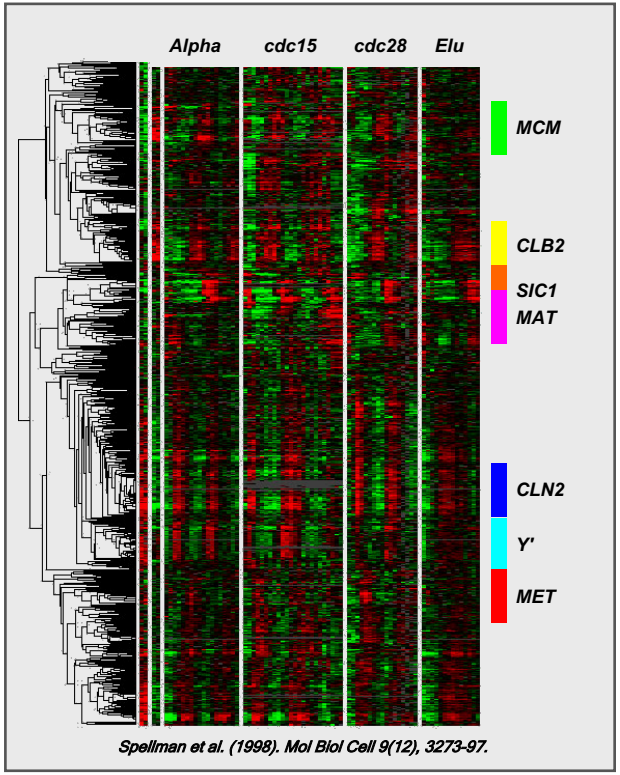
Pattern discovery
String-based approaches

Jacques van Helden
https://orcid.org/0000-0002-8799-8584

Aix-Marseille Université, France
Theory and Approaches of Genome Complexity (TAGC)

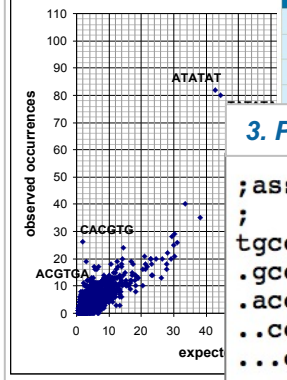
Institut Français de Bioinformatique (IFB)
http://www.france-bioinformatique.fr

Motif discovery in promoters of co-expressed genes



Spellman et al. (1998). Mol Biol Cell 9(12), 3273-97.

1. Count k-mer o



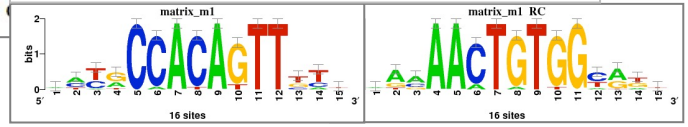
2. Over-representation statistics

seq	identifier	exp_freq	occ	exp_occ	occ_P	occ_E	occ_sig	rank
ccacag	ccacag|ctgtgg	0.0002365577659	18	1.97	4.8e-12	1.0e-08	8.00	1
cacgtg	cacgtg|cacgtg	0.0001569968432	14	1.31	1.4e-10	3.0e-07	6.52	2
actgtg	actgtg|cacagt	0.0003762020409	20	3.13	1.7e-10	3.6e-07	6.45	3
aactgt	aactgt|acagtt	0.0006168655788	21	5.14	1.2e-07	2.6e-04	3.59	4
		0.0002897084237	14	2.41	2.8e-07	5.7e-04	3.24	5
		0.0003355962588	15	2.80	2.8e-07	5.8e-04	3.24	6
					1.7e-05	3.6e-02	1.44	7
					2e-05	4.2e-02	1.37	8
					3.5e-05	7.2e-02	1.14	9
					4.8e-05	1.0e-01	1.00	10
					0.00016	3.3e-01	0.48	11
					0.00016	3.3e-01	0.48	12
					0.00019	3.9e-01	0.41	13
					0.00022	4.6e-01	0.34	14
					0.00035	7.2e-01	0.14	15

3. Pattern assembly

```
;assembly # 1    seed: ccacag      7 words
;     alignt          rev_cpl      score
tgccac.....     .....gtggca      0.48
.gccaca....     ....tgtggc.      3.24
.accaca....     ....tgtggt.      1.37
..ccacag...     ...ctgtgg..      8.00
...cacagt..     ..actgtg...      6.45
....acagtt.
.....cagttg
tgccacagttg
```

4. Extract matrix from patterns in sequence



Bruno André (ULB, Bruxelles, Belgium)

ULB

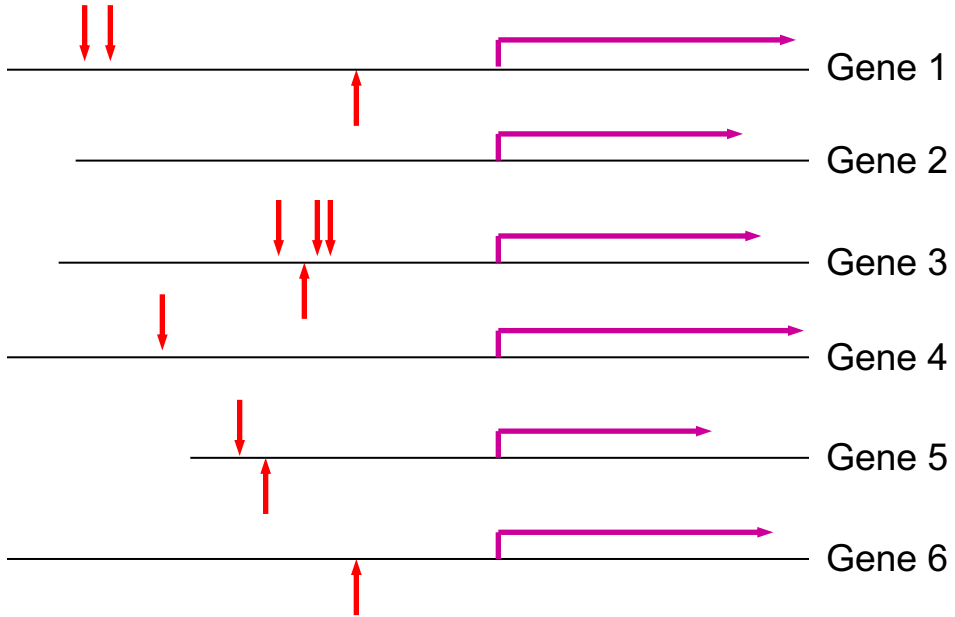
Julio Collado-Vides (CCG, Cuernavaca – Mexico)

CCG

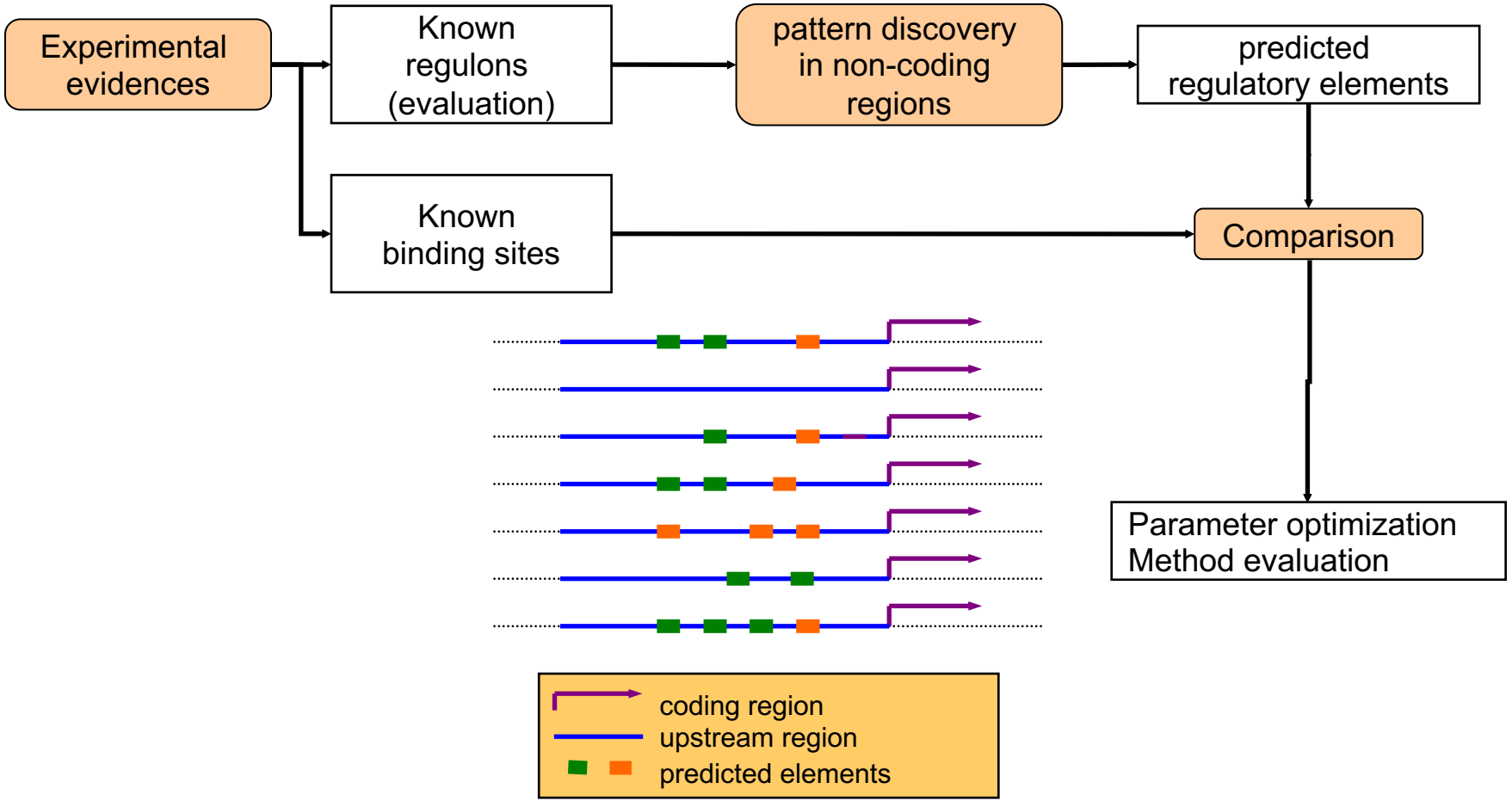
- van Helden, J., Andre, B. and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281, 827-42.
- van Helden, J., Andre, B. and Collado-Vides, J. (2000). A web site for the computational analysis of yeast regulatory sequences. Yeast 16, 177-87.
- van Helden, J., Rios, A. F. and Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res 28, 1808-18.

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
- ...

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

Background model

- In order to detect over-represented patterns, the observed occurrences are compared to the random expectation.
- The random expectation can be estimated according to different models
  - Bernouilli model, with a specific probability for each nucleotide.
  - Markov model, estimated on the basis of the input sequence itself.
  - External background : occurrences for the same pattern in a reference data set
    - whole genome
    - intergenic sequences
    - set of all upstream sequences for the organism considered

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

aaaaaa|tttttt	51
aaaaag|ctttttt	15
aagaaa|tttctt	14
gaaaaa|tttttc	13
tgccaa|ttggca	12
aaaaat|attttt	12
aaatta|taattt	12
agaaaa|ttttct	11
caagaa|ttcttg	11
aaacgt|acgttt	11
aaagaa|ttcttt	11
acgtgc|gcacgt	10

MET

aaaaaa|tttttt	105
atatat|atatat	41
gaaaaa|tttttc	40
tatata|tatata	40
aaaaat|attttt	35
aagaaa|tttctt	29
agaaaa|ttttct	28
aaaata|tatttt	26
aaaaag|ctttttt	25
agaaat|atttct	24
aaataa|ttattt	22
taaaaa|ttttta	21

NIT

aaaaaa|tttttt	80
cttatc|gataag	26
tatata|tatata	22
ataaga|tcttat	20
aagaaa|tttctt	20
gaaaaa|tttttc	19
atatat|atatat	19
agataa|ttatct	17
agaaaa|ttttct	17
aaagaa|ttcttt	16
aaaaca|tgtttt	16
aaaaag|ctttttt	15

GAL

aaaaaa|tttttt	47
aaaaat|attttt	17
aatata|tatatt	17
aaaatt|aatttt	16
aaaata|tatttt	15
attttc|gaaaat	13
aaataa|ttattt	13
aaatat|atattt	13
ataaaa|ttttat	12
atatta|taatat	12
atatat|atatat	11
tgaaaa|ttttca	11

A more relevant criterion for over-representation

- The most frequent patterns do not reveal the motifs specifically bound by specific transcription factors.
- They merely reflect the compositional biases of upstream sequences.
- A more relevant criterion for over-representation is to detect patterns which are more frequent in the upstream sequences of the selected genes (co-regulated) than the random expectation.
- The random expectation is calculated by counting the frequency of each pattern in the complete set of upstream sequences (all genes of the genome).

Estimation of word-specific expected frequencies with a Markov model

- In a Markov model, the probability to find a letter at position *i* depends on the residues found at the m preceding residues.
- The tables represent the transition matrices for Markov chain models of order 1 (top) and 2 (bottom).
- Expected frequencies can be estimated
  - On the basis of a set of ***background sequences*** (e.g. the whole set of upstream sequences of the considered organism).
  - On the basis of the ***input sequence set*** itself: the probability of larger words is estimated from the observed frequencies of the sub-words that compose them.

$$P(S,m) = P(S_{1,m}) \prod_{i=m+1}^{L} P\left(r_i \mid S_{i-m,i-1}\right)$$

Transition matrix, order 1

	g	a	c	t
a	0.178	0.369	0.165	0.288
c	0.166	0.327	0.191	0.316
g	0.190	0.313	0.211	0.286
t	0.175	0.273	0.180	0.372

Transition matrix, order 2

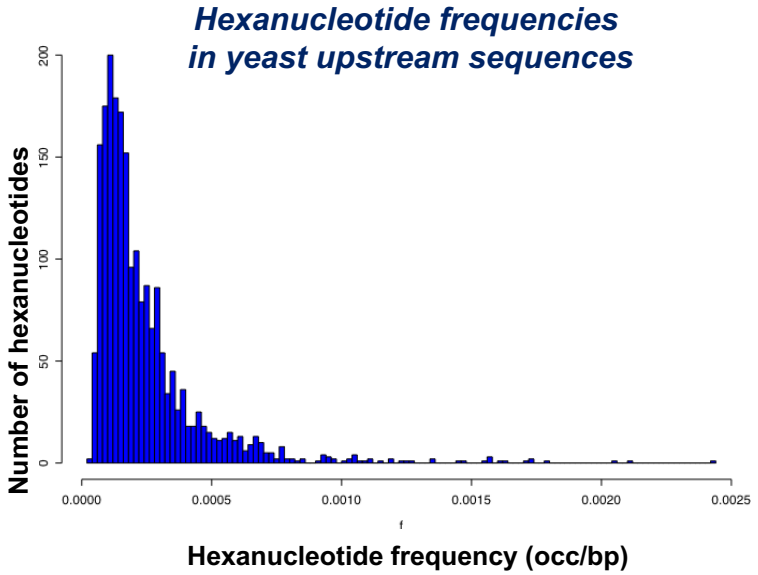
	g	a	c	t
aa	0.185	0.411	0.152	0.252
ac	0.171	0.348	0.186	0.296
ag	0.193	0.337	0.201	0.269
at	0.163	0.343	0.167	0.326
ca	0.181	0.344	0.184	0.291
cc	0.168	0.313	0.198	0.321
cg	0.194	0.283	0.227	0.295
ct	0.187	0.240	0.189	0.384
ga	0.186	0.407	0.145	0.262
gc	0.180	0.331	0.194	0.295
gg	0.192	0.318	0.216	0.274
gt	0.199	0.305	0.159	0.338
ta	0.160	0.304	0.182	0.354
tc	0.151	0.313	0.192	0.344
tg	0.184	0.302	0.210	0.304
tt	0.168	0.220	0.195	0.417

Estimation of word-specific expected frequencies from a set of background sequences

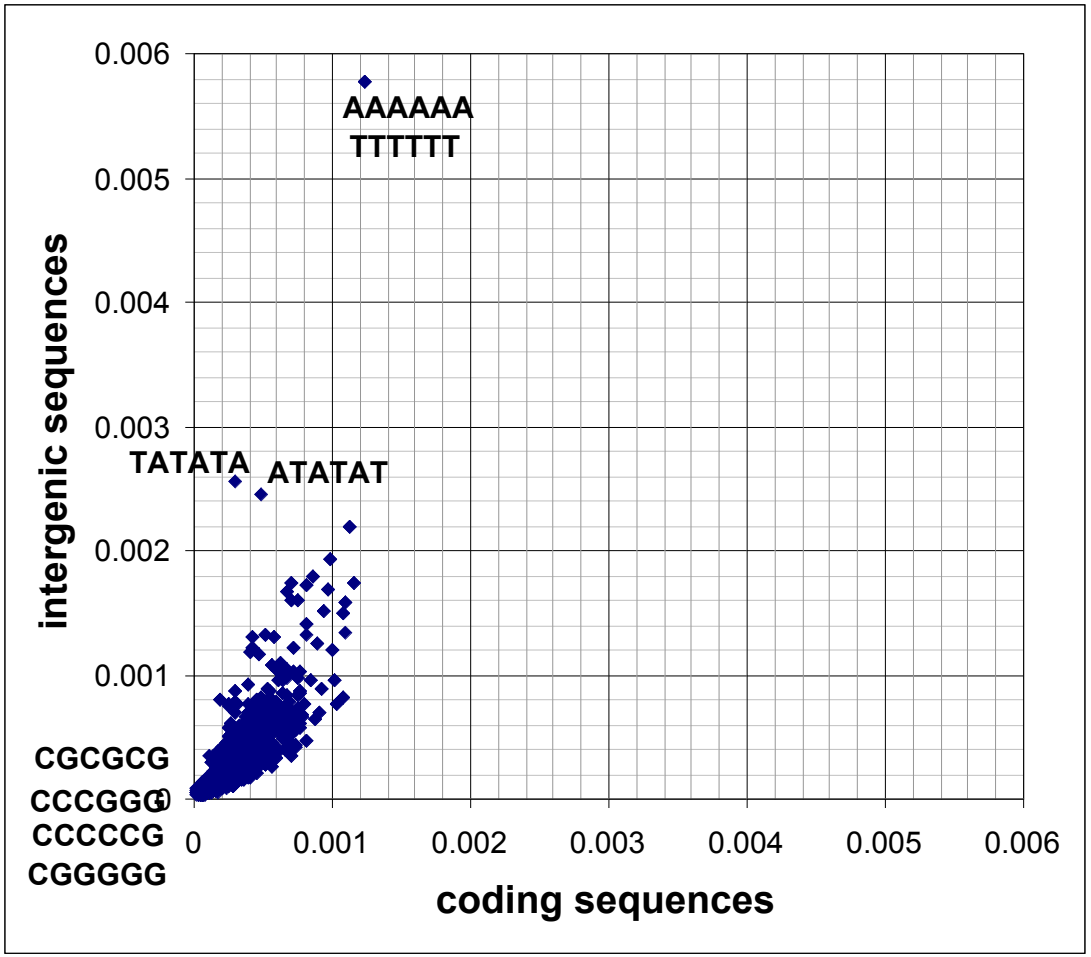
Example:
6nt frequencies in the whole set of yeast upstream sequences
Words are grouped by pairs of reverse complements.

