Position-specific scoring matrices (PSSM)

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

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

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

Introduction

- In the biological literature, the binding specificity of a transcription factor is often represented with a consensus string, which can be strict (e.g. CAGTGggg) or include some ambiguous residues (e.g. CACGTW).
- This representation is convenient to speak about a TF binding specificity, but it is by no way operational to predict TFBS.
- We describe in the following slides the theoretical grounds of the most commonly used representation models for transcription factor binding specificity: position-specific scoring matrices (TFBM).

Consensus representation

- The TRANSFAC database contains 8 binding sites for the yeast transcription factor Pho4p
  - 5/8 contain the core of high-affinity binding sites (CACGTG)
  - 3/8 contain the core of medium-affinity binding sites (CACGTT)
- The IUPAC ambigous nucleotide code allows to represent variable residues.
- 15 letters to represent any possible combination between the 4 nucleotides (24 – 1=15).
- This representation however gives a poor idea of the relative importance of residues.

```
R06098   \TCACACGTGGGA\
R06099   \GGCCACGTGCAG\
R06100   \TGACACGTGGGT\
R06102   \CAGCACGTGGGG\
R06103   \TTCCACGTGCGA\
R06104   \ACGCACGTTGGT\
R06097   \CAGCACGTTTTC\
R06101   \TACCACGTTTTC\

Cons       nnVCACGTKBDn
```

IUPAC ambiguous nucleotide code

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

- TRANSFAC pulic version: http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/search.cgi

From alignments to weights

Building a position-specific scoring matrix from a collection of sites

Alignment of Pho4p binding sites (TRANSFAC annotations)

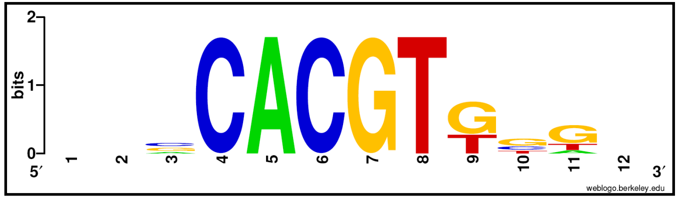
R06098	T	C	A	C	A	C	G	T	G	G	G	A
R06099	G	G	C	C	A	C	G	T	G	C	A	G
R06100	T	G	A	C	A	C	G	T	G	G	G	T
R06102	C	A	G	C	A	C	G	T	G	G	G	G
R06103	T	T	C	C	A	C	G	T	G	C	G	A
R06104	A	C	G	C	A	C	G	T	T	G	G	T
R06097	C	A	G	C	A	C	G	T	T	T	T	C
R06101	T	A	C	C	A	C	G	T	T	T	T	C

Count matrix (TRANSFAC matrix F$PHO4_01)

Residue\position	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
Sum	8	8	8	8	8	8	8	8	8	8	8	8

Tom Schneider's sequence logo

(generated with Web Logo http://weblogo.berkeley.edu/logo.cgi)

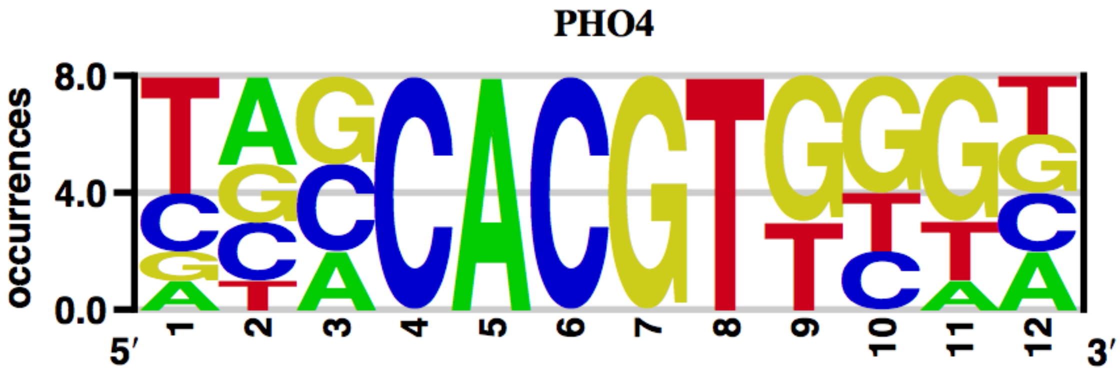


Residue count matrix

Count matrix (TRANSFAC matrix F$PHO4_01)

