The Analysis of Regulatory Sequences

Pattern comparison

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

Patterns representing regulatory signals (in DNA)

- String based description
  - DNA alphabet
  - IUPAC alphabet
  - Regular expressions
- Matrix based description
  - PSSM: Position-Specific Scoring Matrices

Gal4p binding sites (source: SCPD)

```
>YBR018C      AGCGCTCGGACAACTGTTGACC
>YBR018C      ATACTTCGGAGCACTGTTGAGCG
>YBR019C      ......CGGCGGCTTCTAATCCG
>YBR019C      .....TCGGAGGGCTGTCGCCCG
>YBR019C      ......CGGAGGAGAGTCTTCCG
>YBR019C      ATTGTTCGGAGCAGTGCGGCGCG
>YBR020W      ...CGCGCCGCACTGCTCCGAACAAT
>YBR020W      ......CGGAAGACTCTCCTCCG
>YBR020W      ......CGGGCGACAGCCCTCCGA
>YBR020W      ......CGGATTAGAAGCCGCCG
>YLR081W      ...TATCGGGGCGGATCACTCCGAAC
>YLR081W      ...CACCGGCGGTCTTTCGTCCGTGC
>YML051W      ......CGGCGCACTCTCGCCCG
>YOR120W      .....TCGGGGCAGACTATTCCGG
Consensus     ......CGGnnnnnnnnnnnCCG
```

Gal4p matrix (source: SCPD)

Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	0	0	0	4	1	1	7	0	5	1	0	2	0	2	0	0	0
T	0	0	0	0	1	1	1	0	5	2	7	2	1	6	0	0	0
G	0	10	9	4	5	3	2	3	0	3	1	1	4	1	1	0	10
C	10	0	1	2	3	5	0	7	0	4	2	5	5	1	9	10	0

Weights, information and consensus

```
; convert-matrix  -v 1 -i GAL4_matrix_SCPD.tab -format tab -return counts,weights,parameters,information -decimals 2
;
; Matrix type: counts
; Pos 1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17
; -----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
A     0     0     0     4     1     1     7     0     5     1     0     2     0     2     0     0     0
T     0     0     0     0     1     1     1     0     5     2     7     2     1     6     0     0     0
G     0    10     9     4     5     3     2     3     0     3     1     1     4     1     1     0    10
C    10     0     1     2     3     5     0     7     0     4     2     5     5     1     9    10     0
;
; Matrix type: weights
; Pos 1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17
; -----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
A -2.40 -2.40 -2.40  0.44 -0.79 -0.79  0.97 -2.40  0.65 -0.79 -2.40 -0.20 -2.40 -0.20 -2.40 -2.40 -2.40
T -2.40 -2.40 -2.40 -2.40 -0.79 -0.79 -0.79 -2.40  0.65 -0.20  0.97 -0.20 -0.79  0.82 -2.40 -2.40 -2.40
G -2.40  1.32  1.21  0.44  0.65  0.17 -0.20  0.17 -2.40  0.17 -0.79 -0.79  0.44 -0.79 -0.79 -2.40  1.32
C  1.32 -2.40 -0.79 -0.20  0.17  0.65 -2.40  0.97 -2.40  0.44 -0.20  0.65  0.65 -0.79  1.21  1.32 -2.40
;
; Matrix type: information
; Pos 1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17
; -----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
A -0.05 -0.05 -0.05  0.17 -0.09 -0.09  0.64 -0.05  0.31 -0.09 -0.05 -0.04 -0.05 -0.04 -0.05 -0.05 -0.05
T -0.05 -0.05 -0.05 -0.05 -0.09 -0.09 -0.09 -0.05  0.31 -0.04  0.64 -0.04 -0.09  0.47 -0.05 -0.05 -0.05
G -0.05  1.23  1.02  0.17  0.31  0.05 -0.04  0.05 -0.05  0.05 -0.09 -0.09  0.17 -0.09 -0.09 -0.05  1.23
C  1.23 -0.05 -0.09 -0.04  0.05  0.31 -0.05  0.64 -0.05  0.17 -0.04  0.31  0.31 -0.09  1.02  1.23 -0.05
;
; Alphabet  A T G C
;           A           0.25
;           C           0.25
;           G           0.25
;           T           0.25
; pseudo                          1
; total.information               9.28907
; information.per.column          0.546416
; consensus.strict                CGGggcactctcctCCG
; consensus.IUPAC                 CGGrssaswstcstCCG
; consensus.regexp                CGG[ag][cg][cg]a[cg][at][cg]tc[cg]tCCG
```

Sources of regulatory patterns

- Patterns annotated in transcription factor databases (e.g. TRANSFAC, YSCPD, RegulonDB)
  - Collections of experimentally proven cis-acting elements (sites) for a given transcription factor. The site is described by its sequence + position relative to the cis-regulated gene.
  - Matrices (PSSM) built from these collections.
  - Consensus sequences (IUPAC) built from these collections
- Patterns discovered by the analysis of non-coding regulatory regions
  - Clusters of genes in a single organism
    - over-represented motifs in promoters of of co-expressed genes.
  - Clusters of genes in a single organism, genes selected by orthology
    - Regulon in reference organism -> orthologs in query organism -> motifs in promoters
  - Phylogenetic footprinting: single gene, multiple organisms
    - Set of orthologous genes -> promoters -> over-represented motifs

Applications of pattern comparisons

- Interpretation of discovered patterns (e.g.from microarray clusters)
  - Compare discovered patterns with annotated cis-acting elements in order to predict potential trans-acting factors.
- Compare patterns discovered in different data sets (e.g. co-expressed clusters)
- Compare patterns discovered in different organisms
  - Apply pattern discovery in orthologs of regulons for a reference organism.
- Phylogenetic footprinting
  - Discover patterns in upstream sequences of sets of orthologous genes.
  - Then compare patterns found for different genes in order to build putative sets of co-expressed genes.

