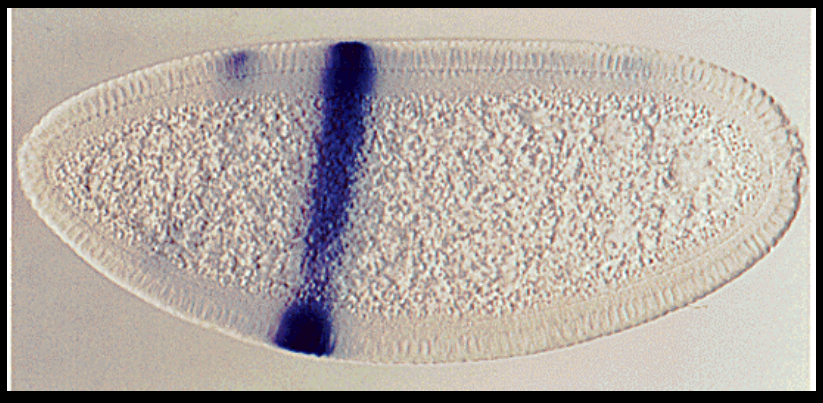
Expression profiles are a superposition of elementary patterns.

Stretches of a few hundred bases in the vicinity of the regulated gene yield those patterns by a strongly combinatorial regulation of transcription.

