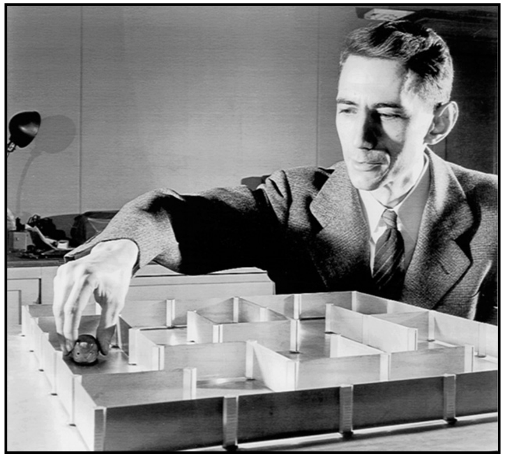
Information content

The information content is related to the number of *binary decisions* required to find the information.

The number of binary decisions (= number of questions whose answer is yes/no) required to find the correct element in a set of N elements is:

$$n_q = \log_2 N$$



Claude Elwood Shannon
(1916-2001)

Example: consider the 4 nucleotides, {A,C,G,T}.

John: I have a nucleotide in mind. Guess which one!
Mike: is it A or T?
John: no!
Mike: is it G?
John: yes!

Thus, $\log_2 4$ = 2 questions where necessary to find unambiguously the right nucleotide.

If we had *N* elements, we first ask if the hidden element is among the N/2 first elements, then in the N/4 left elements, then in N/8 elements, etc, till we reach the 1 correct element. There are thus

$$\begin{aligned} n_q &= 1(N/2?) + 1(N/4?) + 1(N/8?) +\dots 1(N/N) \\ &= 1(N/2?)+1(N/2^2?)+1(N/2^3?)+ \dots 1 (N/2^{\log_2 N}) \\ &= \log_2 N \end{aligned}$$

questions to ask.

Information content

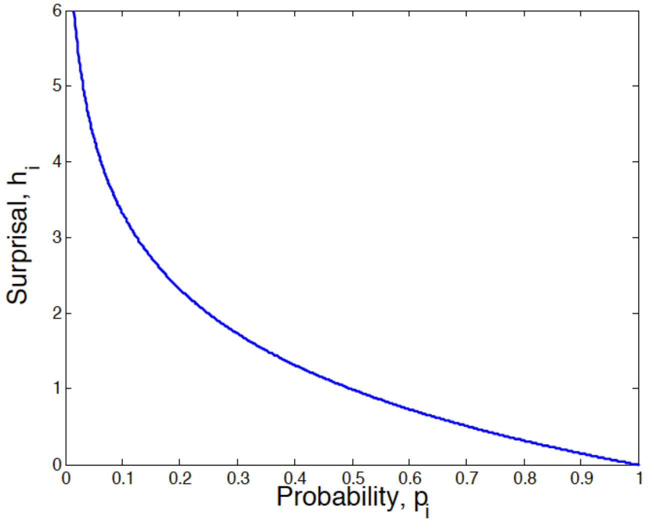
If each element has the same probability, then this probability is $p=1/N$, and

$$n_q = \log_2 N = -\log_2 p$$

In general the elements are not equally likely; they have different probabilities, p_i.

Tribus (1961) then generalizes the formula here above by introducing the concept of *"surprisal"* h_i:

$$h_i = -\log_2 p_i$$



Interpretation:

- if p_i approaches 0, then we will be very surprised to see the ith symbol (since it should almost never appear), and the formula says h_i approaches ∞.
- if p_i=1, then we won't be surprised at all to see the ith symbol (because it should always appear) and $h_i = 0$.

Source: Schneider (1997)

Information content

If each element has the same probability, then this probability is *p=1/N*, and

$$n_q = \log_2 N = -\log_2 p$$

In general the elements are not equally likely; they have different probabilities, p_i.

Tribus (1961) then generalizes the formula here above by introducing the concept of *"surprisal"* h_i:

$$h_i = -\log_2 p_i$$

On this basis, Shannon introduced the "*uncertainty measure*" (also called *"**entropy**"*), which is the average of all suprisals h_i weighted by their occurrence p_i:

$$H = \sum_i p_i h_i = -\sum_i p_i \log_2 p_i$$

Special cases of uncertainty (shown here for a 4 letter alphabet, {A,C,G,T})

p={1,0,0,0}
H=min(H)=0
No uncertainty at all: the nucleotide is completely specified.

