Lecture 09 - Motif Discovery

The foundations

<table>
<thead>
<tr>
<th>Lect</th>
<th>Fundamental bio problem</th>
<th>Fundamental comp. tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Sequence alignment</td>
<td>Dynamic programming</td>
</tr>
<tr>
<td>3</td>
<td>Database search</td>
<td>Hashing</td>
</tr>
<tr>
<td>4</td>
<td>Microarray analysis</td>
<td>Clustering / EM</td>
</tr>
<tr>
<td>5</td>
<td>Cellular states</td>
<td>Classification / Modeling</td>
</tr>
<tr>
<td>6,7</td>
<td>Modeling biological signals</td>
<td>HMMs/Modeling/Learning/EM</td>
</tr>
<tr>
<td>8</td>
<td>Modern gene identification</td>
<td>Conditional Random Fields</td>
</tr>
<tr>
<td>9</td>
<td>Regulatory motifs</td>
<td>Information/Gibbs Sampling/EM</td>
</tr>
<tr>
<td>10</td>
<td>Natural Selection</td>
<td>Statistical inference</td>
</tr>
<tr>
<td>11</td>
<td>Gene trees and species trees</td>
<td>Phylogenetics/Bayesian inference</td>
</tr>
<tr>
<td>12</td>
<td>Inferring population history</td>
<td>Modeling selection</td>
</tr>
</tbody>
</table>

Midterm and switch to final project mode! (Tue 10/27 at 11am in class)

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
2. Expectation maximization approach
3. Extensions to EM approach
4. Gibbs Sampling
5. Motif representation and information content
6. Applications

The regulatory code: All about regulatory motifs

- The parts list: ~20-30k genes
  - Protein-coding genes, RNA genes (tRNA, microRNA, snRNA)
- The circuitry: constructs controlling gene usage
  - Enhancers, promoters, splicing, post-transcriptional motifs
- The regulatory code, complications:
  - Combinatorial coding of 'unique tags'
  - Data-centric encoding of addresses
  - Overlaid with 'memory' marks
    - Large-scale on/off states
  - Modulation of the large-scale coding
    - Post-transcriptional and post-translational information
- Today: discovering motifs in co-regulated promoters

Challenges in Computational Biology

- DNA
  - Genome Assembly
  - Gene Finding
  - Comparative Genomics
  - Evolutionary Theory
  - Gene expression analysis
  - Protein network analysis
- RNA transcript
  - Regulatory network inference
  - Emerging network properties

Previously: Clustering / Classification → Discover groups of co-regulated genes.
Today: Find common regulatory motifs within them.
Regulatory motif discovery

- **Regulatory motifs**
  - Genes are turned on/off in response to changing environments
  - No direct addressing: subroutines (genes) contain sequence tags (motifs)
  - Specialized proteins (transcription factors) recognize these tags

- **What makes motif discovery hard?**
  - Motifs are short (6-8 bp), sometimes degenerate
  - Can contain any set of nucleotides (no ATG or other rules)
  - Act at variable distances upstream (or downstream) of target gene

How Transcription Factors actually recognize motifs

- **Proteins 'feel' DNA**
  - Read chemical properties of bases
  - Do NOT open DNA (no base complementarity)

- **3D Topology dictates specificity**
  - Fully constrained positions:
    - every atom matters
  - 'Ambiguous / degenerate' positions
    - loosely contacted

- **Other types of recognition**
  - MicroRNAs: complementarity
  - Nucleosomes: GC content
  - RNAs: structure/seqn combination

Motifs summarize TF sequence specificity

- **Summarize information**
- **Integrate many positions**
- **Measure of information**
- **Distinguish motif vs. motif instance**
- **Assumptions:**
  - Independence
  - Fixed spacing

Motifs are not limited to DNA sequences

- **Splicing Signals at the RNA level**
  - Splice junctions
  - Exonic Splicing Enhancers (ESE)
  - Exonic Splicing Surpressors (ESS)

- **Domains and epitopes at the Protein level**
  - Glycosylation sites
  - Kinase targets
  - Targeting signals
  - MHC binding specificities

- **Recurring patterns at the physiological level**
  - Expression patterns during the cell cycle
  - Heart beat patterns predicting cardiac arrest
  - Any probabilistic recurring pattern

How would you go about it?

