4. Motif discovery

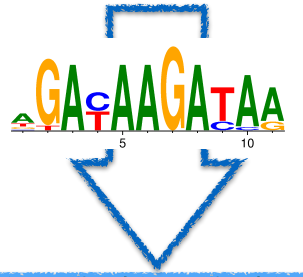
Finding unknown motifs in sequences

IPMB
Institut für Pharmazie und
Molekulare Biotechnologie

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Pattern matching vs. Motif discovery

Consider a particular TF of interest (Evi1)



Where are TFBS ?
What are the potential target genes ?

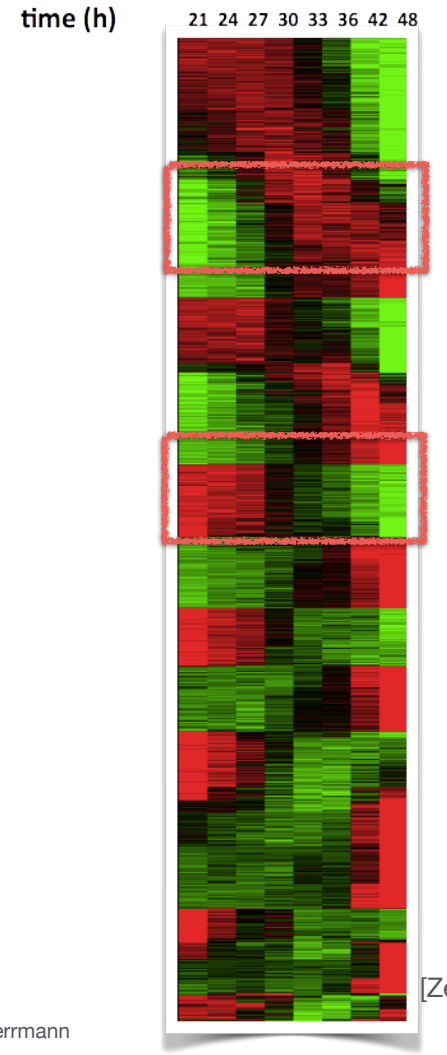
Pattern Matching

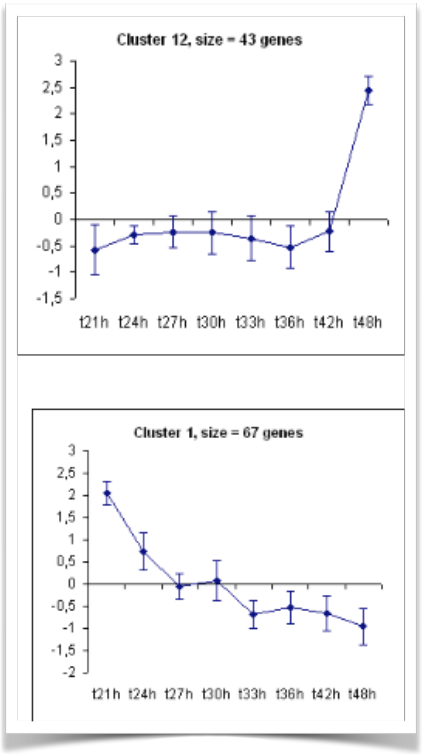
Consider a set of regions of interest (e.g. promoters of co-expressed genes)

What is their potential common regulator ?

Motif discovery

Motif discovery for co-expressed genes



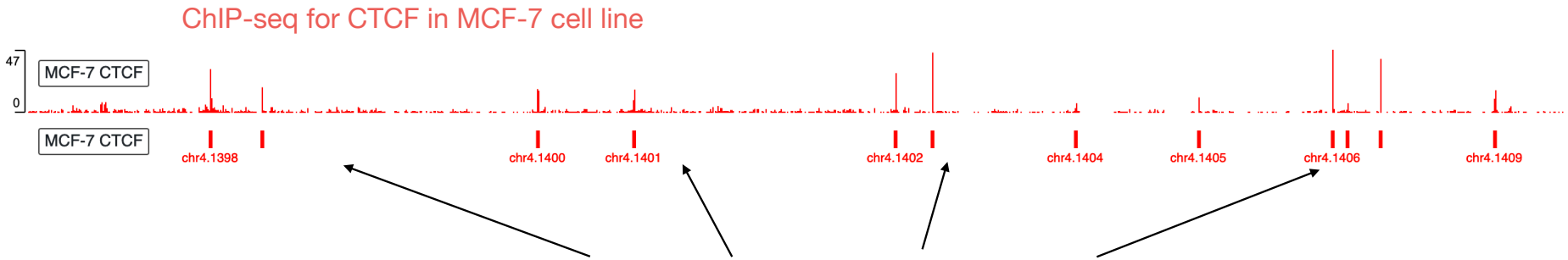


- clusters of co-expressed genes during cardiac remodelling in Drosophila
- ***Are these cluster of genes co-regulated ?***
- ***If so, what is their common regulator ?***

[Zeitouni (2007)]

ChIP-seq

IPMB
Institut für Pharmazie und Molekulare Biotechnologie
UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386



all these ChIP-seq peak sequences should contain a common motif (CTCF, but maybe other motifs too?)

```
>MA0139.1       CTCF    1
acttgcttgtaagcatgggggggggggtttgatgtgggttgggtttttatgtaCTACAGGTGGTcaagtatttatggtaccgtacaatattcatggtggctggcagtaatgtacgaaatacatagcggttgttgatgggtgagtcaatacttgggtggtaccca
>MA0139.1       CTCF    2
tatgtactgttaagggtgggtaggtttgttggtatcctagtgggtgaggggtggctttggagttgcagttgatgtgtgatagttgaaggttgattgctgtacttgcttgtaagcatgggggggggggggtttgatgtgggttgggtttttatgtaCTACAGGTGGT
>MA0139.1       CTCF    3
ggatgacccccctcagatagggtcccttgaccaccatcctccgtgaaatcaatatcccgCACAAGAGTGCtactctcctcgctccgggcccataacacttgggggtagctaaagtgaactgtatccgacatctggttcctacttcagggccataaagcctaaat
>MA0139.1       CTCF    4
atggcttccatgaggctctaagctccccaagctgggttcaccttccgtcgctacttcgcccttttcttccaggttgtcctgggcactgatggtCACCAGGTGGCgatgttgcccaggaggacgctctgttgccatctcatggccaaaattcaacactacactcttc
>MA0139.1       CTCF    5
aatccagggcaccgagaaagtccagcagcgcctgagcctgctgcacttgcagtagcagggactggcCGGGAGATGGCagccctgtcccaggtcgggtggtgcctcagccgcagttcctgtcacggccatcgggaggcagtgcaggaacactcactgatgtgtagc
>MA0139.1       CTCF    6
nnnnnnnnnnnnnnnnnnnnnnnnnnnnaggagagaagacaagattctgggtgaacaggaacagtaagttggctcaggactagggtaagccggtaggaggcgccatttctccacaatccagcggcagcatggccctctgaagcCACTTGAGGGCgtgttacgtccga
>MA0139.1       CTCF    7
aggtgccggggactgaggctaccctgctgtgggctgcagcttgctctcttctcaacttcccagctcccacctgggggcgccagagggcacatatactgctgtactcgtttgtgcCACCAGAGGGCagcctccgcacatcgatctctactgcttccgagaactgt
>MA0139.1       CTCF    8
ggacactacccgctgggctacccagaccggagcgcggcgcgagaggcgtgagtcgatgacctctagcgagggaggccacctagcggccaggccaagtctggaggatacttccgttgggcCGCCAGGGGGCagtaacgccagatcatctgtctcccggaagtaatg
>MA0139.1       CTCF    9
cttggcaagggggcgccgctcaccagctggggtggatgaggagcgtgaagacagctcgctccagtggccgcgggggggcgctcgcccactcagagcatggcCGCCAGATGGCgccggctcgcccgctctccacaaggctcttccaaccccttaccctgtcccacac
>MA0139.1       CTCF    10
accgagtccttgtccccagcagaaaccgtggtccctttgctgttccagatgctctggaccctgggtactggctccccctggcggtcatgataggaagcgcccttccctctggctcCTGTAGAGGGCgtcgcggatgcattgatccttttagttccttcactgtcc
>MA0139.1       CTCF    11
tcggcccaaacctctggcggccgcgtcactcggctccaccagtctgtgccgcctgcaagcgcccctagcggcctcgcagccatcgcgcctcagcagcgccgtggcgtcctgttcggcagtgtcCATTAGAGGGCgccagcgcgcctcagaagagggaccgtgacg
>MA0139.1       CTCF    12
cagtgcctaagtgtctagagagctgagtgggttgcctcggatcggccaggcctggatatcagcgaacagatgcggatttcCACCAGGGGGCgctaggacgccacgaccttcactatgtgcaccaggggcgctgccgcccgccttggccatcccgggctgggaa
>MA0139.1       CTCF    13
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnngcagcggaggtggcCGCCAGGGGACgctcgggagccccgtgggccgtcaggctgtggtgcagcagctcctcctcctggctggagccgaaactgcacaggccgcacttccagggc
>MA0139.1       CTCF    14
ttacggatcggcttactccgcggagttggcctcatttctgcagtcggcgctccctgtagtttctcctctcgaaCGCCAGGTGGAgcaaccggccggataccgccacagccctggcaggcggcgctgtgatgcctgagctgatcctgtatgttgcaatcactctat
>MA0139.1       CTCF    15
ttcagcagcccagcggatttcagcgtttgacaacctggaagttctgctgccccctcacggtgggagaaaagcggtgggtcctgcacaggccgaagaacCACTAGAGGGCagcagcacacgggctccataagacttgctgcgtgtcgaccctggctcatacagtgc
>MA0139.1       CTCF    16
cccgccaagccggccccagcaagccgcaggaagccgtccgctgaacctcaggttcttgggttttcctatgccccctggaggcatctgcgaacctctcggcatcgctccaccacacctcccacCAGCAGGTGGCactgttgccatctccataacatggactggcct
```

Gene regulation

Follow their sight…

… you'll find their common regulator !

