Motif discovery

Morgane Thomas-Chollier

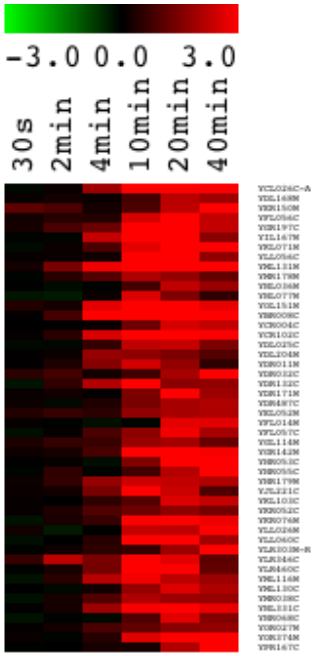
Computational systems biology - IBENS

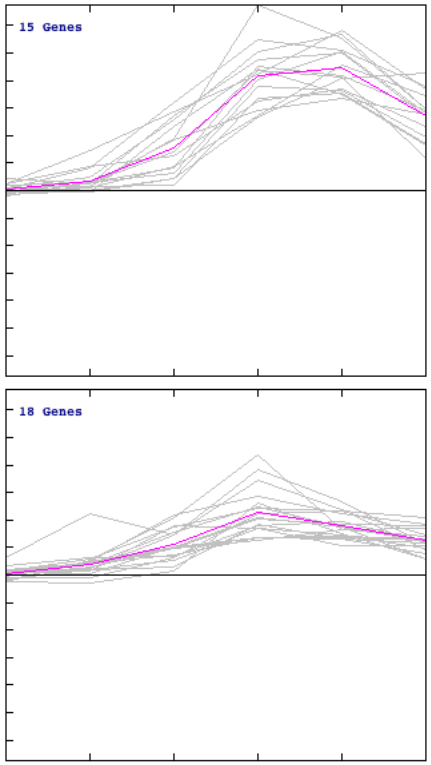
mthomas@biologie.ens.fr

M2 – Computational analysis of cis-regulatory sequences 2015/2016

Denis Thieffry, Jacques van Helden and Carl Herrmann kindly shared some of their slides.

Co-expressed genes



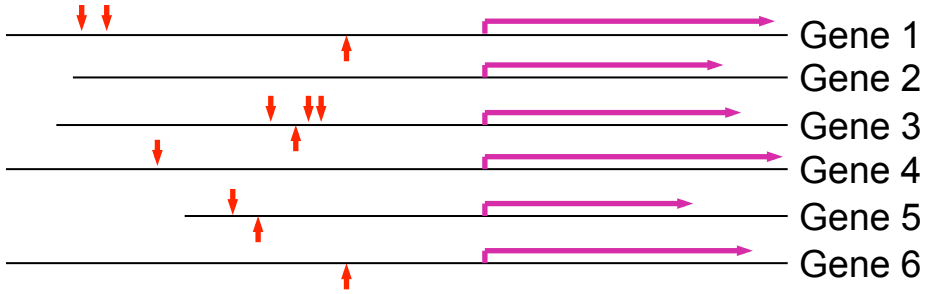


clusters of **co-expressed genes** during oxidative stress in yeast

Are they co-regulated ?
If so, what is the TF ?

Motif discovery

1 - Understand what is a motif discovery problem



2 – Motif discovery approaches

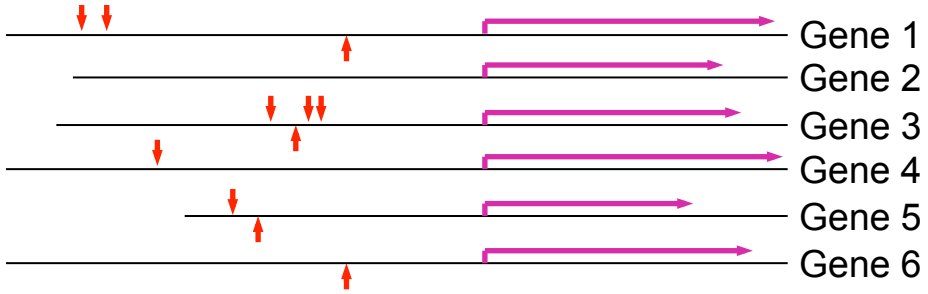
- Word counting
- Gibbs sampling

3 – Important parameters

Aim of the course

Motif discovery

1 - Understand what is a motif discovery problem



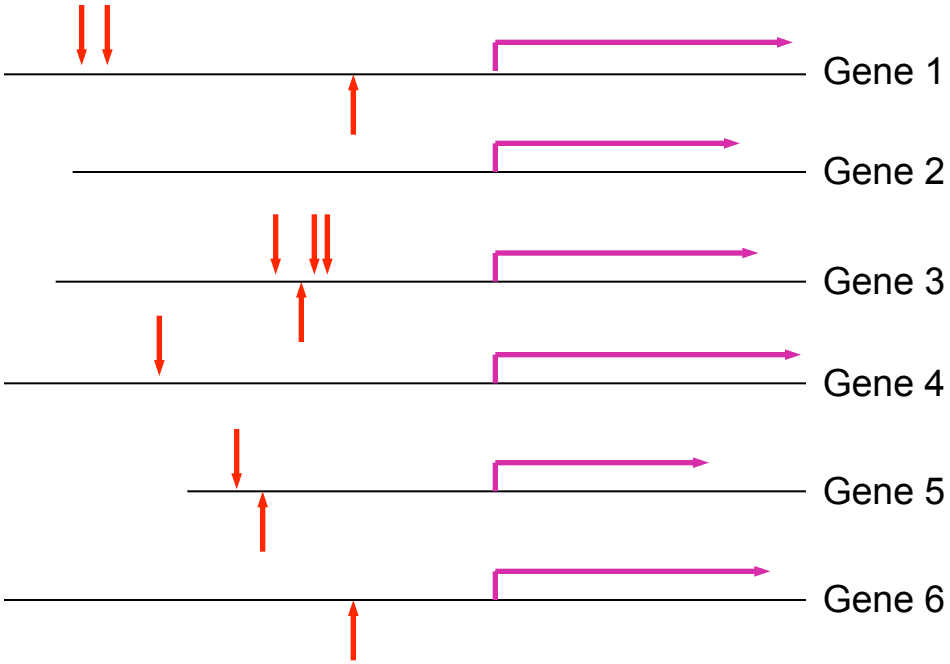
2 – Motif discovery approaches

- Word counting
- Gibbs sampling

3 – Important parameters

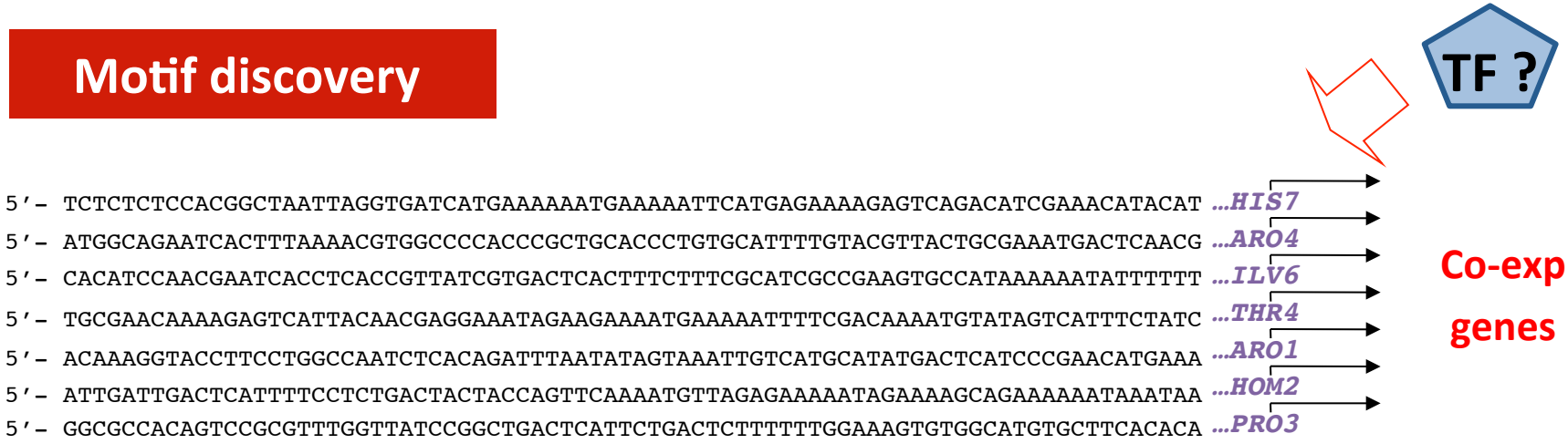
Co-expressed genes

Knowing that a set of genes are co-regulated, one can expect that their upstream regions contains some regulatory signal.