- Given a set of co-regulated/functionally related genes, find common motifs in their promoter regions
- Align the promoters to each other using local alignment
- Use expert knowledge for what motifs should look like
- Find ‘median’ string by enumeration (motif/sample driven)
- Start with conserved blocks in the upstream regions

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
2. Expectation maximization approach
3. Extensions to EM approach
4. Gibbs Sampling
5. Motif representation and information content
6. Applications
### Starting positions ⇔ Motif matrix

- Given aligned sequences → easy to compute profile matrix

<table>
<thead>
<tr>
<th>Motif matrix sequence positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 0.1 0.3 0.1 0.2 0.4 0.3 0.1</td>
</tr>
<tr>
<td>C 0.5 0.2 0.1 0.1 0.6 0.1 0.2 0.7</td>
</tr>
<tr>
<td>G 0.2 0.2 0.6 0.5 0.1 0.2 0.2 0.1</td>
</tr>
<tr>
<td>T 0.2 0.3 0.2 0.2 0.1 0.3 0.3 0.1</td>
</tr>
</tbody>
</table>

**Given profile matrix**: easy to find starting position probabilities

Key idea: Iterative procedure for estimating both, given uncertainty

(learning problem with hidden variables: the starting positions)

### Representing Motif (p_{ck}) and Background (p_{0c})

- Assume motif has fixed width, W
- Motif represented by matrix of probabilities: \( p_{ck} \)
  - the probability that the \( k \)th character of motif is letter \( c \)

\[
 p = \begin{bmatrix}
 1 & 2 & 3 \\
 A & 0.1 & 0.5 & 0.2 \\
 C & 0.4 & 0.2 & 0.1 \\
 G & 0.3 & 0.1 & 0.6 \\
 T & 0.2 & 0.2 & 0.1 
\end{bmatrix}
\]

- Background represented by \( p_{0c} \), frequency of each base

\[
 p_0 = \begin{bmatrix}
 0 & A & 0.26 & C & 0.24 & G & 0.23 & T & 0.27 
\end{bmatrix}
\]

### Representing the starting position probabilities (Z_{ij})

- The element \( Z_{ij} \) of the matrix \( Z \) represents the probability that the motif starts in position \( j \) in sequence \( i \)

\[
 Z = \begin{bmatrix}
 1 & 2 & 3 & 4 \\
 seq1 & 0.1 & 0.1 & 0.2 & 0.6 \\
 seq2 & 0.4 & 0.2 & 0.1 & 0.3 \\
 seq3 & 0.3 & 0.1 & 0.5 & 0.1 \\
 seq4 & 0.1 & 0.5 & 0.1 & 0.3 
\end{bmatrix}
\]

Some examples:

- \( Z_1 \): uniform
- \( Z_2 \): one big winner
- \( Z_3 \): two candidates
- \( Z_4 \): no clear winner

### Starting positions \( (Z_{ij}) \) ⇔ Motif matrix \( p_{ck} \)

- \( Z_{ij} \): Probability that on sequence \( i \), motif start at position \( j \)
- \( p_{ck} \): Probability that \( k \)th character of motif is letter \( c \)

### Three examples for Greedy, Gibbs Sampling, EM

- **Greedy** always picks maximum
  - Gibbs sampling picks one at random
  - EM uses both in estimating motif

- All methods agree

- **Greedy ignores most of the probabilities**
  - Gibbs sampling rapidly converges to some choice
  - EM averages over the entire sequence (no preference)
E-step: Calculating Zij from motif

