Matching PWMs to the Genome

part of “Genomik der Genregulation”

Sonja Prohaska

Computational EvoDevo
University Leipzig

Leipzig, WS 2011/12

Position Weigth Matrix (PWM)

= Position-Specific Weight Matrix (PSWM) = Position-Specific Scoring Matrix (PSSM), is a common representation of a sequence motif or pattern in biological sequences.

d Position weight matrix (PWM)

A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

$$M = \sum_{i=1}^{w} m_{s_i,i}$$

- Given a PWM for a motif of length w and a substring $s = s_1 s_2 ... s_w$ of length w, calculate the match score of the PWM to the substring.
- i ... position in substring s
- s_i ... character at position i in the substring s
- $m_{a,i}$ ... score in row a, column i of the PWM

Data Collection for PWMs

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

observed binding sites for MEF2

- take experimentally verified binding sites
- align the sequences
- the variability among sites strongly affects the PWM and motif finding

Derivation of Position Frequency Matrices (PFMs)

c Position frequency matrix (PFM)

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

- the first column in the alignment (last slide) consists of no As, three Cs, two Gs and three Ts
- this results in in the corresponding first column in the PFM
- an entry in the PFM is $f_{a,i}$ for the character a at position i in the matrix

Derivation of Position Probability Matrices (PPMs)

PPMs give the relative frequencies that can be viewed as probabilities to observe a character a at position i.

$$p_{a,i} = \frac{f_{a,i}}{\sum_{b \in A,C,G,T} f_{b,i}} \tag{1}$$

What about the probability to see a character a at position i if a was not observed among the collected sites?

Pseudocounts: handling zeros in the PFMs

$$p_{a,i} = \frac{f_{a,i} + B/4}{\sum_{b \in A,C,G,T} f_{b,i} + B} \tag{2}$$

- $f_{a,i}$ ... is the observed absolute frequency of character a at position i in the motif
- B ... is the pseudocount (e.g.: 1)

This makes sure that a character a at position i has a very small probability (but greater than zero) to occure even though it was not among the collected sites.

Derivation of Position Weigth Matrices (PWMs)

d Position weight matrix (PWM)

A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	0.66	−1.93	−1.93	1.07	0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

$$m_{a,i} = \log_2 \frac{p_{a,i}}{p_a} \tag{3}$$

- the weigths depend on the nucleotide distribution in the sequence that is searched p_a, the background probability of a
- individual positions in the motif are assumed to be independent

Scoring matches

- given a genomic sequence G of length L, take all $1 <= k <= L - w + 1$ subsequence of length w, $S_k = s_1 s_2 ... s_w$
- score them according to the following equation
- the higher the match score M_k for a subsequence k the better the match
- the match score is proportional to the binding energy contribution

$$M = \sum_{i=1}^{w} m_{s_i,i} \tag{4}$$

e Site scoring

0.45	–0.66	0.79	1.68	0.45	–0.66	0.79	0.45	–0.66	0.79	0.00	1.68	–0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

$\Sigma = 5.23$, 78% of maximum

The meaning of the match score

What you should know from school

$$\log(a \times b) = \log a + \log b$$

$$\log \prod_{i=1}^{n} a_i = \sum_{i=1}^{n} \log a_i$$

$$M = \sum_{i=1}^{w} m_{s_i,i} = \sum_{i=1}^{w} \log_2 \frac{p_{a,i}}{p_a} = \log_2 \prod_{i=1}^{w} \frac{p_{a,i}}{p_a} \tag{5}$$

$$2^M = \prod_{i=1}^{w} \frac{p_{a,i}}{p_a} = \frac{\prod_{i=1}^{w} p_{a,i}}{\prod_{i=1}^{w} p_a} \tag{6}$$

The meaning of the match score

- $\prod_{i=1}^{w} p_{a,i}$ ... probability that the subsequence matches the PPM
- $\prod_{i=1}^{w} p_a$ ... probability that the subsequence occures randomly
- therefore, 2^M is the fraction by which the subsequence is more likely derived from the PPM then the background by chance

What do negative match scores mean?

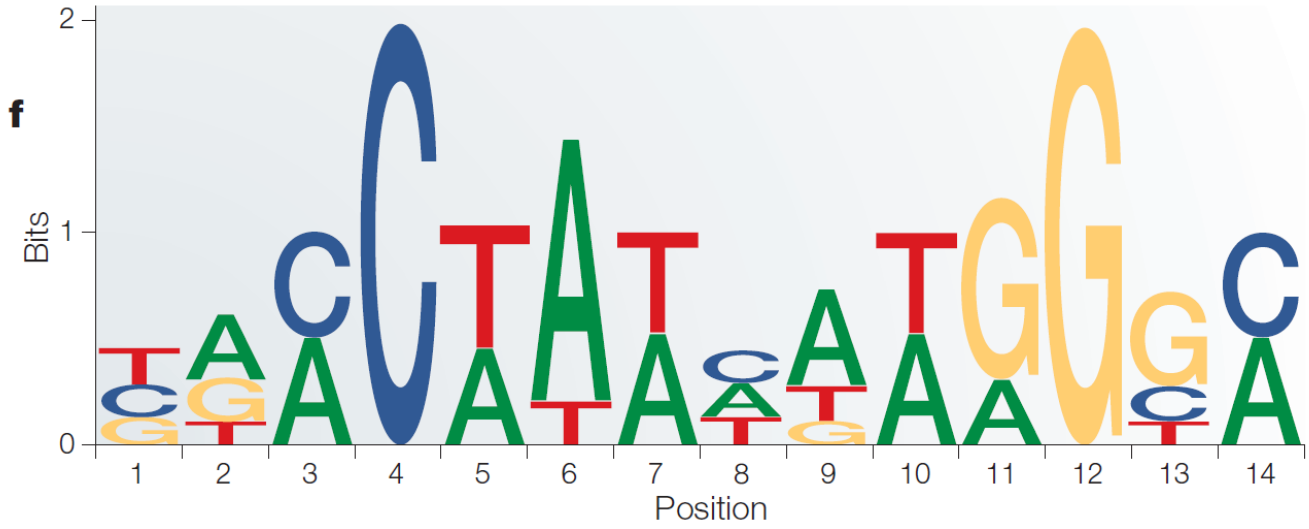
What is the maximal score for a given PWM?

Sequence Logo

The Information Content IC at position i is given by

$$IC_i = 2 + \sum_{binA,C,G,T} p_{b,i} \log_2 p_{b,i} \tag{7}$$

The IC_i multiplied by the relative frequency of a nucleotide a in position i specified by the PFM gives the heigth in bits in the sequence logo.



Literature

- [Stormo, 2000] Gary D. Stormo. *DNA binding sites: representation and discovery*. Bioinformatics. 2000. 16:1, p16-23
- [Kel, 2003] A.E. Kel, E. Gößling, I. Reuter, E. Cheremushkin, O.V. Kel-Margoulis and E. Wingender. *MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences*. Nucl. Acid Res. 2003. 31:13, p3576-3579.
- [Wasserman, 2004] Wyeth W. Wasserman and Albin Sandelin. *Applied Bioinformatics for the identification of regulatory elements*. Nat Rev Genet. 2004. 5:4, p276-287.