Regulatory Sequence Analysis

Regulatory regions and regulatory elements

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

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

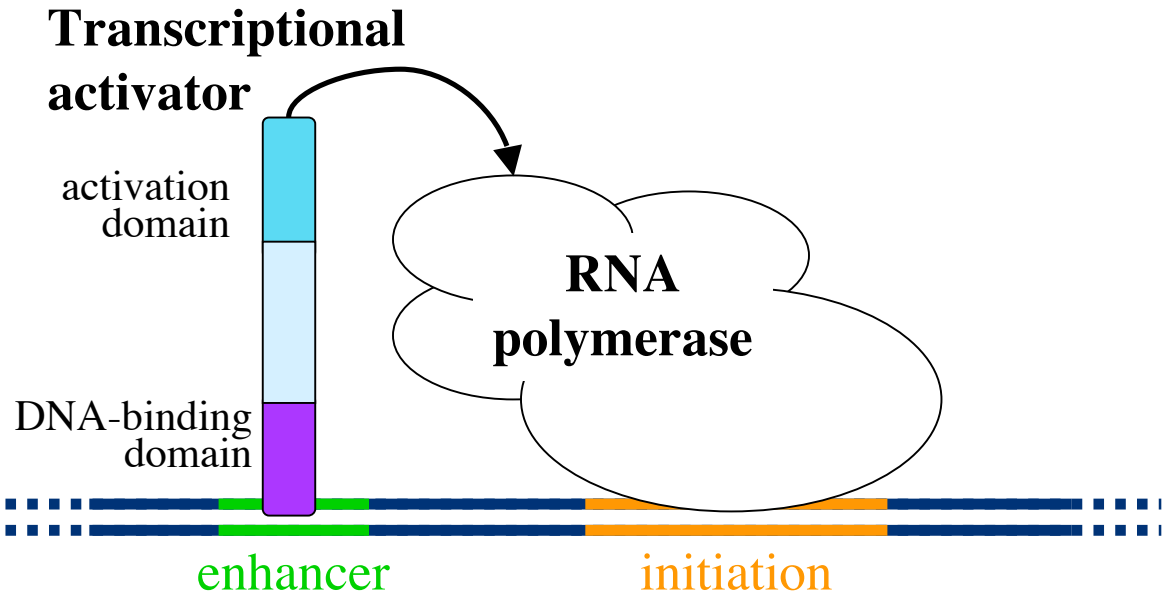
The non-coding genome

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

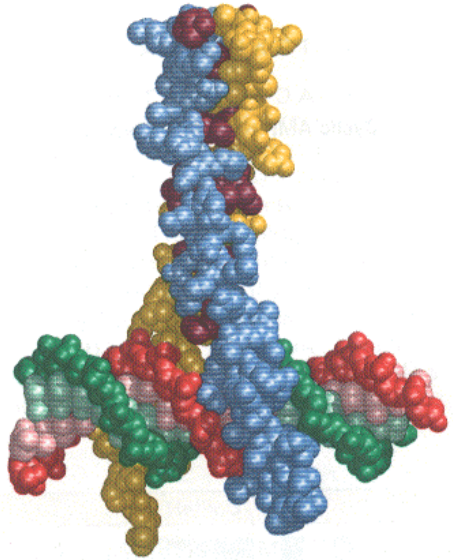
Genome sizes - some examples

Nom d'espèce	Nom commun	Année de publication	Taille du génome Mb	Nombre de gènes	Distance moyenne entre gènes Kb	Fraction couverte par des gènes codants %	Fraction non-codante %	Fraction répétitive %	Fraction transcrite %	Remarques
Bactérie										
Mycoplasma genitalium	*Mycoplasma*	1995	0.6	481	1.2	90	10			Petit génome (intracellulaire)
Haemophilus influenzae		1995	1.8	1 717	1.0	86	14			Premier génome bactérien séquencé
Escherichia coli	*Entérobactérie*	1997	4.6	4 289	1.1	87	13			
Levures										
Saccharomyces cerevisiae	*Levure du boulanger*	1996	12	6 286	1.9	72	28			Premier génome eucaryote
Animaux										
Caenorhabditis elegans	*Ver nématode*	1998	97	19 000	5	27	73			Premier génome de métazoaire
Drosophila melanogaster	*Mouche à vinaigre*	2000	165	16 000	10	15	85			
Ciona intestinalia			174	14 180	12					
Danio rerio	*Poisson zèbre*		1 527	18 957	81					
Xenopus laevis	*Xénope (amphibien)*		1 511	18 023	84					
Gallus gallus	*Poule*		2 961	16 736	177					
Ortnithorynchus anatinus	*Ornithorynque*		1 918	17 951	107					
Mus musculus	*Souris*	2002	3 421	23 493	146					
Pan troglodytes	*Chimpanzé*		2 929	20 829	141					
Homo sapiens	*Humain*	2001	3 200	21 528	149	2	98	46	28	Version "brouillon"
1000 génomes humains		> 2008								Projet annoncé en janvier 2008
Plantes										
Arabidiopsis thaliana	*Arabette*	2001	120	27 000	4	30	70			Premier génome de plante
Oryza sativa	*Riz*		390	37 544	10					
Zea mais	*Maïs*		2 500	50 000	50			50		Nb de gènes approximatif
Triticum aestivum	*Blé*		16 000							Génome hexaploÏde
Lilium	*Lys*		120 000							
Psilotum nudum			250 000							

Transcriptional activation

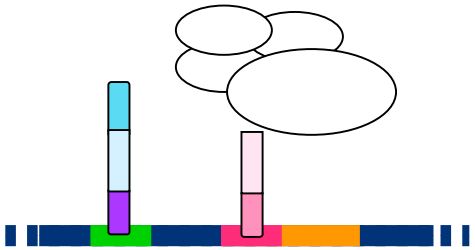


GCN4
(leucine-zipper)
binding to DNA

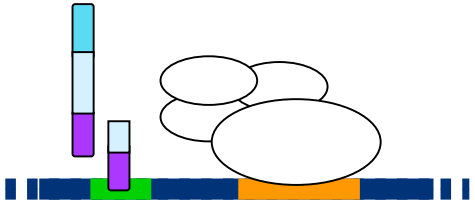


Source: L.Stryer, (1995). Biochemistry. p1003

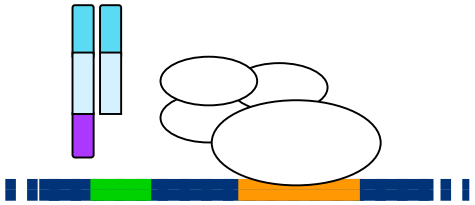
Transcriptional repression



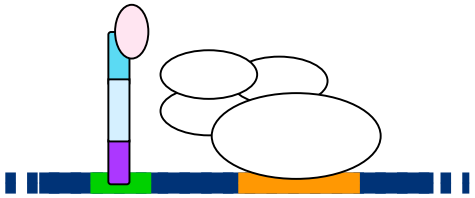
Prevent RNA polymerase from accessing DNA