Issues for pattern comparisons

- Types of comparisons
  - String-based versus string-based
  - Matrix-based versus matrix-based
  - Comparison between string-based and matrix-based patterns
- Scoring the matching
  - Boolean matching (TRUE or FALSE)
  - Count of matching residues
  - P-value to estimate the significance of the matching
  - Information content

String-string comparions

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

Matching between residues

- The simplest way to compare two patterns is to count the number of matches between residues.

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Pattern 1	G	C	A	C	G	T	G	G	G
Pattern 2	T	C	A	C	G	T	G	A	
Match	0	1	1	1	1	1	1	0	0
Matches	6	/	9						

Matching between IUPAC residues

- The matching table can be extended to include the IUPAC code for ambiguous nucleotides

		A	C	G	T
A	A	1	0	0	0
C	C	0	1	0	0
G	G	0	0	1	0
T	T	0	0	0	1
A or G	R	1	0	1	0
C or T	Y	0	1	0	1
A or T	W	1	0	0	1
C or G	S	0	1	1	0
A or C	M	1	1	0	0
G or T	K	0	0	1	1
A, C or T	H	1	1	0	1
C, G or T	B	0	1	1	1
A, C or G	V	1	1	1	0
A, G or T	D	1	0	1	1
A, C, G or T	N	1	1	1	1

A	A	**A**denine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
R	A or G	pu**R**ine
Y	C or T	p**Y**rimidine
W	A or T	**W**eak hydrogen bonding
S	G or C	**S**trong hydrogen bonding
M	A or C	a**M**ino group at common position
K	G or T	**K**eto group at common position
H	A, C or T	not **G**
B	G, C or T	not **A**
V	G, A, C	not **T**
D	G, A or T	not **C**
N	G, A, C or T	a**N**y

Pattern 1	**G**	**C**	**A**	**C**	**G**	**T**	**G**	**C**	**G**
Pattern 2	**S**	**C**	**A**	**C**	**G**	**T**	**K**	**K**	**K**
Match	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**
Matches	**8**	**/**	**9**						

Matching between IUPAC residues

- We can also compare two patterns containing ambiguous nucleotides

		A	C	G	T	R	Y	W	S	M	K	H	B	V	D	N
A	A	1	0	0	0	1	0	1	0	1	0	1	0	1	1	1
C	C	0	1	0	0	0	1	0	1	1	0	1	1	1	0	1
G	G	0	0	1	0	1	0	0	1	0	1	0	1	1	1	1
T	T	0	0	0	1	0	1	1	0	0	1	1	1	0	1	1
A or G	R	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1
C or T	Y	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1
A or T	W	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1
C or G	S	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1
A or C	M	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1
G or T	K	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1
A, C or T	H	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
C, G or T	B	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A, C or G	V	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
A, G or T	D	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
A, C, G or T	N	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

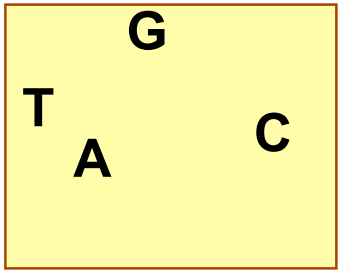
Pattern 1	**T**	**C**	**A**	**C**	**G**	**T**	**A**	**V**	**M**	**W**
Pattern 2	**S**	**C**	**A**	**C**	**G**	**T**	**B**	**B**	**B**	**N**
Match	**0**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**
Matches	**7**	**/**	**10**							

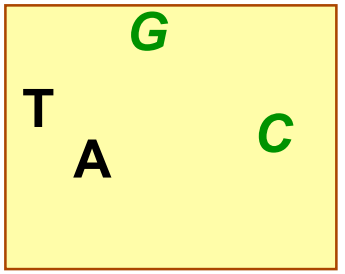
Uninformative and poorly informative matches

- When patterns contain IUPAC degenerated letters, the simple count of matches provides a poor estimation of the similarity between two patterns.
- For example
  - **TCACGTACA**
    NCANNTANN
    - 9 matching letters, but only 4 of them are informative
  - **TCACGTACA**
    TCACGTAGC
    - 7 matching letters, which are all informative
  - **TCACGTAVM**
    TCACGTAGC
    - 9 matching letters
    - M could accept a match with 2 letters (A or C)
    - V could accept a match with 3 letters (A, C or G)
    - The 7 first ones are thus more informative than the last two last ones.

Probability of a match between two IUPAC letters

- How can we estimate the statistical significance of a match between two IUPAC letters
  - S = C or G
  - K = G or T
- We can model this situation as an urn containing labelled (green) and non-labelled (black) balls.
- In total, the urn contains 4 balls (the nucleotide alphabet).
- We label in green the nucleotides matched by the first IUPAC letter (C and G).
- The choice of the second letter amounts to select 2 letters in this urn. What is the probability to obtain at least one green ball in the selection ?



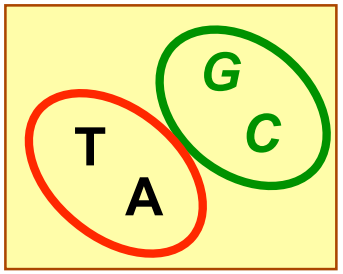


The hypergeometric distribution

$$P(X = x) = \frac{C_m^x C_n^{k-x}}{C_{m+n}^k}$$

$$Pval = \sum_{i=1}^{\min(m,k)} P(X = i) = \sum_{i=1}^{\min(m,k)} \frac{C_m^i C_n^{k-i}}{C_{m+n}^k}$$

- The hypergeometric distribution represents the probability to observe x successes in a sampling without replacement
  - *m* number of possible successes (labelled balls in the urn)
  - *n* number of possible failures (non-labelled balls in the urn)
  - *k* sample size
  - *x* number of successes (labelled balls) in the sample
- We want to calculate the P-value, i.e. the probability to have at least one common letter.