p={0.5,0,0,0.5}
H=1
Uncertainty between two letters

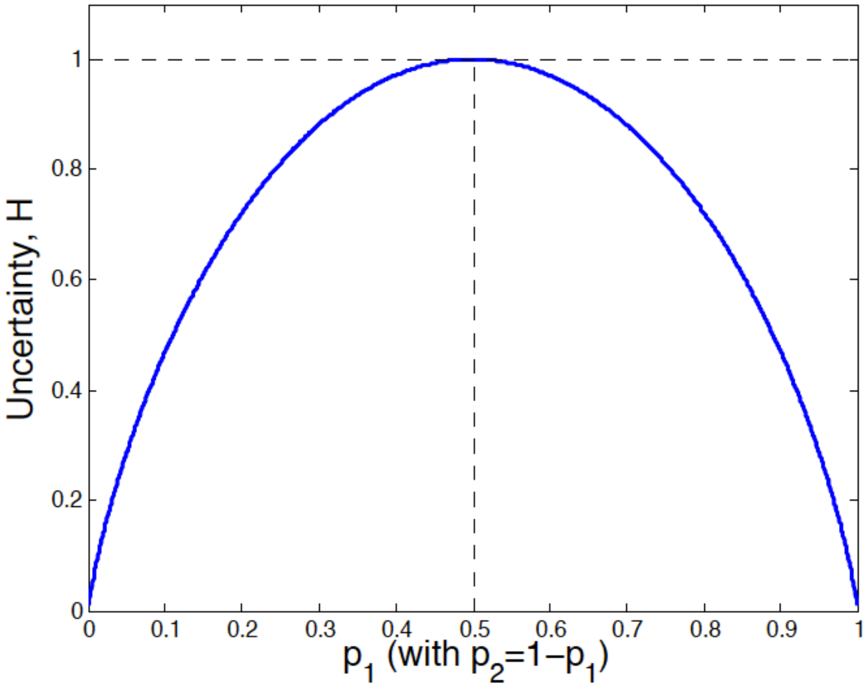
p={0.25,0.25,0.25,0.25}
H=max(H)=2
Complete uncertainty

Source: Schneider (1997)

Information content

Case of 2 symbols:

Uncertainty H for the case of 2 symbols, as a function of the probability p_1 of one symbol (with p_2=1-p_1).



Exercise

Calculate the uncertainty of the following probability sets:

p={1/2,1/4,1/8,1/8}

p={1/2,1/6,1/6,1/6}

NB: For the second case, you will "certainly" need a computer...

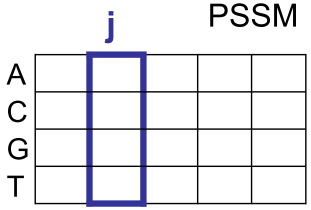
Check that the uncertainty is indeed comprised between 0 and 2 (= maximum uncertainty calculated above).

Source: Schneider (1997)

Information content

Shannon uncertainty applied to PSSM

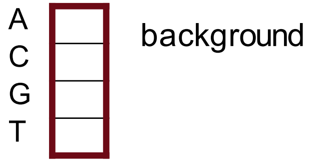
Uncertainty of a column *j* of a PSSM:



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

A is the number of possible elements (4 letter for DNA, 20 for aa) and f_{ij} *is the frequency of letter i in column j.*

Uncertainty of the background (e.g. the GC content of a genome):



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

p_i *is the background frequency of letter i.*

Schneider (1986) defines an *information content* based on Shannon's uncertainty.

$$R_{seq}(j) = H_g - H_s(j)$$ for column *j*

$$R_{seq} = \sum_{j=1}^{w} R_{seq}(j)$$ for the PSSM

For skewed genomes (i.e. unequal residue probabilities), Schneider recommends an alternative formula for the information content.

$$R^*_{seq}(j) = \sum_{i=1}^{A} f_{ij} \log_2\left(\frac{f_{ij}}{p_i}\right)$$ for column *j*

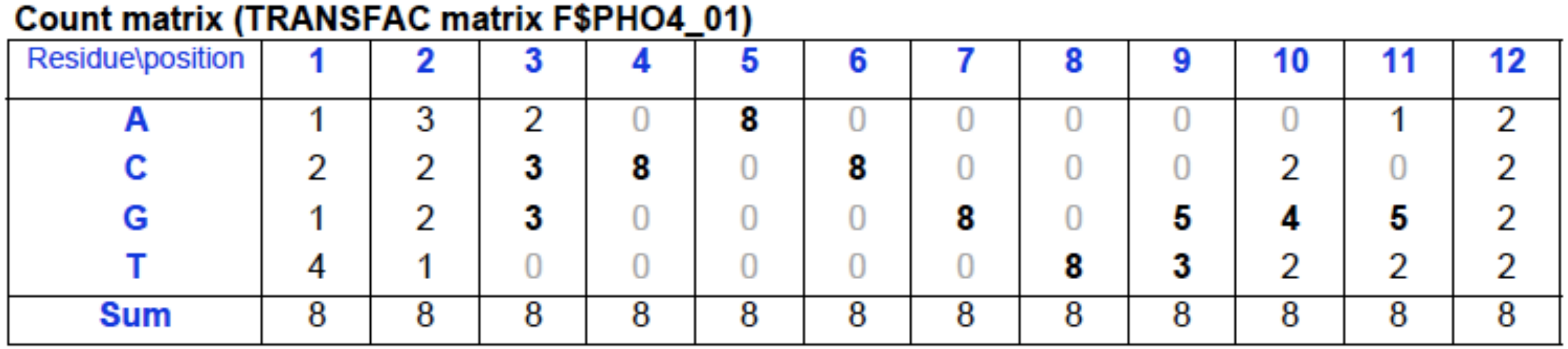
$$R^*_{seq} = \sum_{j=1}^{w} R^*_{seq}(j)$$ for the PSSM

