Regulatory Sequence Analysis

Introduction to cis-regulation

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univ-amu.fr/

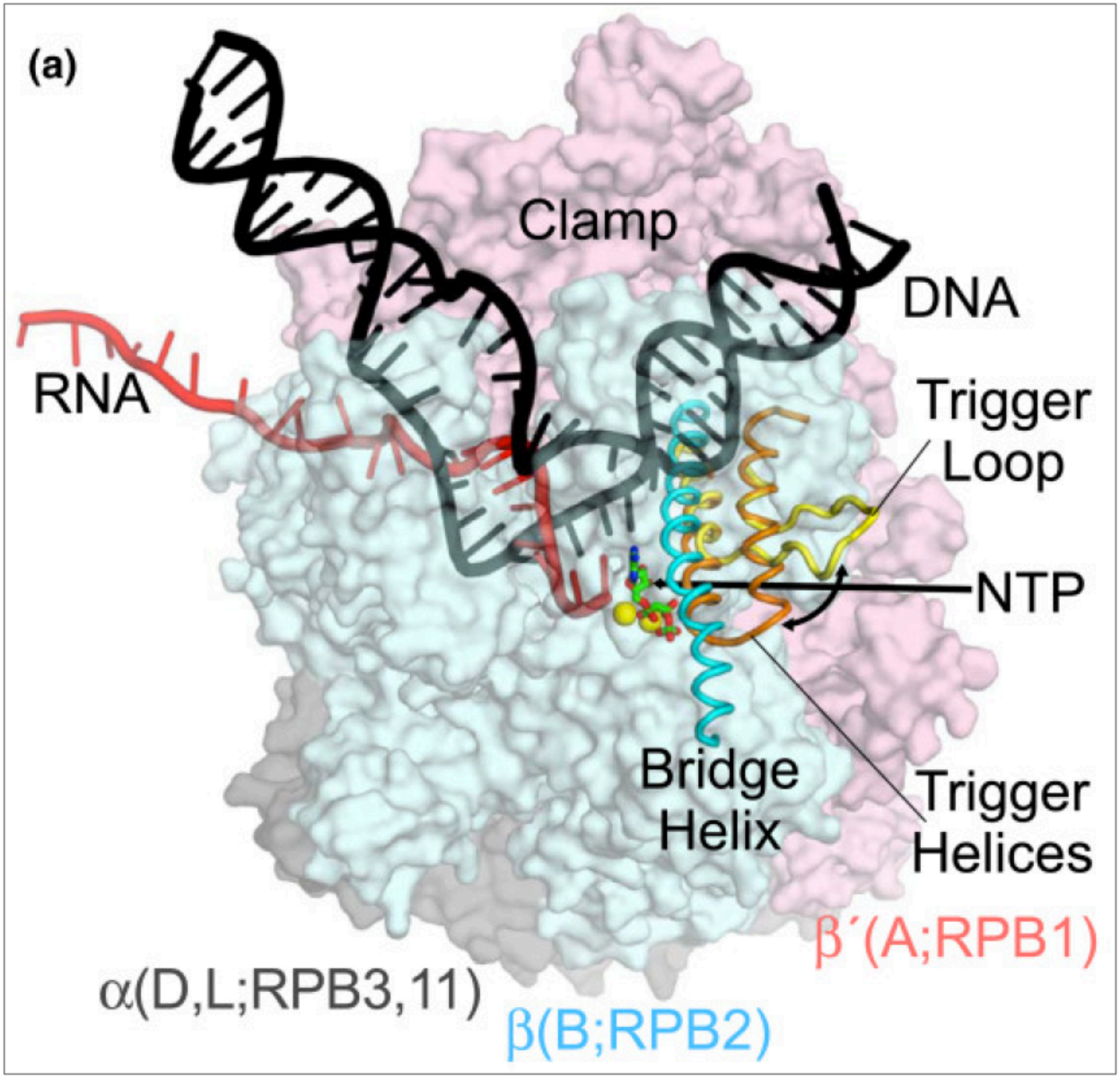
FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/

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

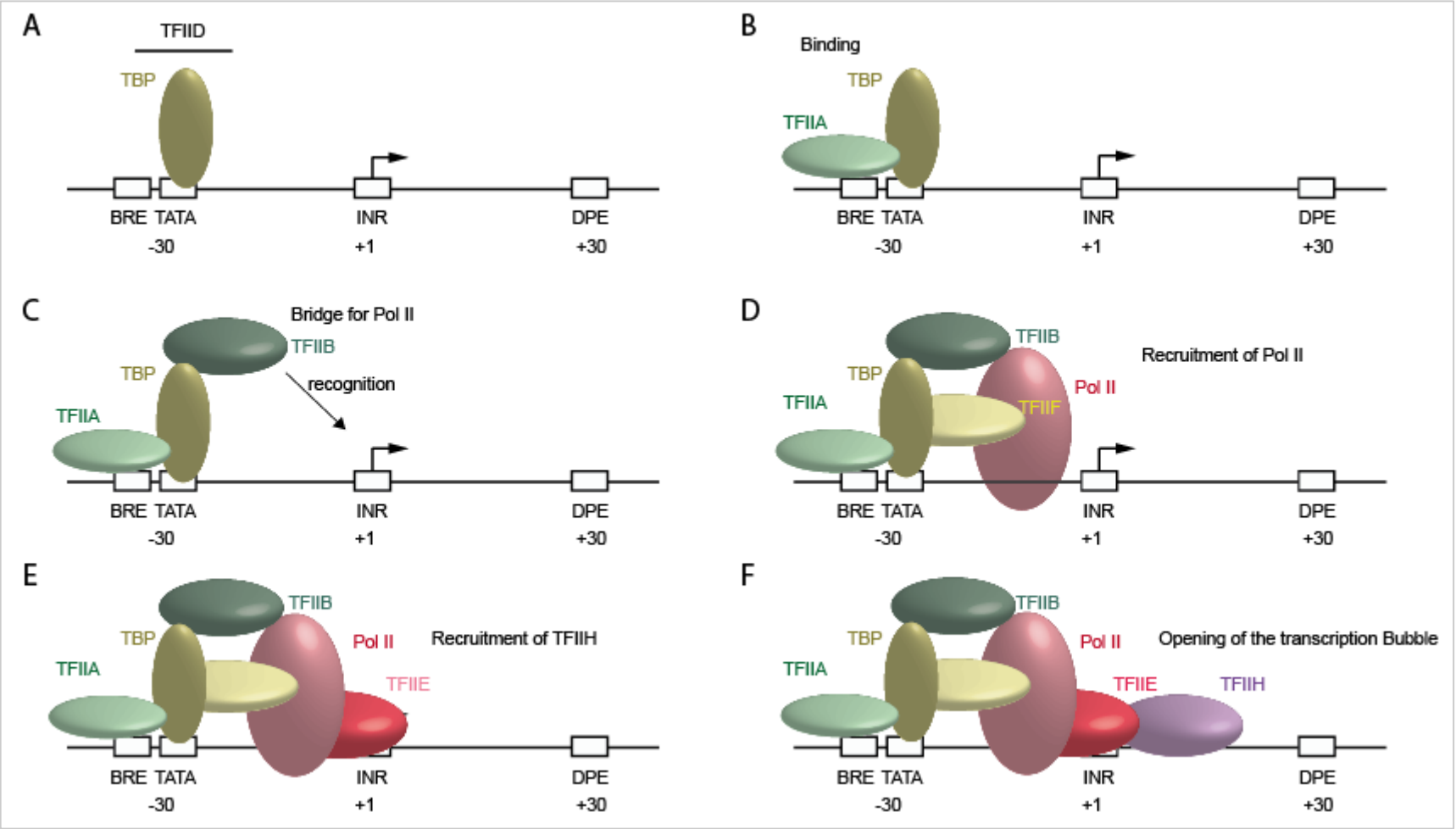
Transcriptional activation and repression

RNA polymerase 3D model



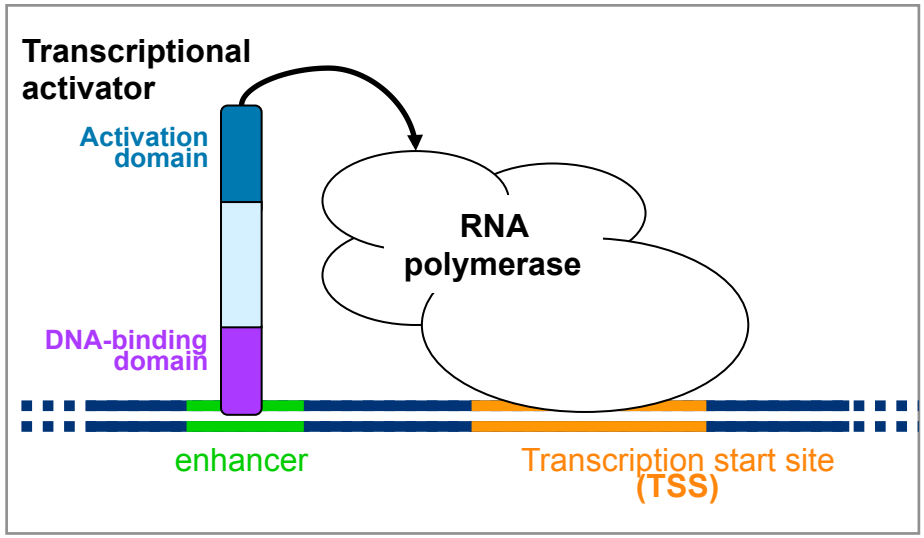
- Hein and Landick. The bridge helix coordinates movements of modules in RNA polymerase. BMC Biol (2010) vol. 8 pp. 141, Figure 1

RNA pol II assembly

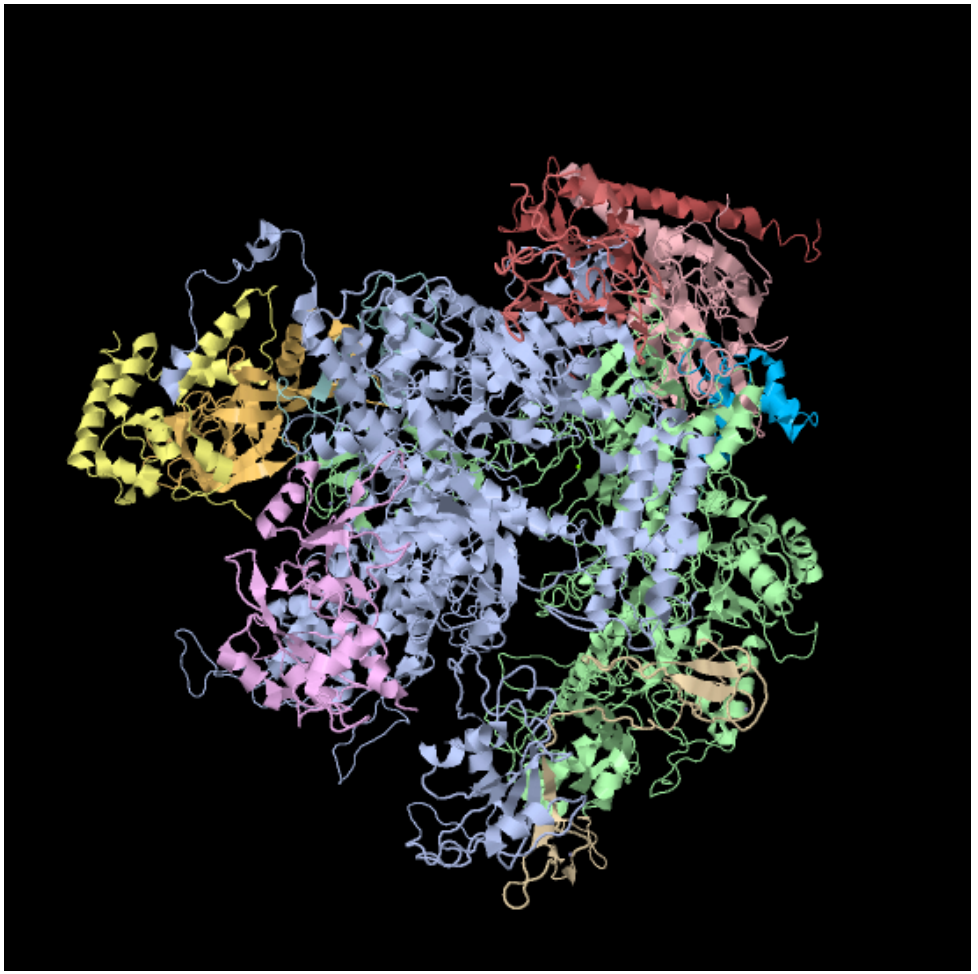


- Figure: Courtesy from Nicolas Descostes (from his PhD thesis manuscript, Dec. 2014).

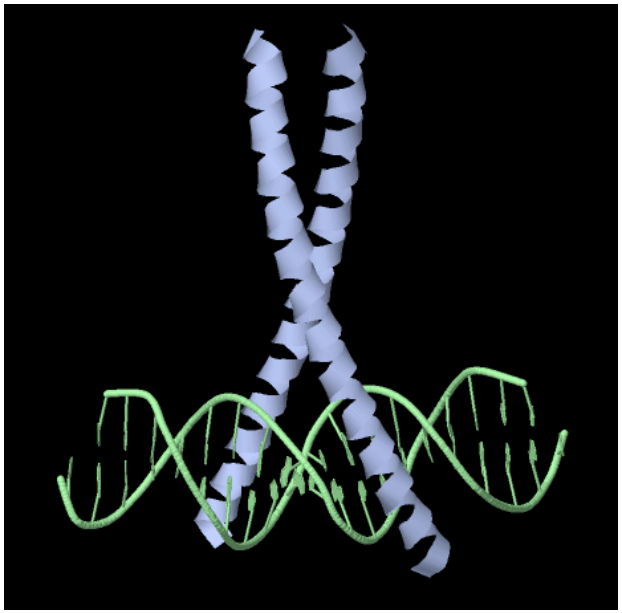
Transcriptional activation



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

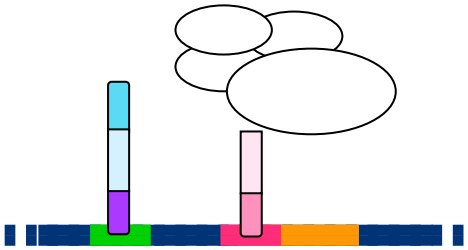


Gcn4p from Saccharomyces cerevisiae
*PDB **2DGC*** http://www.rcsb.org/pdb/explore.do?structureId=2DGC

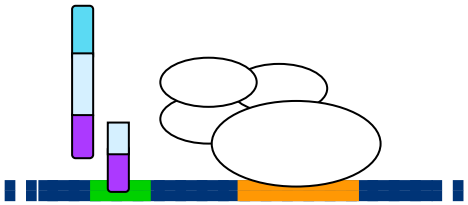


Transcriptional repression

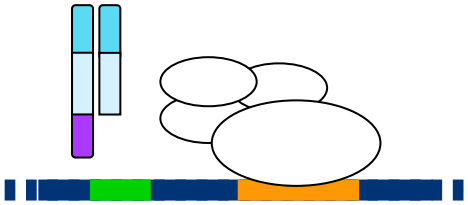
- The concept of transcriptional repression encompasses a variety of molecular mechanisms.



