Characterizing DNA binding sites – high throughput approaches
Biol4230 Tues, April 24, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

- Reviewing sites: affinity and specificity
  - representation
  - binding and specificity
  - (equilibria and competition)
- Comprehensive site identification
  - binding, consensus, and conservation
- What does complete understanding look like?
  - have DNA sequence, identify binding affinity/occupancy
  - have protein sequence of binding domain, identify DNA target

To learn more:

DNA-Protein interaction: binding vs specificity

Dynamic questions:
• Is DNA site $S$ bound to a transcription factor $TF$?
• Is the site bound frequently enough to affect transcription?
• Where is most of the $TF$ binding?
  – on specific DNA sites
  – on non-specific sites
  – on all sites with $K_d < 10^{-x}$
  – there are typically $10^6$ more non-specific than specific sites (but are all accessible)
• what happens when the $TF$ changes state?
  – higher concentration
  – more active (tighter binding) because of co-factor/modification

fasta.bioch.virginia.edu/biol4230

DNA-Protein interaction: binding vs specificity

Binding

Specificity

$TF + S \overset{k_{on}}{\rightleftharpoons} TF
d K_{off} = \frac{[TF][S]}{[TF] + [S]} = \frac{RT ln K_d}{RT ln K_d}$

$P(S \text{ bound}) = \frac{[TF
d][S]}{[TF] + [S]} = \frac{TF}{TF + K_d}$

$Spec = \sum_S \frac{K_d(S)}{\sum_S K_d(S)} ln \frac{K_d(S)}{K_d(S)}$

Terminology: Sites vs Motifs

\{\text{Sites}\} \leftrightarrow \text{Motif}

Think restriction sites:

\text{EcoRI: \{GAATTC\} \leftrightarrow GAATC}
\text{HincII \{GTAAAC, GTTGAC, GTCAAC, GTCGAC\} \leftrightarrow GTYRAC}

Transcription factor motifs should be quantitative, give different scores to different sites, reflecting differences in binding affinity.

\text{Also: site is specific location in genome}

Representations/Models of Protein-DNA binding

\begin{itemize}
  \item Transcription factors don’t bind to just one sequence
  \item A “Consensus sequence” is usually the preferred site, but similar sequences also bind well
  \item Not all variants bind equally well; some positions contribute more to the specificity than others
\end{itemize}
Position Weight Matrix Model  
(PWM, also PSSM)

<table>
<thead>
<tr>
<th></th>
<th>log(2)-odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-2.76 1.82 0.06 1.23 0.96 -2.92</td>
</tr>
<tr>
<td>C</td>
<td>-1.46 -3.11 -1.22 -1.00 -0.22 -2.21</td>
</tr>
<tr>
<td>G</td>
<td>-1.76 -5.00 -1.06 -0.67 -1.06 -3.58</td>
</tr>
<tr>
<td>T</td>
<td>1.67 -1.66 1.04 -1.00 -0.49 1.84</td>
</tr>
</tbody>
</table>

DNA-Protein interaction: binding vs specificity

Binding : One sequence

NO COMPETITION

DNA-Protein interaction: binding vs specificity

Dynamic questions:
• Is DNA site $S$ bound to a transcription factor $TF$?
• Is the site bound frequently enough to affect transcription?
• Where is most of the $TF$ binding?
  – on specific DNA sites
  – on non-specific sites
  – on all sites with $K_d < 10^{-x}$
  – there are typically $10^6$ more non-specific than specific sites (but are all accessible)
• What happens when the $TF$ changes state?
  – higher concentration
  – more active (tighter binding) because of co-factor/modification


fasta.bioch.virginia.edu/biol4230
Transcription factor binding – modern approaches

- Have (functional) protein?
  - measure affinities of protein against large sets of random DNA sequences (chromatin?)
  - transform protein into cells, look at reporter genes

- Have antibody to protein?
  - ChIP-Chip/ChIP-seq – measure where the factor is on chromosomal DNA (in specific states)
  - peak width ALWAYS larger than binding sites
  - isolate surrounding DNA sequence, use consensus strategies (meme) also works with other chromatin modifications

- Have co-expressed sets of genes?
  - identify the genes, isolate sequences near promoters (enhancers?)
  - use consensus strategies (meme)

Transcription factor binding – direct measurements

Fig. 4. In vivo function prediction for Pho4p and Cbf1p. (A and B) Genes with regulatory sequences determined to be bound by our in silico method. All genes shown here have a Pocc of above 0.2 and a sensu stricto conservation score of 25% or above. Pie charts show the functional distribution of the gene sets. (C and D) Venn diagrams comparing our predicted gene sets to gene sets determined with use of expression microarrays and ChIP-chip.


Transcription factor binding – direct measurements

A. Original
B. BEEML (Binding Energy Est. ML) with NS energy
C. BEEML w/ NS, di-nucleotide mutation

Transcription factor binding –
direct measurements

Protein Binding Microarray PBM

HT-Sele

fasta.bioch.virginia.edu/biol4230

Transcription factor binding –
direct (reporter) measurements

Bacterial one-hybrid
1. Provide E. coli with essential yeast gene (HIS3) under control of a weak promoter behind a randomized binding site
2. Transfect (add externally) transcription factor (sometimes linked to RNA-Pol subunit)
3. Plate out colonies expressing HIS3
4. Sequence everything that is still present
5. Most abundant (sequence seen most often) randomized region sequences grew the best, thus most binding

fasta.bioch.virginia.edu/biol4230
High-throughput \textit{in vitro} binding site analyses

