CSEP 527
Spring 2016
5 – Motifs: Representation & Discovery
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 \textit{expressed} means that it:

- is \textit{transcribed} from DNA to RNA
- the mRNA is \textit{processed} in various ways
- is \textit{exported} from the nucleus (eukaryotes)
- is \textit{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
The *lac* Operon and its Control Elements

- **DNA**:
  - *lacI*
  - CAP
  - *P*
  - O
  - *lacZ*
  - AUG
  - *lacY*
  - AUG
  - *lacA*
  - AUG

- **mRNA**:
  - AUG
  - AUG
  - AUG

- **RNA**:
  - AUG
  - AUG
  - AUG

**Gene Expression**

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

**Regulation**

- **cAMP Activator Protein**
- **RNA Polymerase**
- **High (constitutive) level of expression**

- **lacI repressor**
- **X**
- **X**
- **X**

**Expression Levels**

- **Low (basal) level of expression**
  - CAP
  - P
  - O
1965 Nobel Prize
Physiology or Medicine

François Jacob, Jacques Monod, André Lwoff
The sea urchin *Strongylocentrotus purpuratus*
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-
In the groove

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

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

(A) recognition helix

(B) NH₂ COOH
H-T-H Dimers

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

Homo-/hetero-dimers and combinatorial control

Figure 7-19 Molecular Biology of the Cell 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?
Wildcards like R,Y? (\{A,G\}, \{C,T\}, resp.)
**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>

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

\[
\begin{array}{cccccc}
A & -36 & 19 & 1 & 12 & 10 & -46 \\
C & -15 & -36 & -8 & -9 & -3 & -31 \\
G & -13 & -46 & -6 & -7 & -9 & -46 \\
T & 17 & -31 & 8 & -9 & -6 & 19 \\
\end{array}
\]

\[= -90 \]

\[
\begin{array}{cccccc}
A & -36 & 19 & 1 & 12 & 10 & -46 \\
C & -15 & -36 & -8 & -9 & -3 & -31 \\
G & -13 & -46 & -6 & -7 & -9 & -46 \\
T & 17 & -31 & 8 & -9 & -6 & 19 \\
\end{array}
\]

\[= 85 \]

\[
\begin{array}{cccccc}
A & -36 & 19 & 1 & 12 & 10 & -46 \\
C & -15 & -36 & -8 & -9 & -3 & -31 \\
G & -13 & -46 & -6 & -7 & -9 & -46 \\
T & 17 & -31 & 8 & -9 & -6 & 19 \\
\end{array}
\]

\[= -91 \]

Scanning for TATA

See also slide 60
TATA Scan at 2 genes

Score

LacI

LacZ

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

Might as well look at likelihood ratio:

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

(or log likelihood ratio)
Score Distribution
(Simulated)

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

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

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

Answer: like coin flips or dice rolls, count frequencies per position. (Possible HW?)
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></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:

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

(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

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

<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>-∞</td>
<td>-∞</td>
<td>-∞</td>
</tr>
<tr>
<td>C</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>1.42</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>-∞</td>
<td>3</td>
<td>-∞</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

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

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

On average, foreground model scores > background by 11.8 bits (score difference of 118 on 10x scale used in examples above).
For a WMM:

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

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

Proof: exercise

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

<table>
<thead>
<tr>
<th></th>
<th>Freq.</th>
<th>Col 1</th>
<th>Col 2</th>
<th>Col 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.625</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0.25</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>

### LLR

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

### Uniform

**RelEnt**

|   | 0.7 | 2   | 2   | 4.7 |

### Non-uniform

**RelEnt**

|   | 0.51| 1.42| 3   | 4.93|
Pseudocounts

Are the $-\infty$’s a problem?