Residue\position	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
Sum	8	8	8	8	8	8	8	8	8	8	8	8

Tom Schneider's sequence logo
(generated with Web Logo http://weblogo.berkeley.edu/logo.cgi)

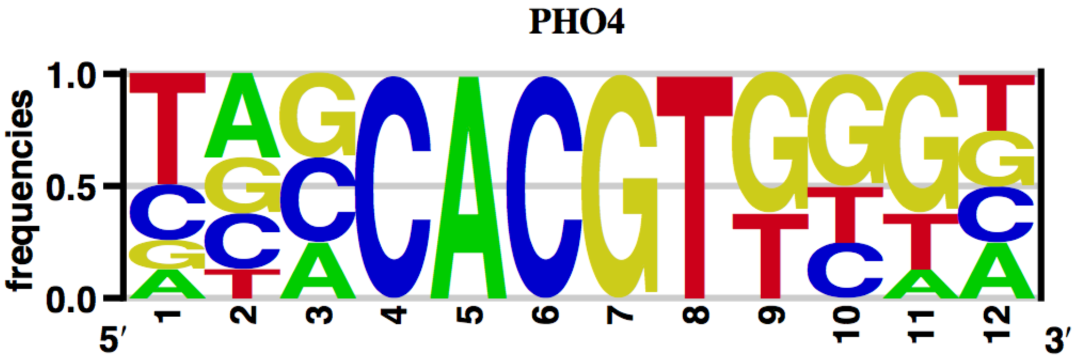


Frequency matrix

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12
A	0,125	0,375	0,250	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000	0,125	0,250
C	0,250	0,250	**0,375**	**1,000**	0,000	**1,000**	0,000	0,000	0,000	0,250	0,000	0,250
G	0,125	0,250	**0,375**	0,000	0,000	0,000	**1,000**	0,000	**0,625**	**0,500**	**0,625**	0,250
T	0,500	0,125	0,000	0,000	0,000	0,000	0,000	**1,000**	**0,375**	0,250	0,250	0,250
Sum	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

$$f_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^{A} n_{i,j}}$$

A	*alphabet size (=4)*
$n_{i,j,}$	*occurrences of residue i at position j*
p_i	*prior residue probability for residue i*
$f_{i,j}$	*relative frequency of residue i at position j*



Count matrix with pseudo-count

1st option: identically distributed pseudo-weight (equiprobable residue priors)

Count matrix with pseudo-count — **k= 1** — **Equiprobable residues**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	1,25	3,25	2,25	0,25	**8,25**	0,25	0,25	0,25	0,25	0,25	1,25	2,25	0,25
C	2,25	2,25	**3,25**	**8,25**	0,25	**8,25**	0,25	0,25	0,25	2,25	0,25	2,25	0,25
G	1,25	2,25	**3,25**	0,25	0,25	0,25	**8,25**	0,25	**5,25**	**4,25**	**5,25**	2,25	0,25
T	4,25	1,25	0,25	0,25	0,25	0,25	0,25	**8,25**	**3,25**	2,25	2,25	2,25	0,25
Sum	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	1,00

$$f'_{i,j} = \frac{n_{i,j} + k/A}{\sum_{i=1}^{A} n_{i,j} + k}$$

2nd option: pseudo-weights distributed according to residue-specific priors

Count matrix with pseudo-count — **k= 1** — **Specific nucleotide frequencies**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	1,33	3,33	2,33	0,33	**8,33**	0,33	0,33	0,33	0,33	0,33	1,33	2,33	0,33
C	2,17	2,17	**3,17**	**8,17**	0,17	**8,17**	0,17	0,17	0,17	2,17	0,17	2,17	0,17
G	1,17	2,17	**3,17**	0,17	0,17	0,17	**8,17**	0,17	**5,17**	**4,17**	**5,17**	2,17	0,17
T	4,33	1,33	0,33	0,33	0,33	0,33	0,33	**8,33**	**3,33**	2,33	2,33	2,33	0,33
Sum	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	9,00	1,00