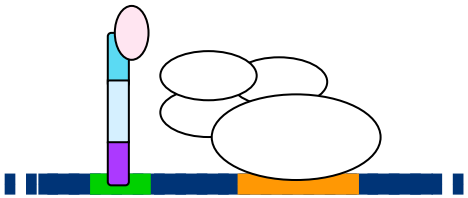
Prevent RNA polymerase from accessing DNA
(e.g. many bacterial repressors)



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



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

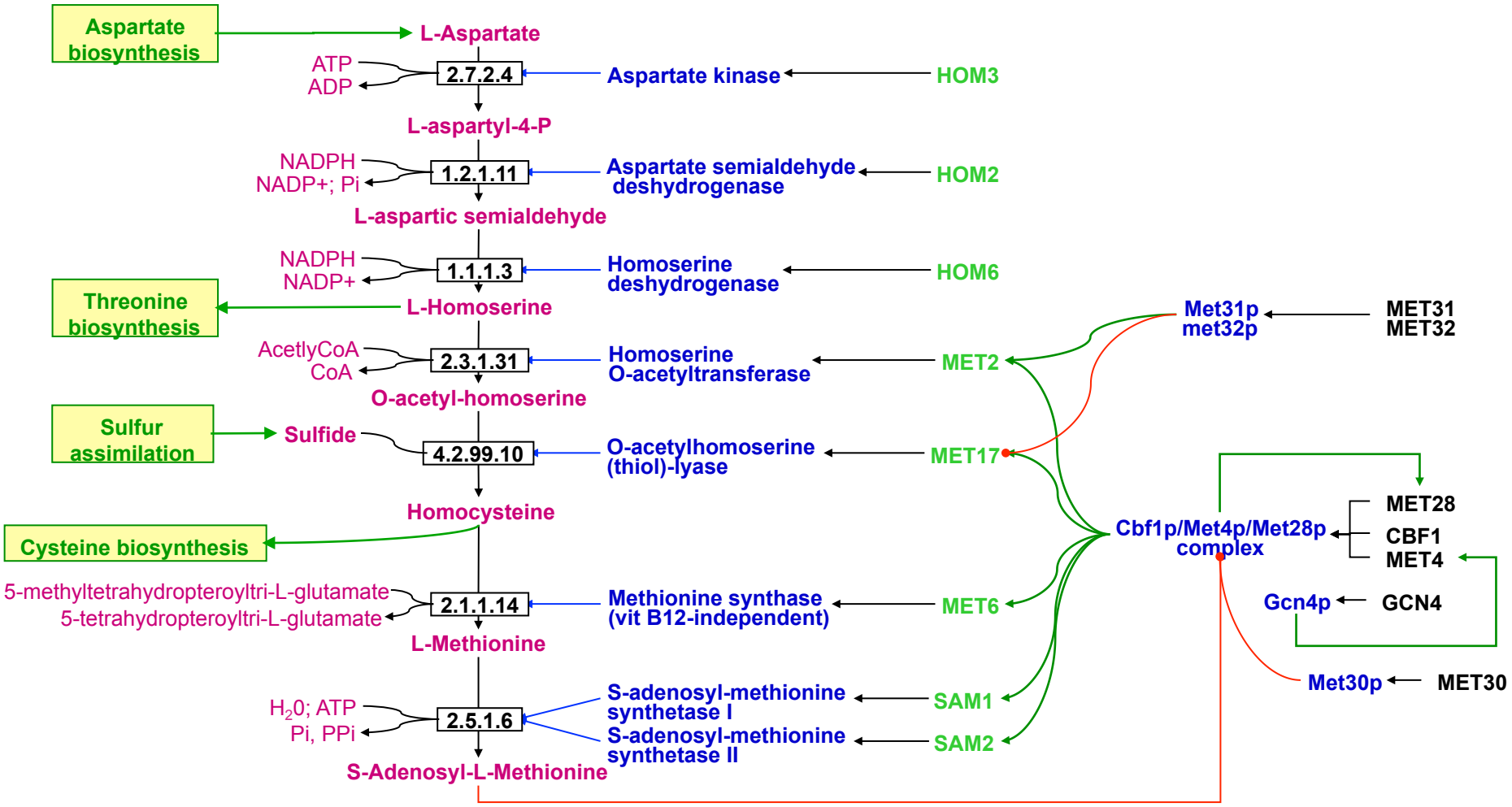


Modify transcription factor conformation -> prevent it from interacting with RNA-polymerase
(e.g. yeast Gal80p)

Cis-regulation of biological processes : some examples

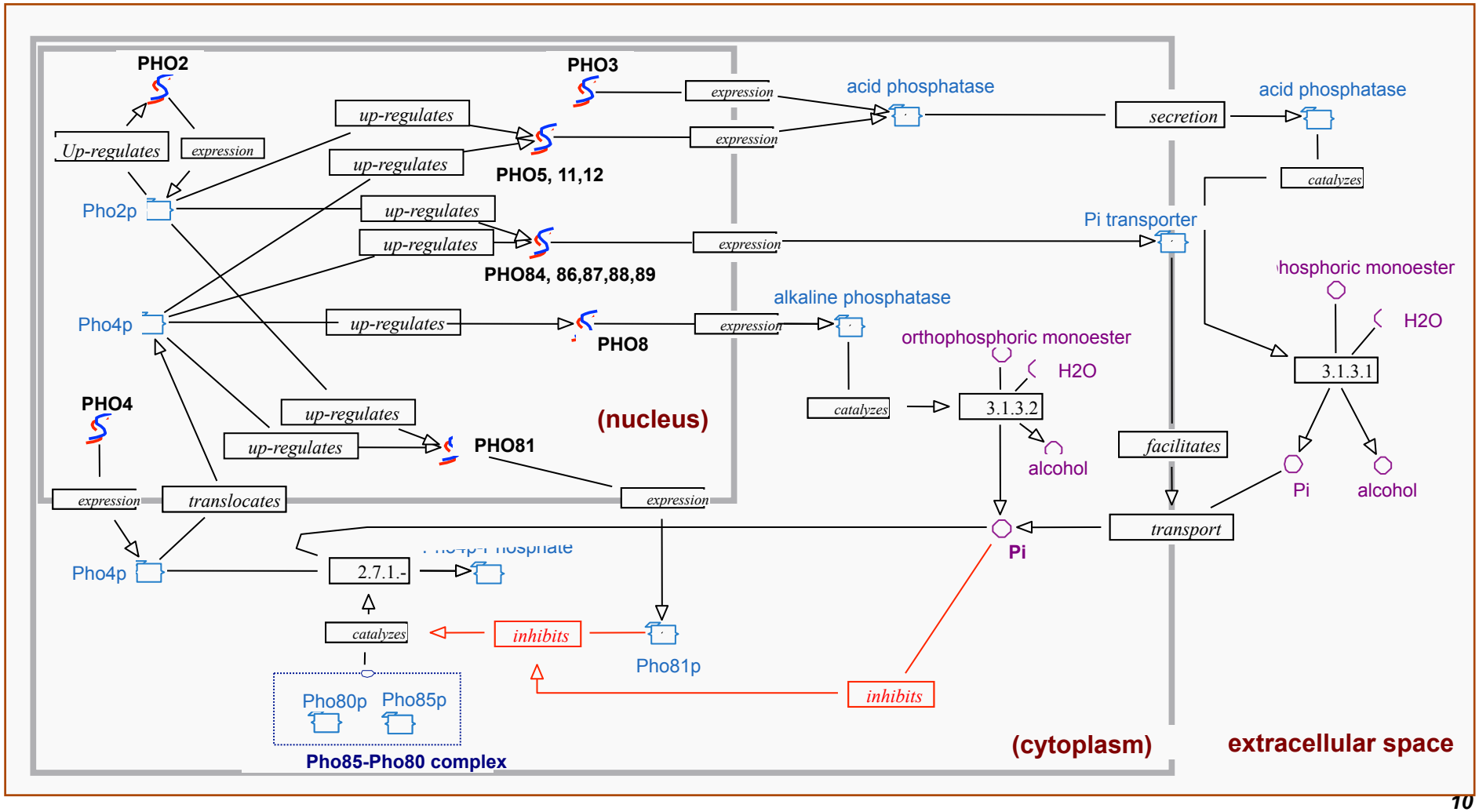
Methionine Biosynthesis in Saccharomcyes cerevisiae

- 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: alcaline (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.

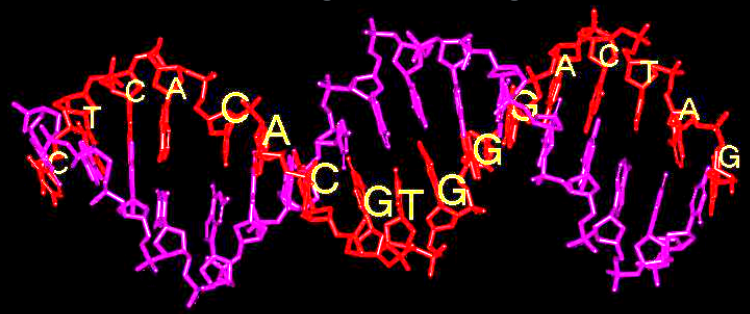


Transcription factors (TF) and their binding sites (TFBS)

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

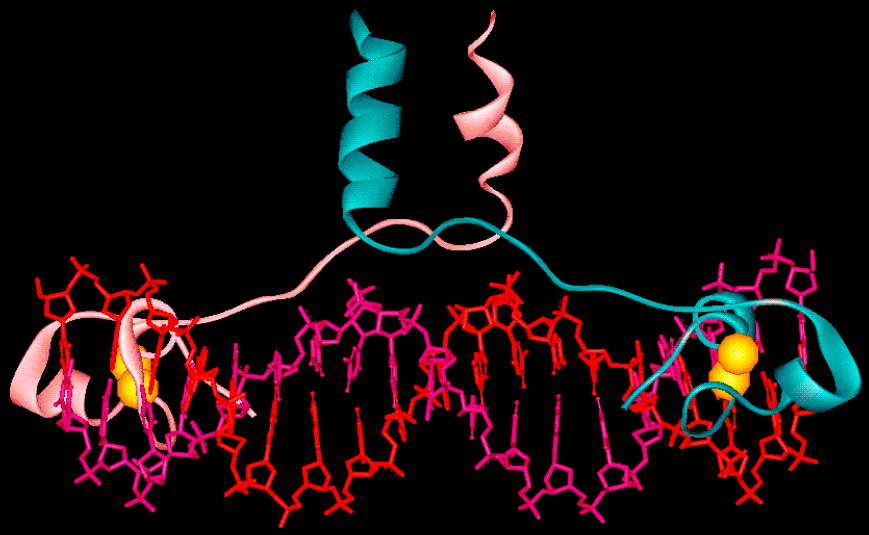
Pho4p (yeast)

Pho4p DNA binding site (oligonucleotide)

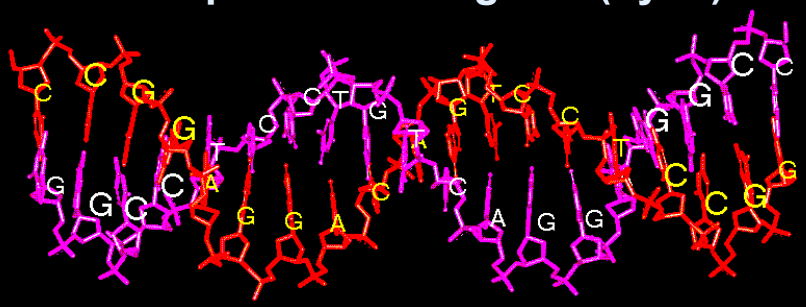


Interface between the yeast Gal4p protein and one of its binding sites

Gal4p (yeast)



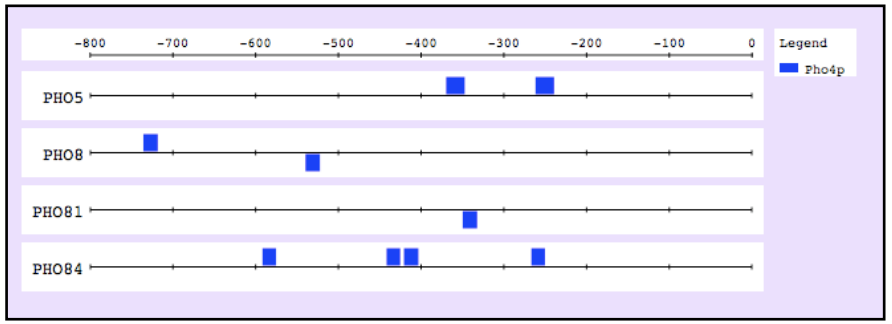
Gal4p DNA binding site (dyad)



S.cerevisiae Pho4p binding sites (TFBS)

Gene	Ft_type	Factor	Strand	left	right	Sequence
PHO5	site	Pho4p	D	-370	-347	TAAATTAG**CACGTTTT**CGCATAGA
PHO5	site	Pho4p	D	-262	-239	TGGCACTCA**CACGTGGG**ACTAGCA
PHO8	site	Pho4p	R	-540	-522	ATCGCTG**CACGTGGCCCG**A
PHO8	site	Pho4p	D	-736	-718	ATATTAAGCGTGCGGGTAA
PHO81	site	Pho4p	R	-350	-332	TTATTCG**CACGTGCC**ATAA
PHO84	site	Pho4p	D	-592	-575	TTACG**CACGTT**GGTGCTG
PHO84	site	Pho4p	D	-421	-403	TTTCCAG**CACGTGGGGCGG**
PHO84	site	Pho4p	D	-442	-425	TAGTTC**CACGTGG**ACGTG
PHO84	site	Pho4p	DR	-879	-874	aaaagtg**t**CACGTG**a**taaaaat
PHO84	site	Pho4p	D	-267	-250	TAATACG**CACGTTTTT**AA

- 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.



From binding sites to binding motifs

Definition: transcription factor binding site (TFBS)

- Transcription factor binding site
  - *"Position on a DNA molecule where a transcription factor (TF) specifically binds. By extension, the sequence of the bound DNA segment. Note that there is a frequent confusion in the literature between the concepts of binding site and binding motif. We recommend to reserve the term "site" to denote the particular sequence (genomic or artificial) where a factor binds, and the term "motif" for the generic description of the binding specificity, obtained by summarizing the information provided by a collection of sites."*

- Jacques van Helden, in *Concise Encyclopaedia of Bioinformatics and Computational Biology*, 2nd Edition. John M. Hancock (Editor), Marketa J. Zvelebil (Editor). ISBN: 978-0-470-97871-9

Definition: transcription factor binding motif (TFBM)

- Transcription factor binding motif
  - "*Representation of the binding specificity of a transcription factor, generally obtained by summarizing the conserved and variable positions of a collection of binding sites. Several modes or representation can be used to describe TFBM: consensus, position-specific scoring matrices, Hidden Markov Models (HMM).*"
  - 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. **`CACGT[GT][GT][GT]`**
    - Position-specific scoring matrix (PSSM)
    - Logo representation (Schneider, 1986)
    - Hidden Markov Models (HMM)

- Jacques van Helden, in *Concise Encyclopaedia of Bioinformatics and Computational Biology*, 2nd Edition. John M. Hancock (Editor), Marketa J. Zvelebil (Editor). ISBN: 978-0-470-97871-9

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 sifficient 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**GCACGTGC**G**A**ATAA--	high
PHO8	Proximal	GTGATCGCT**GCACGTGGC**CCGA---	high
group 1	consensus	---------**gCACGTGgg**--------	high
PHO5	UASp1	--TAAATTA**GCACGTTTT**CGC----	medium
PHO84	Site E	----AATAC**GCACGTTTT**TAATCTA	medium
group 2	consensus	--------**cgCACGTTt**-------	medium
Degenerate consensus		---------**GCACGTKKk**-------	high-med

Non-binding sites

PHO5	UASp3	--TAATTTG**GCA**T**GTGCG**ATCTC--	No binding
PHO84	Site C	-----ACGTC**CACGTGG**A**A**CTAT--	No binding
PHO84	Site A	-----TTTAT**CACGTG**A**A**CACTTTTT	No binding
PHO84	Site B	-----TTAC**GCACGTT**G**G**TGCTG--	No binding
PHO8	Distal	---TTACCC**GCACG**C**TT**AATAT---	No binding

IUPAC ambiguous nucleotide code

A	**A**	**A**denine
C	**C**	**C**ytosine
G	**G**	**G**uanine
T	**T**	**T**hymine
R	**A** or **G**	pu**R**ine
Y	**C** or **T**	p**Y**rimidine
W	**A** or **T**	**W**eak hydrogen bonding
S	**G** or **C**	**S**trong hydrogen bonding
M	**A** or **C**	a**M**ino group at common position
K	**G** or **T**	**K**eto 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**	a**N**y

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 (2 – 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   \TACCACGTTTTC\

Cons      yvvCACGTkbkn
```

IUPAC ambiguous nucleotide code

A	**A**	**A**denine
C	**C**	**C**ytosine
G	**G**	**G**uanine
T	**T**	**T**hymine
R	**A** or **G**	pu**R**ine
Y	**C** or **T**	p**Y**rimidine
W	**A** or **T**	**W**eak hydrogen bonding
S	**G** or **C**	**S**trong hydrogen bonding
M	**A** or **C**	a**M**ino group at common position
K	**G** or **T**	**K**eto 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**	a**N**y

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

Alignment of Pho4p binding sites (TRANSFAC annotations)

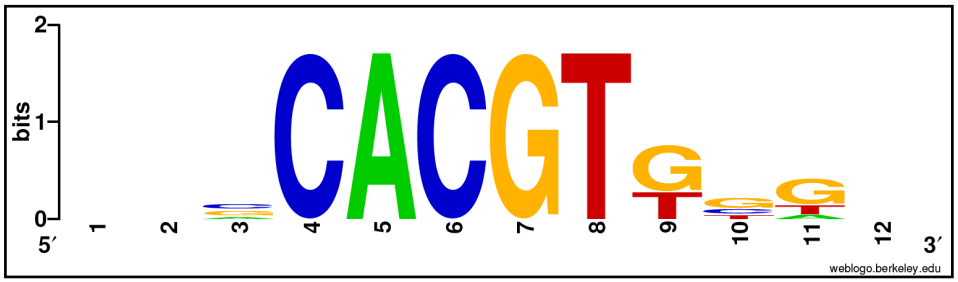
R06098	T	C	A	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**G**	A
R06099	G	G	**C**	**C**	**A**	**C**	**G**	**T**	**G**	C	A	G
R06100	T	G	A	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**G**	T
R06102	C	A	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**G**	G
R06103	T	T	**C**	**C**	**A**	**C**	**G**	**T**	**G**	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)

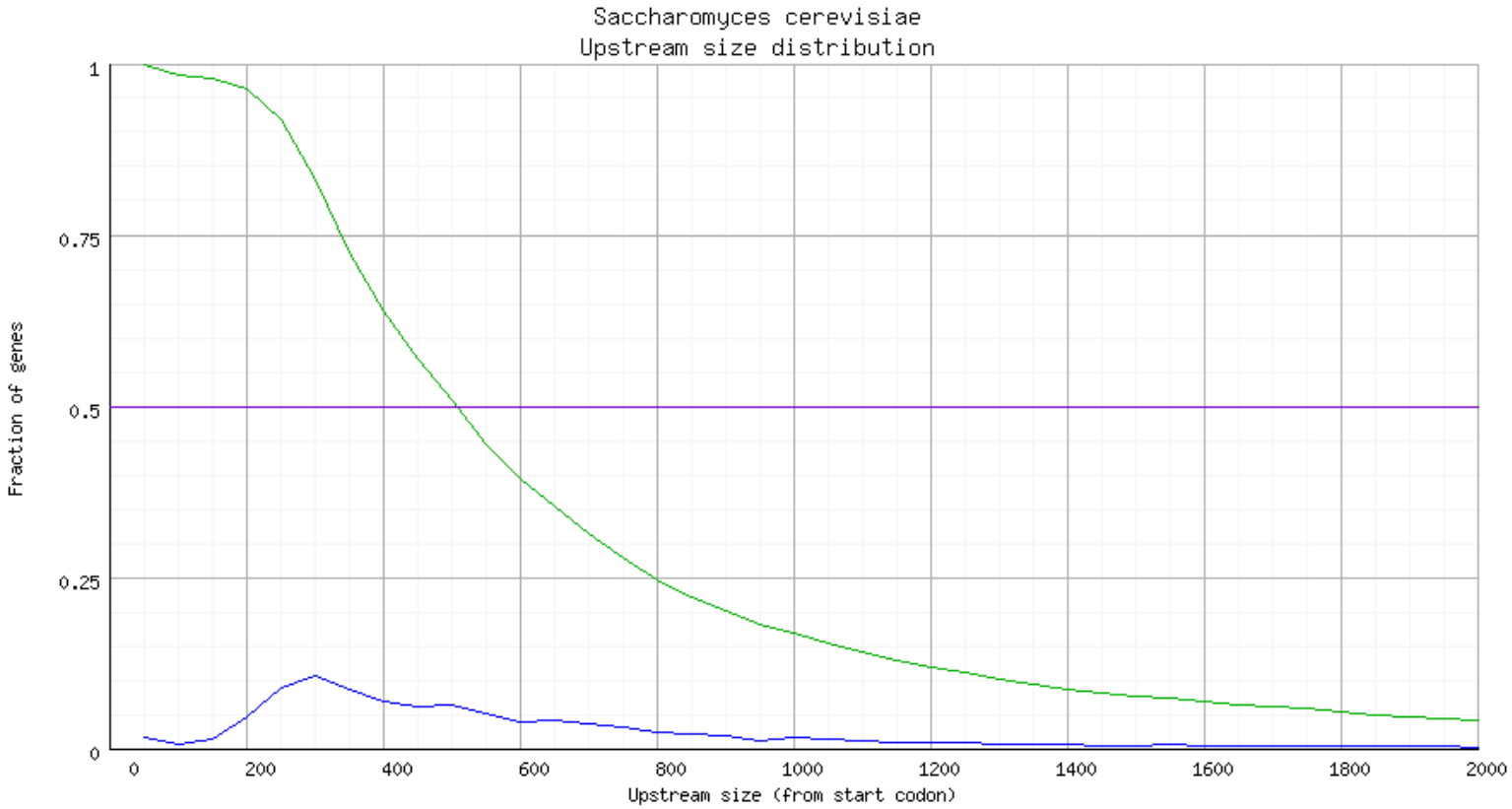


TRANSFAC record for the yeast PHO4 matrix (ID M00064)

