Motif Discovery

Regulatory Motifs
Find promoter motifs associated with co-regulated or functionally related genes

Transcription Factor Binding Sites
• Very Small
• Highly Variable
• ~Constant Size
• Often repeated
• Low-complexity-ish

Motifs Are Degenerate
• Protein-DNA interactions
  – Proteins read DNA by "feeling" the chemical properties of the bases
  – Without opening DNA (not by base complementarity)
• Sequence specificity
  – Topology of 3D contact dictates sequence specificity of binding
  – Some positions are fully constrained, other positions are degenerate
  – “Ambiguous / degenerate” positions are loosely contacted by the transcription factor

Other “Motifs”
• Splicing Signals
  – Splice junctions
  – Exonic Splicing Enhancers (ESE)
  – Exonic Splicing Suppressors (ESS)
• Protein Domains
  – Glycosylation sites
  – Kinase targets
  – Targetting signals
• Protein Epitopes
  – MHC binding specificities

Essential Tasks
• Modeling Motifs
  – How to computationally represent motifs
• Visualizing Motifs
  – Motif “Information”
• Predicting Motif Instances
  – Using the model to classify new sequences
• Learning Motif Structure
  – Finding new motifs, assessing their quality
**Modeling Motifs**

### Consensus Sequences

- **Useful for publication**
  - HEM13: CCCATT TTTCCTC
  - HEM13: TTTCT GTTTCCTC
  - HEM13: TCAATTTT TTAAG
  - ANB1: CTCATT TTTCCTC
  - ANB1: TCCATT TTTCCTC
  - ANB1: CTCATT TTTCCTC
  - ANB1: TCCATT TTTCCTC
  - ROX1: CCAATTTTTTTTC

- **IUPAC symbols for degenerate sites**
  - Not very amenable to computation

---

### Probabilistic Model

**Position Frequency Matrix (PFM)**

- **Count frequencies**
  - Add pseudocounts

<table>
<thead>
<tr>
<th></th>
<th>HEM13</th>
<th>HEM13</th>
<th>HEM13</th>
<th>ANB1</th>
<th>ANB1</th>
<th>ANB1</th>
<th>ANB1</th>
<th>ROX1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>T</td>
<td>A</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Position Weight Matrix (PWM)**

- **Scoring a Sequence**
  - To score a sequence, compare to a null model
  - Log likelihood ratio
  - Score = \( \log \frac{P(S | PFM)}{P(S | B)} \)

**Position Weight Matrix (PWM)**

- **Background DNA (B)**

- **Represent both base frequency and conservation at each position**

- **Common threshold = 60% of maximum score**

---

### Scoring a Sequence

**Common threshold = 60% of maximum score**

---

### Visualizing Motifs – Motif Logos

- **Represent both base frequency and conservation at each position**

- **Height of letter proportional to frequency of base at that position**

- **Height of stack proportional to conservation at that position**

---

**Nature Biotechnology 24, 423 - 425 (2006)**


**Common threshold = 60% of maximum score**
Motif Information

The height of a stack is often called the motif information at that position measured in bits.

Motif Position Information = \[ 2 - \sum_{b=A,T,G,C} -p_b \log_2 p_b \]

Why is this a measure of information?

Uncertainty and probability

Uncertainty is related to our surprise at an event.

“The sun will rise tomorrow”  Not surprising \( p \approx 1 \)

“The sun will not rise tomorrow”  Very surprising \( p \ll 1 \)

Uncertainty is inversely related to probability of event.

Uncertainty \( \propto \frac{1}{p_{\text{event}}} \)

Average Uncertainty

Two possible outcomes for sun rising

A  “The sun will rise tomorrow”  \( P(A) = p_1 \)

B  “The sun will not rise tomorrow”  \( P(B) = p_2 \)

What is our average uncertainty about the sun rising?

Entropy

Entropy measures average uncertainty

Entropy measures randomness

\[ H(X) = -\sum_i p_i \log_2 p_i \]

If \( \log \) is base 2, then the units are called bits.

Entropy versus randomness

Entropy is maximum at maximum randomness

Example: Coin Toss

\( P(\text{heads}) = 0.1 \)  Not very random

\( H(X) = 0.47 \) bits

\( P(\text{heads}) = 0.5 \)  Completely random

\( H(X) = 1 \) bits

Entropy Examples

\[ H(X) = -\left[ 0.25 \log_2(0.25) + 0.25 \log_2(0.25) + 0.25 \log_2(0.25) + 0.25 \log_2(0.25) \right] \]

\( = 2 \) bits

\[ H(X) = -\left[ 0.1 \log_2(0.1) + 0.1 \log_2(0.1) + 0.75 \log_2(0.75) \right] \]

\( = 0.63 \) bits
Information Content

Information is a decrease in uncertainty

Once I tell you the sun will rise, your uncertainty about the event decreases

Information = \( H_{\text{before}}(X) - H_{\text{after}}(X) \)

Information is difference in entropy after receiving information

Motif Information

Motif Position Information = \( 2 - \sum_{b \in \{A,T,G,C\}} -p_b \log p_b \)

Prior uncertainty about nucleotide

Uncertainty after learning it is position i in a motif

Background DNA Frequency

The definition of information assumes a uniform background DNA nucleotide frequency

What if the background frequency is not uniform?