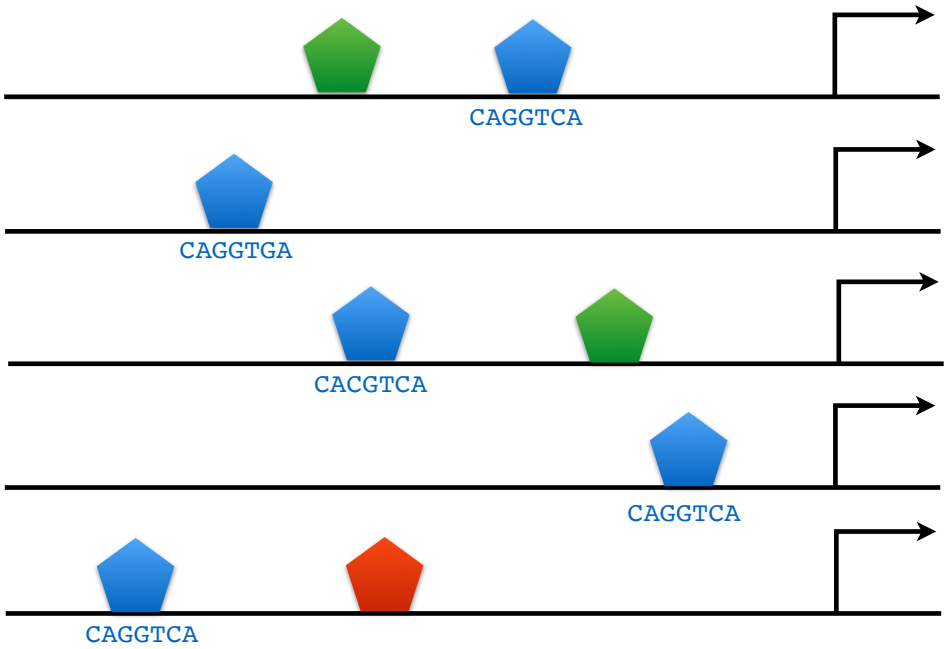
Motif discovery using word counting

IPMB Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

motifs corresponding to binding sites are **repeated**

→ **capture this statistical signal**



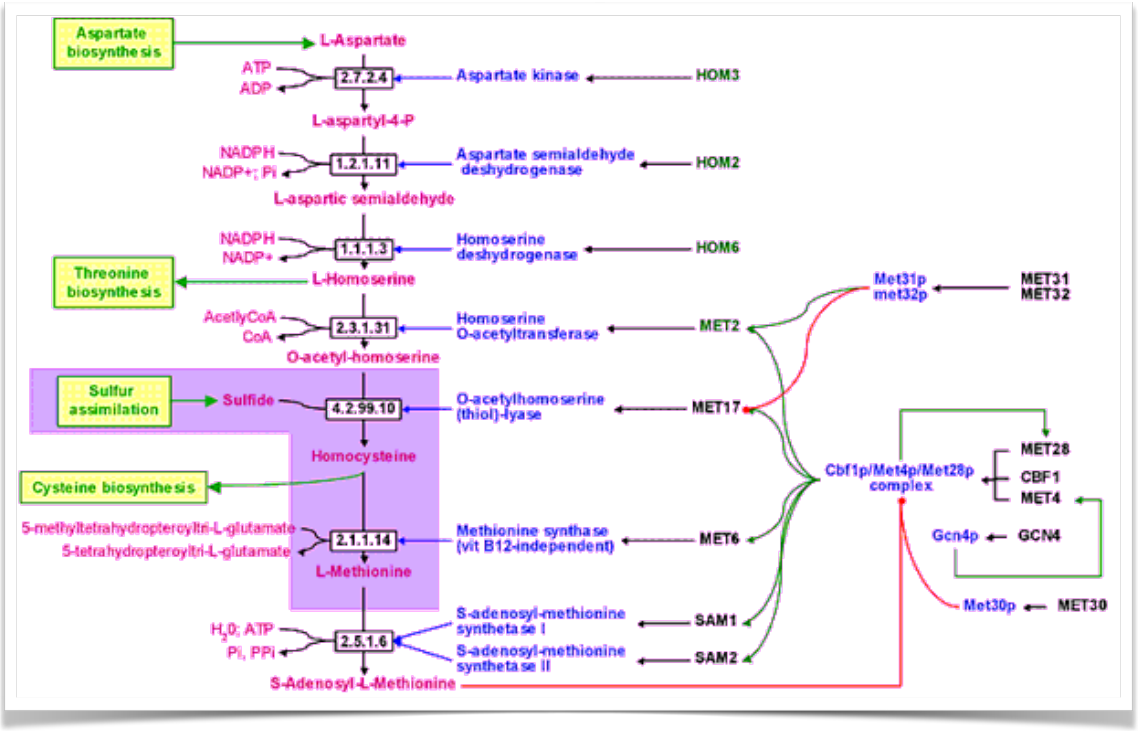
Motif discovery using word counting

- **Algorithm**
  1. count **observed number of occurrences** of all k-mers in a set of related sequences (promoters of co-expressed genes, in ChIP bound regions,...)
  2. build a theoretical **background model** (e.g. Markov Model)
  3. estimate the **expected number of occurrences** in background model
  4. **statistical significance** of the deviation observed (Observed / Expected)

Example

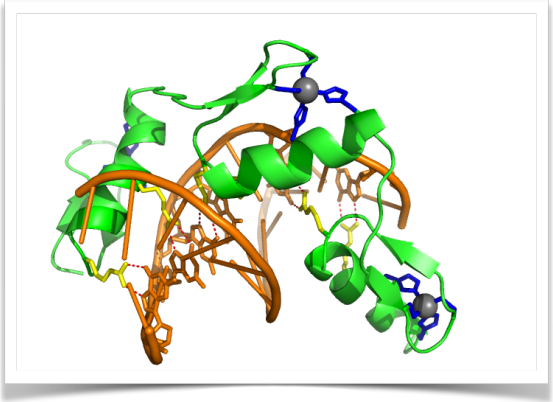
- Are they co-regulated ?
- Do they share common regulatory motifs ?
- Principle
  - Count occurrences of k=6 mers in the 800 bp upstream of the TSS (!! on both strands !!)
  - 9000 possible positions
  - compare observed and expected number of occurences

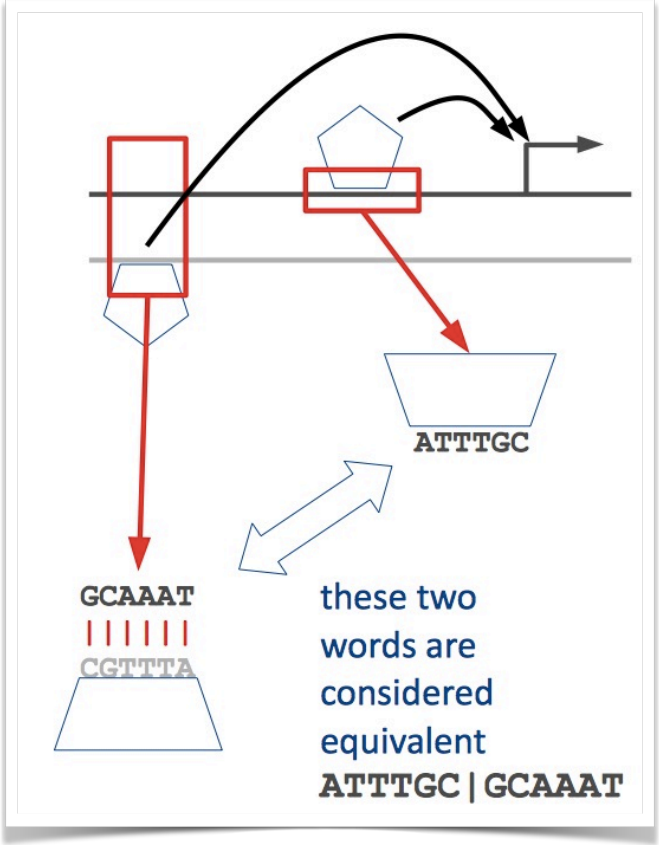
Methionine synthesis pathway in *S. cerevisae*