- Can give good, quantitative models of intrinsic binding specificity
- More data alone isn't sufficient to give better models, also need good analysis methods
- Log-odds method is based on assumptions (independence) that may not be true
- Energetic models can give better descriptions
  - Non-linear relationship between binding affinity and binding probability at high TF concentration

Transcription factor binding –

direct measurements

A. Original
B. BEEM with NS energy
C. BEEM w/ NS, di-nucleotide


Transcription factor binding –

modern approaches

• Have (functional) protein?
  – measure affinities of protein against large sets of random DNA sequences (chromatin?)
  – transform protein into cells, look at reporter genes

• Have antibody to protein?
  – ChIP-Chip/ChIP-seq – measure where the factor is on chromosomal DNA (in specific states)
  – peak width ALWAYS larger than binding sites
  – isolate surrounding DNA sequence, use consensus strategies (meme) also works with other chromatin modifications

• Have co-expressed sets of genes?
  – identify the genes, isolate sequences near promoters (enhancers?)
  – use consensus strategies (meme)
Identifying regulatory sites in chromatin

Regulatory sites in chromatin: GSTM1
Regulatory sites in chromatin: MAP3K3


Regulatory sites in chromatin

A.
Chromatin ImmunoPrecipitation - Sequencing

formaldehyde cross-linking


What do ChIP-Seq signals look like?

ChIP-Seq signals should be "complex" (map across a region, with a peak)


ChIP-Seq counts on known muscle genes have a wide dynamic range

Good ChIP-seq peaks have offset reads on the two strands
What do ChIP-Seq signals look like?

There are typically 100 – 1,000X as many motif/PWM matches as detectable binding sites

But "sites" are much more concentrated at ChIP-seq peaks

Given a set of intervals from peaks, find sites with consensus methods (meme)
ChIP-seq summary:

- Result quality depends on antibody, immunoprecipitation, negative controls – look for reproducible peaks
- Most reads (signal) do not come from peaks
- Many more PWM sites than peaks, but sites more concentrated near peaks
- High peaks ≠ large effect
- Qualitative – enriches regions of interest

Transcription factor binding – modern approaches

- Have (functional) protein?
  - measure affinities of protein against large sets of random DNA sequences (chromatin?)
  - transform protein into cells, look at reporter genes
- Have antibody to protein?
  - ChIP-Chip/ChIP-seq – measure where the factor is on chromosomal DNA (in specific states)
  - binding sites ALWAYS larger than peak width
  - isolate surrounding DNA sequence, use consensus strategies (meme) also works with other chromatin modifications
- Have co-expressed sets of genes?
  - identify the genes, isolate sequences near promoters (enhancers?)
  - use consensus strategies (meme)
Transcription factor binding – position independence

Binding energy model including interactions makes more accurate predictions of in vitro binding specificity than the PWM for Hnf4a. (A) Graphical representation of Hnf4a binding energies estimated from PBM data under the PWM model (Supporting Information, Figure S1). Negatives of binding energy (in units of RT) are plotted on the y-axis. Energies are normalized such that the average energy at each position is 0. This energy logo is equivalent to the “affinity logo” from Foat et al. (2006). (B) Performance of model shown in A on test PBM data. (C) Binding energy model estimated from the same training data but including interaction energies between positions 4 and 5 (Figure S2). (D) Performance of the energy model including interactions on test PBM data.


(From abstract): We find that the specificity of most TFs is well fit with the simple PWM model, but in some cases more complex models are required. We introduce a binding energy model (BEM) that can include energy parameters for nonindependent contributions to binding affinity. We show that in most cases where a PWM is not sufficient, a BEM that includes energy parameters for adjacent dinucleotide contributions models the specificity very well.

How well do methods work?

In vitro defined PWM's accurately predict in vivo binding


fasta.bioch.virginia.edu/biol4230
Information content vs accuracy

**Figure 4** Characteristics of Klf9 motifs produced by the eight PWM-based algorithms evaluated in this study. The algorithms are ranked top to bottom in order of the overall score of their PWM for this TF in our evaluation scheme. Two popular visualization methods of the PWMs produced by each algorithm are depicted. On the left are traditional sequence logos39,40, which display the information content of each nucleotide at each position; the total information content (I.C.) of the PWM is given to the left of this logo. On the right are frequency logos, in which the height of each nucleotide corresponds to its frequency of occurrence at the given position40.


DNA-Protein interaction: what is complete understanding?

1. Understand the DNA binding site
2. Identify the amino-acids that *read* the DNA sequence
3. understand how changes in the protein change the DNA binding site
4. *predict* DNA binding site preferences from protein sequence (engineering)

DNA-Protein interaction (homeobox): what is complete understanding?


fasta.bioch.virginia.edu/biol4230
DNA-Protein interaction (homeobox): what is complete understanding?

Comparison of the Predicted and Determined Recognition Motifs for Six Human Homeodomains: The specificities of the human factors were determined with the B1H system. In each case, the "determined" compares favorably with the "predicted" motif generated with our algorithm.

For the homeobox family, it is possible to predict the DNA binding site from the amino-acid sequence


Characterizing DNA binding sites – high throughput approaches

- Affinity and specificity
  - transcription factors have higher affinity for their specific binding site than non-specific sites
  - but there are $10^6 - 10^7$ more non-specific sites
  - ratios of specific/non-specific binding are $< 10^6$
  - a large fraction of transcription factor binding is non-specific

- High-throughput in vitro methods provide accurate binding constants
  - PWM (independent positions) usually provides accurate model of binding
  - for a fraction of sites, a binding energy term that includes non-independence helps

- ChIP-Seq provides large lists of binding sites
  - but small fraction of motif matches

- For large, highly studied families (homeobox), the amino-acid recognition code is understood