DNA binding factors

Bio5488
Ting Wang
2.17/2019
Protein-DNA Interactions

• Lots of functions:
  • Replication, Repair, Recombination

• Control of expression of the information in the genome

• Specificity of binding sites for regulatory proteins
DNA-binding proteins exhibit a range of specificities

- **Restriction enzymes**
  - Absolute specificity
  - Discriminate methylations

- **Transcription factors**
  - Sequence-specific, but
  - Degenerate
  - Range of affinities may be utilized

- **Nucleosomes**
  - Sequence preferences
  - Translocatable by modeling complexes
Protein-DNA specificity arises from standard molecular contacts:
• Electrostatics, hydrogen bonds, water-mediated contacts, and hydrophobic packing
• In addition, sequence-specific DNA deformations (indirect readout) is often important
Recognition Codes

• In the 1970s, a recognition code for protein-DNA interactions was postulated.

• It has become clear there is no universal code. The interactions are degenerate in both directions.

• However, for a fixed mode of interaction (a single structural family of DNA-binding proteins), there is hope that partial weight matrices may be associated with key amino acid positions in the protein.

• This will require the determination of the binding preferences for many members of a family of TFs.

Seeman, Rosenberg, and Rich, PNAS (1976)
Sequence Motifs

• Motif: subsequence with some specific function

• May be in DNA, RNA, protein

• Function may be context dependent
  • Ribosome binding site must be transcribed
  • RNA, protein motifs may depend on structure

• May be gapped or ungapped

• Use model to search for (predict) new sites
  • Models may be simple sequences (regular expressions) or probabilistic patterns

• Modeling approach depends on data available
  • Quantitative/qualitative
Types of Motifs

Motif: Consensus Sequence Pattern
  – May include degenerate bases and allow for mismatches
  – *Search space is over possible patterns*

Weight Matrix (PWM, Profile, PSSM)
  – Might go to higher order models
  – *Search space is over possible alignments*
Pattern based algorithms

• Motif length $l$, mismatches $m$; $N$ seqs, $L$ long

• $4^l$ patterns, search for most common (or most significant) allowing up to $m$ mismatches
  • $P$-value from background distribution
  • Can allow for $m$ mismatches
  • Can allow degenerate positions: $15^l$ patterns
  • Can just search using existing $l$-mers

• Can use suffix tree for efficient search of patterns allowing mismatches
Consensus Sequence Pattern

- TACGAT
- TATAAT
- TATAAT
- GATACT
- TATGAT
- TATGTT

• Difficult to obtain an optimal consensus for identifying novel sites

• Relative frequency of bases at each positions lost
## Weight Matrix Model

<table>
<thead>
<tr>
<th></th>
<th>A:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TACGAT</td>
<td>-8</td>
<td>10</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td>-8</td>
</tr>
<tr>
<td>TATAAT</td>
<td>-10</td>
<td>-9</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-12</td>
</tr>
<tr>
<td>TATAAT</td>
<td>-7</td>
<td>-9</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
<td>-9</td>
</tr>
<tr>
<td>GATACT</td>
<td>10</td>
<td>-6</td>
<td>9</td>
<td>0</td>
<td>-1</td>
<td>11</td>
</tr>
</tbody>
</table>

- More information than a consensus sequence
- Many ways to determine the weights
- Assumes positional independence
- Requires significant data
Score a site

-24

....A C T A T A A A T G T ...

<table>
<thead>
<tr>
<th></th>
<th>A:</th>
<th>C:</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-8</td>
<td>10</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td>-8</td>
</tr>
<tr>
<td></td>
<td>-10</td>
<td>-9</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-12</td>
</tr>
<tr>
<td></td>
<td>-7</td>
<td>-9</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
<td>-9</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>-6</td>
<td>9</td>
<td>0</td>
<td>-1</td>
<td>11</td>
</tr>
</tbody>
</table>
Score a site

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>A</th>
<th>T</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-8</td>
<td>10</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td>-8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-10</td>
<td>-9</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-7</td>
<td>-9</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
<td>-9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>10</td>
<td>-6</td>
<td>9</td>
<td>0</td>
<td>-1</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\( N(b,i) \)

\( F(b,i) \)