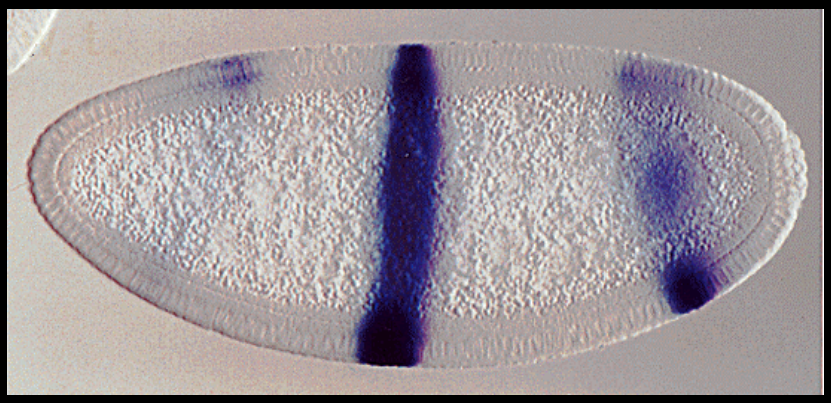
Proximal Eve Promoter



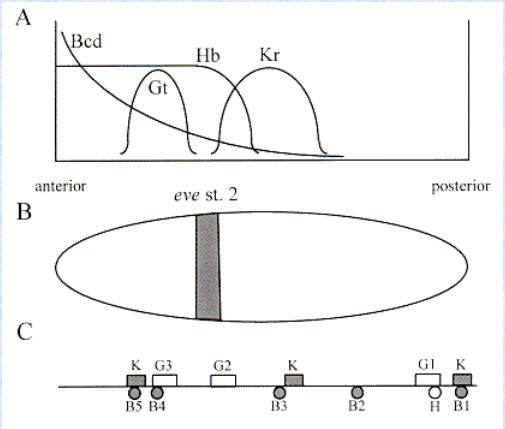
Eve 2 module



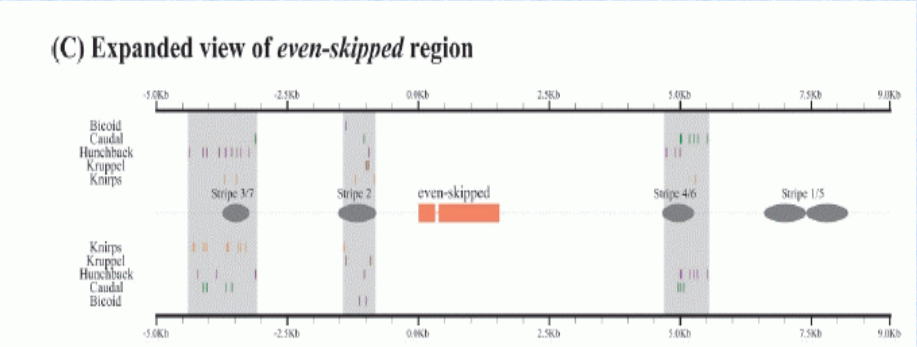
Eve 3-7 module



Binding sites eve 2 module

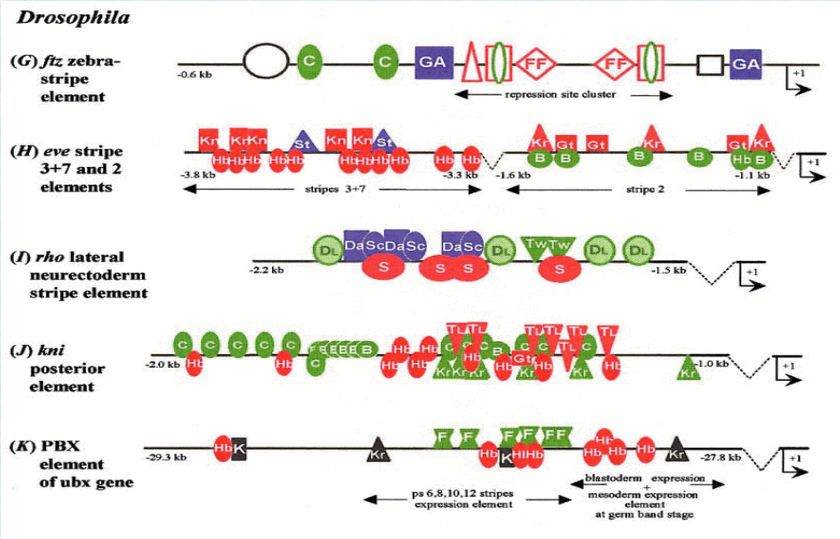


In silico prediction of modules



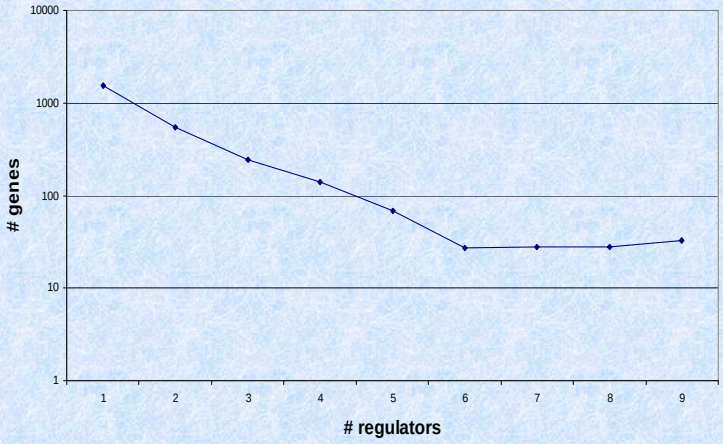
The size of the regions to explore calls for cheaper and faster clues on their location.

Combinatorial control of regulation



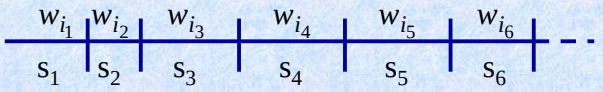
Multiple TF's binding in *S. cerevisiae*

ChIP genome-wide data (Lee et al. *Science*, **298**, 799, 2002)



The probabilistic model (I)

Goal: maximize the likelihood to observe the input sequence given the list of bricks: TF's (as weight matrices w_k) and the background model (as transition probabilities of a Markov chain w_0).



Tiling T

$$P(T) = \prod_{k=1}^{N(T)} p_{w_{i_k}} m(s_k | w_{i_k})$$

The probabilistic model (II)

- The competition among factors for overlapping sites is brought in by the partition function

$$Z = \sum P(T)$$

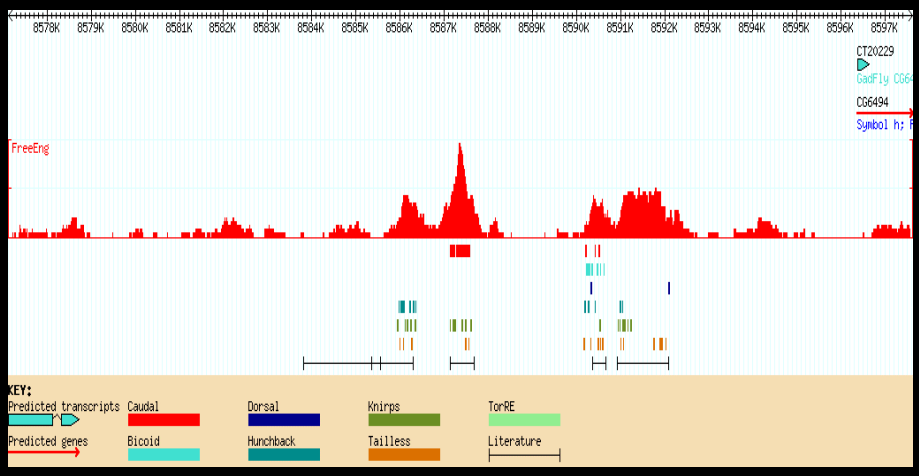
- Maximize Z over the p_w 's to get the maximum likelihood Z_{max} via the transfer matrix method