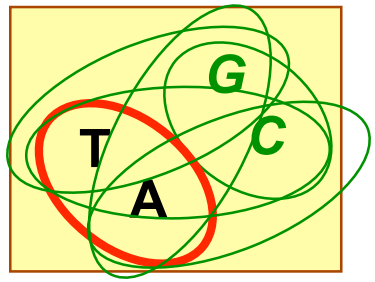


The hypergeometric distribution

$$P(X = x) = \frac{C_2^x C_2^{2-x}}{C_4^2}$$

$$Pval = \sum_{i=1}^{2} P(X = i) = \sum_{i=1}^{2} \frac{C_2^i C_2^{2-i}}{C_4^2}$$

- In our case
  - m = 2 nucleotides matched by the first IUPAC letter (S)
  - n = 2 nucleotides not matched by the first IUPAC letter.
  - k = 2 nucleotides matched by the second IUPAC letter



- *Pval*=0.8333333
- This is quite intuitive
  - There are 6 possible ways to select 2 letters among 4.
  - Among these 6 possibilities, there is only 1 way to have not a single match with {C, G} : for this, you need to select {A, T}.
  - The probability to hav e at least one match is thus Pval=5/6=0.833333

From P-value to significance

$$P(X = x) = \frac{C_m^x C_n^{k-x}}{C_{m+n}^k}$$

$$Pval = \sum_{i=1}^{\min(m,k)} P(X = i) = \sum_{i=1}^{\min(m,k)} \frac{C_m^i C_n^{k-i}}{C_{m+n}^k}$$

$$sig = -\log_4(Pval)$$

- We can now define the significance as the negative logarithm of the P-value.
- It is convenient to use a logarithm in base 4, since this is the alphabet size.
- The significance represents the number of matching letters, with fractional values for degenerate matches.
- Examples
  - Most significant match (1 letter against 1 letter)
    - A against A — Pval=0.25 — sig=1
  - Non-degenerate against degenerate
    - S (G, C) against G — Pval=0.5 — sig=0.5
    - B (C, G,T) against G — Pval=0.75 — sig=0.21
    - N (A, C, G, T) against A — Pval=1 — sig=0 — trivial
  - Degenerate against degenerate
    - S (G, C) against K (G, T) — Pval=0.83 — sig=0.13
    - H (A, C, T) against S (G, C) — Pval=1 — sig=0 — unavoidable

P-value and significance for sequences

- We can now extend the two concepts to estimate the P-value and significance of matches between two strings.
  - The probability of several matches between letters is the product of probabilities of the pairs of aligned letters.
  - Consequently, the significance of the match between strings is the sum of significances of the pairs of aligned letters.
- Examples
  - **TCACGTACA**
    NCANNTANN
    - 9 matches, **4** significant Pval=0 39: sig=**4**
  - **TCACGTACA**
    TCACGTAGC
    - 7 matching letters Pval=6.1e-5; sig=7
  - **TCACGTAVM**
    TCACGTAGC
    - 9 matching letters
      - 7 first letters Pval=0.257=6.10e-5; sig=7
      - V against G Pval=0.75; sig=0.21
      - M against C Pval=0.5; sig=0.5
    - Whole alignment: Pval=2.3e-5; sig=7.71

Match table

- We can calculate the number of matches between each annotated binding site (rows) and each discovered pattern (column), and represent it in a table.

Annotated sites \ discovered patterns	ccctcc	cctgga	ctcccc	agggca	cagaca	ctcctc	ggagga	cctccc	agggag	gaggga	cacaga	Perfect matches	At most one substitution
CCCCCACTGAACCCTTGACCCCTGCCC	5	5	5	5	3	4	4	5	4	4	5	**0**	**6**
aggGTTCACcgaaaGTTCACtcgca	4	3	4	4	4	3	4	4	4	5	5	**0**	**2**
GGGTCAggaggAGGTGA	5	4	5	4	4	**6**	**6**	5	4	4	3	**2**	**5**
GTCCCCGCCTC	5	4	5	4	3	5	4	4	4	4	3	**0**	**3**
GGGTTTGACCTTTCTCTCCGGGTAAAGGTGAAGG	5	5	5	5	4	4	4	4	4	5	4	**0**	**5**
gtagggtgtAGGGAGattGGTTCAatgtccaat	5	4	5	4	4	4	4	5	**6**	5	4	**1**	**5**
TGCCCTtccttatggGGTTCA	5	3	4	**6**	3	4	4	5	5	4	4	**1**	**4**
gatccaactgaGGGTCAgTGACCCaagtga	5	4	4	4	4	4	4	4	4	5	5	**0**	**3**
GGGCTCCGGTGAGTCAGGGCGCGTTATGCA	4	4	5	5	4	4	4	4	5	4	4	**0**	**3**
tccactgTGCCCGaggcTGTCCTggaggta	5	**6**	4	5	4	4	5	5	4	5	5	**1**	**7**
Perfect matches	**0**	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**5**	
At most one substitution	**8**	**3**	**6**	**5**	**0**	**2**	**2**	**5**	**3**	**5**	**4**		**43**

Multiple string comparisons

- String-based pattern discovery programs typically return several string-based patterns.
- Transcription factor databases hopefully contain several binding sites for each transcription factor.
- Each discovered pattern can be compared to each annotated site in a contingency table

