The analysis of regulatory sequences

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Team composition

SCMBB - Bruxelles - Belgium

Unit "Analysis of Biochemical, Regulation and Interaction Networks"

Olivier Sand	Postdoc	evaluation of pattern discovery approaches
Rekin's Janky	DEA 2002 + PhD	evolution of bacterial metabolism and regulation
Sylvain Brohée	PhD	networks of interactions between membrane proteins
Jean-Valéry Turatsinze	DEA 2005 + PhD	evaluation of pattern matching approaches in human ; regulatory elements involved in development
Dimitrios Paraskevas	DEA	phylogenetic footprints
Carlos Moreira Conde da Silva	master	unsupervised classification of regulatory sequences
Kevin Wielemans	master	evolution of nitrogen regulation in fungi

Previous researchers

Shoshana Wodak	lab director	transcriptional regulation of protein complexes
Nicolas Simonis	PhD 2005	transcriptional regulation of protein complexes
Didier Croes	PhD 2005	Path finding in metabolic networks
Joseph Tran	DEA 2002 + PhD	Inference of metbaolic pathways from microarrays
Didier Gonze	DEA 2001 + Postdoc 2003	Supervised classification of regulatory sequences

Previous students

Ahmed Essaghir	DEA 2005	analysis of microarrays and promoters in Plasmodium
Raymond Kalimunda	DEA 2005	development of a pattern assembly algorithm
Mehdi Jbel	DEA 2005	evolution of nitrogen regulation in fungi
Pierre Jonniaux	DEA 2004	Test of phylogenetic footprint approaches
Fabian Couche	graduate 2003	algorithms to find the k shortest paths in weighted graphs
Patrice Chagnaud	DEA 2001	Prediction of Nitrogen-regulated genes in yeast
Hassan Ghazal	DEA 2001	Analysis of plant promoters
Magali Lescot	DEA 2001	Pattern discovery in plant promoters

The analysis of regulatory sequences

Regulation of biological processes

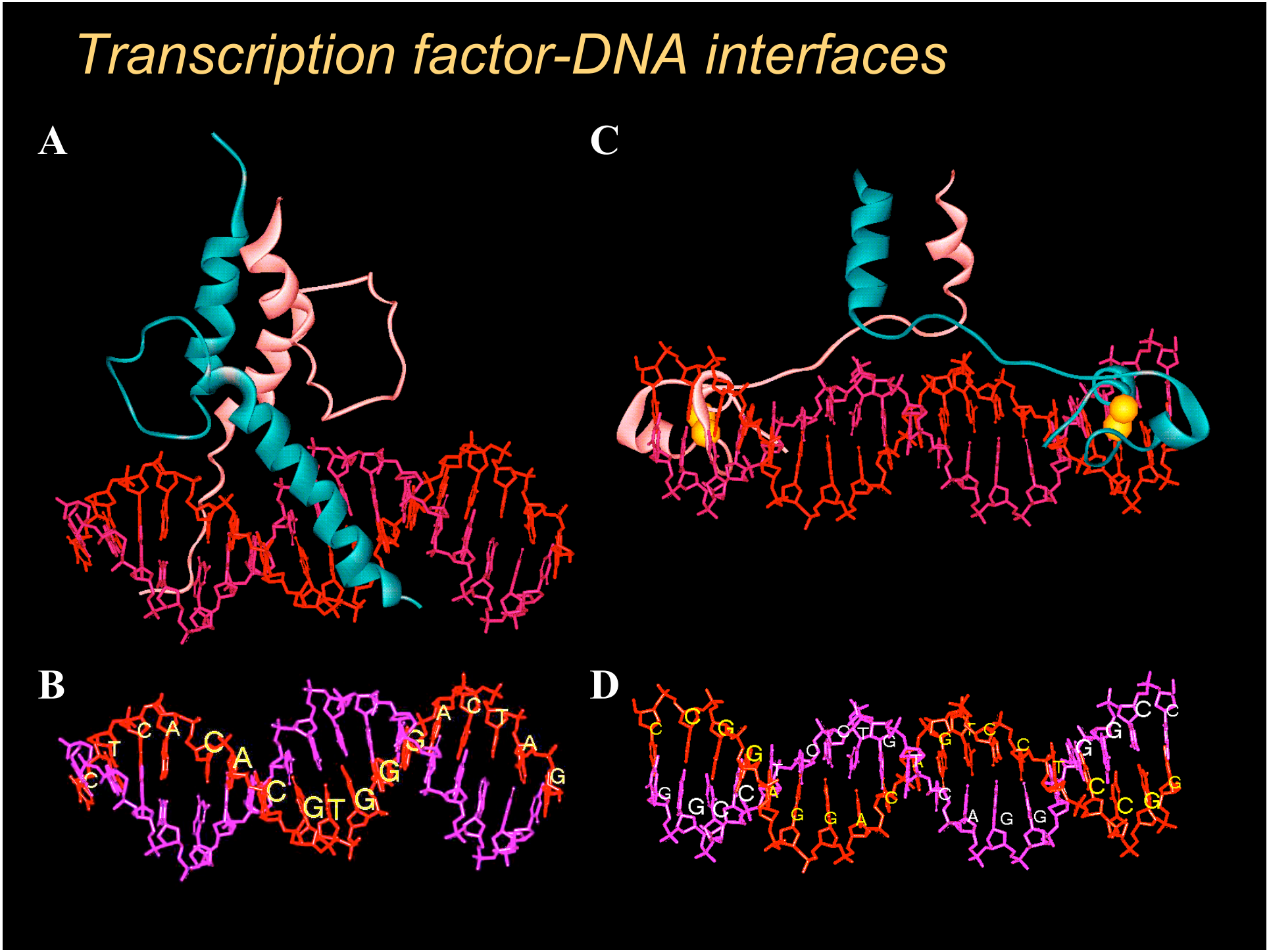
Jacques van Helden
Jacques.van.Helden@ulb.ac.be

The non-coding genome

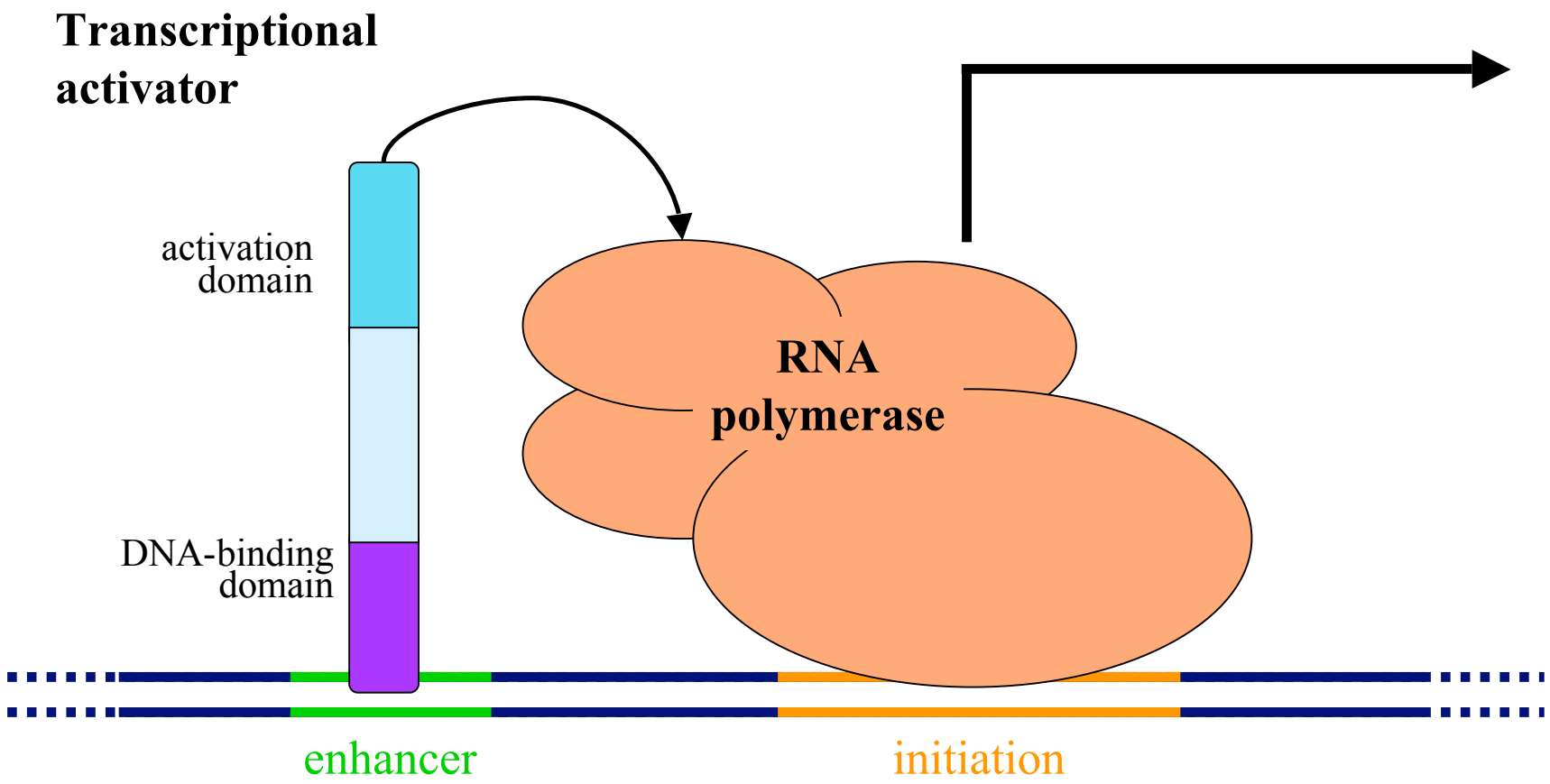
Organism	Year	Size Mb	Genes	genes/Mb	coding %	non-coding %	repetitive %	Transcribed %
Mycoplasma genitalium	1995	0.6	481	802	90	10		
Haemophilus influenzae	1995	1.8	1 717	954	86	14		
Escherichia coli	1997	4.6	4 289	932	87	13		
Saccharomyces cerevisiae	1996	12	6 286	524	72	28		
Arabidiopsis thaliana	2001	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

The genome challenge

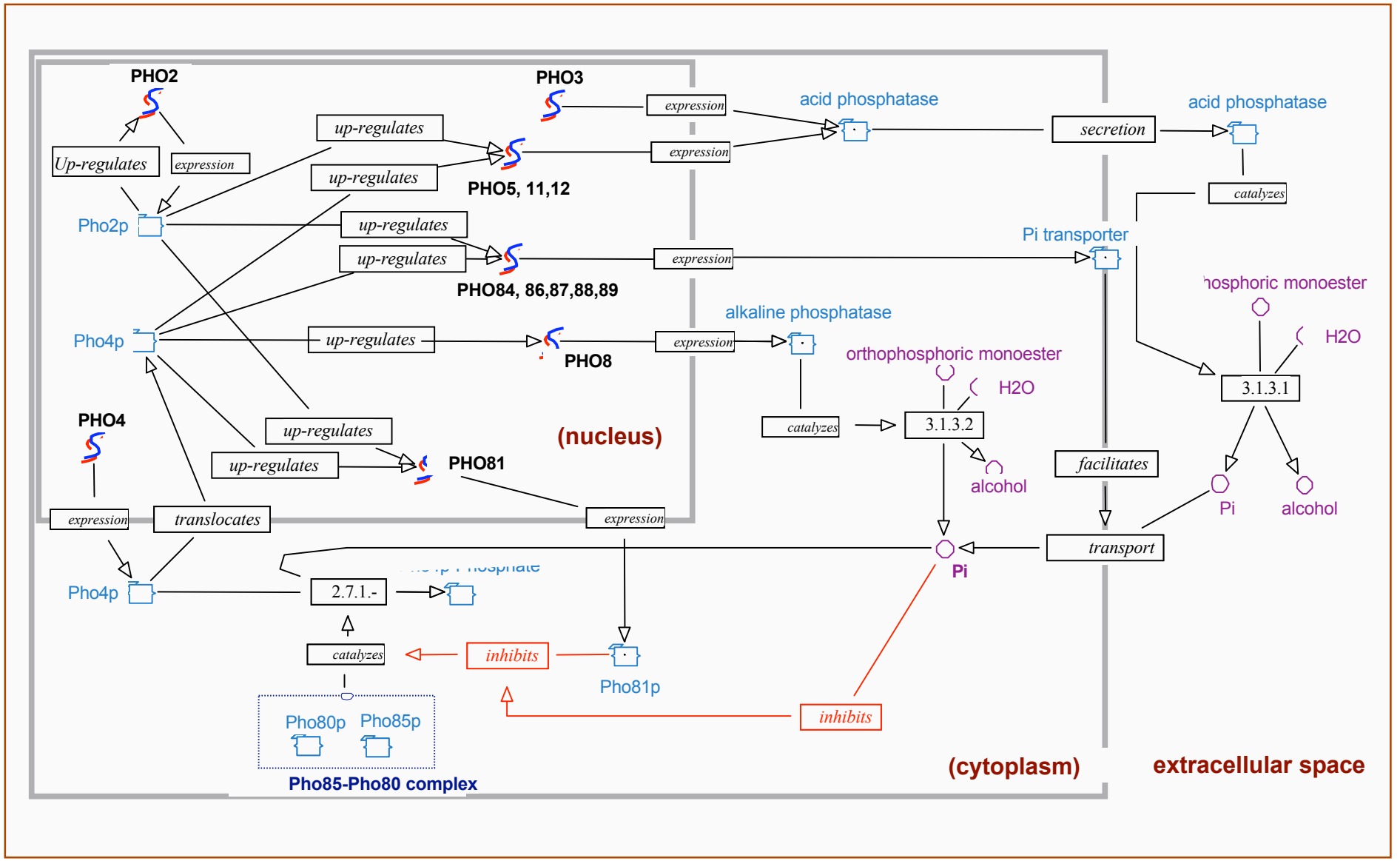
Transcription factor-DNA interfaces



Transcriptional activation



Phosphate utilisation in yeast



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	---aCtCaCACACGTGGGACTAGC-	high
PHO84	Site D	---TTTCCAGCACGTGGGGCGGA--	high
PHO81	UAS	----TTATGGCACGTGCGAATAA--	high
PHO8	Proximal	GTGATCGCTGCACGTGGCCCGA---	high
PHO5	UASp3	--TAATTTGGCATGTGCGATCTC--	low
PHO84	Site C	-----ACGTCCACGTGGAACTAT--	low
PHO84	Site A	-----TTTATCACGTGACACTTTTT	low
group 1	consensus	---------gCACGTGggac-----	high-low
PHO5	UASp1	--TAAATTAGCACGTTTTCGC----	medium
PHO84	Site E	----AATACGCACGTTTTTAATCTA	medium
PHO84	Site B	-----TTACGCACGTTGGTGCTG--	low
PHO8	Distal	---TTACCCGCACGCTTAATAT---	low
group 2	consensus	--------cgCACGTTt--------	med-low
Degenerate consensus		---------GCACGTKKk-------	

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

Regulatory sites : matrix description

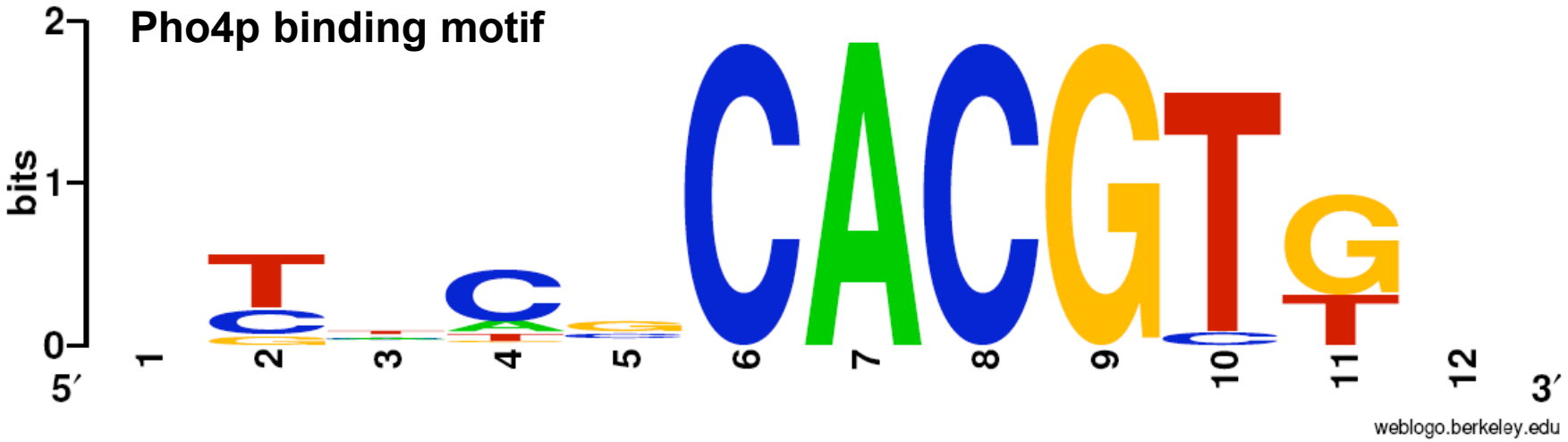
Position-specific scoring matrix (PSSM)

Pos Base	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
			V	C	A	C	G	T	K	B		

Binding motif for the yeast Pho4p transcription factor

(Source : Transfac matrix F$PHO4_01)

Sequence logo



This logo drawing was realized with the program Weblogo (http://weblogo.berkeley.edu/)

Questions and approaches

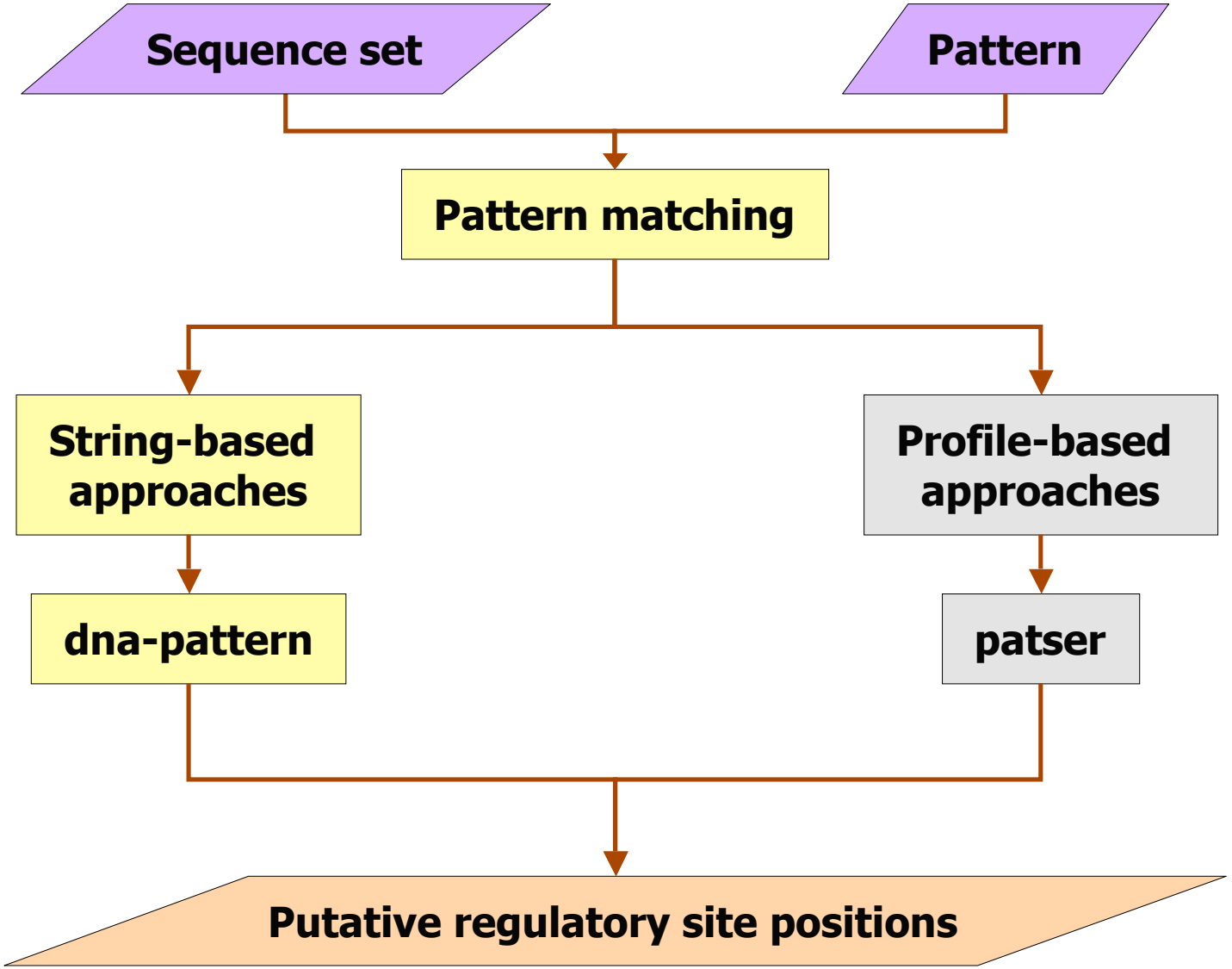
1. If we know the consensus for a given transcription factor, can we predict its binding sites in a DNA sequence ?
   - **Pattern matching** against a sequence
2. Can we scan a sequence for matches with the consensus of all he currently known transcription factor ?
   - **Matching a library** of patterns against a sequence
3. Starting from a set of co-regulated genes, can we predict cis-acting elements involved in their transcriptional regulation ?
   - **Pattern discovery** within a sequence set
4. Can we detect regulatory signals by searching conserved elements in non-coding sequences ?
   - **Phylogenetic footprinting**
5. Can we classify genes on the basis of the presence of regulatory motifs in their regulatory regions ?
   - **Gene classification** on the basis of pattern scores

The analysis of regulatory sequences

Pattern matching

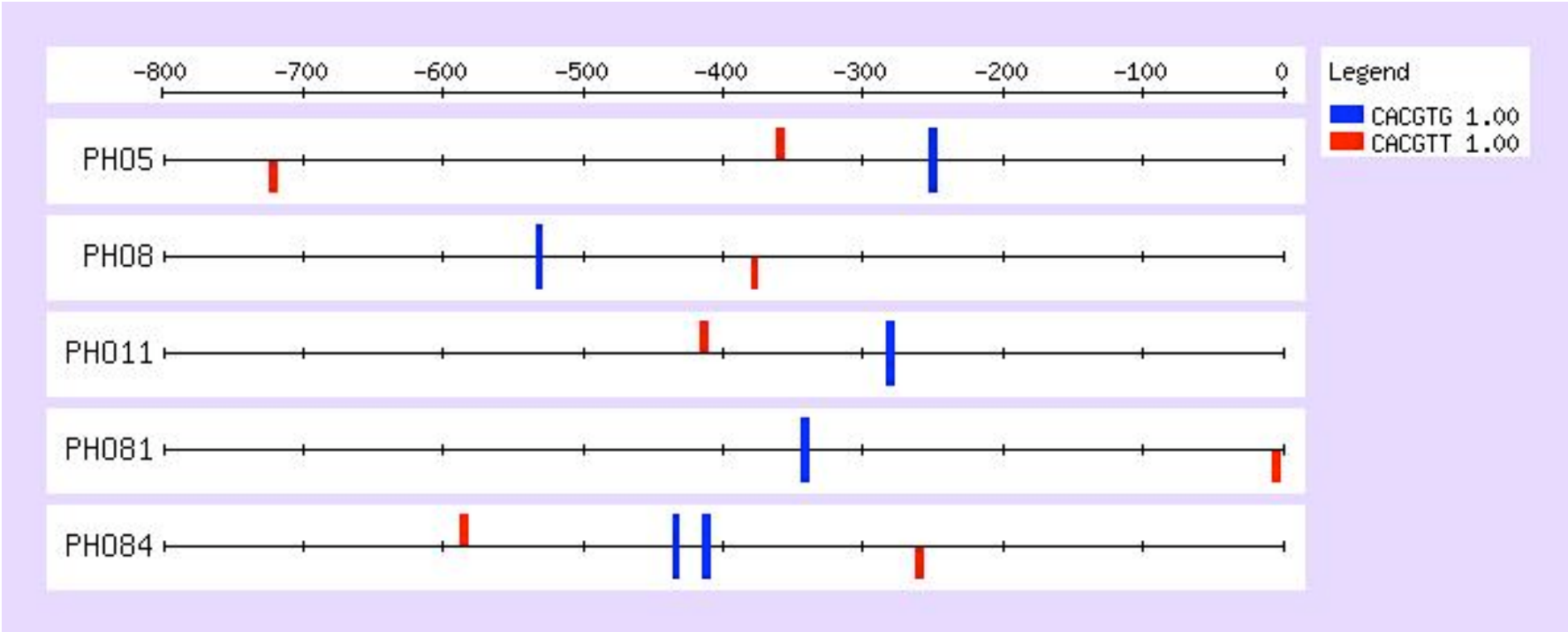
Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Pattern matching in regulatory sequences



Matching simple patterns

- A simple string-based pattern matching is usually poorly informative.
  - spurious matches are expected to be found anywhere
  - the presence of the consensus does not necessarily mean that the factor binds
  - some patterns have a higher significance than other ones (e.g. the core of the consensus).
  - more interesting results can be obtained by assigning a specific score to each pattern, as will be shown a bit later



Regulatory sites : matrix description

Alignment matrix

Pos Base	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
			V	C	A	C	G	T	K	B		

Binding site for the yeast Pho4p transcription factor

(Source : Transfac matrix F$PHO4_01)

Position-weight matrix

Prior	Pos	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00
0.33	A	-0.79	0.13	-0.23	-2.20	**1.05**	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.18	C	0.32	0.32	**0.70**	**1.65**	-2.20	**1.65**	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.18	G	-0.29	0.32	**0.70**	-2.20	-2.20	-2.20	**1.65**	-2.20	**1.19**	**0.97**	**1.19**	0.32
0.33	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	**1.05**	0.13	-0.23	-0.23	-0.23
1	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.19

$$W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}$$

A *alphabet size (=4)*

p_i *prior residue probability for residue i*

$f_{i,j}$ *relative frequency of residue i at position j*

k *pseudo weight (arbitrary, 1 in this case)*

$f'_{i,j}$ *corrected frequency of residue i at position j*

Weight of a sequence segment

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	**-0.79**	0.13	-0.23	-2.20	1.05	-2.20	**-2.20**	**-2.20**	**-2.20**	-2.20	-0.79	-0.23
C	0.32	0.32	0.70	**1.65**	-2.20	1.65	-2.20	-2.20	-2.20	0.32	**-2.20**	0.32
G	-0.29	0.32	**0.70**	-2.20	**-2.20**	-2.20	1.65	-2.20	1.19	**0.97**	1.19	0.32
T	0.39	**-0.79**	-2.20	-2.20	-2.20	**-2.20**	-2.20	1.05	0.13	-0.23	-0.23	**-0.23**
residue r	A	T	G	C	G	T	A	A	A	G	C	T
W(r)	-0.79	-0.79	0.70	1.65	-2.20	-2.20	-2.20	-2.20	-2.20	0.97	-2.20	-0.23
Weight	**-11.67**		**=SUM[W(r)]**									

$$W_S = \ln\left(\frac{P(S \mid M)}{P(S \mid B)}\right) = \ln\left(\frac{\prod_{j=1}^{w} f'_{r_j j}}{\prod_{j=1}^{w} p_{r_j}}\right) = \ln\left(\prod_{j=1}^{w} \frac{f'_{r_j j}}{p_{r_j}}\right) = \sum_{j=1}^{w} \ln\left(\frac{f'_{r_j j}}{p_{r_j}}\right) = \sum_{j=1}^{w} W_{r_j j}$$

Ws	weight of sequence segment S
$P(S\|M)$	probability of the sequence segment, given the matrix
$P(S\|B)$	probability of the sequence segment, given the background
j	position within the segment and within the matrix
r_j	residue at position j of the sequence segment
p_{rj}	prior probability of residue r_j
f'_{rjj}	probability of residue r_j at position j of the matrix

- The ***weight*** of a sequence segment is defined as the log-ratio of
  - $P(S|M)$, the sequence probability under the model described by the PSSM, and
  - $P(S|B)$, the sequence probability under the background model.
- The weight represents the likelihood that this segment is an occurrence of the motif rather than being issued from the background model.
- The weight matrix W_{ij} allows to easily calculate segment weights.

Scanning a sequence with a profile matrix

Ex: sequence . . . . . **GCTGCACGTGGCCC** . .