This is the formula that is nowadays used.

Source: J. van Helden

Information content

Shannon uncertainty applied to PSSM: example



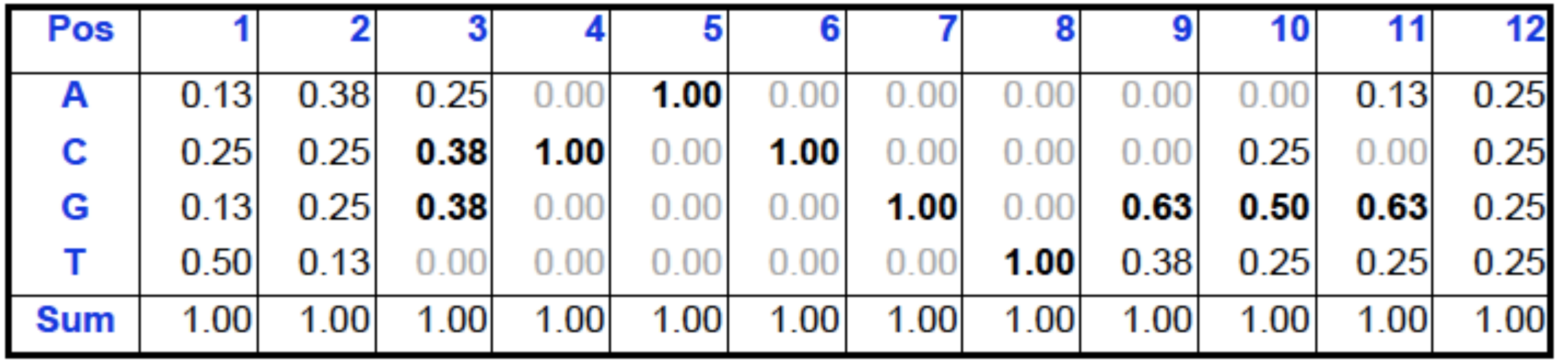
Count matrix (TRANSFAC matrix F$PHO4_01)

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

Counts

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

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*



Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.13	0.38	0.25	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.13	0.25
C	0.25	0.25	**0.38**	**1.00**	0.00	**1.00**	0.00	0.00	0.00	0.25	0.00	0.25
G	0.13	0.25	**0.38**	0.00	0.00	0.00	**1.00**	0.00	**0.63**	**0.50**	**0.63**	0.25
T	0.50	0.13	0.00	0.00	0.00	0.00	0.00	**1.00**	0.38	0.25	0.25	0.25
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

Frequency

Source: J. van Helden

Information content

Shannon uncertainty applied to PSSM: example

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.13	0.38	0.25	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00	0.13	0.25
C	0.25	0.25	**0.38**	**1.00**	0.00	**1.00**	0.00	0.00	0.00	0.25	0.00	0.25
G	0.13	0.25	**0.38**	0.00	0.00	0.00	**1.00**	0.00	**0.63**	**0.50**	**0.63**	0.25
T	0.50	0.13	0.00	0.00	0.00	0.00	0.00	**1.00**	0.38	0.25	0.25	0.25
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

Frequency

1st option: identically distributed pseudo-weight

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

2nd option: pseudo-weight distributed according to residue priors

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

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*

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.15	0.37	0.26	0.04	**0.93**	0.04	0.04	0.04	0.04	0.04	0.15	0.26
C	0.24	0.24	**0.35**	**0.91**	0.02	**0.91**	0.02	0.02	0.02	0.24	0.02	0.24
G	0.13	0.24	**0.35**	0.02	0.02	0.02	**0.91**	0.02	**0.58**	**0.46**	**0.58**	0.24
T	0.48	0.15	0.04	0.04	0.04	0.04	0.04	**0.93**	0.37	0.26	0.26	0.26
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

Frequency
(+pseudo-counts)