```
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 regions

- 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

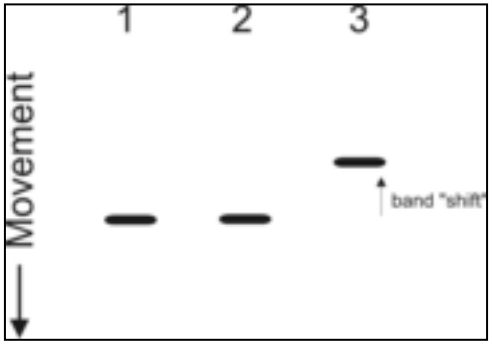


Experimental methods for characterizing cis-regulatory elements

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 regins: 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.

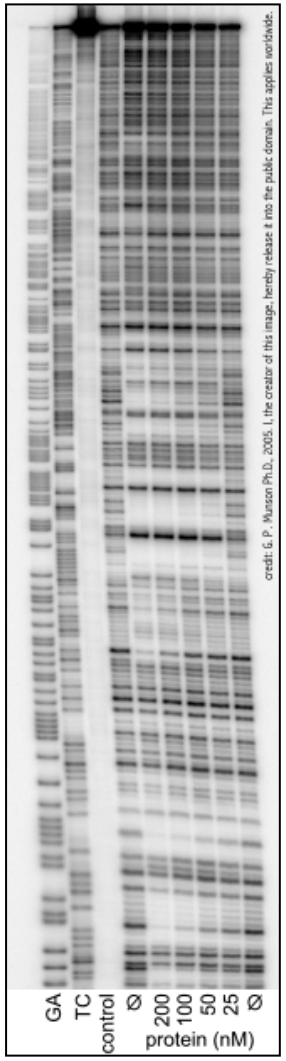
Gel shift (EMSA)



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.

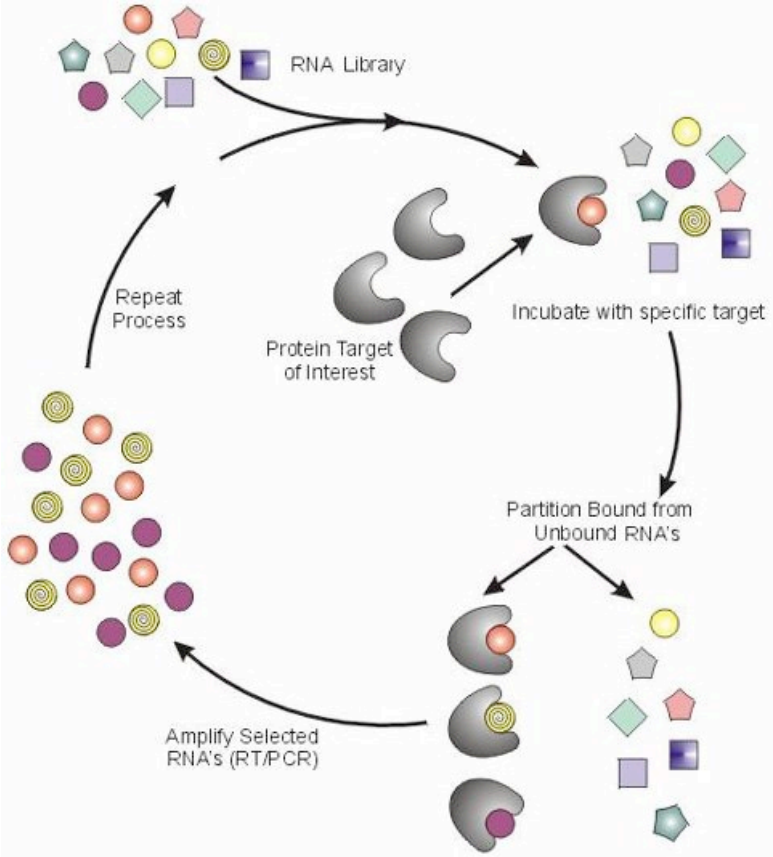
Source: Wikipedia

DNAse footprint



SELEX

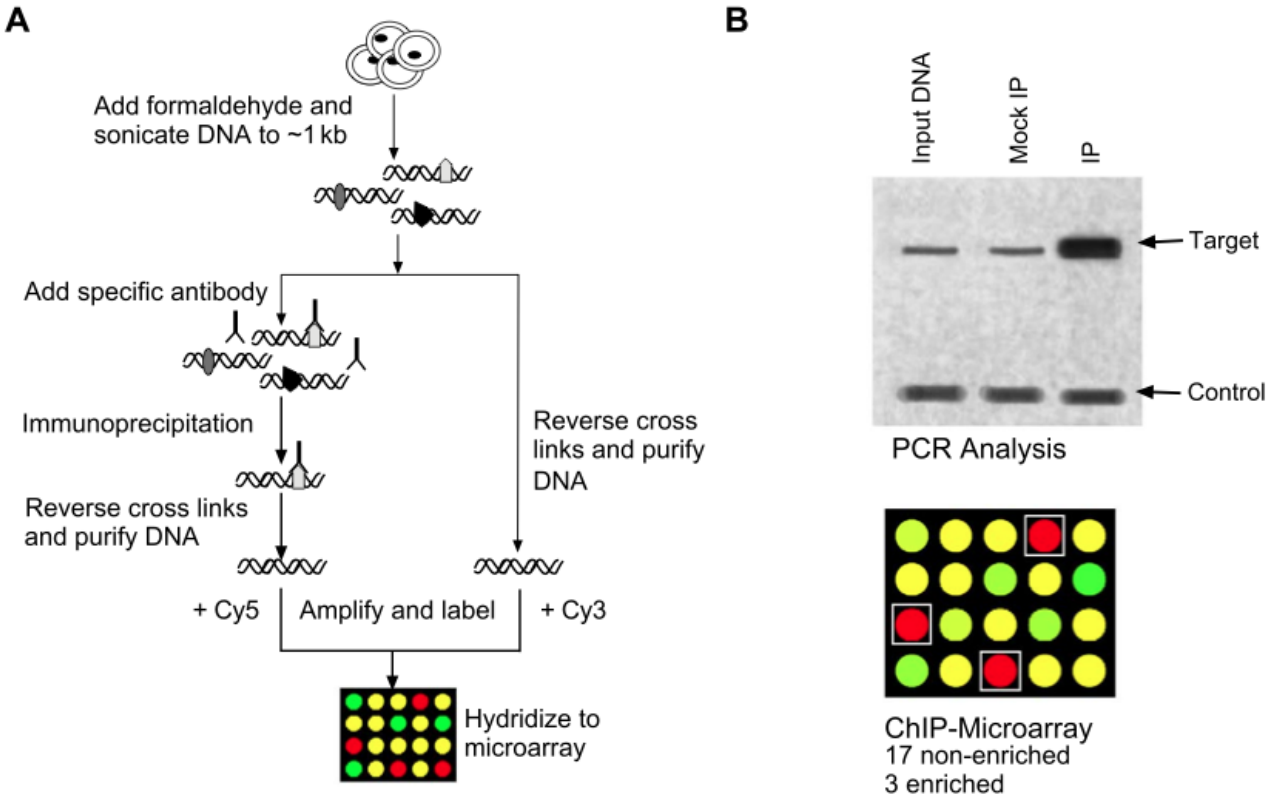
- Systematic Evolution of Ligands by EXponential enrichment.



http://www.molgen.mpg.de/~in-vitro/technology.html

ChIP-on-chip

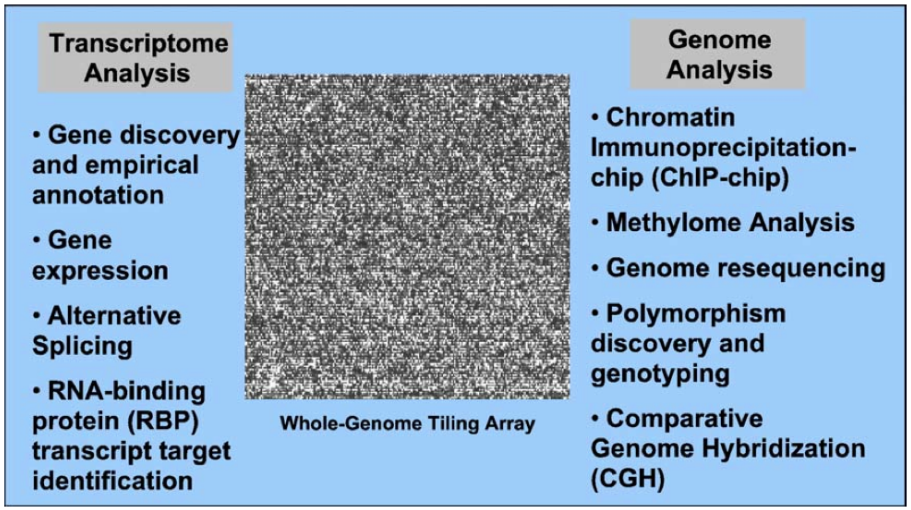
- The ChIP-chip method combines
  - Chromatin Immunoprecipitation (ChIP) to select genome fragments bound to a tagged transcription factor.
  - DNA microarrays (chip) spotted with several thousands of genome fragments (typically all the intergenic regions of agiven organism) are used to detect the relative enrichment: immuno-precipitated (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).

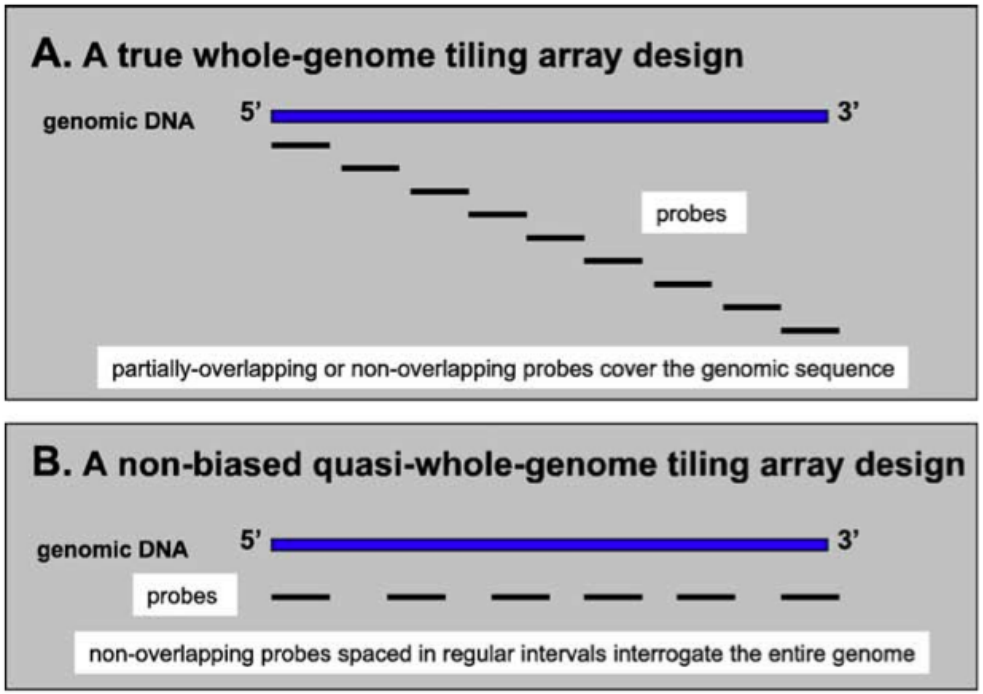


Buck and Lieb. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics (2004) vol. 83 (3) pp. 349-60

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.





Mockler et al. Applications of DNA tiling arrays for whole-genome analysis. Genomics (2005) vol. 85 (1) pp. 1-15

Universal protein-binding microarrays

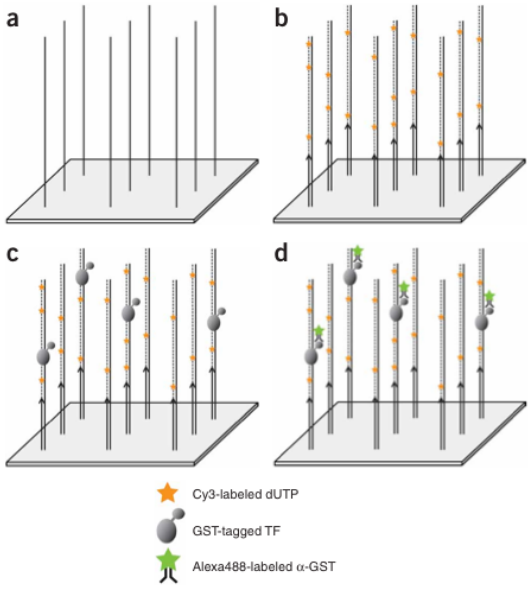
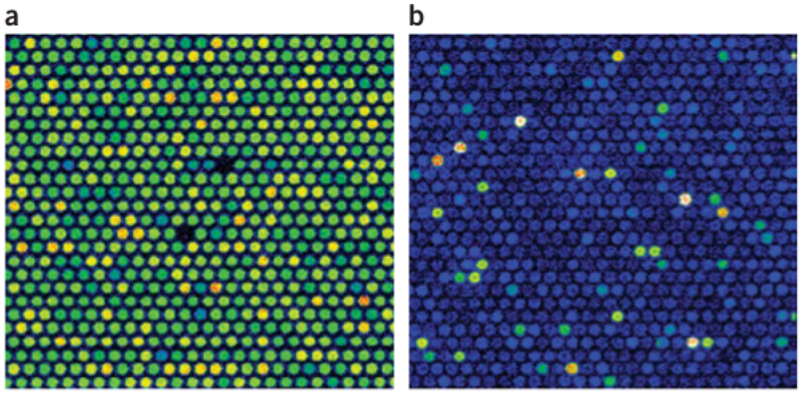


Figure 1 | Schematic of universal PBM experiments. (**a**) A commercially synthesized single-stranded DNA microarray is double-stranded by (**b**) solid-phase primer extension using a small amount of spiked-in fluorescently labeled dUTP. (**c**) An epitope-tagged TF is bound directly to the DNA on the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



Berger and Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nature Protocols (2009) vol. 4 (3) pp. 393-411

The “next generation sequencing” (NGS) era

- When sequencing costs followed Moore’s law
  - The cost of sequencing decreased exponentially since the end of the 1990s, due to the improvements and automation of sequencing, stimulated by the genome sequencing projects.
  - This decrease was more or less proportional to the exponential decrease of storage and computing costs (Moore’s law).
- ***Next Generation Sequencing***
  - In 2007, several companies proposed new technologies enabling a much faster sequencing.
  - The cost of sequences now decreases much faster than the cost of computers.
  - We can foresee real problems for storing and analyzing the massive amounts of sequences to be produced.