19 genes from *S. cerevisiae* involved in methionine biosynthesis pathway

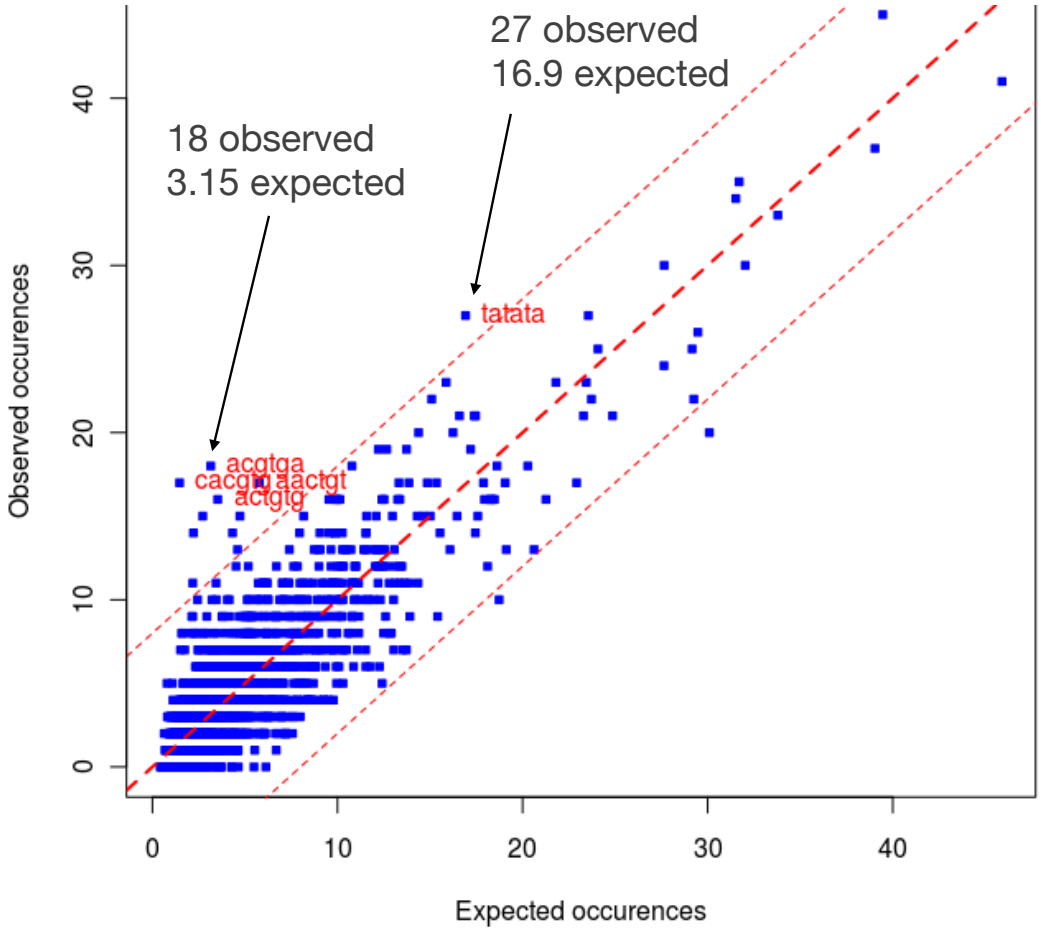
Word counting technicalities





- TFs can bind on both strands
- however, we only work with single stranded sequences
- if the binding site consensus is `ATTTGCA` on the reference strand, `TGCAAAT` corresponds to the same binding site, but on the reverse strand !
- hence $4^6 = 4096$ 6-mers, but only 2080 pairs of 6-mers must be considered
  (why not 2048?)

Motif discovery using word counting



How to evaluate expected number of occurrences ?

Could these be statistical fluctuations ?

Defining a background model

IPMB Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Estimated frequency of ACGTGA in S. cerevisae ?

- **Possibility 1** : based on the actual, observed frequency of this word in the whole genome
  - all **intergenic sequences** in the genome:
    1026 occurrences for 3,310,685 positions → **p = 3.1e-4 (2.79)**

  - only **promoter sequences**:
    921 occurrences for 2,804,964 positions → **p = 3.3e-4 (2.95)**

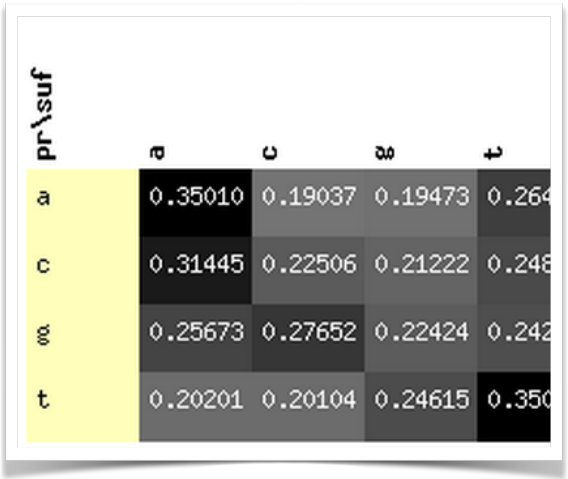
Defining a background model

Estimated frequency of ACGTGA in S. cerevisae ?

- **Possibility 2** :estimate the frequency using a statistical model
  - *Bernouilli model (MM0)*: p(A), p(C), p(G), p(T)

$$p(ACGTGA) = p(A)^2\, p(C)\, p(G)^2\, p(T) \rightarrow p = 3.94e\text{-}4\ (3.70)$$

  - *Markov models*



pr\suf	a	c	g	t
a	0.35010	0.19037	0.19473	0.264
c	0.31445	0.22506	0.21222	0.248
g	0.25673	0.27652	0.22424	0.242
t	0.20201	0.20104	0.24615	0.350

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) p(G|AC) p(T|CG) p(G|GT) p(A|TG)

Markov model order 3 : p = 7.4e-4 (6.96)
p(ACGTGA) = p(ACG) p(T|ACG) p(G|CGT) p(A|GTG)

Defining a background model

Estimated frequency of ACGTGA in S. cerevisae ?

	Method	Frequency (p)	Occurences
Observation	**observed in the dataset**		**18**
Estimations	intergenic frequency	3.25e-4	3.05
	promoter frequency	3.35e-4	3.15
	Markov order 0	3.94e-4	3.70
	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

Statistical evaluation

- At each position in the sequence, there is a probability **p** that the word starting at this position is ACGTGA
- I 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
  - k = 18

Statistical evaluation

- Using the **binomial distribution**
  - number of possible configurations ("In how many ways can I place k=8 words in n=35 positions ?")

$$N = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

  - probability of each configuration

$$P = p^k (1-p)^{n-k}$$

  - probability of the observed configuration

$$P = \binom{n}{k} p^k (1-p)^{n-k}$$

  - probability of observing this number or more

$$P(x \geq k) = 1 - \sum_{i=0}^{k-1} P(i)$$

n=35 *k=8*

0 0 0 0 **1** 0 0 **1** 0 0 0 0 0 **1** 0 0 **1** 0 0 0 0 **1** **1** 0 0 0 0 **1** 0 0 0 0 **1** 0 0

occurence of word

Statistical evaluation

	Method	Frequence (p)	Occurences	P-value
Observation	**observed in the dataset**		**18**	
Estimations	intergenic frequency	3.25e-4	3.05	**4.6e-9**
	promoter frequency	3.35e-4	3.15	**7.3e-9**
	Markov order 0	3.94e-4	3.70	**8e-8**
	Markov order 1	3.70e-4	3.48	**3.2e-8**
	Markov order 2	5.19e-4	4.87	**3.8e-6**
	Markov order 3	7.42e-4	6.96	**3.4e-4**
	promoter frequency in **human**	1.63e-4	1.53	**7.9e-14**

Example

seq	identifier	exp_freq	occ	exp_occ	Pvalue	E-value	occ_sig	rank
cacgtg	cacgtg|cacgtg	0.0001569968432	17	1.47	5e-13	1.0e-09	8.98	1
acgtga	acgtga|tcacgt	0.0003355962588	18	3.15	7.3e-09	1.5e-05	4.82	2
ccacag	ccacag|ctgtgg	0.0002365577659	14	2.22	1e-07	2.1e-04	3.68	3
gccaca	gccaca|tgtggc	0.0002897084237	15	2.72	2e-07	4.1e-04	3.39	4
actgtg	actgtg|cacagt	0.0003762020409	16	3.53	1e-06	2.1e-03	2.68	5
cgtgca	cgtgca|tgcacg	0.0002325962261	11	2.18	1.8e-05	3.8e-02	1.42	6

- **P-value** : what is the risk you take by rejecting the null hypothesis for one particular event (i.e. consider it to be significant)
- but you are testing 2080 possible hexanucleotides *("multiple testing")*
- if you are taking 2080 times a risk of p=1e-4, on average, in 2080*1e-4=0.208 of these cases, you will be wrong
  → ***E-value***

From words to logo

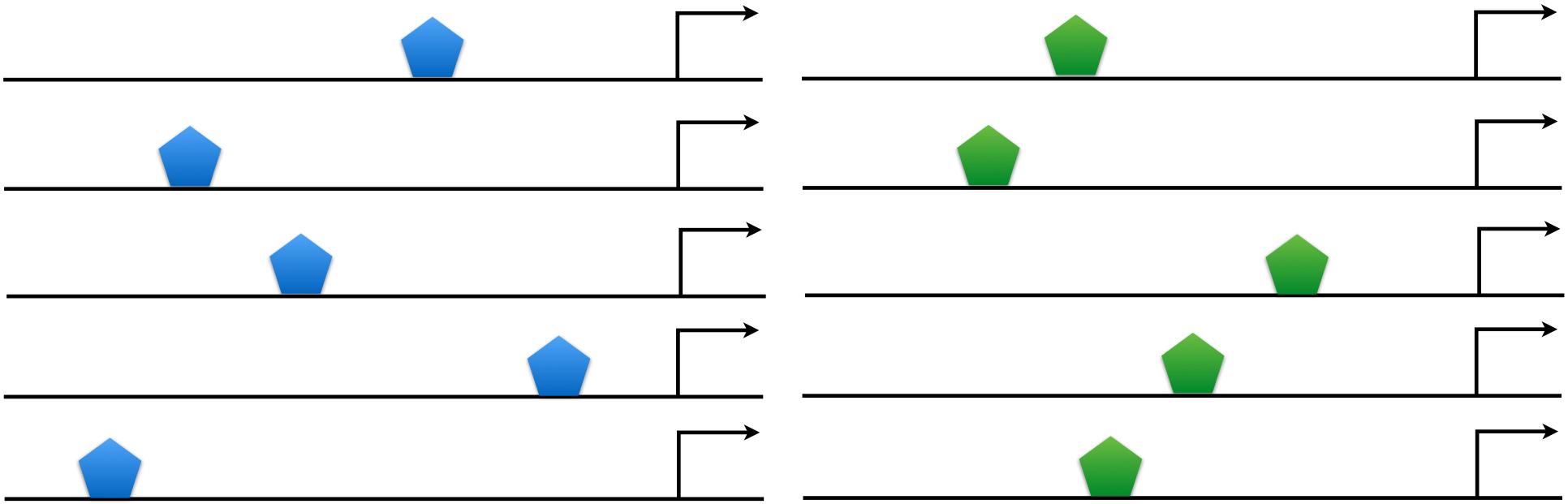
seq	identifier	exp_freq	occ	exp_occ	occ_P	occ
cacgtg	cacgtg|cacgtg	0.0001569968432	17	1.47	5e-13	1.0e
acgtga	acgtga|tcacgt	0.0003355962588	18	3.15	7.3e-09	1.5e
ccacag	ccacag|ctgtgg	0.0002365577659	14	2.22	1e-07	2.1e
gccaca	gccaca|tgtggc	0.0002897084237	15	2.72	2e-07	4.1e
actgtg	actgtg|cacagt	0.0003762020409	16	3.53	1e-06	2.1e
cgtgca	cgtgca|tgcacg	0.0002325962261	11	2.18	1.8e-05	3.8e
aactgt	aactgt|acagtt	0.0006168655788	17	5.78	0.00011	2.4e
agtcat	agtcat|atgact	0.0005039616969	15	4.73	0.00012	2.6e
tagtca	tagtca|tgacta	0.0004613751449	14	4.33	0.00017	3.5e
agccac	agccac|gtggct	0.0002599968758	10	2.44	0.00023	4.7e
cgtgac	cgtgac|gtcacg	0.0001695417189	8	1.59	0.00025	5.2e
cgcgca	cgcgca|tgcgcg	0.0001715224888	8	1.61	0.00027	5.5e
acgtgc	acgtgc|gcacgt	0.0002276443015	9	2.13	0.00038	7.9e
gactca	gactca|tgagtc	0.0002319359695	9	2.18	0.00043	9.0e


```
;assembly # 1    seed:
;    alignt         rev
gtcacg....         ....cg
.tcacgt...         ...acg
..cacgtg..         ..cacg
...acgtga.         .tcacg
....cgtgac         gtcacg
gtcacgtgac         gtcacg

;assembly # 2    seed:
;    alignt         rev
agccac....         ....gt
.gccaca...         ...tgt
..ccacag..         ..ctgt
...cacagt.         .actgt
....acagtt         aactgt
agccacagtt         aactgt

;assembly # 3    seed:
;    alignt         rev
gtcacg....         ....cg
.tcacgt...         ...acg
..cacgtg..         ..cacg
...acgtgc.         .gcacg
```