Source: J. van Helden

Information content

Shannon uncertainty applied to PSSM: example

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.15	0.37	0.26	0.04	**0.93**	0.04	0.04	0.04	0.04	0.04	0.15	0.26
C	0.24	0.24	**0.35**	**0.91**	0.02	**0.91**	0.02	0.02	0.02	0.24	0.02	0.24
G	0.13	0.24	**0.35**	0.02	0.02	0.02	**0.91**	0.02	**0.58**	**0.46**	**0.58**	0.24
T	0.48	0.15	0.04	0.04	0.04	0.04	0.04	**0.93**	0.37	0.26	0.26	0.26
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

Frequency

(+pseudo-counts)

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

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

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

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.79	0.13	-0.23	-2.20	**1.05**	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.175	C	0.32	0.32	**0.70**	**1.65**	-2.20	**1.65**	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.175	G	-0.29	0.32	**0.70**	-2.20	-2.20	-2.20	**1.65**	-2.20	**1.19**	**0.97**	**1.19**	0.32
0.325	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	**1.05**	0.13	-0.23	-0.23	-0.23
1.000	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.19

Position-weight matrix

log (frequency / background

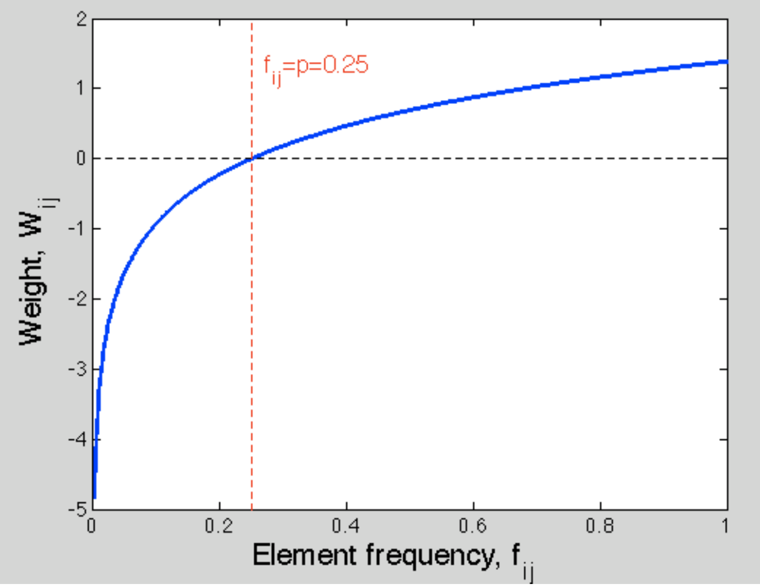
Source: J. van Helden

Information content

Shannon uncertainty applied to PSSM: example

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.79	0.13	-0.23	-2.20	**1.05**	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.175	C	0.32	0.32	**0.70**	**1.65**	-2.20	**1.65**	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.175	G	-0.29	0.32	**0.70**	-2.20	-2.20	-2.20	**1.65**	-2.20	**1.19**	**0.97**	**1.19**	0.32
0.325	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	**1.05**	0.13	-0.23	-0.23	-0.23
1.000	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.19

Position-weight matrix



$$W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_i}\right) \qquad f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k} \qquad \sum_{i=1}^{A} f'_{i,j} = 1$$

The weight W_{ij} 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)

Source: J. van Helden

Information content

Shannon uncertainty applied to PSSM: example

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.79	0.13	-0.23	-2.20	**1.05**	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.175	C	0.32	0.32	**0.70**	**1.65**	-2.20	**1.65**	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.175	G	-0.29	0.32	**0.70**	-2.20	-2.20	-2.20	**1.65**	-2.20	**1.19**	**0.97**	**1.19**	0.32
0.325	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	**1.05**	0.13	-0.23	-0.23	-0.23
1.000	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.19

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

Element (i,j) Information Content

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

Column (j) Information Content

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

PSSM Information Content

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.12	0.05	-0.06	-0.08	**0.97**	-0.08	-0.08	-0.08	-0.08	-0.08	-0.12	-0.06
0.175	C	0.08	0.08	**0.25**	**1.50**	-0.04	**1.50**	-0.04	-0.04	-0.04	0.08	-0.04	0.08
0.175	G	-0.04	0.08	**0.25**	-0.04	-0.04	-0.04	**1.50**	-0.04	**0.68**	**0.45**	**0.68**	0.08
0.325	T	0.19	-0.12	-0.08	-0.08	-0.08	-0.08	-0.08	**0.97**	0.05	-0.06	-0.06	-0.06
1.000	Sum	0.11	0.09	**0.36**	**1.29**	**0.80**	**1.29**	**1.29**	**0.80**	**0.61**	**0.39**	**0.47**	0.04

