Previously: Learning from data
  MLE: Max Likelihood Estimators
  EM: Expectation Maximization (MLE w/hidden data)
These Slides:
  Bio: Expression & regulation
    Expression: creation of gene products
    Regulation: when/where/how much of each gene product; complex and critical
  Comp: using MLE/EM to find regulatory motifs in biological sequence data
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 is 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)
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 genes are highly transcribed, some are not transcribed at all, some transcripts can be sequestered then released, or rapidly degraded, some are weakly translated, some are very actively translated, ...

Below, focus on 1st step only:
transcriptional regulation
E. coli growth on glucose + lactose

http://en.wikipedia.org/wiki/Lac_operon
The *lac* Operon and its Control Elements

**lacI**  **lacZ**  **lacY**  **lacA**

mRNA  

![Diagram of the *lac* Operon](image)

**cAMP Activator Protein**  **RNA Polymerase**

High (constitutive) level of expression  Low glucose  Lactose available

High glucose  Lactose unavailable

Low glucose  Lactose unavailable

Low (basal) level of expression  High glucose  Lactose available
1965 Nobel Prize
Physiology or Medicine

François Jacob, Jacques Monod, André Lwoff
Sea Urchin - Endo16
### Module B

- **CY & CB1**
  - if true, \( i_1 = 1 \)
  - if false, \( i_1 = 0.5 \)
- **R**
  - if true, \( i_2 = i_1 \cdot U_l(t) \)
  - if false, \( i_3 = CB_2(t) \)
  - \( i_3 = k \cdot CB_2(t) \) for \( 1 < k < 2 \)
- **P & CG1 & CB2**
  - if true, \( i_4 = 2 \)
  - if false, \( i_4 = 0 \)
- **UI(t) > threshold & R & i4 > 0**
  - if true, \( i_5 = 1 \)
  - if false, \( i_5 = 0 \)
  - \( i_6 = i_4 \cdot (i_2 + i_3) \)

### Module A

- **CG1, P, OTX, Z, CG2, SPGCF1, CG3, CG4**
- **BP**

### Equations

- **Module B**
  - \( i_2 = i_1 \cdot U_l(t) \)
  - \( i_3 = CB_2(t) \)
  - \( i_3 = k \cdot CB_2(t) \) for \( 1 < k < 2 \)
  - \( i_4 = 2 \)
  - \( i_4 = 0 \)
  - \( i_5 = 1 \)
  - \( i_5 = 0 \)
  - \( i_6 = i_4 \cdot (i_2 + i_3) \)

- **Module A**
  - \( i_7 = OTX(t) \)
  - \( i_7 = 0 \)
  - \( i_8 = i_6 + i_7 \)
  - \( i_9 = 1 \)
  - \( i_9 = 0 \)
  - \( i_10 = 0 \)
  - \( i_10 = i_8 \)
  - \( i_11 = 2 \)
  - \( i_11 = 1 \)
  - \( i_12 = i_11 \cdot i_10 \)
DNA Binding Proteins

A variety of DNA binding proteins (so-called “transcription factors”; a significant fraction, perhaps 5-10%, of all human proteins) modulate transcription of protein coding genes
As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b) is an uncoiled fragment of (a) three complementary base pairs. From a chemist’s viewpoint, each strand is a polymer made up of four re-
called deoxyribonucleotides.
In the groove

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

Figure 7-7 Molecular Biology of the Cell 5/e (© Garland Science 2008)
Helix-Turn-Helix DNA Binding Motif

(A) recognition helix

(B) DNA helix
H-T-H Dimers

Bind 2 DNA patches, ~ 1 turn apart
Increases both specificity and affinity
Zinc Finger Motif
Leucine Zipper Motif

Homo-/hetero-dimers and combinatorial control

Figure 7-19 Molecular Biology of the Cell S/e (© Garland Science 2008)

Figure 7-20 Molecular Biology of the Cell S/e (© Garland Science 2008)
MyoD

http://www.rcsb.org/pdb/explore/jmol.do?structureId=1MDY&bionumber=1
Some Protein/DNA interactions well-understood
But the overall DNA binding “code” still defies prediction
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
"TATA Box" ~ 10bp upstream of transcription start

How to define it?

