

Introduction to cis-regulation

Jacques.van.Helden@ulb.ac.be

Université Libre de Bruxelles, Belgique

Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)

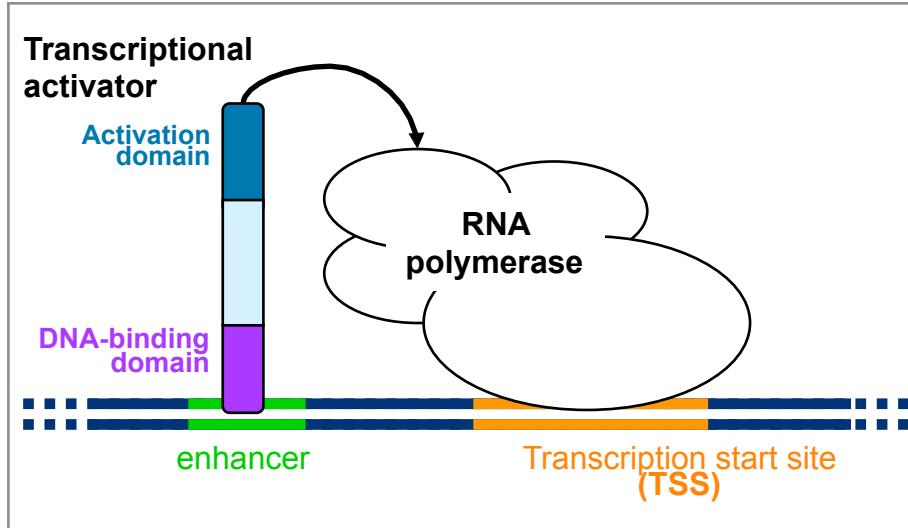
<http://www.bigré.ulb.ac.be/>

Genome sizes - some examples

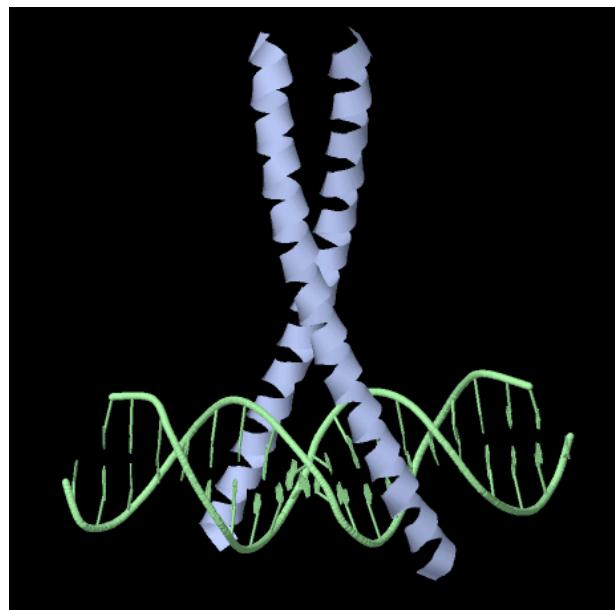
Species name	Common name	Genome completion	Genome size Mb	Number of genes	Average distance between genes Kb	Coding fraction %	Non-coding fraction %	Repeats %	Transcribed %	Remarks
Bacteria										
<i>Mycoplasma genitalium</i>	<i>Mycoplasma</i>	1995	0.6	481	1.2	90	10			Small igenome (intracellular parasite)
<i>Haemophilus influenzae</i>		1995	1.8	1 717	1.0	86	14			First sequenced bacterial genome
<i>Escherichia coli</i>	<i>Enterobacteria</i>	1997	4.6	4 289	1.1	87	13			
Yeasts										
<i>Saccharomyces cerevisiae</i>	<i>Budding yeast</i>	1996	12	6 286	1.9	72	28			First sequenced eukaryote genome
Animals										
<i>Caenorhabditis elegans</i>	<i>Nematod worm</i>	1998	97	19 000	5	27	73			First sequenced metazoan genome
<i>Drosophila melanogaster</i>	<i>Fruit fly</i>	2000	165	16 000	10	15	85			
<i>Ciona intestinalia</i>			174	14 180	12					
<i>Danio rerio</i>	<i>Zebrafish</i>		1 527	18 957	81					
<i>Xenopus laevis</i>	<i>Xenopus (amphibian)</i>		1 511	18 023	84					
<i>Gallus gallus</i>	<i>Chicken</i>		2 961	16 736	177					
<i>Ornithorhynchus anatinus</i>	<i>Platypus</i>		1 918	17 951	107					
<i>Mus musculus</i>	<i>Mouse</i>	2002	3 421	23 493	146					
<i>Pan troglodytes</i>	<i>Chimp</i>		2 929	20 829	141					
<i>Homo sapiens</i>	<i>Human</i>	2001	3 200	21 528	149	2	98	46	28	(20001=draft version)
1000 génomes humains		> 2008								Project launched January 2008
Plants										
<i>Arabidopsis thaliana</i>		2001	120	27 000	4	30	70			First plant genome
<i>Oryza sativa</i>	<i>Rice</i>		390	37 544	10					
<i>Zea mays</i>	<i>Maize</i>		2 500	50 000	50		50			Approximate number of genes
<i>Triticum aestivum</i>	<i>Wheat</i>		16 000							Hexaploid genome
<i>Lilium</i>	<i>Lilium</i>		120 000							
<i>Psilotum nudum</i>	<i>Fern-like plant</i>		250 000							

Transcriptional activation and repression

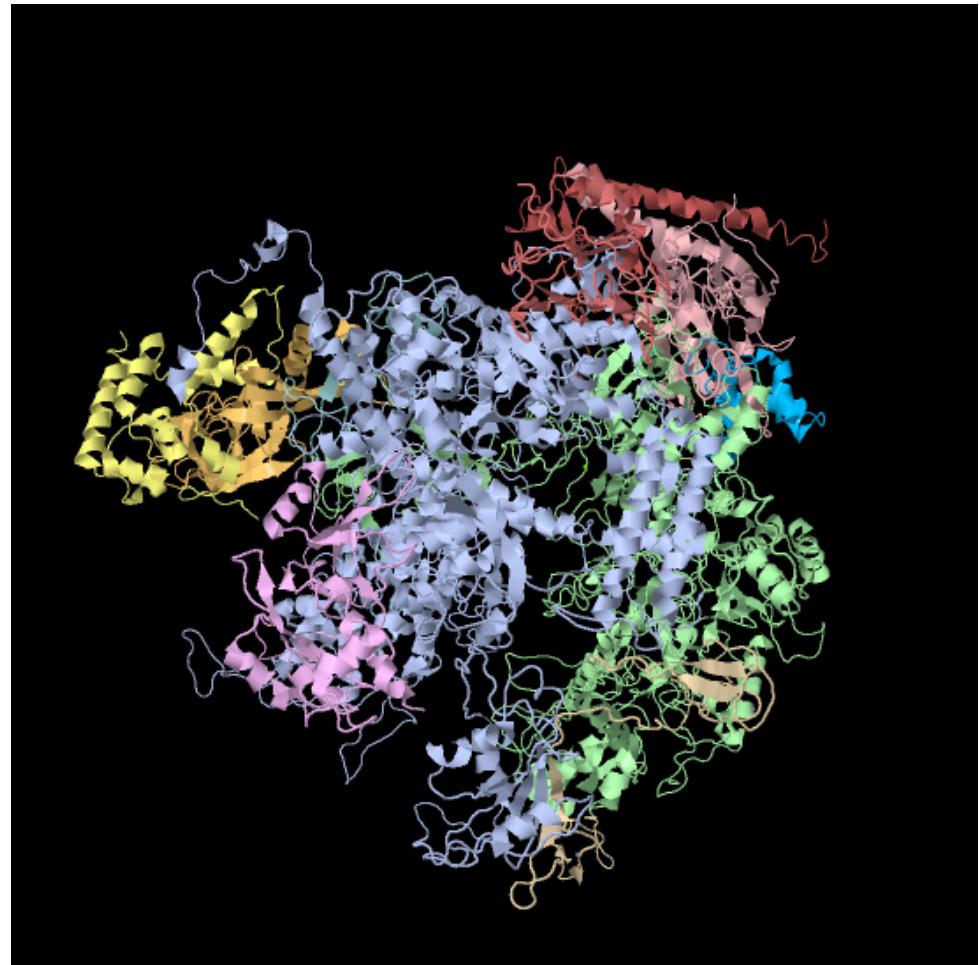
Transcriptional activation



Gcn4p from *Saccharomyces cerevisiae*
PDB 2DGC <http://www.rcsb.org/pdb/explore.do?structureId=2DGC>

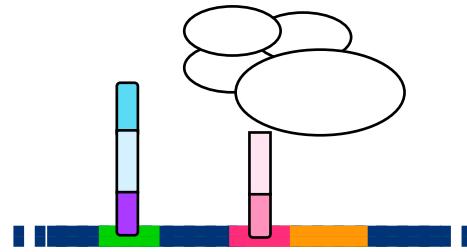


RNA polymerase II from *Schizosaccharomyces pombe*.
PDB 3H0G <http://www.rcsb.org/pdb/explore.do?structureId=3H0G>

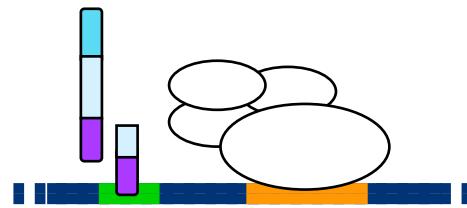


Transcriptional repression

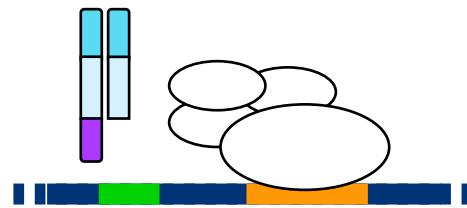
- The concept of transcriptional repression encompasses a variety of molecular mechanisms.



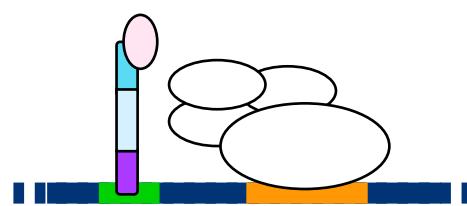
Prevent RNA polymerase from accessing DNA
(e.g. many bacterial repressors)



Competition for factor binding site
(e.g. yeast GATA factors)



Factor titration
(e.g. Drosophila Helix-loop-helix)

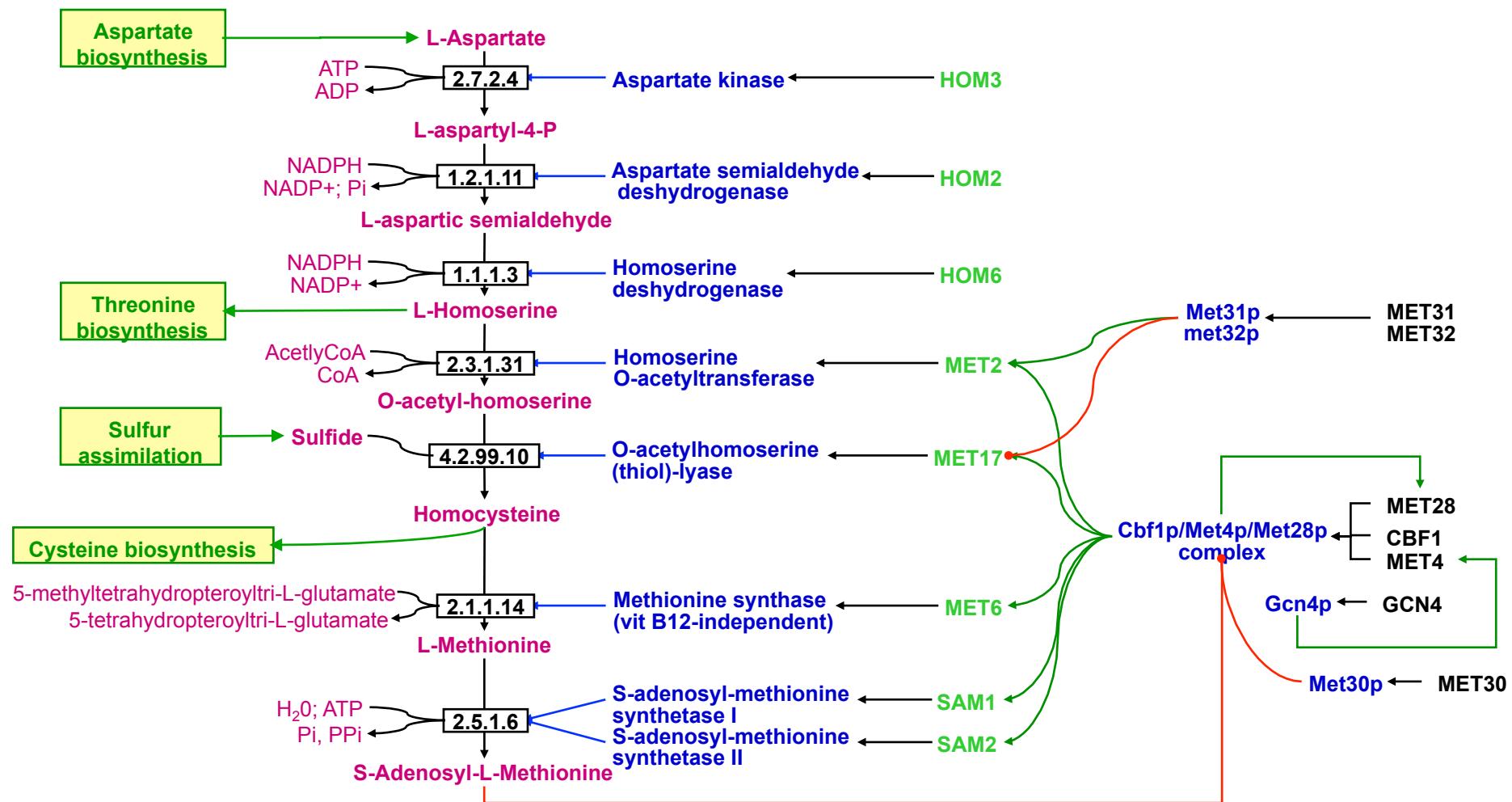


Modify transcription factor conformation -> prevent it from interacting with RNA-polymerase
(e.g. yeast Gal80p)

Cis-regulation of biological processes : some examples

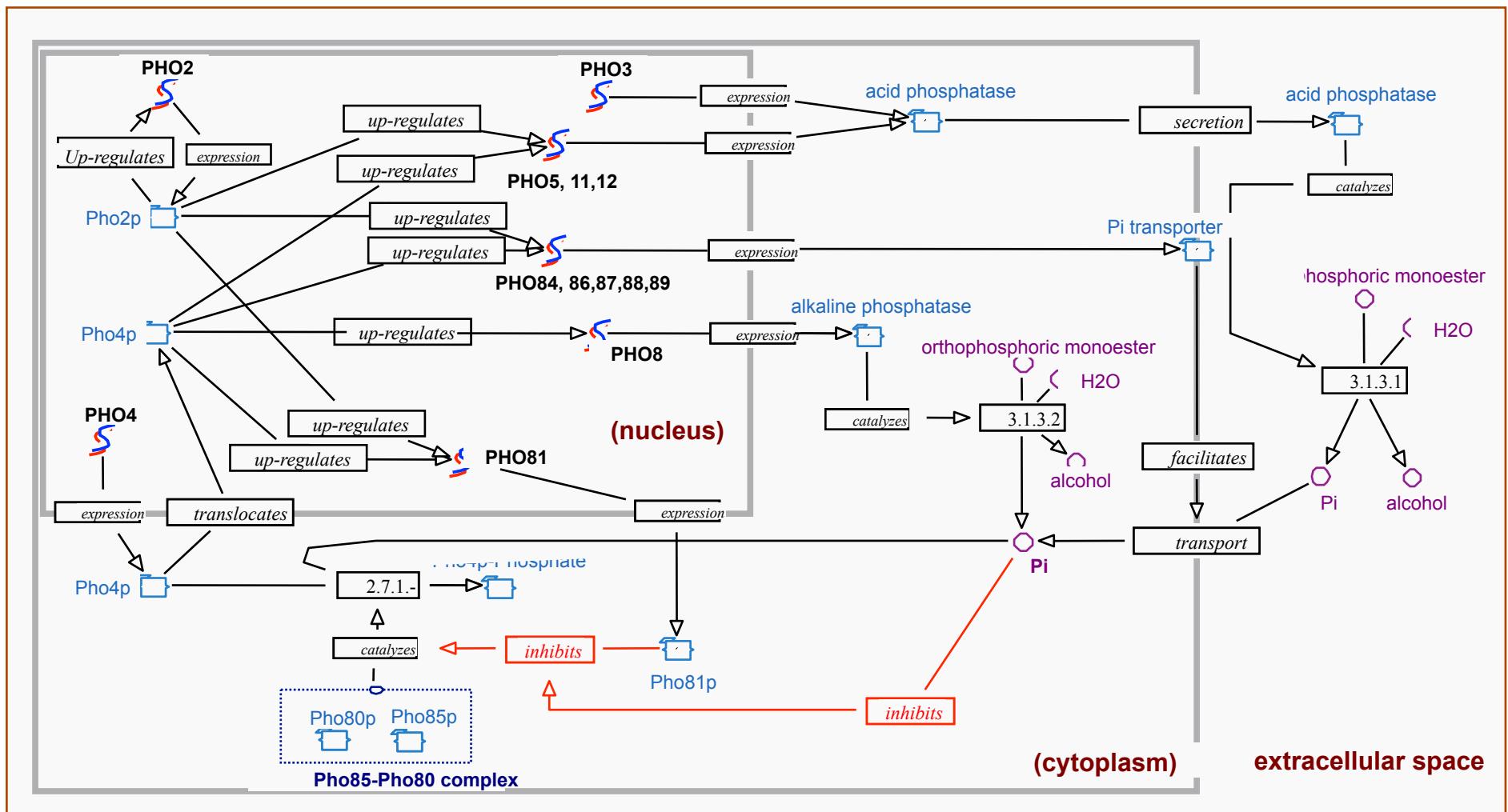
Methionine Biosynthesis in *Saccharomyces cerevisiae*

- In the budding yeast, the enzymes involved in methionine biosynthesis are cis-regulated by various transcription factors.
- Those factors are themselves trans-regulated by the end product, thereby creating a negative feed-back loop that ensures homeostasis.