;seq	identifier	observed_freq	occ
aaaaaa	aaaaaa|ttttt	0.00510699	14555
aaaaac	aaaaac|gtttt	0.00207402	5911
aaaaag	aaaaag|ctttt	0.00375191	10693
aaaaat	aaaaat|atttt	0.00423577	12072
aaaaca	aaaaca|tgttt	0.0019828	5651
aaaacc	aaaacc|ggttt	0.00088526	2523
aaaacg	aaaacg|cgttt	0.00090105	2568
aaaact	aaaact|agttt	0.0014621	4167
aaaaga	aaaaga|tcttt	0.00323016	9206
aaaagc	aaaagc|gcttt	0.00135824	3871
aaaagg	aaaagg|cctttt	0.0017849	5087
aaaagt	aaaagt|acttt	0.0019035	5425
aaaata	aaaata|tattt	0.00336805	9599
aaaatc	aaaatc|gattt	0.00131368	3744
aaaatg	aaaatg|cattt	0.00185648	5291
aaaatt	aaaatt|aattt	0.00269156	7671
aaacaa	aaacaa|ttgtt	0.00209999	5985
aaacac	aaacac|gtgtt	0.00071684	2043
aaacag	aaacag|ctgtt	0.00096491	2750
aaacat	aaacat|atgtt	0.00108982	3106
aaacca	aaacca|tggtt	0.00074421	2121

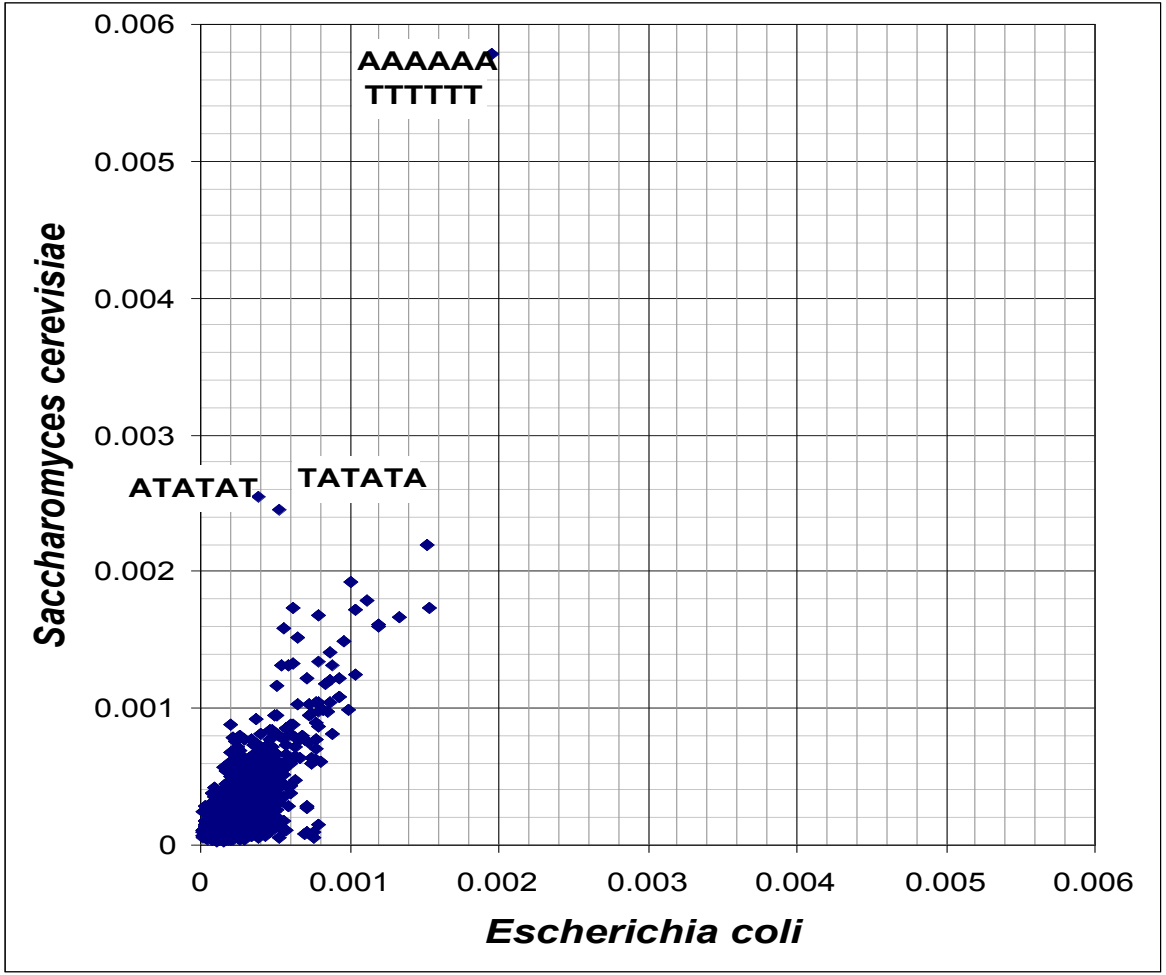
- Hexanucleotide frequencies of have been measured in the whole set of 6000 yeast upstream sequences.
- Some words are very frequent, others are rare.
  - range $4.5e^{-5}$ to $1.2e^{-2}$
  - Ratio between the most frequent and less frequent hexanucleotide:
    - max(f)/min(f)=268



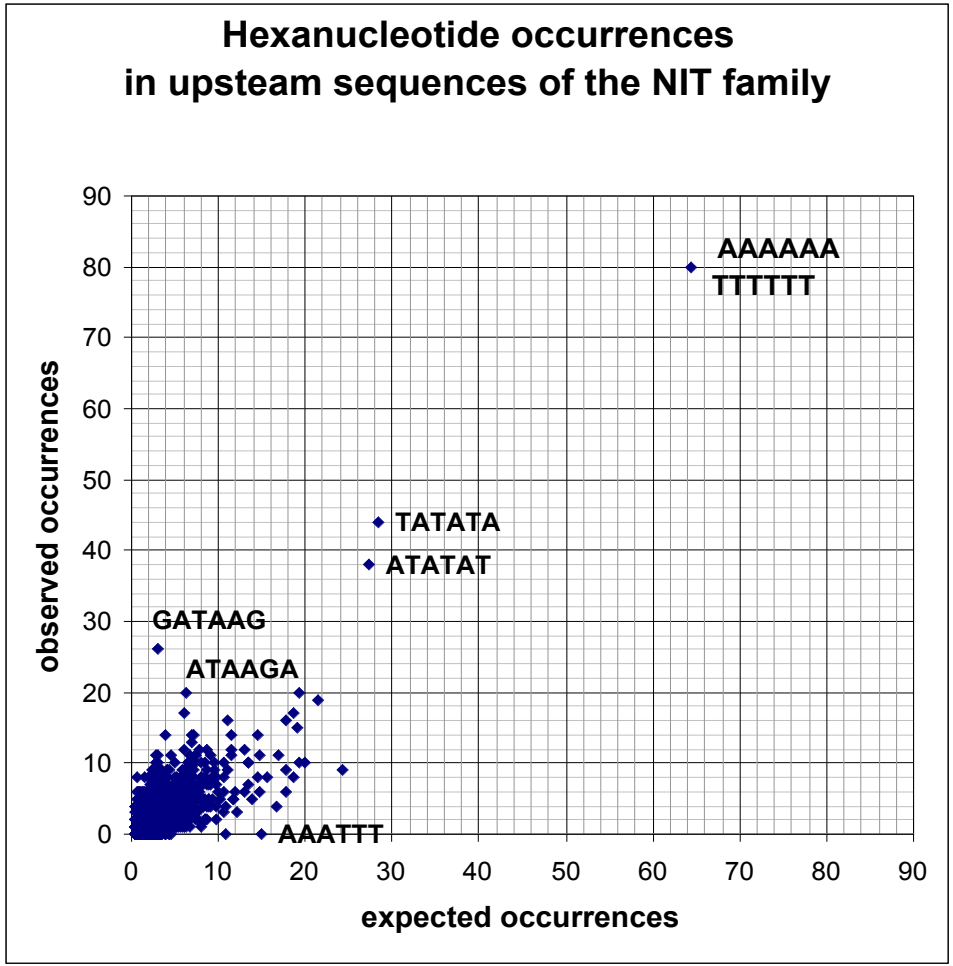
6nt frequencies differ between coding and non-coding sequences



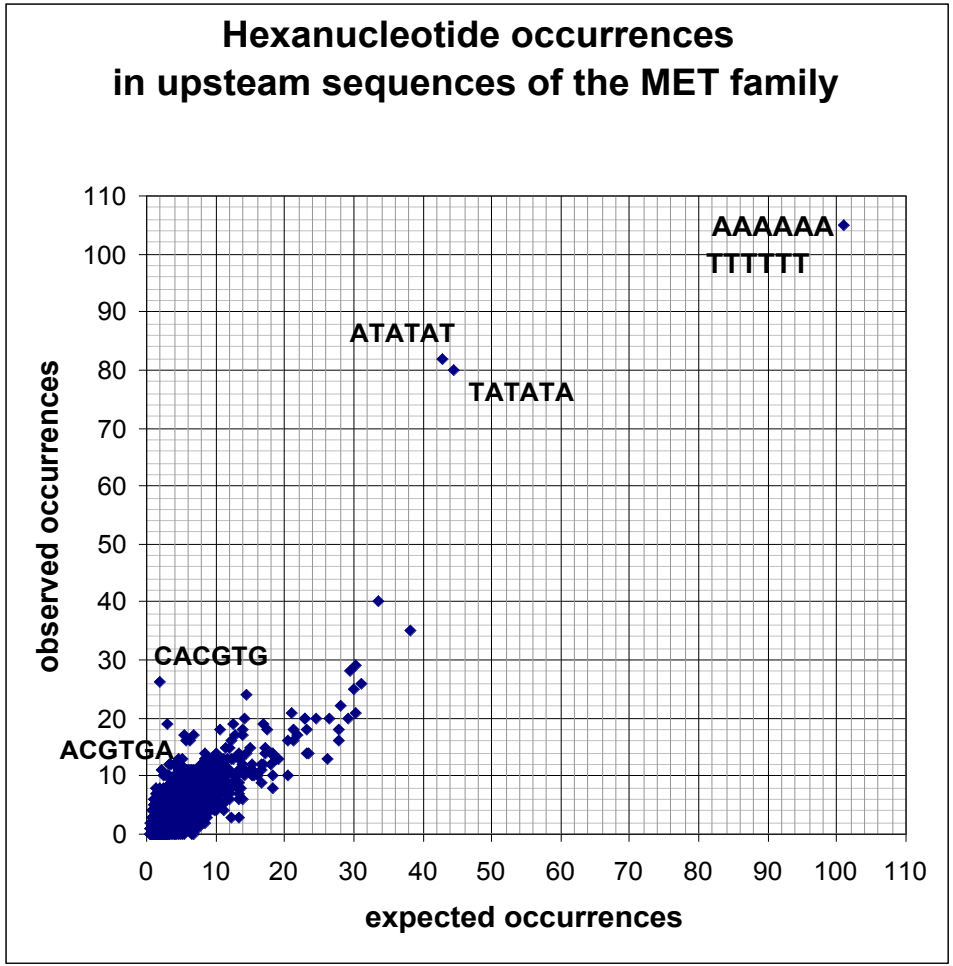
Inter-species variations in intergenic 6nt frequencies



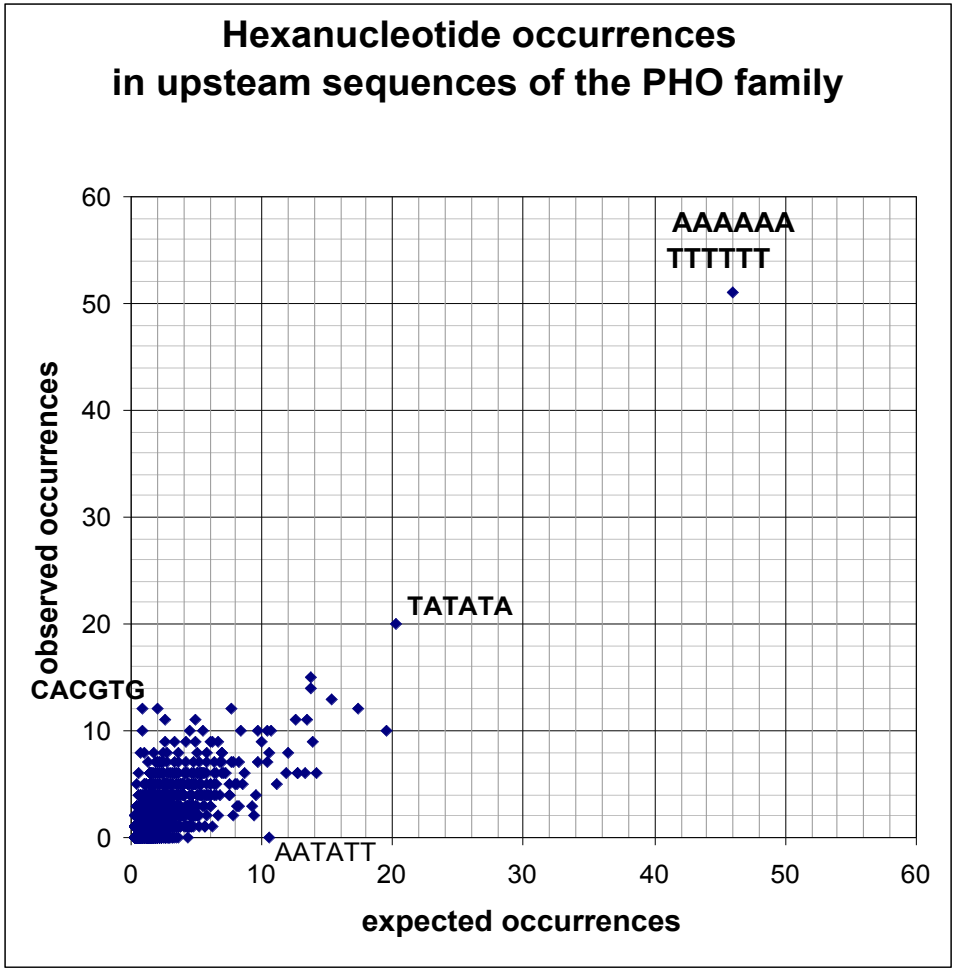
Hexanucleotide occurrences in upstream sequences of NIT genes



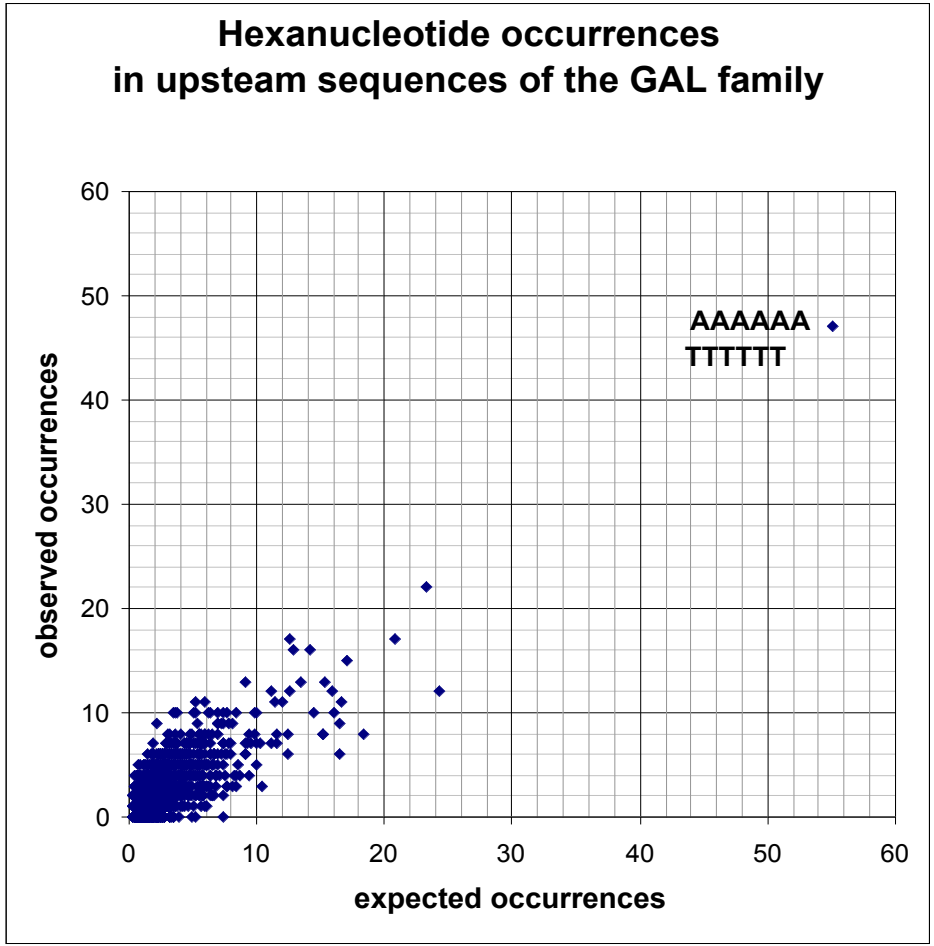
Hexanucleotide occurrences in upstream sequences of MET genes



