Outline

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
• is exported from the nucleus (eukaryotes)
• 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)
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 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, ...

All are important, but 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

**DNA**

```
  5' → binding site → CAP → P → +1 → O → lacZ → lacY → lacA → 3'
```

**mRNA**

```
  mRNA AUG AUG AUG
  S   S   S
```

**RNA**

```
  (RNA)
```

**Expression Levels**

- **High expression**
  - Low glucose
  - Lactose available

- **No expression**
  - High glucose
  - Lactose unavailable
  - Low glucose
  - Lactose unavailable

- **Low (“basal”) expression**
  - High glucose
  - Lactose available
1965 Nobel Prize
Physiology or Medicine

François Jacob, Jacques Monod, André Lwoff
The sea urchin *Strongylocentrotus purpuratus*
Sea Urchin - Endo16
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
The Double Helix

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). Shown are three complementary base pairs, each from a different chemist's viewpoint, each strand made up of four repeating units 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

Figure 7-10 Molecular Biology of the Cell 5/e (© Garland Science 2008)
H-T-H Dimers

Bind 2 DNA patches, ~1 turn apart
Increases both specificity and affinity
LacI Repressor + DNA
(a tetrameric HTH protein)

Two operator DNA sequences bound

Tetramerization region

Zinc Finger Motif
Overheard at the Halloween Party

i'm a "zinc finger"
i'm just here for the free beer.
Leucine Zipper Motif

Homo-/hetero-dimers and combinatorial control
MyoD

http://www.rcsb.org/pdb/explore/jmol.do?structureId=1MDY&bionumber=1
We understand some Protein/DNA interactions

Figure 7-9 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Figure 7-25 Molecular Biology of the Cell 5/e (© Garland Science 2008)
But the overall DNA binding “code” still defies prediction

CAP, LacI repressor, and others bend DNA
Summary

Proteins can “bind” DNA to regulate gene expression (i.e., production of proteins, including themselves)

This is widespread

Complex, combinatorial control is both possible and commonplace
Sequence Motifs
Sequence Motifs

*Motif:* “a recurring salient thematic element”

Last few slides described *structural* motifs in *proteins*

Equally interesting are the *sequence* motifs in *DNA* 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 (dimer symmetry)
Not perfect matches
Example: *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?

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</th>
<th>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></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></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></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></td>
<td>17</td>
<td>-31</td>
<td>8</td>
<td>-9</td>
<td>-6</td>
<td>19</td>
</tr>
</tbody>
</table>

\[
\text{score} = 10 \log_2 \text{foreground:background odds ratio, rounded}
\]

(?) Arbitrary
### 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>

**Scoring:**

- Scoring for ACGT = -90
- Scoring for CAT = 85
- Scoring for ACT = -91

Scanning for TATA

See also slide 66
TATA Scan at 2 genes

See slide 47
Score Distribution
(Simulated)

$10^4$ random 6-mers from foreground (green) or uniform background (red)
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 \( S = B_1B_2...B_6 \):

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

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

Might as well look at likelihood ratio:

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

(or log likelihood ratio)
Score Distribution
(Simulated)

$10^4$ random 6-mers from foreground (green) or uniform background (red)
What’s best WMM?

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

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

Answer: like coin flips or dice rolls, count frequencies per position. (Possible HW?)
Weight Matrices: Biophysics

Experiments show ~80% correlation of log likelihood weight matrix scores to measured binding energies [Fields & Stormo, 1994]

I.e.,
- log prob $\propto$ energy
- “independence assumption” $\Rightarrow$
  - probabilities multiply & energies are additive
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>0.25</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:

\[
\log_2 \frac{f_{x_i,i}}{f_{x_i}} = \frac{1}{4} \quad \text{(uniform background)}
\]
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).

\[
\begin{align*}
  f_A &= f_T = \frac{3}{8} \\
  f_C &= f_G = \frac{1}{8}
\end{align*}
\]