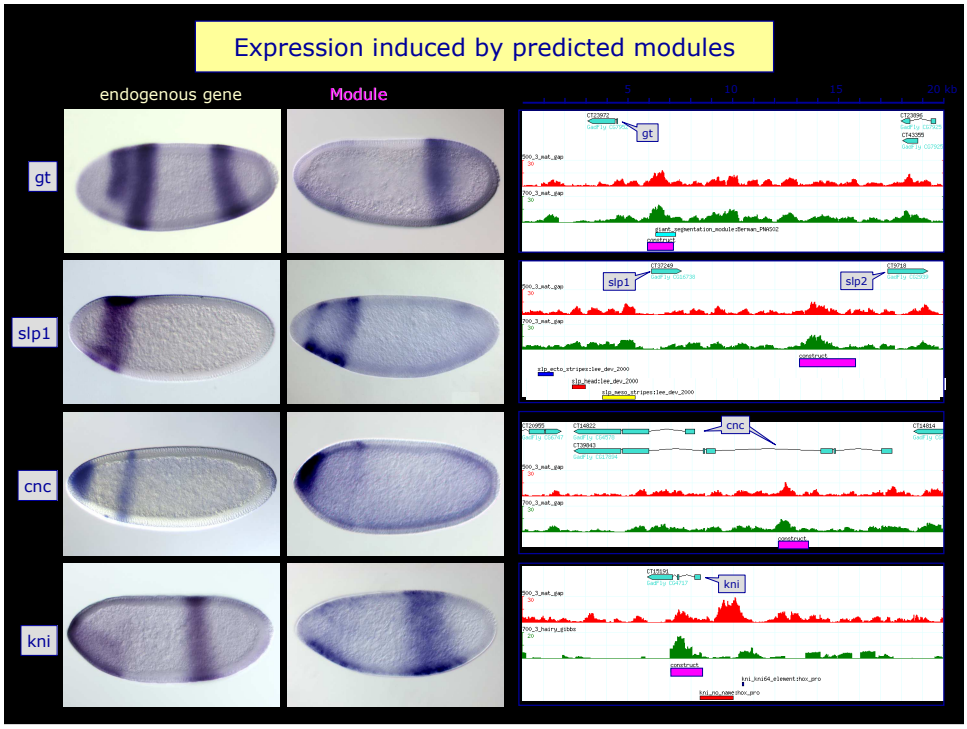
$$Z(1,i) = \sum_{k=0}^{N_w} Z(i - l_k)\, p_{w_k}\, m\left(s_{i_k} \middle| w_k\right)$$