Phosphate utilization in *Saccharomyces cerevisiae*

- The budding yeast responds to a phosphate stress by expressing
 - Two types of phosphatases: alkaline (Pho8p) and acid (Pho5p, Pho11p, Pho12p).
 - Several phosphate transporters (Pho84p, Pho86p, Pho87p, Pho88p, Pho89p).
 - Regulatory proteins (Pho81p) ensuring a negative feedback loop
- When Phosphate concentration is high, the transcriptional activator (Pho4p) is inactivated.

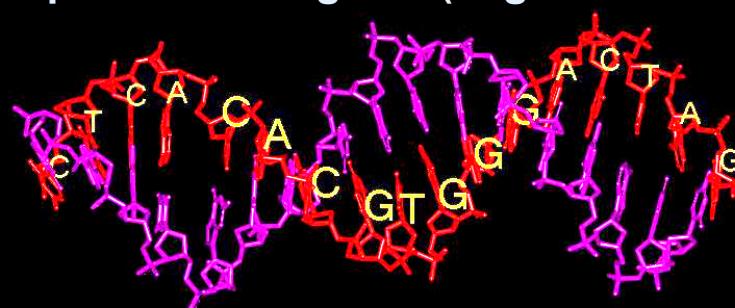


Transcription factors (TF) and their binding sites (TFBS)

Interface between the yeast Pho4p protein and one of its binding sites

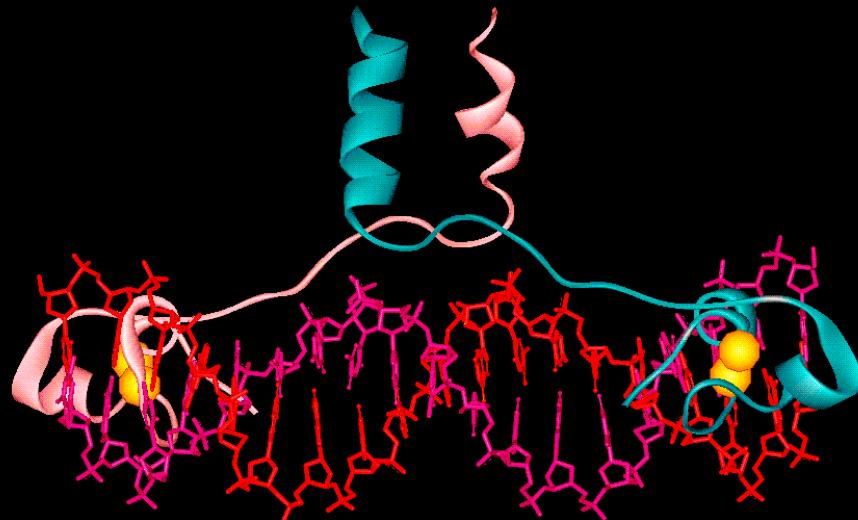


Pho4p DNA binding site (oligonucleotide)

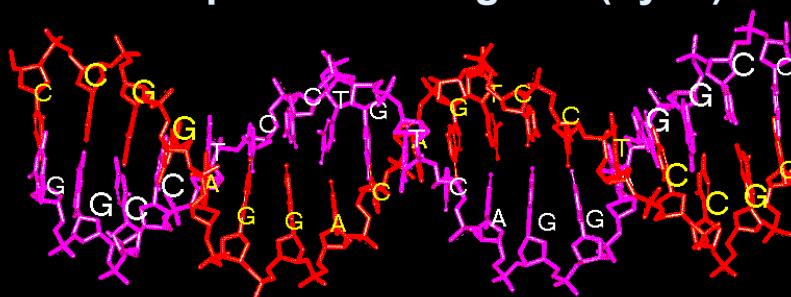


Interface between the yeast Gal4p protein and one of its binding sites

Gal4p (yeast)



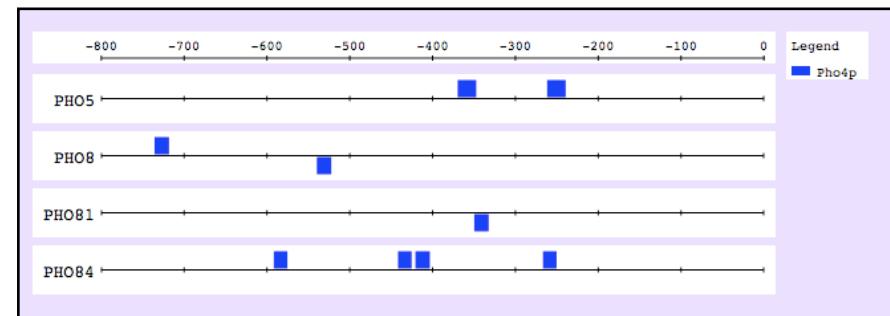
Gal4p DNA binding site (dyad)



S.cerevisiae Pho4p binding sites (TFBS)

Gene	Ft_type	Factor	Strand	left	right	Sequence
PHO5	site	Pho4p	D	-370	-347	TAAATTAGC ACGT TTTCGCATAGA
PHO5	site	Pho4p	D	-262	-239	TGGCACTCAC ACGT GGGACTAGCA
PHO8	site	Pho4p	R	-540	-522	ATCGCTG CACGT GGCCCGA
PHO8	site	Pho4p	D	-736	-718	ATATTAAAGCGTGCGGGTAA
PHO81	site	Pho4p	R	-350	-332	TTATTGC CACGT GCCATAA
PHO84	site	Pho4p	D	-592	-575	TTACG CACGT TTGGTGCTG
PHO84	site	Pho4p	D	-421	-403	TTTCCAGC ACGT GGGGCGG
PHO84	site	Pho4p	D	-442	-425	TAGTTCC CACGT GGACGTG
PHO84	site	Pho4p	DR	-879	-874	aaaagtgt CACGT Gataaaaaat
PHO84	site	Pho4p	D	-267	-250	TAATACG CACGT TTTTAA

- A *transcription factor binding site (TFBS)* is a **location** within a sequence, where a transcription factor binds specifically.
- The site is characterized by
 - a position (start, end, strand) relative to some reference (chromosome start, gene TSS, ...).
 - a sequence
- A site can be
 - experimentally proven(known site)
 - inferred by some algorithm (predicted site)
- Example
 - binding sites for the yeast transcription factor Pho4p. Coordinates are relative to the start codon.



Transcription factor binding motifs

Motif / pattern

- We use the term ***motif*** (or ***pattern***) in the sense of a model representing the specificity of binding for a transcription factor.
- A motif is generally built from a collection of transcription binding sites.
- A motif can be described using different formalisms.
 - Consensus string
 - nucleotide alphabet **CACGTGGG**
 - IUPAC alphabet **CACGTGKK**
 - regular expressions. **CACGTG[GT] [GT]**
 - Position-specific scoring matrix (PSSM)
 - Logo representation (Schneider, 1986)
 - Hidden Markov Models (HMM)

Binding specificity

- The binding specificity of Pho4p has been pretty well described (Source : Oshima et al. Gene 179, 1996; 171-177)
- High-affinity sites have the core CACGTG, followed by a few Gs or Cs
- Medium-affinity sites have the core CACGTT, followed by a few Ts.
- Some single-nucleotide mutations are sufficient to prevent the binding.

Gene	Site Name	Sequence	Affinity
PHO5	UASp2	---aCtCaCA CACGTGGG ACTAGC-	high
PHO84	Site D	---TTTCCA GCACGTGGG GCGGA--	high
PHO81	UAS	----TTATG GCACGTGCG AATAA--	high
PHO8	Proximal	GTGATCGCT GCACGTGG CCCGA---	high
group 1	consensus	----- gCACGTGgg ------	high
PHO5	UASp1	--TAAATTAG GCACGTTT TTCGC---	medium
PHO84	Site E	----AATA GCACGTTT TTAACCTA	medium
group 2	consensus	----- cgCACGTTt t-----	medium
Degenerate consensus		----- GCACGTKKk -----	high-med

IUPAC ambiguous nucleotide code		
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A or G	puRine
Y	C or T	pYrimidine
W	A or T	Weak hydrogen bonding
S	G or C	Strong hydrogen bonding
M	A or C	aMino group at common position
K	G or T	Keto group at common position
H	A, C or T	not G
B	G, C or T	not A
V	G, A, C	not T
D	G, A or T	not C
N	G, A, C or T	aNy

Non-binding sites

PHO5	UASp3	--TAATTGG CATGTGCG ATCTC--	No binding
PHO84	Site C	----ACGT CCACGTGGA ACTAT--	No binding
PHO84	Site A	----TTTAT CACGTGA CACTTTT	No binding
PHO84	Site B	----TTAC GCACGTT GGTGTG--	No binding
PHO8	Distal	---TTACCC GCACGC TTAATAT---	No binding

Consensus representation

- The TRANSFAC database contains 8 binding sites for the yeast transcription factor Pho4p
 - 5/8 contain the core of high-affinity binding sites (CACGTG)
 - 3/8 contain the core of medium-affinity binding sites (CACGTT)
- The IUPAC ambiguous nucleotide code allows to represent variable residues.
- 15 letters to represent any possible combination between the 4 nucleotides ($24 - 1 = 15$).
- This representation however gives a poor idea of the relative importance of residues.

R06098 \TCACACGTGGGA\
R06099 \GGCCACGTGCAG\
R06100 \TGACACGTGGGT\
R06102 \CAGCACGTGGGG\
R06103 \TTCCACGTGCGA\
R06104 \ACGCACGTTGGT\
R06097 \CAGCACGTTTTC\
R06101 \TACCAACGTTTTC\

Cons **nnVCACGTKDn**

IUPAC ambiguous nucleotide code

A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A or G	puRine
Y	C or T	pYrimidine
W	A or T	Weak hydrogen bonding
S	G or C	Strong hydrogen bonding
M	A or C	aMino group at common position
K	G or T	Keto group at common position
H	A, C or T	not G
B	G, C or T	not A
V	G, A, C	not T
D	G, A or T	not C
N	G, A, C or T	aNy

Building a position-specific scoring matrix from a collection of sites

Alignment of Pho4p binding sites (TRANSFAC annotations)

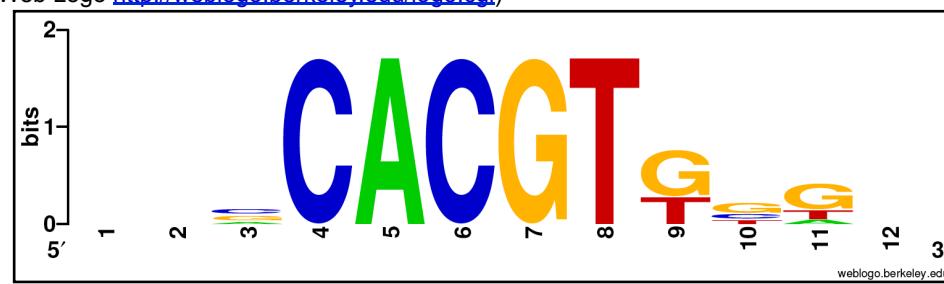
R06098	T	C	A	C	A	C	G	T	G	G	G	A
R06099	G	G	C	C	A	C	G	T	G	C	A	G
R06100	T	G	A	C	A	C	G	T	G	G	G	T
R06102	C	A	G	C	A	C	G	T	G	G	G	G
R06103	T	T	C	C	A	C	G	T	G	C	G	A
R06104	A	C	G	C	A	C	G	T	T	G	G	T
R06097	C	A	G	C	A	C	G	T	T	T	T	C
R06101	T	A	C	C	A	C	G	T	T	T	T	C

Count matrix (TRANSFAC matrix F\$PHO4_01)

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12
A	1	3	2	0	8	0	0	0	0	0	1	2
C	2	2	3	8	0	8	0	0	0	2	0	2
G	1	2	3	0	0	0	8	0	5	4	5	2
T	4	1	0	0	0	0	0	8	3	2	2	2
Sum	8	8	8	8	8	8	8	8	8	8	8	8

Tom Schneider's sequence logo

(generated with Web Logo <http://weblogo.berkeley.edu/logo.cgi>)



TRANSFAC record for the yeast PHO4 matrix (ID M00064)

AC M00064

XX

ID F\$PHO4_01

XX

DT 13.04.1995 (created); hiwi.

DT 18.07.2000 (updated); ewi.

CO Copyright (C), Biobase GmbH.

XX

NA PHO4

XX

DE PHO4

XX

BF T00690 PHO4; Species: yeast, *Saccharomyces cerevisiae*.

XX

PO	A	C	G	T	
01	1	2	1	4	N
02	3	2	2	1	N
03	2	3	3	0	V
04	0	8	0	0	C
05	8	0	0	0	A
06	0	8	0	0	C
07	0	0	8	0	G
08	0	0	0	8	T
09	0	0	5	3	K
10	0	2	4	2	B
11	1	0	5	2	G
12	2	2	2	2	N

XX

BA 8 binding sites from 4 genes

XX

CC compiled sequences

XX

RN [1]; RE0002931.

RX PUBMED: 1327757.

RA Fisher F., Goding C. R.