; sequence	aaacgt	aacgtg	acgtgc	acgtgg	cacgtg	cccacg	cgcacg	ctgcac	tgccaa
TAAATTAGCACGTTTTCGCATAGA	4	4	4	4	5	4	5	4	3
TGGCACTCACACGTGGGACTAGCA	5	5	5	6	6	5	5	4	4
TCGGGCCACGTGCAGCGAT	4	5	6	5	6	5	4	4	4
ATATTAAGCGTGCGGGTAA	5	5	5	4	4	3	4	3	3
TTATGGCACGTGCGAATAA	4	5	6	5	6	4	5	4	5
TTACGCACGTTGGTGCTG	4	4	4	5	5	5	6	4	3
TTTCCAGCACGTGGGGCGG	4	5	5	6	6	4	5	5	4
TAGTTCCACGTGGACGTG	4	5	5	6	6	5	4	4	4
aaaagtgtCACGTGataaaaat	5	5	5	5	6	4	4	3	4
TTAAAAACGTGCGTATTA	6	6	6	5	5	3	4	3	4
GCGTTCACACGTGGGTTTA	5	5	5	6	6	5	5	4	3
GCGTTCACACGTGGGTTTA	5	5	5	6	6	5	5	4	3
AATGCAGCACGTGGGAGAC	4	5	5	6	6	4	5	5	3
GCGCCCGCACGTGCTCTTT	4	5	6	5	6	5	6	5	3
TTTTGCTCACGTGACCGAC	4	5	5	5	6	5	5	4	4
ATGTACGCACGTGGGCGAA	4	5	5	6	6	5	6	4	4
CTTTTCCCACGTGCTCCGC	4	5	6	5	6	6	5	4	4
CTGCAGCCACGTGCCTAGA	4	5	6	5	6	5	4	5	5
AAATTACCACGTTTTCGCA	4	4	4	4	5	5	4	3	4
AAATTACCACGTTTTCGCA	4	4	4	4	5	5	4	3	4
TCATCCCCACGTTGTGCCA	4	4	4	5	5	6	5	4	5
ACACACACACGTTAAGAGA	5	4	4	4	5	5	5	3	3

Multiple string comparisons - matches

- String-based pattern discovery programs typically return several string-based patterns.
- Transcription factor databases hopefully contain several binding sites for each transcription factor.
- Each discovered pattern can be compared to each annotated site in a contingency table
- Example: Gal4p annotated sites versus discovered dyads.

; sequence	cggn{11}ccg	cggn{12}cga	cggn{10}tcc	ccgn{12}ccg	ccgn{1}gcg
AGCGCTCGGACAACTGTTGACC	12	13	11	12	1
ATACTTCGGAGCACTGTTGAGCG	11	12	10	12	1
CGGCGGCTTCTAATCCG	17	17	16	15	4
TCGGAGGGCTGTCGCCCG	14	14	13	17	4
CGGAGGAGAGTCTTCCG	17	17	16	15	4
ATTGTTCGGAGCAGTGCGGCGCG	11	12	10	14	1
TATCGGGGCGGATCACTCCGAAC	12	13	11	13	3
CACCGGCGGTCTTTCGTCCGTGC	13	13	13	13	3
CGGCGCACTCTCGCCCG	17	17	15	15	5
TCGGGGCAGACTATTCCGG	13	14	13	17	4

Multiple string comparisons - significance

- Since the dyads contain trivial matches (N), the significance is more indicative than the number of matching letters.

; sequence	cggn{11}ccg	cggn{12}cga	cggn{10}tcc	ccgn{12}ccg	ccgn{1}gcg
AGCGCTCGGACAACTGTTGACC	1	1	1	0	0
ATACTTCGGAGCACTGTTGAGCG	0	0	0	0	0
CGGCGGCTTCTAATCCG	6	5	6	3	3
TCGGAGGGCTGTCGCCCG	3	2	3	5	3
CGGAGGAGAGTCTTCCG	6	5	6	3	3
ATTGTTCGGAGCAGTGCGGCGCG	0	0	0	2	0
TATCGGGGCGGATCACTCCGAAC	1	1	1	1	2
CACCGGCGGTCTTTCGTCCGTGC	2	1	3	1	2
CGGCGCACTCTCGCCCG	6	5	5	3	4
TCGGGGCAGACTATTCCGG	2	2	3	5	3

Multiple string comparisons

- A threshold can be applied to display significant matches only

; sequence	aaacgt	aacgtg	acgtgc	acgtgg	cacgtg	cccacg	cgcacg	ctgcac	tgccaa
TAAATTAGCACGTTTTCGCATAGA					5		5		
TGGCACTCACACGTGGGACTAGCA	5	5	5	6	6	5	5		
TCGGGCCACGTGCAGCGAT		5	6	5	6	5			
ATATTAAGCGTGCGGGTAA	5	5	5						
TTATGGCACGTGCGAATAA		5	6	5	6		5		5
TTACGCACGTTGGTGCTG				5	5	5	6		
TTTCCAGCACGTGGGGCGG		5	5	6	6		5	5	
TAGTTCCACGTGGACGTG		5	5	6	6	5			
aaaagtgtCACGTGataaaaat	5	5	5	5	6				
TTAAAAACGTGCGTATTA	6	6	6	5	5				
GCGTTCACACGTGGGTTTA	5	5	5	6	6	5	5		
GCGTTCACACGTGGGTTTA	5	5	5	6	6	5	5		
AATGCAGCACGTGGGAGAC		5	5	6	6		5	5	
GCGCCCGCACGTGCTCTTT		5	6	5	6	5	6	5	
TTTTGCTCACGTGACCGAC		5	5	5	6	5	5		
ATGTACGCACGTGGGCGAA		5	5	6	6	5	6		
CTTTTCCCACGTGCTCCGC		5	6	5	6	6	5		
CTGCAGCCACGTGCCTAGA		5	6	5	6	5		5	5
AAATTACCACGTTTTCGCA					5	5			
AAATTACCACGTTTTCGCA					5	5			
TCATCCCCACGTTGTGCCA				5	5	6	5		5
ACACACACACGTTAAGAGA	5				5	5	5		