\( S(b,i) = \log\left[\frac{F(b,i)}{P(b)}\right] \)

\( I(i) = \sum F(b,i)S(b,i) \)

G. Stormo
**Information Content**

Matrix of Frequencies

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>G</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>T</td>
<td>0.7</td>
<td>0.1</td>
<td>0.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

\[
I_{seq} = \sum_j \sum_b f(b,j) \log_2 \frac{f(b,j)}{p(b)}
\]

Sum is over columns j (the positions), and rows b (the bases)

*aka Relative Entropy, Kullbach-Liebler Distance*
## Information Content

<table>
<thead>
<tr>
<th>EcoR1</th>
<th>Random</th>
<th>Rap1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAATTC</td>
<td>GCCTAC</td>
<td>TGTATGGGTTG</td>
</tr>
<tr>
<td>GAATTC</td>
<td>ACATTC</td>
<td>TGTTCCGGATT</td>
</tr>
<tr>
<td>GAATTC</td>
<td>TCAATTC</td>
<td>TGCATGGGTTG</td>
</tr>
<tr>
<td>GAATTC</td>
<td>CGACTC</td>
<td>TGTACAGGTTG</td>
</tr>
<tr>
<td>GAATTC</td>
<td>GAATTC</td>
<td>TGTATGGATG</td>
</tr>
<tr>
<td>GAATTC</td>
<td>ATATCG</td>
<td>TGTTCCGGTT</td>
</tr>
<tr>
<td>GAATTC</td>
<td>GAAATG</td>
<td>TGTATGGGTTG</td>
</tr>
</tbody>
</table>

---

![GAATTC](image1)

![TGTATGGGTTG](image2)
Pseudocounts

Entries of zero in the count matrix cause big problems

• The log(0) is undefined (infinitely negative)

• Not enough observations to observe all possibilities
## Pseudocounts

<table>
<thead>
<tr>
<th></th>
<th>Original count matrix</th>
<th>eg.1 Add 1 pseudocount per column</th>
<th>eg.2 Add 1 pseudocount per column according to background nucleotide frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Assume %A=%T=20%</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>10</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

|   | 0.25                  | 17.25                           | 5.25                             | 3.25                            |
| T | 10.25                 | 0.25                            | 5.25                             | 2.25                            |
| G | 4.25                  | 3.25                            | 5.25                             | 5.25                            |
| C | 6.25                  | 0.25                            | 5.25                             | 10.25                           |

|   | 0.2                   | 17.2                             | 5.2                              | 3.2                             |
| T | 10.2                  | 0.2                             | 5.2                              | 2.2                             |
| G | 4.3                   | 3.3                             | 5.3                              | 5.3                             |
| C | 6.3                   | 0.3                             | 5.3                              | 10.3                            |
How to Set the Matrix Elements

• Could be done experimentally. Synthesize different DNA sequences and measure the protein’s affinity to each. Determine matrix parameters that provide the best fit.
  - Laborious experimental work, slow
  - Next-gen sequencing transformed this strategy

• Statistical treatment of known sites. Need a reasonable sample size. Some assumptions about how the sample is obtained.
  - Can be fast, reasonably accurate
Limitations of standard PWMs

• Non-specific binding sets lower limit on affinity
  • Typically much higher than if all bases considered

• Non-additive contributions to specific binding
  • Binding energy of a base may depend on context (neighboring bases)
  • Can use more complex models

• Most models, such as log-odds, assume binding sites come from a Boltzmann distribution
Motif Finding Problem

• A fundamental problem in molecular biology
  • Specific protein and DNA binding
  • Transcription factor binding sites recognition

• Statistical definition:
  • Given some sequences, find over-represented substrings (motif discovery)

• Biological definition:
  • Given some co-regulated promoters, find transcription factor binding model
  • How do we use biology to improve motif finding algorithms?

• Many algorithms/programs developed
  • consensus, gibbs sampling, EM, projection, phylogenetic footprinting, etc.
Motif Finding Algorithms Class I

Single species, multiple genes (planted motif problem)

- random background sequences
- a proper description of a consensus motif → better models
- randomly plant copies of the motif into sequences
- define an objective function, and use a search algorithm to find the copies that give a good score
The Data Set: Sequences containing sites for cAMP receptor protein (CRP)

For this case, there are 18 sequences of length 105 bp and we are looking for a motif of width 20 bp. There are 86 different 20 bp subsequences per example and ~7×10^{34} alignments to check.
An (intractable) solution

(Exhaustive algorithm)
Construct every possible combination of alignments and keep the one with the highest information content.