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}$$

A *alphabet size (=4)*

$n_{i,j,}$ *occurrences of residue i at position j*

p_i *prior residue probability for residue i*

$f_{i,j}$ *relative frequency of residue i at position j*

k *pseudo weight (arbitrary, 1 in this case)*

$f'_{i,j}$ *corrected frequency of residue i at position j*

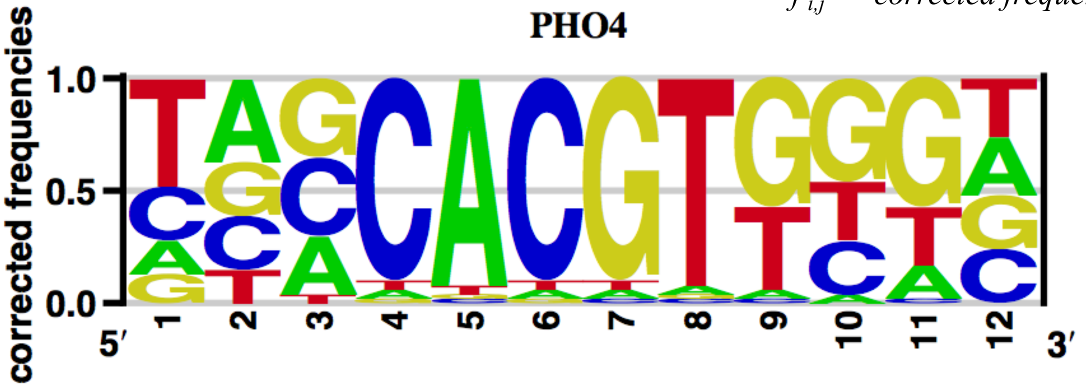
Corrected frequency matrix

Frequency matrix corrected with pseudo-count **k= 1** **Specific nucleotide frequencies**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	0,148	0,370	0,259	0,037	**0,926**	0,037	0,037	0,037	0,037	0,037	0,148	0,259	0,33
C	0,241	0,241	**0,352**	**0,908**	0,019	**0,908**	0,019	0,019	0,019	0,241	0,019	0,241	0,17
G	0,130	0,241	**0,352**	0,019	0,019	0,019	**0,908**	0,019	**0,574**	**0,463**	**0,574**	0,241	0,17
T	0,481	0,148	0,037	0,037	0,037	0,037	0,037	**0,926**	**0,370**	0,259	0,259	0,259	0,33
Sum	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,00

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}$$

- A — *alphabet size (=4)*
- $n_{i,j}$, — *occurrences of residue i at position j*
- p_i — *prior residue probability for residue i*
- $f_{i,j}$ — *relative frequency of residue i at position j*
- k — *pseudo weight (arbitrary, 1 in this case)*
- $f'_{i,j}$ — *corrected frequency of residue i at position j*



Weight matrix (Bernoulli model)

Weight matrix — **k= 1** — **Specific nucleotide frequencies**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (p_i)
A	-0,35	0,05	-0,11	-0,95	**0,45**	-0,95	-0,95	-0,95	-0,95	-0,95	-0,35	-0,11	0,33
C	0,15	0,15	**0,32**	**0,73**	-0,95	**0,73**	-0,95	-0,95	-0,95	0,15	-0,95	0,15	0,17
G	-0,12	0,15	**0,32**	-0,95	-0,95	-0,95	**0,73**	-0,95	**0,53**	**0,44**	**0,53**	0,15	0,17
T	0,16	-0,35	-0,95	-0,95	-0,95	-0,95	-0,95	**0,45**	**0,05**	-0,11	-0,11	-0,11	0,33
Sum	-0,150	0,004	-0,427	-2,135	-2,415	-2,135	-2,135	-2,415	-1,330	-0,472	-0,880	0,093	1,00