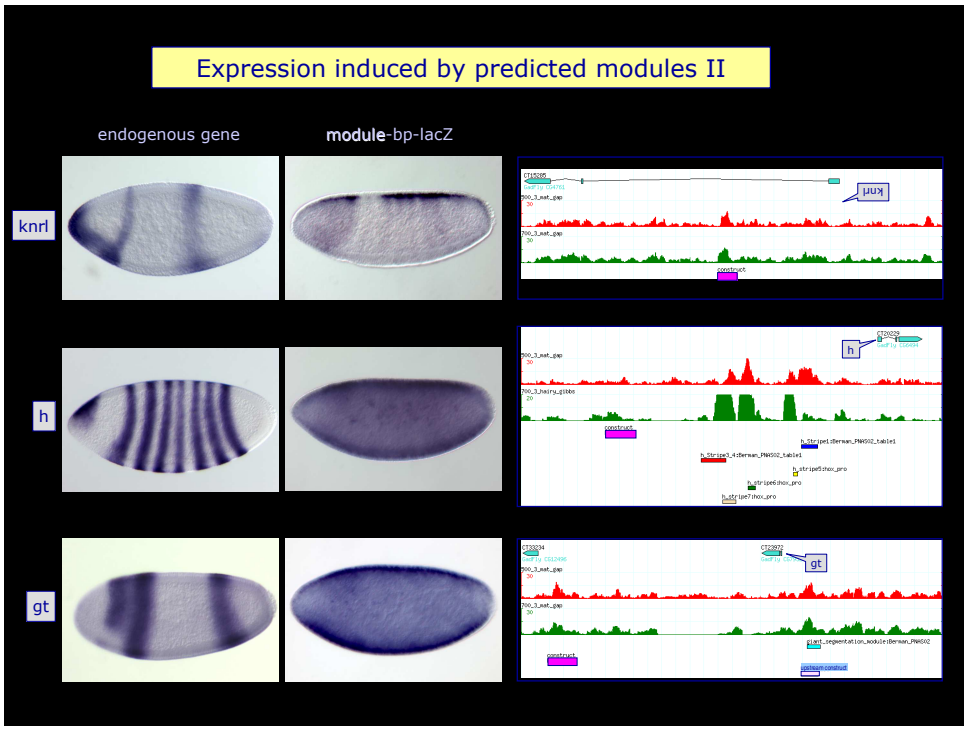
- Score of the sequence

$$R = \log(Z_{\max} / Z_{back})$$

Scores in hairy's proximal region

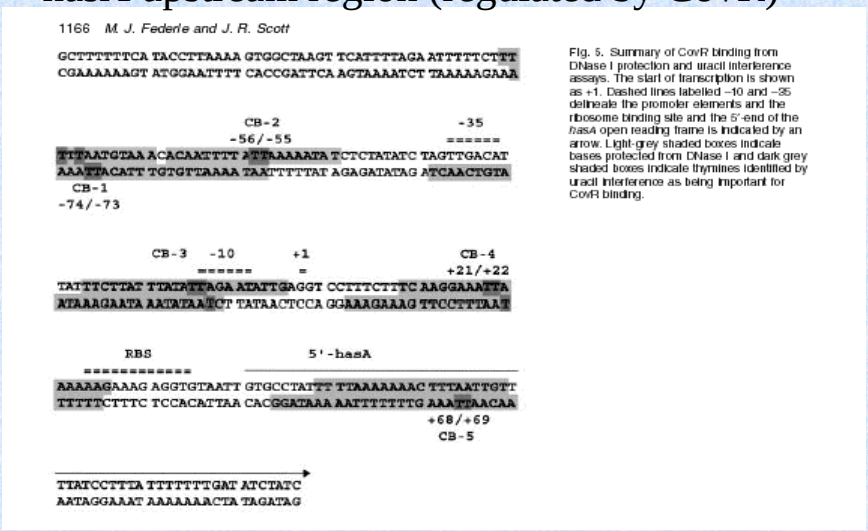



Expression induced by predicted modules
endogenous gene
Module
gt
slp1
cnc
kni
slp2


Expression induced by predicted modules II
endogenous gene
module-bp-lacZ
knrl
h
gt

Streptococcus agalactiae CovR (Control Of Virulence)

hasA upstream region (regulated by CovR)



1166 *M. J. Federle and J. R. Scott*

GCTTTTTTCA TACCTTAAAA GTGGCTAAGT TCATTTTAGA ATTTTTCTTT
CGAAAAAAGT ATGGAATTTT CACCGATTCA AGTAAAATCT TAAAAAGAAA

CB-2 -56/-55 -35
TTTAATGTAA ACACAATTTT ATTAAAAATA TCTCTATATC TAGTTGACAT
AAATTACATT TGTGTTAAAA TAATTTTTAT AGAGATATAG ATCAACTGTA
CB-1 -74/-73

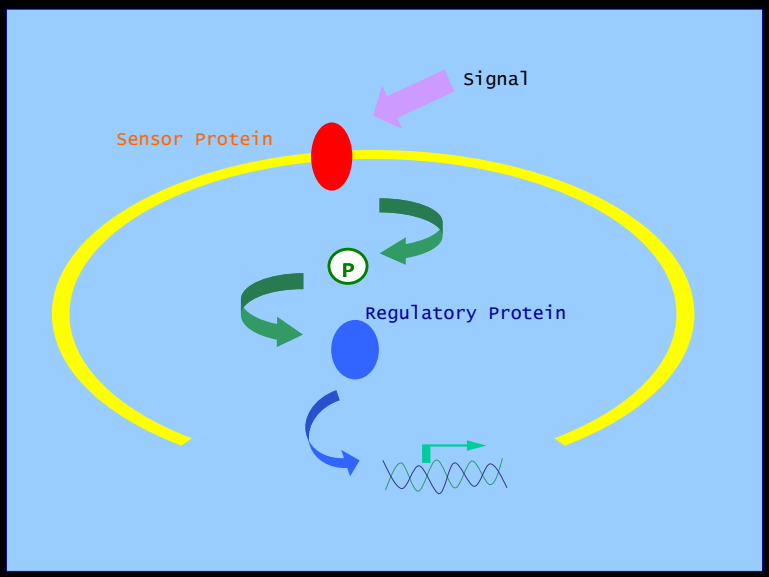
CB-3 -10 +1 CB-4 +21/+22
TATTTCTTAT TTATATTAGA ATATTGAGGT CCTTTCTTTC AAGGAAATTA
ATAAAGAATA AATATAATCT TATAACTCCA GGAAAGAAAG TTCCTTTAAT

RBS 5'-hasA
AAAAAGAAAG AGGTGTAATT GTGCCTATTT TTAAAAAAAC TTTAATTGTT
TTTTTCTTTC TCCACATTAA CACGGATAAA AATTTTTTTG AAATTAACAA
+68/+69 CB-5

TTATCCTTTA TTTTTTTGAT ATCTATC
AATAGGAAAT AAAAAAACTA TAGATAG

Fig. 5. Summary of CovR binding from DNase I protection and uracil interference assays. The start of transcription is shown as +1. Dashed lines labelled –10 and –35 delineate the promoter elements and the ribosome binding site and the 5′-end of the *hasA* open reading frame is indicated by an arrow. Light-grey shaded boxes indicate bases protected from DNase I and dark grey shaded boxes indicate thymines identified by uracil interference as being important for CovR binding.