Multiple string comparisons

- A threshold can be applied to display significant matches only

; sequence	aaacgt	aacgtg	acgtgc	acgtgg	cacgtg	cccacg	cgcacg	ctgcac	tgccaa
TAAATTAGCACGTTTTCGCATAGA									
TGGCACTCACACGTGGGACTAGCA				6	6				
TCGGGCCACGTGCAGCGAT			6		6				
ATATTAAGCGTGCGGGTAA									
TTATGGCACGTGCGAATAA			6		6				
TTACGCACGTTGGTGCTG							6		
TTTCCAGCACGTGGGGCGG				6	6				
TAGTTCCACGTGGACGTG				6	6				
aaaagtgtCACGTGataaaaat					6				
TTAAAAACGTGCGTATTA	6	6	6						
GCGTTCACACGTGGGTTTA				6	6				
GCGTTCACACGTGGGTTTA				6	6				
AATGCAGCACGTGGGAGAC				6	6				
GCGCCCGCACGTGCTCTTT			6		6		6		
TTTTGCTCACGTGACCGAC					6				
ATGTACGCACGTGGGCGAA				6	6		6		
CTTTTCCCACGTGCTCCGC			6		6	6			
CTGCAGCCACGTGCCTAGA			6		6				
AAATTACCACGTTTTCGCA									
AAATTACCACGTTTTCGCA									
TCATCCCCACGTTGTGCCA						6			
ACACACACACGTTAAGAGA									

Multiple string comparisons

- We can define comparison a pattern coverage for each pattern/site comparison
  - PPV(pattern, site) = sig(pattern,site) / max(sig|pattern)
  - In the case below, max(sig|pattern) = 6 (non-degenerated hexamers)

; sequence	aaacgt	aacgtg	acgtgc	acgtgg	cacgtg	cccacg	cgcacg	ctgcac	tgccaa
TAAATTAGCACGTTTTCGCATAGA									
TGGCACTCACACGTGGGACTAGCA				1 0	1 0				
TCGGGCCACGTGCAGCGAT			1 0		1 0				
ATATTAAGCGTGCGGGTAA									
TTATGGCACGTGCGAATAA			1 0		1 0				
TTACGCACGTTGGTGCTG							1 0		
TTTCCAGCACGTGGGGCGG				1 0	1 0				
TAGTTCCACGTGGACGTG				1 0	1 0				
aaaagtgtCACGTGataaaaat					1 0				
TTAAAAACGTGCGTATTA	1 0	1 0	1 0						
GCGTTCACACGTGGGTTTA				1 0	1 0				
GCGTTCACACGTGGGTTTA				1 0	1 0				
AATGCAGCACGTGGGAGAC				1 0	1 0				
GCGCCCGCACGTGCTCTTT			1 0		1 0		1 0		
TTTTGCTCACGTGACCGAC					1 0				
ATGTACGCACGTGGGCGAA				1 0	1 0		1 0		
CTTTTCCCACGTGCTCCGC			1 0		1 0	1 0			
CTGCAGCCACGTGCCTAGA			1 0		1 0				
AAATTACCACGTTTTCGCA									
AAATTACCACGTTTTCGCA									
TCATCCCCACGTTGTGCCA						1 0			
ACACACACACGTTAAGAGA									

Multiple string comparisons

- We can define a “site coverage” for each pattern/site pair
  - Cov(pattern,site) = sig(pattern, site) / max(sig|site)
  - In the case below, sites can have different lengths; their coverage may thus differ, even for a perfectly matching hexamer.
- Note that sites are larger than hexamers, but some sites are covered by multiple patterns.
- A total coverage could be calculated (site positions covered by at least one pattern)

; sequence	aaacgt	aacgtg	acgtgc	acgtgg	cacgtg	cccacg	cgcacg	ctgcac	tgccaa
TAAATTAGCACGTTTTCGCATAGA									
TGGCACTCACACGTGGGACTAGCA				0.250	0.250				
TCGGGCCACGTGCAGCGAT			0.316		0.316				
ATATTAAGCGTGCGGGTAA									
TTATGGCACGTGCGAATAA			0.316		0.316				
TTACGCACGTTGGTGCTG							0.333		
TTTCCAGCACGTGGGGCGG				0.316	0.316				
TAGTTCCACGTGGACGTG				0.333	0.333				
aaaagtgtCACGTGataaaaat					0.273				
TTAAAAACGTGCGTATTA	0.333	0.333	0.333						
GCGTTCACACGTGGGTTTA				0.316	0.316				
GCGTTCACACGTGGGTTTA				0.316	0.316				
AATGCAGCACGTGGGAGAC				0.316	0.316				
GCGCCCGCACGTGCTCTTT			0.316		0.316		0.316		
TTTTGCTCACGTGACCGAC					0.316				
ATGTACGCACGTGGGCGAA				0.316	0.316		0.316		
CTTTTCCCACGTGCTCCGC			0.316		0.316	0.316			
CTGCAGCCACGTGCCTAGA			0.316		0.316				
AAATTACCACGTTTTCGCA									
AAATTACCACGTTTTCGCA									
TCATCCCCACGTTGTGCCA						0.316			
ACACACACACGTTAAGAGA									

Summary: significance

- The P-value and significance defined above allow to compare degenerated patterns, or patterns containing fixed width spacers (e.g. dyads).
- The sig is intuitive: it counts the number of informative matching letters.
  - The unit corresponds to a perfect match between 2 letters.
  - Less informative matches (degenerated) have values between 0 and 1.
  - 0 means either no match, or an uninformative match.
- The concept can easily be extended to peptides, one should then use the logarithm in base 20.

Sequences with uneven residue probabilities

- The hypergeometric model assumes that each “ball” has the same probability to be selected.
- In other terms, until now we assumed that nucleotides are equiprobable.
- This is usually not the case for biological sequences.
- How should we treat sequences with uneven residue probabilities ?