- Consensus is TATAAT
- BUT all differ from it
- Allow k mismatches?
- Equally weighted?
**E. coli** Promoters

“TATA Box” - consensus TATAAT
   ~10bp upstream of transcription start
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.
### 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**
A “Weight Matrix Model” or “WMM”

<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

<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></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>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-15</td>
<td>-36</td>
<td>-8</td>
<td>-9</td>
<td>-3</td>
<td>-31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-13</td>
<td>-46</td>
<td>-6</td>
<td>-7</td>
<td>-9</td>
<td>-46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>17</td>
<td>-31</td>
<td>8</td>
<td>-9</td>
<td>-6</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

= -90

<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></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>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-15</td>
<td>-36</td>
<td>-8</td>
<td>-9</td>
<td>-3</td>
<td>-31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-13</td>
<td>-46</td>
<td>-6</td>
<td>-7</td>
<td>-9</td>
<td>-46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>17</td>
<td>-31</td>
<td>8</td>
<td>-9</td>
<td>-6</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

= 85

<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></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>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-15</td>
<td>-36</td>
<td>-8</td>
<td>-9</td>
<td>-3</td>
<td>-31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-13</td>
<td>-46</td>
<td>-6</td>
<td>-7</td>
<td>-9</td>
<td>-46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>17</td>
<td>-31</td>
<td>8</td>
<td>-9</td>
<td>-6</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

= -91

Scanning for TATA
TATA Scan at 2 genes

LacI

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

Might as well look at likelihood ratio:

$$\frac{f(x_1, x_2, ..., x_n | \theta_1)}{f(x_1, x_2, ..., x_n | \theta_2)} > \tau$$

(or log likelihood ratio)
Score Distribution
(Simulated)
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 $6 \times (4-1)$ 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>A</td>
<td>0.625</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0.250</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.125</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Log-Likelihood Ratio:

\[
\log_2 \frac{f_{x_i,i}}{f_{x_i}} \quad f_{x_i} = \frac{1}{4}
\]
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 e.g., G in col 3 is 8 x more likely via WMM than background, so \((\log_2)\) score = 3 (bits).

<table>
<thead>
<tr>
<th></th>
<th>LLR</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.74</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>G</td>
<td>1.00</td>
<td>-∞</td>
<td>3.00</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-1.58</td>
<td>1.42</td>
<td>-∞</td>
<td></td>
</tr>
</tbody>
</table>
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:

Let $P(x) \log \frac{P(x)}{Q(x)} = 0$ if $P(x) = 0$ [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

Expected score difference: $H(P||Q) + H(Q||P)$
On average, foreground model scores > background by 11.8 bits
(score difference of 118 on 10x scale used in examples above).

\[ H(P \parallel Q) = 5.0 \]

\[-H(Q \parallel P) = -6.8 \]
For a WMM:

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

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

Proof: exercise

Hint: Use the assumption of independence between WMM columns
WMM Example, cont.

<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>A</td>
<td>0.625</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0.250</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.125</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LLR</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.32</td>
<td>-∞</td>
<td>-∞</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-∞</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-1.00</td>
<td>2.00</td>
<td>-∞</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>RelEnt</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.70</td>
<td>2.00</td>
<td>2.00</td>
<td>4.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LLR</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.74</td>
<td>-∞</td>
<td>-∞</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1.00</td>
<td>-∞</td>
<td>3.00</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-1.58</td>
<td>1.42</td>
<td>-∞</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>RelEnt</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.51</td>
<td>1.42</td>
<td>3.00</td>
<td>4.93</td>
</tr>
</tbody>
</table>
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 *ad hoc*; there is a Bayesian justification
WMM Summary

Weight Matrix Model (aka Position Weight Matrix, PWM, Position Specific Scoring Matrix, PSSM, “possum”, 0th order Markov model)

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 $s_1, s_2, ..., s_k$ ($nk$ sets)

Extend to pairs: len $L$ subseqs of each pair of seqs ($n^2(k)$ sets)

Then triples: len $L$ subseqs of each triple of seqs ($n^3(\binom{k}{3})$ sets)

Repeat until all have $k$ sequences ($n^k(\binom{k}{k})$ sets)

Compute relative entropy of each; pick best

Problem: astronomically sloooow
Example

Three sequences (A, B, C), each with two possible motif positions (0, 1)
Greedy Best-First
[Hertz, Hartzell & Stormo, 1989, 1990]

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

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

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

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

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

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

(what is the motif?)

