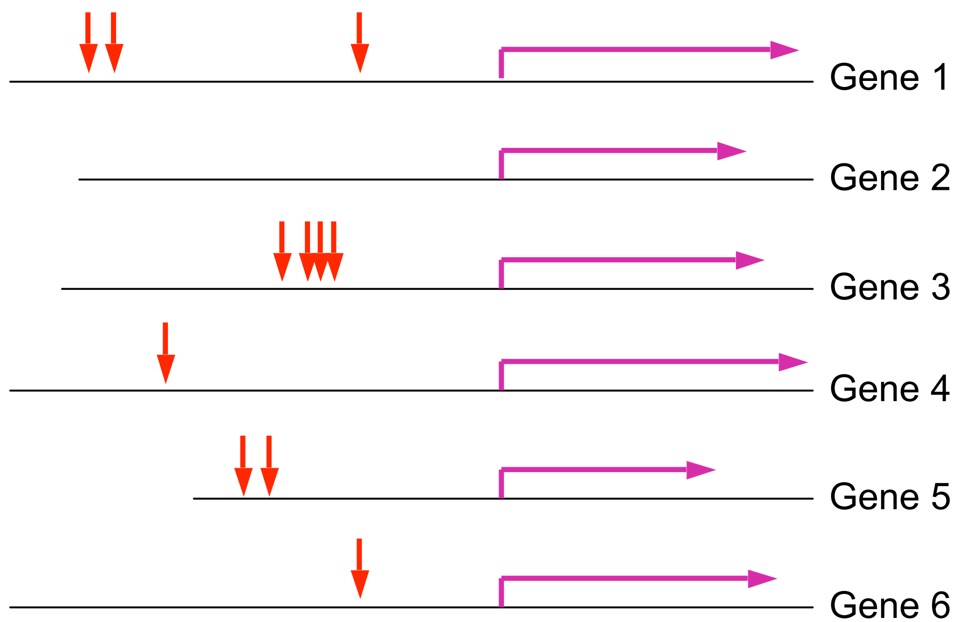
Regulatory Sequence Analysis

Pattern discovery
String-based approaches

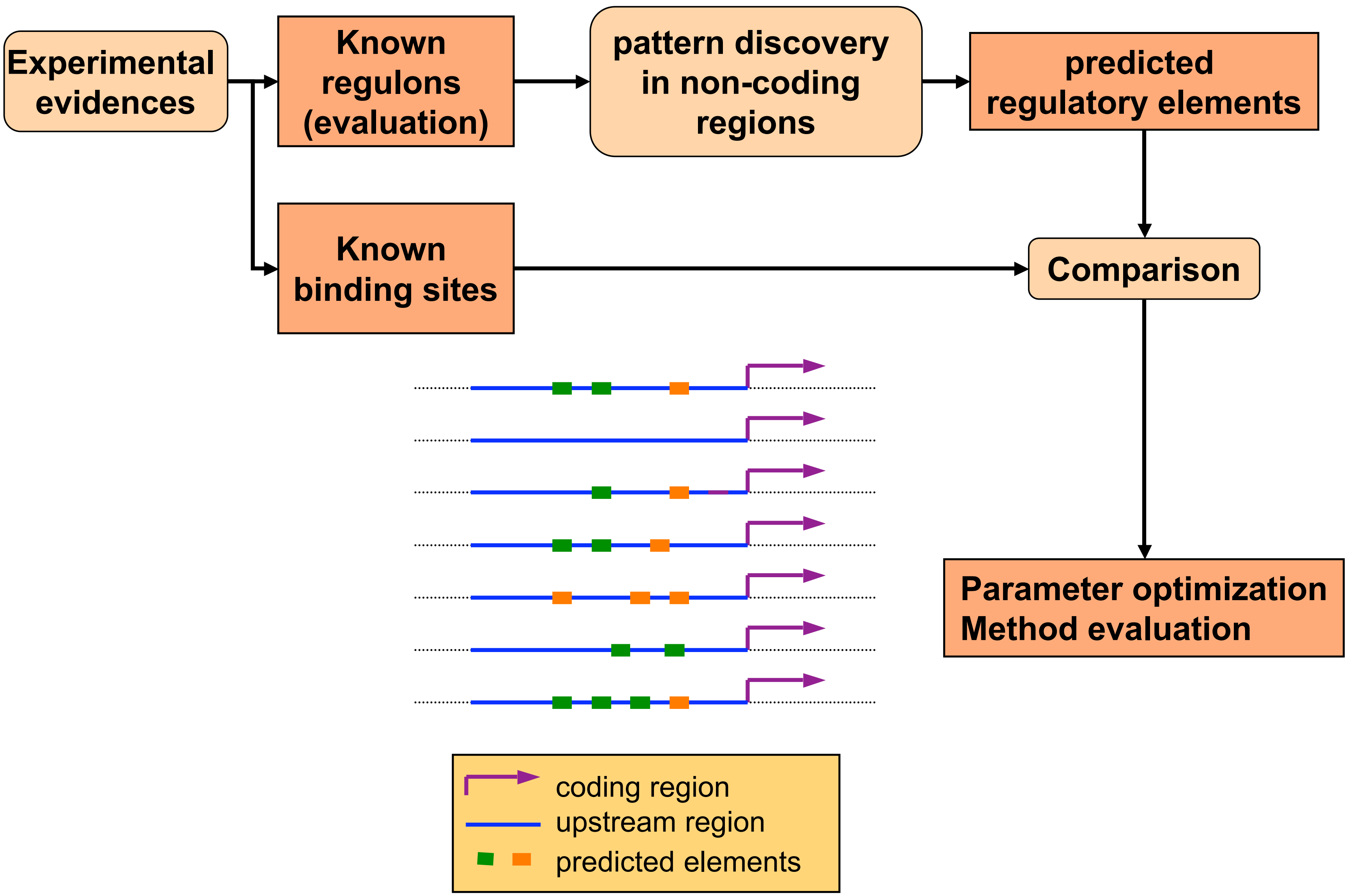
Jacques van Helden
Jacques.van.Helden@ulb.ac.be

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
- ...

Pattern discovery: string-based algorithms

- 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

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, calibrated 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.

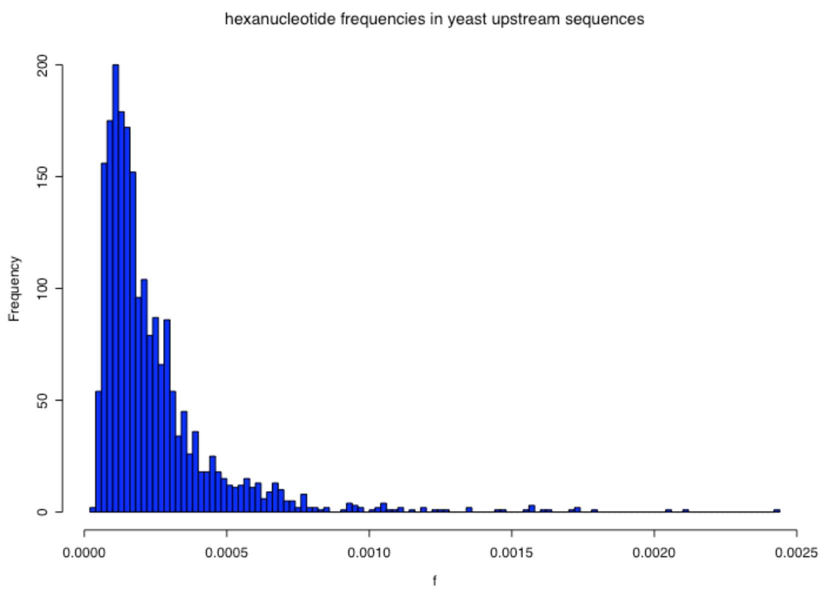
PHO		MET		NIT		GAL	
aaaaaa|tttttt	51	aaaaaa|tttttt	105	aaaaaa|tttttt	80	aaaaaa|tttttt	47
aaaaag|cttttt	15	atatat|atatat	41	cttatc|gataag	26	aaaaat|attttt	17
aagaaa|tttctt	14	gaaaaa|tttttc	40	tatata|tatata	22	aatata|tatatt	17
gaaaaa|tttttc	13	tatata|tatata	40	ataaga|tcttat	20	aaaatt|aatttt	16
tgccaa|ttggca	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|ttttttg	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

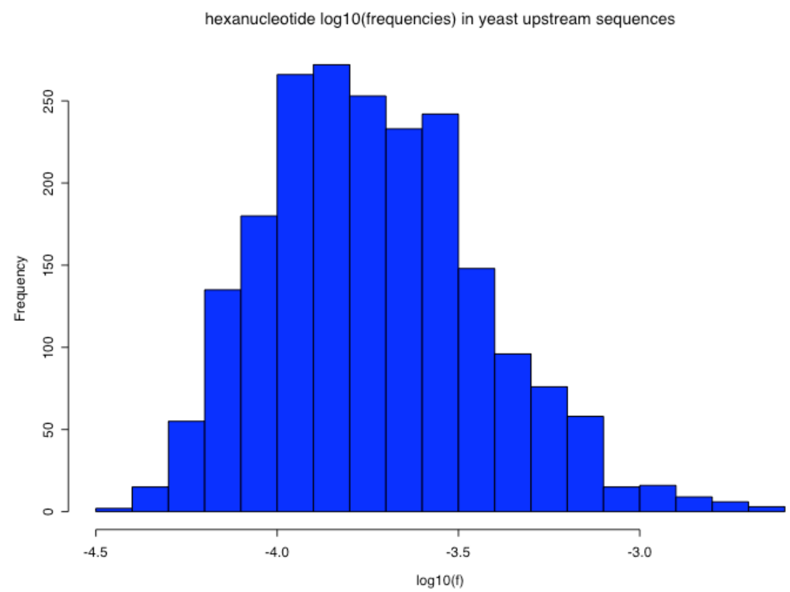
A more relevant criterion for over-representation

- 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).