Two-component regulatory systems

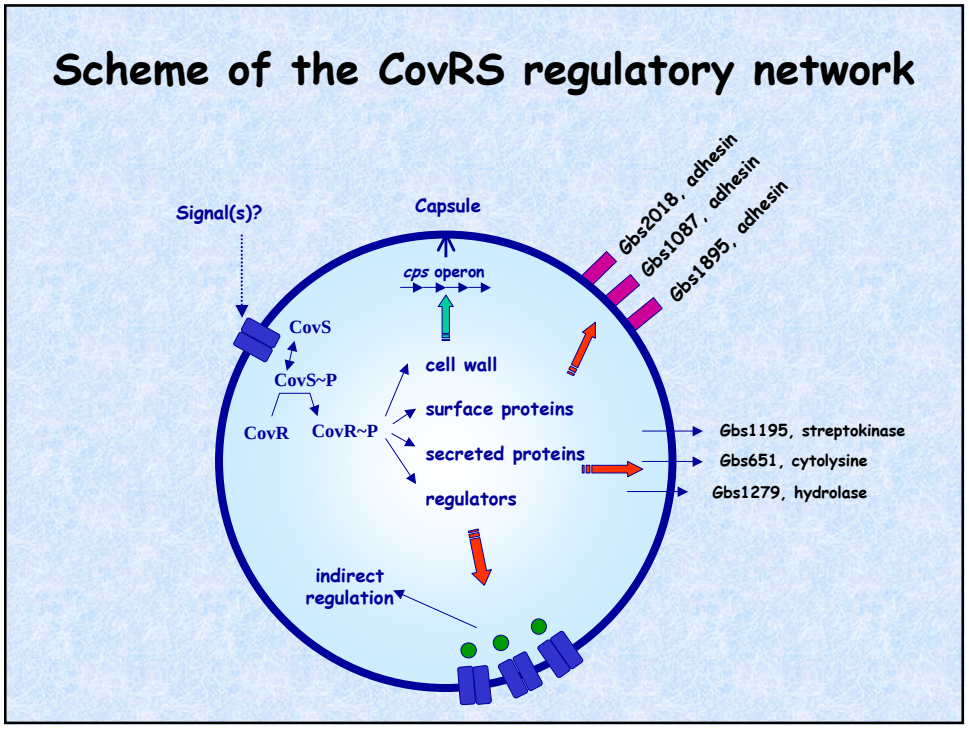


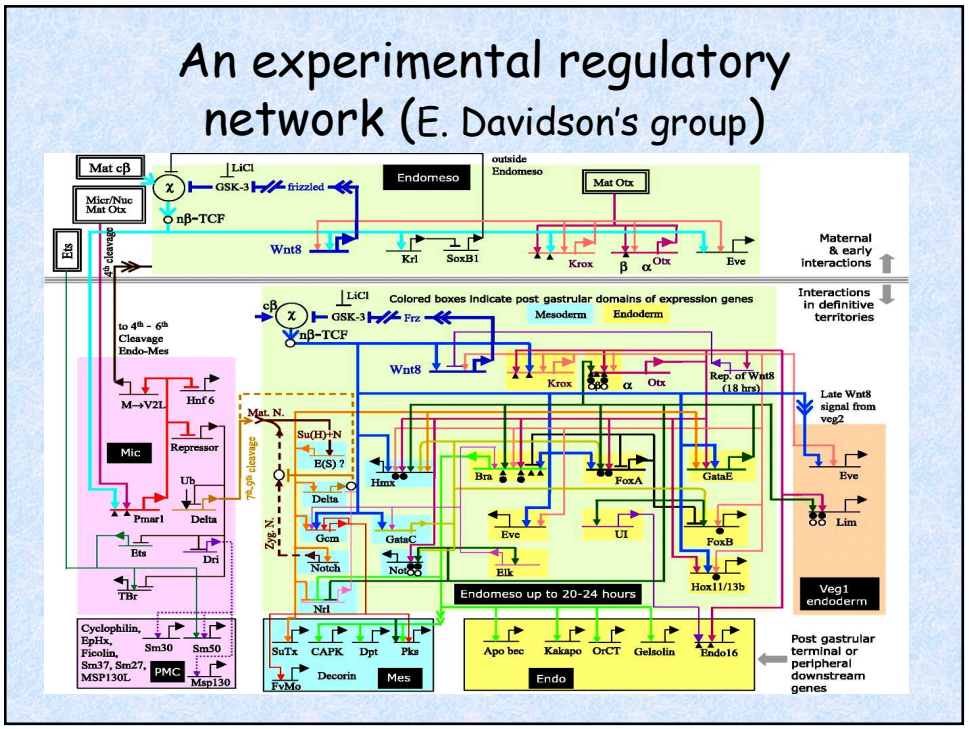
Inactivation of CovRS alters gene expression at a global level