RT Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif

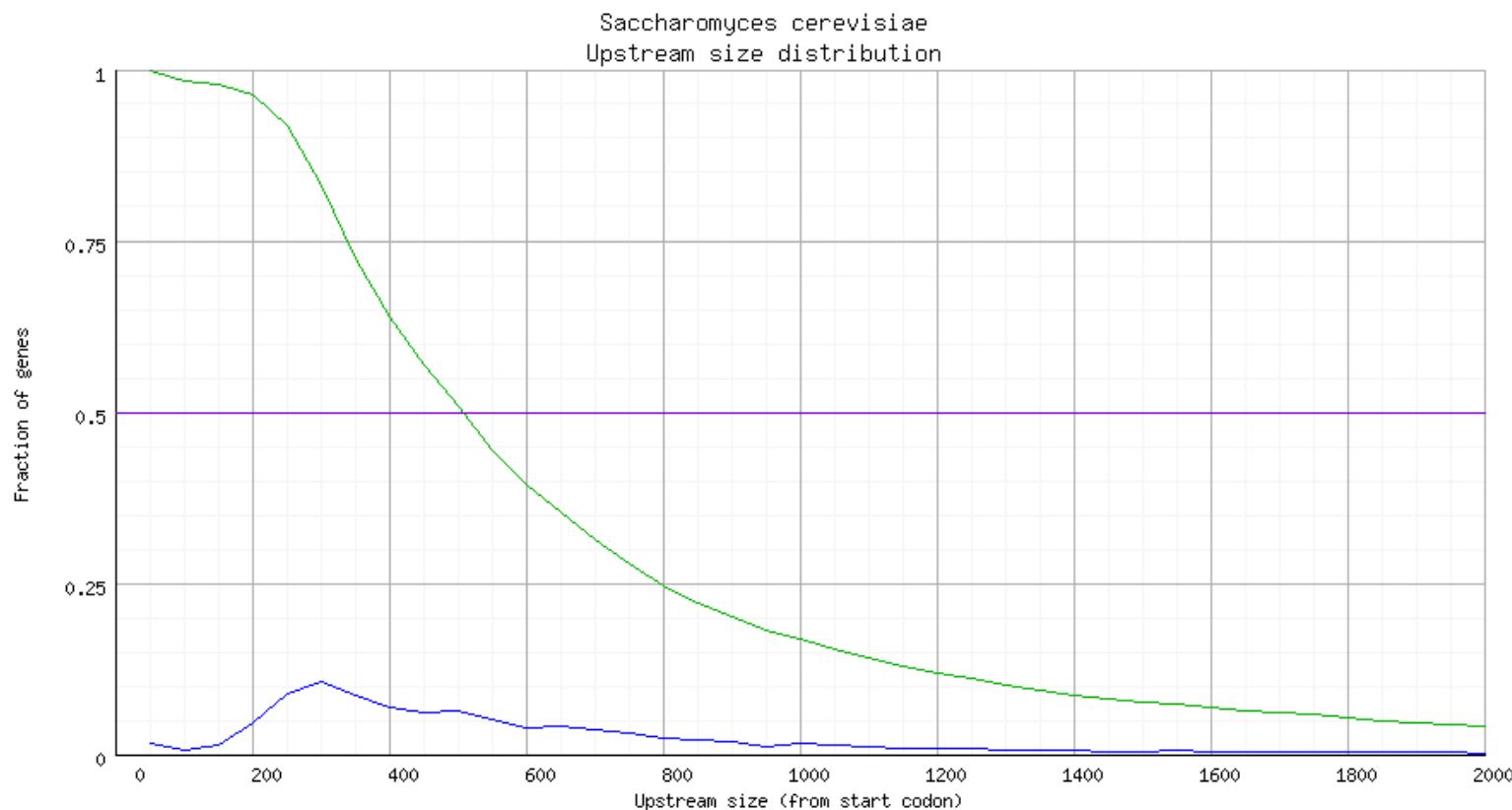
RL EMBO J. 11:4103-4109 (1992).

XX

//

Characteristics of yeast regulatory elements

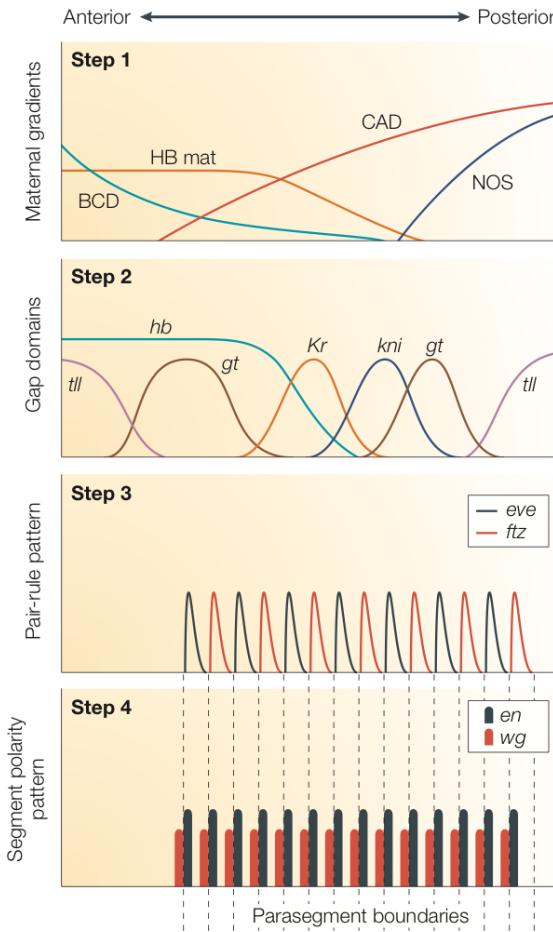
- In the yeast *Saccharomyces cerevisiae*
 - Cis-regulatory elements are located in the non-coding region upstream the regulated gene
 - Strand-insensitive
 - Activity does not depend on the strand
 - Within ~800 bp from the start codon
 - Activity does not depend on precise position



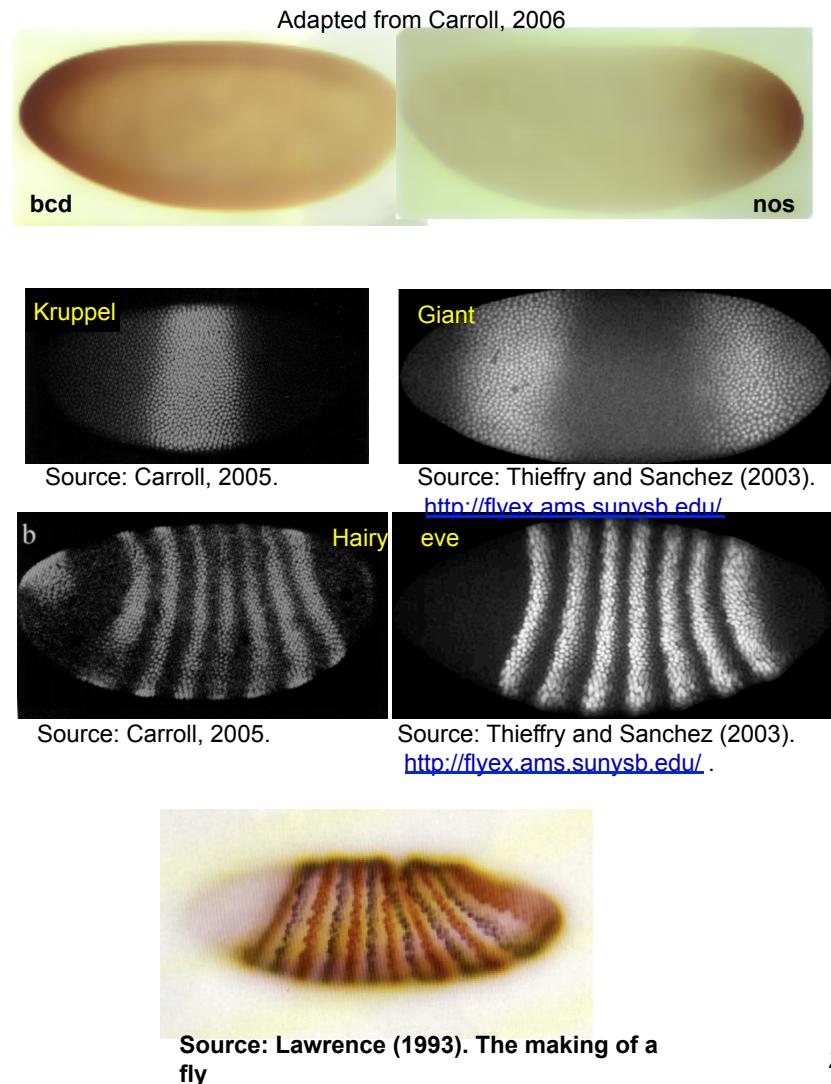
*Pattern formation and *cis*-regulatory modules (CRM)*

Drosophila Antero-Posterior (AP) segmentation – expression domains

- Establishment of expression domains
 - Maternal genes: gradients of mRNAs coding for transcription factors.
 - Gap genes: broad domains.
 - Pair-rule genes: expressed every other segment (odd or even segments).
 - Segment polarity genes: expressed with a segmental periodicity, across 2-3 cells wide bands.

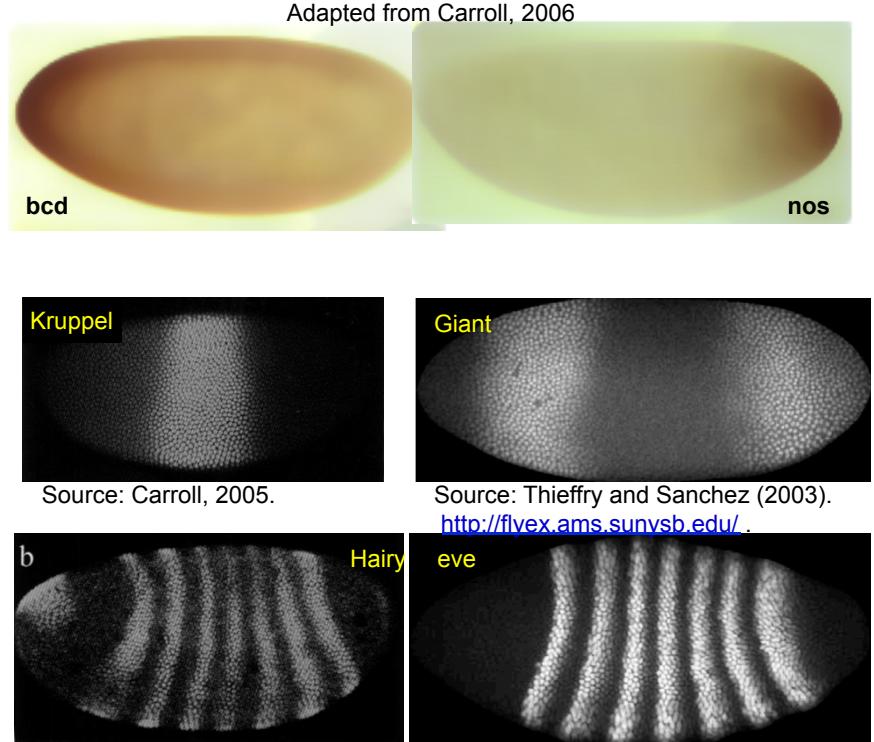
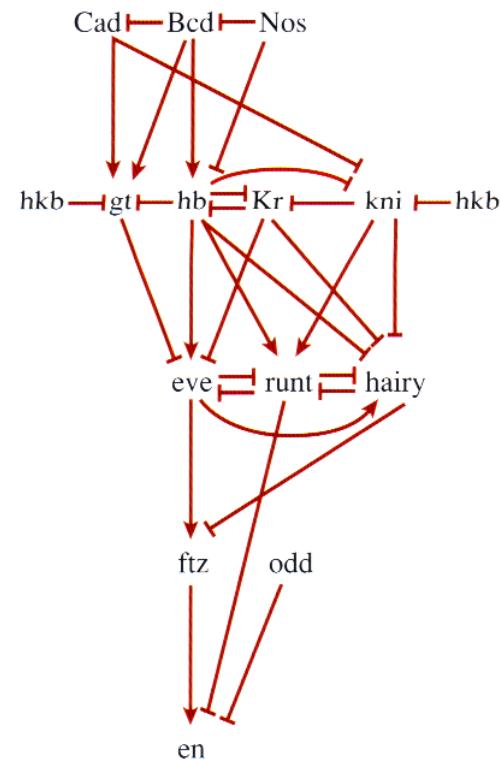
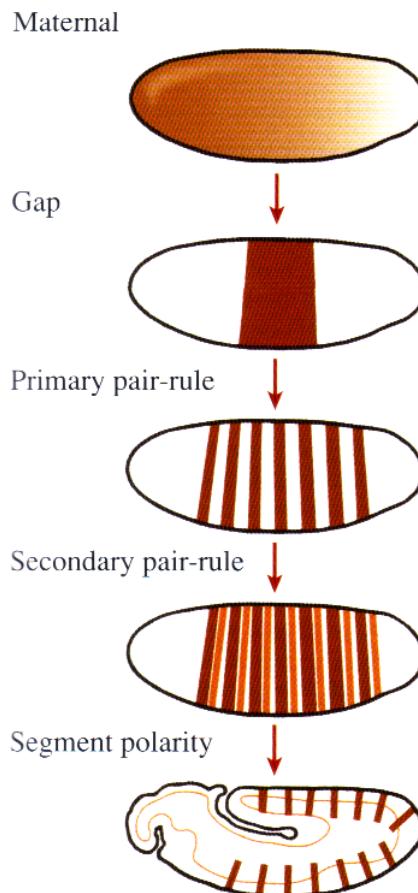


Peel et al. Arthropod segmentation: beyond the Drosophila paradigm.
NATURE REVIEWS GENETICS (2005) vol. 6 (12) pp. 905-16



Drosophila Antero-Posterior (AP) segmentation – regulatory network

- The establishment of expression domains relies on a modular network of transcriptional regulations.
- Hierarchy: Maternal genes -> Gap -> Primary pair-rule -> Secondary pair rule -> Segment polarity.



Source: Lawrence (1993). The making of a fly

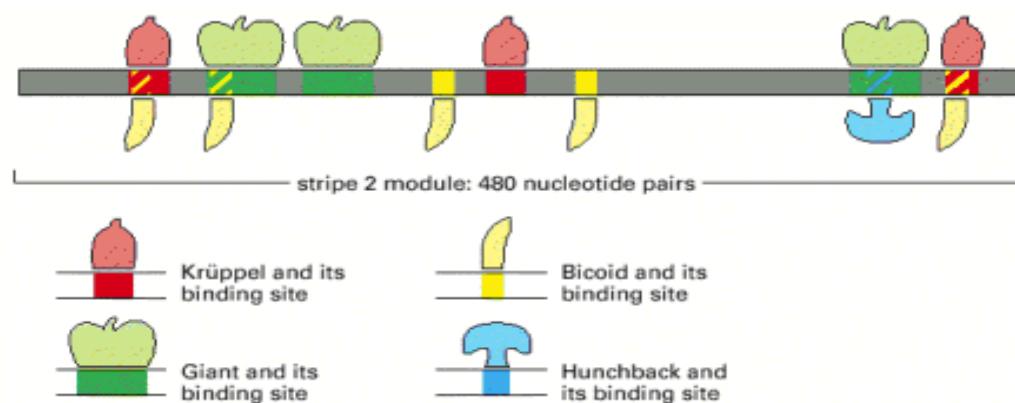
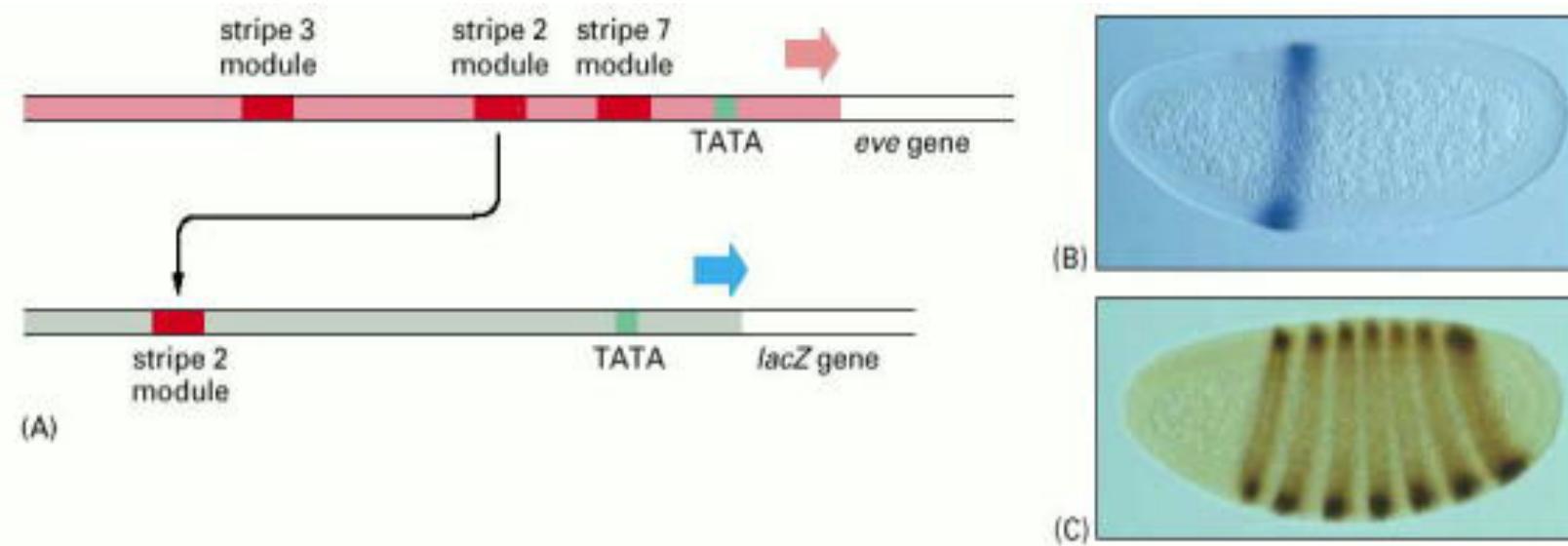
Figure 3.5
The segmentation genetic regulatory hierarchy

(left) The expression patterns of five classes of anteroposterior axis patterning genes are depicted in embryos at different stages.
(right) Selected members of these classes are shown and the regulatory interactions between these genes are indicated. An arrow indicates a positive regulatory interaction; a line crossed at its end indicates a negative repressive regulatory relationship.