Inform. content

SUM = 7.54

Source: J. van Helden

Information content

Shannon uncertainty applied to PSSM: example

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.12	0.05	-0.06	-0.08	**0.97**	-0.08	-0.08	-0.08	-0.08	-0.08	-0.12	-0.06
0.175	C	0.08	0.08	**0.25**	**1.50**	-0.04	**1.50**	-0.04	-0.04	-0.04	0.08	-0.04	0.08
0.175	G	-0.04	0.08	**0.25**	-0.04	-0.04	-0.04	**1.50**	-0.04	**0.68**	**0.45**	**0.68**	0.08
0.325	T	0.19	-0.12	-0.08	-0.08	-0.08	-0.08	-0.08	**0.97**	0.05	-0.06	-0.06	-0.06
1.000	Sum	0.11	0.09	**0.36**	**1.29**	**0.80**	**1.29**	**1.29**	**0.80**	**0.61**	**0.39**	**0.47**	0.04

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

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

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

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

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*

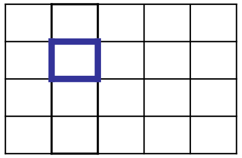
Reference: Hertz (1999).

Bioinformatics 15:563-577$_{13}$

Source: J. van Helden

Information content

Properties of the information content

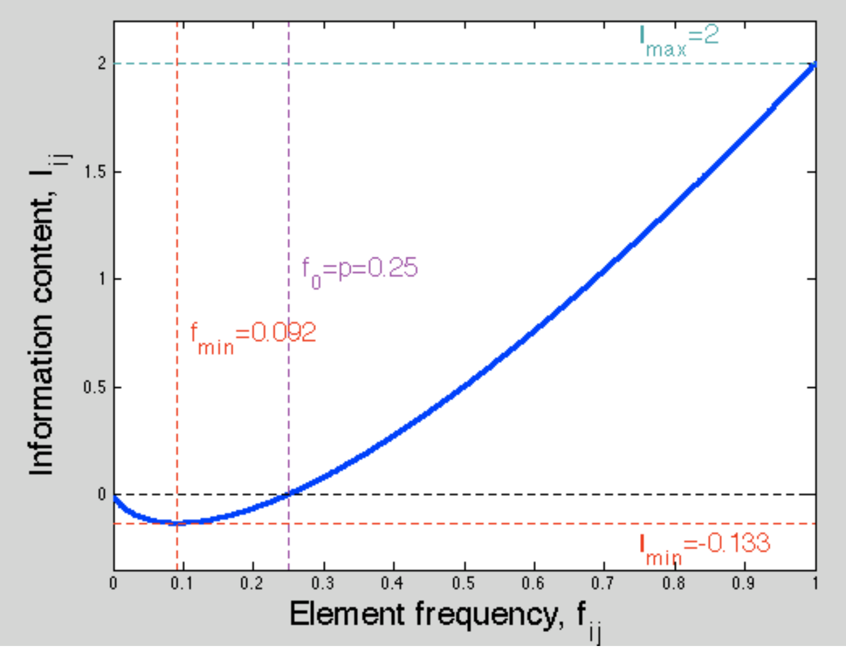


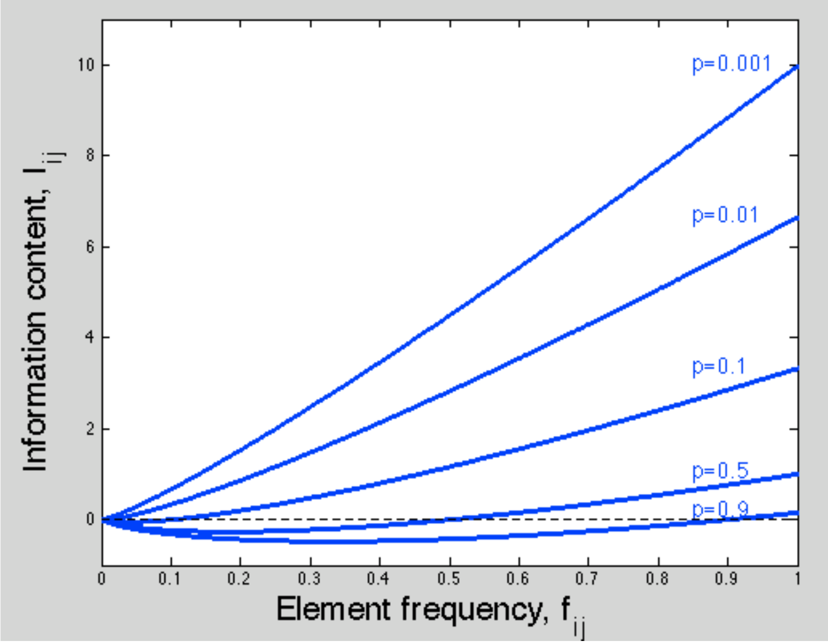
$$I_{i,j} = f'_{i,j} \ln\left(\frac{f'_{i,j}}{p_i}\right)$$