Functional classification	OD 0,3	OD 0,6	OD 0,9
Cell wall (42)	15- 3+	26- 3+	7+
Cell surface proteins (22)	2- 7+	8- 5+	3- 5+
Regulation (35)	3- 7+	8- 19+	4- 3+
Adaptation to atypical conditions (12)	1-	1- 9+	6- 1+
Phage-related functions and Transposons (16)	1- 1+	3- 5+	5- 5+

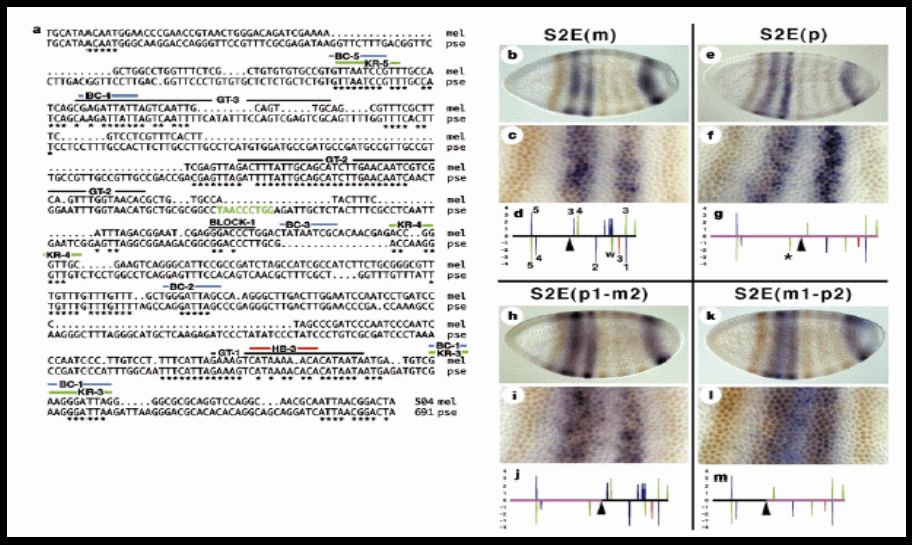
Genes under direct control of covR as predicted *in silico*

Lf	Gene	Remarks	Function
250	gbs0042	Transcriptome	Similar to phosphoribosylamine-glycine ligase
250	gbs0358	Transcriptome	Unknown
1000	gbs0493	Transcriptome	Similar to unknown protein
1000	gbs0595	Pathogenicity island	Similar to ABC transporter (ATP-binding protein)
1000	gbs1087	Transcriptome	Highly repetitive peptidoglycan bound protein (LPTX motif)
1000	gbs1194	Transcriptome	Similar to unknown protein
500	gbs1279	Transcriptome	Similar to putative hydrolytic protein
1000	gbs1312	Pathogenicity island	Exonuclease motif predicted by PFAM
250	gbs1506	Transcriptome	Similar to glycerol (sugar)-3-phosphate transporter
1000	gbs1609	Transcriptome	Unknown
100	gbs1919	Transcriptome	Similar to neuramidinase
1000	gbs1972/ gbs1973	Pathogenicity island	Similar to transcriptional regulator (phage related) /unknown
250	gbs2018	Transcriptome	Putative peptidoglycan linked protein (LPTX motif)
1000	gbs2074	Pathogenicity island	Unknown

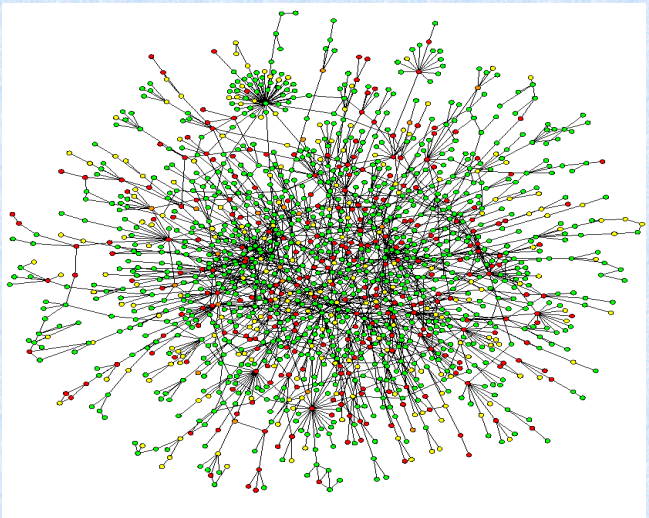

Scheme of the CovRS regulatory network
Signal(s)?
Capsule
cps operon
CovS
CovS~P
CovR
CovR~P
cell wall
surface proteins
secreted proteins
regulators
indirect regulation
Gbs2018, adhesin
Gbs1087, adhesin
Gbs1895, adhesin
Gbs1195, streptokinase
Gbs651, cytolysine
Gbs1279, hydrolase