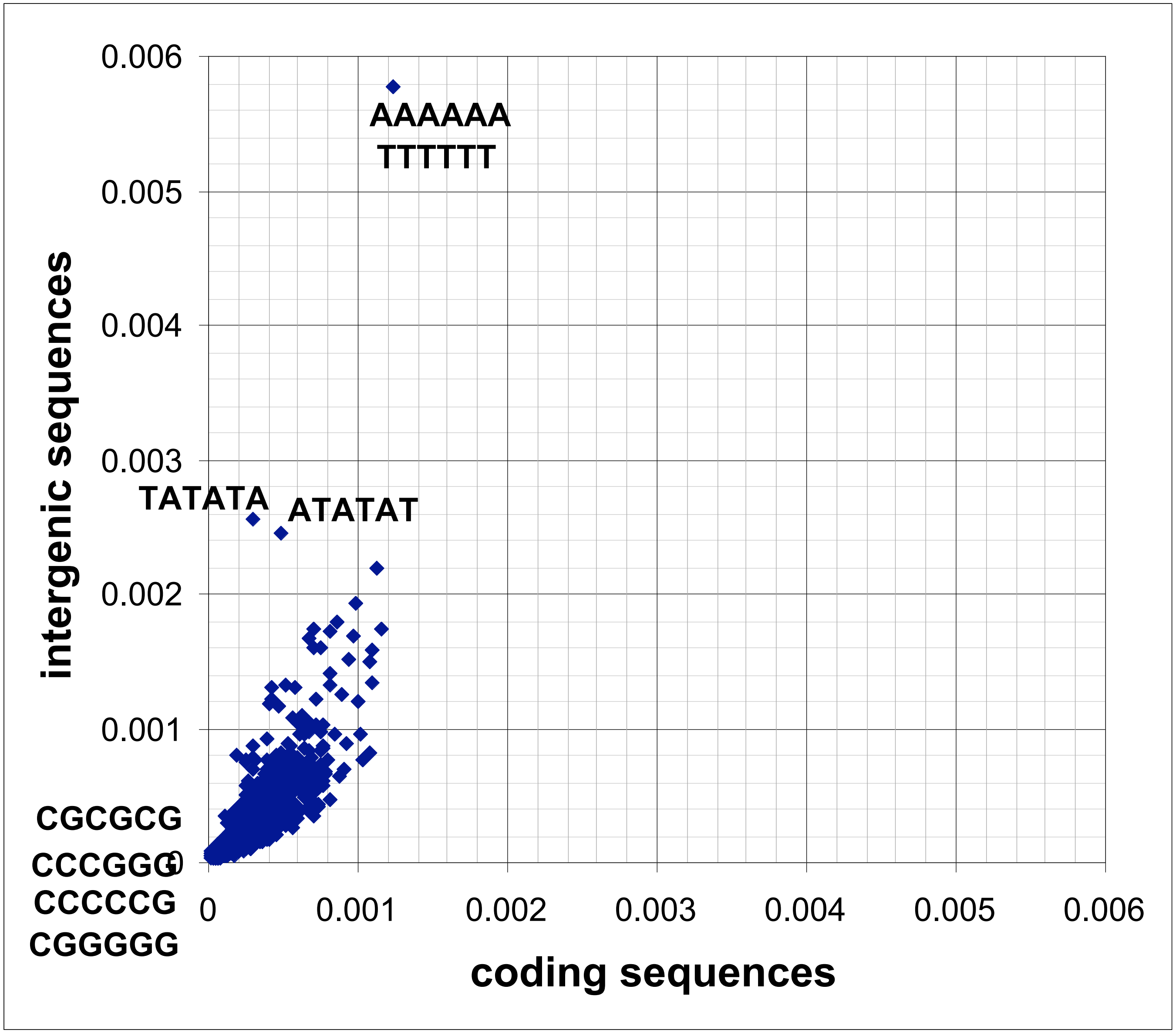
Hexanucleotide frequencies in all upstream sequences

- Hexanucleotide frequencies were measured in the whole set of 6000 yeast upstream sequences
  - range 4.5E-5 to 1.2E-2
  - max(f)/min(f)=268