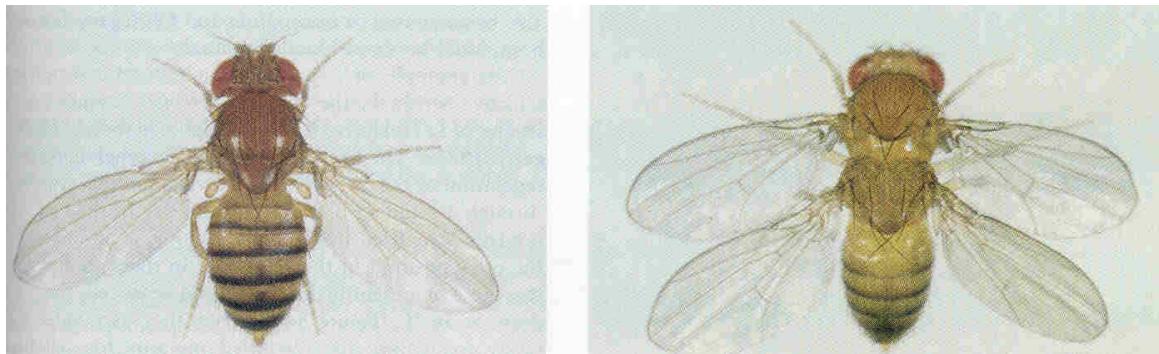
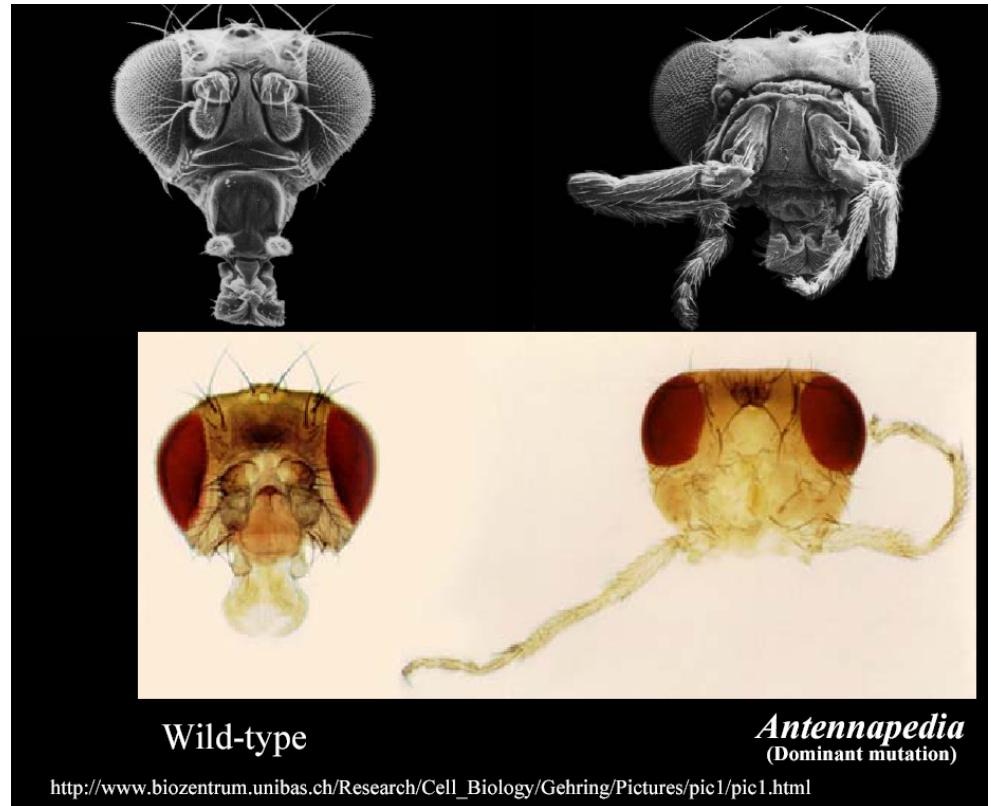
Source: Carroll, 2005. From DNA to diversity (2nd edition). Blackwell Publishing.

The stripe-specific enhancers of eve

- Each one of the 7 stripes of *even-skipped* expression is activated by a specific enhancer.
- The stripe 2 module (enhancer) contains a density of sites for Kr, bcd, Hb and Gt.



Homeotic mutations

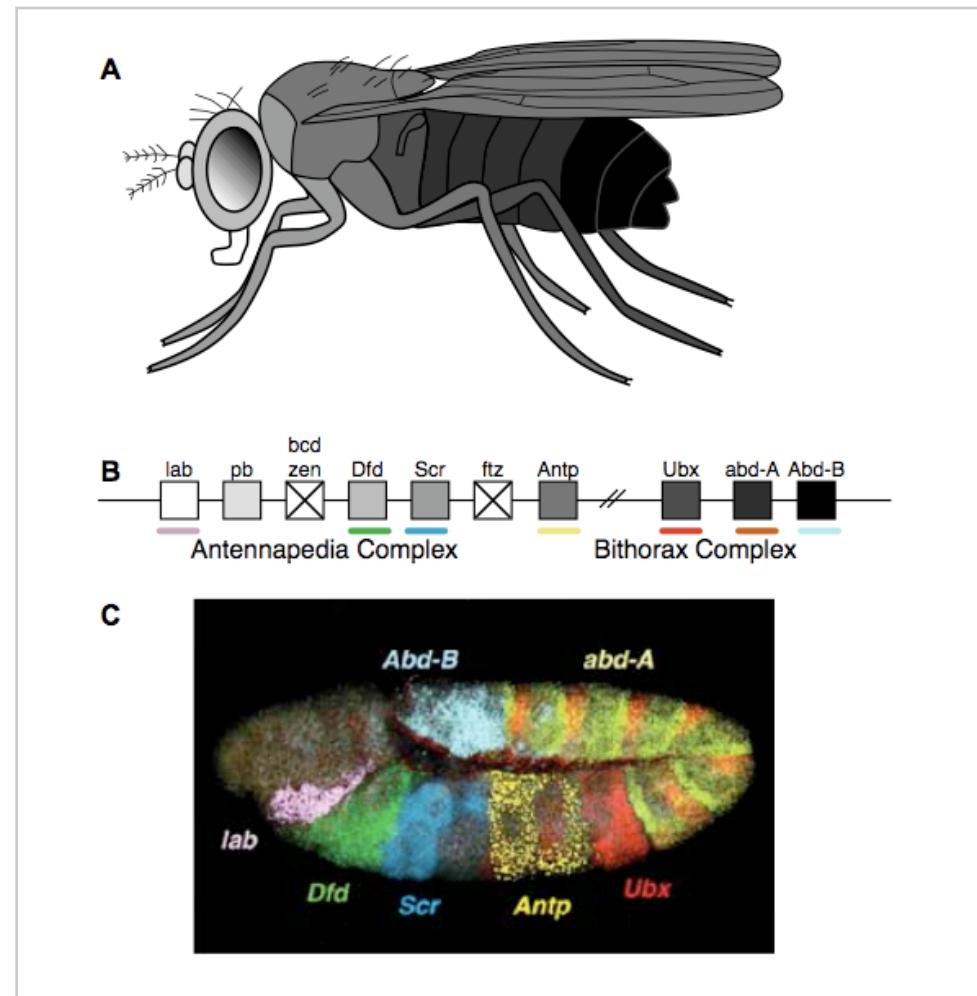


www.cb.ku.ac.th/down_load/drsermsiri.ppt

- Mutations of the Hox genes modify the segmental identity.
- *Antennapedia* mutant fly: legs develop at the location of antennae.
- *Bithorax complex* (triple mutant): the 3rd thoracic segment (metathorax) develops as a copy of the second segment (mesothorax), with wings instead of halterae.

Specification of segmental identity

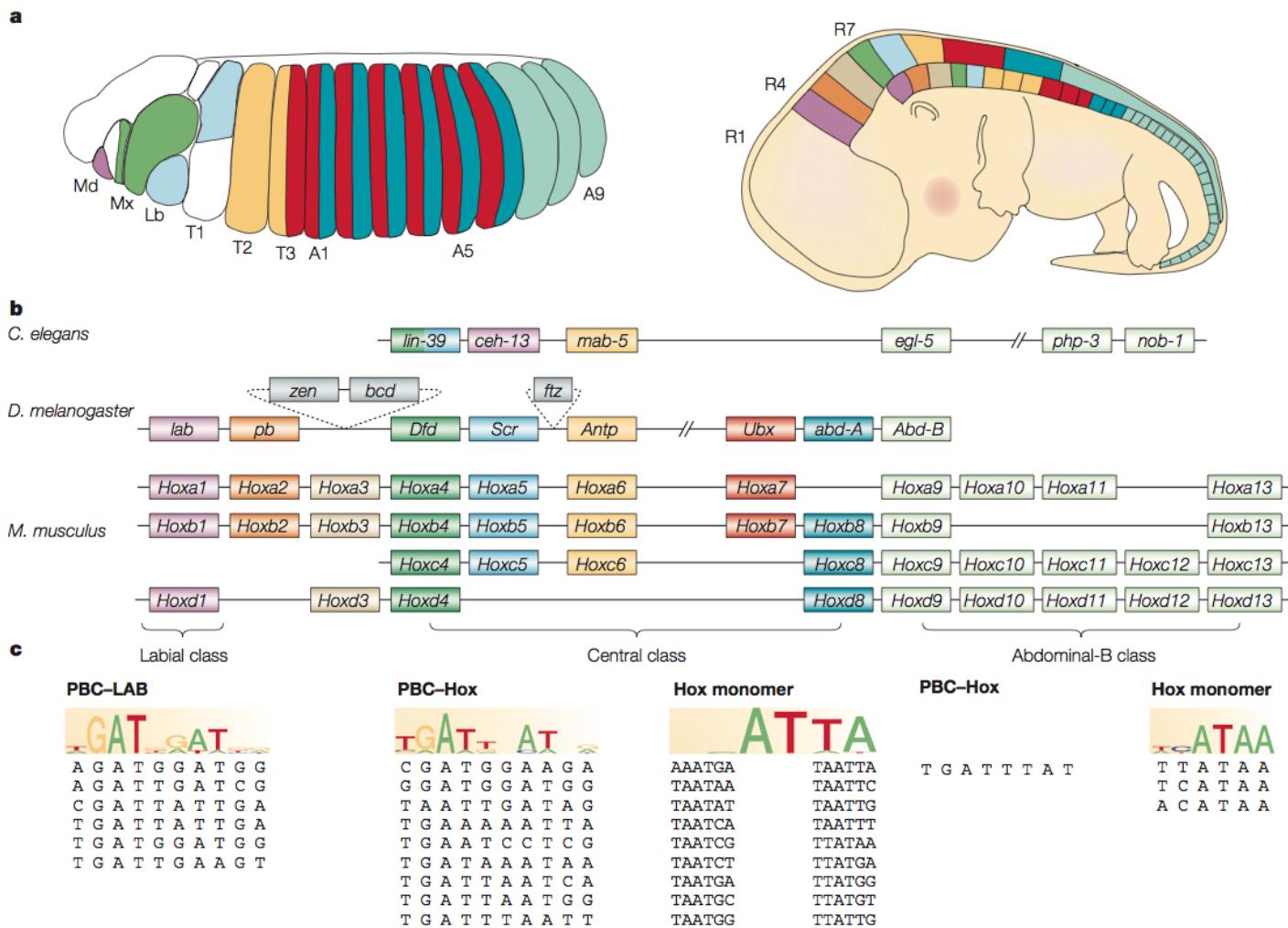
- After segmentation, each segment is committed to a particular « identity »: head, thorax, abdomen, ...
- This identify is specified by transcription factors belonging to the Hox family.
 - Bithorax complex
 - Antennapedia complex
- Each factor is expressed in a specific antero-posterior domain.



Sources de la figure:

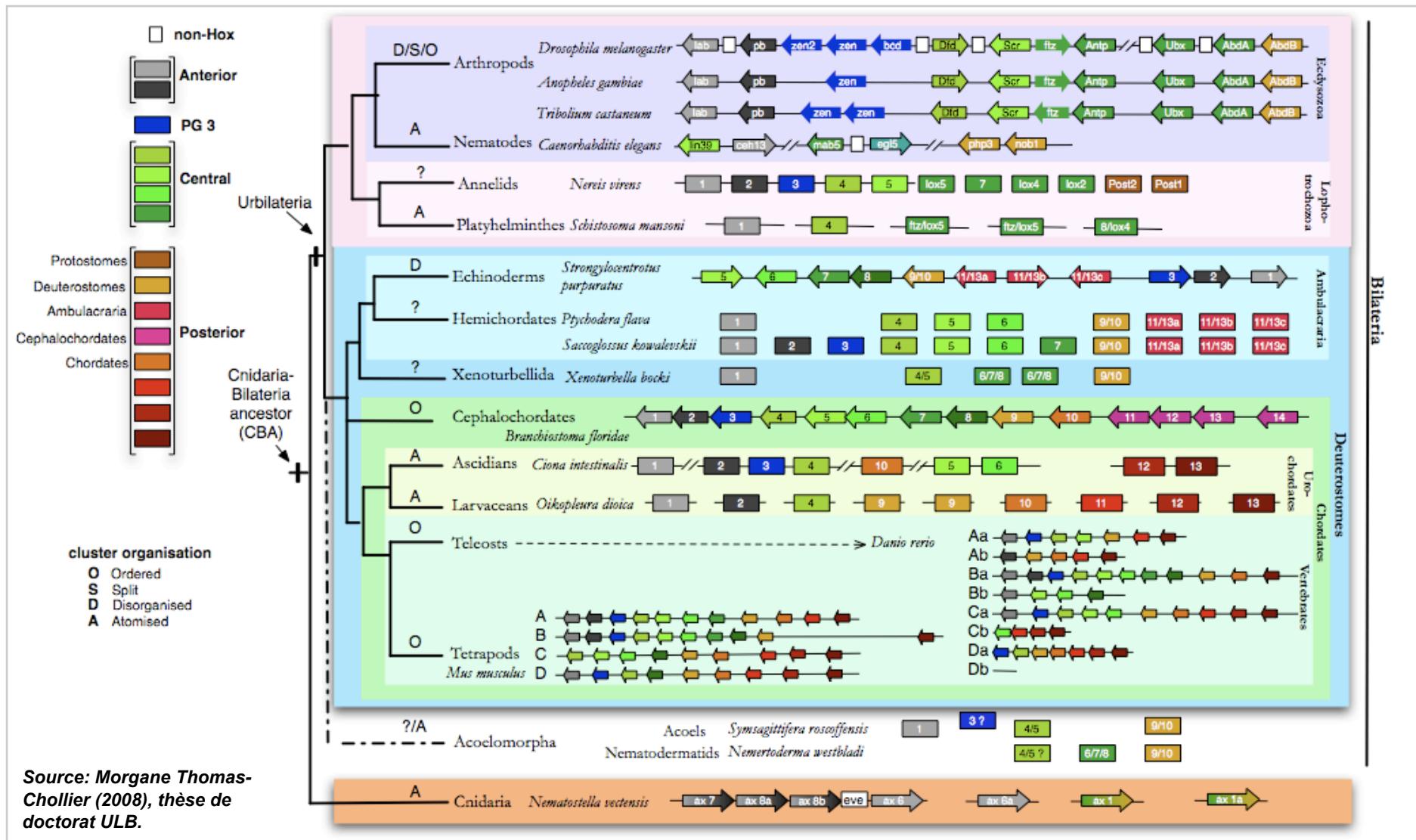
- Morgane Thomas-Chollier (2008). Thèse de doctorat ULB
- Lemons & McGinnis (2006).