Saccharomyces cerevisiae genome

seq	freq	occ
a	0.310	3766125
c	0.191	2320448
g	0.191	2316917
t	0.310	3752811

Plasmodium falciparum genome

seq	freq	occ
A	0.403	9196713
C	0.097	2210881
G	0.097	2210846
T	0.403	9194101

The Analysis of Regulatory Sequences

Matrix-matrix comparisons

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

Matrix against matrix

- How can we quantify the similarity between two position-specific scoring matrices ?
  - Chi-squared statistics can usually not be used, because it requires $n_{expected}$ >= 5 for each cell of the matrix.
  - Information theory (Kullback-Leibler distance)

Transfac matrix for yeast Pho4p

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

Matrix discovered by consensus

Pos	1	2	3	4	5	6	7	8	9	10
A	1	2	0	**5**	0	0	0	0	0	0
C	3	0	**5**	0	**5**	0	0	0	1	2
G	0	3	0	0	0	**5**	0	**5**	4	3
T	1	0	0	0	0	0	**5**	0	0	0

Motif comparison - mutual information

$$d(M_1, M_2) = \frac{1}{2W} \sum_{j=1}^{W} \sum_{b=k}^{T} \left(M_1(b,j) \log\left(\frac{M_1(b,j)}{M_2(b,j)} \right) + M_2(b,j) \log\left(\frac{M_2(b,j)}{M_1(b,j)} \right) \right)$$

M_1 *frequency matrix for the query motif (PSSM)*
M_2 *frequency matrix for the target motif (PSSM)*
W *width of the alignment between the two matrices*
T *number of residues in the alphabet (for DNA motifs, T=4)*

- Several authors used the ***mutual information*** to compare a query and a target motif.
  - Geert This (PhD thesis, 2003),
  - Stein Aerts (Bioinformatics, 2003)
  - Gary Stormo (ref ?)
- The mutual information is based on the Kullback-Leiber distance, calculated in both directions between the query and target motifs. The mutual information provides a symmetrical distance (contrarily to the Kullback-Leiber distance).
- Problem: how can we define a threshold on the distance to decide whether two matrices are similar or not ?

Motif comparison - mutual information

$$d(M_1, M_2) = \frac{1}{2W} \sum_{j=1}^{W} \sum_{b=k}^{T} \left(M_1(b,j) \log\left(\frac{M_1(b,j)}{M_2(b,j)} \right) + M_2(b,j) \log\left(\frac{M_2(b,j)}{M_1(b,j)} \right) \right)$$

- Note: frequency matrices must be corrected with a pseudo-weight, in order to avoid 0 values.

Transfac matrix for yeast Pho4p pseudo-weight 1

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.14	**0.36**	0.25	0.03	**0.92**	0.03	0.03	0.03	0.03	0.03	0.14	0.25
C	0.25	0.25	**0.36**	**0.92**	0.03	**0.92**	0.03	0.03	0.03	0.25	0.03	0.25
G	0.14	0.25	**0.36**	0.03	0.03	0.03	**0.92**	0.03	**0.58**	**0.47**	**0.58**	0.25
T	**0.47**	0.14	0.03	0.03	0.03	0.03	0.03	**0.92**	0.36	0.25	0.25	0.25

Matrix discovered by consensus

Pos	1	2	3	4	5	6	7	8	9	10
A	0.21	**0.38**	0.04	**0.88**	0.04	0.04	0.04	0.04	0.04	0.04
C	**0.54**	0.04	**0.88**	0.04	**0.88**	0.04	0.04	0.04	0.21	**0.38**
G	0.04	**0.54**	0.04	0.04	0.04	**0.88**	0.04	**0.88**	**0.71**	**0.54**
T	0.21	0.04	0.04	0.04	0.04	0.04	**0.88**	0.04	0.04	0.04

Motif comparison - mutual information

$$d(M_1, M_2) = \frac{1}{2W} \sum_{j=1}^{W} \sum_{b=k}^{T} \left(M_1(b,j) \log\left(\frac{M_1(b,j)}{M_2(b,j)} \right) + M_2(b,j) \log\left(\frac{M_2(b,j)}{M_1(b,j)} \right) \right)$$

The query motif is shifted 2 steps left compared to the reference motif

Transfac matrix for yeast Pho4p **pseudo-weight 1**

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.14	**0.36**	0.25	0.03	**0.92**	0.03	0.03	0.03	0.03	0.03	0.14	0.25
C	0.25	0.25	**0.36**	**0.92**	0.03	**0.92**	0.03	0.03	0.03	0.25	0.03	0.25
G	0.14	0.25	**0.36**	0.03	0.03	0.03	**0.92**	0.03	**0.58**	**0.47**	**0.58**	0.25
T	**0.47**	0.14	0.03	0.03	0.03	0.03	0.03	**0.92**	**0.36**	0.25	0.25	0.25

Matrix discovered by the program "consensus"

Pos	2	3	4	5	6	7	8	9	10			
A	0.25	0.03	**0.58**	0.03	0.03	0.03	0.03	0.03	0.03			
C	0.03	**0.58**	0.03	**0.58**	0.03	0.03	0.03	0.14	0.25			
G	**0.36**	0.03	0.03	0.03	**0.58**	0.03	**0.58**	**0.47**	**0.36**			
T	0.03	0.03	0.03	0.03	0.03	**0.58**	0.03	0.03	0.03			

Distance **0.5240**

Pos	2	3	4	5	6	7	8	9	10			
A	0.0142	0.1857	0.0613	0.0000	0.6749	0.0000	0.0000	0.0000	0.0000			
C	0.1060	0.0613	0.1857	0.0327	0.0000	0.6749	0.0000	0.0388	0.1060			
G	0.0461	0.1060	0.1857	0.0000	0.3673	0.0000	0.0327	0.2734	0.0231			
T	0.2734	0.0388	0.0000	0.0000	0.0000	0.3673	0.0000	0.6749	0.1857			
Sum	**0.4397**	**0.3918**	**0.4326**	**0.0327**	**1.0422**	**1.0422**	**0.0327**	**0.9872**	**0.3148**			

Motif comparison - mutual information

$$d(M_1, M_2) = \frac{1}{2W} \sum_{j=1}^{W} \sum_{b=k}^{T} \left(M_1(b,j) \log\left(\frac{M_1(b,j)}{M_2(b,j)} \right) + M_2(b,j) \log\left(\frac{M_2(b,j)}{M_1(b,j)} \right) \right)$$