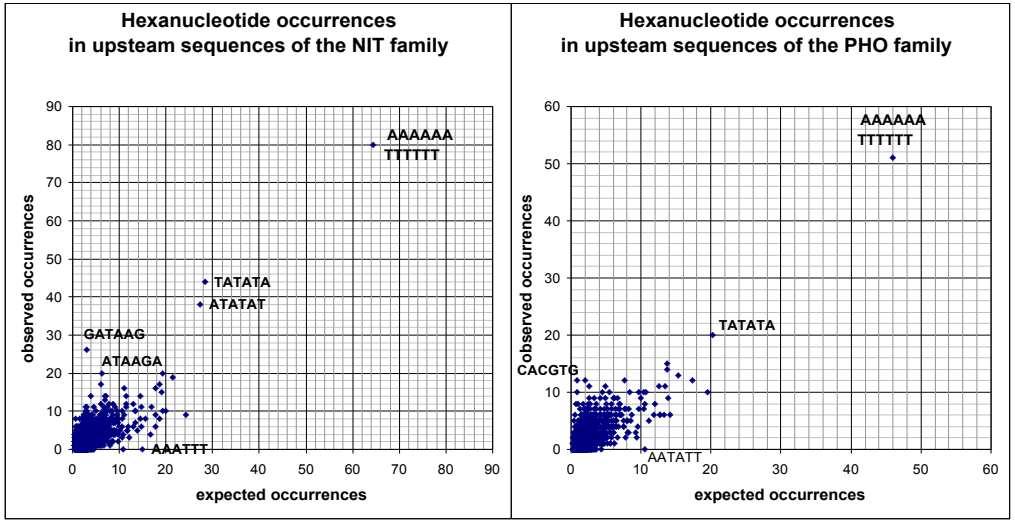
Hexanucleotide occurrences in upstream sequences of PHO genes

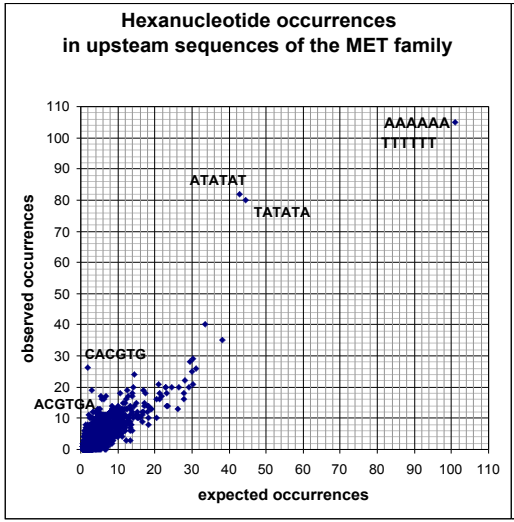


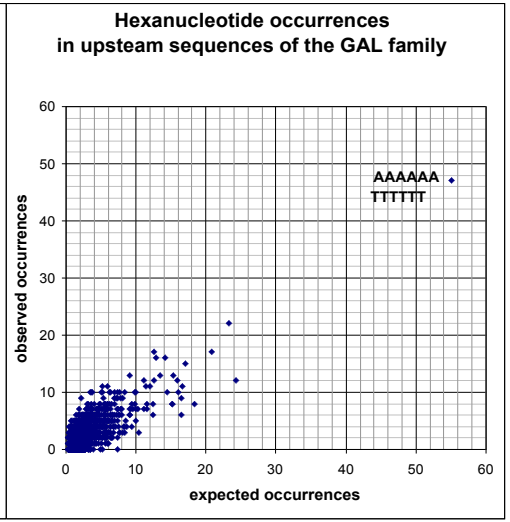
Hexanucleotide occurrences in upstream sequences of GAL genes



Hexanucleotide occurrences in yeast NIT, PHO, MET and GAL upstream sequences

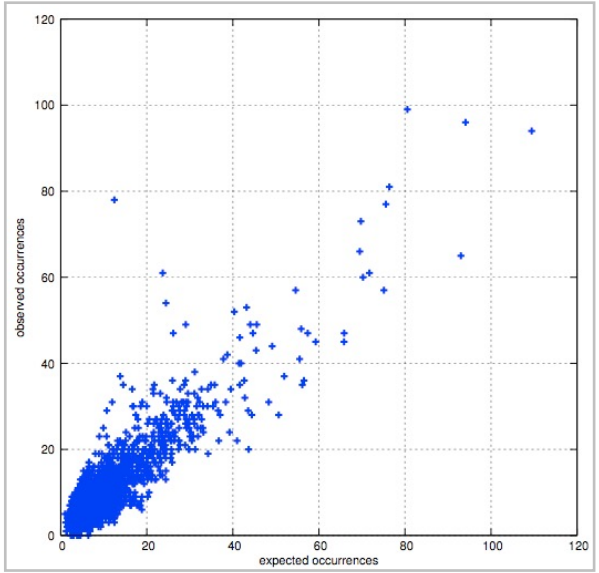


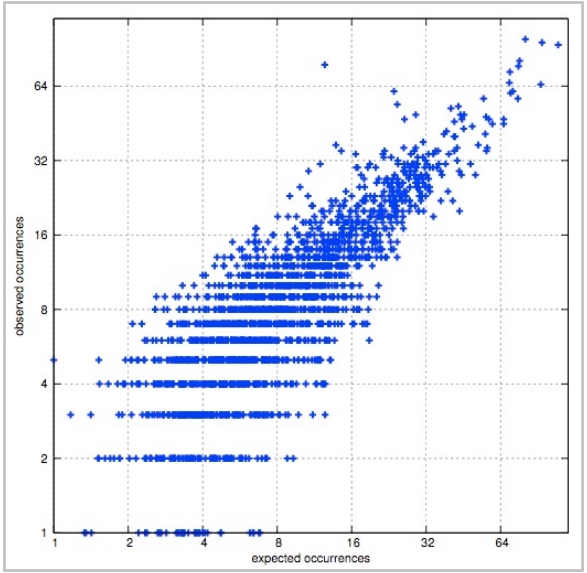




Hexanucleotide occurrences in an extended NIT regulon

- We analyzed here an extended set of 41 NIT genes (taken from Godard et al., 2006).
- The number of genes affects the dispersion around the diagonal on the plot of observed versus expected occurrences.
- The signal-to-noise separation increases when more genes are analyzed.
- The logarithmic axes better emphasize the words with low expected and observed occurrences but does not allow to display words with 0 occurrences.
- Words with very low expected frequencies are sensitive to low-number fluctuations. For such cases, the observed/expected ratio is misleading (e.g. exp=1, obs=4).





Scoring statistics

- Several scoring statistics have been used to assess the statistical significance of word over-representation
- Observed/expected ratio
  - Never use this statistics !
  - The ratio can be misleading, because it over-emphasizes the patterns with a very low number of expected number of occurrences
  - Example:
    - x_{obs}/x_{exp} = 10/1 is quite significant, but x_{obs}/x_{exp} =1/0.1 is not.
- Log-likelihood ratio
  - $LLR=F_{obs}*log(F_{obs}/F_{exp})$
- Z-score (Matthieu Blanchette)
  - $Z\text{-}score = (x_{obs}-x_{exp})/s_X$
  - Only valid for very large sequences (exp >> 10 for each word)
- Poisson (Andreas Wagner)
- Compound Poisson (Sophie Schbath)
- Binomial (Jacques van Helden)

Scoring scheme - Binomial

- Advantages
  - Allows to estimate a P-value.
  - Appropriate for small sequence sets, where some words have a very low expected number of occurrences (<1).
  - Allows to detect over- and under-representation.
- Weaknesses
  - Bias for self-overlapping words (but this can be circumvented by preventing the counting of overlapping occurrences).
  - Assumes that sequence length is much larger than word length

- Probability to observe exactly x occurrences

$$P(X = x) = \frac{T!}{x!(T-x)!} p^x (1-p)^{T-x}$$

- Probability to observe at least s occurrences

$$P(X \geq x) = \sum_{i=x}^{T} \frac{T!}{i!(T-i)!} p^i (1-p)^{T-i}$$

Where

x = observed occurrences

$T = Sum_{i=1->n}(L_i\text{-}k+1)$ = number of possible positions for a word of length k in a sequence of n sequences of length L_i

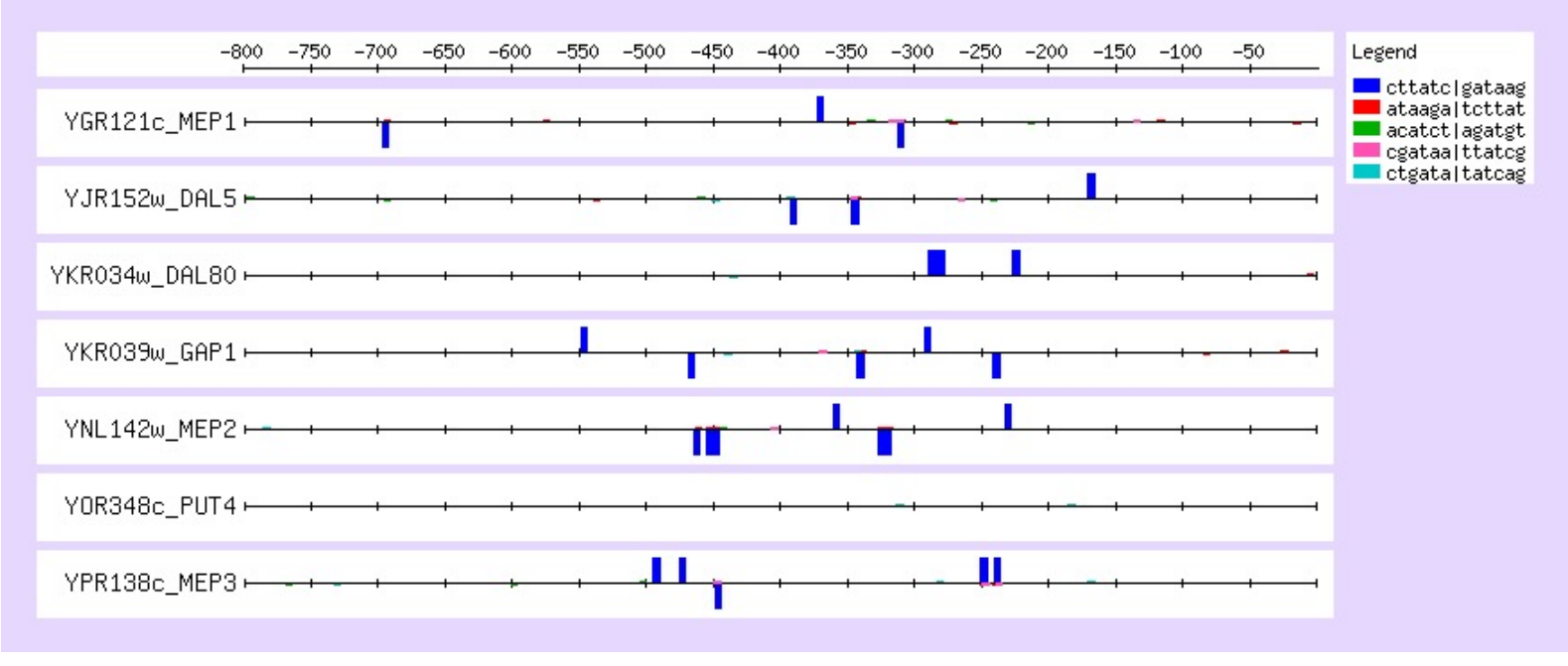
p = word probability

Hexanucleotide analysis in sequences upstream of the NIT regulon

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

Feature-map of discovered patterns - NIT regulon

- Typical features of yeast GATA-boxes
  - Multiple occurrences per sequences.
  - Occurrences generally appear clustered (at least two with a spacing of 0-60bp).
  - This probably stimulates synergic effects.
- Remark: PUT4 does not contain a single optimal motif



Hexanucleotide analysis of the PHO regulon

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

Genes *PHO5, PHO8, PHO11, PHO84, PHO81*

Known motifs *Factors*

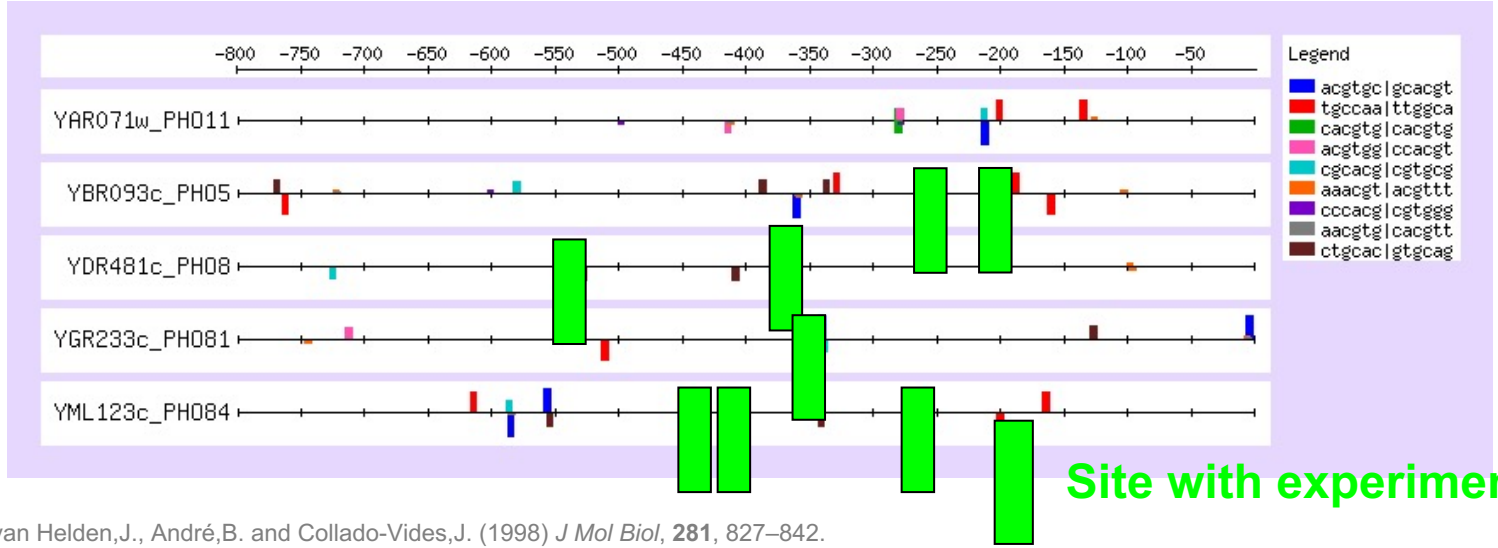
CACGTGGG *Pho4p (high affinity)*

CACGTTTT *Pho4p (medium affinity)*

- van Helden,J., André,B. and Collado-Vides,J. (1998) *J Mol Biol*, **281**, 827–842.

Feature-map of discovered patterns - PHO regulon

- The feature-map provides a convenient representation of the discovered patterns
  - Each colour represents one pattern.
  - Box height reflects pattern significance.
  - Clusters of mutually overlapping words represent sites larger than 6 bp.
- Green bars were superimposed, to indicate the positions of experimentally proven sites, and compare predictions with experimental knowledge.
  - For PHO11, no site is documented, we can thus not check the predictions.
  - For the other genes, the proven sites are detected as clusters of overlapping words

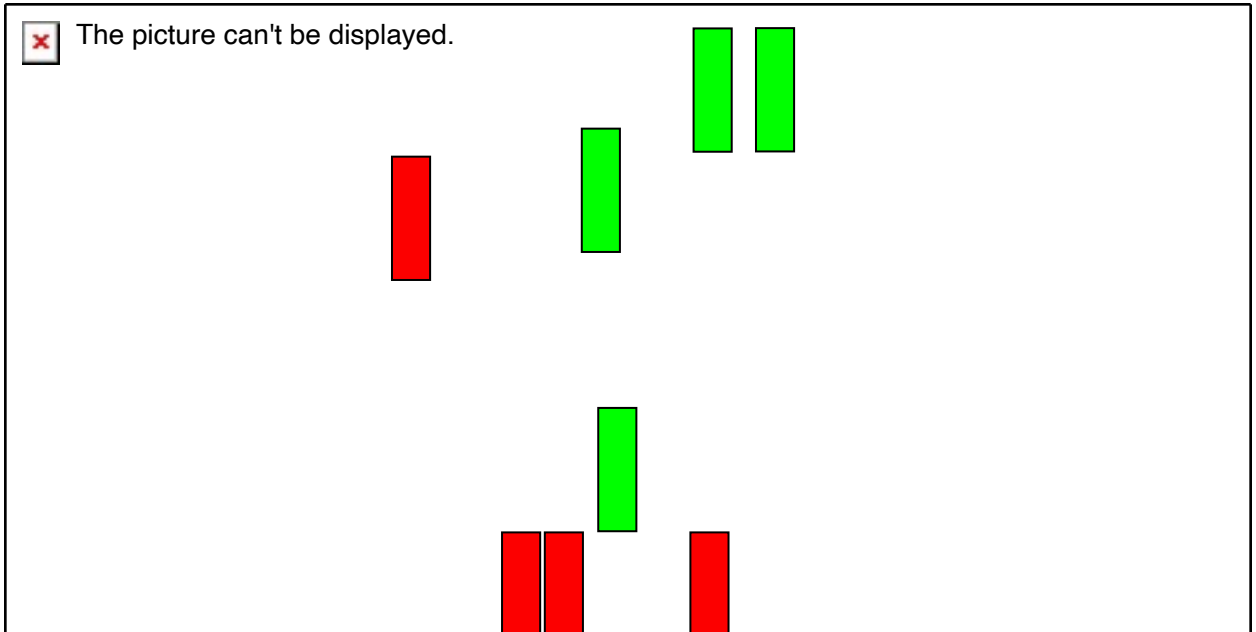


- van Helden,J., André,B. and Collado-Vides,J. (1998) *J Mol Biol*, **281**, 827–842.

Clipping of upstream coding sequences

In the particular case of the the yeast Saccharomyces cerevisiae, the initial annotations were over-predictive, and contained many false ORFs.

- Clipping upstream ORFs sometimes results in a loss of information.
- In the case of the PHO family, half of the know sites would be clipped, and the pattern discovery program would not identify any significant motif anymore.
- This problem has recently been solved, with the new annotations based on comparative genomics.



