Statistically Significant Patterns in DNA Sequences
part of “Genomik der Genregulation”

Sonja Prohaska
Computational EvoDevo
University Leipzig

Leipzig, SS 2011
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 separated by half a turn.
### 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)

\[
\text{AGGTCA}
\]

### multiple sequence motifs

Homologs of nuclear receptors, e.g. RAR-\(\alpha\), RAR-\(\beta\), RAR-\(\gamma\), RXR-\(\alpha\), RXR-\(\beta\), RXR-\(\gamma\), have slightly different binding motifs.

\[
\begin{align*}
\text{AGGTCA} \\
\text{GGTTCA} \\
\text{TGTTCGA} \\
\text{GTTTCGA} \\
\text{GGGTCGA}
\end{align*}
\]

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**  
\[
\text{RGKTCA} [N] \text{5RGKTCA}
\]

This representation 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 occurrences in genomic sequences using **regular expressions**.
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.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7 - 11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>$[N]_5$</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>$[N]_5$</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>9</td>
<td>17</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>$[N]_5$</td>
<td>1</td>
<td>17</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>$[N]_5$</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Frequencies can be **absolut frequencies** or **relative frequencies**.
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).
Shannon Entropy

Claude E. Shannon introduced the term *entropy* in his 1948 paper *A Mathematical Theory of Communication*. He addressed a question to (English) language, namely:

How much information is given by any single character in a message?

With other words:
- How well can we predict the next character?
- What is the expected surprisal for a coin flip?
Example: Flipping a coin

The outcome of flipping a coin can be either heads or tails!

Question: What is the maximum suprisal for a coin flip?
Shannon Entropy and Information Content

**Shannon Entropy**

**Definition:**

\[
H(X) = - \sum_{i} P(x_i) \log P(x_i)
\]

The entropy is maximized if all \( P(x_i) \) are equal!

**Question:** What is the maximum surprisal for a coin flip?
Shannon Entropy and Information Content

**Shannon Entropy**

**Definition:**

$$H(X) = - \sum_i P(x_i) \log P(x_i)$$

The entropy is maximized if all $P(x_i)$ are equal!

**Question:** What is the maximum surprisal for a coin flip?

**Answer depends on which log is used for the calculation!**
Shannon Entropy and Information Content

Shannon Entropy

Definition:

\[ H(X) = - \sum_{i} P(x_i) \log P(x_i) \]

The entropy is maximized if all \( P(x_i) \) are equal!

Question: What is the maximum surprisal for a coin flip?
Answer depends on which log is used for the calculation!

Interpretation of Shannon Entropy

The entropy tells us our uncertainty for the next observation. If we are told the outcome, this uncertainty is reduced to zero (we know exactly what will be observed).
Therefore \textit{entropy is related with information}!
Information Content

The information content of a message can be seen as the difference between what we knew before and what we know after reading the message:

\[ I(X) = H_{before}(X) - H_{after}(X) \]

Often, \( H_{before} \) reflects the fact that we have no information. (all \( P(x_i) \) are equal)

What is \( H_{before} \) for a position in a DNA sequence?

What is \( H_{before} \) for a position in a protein sequence?
Example: Information Content

Given an alignment of DNA binding sites (our message). Without prior knowledge and using log\(_2\) we have

\[ H_{\text{before}} = 2 \text{ bits.} \]

At a specific position of the alignment we observed only A’s and G’s with:

\[ P(A) = 0.7 \text{ and } P(G) = 0.3. \]

\[ H_{\text{after}} = -0.7 \log_2(0.7) - 0.3 \log_2(0.3) = 0.88 \text{ bits} \]

\[ I(X) = H_{\text{before}} - H_{\text{after}} = 2 - 0.88 = 1.12 \text{ bits} \]
Shannon Entropy and Information Content

**Information in sequence logo’s**

In sequence logo’s the value of the information content

\[ I(X) = H_{\text{before}} - H_{\text{after}} \]

is used to scale each stack of letter frequencies. Each letter \( x_i \) has a height of \( P(x_i) \cdot I(X) \).

Thus, the sequence logo tells us which positions are informative and which are not.

Note: positions with low information content might still be important for the binding of the protein to the DNA!

A more general approach is given by the relative entropy where probabilities \( P(x_i) \) in \( H_{\text{before}} \) are not necessarily equal.
Motif Finding

Task: Given a sequence motif, find all occurrences 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?
possible assumptions:

- the genomic sequence is short, a random occurrence of the motif is not expected
- the genomic sequence is long, random occurrences of motifs are expected, however, functional sites occur clustered resulting in local overrepresentation 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 occurrences are expected?
Given:

- long, random genome with \( f_x = 0.25 \) for \( x \in \{A, C, G, T\} \)

Expected frequency of motifs \( \text{AGGTCA} \) and \( \text{ATATAT} \) with length \( k = 3 \):

\[
E_k = f_x^k (N - k + 1) \quad (1)
\]

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 occurrences 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** $\text{AGGTCA}$

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

**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 $N = 50$: 
ACGTTGACCTTGAGCGCTAGCACTCGAAGGGTCCAGGTCACGGTAGCACT

Motif of length $k = 4$: GCGC

nucleotide counts: 
\[
\begin{align*}
c_A &= 11 \\
c_C &= 14 \\
c_G &= 16 \\
c_T &= 9 \\
\end{align*}
\]

di-nucleotide frequencies: 
\[
\begin{align*}
f_{CC} &= 0.0408 \\
f_{CG} &= 0.0816 \\
f_{GC} &= 0.0816 \\
f_{GG} &= 0.1020 \\
\end{align*}
\]

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