Given a motif of width $w$, and $k$ sequences of length $l$, there are $L = (l-w+1)$ possible locations in each sequence, and $L^k$ alignments to check.
Alignment (Profile) based methods

- Greedy algorithm (Consensus)
- Expectation Maximization (MEME)
- Gibbs Sampler
- Regression (MatrixReduce)

- Can use phylogenetic conservation
Greedy Algorithm (Consensus)

• Simple version: assume every sequence contains at least one true binding site
• Using each l-mer find best match to generate 2-seq alignments
• Using top K PWMs to search remaining sequences to include a new sequence
• Repeat until all seqs contribute
  • Or objective function is maximized (IC, p-value)
Expectation Maximization (MEME)

- Initial PWM (at random or from average over all potential sites)
- Using current PWM determine probability of all positions being sites
- Re-estimate PWM based on those probabilities
- Continue until convergence – always convergences
- Objective is LLR
**Gibbs Sampling**

- Similar to EM, but some important differences
- At each iteration pick one site on each seq, chosen by its probability, to update PWM
- Not guaranteed to converge, but tends to increase objective (IC) and plateau
- Can escape local optima
  - Other MCMC algorithms
    - Metropolis
    - Simulated annealing
Gibbs’ Sampling Approach to Motif Discovery

Basic Idea:
- Given “sites”, estimate pattern matrix
- Given “matrix”, pick likely sites according to their probability
- Iterate between those steps until “convergence”

Important details:
- Use “pseudocounts” to avoid prob. = 0
- Sample sites from estimated prob. distrib.
ratios. Equivalently, one may maximize $F$, the sum of the logarithms of these ratios. In the notation developed above, $F$ is given by the formula

$$F = \sum_{i=1}^{w} \sum_{j=1}^{20} c_{i,j} \log \frac{q_{i,j}}{p_j}$$

(2)

where the $c_{i,j}$ and $q_{i,j}$ are calculated from the complete alignment (Fig. 1).

should be supplemented with residue-dependent “pseudocounts” $b_j$ to yield pattern probabilities

$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + B}$$

(1)

where $B$ is the sum of the $b_j$. The $p_j$ are

Gibbs Sampling

Initialization:
Random assignment of motif locations $a_1$-$a_k$

Construct initial matrix $S$ from this alignment
Gibbs Sampling

Update:
Score all possible motif locations in the sequence to be updated with current matrix

Select a new motif placement randomly, but biased toward high scoring regions

Create a new scoring matrix from the new alignment
Gibbs Sampling

How does it end? Eventually you nucleate a few correct placements.

The matrix has weak but sufficient scoring power.
Gibbs Sampling

An approximately correct matrix rapidly converges, with the subsequent alignments possessing more information content and making better motif window placements.

But notice two suboptimal results: we have one sequence with a placement but no genuine site, and one sequence with two sites but one placement. This is common enough to merit special treatment.
Summary

- The genome encodes much of its own regulation in protein binding sites
- A full description of the regulatory networks will require identifying these sites
- Compact descriptions of the DNA-binding preferences of TFs is afforded by weight matrices
- The information content of an alignment is a measure of specificity
- Weight matrix information for a TF is not enough to rule out false positives
- Multiple experimental techniques exist for identifying sequences harboring binding sites
- A variety of algorithms can be used to identify motifs in unaligned data
Evolution and comparative analysis

• Evolution:
  • All species evolved from a single life form
  • Variation within a species occurs randomly
  • Natural selection: positive, negative, drift
  • Evolutionary change is gradual: change of allele frequency

• Comparative analysis
  • Find sequences that have diverged less than we expect
  • Find sequences that have diverged more than we expect
Motif finding by using multiple genomes

• Key parameters:
  • Which species to choose?
  • Which sequences?
  • What model to describe evolving sequences?
  • What evolutionary signature to go after?