Position-Specific Scoring Matrix

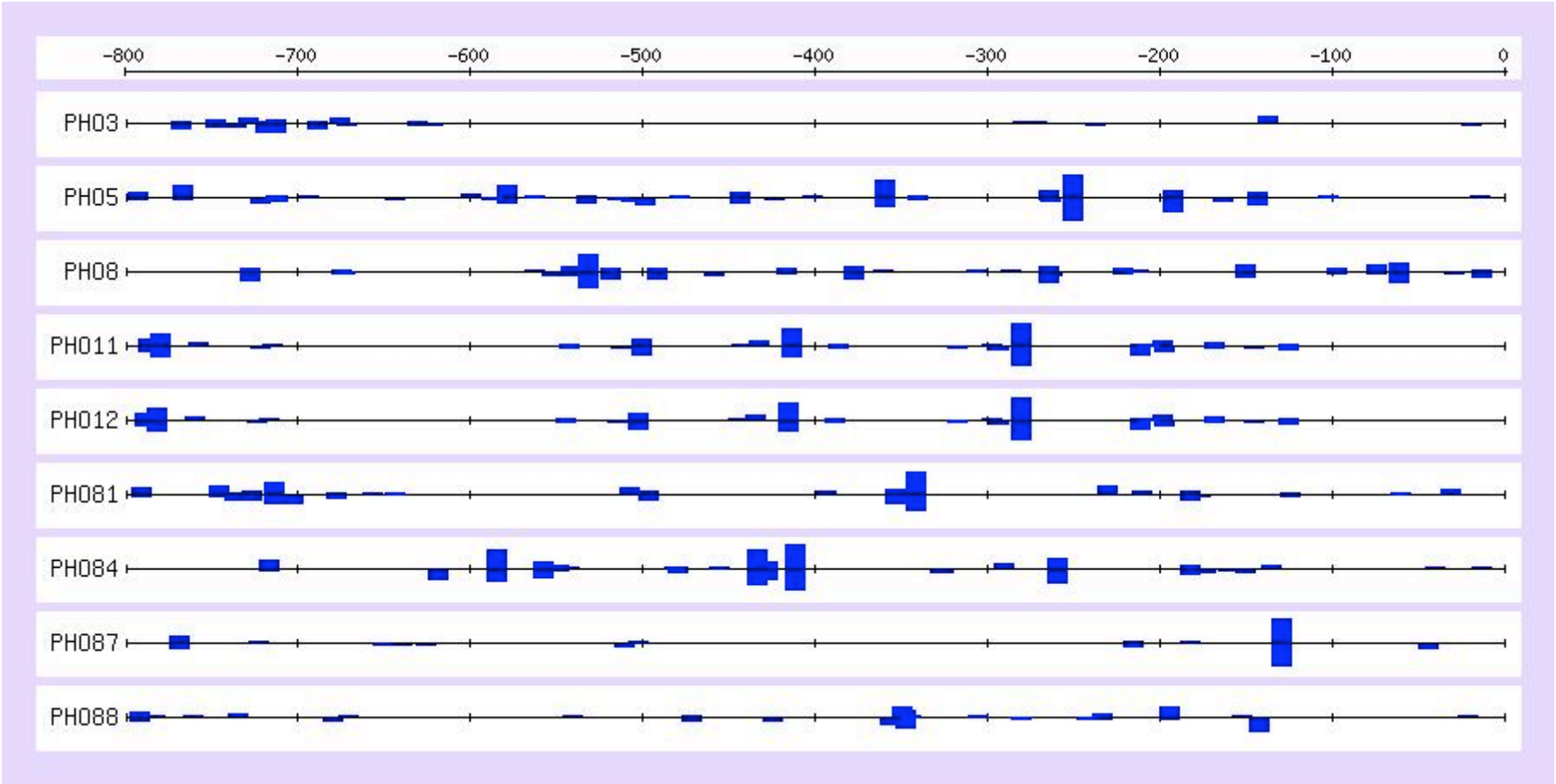
	1	2	3	4	5	6	7	8	9	10	11	12
A	-0,8	0,1	-0,2	-2,2	1,0	-2,2	-2,2	-2,2	-2,2	-2,2	-0,8	-0,2
C	0,3	0,3	0,7	1,6	-2,2	1,6	-2,2	-2,2	-2,2	0,3	-2,2	0,3
G	-0,3	0,3	0,7	-2,2	-2,2	-2,2	1,6	-2,2	1,2	1,0	1,2	0,3
T	0,4	-0,8	-2,2	-2,2	-2,2	-2,2	-2,2	1,0	0,1	-0,2	-0,2	-0,2

Scanning

1	**SUM**	**G**	**C**	**T**	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**C**	**C**	**C**
	-10,54	**-0,3**	**0,3**	**-2,2**	**-2,2**	**-2,2**	**-2,2**	**-2,2**	**-2,2**	**0,1**	**1,0**	**1,2**	**0,3**		
2		**C**	**T**	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**C**	**C**	**C**	
	7,55	**0,3**	**-0,8**	**0,7**	**1,6**	**1,0**	**1,6**	**1,6**	**1,0**	**1,2**	**1,0**	**-2,2**	**0,3**		
3		**T**	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**G**	**C**	**C**	**C**		
	-9,93	**0,4**	**0,3**	**0,7**	**-2,2**	**-2,2**	**-2,2**	**-2,2**	**-2,2**	**1,2**	**0,3**	**-2,2**	**0,3**		

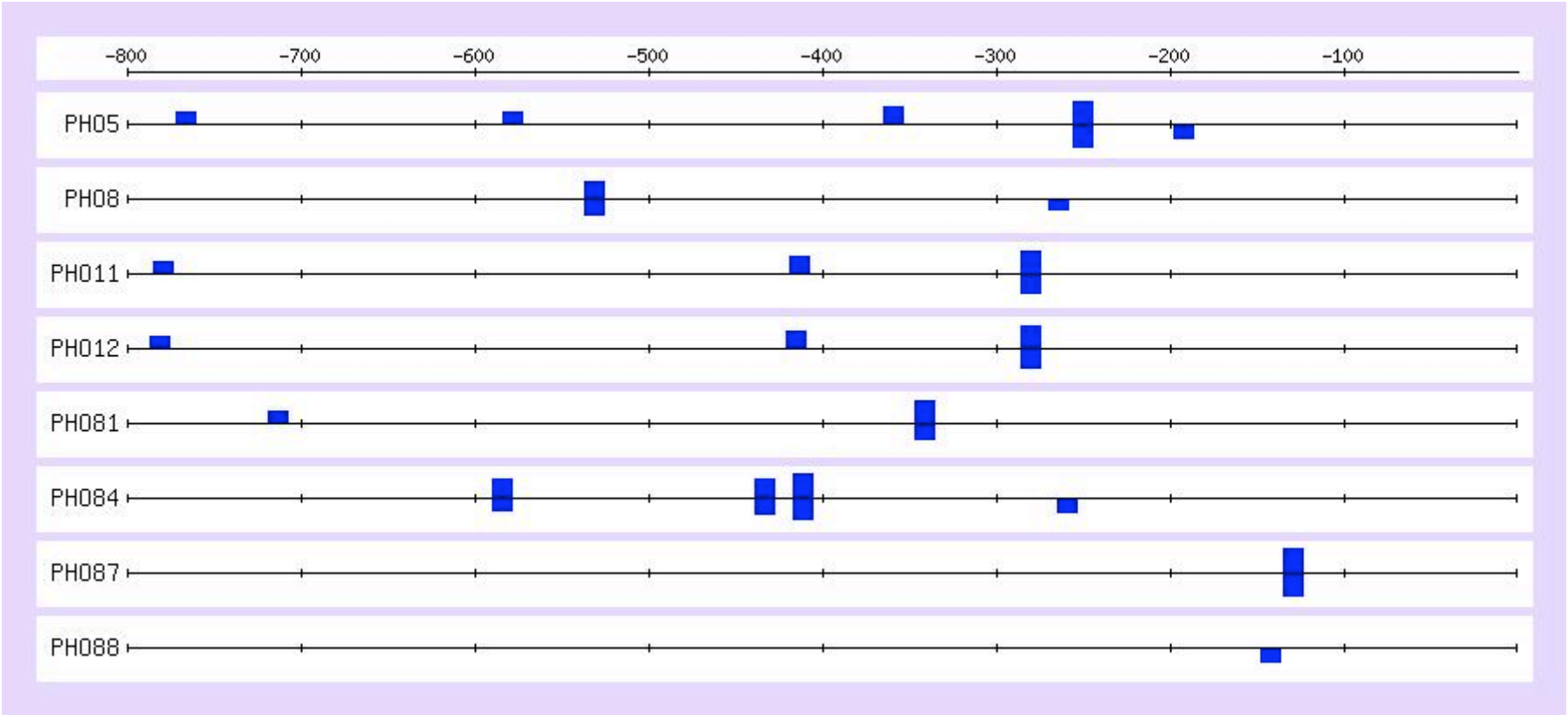
Matrix search : matching positions

- Matrix-based pattern matching is more sensitive than string-based pattern matching.
- How to choose the threshold ?

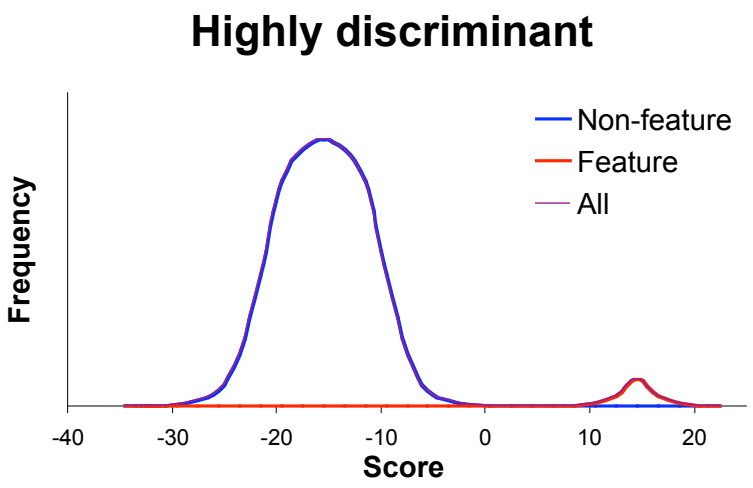


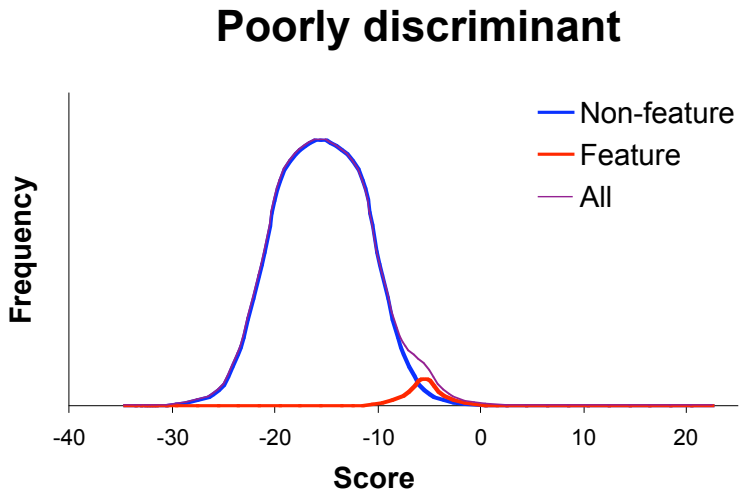
Matrix search : threshold selection

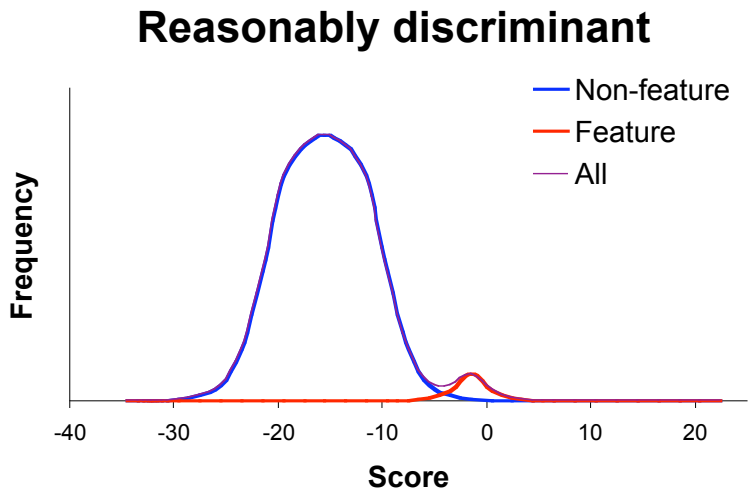
- Patser includes an option to automatically select a threshold on the basis of
  - the information content of the matrix
  - the length of the sequence to be scanned
- Note : the gene PHO3 is not displayed because there was not a single match. This gene is indeed not regulated by phosphate.
- Another approach would be to select the threshold on the basis of scores returned when the matrix is used to scan known binding sites for the factor.



Discrimination power of a matrix



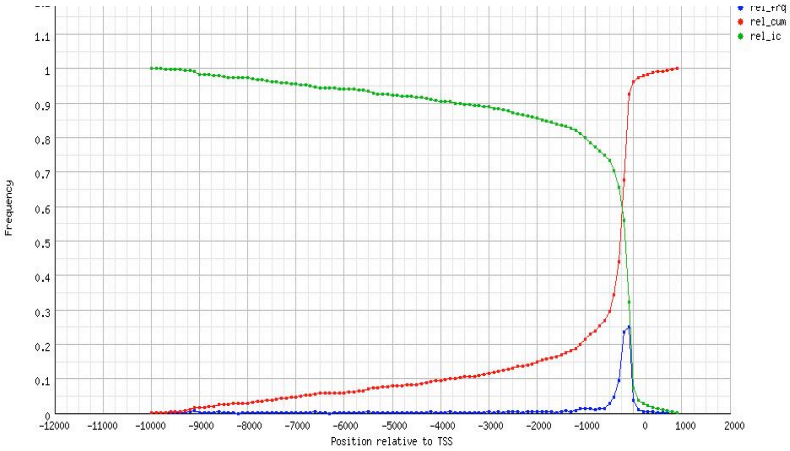




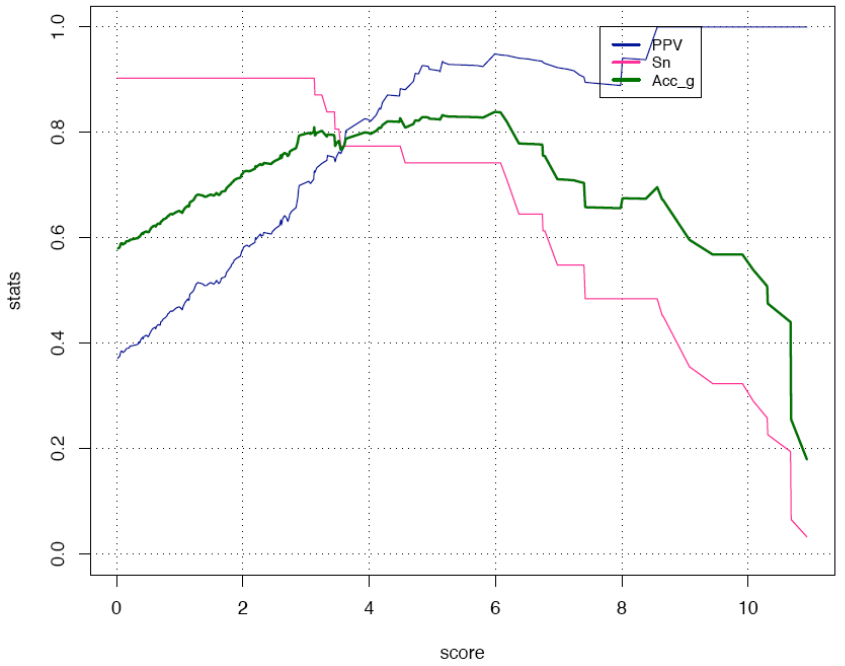
Evaluation of pattern matching results

- Evaluation, at the site level, for pattern matching results in human promoters with an NFkB matrix.
- Statistics
  - Sensitivity
    - *Sn=TP/(TP+FN)=(true predictions)/(annotated sites)*
  - Positive Predictive Value
    - *PPV=TP/(TP+FP)=(true predictions)/(total predictions)*
  - Accuracy
    - *Acc.a = (Sn + PPV)/2*
    - *Acc.g = sqrt(Sn * PPV)*
- Notes
  - The predictions were restricted to 500bp, because this is the best annotated interval in the reference database (TRANSPRO).
  - This is an illustration only, for one of the best examples.
  - NFkB is one of the best annotated factors in TRANSPRO.
  - It is not representative of overall performances.
  - Predictions give better results for NFkB than for other factors (*in preparation*).

Positions of human sites annotated in TRANSPRO (relative to TSS).

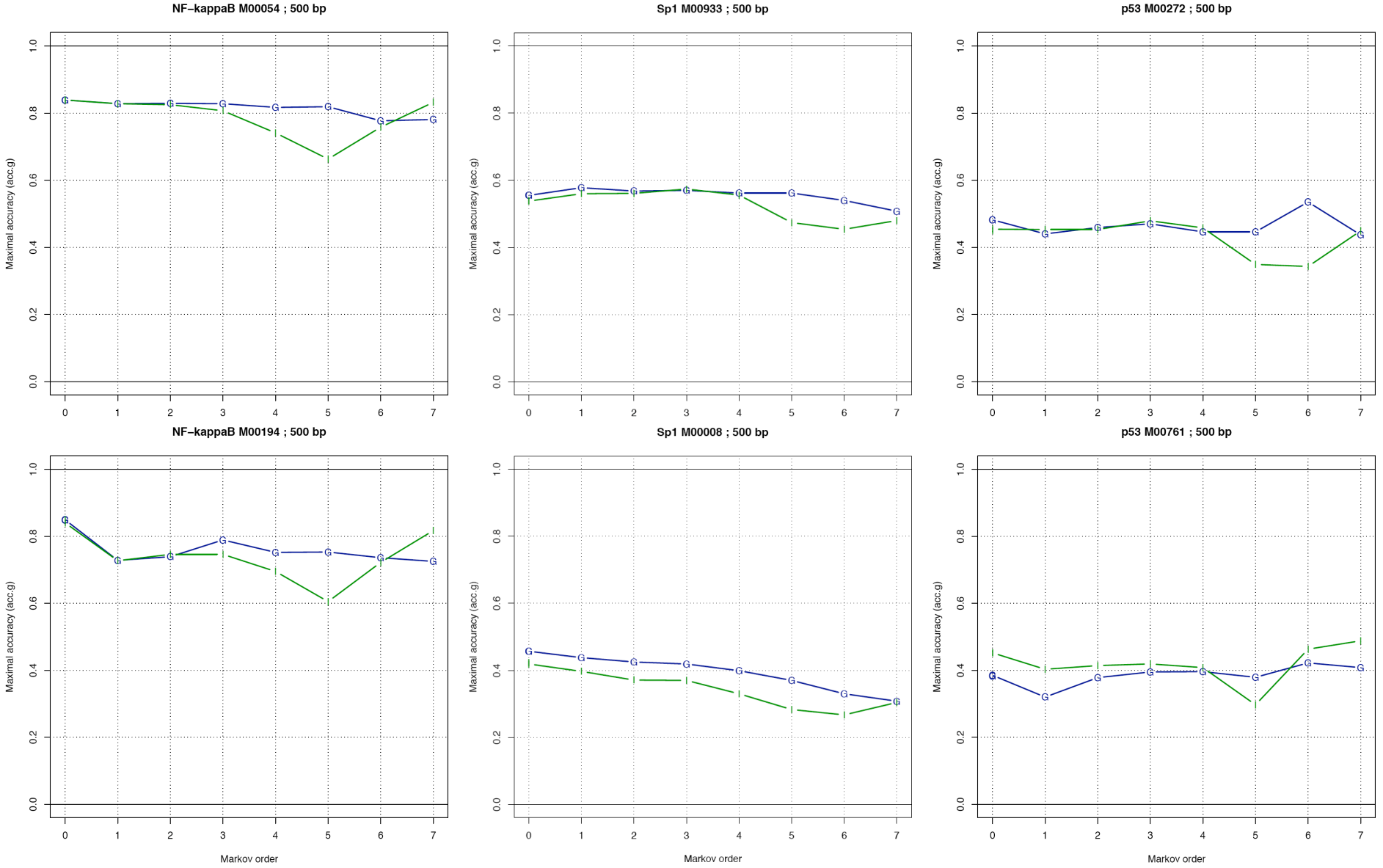


Patser predictions: NKkB, 500bp upstream of TSS



Jean Valéry Turatsinze (2005). Graduate thesis.

Effect of the matrix and of the background model



Jean Valéry Turatsinze (2005). Graduate thesis.

The analysis of regulatory sequences

Matching a sequence with a library of patterns

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Match a sequence with a library of patterns

- Goal : given a sequence, find matches for any known regulatory site
  - → identify transcription factors that could regulate the gene
- Strategy: apply systematically pattern search with all patterns stored in the library
- Problem: how to set the threshold for the different patterns ?
- Warning : generates many false positive

Transfac Matsearch result - PHO5 upstream region

```
Inspecting sequence PHO5_4 [?] (1 - 816):
  F$NIT2_01            |      141 (+) |  1.000 |  0.995 | TATCtc
  F$PHO4_01            |      561 (+) |  1.000 |  0.990 | tcaCACGtggga
  F$PHO4_01            |      561 (-) |  1.000 |  0.982 | tccCACGtgtga
  F$NIT2_01            |      634 (+) |  1.000 |  0.972 | TATCaa
  F$NIT2_01            |      543 (-) |  1.000 |  0.967 | TATCga
  F$NIT2_01            |      676 (-) |  1.000 |  0.945 | TATCcc
  F$NIT2_01            |       31 (-) |  1.000 |  0.937 | TATCag
  F$PHO4_01            |      452 (+) |  1.000 |  0.935 | tagCACGttttc
  F$MCM1_01            |      666 (-) |  0.961 |  0.929 | tatCCCAaatgggtat
  F$MATA1_01           |      202 (+) |  1.000 |  0.926 | tGATGtcagt
  F$GCR1_01            |      323 (-) |  1.000 |  0.922 | gaCTTCcaa
  F$GCN4_C             |      536 (+) |  0.837 |  0.902 | aaaTGAATcg
  F$ABAA_01            |      292 (-) |  1.000 |  0.889 | atttgcgCATTcttgttga
  F$ABF_C              |      205 (+) |  0.887 |  0.885 | tgtcagtccccACGC
  F$MATA1_01           |      727 (+) |  1.000 |  0.882 | tGATGttttg
  F$MIG1_01            |      210 (-) |  1.000 |  0.881 | gctattagcgtGGGGac
  F$GCR1_01            |       69 (+) |  0.826 |  0.880 | ggCATCcaa
  F$PHO4_01            |       90 (-) |  1.000 |  0.879 | ggtCACGtttct
  F$MAT1MC_02          |      696 (+) |  1.000 |  0.875 | tgaaTTGTcg
  F$GCN4_C             |      589 (+) |  0.882 |  0.862 | ttaTGATTct
  F$STE11_01           |      415 (+) |  1.000 |  0.860 | ctttttCTTTgtctgcac
  F$GCR1_01            |      249 (-) |  0.783 |  0.859 | ggCGTCctg
  F$STE11_01           |      425 (-) |  1.000 |  0.859 | atatttCTTTgtgcagac
  F$MCM1_01            |      484 (+) |  0.831 |  0.855 | atgCCAAaaaaagtaa
```

Transfac Matsearch result - random sequence (mkv 5)

```
Inspecting sequence random mkv5 [?] (1 - 817):
  F$NIT2_01            |      176 (+) |  1.000 |  1.000 | TATCta
  F$NIT2_01            |      656 (+) |  1.000 |  1.000 | TATCta
  F$NIT2_01            |      275 (+) |  1.000 |  0.995 | TATCtc
  F$NIT2_01            |      455 (+) |  1.000 |  0.995 | TATCtc
  F$NIT2_01            |      298 (-) |  1.000 |  0.980 | TATCtt
  F$MATA1_01           |      506 (-) |  1.000 |  0.980 | tGATGtatgt
  F$ABF_C              |       84 (+) |  0.991 |  0.973 | aatcattcttgACGT
  F$MIG1_01            |      264 (-) |  1.000 |  0.958 | gagataaaactGGGGtt
  F$NIT2_01            |      701 (+) |  1.000 |  0.947 | TATCgt
  F$NIT2_01            |      802 (-) |  1.000 |  0.947 | TATCgt
  F$ABF1_01            |       81 (+) |  0.976 |  0.944 | gtaaatcattcttgACGTtttt
  F$MAT1MC_02          |      665 (-) |  1.000 |  0.918 | cctaTTGTga
  F$NIT2_01            |      280 (-) |  1.000 |  0.915 | TATCcg
  F$ABAA_01            |       42 (+) |  1.000 |  0.902 | tccccatCATTctaacagt
  F$PACC_01            |      331 (-) |  1.000 |  0.897 | acgaGCCAagaaaagtt
  F$ABAA_01            |      201 (+) |  1.000 |  0.883 | accatagCATTctggatct
  F$MAT1MC_02          |      442 (-) |  1.000 |  0.882 | tataTTGTat
  F$ABF_C              |      638 (-) |  0.991 |  0.882 | agtcaaatgaaACGT
  F$ABF_C              |      609 (-) |  0.949 |  0.874 | tttctttaaacACGG
  F$MATA1_01           |      558 (-) |  1.000 |  0.868 | tGATGgaaga
  F$HSF_03             |      713 (-) |  1.000 |  0.859 | AGAAttgaaattttt
  F$MAT1MC_02          |      134 (-) |  1.000 |  0.858 | cacaTTGTgt
  F$ABAA_01            |       80 (+) |  1.000 |  0.856 | agtaaatCATTcttgacgt
  F$HAP234_01          |      332 (-) |  1.000 |  0.851 | acgagCCAAgaaaagt
```

Transfac Matsearch result - random sequence (iid)

```
Inspecting sequence random iid [?] (1 - 817):
  F$NIT2_01            |      534 (-) |  1.000 |  1.000 | TATCta
  F$NIT2_01            |      294 (+) |  1.000 |  0.995 | TATCtc
  F$NIT2_01            |      634 (-) |  1.000 |  0.972 | TATCaa
  F$NIT2_01            |      216 (-) |  1.000 |  0.965 | TATCtg
  F$STUAP_01           |      808 (-) |  1.000 |  0.959 | attCGCGtct
  F$NIT2_01            |       24 (+) |  1.000 |  0.952 | TATCat
  F$NIT2_01            |      343 (+) |  1.000 |  0.952 | TATCat
  F$NIT2_01            |      413 (-) |  1.000 |  0.952 | TATCat
  F$STUAP_01           |      441 (+) |  1.000 |  0.930 | aagCGCGcct
  F$NIT2_01            |      244 (-) |  1.000 |  0.930 | TATCct
  F$STUAP_01           |      808 (+) |  1.000 |  0.926 | agaCGCGaat
  F$GCR1_01            |      499 (+) |  1.000 |  0.922 | gaCTTCcta
  F$PACC_01            |      647 (-) |  1.000 |  0.920 | ctccGCCAggcactgaa
  F$NIT2_01            |      475 (+) |  1.000 |  0.915 | TATCcg
  F$ABF_C              |      235 (-) |  0.949 |  0.904 | tatcctgcaacACGG
  F$PHO4_01            |      246 (-) |  1.000 |  0.882 | gctCACGttatc
  F$GCR1_01            |      763 (-) |  1.000 |  0.866 | acCTTCcgc
  F$STUAP_01           |      441 (-) |  1.000 |  0.859 | aggCGCGctt
  F$MIG1_01            |      371 (+) |  1.000 |  0.857 | accgaaacagtGGGGtt
  F$MAT1MC_02          |      375 (-) |  0.769 |  0.855 | cccaCTGTtt
```

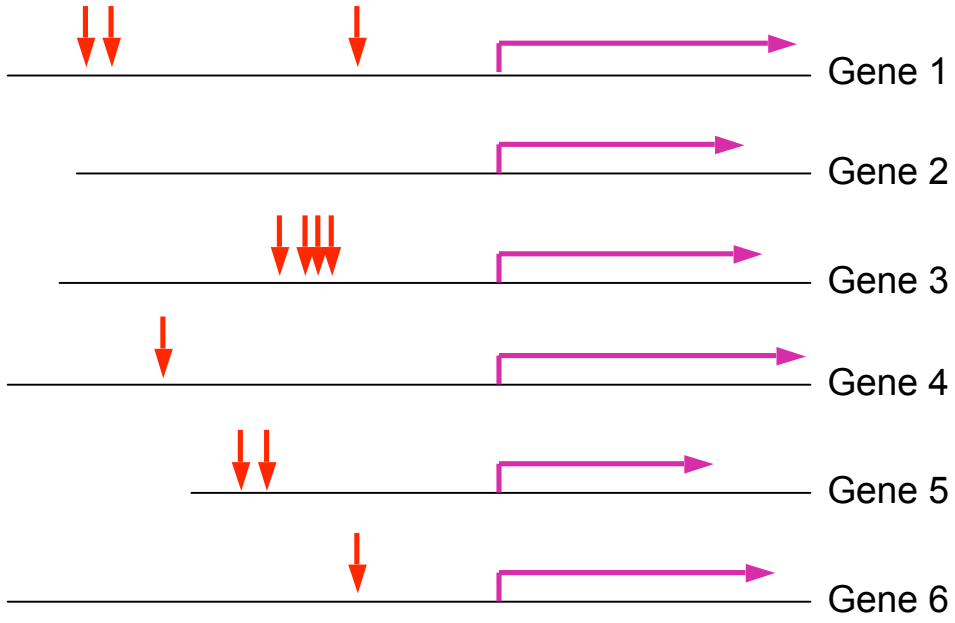
The analysis of regulatory sequences

Pattern discovery

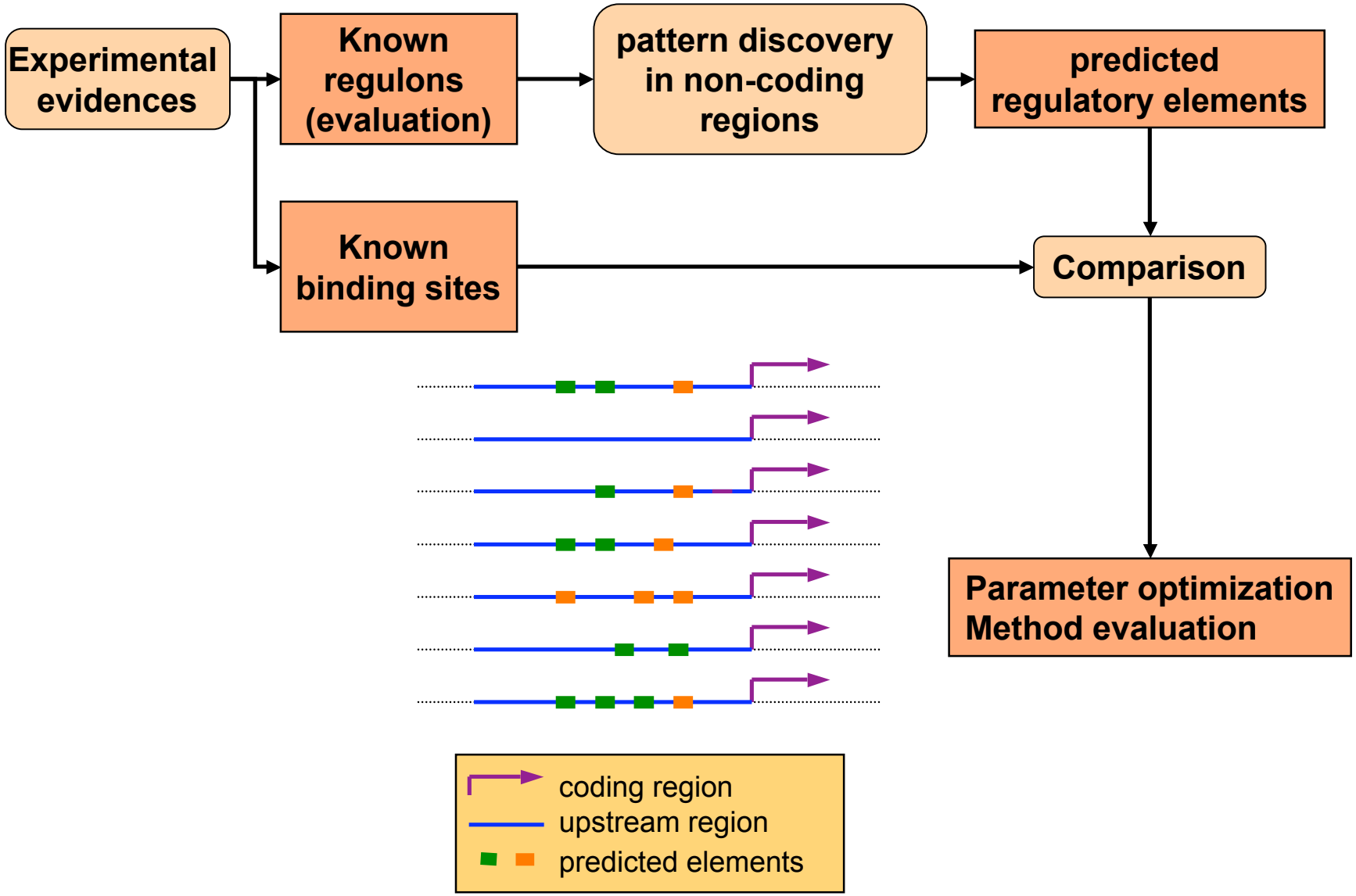
Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Detection of over-represented patterns

- Knowing that a set of genes are co-regulated, one can expect that their upstream regions contains some regulatory signal.
- This signal is likely to be more frequent in the upstream regions of the co-regulated genes than in a random selection of genes.
- In order to discover signals responsible for the co-regulation of a group of genes, we will thus detect over-represented patterns in their upstream sequences.