Motif discovery using expect.max.

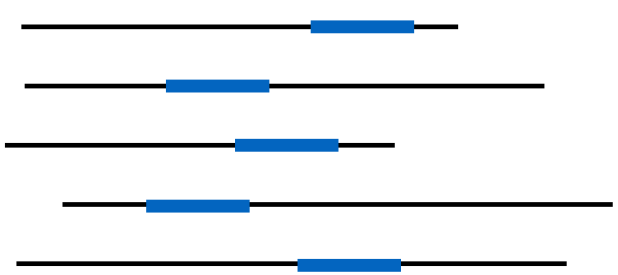


- What is the most likely situation, given the sequences?
- Can we guess the most likely motif, given the sequences?

Motif discovery

PFM θ_1

PFM θ_2



sequences X

?

$P(X \mid \theta_1)$ < > $P(X \mid \theta_2)$

likelihood of the sequences X, if they all contain motif θ_1

likelihood of the sequences X, if they all contain motif θ_2

Motif discovery using Likelihood Maximization

Maximize the **likelihood** that a set of sequences share common motifs rather than not

→ **search the motif that maximizes this likelihood**

→ ***Expectation-Maximization***

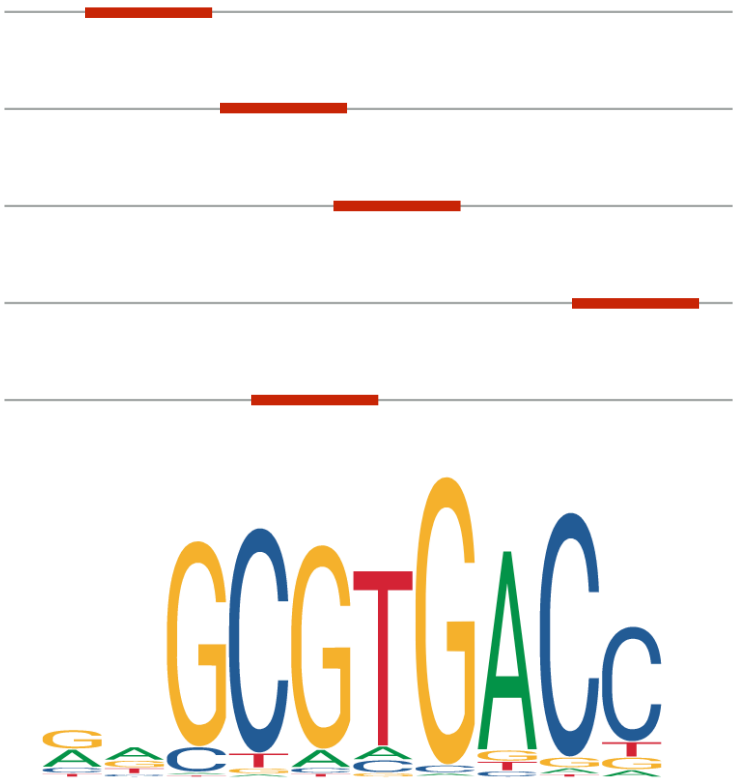
$$\max_{\theta} P(X \mid \theta)$$

sequences

motif

EM for motif discovery

- we have a **dataset** ***(X)***
  of sequences possibly sharing common motifs
- ... but there is also **missing data** ***(Z)***:
  we don't know the position of the motif in the sequences
- ... and we have unknown parameters ***(θ)***
  position weight matrix



Principles of EM

- Back to our 2 biased **blue** and **red** coins …
- We want to estimate the **emission probabilities** by observing **5 series of 10 tosses** (each sequence obtained with one coin) (= X)

- What we don't know:
  - emissions probabilities (= θ)
  - which series was played with which coin (= Z)

Principles of EM

If we know the color of the coin used …

Sequence	Prob Red	Prob Blue	Number of 1's	Number of 2's	Count for Red	Counts for Blue
1 2 2 2 1 1 2 1 2 1	1	0	5	5	5 1's 5 2's	
1 1 1 1 2 1 1 1 1 1	0	1	9	1		9 1's ; 1 2's
1 2 1 1 1 1 1 2 1 1	0	1	8	2		8 1's ; 2 2's
1 2 1 2 2 2 1 1 2 2	1	0	4	6	4 1's ; 6 2's	
2 1 1 1 2 1 1 1 2 1	0	1	7	3		7 1's ; 3 2's
					9 1's ; 11 2's	**24 1's ; 6 2's**

Probability of getting a 1:

$\theta_{Red} = 9 / (9 + 11) = 0.45$

$\theta_{Blue} = 24 / (24 + 6) = 0.8$

Principles of EM

- Back to our 2 biased **blue** and **red** coins …
- We want to estimate the emission probabilities by observing 5 series of 10 tosses ***(with unknown coin !!)***

Initialization

blue or red ?	**1 2 2 2 1 1 2 1 2 1**
blue or red ?	1 1 1 1 2 1 1 1 1 1
blue or red ?	1 2 1 1 1 1 1 2 1 1
blue or red ?	1 2 1 2 2 2 1 1 2 2
blue or red ?	2 1 1 1 2 1 1 1 2 1

- We **randomly initialize the emission probabilities** (i.e. probability that the red and blue coin yields a 1)
  $\theta^{(0)}_{Blue} = 0.6$; $\theta^{(0)}_{Red} = 0.5$
- What is the probability that the first series was made by
  - blue coin = $0.6^5 * 0.4^5 = 0.00079$
  - red coin = $0.5^5 * 0.5^5 = 0.000976$
- We normalize the probabilities to 1
  - blue coin = 0.45
  - red coin = 0.55

Principles of EM

- First iteration -

Initial parameters : $\theta^{(0)}_{Blue} = 0.6$; $\theta^{(0)}_{Red} = 0.5$

Sequence	Prob Red	Prob Blue	Number of 1's	Number of 2's	Count for Red	Counts for Blue
1 2 2 2 1 1 2 1 2 1	0.55	0.45	5	5		
1 1 1 1 2 1 1 1 1 1	0.20	0.80	9	1		
1 2 1 1 1 1 1 2 1 1	0.27	0.73	8	2		
1 2 1 2 2 2 1 1 2 2	0.65	0.35	4	6		
2 1 1 1 2 1 1 1 2 1	0.35	0.64	7	3		

Principles of EM

- First iteration -

Initial parameters : $\theta^{(0)}_{Blue} = 0.6$; $\theta^{(0)}_{Red} = 0.5$

Sequence	Prob Red	Prob Blue	Number of 1's	Number of 2's	*Expectation step* Count for Red	Counts for Blue
1 2 2 2 1 1 2 1 2 1	0.55	0.45	5	5	2.8 1's; 2.8 2's	2.2 1's; 2.2 2's
1 1 1 1 2 1 1 1 1 1	0.20	0.80	9	1	1.8 1's 0.2 2's	7.2 1's 0.8 2's
1 2 1 1 1 1 1 2 1 1	0.27	0.73	8	2	2.1 1's 0.5 2's	5.9 1's 1.5 2's
1 2 1 2 2 2 1 1 2 2	0.65	0.35	4	6	2.6 1's 3.9 2's	1.4 1's 2.1 2's
2 1 1 1 2 1 1 1 2 1	0.35	0.64	7	3	2.5 1's 1.1 2's	4.5 1's 1.9 2's

In the expectation step, the counts are weighted by the likelihood of the configuration

Principles of EM

- First iteration -

Initial parameters : $\theta^{(0)}_{Blue} = 0.6$; $\theta^{(0)}_{Red} = 0.5$

Maximization step

Sequence	Prob Red	Prob Blue	Number of 1's	Number of 2's	Count for Red	Counts for Blue
1 2 2 2 1 1 2 1 2 1	0.55	0.45	5	5	2.8 1's; 2.8 2's	2.2 1's; 2.2 2's
1 1 1 1 2 1 1 1 1 1	0.20	0.80	9	1	1.8 1's 0.2 2's	7.2 1's 0.8 2's
1 2 1 1 1 1 1 2 1 1	0.27	0.73	8	2	2.1 1's 0.5 2's	5.9 1's 1.5 2's
1 2 1 2 2 2 1 1 2 2	0.65	0.35	4	6	2.6 1's 3.9 2's	1.4 1's 2.1 2's
2 1 1 1 2 1 1 1 2 1	0.35	0.64	7	3	2.5 1's 1.1 2's	4.5 1's 1.9 2's