Hexanucleotide analysis of the MET regulon

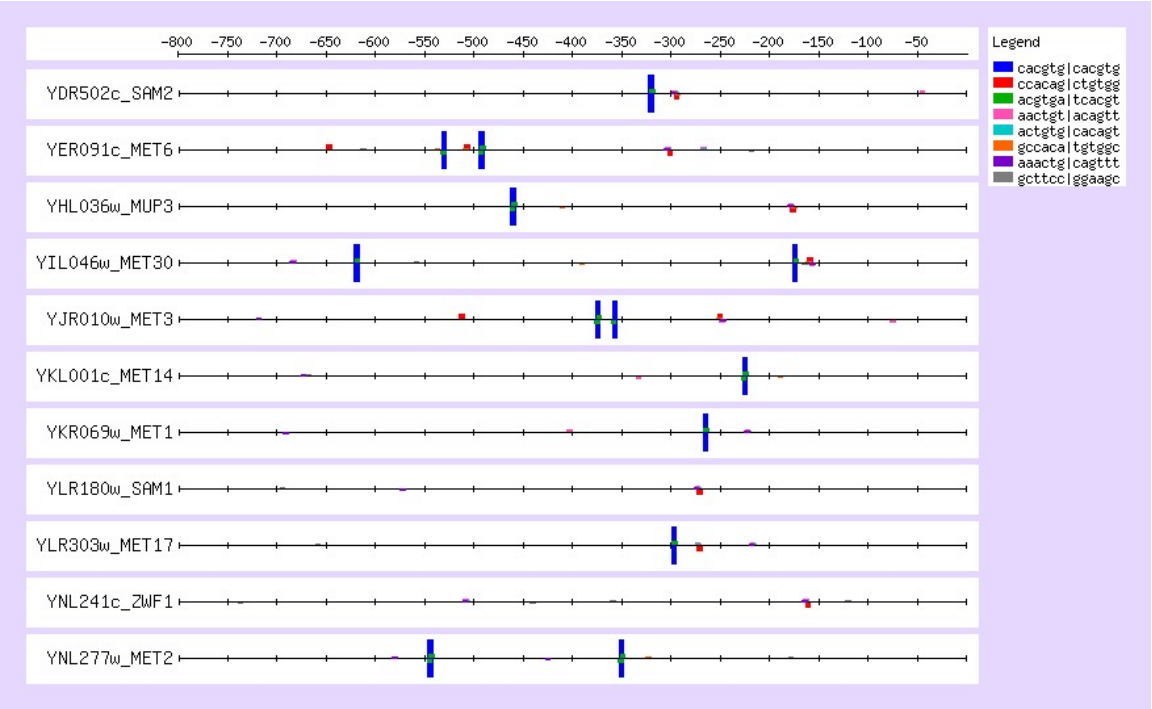
Sequence	exp freq	occ	exp occ	P-value	E-value	sig	matching sequences
..ACGTGa	0.00033	13	2.9	1.00E-05	2.20E-02	1.67	9
.CACGTG.	**0.00012**	**13**	**1.0**	**6.90E-11**	**1.40E-07**	**6.84**	**9**
tCACGTG.	0.00033	13	2.9	1.00E-05	2.20E-02	1.67	9
tCACGTGa	consensus						
....TGTGGc	0.00027	10	2.3	1.50E-04	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

Genes	*SAM2, MET6, MUP3, MET30, MET3, MET14, MET1, SAM1, MET17, ZWF1, MET2*
Known motifs	*Factors*
TCACGTG	*Cbf1p/Met4p/Met28p*
AAAACTGTGG	*Met31p; Met32p*

- van Helden,J., André,B. and Collado-Vides,J. (1998) *J Mol Biol*, **281**, 827–842.

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.



Expected frequency calibration

- 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	Bernoulli	equiprobable
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)	**aTTACTAA**	-	-	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

- 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 ?

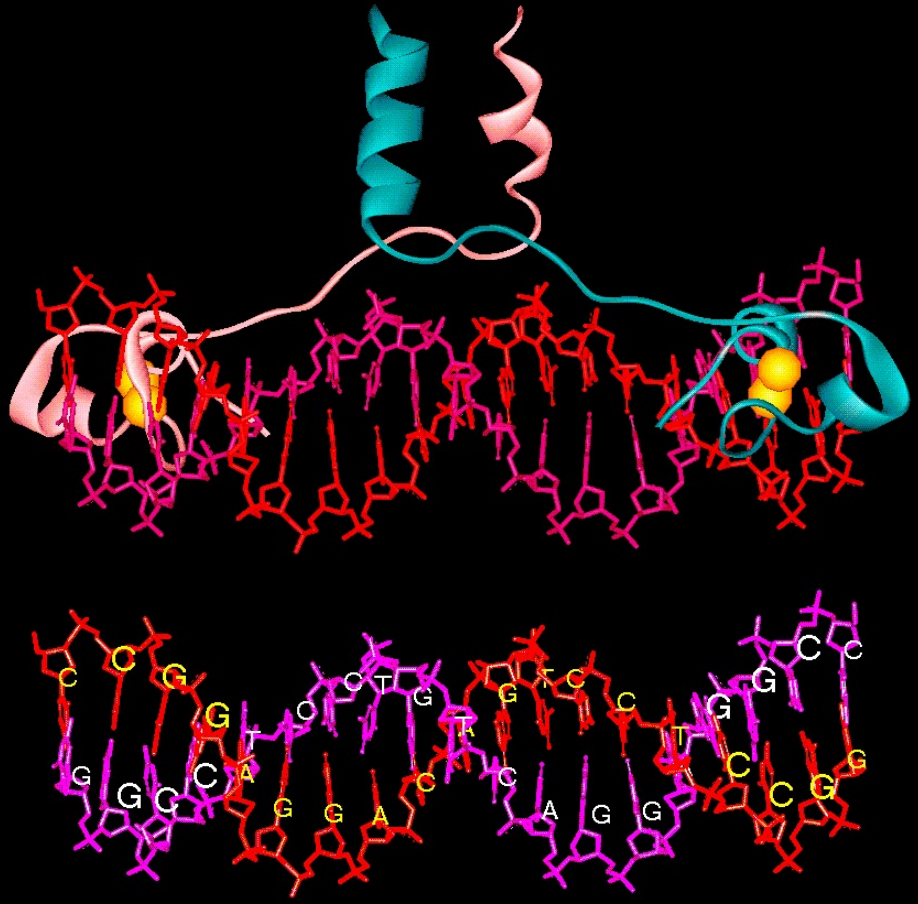
Sequence	exp freq	occ	exp occ	P-value	E-value	sig	matching sequences
agacat	0.00044	9	2.1	0.00033	0.69	0.16	4

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

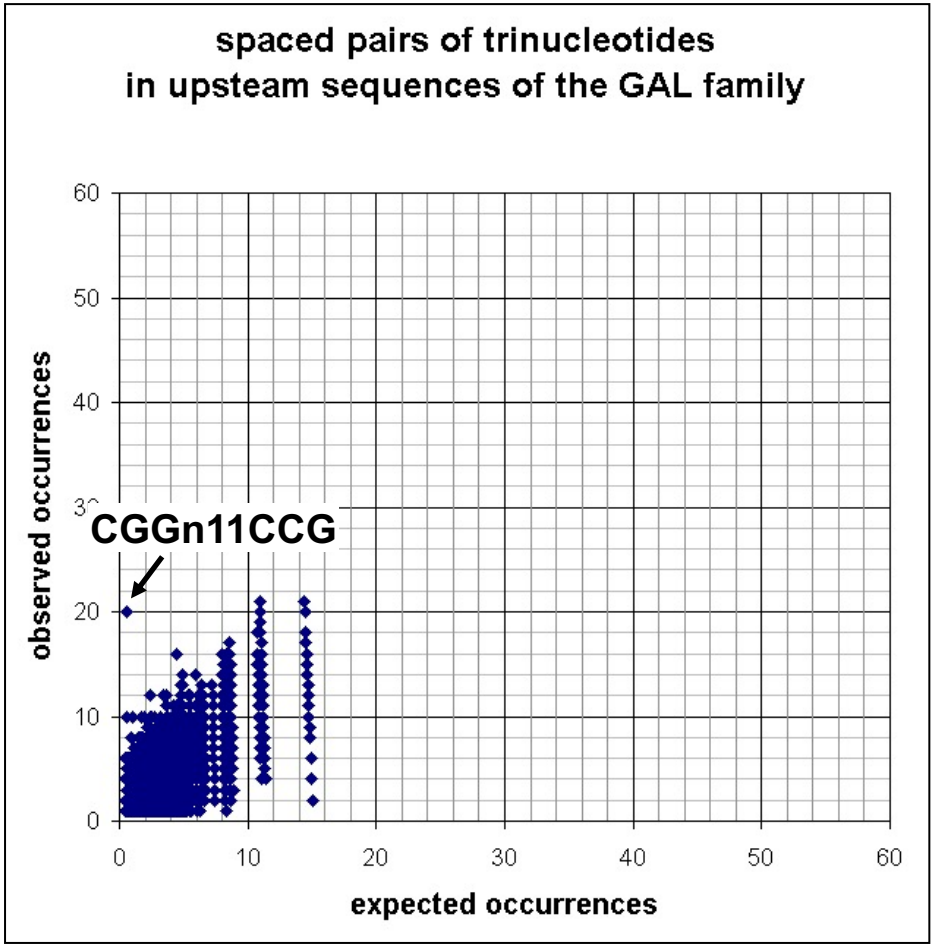
Known motifs *Factors*

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

DNA/protein interface of the yeast transcription factor Gal4p



Occurrences of 3nt dyads in the GAL family



Dyad analysis of the GAL family

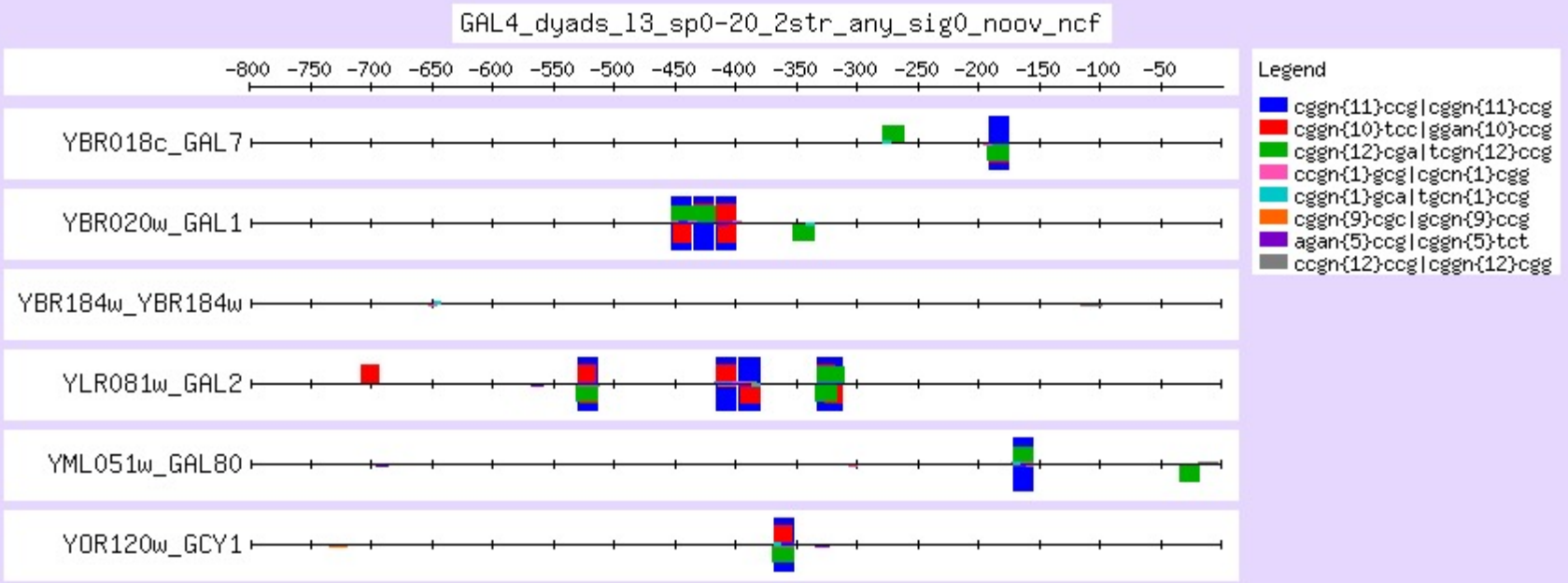
		occ	occ			
..GGa.........CCG.	0.00006	10	0.5	2.70E-10	1.20E-05	4.92
.CGG...........Cga	0.00006	10	0.5	4.80E-10	2.10E-05	4.68
.CGG...........CCG.	**0.00007**	**20**	**0.6**	**2.10E-12**	**9.20E-08**	**7.03**
.CGG.........tCC..	0.00006	10	0.5	2.70E-10	1.20E-05	4.92
.CGG.........cgC...	0.00004	6	0.4	5.30E-06	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

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

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).

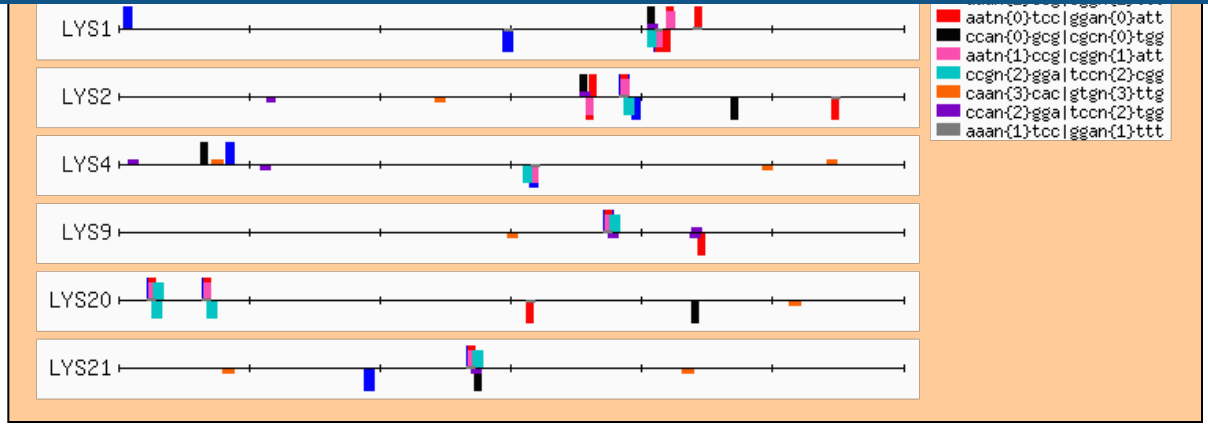
Feature map of discovered patterns - GAL family



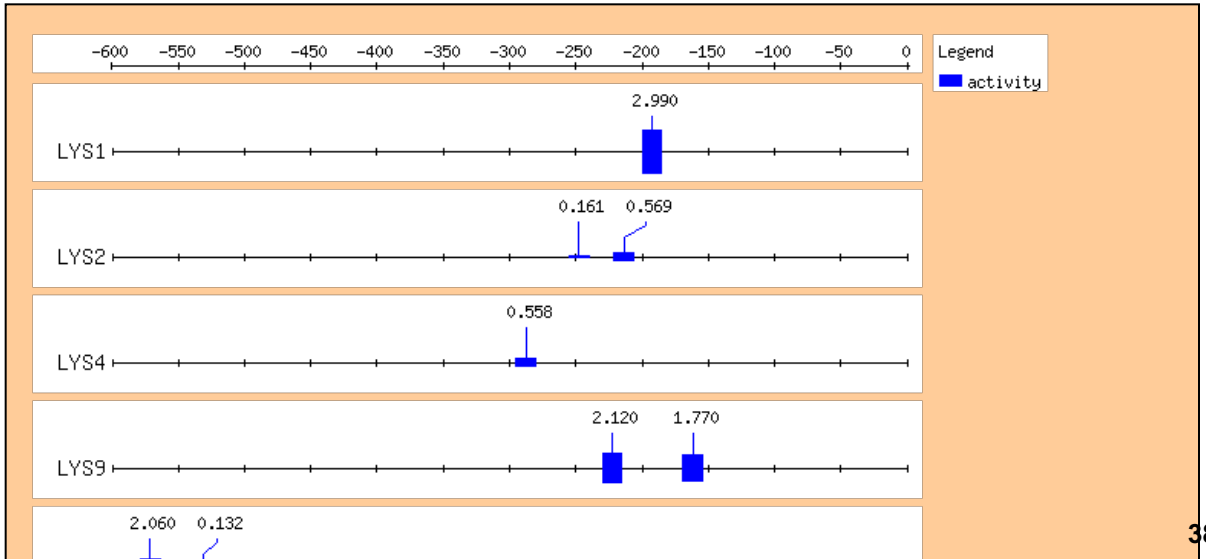
FACTOR	# genes	KNOWN MOTIFS	DYADS	REVERSE DYADS	SCORE
GAL4	6	CGGn$_{11}$CCG	TCGGAn$_9$TCCGG TCGGCGCAGAn$_4$TCCGG	CCGGAn$_9$TCCGA CCGGAn$_4$TCTGCGCCGA	7.8 7.8
HAP1	9	CGGnnntanCGG	GGAn$_5$CGGC GGGGGn$_{12}$GGC CCTn$_{10}$GGC	GCCGn$_5$TCC GCCn$_{12}$CCCCC GCCn$_{10}$AGG	1.8 1.4 1.1
LEU3	5	RCCggnnccGGY	CCGn$_3$CCG	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

Comparison of discovered patterns with known sites (LYS family)

