Interaktionen von RNAs und Proteinen

Sonja Prohaska

Computational EvoDevo
Universitaet Leipzig

May 7, 2018

DNA Binding Motifs and Binding Site

binding motif – what it is (an abstract model)
binding site (BS) – where it is (position in a sequence)

- **motif finding**

  given: a motif representation

  goal: find all occurences of the motif in sequence S
- **motif detection** or **binding site prediction**

  given: a motif model (PPM/PWM)

  goal: detection and score of occurences in sequence S
- **motif discovery**

  given: a set of sequences $\mathfrak{S}$

  goal: find a motif that is common to all sequences in (a subset of) $\mathfrak{S}$

How to represent a motif?

Given an alignment of N motifs of length k

- **consensus string**

  most frequent nucleotide per position 5'-TGAGTCA-3'

- **consensus pattern**

  all possible nucleotides per position
  5'-T(G|T)(A|T)(G|C)TCA-3'

  $K = \{G, T\}$, $W = \{A, T\}$, $S = \{G, C\} \rightarrow$ 5'-TKWSTCA-3'

- **position frequency matrix – PFM**

  Alphabeth $= \{A, C, G, T\}$, size of Alphabet $= a$, matrix: $a \times k$

- **motif logo**

How to derive a motif model?

aligned motif sequences

- # of sequences $N = 5$
- motif size $l = 5$

Position Frequency Matrix

- absolute frequency $f_a(i, j)$ of nucleotide i at position j
- check: column sum is N

Position Probability Matrix

- relative frequency $f_r(i, j)$ of nucleotide i and position j
- check: column sum is 1

GGCCT
GGGTT
TGCCA
CGCCA
AGCCT

PFM

	1	**2**	**3**	**4**	**5**
A	1	0	0	0	...
C	1	0	4	4	...
G	2	5	1	0	...
T	...	...	...	...	...
	5	5	5	5	5

PPM

	1	**2**	**3**	**4**	**5**
A	0.2	0	0	0	...
C	0.2	0	0.8	0.8	...
G	0.4	1.0	0.2	0	...
T	...	...	...	...	...
	1	1	1	1	1

Pseudocounts (get ride of zeros)

distribute a small "pseudocount" equally

- add the values in the pseudocount matrix to the PFM
- re-calculate the PPM $\rightarrow$ now PPM*
  (Caution! Devide by 5+1!)

PFM

	1	2	3	4	5
A	1	0	0	0	...
C	1	0	4	4	...
G	2	5	1	0	...
T	...	...	...	...	...
	5	5	5	5	5

+

pseudocount (matrix)

	1	2	3	4	5
A	1/4	1/4	1/4	1/4	1/4
C	1/4	1/4	1/4	1/4	1/4
G	1/4	1/4	1/4	1/4	1/4
T	1/4	1/4	1/4	1/4	1/4
	1	1	1	1	1

=

PPM*

	1	2	3	4	5
A	0.208	0.042	0.042	0.042	...
C	0.208	0.042	0.708	0.708	...
G	0.375	0.875	0.208	0.042	...
T	...	...	...	...	...
	1	1	1	1	1

Now that there are no zeros in the matrix we can take the probabilities $f_r(i,j)$ and

- multiply them: $\prod_j f_r(i,j)$ or
- add their logarithms: $\sum_i \log(f_r(i,j))$.

Position Weight Matrix (PWM) for Scoring

- **forground**: the position probability matrix
- **background**: (for simplicity) equal distribution
- i.e. $f(A) = f(C) = f(G) = f(T) = 0.25$

forground

PPM*

	1	2	3	4	5
A	0.208	0.042	0.042	0.042	...
C	0.208	0.042	0.708	0.708	...
G	0.375	0.875	0.208	0.042	...
T	...	...	...	...	...
	1	1	1	1	1

background

A	0.25
C	0.25
G	0.25
T	0.25

scoring matrix

PWM*

	1	2	3	4	5
A	–0.080	–0.775	–0.775	–0.775	...
C	–0.080	–0.775	0.452	0.452	...
G	0.176	0.544	–0.080	–0.775	...
T	...	...	...	...	...

$$M_{i,j}^{PWM} = log(f_r(i,j)/f(i)) = log(M_{i,j}^{PPM}/f(i))$$

"Now, let's search for the binding sites in the genome!"

Best and Worst Motif Scores

The best score S^{max}

$$S^{max} = \sum_j max_i(M_{i,j})$$

PWM

	1	2	3	4	5
A	–0.080	–0.775	–0.775	–0.775	...
C	–0.080	–0.775	0.452	0.452	...
G	0.176	0.544	–0.080	–0.775	...
T	...	...	...	...	...

The worst score S^{min}

$$S^{min} = \sum_j min_i(M_{i,j})$$

PWM

	1	2	3	4	5
A	–0.080	–0.775	–0.775	–0.775	...
C	–0.080	–0.775	0.452	0.452	...
G	0.176	0.544	–0.080	–0.775	...
T	...	...	...	...	...