Starting positions: Zij  Motif: \( p_{ij} \)

Calculating P(Xi) when motif position is known

\[
Pr(X_i | Z_{ij} = 1, p) = \prod_{i=1}^{n} \prod_{j=0}^{W-1} \prod_{k=1}^{8} \prod_{c=A}^{T} p_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_{ij} k_{ij} c_
M-step example: Estimating $p_{cr}$ from $Z_{ij}$

$X_1 = \begin{bmatrix} A & C & A & G & C & A \end{bmatrix}$
$Z_1 = \begin{bmatrix} 0.1 & 0.7 & 0.1 & 0.1 \end{bmatrix}$

$X_2 = \begin{bmatrix} A & G & G & C & A \end{bmatrix}$
$Z_2 = \begin{bmatrix} 0.4 & 0.1 & 0.1 & 0.4 \end{bmatrix}$

$X_3 = \begin{bmatrix} T & C & A & G & T & C \end{bmatrix}$
$Z_3 = \begin{bmatrix} 0.2 & 0.6 & 0.1 & 0.1 \end{bmatrix}$

- $EM$: sum over full probability
  - $n_{A,1} = 0.1+0.1+0.4+0.1 = 0.7$
  - $n_{C,1} = 0.7+0.4+0.6 = 1.7$
  - $n_{G,1} = 0.1+0.1+0.1+0.1 = 0.4$
  - $n_{T,1} = 0.2 = 0.2$
  - Total: $T = 0.7+1.7+0.4+0.2 = 3.0$

- Normalize and add pseudo-counts
  - $P_{A,1} = (0.7+1)/(T+4) = 1.7/7 = 0.24$
  - $P_{C,1} = (1.7+1)/(T+4) = 2.7/7 = 0.39$
  - $P_{G,1} = (0.4+1)/(T+4) = 1.4/7 = 0.2$
  - $P_{T,1} = (0.2+1)/(T+4) = 1.2/7 = 0.17$

$P_{X_1} = Z_1 = Z_3 / (Z_1 + Z_2 + Z_3 + 1)$

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>A</td>
</tr>
</tbody>
</table>

Gibbs sampling: Pick one, Greedy: Pick max

### The EM Algorithm

- EM converges to a local maximum in the likelihood of the data given the model:
  $$\prod_i Pr(X_i | p)$$

- Usually converges in a small number of iterations
- Sensitive to initial starting point (i.e. values in $p$)

### P(Seq|Model) Landscape

**EM searches for parameters to increase $P$s**

Useful to think of $P$s as a function of parameters

EM starts at an initial set of parameters
And then "climbs uphill" until it reaches a local maximum

Where EM starts can make a big difference

### 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

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

### Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
2. Expectation maximization approach
3. Extensions to EM approach
4. Gibbs Sampling
5. Motif representation and information content
6. Applications
3a. Extending the basic one-motif-per-sequence model

**Basic Promoter Model: One instance per promoter**

![Diagram](image)

The same motif model in all promoters

**The ZOOPS Model**

- (zero-or-one-occurrence-per-sequence)
  - 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)

$$
\lambda \quad \text{prior prob that any position in a sequence is the start of a motif}
$$

$$
\gamma = (L-W+1)\lambda \quad \text{prior prob of a sequence containing a motif}
$$

**M-step in the ZOOPS Model**

- update $p$ same as before
- update $\lambda, \gamma$ as follows

$$
\lambda^{(i+1)} = \frac{\gamma^{(i+1)}}{(L-W+1)} = \frac{1}{n(L-W+1)} \sum_{j=1}^{n} \sum_{t=1}^{N} Z_{t,i}^{(i)}
$$