<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>-\infty</td>
<td>-\infty</td>
<td>-\infty</td>
</tr>
<tr>
<td>C</td>
<td>-\infty</td>
<td>-\infty</td>
<td>-\infty</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>-\infty</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>-1.58</td>
<td>1.42</td>
<td>-\infty</td>
<td></td>
</tr>
</tbody>
</table>
Relative entropy
Relative Entropy

AKA Kullback-Liebler 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 $0 = Q(x) < P(x)$
Relative Entropy

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

- Intuition: A quantitative measure of how much P “diverges” from Q. (Think “distance,” but note it’s not symmetric.)
  - If \( P \approx Q \) everywhere, then \( \log(P/Q) \approx 0 \), so \( H(P||Q) \approx 0 \)
  - But as they differ more, sum is pulled above 0 (next 2 slides)
- What it means quantitatively: Suppose you sample \( x \), but aren’t sure whether you’re sampling from \( P \) (call it the “null model”) or from \( Q \) (the “alternate model”). Then \( \log(P(x)/Q(x)) \) is the log likelihood ratio of the two models given that datum. \( H(P||Q) \) is the expected per sample contribution to the log likelihood ratio for discriminating between those two models.
- Exercise: if \( H(P||Q) = 0.1 \), say. Assuming \( Q \) is the correct model, how many samples would you need to confidently (say, with 1000:1 odds) reject \( P \)?
\[ \ln x \leq x - 1 \]

\[ -\ln x \geq 1 - x \]

\[ \ln \left(\frac{1}{x}\right) \geq 1 - x \]

\[ \ln y \geq 1 - \frac{1}{y} \]

\((y = \frac{1}{x})\)
Theorem: \( H(P \parallel Q) \geq 0 \)

\[
H(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \\
\geq \sum_x P(x) \left(1 - \frac{Q(x)}{P(x)}\right) \\
= \sum_x (P(x) - Q(x)) \\
= \sum_x P(x) - \sum_x Q(x) \\
= 1 - 1 \\
= 0
\]

Idea: if \( P \neq Q \), then

\( P(x) > Q(x) \Rightarrow \log(P(x)/Q(x)) > 0 \)

and

\( P(y) < Q(y) \Rightarrow \log(P(y)/Q(y)) < 0 \)

Q: Can this pull \( H(P \parallel Q) < 0 \)?

A: No, as theorem shows.

Intuitive reason: sum is weighted by \( P(x) \), which is bigger at the positive log ratios vs the negative ones.

Furthermore: \( H(P \parallel Q) = 0 \) if and only if \( P = Q \)

Bottom line: “bigger” means “more different”
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:

$$H(P \parallel Q) = \sum_{x \in \Omega} P(x) \log_2 \frac{P(x)}{Q(x)}$$

$H(P \parallel Q)$ is expected log likelihood score of a sequence randomly chosen from WMM (wrt background);

$-H(Q \parallel P)$ is expected score of Background (wrt WMM)

Expected score difference: $H(P \parallel Q) + H(Q \parallel P)$
WMM Scores vs Relative Entropy

On average, foreground model scores > background by 11.8 bits (score difference of 118 on 10x scale used in examples above). $2^{11.8} \approx 3566$, which is good, since many more non-TATA than TATA
For a WMM:

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

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>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>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0.25</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>0.125</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Uniform</th>
<th>Non-uniform</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<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>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-∞</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>-1</td>
<td>2</td>
<td>-∞</td>
</tr>
<tr>
<td>RelEnt</td>
<td>0.7</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<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>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>-∞</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>-1.58</td>
<td>1.42</td>
<td>-∞</td>
</tr>
<tr>
<td>RelEnt</td>
<td>0.51</td>
<td>1.42</td>
<td>3</td>
</tr>
</tbody>
</table>
Pseudocounts

Are the $-\infty$’s a problem?