Evaluation with known regulons



Testing the performances with known regulons

- NIT
  - 7 genes expressed under low nitrogen conditions
- MET
  - 10 genes expressed in absence of methionine
- PHO
  - 5 genes expressed under phosphate stress
- GAL
  - 6 genes expressed in presence of galactose
- ...

The most frequent oligonucleotides are not informative

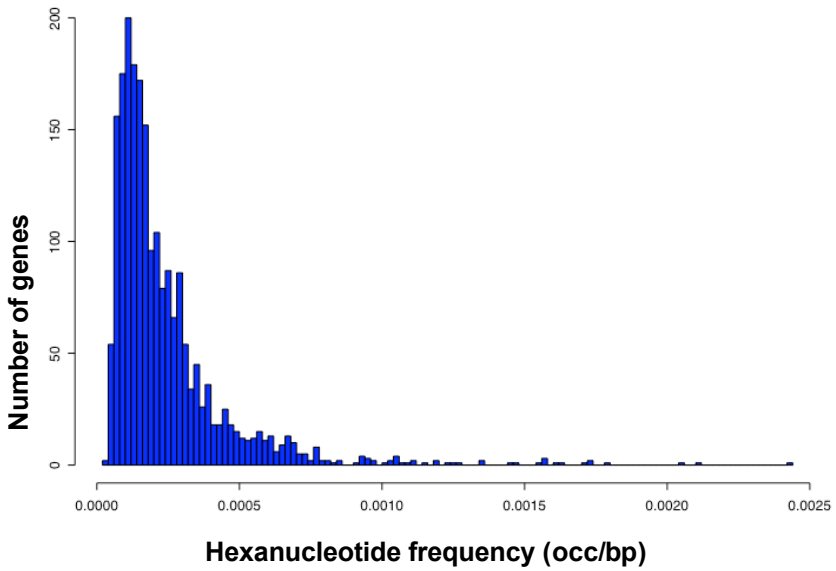
- A (too) simple approach would consist in detecting the most frequent oligonucleotides (for example hexanucleotides) for each group of upstream sequences.
- This would however lead to deceiving results.
  - In all the sequence sets, the same kind of patterns are selected: AT-rich hexanucleotides.

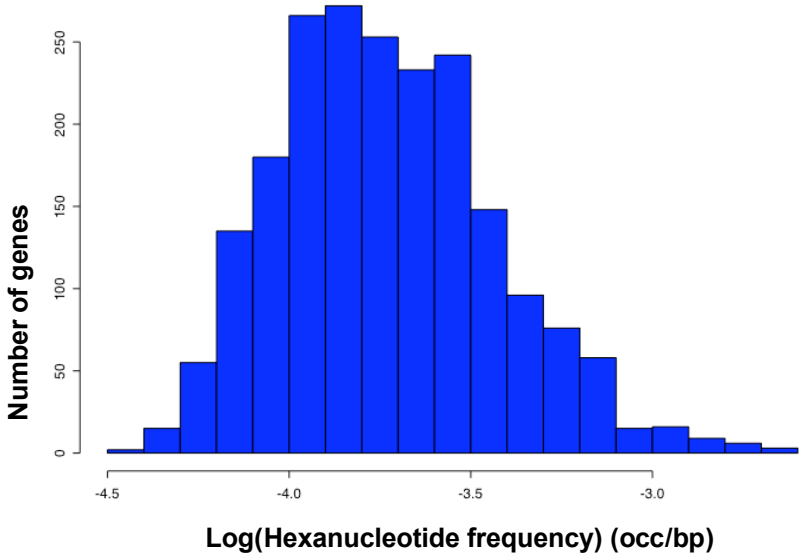
PHO		MET		NIT		GAL	
aaaaaa|tttttt	51	aaaaaa|tttttt	105	aaaaaa|tttttt	80	aaaaaa|tttttt	47
aaaaag|cttttt	15	atatat|atatat	41	cttatc|gataag	26	aaaaat|attttt	17
aagaaa|tttctt	14	gaaaaa|ttttтc	40	tatata|tatata	22	aatata|tatatt	17
gaaaaa|tttttc	13	tatata|tatata	40	ataaga|tcttat	20	aaaatt|aatttt	16
tgccaa|ttggca	12	aaaaat|attttt	35	aagaaa|tttctt	20	aaaata|tatttt	15
aaaaat|attttt	12	aagaaa|tttctt	29	gaaaaa|tttttc	19	attttc|gaaaat	13
aaatta|taattt	12	agaaaa|ttttct	28	atatat|atatat	19	aaataa|ttattt	13
agaaaa|ttttct	11	aaaata|tatttt	26	agataa|ttatct	17	aaatat|atattt	13
caagaa|ttcttg	11	aaaaag|cttttt	25	agaaaa|ttttct	17	ataaaa|ttttat	12
aaacgt|acgttt	11	agaaat|atttct	24	aaagaa|ttcttt	16	atatta|taatat	12
aaagaa|ttcttt	11	aaataa|ttattt	22	aaaaca|tgtttt	16	atatat|atatat	11
acgtgc|gcacgt	10	taaaaa|ttttta	21	aaaaag|cttttt	15	tgaaaa|ttttca	11
aataat|attatt	10	tgaaaa|ttttca	21	agaaga|tcttct	14	caaaaa|tttttg	11
aagaag|cttctt	10	ataata|tattat	20	tgataa|ttatca	14	taaaaa|ttttta	11
atataa|ttatat	10	atataa|ttatat	20	atataa|ttatat	14	agatat|atatct	11

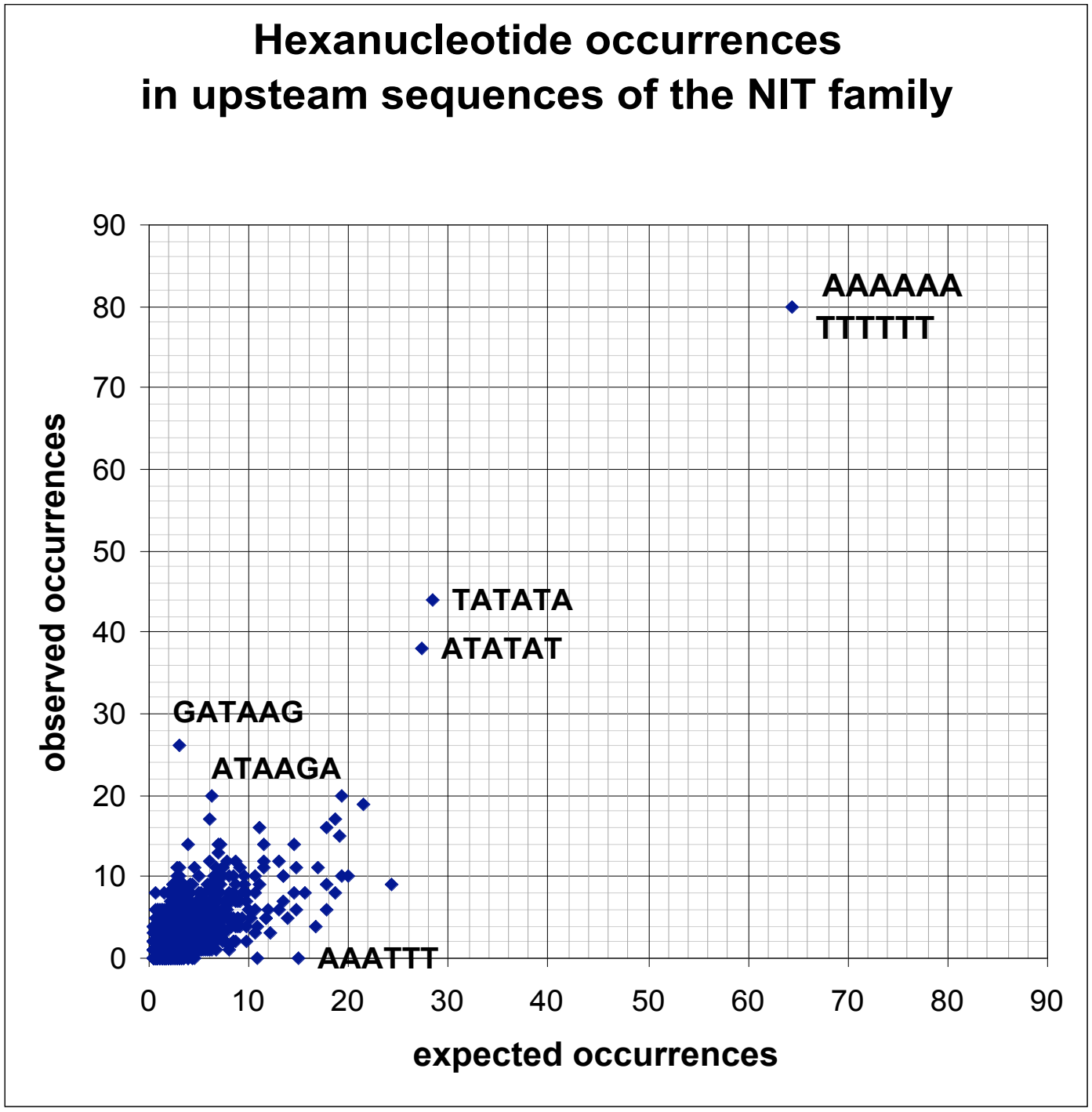
Hexanucleotide frequencies in all upstream sequences

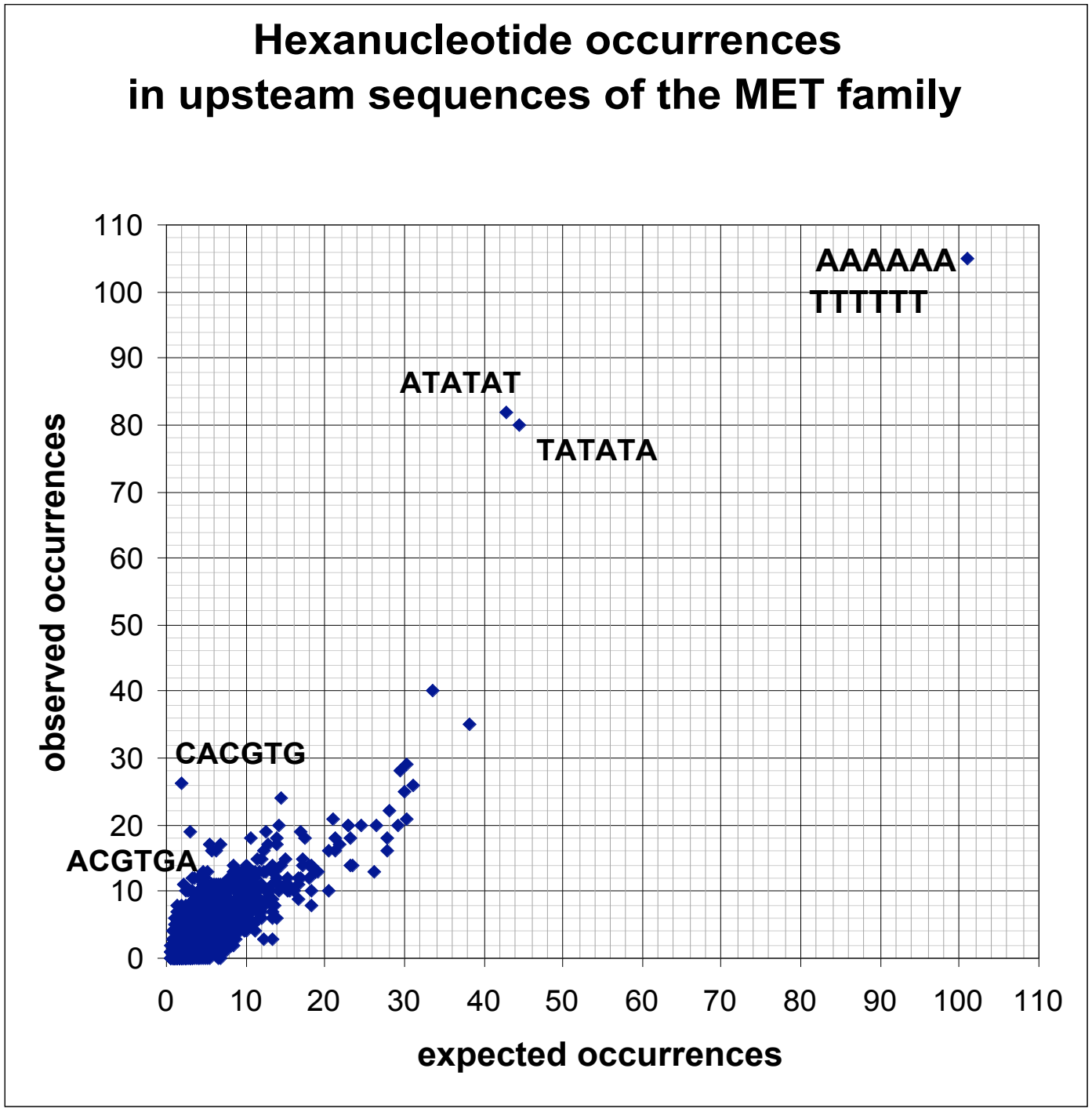
- The frequencies of all the hexanucleotides have been measured in the whole set of 6000 yeast upstream sequences
- Some words are very frequent, others are rare.
  - range 4.5E-5 to 1.2E-2
  - max(f)/min(f)=268

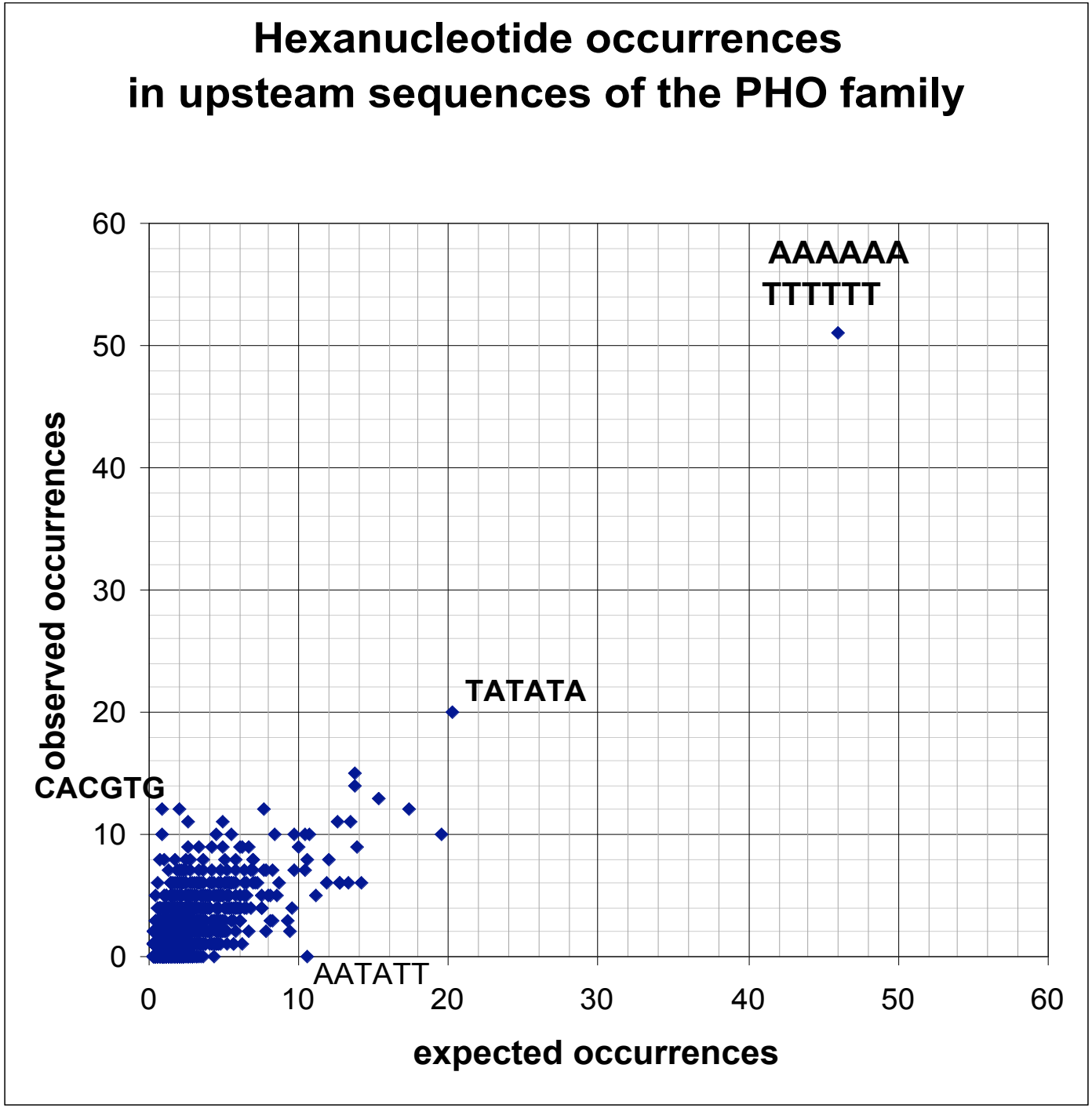
Hexanucleotide frequenies in yeast upstream sequences

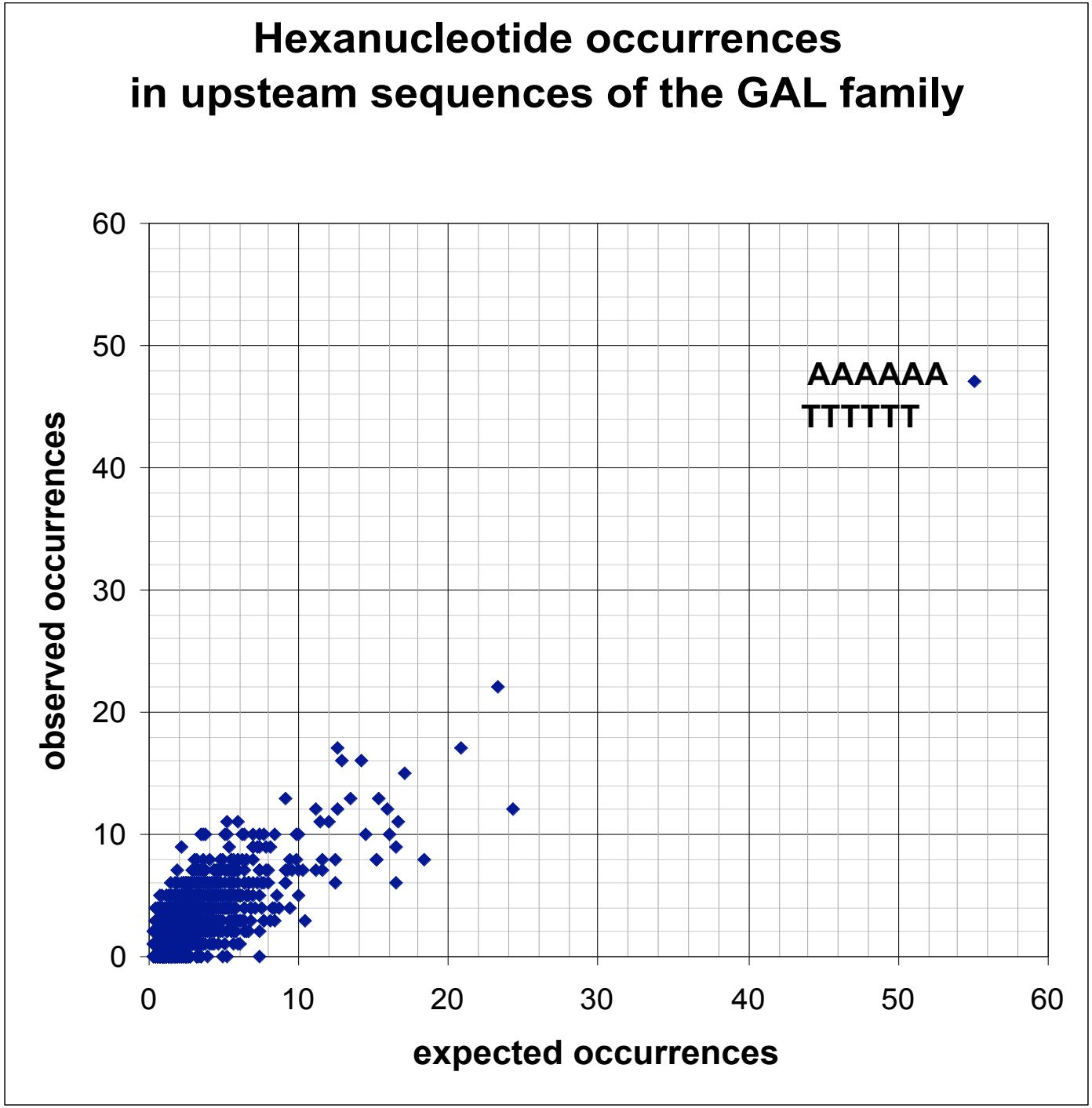





Hexanucleotide occurrences
in upsteam sequences of the NIT family
observed occurrences
expected occurrences
0
10
20
30
40
50
60
70
80
90
AAAAAA
TTTTTT
TATATA
ATATAT
GATAAG
ATAAGA
AAATTT


Hexanucleotide occurrences
in upsteam sequences of the MET family
observed occurrences
expected occurrences
0
10
20
30
40
50
60
70
80
90
100
110
AAAAAA
TTTTTT
ATATAT
TATATA
CACGTG
ACGTGA


Hexanucleotide occurrences
in upsteam sequences of the PHO family
observed occurrences
expected occurrences
0
10
20
30
40
50
60
AAAAAA
TTTTTT
TATATA
CACGTG
AATATT


Hexanucleotide occurrences
in upsteam sequences of the GAL family
observed occurrences
expected occurrences
0
10
20
30
40
50
60
AAAAAA
TTTTTT

Hexanucleotide analysis of the NIT family

Genes	*DAL5, DAL80, GAP1, MEP1, MEP2, MEP3, PUT4*
Known motifs	*Factors*
GATAAg	*Gln3p; Nil1p; Gzf3p; Uga43p*

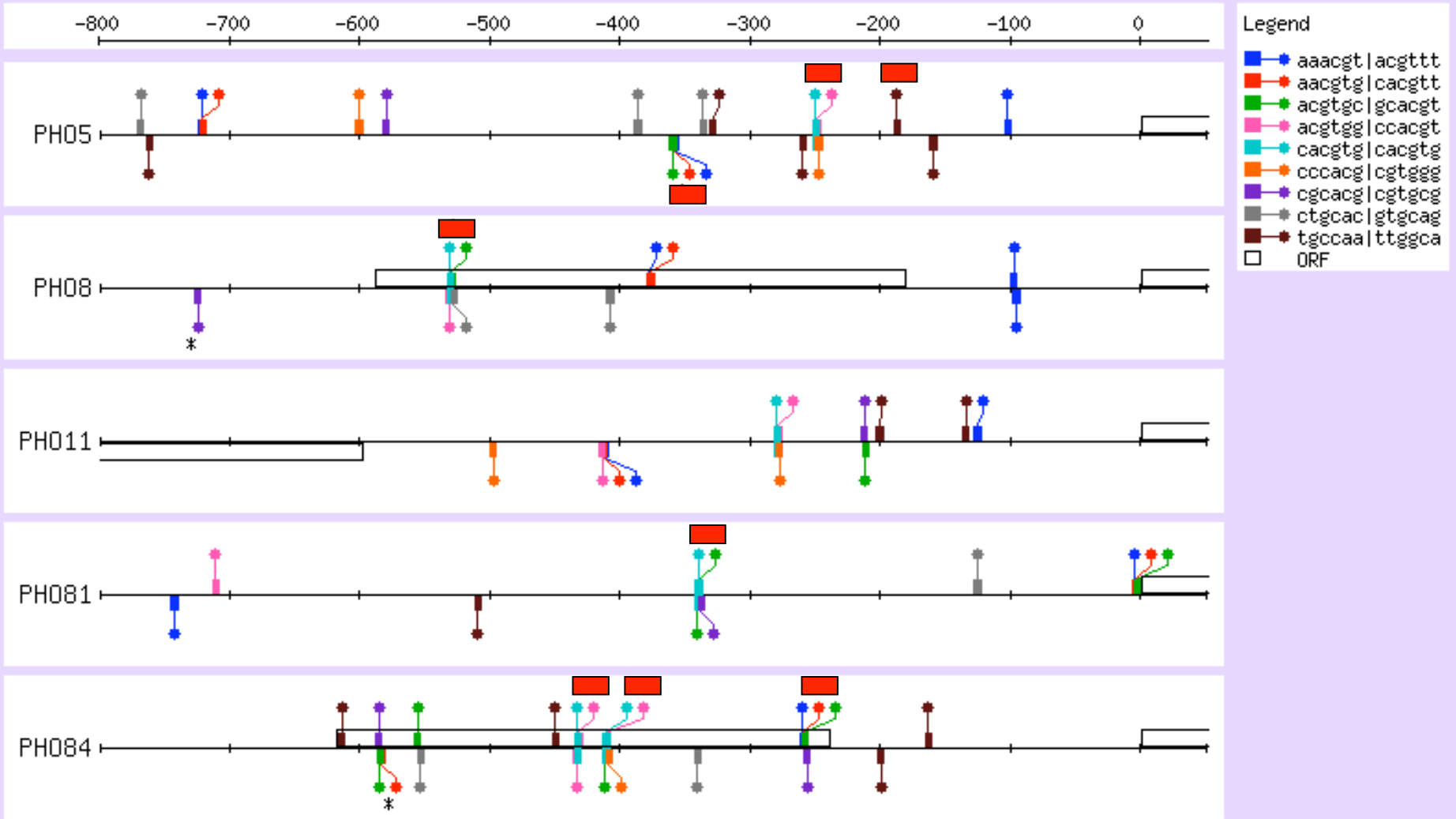
Sequence	exp freq	occ	exp occ	P-value	E-value	sig	matching sequences
...ATAAGa	0.00110	18	6.1	6.20E-05	1.30E-01	0.89	6
..GATAAG.	**0.00053**	**24**	**2.9**	**1.20E-14**	**2.60E-11**	**10.59**	**6**
.cGATAA..	0.00048	10	2.7	0.00044	9.20E-01	0.04	5
ctGATA...	0.00052	11	2.9	0.00019	4.00E-01	0.4	6
acatct	0.00051	11	2.8	0.00016	3.40E-01	0.47	4

Hexanucleotide analysis of the PHO family

Genes	*PHO5, PHO8, PHO11, PHO84, PHO81*
Known motifs	*Factors*
CACGTGGG	*Pho4p (high affinity)*
CACGTTTT	*Pho4p (medium affinity)*

Sequence	exp freq	occ	exp occ	P-value	E-value	sig	matching sequences
.....CGTGGG	0.00013	5	0.5	0.00021	4.30E-01	0.36	3
....ACGTGc.	**0.00021**	**9**	**0.8**	**2.50E-07**	**5.20E-04**	**3.29**	**5**
....ACGTGG.	0.00018	7	0.7	9.00E-06	1.90E-02	1.73	5
...CACGTG..	0.00012	6	0.5	8.90E-06	1.90E-02	1.73	5
.cgCACG....	0.00013	6	0.5	1.40E-05	2.90E-02	1.54	5
ctgCAC...	0.00024	8	1.0	7.80E-06	1.60E-02	1.79	4
....ACGTTT.	0.00061	10	2.4	0.00019	3.90E-01	0.41	5
...CACGTT..	0.00030	7	1.2	0.00024	5.00E-01	0.3	5
tgccaa	0.00048	12	1.9	7.40E-07	1.50E-03	2.81	4

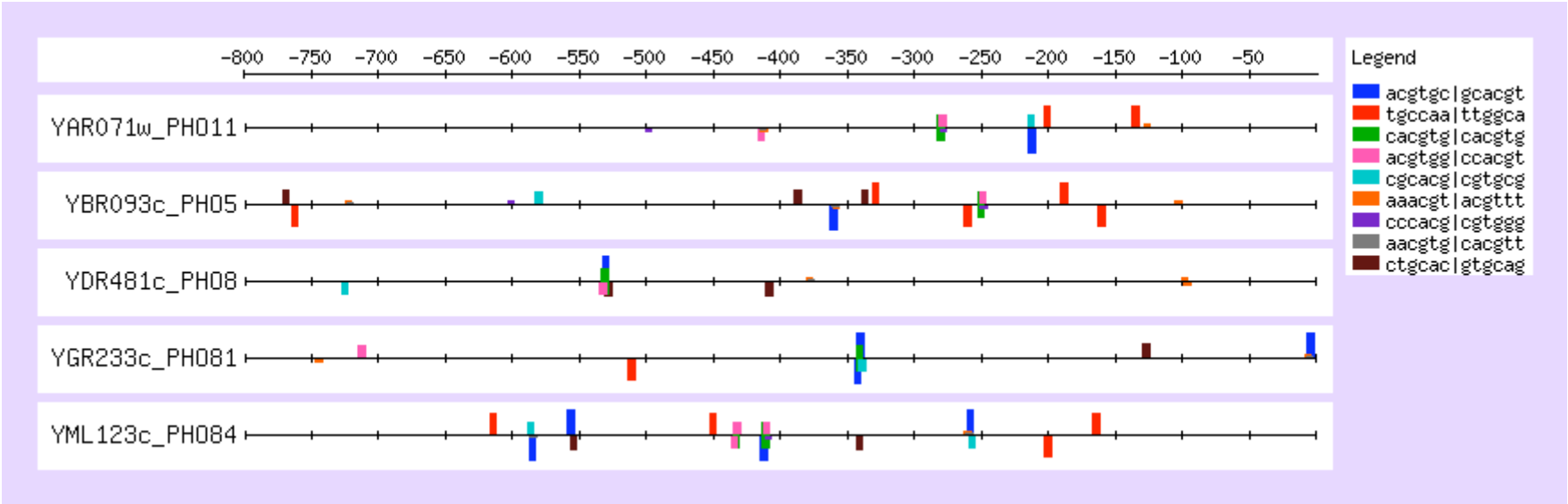
Comparison with known binding sites



Site with experimental evidence

Feature-map of discovered patterns - PHO family

- Each colour represents one pattern.
- Box height represents pattern significance.
- Clusters of mutually overlapping words represent sites larger than 6 bp.