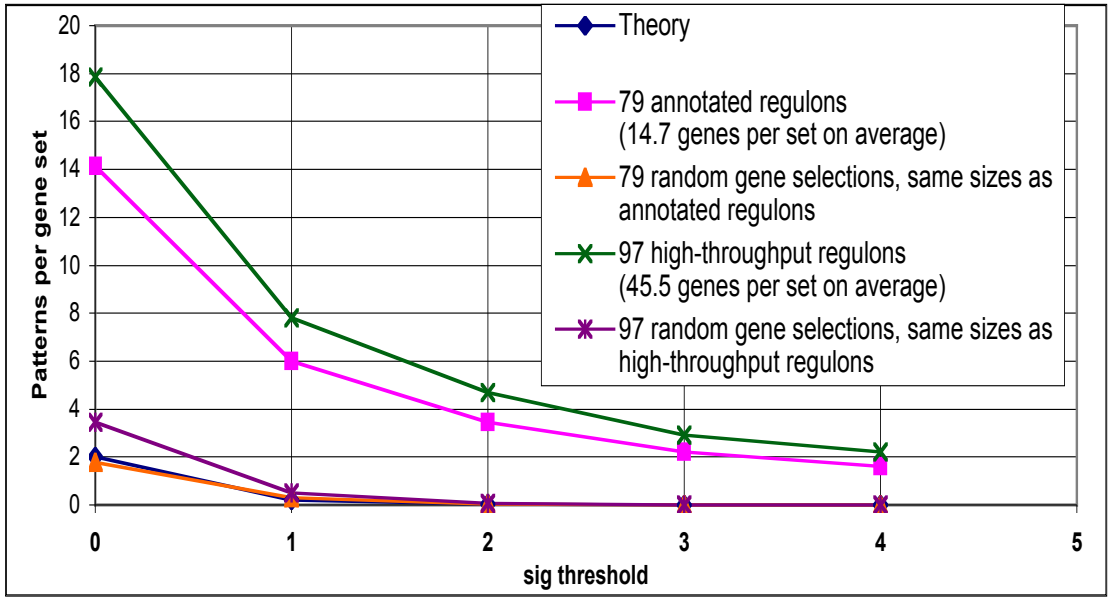
Patterns discovered by dyad analysis

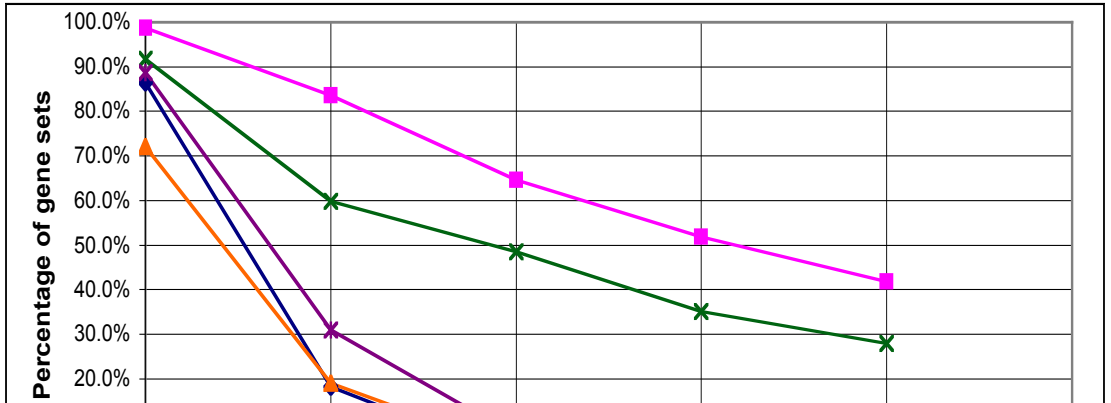


Experimental measurement of activity

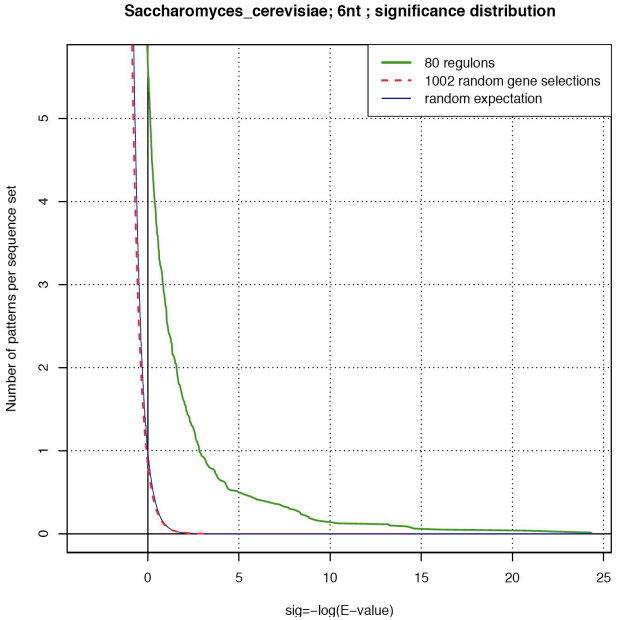


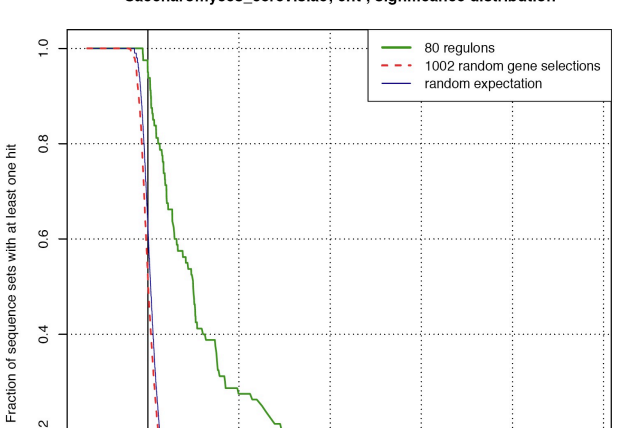
Quantitative evaluation of pattern discovery results





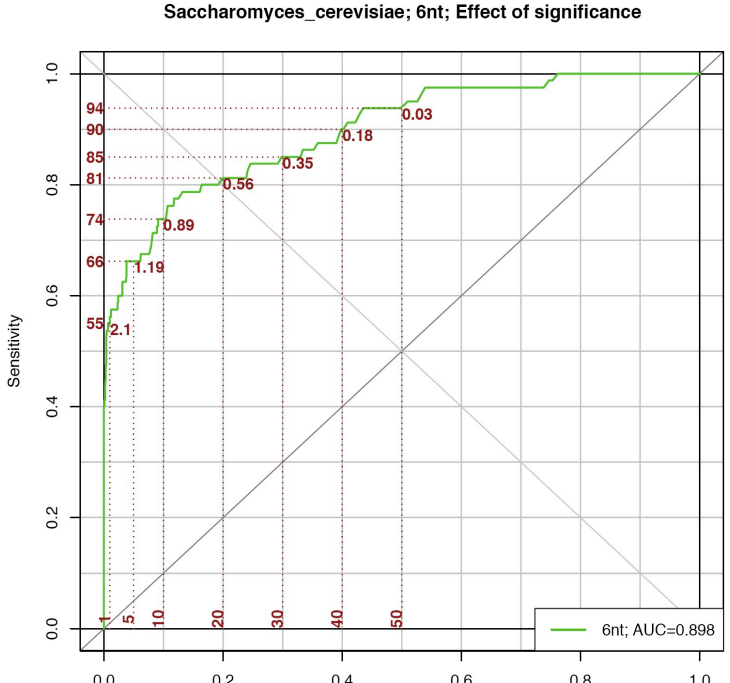
- These figures regroup patterns detected with
  - *oligo-analysis*
  - *dyad-analysis*
- Regulons were collected from TRANSFAC and aMAZE.
- All the regulons with ≥5 genes were analyzed.
  - Significant patterns (sig ≥ 2) are detected in 65% of the regulons.
- As a negative control, sets of random genes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





- **in regulons** (*positive control*)
- **in random gene selections** (*negative control*)

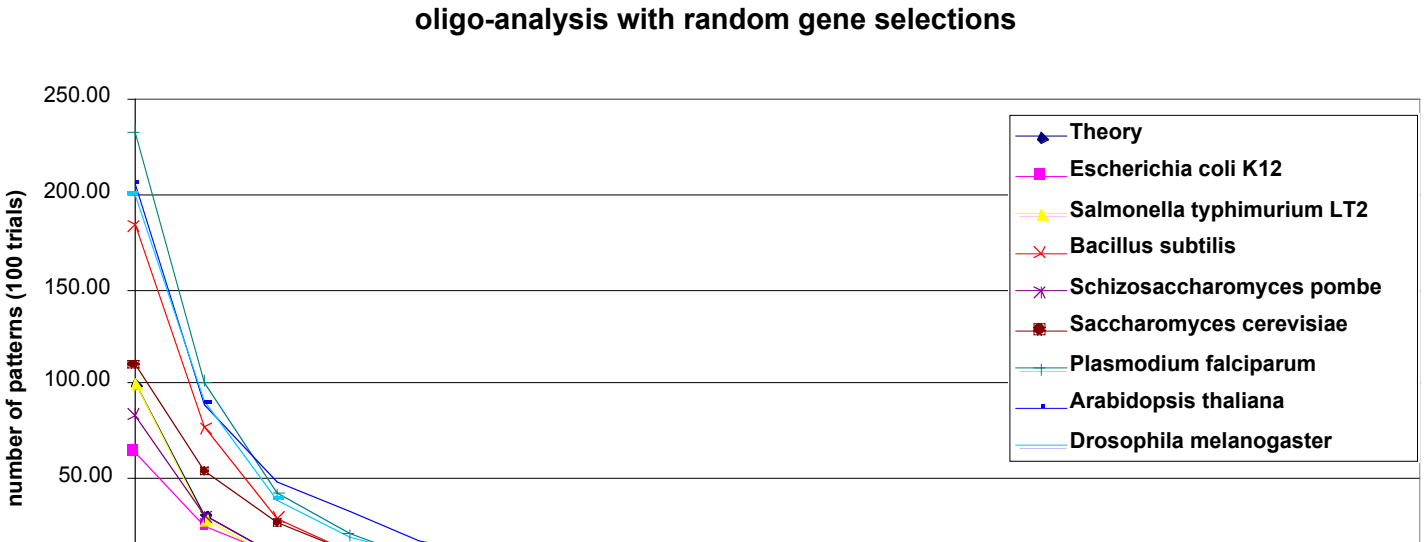
- In the yeast *Saccharomyces cerevisiae*
  - FPR fits remarkably well the binomial P-value.
  - When the significance threshold increases,
    - sensitivity decreases (less patterns found in regulons)
    - specificity increases (less patterns in random selections)



Rate of false positive in different organisms

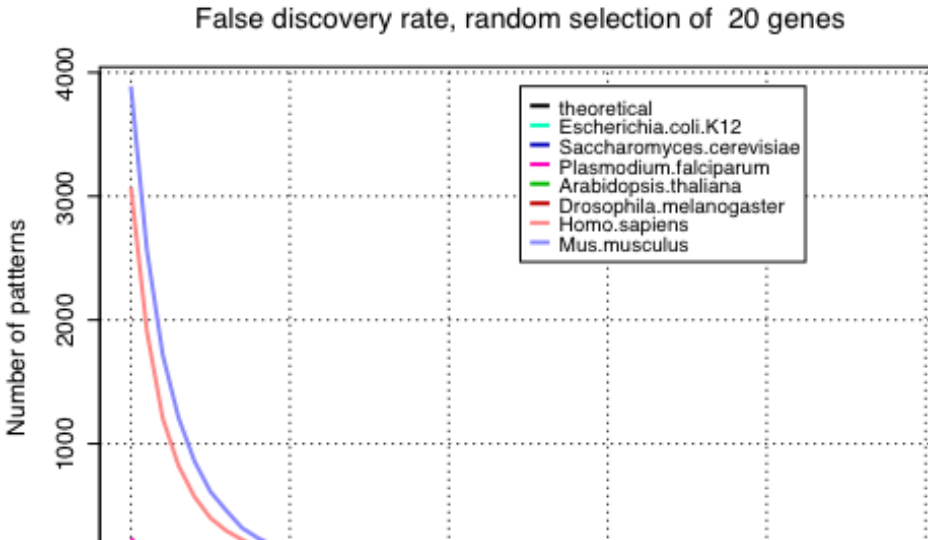
The analysis of random gene selections allows to evaluate the rate of false positive returned by a pattern discovery program.

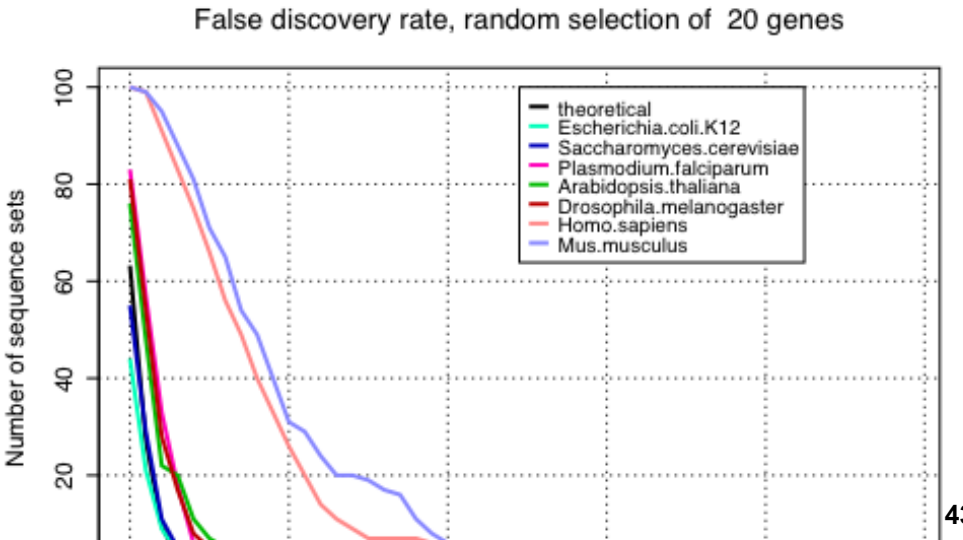
- The rate of false positive is good for microbes (bacteria, yeasts, ...), but increases for multicellular organisms (e.g. the fly Drosophila, the plant Arabidopsis thaliana, …).
- The rate of false positive is also higher in the protozoan Plasmodium falciparum (the agent of the malaria) than in bacteria and yeast.

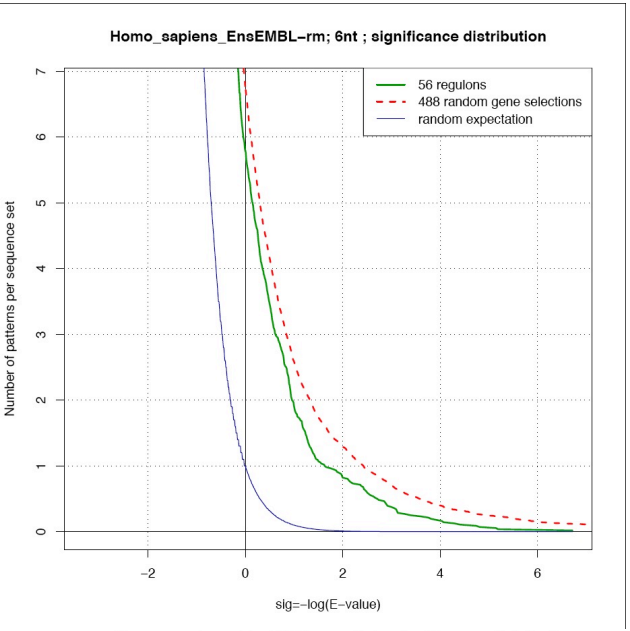


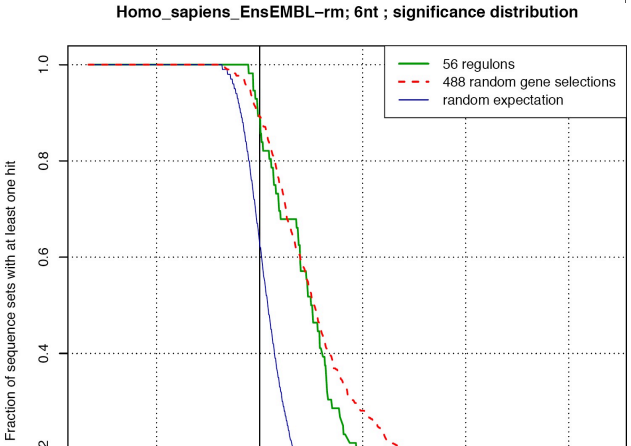
Rate of false positive in higher organisms

- 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.

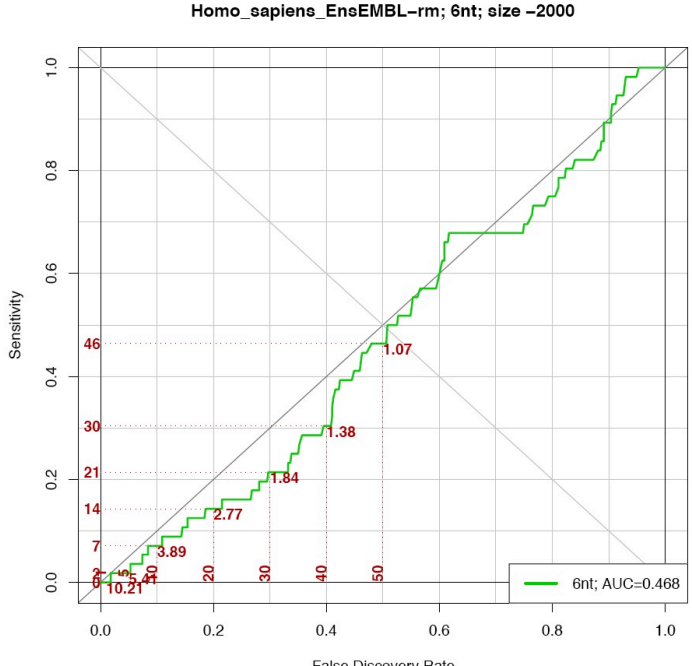








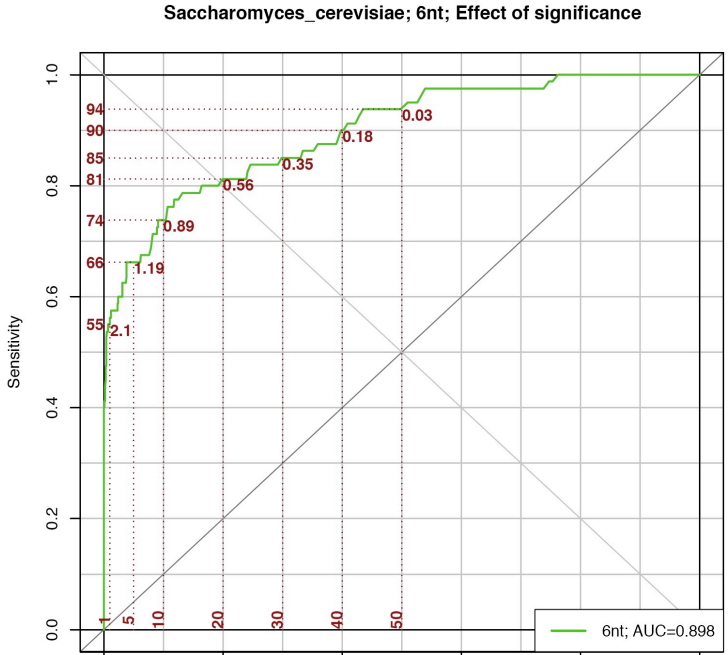
- In *Homo sapiens*
  - False positive rate (FPR) much higher than theoretical expectation
  - Significance score is quite inefficient to distinguish between reliable motifs and false positives.
  - Reasons:
    - Inadequacy of background models.
    - Actual TFBS are not restricted to proximal promoters



discovered in *Saccharomyces cerevisiae*

- **in regulons** (*positive control*)
- **in random gene selections** (*negative control*)

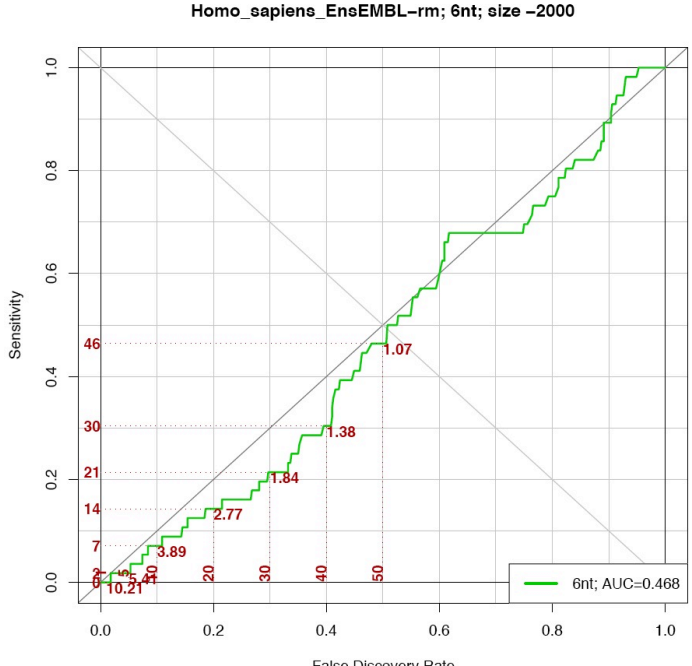
- In the yeast *Saccharomyces cerevisiae*
  - FPR fits remarkably well the binomial P-value.
  - When the significance threshold increases,
    - sensitivity decreases (less patterns found in regulons)
    - specificity increases (less patterns in random selections)