A motif discovery problem

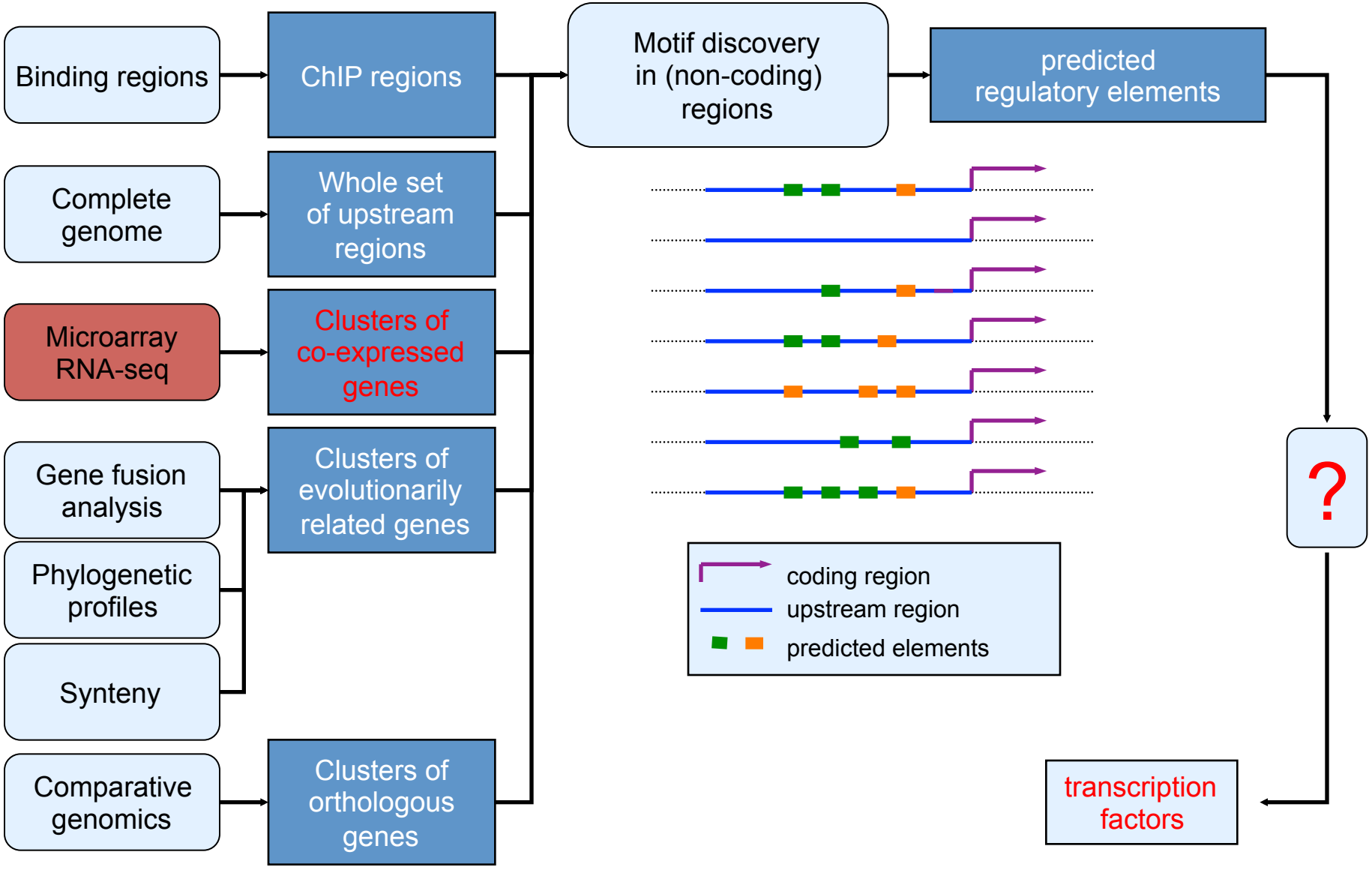
Motif discovery



Problem : If there is a common regulating factor, can we discover its motif (some signal) on the basis of these sequences ONLY ?

- We have a set of sequences
- We suspect that they share some functional signal
- We ignore the transcription factors involved in this regulation.
- We ignore the cis-acting elements