In the example:

$S^{max} = 0.176 + 0.544 + 0.452 + 0.452 + ... = 1.96$

$S^{min} = +(-0.08) + (-0.775) + (-0.775) + (-0.775) + ... = -3,18$

Motif Detection: Site Scoring with PWMs

Given a PWM, and a genomic sequence search for the most probable binding sites.

	1	2	3	4	5	6	7	8	9	10
Genome	A	A	G	G	C	C	A	T	T	G

Motif Detection: Site Scoring with PWMs

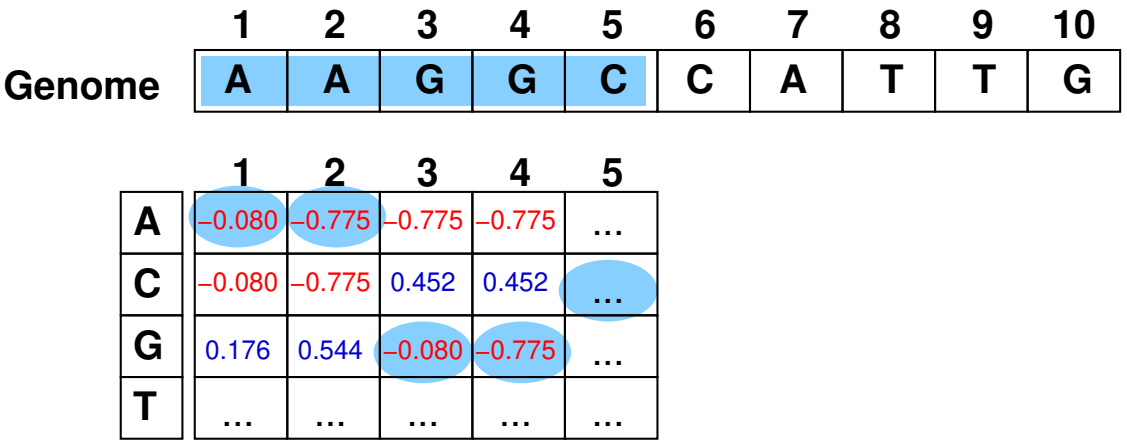
Given a PWM, and a genomic sequence search for the most probable binding sites.

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

	1	2	3	4	5
A	–0.080	–0.775	–0.775	–0.775	...
C	–0.080	–0.775	0.452	0.452	...
G	0.176	0.544	–0.080	–0.775	...
T	...	...	...	...	...

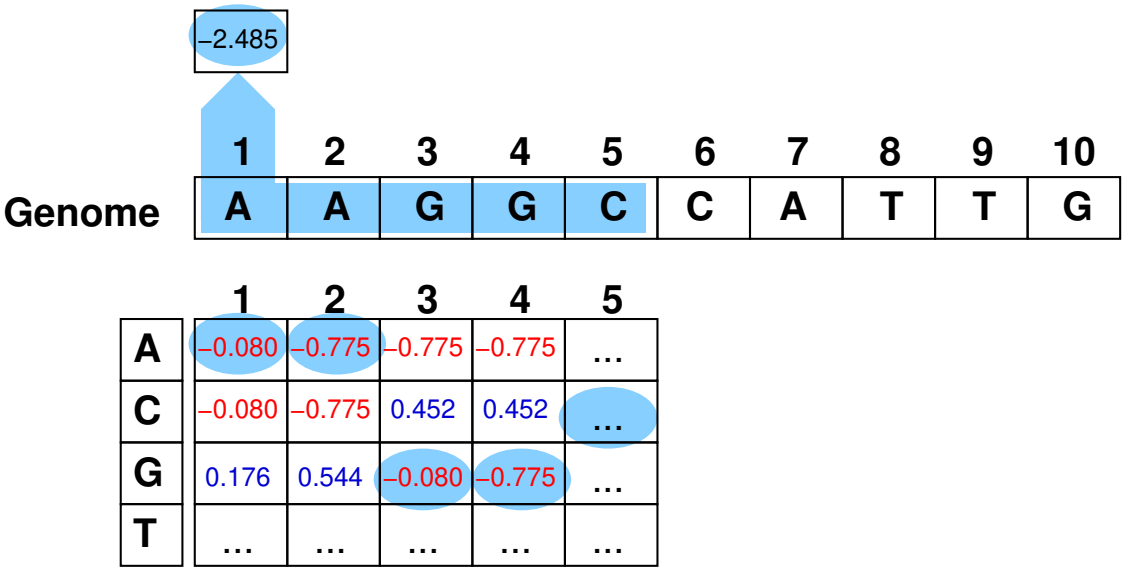
Motif Detection: Site Scoring with PWMs

Given a PWM, and a genomic sequence search for the most probable binding sites.