- ***False positive rate (FPR) much higher than theoretical expectation.***
- ***Significance score cannot distinguish between reliable motifs and false positives.***
- ***Reasons:***
  - ***Inadequacy of background models.***
  - ***Actual TFBS are not restricted to proximal promoters.***

Olivier Sand

Jean Valéry Turatsinze



String-based pattern discovery: strengths

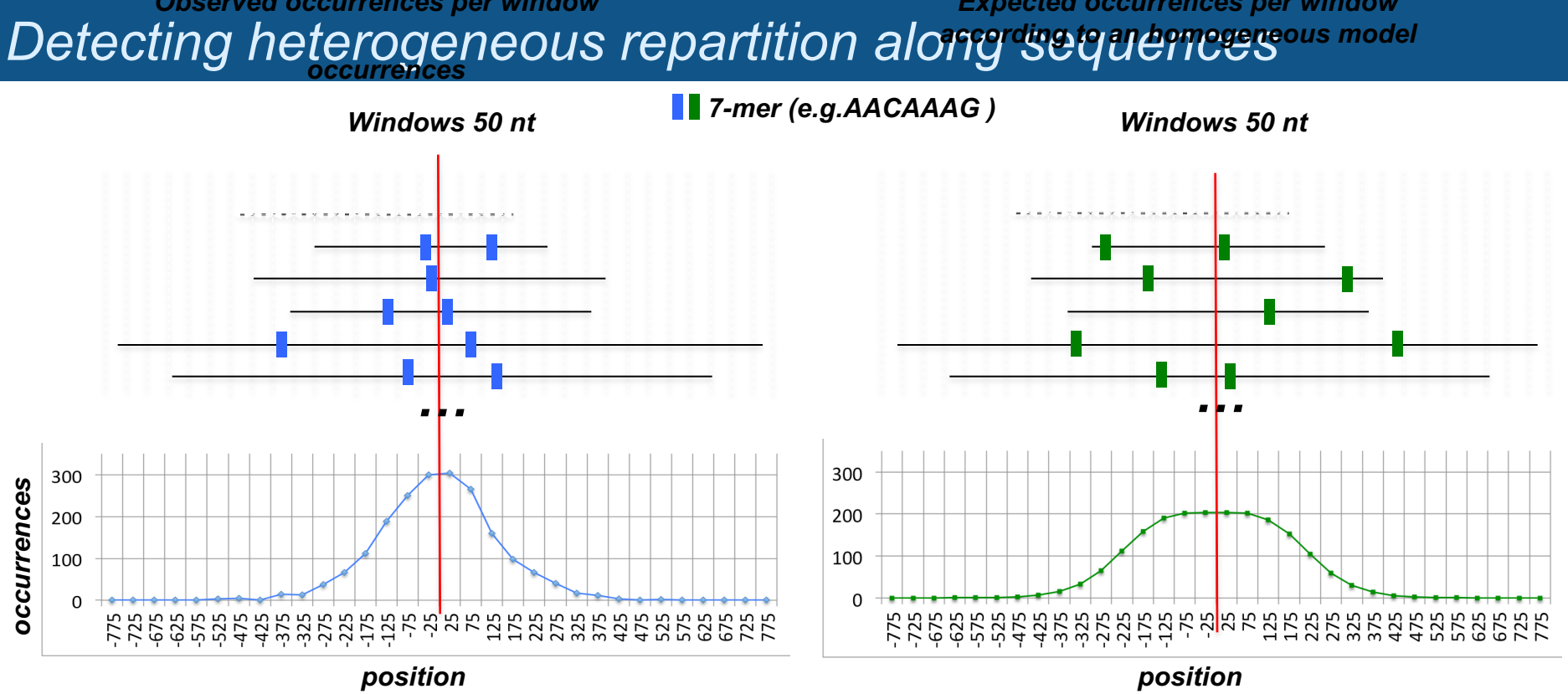
- Deterministic (not heuristic) and exhaustive
  - all possible words/dyads are tested
  - ability to return several patterns in a single run
- Speed
  - co-expression clusters are treated within seconds
- Time increases linearly with sequence set
  - Can be applied to very large sequence sets (full genomes)
  - Realistic application: ChIP-seq peaks generally cover several Mb or even tens of Mb. Such files are treated in a few minutes on a personal laptop.
- Ability to return a negative answer
  - "not a single over-represented pattern in this sequence set"
  - Corollary: very low false positive rate
- Ability to detect over-represented, but also under-represented motifs
  - (e.g. restriction sites in bacterial genomes)
- Pattern assembly refines the result
  - ability to detect some level of degeneracy
    (result contains words differing by single substitutions)
  - ability to detect motifs larger than the oligonucleotide size
    (result contains strongly overlapping words)

String-based pattern discovery: weaknesses

- No direct treatment of pattern degeneracy
  - NB: degenerated words can be analyzed with similar statistics, but it is not tractable due to the increase of the number of patterns: 15k possible words of length k.
- String patterns are poor descriptions for genome-scale pattern matching.
  - Matrices are more appropriate to describe the weight of each substitution at a given position.
- Solution
  - string-based approach for pattern discovery (RSAT programs *oligo-analysis, dyad-analysis*, *position-analysis*, *local-words*).
  - use discovered strings as seeds for building a matrix, which can be used for pattern search (RSAT program *matrix-from-patterns*)

Position-analysis

Detecting heterogeneous repartition along sequences



Drawing by Elodie Darbo

position-analysis method

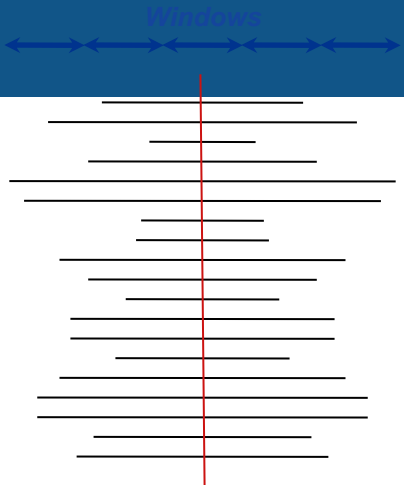
- *van Helden, J., del Olmo, M. and Perez-Ortin, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res 28, 1000-10.*

Detecting biases in word positions

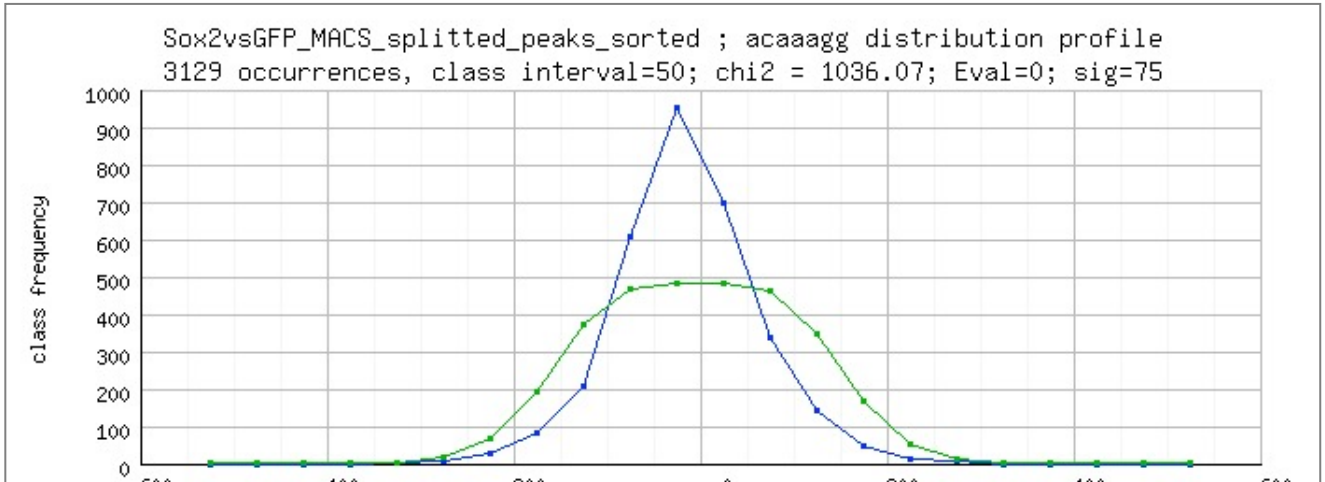
heterogeneous distribution of occurrences across a set of input sequences.

Principle: for each word

- Compute the number of occurrences in non-overlapping windows starting from a reference point (sequence start, center or end).
- Compute the expected occurrences in each window according to a homogeneous distribution model.
- Compute the difference between the observed and expected positional distribution (chi2 test for goodness of fit).



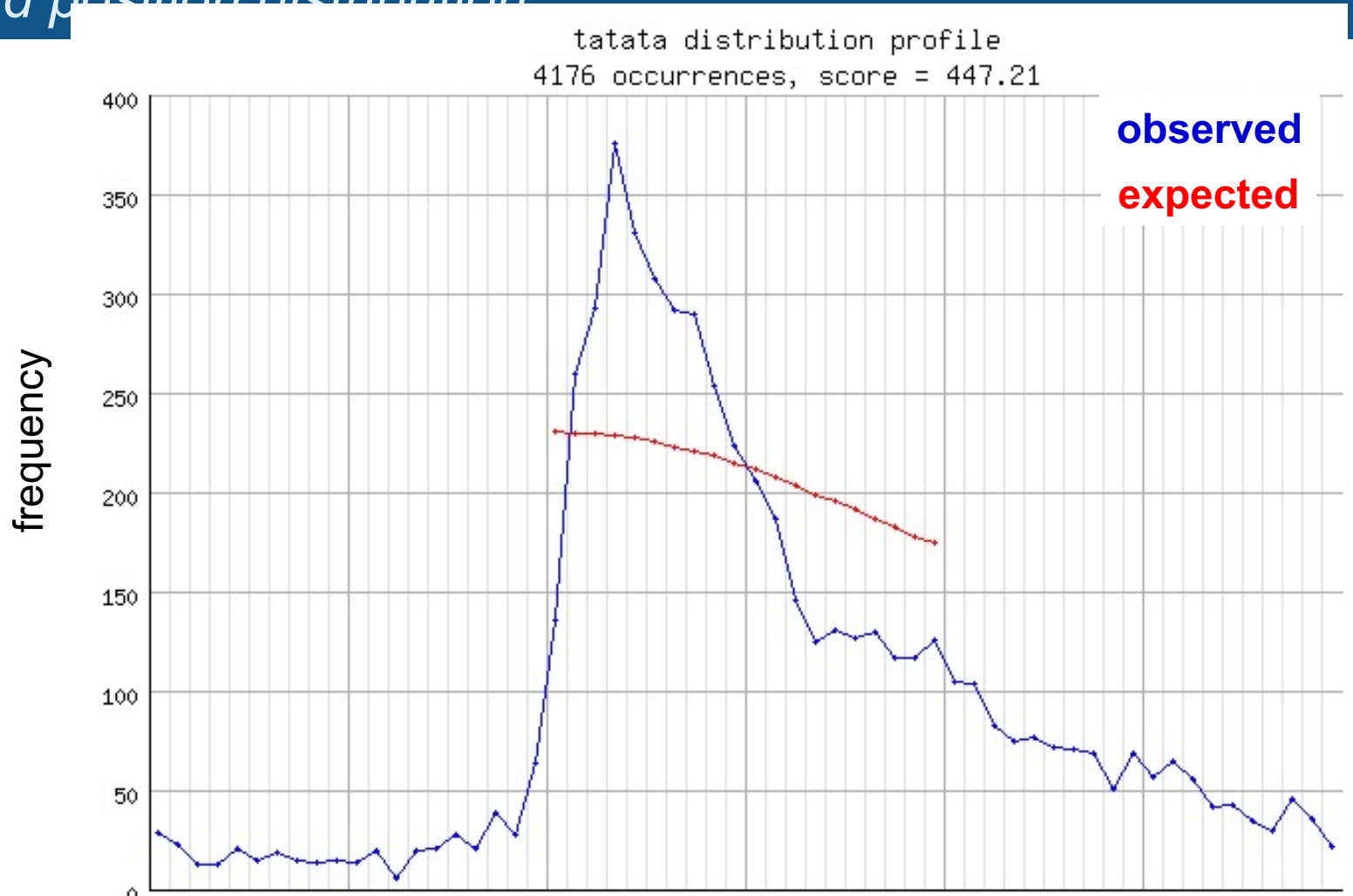
- Example: Sox2 peaks from Chen, 2008
  - 10,929 peaks of size between 60 and 1,059 bp
  - Word length k=7
  - Reference position: the center of each peak.
  - The most significant word is ACAAAGG, which corresponds to the Sox2 consensus.



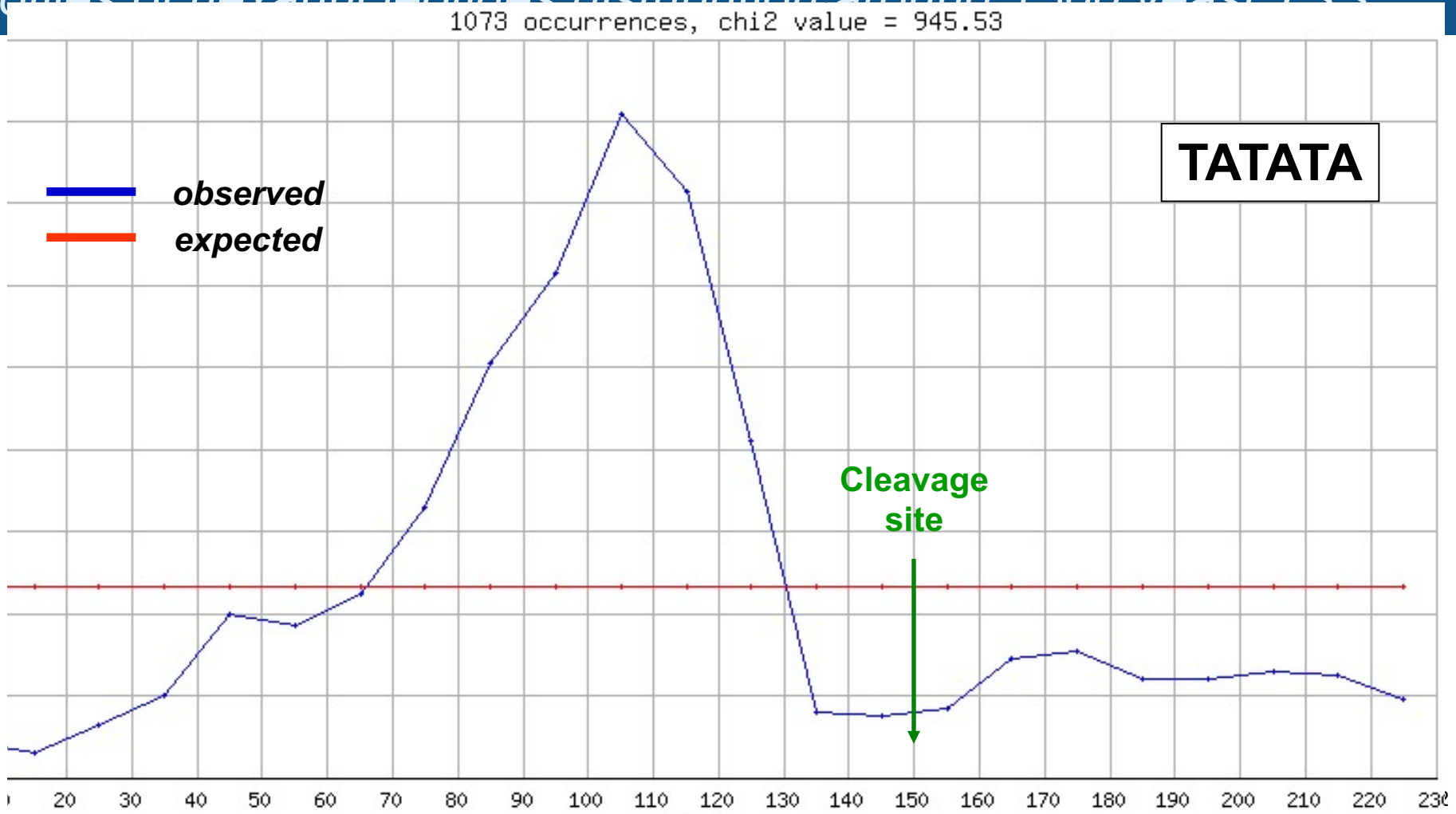
- ***Green: expected occurrences***
  - ***Note: the expectation decreases with the distance to peak center because peaks have variable lengths.***
- ***Blue: observed occurrences***
  - ***The word***