Competition for factor binding site

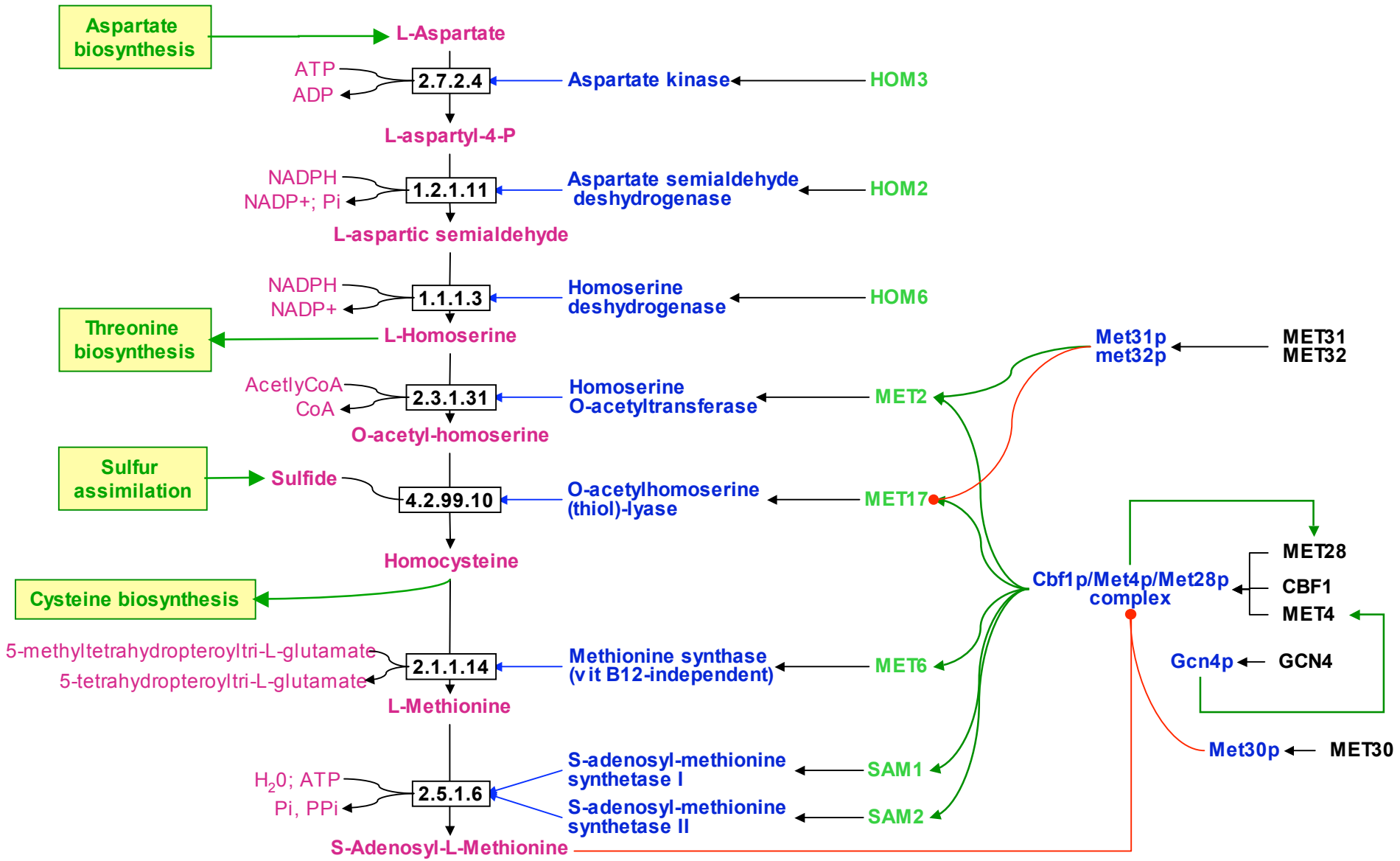


Factor titration

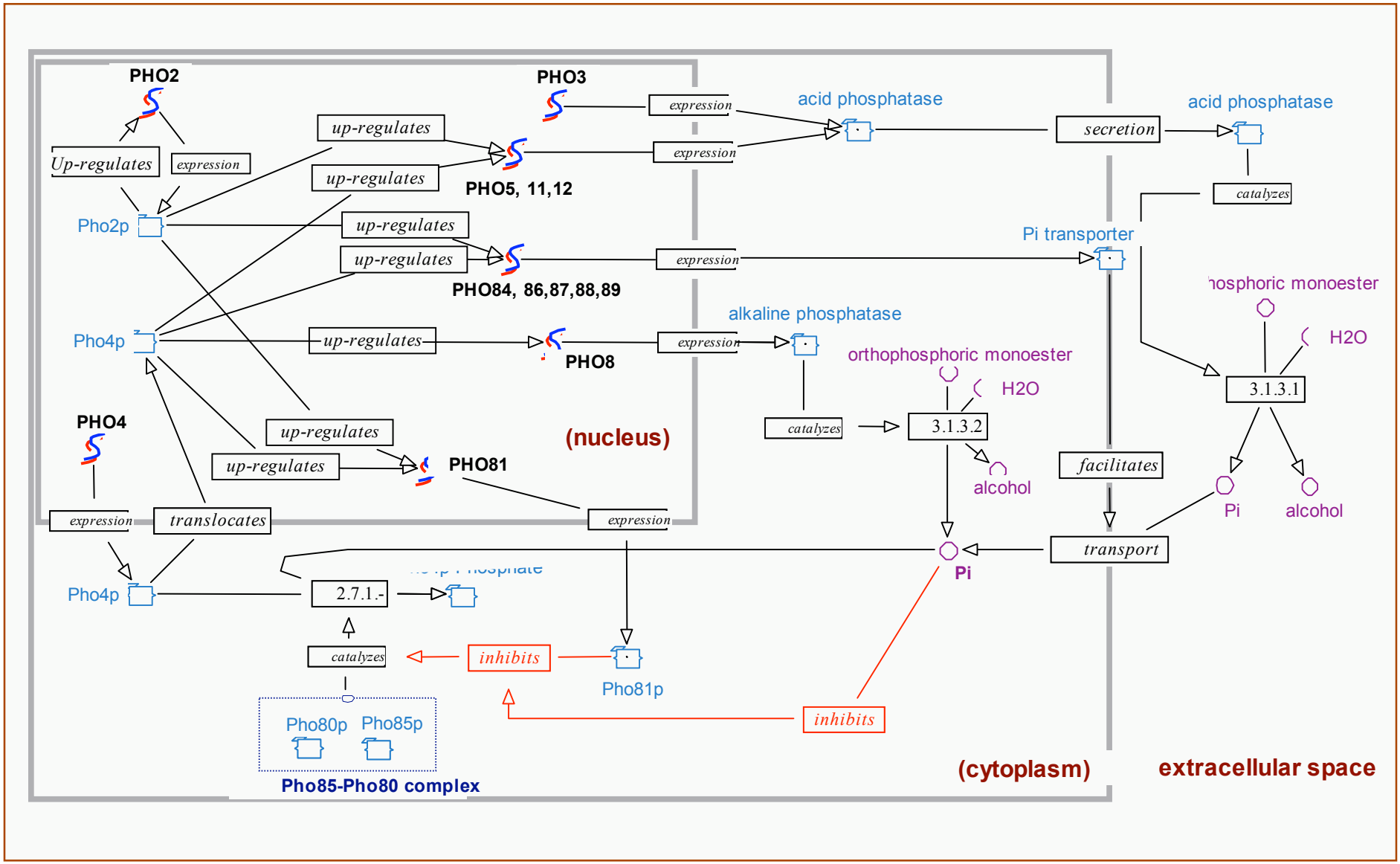


Prevent transcription factor from interacting with RNA-polymerase (bind with activation domain)

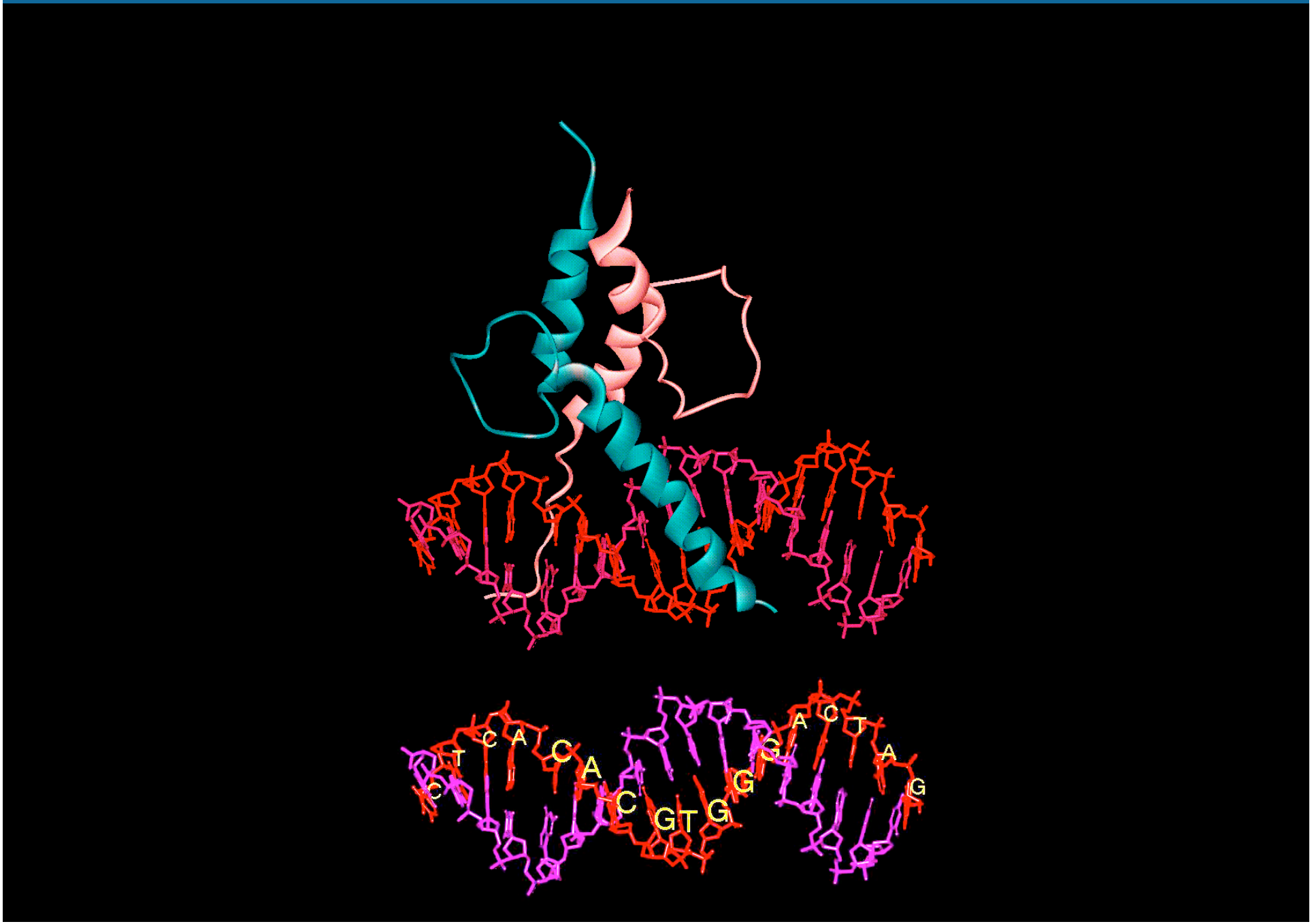
Methionine Biosynthesis in S.cerevisiae



Phosphate utilization in yeast



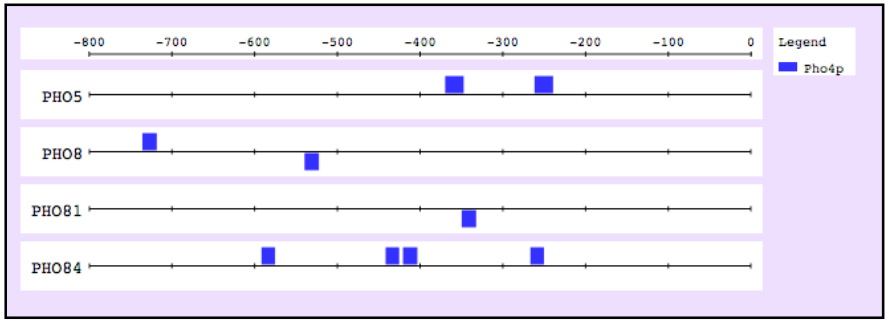
Interface between the yeast Pho4p protein and one of its binding sites



Transcription factor binding site (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	aaaagtgt**CACGTG**ataaaaat
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.
- A site can be
  - characterized experimentally (known site)
  - inferred by some algorithm (predicted site)
- Example
  - binding sites for the yeast transcription factor Pho4p. Coordinates are relative to the start codon.