Find $\theta$ maximizing expected value:

$$Q(\theta \mid \theta^t) = E_{Y \sim \theta^t}[\log P(s, Y \mid \theta)]$$

$$= E_{Y \sim \theta^t}[\log \prod_{i=1}^{k} P(s_i, Y_i \mid \theta)]$$

$$= E_{Y \sim \theta^t}[\sum_{i=1}^{k} \log P(s_i, Y_i \mid \theta)]$$

$$= E_{Y \sim \theta^t}[\sum_{i=1}^{k} \sum_{|s_i|-l+1}^{Y_{i,j}} \log P(s_i, Y_{i,j} = 1 \mid \theta)]$$

$$= E_{Y \sim \theta^t}[\sum_{i=1}^{k} \sum_{j=1}^{\hat{Y}_{i,j}} \log(P(s_i \mid Y_{i,j} = 1, \theta)P(Y_{i,j} = 1 \mid \theta))]$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{|s_i|-l+1} E_{Y \sim \theta^t}[Y_{i,j}] \log P(s_i \mid Y_{i,j} = 1, \theta) + C$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{|s_i|-l+1} \hat{Y}_{i,j} \log P(s_i \mid Y_{i,j} = 1, \theta) + C$$
**M-Step (cont.)**

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

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.

\[
\begin{align*}
  s_1 & : \text{ ACGGATT...} \\
  \vdots & \\
  s_k & : \text{ GC... TCGGAC} \\
  \hat{Y}_{1,1} & : \text{ ACGG} \\
  \hat{Y}_{1,2} & : \text{ CGGA} \\
  \hat{Y}_{1,3} & : \text{ GGAT} \\
  \vdots & \\
  \hat{Y}_{k,l-1} & : \text{ CGGA} \\
  \hat{Y}_{k,l} & : \text{ GGAC}
\end{align*}
\]
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):

http://meme.sdsc.edu/
Another Motif Discovery Approach

The Gibbs Sampler

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>94</td>
<td>222</td>
<td>265</td>
<td>137</td>
<td>9</td>
<td>9</td>
<td>137</td>
<td>137</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>52</td>
<td>222</td>
<td>94</td>
<td>94</td>
<td>9</td>
<td>265</td>
<td>606</td>
</tr>
<tr>
<td>Lys</td>
<td>9</td>
<td>133</td>
<td>442</td>
<td>380</td>
<td>9</td>
<td>71</td>
<td>380</td>
<td>194</td>
<td>9</td>
<td>133</td>
<td>9</td>
<td>71</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>71</td>
<td>256</td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>53</td>
<td>9</td>
<td>96</td>
<td>401</td>
<td>9</td>
<td>9</td>
<td>140</td>
<td>140</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>53</td>
<td>140</td>
<td>140</td>
<td>9</td>
<td>9</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>67</td>
<td>9</td>
<td>9</td>
<td>473</td>
<td>9</td>
<td>9</td>
<td>299</td>
<td>125</td>
<td>9</td>
<td>67</td>
<td>9</td>
<td>67</td>
<td>67</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>9</td>
<td>600</td>
<td>224</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>224</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>228</td>
<td>63</td>
<td>278</td>
<td>9</td>
<td>9</td>
<td>170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>240</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>125</td>
<td>125</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>125</td>
<td>125</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>240</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>168</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>168</td>
<td>89</td>
<td>9</td>
<td>89</td>
<td>248</td>
<td>9</td>
<td>168</td>
<td>89</td>
<td>9</td>
<td>89</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>117</td>
<td>9</td>
<td>117</td>
<td>117</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>819</td>
<td>63</td>
<td>387</td>
<td>63</td>
<td>9</td>
<td>819</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>151</td>
<td>9</td>
<td>56</td>
<td>9</td>
<td>9</td>
<td>151</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>1141</td>
<td>9</td>
<td>151</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>56</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Ala</td>
<td>9</td>
<td>9</td>
<td>112</td>
<td>43</td>
<td>181</td>
<td>901</td>
<td>43</td>
<td>181</td>
<td>215</td>
<td>9</td>
<td>43</td>
<td>9</td>
<td>43</td>
<td>181</td>
<td>112</td>
<td>9</td>
<td>43</td>
<td>78</td>
</tr>
<tr>
<td>Thr</td>
<td>915</td>
<td>130</td>
<td>130</td>
<td>9</td>
<td>251</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>311</td>
<td>130</td>
<td>70</td>
<td>855</td>
<td>9</td>
<td>130</td>
<td>9</td>
</tr>
<tr>
<td>Pro</td>
<td>76</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>295</td>
<td>581</td>
<td>295</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>58</td>
<td>107</td>
<td>9</td>
<td>9</td>
<td>500</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>156</td>
<td>9</td>
<td>598</td>
<td>9</td>
<td>205</td>
<td>58</td>
<td>9</td>
<td>746</td>
<td>9</td>
<td>58</td>
</tr>
<tr>
<td>Leu</td>
<td>9</td>
<td>121</td>
<td>9</td>
<td>9</td>
<td>149</td>
<td>9</td>
<td>9</td>
<td>149</td>
<td>458</td>
<td>9</td>
<td>149</td>
<td>9</td>
<td>37</td>
<td>37</td>
<td>9</td>
<td>177</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Ile</td>
<td>9</td>
<td>166</td>
<td>114</td>
<td>61</td>
<td>323</td>
<td>9</td>
<td>114</td>
<td>166</td>
<td>9</td>
<td>9</td>
<td>427</td>
<td>9</td>
<td>61</td>
<td>61</td>
<td>9</td>
<td>427</td>
<td>9</td>
<td>61</td>
</tr>
<tr>
<td>Met</td>
<td>9</td>
<td>104</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>198</td>
<td>198</td>
<td>9</td>
<td>104</td>
<td>9</td>
<td>9</td>
<td>198</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>9</td>
<td>9</td>
<td>136</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>262</td>
<td>262</td>
<td>9</td>
<td>136</td>
<td>136</td>
<td>9</td>
<td>262</td>
<td>9</td>
<td>262</td>
<td>136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>108</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>366</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>366</td>
<td></td>
</tr>
</tbody>
</table>
Some History