Hexanucleotide analysis of the MET family

Genes *SAM2, MET6, MUP3, MET30, MET3, MET14, MET1, SAM1, MET17, ZWF1, MET2*

Known motifs *Factors*

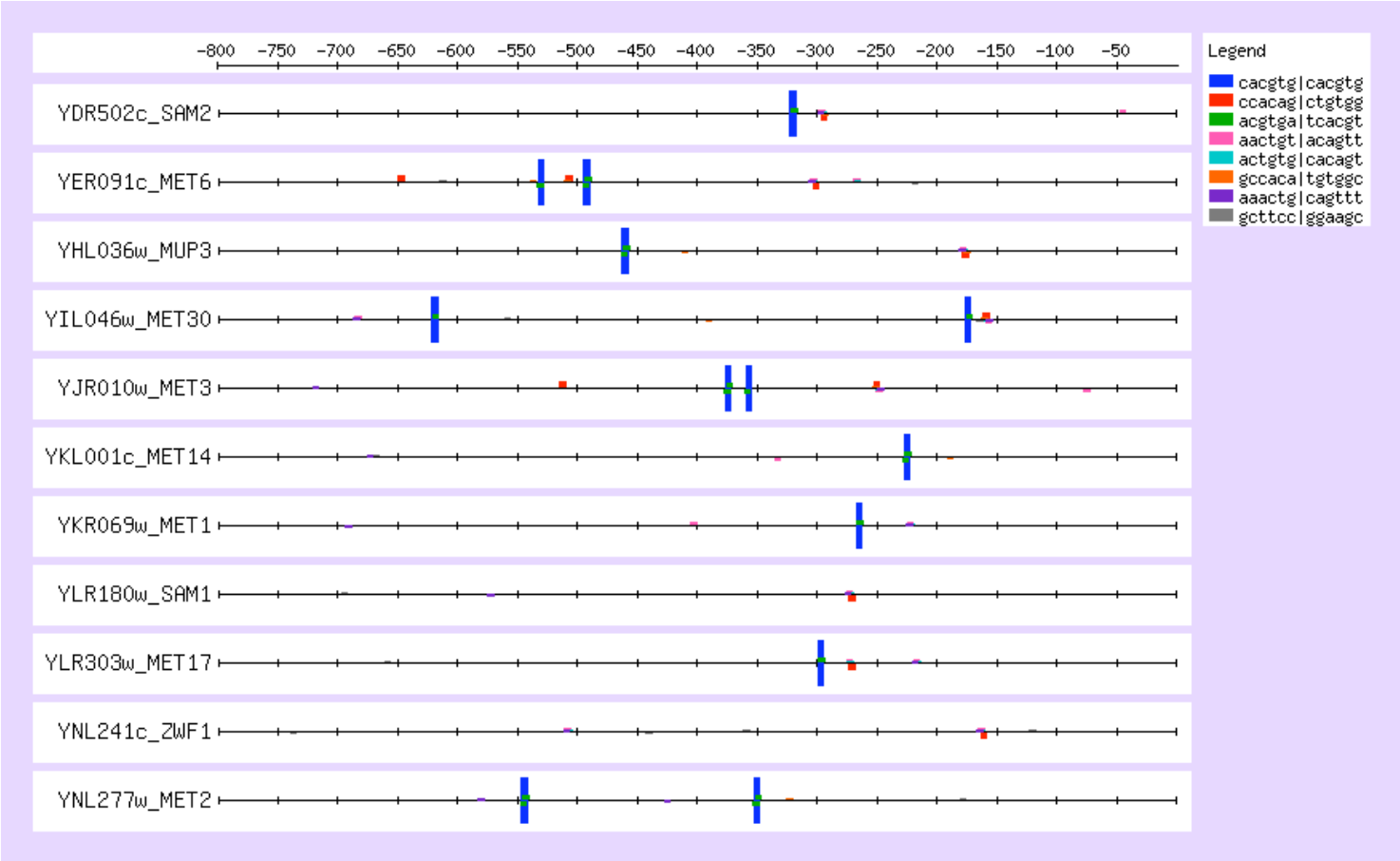
TCACGTG *Cbf1p/Met4p/Met28p*

AAAACTGTGG *Met31p; Met32p*

Sequence	exp freq	occ	exp occ	P-value	E-value	sig	matching sequences
..ACGTGa	0.00033	13	2.9	1.00E-05	2.20E-02	1.67	9
.CACGTG.	**0.00012**	**13**	**1.0**	**6.90E-11**	**1.40E-07**	**6.84**	**9**
tCACGTG.	0.00033	13	2.9	1.00E-05	2.20E-02	1.67	9
tCACGTGa	consensus						
....TGTGGc	0.00027	10	2.3	1.50E-04	3.20E-01	0.49	7
...CTGTGG.	**0.00022**	**11**	**1.9**	**4.30E-06**	**8.90E-03**	**2.05**	**8**
..aCTGTG..	0.00036	12	3.1	9.90E-05	2.10E-01	0.69	9
.aaCTGT...	0.00063	17	5.4	4.90E-05	1.00E-01	0.99	11
aaaCTG....	0.00074	17	6.4	0.00037	7.60E-01	0.12	11
aaaCTGTGGc	consensus						
gcttcc	0.00039	12	3.4	0.00021	4.50E-01	0.35	7

Feature-map of discovered patterns - MET family

- Two distinct motifs (combinations of words) are apparent.
  - blue-green TCACGTGA Met4p/Met28p/Cbf1p
  - red-violet AAACTGTG Met31p; Met32p
- Multiple clustered motifs ar sometimes found, but not always.



Effect of the background model

- The results of string-based pattern discovery depend drastically on the choice of a background model.
- Taking the MET family as example
  - With 6nt calibration in intergenic sequences, the Met4p binding site appears at rank 1, and Met31p at rank 3
  - With equiprobable nucleotides, Met4p only appears are rank 20, and Met31p at rank 32. In other terms, they will never be considered as the most interesting motifs
  - With a single-nucleotide calibration, the Met4p appears at rank 4 and Met31p at rank 13. The first motif would thus have been easily detected, but not the second one.

		Background model		
pattern	rev compl	intergenic	alpha	iid
atcacg....	....cgtgat	9	44	139
gtcacg....	....cgtgac	5	34	266
.tcacgt...	...acgtga.	2	4	20
..cacgtg..	..cacgtg..	1	3	23
...acgtga.	.tcacgt...	2	4	20
....cgtgac	gtcacg....	5	34	266
....cgtgat	atcacg....	9	44	139
gccaca....	....tgtggc	7	17	164
.ccacag...	...ctgtgg.	3	13	99
..cacagt..	..actgtg..	6	21	75
...acagtt.	.aactgt...	4	19	32
....cagttt	aaactg....	10	18	33
gcttcc	ggaagc	8	10	77

Effect of oligonucleotide size on the significance

		oligoncleotide length					
Family	**Pattern**	**4**	**5**	**6**	**7**	**8**	**9**
NIT	**aGATAAGa**	1.8	4.1	**9.1**	4.6	0.9	-
MET	**gTCACGTG**	4.4	4.1	7	**8.2**	3.2	-
	AAACTGTGg	1.5	2.3	1.6	4.8	**5.2**	4.9
PHO	**CACGTggg**	4.7	**8.4**	4.4	4.3	4.3	-
	aTGCCAA	**2.6**	1.5	**2.6**	0.6	-	-
	CTGCAC	-	-	1.7	-	-	-
INO	**CAACAAg**	2.9	2.1	**3.7**	1.3	-	-
	cCATGTGAA	-	-	2.7	3.2	**6.4**	0.4
PDR	**tCCGTGGa**	1.5	3.3	**7.4**	6.9	4.2	1.4
	tCCGCGga	6.9	**7.1**	4.5	5.6	1.8	1
GCN4	**GCNgtGACTCa**	5.4	**8.8**	8.2	7.7	4.7	-
	CAGCGGa	3.3	3.5	**4**	0.6	-	-
YAP	**CATTACTAA**	-	-	1	2.3	2.1	**3.2**
	cCGTTCC	0.1	0.5	**3.3**	0.3	-	-
YAP (400bp)	**cATTACTAA**	-	-	0.7	**4.5**	2.5	3.5
	cCGTTCC	0.8	0.5	**2.4**	0.7	0.2	-
TUP	**gtGGGGta**	**10.1**	9	8.6	5.6	3	-
	catAGGCAC	3.3	3.3	**4.3**	2.6	3.3	1.7

oligo-analysis results with known regulons (sig > 1)

Family	Factor	DNA-binding Domain	Known motifs	oligont	reverse oligont	score
NIT	GATA factors	Zn finger	GATAAG	TCTTATCT	AGATAAGA	20.0
MET	Cbf1p/Met4p/Met28p Met31p, Met32p	bHLH/bLZ/bLZ Zn finger	TCACGTG AAAACTGTGG	CACGTGAT CACGTGAC AACTGTGGCG	ATCACGTG GTCACGTG CGCCACAGTT	9.0 9.0 3.6
PHO	Pho4p (high affinity) Pho4p (medium affin.)	bHLH bHLH	GCACGTGGG GCACGTTTT	CCCACGTGCG AAACGTGCG TGCCAA CTGCAC	CGCACGTGGG CGCACGTTT TTGGCA GTGCAG	4.4 4.4 2.6 1.8
PDR	Pdr1p, Pdr3p	Zn_2Cys_6 binuclear cluster	tytCCGYGGary	TCCGTGGAA TCCGCGG	TTCCACGGA CCGCGGA	7.4 4.5
GCN4	Gcn4p	bZip	RRTGACTCTTT	ATGACTCA AGTGACTCA ATGACTCT ATGACTCC ATGACTA CCGCTG GCCGGT	TGAGTCAT TGAGTCACT AGAGTCAT GGAGTCAT TAGTCAT CAGCGG ACCGGC	8.5 8.5 8.5 8.5 3.8 3.7 1.3
INO	Ino2p/Opi1p	bHLH/leucine zipper	CATGTGAAWT	CAACAACG CAACAAG TTCACATG	CGTTGTTG CTTGTTG CATGTGAA	3.8 3.8 2.8
HAP 2/3/4	Hap2/3/4/5p		CCAAY	AGAGAGA	TCTCTCT	2.8
GAL4	Gal4p	Zn_2Cys_6 binucl. cluster	$CGGn_{11}CCG$	no significant pattern		

van Helden et al. (1998). *J Mol Biol* **281**(5), 827-42.

Hexanucleotide analysis of the GAL family

Genes *GAL1, GAL2, GAL7, GAL80, MEL1, GCY1*

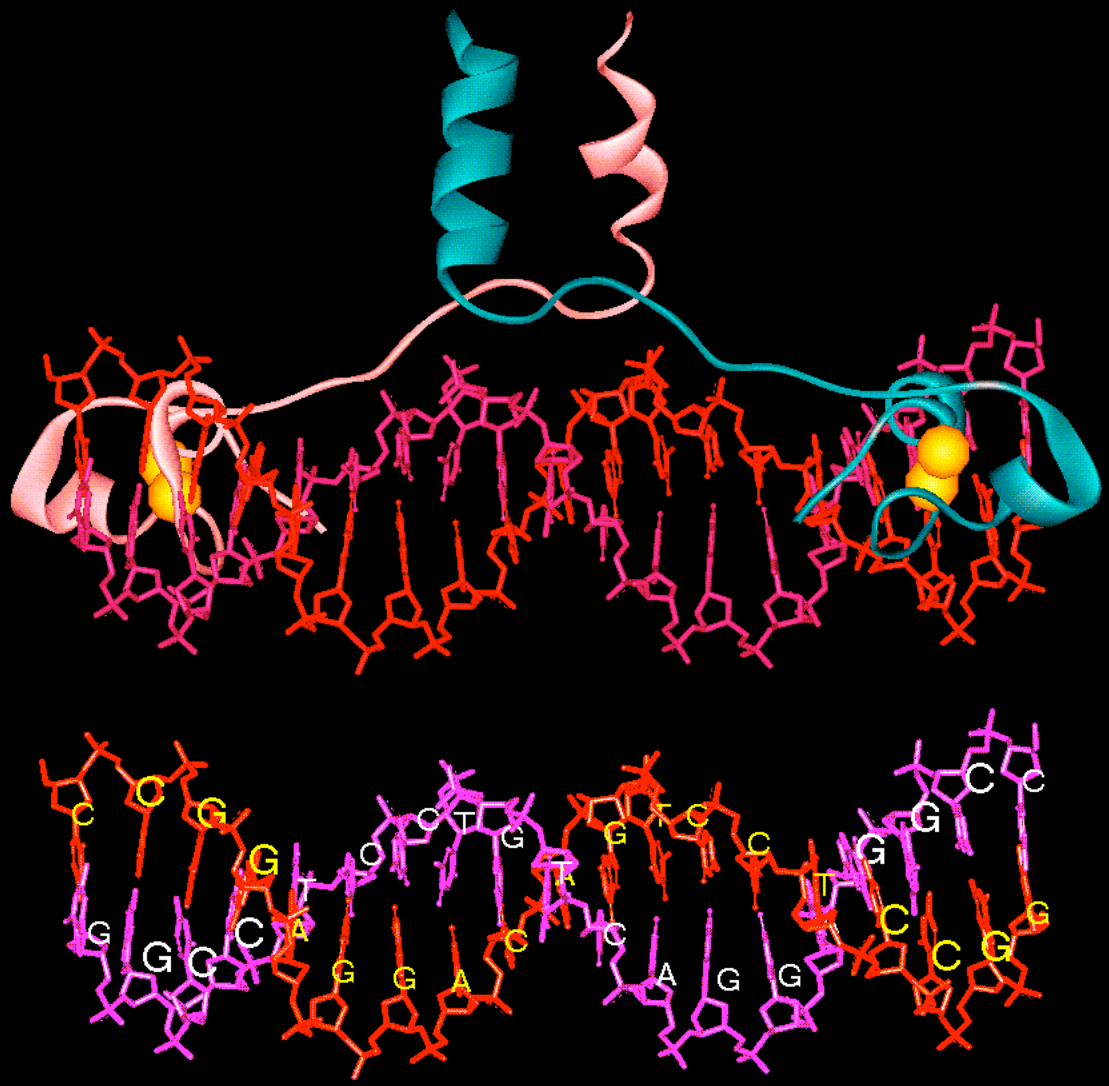
Known motifs *Factors*

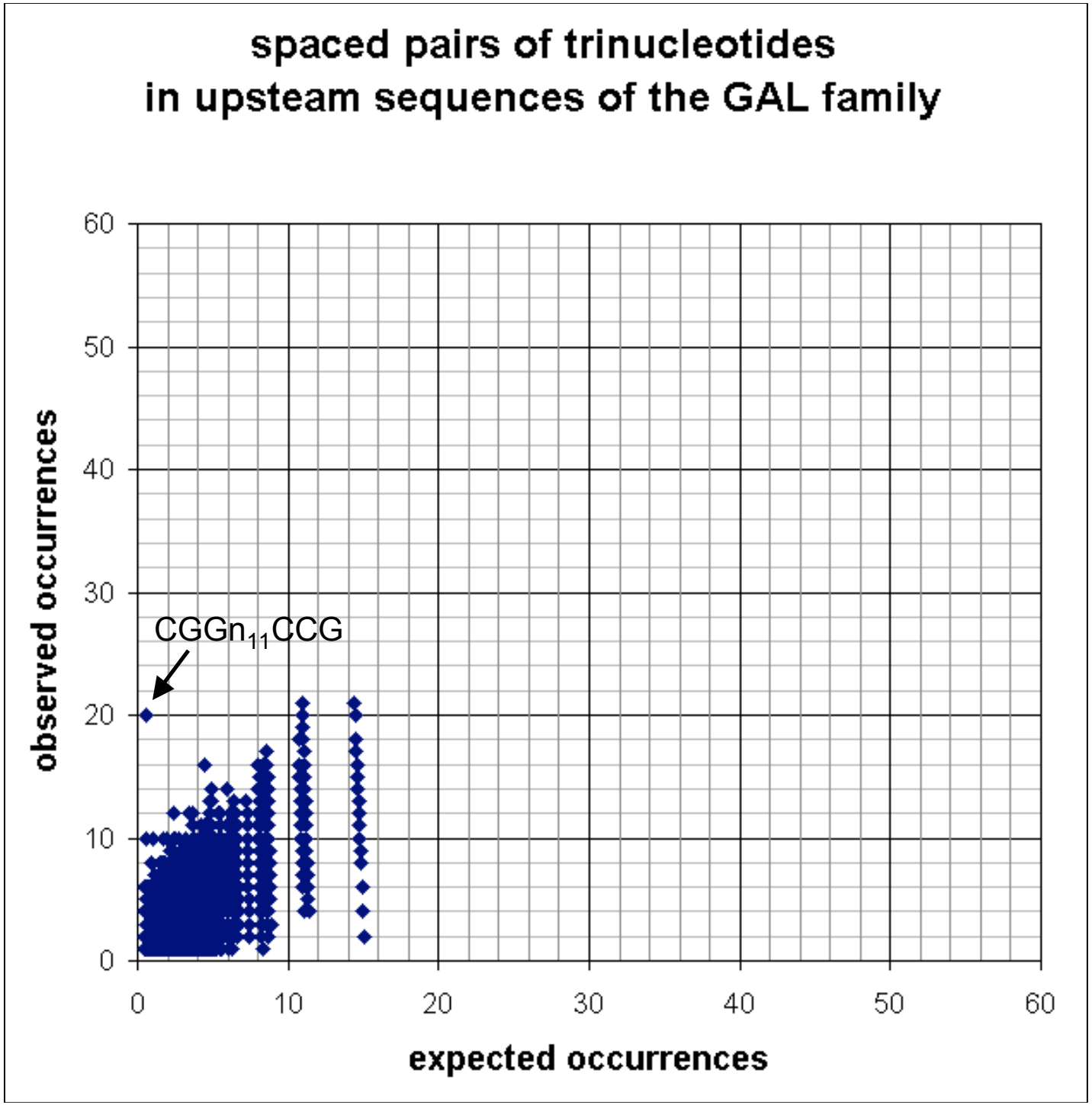
CGGn$_5$wn$_5$CCG *Gal4p*

Sequence	exp freq	occ	exp occ	P-value	E-value	sig	matching sequences
agacat	0.00044	9	2.1	0.00033	0.69	0.16	4

- With the GAL family, the program returns a single pattern.
  - The significance of this pattern is very low.
  - This level of significance is expected at random ~ once per sequence set.
  - This can be considered as a negative result: the program did not detect any really significant pattern.
- Why did the program fail to discover the GAL4 motif ?

Structure of the Gal4p-DNA interface




spaced pairs of trinucleotides
in upsteam sequences of the GAL family
observed occurrences
expected occurrences
CGGn11CCG
0
10
20
30
40
50
60

Dyad analysis of the GAL family

Genes *GAL1, GAL2, GAL7, GAL80, MEL1, GCY1*

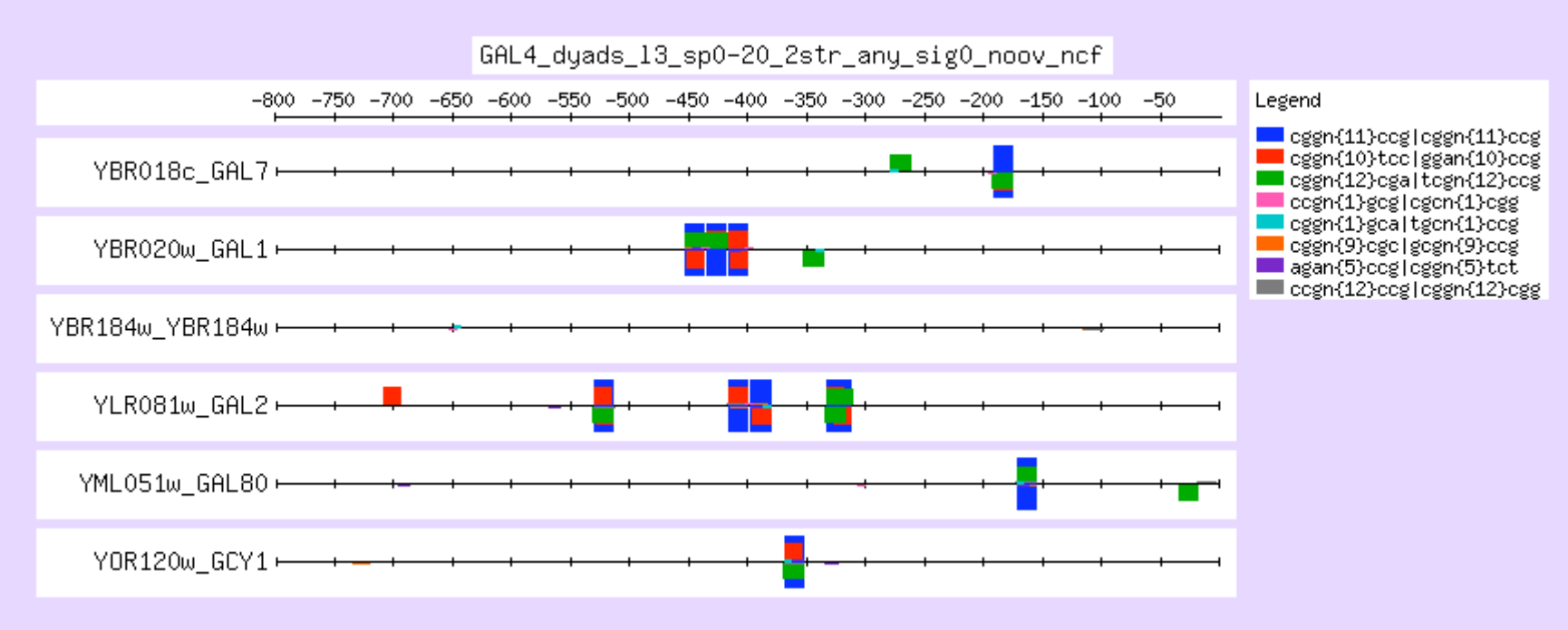
Known motifs *Factors*

CGGn$_5$wn$_5$CCG *Gal4p*

Sequence	exp freq	obs occ	exp occ	P-value	E-value	sig
..GGa..........CCG.	0.00006	10	0.5	2.70E-10	1.20E-05	4.92
.CGG..........Cga	0.00006	10	0.5	4.80E-10	2.10E-05	4.68
.CGG..........CCG.	**0.00007**	**20**	**0.6**	**2.10E-12**	**9.20E-08**	**7.03**
.CGG..........tCC..	0.00006	10	0.5	2.70E-10	1.20E-05	4.92
.CGG..........cgC...	0.00004	6	0.4	5.30E-06	2.30E-01	0.64
tCG..........CCG.	0.00006	10	0.5	4.80E-10	2.10E-05	4.68
cCG..........CCG.	0.00005	6	0.4	6.40E-06	2.80E-01	0.55
yCGGa.......ckCCGa						
AGA.....CCG	0.00010	8	0.9	7.00E-06	3.10E-01	0.51
CCG.GCG	0.00005	6	0.5	9.30E-06	4.00E-01	0.39

Feature-map of discovered patterns - GAL family

- Clusters of overlapping dyads indicates that conservation extends over 3 bp on each side of the dyad.
- Some genes, but not all, contain multiple motifs (synergic effect).



Dyad analysis: regulons of Zn cluster proteins

FACTOR	# genes	KNOWN MOTIFS	DYADS	REVERSE DYADS	SCORE
GAL4	6	CGGn$_{11}$CCG	TCGGAn$_9$TCCGG TCGGCGCAGAn$_4$TCCGG	CCGGAn$_9$TCCGA CCGGAn$_4$TCTGCGCCGA	7.8 7.8
HAP1	9	CGGnnntanCGG	GGAn$_5$CGGC GGGGGn$_{12}$GGC CCTn$_{10}$GGC	GCCGn$_5$TCC GCCn$_{12}$CCCCC GCCn$_{10}$AGG	1.8 1.4 1.1
LEU3	5	RCCggnnccGGY	CCGn$_3$CCG	CGGn$_3$CGG	1.0
LYS	6	wwwTCCrnyGGAwww	AAATTCCG TCCGCTGGA	CGGAATTT TCCAGCGGA	1.9 1.0
PDR	6	tytCCGYGGary	CTCCGTGGAA CTCCGCGGAA	TTCCACGGAG TTCCGCGGAG	6.7 6.7
PPR1	3	wyCGGnnwwykCCGaw		CGGn$_6$CCG	0.5
PUT3	2	yCGGnangcgnannnCCGa	CGGn$_{10}$CCG	CGGn$_{10}$CCG	1.2
UGA3	3	aaarccgcsggcggsawt	CGGn$_{14}$AGG GCCn$_{11}$TCC	CCTn$_{14}$CCG GGAn$_{11}$GGC	1.7 1.0
UME6	25	tagccgccga	TCGGCGGCTA	TAGCCGCCGA	4.9
CAT8	5	CGGnnnnnnGGA	CGGn$_4$ATGGAA	TTCCATn$_4$CCG	6.0

van Helden et al. (2000). *Nucleic Acids Res* **28**(8), 1808-18.

Regulatory sequence analysis

Genome-scale pattern discovery

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Estimation of expected frequencies with Markov models

- Estimation of expected word frequencies
  - On basis of the input sequences themselves
  - Markov chain models: the expected frequency of each k-letter word is estimated on basis of sub-word frequencies.

- Example
  - Estimation of hexanucleotide frequencies with a 4^{th} order Markov chain mode.

$$\text{e.g.: } \exp\{GATAAG\} = \frac{\text{obs}\{GATAA\} \times \text{obs}\{ATAAG\}}{\text{obs}\{ATAA\}}$$

Oligo-analysis with Markov chain models

- Analysis of a set of 6217 downstream sequences, 200bp each
- Detection of over-represented words, and grouping by sequence similarity

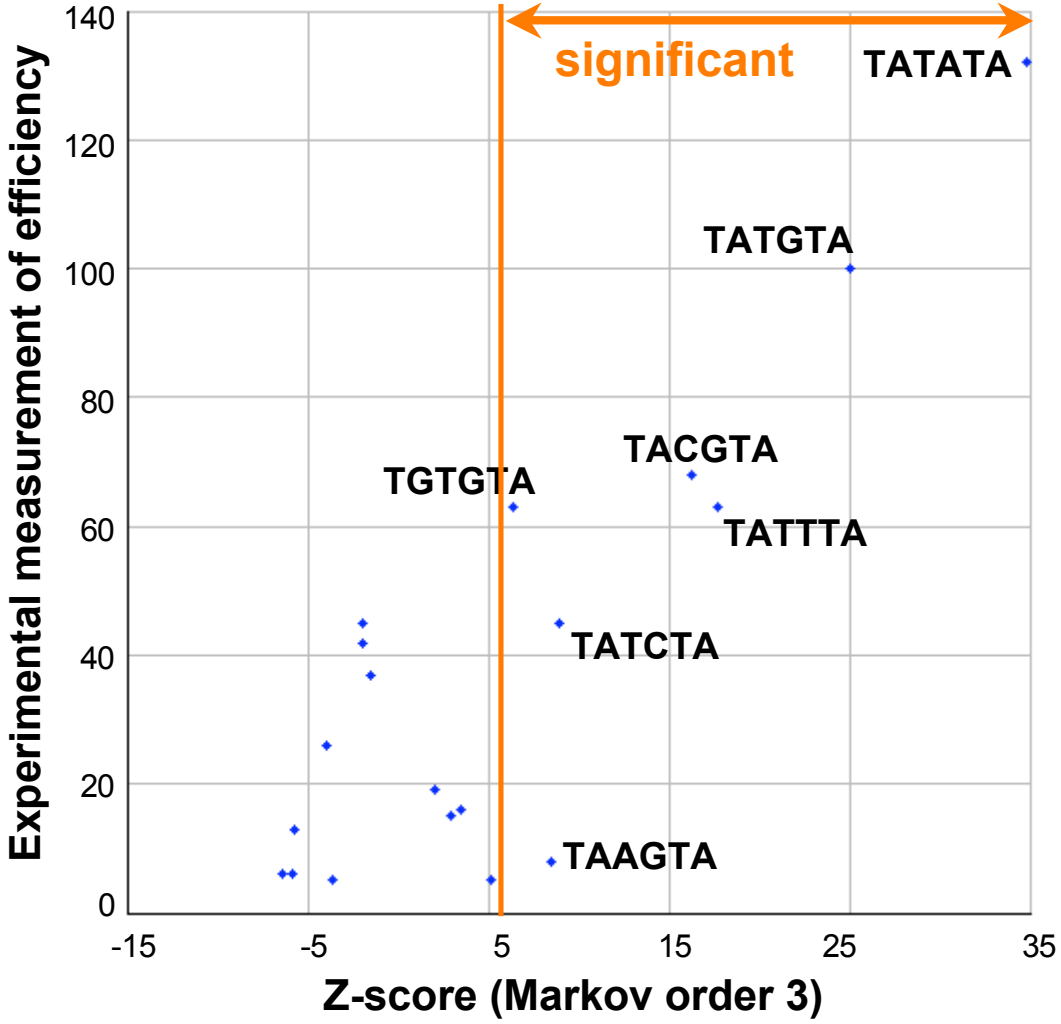
ATATAT.	27.0
ATACAT.	15.5
ATGTAT.	11.9
ATAAAT.	9.9
ATAGAT.	9.9
ATTTAT.	9.8
GTATAT.	8.2
ATATGT.	7.8
ACATAT.	7.7
ATATAC.	7.4
.TATATA	34.9
.TACATA	27.7
.TATGTA	25.0
.TAAATA	22.0
.TATTTA	17.7
.TAGATA	11.9
.TGTATA	8.6
.TATACA	7.3
.CATATA	3.5

AAAAAA	18.28
AAATAA	16.65
AATAAA	14.09
AAGAAA	9.27
AACAAA	9.02
AAAGAA	8.17
AAACAA	7.69

TTTTTT	16.87
TTATTT	16.74
TTTATT	13.25
TTTCTT	9.42
TTTGTT	8.72
TTCTTT	8.46

ACATAC.	12.21
ACACAC.	11.15
.CACACA	13.00
.CATACA	8.81

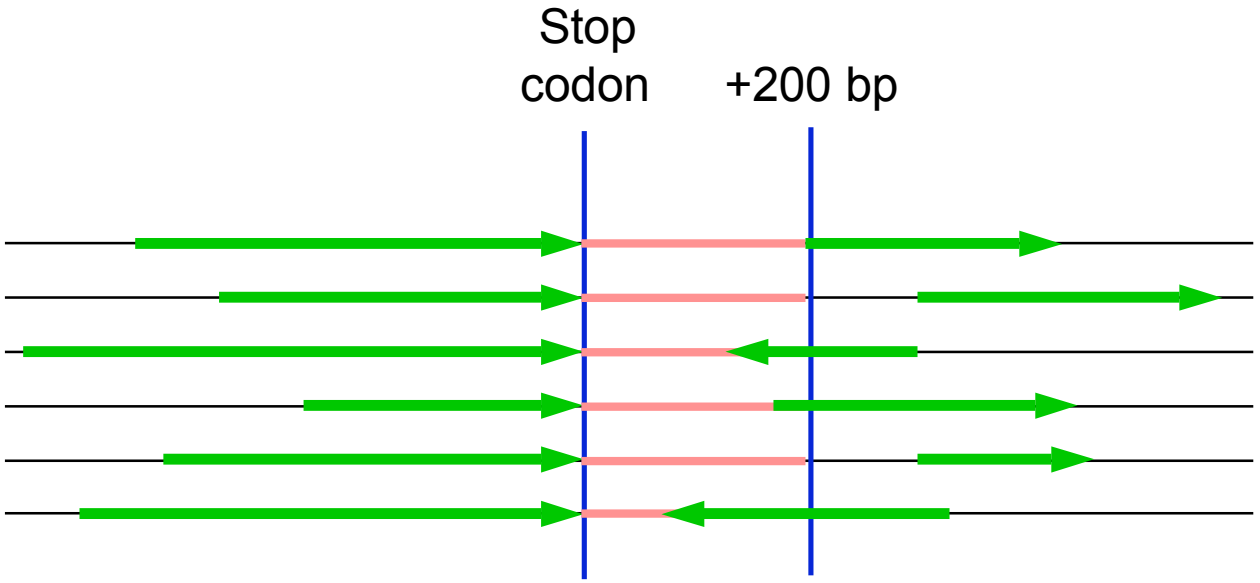
Comparison with experimental values



- Irniger and Braus (1994) performed a saturation mutagenesis and measured the the efficiency of all single-base mutants of TATGTA.
- High Z-score values from Markov 4 model correlate pretty well with experimental efficiency