Are you certain that a given residue never occurs in a given pos? Then $-\infty$ just right. Else, it may be a small-sample artifact.

Typical fix: add a pseudocount to each observed count—small constant (often 1.0; but needn't be).

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 (k\text{th} order MM)
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
Motif Discovery

Based on the above, a natural approach to motif discovery, given, say, unaligned upstream sequences of genes thought to be co-regulated, is to find a set of subsequences of max relative entropy

Unfortunately, this 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 ($n^2 \binom{k}{2}$ 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)
$(n+1)^k$ in total; compute relative entropy of each; pick best
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):
Sequences $s_1, s_2, ..., s_k$; 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
Cartoon Example

xATAxyz

CATGACTAGCATAATCCGAT
TATAATTTCCCAAGGGAATAACA
TaCAATAGGACCATAGAATGCGC

xATAAAz

CATGACTAGCATAATCCGAT
TATAATTTCCCAAGGGAATAACA
TaCAATAGGACCATAGAATGCGC

TATAAT

CATGACTAGCATAATCCGAT
TATAATTTCCCAAGGGAATAACA
TaCAATAGGACCATAGAATGCGC

xATAAaz

CATGACTAGCATAATCCGAT
TATAATTTCCCAAGGGAATAACA
TaCAATAGGACCATAGAATGCGC

xATAAaz

CATGACTAGCATAATCCGAT
TATAATTTCCCAAGGGAATAACA
TaCAATAGGACCATAGAATGCGC

CATAAAT
CATGAC
GATAAAC
TATAAT
CATAGA
TAGAAT
AATAGG

TATAAAT

CATGACTAGCATAATCCGAT
TATAATTTCCCAAGGGAATAACA
TaCAATAGGACCATAGAATGCGC

CATAAT
GATAAAC
TATAAT
CATAGA
TAGAAT
TACAAT

TATAAT

CATGACTAGCATAATCCGAT
TATAATTTCCCAAGGGAATAACA
TaCAATAGGACCATAGAATGCGC

TATAAT
\[ \hat{Y}_{i,j} = E(Y_{i,j} \mid s_i, \theta^t) = P(Y_{i,j} = 1 \mid s_i, \theta^t) = P(s_i \mid Y_{i,j} = 1, \theta^t) \frac{P(Y_{i,j} = 1 \mid \theta^t)}{P(s_i \mid \theta^t)} = cP(s_i \mid Y_{i,j} = 1, \theta^t) = c' \prod_{k=1}^{l} P(s_{i,j+k-1} \mid \theta^t) \]

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

Find $\theta$ maximizing expected log likelihood:

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

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

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

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

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

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

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

From E-step
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. 

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

$s_1: \quad$ ACGGATTT...
$s_2: \quad$ GC...TCGGAC

$\hat{Y}_{1,1} \quad$ ACGG
$\hat{Y}_{1,2} \quad$ CGGA
$\hat{Y}_{1,3} \quad$ GGAT

$\hat{Y}_{k,l-1} \quad$ CGGA
$\hat{Y}_{k,l} \quad$ GGAC
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-suite.org
What Data?

Upstream regions of many genes (find widely shared motifs, like TATA)

Upstream regions of co-regulated genes (find shared, but more specific, motifs involved in that regulation)

ChIP seq data (find motifs bound by specific proteins)  

(slide 90)
Another Motif Discovery Approach
The Gibbs Sampler

<table>
<thead>
<tr>
<th>Protein</th>
<th>Accession</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigma-37</td>
<td>223</td>
<td>IIDLTYIQNK SQKETGDILGISQMHSVRLQRKAVKKLR 240 A25944</td>
</tr>
<tr>
<td>SpoIIIC</td>
<td>94</td>
<td>RFGLDLKEK TQREIAKELGIRSYSYVRIEKRALMKMF 111 A28627</td>
</tr>
<tr>
<td>NahR</td>
<td>22</td>
<td>VVFNQLLVDVR RVSTALENLGLTQPASNLKRLRTSLQ 39 A32837</td>
</tr>
<tr>
<td>Antennapedia</td>
<td>326</td>
<td>FHFNRYLTRR RRIETAILCTERTQIKWFQNRRMKWK 343 A23450</td>
</tr>
<tr>
<td>NtrC (Brady.)</td>
<td>449</td>
<td>LTAALAAATRG NQIRAADDLGLNNRTLRKIRDLDIQVY 466 B26499</td>
</tr>
</tbody>
</table>
| DicA         | 22        | IRYRRNLKHK TQSLAKALKSHVSVQ WERGDEPTG 39 B24328 (BVECD)
| MerD         | 5         | MNAY TVSRLLADAGVSVHVRD YLRRGLRPV 22 C29010 |
| Fis          | 73        | LDMVMQYTRG NQTRAALMGINRGITLRKLLKGYGMN 90 A32142 (DNFCS)
| MAT al       | 99        | FRRKQSLNNSK EKEEVAKCGCTPQLQVRWFINKRMRSK 116 A90983 (JEBY)
| Lambda cII   | 25        | SALLNKiAML GTEKTAEAVGDQKQISR WKRDIKFPS 42 A03579 (QCBP2)
| Crp (CAP)    | 169       | THPDGMQIKI TRQEOQIVGCSRETQVR ILKMLEDQNL 186 A03553 (QRECC)
| Lambda Cro   | 15        | ITLKYAMRF GQKTAKDLQYQSAINK AIHAGKIFL 32 A03577 (RCBPL)
| P22 Cro      | 12        | YKKDVIDHFQ TQRSAKALGIDASAVSQ WKEVIPKDA 29 A25867 (RGBP2)
| AraC         | 196       | ISDHLADSNF DIASYVHQVCLPSRLSH LFROQGLISV 213 A03554 (RGEC)
| Fnr          | 196       | FSPEFRTLM TGRTSGLGMDTETISR LLGFLQKSMG 213 A03552 (RGECF)
| HtpR         | 252       | ARWLDGDNKSG TLQELDRAYGVSPLRQ LEKNNMALKR 269 A00700 (RGECH)
| NtrC (K.a.)  | 444       | LTTLALRTQG HKEQAIRLGGWGRNITLR KLIELGME 461 A03564 (RGKBC)
| CytR         | 11        | MKAKQETAA TMKDVALKAVTSTVSR ALMNPDVKSQ 28 A24963 (RPECT)
| DeoR         | 23        | LQELKRSQKL HLKDAAALLGVESEITRR DLNNHAPVQV 40 A24076 (RPECO)
| GalR         | 3         | MA TIKVATLAGVSATVRVINNSPKASE 20 A03559 (RPEC)
| Laci         | 5         | MKPV TLYDVEYAVGYSQVTSPS RNQASHVSA 22 A03558 (RPECL)
| TetR         | 26        | LLNEVGEIGL TTRKLAQKLVEQPTYWRHVNKNRALLD 43 A03576 (RPECTN)
| TrpR         | 67        | IVEELLQGEM SQRLKELKAGIATIR GSNSLKAAPV 84 A03568 (RPECW)
| NifA         | 495       | LIAALETAGW VQAARKLLIGMTPQVAY RIIQIMDTMP 512 S02513 |
| SpoIIIG      | 205       | RFGLVSGEEK TQKDVADMMSOQSYISREKRIKLR 222 S07337 |
| Pin          | 160       | QAGRLIAAGT PRQKVAITYDYGVSLLYT TFPAGDK 177 S07958 |
| PurR         | 3         | MA TIKDVARALVSTTSVHLVKNTRFVIAE 20 S08477 |
| EbgR         | 3         | MA TLDCAIEAGVSLATVRVNLDDPTTNV 20 S09205 |
| LexA         | 27        | DHISQTMPP TRAEIAQRLLGFRSPNAE EHLKALARKG 44 S11945 |
| P22 CI       | 25        | SSILNRIAIR GQRKVADALGINESQISRGWKGDFIEPKMG 42 B25867 (Z1BPC2) |
|     | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Arg | 94 | 222| 265| 137| 9  | 9  | 137| 137| 9  | 9  | 9  | 52 | 222| 94 | 94 | 9  | 265| 606|
| Lys | 9  | 133| 442| 380| 9  | 71 | 380| 194| 9  | 133| 9  | 9  | 71 | 9  | 9  | 9  | 71 | 256|
| Glu | 53 | 9  | 96 | 401| 9  | 9  | 140| 140| 9  | 9  | 9  | 53 | 140| 140| 9  | 9  | 9  | 53 |
| Asp | 67 | 9  | 9  | 473| 9  | 9  | 299| 125| 9  | 67 | 9  | 67 | 67 | 9  | 9  | 9  | 67 |
| Gln | 9  | 600| 224| 9  | 9  | 224| 9  | 9  | 9  | 9  | 9  | 278| 63 | 278| 9  | 9  | 9  | 170|
| His | 240| 9  | 9  | 9  | 9  | 9  | 125| 125| 9  | 9  | 9  | 125| 125| 125| 9  | 9  | 9  | 240|
| Asn | 168| 9  | 9  | 9  | 9  | 9  | 168| 89 | 9  | 89 | 9  | 248| 9  | 168| 89 | 9  | 89 | 89 |
| Ser | 117| 9  | 9  | 117| 9  | 9  | 9  | 9  | 9  | 9  | 9  | 819| 63 | 387| 63 | 9  | 819| 9  |
| Gly | 151| 9  | 56 | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 151| 9  | 56 | 9  | 9  | 56 | 9  | 9  |
| Ala | 9  | 112| 43 | 181| 901| 9  | 43 | 181| 215| 9  | 43 | 9  | 43 | 181| 112| 43 | 78 | 9  |
| Thr | 915| 130| 130| 9  | 251| 9  | 9  | 9  | 9  | 9  | 9  | 311| 130| 70 | 855| 9  | 130| 9  |
| Pro | 76 | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 210| 210| 9  | 9  | ...| 9  |
| Cys | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  |
| Val | 58 | 107| 9  | 500| 9  | 9  | 9  | 156| 9  | 598| 9  | 9  | 205| 58 | 9  | 746| 9  | 58 |
| Leu | 9  | 121| 9  | 149| 9  | 9  | 9  | 156| 9  | 149| 9  | 9  | 37 | 37 | 9  | 177| 9  | 9  |
| Ile | 9  | 166| 114| 61 | 323| 9  | 114| 166| 9  | 427| 9  | 61 | 61 | 427| 9  | 9  | 61 | 9  |
| Met | 9  | 104| 9  | 9  | 9  | 9  | 9  | 198| 198| 9  | 104| 9  | 9  | 198| 9  | 9  | 9  | 9  |
| Tyr | 9  | 9  | 136| 9  | 9  | 9  | 262| 262| 9  | 9  | 136| 136| 9  | 262| 9  | 262| 136| 9  |
| Phe | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 108| 9  | 9  | 9  | 9  | 9  | 9  | 9  |
| Trp | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 9  | 366| 9  | 9  | 9  | 9  | 9  | 9  | 366|
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:

\( k \) random variables:

Joint distribution (p.d.f.):

Some function:

Want Expected Value:

\[
\begin{align*}
x_1, x_2, \ldots, x_k & \\
P(x_1, x_2, \ldots, x_k) & \\
f(x_1, x_2, \ldots, x_k) & \\
E(f(x_1, x_2, \ldots, x_k)) & 
\end{align*}
\]
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
  
  $P(x_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$

- Algorithm
  
  for $t = 1$ to $\infty$
    
    for $i = 1$ to $k$
      
      $x_{t+1,i} \sim P(x_{t+1,i} | x_{t+1,1}, x_{t+1,2}, \ldots, x_{t+1,i-1}, x_{t,i+1}, \ldots, x_{t,k})$
\( \hat{Y}_{i,j} \)
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? In particular:

- Samples are not independent; may not “move” freely through the sample space
- E.g., may be many isolated modes
“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 Tompa\textsuperscript{1,2}, Nan Li\textsuperscript{1}, Timothy L Bailey\textsuperscript{3}, George M Church\textsuperscript{4}, Bart De Moor\textsuperscript{5}, Eleazar Eskin\textsuperscript{6}, Alexander V Favorov\textsuperscript{7,8}, Martin C Frith\textsuperscript{9}, Yuta\textsuperscript{9}o Fu\textsuperscript{9}, W James Kent\textsuperscript{10}, Vsevolod J Makeev\textsuperscript{7,8}, Andrei A Mironov\textsuperscript{7,11}, William Stafford Noble\textsuperscript{1,2}, Giulio Pavesi\textsuperscript{12}, Graziano Pesole\textsuperscript{13}, Mireille Régnier\textsuperscript{14}, Nicolas Simonis\textsuperscript{15}, Saurabh Sinha\textsuperscript{16}, Gert Thijs\textsuperscript{5}, Jacques van Helden\textsuperscript{15}, Mathias Vandenbogaert\textsuperscript{14}, Zhiping Weng\textsuperscript{9}, Christopher Workman\textsuperscript{17}, Chun Ye\textsuperscript{18} & Zhou Zhu\textsuperscript{4}
Methodology

13 tools

Real ‘motifs’ (Transfac)

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

‘Real’, ‘generic’, ‘Markov’

Expert users, top prediction only

“Blind” – sort of
• \(nTP\) is the number of nucleotide positions in both known sites and predicted sites,
• \(nFN\) is the number of nucleotide positions in known sites but not in predicted sites,
• \(nFP\) is the number of nucleotide positions not in known sites but in predicted sites, and
• \(nTN\) is the number of nucleotide positions in neither known sites nor predicted sites.

• \(sTP\) be the number of known sites overlapped by predicted sites,
• \(sFN\) be the number of known sites not overlapped by predicted sites, and
• \(sFP\) be the number of predicted sites not overlapped by known sites.

At either the nucleotide \((x = n)\) or site \((x = s)\) level, one can then define:

• Sensitivity: \(xSn = xTP / (xTP + xFN)\), and
• Positive Predictive Value: \(xPPV = xTP / (xTP + xFP)\).

Specificity: \(nSP = nTN / (nTN + nFP)\).

Finally, it is enlightening to consider various single statistics that in some sense average (some of) these quantities. Following Pevzner & Sze\(^1\), define the (nucleotide level) performance coefficient as:

\[
nCC = \frac{nTP \cdot nTN - nFN \cdot nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}}
\]

The correlation coefficient \(nCC\) is the Pearson product-moment coefficient of correlation in the particular case of two binary variables, also called the ‘phi coefficient of correlation.’ The two binary variables are the characteristic vectors of the known nucleotide positions and
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
ChIP-seq

Chromatin ImmunoPrecipitation Sequencing
ChIP-seq

http://res.illumina.com/images/technology/chip_seq_assay_lg.gif
Map the reads back to the reference genome

Peak calling

De novo motif analysis
TF Binding Site Motifs
From ChIP-seq

LOTS of data

E.g. $10^3$–$10^5$ sites, hundreds of reads each
(plus perhaps even more nonspecific)

Motif variability

Co-factor binding sites
Discriminative motif analysis of high-throughput dataset

Zizhen Yao¹,* , Kyle L. MacQuarrie¹,², Abraham P. Fong³,⁴, Stephen J. Tapscott¹,³,⁵, Walter L. Ruzzo⁶,⁷,⁸ and Robert C. Gentleman⁹

¹Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, ²Molecular and Cellular Biology Program, University of Washington, Seattle, Washington, 98105, USA, ³Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, ⁴Department of Pediatrics, School of Medicine, ⁵Department of Neurology, School of Medicine, University of Washington, Seattle, Washington, 98105, USA, ⁶Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, ⁷Department of Computer Science and Engineering, ⁸Department of Genome Sciences, University of Washington, Seattle, Washington, 98105, USA and ⁹Bioinformatics and Computational Biology, Genentech, South San Francisco, CA 94080, USA
Method Outline

"Motif" = IUPAC pattern

\[ \text{e.g., GNRA , "N" = any, "R" = A or G} \]

**Discriminative model** –

Does presence/absence of motif "explain" presence/absence of peak?

Logistic regression – \( x = \text{motif count} \)

\[
\log \frac{p}{1-p} = \beta_0 + \beta_1 x
\]

Improve via seed/extend/perturb
Method Outline

Input sequences → Enumerate Nmers → Candidates filtering → Candidates counting → Seed selection

Seed motif → Mask motif match → Logistic regression → improved

Find next motif → perturbation → extension

Not improved → Seed refinement

Motif

Repeat

Repeat
Figure 2

A. The two methods predict similar motifs most of the time. The motifs are shown as dots, with each dot representing a motif from one of the 148 ENCODE datasets. The motifs are color-coded to indicate whether they are predicted by both methods (blue), DREME (orange), or motifRG (green). The x-axis represents the number of motifs predicted by DREME, and the y-axis represents the number of motifs predicted by motifRG. The size of the dots is proportional to the number of motifs predicted by each method.

B. The -log10(p-value) for each motif is shown in the scatter plot, with the x-axis representing the -log10(p-value) for DREME and the y-axis representing the -log10(p-value) for motifRG. The motifs are color-coded to indicate whether they are predicted by both methods (blue), DREME (orange), or motifRG (green). The x-axis represents the number of motifs predicted by DREME, and the y-axis represents the number of motifs predicted by motifRG. The size of the dots is proportional to the number of motifs predicted by each method.

C. The time (in seconds) taken by each method to predict the motifs is shown in the scatter plot, with the x-axis representing the time taken by DREME and the y-axis representing the time taken by motifRG. The motifs are color-coded to indicate whether they are predicted by both methods (blue), DREME (orange), or motifRG (green). The x-axis represents the number of motifs predicted by DREME, and the y-axis represents the number of motifs predicted by motifRG. The size of the dots is proportional to the number of motifs predicted by each method.
Discriminative motif analysis can be applied to any high-throughput sequence datasets besides ChIP-Seq data. We used motif prediction in a wide range of biological samples and applications. Here, we also give evidence that the method is robust to variation in sample size (Fig. 6). To determine how the z-scores of enriched motifs change with sample size, we randomly kept 20, 40 to 100% of the original foreground samples and replaced the remaining labeling with size from 1 to 64K. We plotted the distribution of z-values for all 6mers sampled 1–64K sequences from the combined foreground and background, thus likely to be more informative in the given cell type.

To test the validity of motif prediction, we performed the following experiment on the MyoD ChIP-seq dataset: we randomly sampled 1–64K sequences from the combined foreground and background, thus likely to be more informative in the given cell type. We chose background as the sites that are shared by a set of well-studied cell types and the foreground sites are the sites that do not overlap with the foreground. The predicted motifs are consistently optimized to highlight the distinction between foreground and background, thus likely to be more informative in the given cell type.

We found many key TFs which do not overlap with the foreground. The predicted motifs are consistently optimized to highlight the distinction between foreground and background, thus likely to be more informative in the given cell type.