The Hox complex - from *drosophila* to mammals



Hox evolution: complexification by duplication/divergence

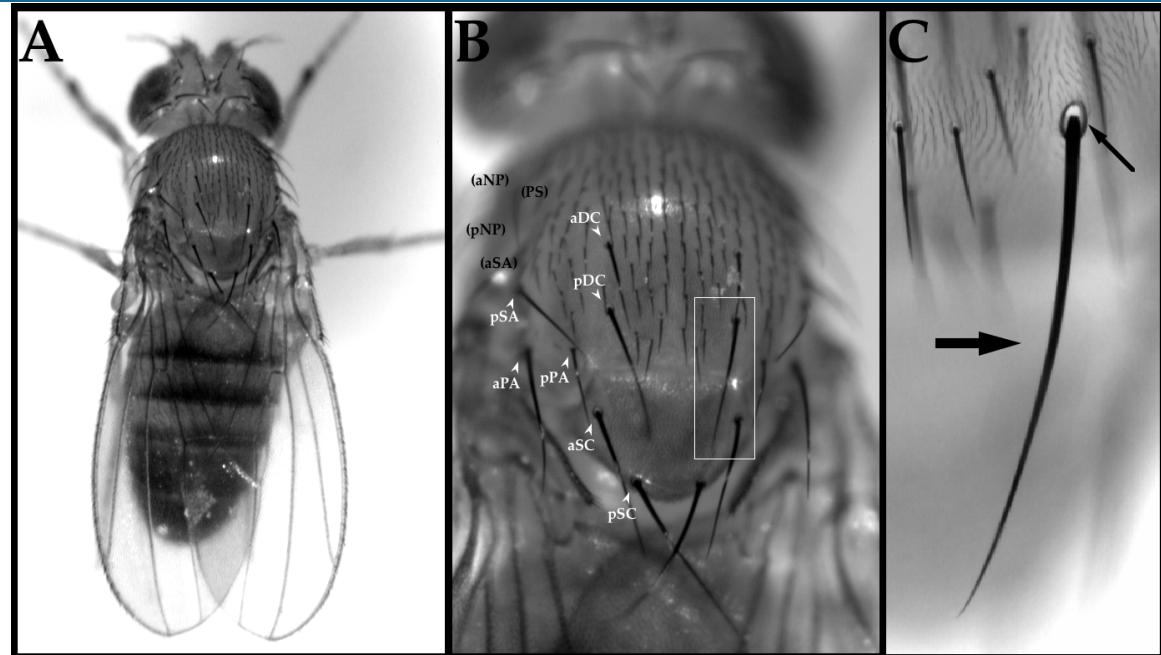
- Hox genes are found in all the Bilaterians, and they determine segmental identity.
- The topological organization of the complex has been partly conserved from invertebrate to vertebrate.
- The whole complex has been duplicated several times during evolution



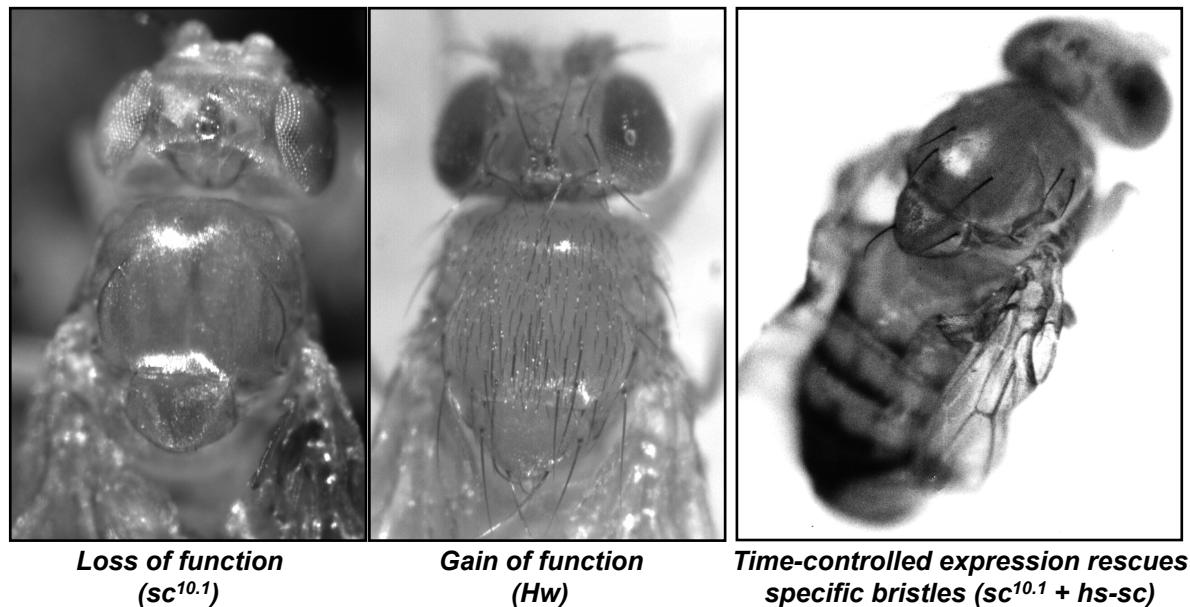
Source: Morgane Thomas-Chollier (2008), thèse de doctorat ULB.

The proneural genes in *Drosophila melanogaster*

- In *Drosophila*, sensory organs are arranged in a species-specific way, identical between individuals of the same species.
- Sensory bristles are determined by the proneural genes *achaete* and *scute*.
- **Loss of function:** *achate-scute* double mutants (*ac⁻ sc⁻*) are devoid of sensory bristles.
- **Gain of function:** an excess of *achate-scute* expression provokes the formation of ectopic bristle.
- **Rescue:** a time-controlled expression of *scute* partly rescues the *achate-scute* loss of function phenotype.

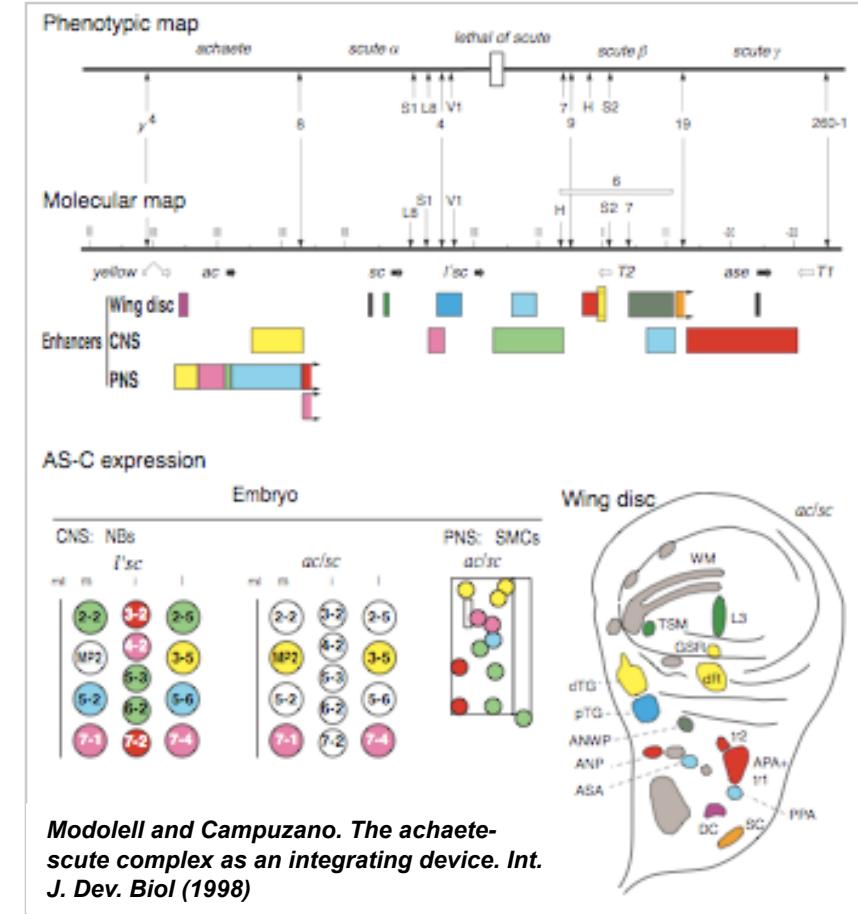
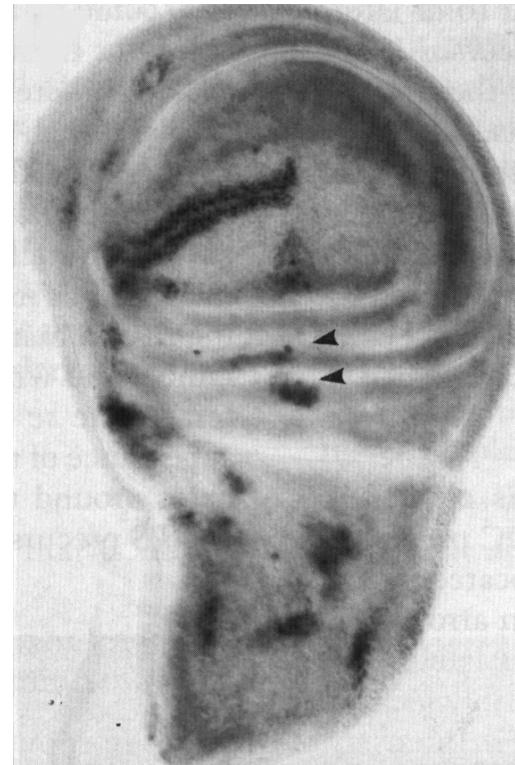


Wild type phenotype



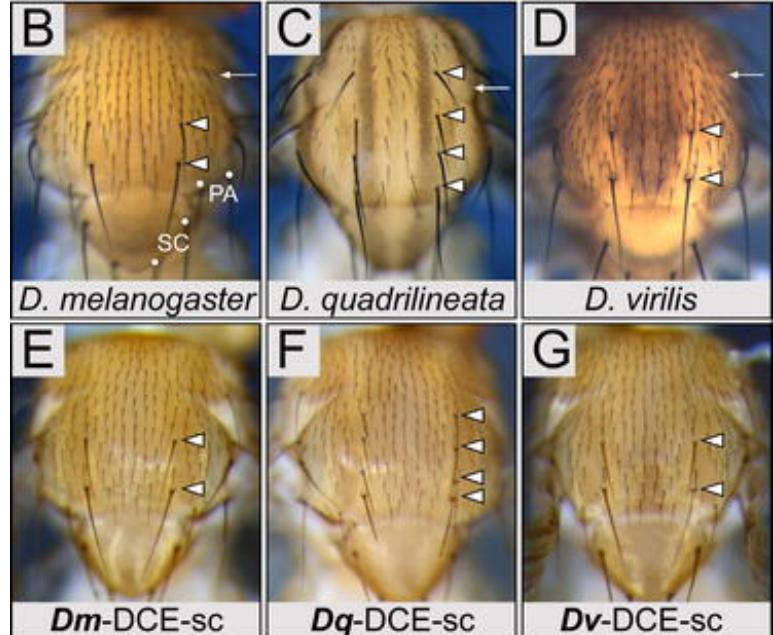
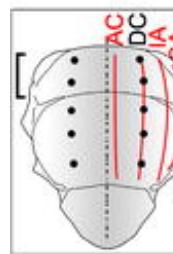
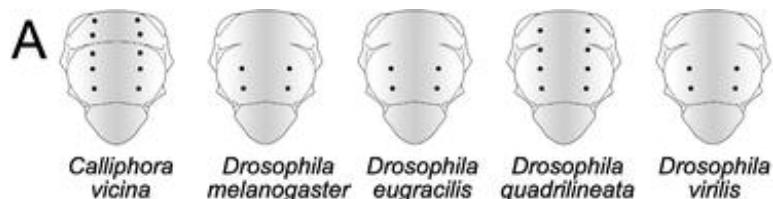
Position-specific enhancers in the achaete-scute complex

- The **achaete-scute complex (ASC)** contains 4 genes coding for paralogous transcription factors.
- Those genes are expressed in specific groups of cells (**proneural groups**) in the wing discs of the larva. A sensory organ mother cell emerges from each proneural cluster, and give rise to a bristle of the adult.
- Most of the complex is made of non-coding sequences containing **position-specific enhancers**.

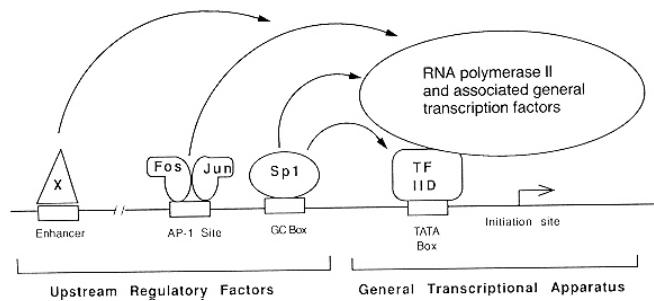


Modolell and Campuzano. The achaete-scute complex as an integrating device. Int. J. Dev. Biol (1998)

Species-specificity of the developmental patterns



Cis-regulatory modules (CRM)



- In Metazoan, some non-coding regions (typically 100-200 bp) contain closely packed binding sites for distinct transcription factors.
- These regions are called **cis-regulatory modules (CRMs)**
- CRMs play the role of integrating devices.
- Depending on the combination of transcription factors present in the cell, they will activate or repress the expression of a target gene.
 - Activation -> enhancers
 - Repression -> silencers

Cis-regulatory elements and their organization

organism	coli	yeast	metazoan
location	upstream overlap. Initiation	upstream	upstream downstream within introns
distance range	-400 to +50 bp	-800 to -1 bp	from several Kbs to several Mb !
position effect	often essential	often irrelevant	often irrelevant
strand	sensitive or symmetric	insensitive	insensitive
most common core	spaced pair of 3nt	~5-8 conserved bp	~5-8 conserved bp
repeated sites	rare	occasional	frequent
cis-regulatory modules (CRMs)			frequent

*Experimental methods for characterizing
cis-regulatory elements*

Experimental methods for characterizing transcription factor binding sites

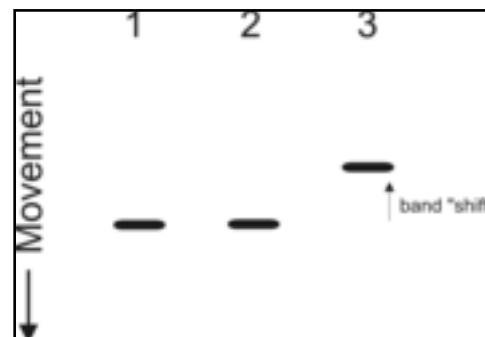
■ DNAse footprint

- Galas & Schmitz (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 30: 1851-1858.
- The residues participating in the DA-TF interface are protected from the DNase.
- Sites are characterized very precisely (typically 6-20bp)

■ EMSA

- Garner & Revzin (1981). A gel electrophoretic method for quantifying the binding of proteins to specific DNA regions: applications to components of the *Escherichia coli* lactose operon regulatory system. Nucleic Acids Res. 5: 3157-3170.
- Electrophoretic mobility shift assay (also called **gel shift**).
- Larger fragments than footprints: sometimes 50bp or more.
- The concept of “binding site” itself can be questioned.
 - Transcription factors have a higher affinity for DNA than for the nucleoplasm.
 - According to some models, they can bind anywhere on DNA, but they spend more time on some sites than on other ones.
 - One could thus consider a continuum of binding affinities.