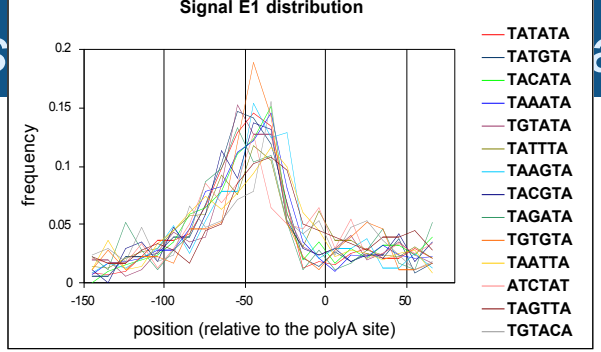
Positions relative to the stop codon

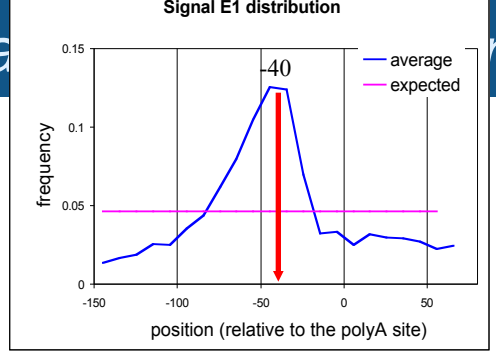
Word position distribution

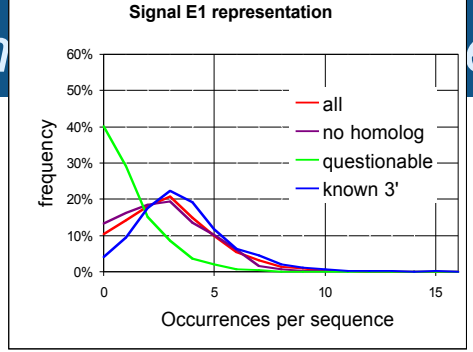


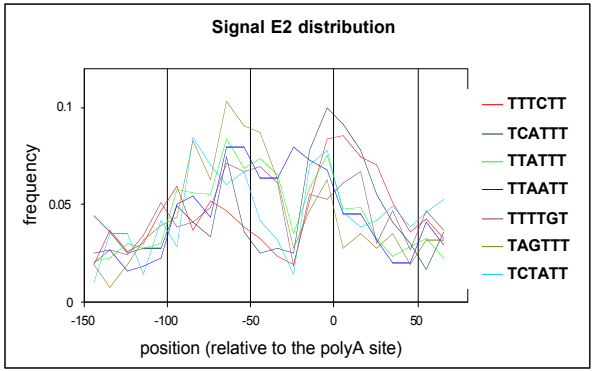
Profiles of hexanucleotides distribution around 1500 yeast TSS

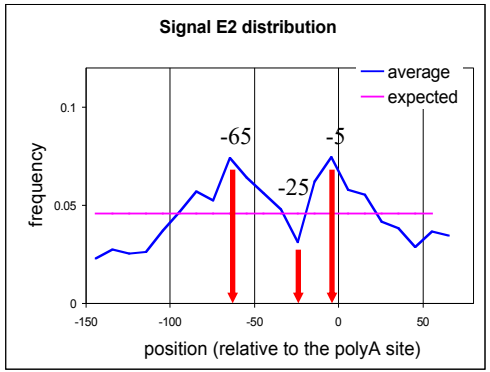


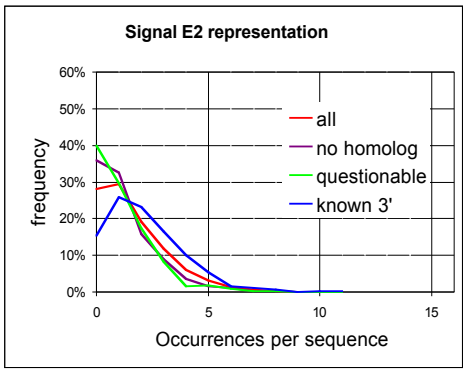

Signal E1 distribution
frequency
position (relative to the polyA site)
TATATA
TATGTA
TACATA
TAAATA
TGTATA
TATTTA
TAAGTA
TACGTA
TAGATA
TGTGTA
TAATTA
ATCTAT
TAGTTA
TGTACA

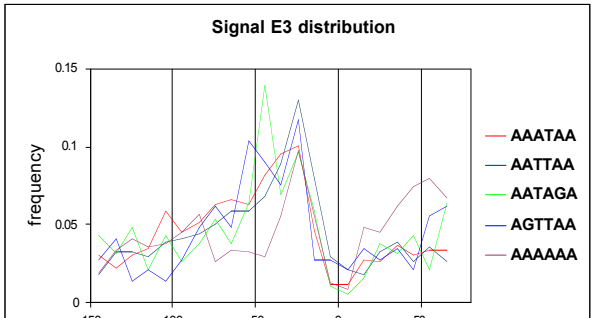

Signal E1 distribution
-40
average
expected
frequency
position (relative to the polyA site)

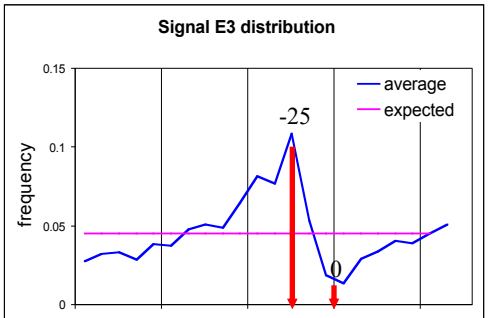

Signal E1 representation
all
no homolog
questionable
known 3'
frequency
Occurrences per sequence

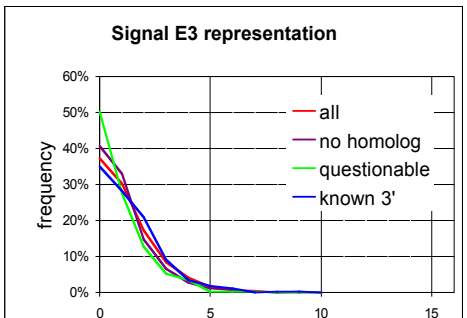

Signal E2 distribution
TTTCTT
TCATTT
TTATTT
TTAATT
TTTTGT
TAGTTT
TCTATT
frequency
position (relative to the polyA site)


Signal E2 distribution
-65
-25
-5
average
expected
frequency
position (relative to the polyA site)


Signal E2 representation
all
no homolog
questionable
known 3'
frequency
Occurrences per sequence

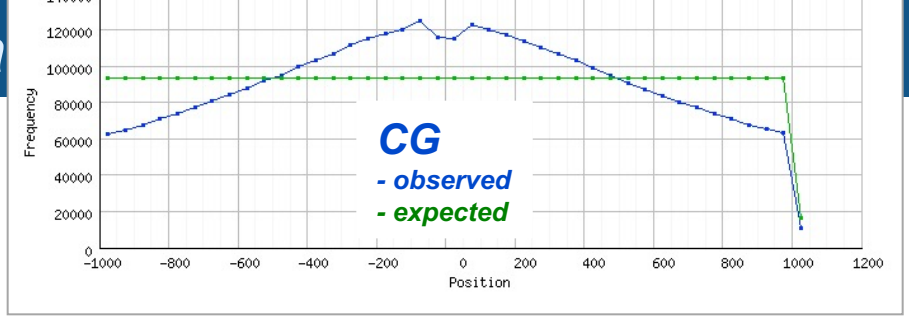

Signal E3 distribution
AAATAA
AATTAA
AATAGA
AGTTAA
AAAAAA
frequency

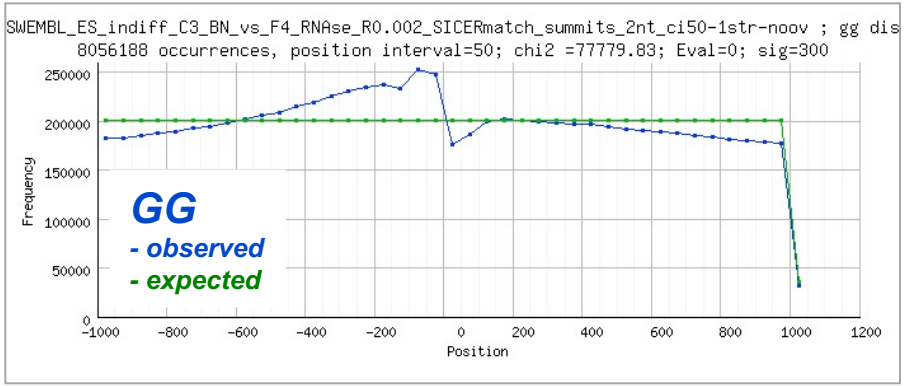

Signal E3 distribution
-25
0
average
expected
frequency

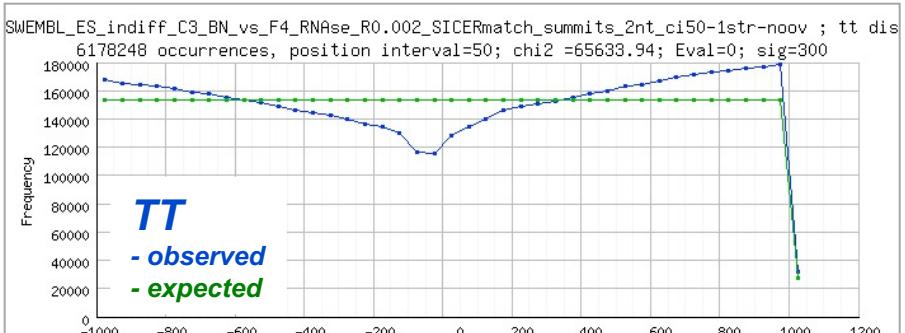

Signal E3 representation
all
no homolog
questionable
known 3'
frequency

Position analysis of dinucleotides a

65,009 peaks
- 2kb on each side of peak summits (130Mb analyzed in total)
- K-mer occurrences per 50bp windows
- **Background model: homogeneous distribution**
- Significance computed with Chi-square conformity test.
- Result: **all** dinucleotides are completely biased, with p-values < 1e-300.







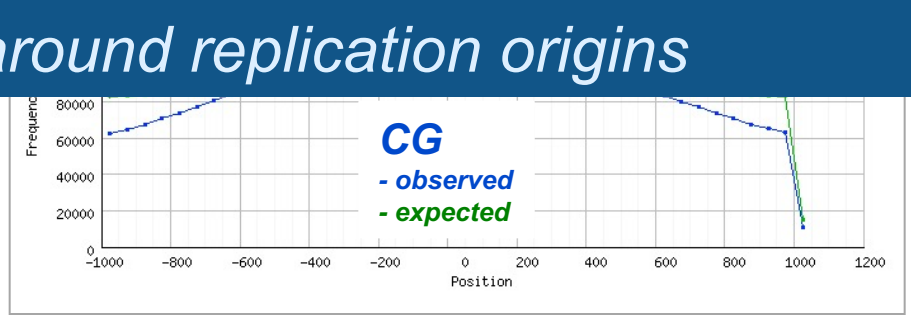
Sequence	ID	Occ	Overlaps	Chi2	df	Pval	Eval	Sig	Rank
cg	cg	3748310	0	164650.8	40	0.0e+00	0	300.0	1
cc	cc	8078476	2609220	78471.5	40	0.0e+00	0	300.0	2
gg	gg	8056188	2595227	77779.8	40	0.0e+00	0	300.0	3
ta	ta	5242474	0	72304.9	40	0.0e+00	0	300.0	4
aa	aa	6153023	2033169	66112.4	40	0.0e+00	0	300.0	5
tt	tt	6178248	2048441	65633.9	40	0.0e+00	0	300.0	6
gc	gc	8512412	0	64740.0	40	0.0e+00	0	300.0	7
at	at	6039429	0	58647.8	40	0.0e+00	0	300.0	8
tc	tc	8137303	0	26051.4	40	0.0e+00	0	300.0	9
ga	ga	8101277	0	25343.6	40	0.0e+00	0	300.0	10
ag	ag	10092541	0	21823.5	40	0.0e+00	0	300.0	11
ct	ct	10113605	0	21797.1	40	0.0e+00	0	300.0	12
ac	ac	6833408	0	15129.5	40	0.0e+00	0	300.0	13

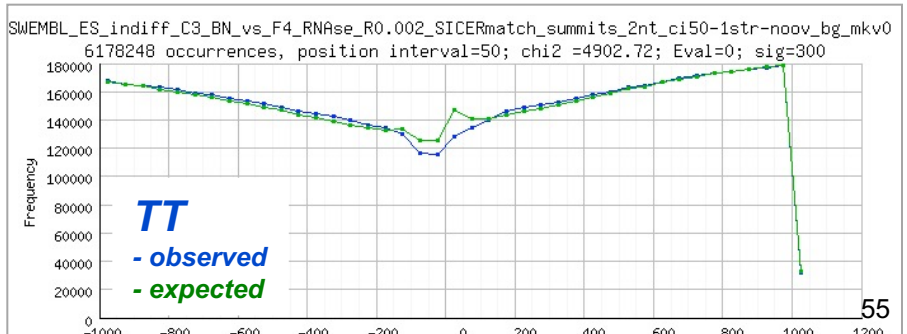
Position analysis of dinucleotides around replication origins

65,009 peaks
2kb on each side of peak summits (130Mb analyzed in total)

- K-mer occurrences per 50bp windows
- **Background model: window-specific estimation based on nucleotide composition.**
- Significance computed with Chi-square conformity test.
- Result: **all** dinucleotides are completely biased, with p-values < 1e-300.

Sequence	ID	Occ	Overlaps	Chi2	df	Pval	Eval	Sig	Rank
cg	cg	3748310	0	67200.5	40	0.0e+00	0	300.0	1
ac	ac	6833408	0	10173.3	40	0.0e+00	0	300.0	2
gt	gt	6841055	0	10001.6	40	0.0e+00	0	300.0	3
ca	ca	9621040	0	8646.5	40	0.0e+00	0	300.0	4
tg	tg	9613852	0	8508.0	40	0.0e+00	0	300.0	5
ta	ta	5242474	0	7453.5	40	0.0e+00	0	300.0	6
tt	tt	6178248	2048441	4902.7	40	0.0e+00	0	300.0	7
aa	aa	6153023	2033169	4628.1	40	0.0e+00	0	300.0	8
gg	gg	8056188	2595227	3843.9	40	0.0e+00	0	300.0	9
cc	cc	8078476	2609220	3773.6	40	0.0e+00	0	300.0	10
at	at	6039429	0	1763.3	40	0.0e+00	0	300.0	11
gc	gc	8512412	0	1447.1	40	0.0e+00	0	300.0	12
ag	ag	10092541	0	1392.3	40	0.0e+00	0	300.0	13

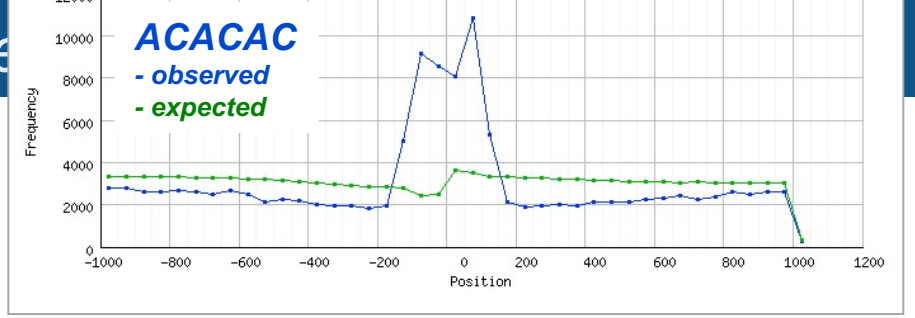


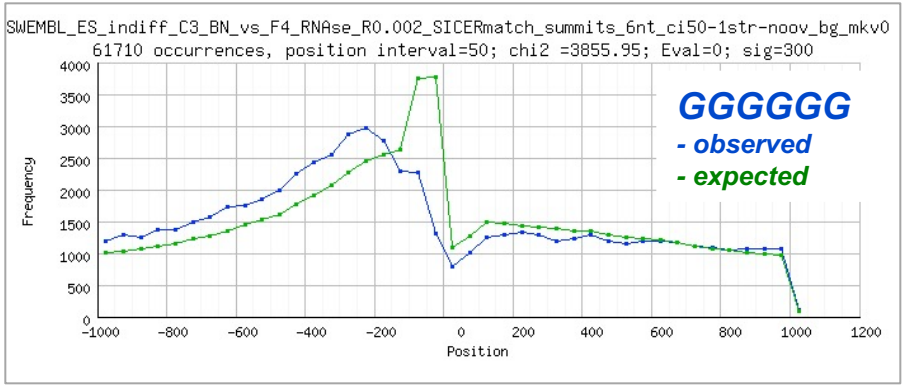


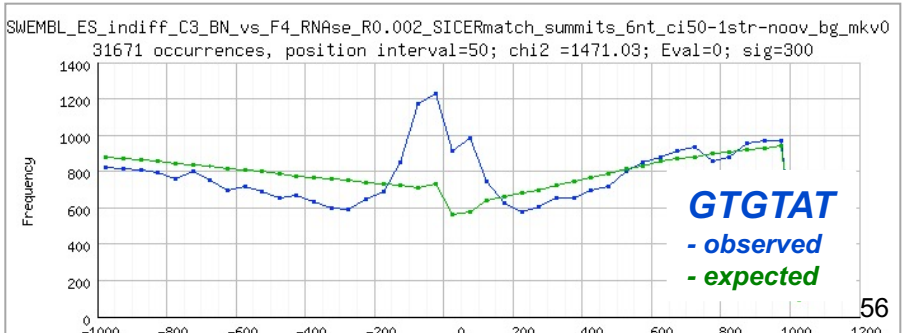
Position analysis of hexanucleotide

- A lot of very highly significant 6-mers.
- Most of them are low-complexity motifs (periodic k-mers).