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{r=1}^{A} n_{r,j} + k}$$

$$W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

A *alphabet size (=4)*
$n_{i,j,}$ *occurrences of residue i at position j*
p_i *prior residue probability for residue i*
$f_{i,j}$ *relative frequency of residue i at position j*
k *pseudo weight (arbitrary, 1 in this case)*
$f'_{i,j}$ *corrected frequency of residue i at position j*
$W_{i,j}$ *weight of residue i at position j*

The use of a weight matrix relies on Bernoulli assumption

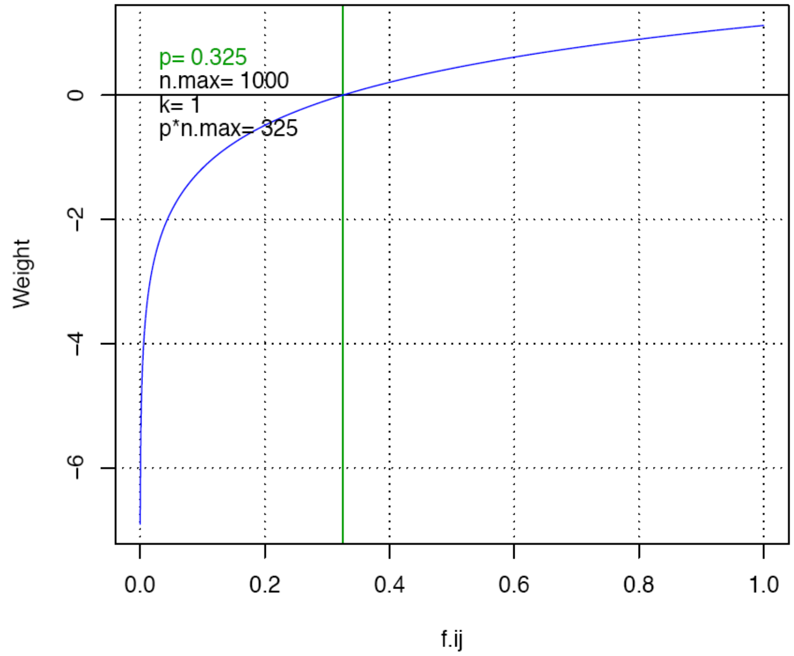
If we assume, for the background model, an independent succession of nucleotides (Bernoulli model), the weight W_S of a sequence segment *S* is simply the sum of weights of the nucleotides at successive positions of the matrix ($W_{i,j}$).

In this case, it is convenient to convert the PSSM into a weight matrix, which can then be used to assign a score to each position of a given sequence.

Properties of the weight function

$$W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k} \qquad \sum_{i=1}^{A} f'_{i,j} = 1$$



- The weight is
  - *positive* when $f'_{i,j} > p_i$ (*favourable* positions for the binding of the transcription factor)
  - *negative* when $f'_{i,j} < p_i$ (*unfavourable* positions)

Information content

Shannon uncertainty

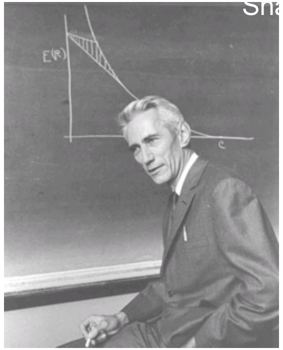
- Shannon uncertainty
  - Hs(j): uncertainty of a column of a PSSM
  - Hg: uncertainty of the background (e.g. a genome)
- Special cases of uncertainty (for a 4 letter alphabet)
  - min(H)=0
    - No uncertainty at all: the nucleotide is completely specified (e.g. p={1,0,0,0})
  - H=1
    - Uncertainty between two letters (e.g. p={0.5,0,0,0.5})
  - max(H) = 2 (Complete uncertainty)
    - One bit of information is required to specify the choice between each alternative (e.g. p={0.25,0.25,0.25,0.25}).
    - Two bits are required to specify a letter in a 4-letter alphabet.
- R_{seq}
  - Schneider (1986) defines an information content based on Shannon's uncertainty.
- R^*_{seq}
  - For skewed genomes (i.e. unequal residue probabilities), Schneider recommends an alternative formula for the information content .
  - This is the formula that is nowadays used.