• Assumptions:
  • functional sequences evolve more slowly than non-functional sequences, as they are subject to selection pressure.
  • TFs and regulatory mechanism remain unchanged
Single gene, multiple species (phylogenetic footprinting)

- orthologous background sequences
- sequences linked by a phylogenetic tree
- identify the “best conserved” motif that is under selective pressure
Motif Finding Algorithms Class III

Multiple genes, multiple species

- combination of phylogenetic data and gene regulation
- use phylogenetic data to reduce search space
- use correlation of motif occurrences among orthologous genes to increase signal strength
Motif: Gal4 – CGGNNNNNNNNNNNCCG
Phylogenetic footprinting

How many species do we need to obtain a footprint?

- Eddy, 2005
Methods

- Wasserman et al. 2000
- CLUSTALW (Cliften et al. 2003)
- MCS (Kellis et al. 2003; Xie et al. 2005)
- PhyloCon (Wang and Stormo 2003)
- PhyloNet (Wang and Stormo 2005)
- Footprinter (Blanchette et al., 2003)
- EMnEM (Moses et al. 2004)
- OrthoMEME (Prakash et al. 2004)
- PhyME (Sinha et al. 2004)
- CompareProspector (Liu et al. 2004)
- PhyloGibbs (Siddharthan et al. 2005)
- Tree Sampler (Li and Wong 2005)
- MultiModule (Zhou and Wong 2005)
- ……
Protein Binding Microarrays

Custom arrays of 60-mer DNA sequences (~44,000 probes)

Contain all possible 10bp sequences

Each probe contains 27 10-mers

8-mers guaranteed to occur 16 times

ChIP-chip/ChIP-seq

Cross-link protein to DNA

Affinity purify protein-DNA complexes:
  - Ab to TF
  - Ab to tag on TF
  - affinity tag on TF

Reverse cross-links

Identify sequence by hybridization to microarray or by high-throughput sequencing

**Bacterial One-hybrid**

**a**

[Diagram showing the process of bacterial one-hybrid]

- **TF** is fused to the α-subunit of RNA polymerase.
- **Randomized region** is screened for autoactivation.
- **Positive marker** and **Negative marker** are used for selection.

**b**

**Construct α-TF expression vector**

**Construct binding site library**

**pB1H1/2**

**pHSU3**

- **Genetic selection:** survival is dependent on DNA-binding.
- **TF of interest** is fused to α-subunit of RNA polymerase.
- Randomized library of binding sites created and screened for autoactivation.

**c**

**Purify binding site library**

- Introduce prey library into selection strain (one prey/cell).
- Eliminate self-activating prey using 5-FQA.
- Isolate surviving prey.

**d**

**Binding site selection procedure**

- Co-transform with TF and select.
- Library complexity is limited by transformation efficiency (~10⁹).

High-throughput SELEX

Randomized Library

Amplification

Selection

Purified protein

High complexity library

Next-gen Sequencing

The impact of sequencing technology

Sanger SELEX -> SAGE-SELEX -> NGS

~7k sequences

Points of discussion

• Which motif finding algorithm is better?
  • Specific hypothesis
  • Binding site model
  • Search/optimization method and objective function
  • Some basic rules in practice

• Evolution of TF binding sites
  • Binding site turn over
  • Evolution by substitution, in/del, duplication, transposition
  • Co-evolution with TF
  • Impact on shaping regulatory networks

• Motif != binding != function
  • Sensitivity and specificity of wet/dry experiments
  • How to validate?
    • Biological function versus biochemical activity

• Species specific regulation
• Beyond primary sequence conservation
Challenge of Specificity

- A 7-mer is expected to occur every 16,384 base pairs by chance
- In human, this means $3 \times 10^9 / 16,384 \approx 180,000$ sites in total
- TFBS are usually degenerative
- Total number of genes $\approx 25,000$
- Most of predicted binding sites are false positives!
- Need other restrictive information to reduce false positives
Challenges and open questions

• How to better integrate large scale genomics data?
  • RNA-seq, ChIP-seq, DNasel-seq,

• How to correlate with chromatin structure and epigenetic marks?

• How to connect predicted motifs to TFs?

• Can we eliminate complexity in some contexts?
  • Limited number of players in cell types?
  • Do open chromatin regions represent a ‘reduced’ genome?