Interaktionen von Nukleinsäuren und Proteinen

Sonja Prohaska

Computational EvoDevo
Universitaet Leipzig

June 9, 2015
DNA is “never” naked in a cell

DNA is usually in association with proteins. In all domains of life there are small, basic **chromosomal architectural proteins** attached to the DNA forming “chromatin” (bacteria – HU, archaea – HU, Alba & histones, eukaryots – histones). Their function is to prevent DNA from agglutination, ensuring stability and flexibility, aid structure formation (“packaging”) and engage in gene regulation. Sequence-specific **transcription factors** associate with specific binding sites. There functions is commonly gene regulation.
general vs. specific binder

**general binder**

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

**specific binder**

- large protein with DNA-binding domain(s) (DBD)
- contact major groove of DNA
- DBDs: e.g. HTH (helix-turn-helix), zinc finger, leucine zipper, helix-loop-helix
- bind to specific sites
- sequence-dependent
Example
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.
How to study DNA-protein binding?
How to study DNA-protein binding?

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 are removed (by affinity chromatography)
- bound sequences are amplified by PCR
- subsequent rounds of selection with increase stringency
- sequencing
- returns consensus motif
How to study DNA-protein binding?

Random DNA oligonucleotides

TF protein binding

Gel mobility shift assay

Immunoprecipitation

Purification

PCR

SELEX round

---

FIG. 3. GCN4-binding sites.
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 \( a \times l \)

- **motif logo**
What to Do With the Motif?

How often would you expect to see a motif $m$ in a sequence $M$?

- assume the motif is a string, e.g. $5'$-GGCCT-3'
- of motif length $l = 5$
- the sequence $M$, e.g. the human HoxA cluster,
- has a length of $L = 163001$
- Alphabet $\{A, C, G, T\}$

1. assume **uniform** nucleotide distribution: $f(x) = 0.25$

2. use **mono**-nucleotide distribution:
   - $fA = 0.2428, fC = 0.2555, f(G) = 0.2552, f(T) = 0.2466$

3. use **di**-nucleotide distribution:
   - $fGG = 0.0812, f(GC) = 0.0677, f(CC) = 0.0815, f(CT) = 0.0751$

Results: $E^{UNI}(m) = 159; E^{MONO}(m) = 171; E^{DI}(m) = 329;$
Which expectation comes closer to the observation?

**Observed sites:** count the motif $m$ in $M!$ $O = 312$

**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.

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

look-up values of the $\chi^2$ distribution for one degree of freedom

- significance at 5% level when $\chi^2 \geq 3.841$
- significance at 1% level when $\chi^2 \geq 6.635$

Only $E^{DI}(m) = 329$ is a good prediction for the number of sites.
How to study DNA-protein binding?

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

in vivo – ChIP-seq

- ChIPseq (Chromatin Immuno-Precipitation followed by sequencing)
- proteins are cross-linked 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
Study DNA-protein binding with ChIP-seq
Do Predictions Fit Observations?

A set of sites $S$ known to contain:

Observation:
- # bound sites $P$
- # unbound sites $N$

Prediction:
- # predicted bound $P'$
- # predicted unbound $N'$

$P + N = P' + N'$

$S$ $- - -$

$N$

$S'$ $- - +$

$P'$ $+ +$

Observation (thorough)

Prediction

True

Negative

TN

True

Negative

TN

False

Positive

FP

False

Positive

TP

False

Negative

FN
Sensitivity and Specificity of the Prediction

- **Sensitivity**: mit welcher Rate wird die Präsenz von BS richtig vorhergesagt?
  
  \[
  \text{True positive rate} \quad \frac{TP}{P} = \frac{TP}{TP+FN}
  \]
  
  Ziel: bloss keine Bindungsstelle überschauen!
  
  
  \[\rightarrow \text{high sensitivity}\]

- **Specificity**: mit welcher Rate wird die Abwesenheit von BS richtig vorhergesagt?
  
  \[
  \text{True negative rate} \quad \frac{TN}{N} = \frac{TN}{TN+FP}
  \]
  
  Ziel: bloss keine falschen Bindungsstellen!
  
  \[\rightarrow \text{high specificity}\]
ROC (curve) to Compare Predictions

Ziel (Idealfall)

high sensitivity = high true positive rate
high specificity = low false positive rate

ROC curve