Interaktionen und Modifikation von RNAs und Proteinen

Prof. Sonja Prohaska

Computational EvoDevo Group
Institute of Computer Science
Universität Leipzig

SS 2019
DNA is “never” naked in a cell

general binder
DNA is usually in association with proteins. In all domains of life there are small, basic chromatin associated proteins (ChAP) attached to the DNA forming higher order structures. Their function is to prevent DNA from agglutination, ensuring stability and flexibility, aid structure formation (“packaging”) and nuclear organization, and engage in gene regulation.

specific binder
Sequence-specific transcription factors associate with specific binding sites. Their functions is commonly gene regulation.
Proteins that are general binders of DNA

**eukaryots**

- the DNA-protein complex: chromatin
- chromatin associating proteins (ChAP): **histones**

**bacteria**

- the DNA-protein complex: nucleoid
- nucleoid associating proteins (NAP): e.g. **HU** – introduces negative supercoiling
  **H-NS** – regulates gene expression

HU, core (red), arms (blue)  histone-like nucleoid-structuring protein (H-NS)
General vs. specific binder

**general binder**
- small basic proteins
- contact negatively charged DNA backbone
- e.g. HU, Alba, histones (double stranded DNA)
- e.g. SSB (single stranded DNA)
- bind everywhere, often bend the DNA

**specific binder**
- large protein with DNA-binding domain(s) (DBD)
- e.g. helix-turn-helix, helix-loop-helix, zinc finger, leucine zipper
- contact the major groove of DNA
- bind to specific sites
- in sequence-dependent manner
DNA-binding proteins – a molecular view

Fig: Electrostatic potential surface of the DNA. General binders are basic and contact the acidic DNA backbone. Specific binders make contact with the nucleobases in the major and minor groove.
Proteins can distinguish all four nucleobase pairs in the major groove. In the minor groove only AT and GC basepairs can be distinguished.
DNA-binding proteins – a molecular view

Example: the estrogen receptor dimer

Homodimeric proteins consist of two identical units that are in contact via the dimer interface. Here each dimer makes contact to 6 base pairs in the major groove in a sequence-specific manner. The motif is **palindromic**.
How to study DNA-protein binding?

DNA-protein Interaction

- EMSA
- SELEX
- motif finding
- motif discovery (meme)
- protein
- ChIP-seq
- motif
- genome localization

Interact SS18
From the DNA-binding protein to its binding motif

Given a protein, which DNA sequence does it bind preferentially?

in vitro selection – SELEX

▶ SELEX (systematic evolution of ligands by exponential enrichment) also known as in-vitro evolution
▶ a very large oligonucleotide library
   ▶ randomly generated sequences of fixed length
▶ is exposed to the target protein
▶ unbound oligos flow through (are removed)
▶ bound sequences are eluted and amplified by PCR
   ▶ another SELEX round with increased stringency is done or
▶ the pool of sequences is finally sequenced
▶ an alignment reveals the **motif** and its variation
From the DNA-binding protein to its binding motif
How to represent a motif?

Given an alignment of $n$ motifs of length $l$
here the result of the SELEX experiment: $n = 43$, $l = 7$

- **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 $a \times l$

- **sequence logo** based on the information content (IC)
How to calculate the information content?

The information content (y-axis) is measured in bits.
The maximal information depends on the size $k$ of the Alphabet $\mathcal{A}$, here $\mathcal{A} = \{A, C, G, T\}$, $a \in \mathcal{A}$, and $k = 4$.
The **Shannon entropy** (uncertainty) $H_i$ at position $i$ of the motif is

$$H_i = - \sum_{a} f_i(a) \log_2 f_i(a) \tag{1}$$

where $f_i(a)$ is the relative frequency of nucleotide $a$ at position $i$.
The **information content** $R_i$ at position $i$ of the motif is

$$R_i = \log_2(k) - H_i \tag{2}$$

(i.e. the maximal information content minus the uncertainty)

The height $h_i(a)$ of a letter $a$ in column $i$ is

$$h_i = f_i(a) \times R_i \tag{3}$$

Note: $\log_2 k = \log_2 4 = 2$
... with small-sample size correction

Small-sample correction $e_n$ is given by

$$e_n = \frac{1}{\ln 2} \times \frac{s - 1}{2n}$$  \hspace{1cm} (4)

where $s$ is the number of sequences in the alignment.

The information content $R_i$ at position $i$ of the motif is then

$$R_i = \log_2(k) - (H_i + e_n)$$  \hspace{1cm} (5)

Explain, what does $e_n$ do?
Search for binding sites in the genomic sequence

Let’s assume you want to apply a word search method to find the motif in a genomic sequence.

- Assume that motif $m$ is a string, e.g. 5′-GGCCT-3′
- Of motif length $l = 5$bp
- The genomic sequence $M$, e.g. the human HoxA cluster,
- Has a length of $L = 163001$bp
- Alphabet $\mathcal{A} = \{A, C, G, T\}$

You find 312 motif sites (on the plus strand).
Search for binding sites in the genomic sequence

How often do you expect to find a motif \( m \) in a sequence \( M \)?