$$H_s(j) = -\sum_{i=1}^{A} f_{i,j} \log_2(f_{i,j})$$

$$H_g = -\sum_{i=1}^{A} p_i \log_2(p_i)$$

$$R_{seq}(j) = H_g - H_s(j) \qquad R_{seq} = \sum_{j=1}^{w} R_{seq}(j)$$

$$R^*_{seq}(j) = \sum_{i=1}^{A} f_{i,j} \log_2\left(\frac{f_{i,j}}{p_i}\right) \qquad R^*_{seq} = \sum_{j=1}^{w} R^*_{seq}(j)$$



Adapted from Schneider et al. Information content of binding sites on nucleotide sequences. J Mol Biol (1986) vol. 188 (3) pp. 415-31. PMID 3525846.

Information content of a PSSM

Information content matrix **k= 1** **Specific nucleotide frequencies**

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12	Prior (pi)
A	-0,12	0,04	-0,06	-0,08	**0,95**	-0,08	-0,08	-0,08	-0,08	-0,08	-0,12	-0,06	0,33
C	0,08	0,08	**0,26**	**1,52**	-0,04	**1,52**	-0,04	-0,04	-0,04	0,08	-0,04	0,08	0,17
G	-0,03	0,08	**0,26**	-0,04	-0,04	-0,04	**1,52**	-0,04	**0,70**	**0,46**	**0,70**	0,08	0,17
T	0,18	-0,12	-0,08	-0,08	-0,08	-0,08	-0,08	**0,95**	**0,04**	-0,06	-0,06	-0,06	0,33
Sum	0,112	0,092	0,370	1,318	0,791	1,318	1,318	0,791	0,620	0,405	0,476	0,043	1,00

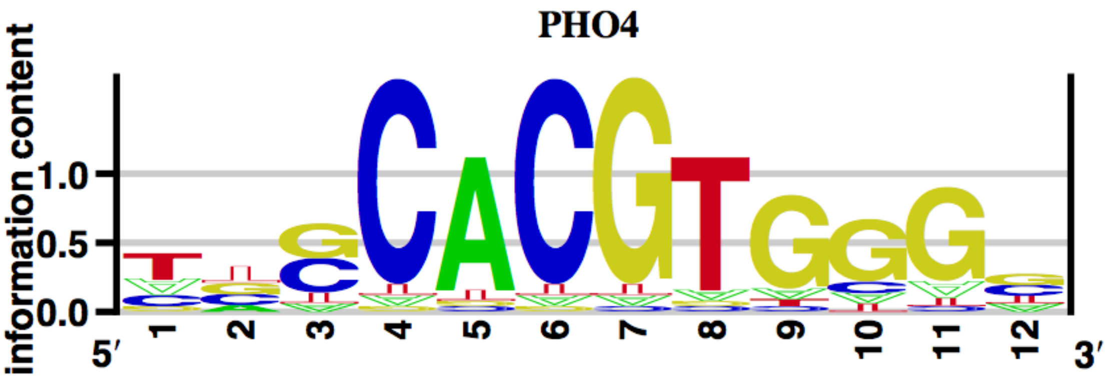
$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}$$