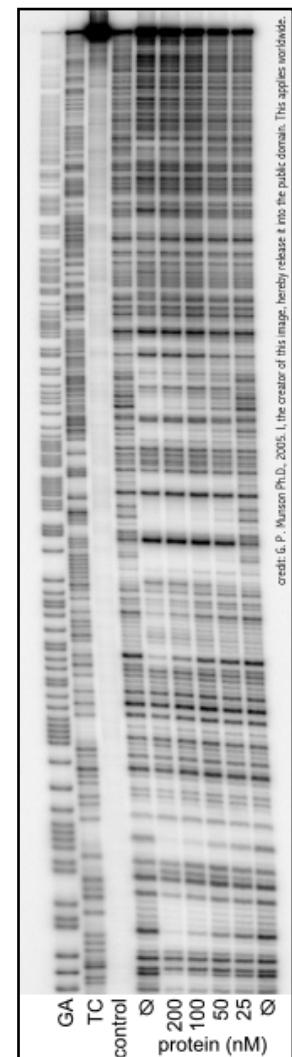
Gel shift (EMSA)



Lane 1 is a negative control, and contains only DNA. Lane 2 contains protein as well as a DNA fragment that, based on its sequence, does not interact. Lane 3 contains protein and a DNA fragment that does react; the resulting complex is larger, heavier, and slower-moving. The pattern shown in lane 3 is the one that would result if all the DNA were bound and no dissociation of complex occurred during electrophoresis. When these conditions are not met a second band might be seen in lane 3 reflecting the presence of free DNA or the dissociation of the DNA-protein complex.

Source: Wikipedia

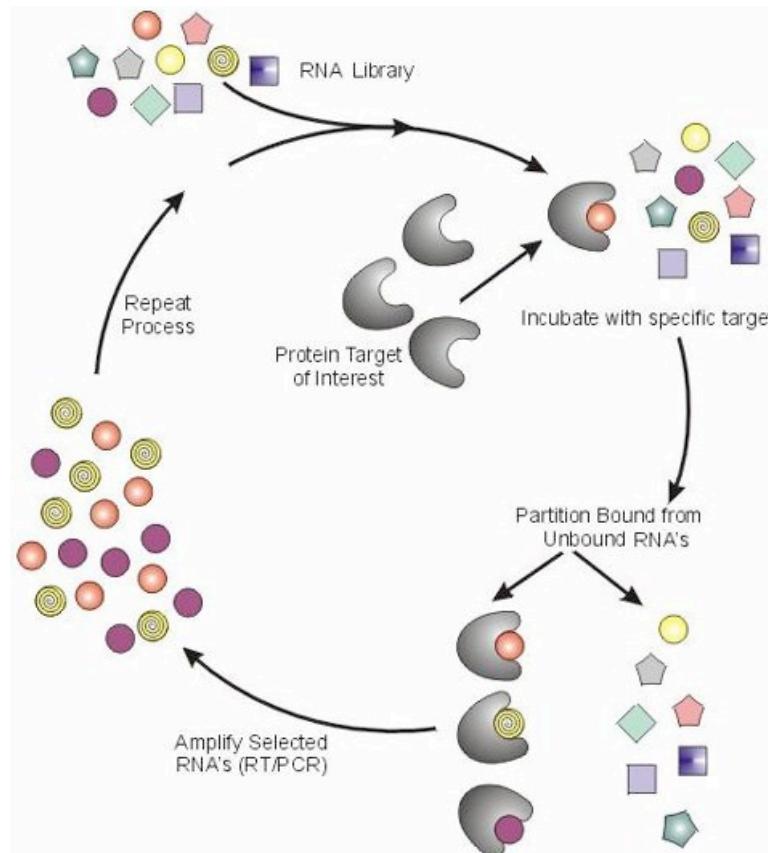
DNAse footprint



Credit: G. P. Watson Ph.D., 2005. I, the creator of this image, hereby release it into the public domain. This applies worldwide.

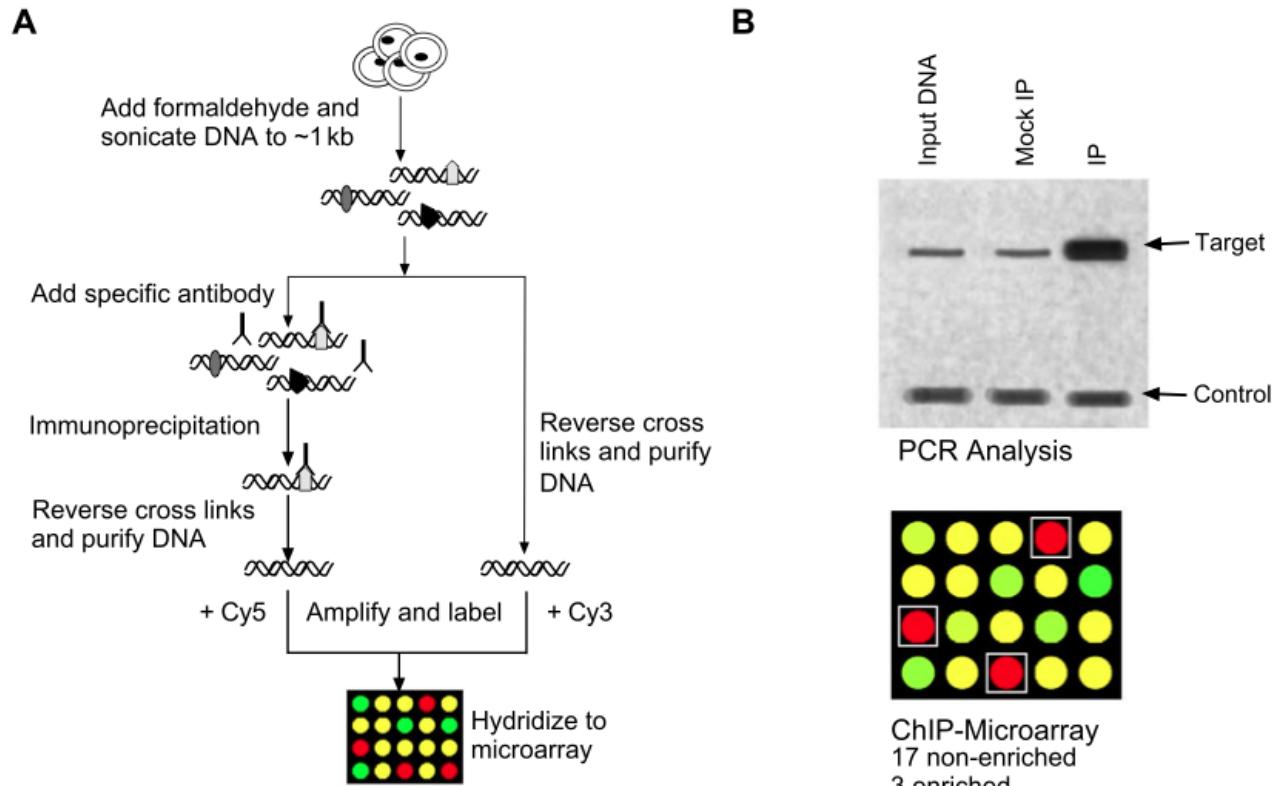
SELEX

- Systematic Evolution of Ligands by EXponential enrichment.



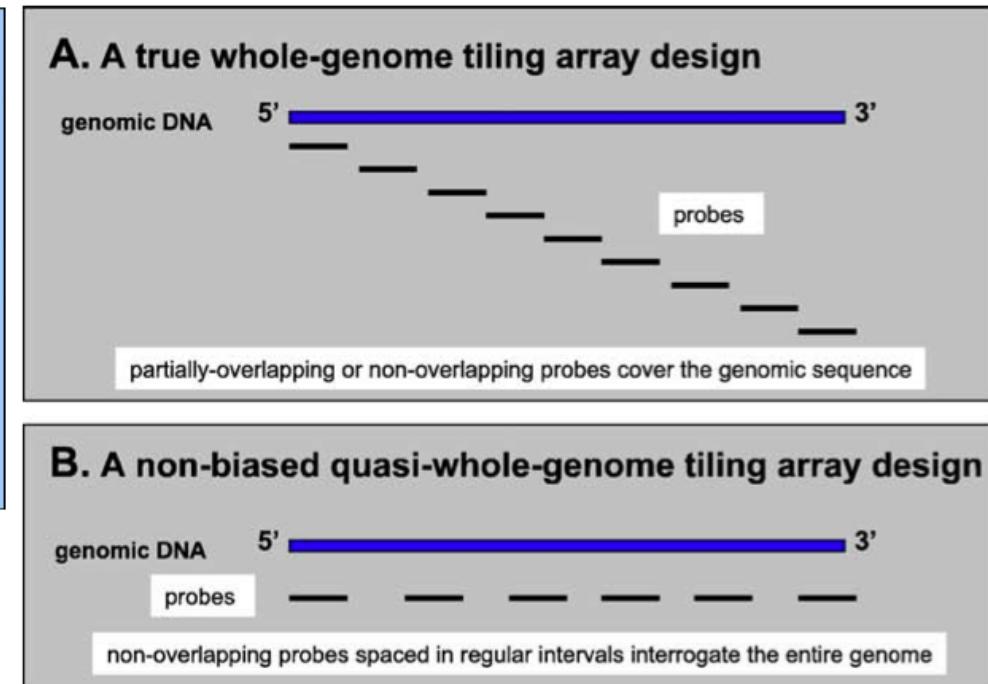
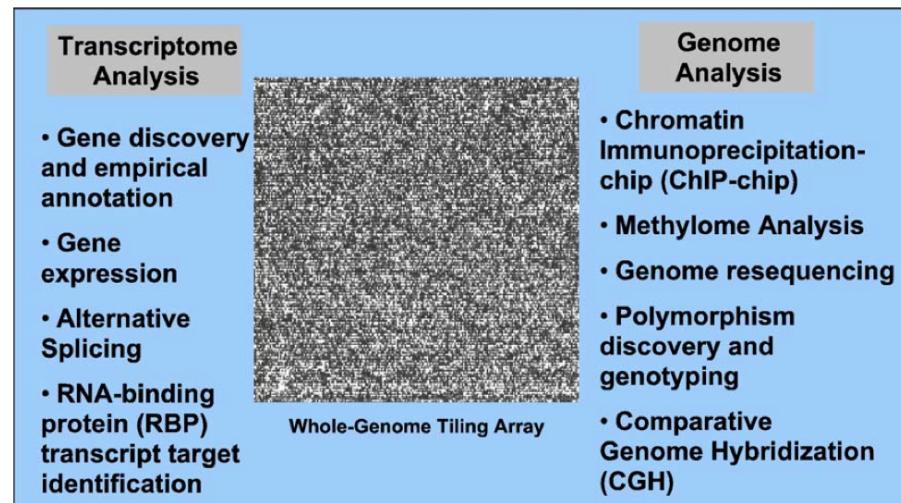
ChIP-chip

- The ChIP-chip method combines
 - Chromatin Immunoprecipitation (ChIP) to select genome fragments bound to a tagged transcription factor.
 - DNA microarrays (chip) spotted with several thousands of genome fragments (typically all the intergenic regions of a given organism) are used to detect the relative enrichment: immunoprecipitated (IP) versus non-precipitated DNA (« mock » IP).
- Strength: genome-wide coverage
- Weakness: fragmentation by sonication -> large variations in DNA fragment sizes (from a few tens of bases to several kbs).



Tiling arrays

- Tiling arrays cover the entirety of a genome, without pre-selection of any particular sequence type (intergenic, coding).
- Can be used to obtain high-coverage mapping of TF binding sites with the ChIP-chip method.
- Number of sequence fragments per array: between 10,000 and 6,000,000.



ChIP-seq

- Combination of
 - Chromatin Immunoprecipitation (ChIP), as for ChIP-chip.
 - Instead of using microarrays, the immunoprecipitated fragments are sequenced
- Strength:
 - no problem of imprecision due to the hybridation of large IP fragments to short spotted features.
 - Thanks to the « next » generation sequencing (NGC) methods, sequencing can be very efficient.
 - Does not require prior sequencing of the genome.
- Weaknesses
 - Variability of fragment sizes obtained by ultrasonication.

Universal protein-binding microarrays

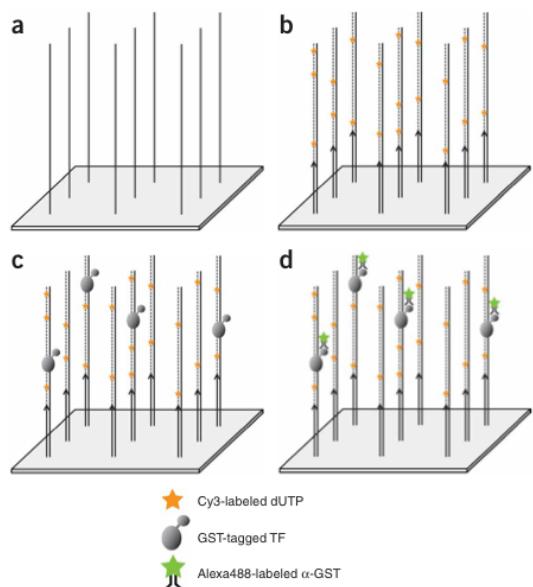
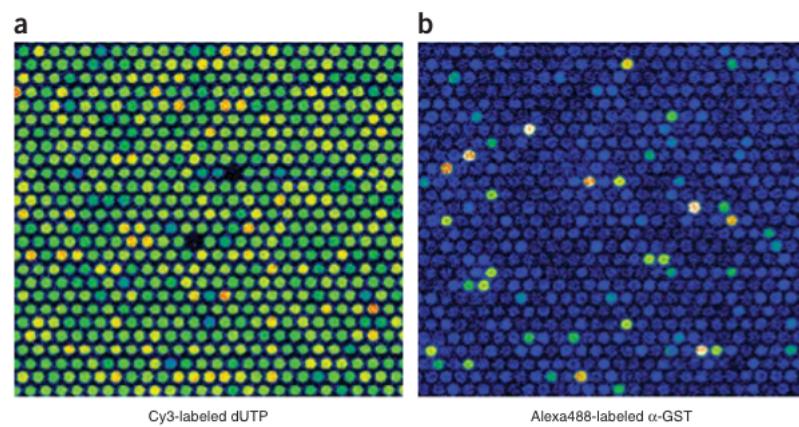


Figure 1 | Schematic of universal PBM experiments. (a) A commercially synthesized single-stranded DNA microarray is double-stranded by (b) solid-phase primer extension using a small amount of spiked-in fluorescently labeled dUTP. (c) An epitope-tagged TF is bound directly to the DNA on the microarray, and the (d) protein-bound array is labeled with a fluorophore-conjugated antibody.

Figure 3 | Zoom-in of a universal PBM scan.
(a) Region of a single subgrid, consisting of just over 1% of the total slide area, scanned to detect relative DNA amounts, as indicated by Cy3-labeled dUTP. (b) The same region of the same microarray, scanned with a different laser to detect protein binding, as indicated by Alexa 488-labeled anti-GST antibody. Intensities are shown in false color, with white indicating saturated signal intensity, yellow indicating high signal intensity, green indicating moderate signal intensity and blue indicating low signal intensity.