Figure from: Sboner et al. The real cost of sequencing: higher than you think!. Genome Biol (2011) vol: 12 (8) pp. 125

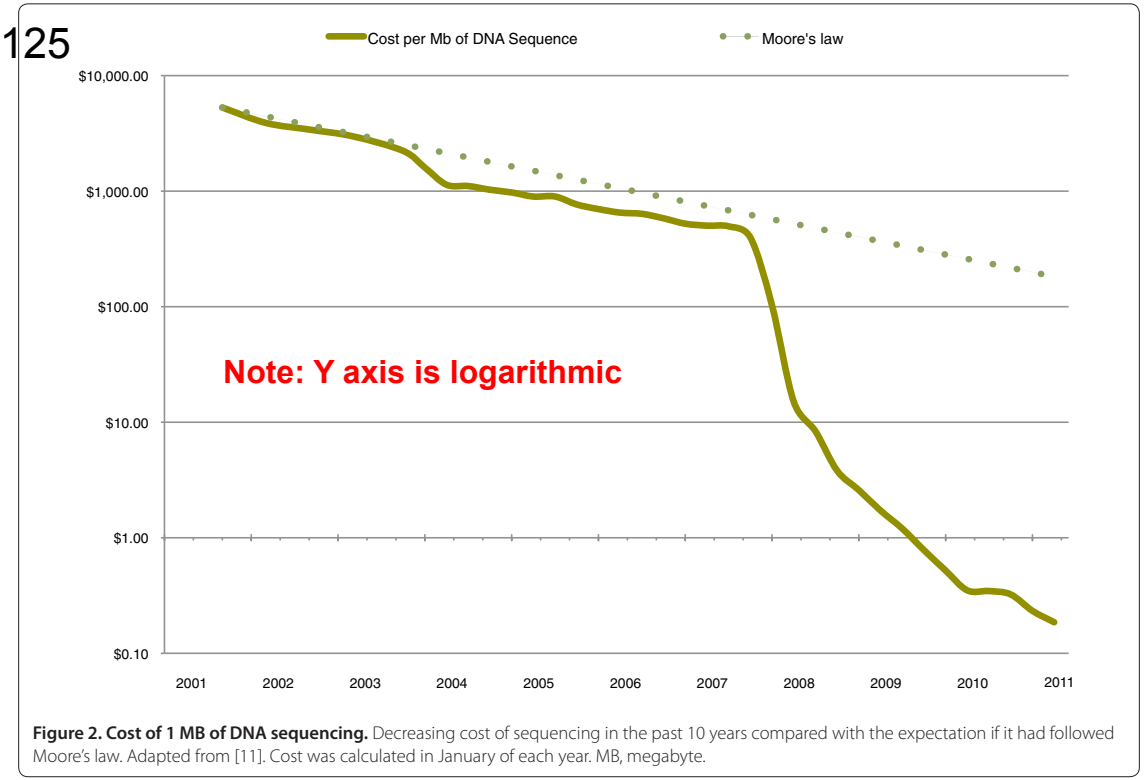


Figure 2. Cost of 1 MB of DNA sequencing. Decreasing cost of sequencing in the past 10 years compared with the expectation if it had followed Moore’s law. Adapted from [11]. Cost was calculated in January of each year. MB, megabyte.

Sboner et al. (2011) The real cost of sequencing: higher than you think!. Genome Biol 12: 125

Cost of sequencing projects

- The decrease of sequencing cost is accompanied by a drastic change in cost repartition, with a relative increase of the pre-processing (sample collection) and post-processing (bioinformatics analysis).
- There is thus an increasing need for bioinformatics know-how in all the laboratories treating next generation sequencing data.

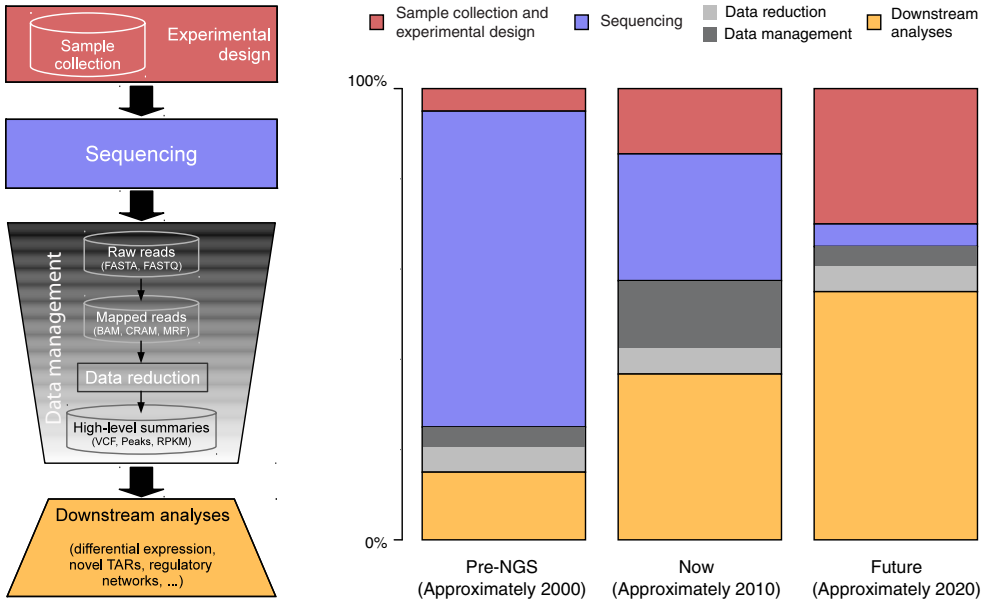
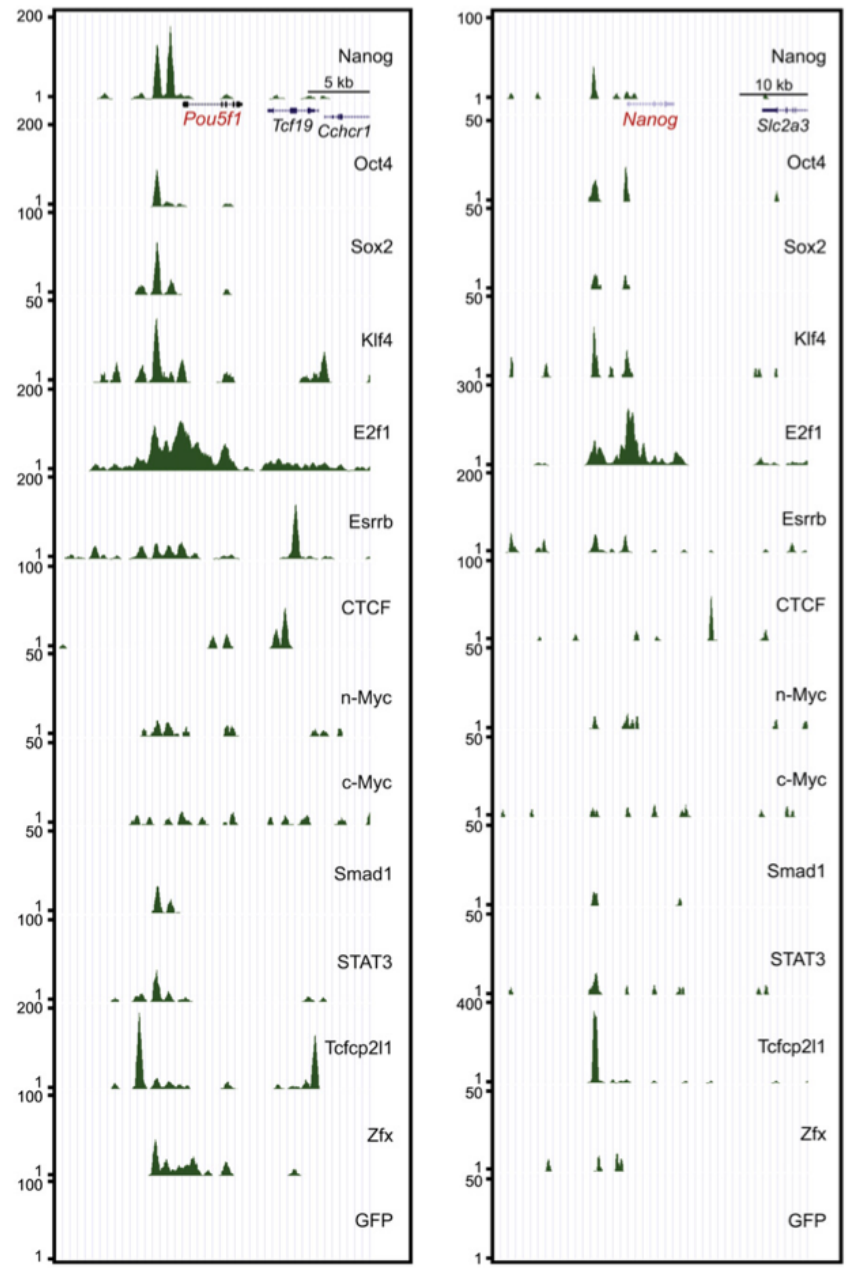


Figure 1. Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.

Sboner et al. (2011) The real cost of sequencing: higher than you think!. Genome Biol 12: 125

ChIP-seq

- Combination of
  - Chromatin Immunoprecipitation (ChIP), as for ChIP-chip.
  - Next Generation Sequencing (NGS) to characterize the immunoprecipitated DNA fragments.
- 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.
  - Detection of relevant peaks (peak calling) is not trivial.



Source: Chen et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell (2008) vol. 133 (6) pp. 1106-17

From transcription factor binding sites to cis-regulatory modules (CRM)

Cis-regulatory modules (CRM)

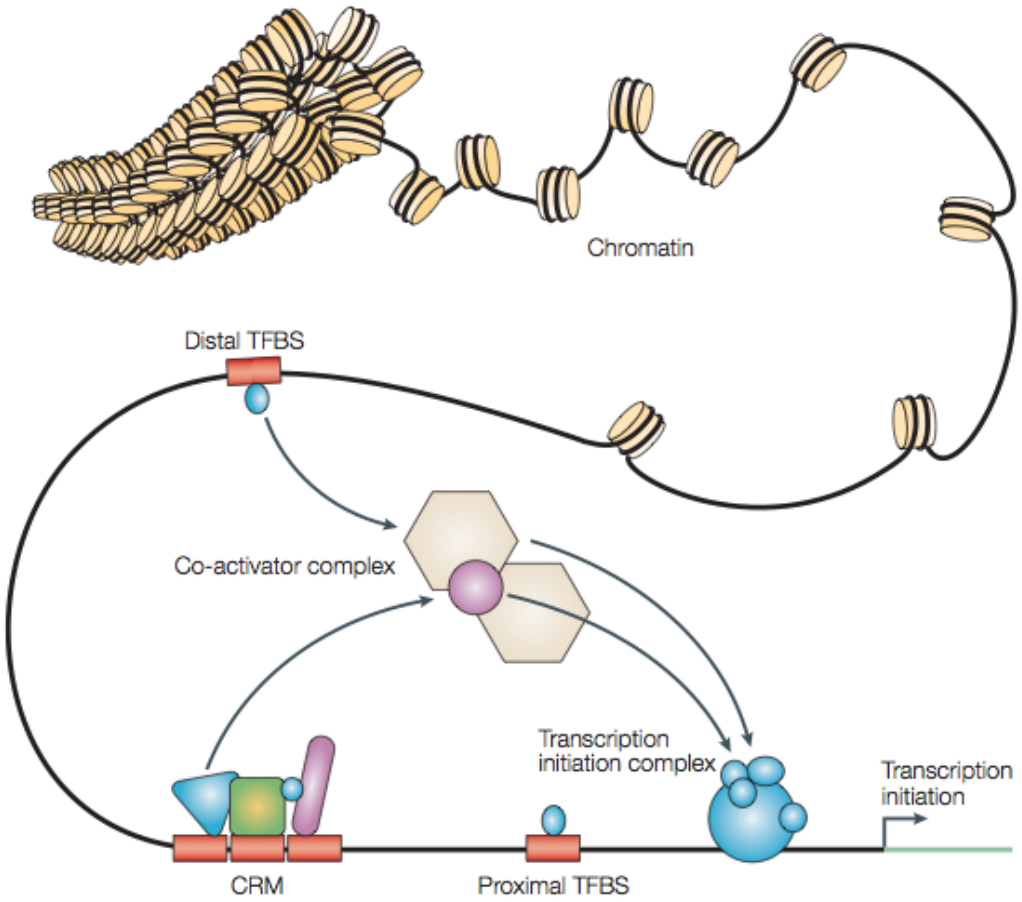


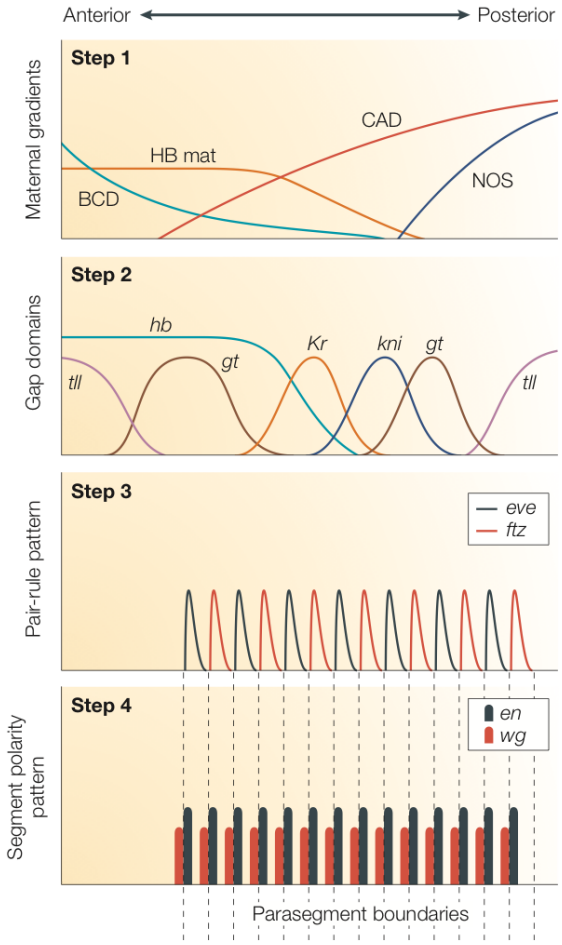
Figure 1 | **Components of transcriptional regulation.** Transcription factors (TFs) bind to specific sites (transcription-factor binding sites; TFBS) that are either proximal or distal to a transcription start site. Sets of TFs can operate in functional *cis*-regulatory modules (CRMs) to achieve specific regulatory properties. Interactions between bound TFs and cofactors stabilize the transcription-initiation machinery to enable gene expression. The regulation that is conferred by sequence-specific binding TFs is highly dependent on the three-dimensional structure of chromatin.

Source: Wasserman and Sandelin. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet (2004) vol. 5 (4) pp. 276-87. PMID 15131651

- In Metazoa, 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

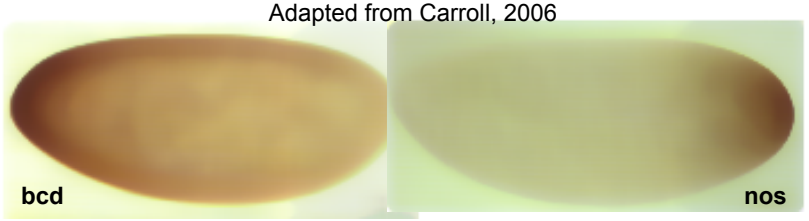
Drosophila Antero-Posterior (AP) segmentation – expresion domains

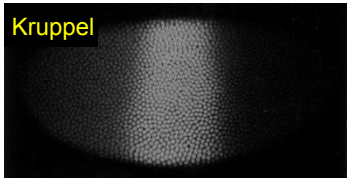
- Establishment of expression domains
  - Maternal genes: gradients of mRNAs coding for transcription factors.
  - Gap genes: broad domaines.
  - Pair-rule genes: expressed every other segment (odd or even segments).
  - Segment polarity genes: expressed with a segmental periodicity, across 2-3 cells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



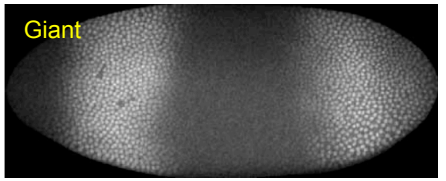
Peel et al. Arthropod segmentation: beyond the Drosophila paradigm. Nature Reviews Genetics (2005) vol. 6 (12) pp. 905-16. PMID 16341071

Adapted from Carroll, 2006

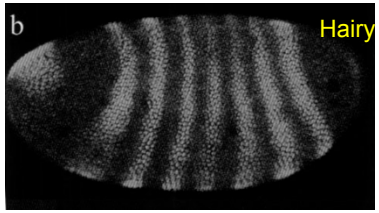




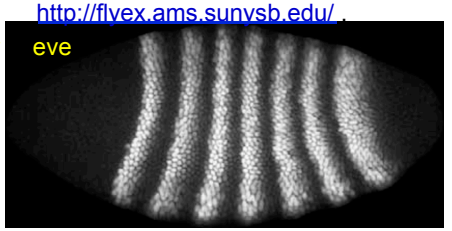
Source: Carroll, 2005.



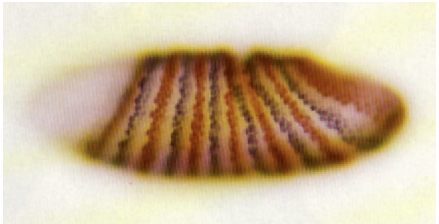
Source: Thieffry and Sanchez (2003). http://flyex.ams.sunysb.edu/ .



Source: Carroll, 2005.



Source: Thieffry and Sanchez (2003). http://flyex.ams.sunysb.edu/ .



Source: Lawrence (1992). *The Making of a Fly: The Genetics of Animal Design. Blackwell Science Ltd. ISBN 0632030488*

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.