Background DNA Frequency

Motif Position Information = \( 1.7 - \sum_{b \in \{A,T,G,C\}} -p_b \log p_b = -0.2 \) bits

Some motifs could have negative information!

A Different Measure

Relative entropy or Kullback-Leibler (KL) divergence

Divergence between a “true” distribution and another

\[ D_{\text{KL}}(P_{\text{true}} \parallel P_{\text{background}}) = \sum_{i \in \{A,T,G,C\}} P_{\text{true}}(i) \log \frac{P_{\text{true}}(i)}{P_{\text{background}}(i)} \]

“True” Distribution Other Distribution

\( D_{\text{KL}} \) is larger the more different \( P_{\text{true}} \) is from \( P_{\text{background}} \)

Comparing Both Methods

Information assuming uniform background DNA

KL Distance assuming 20% GC content (e.g. Plasmodium)
Online Logo Generation

Finding New Motifs

Learning Motif Models

A Promoter Model

Length K

The same motif model in all promoters

Probability of a Sequence

Given a sequence(s), motif model and motif location

Probability of a Sequence

A: 0.25
T: 0.25
G: 0.25
C: 0.25
A: 0.25
T: 0.25
G: 0.25
C: 0.25

Parameterizing the Motif Model

Given multiple sequences and motif locations but no motif model

AATTCG
ATATCG
ATATCG
GATGCA

Finding Known Motifs

Given multiple sequences and motif model but no motif locations

Calculate P(Seqwindow|Motif) for every starting location
Motif Position Distribution $Z_{ij}$

- the element $Z_{ij}$ of the matrix $Z$ represents the probability that the motif starts in position $j$ in sequence $i$.

$$Z = \begin{bmatrix}
\text{seq1} & 0.1 & 0.1 & 0.2 & 0.6 \\
\text{seq2} & 0.4 & 0.2 & 0.1 & 0.3 \\
\text{seq3} & 0.3 & 0.1 & 0.5 & 0.1 \\
\text{seq4} & 0.1 & 0.5 & 0.1 & 0.3 \\
\end{bmatrix}$$

Some examples:

- $Z_1$
- $Z_2$
- $Z_3$
- $Z_4$

Calculating the Z Vector

$$P(Z_0 = 1 | S, M) = \frac{P(S | Z_{ij} = 1, M) P(Z_{ij} = 1)}{P(S)} \quad \text{(Bayes' rule)}$$

$$P(Z_0 = 1 | S, M) = \frac{\prod_{j=1}^{m} P(S | Z_{ij} = 1, M) P(Z_{ij} = 1)}{\sum_{j=1}^{m} \prod_{i=1}^{n} P(S | Z_{ij} = 1, M) P(Z_{ij} = 1)}$$

Assume uniform priors (motif equally likely to start at any position)

Discovering Motifs

Given a set of co-regulated genes, we need to discover
with only sequences

We have neither a motif model nor motif locations
Need to discover both

How can we approach this problem?

Expectation Maximization (EM)

Remember the basic idea!

1. Use model to estimate distribution of missing data
2. Use estimate to update model
3. Repeat until convergence

Model is the motif model
Missing data are the motif locations

EM for Motif Discovery

1. Start with random motif model
2. E Step: estimate probability of motif positions for each sequence
3. M Step: use estimate to update motif model
4. Iterate (to convergence)
The M-Step Calculating the Motif Matrix

• $M_{ck}$ is the probability of character $c$ at position $k$
• With specific motif positions, we can estimate $M_{ck}$:
  \[ M_{ck} = \sum \frac{n_{c,k}}{\sum n_{k,j}} + \frac{d_{c,k}}{\sum d_{k,j}} \]
• But with probabilities of positions, $Z_{ij}$, we average:
  \[ n_{c,k} = \sum_{sequences} \sum \frac{Z_{ij}}{\sum_{sequences} n_{k,j}} \]

MEME

• MEME – implements EM for motif discovery in DNA and proteins
• MAST – search sequences for motifs given a model

P(Seq|Model) Landscape

EM searches for parameters to increase P(seq|parameters)

Useful to think of P(seq|parameters) as a function of parameters
EM starts at an initial set of parameters
And then “climbs uphill” until it reaches a local maximum

Search from Many Different Starts

To minimize the effects of local maxima, you should search multiple times from different starting points

MEME uses this idea

Search from Many Different Starts

Start at many points
Run for one iteration
Choose starting point that got the “highest” and continue