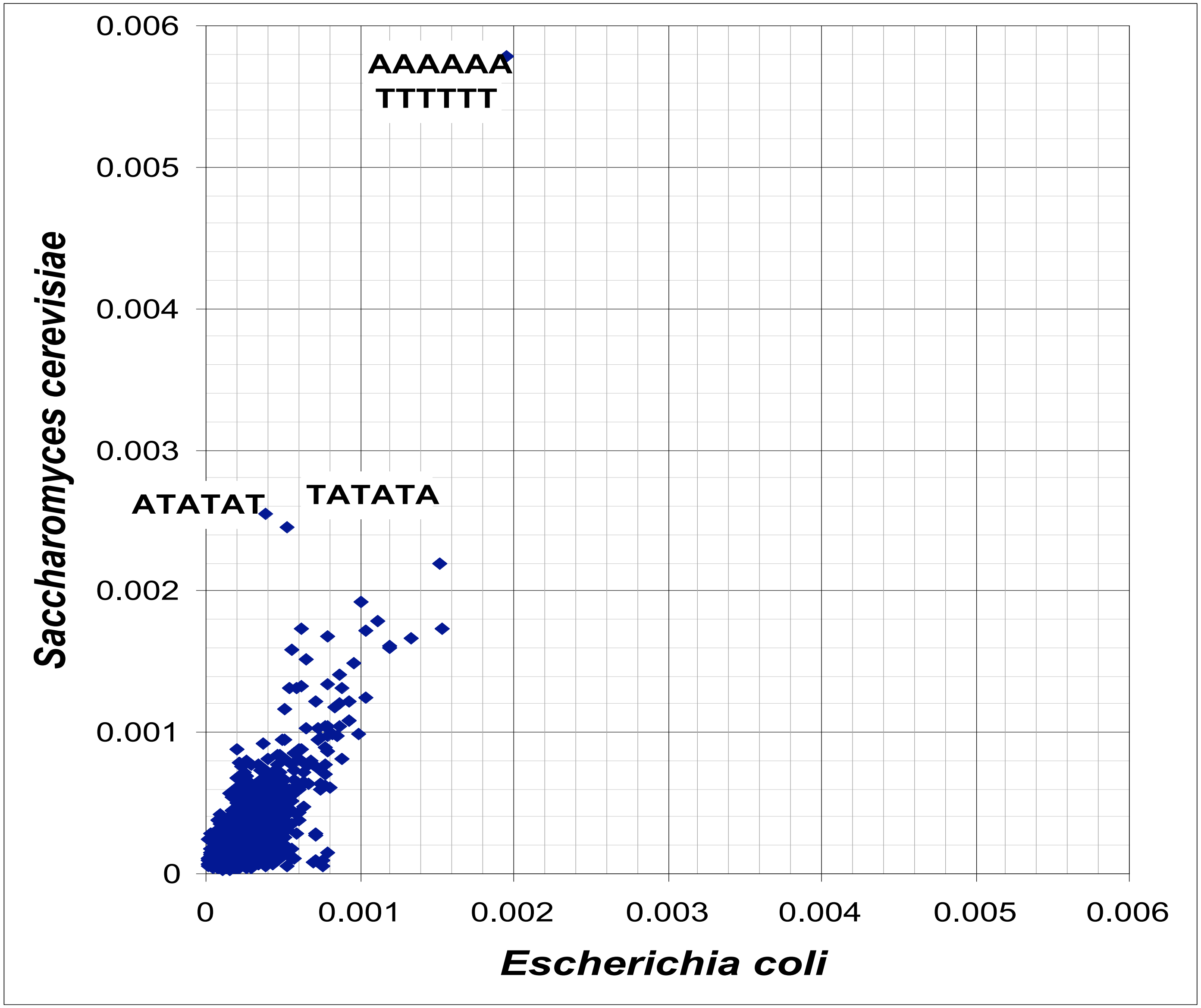


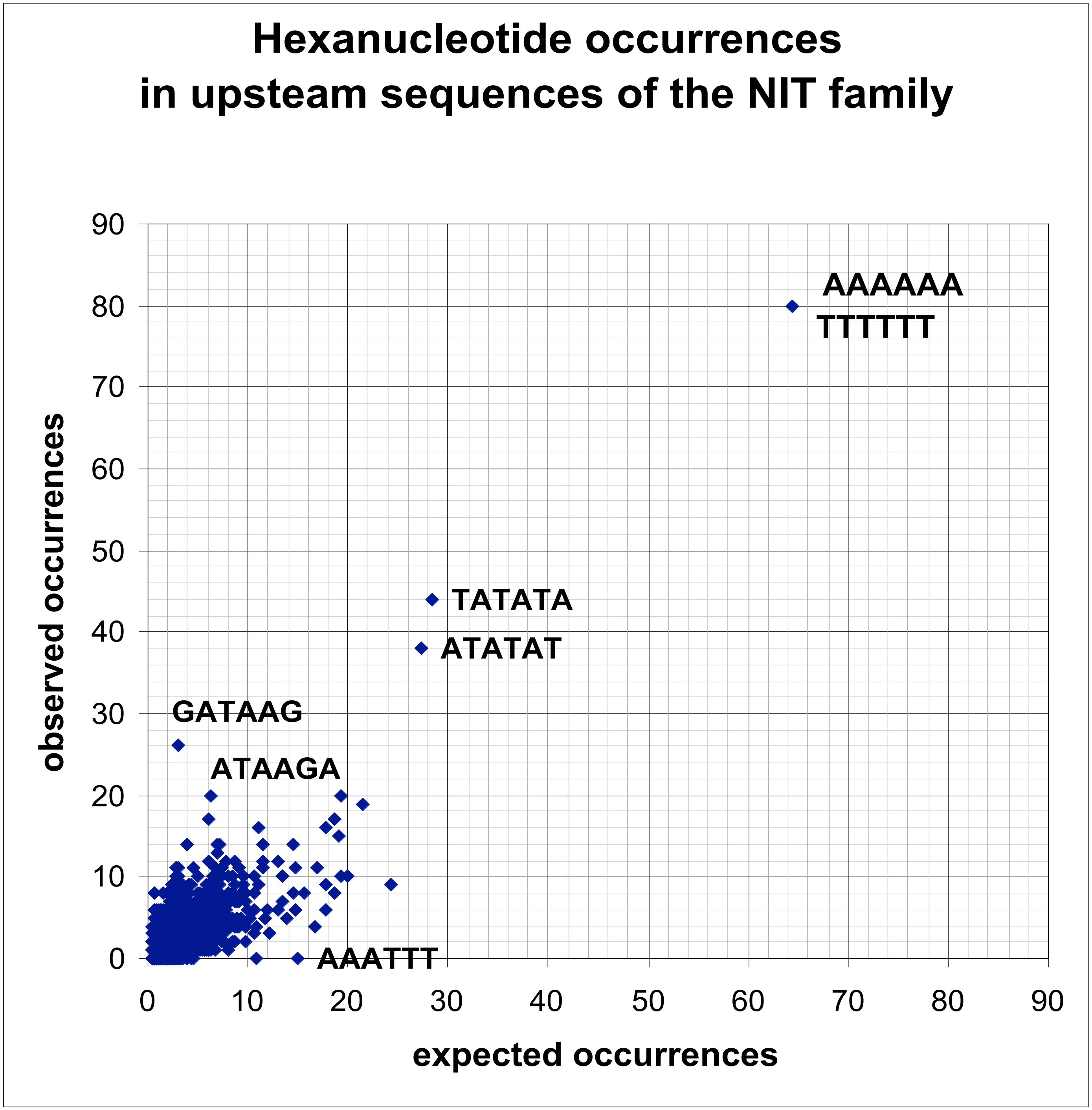


6nt frequencies differ between coding and non-coding sequences

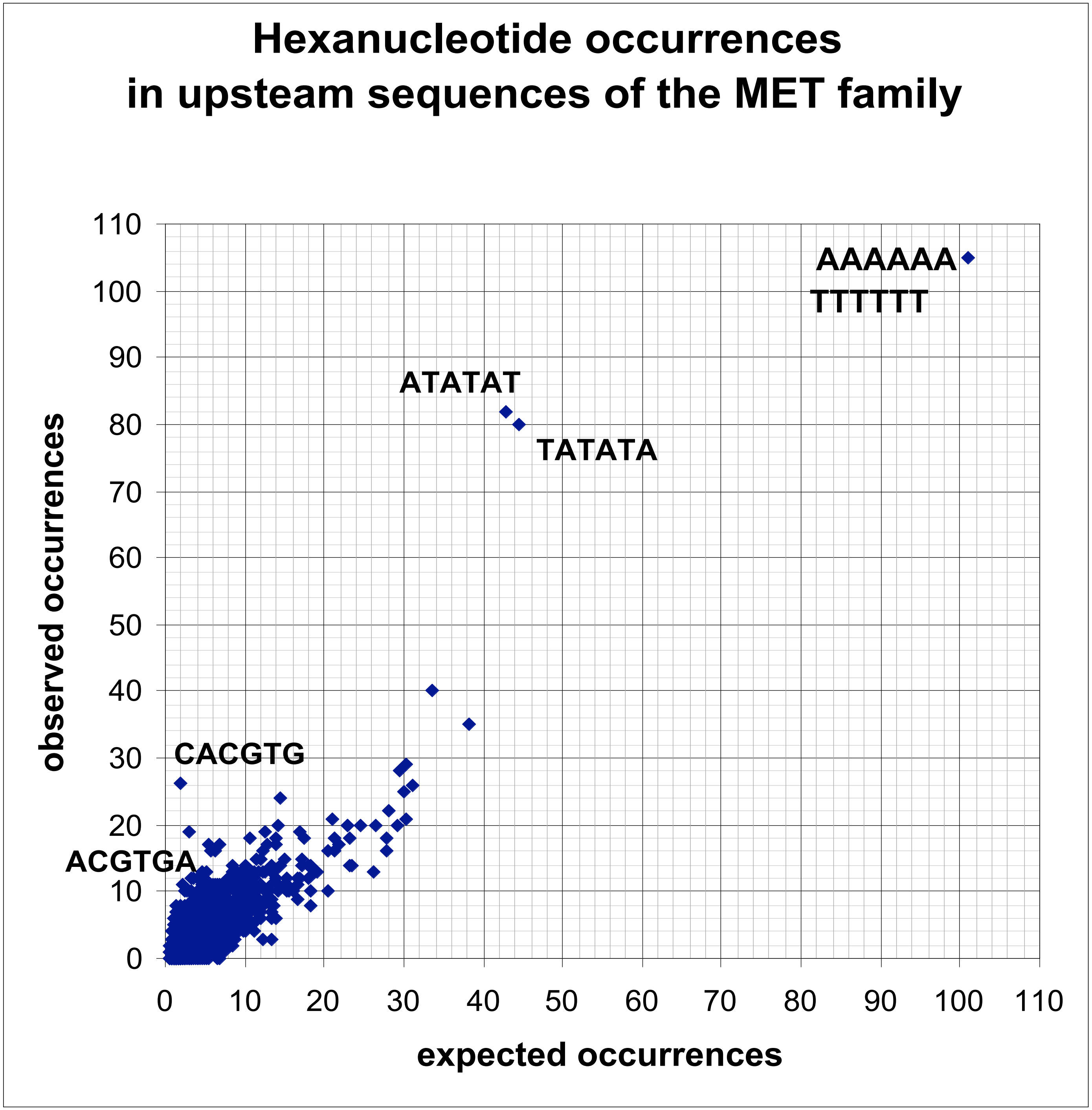


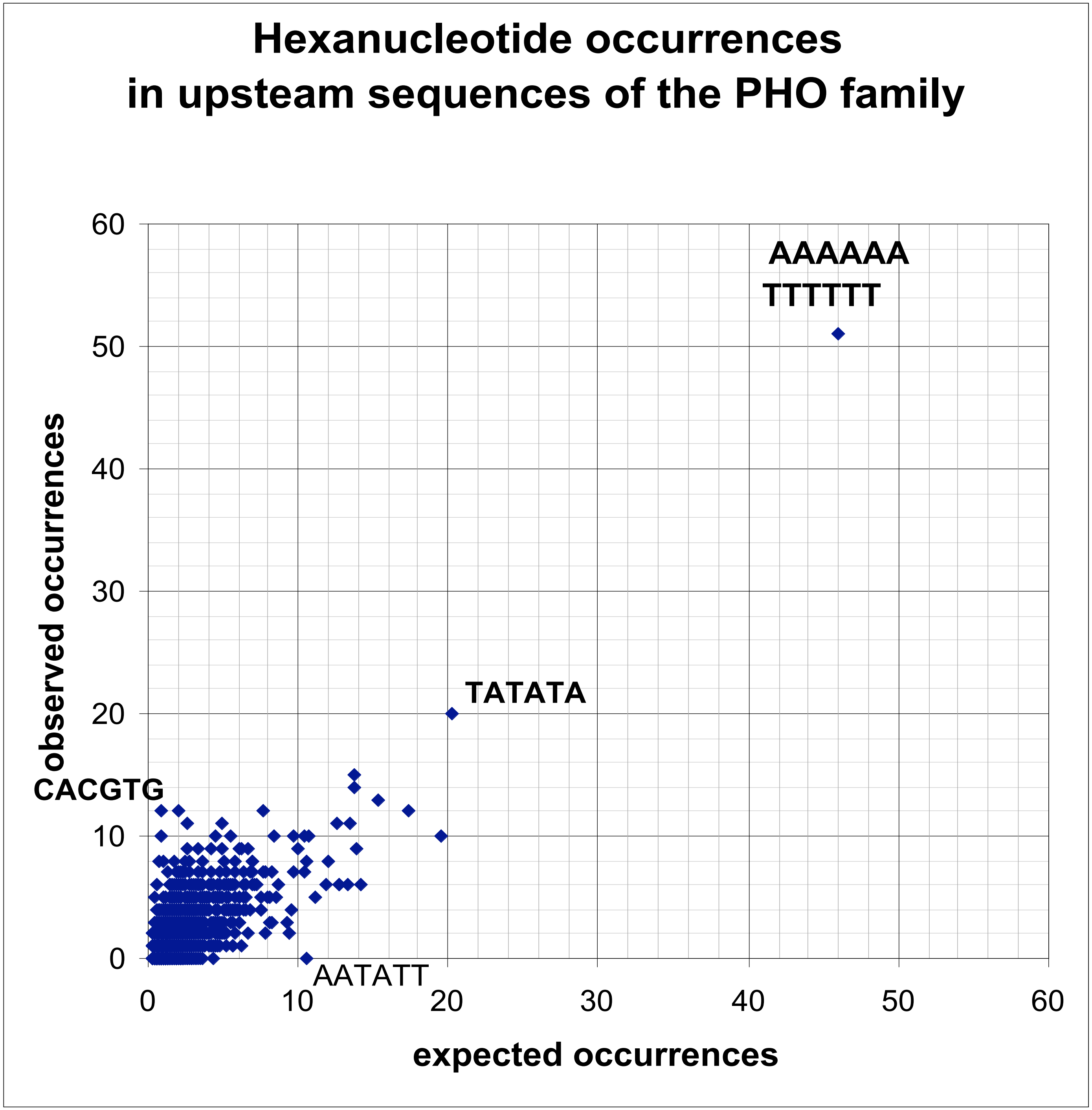
Inter-species variations in intergenic 6nt frequencies