- assume that motif \( m \) is a string, e.g. 5′-GGCCT-3′
- of motif length \( l = 5 \)bp
- the genomic sequence \( M \), e.g. the human HoxA cluster,
- has a length of \( L = 163001 \)bp
- Alphabet \( \mathcal{A} = \{A, C, G, T\} \)

the background model

1. **uniform** nucleotide distribution: \( f(x) = 0.25 \)
2. use **mono**-nucleotide distribution:
   \[ f(A) = 0.2428, f(C) = 0.2555, f(G) = 0.2552, f(T) = 0.2466 \]
3. use **di**-nucleotide distribution:
   \[ f(AG) = 0.0604, f(GG) = 0.0812, f(GC) = 0.0677 \]
   \[ f(CC) = 0.0815, f(CT) = 0.0751 \]
\[ E^{UNI}(m) = f(X)^5 \times (L - (l - 1)) \] (6)

\[ E^{MONO}(m) = (f(G)^2 \times f(C)^2) \times f(T)) \times (L - (l - 1)) \] (7)

\[ E^{DI}(m) = \frac{f(GG) \times f(GC) \times f(CC) \times f(CT)}{f(G) \times f(C)^2} \times (L - (l - 1)) \] (8)

Results: \[ E^{UNI}(m) = 159; \quad E^{MONO}(m) = 171; \quad E^{DI}(m) = 329; \]
Comparison of expectation and observation

With word search you find 312 motif sites (on the plus strand).

**Observed sites:** number of matches of motif $m$ in $M$ is $O = 312$

**Expected sites:** $E^{UNI}(m) = 159; E^{MONO}(m) = 171; E^{DI}(m) = 329;$

Use the **chi-squared test** $(\chi^2)$ to determine whether there is a significant difference between the expected frequency and the observed frequency.

$$\chi^2 = \frac{(O - E)^2}{E}$$

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>0.995</th>
<th>0.975</th>
<th>0.9</th>
<th>0.5</th>
<th>0.1</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.016</td>
<td>0.455</td>
<td>2.706</td>
<td>3.841</td>
<td>5.024</td>
<td>6.635</td>
<td>7.879</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.010</td>
<td>0.051</td>
<td>0.211</td>
<td>1.386</td>
<td>4.605</td>
<td>5.991</td>
<td>7.378</td>
<td>9.210</td>
<td>10.597</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.072</td>
<td>0.216</td>
<td>0.584</td>
<td>2.366</td>
<td>6.251</td>
<td>7.815</td>
<td>9.348</td>
<td>11.345</td>
<td>12.838</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.207</td>
<td>0.484</td>
<td>1.064</td>
<td>3.357</td>
<td>7.779</td>
<td>9.488</td>
<td>11.143</td>
<td>13.277</td>
<td>14.860</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.412</td>
<td>0.831</td>
<td>1.610</td>
<td>4.351</td>
<td>9.236</td>
<td>11.070</td>
<td>12.832</td>
<td>15.086</td>
<td>16.750</td>
<td></td>
</tr>
</tbody>
</table>

Results: $\chi^2(UNI) = 147.226; \chi^2(MONO) = 116.263; \chi^2(DI) = 0.878$;
Comparison of expectation and observation

Results: $\chi^2(UNI) = 147.226; \chi^2(MONO) = 116.263; \chi^2(DI) = 0.878$

**Significance levels**
for one degree of freedom (df)
expected and observed value, $E$ and $O$, are significantly different at 5% level ($p = 0.05$) if $\chi^2 \geq 3.841$
at 1% level ($p = 0.01$) if $\chi^2 \geq 6.635$

**Conclusion**
159, i.e. $E^{UNI}(m)$, is significantly different from 312, i.e. $O$.
172, i.e. $E^{MONO}(m)$, is significantly different from 312, i.e. $O$.
329, i.e. $E^{DI}(m)$, is NOT significantly different from 312, i.e. $O$.

Calculation of $E(m)$ using the di-nucleotide distribution results in an expected number of sites similar to the observed number of sites.
Three things to learn from the above

First...

► the expectation value strongly depends on the background model
  i.e. the choice of nucleotide distribution
► the closer the background model to reality
  the better the estimation
► the di-nucleotide distribution is a good background model

Second...

► there is nothing special about the motif and the region
  why? the number of observed sites is very similar to the number of
  expected sites
► the genomic sequence is neither enriched
  nor depleted in binding sites for the DNA-binding protein with the
  motif 5′-GGCCT-3′
Three things to learn from the above...

Third...

- Did we do something wrong maybe?
- How about the minus strand?

With word search you find 482 motif sites (on the minus strand).

**Exercise**: improve the calculations!

- searching GGCCT on the minus strand is equivalent to searching AGGCC on the plus strand
- use nucleotide frequencies as given (they are given for the plus strand)
- Result: $E_{DI}(m) = 427$ for the minus strand.
- Result: with $\chi^2 >= 7,084$ significant at 1% level
- Result: the motif is enriched in the minus strand
DNA is more than a (plus) strand!

- DNA is double stranded
- the two strands are referred to as **plus and minus strand**
- the **convention** is:
  - the strand which runs from 5’ to 3’ from left to right
  - is the top strand (when both strands are displayed) and
  - it is also the plus strand (the one displayed if only one strand is displayed)
- the sequence of one strand determines the sequence of the other
- the relation between the two strands is the “reverse complement”
Exercises

▶ Calculate how any time you expect to find motif 5′-GGCCT-3′ on the minus strand.
▶ Compare the expected and observed value of binding sites on the minus strand.
▶ Should you add-up values for plus and minus strand? What if the motif is palindromic, e.g. 'GGTACC'?
▶ How different are expectation values for motif 5′-CGTCG-3′ in comparison to 5′-GGCCT-3′?
▶ How many different motifs of length 5 can be derived from the sequence below?
▶ How many binding sites can be found with motif 'GGTCA'?

Interact SS18