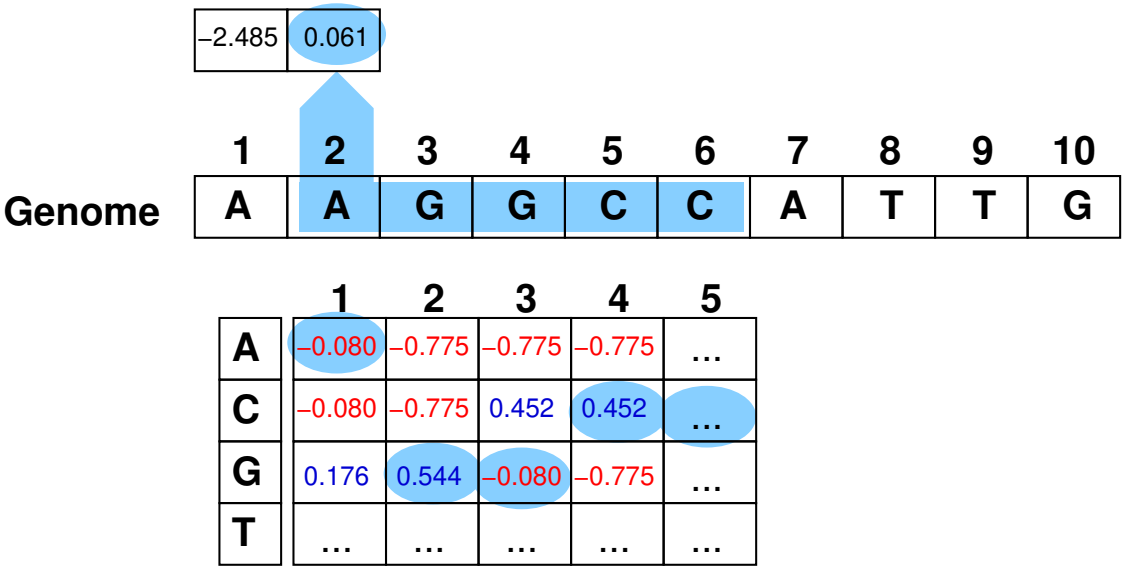
Motif Detection: Site Scoring with PWMs

Given a PWM, and a genomic sequence search for the most probable binding sites.



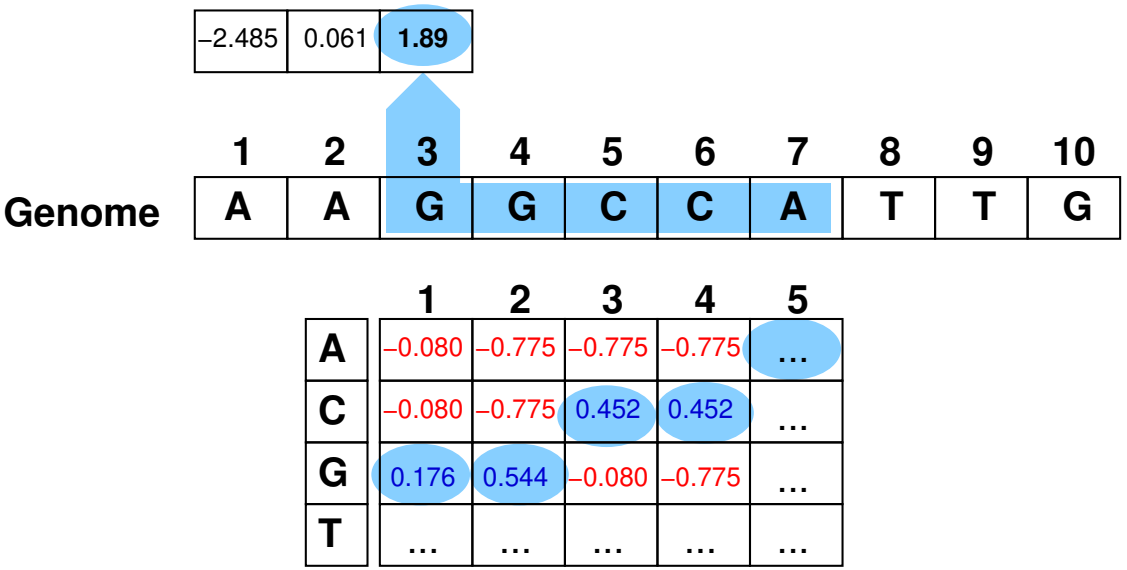
Motif Detection: Site Scoring with PWMs

Given a PWM, and a genomic sequence search for the most probable binding sites.



Motif Detection: Site Scoring with PWMs

Given a PWM, and a genomic sequence search for the most probable binding sites.



Normalizing the Scores

- a large difference ΔS between S_{max} and S_{min} indicates high specificity of the motif model
- nevertheless, normalization may be applied that maps ΔS onto the interval $[0, 1]$

$$S_{norm} = \frac{S - S_{min}}{S_{max} - S_{min}}$$

- mostly used to compare scores of **different motifs**

Motif Discovery

Let's assume...
the protein is known but the binding motif is NOT...

What to do?

- Ask a biologist to do a ChIP-seq experiment.
- Discover the binding motif from a set of sequences computationally.

Can we derive a common binding motif from ChIP-seq data?

discover – entdecken; genau das ist hier gefragt