Position analysis

- Measure the positional distribution of each word
- Perform a test of homogeneity and select all words with a significant bias
- Significance of the non-homogeneity is estimated with a χ^2 test
- Note : in our case, homogeneous is not flat, because sequences are clipped when there is a downstream ORF closer than 200 bp

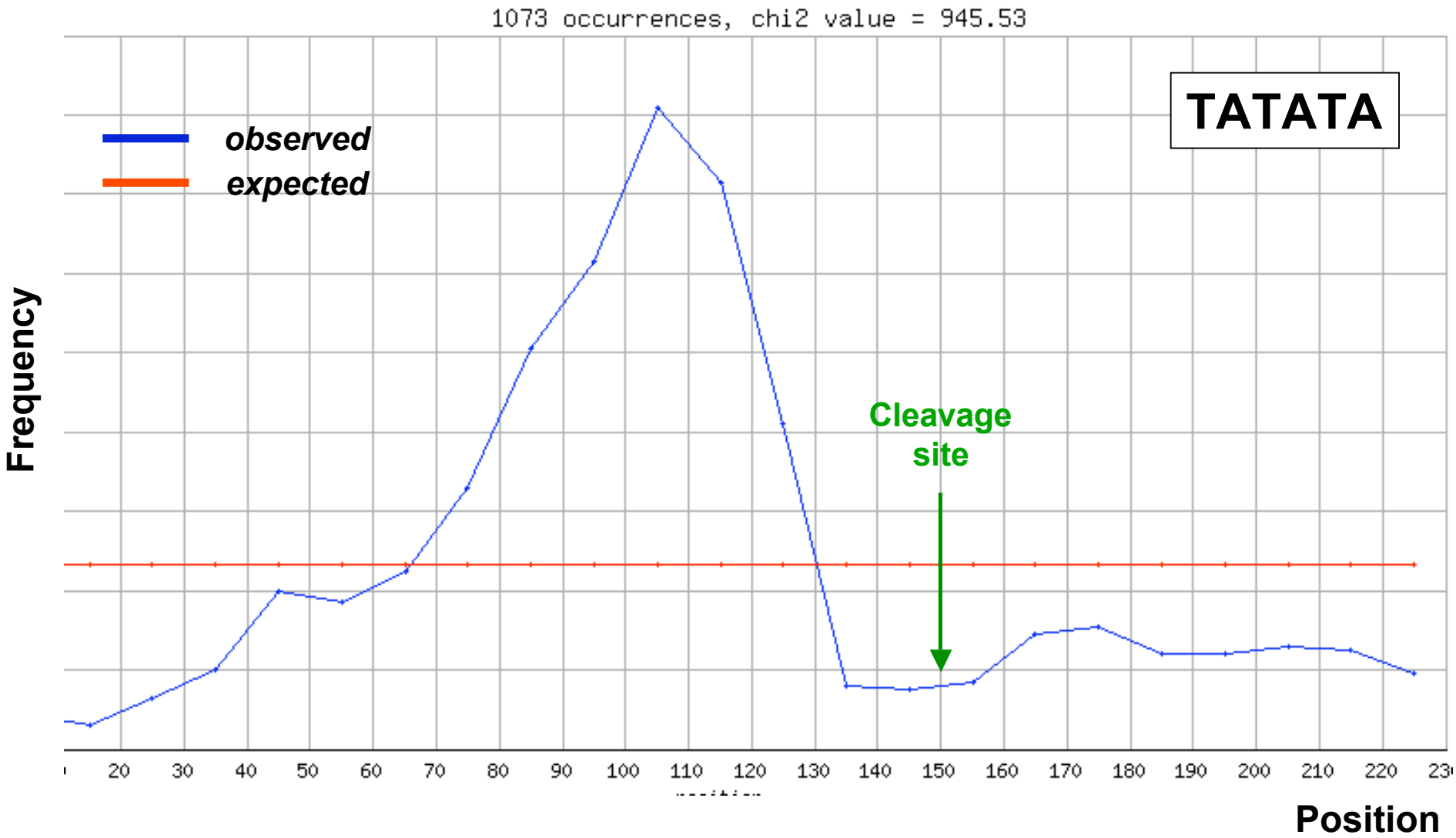


Word position distribution

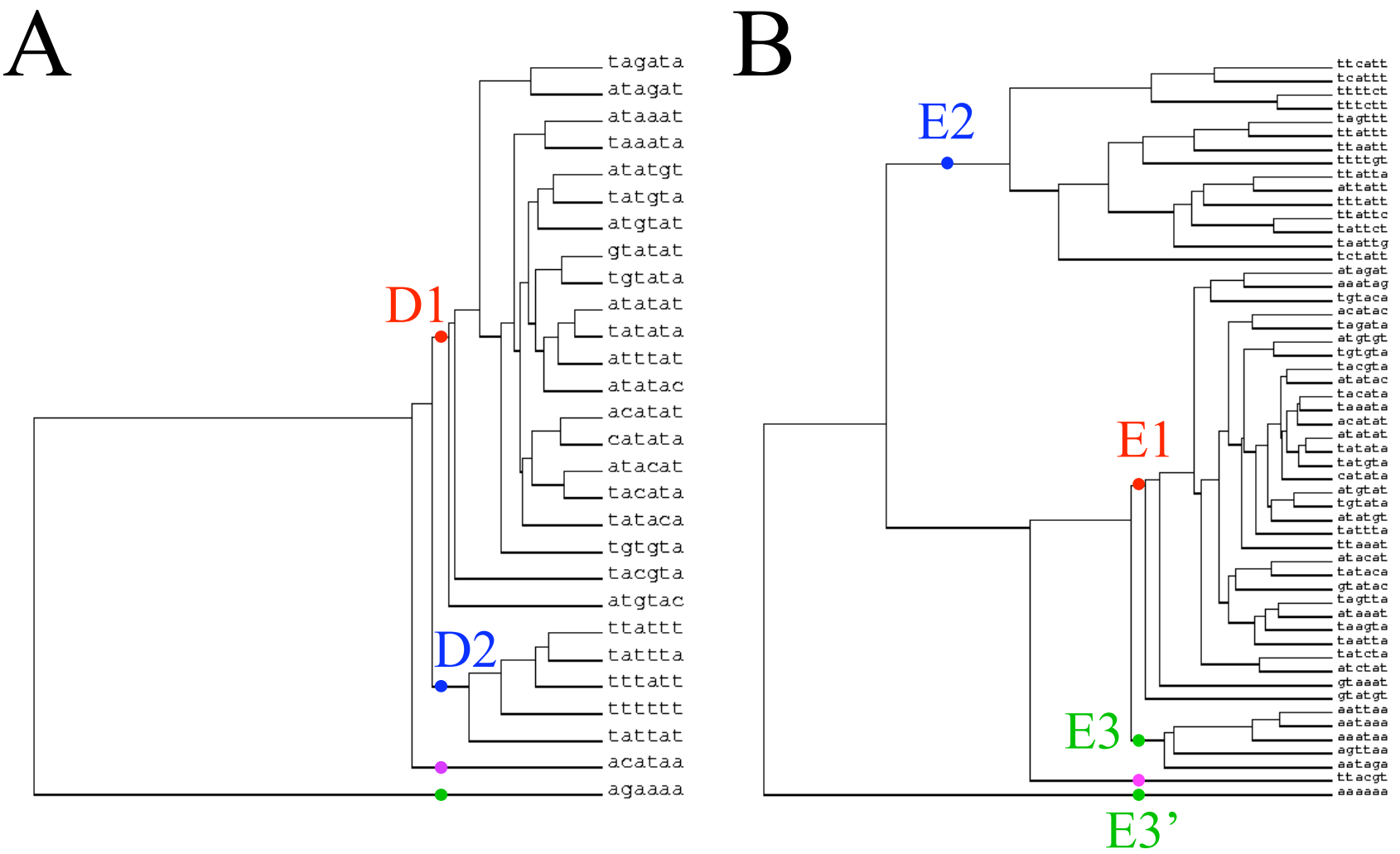
- Positions relative to the stop codon

Position analysis : profiles of word distribution

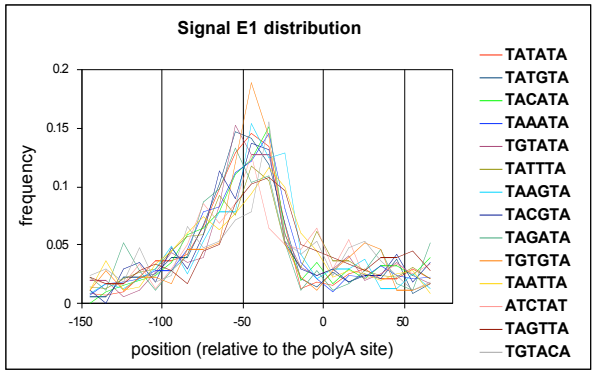
- Positions relative to the cleavage site

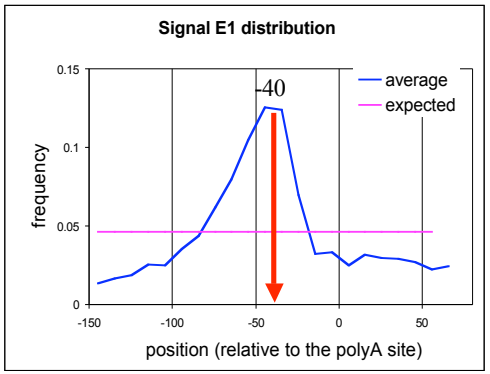


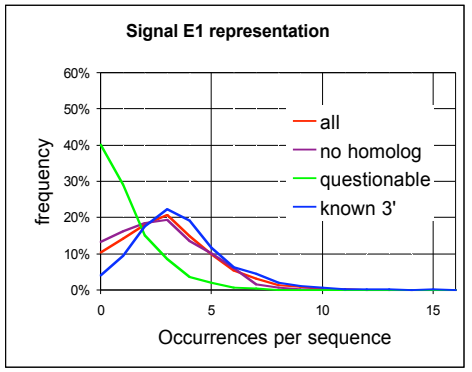
Word clustering according to position profiles

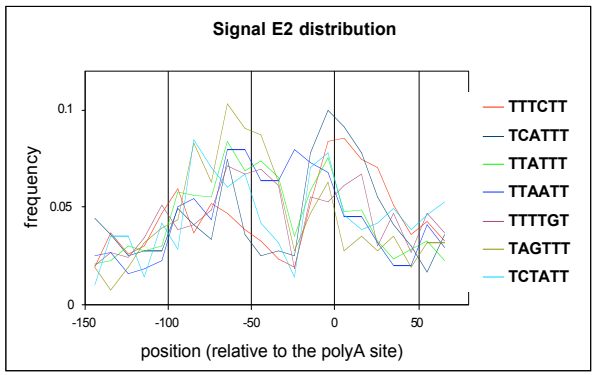


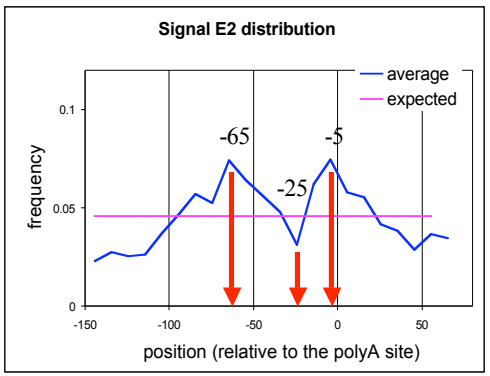
Signal distribution and representation

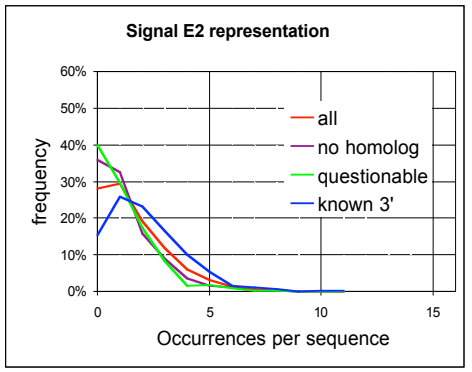


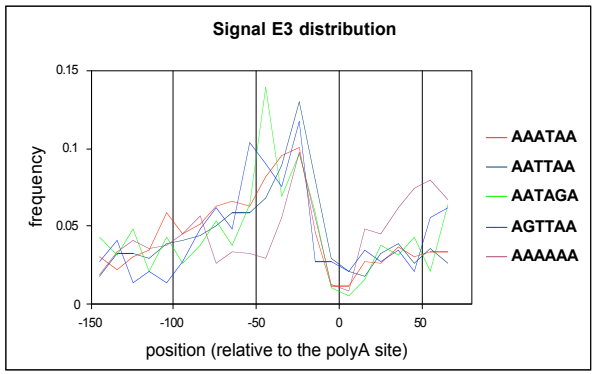


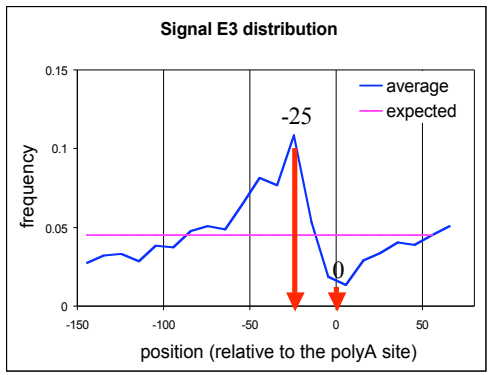


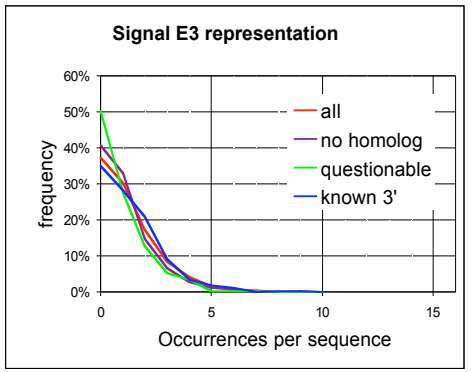










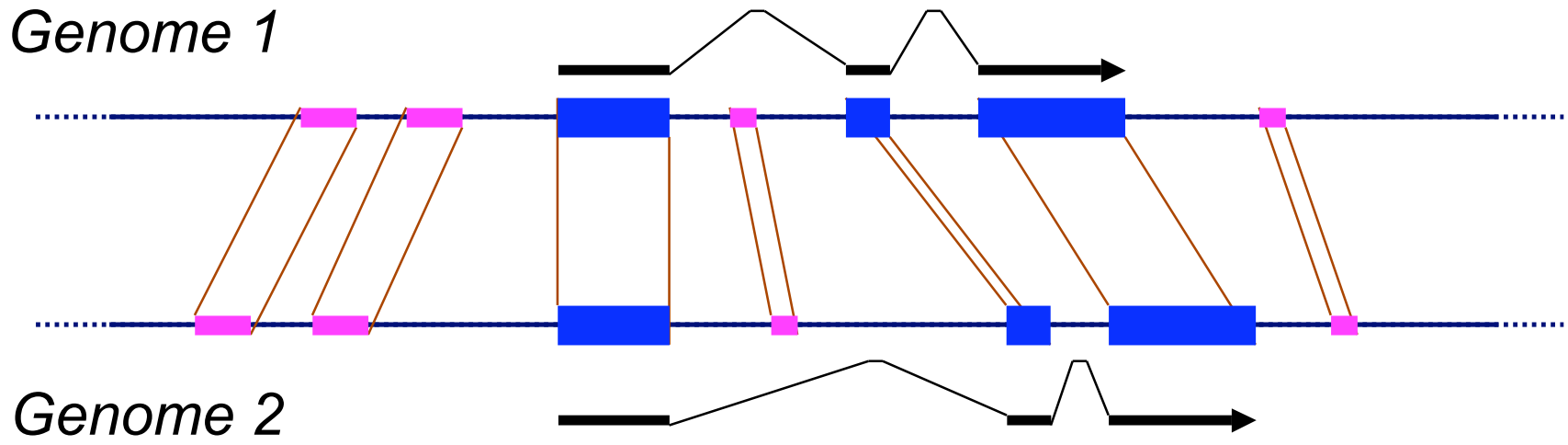


Regulatory Sequence Analysis

Phylogenetic footprinting

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

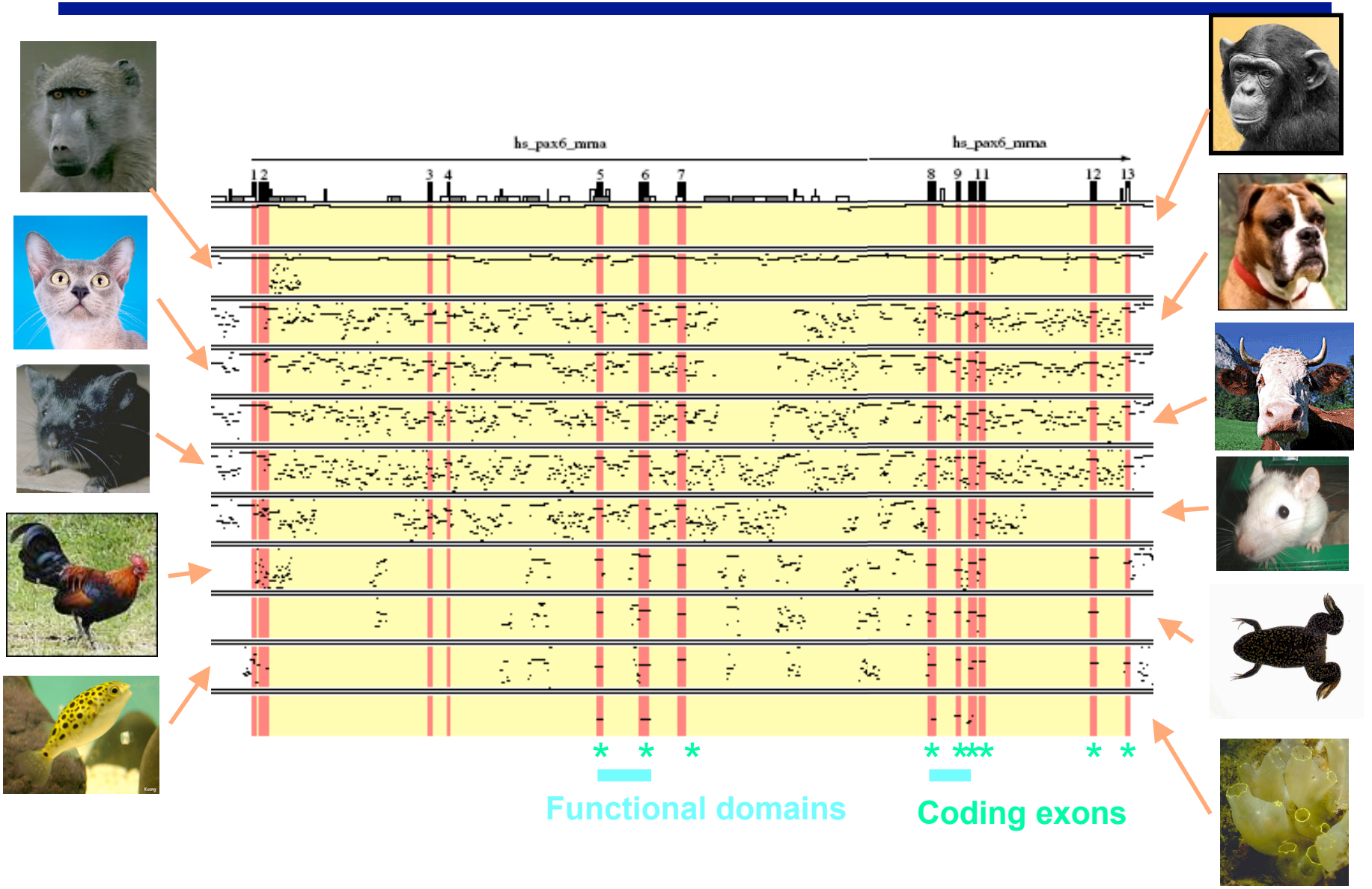
Phylogenetic footprinting to define regulatory regions



conserved exon *conserved non-coding region*

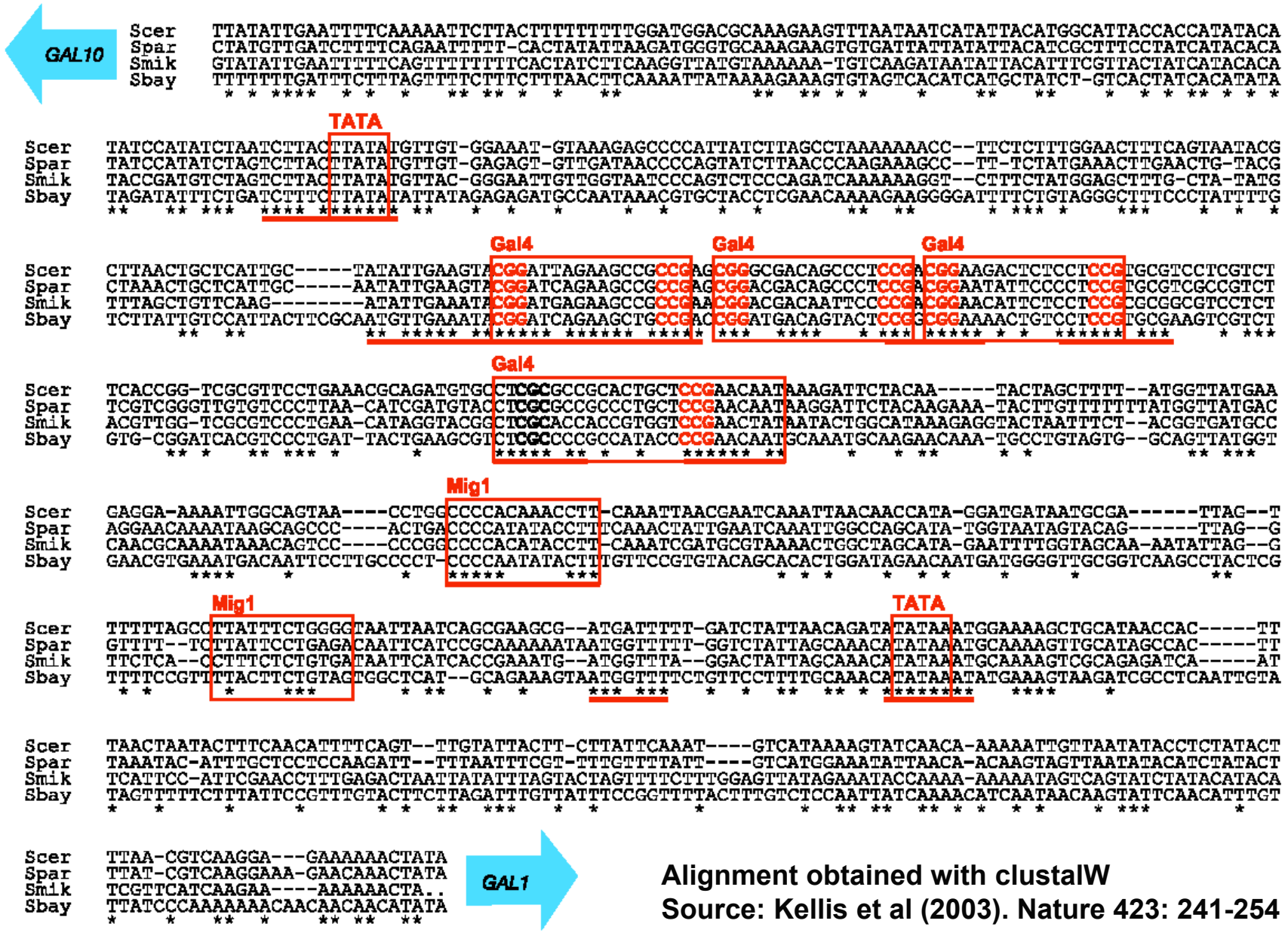
- Within non-coding sequences, regulatory elements evolve slower than their surrounding.
- Conserved non-coding sequences contain a high concentration in regulatory elements.

Phylogenetic footprints for the pax6 gene



Slide from Philippe Gautier

Global alignment of intergenic regions



Alignment obtained with clustalW

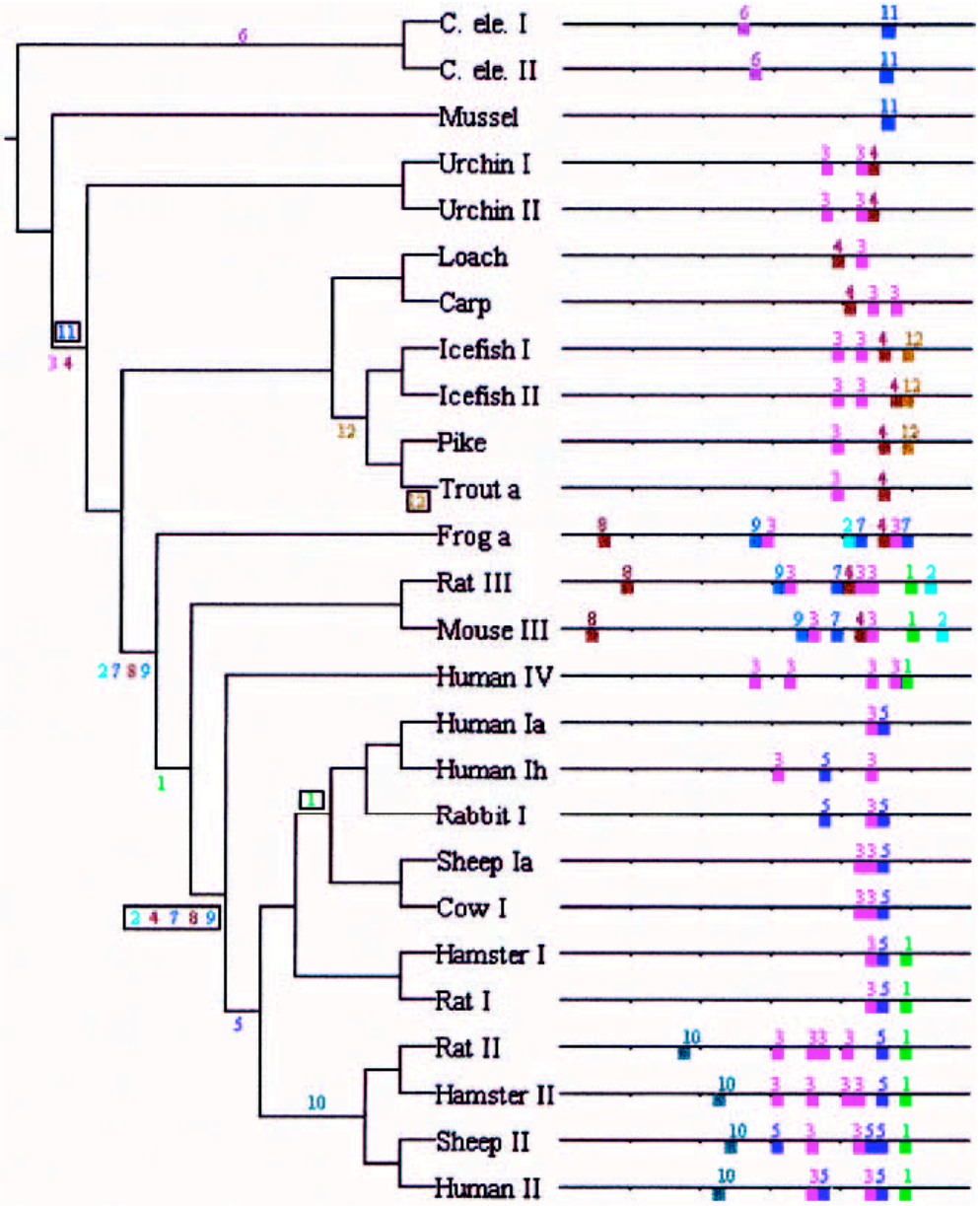
Source: Kellis et al (2003). Nature 423: 241-254

Multiple alignment for another gene in the same genomes

```
GAL80 (YML051W) upstream regions
Scer            ATGGCGCAAGTTTTCCGCTTTGTAATATATATTTATACCCCTTTCTTCTCTCCCCTGCAA
Spar            AGGGGCCAAAGCTCCCGCTCTGTAAAATATATTTATATCCCTTCCTTCTCTCCCCTGCAA
Smik            TAGGGACAAAGCCCGCCTTTTGTAATATATACTTATACCCTCTCCTTCTCTCCCCTGCAA
Sbay            ............................................................
                  **  ***      *  * ***** ***** ***** **  * ****************

Scer            TATAATAGTTTAATTCTAATATTAATAATA---TCCTATATTTTCTTCATTTACCGGCGC
Spar            TATAATAGTTTAATTCTAATATTAATAATA---TCCTATATTTTCCTTACC-ACCGGCGC
Smik            CATAATAGTTAACTCCTAATATTAATAATAATATCCTACAATTTCCTTAGC-ACCGGGGC
Sbay            ............................................................
                 ********* * * ***************   ***** * **** * *   ***** **

Scer            ACTCTCGCCCGAACGACCTCAAAATGTCTGCTACATTCATAATAACCAAAAGCTCATAAC
Spar            ACTCTCGCCCGAACGACCTCAAAATGCTTGCTACATTCATAATAATCAAAAGCTTATAAC
Smik            ACTCTCGCCCGAACGACCTCAAAACGCTTGCTACATCCATAATATTCAGAACTACATCAC
Sbay            ............................................................
                ************************ *  ******** *******  ** **    ** **

Scer            TTTTTTTTT----TGAACCTGAATATATATACATCACATATCACTGCTGGTCCTTGCCGA
Spar            TTTTTTTTTCCTTTGTACCTGAATATATATACATCTCATGTCACTGCTGGTCCTTGCCGG
Smik            TTTTTTTTT-----GTACATAAAAATATATAC--CACATGTCACTGCTGATCCTTGCTGA
Sbay            ............................................................
                *********     * ** * ** ********  * *** ********* ******* *

Scer            CCAGCGTATACAATCTCGATAGTTGGTTT-C-CCGT
Spar            CCAGCGTATACAACCTCGATAGCTGGTTTTC-CCGT
Smik            CGAGCGTATACAAGCTCGATAGCTGGTCTTTACCGT
Sbay            ............................................................
                * ********** ******** **** *   **** *    * ** * ***********
```

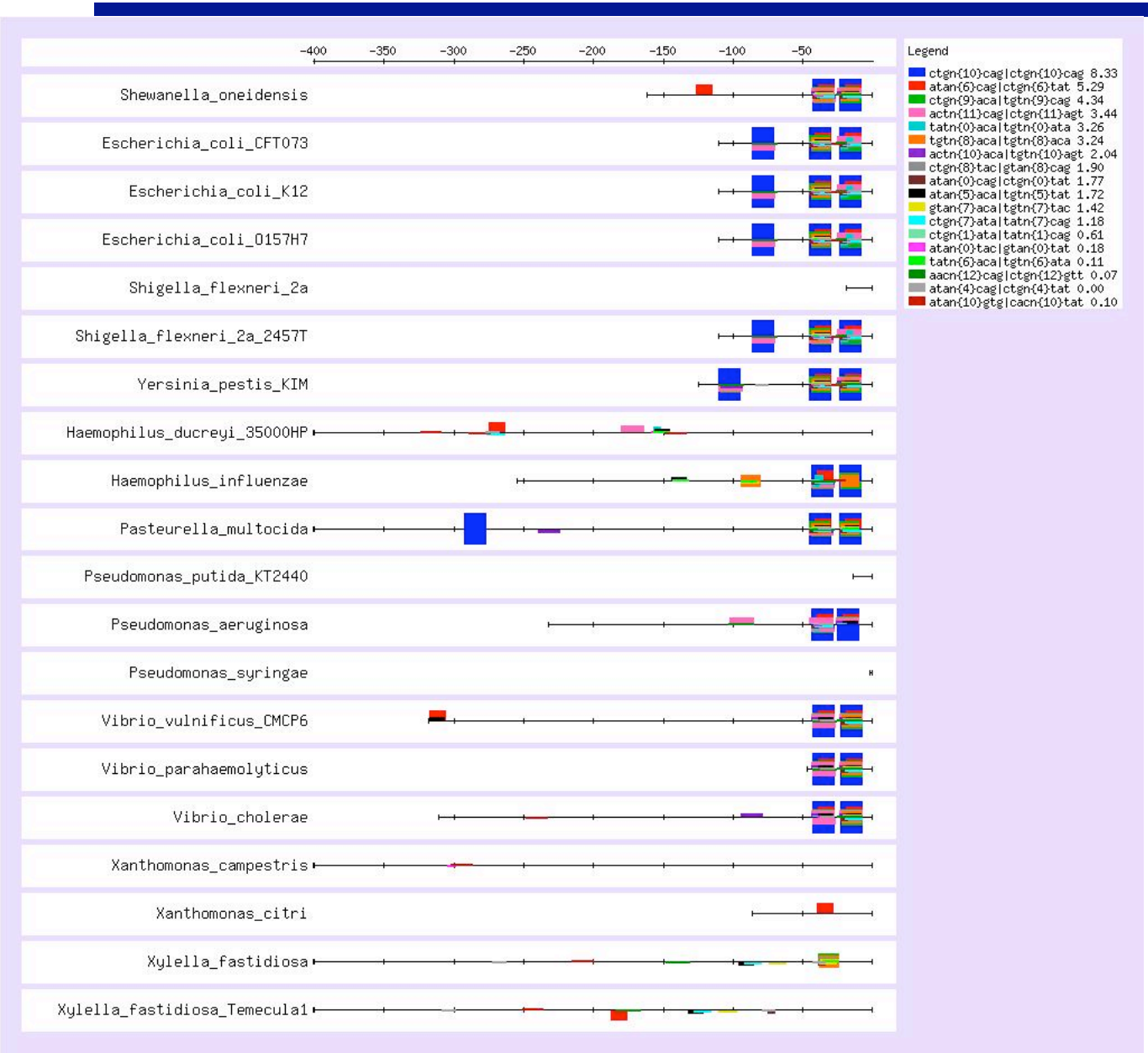
where are the footprints ?
where is the GAL4 binding site ?