Information content of a given cell (i,j)

For a given cell (i,j) of the matrix:

- $I_{ij} > 0$ when $f'_{ij} > p_i$ (i.e. when residue i is more frequent at position j than expected by chance)
- $I_{ij} < 0$ when $f'_{ij} < p_i$ (i.e. when residue i is less frequent at position j than expected by chance)
- I_{ij} tends towards 0 when f'_{ij} -> 0 (because limit x->0 x*ln(x) = 0)



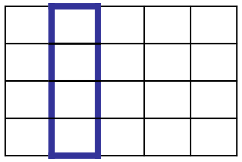


Remark: The upper bound of I_{ij} increases when p_i decreases (I_{ij} -> Inf when p_i -> 0). The information content, as defined by Gerald Hertz, has thus no upper bound.

Source: J. van Helden

Information content

Properties of the information content



For a given column *j* of the matrix:

- The information content of the column (I_j) is the sum of information of its cells.
- I_j is always positive ($I_j > 0$)
- I_j is 0 when the frequency of all residues equal their prior probability ($I_j = 0$ if $f_{ij} = p_i$)
- I_j is maximal when the residue with the lowest prior probability has a frequency of 1 (all other residues have a frequency of 0); the pseudo-weight is 0

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

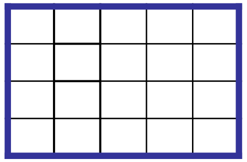
Information content of a given column (j)

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

Source: J. van Helden

Information content

Properties of the information content



- 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) optimizes 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.

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

Information content of the PSSM

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

Source: J. van Helden

Information content

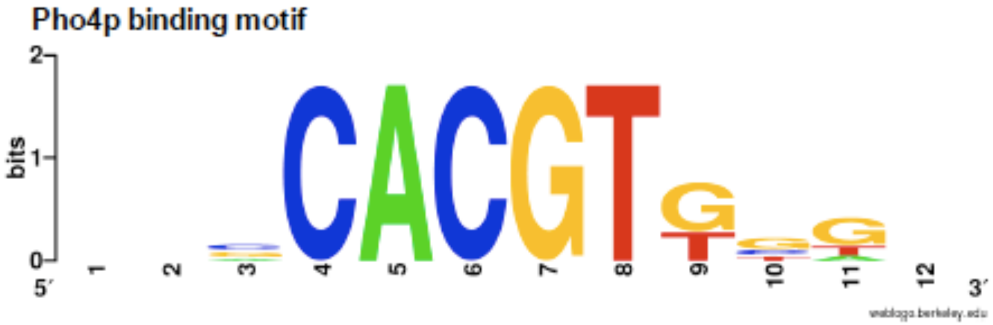
Sequences logo based on information content

Schneider (1990) proposes a graphical representation (logo) 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 R_{seq}:

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

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

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


Pho4p binding motif

$H_s(j)$ = uncertainty of column j
$R_{seq}(j)$ = “information content” of column j (NB: this definition differs from Hertz’ information content)
e(n) = correction for small samples (pseudo-weight)

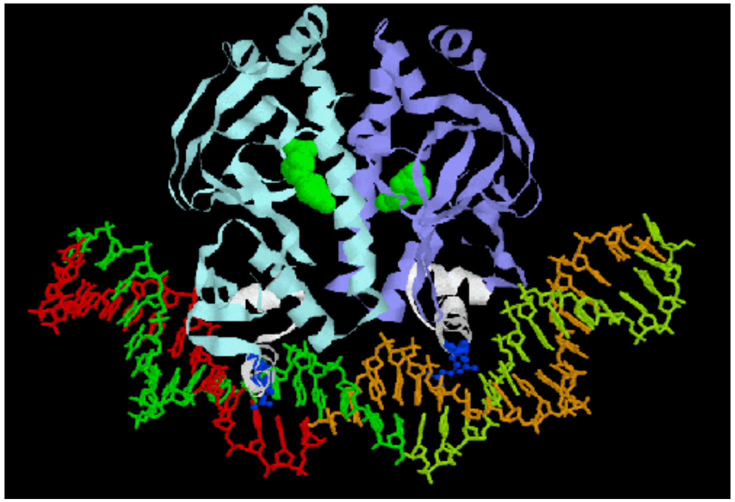
Remarks
This information content does not include any correction for the prior residue probabilities (p_i)
This information content is expressed in bits.