The ZOOPS Model

• The approach as we’ve outlined it, assumes that each sequence has exactly one motif occurrence per sequence; this is the OOPS model
• The ZOOPS model assumes zero or one occurrences per sequence

E-step in the ZOOPS Model

• We need to consider another alternative: the $i$th sequence doesn’t contain the motif
• We add another parameter (and its relative)
  \[ \lambda \]
  • prior prob that any position in a sequence is the start of a motif
  \[ \gamma = (L - W + 1)\lambda \]
  • prior prob of a sequence containing a motif
**E-step in the ZOOPS Model**

\[ P(Z_g = 1) = \frac{\Pr(S_g \mid Z_g = 1, M) \lambda}{\Pr(S_g \mid Z_g = 0, M)(1 - \gamma) + \sum_{i=1}^{M} \Pr(S_g \mid Z_{g_i} = 1, M) \lambda} \]

- Here \( Q_g \) is a random variable that takes on 0 to indicate that the sequence doesn’t contain a motif occurrence.

\[ Q_g = \sum_{j=1}^{L-W+1} Z_{g,j} \]

**M-step in the ZOOPS Model**

- Update \( \rho \) same as before.
- Update \( \lambda, \gamma \) as follows.

\[ Z^{(i+1)} = \frac{\gamma^{(i+1)} (L-W+1)}{n(L-W+1)} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{g,i,j} \]

- Average of \( Z_{g,i,j}^{(i)} \) across all sequences, positions.

**The TCM Model**

- The TCM (two-component mixture model) assumes zero or more motif occurrences per sequence.

**Likelihood in the TCM Model**

- The TCM model treats each length \( W \) subsequence independently.
- To determine the likelihood of such a subsequence:

\[ \Pr(S_g \mid Z_g = 1, M) = \prod_{k=j}^{j+W-1} M_{g,k} \text{ assuming a motif starts there} \]

\[ \Pr(S_g \mid Z_g = 0, p) = \prod_{k=j}^{j+W-1} P(c_k \mid B) \text{ assuming a motif doesn’t start there} \]

**E-step in the TCM Model**

\[ Z_g = \frac{\Pr(S_g \mid Z_g = 1, M) \lambda}{\Pr(S_g \mid Z_g = 0, B)(1 - \lambda) + \Pr(S_g \mid Z_g = 1, M) \lambda} \]

- M-step same as before.

**Gibbs Sampling**

A stochastic version of EM that differs from deterministic EM in two key ways.

1. At each iteration, we only update the motif position of a single sequence.
2. We may update a motif position to a “suboptimal” new position.

---

8
Gibbs Sampling

1. Start with random motif locations and calculate a motif model
2. Randomly select a sequence, remove its motif and recalculate temporary model
3. With temporary model, calculate probability of motif at each position on sequence
4. Select new position based on this distribution
5. Update model and iterate

Gibbs Sampling and Climbing

Because gibbs sampling does not always choose the best new location, it can move to another place not directly uphill.

In theory, Gibbs Sampling less likely to get stuck a local maxima.

AlignACE

- Implements Gibbs sampling for motif discovery
  - Several enhancements
- ScanAce – look for motifs in a sequence given a model
- CompareAce – calculate “similarity” between two motifs (i.e. for clustering motifs)

http://atlas.med.harvard.edu/cgi-bin/alignace.pl

Antigen Epitope Prediction

Antigens and Epitopes

- Antigens are molecules that induce immune system to produce antibodies
- Antibodies recognize parts of molecules called epitopes

Genome to “Immunome”

Pathogen genome sequences provide define all proteins that could illicit an immune response

- Looking for a needle...
  - Only a small number of epitopes are typically antigenic
- In a very big haystack
  - Vaccinia virus (258 ORFs): 175,716 potential epitopes (8-, 9-, and 10-mers)
  - M. tuberculosis (~4K genes): 433,206 potential epitopes
  - A. nidulans (~9K genes): 1,579,000 potential epitopes

Can computational approaches predict all antigenic epitopes from a genome?
Antigen Processing & Presentation

Modeling MHC Epitopes
- Have a set of peptides that have been associated with a particular MHC allele
- Want to discover motif within the peptide bound by MHC allele
- Use motif to predict other potential epitopes

Motifs Bound by MHCs
- **MHC 1**
  - Closed ends of groove
  - Peptides 8-10 AAs in length
  - Motif is the peptide

- **MHC 2**
  - Grove has open ends
  - Peptides have broad length distribution: 10-30 AAs
  - Need to find binding motif within peptides

MHC 2 Motif Discovery
- Use Gibbs Sampling!
- 462 peptides known to bind to MHC II
  - HLA-DR4(B1*0401)
- 9-30 residues in length
- Goal: identify a common length 9 binding motif

Vaccinia Epitope Prediction
- Predict MHC1 binding peptides
- Using 4 matrices for H-2 Kb and Db
- Top 1% predictions experimentally validated
- 49 validated epitopes accounting for 95% of immune response