CSE P 590 A
Autumn 2008
Lecture 5
Motifs: Representation & Discovery

Outline

Last week: Learning from data:
- MLE: Max Likelihood Estimators
- EM: Expectation Maximization (MLE w/hidden data)

Expression & regulation
- Expression: creation of gene products
- Regulation: when/where/how much of each gene product; complex and critical

Next: using MLE/EM to find regulatory motifs in biological sequence data

George Palade
Nov. 19, 1912 -- Oct 8, 2008

1966 Albert Lasker Award for Basic Medical Research
1974 Nobel Prize in Physiology or Medicine (with Albert Claude and Christian de Duve)
Identified the function of mitochondria, ribosomes and cellular secretion

Gene Expression & Regulation
Gene Expression

Recall a gene is a DNA sequence for a protein. To say a gene is expressed means that it is transcribed from DNA to RNA. The mRNA is processed in various ways. It is exported from the nucleus (eukaryotes) and translated into protein. A key point: not all genes are expressed all the time, in all cells, or at equal levels.

RNA Transcription

Some genes heavily transcribed (many are not).

Regulation

In most cells, pro- or eukaryote, easily a 10,000-fold difference between least- and most-highly expressed genes. Regulation happens at all steps. E.g., some transcripts can be sequestered then released, or rapidly degraded, some are weakly translated, some are very actively translated, some are highly transcribed, some are not transcribed at all. Below, focus on 1st step only: transcriptional regulation.

E. coli growth on glucose + lactose

http://en.wikipedia.org/wiki/Lac_operon
DNA Binding Proteins

A variety of DNA binding proteins ("transcription factors"; a significant fraction, perhaps 5-10%, of all human proteins) modulate transcription of protein coding genes.
In the groove

Different patterns of potential H bonds at edges of different base pairs, accessible esp. in major groove.

H-T-H Dimers

Bind 2 DNA patches, ~ 1 turn apart
Increases both specificity and affinity.

Helix-Turn-Helix DNA Binding Motif

Zinc Finger Motif
Leucine Zipper Motif

Homo-/hetero-dimers and combinatorial control

Some Protein/DNA interactions well-understood

But the overall DNA binding “code” still defies prediction

CAP
Bacterial Met Repressor

Negative feedback loop:
high Met level ⇒ repress Met synthesis genes
(a beta-sheet DNA binding domain)

Summary

Proteins can bind DNA to regulate gene expression (i.e., production of other proteins & themselves)
This is widespread
Complex combinatorial control is possible

Sequence Motifs

Motif: “a recurring salient thematic element”

Last few slides described structural motifs in proteins

Equally interesting are the DNA sequence motifs to which these proteins bind - e.g., one leucine zipper dimer might bind (with varying affinities) to dozens or hundreds of similar sequences

DNA binding site summary

Complex “code”
Short patches (4-8 bp)
Often near each other (1 turn = 10 bp)
Often reverse-complements
Not perfect matches
E. coli Promoters

“TATA Box” ~ 10bp upstream of transcription start

How to define it?
- Consensus is TATAAT
- BUT all differ from it
- Allow k mismatches?
- Equally weighted?
- Wildcards like R,Y? (\{A,G\}, \{C,T\}, resp.)

Not exact: of 168 studied (mid 80’s)
- nearly all had 2/3 of TAxyzT
- 80-90% had all 3
- 50% agreed in each of x,y,z
- no perfect match

Other common features at -35, etc.