11.7 1's 8.4 2's **21.3 1's 8.6 2's**

Maximization step
new estimation
for the parameters:

$\theta^{(1)}_{Red} = 11.7 / (11.7 + 8.4) = 0.58$

$\theta^{(1)}_{Blue} = 21.3 / (21.3 + 8.6) = 0.71$

Principles of EM

- Second iteration -

Parameters after 1 iteration

$\theta^{(1)}_{Blue} = 0.71$; $\theta^{(1)}_{Red} = 0.58$

Expectation step

Sequence	Prob Red	Prob Blue	Number of 1's	Number of 2's	Count for Red	Counts for Blue
1 2 2 2 1 1 2 1 2 1			5	5		
1 1 1 1 2 1 1 1 1 1			9	1		
1 2 1 1 1 1 1 2 1 1			8	2		
1 2 1 2 2 2 1 1 2 2			4	6		
2 1 1 1 2 1 1 1 2 1			7	3		

13.8 1's 10.4 2's 19.2 1's 6.6 2's

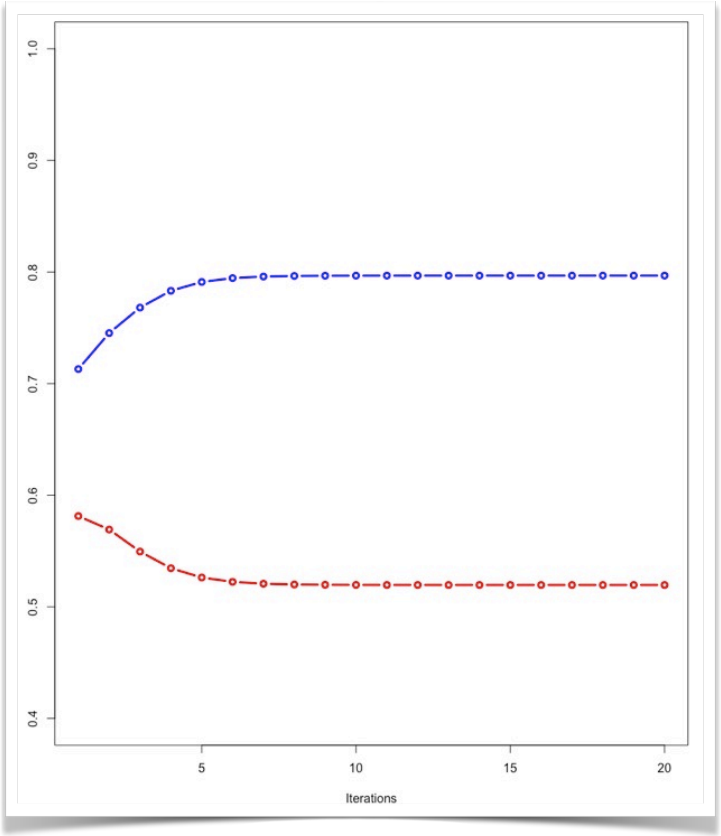
Maximization step
new estimation
for the parameters:

$\theta^{(2)}_{Red} = 13.8 / (13.8 + 10.4)$
$= 0.57$

$\theta^{(2)}_{Blue} = 19.2 / (19.2 + 6.6)$
$= 0.75$

Principle of EM

- Parameters converge after a number of iterations

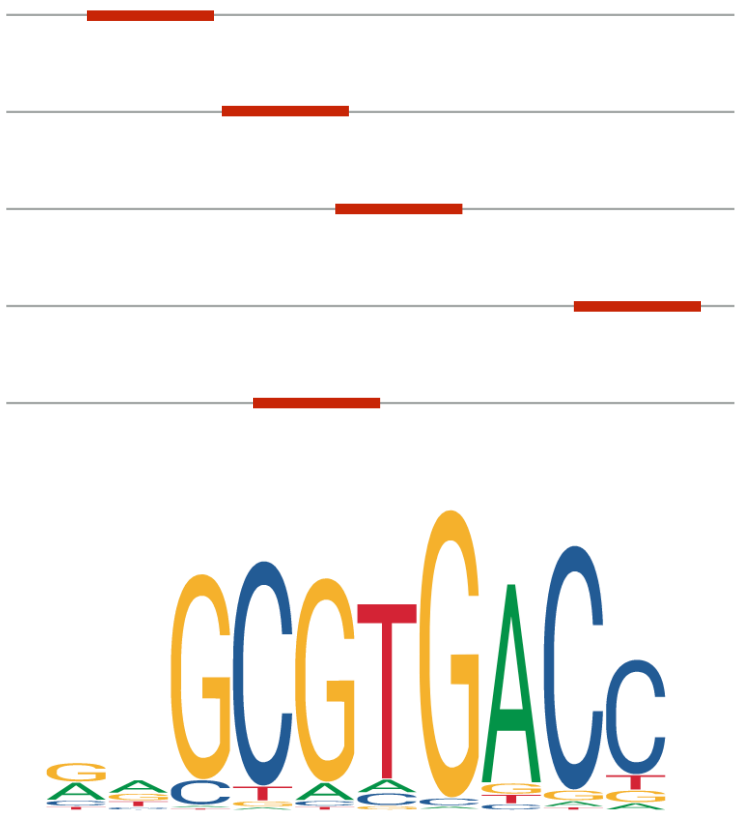


0.8 $\theta^{(i)}_{Blue}$

0.52 $\theta^{(i)}_{Red}$

EM for motif discovery

- we have a **dataset** ***(X)***
  of sequences possibly sharing common motifs
- ... but there is also **missing data** ***(Z)***:
  we don't know the position of the motif in the sequences
- ... and we have unknown parameters ***(θ)***
  position weight matrix



EM for motif discovery

- We want to maximize the likelihood of the data

$$\max_\theta P(X \mid \theta) = \max_\theta \left(\sum_Z P(X \mid Z) P(Z \mid \theta) \right)$$

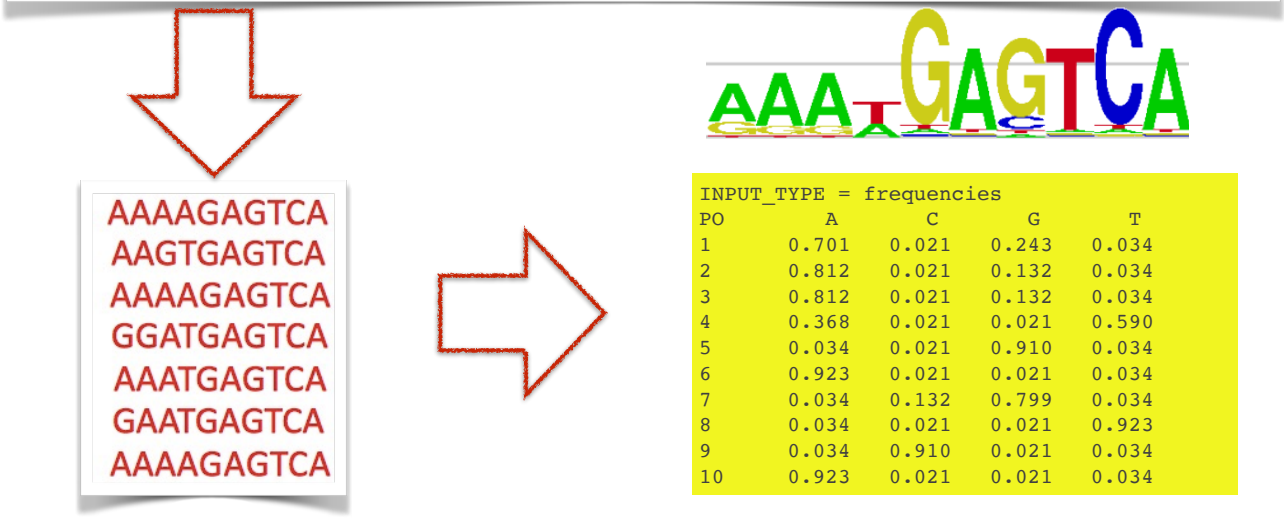
- *P(X|Z) :* likelihood of the sequences X given that they contain a motif at position Z
- *P(Z|θ)* : likelihood of the positions Z given the motif θ

EM for motif discovery

We look for θ which maximizes P(X|θ)

- if we knew Z (position of motifs), we would know θ ...

```
5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT ...HIS7
5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG ...ARO4
5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT ...ILV6
5′- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC ...THR4
5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA ...ARO1
5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA ...HOM2
5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA ...PRO3
```

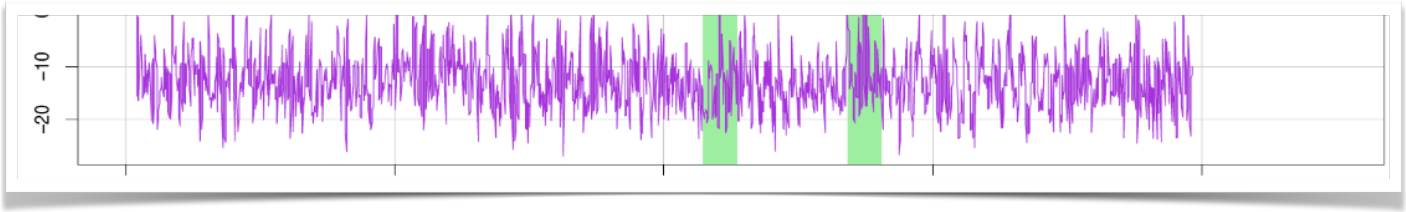



```
INPUT_TYPE = frequencies
PO        A        C        G        T
1       0.701    0.021    0.243    0.034
2       0.812    0.021    0.132    0.034
3       0.812    0.021    0.132    0.034
4       0.368    0.021    0.021    0.590
5       0.034    0.021    0.910    0.034
6       0.923    0.021    0.021    0.034
7       0.034    0.132    0.799    0.034
8       0.034    0.021    0.021    0.923
9       0.034    0.910    0.021    0.034
10      0.923    0.021    0.021    0.034
```

EM for motif discovery

We look for θ which maximizes P(X|θ)