Boundaries
$min(R_{seq})=0$ <=> equiprobable residues
$max(R_{seq})=2$ <=> 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/

Information content

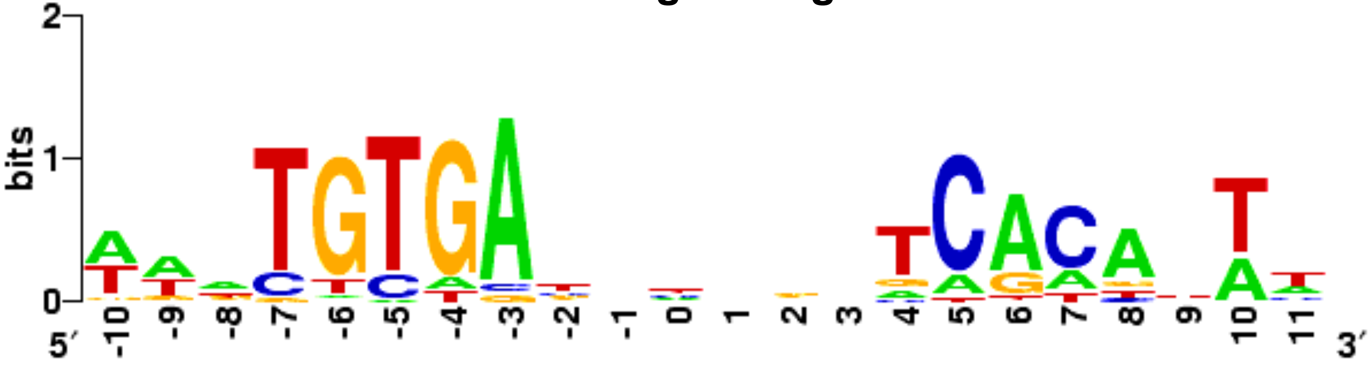
Example: CRP binding site.



- cAMP Receptor Protein (CRP) is a dimer of two identical subunits.
- cAMP-CRP activates expression of many genes in *E. coli*, by binding to specific sites on the DNA where it directly interacts with RNA Polymerase.
- Nucleotide sequencing and analysis of CRP binding sites established a consensus binding sequence consisting of an imperfect 5 bp palindrome:

  TGTGA---N6---TCACA

CRP binding site logo



Information content

Example: CRP binding site.

A

Col E1 site 2	T	T	T	T	G	T	G	G	C	A	T	C	G	G	G	C	G	A	G	A	A	T
Col E1 site 1	T	T	T	T	T	T	G	A	T	C	G	T	T	T	T	C	A	C	A	A	A	A
ara site 2	A	A	A	A	G	T	G	T	C	T	A	T	A	A	T	C	A	C	G	G	C	A
ara site 1	T	T	A	T	T	T	G	C	A	C	G	G	C	G	T	C	A	C	A	C	T	T
bgl R mut	A	A	C	T	G	T	G	A	G	C	A	T	G	G	T	C	A	T	A	T	T	T
crp	G	T	A	T	G	C	A	A	A	G	G	A	C	G	T	C	A	C	A	T	T	A
cya	A	G	G	T	G	T	T	A	A	A	T	T	G	A	T	C	A	C	G	T	T	T
deo P2 site 2	T	T	A	T	T	T	G	A	A	C	C	A	G	A	T	C	G	C	A	T	T	A
deo P2 site 1	A	A	T	T	G	T	G	A	T	G	T	G	T	A	T	C	G	A	A	G	T	G
gal	T	A	A	T	T	T	A	T	T	C	C	A	T	G	T	C	A	C	A	C	T	T
ilv B	A	A	A	C	G	T	G	A	T	C	A	A	C	C	C	C	T	C	A	A	T	T
lac site 1	T	A	A	T	G	T	G	A	G	T	T	A	G	C	T	C	A	C	T	C	A	T
lac site 2	A	A	T	T	G	T	G	A	G	C	G	G	A	T	A	A	C	A	A	T	T	T
mal E	T	T	C	T	G	T	A	A	C	A	G	A	G	A	T	C	A	C	A	C	A	A
mal K	T	T	C	T	G	T	G	A	A	C	T	A	A	A	C	C	G	A	G	G	T	C
mal T	A	A	T	T	G	T	G	A	C	A	C	A	G	T	G	C	A	A	A	T	T	C
omp A	A	T	G	C	C	T	G	A	C	G	G	A	G	T	T	C	A	C	A	C	T	T
tna A	G	A	T	T	G	T	G	A	T	T	C	G	A	T	T	C	A	C	A	T	T	T
uxu AB	T	G	T	T	G	T	G	A	T	G	T	G	C	T	T	A	A	C	C	C	A	A
pBR P4	C	G	G	T	G	T	G	A	A	A	T	A	A	C	G	C	A	C	A	G	A	T
cat site 2	A	C	C	T	G	T	G	A	C	G	T	G	A	G	A	T	C	A	C	T	T	C
cat site 1	A	A	A	T	G	A	G	A	C	G	T	T	G	A	T	C	G	G	C	A	C	G
tdc	A	T	T	T	G	T	G	A	G	T	G	G	T	C	G	C	A	C	A	T	A	T