Certainly 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\textsuperscript{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 \textit{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 $s_1, s_2, \ldots, s_k$ (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 $s_1, s_2, \ldots, s_k$ ($n^k$ such tuples)
Compute relative entropy of each
Pick best
Brute Force, II

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

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

xATAyz

CATGACTAGCATAATCCGAT
TATAATTTCCCGAGGGGATAGCA
TACAATAGGACCATAAGAATGCGC

xATAAz

CATGACTAGCATAATCCGAT
TATAATTTCCCGAGGGGATAGCA
TACAATAGGACCATAAGAATGCGC

TATATAT

CATGACTAGCATAATCCGAT
TATAATTTCCCGAGGGGATAGCA
TACAATAGGACCATAAGAATGCGC
Expectation Step
(where are the motif instances?)

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

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

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

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

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

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

\[ E = 0 \cdot P(0) + 1 \cdot P(1) \]
Maximization Step

(what is the motif?)

Find $\theta$ maximizing expected log likelihood:

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

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

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

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

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

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

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

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.
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>Protein</th>
<th>Accession 1</th>
<th>Accession 2</th>
<th>Accession 3</th>
<th>Accession 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigma-37</td>
<td>IIDLTYIQNK</td>
<td>SQKETGDILGISQMHVSR</td>
<td>LQRKAVKKLR</td>
<td>240 A25944</td>
</tr>
<tr>
<td>SpoIIIC</td>
<td>RFGDLKKEK</td>
<td>TQREIAKELGISRSYVSR</td>
<td>IKEKRALMKMF</td>
<td>111 A28627</td>
</tr>
<tr>
<td>NahR</td>
<td>VVFQNQLVDR</td>
<td>RVSITAENGLTQPASVRN</td>
<td>ALKRLRTSLQ</td>
<td>39 A32837</td>
</tr>
<tr>
<td>Antennapedia</td>
<td>FHFNYRTRR</td>
<td>RRIEIAHALCITQREIKI</td>
<td>WFQNRMRMKW</td>
<td>343 A23450</td>
</tr>
<tr>
<td>NtrC (Brady.)</td>
<td>LTAALAATRG</td>
<td>NQIRAADLGLNRTLQR</td>
<td>KIRDLDIASVY</td>
<td>466 B26499</td>
</tr>
<tr>
<td>DicA</td>
<td>IRYRKNKLNH</td>
<td>TQRSLAKLHSHVSQSV</td>
<td>WERGDSEPTG</td>
<td>39 B24328 (BVECD)</td>
</tr>
<tr>
<td>MerD</td>
<td>5 MNAY</td>
<td>TVSRLADAVSVHTRVD</td>
<td>YLRRGLRNPV</td>
<td>22 C29010</td>
</tr>
<tr>
<td>Fis</td>
<td>LDMVMQYTRG</td>