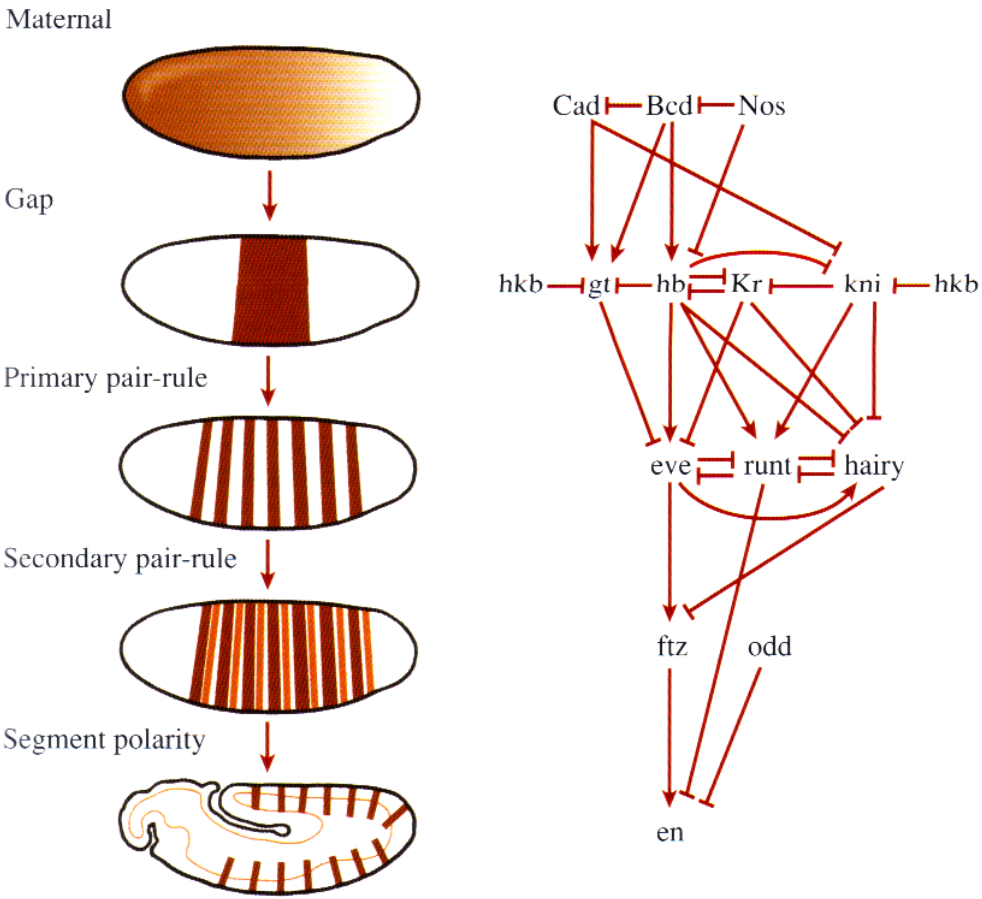
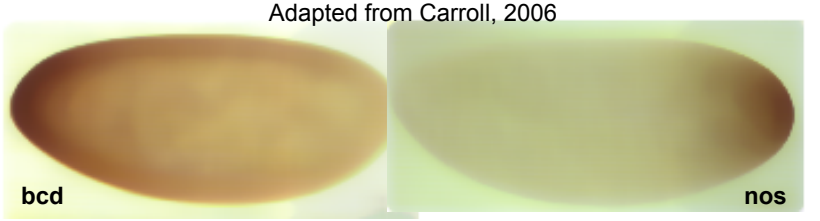


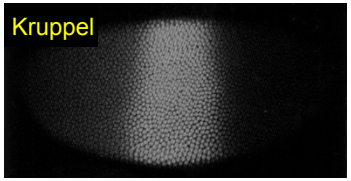
Figure 3.5
The segmentation genetic regulatory hierarchy

(**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.

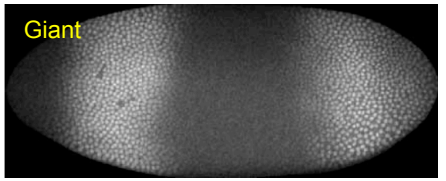
Source: Carroll, 2005. From DNA to diversity (2nd edition). Blackwell Publishing.

Adapted from Carroll, 2006

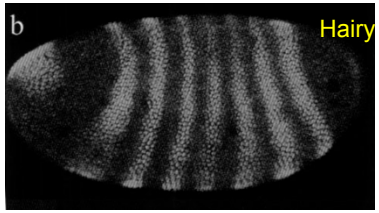




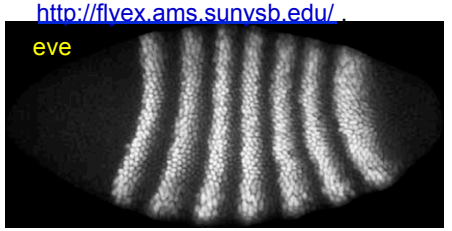
Source: Carroll, 2005.



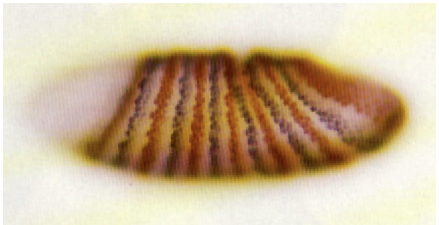
Source: Thieffry and Sanchez (2003). http://flyex.ams.sunysb.edu/ .



Source: Carroll, 2005.



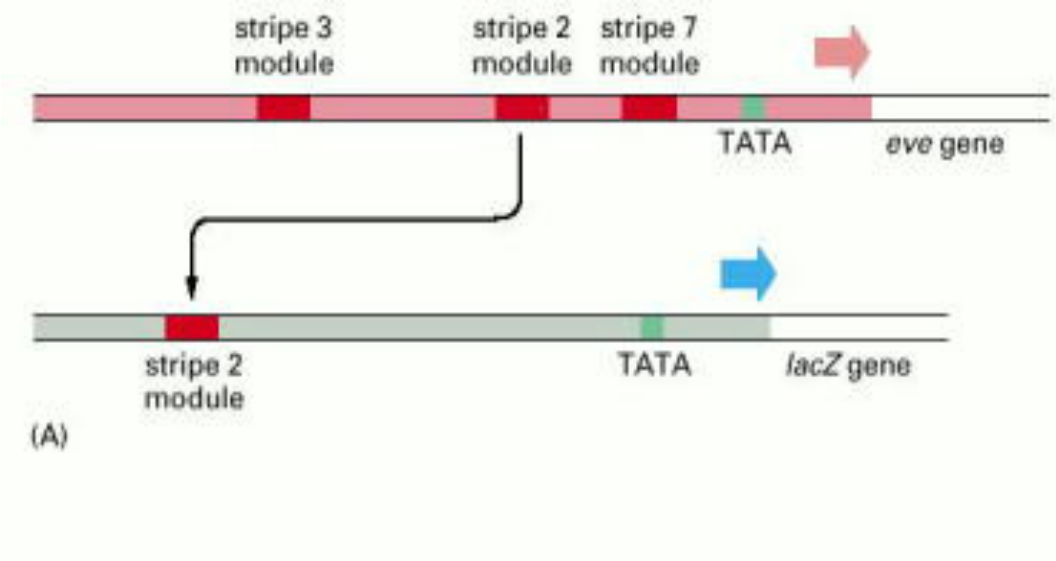
Source: Thieffry and Sanchez (2003). http://flyex.ams.sunysb.edu/ .

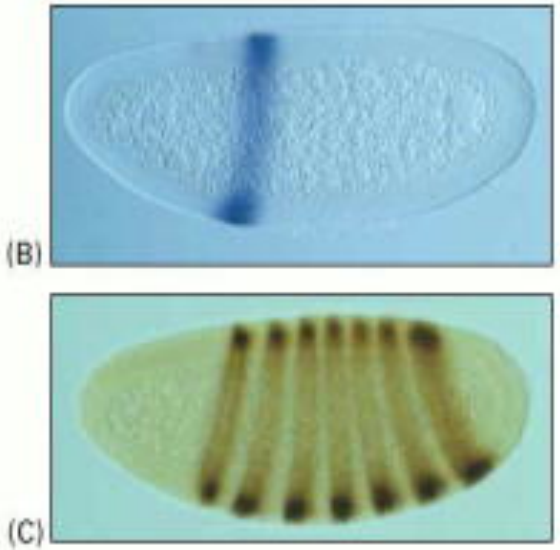


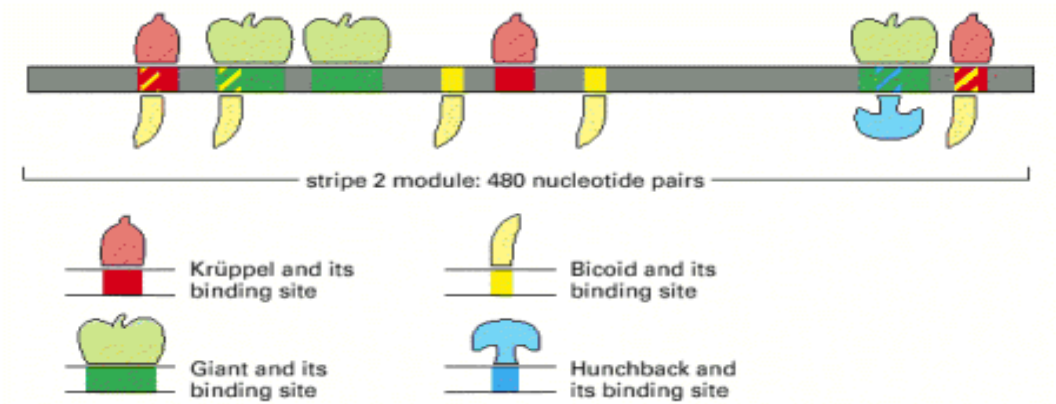
Source: Lawrence (1992). *The Making of a Fly: The Genetics of Animal Design. Blackwell Science Ltd. ISBN 0632030488*

The stripe-specific enhancers of Drosophila even-skipped (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.

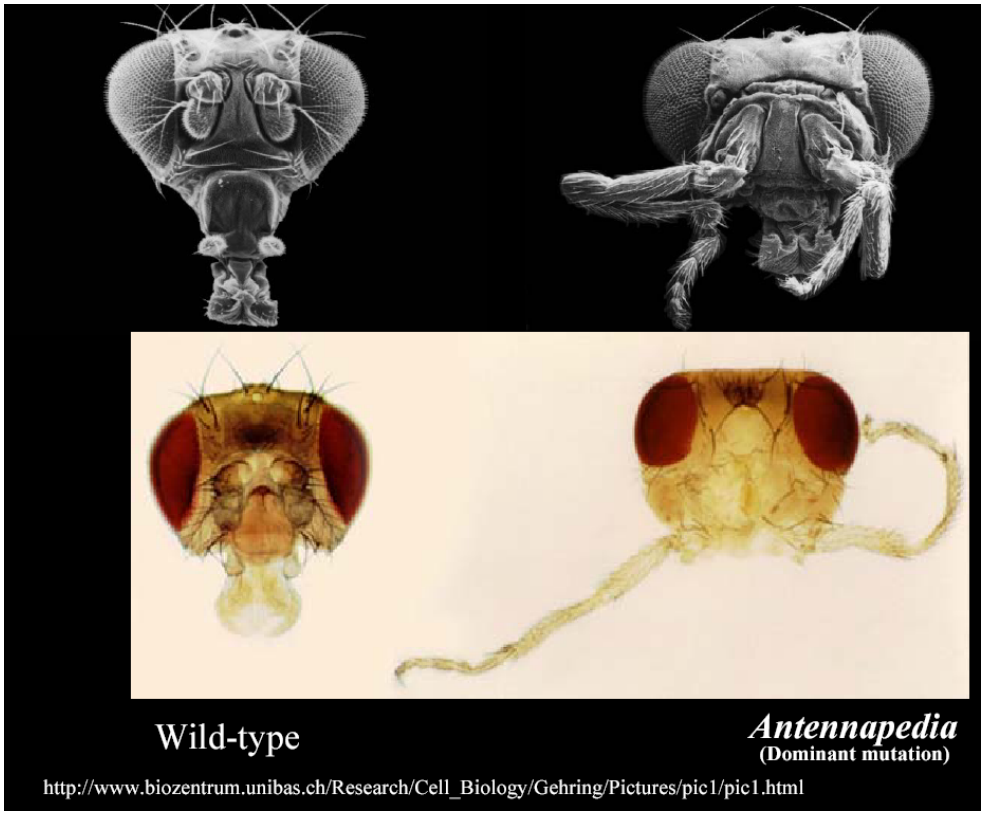




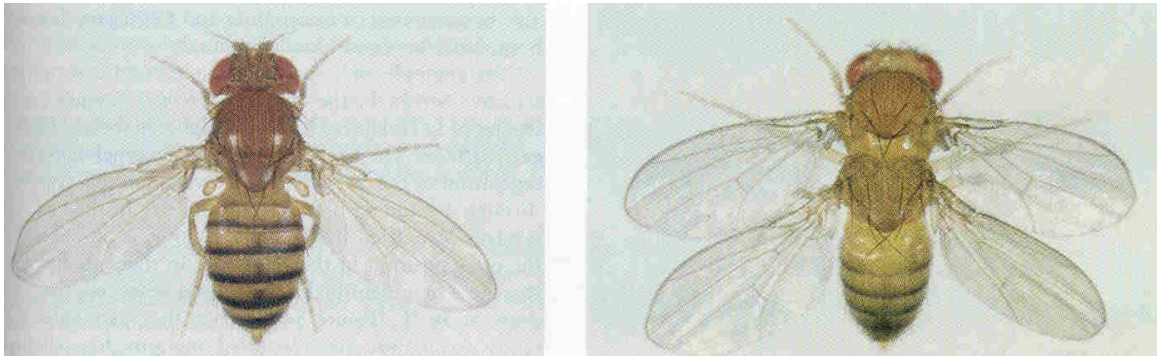


Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular Biology of the Cell (2002).

Homeotic mutations



- Mutations of the Hox genes modify the segmental identity.
- Antennapedia mutant fly: legs develop at the location of antennae.

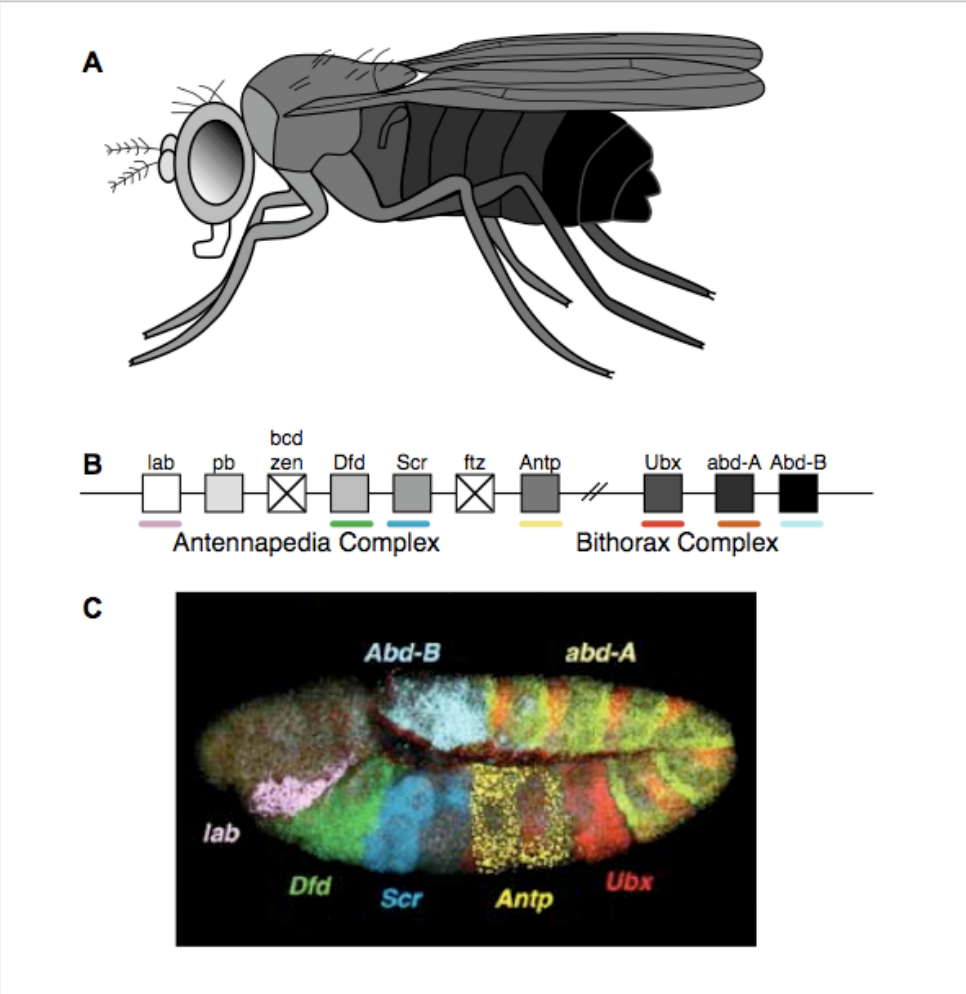


http://www.cb.ku.ac.th/down_load/drsermsiri.ppt

- Bithorax complex (triple mutant): the 3rd thoracic segment (metathorax) develops as a copy of the second segment (mesothorax), with wings instead of haltera.