The query motif is shifted 1 step left compared to the reference motif

Transfac matrix for yeast Pho4p **pseudo-weight 1**

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.14	**0.36**	0.25	0.03	**0.92**	0.03	0.03	0.03	0.03	0.03	0.14	0.25
C	0.25	0.25	**0.36**	**0.92**	0.03	**0.92**	0.03	0.03	0.03	0.25	0.03	0.25
G	0.14	0.25	**0.36**	0.03	0.03	0.03	**0.92**	0.03	**0.58**	**0.47**	**0.58**	0.25
T	**0.47**	0.14	0.03	0.03	0.03	0.03	0.03	**0.92**	**0.36**	0.25	0.25	0.25

Matrix discovered by the program "consensus"

Pos	1	2	3	4	5	6	7	8	9	10		
A	0.14	0.25	0.03	**0.58**	0.03	0.03	0.03	0.03	0.03	0.03		
C	**0.36**	0.03	**0.58**	0.03	**0.58**	0.03	0.03	0.03	0.14	0.25		
G	0.03	**0.36**	0.03	0.03	0.03	**0.58**	0.03	**0.58**	**0.47**	**0.36**		
T	0.14	0.03	0.03	0.03	0.03	0.03	**0.58**	0.03	0.03	0.03		

Distance **0.6167**

Pos	1	2	3	4	5	6	7	8	9	10		
A	0.0000	0.0089	0.1060	0.3673	0.6749	0.0000	0.0000	0.0000	0.0000	0.0000		
C	0.0089	0.1060	0.0231	0.6749	0.3673	0.6749	0.0000	0.0000	0.0388	0.0000		
G	0.0388	0.0089	0.1857	0.0000	0.0000	0.3673	0.6749	0.3673	0.0051	0.0065		
T	0.0886	0.0388	0.0000	0.0000	0.0000	0.0000	0.3673	0.6749	0.1857	0.1060		
Sum	**0.1363**	**0.1626**	**0.3148**	**1.0422**	**1.0422**	**1.0422**	**1.0422**	**1.0422**	**0.2296**	**0.1125**		

Motif comparison - mutual information

$$d(M_1, M_2) = \frac{1}{2W} \sum_{j=1}^{W} \sum_{b=k}^{T} \left(M_1(b,j) \log\left(\frac{M_1(b,j)}{M_2(b,j)} \right) + M_2(b,j) \log\left(\frac{M_2(b,j)}{M_1(b,j)} \right) \right)$$

The query motif is aligned with the reference motif

Transfac matrix for yeast Pho4p **pseudo-weight 1**

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.14	**0.36**	0.25	0.03	**0.92**	0.03	0.03	0.03	0.03	0.03	0.14	0.25
C	0.25	0.25	**0.36**	**0.92**	0.03	**0.92**	0.03	0.03	0.03	0.25	0.03	0.25
G	0.14	0.25	**0.36**	0.03	0.03	0.03	**0.92**	0.03	**0.58**	**0.47**	**0.58**	0.25
T	**0.47**	0.14	0.03	0.03	0.03	0.03	0.03	**0.92**	**0.36**	0.25	0.25	0.25

Matrix discovered by the program "consensus"

Pos		1	2	3	4	5	6	7	8	9	10	
A		0.14	0.25	0.03	**0.58**	0.03	0.03	0.03	0.03	0.03	0.03	
C		**0.36**	0.03	**0.58**	0.03	**0.58**	0.03	0.03	0.03	0.14	0.25	
G		0.03	**0.36**	0.03	0.03	0.03	**0.58**	0.03	**0.58**	**0.47**	**0.36**	
T		0.14	0.03	0.03	0.03	0.03	0.03	**0.58**	0.03	0.03	0.03	

Distance **0.1090**

Pos		1	2	3	4	5	6	7	8	9	10	
A		0.0461	0.0000	0.0000	0.0327	0.0000	0.0000	0.0000	0.0000	0.0000	0.0388	
C		0.0089	0.1857	0.0327	0.0000	0.0327	0.0000	0.0000	0.0000	0.0142	0.1060	
G		0.1060	0.0000	0.0000	0.0000	0.0000	0.0327	0.0000	0.0000	0.0000	0.0231	
T		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0327	0.1857	0.1060	0.1060	
Sum		**0.1610**	**0.1857**	**0.0327**	**0.0327**	**0.0327**	**0.0327**	**0.0327**	**0.1857**	**0.1202**	**0.2740**	

Motif comparison - mutual information

$$d(M_1, M_2) = \frac{1}{2W} \sum_{j=1}^{W} \sum_{b=k}^{T} \left(M_1(b,j) \log\left(\frac{M_1(b,j)}{M_2(b,j)} \right) + M_2(b,j) \log\left(\frac{M_2(b,j)}{M_1(b,j)} \right) \right)$$

The query motif is shifted one step right to the reference motif

Transfac matrix for yeast Pho4p **pseudo-weight 1**

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.14	**0.36**	0.25	0.03	**0.92**	0.03	0.03	0.03	0.03	0.03	0.14	0.25
C	0.25	0.25	**0.36**	**0.92**	0.03	**0.92**	0.03	0.03	0.03	0.25	0.03	0.25
G	0.14	0.25	**0.36**	0.03	0.03	0.03	**0.92**	0.03	**0.58**	**0.47**	**0.58**	0.25
T	**0.47**	0.14	0.03	0.03	0.03	0.03	0.03	**0.92**	**0.36**	0.25	0.25	0.25

Matrix discovered by the program "consensus"

Pos			1	2	3	4	5	6	7	8	9	10
A			0.14	0.25	0.03	**0.58**	0.03	0.03	0.03	0.03	0.03	0.03
C			**0.36**	0.03	**0.58**	0.03	**0.58**	0.03	0.03	0.03	0.14	0.25
G			0.03	**0.36**	0.03	0.03	0.03	**0.58**	0.03	**0.58**	**0.47**	**0.36**
T			0.14	0.03	0.03	0.03	0.03	0.03	**0.58**	0.03	0.03	0.03