- if we knew θ (motif matrix), we would know Z (the position) ...



highest scoring
binding sites

EM for motif discovery

2 steps

Step 1 :
suppose **we know the matrix θ**
and determine the binding
site positions Z in the sequences
P(Z|θ)

Step 2 :
suppose
we know the positions Z
of the binding sites in the
sequences
and determine the matrix θ

EM for motif discovery

2 steps

"Expectation Step"

Step 1 :
suppose **we know the matrix θ**
and determine the binding
site positions Z in the sequences
P(Z|θ)

"Maximization Step"

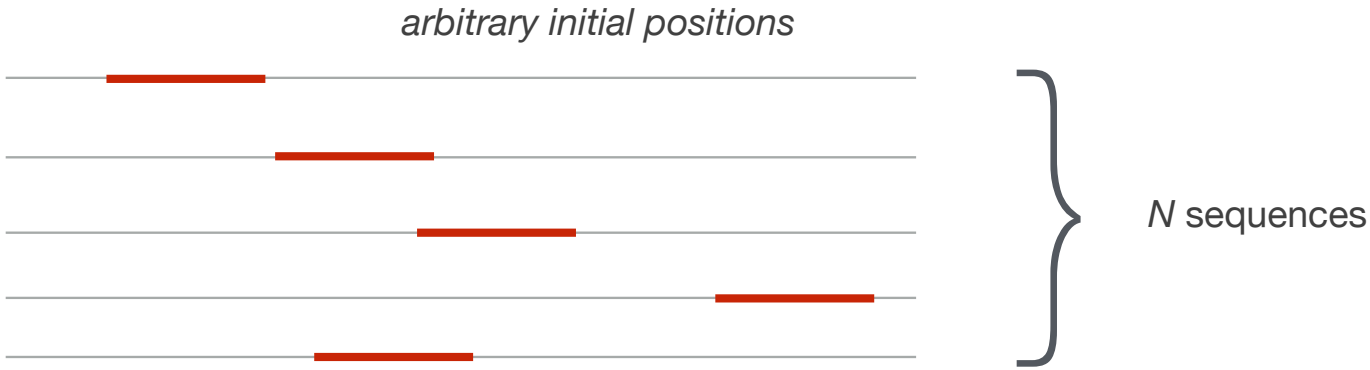
Step 2 :
suppose
we know the positions Z
of the binding sites in the
sequences
and determine the matrix θ

- each step takes as input the estimations of the previous step
- the maximization step by definition results in an increase to *L*
- converge to a local maximum of

$$\max_\theta P(X \mid \theta)$$

EM algorithm

- **Step 0** : parameter initialization + background model



initial matrix + background model

$$\theta \quad : \quad f_{ij} = \frac{n_{ij}}{N}$$

Background: p_i

a|	8	0	14	0	14	14	0	14	0	13	11
c|	1	0	0	7	0	0	0	0	3	1	0
g|	2	13	0	0	0	0	14	0	0	0	3
t|	3	1	0	7	0	0	0	0	11	0	0

EM algorithm : expectation step

Given our current estimate of the matrix , determine the most likely position of the motif

probability of motif

$$p_\theta(k) = f_{C,1} \cdot f_{G,2} \cdot f_{G,3} \cdot f_{G,4} \cdot f_{A,5} \cdot f_{T,6} \cdot f_{T,7}$$

$f_{A,5}$ = frequency of base A at position 5 of the PFM θ

k

A C G G A T C G G A T T C G G G A T T C C C G T A G G C A G C T T A G

probability of background

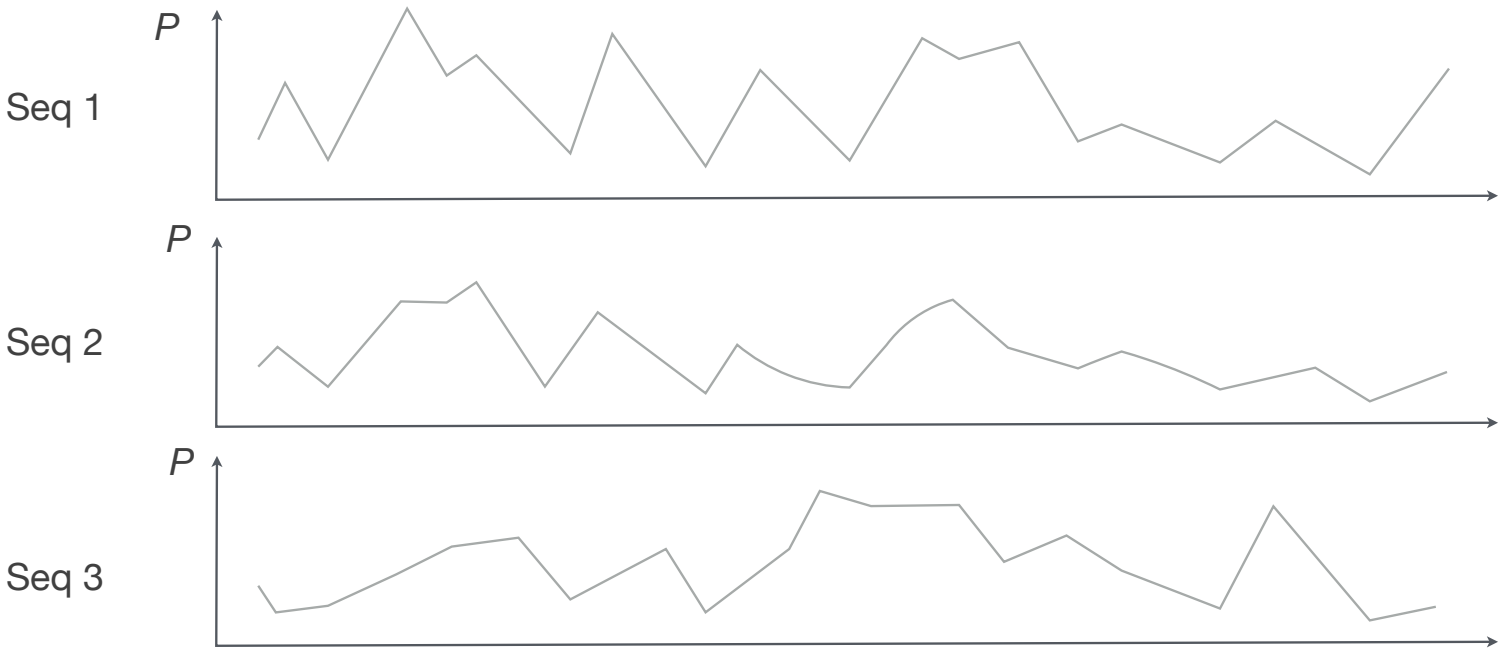
$$p_b(k) = p_A^6 \cdot p_C^7 \cdot p_G^9 \cdot p_T^6$$

Likelihood of X given θ given $Z = k$:

$$P(X|Z_k) \cdot P(Z_k|\theta) = p_b(k) p_\theta(k)$$

$$P(X|\theta) = \sum_k P(X|Z_k) \cdot P(Z_k|\theta)$$

EM algorithm : expectation step

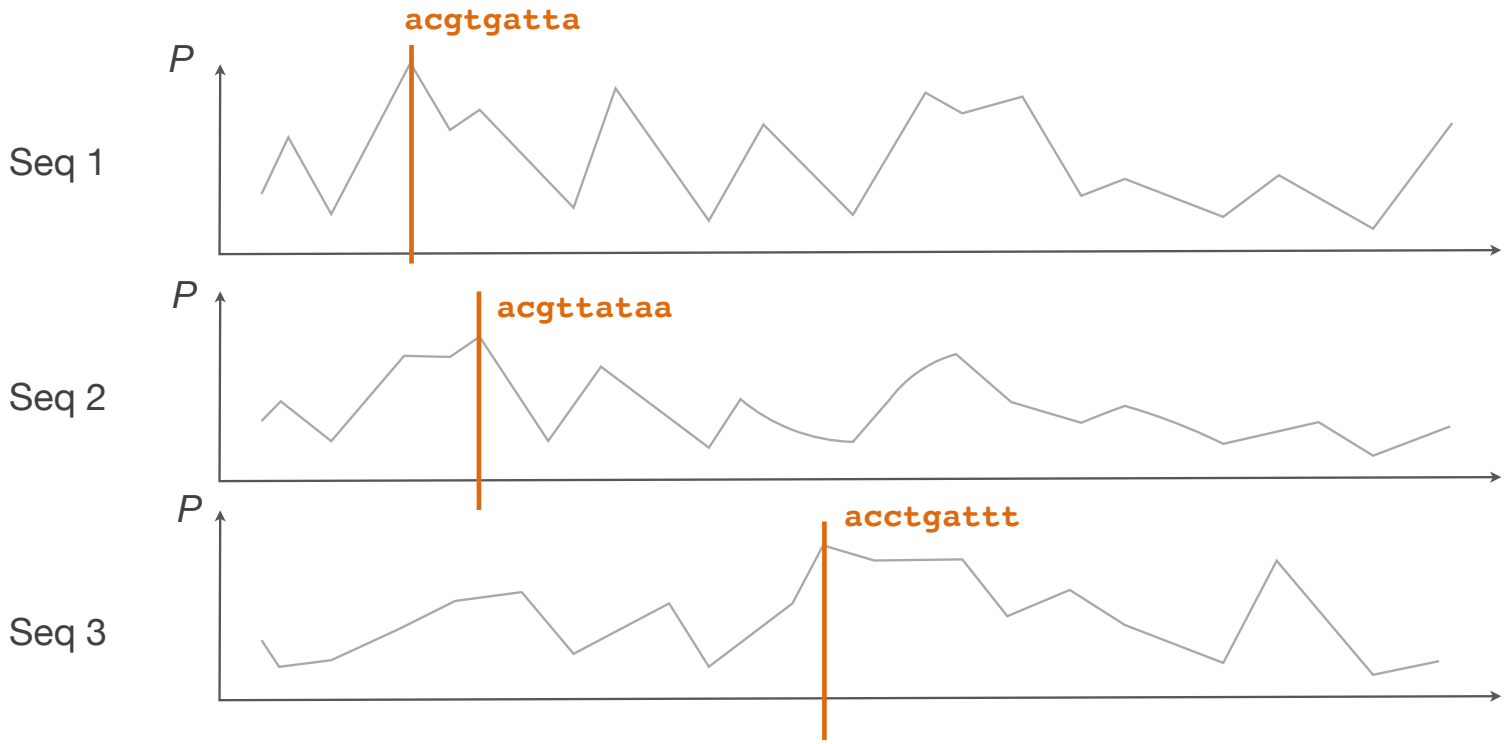


plots of $P = P(X|Z_k)P(Z_k|\theta)$

→ probability that the subsequences at position k matches the matrix θ

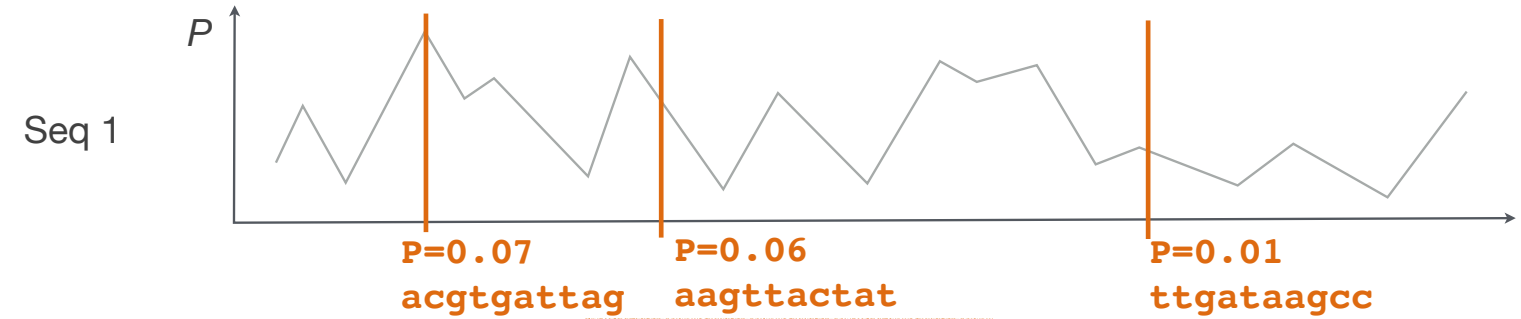
EM algorithm : maximization step

First possibility : add these new site to the current matrix to improve it



EM algorithm : maximization step

Second possibility : update matrix according to the profiles




```
add
0.07 to A at position 1
0.07 to C at position 2
0.07 to G at position 3
......
```

```
add
0.06 to A at position 1
0.06 to A at position 2
0.06 to G at position 3
......
```

```
add
0.01 to T at position 1
0.01 to T at position 2
0.01 to G at position 3
......
```

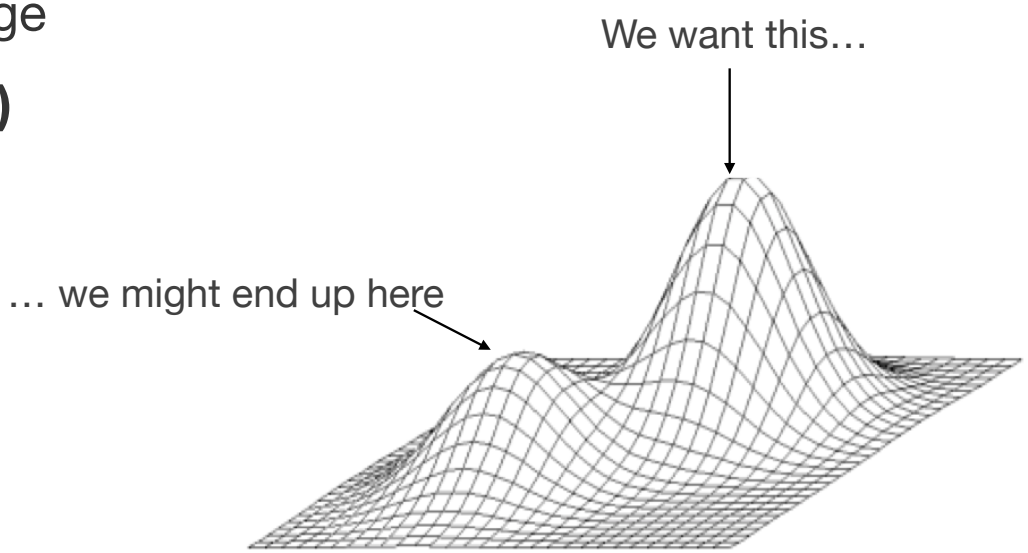
- same at all positions
- same for all sequences
- renormalize matrix columns

PO	A	C	G	T
1	0.701	0.021	0.243	0.034
2	0.812	0.021	0.132	0.034
3	0.812	0.021	0.132	0.034
4	0.368	0.021	0.021	0.590
5	0.034	0.021	0.910	0.034
6	0.923	0.021	0.021	0.034
7	0.034	0.132	0.799	0.034
8	0.034	0.021	0.021	0.923
9	0.034	0.910	0.021	0.034
10	0.923	0.021	0.021	0.034

New estimation of θ : iterate again the expectation step

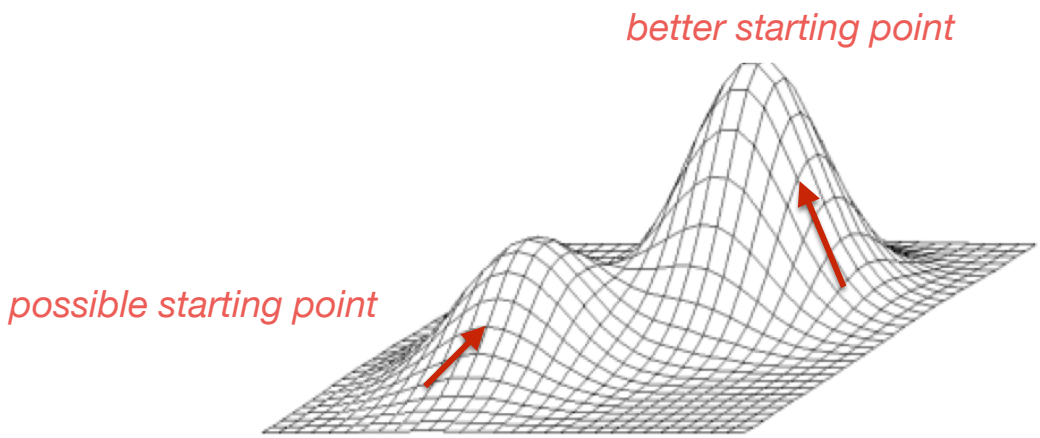
Expectation-Maximization algorithm (EM)

- initiate a new E-step using the updated matrix
- stop iterations when
  - max. number of iterations ; or
  - parameters do not change
- **(local) maximum of P(X|θ)**



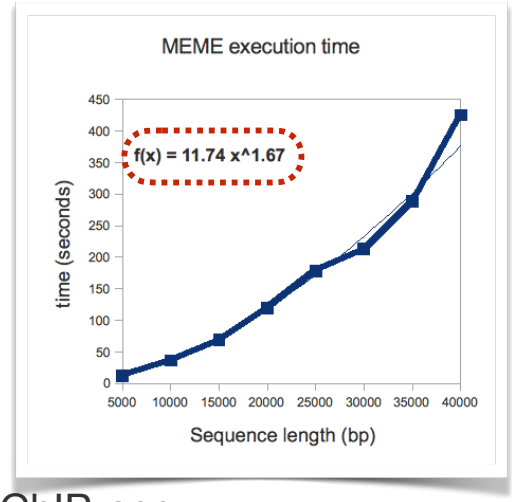
Expectation-Maximization algorithm (EM)

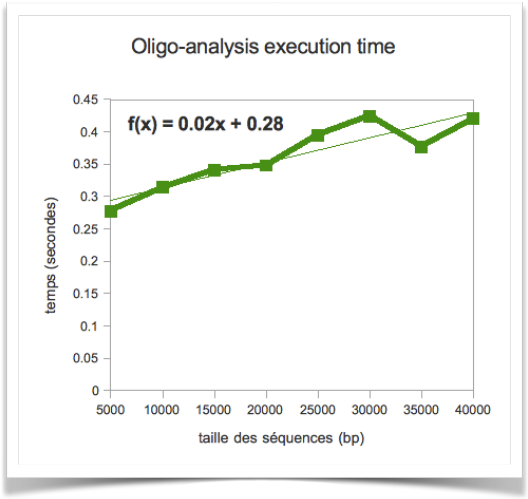
- starting point of the algorithm will influence the type of maxima
- Solution : test all subsequences of length W for the matrix initialization
- selects the set of initial sequences that improve Pr(X|θ) most after one iteration



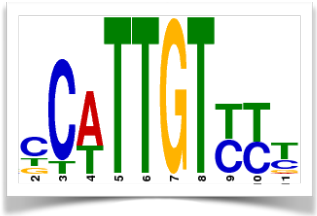
Expectation maximization

- Algorithmic complexity: EM is quadratic 😞 Word counting linear 😊

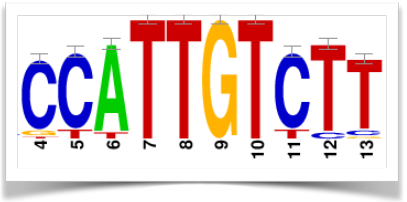




- Example: SOX6 ChIP-seq



MEME (expectation-max.)

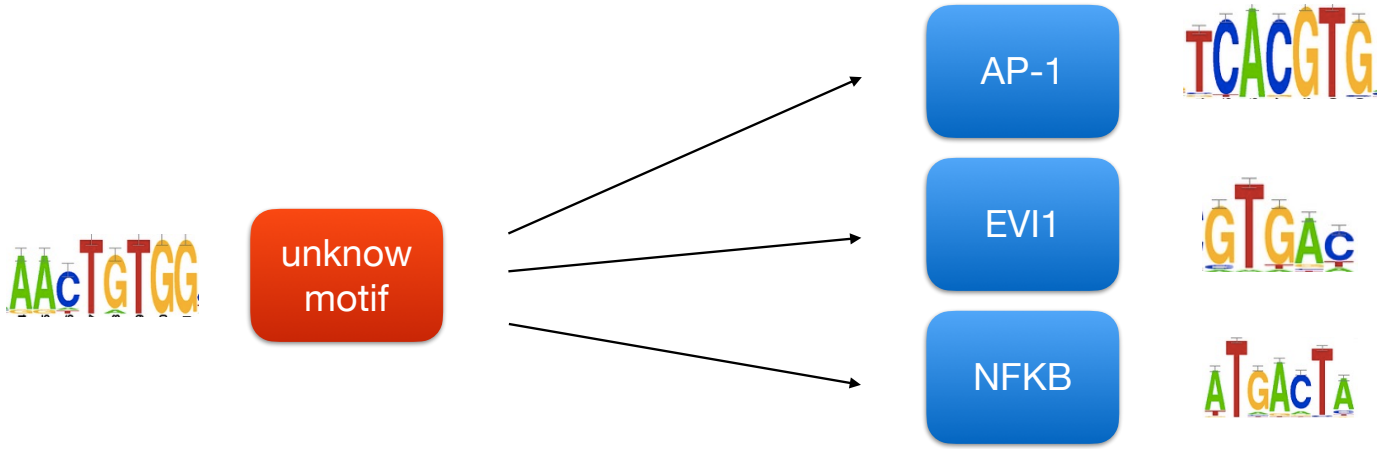


oligo-analysis (word counting)

Identifying novel motifs

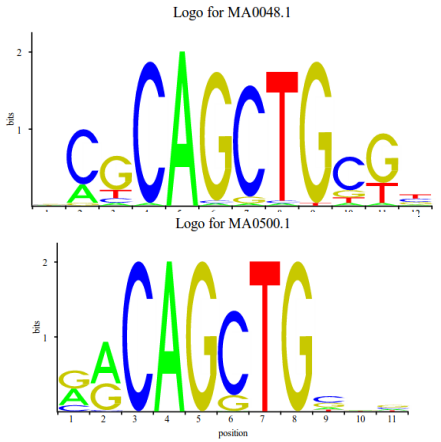
- Using motif discovery we obtain a significant motif
  → ***to which TF could it correspond ?***

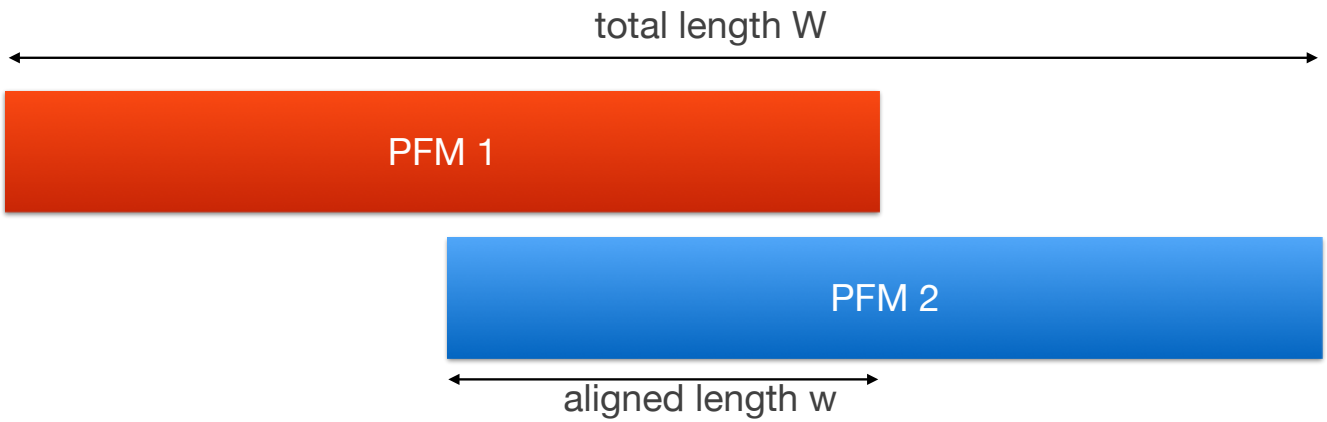
- we need a measure to compare motifs through their
  - position frequency matrix
  - logo



Comparing binding profiles

- matrices are compared by considering all possible alignments of one other (« local alignment »)
- a mesure of similarity is computed for each alignment
- the configuration with the highest similarity is retained

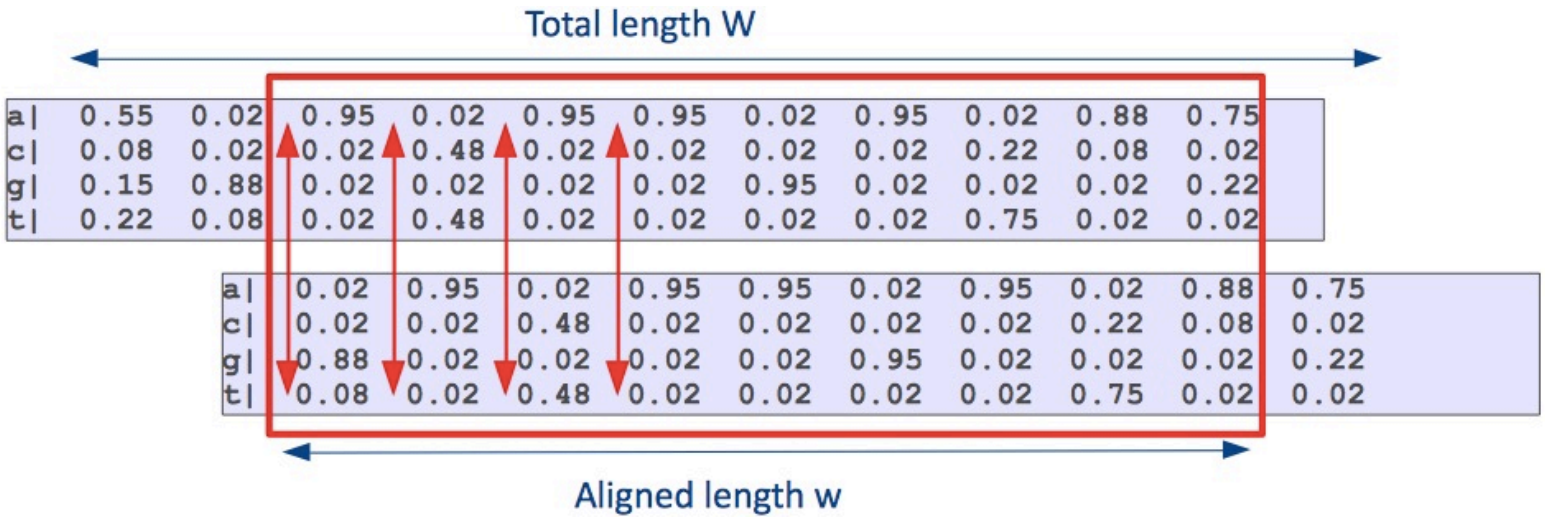




Comparing binding profiles

- Correlation/covariance coefficient
- Correlation is normalized to the length of the alignment to avoid perfect correlation over a single column.

$$Ncor = \frac{w}{W} \times cor$$

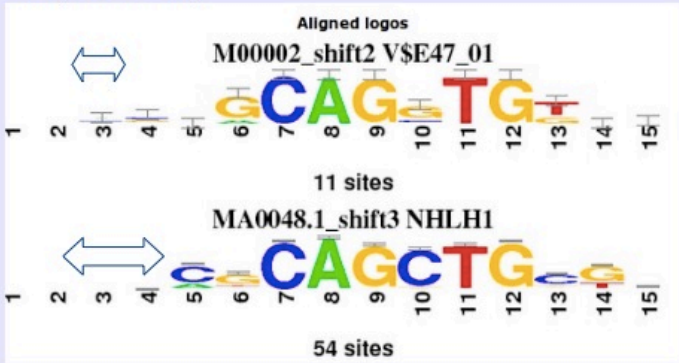


Comparing binding profiles

Other measure : sum of squared distances

$$SSD = \sum_{i=1}^{w} \sum_{j=1}^{4} (f_{i,j} - g_{i,j})^2$$

```
; M00002 (V$E47_01); m=0 (reference); ncol1=15; shift=2; ncol=17;
; Alignment reference
a | 0 0 4 2 3 2 0 11 0 1 0 0 0 1 1 1 1
c | 0 0 4 5 2 0 11 0 0 2 0 0 0 4 6 4 4
g | 0 0 3 4 4 9 0 0 11 8 0 11 4 3 2 4 2
t | 0 0 0 0 2 0 0 0 0 0 11 0 7 3 2 2 3
//
; M00002 versus MA0048.1 (NHLH1); m=1/4; ncol2=12
; w=12; offset=1; strand=D; shift=3; score=     3.6;
; SSD=2.305; SW=19.695; NSW=0.895;
a | 0 0 0 13 13 3 1 54 1 1 1 0 3 2 5 0 0
c | 0 0 0 13 39 5 53 0 1 50 1 0 37 0 17 0 0
g | 0 0 0 17 2 37 0 0 52 3 0 53 8 37 12 0 0
t | 0 0 0 11 0 9 0 0 0 0 52 1 6 15 20 0 0
```



Summary

- **Motif discovery**: looking for shared or over-represented common motif in a set of sequences (co-expressed promoters, ChIP-seq sequences,...)
- unlike **Pattern matching**: know motif → look for possible occurences
- Different **statistical strategies** to find this signal (i.e. motif)
  - **word-counting**: find over-represented words / assemble them → motif
  - **expectation maximization**: find the motif(s) that maximizes the likelihood of the set of sequences given this motif
- Motif discovery yields an **unknown motif**: need to identify it comparing with known motifs (e.g. JASPAR database)