$$I_{i,j} = f'_{i,j} \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

$$I_j = \sum_{i=1}^{A} I_{i,j}$$

$$I_{matrix} = \sum_{j=1}^{w} \sum_{i=1}^{A} I_{i,j}$$

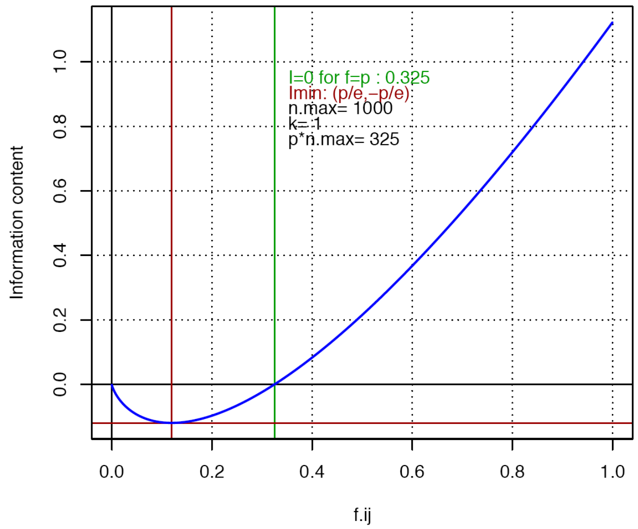
A	*alphabet size (=4)*
$n_{i,j,}$	*occurrences of residue i at position j*
w	*matrix width (=12)*
p_i	*prior residue probability for residue i*
$f_{i,j}$	*relative frequency of residue i at position j*
k	*pseudo weight (arbitrary, 1 in this case)*
$f'_{i,j}$	*corrected frequency of residue i at position j*
$W_{i,j}$	*weight of residue i at position j*
$I_{i,j}$	*information of residue i at position j*



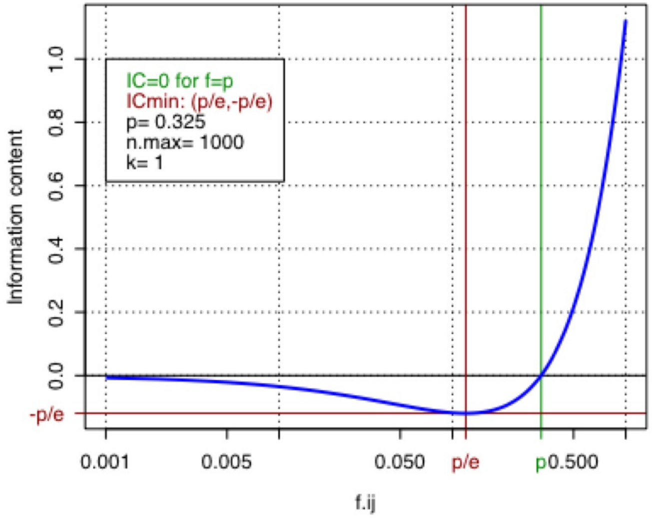
Information content I_{ij} of a cell of the matrix

- For a given cell of the matrix
  - I_{ij} is positive when $f'_{ij} > p_i$
    (i.e. when residue *i* is more frequent at position *j* than expected by chance)
  - I_{ij} is negative when $f'_{ij} < p_i$
  - I_{ij} tends towards 0 when $f'_{ij} \to 0$
    because $limit_{x \to 0} (x \ln(x)) = 0$

Information content as a function of residue frequency



Information content as a function of residue frequency (log scale)



Information content of a column of the matrix

- For a given column i of the matrix
  - The information of the column (I_j) is the sum of information of its cells.
  - I_j is always positive
  - I_j is 0 when the frequency of all residues equal their prior probability ($f_{ij}=p_i$)
  - I_j is maximal when
    - the residue i_m with the lowest prior probability has a frequency of 1
      (all other residues have a frequency of 0)
    - and the pseudo-weight is null ($k=0$).