Distance **0.6391**

Pos			1	2	3	4	5	6	7	8	9	10
A			0.0142	0.1060	0.6749	0.3673	0.0000	0.0000	0.0000	0.0000	0.0388	0.1060
C			0.0000	0.6749	0.3673	0.6749	0.3673	0.0000	0.0000	0.1060	0.0388	0.0000
G			0.1857	0.1857	0.0000	0.0000	0.6749	0.3673	0.3673	0.0051	0.0051	0.0089
T			0.0388	0.0000	0.0000	0.0000	0.0000	0.6749	0.0231	0.1060	0.1060	0.1060
Sum			**0.2387**	**0.9666**	**1.0422**	**1.0422**	**1.0422**	**1.0422**	**0.3904**	**0.2172**	**0.1888**	**0.2209**

Motif comparison - mutual information

$$d(M_1, M_2) = \frac{1}{2W} \sum_{j=1}^{W} \sum_{b=k}^{T} \left(M_1(b,j) \log\left(\frac{M_1(b,j)}{M_2(b,j)} \right) + M_2(b,j) \log\left(\frac{M_2(b,j)}{M_1(b,j)} \right) \right)$$

The query motif is shifted one step right to the reference motif

Transfac matrix for yeast Pho4p **pseudo-weight 1**

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.14	**0.36**	0.25	0.03	**0.92**	0.03	0.03	0.03	0.03	0.03	0.14	0.25
C	0.25	0.25	**0.36**	**0.92**	0.03	**0.92**	0.03	0.03	0.03	0.25	0.03	0.25
G	0.14	0.25	**0.36**	0.03	0.03	0.03	**0.92**	0.03	**0.58**	**0.47**	**0.58**	0.25
T	**0.47**	0.14	0.03	0.03	0.03	0.03	0.03	**0.92**	**0.36**	0.25	0.25	0.25

Matrix discovered by the program "consensus"

Pos				1	2	3	4	5	6	7	8	9
A				0.14	0.25	0.03	**0.58**	0.03	0.03	0.03	0.03	0.03
C				**0.36**	0.03	**0.58**	0.03	**0.58**	0.03	0.03	0.03	0.14
G				0.03	**0.36**	0.03	0.03	0.03	**0.58**	0.03	**0.58**	**0.47**
T				0.14	0.03	0.03	0.03	0.03	0.03	**0.58**	0.03	0.03

Distance **0.4121**

Pos				1	2	3	4	5	6	7	8	9
A				0.0388	0.1881	0.0000	0.3673	0.0000	0.0000	0.0000	0.0388	0.1060
C				0.1124	0.0000	0.0327	0.0000	0.3673	0.0000	0.1060	0.0000	0.0142
G				0.0000	0.1857	0.0000	0.6749	0.0000	0.0000	0.2734	0.0000	0.0307
T				0.0388	0.0000	0.0000	0.0000	0.6749	0.1857	0.0613	0.1060	0.1060
Sum				**0.1900**	**0.3737**	**0.0327**	**1.0422**	**1.0422**	**0.1857**	**0.4408**	**0.1449**	**0.2569**

Matrix-to-matrix comparison: software

- T-Reg Comparator
  - Roepcke et al. (2005). NAR 33 (Web Server Issue):W438-441.
  - Compares a query matrix against all matrices annotated in TRANSFAC and JASPAR.
  - http://treg.molgen.mpg.de/

Dissimilarity between matrices

- Sandelin et al (2003) define a dissimilarity between two columns of a frequency matrix (formulae are adapted from Sandelin)
  - A,B two frequency matrices (PSSM)
  - $A_{r,i}$ frequency of residue b in the i^{th} column of matrix A
  - $B_{r,j}$ frequency of residue b in the j^{th} column of matrix B
  - $S_{Ai,Bj}$ Dissimilarity score between column i of matrix A and column j of matrix B
  - w shortest width of the two aligned patterns
  - F total score for an alignment of w columns, with an offset of g for matrix A and h for matrix B
  - N normalized score

$$S_{A_i,B_j} = 2 - \sum_{r\in\{A,C,G,T\}} \left(A_{r,i} - B_{r,j}\right)^2$$

$$F = \sum_{k=0}^{w+1} S_{A_{g+k},B_{h+k}}$$

$$N = F/2w$$

Sandelin, A., A. Hoglund, B. Lenhard, and W.W. Wasserman. 2003. Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct Integr Genomics* **3:** 125-134.

References

- Matrix-matrix comparisons
  - Stormo, G.D. 2000. DNA binding sites: representation and discovery. *Bioinformatics* **16:** 16-23.
  - Sandelin, A., A. Hoglund, B. Lenhard, and W.W. Wasserman. 2003. Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct Integr Genomics* **3:** 125-134.

The Analysis of Regulatory Sequences

Matrix-string comparisons

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

Scoring a site with a matrix

- A simple way to compare a matrix-based motif (PSSM) with a string-based motif is to calculate the weight of the site according to the PSSM (see chapter on matrix-based pattern matching)
- This however returns continuous values, and we still need a criterion to decide whether the considered matrix and a site do or not represent the same motif.

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

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

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

Scoring a site with a matrix

- Actually, we would like to estimate P(M|S), the probability for sequence S to be an instance of the motif M, rather than P(S|M)
- This can be calculated with Bayes' rule

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{P(S \mid M)P(M)}{P(S \mid M)P(M) + P(S \mid B)P(B)}$$

$P(S\|M)$	probability of the sequence segment, given the motif
$P(S\|B)$	probability of the sequence segment, given the background
$P(S)$	probability of the sequence segment (given the motif and the background)
$P(B)$	prior probability of the background
$P(M)$	prior probability of the motif

- However, for this we need to estimate
  - *P(B)*, the prior probabilities of the background
  - *P(M)*, the prior probabilities of the motif
- We thus need to estimate the frequency of occurrences of the motif in the sequence.
- Geert Thijs proposes an approach to estimate these priors (Thijs 2003, PhD thesis)