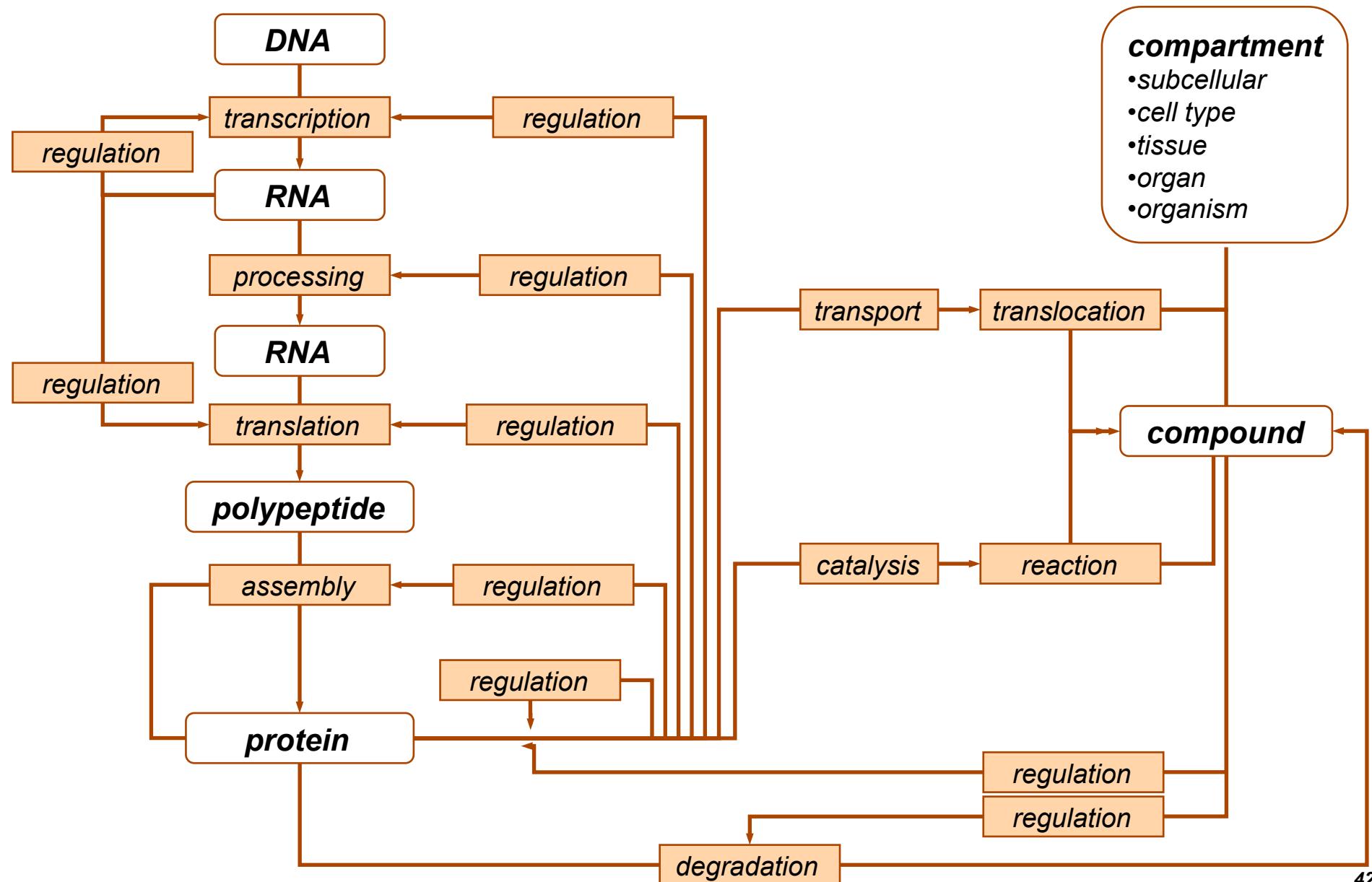
Regulatory Sequence Analysis

Questions and approaches

Questions and approaches

- Pattern matching
 - If we know the consensus for a given transcription factor, can we predict its binding sites in a DNA sequence ?
- Matching a library of patterns
 - Can we scan a sequence for matches with the consensus of all he currently known transcription factor ?
- Pattern discovery
 - Starting from a set of co-regulated genes, can we predict cis-acting elements involved in their transcriptional regulation ?
- Phylogenetic footprinting
 - Can we detect regulatory signals by searching conserved elements in non-coding sequences of orthologous genes ?
- Network inference
 - Can we infer groups networks of regulation from cis-regulatory elements ?
- Gene classification on the basis of pattern scores
 - Can we classify genes on the basis of the presence of regulatory motifs in their regulatory regions ?
 - Unsupervised classification (clustering): regroup elements (genes) in clusters without a priori knowledge about these clusters. The clusters are “discoverd” during the clustering process.
 - Supervised classification: use pre-defined groups of genes (training sets) to train a program, and then use this programs to assign new elements (genes) to one of the pre-defined groups.

Molecular networks (shamefully simplified)

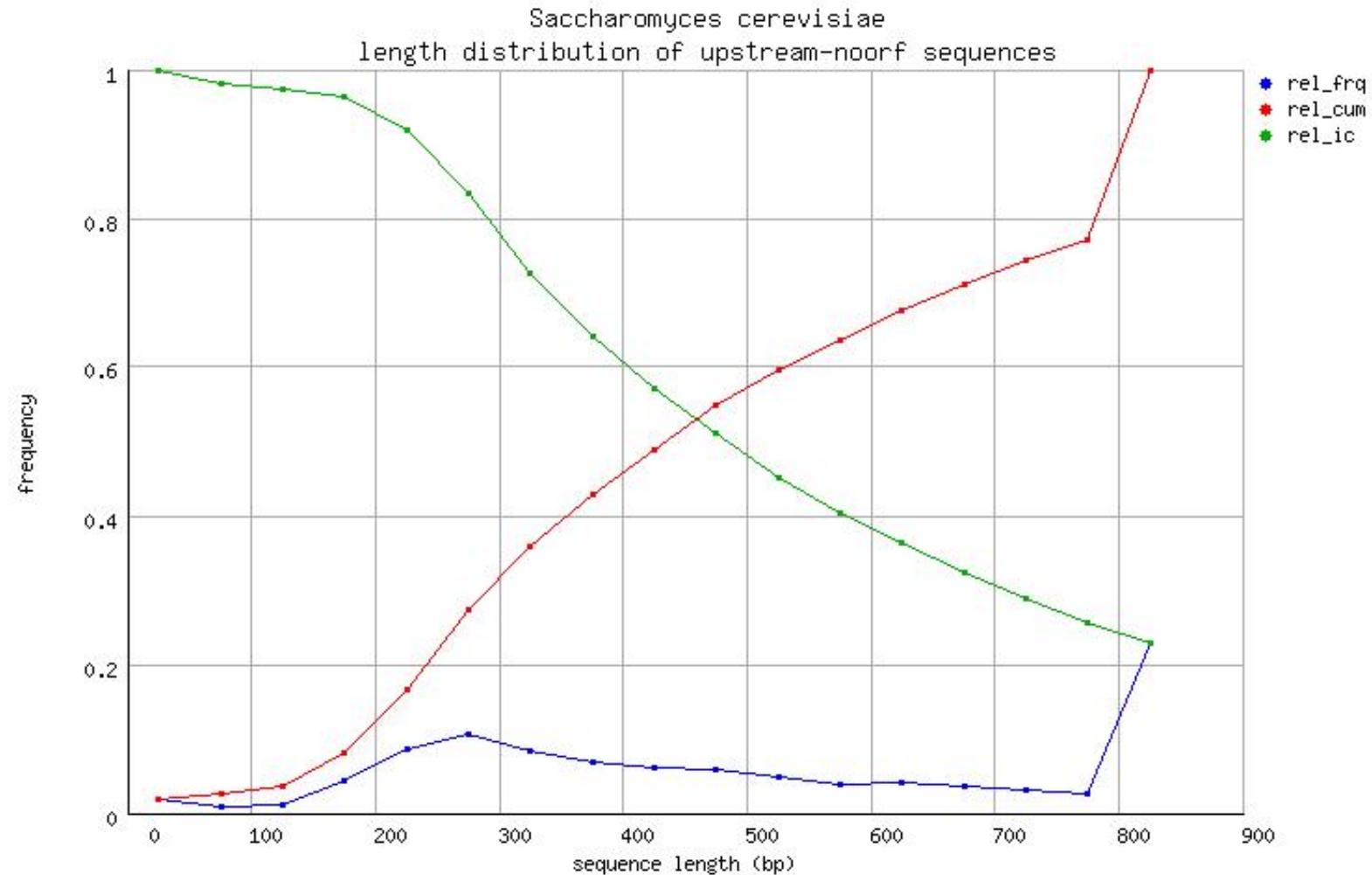


Regulatory Sequence Analysis

Supplementary material

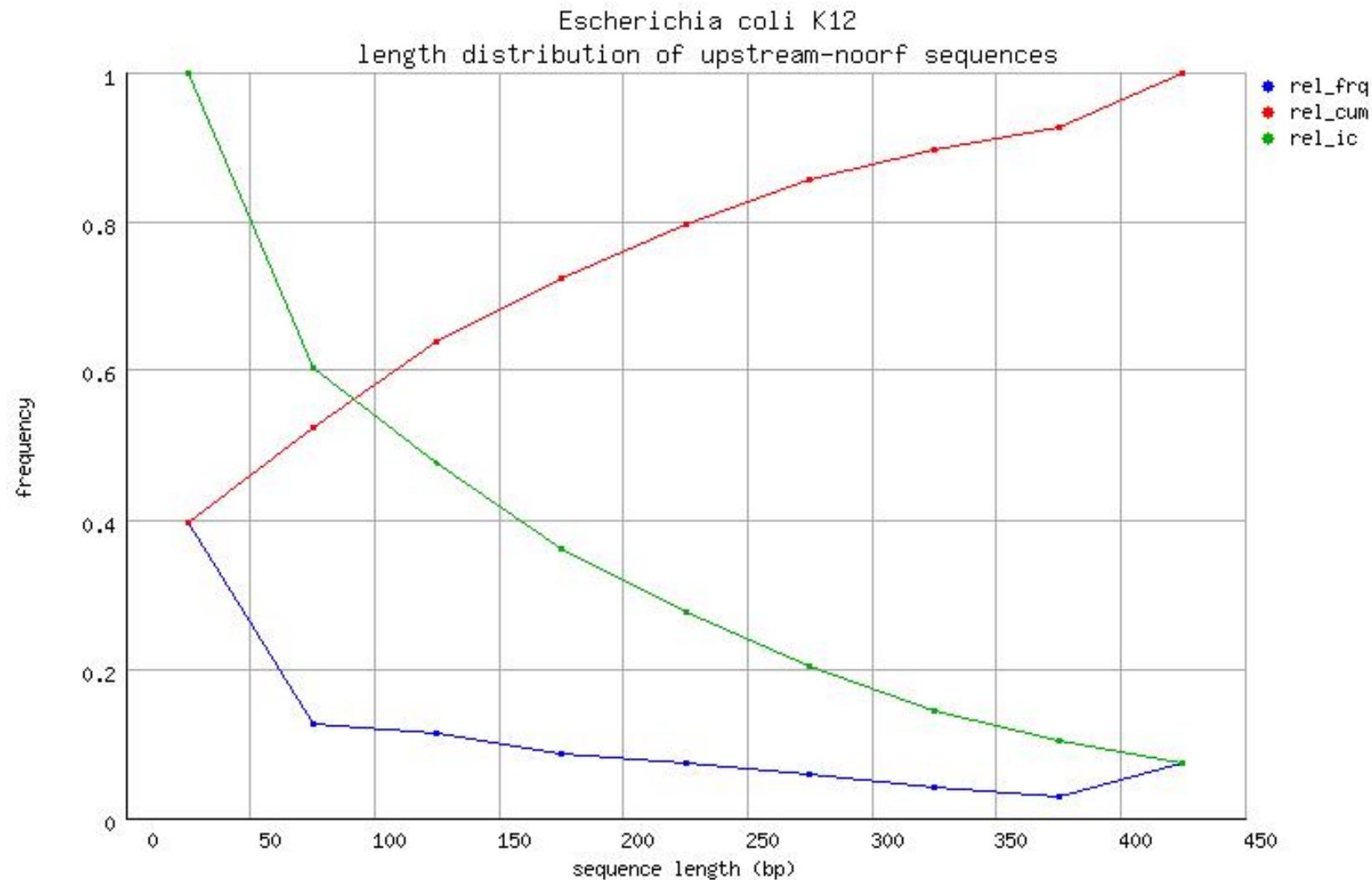
Distribution of upstream sequence lengths

Saccharomyces cerevisiae

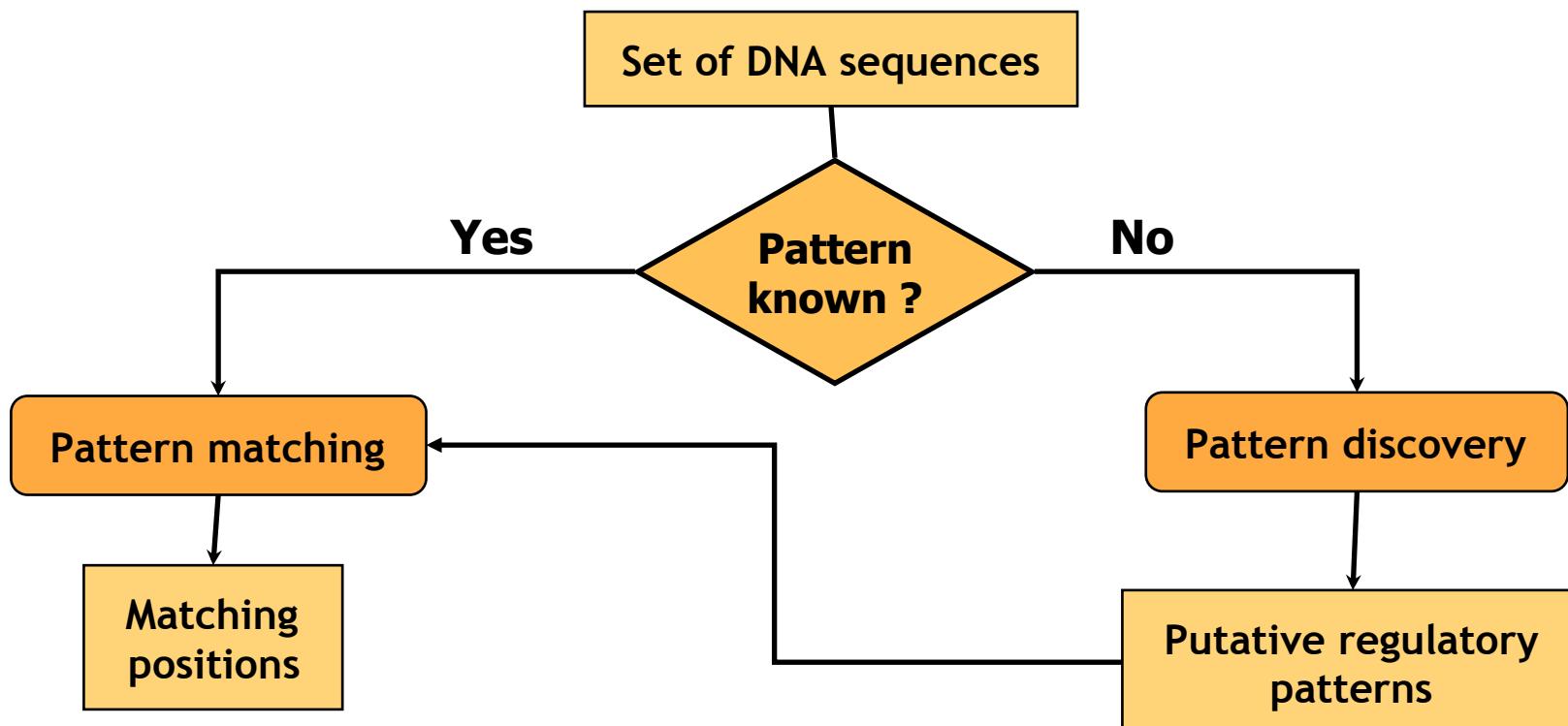


Distribution of upstream sequence lengths

Escherichia coli K12



Pattern matching vs pattern discovery



Other examples of sequence logos

A T G T A T G G

Rap1

G G T G G C A A A

Rpn4

A A T G A T C A

Gcn4

G A A T T C A G A A

HSE

G G G G A A C C

Mig1

C A C C G T G

Met4/Cbf1

Typical situations : pattern discovery

- Selected sequence set
 - e.g. family of 20 co-regulated genes, obtained from DNA chip experiment
→ identify putative regulatory sites
- Genome-scale pattern discovery
 - e.g. all upstream sequences
→ identify transcription initiation signals
 - e.g. all downstream sequences
→ identify 3' maturation signals

Typical situations : pattern matching

- Selected genes, selected patterns
 - e.g. 10 genes known to be regulated by a factor
→ search matching positions
- Selected genes, library of patterns
 - → infer putative action of any previously known transcription factor
- All genes, selected patterns
 - → classify all the genes of a genome according to putative regulatory properties

Met4p binding sites

gene	start	end	sequence
MET3	-367	-349	GAAAAGTCACGT GTA ATT
MET3	-384	-366	AAAAGGT CACGTG A CCAGA
MET14	-235	-217	CTAATT T CACGT GAT CAAT
MET16	-185	-167	ATCATT T CACGT GG GCTAGT
ECM17	-311	-293	ATTTCAT T CACGT G CGTATT
ECM17	-339	-321	.TTTGT CACGTG A TATTC
MET10	-255	-237	.CCACAC CACGTG A GCTT TAT
MET10	-237	-219	.TAGAAG CACGTG A CCACAA
MET2	-360	-342	GTATT T T CACGTG A TGC GC
MET2	-554	-536	TAATAAT T CACGT GAT ATTT
MET17	-306	-288	.AAATGG CACGTG A AGCT GT
MET17	-332	-314	TTGAGGT CACATG A T CGCA
MET6	-540	-522	GCCACAT T CACGT GCA CATT
MET6	-502	-484	AATATT T CACGT G A CTTAC
SAM2	-329	-311	.TCTAC CACGTG A CTATAA
SAM2	-381	-363	.TCTTCAC CATG T GATT CATC