An experimental regulatory network (E. Davidson's group)
Colored boxes indicate post gastrular domains of expression genes
Mesoderm
Endoderm
Endomeso up to 20-24 hours
Maternal & early interactions
Interactions in definitive territories
Late Wnt8 signal from veg2
Post gastrular terminal or peripheral downstream genes
Mat cβ
Micr/Nuc Mat Otx
Ets
LiCl
GSK-3
frizzled
nβ-TCF
Wnt8
Endomeso
outside Endomeso
Mat Otx
Krl
SoxB1
Krox
Otx
Eve
Frz
Rep. of Wnt8 (18 hrs)
to 4th - 6th Cleavage Endo-Mes
M→V2L
Hnf 6
Mic
Repressor
Ub
Pmar1
Delta
Bts
Dri
TBr
Mat. N.
Su(H)+N
E(S) ?
Hmx
Gcm
GataC
Notch
Nrl
Bra
FoxA
GataE
Eve
Lim
UI
FoxB
Hox11/13b
Elk
Veg1 endoderm
Cyclophilin, EpHx, Ficolin, Sm37, Sm27, MSP130L
Sm30
Sm50
PMC
Msp130
SuTx
CAPK
Dpt
Pks
FvMo
Decorin
Mes
Apo bec
Kakapo
OrCT
Gelsolin
Endo16
Endo

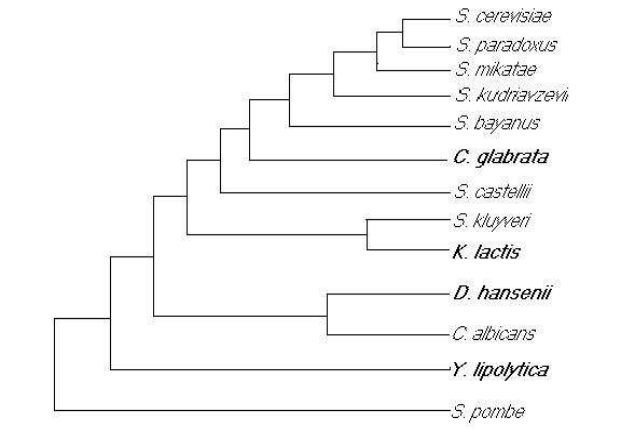
Co-evolution dynamics
(M.Z. Ludwig et al. *Nature* **403**, 564, 2000)



Co-evolutive patterns in yeast protein complexes



S. cerevisiae compared to *C. glabrata, K. lactis, D. hansenii, Y. lipolytica*, sequenced by Génolevures (B. Dujon et al. Nature 2004).



Comparative analysis

• List of putative orthologs built by bi-directional best hits to reduce the role of paralogs.

• For each organism of comparison, evolutive ranks r_i for proteins of *S. cerevisiae* obtained by ordering them in increasing order of the Blast e-value of the alignment to their best hit.

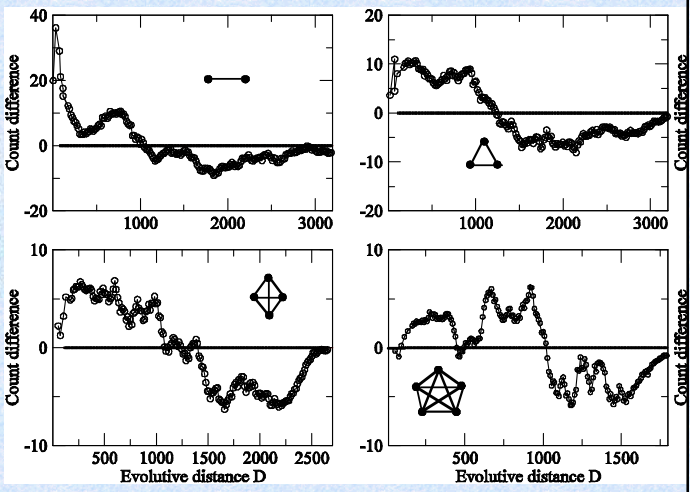
• Evolutive distance $D=\sqrt{\sum_{i<j}\left(r_i - r_j\right)^2}$ computed for cliques of proteins.