E. coli Promoters

“TATA Box” - consensus TATAAT
~10bp upstream of transcription start

TATA Box Frequencies

<table>
<thead>
<tr>
<th>pos base</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>95</td>
<td>26</td>
<td>59</td>
<td>51</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
<td>2</td>
<td>14</td>
<td>13</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>1</td>
<td>16</td>
<td>15</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>79</td>
<td>3</td>
<td>44</td>
<td>13</td>
<td>17</td>
<td>96</td>
</tr>
</tbody>
</table>

TATA Scores

<table>
<thead>
<tr>
<th>pos base</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-36</td>
<td>19</td>
<td>1</td>
<td>12</td>
<td>10</td>
<td>-46</td>
</tr>
<tr>
<td>C</td>
<td>-15</td>
<td>-36</td>
<td>-8</td>
<td>-9</td>
<td>-3</td>
<td>-31</td>
</tr>
<tr>
<td>G</td>
<td>-13</td>
<td>-46</td>
<td>-6</td>
<td>-7</td>
<td>-9</td>
<td>-46</td>
</tr>
<tr>
<td>T</td>
<td>17</td>
<td>-31</td>
<td>8</td>
<td>-9</td>
<td>-6</td>
<td>19</td>
</tr>
</tbody>
</table>
Scanning for TATA

Score Distribution (Simulated)

Weight Matrices:
Statistics

Assume:

\[ f_{b,i} = \text{frequency of base } b \text{ in position } i \text{ in TATA} \]
\[ f_b = \text{frequency of base } b \text{ in all sequences} \]

Log likelihood ratio, given \( S = B_1B_2...B_6 \):

\[
\log \left( \frac{P(S|\text{"tata"})}{P(S|\text{"non-tata"})} \right) = \log \frac{\prod_{i=1}^{6} f_{B_{i,i}}}{\prod_{i=1}^{6} f_{B_i}} = \sum_{i=1}^{6} \log \frac{f_{B_{i,i}}}{f_{B_i}}
\]

Assumes independence
Neyman-Pearson

Given a sample \( x_1, x_2, \ldots, x_n \) from a distribution \( f(\cdot|\Theta) \) with parameter \( \Theta \), want to test hypothesis \( \Theta = \theta_1 \) vs \( \Theta = \theta_2 \).

Might as well look at likelihood ratio:

\[
\frac{f(x_1, x_2, \ldots, x_n|\theta_1)}{f(x_1, x_2, \ldots, x_n|\theta_2)} > \tau
\]

What’s best WMM?

Given, say, 168 sequences \( s_1, s_2, \ldots, s_k \) of length 6, assumed to be generated at random according to a WMM defined by 6 x (4-1) parameters \( \theta \), what’s the best \( \theta \)?

E.g., what’s MLE for \( \theta \) given data \( s_1, s_2, \ldots, s_k \)?

Answer: like coin flips or dice rolls, count frequencies per position (see HW).

Weight Matrices: Chemistry

Experiments show ~80% correlation of log likelihood weight matrix scores to measured binding energy of RNA polymerase to variations on TATAAT consensus.

[Stormo & Fields]
Another WMM example

### 8 Sequences:

<table>
<thead>
<tr>
<th></th>
<th>Freq.</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATG</td>
<td>A 0.625</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ATG</td>
<td>C 0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ATG</td>
<td>G 0.250</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ATG</td>
<td>T 0.125</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>GTG</td>
<td>A 0.625</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>GTG</td>
<td>C 0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>GTG</td>
<td>G 0.250</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GTG</td>
<td>T 0.125</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

### Log-Likelihood Ratio:

\[
\log_2 \frac{f_{x,i}}{f_{x,i}} = \frac{1}{4}
\]

Non-uniform Background

- *E. coli* - DNA approximately 25% A, C, G, T
- *M. jannaschi* - 68% A-T, 32% G-C

LLR from previous example, assuming:

\[
f_A = f_T = 3/8 \\
f_C = f_G = 1/8
\]

e.g., G in col 3 is 8 x more likely via WMM than background, so \((\log_2)\) score = 3 (bits).

Relative Entropy

AKA Kullback-Liebler Distance/Divergence, AKA Information Content

Given distributions \(P, Q\):

\[
H(P||Q) = \sum_{x \in \Omega} P(x) \log \frac{P(x)}{Q(x)} \geq 0
\]

Notes:

- \(P(x) \log \frac{P(x)}{Q(x)} = 0\) if \(P(x) = 0\) \(\text{[since } \lim_{y \to 0} y \log y = 0]\)
- Undefined if \(0 = Q(x) < P(x)\)

WMM: How “Informative”?

Mean score of site vs bkg?

For any fixed length sequence \(x\), let

\[
P(x) = \text{Prob. of } x \text{ according to WMM} \\
Q(x) = \text{Prob. of } x \text{ according to background}
\]

Relative Entropy:

\[
H(P||Q) = \sum_{x \in \Omega} P(x) \log_2 \frac{P(x)}{Q(x)}
\]

\(H(P||Q)\) is expected log likelihood score of a sequence randomly chosen from WMM;

\(-H(Q||P)\) is expected score of Background
For WMM, you can show (based on the assumption of independence between columns), that:

\[ H(P||Q) = \sum_i H(P_i||Q_i) \]

where \( P_i \) and \( Q_i \) are the WMM/background distributions for column \( i \).

**Pseudocounts**

Are the \(-\infty\)'s a problem?

- Certain that a given residue never occurs in a given position? Then \(-\infty\) just right
- Else, it may be a small-sample artifact

Typical fix: add a pseudocount to each observed count—small constant (e.g., .5, 1)

Sounds \textit{ad hoc}; there is a Bayesian justification
WMM Summary

Weight Matrix Model (aka Position Specific Scoring Matrix, PSSM, “possum”, 0th order Markov models)
Simple statistical model assuming independence between adjacent positions
To build: count (+ pseudocount) letter frequency per position, log likelihood ratio to background
To scan: add LLRs per position, compare to threshold
Generalizations to higher order models (i.e., letter frequency per position, conditional on neighbor) also possible, with enough training data

How-to Questions

Given aligned motif instances, build model?
  Frequency counts (above, maybe w/ pseudocounts)
Given a model, find (probable) instances
  Scanning, as above
Given unaligned strings thought to contain a motif, find it? (e.g., upstream regions of co-expressed genes)
  Hard ... rest of lecture.

Motif Discovery

Unfortunately, finding a site of max relative entropy in a set of unaligned sequences is NP-hard [Akutsu]

Motif Discovery: 4 example approaches

Brute Force
Greedy search
Expectation Maximization
Gibbs sampler
Brute Force

Input:
Motif length $L$, plus sequences $s_1, s_2, ..., s_k$ (all of length $n+L-1$, say), each with one instance of an unknown motif

Algorithm:
Build all $k$-tuples of length $L$ subsequences, one from each of $s_1, s_2, ..., s_k$ ($n^k$ such tuples)
Compute relative entropy of each
Pick best

Brute Force, II

Input:
Motif length $L$, plus seqs $s_1, s_2, ..., s_k$ (all of length $n+L-1$, say), each with one instance of an unknown motif

Algorithm in more detail:
Build singletons: each len $L$ subseq of each $s_1, s_2, ..., s_k$ ($n^k$ sets)
Extend to pairs: len $L$ subseqs of each pair of seqs ($n^2k$ sets)
Then triples: len $L$ subseqs of each triple of seqs ($n^3k$ sets)
Repeat until all have $k$ sequences ($n^k$ sets)
Compute relative entropy of each; pick best

Greedy Best-First
[Hertz & Stormo]

Input:
Sequences $s_1, s_2, ..., s_k$; motif length $L$;
“breadth” $d$, say $d = 1000$

Algorithm:
As in brute, but discard all but best $d$
relative entropies at each stage

Expectation Maximization
[MEME, Bailey & Elkan, 1995]

Input (as above):
Sequence $s_1, s_2, ..., s_k$; motif length $L$; background model; again assume one instance per sequence (variants possible)

Algorithm: EM
Visible data: the sequences
Hidden data: where’s the motif

$$ Y_{i,j} = \begin{cases} 1 & \text{if motif in sequence } i \text{ begins at position } j \\ 0 & \text{otherwise} \end{cases} $$

Parameters $\theta$: The WMM
MEME Outline

Typical EM algorithm:

Parameters $\theta^t$ at $t^{th}$ iteration, used to estimate where the motif instances are (the hidden variables)

Use those estimates to re-estimate the parameters $\theta$ to maximize likelihood of observed data, giving $\theta^{t+1}$

Repeat

Key: given a few good matches to best motif, expect to pick out more

Expectation Step

(where are the motif instances?)

$$\hat{Y}_{i,j} = E(Y_{i,j} \mid s_i, \theta^t)$$

$$= P(Y_{i,j} = 1 \mid s_i, \theta^t)$$

$$= P(s_i \mid Y_{i,j} = 1, \theta^t) \frac{P(Y_{i,j} = 1 \mid s_i, \theta^t)}{P(s_i \mid \theta^t)}$$

$$= cP(s_i \mid Y_{i,j} = 1, \theta^t)$$

$$= c' \prod_{k=1}^l P(s_{i,j+k-1} \mid \theta^t)$$

where $c'$ is chosen so that $\sum_j \hat{Y}_{i,j} = 1.$

Maximization Step

(what is the motif?)

Find $\theta$ maximizing expected value:

$$Q(\theta \mid \theta^t) = E_Y_{i=1} \log P(s_i, Y_i \mid \theta)$$

$$= E_{Y_{i=1}} \log \prod_{i=1}^k P(s_i, Y_{i} \mid \theta)$$

$$= E_{Y_{i=1}} \log \sum_{i=1}^{k} P(s_i, Y_i \mid \theta)$$

$$= E_{Y_{i=1}} \sum_{i=1}^{k} \sum_{j=1}^{\mid s_i \mid + l + 1} Y_{i,j} \log P(s_i, Y_{i,j} = 1 \mid \theta)$$

$$= E_{Y_{i=1}} \sum_{i=1}^{k} \sum_{j=1}^{\mid s_i \mid + l + 1} Y_{i,j} \log \left(P(s_i \mid Y_{i,j} = 1, \theta)P(Y_{i,j} = 1 \mid \theta) \right)$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{\mid s_i \mid + l + 1} E_{Y_{i=1}} \log P(s_i \mid Y_{i,j} = 1, \theta) + C$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{\mid s_i \mid + l + 1} \hat{Y}_{i,j} \log P(s_i \mid Y_{i,j} = 1, \theta) + C$$

M-Step (cont.)

$$Q(\theta \mid \theta^t) = \sum_{i=1}^{k} \sum_{j=1}^{\mid s_i \mid + l + 1} \hat{Y}_{i,j} \log P(s_i \mid Y_{i,j} = 1, \theta) + C$$

Exercise: Show this is maximized by “counting” letter frequencies over all possible motif instances, with counts weighted by $\hat{Y}_{i,j},$ again the “obvious” thing.
Initialization

1. Try every motif-length substring, and use as initial $\theta$ a WMM with, say 80% of weight on that sequence, rest uniform
2. Run a few iterations of each
3. Run best few to convergence

(Having a supercomputer helps)

Another Motif Discovery Approach

The Gibbs Sampler

Some History

Geman & Geman, IEEE PAMI 1984
Hastings, Biometrika, 1970
Josiah Williard Gibbs, 1839-1903, American physicist, a pioneer of thermodynamics

How to Average

An old problem:
\[ n \text{ random variables:} \quad x_1, x_2, \ldots, x_k \]
Joint distribution (p.d.f.): \( P(x_1, x_2, \ldots, x_k) \)
Some function: \( f(x_1, x_2, \ldots, x_k) \)
Want Expected Value:
\[ E(f(x_1, x_2, \ldots, x_k)) = \]
\[ \int_{x_1} \int_{x_2} \cdots \int_{x_k} f(x_1, x_2, \ldots, x_k) \cdot P(x_1, x_2, \ldots, x_k) \, dx_1 \, dx_2 \cdots dx_k \]

Approach 1: direct integration
(rarely solvable analytically, esp. in high dim)
Approach 2: numerical integration
(often difficult, e.g., unstable, esp. in high dim)
Approach 3: Monte Carlo integration

sample \( \bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(n)} \sim P(\bar{x}) \) and average:
\[ E(f(\bar{x})) \approx \frac{1}{n} \sum_{i=1}^{n} f(\bar{x}^{(i)}) \]

Markov Chain Monte Carlo (MCMC)

- Independent sampling also often hard, but not required for expectation
- MCMC \( \tilde{X}_{t+1} \sim P(\tilde{X}_{t+1} | \tilde{X}_t) \) w/ stationary dist \( \not= \) \( \bar{\pi} \)
- Simplest & most common: Gibbs Sampling
\[ P(x_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k) \]
- Algorithm
for \( t = 1 \) to \( \infty \)
for \( i = 1 \) to \( k \) do:
\[ x_{t+1,i} \sim P(x_{t+1,i} | x_{t+1,1}, x_{t+1,2}, \ldots, x_{t+1,i-1}, x_{t,i+1}, \ldots, x_{t,k}) \]
Input: again assume sequences $s_1, s_2, ..., s_k$ with one length $w$ motif per sequence

Motif model: WMM

Parameters: Where are the motifs? for $1 \leq i \leq k$, have $1 \leq x_i \leq |s_i| - w + 1$

“Full conditional”: to calc

$P(x_i = j \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$

build WMM from motifs in all sequences except $i$, then calc prob that motif in $i^{th}$ seq occurs at $j$ by usual “scanning” alg.

Input: again assume sequences $s_1, s_2, ..., s_k$ with one length $w$ motif per sequence

Motif model: WMM

Parameters: Where are the motifs? for $1 \leq i \leq k$, have $1 \leq x_i \leq |s_i| - w + 1$

“Full conditional”: to calc

$P(x_i = j \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$

build WMM from motifs in all sequences except $i$, then calc prob that motif in $i^{th}$ seq occurs at $j$ by usual “scanning” alg.

Overall Gibbs Alg

Randomly initialize $x_i$’s

for $t = 1$ to $\infty$

for $i = 1$ to $k$

    discard motif instance from $s_i$

    recalc WMM from rest

    for $j = 1 \ldots |s_i| - w + 1$

        calculate prob that $i^{th}$ motif is at $j$:

        $P(x_i = j \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$

        pick new $x_i$ according to that distribution

Similar to MEME, but it would average over, rather than sample from

Issues

Burnin - how long must we run the chain to reach stationarity?

Mixing - how long a post-burnin sample must we take to get a good sample of the stationary distribution? (Recall that individual samples are not independent, and may not “move” freely through the sample space. Also, many isolated modes.)
Variants & Extensions

“Phase Shift” - may settle on suboptimal solution that overlaps part of motif. Periodically try moving all motif instances a few spaces left or right.

Algorithmic adjustment of pattern width: Periodically add/remove flanking positions to maximize (roughly) average relative entropy per position

Multiple patterns per string
Methodology

13 tools
Real ‘motifs’ (Transfac)
56 data sets (human, mouse, fly, yeast)
‘Real’, ‘generic’, ‘Markov’
Expert users, top prediction only

Lessons

Evaluation is hard (esp. when “truth” is unknown)

Accuracy low

partly reflects limitations in evaluation methodology (e.g. ≤ 1 prediction per data set; results better in synth data)

partly reflects difficult task, limited knowledge (e.g. yeast > others)

No clear winner re methods or models
Motif Discovery
Summary

Important problem: a key to understanding gene regulation
Hard problem: short, degenerate signals amidst much noise
Many variants have been tried, for representation, search, and discovery. We looked at only a few:

- Weight matrix models for representation & search
- Greedy, MEME and Gibbs for discovery

Still much room for improvement. Comparative genomics, i.e. cross-species comparison is very promising