Sequence	ID	Occ	Overlaps	Chi2	df	Pval	Eval	Sig	Ran
acacac	acacac	125516	128532	65433.9	40	0.0e+00	0	300.0	1
gtgtgt	gtgtgt	125740	127945	62813.5	40	0.0e+00	0	300.0	2
cacaca	cacaca	143816	131933	57978.6	40	0.0e+00	0	300.0	3
tgtgtg	tgtgtg	143246	130948	56183.5	40	0.0e+00	0	300.0	4
gggggg	gggggg	61710	64984	3855.9	40	0.0e+00	0	300.0	5
cccccc	cccccc	61215	65010	3540.8	40	0.0e+00	0	300.0	6
tttttt	tttttt	73687	122406	2182.5	40	0.0e+00	0	300.0	7
aaaaaa	aaaaaa	73293	122290	2055.5	40	0.0e+00	0	300.0	8
gggagg	gggagg	141834	13752	1957.4	40	0.0e+00	0	300.0	9
tggggg	tggggg	100645	0	1936.2	40	0.0e+00	0	300.0	10
ggggga	ggggga	79273	0	1932.2	40	0.0e+00	0	300.0	11
tccccc	tccccc	80747	0	1923.7	40	0.0e+00	0	300.0	12
ggaggg	ggaggg	127623	13258	1919.7	40	0.0e+00	0	300.0	13
ccccca	ccccca	101109	0	1872.0	40	0.0e+00	0	300.0	14
tgtgta	tgtgta	49360	0	1850.7	40	0.0e+00	0	300.0	15
cctccc	cctccc	142300	13511	1823.3	40	0.0e+00	0	300.0	16
atgtgt	atgtgt	51630	0	1812.9	40	0.0e+00	0	300.0	17
gggtgg	gggtgg	104120	6751	1799.7	40	0.0e+00	0	300.0	18
acacat	acacat	51430	0	1799.4	40	0.0e+00	0	300.0	19
tacaca	tacaca	48626	0	1782.6	40	0.0e+00	0	300.0	20
ccctcc	ccctcc	128878	12851	1755.2	40	0.0e+00	0	300.0	21
ccaggg	ccaggg	93739	0	1706.5	40	0.0e+00	0	300.0	22







Examples of applications

family	oligo-analysis				dyad-analysis (non-coding dyad frequency calibration)			
	word	reverse clpt	sig	remark	dyad	reverse clpt	sig	remark
CLN2	TACGCGAA.	.TTCGCGTA	30.5	MBF; SBF variant	TTTACGCGAAAA	TTTTCGCGTAAA	29.0	MBF; SBF variant
	TACGCGTA.	.TACGCGTA	30.5	MBF; SBF	GAAAACGCGTAAA	TTTACGCGTTTC	29.0	MBF; SBF
	TTCGCGTCG	CGACGCGAA	30.5	MBF; SBF variant	TTTTCGCGTCA.	.TGACGCGAAAA	29.0	MBF; SBF variant
	AAACGCGAA.	.TTCGCGTTT	30.5	MBF; SBF variant	TTTACGCGTCA.	.TGACGCGTAAA	29.0	MBF; SBF
	TTCGCGTCA.	TGACGCGAA	30.5	MBF; SBF variant	CGACGCGAAAA	TTTTCGCGTCG	29.0	MBF; SBF variant
	TGCCAA	TTGGCA	1.8		GAAAACGCGTCA.	.TGACGCGTTTC	8.1	MBF; SBF
	ATCAAG	CTTGAT	1.3		AAAn8CGC	GCGn8TTT	1.9	
					CAAn5CGC	GCGn5TTG	1.1	
Y' (purged)	CTCGTC	GACGAG	1.8		AGTnGAG	CTCnACT	3.0	
	AGTATC	GATACT	1.2		CAGn{10}ATC	GATn{10}CTG	2.0	
					ATCn{12}GAG	CTCn{12}GAT	1.2	
histone (purged)	CGCCCG	CGGGCG	2.6		GCGn8AGAAC	GTTCTn8CGC	3.0	
	CCAGAA	TTCTGG	1.7	Mcm1	CGCCCG	CGGGCG	1.3	
					ATTn2GCG	CGCn2AAT	1.3	
Cell cycle	TGCCACAGTT	AACTGTGGCA	10.1	Met31; Met32	GCCACAGTT	AACTGTGGC	8.6	Met31; Met32
MET	TCACGTGA	TCACGTGA	10.1	Met4/Met28/Cbf1	GTCACGTGAC	GTCACGTGAC	6.9	Met4/Met28/Cbf1
	ACAGAG	CTCTGT	1.9					
	GACTCA	TGAGTC	0.9					
CLB2	CCAAAG	CTTTGG	1.3		CCCn6GAA	TTCn6GGG	2.5	ECB
	CCTTCA	TGAAGG	0.9	NEG	CAAn13GCC	GGCn13TTG	0.9	
					ACCn14AAT	ATTn14GGT	0.9	
MCM	AGAGCA	TGCTCT	1.4		TCCCn4GGGA	TCCCn4GGGA	3.9	ECB variant
	TCCTAA	TTAGGA	1.0	Mcm1	AAAnAGG	CCTnTTT	2.8	ECB ?
					AGGn10ACT	AGTn10CCT	1.2	
SIC1	AACCAGCAA	TTGCTGGTT	20.0	Swi5;Ace2	AACCAGCA.	. . . .TGCTGGTT	20.0	Swi5; Ace2

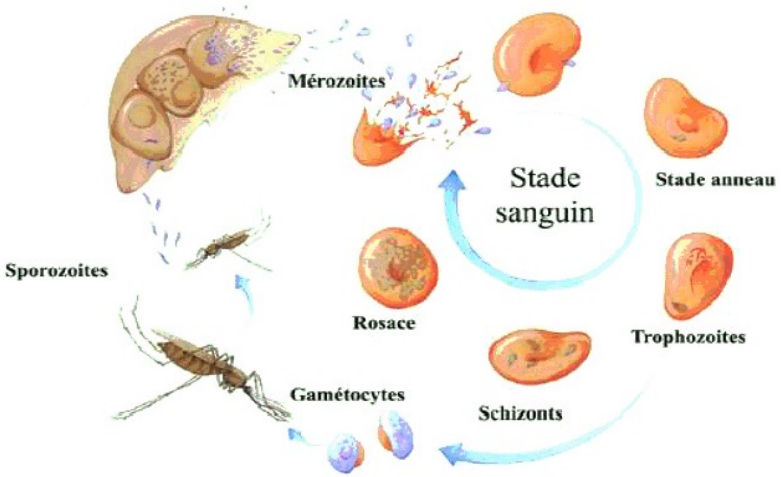
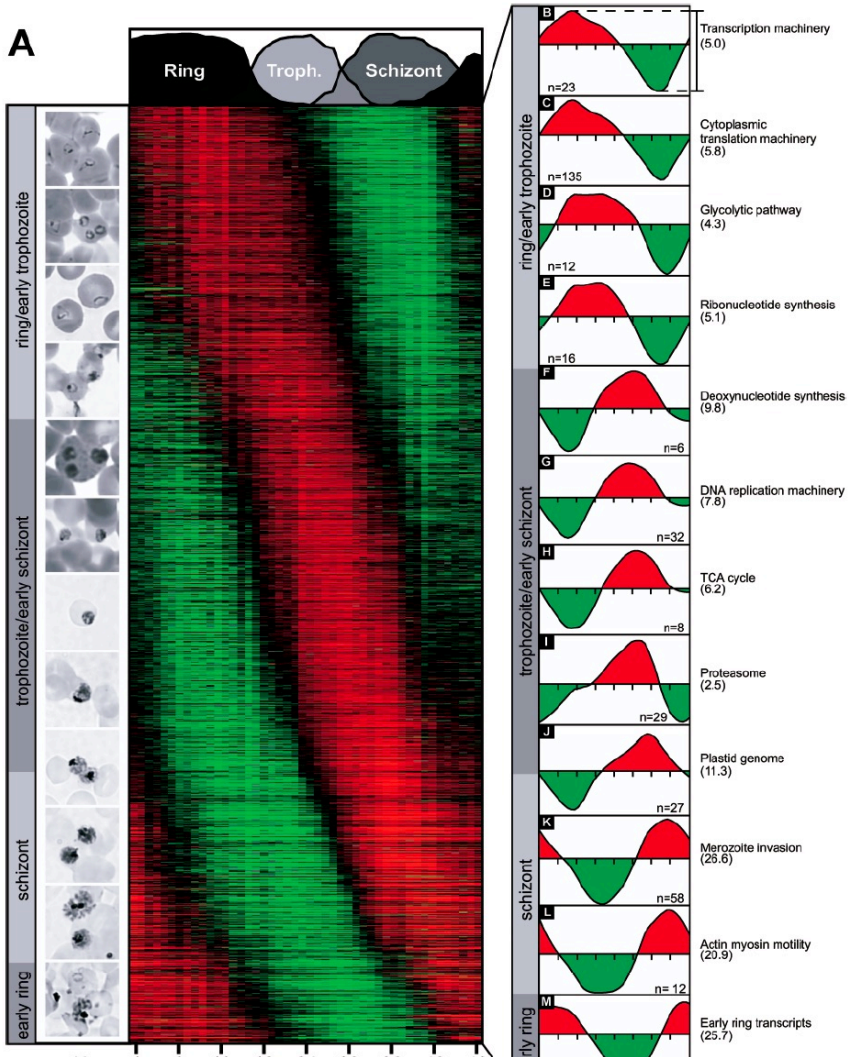
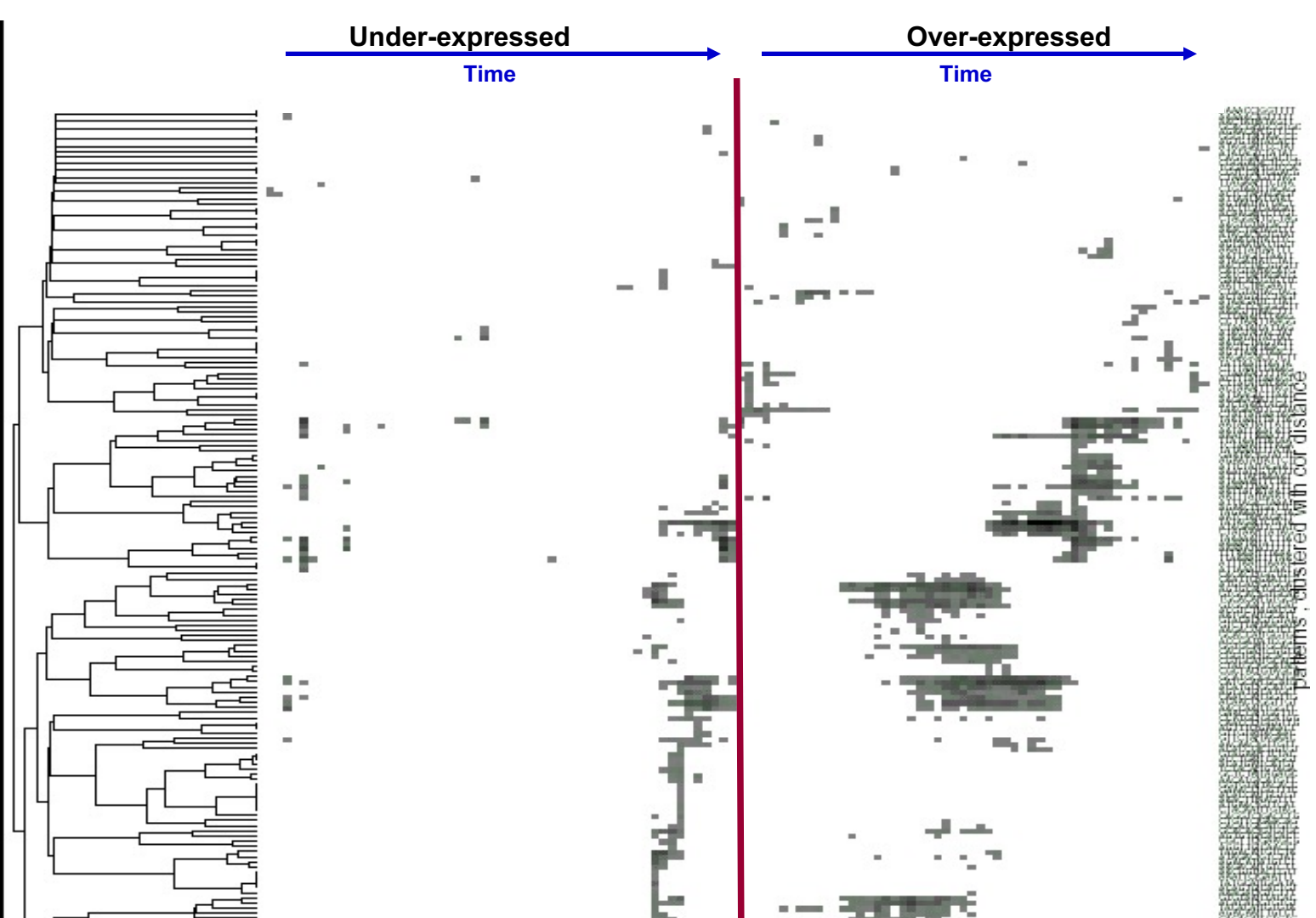
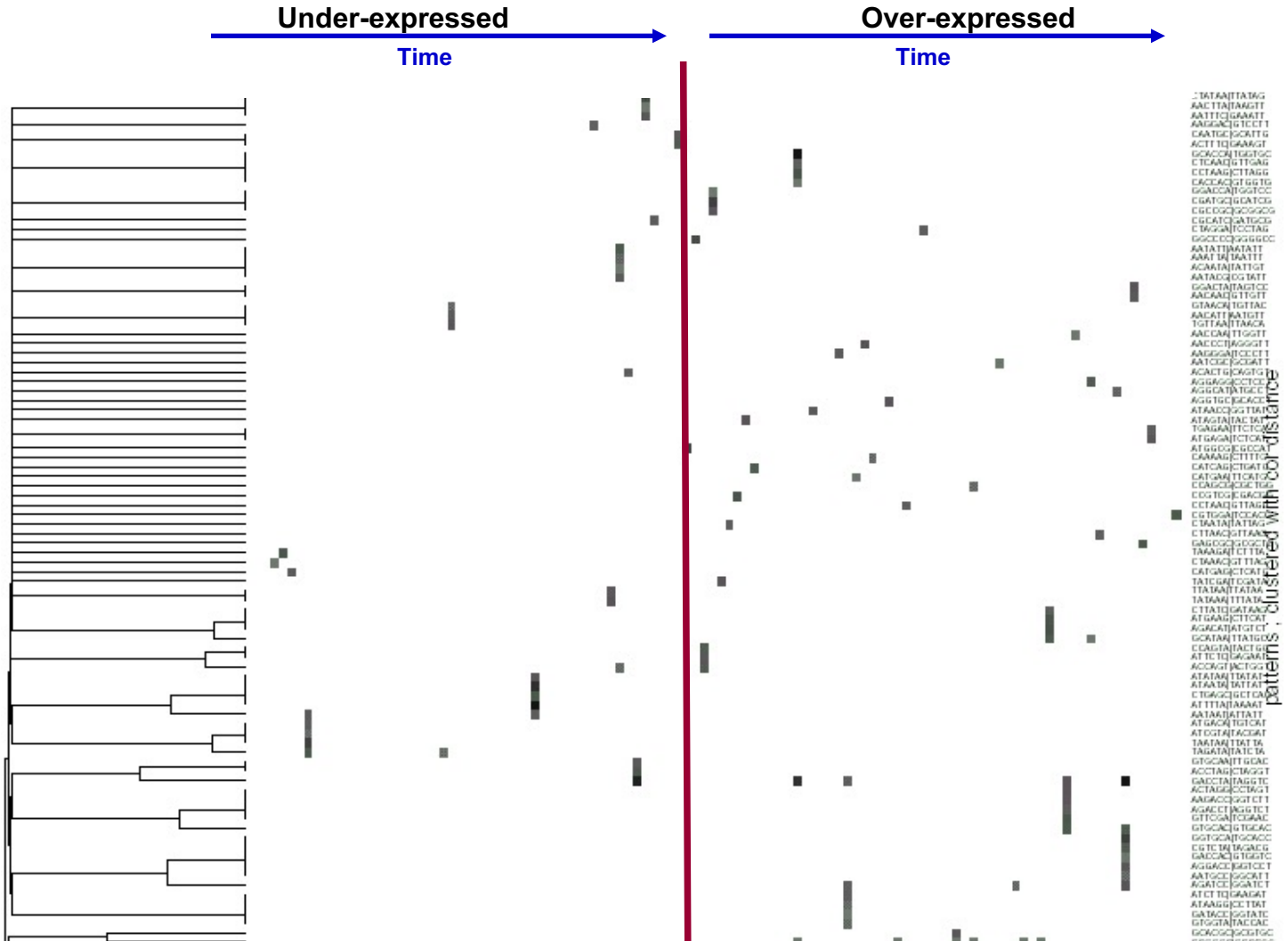


figure 2: Cycle de vie de Plasmodium falciparum (source :institut pasteur (France) page web)




Under-expressed
Time
Over-expressed
Time
patterns, clustered with cor distance

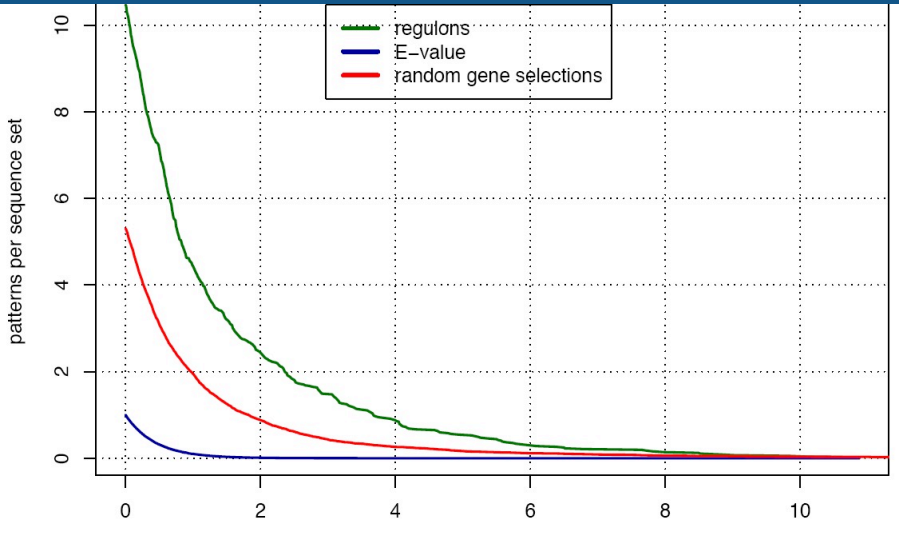

Under-expressed
Time
Over-expressed
Time
patterns , clustered with correlation distance

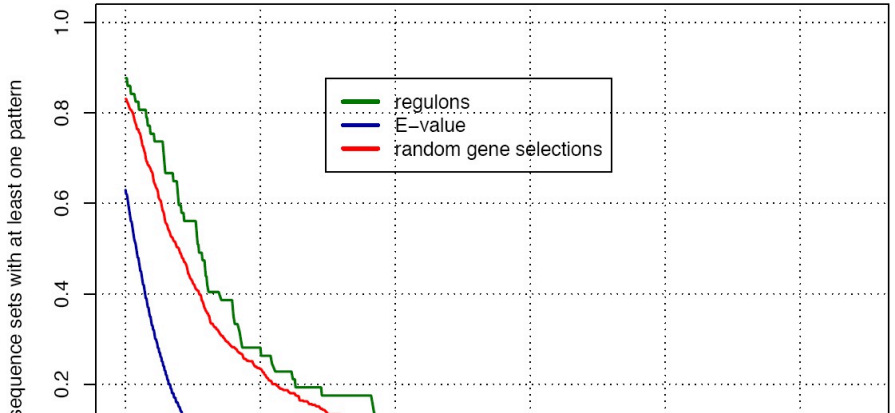
Supplementary material

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/

- Count occurrences observed for each word
- Calculate expected word frequencies
  - Choice of a model :
    - independently distributed nucleotides
      (equiprobable or biased alphabet utilization)
    - Markov chain : on basis of subword frequencies
    - External reference (e.g. word frequencies observed in the whole set of upstream sequences)
- Calculate a score for each word
    - obs/exp ratio (very bad)
    - log-likelihood
    - Z-value
    - binomial probability
- Select all words above a defined threshold
  - Statistical criterion for establishing the threshold

Pattern significance in regulons - Homo sapiens





in Homo sapiens

- The rate of false positive is much higher than the theoretical expectation
- The number of patterns detected in regulons is still higher, but the significance score is quite inefficient to distinguish between reliable motifs and false positives.
- This indicates that the background model is inadequate to treat the complexity of human promoters.