<td>NQTRAALMGSGRITLQR</td>
<td>KLKGYMN</td>
<td>90 A32142 (DNECS)</td>
</tr>
<tr>
<td>MAT al</td>
<td>FRRKQSLNSK</td>
<td>EEVEAKCACGIPTLQVRV</td>
<td>WFINKRMRSK</td>
<td>116 A90983 (JEBY1)</td>
</tr>
<tr>
<td>Lambda cII</td>
<td>SALLNKIAML</td>
<td>GTEKTAEAVGDKQISER</td>
<td>WKREDWIPKFS</td>
<td>42 A03579 (QCBP2L)</td>
</tr>
<tr>
<td>Crp (CAP)</td>
<td>TPHDMQIHIK</td>
<td>TQREIQGTVGCSRETQG</td>
<td>ILKMLSDQNL</td>
<td>186 A03553 (QRECC)</td>
</tr>
<tr>
<td>Lambda Cro</td>
<td>TTLKYAMRF</td>
<td>GQGTKADUGVYQSAIN</td>
<td>AHAHGRKFL</td>
<td>32 A03577 (RCBPL)</td>
</tr>
<tr>
<td>P22 Cro</td>
<td>YKKDVIDHFG</td>
<td>TQRVALAKGLISDAVASQ</td>
<td>WKKEVIPEKA</td>
<td>29 A03554 (RGBP22)</td>
</tr>
<tr>
<td>AraC</td>
<td>ISDHLADSNF</td>
<td>DIASAQVHCLSPRSRLSH</td>
<td>LFRQQLGISV</td>
<td>213 A03554 (RGECA)</td>
</tr>
<tr>
<td>Fnr</td>
<td>FSPREFRLTM</td>
<td>TRGDIGNYLGLTETLISR</td>
<td>LLGRFQKSGM</td>
<td>213 A03552 (RGECF)</td>
</tr>
<tr>
<td>HtpR</td>
<td>ARWLDENDNKS</td>
<td>TLQELADRYVSAERVQ</td>
<td>LEKNAMKLR</td>
<td>269 A00700 (RGECH)</td>
</tr>
<tr>
<td>NtrC (K.a.)</td>
<td>LTTALRHTQG</td>
<td>HKQAEARLLGWRRNTLRT</td>
<td>KLKELGME</td>
<td>461 A03564 (RGKBCP)</td>
</tr>
<tr>
<td>CytR</td>
<td>MKKAKQTERA</td>
<td>TMKDVAALKVASTVSR</td>
<td>ALMNPDKVQ</td>
<td>28 A24963 (RPECCT)</td>
</tr>
<tr>
<td>DeoR</td>
<td>LQELKRSKDL</td>
<td>HLKDAALLGVSEMTT</td>
<td>DLNNHSAVPV</td>
<td>40 A24076 (RPECDO)</td>
</tr>
<tr>
<td>GalR</td>
<td>MA TIKDCVARAGSVATSVR</td>
<td>VINNSPKASE</td>
<td>20 A03559 (RPECQ)</td>
<td></td>
</tr>
<tr>
<td>LacI</td>
<td>MKPV TLYDDAYAGVSYQTQVSR</td>
<td>VNVQASHVSA</td>
<td>22 A03558 (RPECL)</td>
<td></td>
</tr>
<tr>
<td>TetR</td>
<td>LLNEVGS IG</td>
<td>TTRKLAQKLGVEQPTLYW</td>
<td>HVKNKRALLD</td>
<td>43 A03576 (RPECTN)</td>
</tr>
<tr>
<td>TrpR</td>
<td>67 IVEELLGEM SQRELKINLGAIGATITR</td>
<td>GSNSLKAAPV</td>
<td>84 A03568 (RPECW)</td>
<td></td>
</tr>
<tr>
<td>NifA</td>
<td>LIAALEKAGW</td>
<td>VQAACKIALLLGMPRQVAY</td>
<td>RIQIMDTMP</td>
<td>512 S02513</td>
</tr>
<tr>
<td>SpoIIG</td>
<td>205 RFGLVGEEEK</td>
<td>TQKDVADMMGISOQYISR</td>
<td>LEKRIKLR</td>
<td>222 S07337</td>
</tr>
<tr>
<td>Pin</td>
<td>160 QAGRLIAAGT</td>
<td>PRQKVAYYDVGVSTLYK</td>
<td>TFPAGDK</td>
<td>177 S07958</td>
</tr>
<tr>
<td>PurR</td>
<td>3 MA TIKDCVARANVSTTVSH</td>
<td>VINKTRFVAE</td>
<td>20 S08477</td>
<td></td>
</tr>
<tr>
<td>EbgR</td>
<td>3 MA TLKDAIIEAGVSLATVSR</td>
<td>VLNDPRTLNV</td>
<td>20 S09205</td>
<td></td>
</tr>
<tr>
<td>LexA</td>
<td>27 DHISQTGMPP</td>
<td>TRAEIAQRLLGFRSNNAE</td>
<td>EHLKALARKG</td>
<td>44 S11945</td>
</tr>
<tr>
<td>P22 cI</td>
<td>25 SSILNRIAIR</td>
<td>GQRKVADALGINESQISR</td>
<td>WKGDIFPKMG</td>
<td>42 B25867 (Z1BPC2)</td>
</tr>
</tbody>
</table>