Footprinter example metallothionein



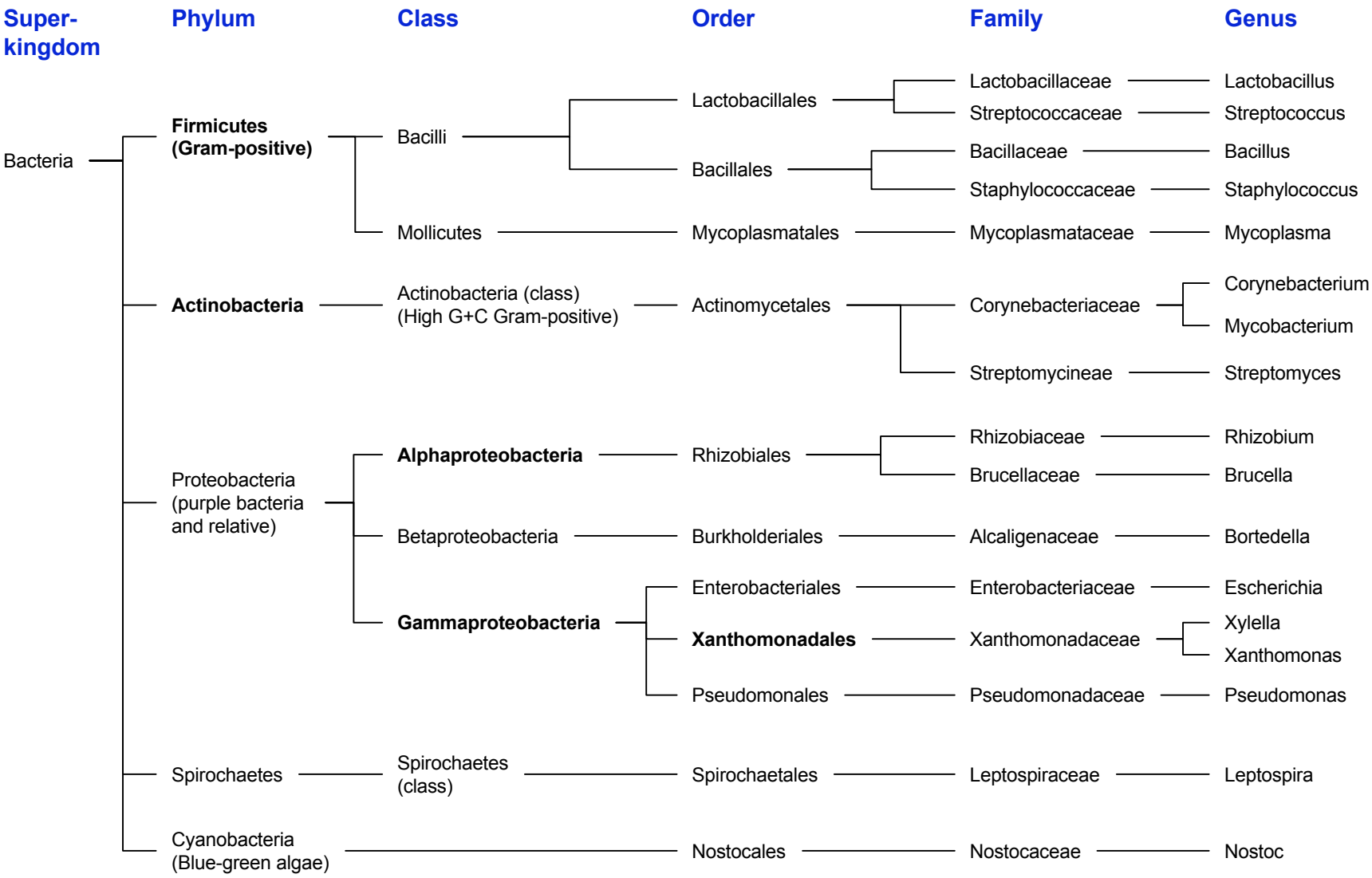
- 590 bp upstream of the same gene (methallothionein) in different species.
- 12 highly conserved motifs are detected.
- Each motif can be associated to a given internal node of the phylogenetic tree.
- Source: Blanchette and Tompa (2002). Genome Research. 12, 739–748.

Pattern discovery in upstream regions of COGs

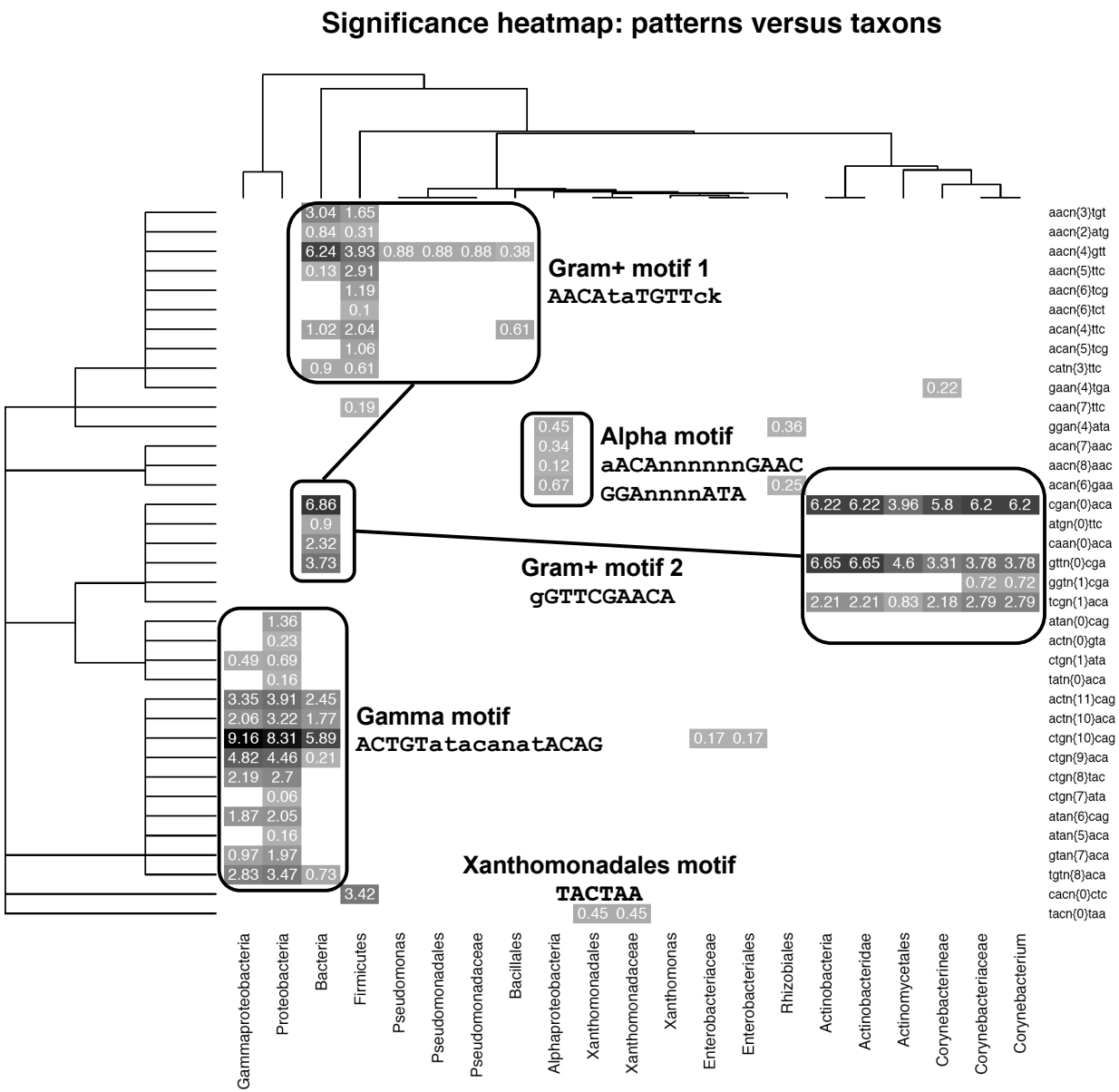


- Sequences
  - Upstream sequences of the cluster of orthologs for the gene lexA in Gamma-bacteria
- Pattern discovery
  - dyad-analysis
- A very highly significant signal is detected. It is found in two conserved positions in most genes of the group, and in 3 positions for certain bacteria.
- This motif corresponds to the known lexA binding site

Bacteria taxonomy

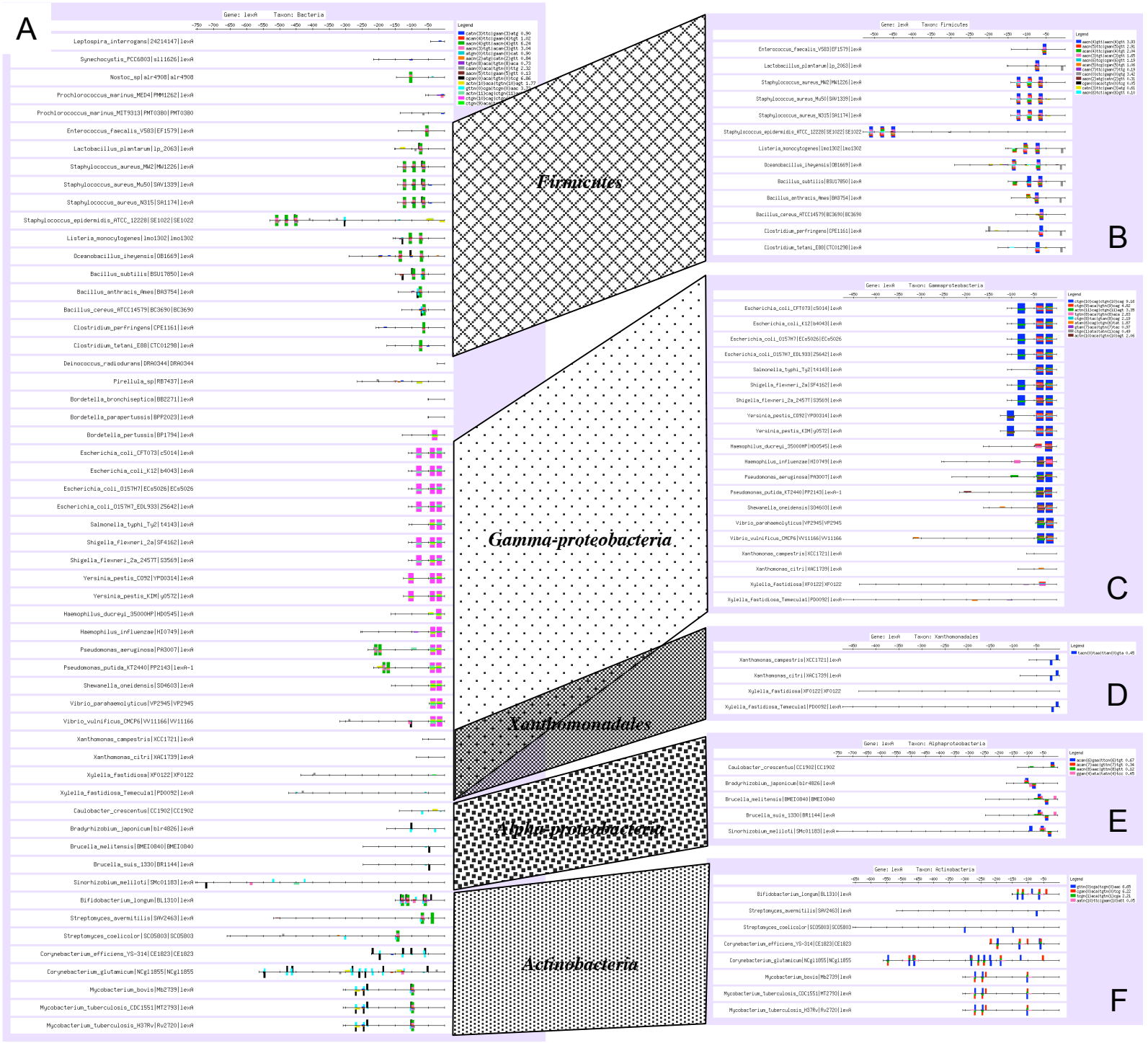


Heat map - pattern sinificance per taxon



- The color in each cell indicates the significance of one pattern (row) in one taxon (column)
  - Rows are clustered by similarity between patterns.
  - Taxons are clustered by similarity between profiles of pattern significance.
- Several coherent groups of patterns are detected

Janky, R. & van Helden, J., in prep.


A
Gene: lexA Taxon: Bacteria
Leptospira_interrogans|24214147|lexA
Synechocystis_PCC6803|sll1626|lexA
Nostoc_sp|alr4908|alr4908
Prochlorococcus_marinus_MED4|PMM1262|lexA
Prochlorococcus_marinus_MIT9313|PMT0380|PMT0380
Enterococcus_faecalis_V583|EF1579|lexA
Lactobacillus_plantarum|lp_2063|lexA
Staphylococcus_aureus_MW2|MW1226|lexA
Staphylococcus_aureus_Mu50|SAV1339|lexA
Staphylococcus_aureus_N315|SA1174|lexA
Staphylococcus_epidermidis_ATCC_12228|SE1022|SE1022
Listeria_monocytogenes|lmo1302|lmo1302
Oceanobacillus_iheyensis|OB1669|lexA
Bacillus_subtilis|BSU17850|lexA
Bacillus_anthracis_Ames|BA3754|lexA
Bacillus_cereus_ATCC14579|BC3690|BC3690
Clostridium_perfringens|CPE1161|lexA
Clostridium_tetani_E88|CTC01298|lexA
Deinococcus_radiodurans|DRA0344|DRA0344
Pirellula_sp|RB7437|lexA
Bordetella_bronchiseptica|BB2271|lexA
Bordetella_parapertussis|BPP2023|lexA
Bordetella_pertussis|BP1794|lexA
Escherichia_coli_CFT073|c5014|lexA
Escherichia_coli_K12|b4043|lexA
Escherichia_coli_O157H7|ECs5026|ECs5026
Escherichia_coli_O157H7_EDL933|Z5642|lexA
Salmonella_typhi_Ty2|t4143|lexA
Shigella_flexneri_2a|SF4162|lexA
Shigella_flexneri_2a_2457T|S3569|lexA
Yersinia_pestis_CO92|YPO0314|lexA
Yersinia_pestis_KIM|y0572|lexA
Haemophilus_ducreyi_35000HP|HD0545|lexA
Haemophilus_influenzae|HI0749|lexA
Pseudomonas_aeruginosa|PA3007|lexA
Pseudomonas_putida_KT2440|PP2143|lexA-1
Shewanella_oneidensis|SO4603|lexA
Vibrio_parahaemolyticus|VP2945|VP2945
Vibrio_vulnificus_CMCP6|VV11166|VV11166
Xanthomonas_campestris|XCC1721|lexA
Xanthomonas_citri|XAC1739|lexA
Xylella_fastidiosa|XF0122|XF0122
Xylella_fastidiosa_Temecula1|PD0092|lexA
Caulobacter_crescentus|CC1902|CC1902
Bradyrhizobium_japonicum|blr4826|lexA
Brucella_melitensis|BMEI0840|BMEI0840
Brucella_suis_1330|BR1144|lexA
Sinorhizobium_meliloti|SMc01183|lexA
Bifidobacterium_longum|BL1310|lexA
Streptomyces_avermitilis|SAV2463|lexA
Streptomyces_coelicolor|SCO5803|SCO5803
Corynebacterium_efficiens_YS-314|CE1823|CE1823
Corynebacterium_glutamicum|NCgl1855|NCgl1855
Mycobacterium_bovis|Mb2739|lexA
Mycobacterium_tuberculosis_CDC1551|MT2793|lexA
Mycobacterium_tuberculosis_H37Rv|Rv2720|lexA
Firmicutes
Gamma-proteobacteria
Xanthomonadales
Alpha-proteobacteria
Actinobacteria
B
Gene: lexA Taxon: Firmicutes
C
Gene: lexA Taxon: Gammaproteobacteria
D
Gene: lexA Taxon: Xanthomonadales
E
Gene: lexA Taxon: Alphaproteobacteria
F
Gene: lexA Taxon: Actinobacteria

Summary - phylogenetic approaches

- Matching conserved sites for known transcription factors
  - Consite: http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/
  - Lenhard et al. (2003). J.Biology 2:13.
- Global alignment of promoters of orthologous genes
  - clustalW
  - e.g.: Kellis et al (2003). Nature 423: 241-254.
- Pattern discovery in promoters of orthologous genes
  - Footprinter: http://bio.cs.washington.edu/software.html
  - Blanchette and Tompa (2002). Genome Research. 12, 739–748.

The Analysis of Regulatory Sequences

Classification of upstream sequences on the basis of regulatory signals

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Data - pattern counts

- 94 sequences
  - NIT (31 upstream sequences); PHO (13 upstream sequences); MET (20 upstream sequences); RAND (30 random sequences Markov 5)
- 44 patterns
  - Hexanucleotides and dyads involved in the regulation of the MET, PHO and NIT genes.
  - Some of these patterns are very well conserved in the core of the binding site (e.g. CACGTG, CACGTT, ...) whereas some other represent partial conservation of the flanking nucleotides (e.g. ACGTGg, ACGTTt, ...).
  - The data is presented in a multi-variate table, with one row per gene, and one column per pattern.

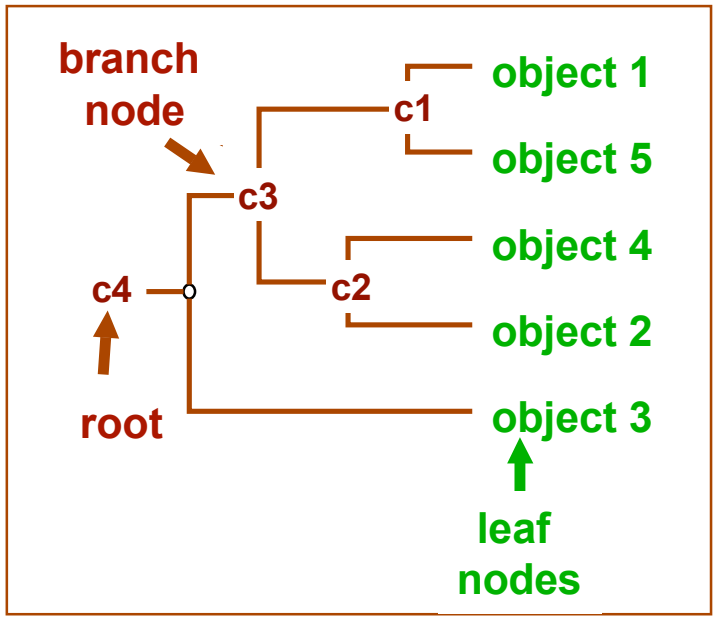
Gene	aaacgt|acgttt	aacgtg|cacgtt	aacn{1}gtg|cacn{1}gtt	aactgt|acagtt	acan{14}tgc|gcan{14}tgt	acan{15}gca|tgcn{15}tgt	acan{6}gca|tgcn{6}tgt	acatct|agatgt	acgn{1}gcg|cgcn{1}cgt	acgn{6}agc|gctn{6}cgt	acgtga|tcacgt	acgtgc|gcacgt	acgtgg|ccacgt	actgtg|cacagt	agataa|ttatct	ataaga|tcttat	atcacg|cgtgat	cacgcc|ggcgtg	cacgtg|cacgtg	cacn{15}ggc|gccn{15}gtg	cacn{1}tga|tcan{1}gtg	cacn{2}gac|gtcn{2}gtg	cagn{2}cgg|ccgn{2}ctg	cagn{7}atc|gatn{7}ctg	ccacag|ctgtgg	cccacg|cgtggg	cccatc|gatggg	ccgcgc|gcgcgg
GDH3	0	0	0	4	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	2	0	0	0	0	0	0
YBR043C	2	0	0	0	0	0	0	2	0	2	0	0	0	2	4	2	0	2	0	0	0	0	0	4	0	0	2	0
APG14	0	0	0	4	0	0	0	0	0	0	0	0	0	0	4	2	0	0	0	0	0	2	0	0	0	0	0	0
AGP1	0	0	2	2	0	0	0	0	0	0	0	0	2	2	4	2	2	0	0	0	0	0	0	2	2	4	0	0
CHA1	0	0	2	2	0	0	4	0	0	2	0	0	0	0	6	6	2	0	0	0	0	0	0	2	2	0	0	2
UGA4	4	0	0	0	0	0	0	2	0	0	0	2	0	2	0	6	0	0	0	0	0	0	0	0	0	0	0	0
PRB1	0	0	2	0	0	0	0	0	0	0	0	0	0	2	4	4	2	2	0	0	0	4	0	0	0	0	0	0
CAN1	0	2	0	0	0	0	2	0	0	4	0	0	0	0	6	2	0	0	0	0	0	0	0	0	2	0	0	0
GAT1	0	0	0	0	2	0	0	0	0	2	2	0	2	0	8	6	0	0	0	0	0	2	2	0	2	0	0	0
UGA1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0	2	0	0	0	2	4	0	0	0
MEP1	2	0	0	0	0	2	0	6	0	0	0	0	0	0	6	14	0	0	0	0	2	0	4	4	2	0	0	0

Hierarchical clustering

Distance matrix

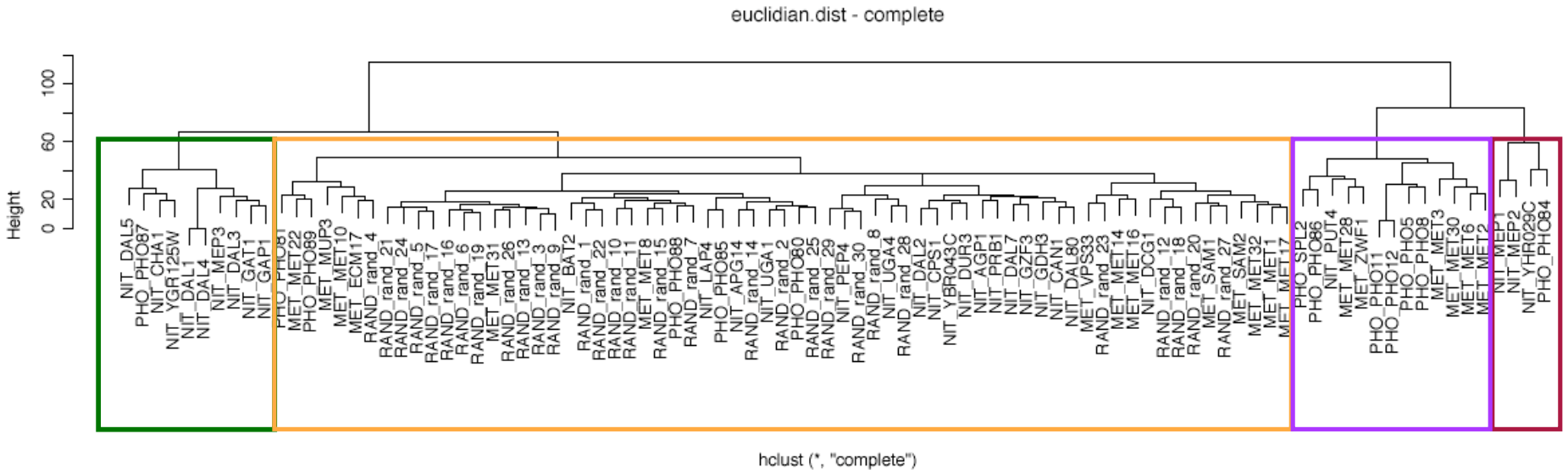
	object 1	object 2	object 3	object 4	object 5
object 1	0.00	4.00	6.00	3.50	1.00
object 2	4.00	0.00	6.00	2.00	4.50
object 3	6.00	6.00	0.00	5.50	6.50
object 4	3.50	2.00	5.50	0.00	4.00
object 5	1.00	4.50	6.50	4.00	0.00

Tree representation



- Hierarchical clustering is an aggregative clustering method
- One needs to define a (dis)similarity metric between two groups. There are several possibilities
  - **Average linkage**: the average distance between objects from groups A and B
  - **Single linkage**: the distance between the closest objects from groups A and B
  - **Complete linkage**: the distance between the most distant objects from groups A and B
  - **Ward clustering**: the dissimilarity between two groups is estimated by the moment of inertia of their elements from the gravity center.
- Algorithm
  - (1) Assign each object to a separate cluster.
  - (2) Find the pair of clusters with the shortest distance, and regroup them in a single cluster
  - (3) Repeat (2) until there is a single cluster
- The result is a tree, whose intermediate nodes represent clusters
  - `N objects → N-1 intermediate nodes
- Branch lengths represent distances between clusters

Clustering - Euclidian distance



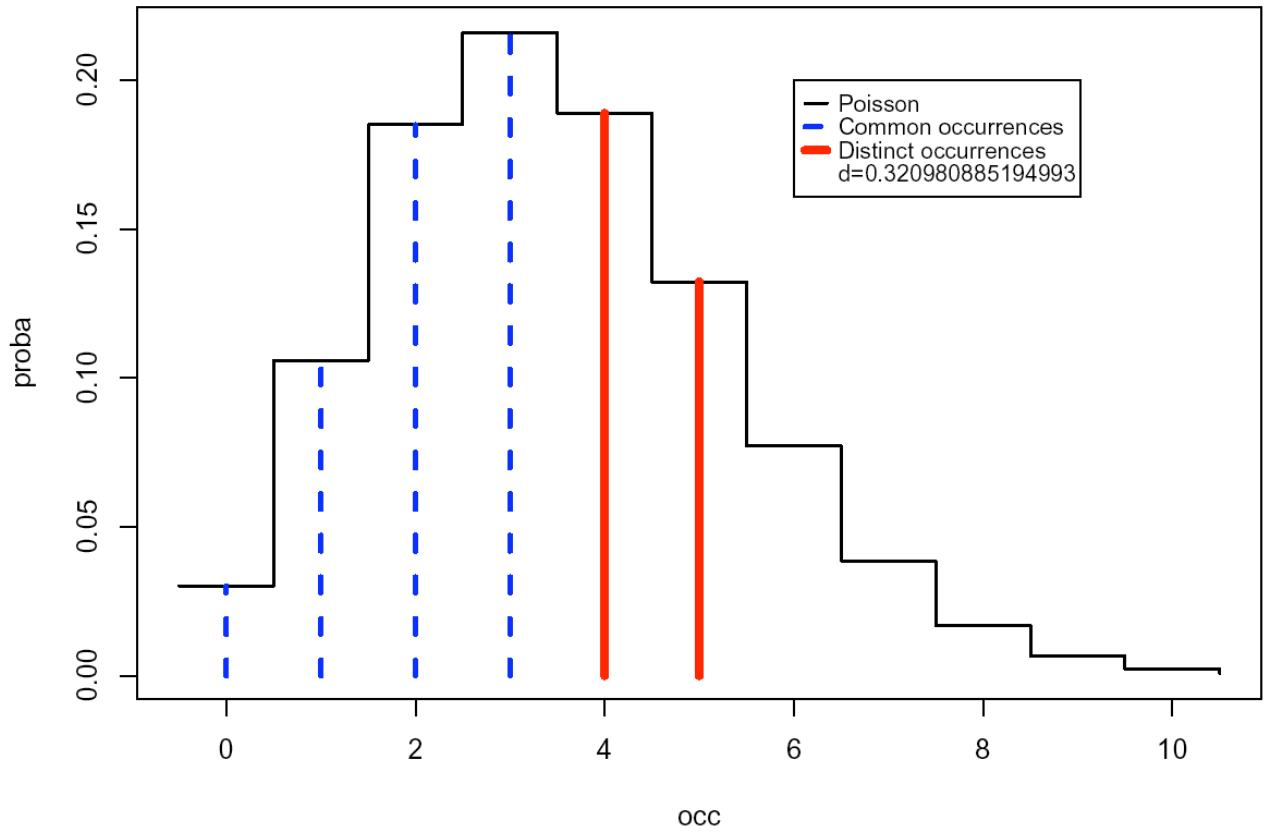
- Sequence clustering on the basis of pattern counts
- Distance metric: Euclidian
- Clustering method: UPGMA (complete)
- The four main clusters do not correspond to the prior functional classes
- Genes from different classes are intermingled

Predicted \ Known	RAND	MET	NIT	PHO	SUM
RAND	30	14	18	5	67
MET	0	6	1	6	13
NIT	0	0	9	1	10
PHO	0	0	3	1	4
SUM	30	20	31	13	94
Errors	48	51.1%			
Correct	46	48.9%			

van Helden, J. 2004. Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20: 399-406.8

Poisson-based similarity and dissimilarity

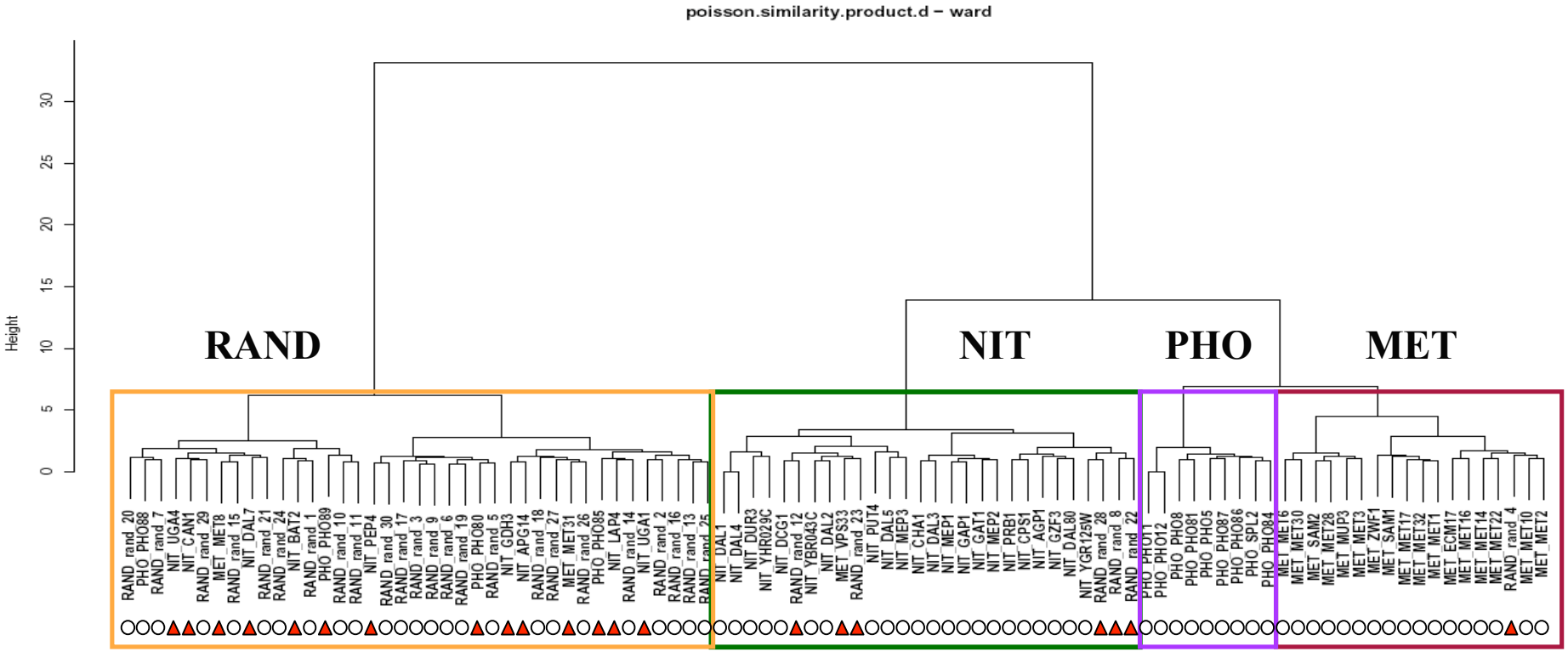
- Let us take a simple example:
  - Sequence *a* contains 3 occurrences of a motif
  - Sequence *b* contains 5 occurrences of the same motif
- We have thus 3 common and 2 distinct occurrences.



Clustering - Poisson-based distance metrics

- Metric: Poisson similarity product;
- Clustering: Ward hierarchical.
- Red triangles below the tree indicate errors
- Most errors consist in false negative.

		Known				
		RAND	MET	NIT	PHO	SUM
Predicted	RAND	24	2	9	4	39
	MET	1	17	0	0	18
	NIT	5	1	22	0	28
	PHO	0	0	0	9	9
	SUM	30	20	31	13	94
	Errors	22	23.40%			
	Correct	72	76.60%			



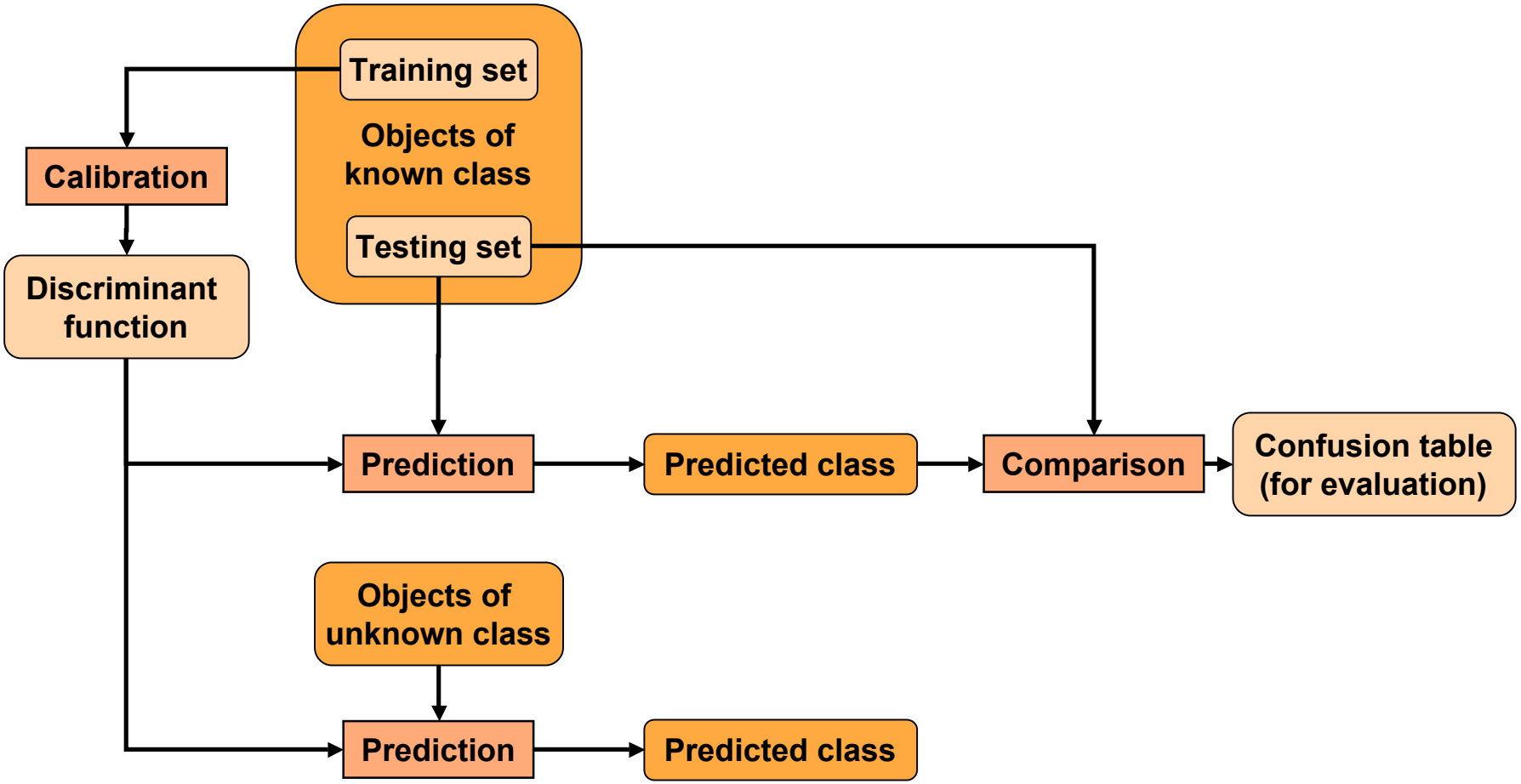
van Helden, J. 2004. Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20: 399-406.8

Results with MET-PHO-NIT genes + random sequences

metric	model	clustering method	MET >NIT	NIT >NIT	PHO >NIT	RAND >NIT	MET >MET	NIT >MET	PHO >MET	RAND >MET	MET >PHO	NIT >PHO	PHO >PHO	RAND >PHO	MET>RAND	NIT > RAND	PHO > RAND	RAND > RAND	EXTERNAL	TRUE	FALSE	hit rate
poisson.similarity.product.d	product	ward	1	22	0	5	17	0	0	1	0	0	9	0	2	9	4	24		72	22	76.6%
poisson.mixed.distinct.d	additive	complete	0	25	0	5	10	0	0	0	0	0	8	0	10	6	5	25		68	26	72.3%
poisson.mixed.over.d	additive	ward	3	21	1	7	15	1	1	1	0	0	8	0	2	9	3	22		66	28	70.2%
poisson.mixed.distinct.d	additive	ward	0	19	0	5	16	1	2	3	0	0	8	0	4	11	3	22		65	29	69.1%
poisson.mixed.distinct.product.d	product	complete	0	15	0	0	12	5	1	1	5	0	8	0	3	11	4	29		64	30	68.1%
poisson.similarity.d	additive	ward	4	25	2	14	15	1	1	0	0	0	8	0	1	5	2	16		64	30	68.1%
poisson.dissimilarity.distinct	additive	ward	4	22	2	12	15	0	1	0	0	0	8	0	1	9	2	18		63	31	67.0%
poisson.mixed.over.product.d	product	ward	3	23	1	9	11	0	0	1	3	0	9	2	3	8	3	18		61	33	64.9%
correlation.coefficient.d	additive	ward	1	26	2	8	10	1	1	2	1	1	8	6	8	3	2	14		58	36	61.7%
poisson.dissimilarity.over	additive	ward	9	16	1	4	8	0	1	0	0	0	8	0	3	15	3	26		58	36	61.7%
poisson.similarity.d	additive	complete	13	25	2	13	6	0	0	0	0	0	8	0	1	6	3	17		56	38	59.6%
correlation.coefficient.d	additive	complete	0	26	3	11	9	1	0	2	1	1	8	5	10	3	2	12		55	39	58.5%
poisson.similarity.product.d	product	complete	0	12	0	0	12	10	1	9	6	0	9	0	2	9	3	21		54	40	57.4%
manhattan.dist	additive	ward	0	10	0	0	10	13	3	4	8	0	7	0	2	8	3	26		53	41	56.4%
park.similarity.d (pruning 5)	additive	ward	4	14	2	9	11	1	0	2	0	0	8	0	1	15	2	19	6	52	42	55.3%
poisson.dissimilarity.over	additive	complete	5	31	4	29	13	0	1	1	0	0	8	0	2	0	0	0		52	42	55.3%
poisson.mixed.distinct.product.d	product	ward	0	20	0	1	14	0	9	0	5	7	2	14	1	4	2	15		51	43	54.3%
poisson.dissimilarity.distinct	additive	complete	5	31	4	30	11	0	1	0	0	0	8	0	4	0	0	0		50	44	53.2%
poisson.mixed.over.d	additive	complete	3	28	3	26	13	1	1	1	2	2	9	3	2	0	0	0		50	44	53.2%
park.similarity.d	additive	ward	5	29	4	28	11	1	0	2	0	0	8	0	4	1	1	0		48	46	51.1%
euclidian.dist	additive	ward	0	6	0	0	12	16	3	8	6	4	7	0	2	5	3	22		47	47	50.0%
mahalanobis.dist	additive	complete	6	16	5	12	9	7	0	1	5	7	7	2	0	1	1	15		47	47	50.0%
euclidian.dist	additive	complete	0	9	1	0	6	1	6	0	0	3	1	0	14	18	5	30		46	48	48.9%
mahalanobis.dist	additive	ward	6	15	5	9	11	11	2	5	3	4	5	1	0	1	1	15		46	48	48.9%
park.similarity.d	additive	complete	15	30	4	30	5	1	1	0	0	0	8	0	0	0	1	0		43	52	45.3%
poisson.mixed.over.product.d	product	complete	5	27	9	24	2	4	1	6	0	0	2	0	0	0	1	0		31	50	38.3%
manhattan.dist	additive	complete	12	22	6	30	6	9	1	0	0	0	6	0	2	0	0	0		34	60	36.2%

van Helden, J. 2004. Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20: 399-406.8

Discriminant analysis



Optimal conditions

- Pattern detection: 3 top scores for 5 position-weight matrices
- Linear Discriminant Analysis
- Forward selection procedure
- Externa d

3 group classification

Predicted \ Known	PHO	MET	CTL	SUM
PHO	7	0	0	7
MET	1	12	1	14
CTL	0	4	79	83
SUM	8	16	80	104
Errors	6	5.77%		
Correct	98	94.23%		

PHO against others

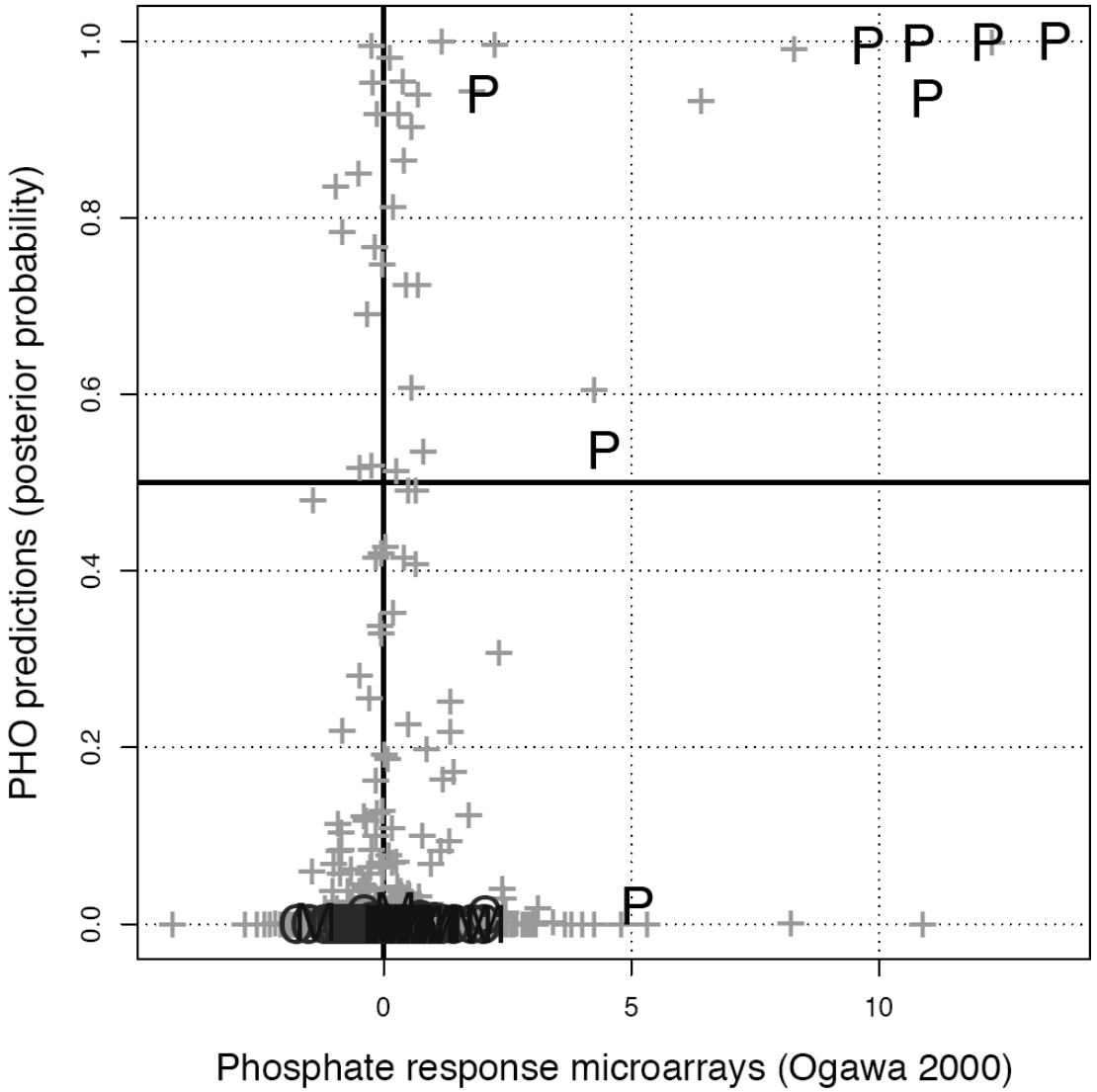
Predicted \ Known	PHO	CTL	SUM
PHO	7	0	7
CTL	1	96	97
SUM	8	96	104
Errors	1	0.96%	
Correct	103	99.04%	

MET against others

Predicted \ Known	MET	CTL	SUM
MET	13	0	13
CTL	3	88	91
SUM	16	88	104
Errors	3	2.88%	
Correct	101	97.12%	

Gonze, D. et al.. 2005. Discrimination of yeast genes involved in methionine and phosphate metabolism on the basis of upstream motifs. *Bioinformatics* **21:** 3490-3500.

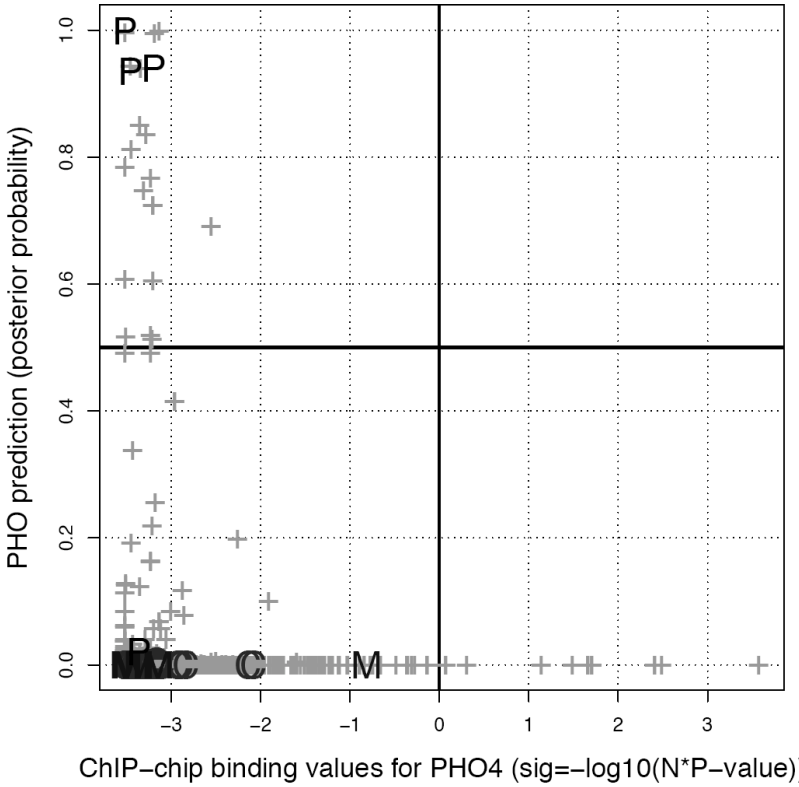
PHO predictions versus microarray data

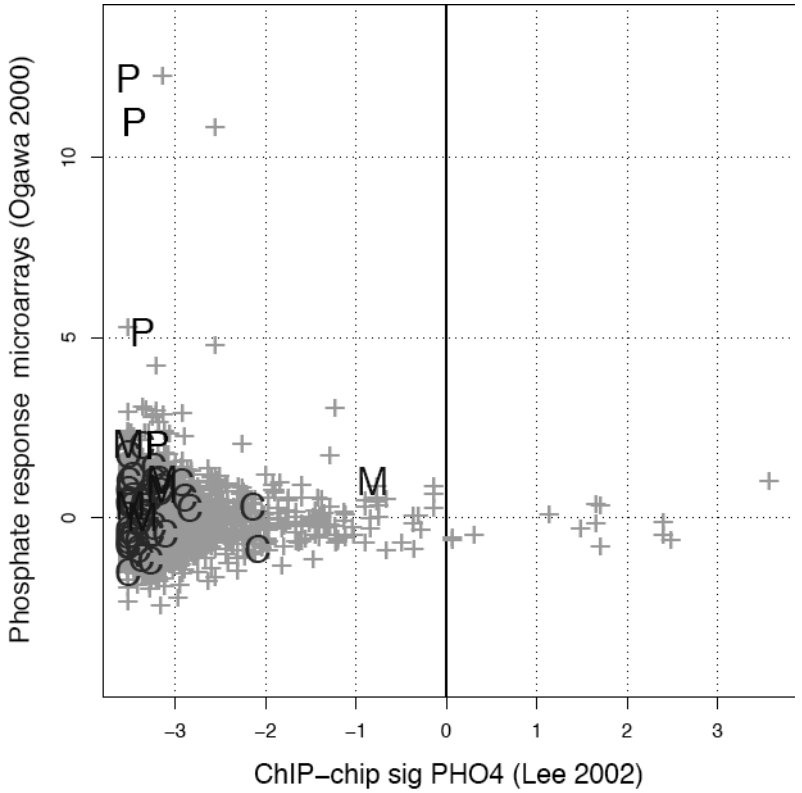


- PHO predictions include most (but not all) of the phosphate-responding genes (microarray data, Ogawa 2000)
- There are many additional predictions which are not detected by microarrays
  - False positives ?
  - Genes responding to different conditions ?

PHO predictions versus Chip-CHIP data (Lee, 2002)

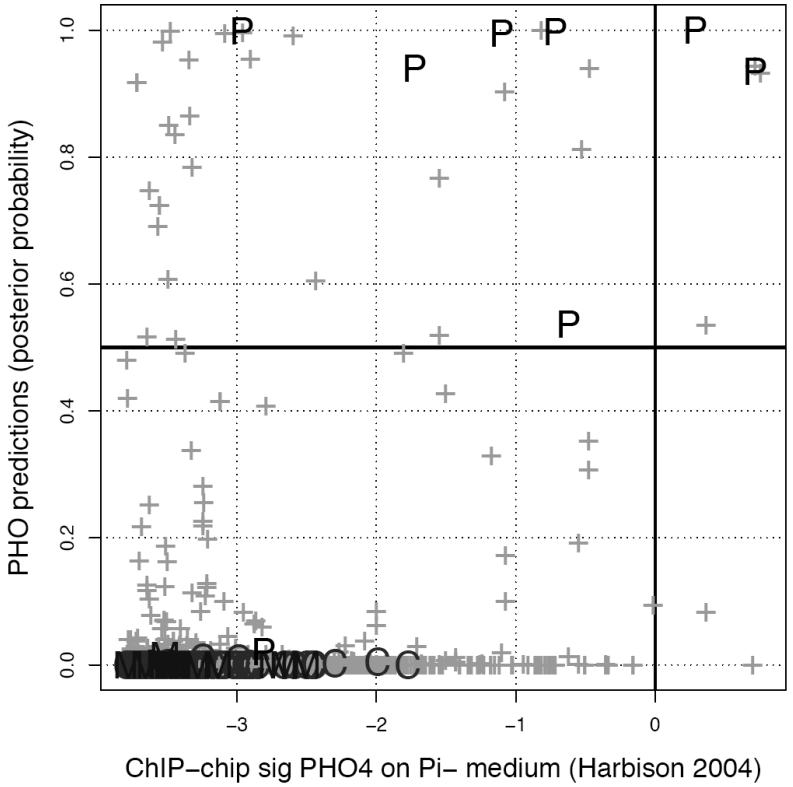
- There is not a single common gene between our PHO predictions and the Pho4p-bound promoters detected with the ChIP-chip technology by Lee et al, 2002)
- However, Lee results for Pho4p fail to detect
  - genes known to be regulated by Pho4p
  - Genes responding to phosphate in Ogawa (2000)
- Problem with the ChIP-chip experiment
  - was performed in rich medium -> ***Pho4p is inactive !!!***

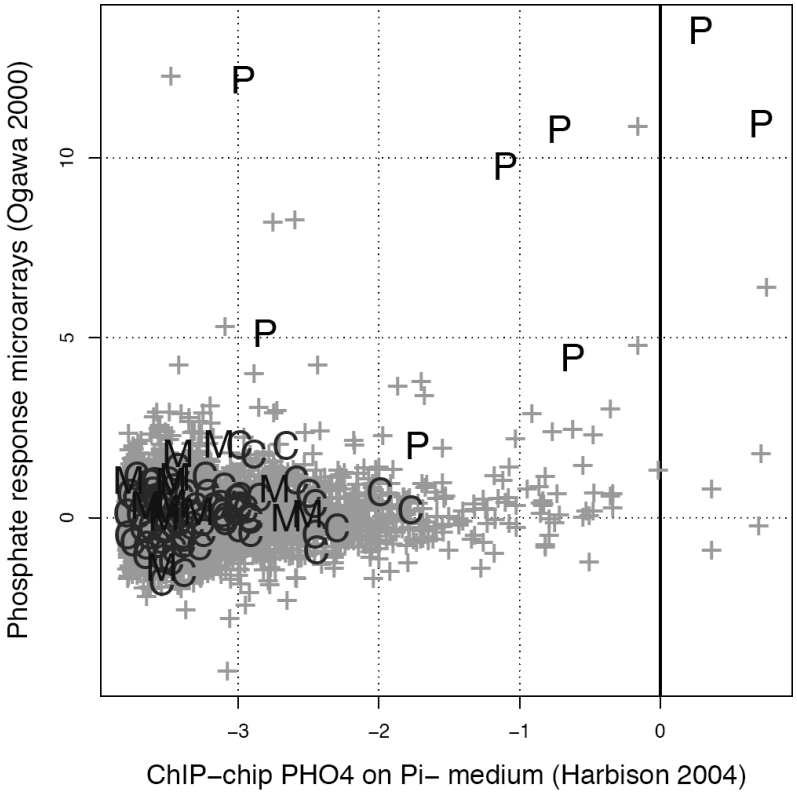




PHO predictions versus Chip-CHIP data (Harbison, 2004)

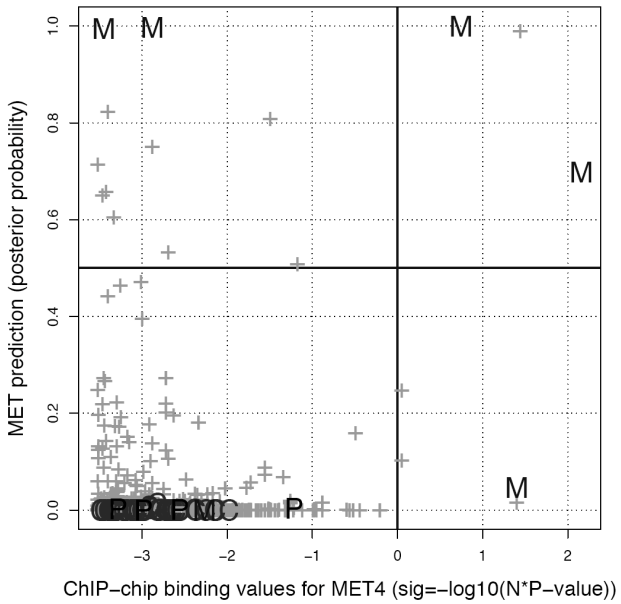
- In 2004, the same group performed new experiments with different environmental conditions (Harbison, 2004)
- There is a slightly better (but far from perfect) correspondence between ChIP-chip results and
  - Our PHO predictions
  - microarray data (Ogawa *et al.*, 2000)
  - Annotated Pho4p target genes (P on the plots)

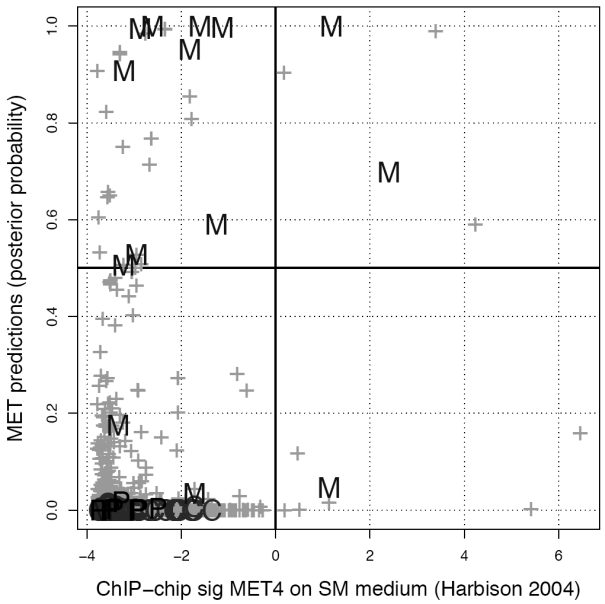


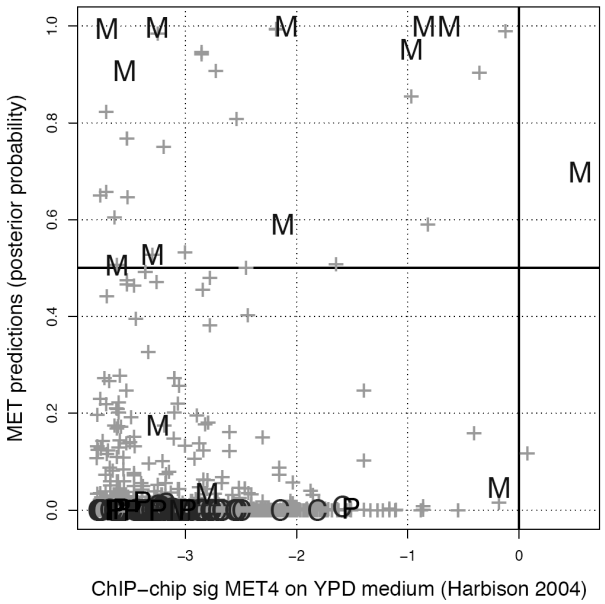


MET predictions versus chip-chip data

- We compared MET predictions with ChIP-chip data
  - Lee (2002): rich medium
  - Harbison (2004): SM medium
  - Harbison (2004): YPD medium
- The correspondences are rather poor
- Even though our predictions contain a rate of false positives, and miss some MET genes, the correspondence with annotated MET is better than for genes detected experimentally with the ChIP-chip method !