Alignment of transcription factor binding sites

Binding sites for the yeast Pho4p transcription factor

(Source : Oshima et al. Gene 179, 1996; 171-177)

Gene	Site Name	Sequence	Affinity
PHO5	UASp2	---aCtCaCA**CACGTGGG**ACTAGC-	high
PHO84	Site D	---TTTCCA**GCACGTGGG**GCGGA--	high
PHO81	UAS	----TTATG**GCACGTGC**G**G**AATAA--	high
PHO8	Proximal	GTGATCGCT**GCACGTGG**C**C**CGA---	high
group 1	consensus	---------**gCACGTGgg**-------	high
PHO5	UASp1	--TAAATTA**GCACGTTTT**CGC----	medium
PHO84	Site E	----AATA**C**G**CACGTTTT**TAATCTA	medium
group 2	consensus	--------**cgCACGTTtt**-------	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**T**G**GTGCTG--	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

From binding sites to count matrix

- 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)

R06098 \TCA**CACGTG**GGA\
R06099 \GGC**CACGTG**CAG\
R06100 \TGA**CACGTG**GGT\
R06102 \CAG**CACGTG**GGG\
R06103 \TTC**CACGTG**CGA\
R06104 \ACG**CACGTT**GGT\
R06097 \CAG**CACGTT**TTC\
R06101 \TAC**CACGTT**TTC\

Count matrix

Alignment of Pho4p binding sites (TRANSFAC annotations)

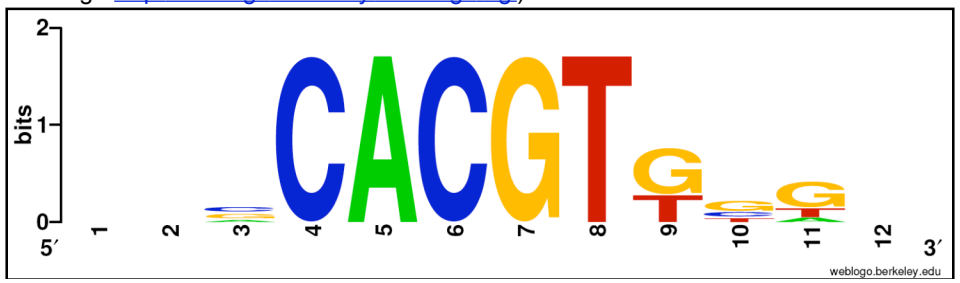
R06098	T	C	A	C	A	C	G	T	G	G	G	A
R06099	G	G	C	C	A	C	G	T	G	C	A	G
R06100	T	G	A	C	A	C	G	T	G	G	G	T
R06102	C	A	G	C	A	C	G	T	G	G	G	G
R06103	T	T	C	C	A	C	G	T	G	C	G	A
R06104	A	C	G	C	A	C	G	T	T	G	G	T
R06097	C	A	G	C	A	C	G	T	T	T	T	C
R06101	T	A	C	C	A	C	G	T	T	T	T	C

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

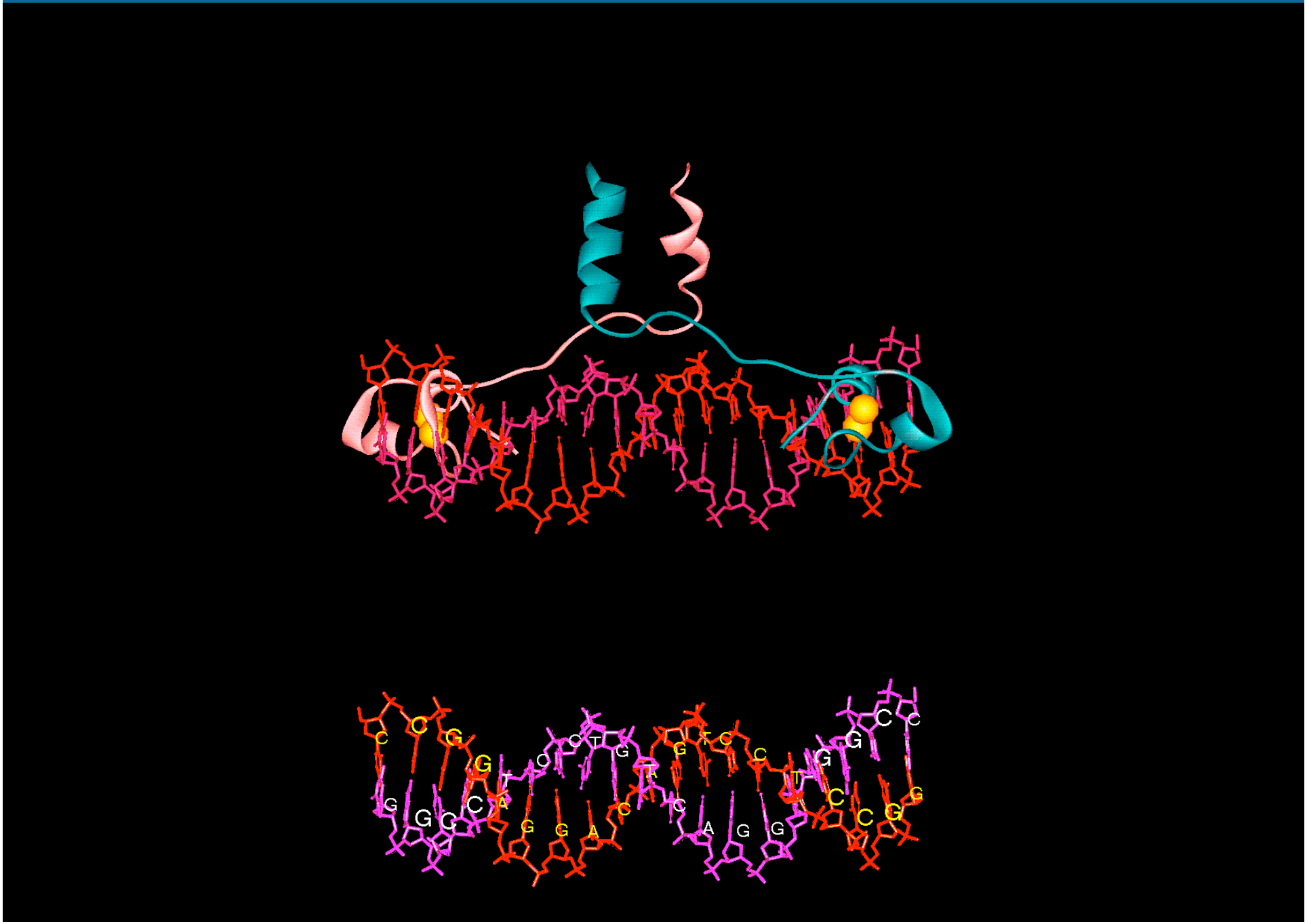
(generated with Web Logo http://weblogo.berkeley.edu/logo.cgi)