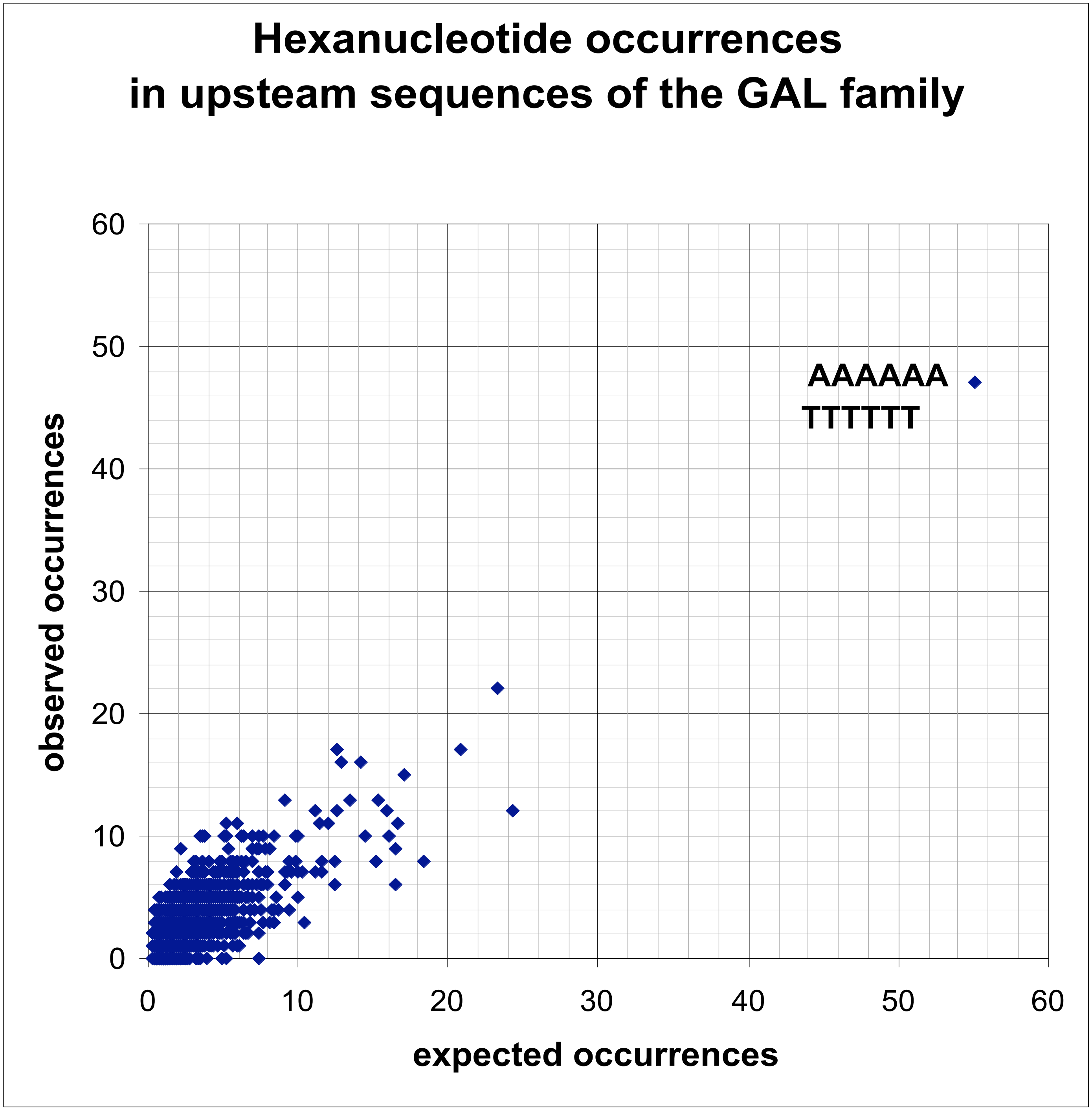


Hexanucleotide occurrences
in upsteam sequences of the NIT family
AAAAAA
TTTTTT
TATATA
ATATAT
GATAAG
ATAAGA
AAATTT
observed occurrences
expected occurrences
0
10
20
30
40
50
60
70
80
90


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



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



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

Scoring scheme - Binomial

- Advantages
  - rigorous probability
  - appropriate for small sequence sets, where some words have a very low expected number of occurrences (<1)
- Weaknesses
  - bias for self-overlapping words

- Probability to observe exactly s occurrences

$$P(X = s) = \frac{n!}{s!(n-s)!} p^s (1-p)^{n-s}$$

- Probability to observe at least s occurrences

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

Where

s = occurrences

n = positions on sequence

p = word probability

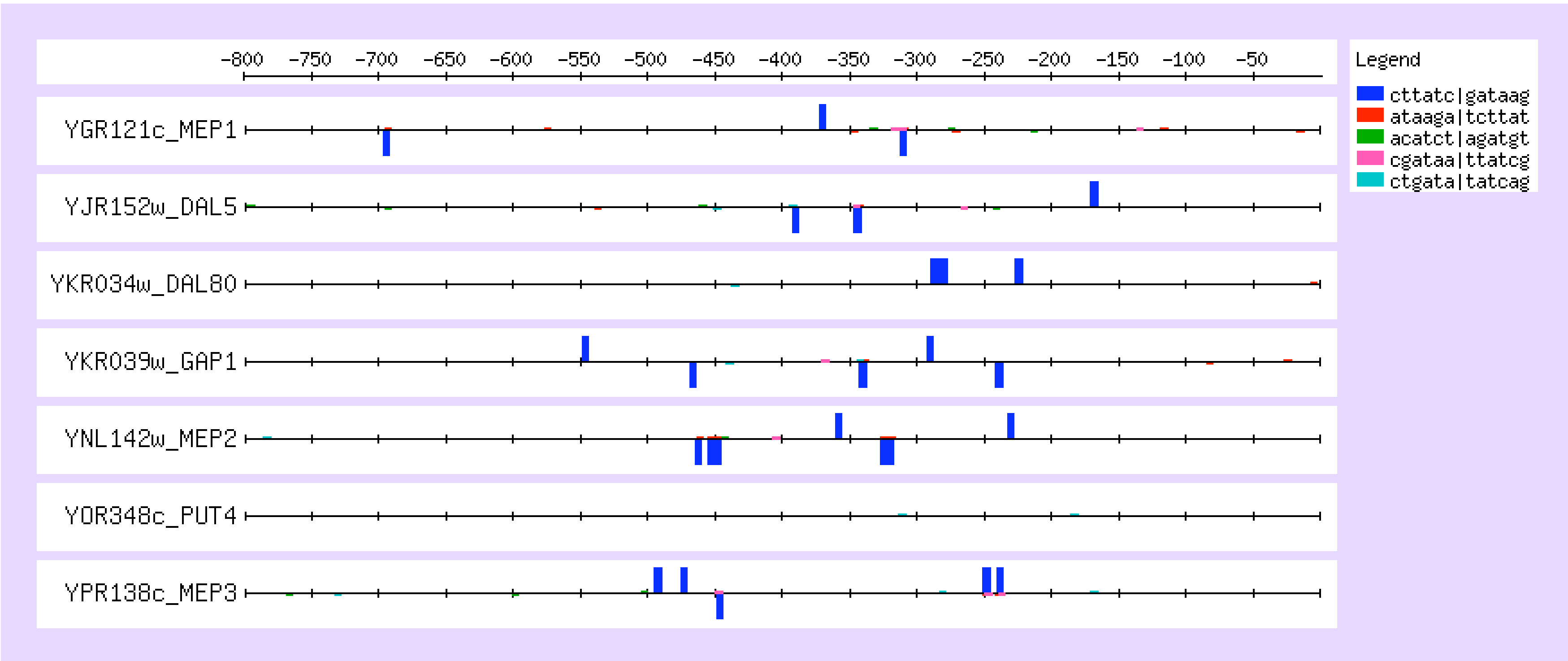
Hexanucleotide analysis of the NIT family

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

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

Feature-map of discovered patterns - NIT family

- 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



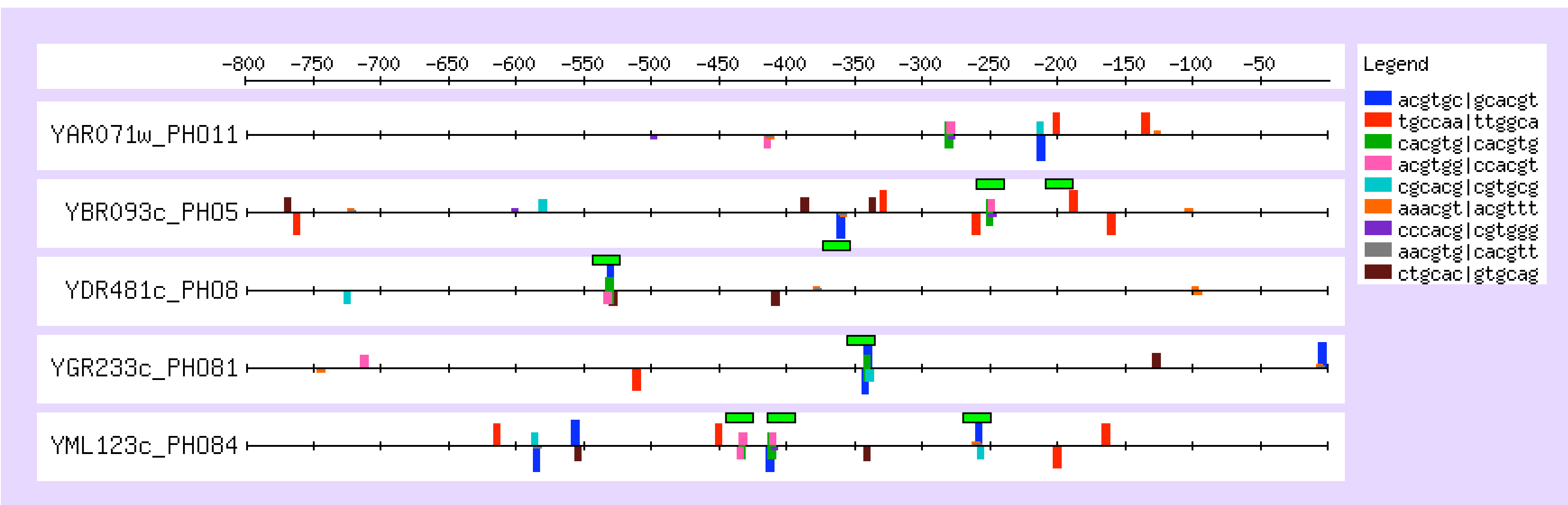
Hexanucleotide analysis of the PHO family