A	13	11	3	3	2	0	16	0	1	0	0	12
C	1	0	0	3	0	16	0	15	0	0	0	0
G	1	1	4	4	4	0	0	0	15	0	16	4
T	1	4	9	6	10	0	0	1	0	16	0	0

Met31p binding sites

gene	start	end	sequence
MET14	-202	-182	CCTC AAAAAA ATGTGGCAATGG
MET2	-313	-293	TGC AAAAAA ATTGTGGATGCAC
MET17	-227	-207	TCATG AAA ACTGTGTAAACATA
MET6	-313	-293	GTCGC AAA ACTGTGGTAGTCA
SAM2	-306	-286	GCTTG AAA ACTGTGGCGTTT
SAM1	-283	-263	ACAGG AAA ACTGTGGTGGCGC
MET19	-173	-153	ATAAG AAA ACTGTGGGTTCAT
MUP3	-188	-168	CGG AAAAA ACTGTGGCGTCGC
MET8	-184	-164	GG AAAAAAA ATGTGAAAATCG
MET1	-232	-212	CATAATA AA ACTGTG A ACGGAC
MET3	-259	-239	ACAAAGCCACACAGTTTACAAC
MET28	-159	-139	CTAACACCACAGTTTGGCG
MET8	-434	-414	TCTTGTCCGCAGTTT T ATCTG
MET30	-168	-148	GGGAAGCCACACAGTTGCGCGG
MET6	-405	-385	CTATCGAA CTCGTT AGTCGC

A	5	11	14	14	14	2	0	0	0	0	2	5
C	2	2	0	0	0	11	0	0	1	0	0	5
G	5	0	0	0	0	0	0	14	0	14	11	1
T	2	1	0	0	0	1	14	0	13	0	1	3

Pho4p binding sites

gene	start	end	sequence
PHO5	-260	-242	.. GCACTCA CACGTGGG ACTA
PHO5	-260	-245	.. GCACTCA CACGTGGG A
PHO5	-262	-239	TGGCACTCA CACGTGGG ACTAGCA
PHO8	-540	-522	... TCGGGC CACGTGC AGCGAT
PHO8	-736	-718	.. ttacccg CACG <u>TT</u> aatat
PHO81	-350	-332	... TTATGG CACGTGCG AAATAA
PHO84	-421	-403	.. TTTCCAG CACGTGGG GCGG
PHO84	-442	-425	... TAGTTC CACGTGG ACGTG
PHO84	-879	-874	.aaaagtgt CACGTG ataaaaat
PHO84	-267	-250	.. taatacg CACGTTTT aa
PHO84	-592	-575	... TTACG CACGTT GGTGCTG
PHO5	-368	-349	... AATTAG CACGTTTT CGCATA
PHO5	-369	-354	.. AAATTAG CACGTTT CTC
PHO5	-370	-347	. TAAATTAG CACGTTTT CGCATAGA

IUPAC ambiguous nucleotide code

A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A or G	puRine
Y	C or T	pYrimidine
W	A or T	Weak hydrogen bonding
S	G or C	Strong hydrogen bonding
M	A or C	aMino group at common position
K	G or T	Keto group at common position
H	A, C or T	not G
B	G, C or T	not A
V	G, A, C	not T
D	G, A or T	not C
N	G, A, C or T	aNy

Pho4p binding specificity - matrix descriptions

C

Pho4p

A	14	0	5	7	6	0	26	0	0	0	0	3
C	2	8	5	16	6	26	0	26	0	1	0	4
G	4	2	1	1	12	0	0	0	26	0	16	12
T	6	16	15	2	2	0	0	0	0	25	10	7

D

Pho4p.cacgtg

A	2	17	0	0	0	0	2	1	8	5	5	13
C	16	0	18	0	0	0	6	3	4	5	0	1
G	0	1	0	18	0	18	9	12	2	5	2	1
T	0	0	0	0	18	0	1	2	4	3	11	3

E

Pho4p.cacgtt

A	7	0	2	5	1	0	8	0	0	0	0	1
C	0	1	1	3	3	8	0	8	0	0	0	0
G	0	0	0	0	4	0	0	0	8	0	0	2
T	1	7	5	0	0	0	0	0	0	8	8	5

Position-specific scoring matrix (PSSM)

Pos	1	2	3	4	5	6	7	8	9	10
A	3	2	0	12	0	0	0	0	1	3
T	1	1	0	0	0	0	11	5	4	4
G	3	7	0	0	0	12	0	7	5	4
C	5	2	12	0	12	0	1	0	2	1

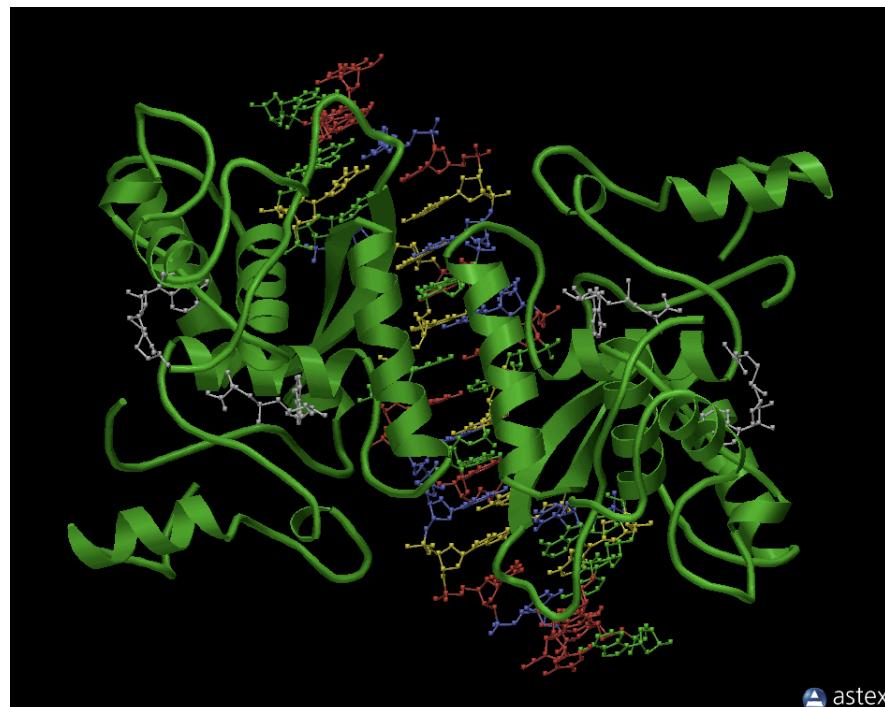
Binding motif for the yeast Pho4p transcription factor

Source : SCPD

<http://rulai.cshl.edu/cgi-bin/SCPD/getfactor?PHO4>

Methionine repressor

- Crystal structure : the methionine repressor of *Escherichia coli*.
- In green: the MetJ protein forms a homodimer which is able to bind DNA.
- Nucleotide structure is coloured by type of nucleotide (A,C,G,T).
- In grey: the repressor is activated by binding of methionine molecules

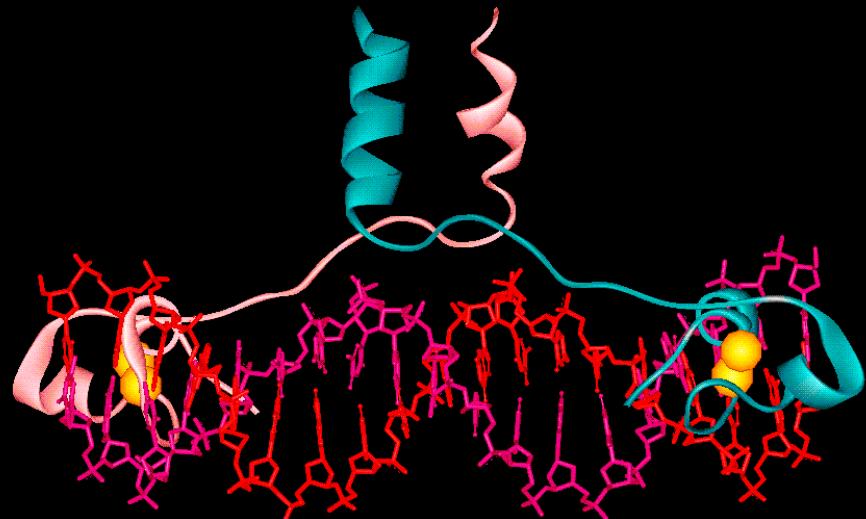


Transcription factor-DNA interfaces

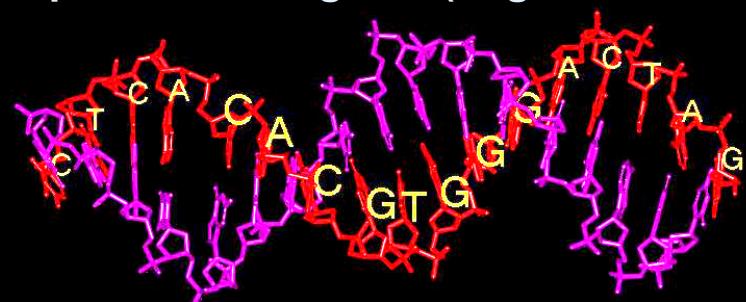
Pho4p (yeast)



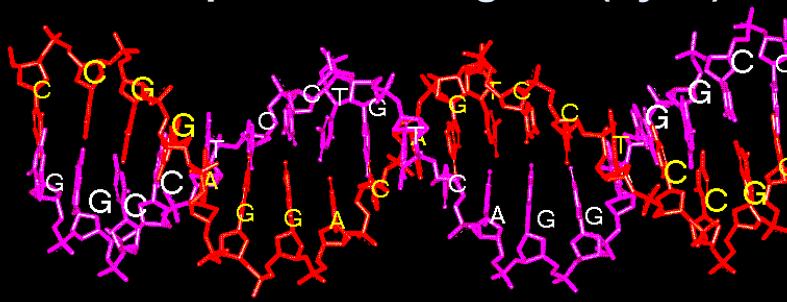
Gal4p (yeast)



Pho4p DNA binding site (oligonucleotide)



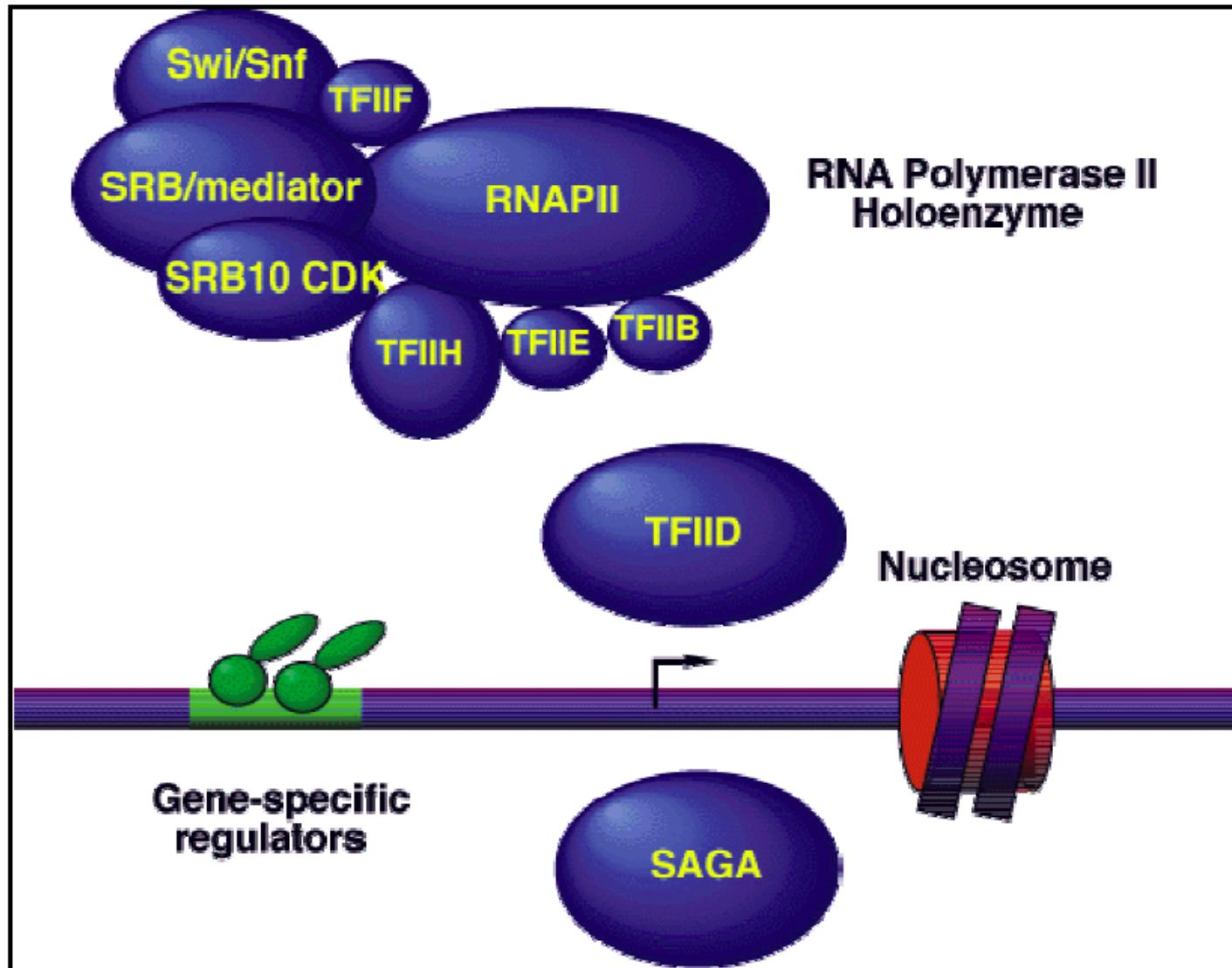
Gal4p DNA binding site (dyad)



The genome challenge



RNA polymerase



The non-coding genome

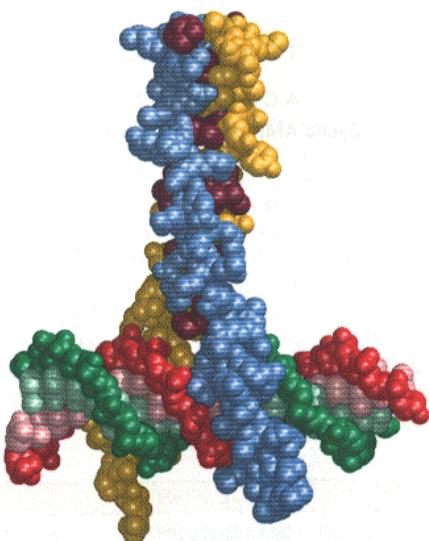
Organism	Year	Size Mb	Genes	genes/Mb	coding		% non-coding	% repetitive	% Transcribed
					%	%			
<i>Mycoplasma genitalium</i>	1995	0.6	481	802	90	10			
<i>Haemophilus influenzae</i>	1995	1.8	1 717	954	86	14			
<i>Escherichia coli</i>	1997	4.6	4 289	932	87	13			
<i>Saccharomyces cerevisiae</i>	1996	12	6 286	524	72	28			
<i>Arabidopsis thaliana</i>	2001	120	27 000	225	30	70			
<i>Caenorhabditis elegans</i>	1998	97	19 000	196	27	73			
<i>Drosophila melanogaster</i>	2000	165	16 000	97	15	85			
<i>Homo sapiens</i>	2001	3 200	31 000	10	3	97	46	28	

Genomic sequences

- A genome G contains a set of n chromosomes.
 - $G = \{S_1, S_2, \dots, S_i, \dots, S_n\}$
- Each chromosome is a molecule of deoxyribonucleic acid (DNA), a polymer of 4 nucleotides
 - A Adenosine
 - C Cytidine
 - G Guanosine
 - T Thymidine
- Each chromosome is represented as a sequence (S_i) of a text written in a 4-letter alphabet (A)
 - $A = \{A, C, G, T\}$
 - $S_i = (s_{i1}, s_{i2}, \dots, s_{ij}, \dots, s_{iL_i})$
 - L_i is the length of the i^{th} chromosome

GCN4 structure from the Stryer textbook

**GCN4 (leucine-zipper)
bound to DNA**



Source: L.Stryer, (1995). Biochemistry. p1003