Geman & Geman, IEEE PAMI 1984

Hastings, Biometrika, 1970


Josiah Williard Gibbs, 1839-1903, American physicist, a pioneer of thermodynamics
An old problem:

- $n$ random variables:
- 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: $E(f(x_1, x_2, \ldots, x_k))$
How to Average

\[ E(f(x_1, x_2, \ldots, x_k)) = \int_{x_1} \int_{x_2} \cdots \int_{x_k} f(x_1, x_2, \ldots, x_k) \cdot P(x_1, x_2, \ldots, x_k) \, dx_1 \, dx_2 \cdots \, dx_k \]

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 \( \vec{X}_{t+1} \sim P(\vec{X}_{t+1} | \vec{X}_t) \) w/ stationary dist = \( P \)

• Simplest & most common: Gibbs Sampling

\[
P(x_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)
\]

• Algorithm

\[
\begin{align*}
&\text{for } t = 1 \text{ to } \infty \\
&\quad \text{for } i = 1 \text{ to } k \text{ do :} \\
&\quad \quad x_{t+1,i} \sim P(x_{t+1,i} \mid x_{t+1,1}, x_{t+1,2}, \ldots, x_{t+1,i-1}, x_t,i+1, \ldots, x_t,k)
\end{align*}
\]
\( \hat{Y}_{i,j} \)
Input: again assume sequences $s_1, s_2, \ldots, s_k$ 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.
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? In particular:

Samples are not independent; may not “move” freely through the sample space
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
Assessing computational tools for the discovery of transcription factor binding sites

Martin Tompa¹,², Nan Li¹, Timothy L Bailey³, George M Church⁴, Bart De Moor⁵, Eleazar Eskin⁶, Alexander V Favorov⁷,⁸, Martin C Frith⁹, Yutao Fu⁹, W James Kent¹⁰, Vsevolod J Makeev⁷,⁸, Andrei A Mironov⁷,¹¹, William Stafford Noble¹,², Giulio Pavesi¹², Graziano Pesole¹³, Mireille Régnier¹⁴, Nicolas Simonis¹⁵, Saurabh Sinha¹⁶, Gert Thijs⁵, Jacques van Helden¹⁵, Mathias Vandenbogaert¹⁴, Zhiping Weng⁹, Christopher Workman¹⁷, Chun Ye¹⁸ & Zhou Zhu⁴
Methodology

13 tools

Real ‘motifs’ (Transfac)

56 data sets (human, mouse, fly, yeast)

‘Real’, ‘generic’, ‘Markov’

Expert users, top prediction only

“Blind” – sort of
Lessons

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

Accuracy low

partly reflects limitations in evaluation methodology (e.g. \( \leq 1 \) 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