- average of $Z_{t,i}^{(i)}$ across all sequences, positions

**E-step in the ZOOPS Model**

- update $Q_i$ as follows

$$
Q_i = \sum_{j=1}^{n} Z_{t,j}
$$
The TCM Model (two-component mixture 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:

$$\Pr(X_{ij} \mid Z_{ij} = 1, p) = \prod_{k=j}^{j+W-1} p_{ij,k-j+1}$$ assuming a motif starts there

$$\Pr(X_{ij} \mid Z_{ij} = 0, p) = \prod_{k=j}^{j+W-1} p_{ij,k}$$ assuming a motif doesn’t start there

E-step in the TCM Model

$$Z_{ij}^{(t)} = \frac{\Pr(X_{ij} \mid Z_{ij} = 1, p^{(t)}) \lambda^{(t)}}{\Pr(X_{ij} \mid Z_{ij} = 0, p^{(t)}) (1 - \lambda^{(t)}) + \Pr(X_{ij} \mid Z_{ij} = 1, p^{(t)}) \lambda^{(t)}}$$

- M-step same as before

Finding Multiple Motifs

- basic idea: discount the likelihood that a new motif starts in a given position if this motif would overlap with a previously learned one
- when re-estimating $$Z_{ij}$$, multiply by

$$\Pr(V_{ij} = 1) = \begin{cases} 1, & \text{no previous motifs in } \{X_{ij}, \ldots, X_{i,j+w-1}\} \\ 0, & \text{otherwise} \end{cases}$$

- $$V_{ij}$$ is estimated using $$Z_{ij}$$ values from previous passes of motif finding

MEME Enhancements to the Basic EM Approach

- MEME builds on the basic EM approach in the following ways:
  - trying many starting points
  - not assuming that there is exactly one motif occurrence in every sequence
  - allowing multiple motifs to be learned
  - incorporating Dirichlet prior distributions

3b. Other enhancements to basic algorithm
Starting Points in MEME

- for every distinct subsequence of length \( W \) in the training set
  - derive an initial \( p \) matrix from this subsequence
  - run EM for 1 iteration
- choose motif model (i.e. \( p \) matrix) with highest likelihood
- run EM to convergence

Using Subsequences as Starting Points for EM

- set values corresponding to letters in the subsequence to \( X \)
- set other values to \((1-X)/(M-1)\) where \( M \) is the length of the alphabet
- example: for the subsequence TAT with \( X=0.5 \)

\[
P = \begin{pmatrix}
1 & 2 & 3 \\
A & 0.17 & 0.5 & 0.17 \\
C & 0.17 & 0.17 & 0.17 \\
G & 0.17 & 0.17 & 0.17 \\
T & 0.5 & 0.17 & 0.5 \\
\end{pmatrix}
\]

Repeats, and a Better Background Model

- Repeat DNA can be confused as motif
  - Especially low-complexity CACACA...AAAAA, etc.

Solution:

- more elaborate background model
  - 0th order: \( B = \{ p_A, p_C, p_G, p_T \} \)
  - 1st order: \( B = \{ P(A|A), P(A|C), ..., P(T|T) \} \)
  - ...
  - \( K \)th order: \( B = \{ P(X | b_1...b_K); X, b_i \in \{A,C,G,T\} \} \)

Has been applied to EM and Gibbs (up to 3rd order)

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
2. Expectation maximization approach
3. Extensions to EM approach
4. Gibbs Sampling
5. Motif representation and information content
6. Applications

Three examples of Greedy, Gibbs Sampling, EM

- Greedy always picks maximum
  - Gibbs sampling picks one at random
  - EM uses both in estimating motif

- All methods agree

- Greedy ignores most of the probability
  - Gibbs sampling rapidly converges to some choice
  - EM averages over the entire sequence (no preference)

Gibbs Sampling

- a general procedure for sampling from the joint distribution of a set of random variables
- \( \Pr(U_1, U_2, ..., U_n) \) by iteratively sampling from
- for each \( j \)
  - \( \Pr(U_j | U_1, ..., U_{j-1}, U_{j+1}, ..., U_n) \)
- application to motif finding: Lawrence et al. 1993
- can view it as a stochastic analog of EM for this task
- less susceptible to local minima than EM
Gibbs Sampling Approach

- in the EM approach we maintained a distribution $Z_i$ over the possible motif starting points for each sequence
- in the Gibbs sampling approach, we'll maintain a specific starting point for each sequence $a_i$, but we'll keep resampling these

Sampling New Motif Positions

- for each possible starting position, $a_j = j$, compute a weight
  $$A_j = \frac{\prod_{k=0}^{j-1} P_{x, k+1} \prod_{k=j+1}^{W-1} P_{x, k-j}}{\prod_{k=0}^{W-1} P_{x, k}}$$
- randomly select a new starting position $a_j$ according to these weights

Gibbs Sampling (AlignACE)

- AlignACE: first statistical motif finder
- BioProspector: improved version of AlignACE

Algorithm (sketch):

1. Initialization:
   a. Select random locations in sequences $x^1, ... , x^n$
   b. Compute an initial model $M$ from these locations

2. Sampling Iterations:
   a. Remove one sequence $x^i$
   b. Recalculate model
   c. Pick a new location of motif in $x^i$ according to probability
   the location is a motif occurrence

Gibbs Sampling (AlignACE)

- Given:
  - $x^1, ... , x^n$
  - motif length $K$
  - background $B$

- Find:
  - $M$: Model $M$
  - Locations $a_1, ... , a_N$ in $x^1, ... , x^n$

Maximizing log-odds likelihood ratio:

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \log \frac{M(k, x^i_{a_k})}{B(x^i_{a_k})}$$

Gibbs Sampling Approach

given: length parameter $W$, training set of sequences
choose random positions for $a$
do
pick a sequence $X_j$
estimate $p$ given current motif positions $a$ (update step)
(using all sequences but $X_j$)
sample a new motif position $a_j$ for $X_j$ (sampling step)
until convergence
return: $p, a$
Gibbs Sampling (AlignACE)

**Predictive Update:**
- Select a sequence \( x = x' \)
- Remove \( x' \), recompute model:

\[
M_{kj} = \frac{1}{(N-1)+B} (\beta_j + \sum_{x_{j+k-1}} (x_{j+k-1} = j))
\]

where \( \beta_j \) are pseudocounts to avoid 0s, and \( B = \sum \beta_j \)

---

Gibbs Sampling (AlignACE)

**Sampling:**
For every K-long word \( x_j, \ldots, x_{j+k-1} \) in \( x \):

- \( Q_j = \text{Prob} \{ \text{word} | \text{motif} \} = M(1, x_j) \times \ldots M(k, x_{j+k-1}) \)
- \( P_j = \text{Prob} \{ \text{word} | \text{background} \} = B(x_j) \times \ldots B(x_{j+k-1}) \)

Let

\[
A_j = \frac{Q_j / P_j}{\sum_{j=1}^{N} Q_j / P_j}
\]

Sample a random new position \( a_i \) according to the probabilities \( A_1, \ldots, A_{N+k+1} \).

---

Gibbs Sampling (AlignACE)

**Running Gibbs Sampling:**
1. Initialize
2. Run until convergence
3. Repeat 1,2 several times, report common motifs

---

Advantages / Disadvantages

- **Advantages:**
  - Very similar to EM
  - Easier to implement
  - Less dependent on initial parameters
  - More versatile, easier to enhance with heuristics

- **Disadvantages:**
  - More dependent on all sequences to exhibit the motif
  - Less systematic search of initial parameter space

---

Gibbs Sampling and Climbing

Because gibbs sampling does 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*

---

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
2. Expectation maximization approach
3. Extensions to EM approach
4. Gibbs Sampling
5. Motif representation and information content
6. Applications
Visualizing Motifs – Motif Logos

Represent both base frequency and conservation at each position

Motif Information

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

Motif Position Information = \( 2 - \sum_{i=A,T,G,C} -p_i \log_2 p_i \)

Why is this a measure of information?

Uncertainty and probability

Uncertainty is related to our surprise at an event

“The sun will rise tomorrow” \( \text{Not surprising (p~1)} \)

“The sun will not rise tomorrow” \( \text{Very surprising (p<<1)} \)

Uncertainty is inversely related to probability of event

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

Uncertainty = \( \log \frac{1}{p_{\text{event}}} \)

Uncertainty = \( - \log p_{\text{event}} \)

Entropy: 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

\[ = P(A)\text{Uncertainty}(A) + P(B)\text{Uncertainty}(B) \]

\[ = -p_1 \log p_1 - p_2 \log p_2 \]

\[ = -\sum p_i \log p_i = \text{Entropy} \]

Entropy

Entropy measures average uncertainty

Entropy measures randomness

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

If \( \log \) is base 2, then the units are called bits
(\( \log_e \) nats … \( \log_{10} \) dits / digits)

Entropy versus randomness

Entropy is maximum at maximum randomness

Example: Coin Toss

P(\text{heads})=0.1 \not \text{very random} \quad H(X)=0.47 \text{bits}

P(\text{heads})=0.5 \text{completely random} \quad H(X)=1 \text{bits}

P(\text{heads})=1/3 \quad H(X) = 0.918

- \text{If heads:}
  \log_2(1/(1/3))=\log_2 3=1.58
  \text{More than 1 bit of information}
### Entropy Examples

**Entropy Examples**

\[ H(X) = 0.25 \log(0.25) + 0.25 \log(0.25) + 0.25 \log(0.25) \]

= 2 bits

\[ H(X) = 0.1 \log(0.1) + 0.1 \log(0.1) + 0.75 \log(0.75) \]

= 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

\[ H_{\text{background}}(X) \]

Prior uncertainty about nucleotide

\[ H_{\text{motif}}(X) \]

Uncertainty after learning it is position i in a motif

Uncertainty at this position has been reduced by 0.37 bits.

**Motif Logo**

Conserved Residue

Reduction of uncertainty of 2 bits

Little Conservation

Minimal reduction of uncertainty

---

**Background DNA Frequency**

The definition of information assumes a uniform background DNA nucleotide frequency.

What if the background frequency is not uniform?

\[ H_{\text{background}}(X) \]

H(X)=2 bits

\[ H_{\text{motif}}(X) \]

H(X)=0.63 bits

Motif Position Information

\[ \sum_{i={A,T,G,C}} -p_i \log p_i \]

= -0.2 bit

Some motifs could have negative information!

**A Different Measure**

Relative entropy or Kullback-Leibler (KL) divergence

Divergence between a "true" distribution and another

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

\( D_{KL} \) 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

http://weblogo.berkeley.edu/
http://bioinfoугл.р.р.о.у.т.е/енологос/enologos.cgi

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
2. Expectation maximization approach
3. Extensions to EM approach
4. Gibbs Sampling
5. Motif representation and information content
6. Applications

6a. Motif discovery in yeast

Example Application: Motifs in Yeast

Group:
Tavazoie et al. 1999, G. Church’s lab, Harvard

Data:
- Microarrays on 6,220 mRNAs from yeast Affymetrix chips (Cho et al.)
- 15 time points across two cell cycles

Processing of Data

1. Selection of 3,000 genes
   - Genes with most variable expression were selected
2. Clustering according to common expression
   - K-means clustering
   - 30 clusters, 50-190 genes/cluster
   - Clusters correlate well with known function
3. AlignACE motif finding
   - 600-long upstream regions
   - 50 regions/trial
6b. Antigen Epitope Prediction

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 Epitope Prediction

Antigens and Epitopes

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

- Have a set of peptides that have been associate 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 grove
  - 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

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
2. Expectation maximization approach
3. Extensions to EM approach
4. Gibbs Sampling
5. Motif representation and information content
6. Applications

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