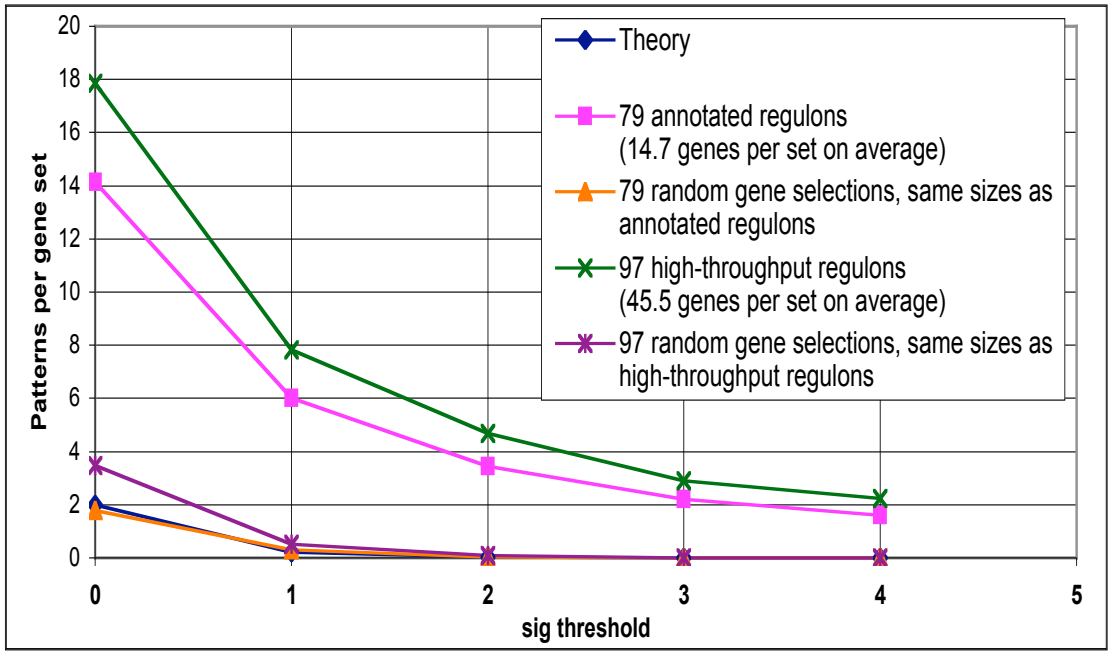


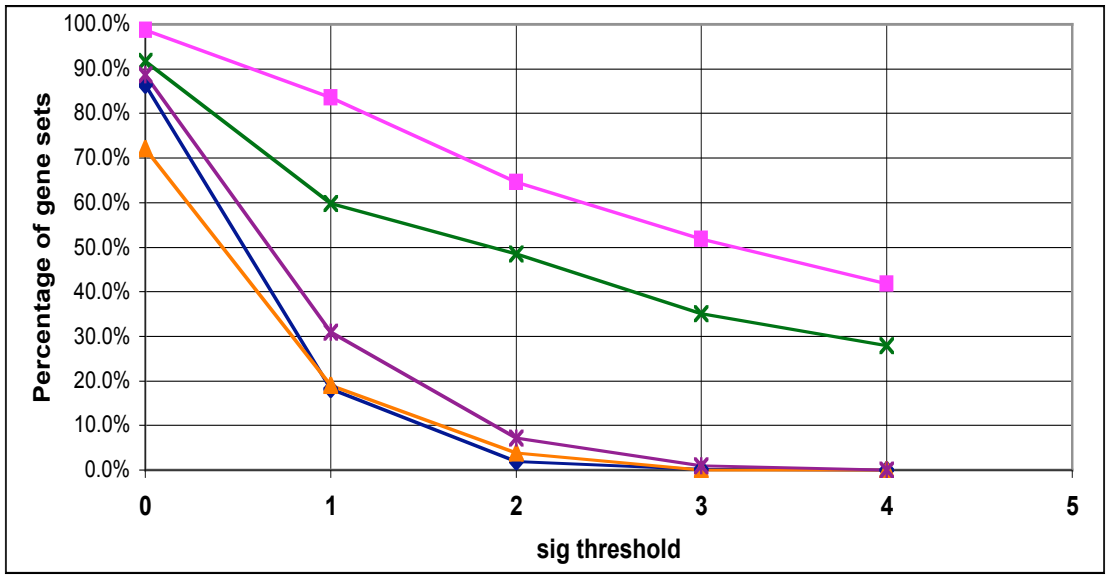
The Analysis of Regulatory Sequences

Assessment of pattern discovery results

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

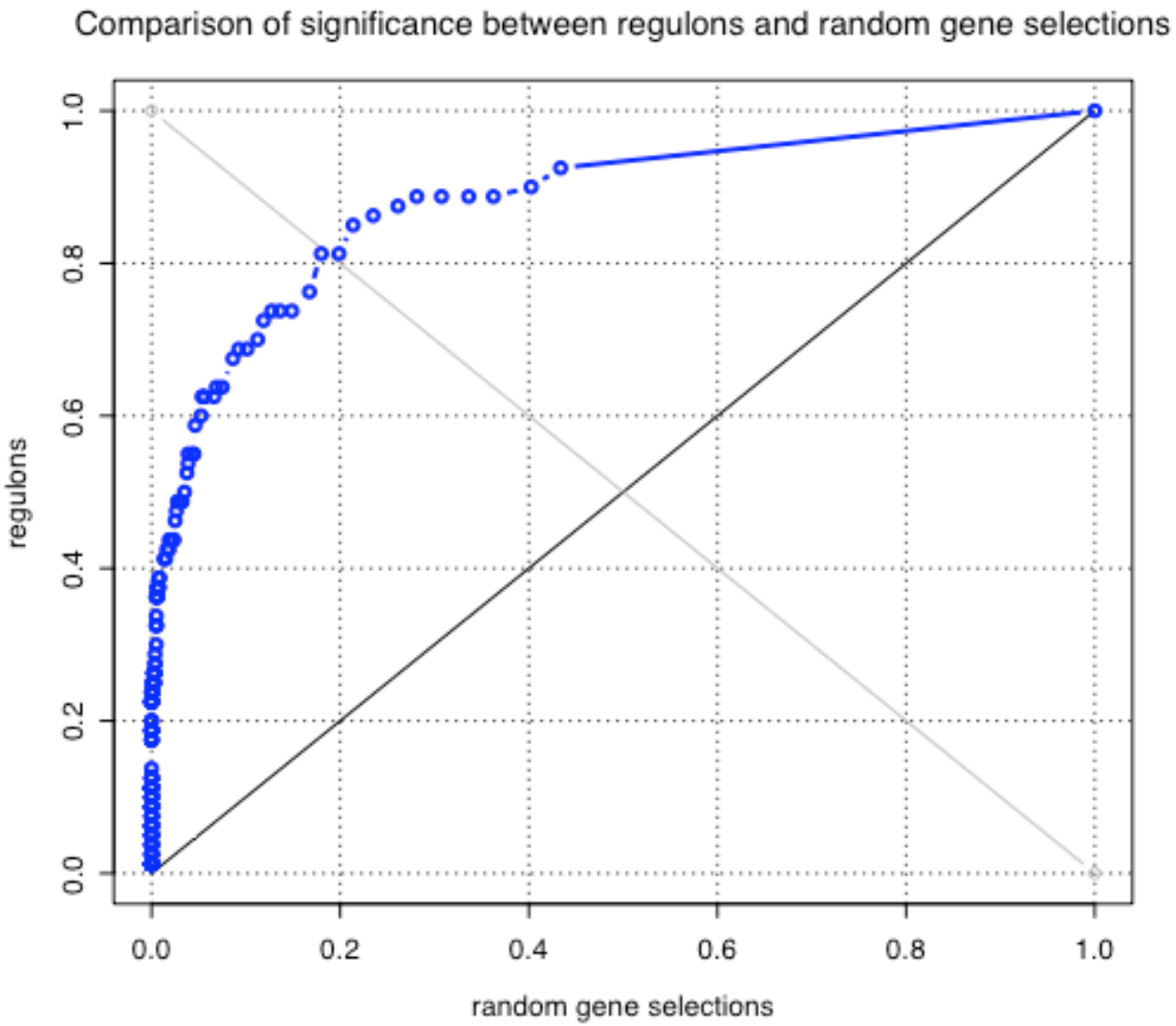
Validation of pattern discovery with yeast regulons





- Regulons were collected from TRANSFAC and aMAZE.
- All the regulons with ≥5 genes were analysed.
  - Significant patterns (sig≥ 2) are detected in 65% of the regulons.
- As a negative control, sets of random genes were analysed.
  - The rate of false positive follows pretty well the statistical expectation.

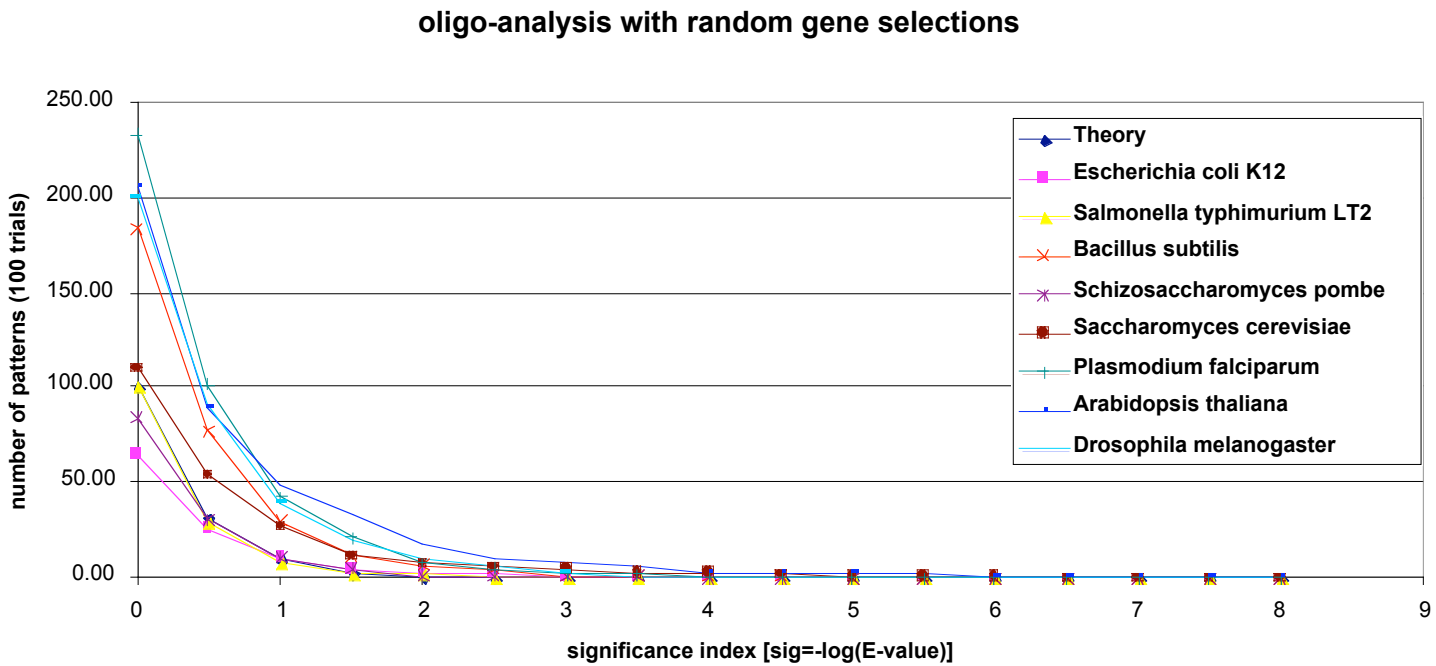
ROC curve representation



- X axis: 1 - specificity
  - Significance in random gene selections
- Y axis: sensitivity
  - Significance in regulons
- The surface below the ROC curve indicates the accuracy of the predictions.

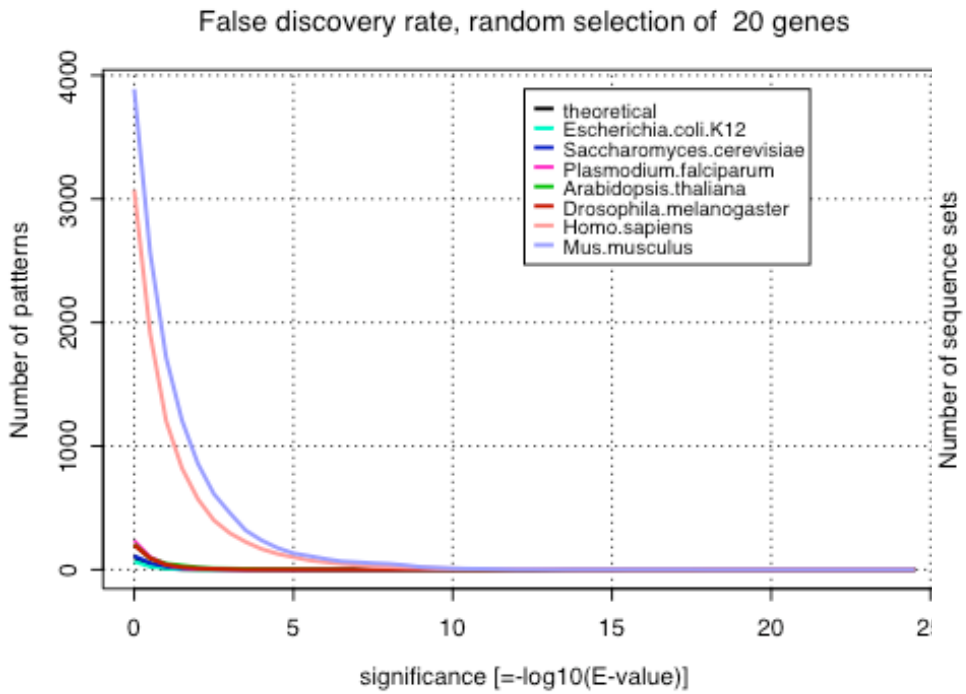
Rate of false positive in different organisms

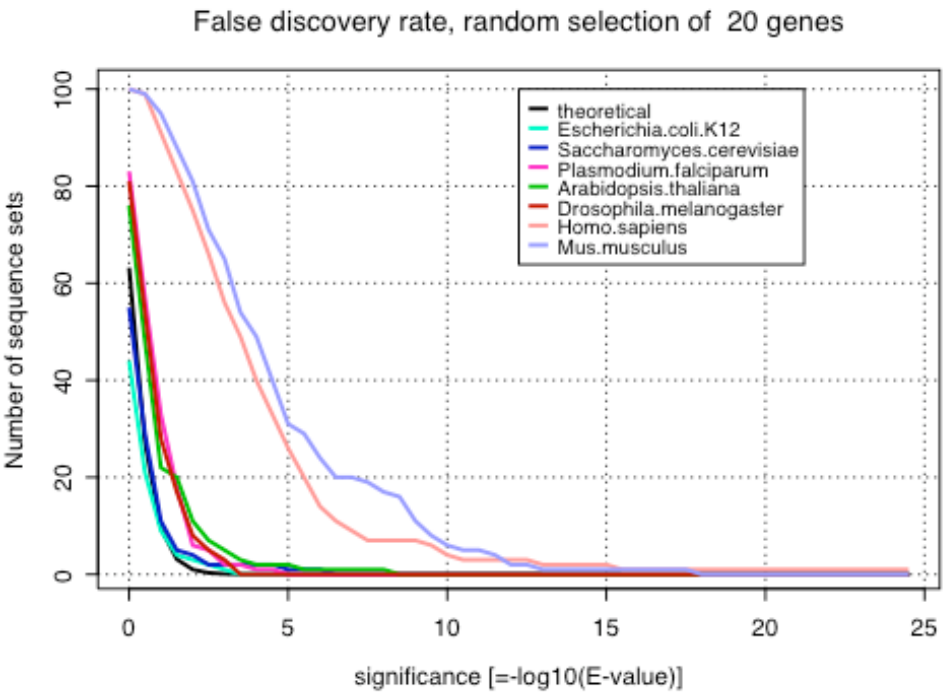
- The analysis of random gene selections allows to evaluate the rate of false positive returned by a pattern discovery program.
- The rate of false positive is good for microbes (bacteria and yeasts), but increases for higher organisms.
- This is likely to result from the higher heterogeneity of genomic sequences in these organisms. We are currently developing more elaborate background models to treat this problem.



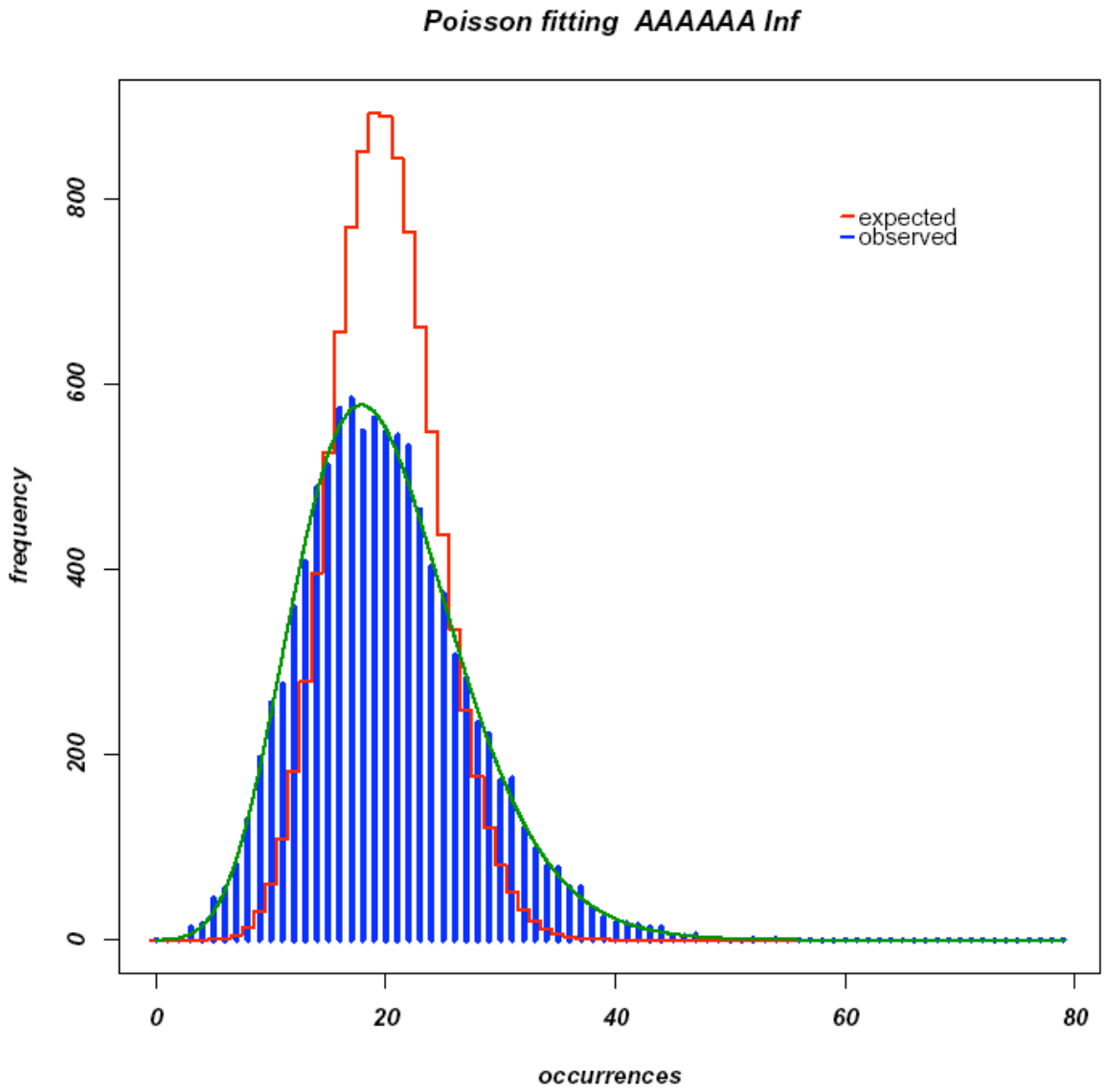
Rate of false positive in higher organisms

- The rate of false positive increases dramatically with higher organisms.
- This is likely to come from
  - a bad treatment of repetitive elements : genome-scale calibration does not account for local frequencies
  - positional heterogeneities : oligonucleotide frequencies depend on the distance from the gene
  - the higher heterogeneity of genomic sequences in these organisms (GC-rich vs AT-rich promoters)
- We are currently developing more elaborate background models to treat this problem.





Fitting a negative binomial on word distributions



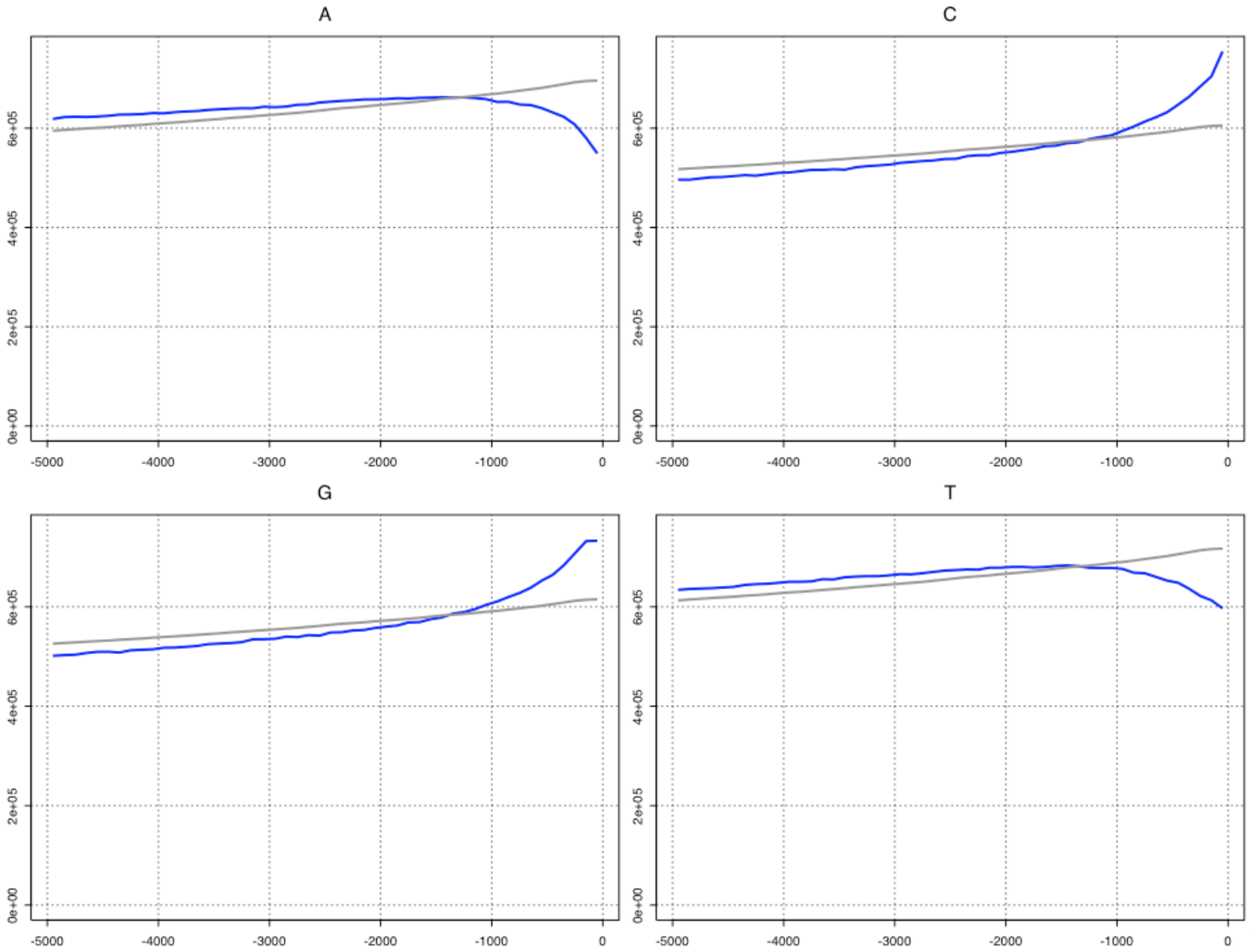
The properties of the negative binomial allow to fit it on an observed distribution where a variance larger than the mean.

Goodness of fit -all upstream regions

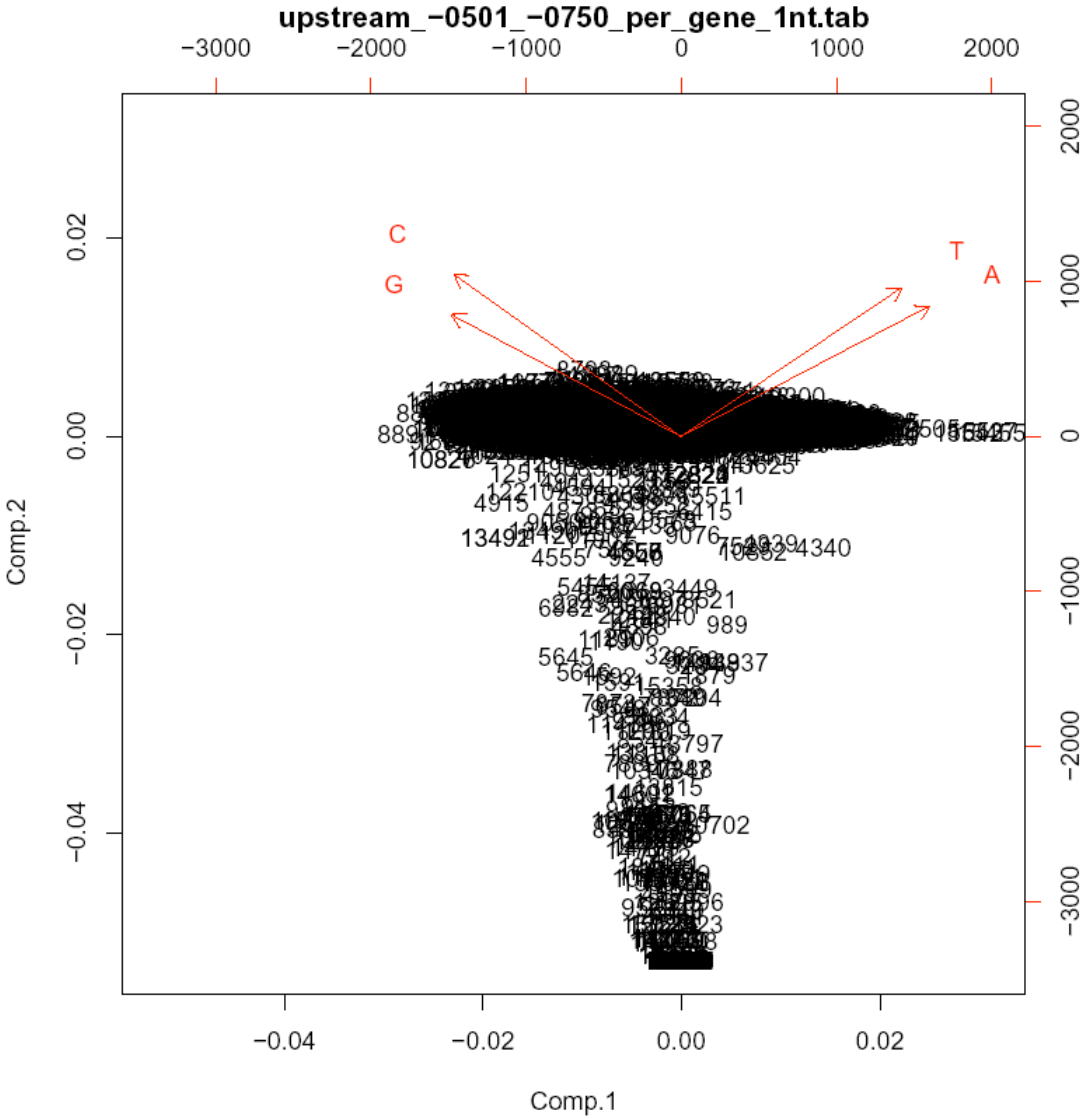
	N		Poisson		negbin	
Organism		**Length**	**fit**	**no fit**	**fit**	**no fit**
E.coli	***4000***	**200**	**4073**	**17**	**4090**	**0**
S.cerevisiae	***6000***	**800**	**3911**	**185**	**4058**	**38**
A.thaliana	***27000***	**1000**	**800**	**3296**	**3746**	**350**
Homo sapiens	***30000***	**1000**	**797**	**3299**	**3233**	**863**

- We tested the goodness of fit
  - for each hexanucleotide,
  - On the whole set of promoters of different organisms
  - with the Poisson and the negative binomial distributions, respectively.

Position analysis - single nucleotides



PCA - nucleotide composition -750 to -500



Acknowledgements

SCMBB - Bruxelles - Belgium

Analysis of Biochemical, Regulation and Interaction Networks

Olivier Sand (postdoc)
Rekin's Janky (DEA 2002 + PhD)
Sylvain Brohée (PhD)
Jean-Valéry Turatsinze (graduate 2005 - PhD)
Dimitrios Paraskevas (DEA 2006)
Carlos Moreira Conde da Silva (graduate 2006)
Kevin Wielemans (graduate 2006)

Previous researchers

Shoshana Wodak (lab director)
Nicolas Simonis (PhD 2005)
Didier Croes (PhD 2005)
Didier Gonze (DEA 2001 + postdoc 2003)
Joseph Tran (DEA 2002 + PhD)

Previous students

Ahmed Essaghir (DEA 2005)
Raymond Kalimunda (DEA 2005)
Mehdi Jbel (DEA 2005)
Pierre Jonniaux (DEA 2004)
Fabian Couche (graduate 2003)
Hassan Ghazal (DEA 2001)
Magali Lescot (DEA 2001)
Patrice Chagnaud (DEA 2001)

Collaborations
(on regulatory sequences)

CIFN - Cuernavaca - Mexico
Julio Collado-Vides

ULB - Bruxelles - Belgium
Bruno André
Valérie Ledent
Morgane Thomas-Chollier

LIFL - Lille - France
Matthieu Defrance
Hélène Touzet

Université Paris IV (Jussieu)
Alessandra Carbone
Catherine Vaquero

LIRMM - Montpellier - France
Olivier Gascuel
Sylvie Pinloche

ENS - Paris - France
Philippe Marc

Univ Valencia - Spain
José Perez-Ortin
José Garcia-Martinez

Links and publications on regulatory sequences

Regulatory sequence analysis tools http://rsat.scmbb.ulb.ac.be/rsat/

Reprints and preprints http://www.scmbb.ulb.ac.be/Users/jvanheld/papers/

1. van Helden, J., Andre, B. & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281(5), 827-42.
2. van Helden, J., Rios, A. F. & Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res 28(8), 1808-18.
3. van Helden, J., Andre, B. & Collado-Vides, J. (2000). A web site for the computational analysis of yeast regulatory sequences. Yeast 16(2), 177-87.
4. van Helden, J., del Olmo, M. & Perez-Ortin, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res 28(4), 1000-10.
5. van Helden, J., Gilbert, D., Wernisch, L., Schroeder, M. & Wodak, S. (2001). Applications of regulatory sequence analysis and metabolic network analysis to the interpretation of gene expression data. Lecture Notes in Computer Sciences 2066, 155-172.
6. van Helden, J. (2003). Prediction of transcriptional regulation by analysis of the non-coding genome. Current Genomics 4(3), 217-224.
7. van Helden, J. (2003). Regulatory sequence analysis tools. Nucleic Acids Res 31(13), 3593-6.
8. van Helden, J. (2004). Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 2004 20(3):399-406.
9. Simonis, N., S.J. Wodak, G.N. Cohen, and J. van Helden. 2004. Combining pattern discovery and discriminant analysis to predict gene co-regulation. Bioinformatics 20: 2370-2379.
10. Simonis, N., J. Van Helden, G.N. Cohen, and S.J. Wodak. 2004. Transcriptional regulation of protein complexes in yeast. Genome Biol 5: R33.
11. Tompa, M., N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, V.J. Makeev, A.A. Mironov, W.S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137-144.
12. Gonze, D., S. Pinloche, O. Gascuel, and J. van Helden. 2005. Discrimination of yeast genes involved in methionine and phosphate metabolism on the basis of upstream motifs. *Bioinformatics* **21:** 3490-3500..
13. van Helden, J. 2005. The Analysis of Regulatory Sequences. In Multiple aspects of DNA and RNA: From Biophysics to Bioinformatics (ed. R. Monasson). Les Houches. *In press.*
14. Sand, O and van Helden, J. Evaluation of pattern discovery results. *In prep.*