Real cliques *vs* randomized

Small D values are indicative of multi-point co-evolutive correlations.

$$D=\sqrt{\sum_{i<j}\left(r_i - r_j\right)^2}$$

r_i evolutive ranks of the n proteins composing the clique.

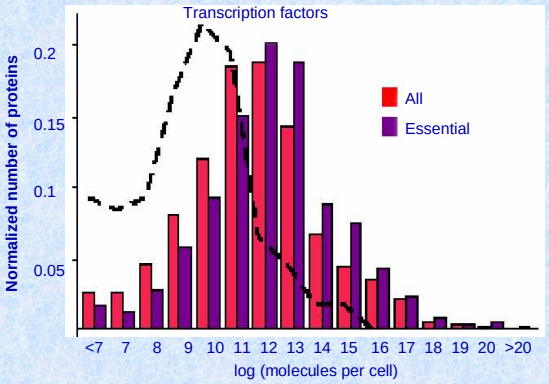


Wilcoxon tests of similarity

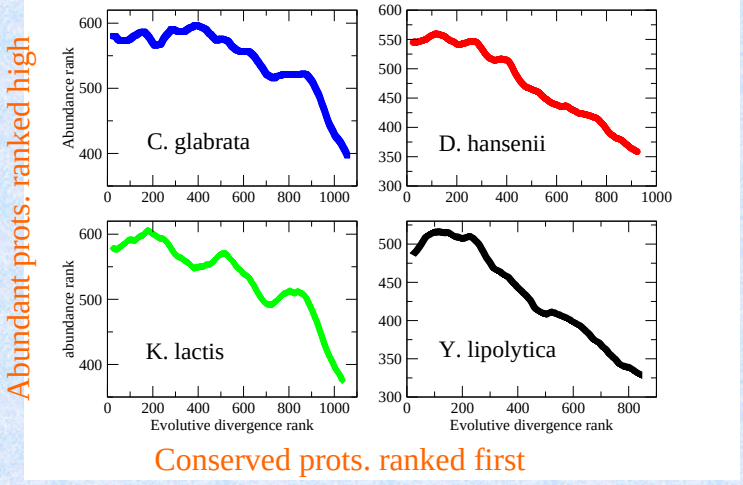
	C. glabrata				*K. lactis*			
# proteins interlinked	2	3	4	5	2	3	4	5
# groups	11006	5861	4526	3202	10652	5713	4337	3045
# proteins	3420	1451	716	434	3316	1420	711	413
z-score	8.1 ± 0.3	12.8 ± 0.2	12.1 ± 0.1	10.5 ± 0.1	7.8 ± 0.3	12.6 ± 0.2	12.3 ± 0.1	10.1 ± 0.1
probability	$2.8\ 10^{-16}$	$8.2\ 10^{-38}$	$5.3\ 10^{-34}$	$4.3\ 10^{-26}$	$3.1\ 10^{-15}$	$1.1\ 10^{-36}$	$4.5\ 10^{-35}$	$2.8\ 10^{-24}$
	D. hansenii				*Y. lipolytica*			
# proteins interlinked	2	3	4	5	2	3	4	5
# groups	8290	4656	3467	2380	6642	3592	2648	1875
# proteins	2662	1156	612	363	2248	970	514	289
z-score	8.0 ± 0.3	8.6 ± 0.1	7.4 ± 0.1	6.0 ± 0.1	6.1 ± 0.3	8.5 ± 0.1	6.2 ± 0.1	3.8 ± 0.1
probability	$6.2\ 10^{-16}$	$4.0\ 10^{-18}$	$6.8\ 10^{-14}$	$9.9\ 10^{-10}$	$5.3\ 10^{-10}$	$9.5\ 10^{-18}$	$2.8\ 10^{-10}$	$7.2\ 10^{-5}$

Interacting proteins have similar abundances

	Abundance			
# proteins interlinked	2	3	4	5
# groups	9114	4431	3207	2112
# proteins	3029	1232	596	344
z-score	8.0 ± 0.3	8.6 ± 0.1	7.4 ± 0.1	6.0 ± 0.1



S. Ghaemmaghami *et al. Nature* **425**, 737 (2003)



Same behavior found in Pál, C. *et al. Genetics* **158** 927 (2001)
for the comparison with *C. albicans* based on transcript levels.

Independence of co-evolutive and abundance correlations

Acknowledgements

U. Gaul, **E. Siggia** (Rockefeller Univ.);

N. Rajewsky (New York Univ.)

(BMC Bioinformatics, **3**:30)

P. Glaser, F. Kunst, T. Msadek, P. Trieu-Cuot

(Molecular Microb., to appear)

B. Dujon (Inst. Pasteur);

A. Vespignani (Indiana Univ.)

(Proteomics, to appear)