Statistically Significant Patterns in DNA Sequences

part of “Genomik der Genregulation”

Sonja Prohaska

Computational EvoDevo
University Leipzig

Leipzig, SS 2011

Searching for Sequence Motifs

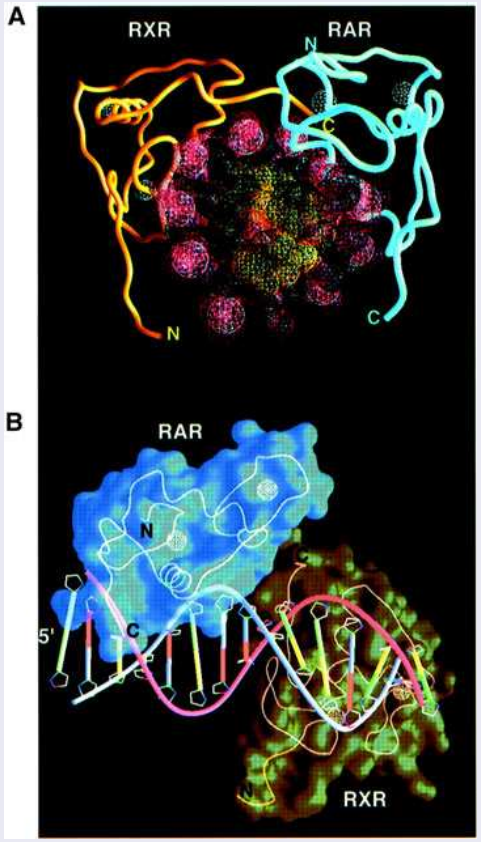
Why?

Proteins that bind to DNA make contact with the bases via the major groove of the double helix. The protein is said to bind to the DNA in a sequence-specific manner. A single binding domain usually contacts 4-8 basepairs.

RAR ... retinoic acid receptor
RXR ... retinoid X receptor

A) view along the DNA helix. RAR and RXR contact the DNA from opposite sides (top left and top right).

B) view from the side. The contact sites from RAR and RXR are seperated by half a turn.



Representation of DNA Sequence Motifs

single sequence motif

A single DNA sequence motif is given in form of a string over the same alphabet as the genomic DNA sequence.
Example: binding site of a nuclear receptor (e.g. RAR or RXR)

AGGTCA

multiple sequence motifs

Homologs of nuclear receptors, e.g. RAR-α, RAR-β, RAR-γ, RXR-α, RXR-β, RXR-γ, have slightly different binding motifs.

AGGTCA
GGTTCA
TGTTCA
GGGTCA

Such motifs can be found in sequences using string search.

Representation of DNA Sequence Motifs

consensus sequence motif

subsumes a set of similar/aligned sequences in one string.
Example: RARE means “retinoic acid response element”, it binds the heterodimer built from RAR (retinoic acid receptor) and RXR (retinoid X receptor). The motif exists of two “half-sites”: a binding site for RAR, and a binding site for RXR in a specific distance.

DR5-type RARE $\texttt{RGKTCA[N]}_5\texttt{RGKTCA}$

This represnetation uses wildcards (R for puRines (A or G), K for G or T, N for any nucleotide).

Consensus motifs can be used to find motif occurences in genomic sequences using regular expressions.

Representation of DNA Sequence Motifs

position frequency matrix (PFM)

Rows in the matrix stand for A, C, G, T (in this order) and columns for the positions in the motif. The entry at $M_{x,y}$ indicates the frequency of letter x at position y in the motif.

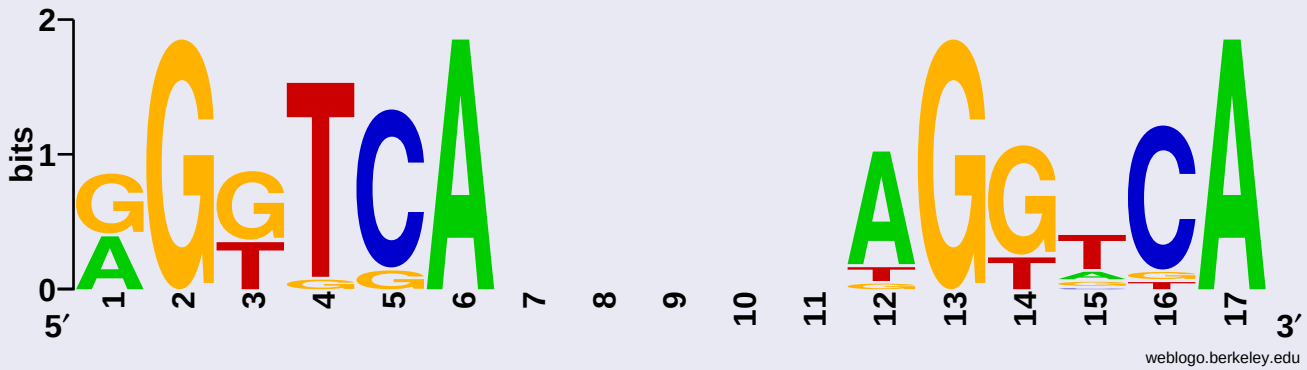
	1	2	3	4	5	6	7 - 11	12	13	14	15	16	17
A	8	0	0	0	0	17	$[N]_5$	14	0	0	3	0	17
C	0	0	0	0	15	0	$[N]_5$	0	0	0	1	15	0
G	9	17	10	1	2	0	$[N]_5$	1	17	13	2	1	0
T	0	0	7	16	0	0	$[N]_5$	2	0	4	11	1	0