23 sites identified as binding sites to CRP in *E. coli*

B

A	0.48	0.48	0.39	0.04	0.00	0.04	0.13	0.83	0.26	0.22	0.13	0.48	0.22	0.31	0.09	0.09	0.65	0.26	0.65	0.17	0.30	0.26
C	0.04	0.00	0.13	0.09	0.04	0.04	0.00	0.04	0.30	0.35	0.17	0.04	0.17	0.17	0.09	0.87	0.09	0.65	0.13	0.26	0.09	0.13
G	0.09	0.13	0.13	0.00	0.78	0.00	0.83	0.04	0.17	0.26	0.35	0.26	0.44	0.26	0.17	0.00	0.22	0.04	0.17	0.17	0.00	0.09
T	0.39	0.39	0.35	0.87	0.17	0.91	0.04	0.09	0.26	0.17	0.35	0.22	0.17	0.26	0.65	0.04	0.04	0.04	0.04	0.39	0.61	0.52

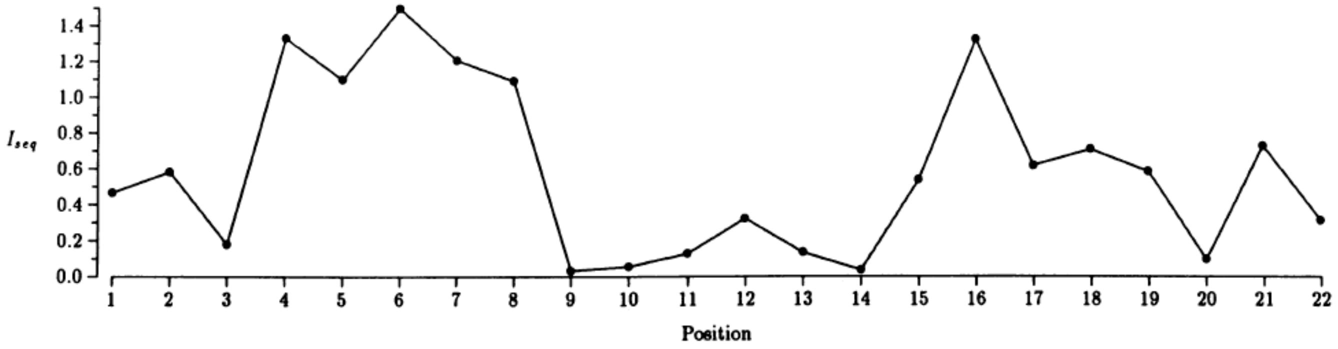
PSSM (frequencies)

C

A	0.94	0.94	0.64	-2.64	-2.75	-2.63	-0.94	1.73	0.07	-0.18	-0.94	0.94	-0.18	0.31	-1.47	-1.47	1.38	0.07	1.39	-0.54	0.26	0.06
C	-2.64	-2.75	-0.94	-1.47	-2.63	-2.63	-2.75	-2.64	0.28	0.49	-0.56	-2.64	-0.56	-0.56	-1.47	1.80	-1.47	1.39	-0.93	0.07	-1.47	-0.94
G	-1.47	-0.94	-0.94	-2.75	1.66	-2.75	1.73	-2.64	-0.54	0.06	0.49	0.06	0.82	0.06	-0.56	-2.75	-0.18	-2.63	-0.54	-0.54	-2.75	-1.47
T	0.64	0.64	0.49	1.80	-0.54	1.88	-2.64	-1.47	0.07	-0.56	0.49	-0.18	-0.56	0.06	1.38	-2.64	-2.64	-2.63	-2.63	0.66	1.29	1.06

PSSM (scores)

D



Information content per position

Stormo & Hartzell, PNAS (1989)

Information content

References

- Shannon CE (1948) A Mathematical Theory of Communication, *Bell System Technical Journal* 27:379-423, 623-656.

- Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol* 188:415-431.

- Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097-6100.

- Schneider TD (1997) Information content of individual genetic sequences. *J Theor Biol* 189:427-441.

- Hertz GZ & Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15:563-577.

- Stormo GD, Hartzell GW (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci USA* 86:1183-7.

See also:

Schneider TD (2012) *Information Theory Primer*, Lecture notes (available at http://schneider.ncifcrf.gov/)