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)*

Feature-map of discovered patterns - PHO family

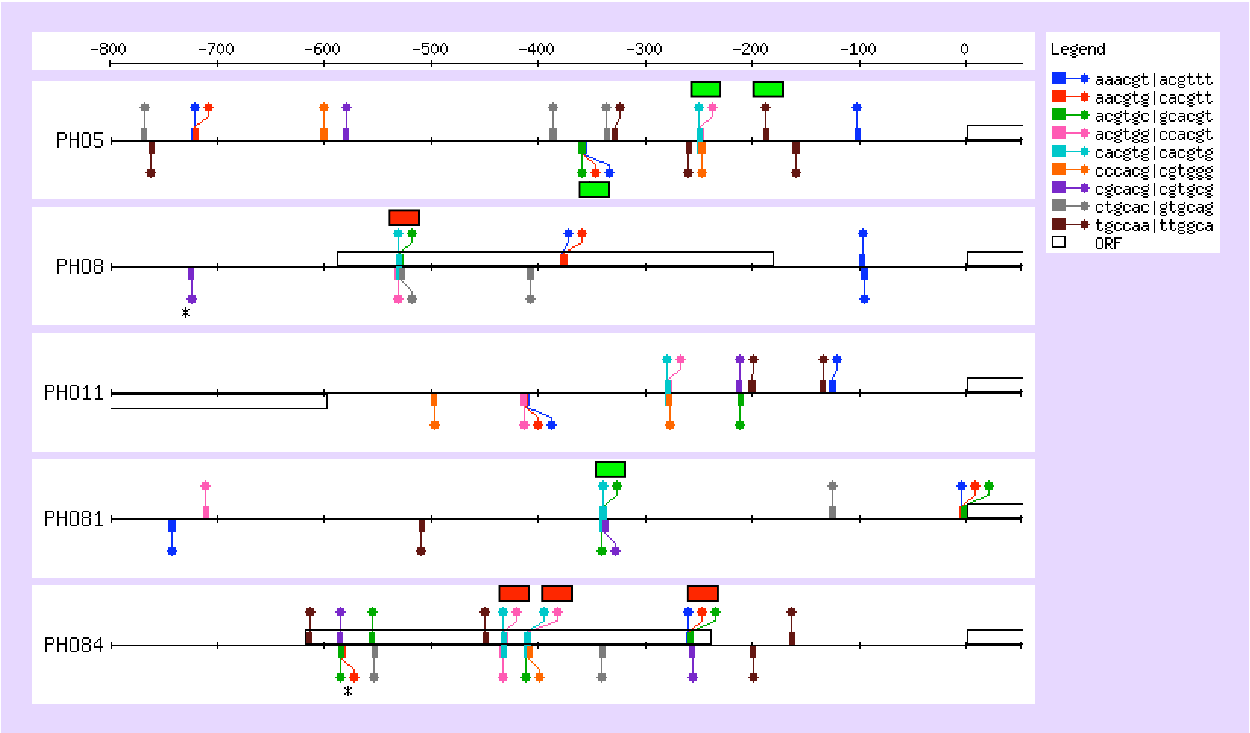
- The feature-map provides a convenient representation of the discovered patterns
  - Each colour represents one pattern.
  - Box height reflects pattern significance.
  - Clusters of mutually overlapping words represent sites larger than 6 bp.
- Green bars were superimposed, to indicate the positions 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



Site with experimental evidence

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.



Real Site located in false ORF predictions

Hexanucleotide analysis of the MET family

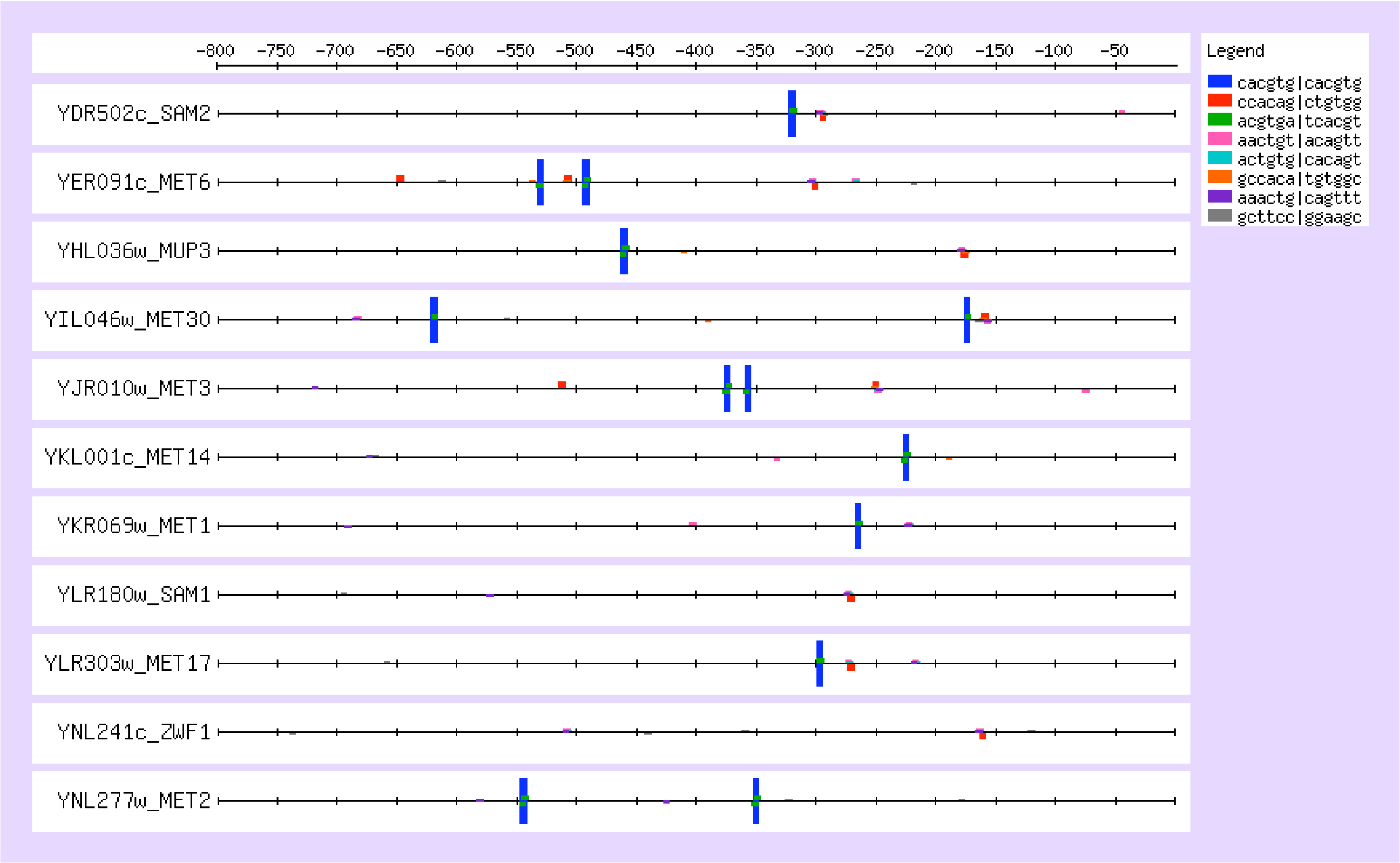
Sequence	exp freq	occ	exp occ	P-value	E-value	sig	matching sequences
..ACGTGa	0.00033	13	2.9	1.00E-05	2.20E-02	1.67	9
.CACGTG.	**0.00012**	**13**	**1.0**	**6.90E-11**	**1.40E-07**	**6.84**	**9**
tCACGTG.	0.00033	13	2.9	1.00E-05	2.20E-02	1.67	9
tCACGTGa	consensus						
....TGTGGc	0.00027	10	2.3	1.50E-04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

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	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.

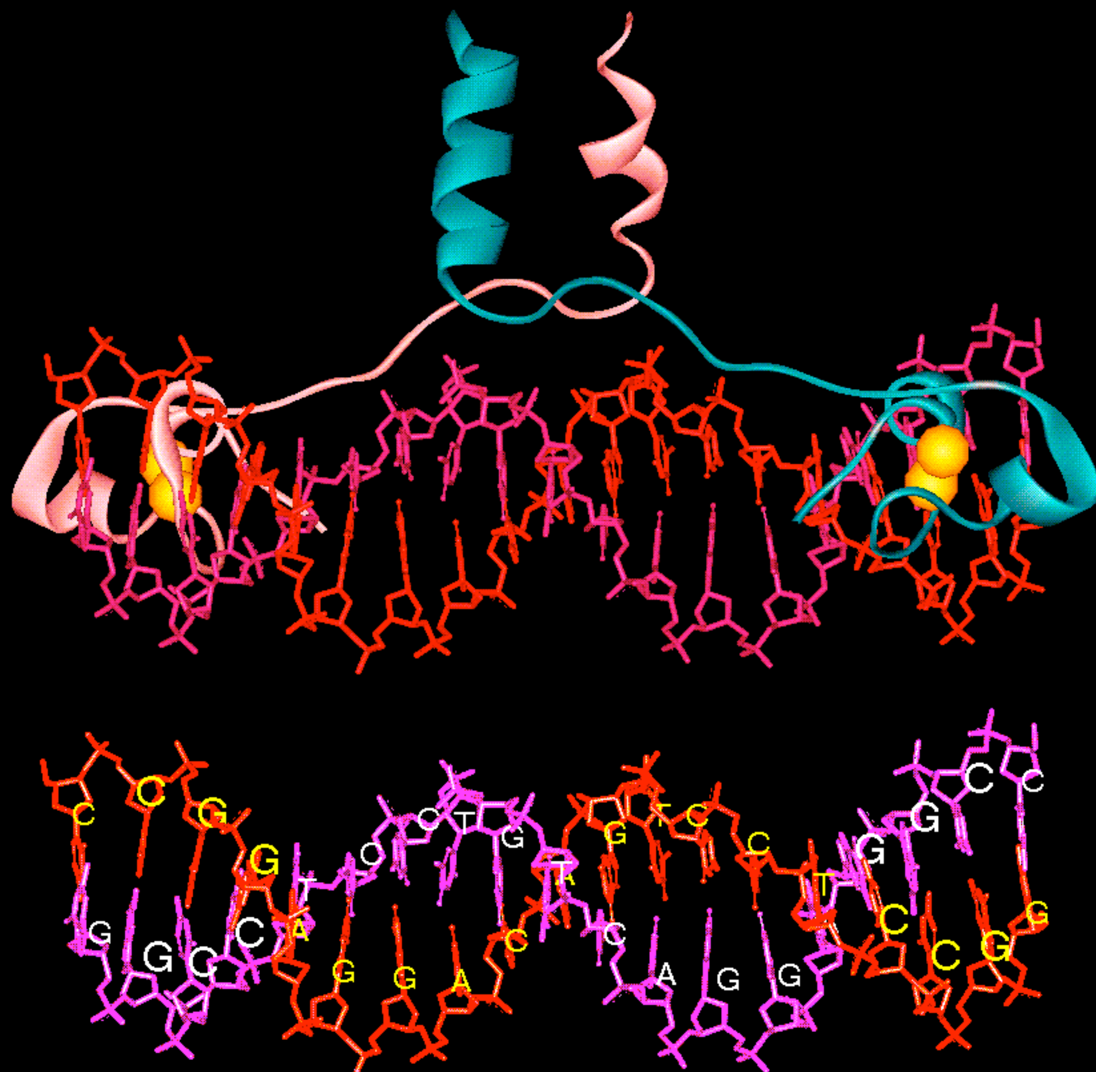
Hexanucleotide analysis of the GAL family

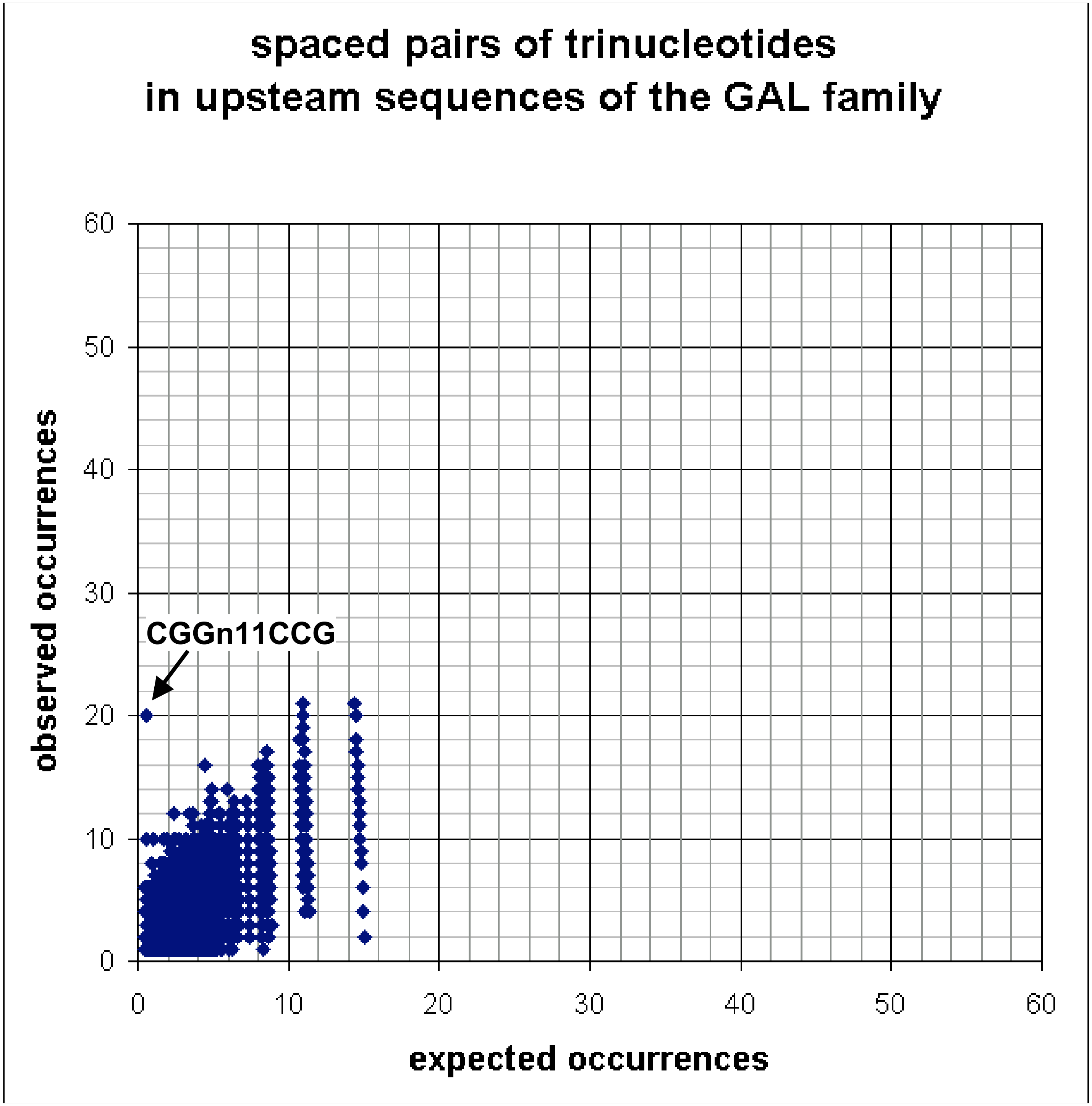
Sequence	exp freq	occ	exp occ	P-value	E-value	sig	matching sequences
agacat	0.00044	9	2.1	0.00033	**0.69**	**0.16**	4

Genes	*GAL1, GAL2, GAL7, GAL80, MEL1, GCY1*
Known motifs	*Factors*
CGG**n_5**w**n_5**CCG	*Gal4p*

- 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
CGGn11CCG
observed occurrences
expected occurrences
0
10
20
30
40
50
60

Dyad analysis of the GAL family

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

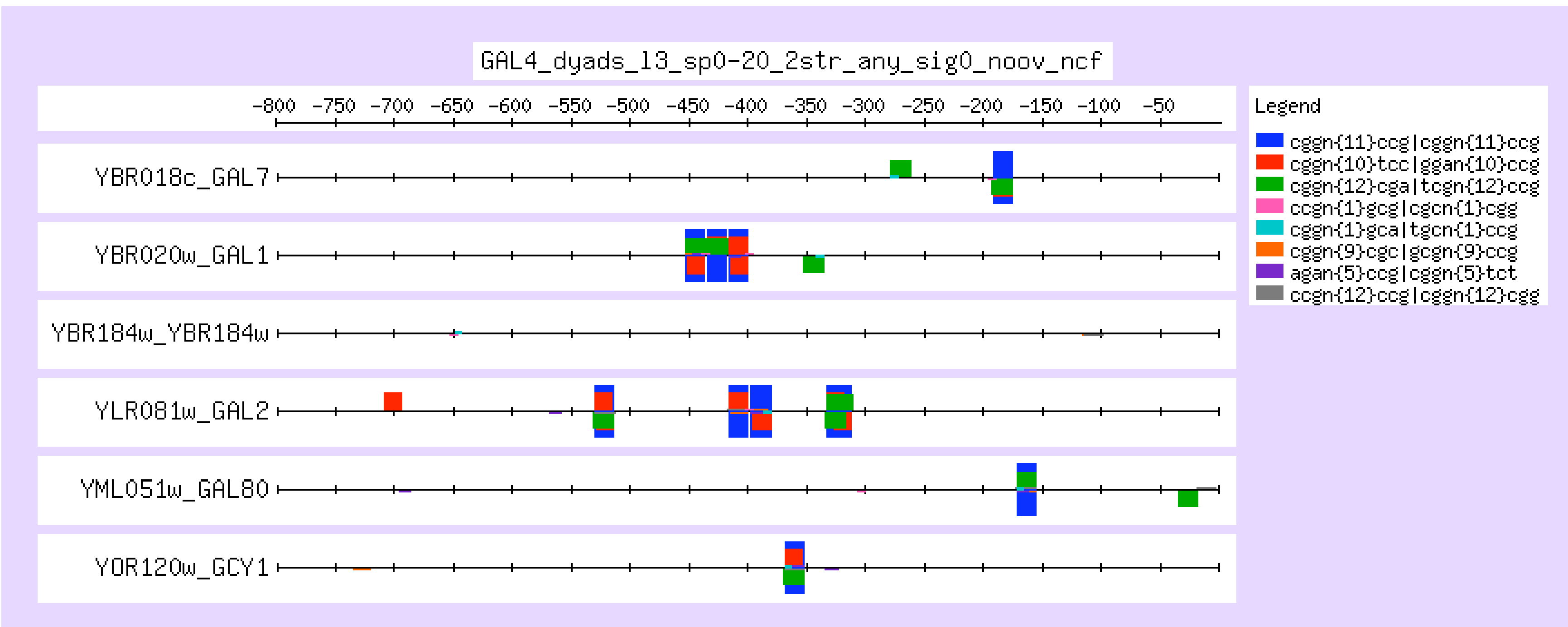
Genes *GAL1, GAL2, GAL7, GAL80, MEL1, GCY1*

Known motifs *Factors*

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

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).



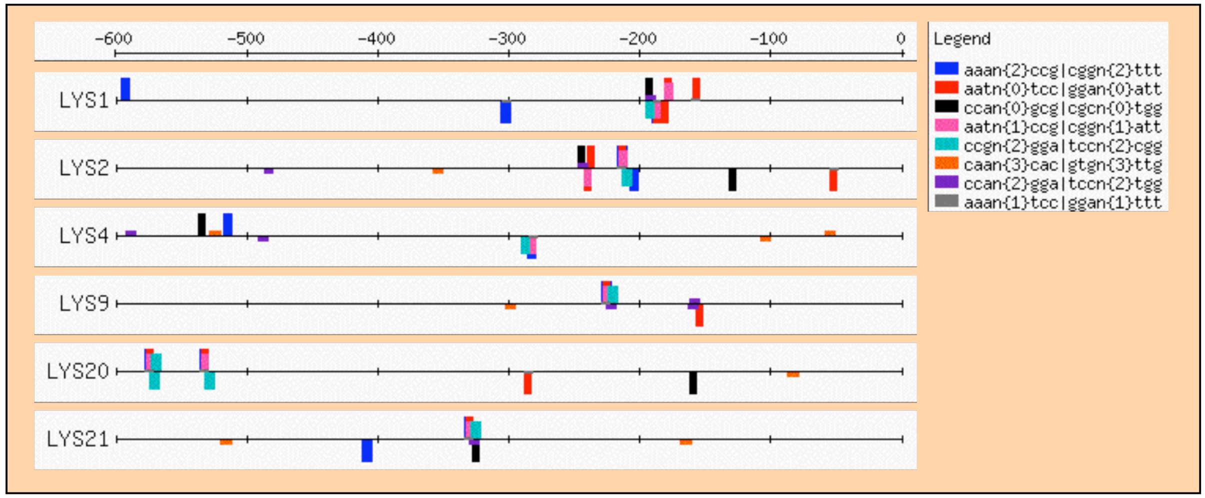
Dyad analysis: regulons of Zn cluster proteins

FACTOR	# genes	KNOWN MOTIFS	DYADS	REVERSE DYADS	SCORE
GAL4	6	$CGGn_{11}CCG$	$TCGGAn_9TCCGG$ $TCGGCGCAGAn_4TCCGG$	$CCGGAn_9TCCGA$ $CCGGAn_4TCTGCGCCGA$	7.8 7.8
HAP1	9	CGGnnntanCGG	$GGAn_5CGGC$ $GGGGGn_{12}GGC$ $CCTn_{10}GGC$	$GCCGn_5TCC$ $GCCn_{12}CCCCC$ $GCCn_{10}AGG$	1.8 1.4 1.1
LEU3	5	RCCggnnccGGY	$CCGn_3CCG$	$CGGn_3CGG$	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		$CGGn_6CCG$	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	$CGGn_4ATGGAA$	$TTCCATn_4CCG$	6.0

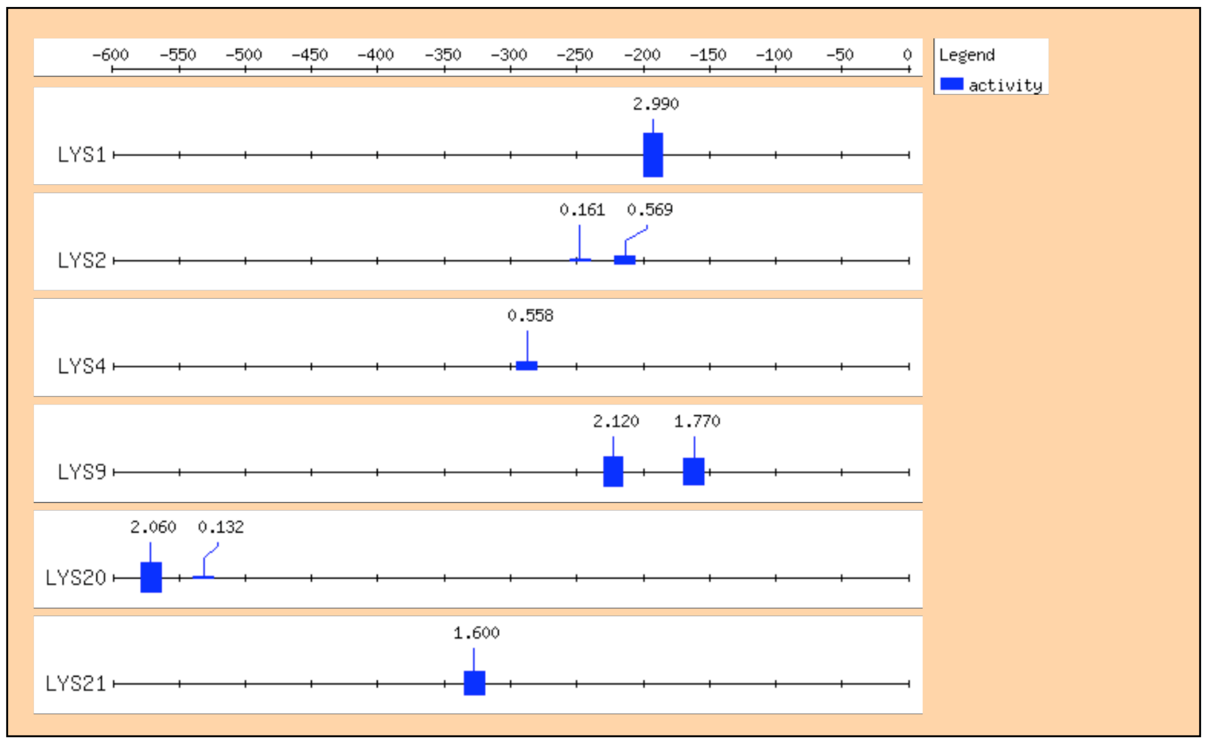
van Helden et al. (2000). *Nucleic Acids Res* **28**(8), 1808-18.

Comparison of discovered patterns with known sites (LYS family)

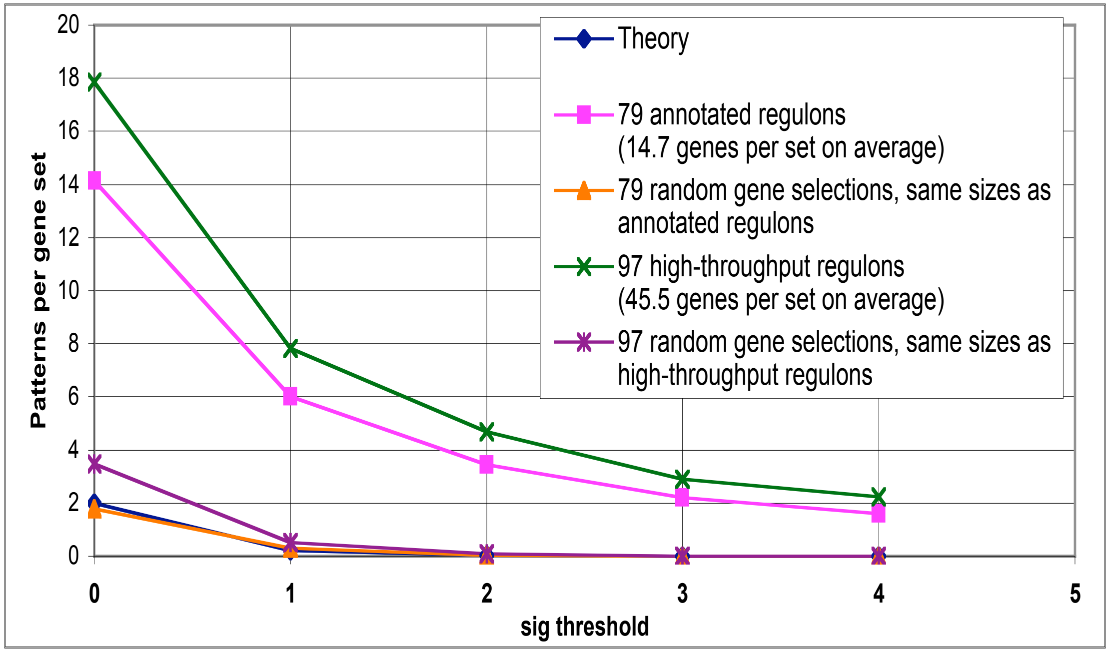
Patterns discovered by dyad analysis

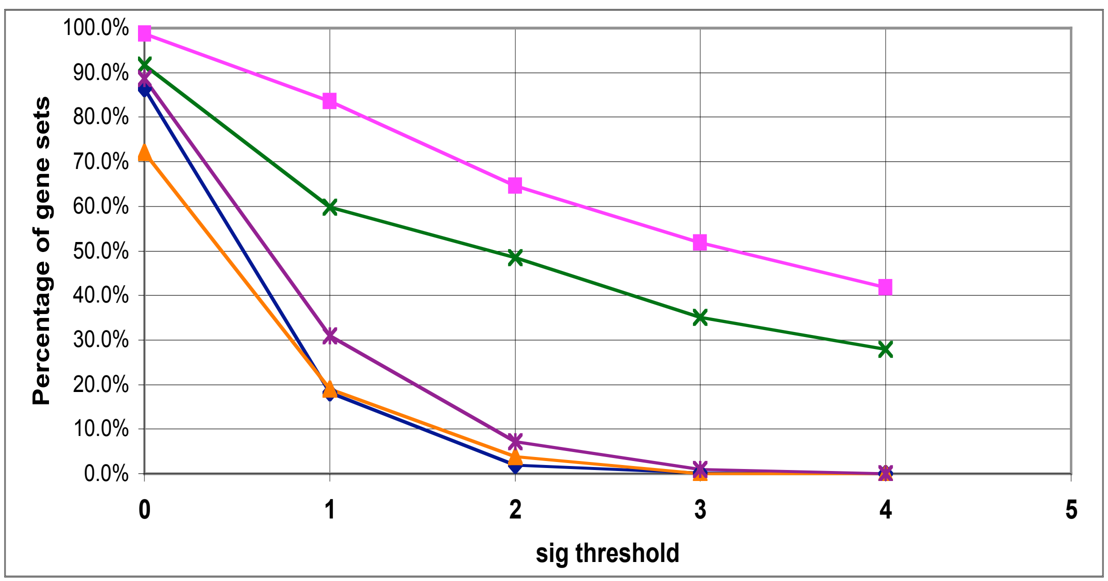


Experimental measurement of activity



Validation of pattern discovery with yeast regulons

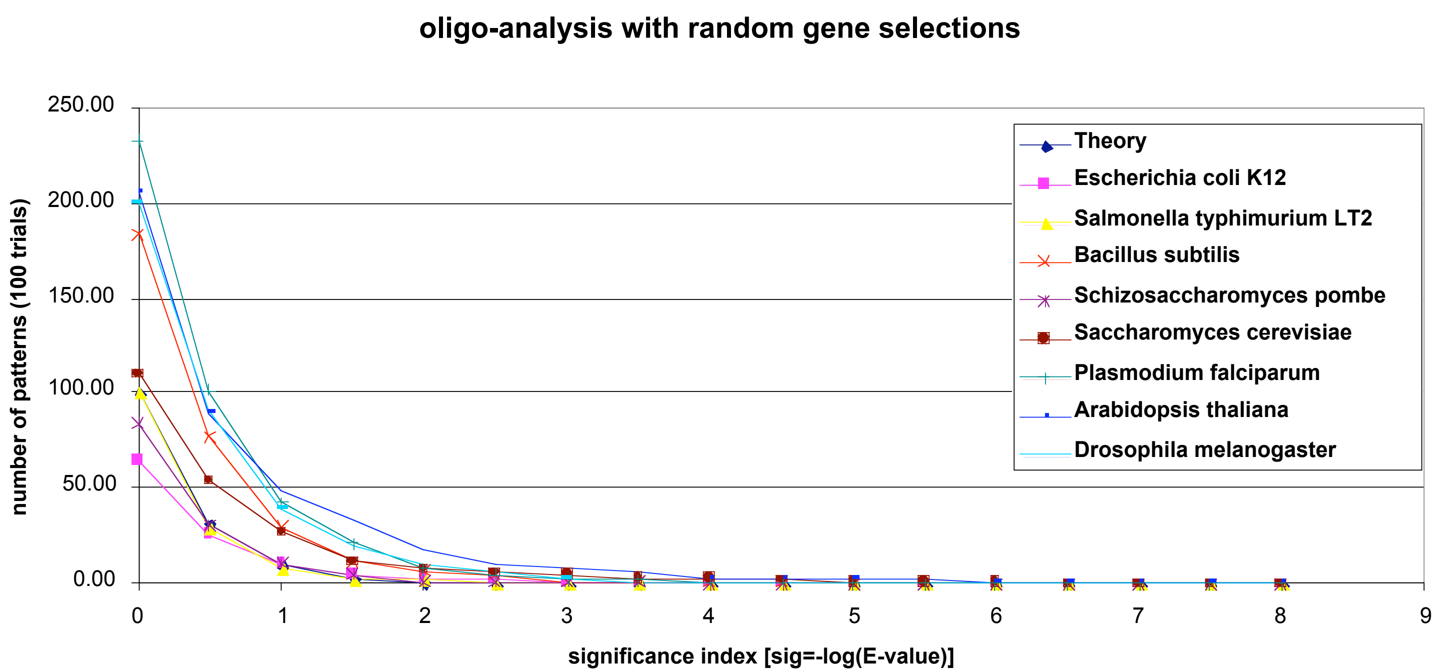




- Regulons were collected from TRANSFAC and aMAZE.
- All the regulons with ≥5 genes were analyzed.
  - Significant patterns (sig ≥ 2) are detected in 65% of the regulons.
- As a negative control, sets of random genes were analyzed.
  - The rate of false positive follows pretty well the statistical expectation.

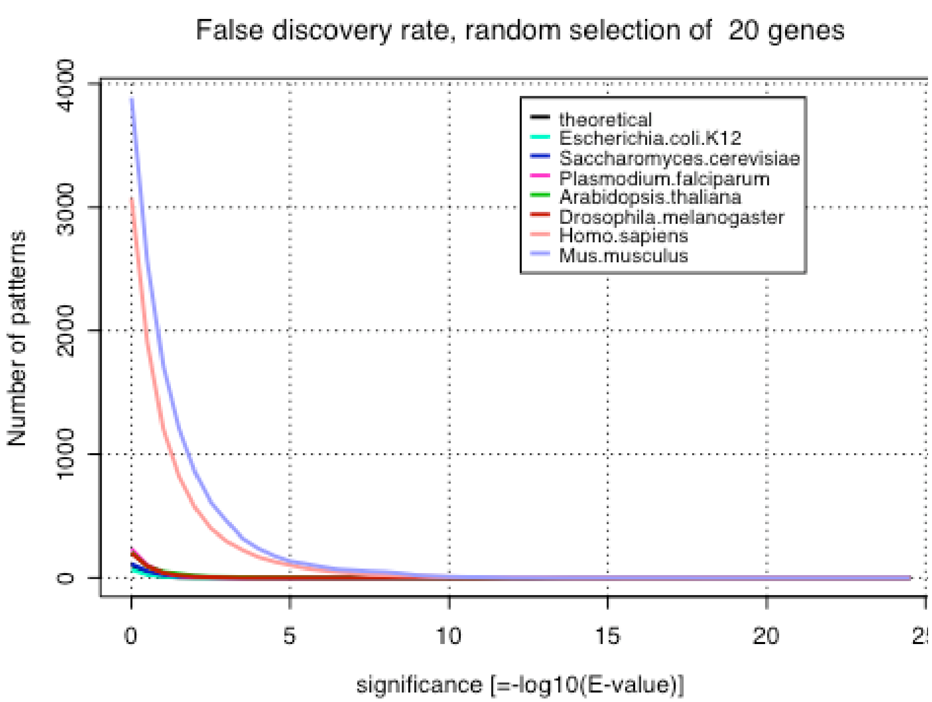
Rate of false positive in different organisms

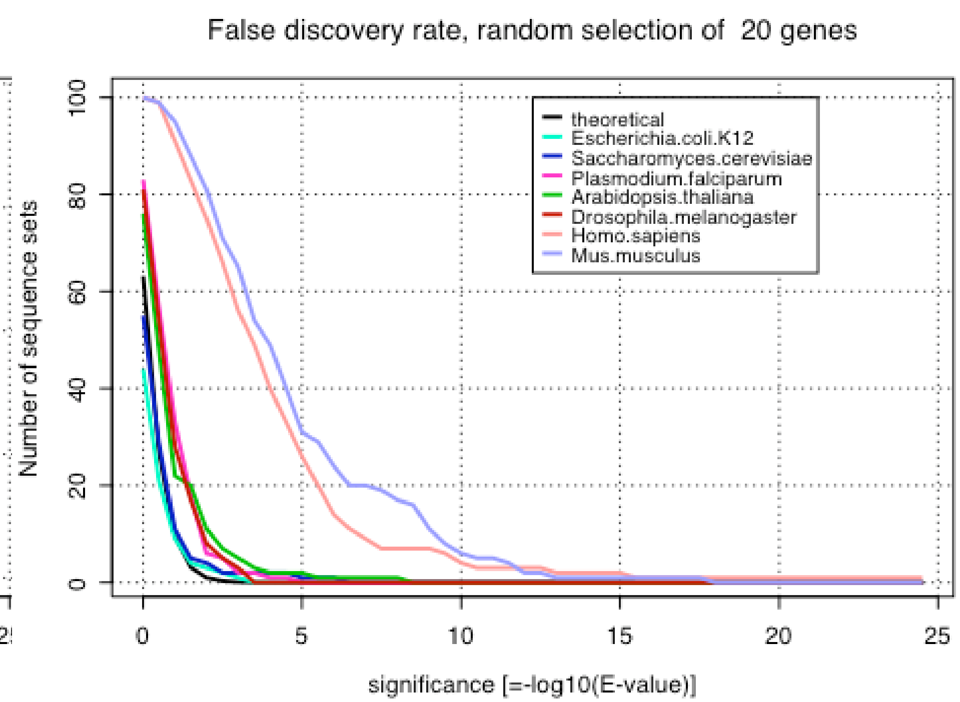
- The analysis of random gene selections allows to evaluate the rate of false positive returned by a pattern discovery program.
- The rate of false positive is good for microbes (bacteria, yeasts, ...), but increases for multicellular organisms (e.g. the fly *Drosophila*, the plant *Arabidopsis thaliana*, …).
- 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.





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
- Fast (2-3 seconds/family)
- Time increases linearly with sequence set
  - Can be applied to very large sequence sets (full genomes)
- Ability to return a negative answer
  - "not a single over-represented pattern in this sequence set"
  - Corollary: very low false positive rate
- 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: 15^k 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-approach for pattern discovery
  - use discovered strings as seeds for building a matrix, which can be used for pattern search