Frequencies can be absolut frequencies or relative frequencies.

Representation of DNA Sequence Motifs

sequence logo

A logo displays the frequencies of bases at each position, as the relative heights of letters, along with the degree of sequence conservation as the total height of a stack of letters, measured in bits of information. The vertical scale is in bits, with a maximum of 2 bits possible at each position for DNA (with 4 bases, $\log_2 4 = 2$ bits per base).



Motif Finding

Task: Given a sequence motif, find all occurences of the motif in a genome.

Let N be the length of the genome and k the length of a motif, there are $N - k + 1$ substrings of length k. Count the number of substrings that match the sequence motif.

What does the number of observed motifs tell us?

Observed versus Expected

possible assumptions:

- the genomic sequence is short, a random occurence of the motif is not expected
- the genomic sequence is long, random occurences of motifs are expected, however, functional sites occure clustered resulting in local overrepresenation of sites
- a sequence or set of sequences is expected to have similar/higher/lower/more regular a.o. distribution of motifs than another sequence or set of sequences

How many motif occurences are expected?

Observed versus Expected – Simple Case

Given:

- long, random genome with $f_x = 0.25$ for $x \in A, C, G, T$

Expected frequency of motifs AGGTCA and ATATAT with length $k = 3$:

$$E_k = {f_x}^k (N - k + 1) \tag{1}$$

The expected number of a specific motif of length k is the probability of the motif (i.e. ${f_x}^k$) multiplied with the number of substrings of length k (i.e. $(N - k + 1)$).

Each word of length k has the same expectation value!

Observed versus Expected – based on single nucleotide frequency

Given:

- long, random genome with nucleotide frequencies f_A, f_C, f_G and f_T
- $\sum_{x \in A,C,G,T} f_x = 1$

Expected frequency of a motifs AGGTCA and ATATAT

$$E_{AGGTCA} = f_A^2 \times f_C^1 \times f_G^2 \times f_T^1 \times (N - k + 1) \quad (2)$$

$$E_{ATATAT} = f_A^3 \times f_T^3 \times (N - k + 1) \quad (3)$$

Exercise: How to handle occurences of motifs on both strands?

Observed versus Expected – based on di-nucleotide frequency

Given:

- long, random genome with a base composition bias on di-nucleotide level
- nucleotide frequencies f_A, f_C, f_G and f_T
- $\sum_{x \in A,C,G,T} f_x = 1$
- di-nucleotide frequencies f_{AA}, f_{AC}, f_{AG} ...
- $\sum_{x \in A,C,G,T} \sum_{y \in A,C,G,T} f_{xy} = 1$

Expected frequency of motif AGGTCA

$$E_{AGGTCA} = \frac{f_{Ag} \times f_{GG} \times f_{GT} \times f_{TC} \times f_{CA}}{{f_G}^2 \times f_T \times f_C} \times (N - k + 1) \qquad (4)$$

Exercise: How can this be generalized to a Markov chain of higher order?

Exercise

Given the following sequence motif and genomic sequence, is the motif more or less frequent than expected?

genomic sequence of length $N = 50$:
ACGGTGACCTTGAGCGCTAGCACTCGAAGGGTCCAGGTCACGGTAGCACT
motif of length $k = 4$: GCGC

nucleotide counts:

$c_A = 11$
$c_C = 14$
$c_G = 16$
$c_T = 9$

di-nucleotide frequencies:

$f_{CC} = 0.0408$
$f_{CG} = 0.0816$
$f_{GC} = 0.0816$
$f_{GG} = 0.1020$

a) assume $f_A = f_C = f_G = f_T = 0.25$
b) use a background model based on observed single-nucleotide frequencies
c) use a background model based on observed single- and di-nucleotide frequencies