$$I_j = \sum_{i=1}^{A} I_{i,j} = \sum_{i=1}^{A} f'_{i,j} \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

$$i_m = \arg\min_i(p_i) \qquad k = 0$$

$$\max(I_j) = 1 \cdot \ln\left(\frac{1}{p_i}\right) = -\ln(p_i)$$

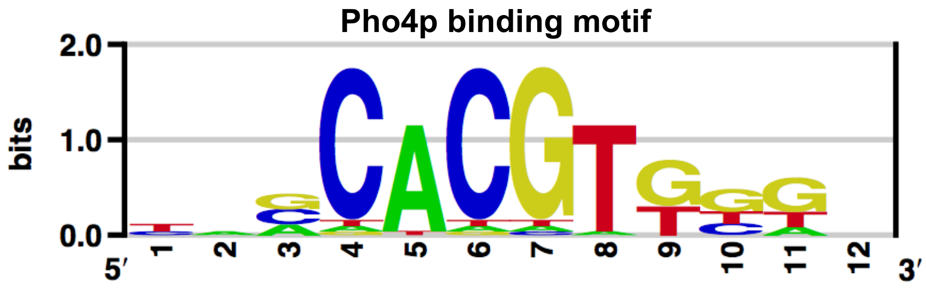
Schneider logos

- Schneider & Stephens(1990) propose a graphical representation based on his previous entropy (H) for representing the importance of each residue at each position of an alignment. He provides a new formula for Rseq
  - Hs(j) uncertainty of column j
  - Rseq(j) "information content" of column j (beware, this definition differs from Hertz' information content)
  - e(n) correction for small samples (pseudo-weight)
- Remarks
  - This information content does not include any correction for the prior residue probabilities (pi)
  - This information content is expressed in bits.
- Boundaries
  - min(Rseq)=0 equiprobable residues
  - max(Rseq)=2 perfect conservation of 1 residue with a pseudo-weight of 0,
- Sequence logos can be generated
  - from aligned sequences on the Weblogo server http://weblogo.berkeley.edu/logo.cgi
  - From matrices or sequences on enologos http://www.benoslab.pitt.edu/cgi-bin/enologos/enologos.cgi

$$H_s(j) = -\sum_{i=1}^{A} f_{ij} \log_2(f_{ij})$$

$$R_{seq}(j) = 2 - H_s(j) + e(n)$$

$$h_{ij} = f_{ij} R_{seq}(j)$$



Information content of the matrix

- The total information content represents the capability of the matrix to make the distinction between a binding site (represented by the matrix) and the background model.
- The information content also allows to estimate an upper limit for the expected frequency of the binding sites in random sequences.
- The pattern discovery program consensus (developed by Jerry Hertz) optimises the information content in order to detect over-represented motifs.
- Note that this is not the case of all pattern discovery programs: the gibbs sampler algorithm optimizes a log-likelihood.

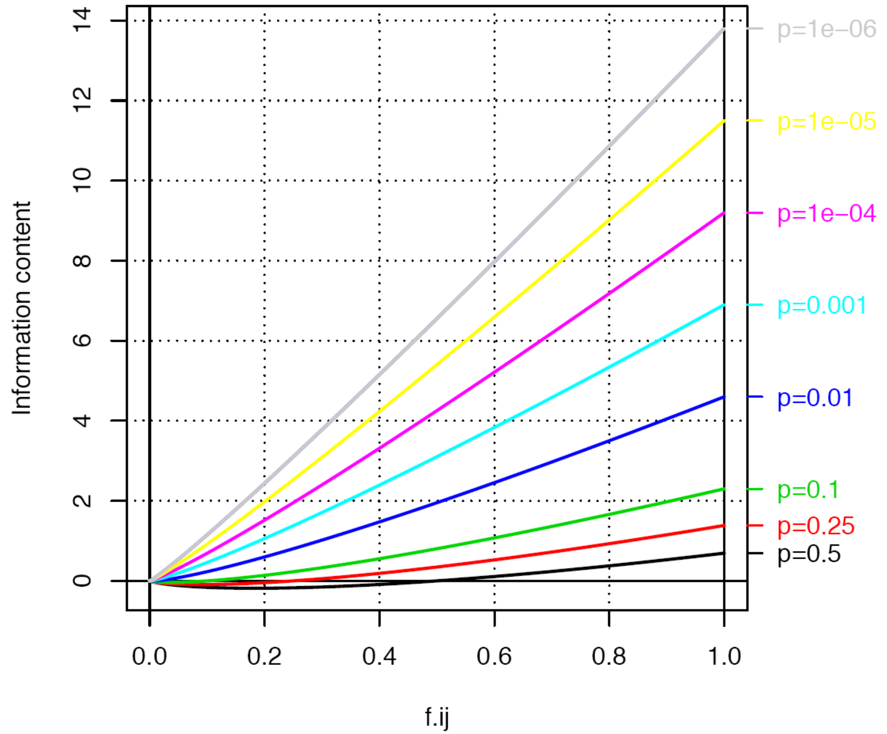
$$I_{matrix} = \sum_{j=1}^{w} \sum_{i=1}^{A} I_{i,j}$$