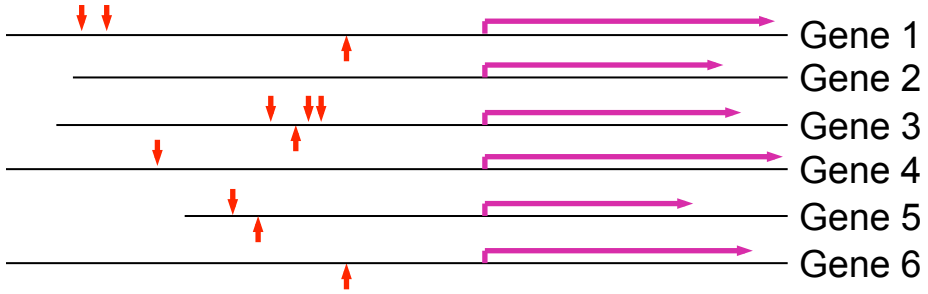
Typical motif discovery problems



Aim of the course

Motif discovery

1 - Understand what is a motif discovery problem

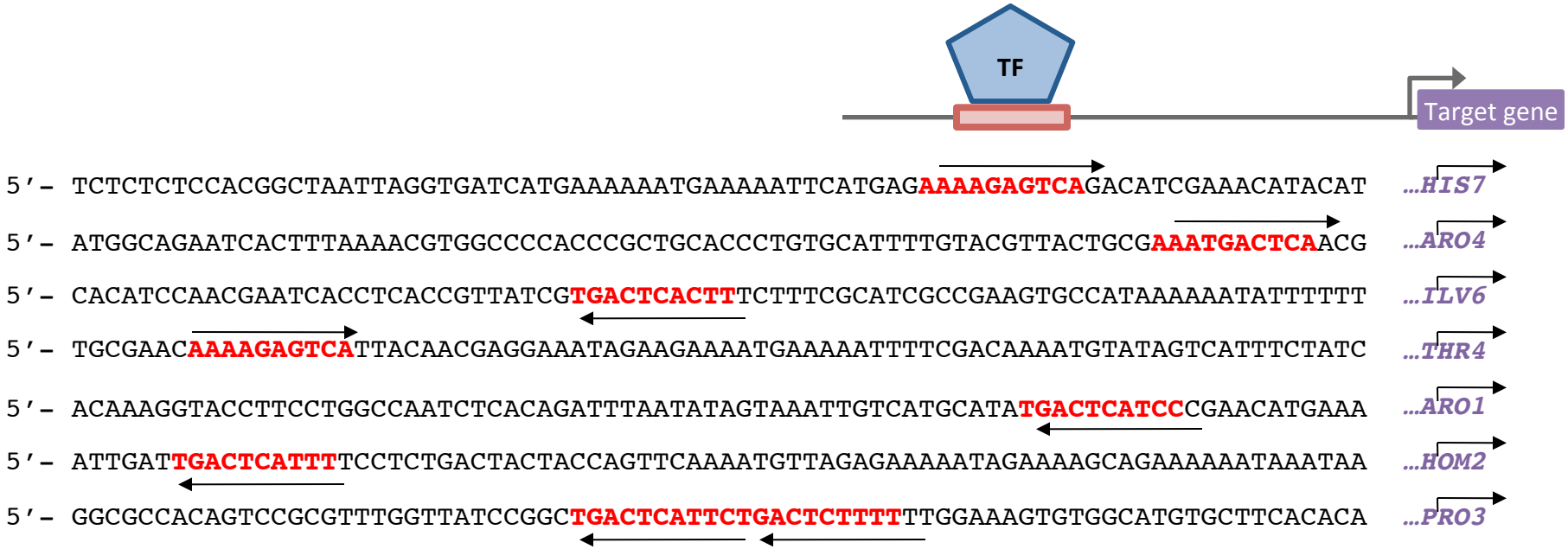


2 – Motif discovery approaches

- Word counting
- Gibbs sampling

3 – Important parameters

Principle: detect unexpected patterns



- Binding sites are represented as “words” = “string”=“k-mer”
  - e.g. **acgtga** is a 6-mer
- Signal is likely to be more frequent in the upstream regions of the co-regulated genes than in a random selection of genes
- We will thus detect over-represented words

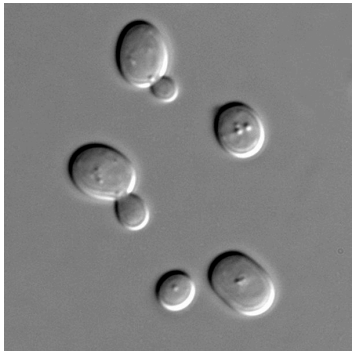
Motif discovery using word counting

Idea:
motifs corresponding to binding sites are generally repeated in the dataset
→ capture this statistical signal

- **Algorithm**
  - count occurrences of **all k-mers** in a set of related sequences (promoters of co-expressed genes, in ChIP bound regions,...)

Let's take an example (yeast Saccharomyces cerevisiae)

- NIT
  - 7 genes expressed under low nitrogen conditions
- MET
  - 10 genes expressed in absence of methionine
- PHO
  - 5 genes expressed under phosphate stress



PHO		
aaaaaa	tttttt	51
aaaaag	ctttttt	15
aagaaa	tttctt	14
gaaaaa	ttttttc	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
aataat	attatt	10
aagaag	cttctt	10
atataa	ttatat	10

MET		
aaaaaa	tttttt	105
atatat	atatat	41
gaaaaa	ttttttc	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
tgaaaa	ttttca	21
ataata	tattat	20
atataa	ttatat	20

NIT		
aaaaaa	tttttt	80
cttatc	gataag	26
tatata	tatata	22
ataaga	tcttat	20
aagaaa	tttctt	20
gaaaaa	ttttttc	19
atatat	atatat	19
agataa	ttatct	17
agaaaa	ttttct	17
aaagaa	ttcttt	16
aaaaca	tgtttt	16
aaaaag	ctttttt	15
agaaga	tcttct	14
tgataa	ttatca	14
atataa	ttatat	14

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	
aaaaaa|tttttt	51	aaaaaa|tttttt	105	aaaaaa|tttttt	80
aaaaag|cttttt	15	atatat|atatat	41	cttatc|gataag	26
aagaaa|tttctt	14	gaaaaa|ttttтc	40	tatata|tatata	22
gaaaaa|ttttтc	13	tatata|tatata	40	ataaga|tcttat	20
tgccaa|ttggca	12	aaaaat|attttt	35	aagaaa|tttctt	20
aaaaat|attttt	12	aagaaa|tttctt	29	gaaaaa|ttttтc	19
aaatta|taattt	12	agaaaa|ttttct	28	atatat|atatat	19
agaaaa|ttttct	11	aaaata|tatttt	26	agataa|ttatct	17
caagaa|ttcttg	11	aaaaag|cttttt	25	agaaaa|ttttct	17
aaacgt|acgttt	11	agaaat|atttct	24	aaagaa|ttcttt	16
aaagaa|ttcttt	11	aaataa|ttattt	22	aaaaca|tgtttt	16
acgtgc|gcacgt	10	taaaaa|ttttta	21	aaaaag|cttttt	15
aataat|attatt	10	tgaaaa|ttttca	21	agaaga|tcttct	14
aagaag|cttctt	10	ataata|tattat	20	tgataa|ttatca	14
atataa|ttatat	10	atataa|ttatat	20	atataa|ttatat	14

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).
  => **“Background”**

Motif discovery using word counting

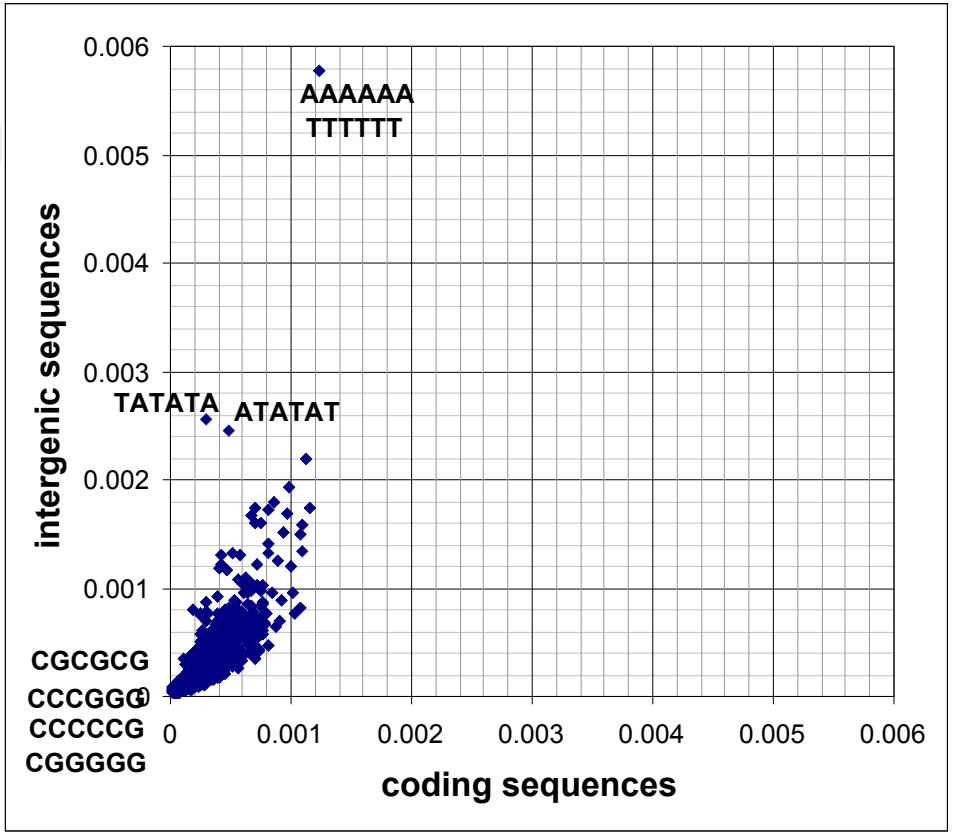
Idea:
motifs corresponding to binding sites are generally repeated in the dataset
→ capture this statistical signal

- **Algorithm**
  - count occurrences of **all k-mers** in a set of related sequences (promoters of co-expressed genes, in ChIP bound regions,...)
  - estimate the **expected number of occurrences** from a background model
    - empirical based on observed k-mer frequencies
    - theoretical background model (Markov Models)

Estimation of word expected frequencies from background sequences

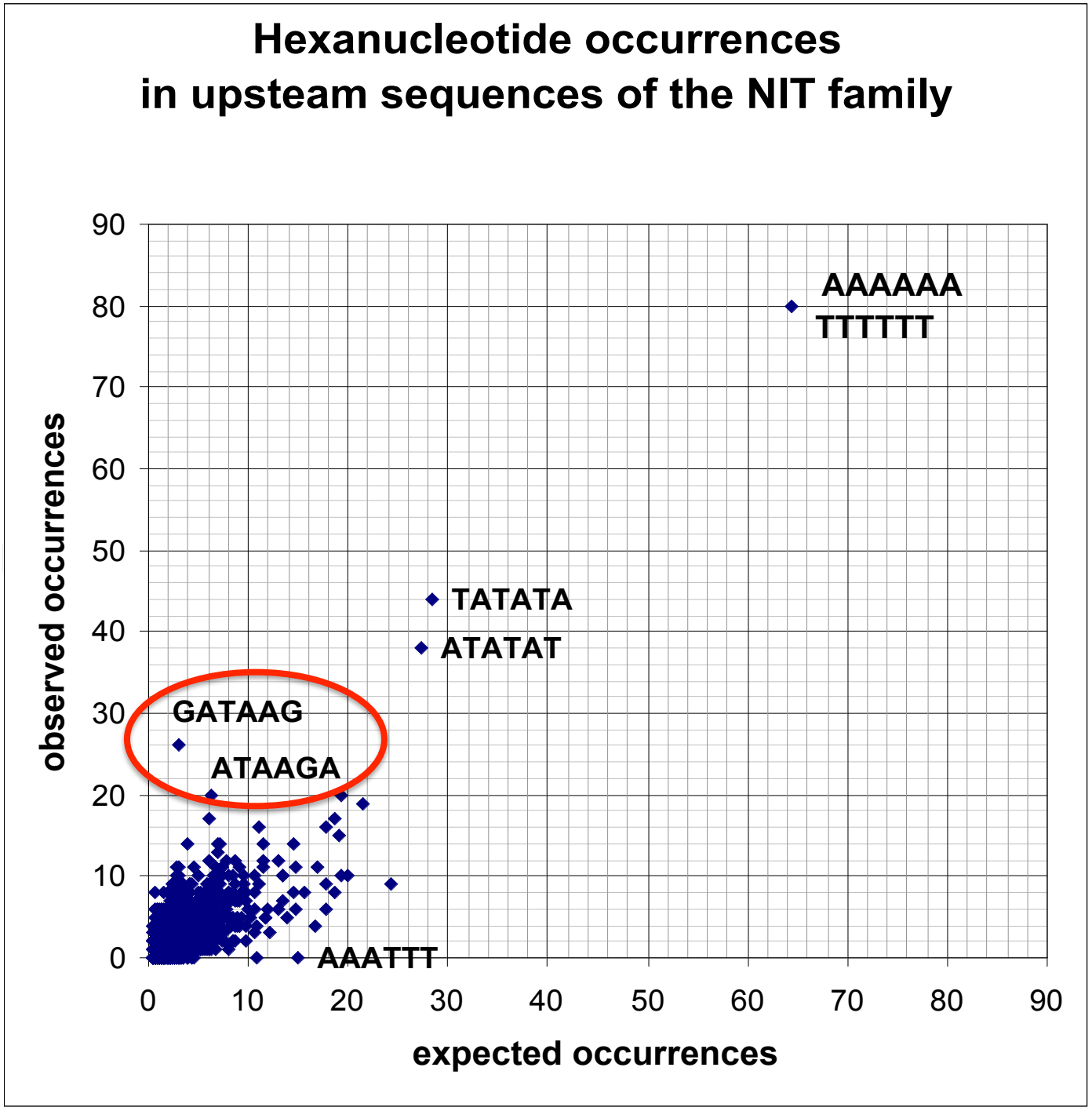
Example:
6nt frequencies in the whole set of 6000 yeast **upstream** sequences

;seq	identifier	observed_freq	occ
aaaaaa	aaaaaa|tttttt	0,00510699	14555
aaaaac	aaaaac|gttttt	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|ccttt	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



6nt frequencies differ between coding and non-coding sequences

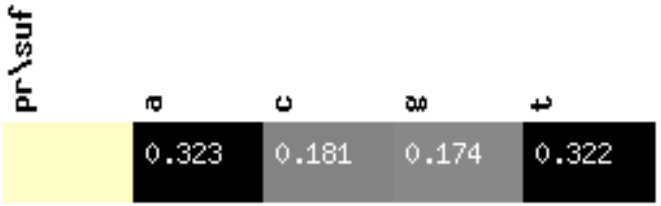
NIT		
aaaaaa	tttttt	80
cttatc	**gataag**	**26**
tatata	tatata	22
ataaga	**tcttat**	**20**
aagaaa	tttctt	20
gaaaaa	ttttтc	19
atatat	atatat	19
agataa	ttatct	17
agaaaa	ttttct	17
aaagaa	ttcttt	16
aaaaca	tgtttt	16
aaaaag	cttttt	15
agaaga	tcttct	14
tgataa	ttatca	14
atataa	ttatat	14



Estimation of background frequencies from a Markov Model

- **Estimate the frequency using a statistical model**
  - **Bernouilli model** (=Markov order 0): p(A), p(C), p(G), p(T)
    Assumes independence between successive nucleotides

    simplest model: p(A)=p(C)=p(G)=p(T) → p=0.25
    => NOT realistic does not reflect biological sequences !!!



frequencies in non-coding upstream regions of S. cerevisiae
p(A)=0.3 p(C)=0.2 p(G)=0.2 p(T)=0.3

  - **Markov model**

The probability of each residue depends on the ***m*** preceding residues.
The parameter ***m*** is called the **order** of the Markov model

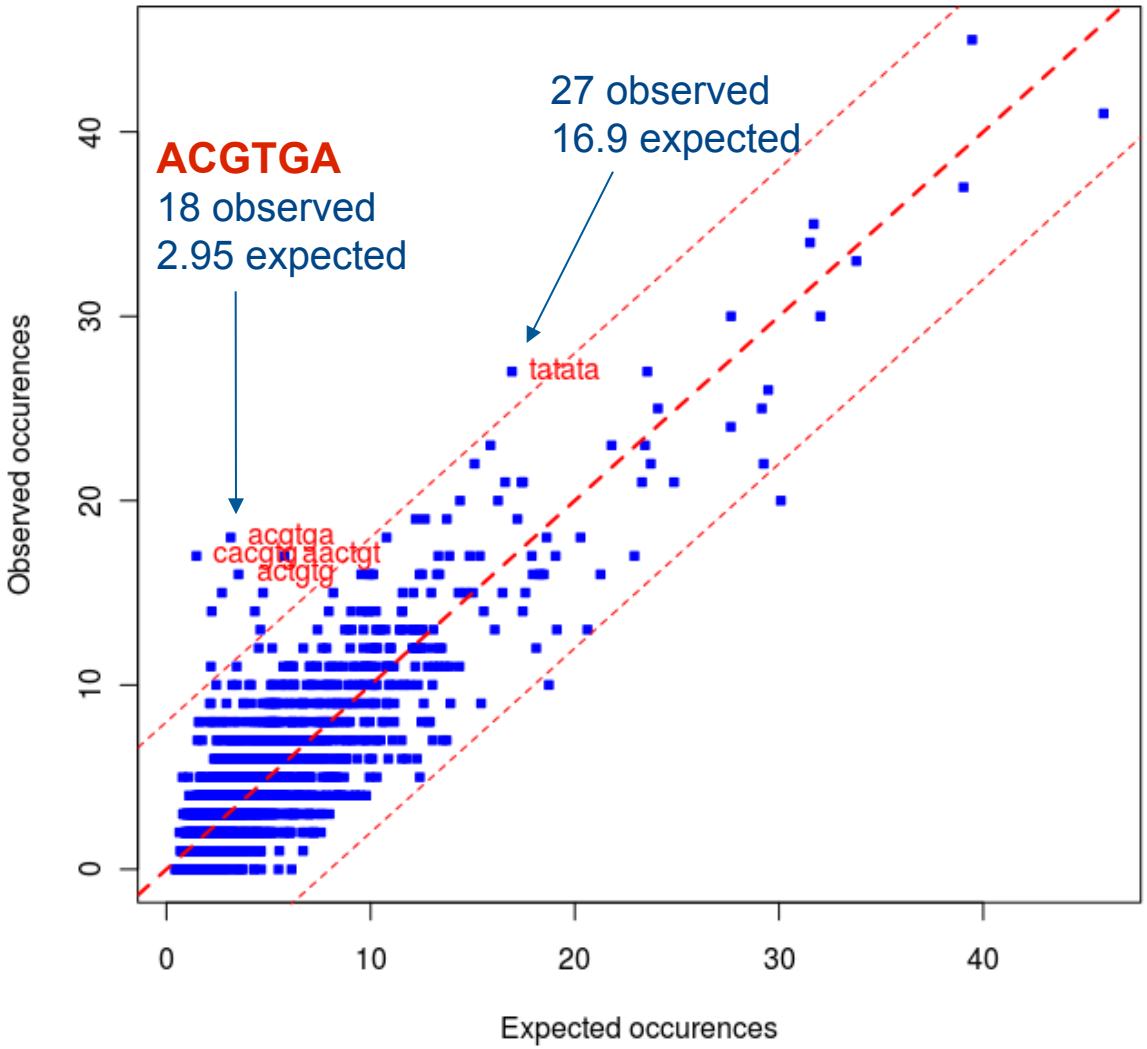
Motif discovery using word counting

- **Example**:
  19 genes from *Saccharomyces cerevisiae* involved in methionine biosynthesis pathway

- *Are they co-regulated ?*
  Do they share common regulatory motifs ?

- **Principle**
  - Count occurrences of k=6 mers in the 800 bp upstream of the TSS (!! on both strands !!)
  - 9000 possible positions
  - compare **observed** vs **expected** occurences

Motif discovery using word counting



How to evaluate expected number of occurrences ?

Empirical background model (frequencies)

Estimated frequency of **ACGTGA in** S. cerevisae ?

- **observed frequency** of this word in the whole genome
  - all **intergenic sequences** in the genome:
    1026 occurrences for 3310685 positions → **p = 3.09e-4 (2.78 expected occurrences for 9000 positions)**

  - all **upstream sequences** in the genome :
    921 occurrences for 2804964 positions → **p = 3.33e-4 (2.95 expected occurrences for 9000 positions)**

Background as a Markov model

Estimated frequency of **ACGTGA in** S. cerevisae ?

- estimate the frequency using a statistical model
  - **Bernouilli model** : p(A), p(C), p(G), p(T)

    $p(ACGTGA) = p(A)^2 \times p(C) \times p(G)^2 \times p(T)$ → **p = 3.94e-4 (3.70)**
  - **Markov models**

pr\suf	a	c	g	t	P_prefix
a	0.35010	0.19037	0.19473	0.26480	0.28
c	0.31445	0.22506	0.21222	0.24827	0.22
g	0.25673	0.27652	0.22424	0.24251	0.22
t	0.20201	0.20104	0.24615	0.35081	0.28

- Markov model order 1 : **p = 3.48e-4 (3.48)**
  p(ACGTGA) = p(A) p(C|A) p(G|C) p(T|G) p(G|T) p(A|G)
- Markov model order 2 : **p = 4.87e-4 (4.87)**
  p(ACGTGA) = p(AC)x p(G|AC)x p(T|CG)x p(G|GT)x p(A|TG)
- Markov model order 3 : **p = 7.4e-4 (6.96)**
  p(ACGTGA) = p(ACG)x p(T|ACG)x p(G|CGT)x p(A|GTG)

Expected occurrences under different background models

Estimated frequency of **ACGTGA in** S. cerevisae ?

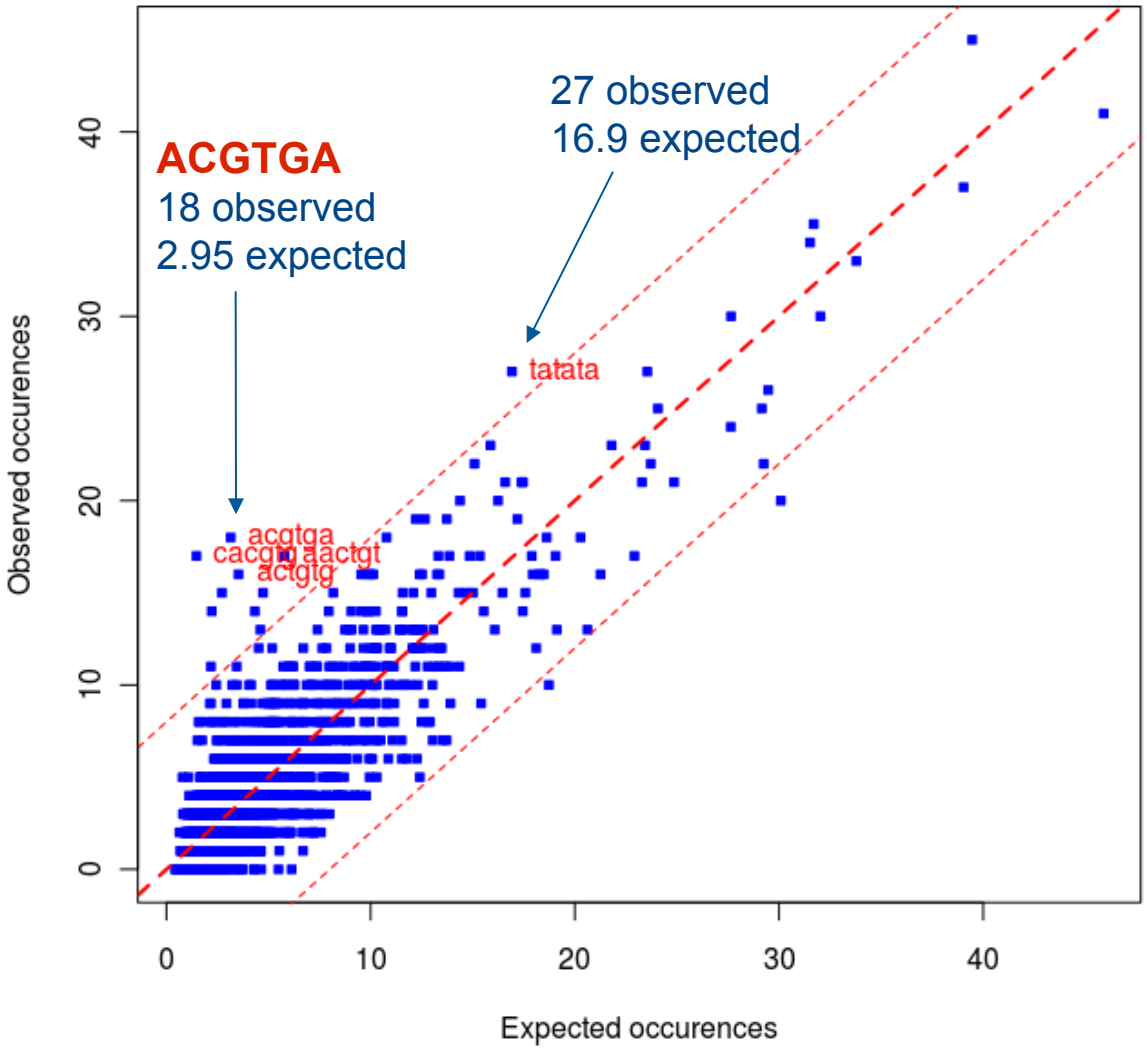
	Method	Frequency (p)	Occurrences for 9000 positions
Observation	**obvserved in the dataset**		**18**
	intergenic frequency	3.25e-4	3.05
	promoter frequency	3.35e-4	3.15
	Markov order 0	3.94e-4	3.70
Estimations	Markov order 1	3.70e-4	3.48
	Markov order 2	5.19e-4	4.87
	Markov order 3	7.42e-4	6.96
	promoter frequency in **human**	1.63e-4	1.53

Motif discovery using word counting

Idea:
motifs corresponding to binding sites are generally repeated in the dataset
→ capture this statistical signal

- **Algorithm**
  - count occurrences of **all k-mers** in a set of related sequences (promoters of co-expressed genes, in ChIP bound regions,...)
  - estimate the **expected number of occurrences** from a background model
    - empirical based on observed k-mer frequencies
    - theoretical background model (Markov Models)
  - **statistical evaluation of the deviation observed** (P-value/E-value)

Statistical evaluation



How « big » is the surprise to observe 18 occurrences when we expect 2.95 ?

Statistical evaluation

How « big » is the surprise to observe 18 occurrences when we expect 2.95 ?

- **at each position in the sequence, there is a probability *p*** that the word starting at this position is ACGTGA
- we consider ***n*** positions
- what is the probability that ***k*** of these ***n*** positions correspond to ACGTGA ?
- **Application** : *p* = 3.4e-4 (intergenic frequencies)
  n = 9000 position
  x = 18 observed occurences

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

Binomial distribution to measure the "surprise"

Statistical evaluation : significance

- We observe x occurrences of a word. Is this word significantly
  - Over-represented ?
  - Under-represented ?
- Choice of a scoring scheme
  - Which theoretical distribution should we use to score this significance ?

Other scoring schemes

Several statistics can be used to score the significance of the observed number of occurrences

- **Ratio** $r = CW / EW$

⇒ overestimates the importance of words with weak expected frequencies, no correction for self-overlapping patterns

⇒ Never use the observed/expected ratio to estimate over/under representation !

- **Log likelihood** $K = FW \ln(FW / PW)$

⇒ no estimation of the P-value

- **Binomial distribution**

⇒ no direct correction for self-overlapping patterns

- **Poisson distribution**
- **Compound Poisson**

⇒ See « DNA,words and model : Statistics of Exceptional Words » Schbath & Robin

Statistical evaluation

seq	identifier	exp_freq	occ	exp_occ	occ_P	occ_E
cacgtg	cacgtg|cacgtg	0.0001569968432	17	1.47	5e-13	1.0e-09
acgtga	acgtga|tcacgt	0.0003355962588	18	3.15	7.3e-09	1.5e-05
ccacag	ccacag|ctgtgg	0.0002365577659	14	2.22	1e-07	2.1e-04
gccaca	gccaca|tgtggc	0.0002897084237	15	2.72	2e-07	4.1e-04
actgtg	actgtg|cacagt	0.0003762020409	16	3.53	1e-06	2.1e-03
cgtgca	cgtgca|tgcacg	0.0002325962261	11	2.18	1.8e-05	3.8e-02

- ***p-value*** : what is the risk you take by rejecting the null hypothesis for one particular event (i.e. consider it to be significant while this is false)

- but you are testing 2080 possible hexanucleotides ("*multiple testing*") for each position !
- if you are taking 2080 times a risk of *p*=1e-7, on average, in 2080*1e-7=2.1e-4 of these cases, you will be wrong → ***E-value***

Motif discovery using word counting

Idea:
motifs corresponding to binding sites are generally repeated in the dataset
→ capture this statistical signal

- **Algorithm**
  - count occurrences of **all k-mers** in a set of related sequences (promoters of co-expressed genes, in ChIP bound regions,...)
  - estimate the **expected number of occurrences** from a background model
    - empirical based on observed k-mer frequencies
    - theoretical background model (Markov Models)
  - **statistical evaluation of the deviation observed** (P-value/E-value)
  - **Select all words above a defined threshold**

Threshold

$$E\text{-}value = P(X >= x) * T$$

$$sig = -\log_{10}(E\text{-}value)$$

Where

T is the number of tested words

- Takes into consideration the dependency of the threshold on word length
  - Different number of possible words T depending on k-mer

- Provides **an intuitive perception** of the level of over-representation

  sig > 0 1 such word at random in each sequence set

  sig > 1 1 such word expected every 10 sequence sets

  sig > 2 1 such word expected every 100 sequence sets

  ...

- This index is very convenient to interpret : higher valuescorrespond to exceptional patterns.

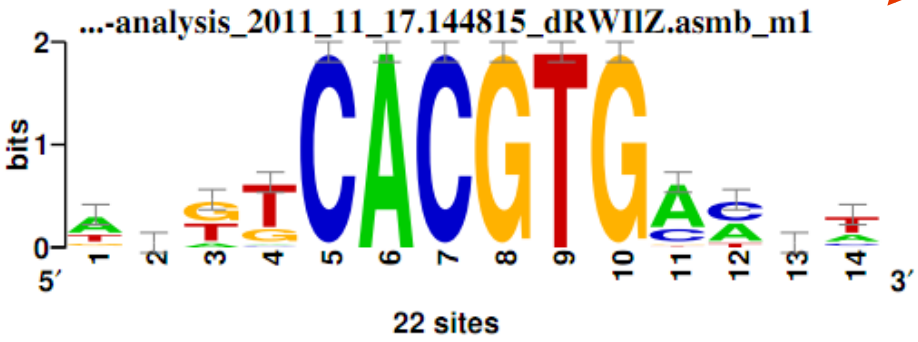
  A significance of 0 corresponds to an E-value of 1.

  A significance of 2 to an E-value of 1e-2 (i.e. one expects no more than 0.01 false positives in the whole collection of patterns).

Assembling overlapping words

Warning : the words are already a result !!!

seq	identifier	exp_freq	occ	exp_occ	occ_P	occ_E
cacgtg	cacgtg|cacgtg	0.0001569968432	17	1.47	5e-13	1.0e-09
acgtga	acgtga|tcacgt	0.0003355962588	18	3.15	7.3e-09	1.5e-05
ccacag	ccacag|ctgtgg	0.0002365577659	14	2.22	1e-07	2.1e-04
gccaca	gccaca|tgtggc	0.0002897084237	15	2.72	2e-07	4.1e-04
actgtg	actgtg|cacagt	0.0003762020409	16	3.53	1e-06	2.1e-03
cgtgca	cgtgca|tgcacg	0.0002325962261	11	2.18	1.8e-05	3.8e-02
aactgt	aactgt|acagtt	0.0006168655788	17	5.78	0.00011	2.4e-01
agtcat	agtcat|atgact	0.0005039616969	15	4.73	0.00012	2.6e-01
tagtca	tagtca|tgacta	0.0004613751449	14	4.33	0.00017	3.5e-01
agccac	agccac|gtggct	0.0002599968758	10	2.44	0.00023	4.7e-01
cgtgac	cgtgac|gtcacg	0.0001695417189	8	1.59	0.00025	5.2e-01
cgcgca	cgcgca|tgcgcg	0.0001715224888	8	1.61	0.00027	5.6e-01
acgtgc	acgtgc|gcacgt	0.0002276443015	9	2.13	0.00038	7.9e-01
gactca	gactca|tgagtc	0.0002319359695	9	2.18	0.00043	9.0e-01



Word assembly to form longer motifs and matrices

```
;assembly # 1    seed: cacgtg
;    alignt          rev_cpl
gtcacg....       ....cgtgac
.tcacgt...       ...acgtga.
..cacgtg..       ..cacgtg..
...acgtga.       .tcacgt...
....cgtgac       gtcacg....
gtcacgtgac       gtcacgtgac

;assembly # 2    seed: ccacag
;    alignt          rev_cpl
agccac....       ....gtggct
.gccaca...       ...tgtggc.
..ccacag..       ..ctgtgg..
...cacagt.       .actgtg...
....acagtt       aactgt....
agccacagtt       aactgtggct

;assembly # 3    seed: cgtgca
;    alignt          rev_cpl
gtcacg....       ....cgtgac
.tcacgt...       ...acgtga.
..cacgtg..       ..cacgtg..
...acgtgc.       .gcacgt...
....cgtgca       tgcacg....
gtcacgtgca       tgcacgtgac
```

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*
$CGGn_5wn_5CCG$	*Gal4p*

With the GAL family, the program returns a single pattern.

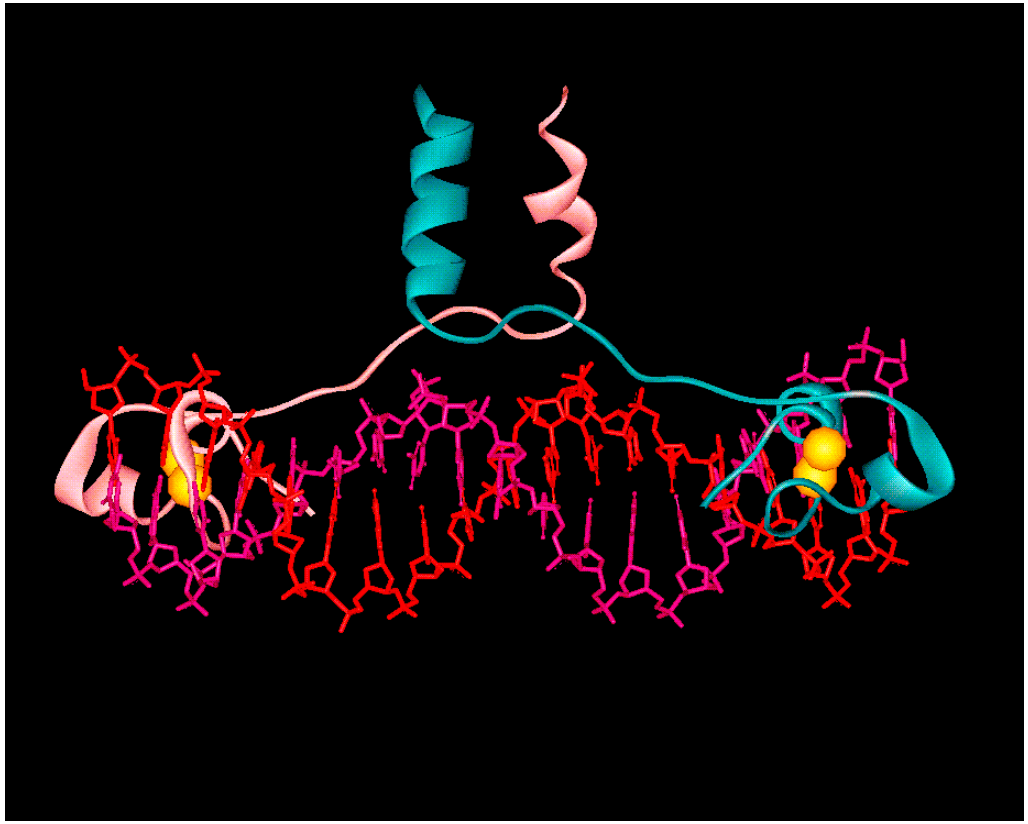
The significance of this pattern is very low.

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 ?

Spaced motif (dyads)

DNA/protein interface of the yeast transcription factor Gal4p

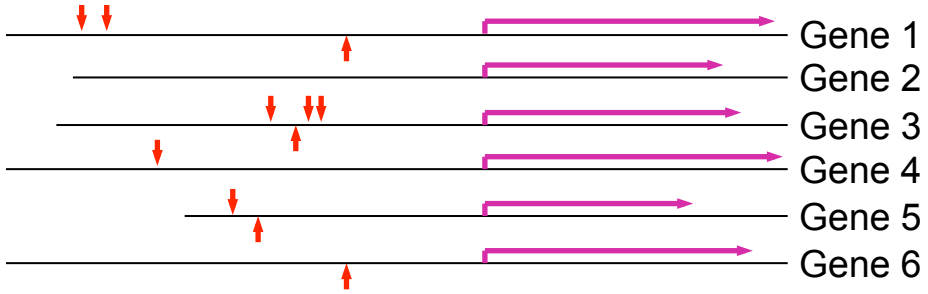


CGG n11 CCG

dyad = pairs of words separated by a spacer

Motif discovery

1 - Understand what is a motif discovery problem

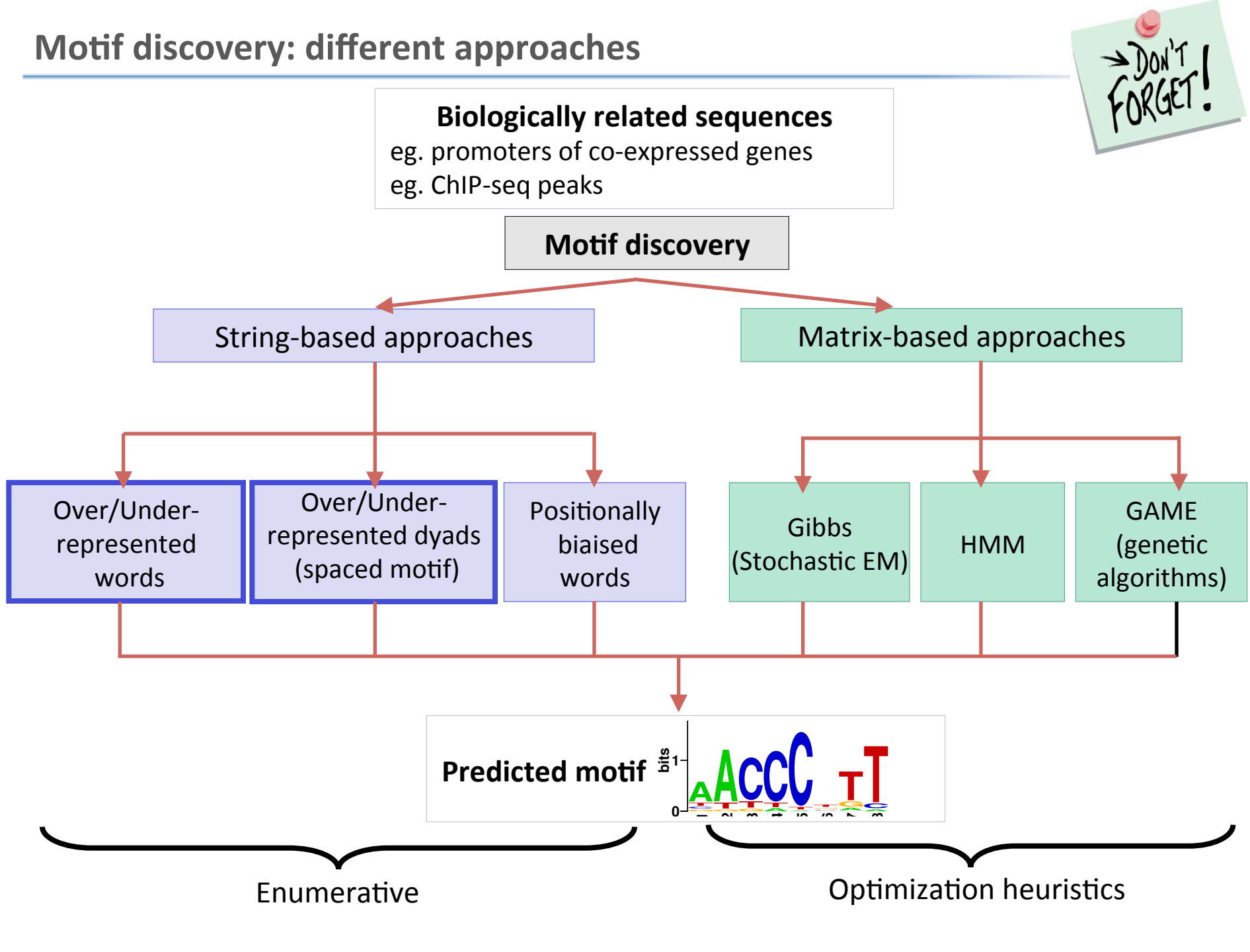


2 – Motif discovery approaches

- Word counting
- Gibbs sampling => for after matrices will be introduced

3 – Important parameters

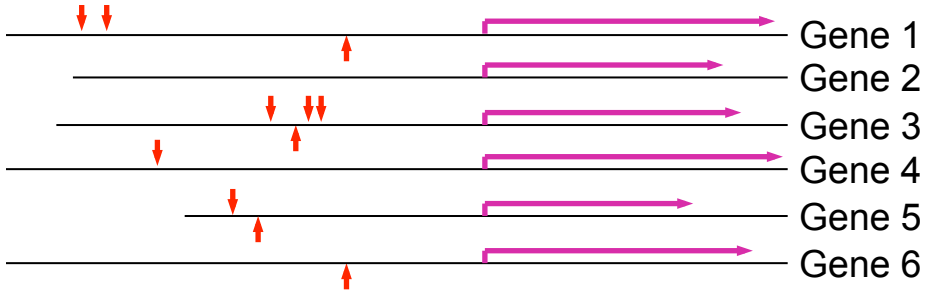
Motif discovery: different approaches



Aim of the course

Motif discovery

1 - Understand what is a motif discovery problem



2 – Motif discovery approaches

- Word counting
- Gibbs sampling

3 – Important parameters

Important parameters

- **Size of upstream sequences**
  - organism-dependent : -400 to +50bp bacteria, -800 to -1 bp fungi
  - in metazoan, regulatory regions are located several kbs to several Mb !!

- **Size of the clusters**
  - Problem of signal/noise ratio.

- **Background**

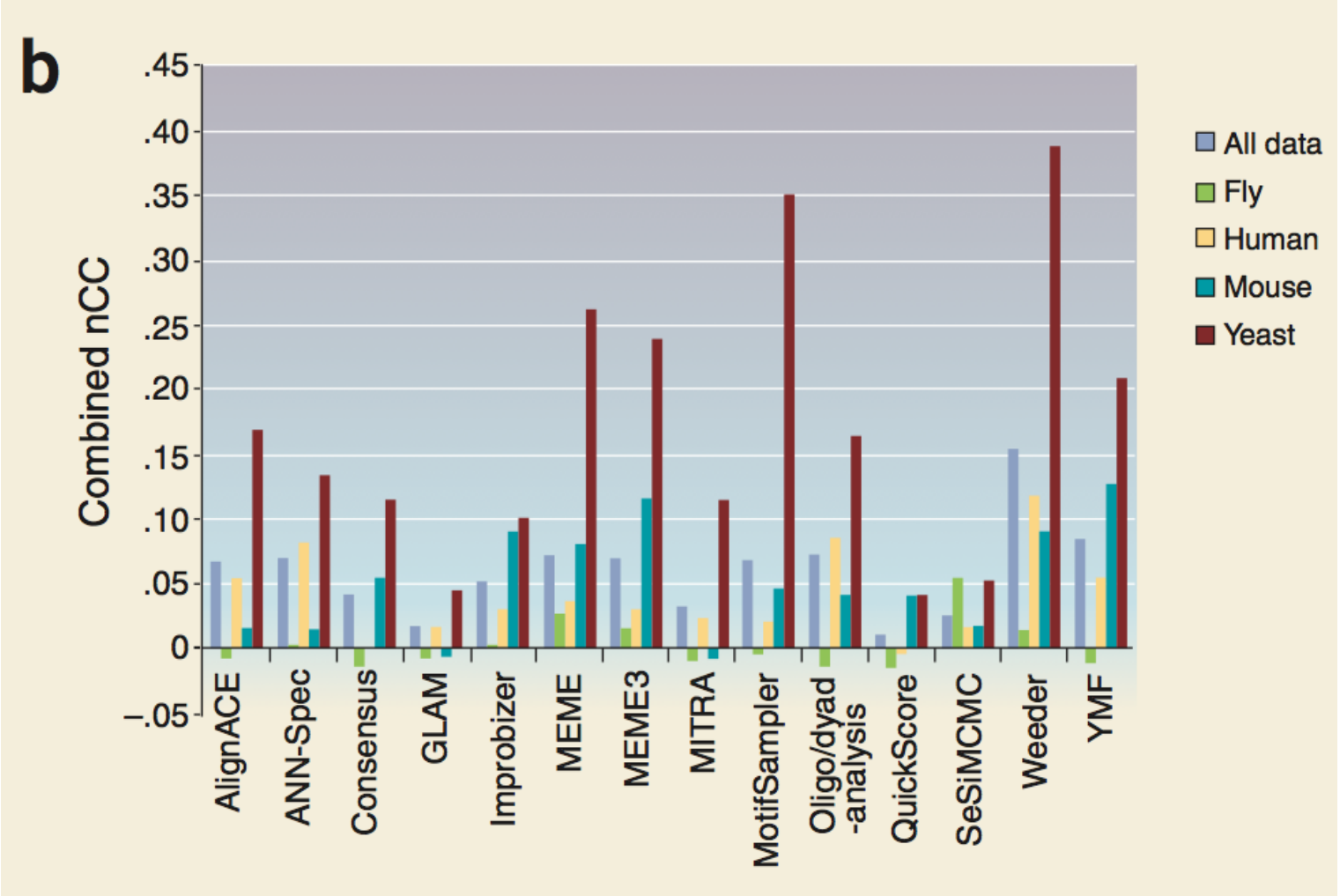
  - problem of heterogeneity of sequences in **vertebrates**. String-based motif discovery yields poor results when using upstream regions of clusters of genes. However, the same approaches provides good results in ChIP-seq datasets

  - Choice of a model :

    Markov chain : on basis of subword frequencies

    External reference (e.g. word frequencies observed in the whole set of upstream sequences)

Pattern-discovery tools poorly perform in human compared to yeast



Tompa et al, Assessing computational tools for the discovery of TFBS, Nat biotech 2005

Technicalities of word counting

- **Self-overlapping words**

  - Stretches of repetitive sequences can **bias countings**
  - Probability of further occurrences of a repetitive motif is dependent of previous occurrences
  - **Solution** : discard overlapping occurrences of the **same** k-mer

Counting all occurrences → 6

```
ATATATATATATATATAT
ATATAT
  ATATAT
    ATATAT
      ATATAT
        ATATAT
          ATATAT
```

Discarding overlapping matches → 2

```
ATATATATATATATATAT
ATATAT
      ATATAT
```

Technicalities of word counting

- **duplicated regulatory regions**

  - Over-representation statistics rely on the independence of successive positions
  - Cases of large sequence duplications
    - recent duplication of a gene along with its upstream sequence
    - intergenic region located between two divergently transcribed genes
      → the **same sequence is taken twice**
  - Bias
    - all the words included in duplicated regions are over-estimated
  - Treatment
    - **sequences have to be purged** before any analysis

Technicalities of word counting

- TFs can bind on **both strands**
- however, we only work with single stranded sequences
- if the BS consensus is **ATTTGCA** on the reference strand, **ACGTTTA** corresponds to the same BS, but on the reverse strand !
- hence 4^6 = 4096 6-mers, but only **2080 pairs of 6-mers** must be considered