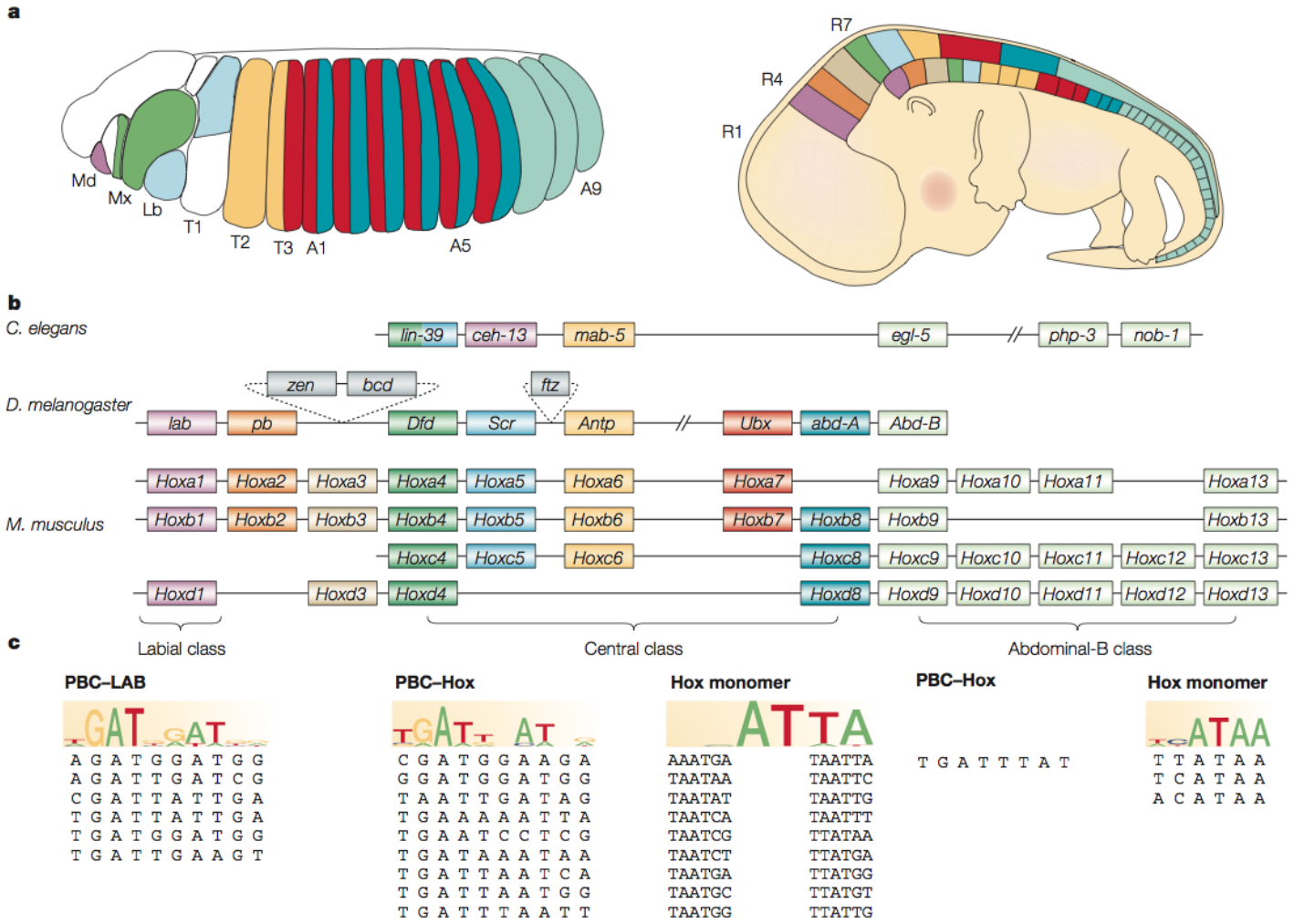
Specification of segmental identity

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



Sources of the Figures:
•Morgane Thomas-Chollier (2008). PhD Thesis, ULB
•Lemons & McGinnis (2006).

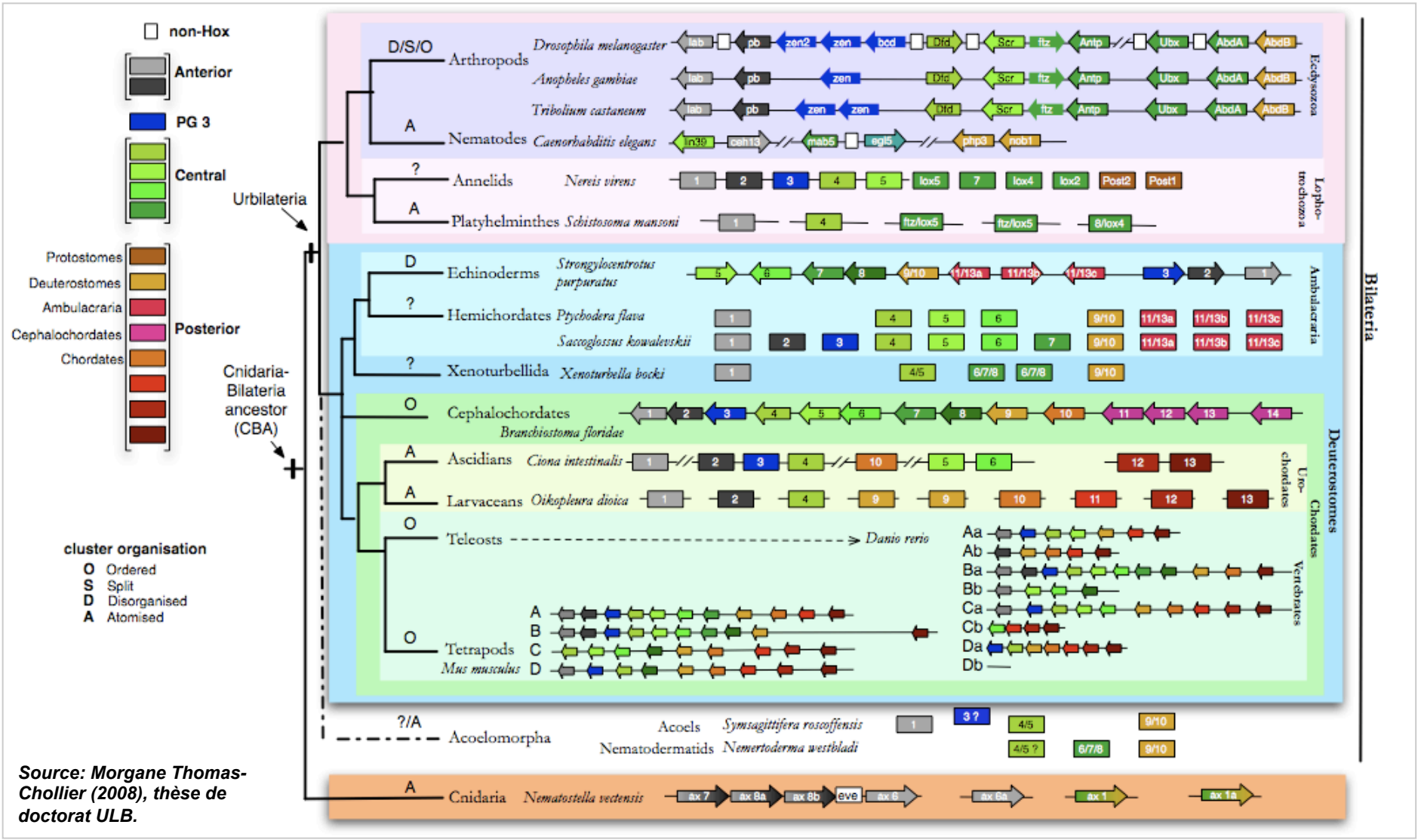
The Hox complex - from drosophila to mammals



•Morgane Thomas-Chollier (2008). PhD Thesis, ULB

Hox evolution: complexification by duplication/divergence

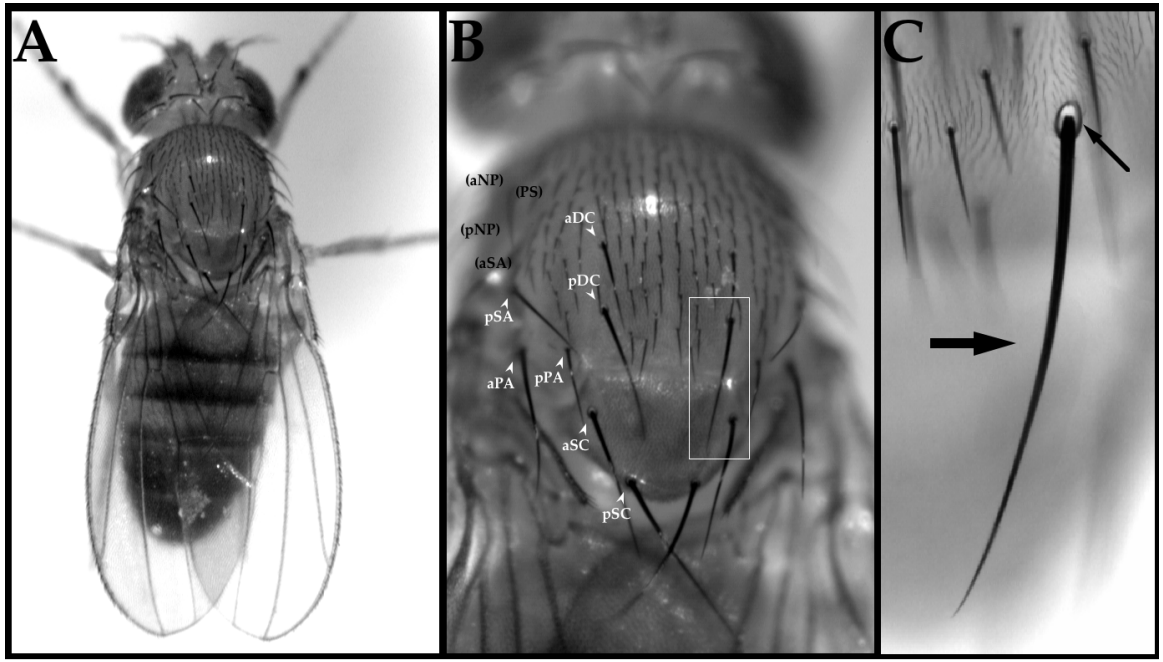
- Hox genes are found in all the Bilaterians, and they determine segmental identity.
- The topological organization of the complex has been partly conserved from invertebrate to vertebrate.
- The whole complex has been duplicated several times during evolution



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.
- Sensry 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

Loss of function (sc$^{10.1}$)

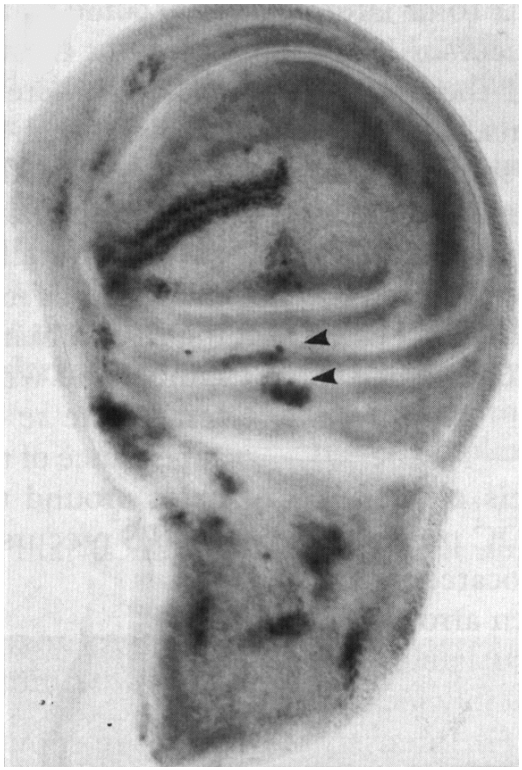
Gain of function (Hw)

Time-controlled expression rescues specific bristles (sc$^{10.1}$ + hs-sc)

Figures: J.van Helden (1995). PhD Thesis. ULB.

Gènes proneuraux : le complexe achaete-scute

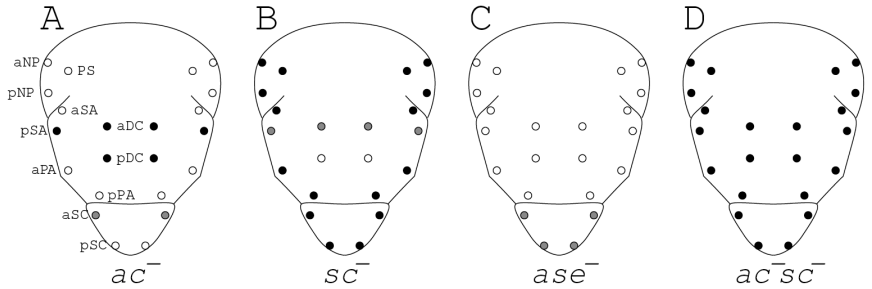
- Le ***complexe achaete-scute*** inclut 4 gènes codant pour des facteurs transcriptionnels paralogues.
- Ces gènes s'expriment dans des groupes de cellules durant le développement: les ***groupes proneuraux***.

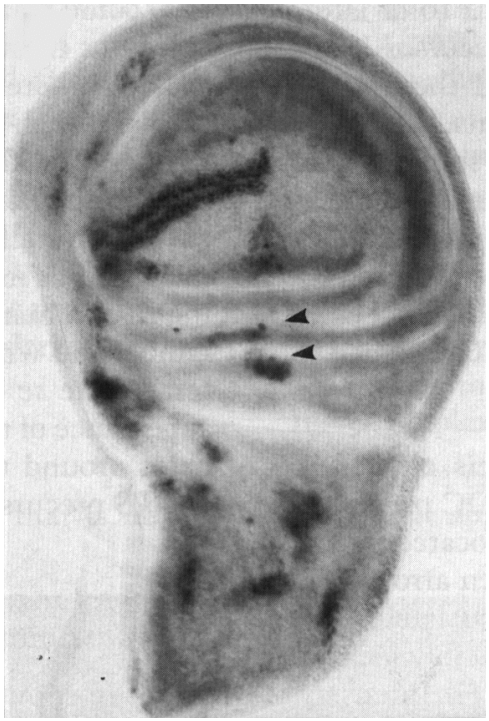


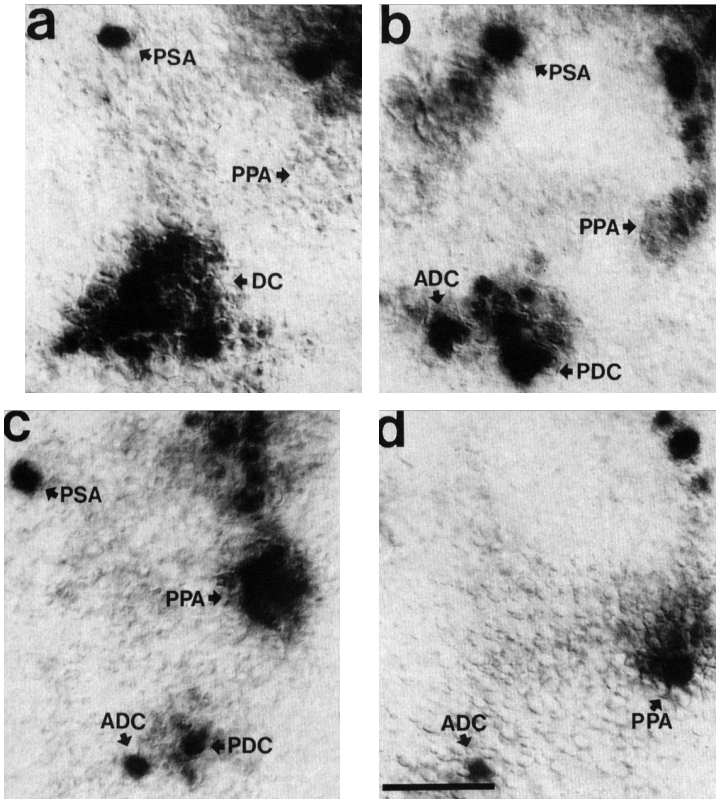
J.van Helden (1995).
Thèse de doctorat ULB.

L'expression d'achate-scute détermine la formation de soies

- The deletion of each specific gene of the achaete-scute genes leads to the absence of a specific subsets of sensory bristles (black dots on the top schemas).
- The simultaneous deletion of both achaete and scute leads to the total absence of sensory bristles.
- In the wing imaginal disc, the specific deletions are characterized by the absence of the corresponding clusters of expression of achaete and scute.



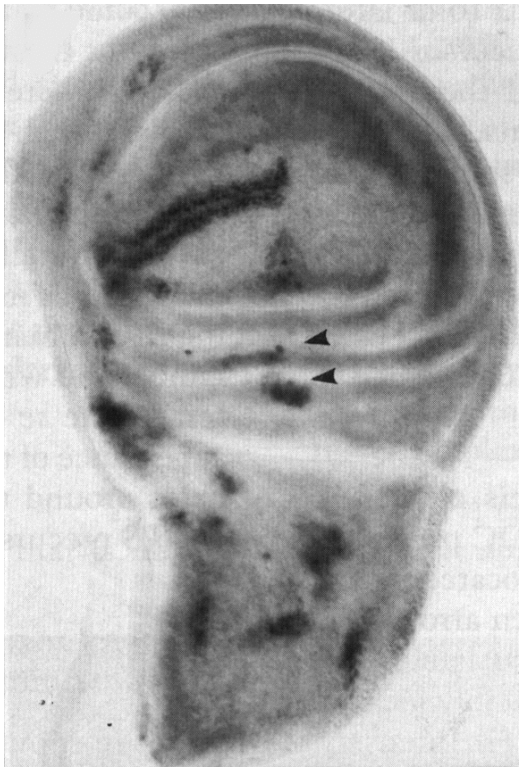


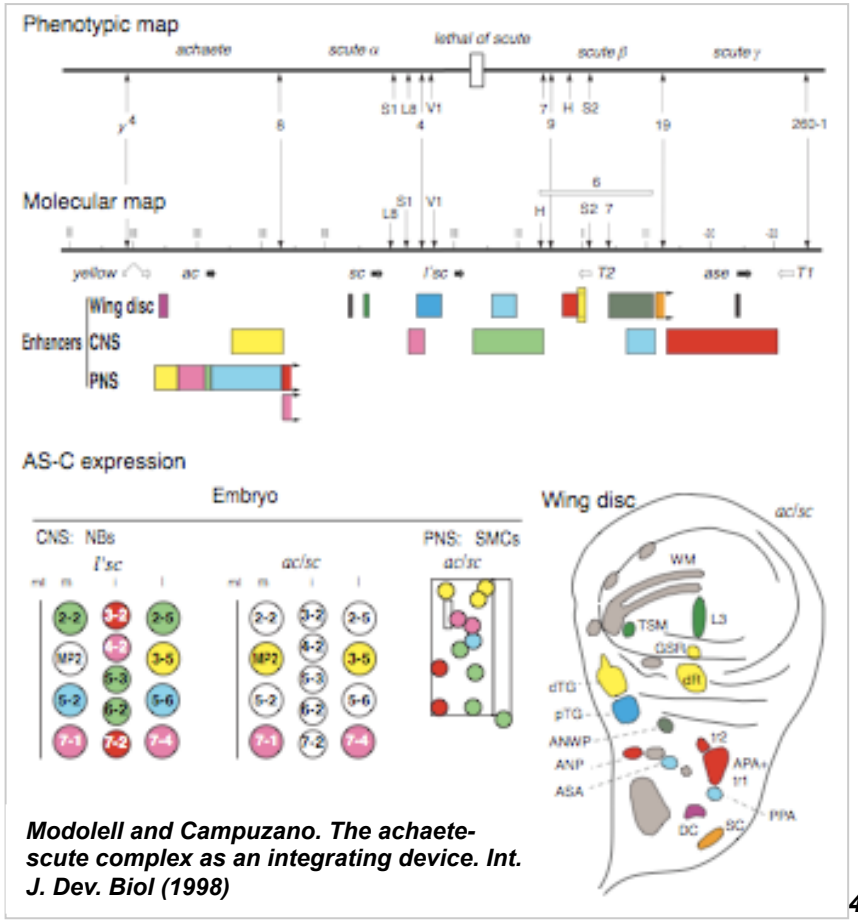


Position-specific enhancers in the achaete-scute complex

- The ***achate-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.
- This extremely complex, precise and reproducible expression pattern is determined by the action f specific cis-regulatory elements located in the 100kb region encompassing the 4 genes of the achaete-scute complex (ac, sc, l'sc and ase). Most of the region is made of non-coding sequences containing ***time- and position-specific enhancers***.