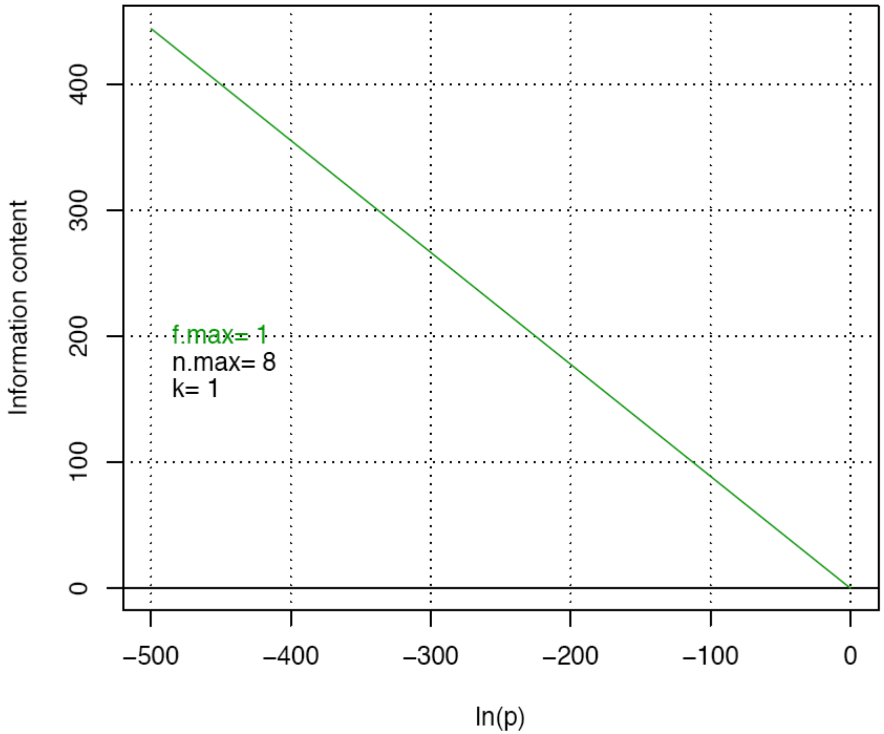
$$P(site) \leq e^{-I_{matrix}}$$

- Hertz and Stormo (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15: 563-77. PMID 1

Information content: effect of prior probabilities

- The upper bound of I_j increases when p_i decreases
  - $I_j \rightarrow Inf$ when $p_i \rightarrow 0$
- The information content, as defined by Gerald Hertz, has thus no upper bound.





References - PSSM information content

- Seminal articles by Tom Schneider
  - Schneider, T.D., G.D. Stormo, L. Gold, and A. Ehrenfeucht. 1986. Information content of binding sites on nucleotide sequences. J Mol Biol 188: 415-431.
  - Schneider, T.D. and R.M. Stephens. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097-6100.
  - Tom Schneider's publications online
    - http://www.lecb.ncifcrf.gov/~toms/paper/index.html
- Seminal article by Gerald Hertz
  - Hertz, G.Z. and G.D. Stormo. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15: 563-577.
- Software tools to draw sequence logos
  - Weblogo
    - http://weblogo.berkeley.edu/logo.cgi
  - Enologos
    - http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi