CSE 427
Fall 2017
Motifs: Representation & Discovery
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)
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

![Diagram of the lac Operon and its control elements](image)

- **DNA:**
  - *lacI* → CAP → P → +1 → O → *lacZ* → *lacY* → *lacA* → 3'

- **mRNA:**
  - AUG → AUG → AUG

- **Gene Expression Levels:**
  - **High (constitutive) level of expression**
  - **Low (basal) level of expression**

- **Conditions:**
  - **Low glucose, Lactose available**
  - **High glucose, Lactose unavailable**

- **Proteins:**
  - cAMP Activator Protein
  - RNA Polymerase
  - *lacI* repressor
1965 Nobel Prize
Physiology or Medicine

François Jacob, Jacques Monod, André Lwoff
The sea urchin *Strongylocentrotus purpuratus*
Sea Urchin - Endo16
if CY & CB1
else
    
    i1 = 1
    i1 = 0.5

    i2 = i1 \cdot UI(t)

if R
else
    
    i3 = \frac{CB2(t)}{1 < k < 2}

if P & CG1 & CB2
else
    
    i4 = 2
    i4 = 0

if UI(t)>\text{threshold} & R & i4 \neq 0
else
    
    i5 = 1
    i5 = 0

    i6 = i4 \cdot (i2+i3)

if i5=0
else
    
    i7 = \text{OTX(t)}
    i7 = 0

    i8 = i6 + i7

if (F or E or DC) & Z
else
    
    i9 = 1
    i9 = 0

if i9=1
else
    
    i10 = 0
    i10 = i8

if (CG2 & CG3 & CG4)
else
    
    i11 = 2
    i11 = 1

    i12 = i11 \cdot i10
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 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

(A)

(B)

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
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

Figure 7-19 Molecular Biology of the Cell 5/e © Garland Science 2008
MyoD

http://www.rcsb.org/pdb/explore/jmol.do?structureId=1MDY&bionumber=1
We understand some Protein/DNA interactions
But the overall DNA binding “code” still defies prediction
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
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 TAxxyzT
– 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>

score = $10 \log_2$ foreground:background odds ratio, rounded
**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>T</th>
<th>C</th>
<th>G</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>
<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>
<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>
<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>
<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>T</th>
<th>C</th>
<th>G</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>
<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>
<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>
<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>
<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>T</th>
<th>C</th>
<th>G</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>
<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>
<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>
<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>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
= -91
\]

---

Scanning for TATA

See also slide 60
TATA Scan at 2 genes

LacI

LacZ
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 \( S = B_1B_2...B_6 \):

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

Assumes independence
Neyman-Pearson

Given a sample $x_1, x_2, \ldots, x_n$, from a distribution $f(...|\Theta)$ 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]
Another WMM example

8 Sequences:

<table>
<thead>
<tr>
<th>Sequence</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 0.25</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></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Log-Likelihood Ratio:

\[ \log_2 \frac{f_{x_i|i}}{f_{x_i}} = \frac{1}{4} \] (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

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

<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>-∞</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</td>
<td>-∞</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>-1.58</td>
<td>1.42</td>
<td>-∞</td>
<td></td>
</tr>
</tbody>
</table>

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 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 (wrt background);

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

Expected score difference: $H(P||Q) + H(Q||P)$
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).

\[ -H(Q||P) = -6.8 \]

\[ H(P||Q) = 5.0 \]
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>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LLR</strong></td>
<td><strong>Col 1</strong></td>
</tr>
<tr>
<td>A</td>
<td>1.32</td>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>-1</td>
</tr>
<tr>
<td><strong>RelEnt</strong></td>
<td><strong>0.7</strong></td>
</tr>
<tr>
<td><strong>LLR</strong></td>
<td><strong>Col 1</strong></td>
</tr>
<tr>
<td>A</td>
<td>0.74</td>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>-1.58</td>
</tr>
<tr>
<td><strong>RelEnt</strong></td>
<td><strong>0.51</strong></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 ($k^{th}$ 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 $k$-tuple
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 ($\binom{n}{k}$ sets)
- $(n+1)^k$ in total; 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):

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
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

xATAyz

CATGACTAGCATAATCCGAT
TATAATTTCCCAGGGATAGCA
TACAAATTAGGACCATAAGATGCGC

xATAAz

CATGACTAGCATAATCCGAT
TATAATTTCCCAGGGATAGCA
TACAAATTAGGACCATAAGATGCGC

TAATAAT

CATGACTAGCATAATCCGAT
TATAATTTCCCAGGGATAGCA
TACAAATTAGGACCATAAGATGCGC
Expectation Step
(where are the motif instances?)

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

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

\[ \sum_{j} \hat{Y}_{i,j} = 1 \]

Recall slide 33
Maximization Step
(what is the motif?)

Find θ maximizing expected log likelihood:

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

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

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

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

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

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

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

From E-Step
M-Step (cont.)

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

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

\begin{align*}
s_1 & : \text{ACGGATT}\ldots \\
\cdots \\
s_k & : \text{GC}\ldots\text{TCGGAC} \\
\hat{Y}_{1,1} & : \text{ACGG} \\
\hat{Y}_{1,2} & : \text{CGGA} \\
\hat{Y}_{1,3} & : \text{GGAT} \\
\vdots & : \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>Gene</th>
<th>Length</th>
<th>Sequence</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigma-37</td>
<td>223</td>
<td>IIDLTYIQNK SQTGTDILGILGQHSLR LQRKAVKRLR</td>
<td>240</td>
</tr>
<tr>
<td>SpoIIIC</td>
<td>94</td>
<td>RFGLDQKEKK TQREIAKELGGRSRYVSR IEKRALMKMF</td>
<td>111</td>
</tr>
<tr>
<td>NahR</td>
<td>22</td>
<td>VVFQQLVQDR RVSITAENLLGTQPAVSN ALKRLRTSLQ</td>
<td>39</td>
</tr>
<tr>
<td>Antennapedia</td>
<td>326</td>
<td>FHFNNRTLRR RRIEIAHALCLTERQIKI WFQNNMKRKW</td>
<td>343</td>
</tr>
<tr>
<td>NtrC (Brady.)</td>
<td>449</td>
<td>LTAALAAATRG NQIRAADLLGLNRNTLRK KIRLDDLQVY</td>
<td>466</td>
</tr>
<tr>
<td>DicA</td>
<td>22</td>
<td>IRYRRNLKIH TQRSLAKILLHVSQ WERGDEPTG</td>
<td>39</td>
</tr>
<tr>
<td>MerD</td>
<td>5</td>
<td>MNAY TVSRLDLADGVHVRD YLLRGDLPRV</td>
<td>22</td>
</tr>
<tr>
<td>Fis</td>
<td>73</td>
<td>LDMVQMYTRG NQTRAALMGGINRGLRKLKKYGMN</td>
<td>90</td>
</tr>
<tr>
<td>MAT al</td>
<td>99</td>
<td>FRRKQSLNSK EKEEAVKCGITPLQVRV WFINKRMRSK</td>
<td>116</td>
</tr>
<tr>
<td>Lambda cII</td>
<td>25</td>
<td>SALLNIKAML GTEKTAEVGVDKQISRF WKRDWIPKFS</td>
<td>42</td>
</tr>
<tr>
<td>Crp (CAP)</td>
<td>169</td>
<td>THPDGQMQIKI TQRONGIVGGCSRETQVRG ILKMLDQNL</td>
<td>186</td>
</tr>
<tr>
<td>Lambda Cro</td>
<td>15</td>
<td>ITLKDAMRF GQTATKDLGYQSAINK AIAHARGKFL</td>
<td>32</td>
</tr>
<tr>
<td>P22 Cro</td>
<td>12</td>
<td>YKKDVIDHFG TQRVAKALGISHDAAVSQ WKĐEVPQD</td>
<td>29</td>
</tr>
<tr>
<td>AraC</td>
<td>196</td>
<td>ISDHLLASNF DIASVAQHVLSPSRLSH LFRQQLGISV</td>
<td>213</td>
</tr>
<tr>
<td>Fnr</td>
<td>196</td>
<td>FSQPRFRLTM TRGDIGNYLGLTVEILSR LLGFRQKSGM</td>
<td>213</td>
</tr>
<tr>
<td>HtpR</td>
<td>252</td>
<td>ARWLDENKNS TLQELAUDRGVSAEVQRE LEKnamklknL</td>
<td>269</td>
</tr>
<tr>
<td>NtrC (K.a.)</td>
<td>444</td>
<td>LTTLARHTQG HKQAEALGILGWRNTLDR KIKELGME</td>
<td>461</td>
</tr>
<tr>
<td>CytR</td>
<td>11</td>
<td>MAAKQCTAAK TMMKVALQKVA2TSTVSR ALMNPDKVSQ</td>
<td>28</td>
</tr>
<tr>
<td>DeoR</td>
<td>23</td>
<td>LQELKRDZKL HLKDAALLGVESEMTTDD DLNNSAPVPV</td>
<td>40</td>
</tr>
<tr>
<td>GalR</td>
<td>3</td>
<td>MAA TIKDVARLAVGVSATVSR VINNSPKASE</td>
<td>20</td>
</tr>
<tr>
<td>LacI</td>
<td>5</td>
<td>MKPV TLYQDASEAVGGSYQTVSR VVNXASHVSA</td>
<td>22</td>
</tr>
<tr>
<td>TetR</td>
<td>26</td>
<td>LLNEVGIEGL TTRKLAQKLGVEQPTLYW HVKNKRALLD</td>
<td>43</td>
</tr>
<tr>
<td>TrpR</td>
<td>67</td>
<td>IVEELLHGHM SQRELKNEGIAIGAIITR GSNNSLKAAPV</td>
<td>84</td>
</tr>
<tr>
<td>NifA</td>
<td>495</td>
<td>LIALEAKAGW VQAARLLGLMTPTQVAY RQIMIDTMPP</td>
<td>512</td>
</tr>
<tr>
<td>SpoIIIG</td>
<td>205</td>
<td>RFGLVFGEER KQDKVADMMQISQYISR LEKRIIKRLR</td>
<td>222</td>
</tr>
<tr>
<td>Pin</td>
<td>160</td>
<td>QAGRLIAAGT TPRQKVAMYDGVVSTLYK TFPAGDK</td>
<td>177</td>
</tr>
<tr>
<td>PurR</td>
<td>3</td>
<td>MAA TIKDVARLAVGSVATVSR VVKNKRALLD</td>
<td>20</td>
</tr>
<tr>
<td>EbgR</td>
<td>3</td>
<td>MAA TLLKDAIEAGVSLATVSR VLNNDDPTLNV</td>
<td>20</td>
</tr>
<tr>
<td>LexA</td>
<td>27</td>
<td>DIHISATGQPP TRAEIAQRLGFRSPNAAE EHLKALAKG</td>
<td>44</td>
</tr>
<tr>
<td>P22 cI</td>
<td>25</td>
<td>SSILNRIAIR GQKVQDALGNESQISR WKGDFIPKMG</td>
<td>42</td>
</tr>
</tbody>
</table>

**Note:**ライフサイエンスデータを提供する。
<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>9</td>
<td>71</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>71</td>
<td>256</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>9</td>
<td>53</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>9</td>
<td>278</td>
<td>63</td>
<td>278</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>170</td>
</tr>
<tr>
<td>His</td>
<td>240</td>
<td>9</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>125</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>240</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>9</td>
<td>248</td>
<td>9</td>
<td>168</td>
<td>89</td>
<td>9</td>
<td>89</td>
<td>89</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>9</td>
<td>819</td>
<td>63</td>
<td>387</td>
<td>63</td>
<td>9</td>
<td>819</td>
<td>9</td>
</tr>
<tr>
<td>Gly</td>
<td>151</td>
<td>9</td>
<td>56</td>
<td>9</td>
<td>151</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>1141</td>
<td>9</td>
<td>151</td>
<td>9</td>
<td>56</td>
<td>9</td>
<td>56</td>
<td>9</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>210</td>
<td>210</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</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>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>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>427</td>
<td>9</td>
<td>61</td>
<td>61</td>
<td>427</td>
<td>9</td>
<td>61</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>198</td>
<td>9</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>9</td>
<td>262</td>
<td>262</td>
<td>9</td>
<td>9</td>
<td>136</td>
<td>136</td>
<td>262</td>
<td>9</td>
<td>262</td>
<td>136</td>
<td>9</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>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>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>9</td>
<td>366</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>366</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
How to Average

An old problem:

\[ n \text{ random variables:} \]

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 \( \vec{x}^{(1)}, \vec{x}^{(2)}, \ldots, \vec{x}^{(n)} \sim P(\vec{x}) \) and average:

\[ E(f(\vec{x})) \approx \frac{1}{n} \sum_{i=1}^{n} f(\vec{x}^{(i)}) \]
Markov Chain Monte Carlo (MCMC)

- *Independent* sampling also often hard, but *not required* for expectation
- **MCMC** \( \vec{X}_{t+1} \sim P(\vec{X}_{t+1} \mid \vec{X}_t) \) 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**
  
  for \( t = 1 \) to \( \infty \)
    
    for \( i = 1 \) to \( k \) do:
      
      \[
x_{t+1,i} \sim P(x_{t+1,i} \mid x_{t+1,1}, x_{t+1,2}, \ldots, x_{t+1,i-1}, x_{t,i+1}, \ldots, x_{t,k})
      \]
**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 MEME 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
“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}, Yutao 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
Overall “Accuracy”

- $ Greed
- * Gibbs
- ^ EM

[Graph showing overall accuracy for various methods such as AlignACE, ANN-Spec, Consensus, GLAM, Improbizer, MEME, MEME3, MITRA, MotifSampler, Oligo/dyad-analysis, QuickScore, SeSiMCMC, Weeder, YMF, with metrics like nSn, nPPV, nPC, nCC, sSn, sPPV, sASP]
- $nTP$ 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:** $xSn = \frac{xTP}{xTP + xFN}$, and
- **Positive Predictive Value:** $xPPV = \frac{xTP}{xTP + xFP}$.

**Specificity:** $nSP = \frac{nTN}{nTN + nFP}$.

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

- $nPC = \frac{nTP}{nTP + nFN + nFP}$.
- $sASP = \frac{(sSn + sPPV)}{2}$.

![Formula](attachment:image.png)

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. \( \leq 1 \) 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
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
ChIP-seq

http://res.illumina.com/images/technology/chip_seq_assay_lg.gif
TF Binding Site Motifs
From ChIPseq

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
Method Outline

**Discriminative model** – foreground/background

Logistic regression: \( x = \text{motif count}, \ p = \text{pr(site)} \)

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

IUPAC patterns – e.g., “R” = A or G

seed/extend/perturb

Z-scores
Method Outline

1. Enumerate Nmers
2. Candidates filtering
3. Candidates counting
4. Seed selection
5. Seed motif refinement
6. Mask motif match
7. Logistic regression
8. Perturbation
9. Extension

Flow:
- Input sequences
- Find next motif
- Repeat
- Not improved
- Seed motif refinement
- Improved
Perturb

Extend
3 RESULTS

Candidate enumeration, evaluation and bootstrap validation can be performed de novo. To find the annotated motif of the ChIP-ed sequence from the flanking regions, randomly chosen from the peak, and with the same width as the peak. We then predicted up to five enriched regions, and with the same width as the peak. We then predicted up to five enriched regions, and with the same width as the peak. We then predicted up to five enriched regions, and with the same width as the peak.

The refinement step can be subject to over-fitting, as a small perturbation may not be meaningful. To improve robustness, we perform the refinement to be more aggressive or conservative (see Section 3 for details). The refinement step first, as we think it is more important to determine the full-length signature of the motif. In the extension step, the maximum number of decisions to standardize/simplify the analysis and be-

The masking step is performed by requiring either an increase of total foreground counts or a decrease of total background counts, so the number of candidates tested is 2 to the power of the length of the motif.

To assess the performance of our method for motif prediction, we compared our results to DREME, which was run on the same sets of foreground and background sequences under the default setting. We claimed success in finding the annotated motif if it exceeded in the datasets where motifRG and/or DREME significantly failed in finding the right motifs in 78% of datasets. In comparison, DREME found annotated motif in 116 datasets, almost 20% more than our method. We failed to find the motifs generally have low AUC, which suggests that the failure of the algorithm may be due to the lack of the TF motif database using Tomtom (Etter et al., 2011; Newburger and Bulyk, 2009). If no exact matches were found, we used the motif of a homolog; e.g. we tested it on 207 ENCODE ChIP-seq datasets collected from two sources, HAIB_TFBS by HudsonAlpha and SYDH_TFBS by Yale and UCD (see Supplementary Fig. S2). This dataset covers 82 unique TFs and 25 cell types with different characteristics: the number of peaks varies from a few to 100 to 1000 nucleotides (Supplementary Table S2). The ranges from 0.40 to 0.66 and median peak width varies from 200 to 1000 nucleotides (Supplementary Fig. S2).

We claimed success in finding the annotated motif if it failed to find the motifs generally have low AUC, which suggests that the failure of the algorithm may be due to the lack of the TF motif database using Tomtom (Etter et al., 2011; Newburger and Bulyk, 2009). If no exact matches were found, we used the motif of a homolog; e.g. we tested it on 207 ENCODE ChIP-seq datasets collected from two sources, HAIB_TFBS by HudsonAlpha and SYDH_TFBS by Yale and UCD (see Supplementary Fig. S2). This dataset covers 82 unique TFs and 25 cell types with different characteristics: the number of peaks varies from a few to 100 to 1000 nucleotides (Supplementary Table S2). The ranges from 0.40 to 0.66 and median peak width varies from 200 to 1000 nucleotides (Supplementary Fig. S2).

To improve the motif prediction, we randomly sample the whole sequence dataset (including positive and negative sequences) with replacement for a few times (default 5 times), calculate the z-values for the new and the original motif for each sampling and compute the z-test on two sets of z-values. Accept the new motif if the difference of means is statistically significant. We claim they have no real effect on the outcome. If the number of bootstraps we performed is small, we found the variance estimate is reasonably accurate and informative to guide the refinement step. The variance estimate is reasonably accurate and informative to guide the refinement step.
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 5/C20 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. A recent ENCODE study (Neph et al., 2012) used motifs in curated databases or de novo predicted motifs to scan accessible regions and compute enrichment in the given cell type. We offer a more direct alternative by combining motif prediction and discriminative analysis. The predicted motifs are consequently optimized to highlight the distinction between foreground and background, thus likely to be more informative in this setting.

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. S5A), 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...
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 CAGCTG patterns. 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. A recent ENCODE study (Neph et al., 2012) used motifs in curated databases or de novo predicted motifs to scan accessible regions and compute enrichment in the given cell type. We offer a more direct alternative by combining motif prediction and discriminative analysis. The predicted motifs are consequently optimized to highlight the distinction between foreground and background, thus likely to be more informative in this setting.

**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...
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 /C20 /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. A recent ENCODE study (Neph et al., 2012) used motifs in curated databases or de novo predicted motifs to scan accessible regions and compute enrichment in the given cell type. We offer a more direct alternative by combining motif prediction and discriminative analysis. The predicted motifs are consequently optimized to highlight the distinction between foreground and background, thus likely to be more informative in this setting.

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. 6 A), and the motif score is linear with the square root of the sample size (Supplementary Fig. S5 B), 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
- Greedy, MEME and Gibbs for discovery

Still room for improvement. E.g., ChIP-seq and Comparative genomics (cross-species comparison) are very promising.