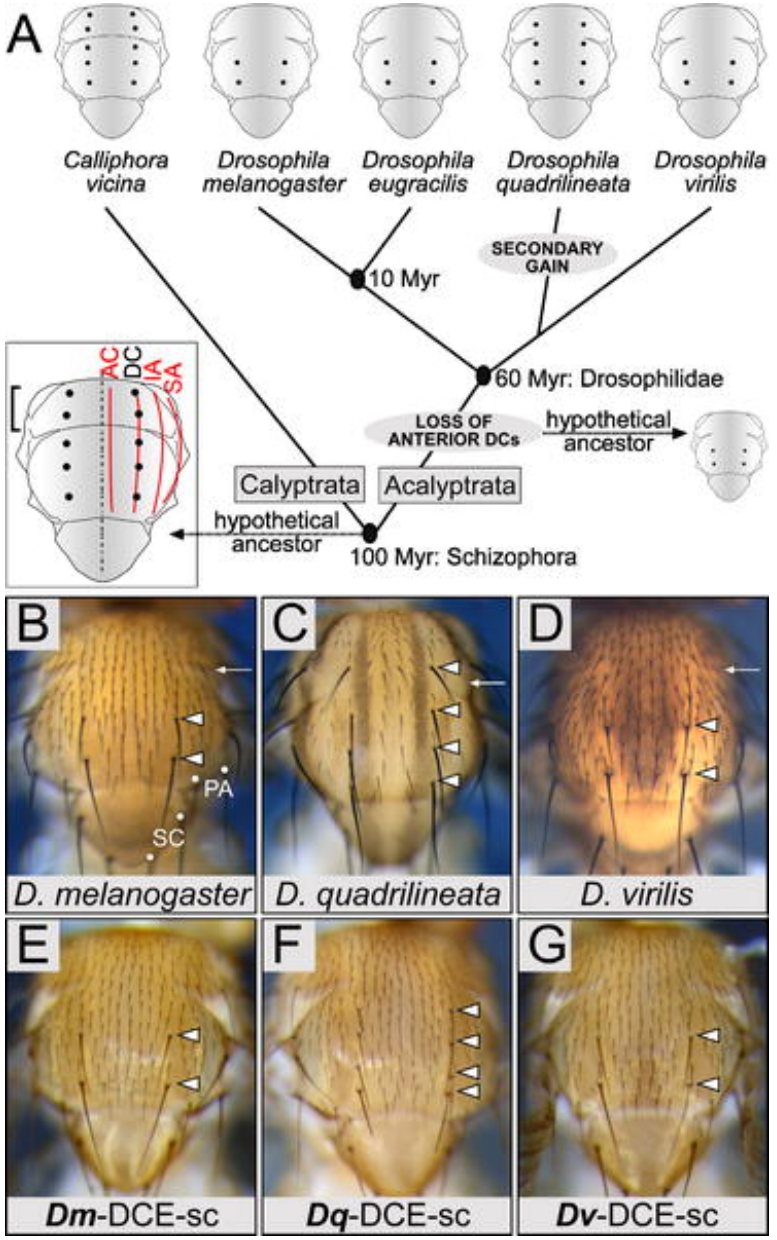
J.van Helden (1995). PhD thesis, ULB, 1995.





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

Species-specificity of the developmental patterns



Marcellini et al. PLoS Biol (2006) vol. 4 (12) pp. e386

Cis-regulatory regions

Cis-regulatory elements and their organization

- The localization of cis-regulatory regions varies depending on the type of organism.

organism	Bacteria	Fungi	Metazoa
location	upstream overlap. Initiation	upstream	upstream downstream intergenic regions 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

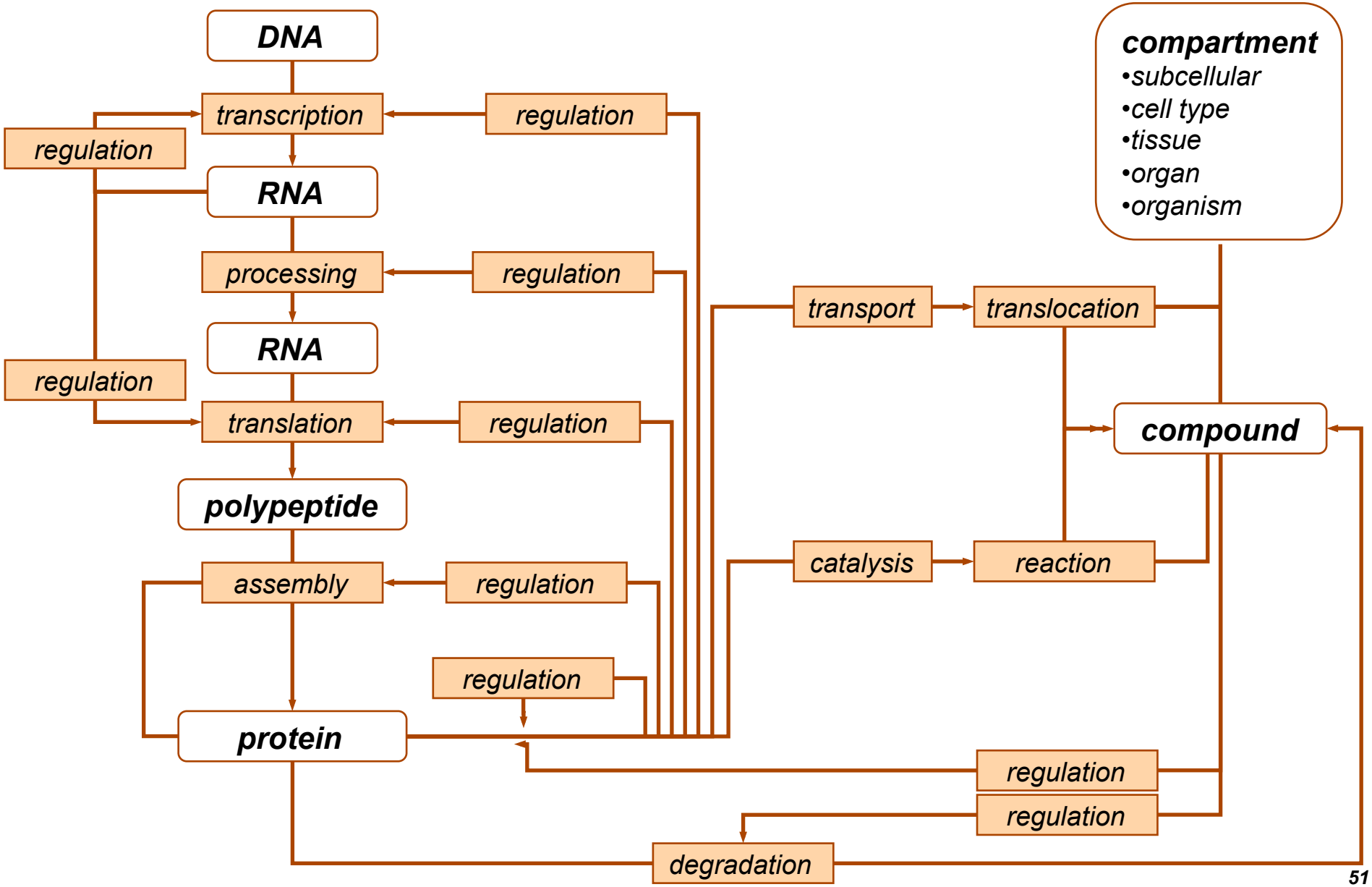
Questions and approaches

Questions and approaches

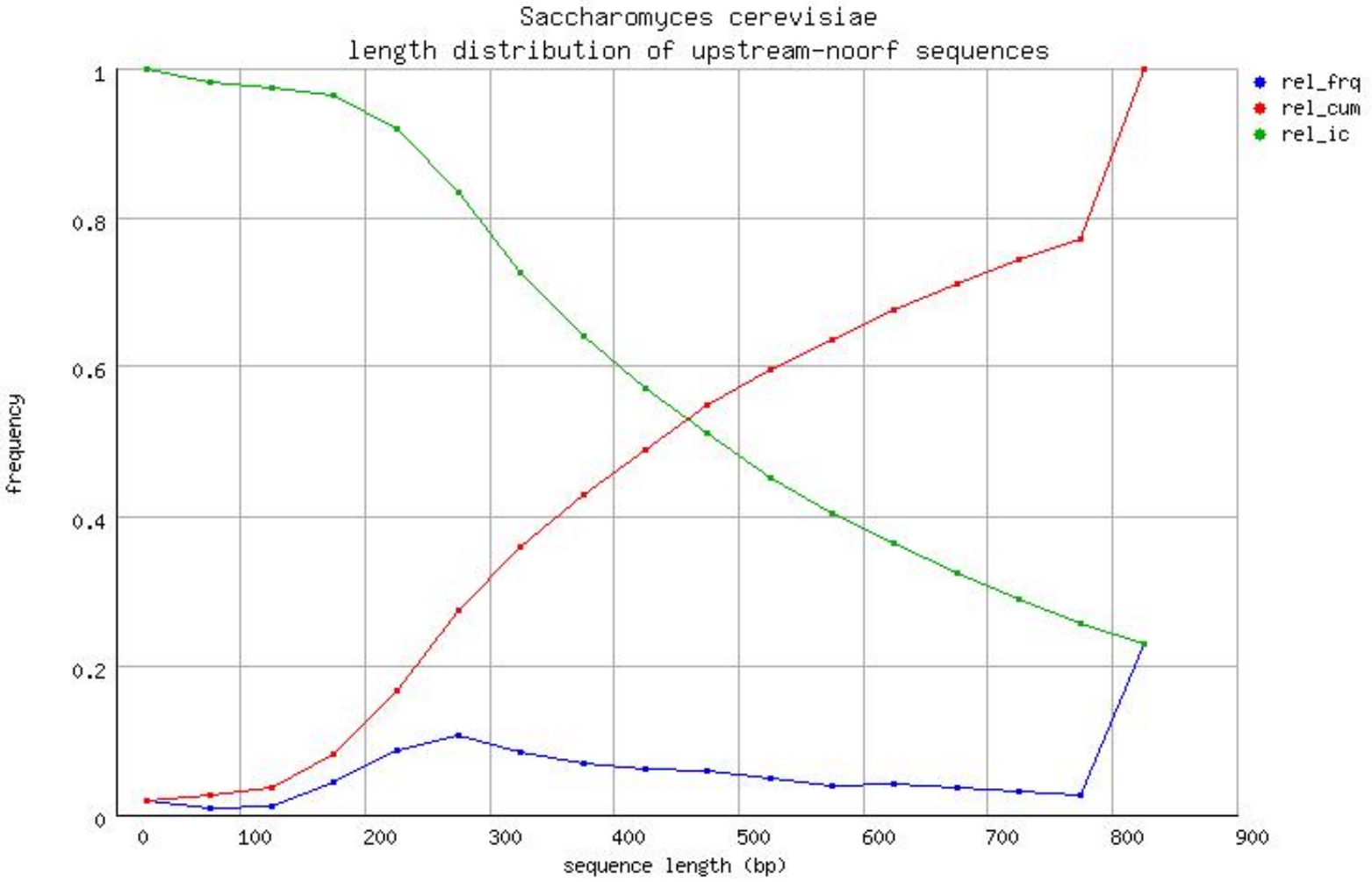
- Pattern matching
  - If we know the consensus for a given transcription factor, can we predict its binding sites in a DNA sequence ?
- Matching a library of patterns
  - Can we scan a sequence for matches with the consensus of all he currently known transcription factor ?
- CRM prediction
  - Detect regions with a higher density of predicted sites than expected by chance (cis-regulatory enriched regions, CRERs).
- Motif discovery
  - Starting from a set of supposedly co-regulated genomic regions (promoters, ChIP-seq peaks), can we predict transcription factor binding motifs 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 ?
- Classification of genomic sequences based on pattern scores
  - Can we classify regulatory regions (promoters, ChIP-seq peaks, enhancers) 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 “discovered” 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.

Supplementary material

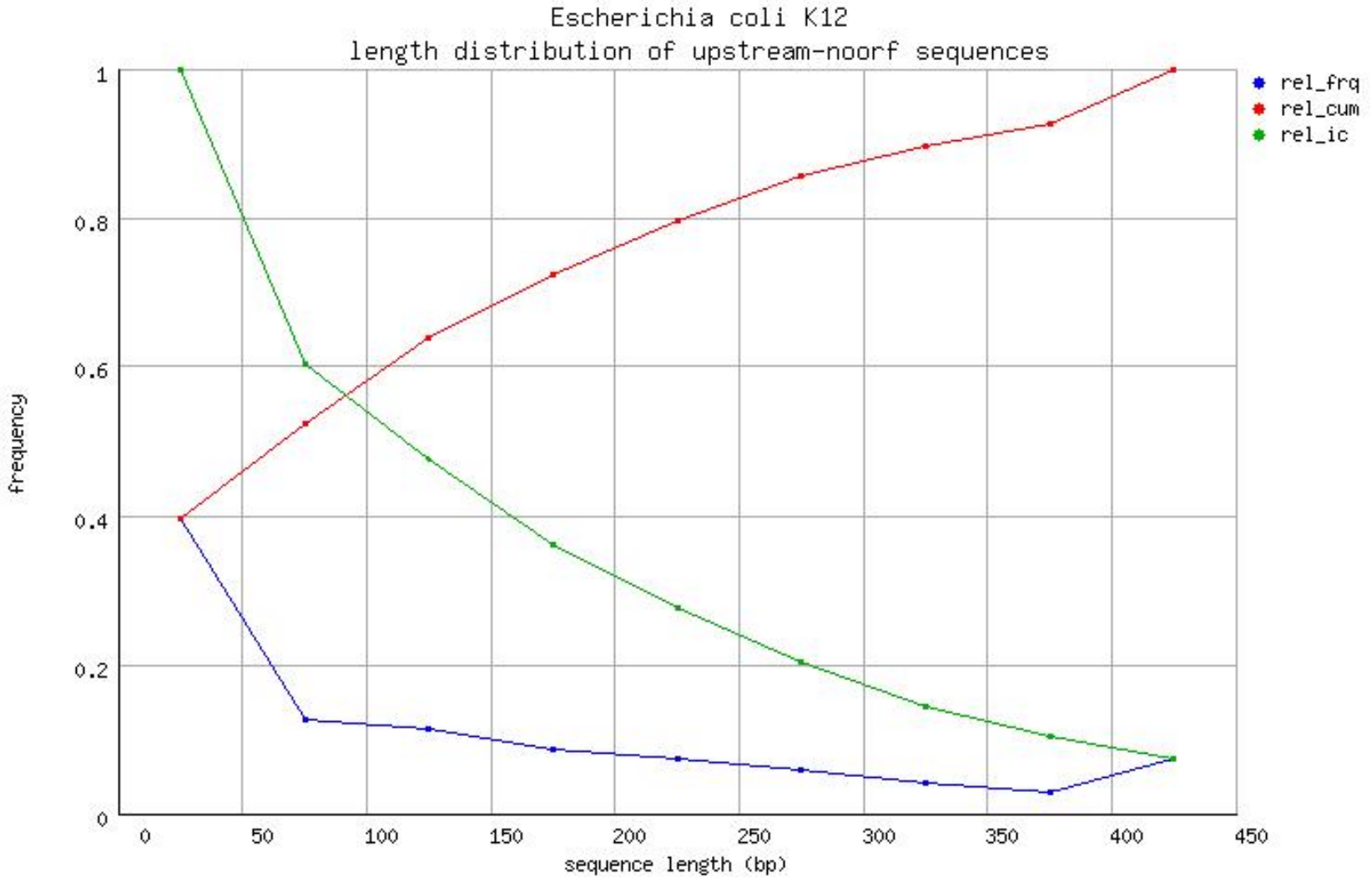
Molecular networks (shamefully simplified)



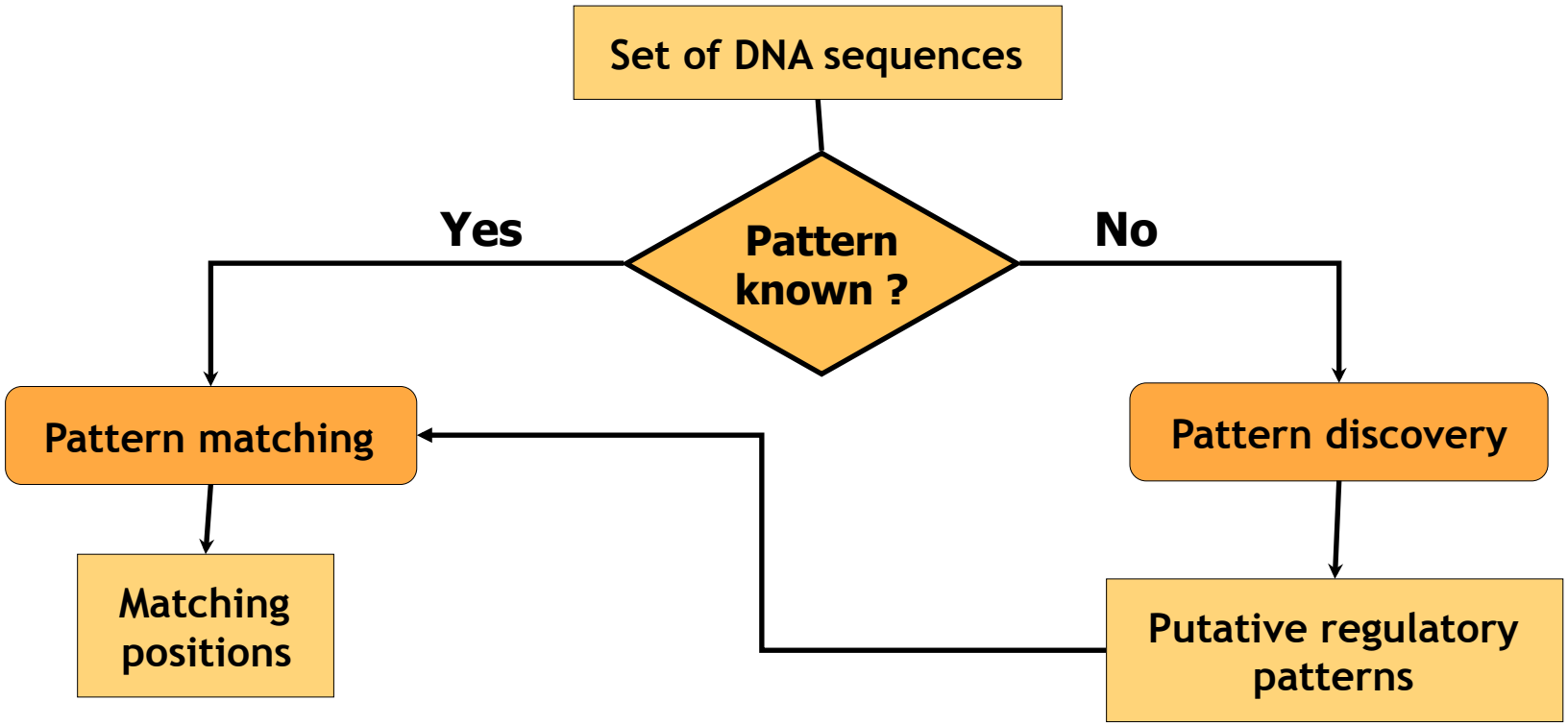
Distribution of upstream sequence lengths
Saccharomyces cerevisiae



Distribution of upstream sequence lengths
Escherichia coli K12



Pattern matching vs pattern discovery



Met4p binding sites

gene	start	end	sequence
MET3	-367	-349	GAAAAG**TCACGTG**TAATTT
MET3	-384	-366	AAAAGG**TCACGTGA**CCAGA
MET14	-235	-217	CTAATT**TCACGTGA**TCAAT
MET16	-185	-167	ATCATT**TCACGTG**GCTAGT
ECM17	-311	-293	ATTTCA**TCACGTG**CGTATT
ECM17	-339	-321	.TTTGTC**CACGTGA**TATTTC
MET10	-255	-237	.CCACAC**CACGTGA**GCTTAT
MET10	-237	-219	.TAGAAG**CACGTGA**CCACAA
MET2	-360	-342	GTATTT**TCACGTGA**TGCGC
MET2	-554	-536	TAATAA**TCACGTGA**TATTT
MET17	-306	-288	.AAATGG**CACGTGA**AGCTGT
MET17	-332	-314	TTGAGG**TCAC**A**TGA**TCGCA
MET6	-540	-522	GCCACA**TCACGTGCA**CATT
MET6	-502	-484	AATATT**TCACGTGA**CTTAC
SAM2	-329	-311	.TCTACC**CACGTGA**CTATAA
SAM2	-381	-363	.TCTTCA**CA**T**GTGA**TTCATC

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**AAAAA**A**TGTGG**CAATGG
MET2	-313	-293	TGC**AAAAAA**T**TGTGG**ATGCAC
MET17	-227	-207	TCATG**AAAACTGTG**TAACATA
MET6	-313	-293	GTCGC**AAAACTGTGG**TAGTCA
SAM2	-306	-286	GCTTG**AAAACTGTGG**CGTTTT
SAM1	-283	-263	ACAGG**AAAACTGTGG**TGGCGC
MET19	-173	-153	ATAAGC**AAACTGTGG**GTTCAT
MUP3	-188	-168	CGG**AAAAAACTGTGG**CGTCGC
MET8	-184	-164	GG**AAAAAAA**A**TGTG**AAAATCG
MET1	-232	-212	CATAAT**AAACTGTG**AACGGAC
MET3	-259	-239	ACAAAG**CCACAGTTTT**ACAAC
MET28	-159	-139	CTAACA**CCACAGTTTT**GGGCG
MET8	-434	-414	TCTTGT**CCGCAGTTTT**ATCTG
MET30	-168	-148	GGGAAG**CCACAGTTT**GCGCGG
MET6	-405	-385	CTATCGAA**CTCGTTT**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**CACGCTT**aatat
PHO81	-350	-332	...TTATGG**CACGTGCG**AATAA
PHO84	-421	-403	..TTTCCAG**CACGTGGG**GCGG
PHO84	-442	-425	...TAGTTC**CACGTGG**ACGTG
PHO84	-879	-874	.aaaagtgt**CACGTG**ataaaaat
PHO84	-267	-250	..taatacg**CACGTTTTT**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**	**A**denine
C	**C**	**C**ytosine
G	**G**	**G**uanine
T	**T**	**T**hymine
R	**A** or **G**	pu**R**ine
Y	**C** or **T**	p**Y**rimidine
W	**A** or **T**	**W**eak hydrogen bonding
S	**G** or **C**	**S**trong hydrogen bonding
M	**A** or **C**	a**M**ino group at common position
K	**G** or **T**	**K**eto 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**	a**N**y

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

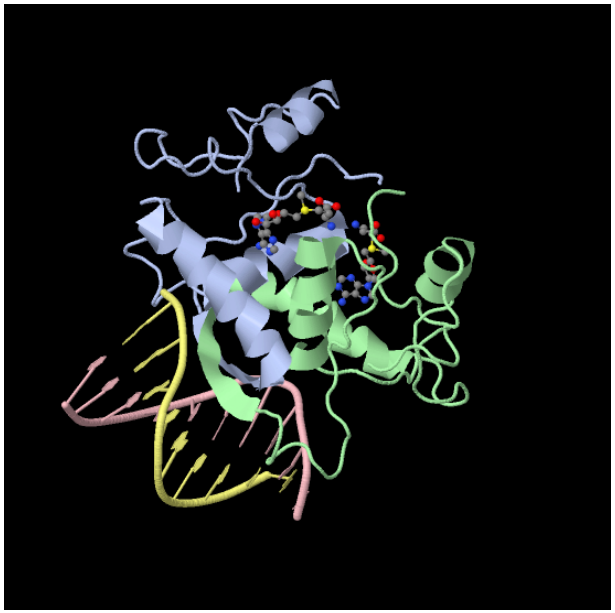
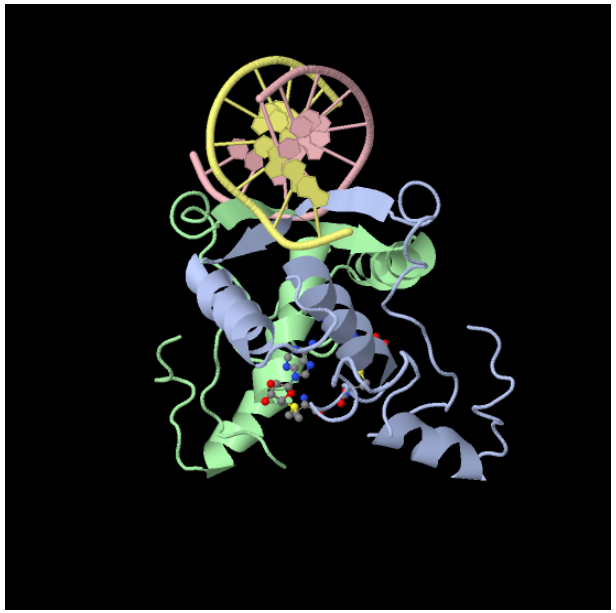
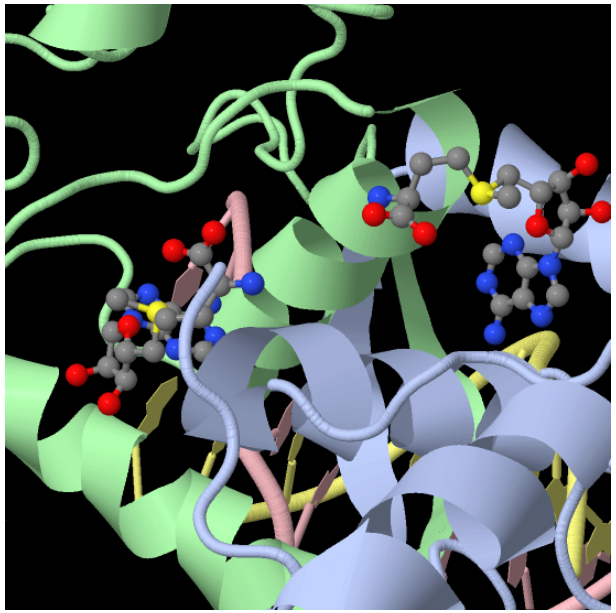
Binding motif for the yeast Pho4p transcription factor

Source : SCPD

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

Methionine repressor

- Crystal structure : the methionine repressor of *Escherichia coli*.
- Green + violet: the MetJ protein forms a homodimer which is able to bind DNA.
- Pink + yellow: the two strands of the DNA binding site
- Detail: the repressor is activated by binding of methionine molecules

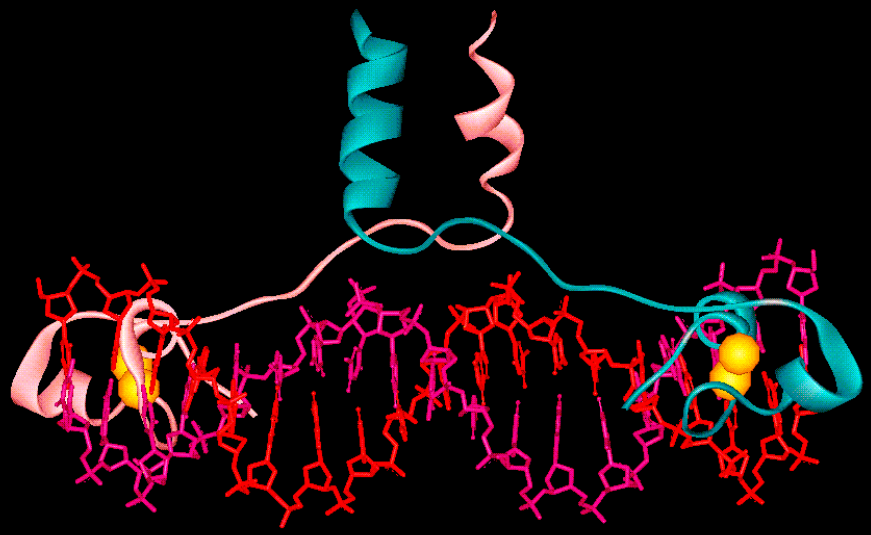




http://www.pdb.org/pdb/explore/jmol.do?structureId=1CMA

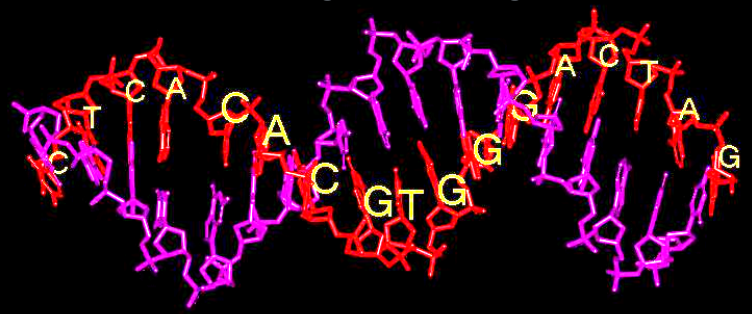
Transcription factor-DNA interfaces

Pho4p (yeast)

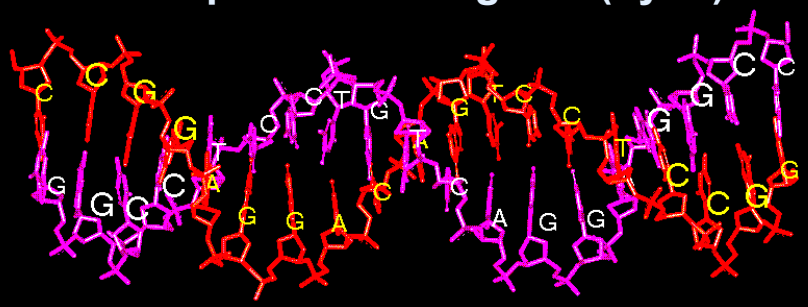
Gal4p (yeast)



Pho4p DNA binding site (oligonucleotide)

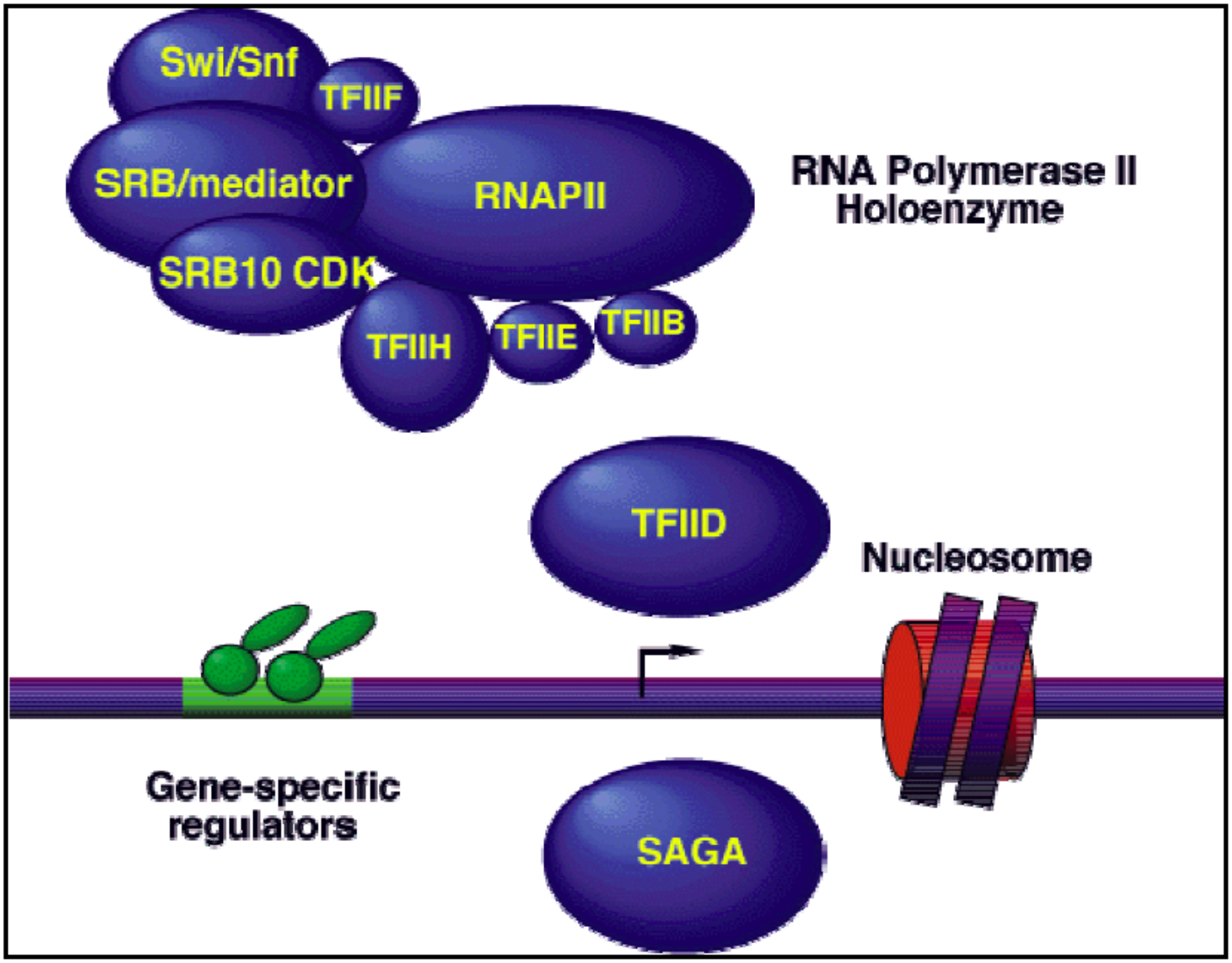


Gal4p DNA binding site (dyad)



The genome challenge

RNA polymerase



The non-coding genome

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

Genomic sequences

- A genome G contains a set of n chromosomes.
  - $G=\{S_1,S_2,...,S_i,...,S_n\}$
- Each chromosome is a molecule of dexyribonucleic 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)
  - $A=\{A,C,G,T\}$
  - $Si=(s_{i1},s_{i2},...,s_{ij},...,S_{iLi})$
  - L_i is the length of the i^{th} chromosome