Interaktionen von RNAs und Proteinen

Sonja Prohaska

Computational EvoDevo
Universitaet Leipzig

June 9, 2015
How to represent a motif?

Given an alignment of $N$ motifs of length $l$

- **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**
  Alphabet = $\{A, C, G, T\}$, size of Alphabet $a$, matrix $a \times l$

- **motif logo**

![Motif Logo](image)
How to derive a motif model?

**aligned motif sequences**
- # of sequences ... $N$
- example: $N = 5$, $l = 5$

**Position Frequency Matrix**
- absolute frequency $f_a(i,j)$ of nucleotide $i$ and position $j$
- example: column sum is $N$

**Position Probability Matrix**
- relative frequency $f_r(i,j)$ of nucleotide $i$ and position $j$
- example: column sum is $N$

---

<table>
<thead>
<tr>
<th></th>
<th>PFM</th>
<th>PPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>C</td>
<td>0.2</td>
<td>0</td>
<td>0.8</td>
<td>0.8</td>
<td>...</td>
</tr>
<tr>
<td>G</td>
<td>0.4</td>
<td>1.0</td>
<td>0.2</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>T</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Interaktionen von RNAs und Proteinen
pseudocounts (get ride of zeros)

distribute a small “pseudocount” equally

- add the values in the pseudocount matrix to the PFM
- re-calculate the PPM → now PPM*

\[
\text{PFM pseudocount (matrix)} =
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
1 & 4 & 4 & 0 & 0 \\
2 & 5 & 5 & 5 & 5 \\
5 & 5 & 5 & 5 & 5
\end{bmatrix}
\]

\[
\text{ppm} =
\begin{bmatrix}
1/4 & 1/4 & 1/4 & 1/4 & 1/4 \\
1/4 & 1/4 & 1/4 & 1/4 & 1/4 \\
1/4 & 1/4 & 1/4 & 1/4 & 1/4 \\
1/4 & 1/4 & 1/4 & 1/4 & 1/4
\end{bmatrix}
\]

Now that there are no more zeros in the matrix we can multiply the probabilities \(f_r(i, j)\), or add their logarithms.

\[
a \times b \times c \rightarrow \log(a) + \log(b) + \log(c)
\]
Position Weight Matrix (PWM) for Scoring

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

\[
\begin{align*}
M_{i,j}^{PWM} &= \log(f_r(i,j)/f(i)) = \log(M_{i,j}^{PPM}/f(i))
\end{align*}
\]

“Now, let’s search for the BSs in the genome!”
Motif Detection: Site Scoring with PWMs

Now we can make good predictions!
Motif Discovery

Known protein, binding motif \textbf{NOT} known...
... biologist brings ChIP-seq data.

What to do?

discover – entdecken; genau das ist hier gefragt
ChIP-seq

Given a protein, where does it bind to the genome?

ChIP-seq (in vivo method)

- ChIPseq (Chromatin Immuno-Precipitation followed by sequencing)
- proteins are cross-liked to DNA in vivo
- chromatin is isolated
- sonicate to obtain chromatin fragments
- immunoprecipitate with protein-specific antibodies
- purify immunocomplexes (remove unbound chromatin fragments)
- reverse cross-linking
- purify DNA from chromatin fragments
- sequence DNA fragments
ChIP-seq method

Cross-link cells with formaldehyde. Isolate genomic DNA and sonicate to shear chromatin.

Add an antibody specific to the protein of interest.

Perform immunoprecipitation to isolate DNA bound by the factor of interest. Reverse cross-links and purify isolated DNA.
ChIP-seq output

... after peak calling

- high number of regions/peaks
- size 200-1000nt (or so)
- supposed to bear BSs for protein of interest
- binding sites and binding motifs 4-12nt

Can we derive the motif from those sequences?
Multiple Expectation Maximization for Motif Elicitation

**Problem:** Given a set of (unaligned) sequence of variable length
- characterize – derive a motif model
- identify – matching positions in the sequences

**Assumptions:**
- motif is contiguous
  - no insertions or deletions
  - all instances of a motif have the same length
- motif positions are independent

find the most probable non-overlapping shared motif(s).
Types of Possible Motif Models

Motif occurrences within a set of sequences

**OOPS**
- exactly one per sequence
- motif in all sequences

**ZOOPS**
- zero or one per sequence
- motif in subset of sequences

**Two-Component Mixture – TCM**
- any number of motifs
- motif in subset of sequences
Starting motif model

- choose an initial motif model wisely
- one way to go
  - list all $k$-mers while $k$ is the length of the desired motif
  - use the most frequent $k$-mer to set up a model (with pseudocounts)
- $\rightarrow$ a PPM* model
Step 1: Expectation step

- score every position in the sequence and
- every sequence with the current model
- record the scores/probabilities
Step 2: Maximization step

- use the probabilities to improve/update the pattern
- don’t forget pseudocounts and normalization
Discover multiple motifs

- Iterate E an M step until motif model does not change anymore.
- Use the model to find the highest scoring motif site for each sequence.

What if you want to find a second or third motif?

- x-out motif sites obtained by the previous run
- run again
What else?

Consider

- what’s the best initial guess?
- What’s the best motif size?
- consider ZOOPS, TCM
- add a prior probability (additional knowledge) to the probability distribution

MEME will always return something! from almost nothing!
What could be the causes for the discrepancies between prediction (black bars) and experimental data (red peaks)?

- boring case: prediction or experiment to stringent/sensitive
- direct interaction of a second binding domain and corresponding binding site on the DNA
- indirect interaction via an RNA or protein mediator/cofactor
- DNA modifications that stimulate/inhibit the binding
- binding site is present but blocked or not accessible
- binding is transient
- repetitive regions are not mapped