For various different cell types, we found motifs for E2A (annotated as TCFE2A), Runx family, Zic2 motifs in nervous system, and MyoD motifs in skeletal muscle. A recent ENCODE study (Z.Yao et al., 2012) used profiles. To predict motifs in de novo experimental data, we used an automated approach based on discriminative analysis. The predicted motifs are consistently optimized to highlight the distinction between foreground and background, thus likely to be more informative in the given cell type.

For various different cell types, we found motifs for E2A (annotated as TCFE2A), Runx family, Zic2 motifs in nervous system, and MyoD motifs in skeletal muscle. A recent ENCODE study (Z.Yao et al., 2012) used profiles. To predict motifs in de novo experimental data, we used an automated approach based on discriminative analysis. The predicted motifs are consistently optimized to highlight the distinction between foreground and background, thus likely to be more informative in the given cell type.
flanking preference of CAGCTG E-box of these factors (Fig. 4).

### 3.3 Application to cell type specific accessible sites

Discriminative motif analysis can be applied to any high-throughput sequence datasets besides ChIP-Seq data. We used this method to identify key TFs that are involved in regulation of cell type-specific chromatin remodeling using DNaseI hypersensitivity data. We collected 211 DNaseI hypersensitivity datasets from the ENCODE Web site. Combining highly similar ones yielded 77 profiles. We defined cell type-specific accessible sites as the sites that are shared by more than 5 profiles. To predict motifs in each set of cell type specific sites, we chose background as the random sampling of the cell type-specific sites in all cell types that do not overlap with the foreground. The predicted motifs for a set of well-studied cell types are shown in Figure 5 (full results in Supplementary Table S4). We found many key TFs that are known to regulate the given cell type. For example, we found motifs for Oct4 (annotated as Pou2f2), Sox2 and GC-rich motifs that mimic KLF4 (annotated as MZF1 and ASCL2) in Nt2d1, an embryonic cell line. All these factors are well known to be markers of cell pluripotency. We found motifs for IRF1 in B cells, a key factor for immune response. In various lymphocyte cells, we found motifs for E2A (annotated as TCFE2A), Runx and ETS family TFs (annotated as SPIB, ELF5, SFPI1), all of which are critical immune system regulators. For various differential epithelial cells in kidney, colon, lung, breast, pancreas and prostate, FOX family motifs are dominant and motifs for HNF family are enriched in kidney and colon. Similarly, we found significant enrichment of various Homeobox, NeuroD and Zic2 motifs in nervous system and MyoD motifs in skeletal muscle.

### 3.4 Motif significance and sample size

We have shown that our method can discover biologically relevant motifs in a wide range of biological samples and application settings. Here, we also give evidence that the $z$-value calculated by our software is a true indication of a motif's statistical significance, and that the method is robust to variation in sample size and motif enrichment level. To quantify motif significance, we use the $z$-value statistic from the logistic regression model as the 'motif score.' To test its validity, we performed the following experiment on the MyoD ChIP-seq dataset: we randomly sampled 1–64K sequences from the combined foreground and background datasets and then randomly permuted the class labels within each sample. We repeated the permutation 5 times. The $z$-values for all enumerated 6mers in each permutation are approximately normally distributed, as shown by quantile–quantile plots (Supplementary Fig. S5), indicating an accurate reflection of true statistical significance. To determine how the z-scores of enriched motifs change with sample size, we plotted the distribution of z-values for all 6mers using samples from 1 to 64K, and highlight CAGCTG, which is identified as the most significant 6mer using all the data. CAGCTG is consistently the most significant motif for each sample size (Fig. 6A), and the motif score is linear with the square root of the sample size (Supplementary Fig. S5B), in accord with the central limit theorem. We also tested how the motif scores correlate with motif enrichment level. For each sampling with size from 1 to 64K, we randomly kept 20, 40 to 100% of the original foreground samples and replaced the remaining foreground sequences with background sequences while keeping...
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
- Relative Entropy for evaluation/comparison
- Greedy, MEME and Gibbs for discovery

Still room for improvement. E.g., ChIP-seq and Comparative