Motif / pattern

- We use the term ***motif*** (or ***pattern***) in the sense of a model used to represent the specificity of binding for a transcription factor.
- 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)

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

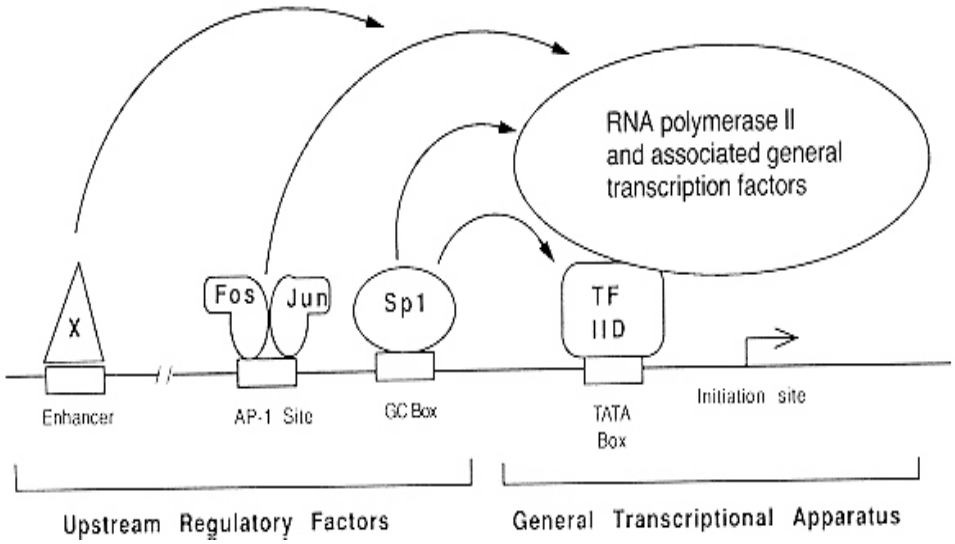


Characteristics of yeast regulatory elements

- Short DNA sequences (5-30 bp)
  - Highly conserved core (5-8 bp), with partly conserved flanking nucleotides
  - Pair of very shot oligonucleotides (3 nt) separated by a non-conserved segment (0-20 bp)

- In the yeast Saccharomyces cerevisiae
  - Located upstream the regulated gene
  - Strand-insensitive
    - Activity does not depend on the strand
  - Wihtin 800 bp from the start codon
    - Activity does not depend on precise position

Cis-regulatory modules (CRM)



- In higher organisms, 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.

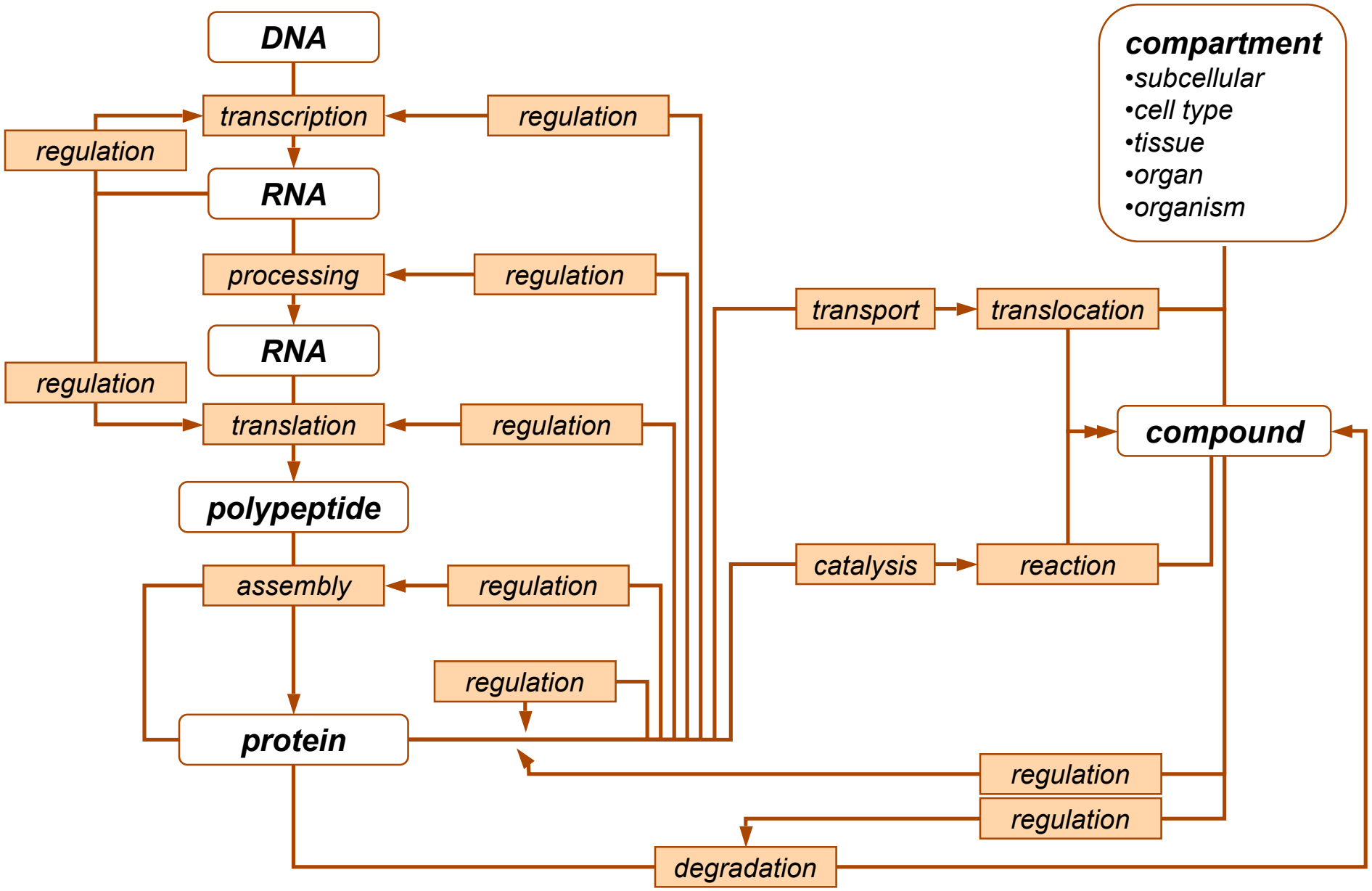
Regulatory regions

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

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)

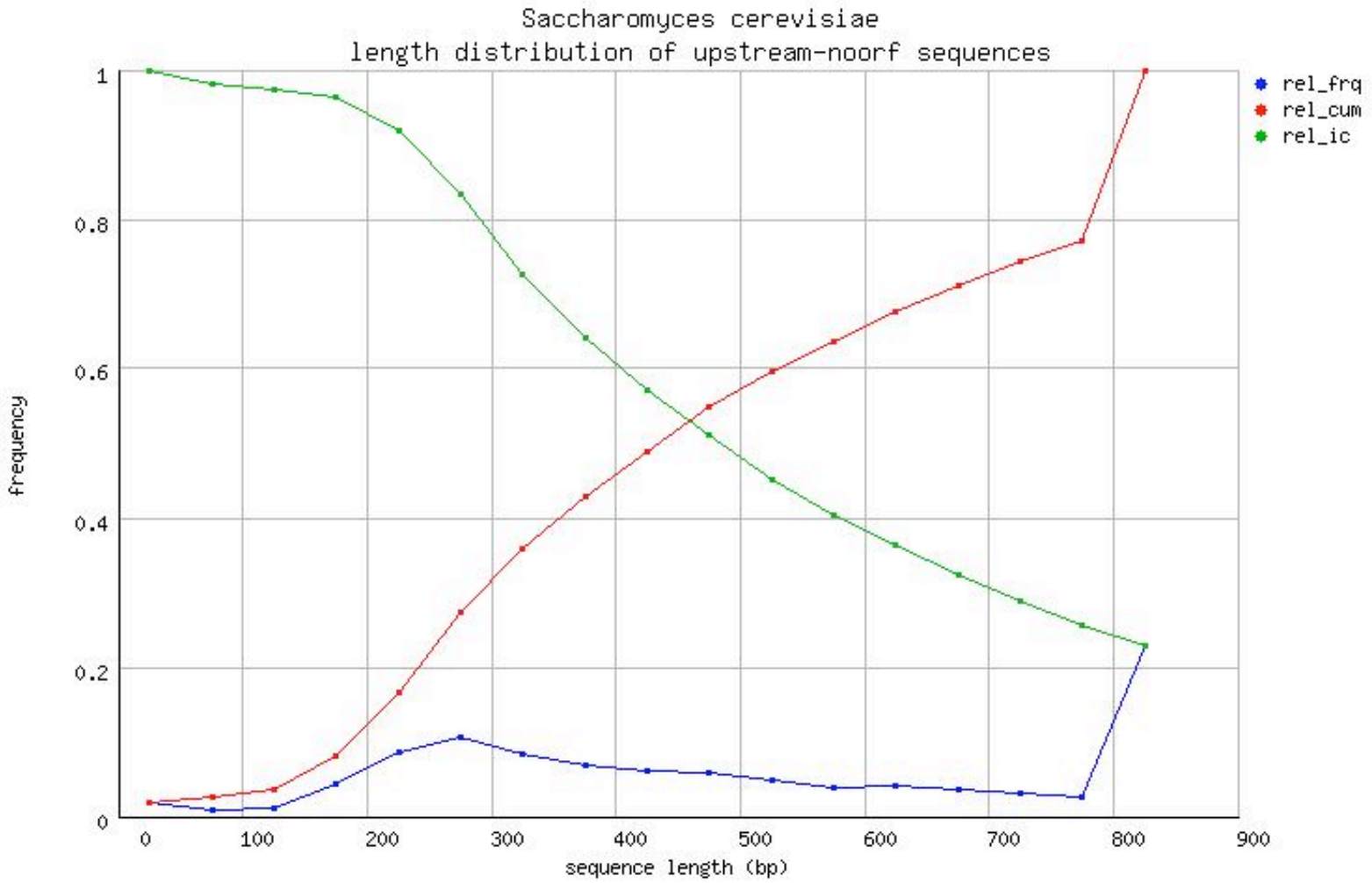


Regulatory Sequence Analysis

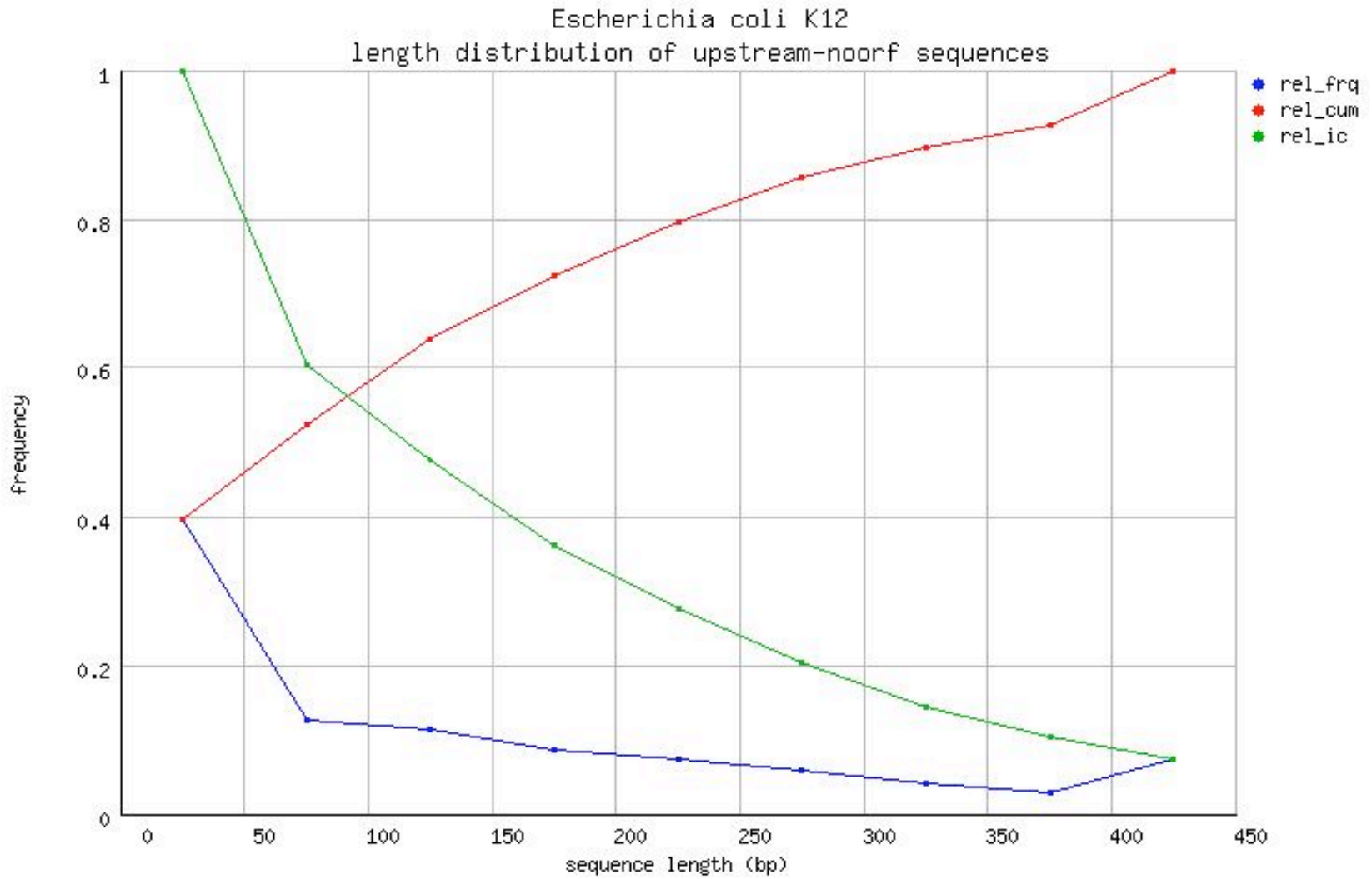
Supplementary material

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

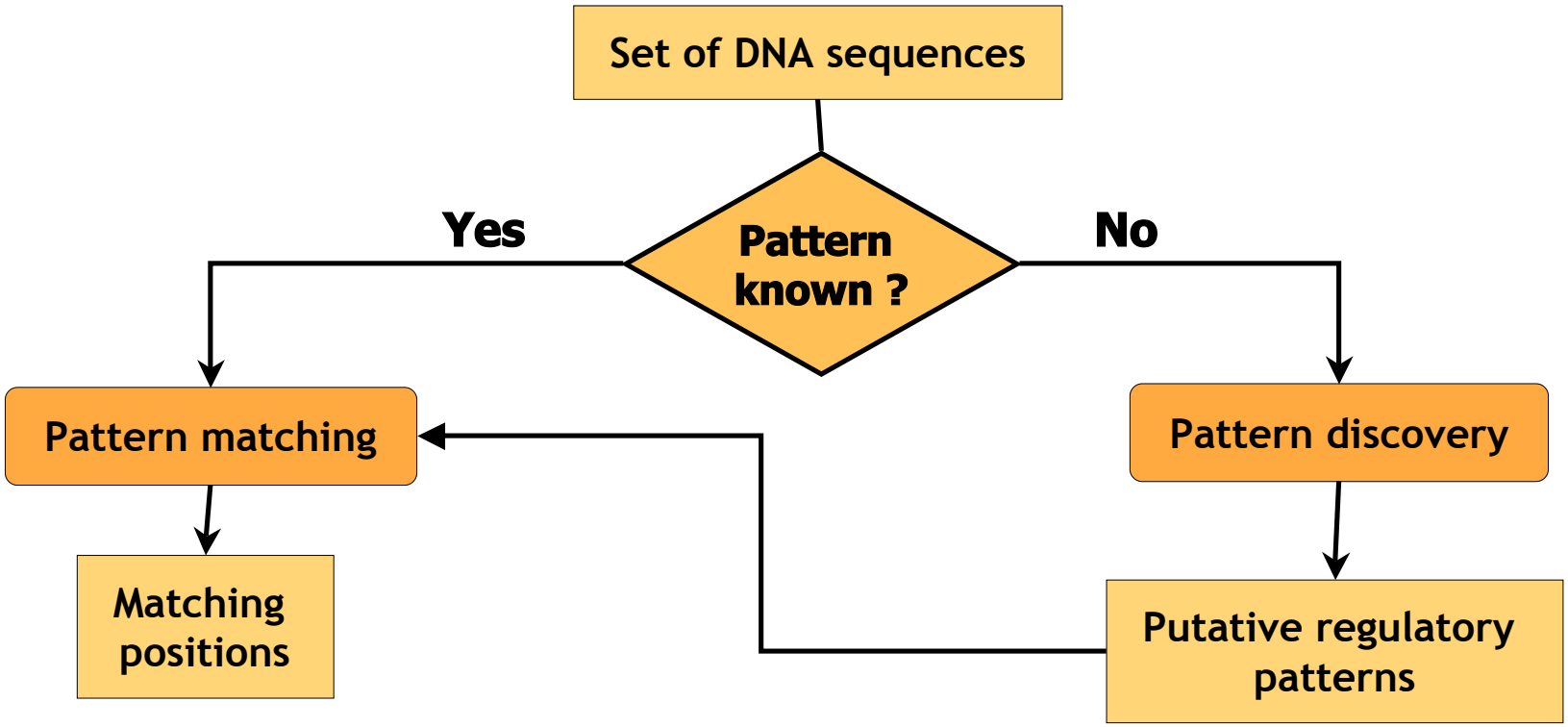
Distribution of upstream sequence lengths
Saccharomyces cerevisiae



Distribution of upstream sequence lengths
Escherichia coli K12



Pattern matching vs pattern discovery



Other examples of sequence logos

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	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

Regulatory sites : matrix description

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

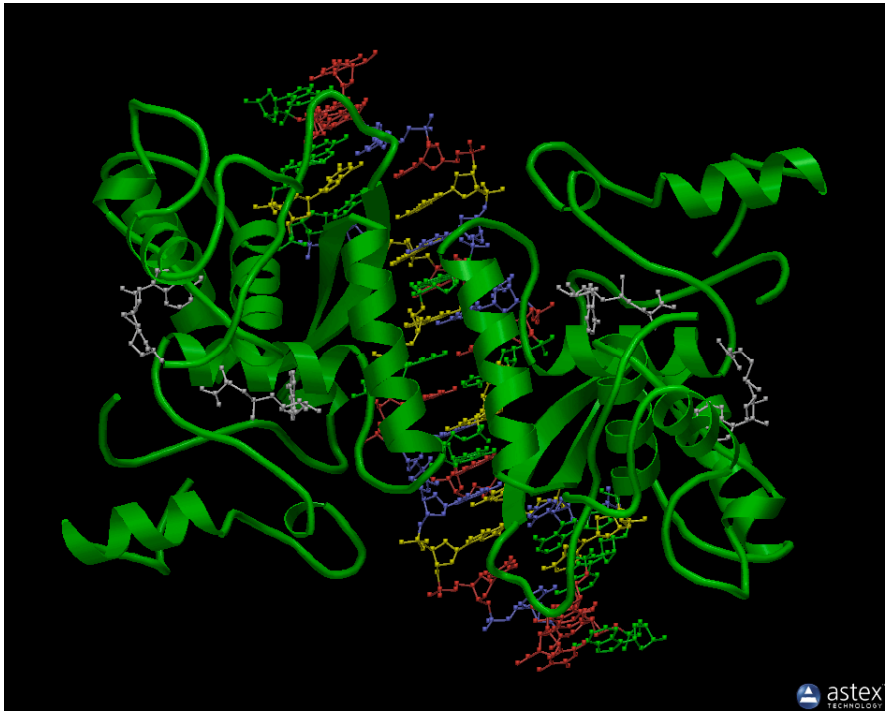
Binding motif for the yeast Pho4p transcription factor

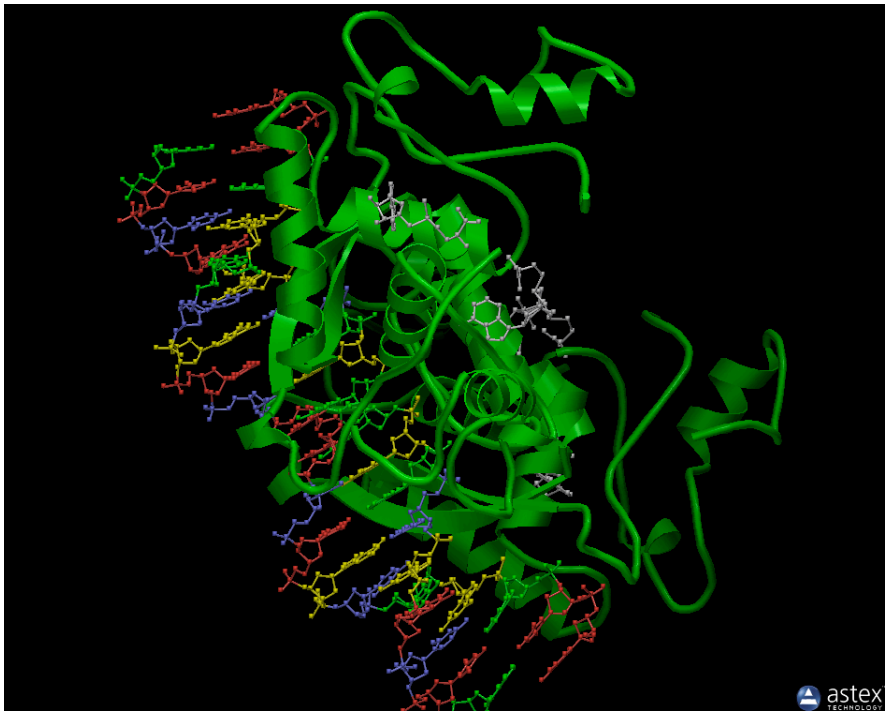
Source : SCPD

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

Methionine repressor

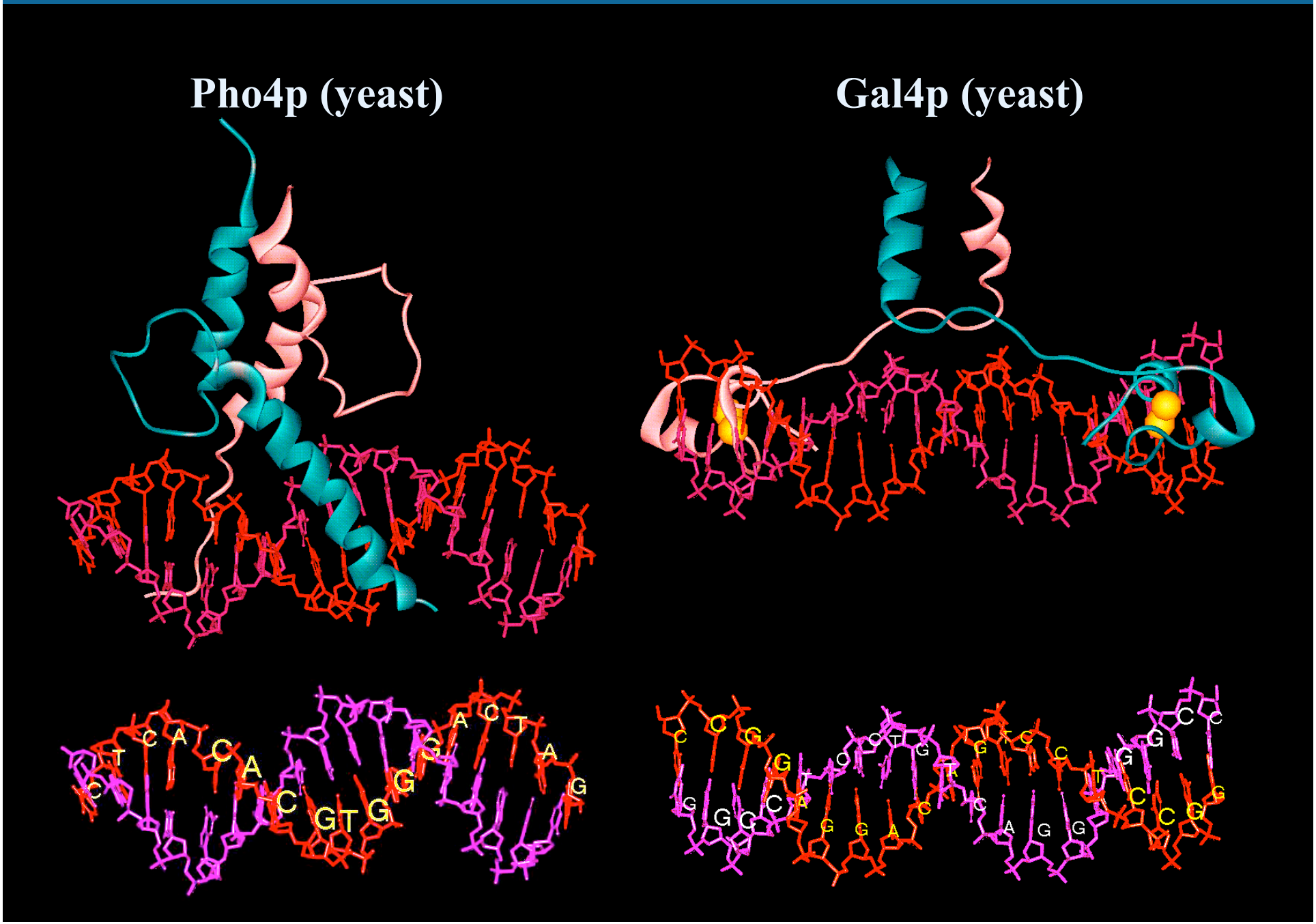
- Crystal structure of the methionine repressor from Escherichia coli.
- In green: the MetJ protein forms a homodimer which is able to bind DNA.
- Nucleotide structure is coloured by type of nucleotide (A,C,G,T).
- In grey: the repressor is activated by binding of methionine molecules





http://www.ebi.ac.uk/msd-srv/msdlite/apps/query?id_type=swissProtId&id_value=P0A8U6

Transcription factor-DNA interfaces



The genome challenge

RNA polymerase