*************** ***
<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>224</td>
<td>9</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>170</td>
<td></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>240</td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>168</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>168</td>
<td>89</td>
<td>9</td>
<td>89</td>
<td>9</td>
<td>248</td>
<td>9</td>
<td>168</td>
<td>89</td>
<td>9</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>117</td>
<td>9</td>
<td>117</td>
<td>117</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>819</td>
<td>63</td>
<td>387</td>
<td>63</td>
<td>9</td>
<td>819</td>
<td></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>1141</td>
<td>9</td>
<td>151</td>
<td>9</td>
<td>56</td>
<td>9</td>
<td>9</td>
<td>56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ala</td>
<td>9</td>
<td>9</td>
<td>112</td>
<td>43</td>
<td>181</td>
<td>901</td>
<td>43</td>
<td>181</td>
<td>215</td>
<td>9</td>
<td>43</td>
<td>9</td>
<td>43</td>
<td>181</td>
<td>112</td>
<td>9</td>
<td>43</td>
<td></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></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>210</td>
<td>210</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>295</td>
<td>581</td>
<td>295</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>58</td>
<td>107</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>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>9</td>
<td>121</td>
<td>9</td>
<td>149</td>
<td>9</td>
<td>93</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>
<td></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></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>9</td>
<td>104</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>198</td>
<td>198</td>
<td>9</td>
<td>104</td>
<td>9</td>
<td>9</td>
<td>198</td>
<td>9</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>9</td>
<td>9</td>
<td>136</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>262</td>
<td>262</td>
<td>9</td>
<td>262</td>
<td>9</td>
<td>136</td>
<td>136</td>
<td>9</td>
<td>262</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>108</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>366</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>366</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Position in site**
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 random variables:

Joint distribution (p.d.f.):

Some function:

Want Expected Value:

\[ x_1, x_2, \ldots, x_k \]

\[ P(x_1, x_2, \ldots, x_k) \]

\[ f(x_1, x_2, \ldots, x_k) \]

\[ E(f(x_1, x_2, \ldots, x_k)) \]
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 \( \vec{X}_{t+1} \sim P(\vec{X}_{t+1} \mid \vec{X}_t) \) 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
  
  \[
  \text{for } t = 1 \text{ to } \infty \\
  \quad \text{for } i = 1 \text{ to } k \text{ do :} \\
  \quad \quad x_{t+1,i} \sim P(x_{t+1,i} \mid x_{t+1,1}, x_{t+1,2}, \ldots, x_{t+1,i-1}, x_{t,i+1}, \ldots, x_{t,k})
  \]
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
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 Tompa¹,², Nan Li¹, Timothy L Bailey³, George M Church⁴, Bart De Moor⁵, Eleazar Eskin⁶, Alexander V Favorov⁷,⁸, Martin C Frith⁹, Yutao Fu⁹, W James Kent¹⁰, Vsevolod J Makeev⁷,⁸, Andrei A Mironov⁷,¹¹, William Stafford Noble¹,², Giulio Pavesi¹², Graziano Pesole¹³, Mireille Régnier¹⁴, Nicolas Simonis¹⁵, Saurabh Sinha¹⁶, Gert Thijs⁵, Jacques van Helden¹⁵, Mathias Vandenbogaert¹⁴, Zhiping Weng⁹, Christopher Workman¹⁷, Chun Ye¹⁸ & Zhou Zhu⁴
Methodology

13 tools

Real ‘motifs’ (Transfac)

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

‘Real’, ‘generic’, ‘Markov’

Expert users, top prediction only

“Blind” – sort of
Greedy Gibbs EM
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
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 – $x = \text{motif count}$

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

IUPAC patterns – e.g., “R” = A or G

seed/extend/perturb

Z-scores
Method Outline

Enumerate Nmers → Candidates filtering → Candidates counting → Seed selection

Find next motif → Mask motif match

Seed motif → Logistic regression

Not improved → Seed refinement

Repeat → extension → perturbation → Repeat

Input sequences
extend the motif as long as needed. The allowed candidates is a lot smaller. Using this strategy we can afford to filter the candidates by requiring either an increase of total foreground turbed patterns must be compatible with the initial seed motif, and we each sampling and compute the 5 times), calculate the z-values for the new and the original motif for positive and negative sequences) with replacement for a few times (default improvement: randomly sample the whole sequence dataset (including motif databases Jaspar ([et al., 2008]) and Uniprobe ([et al., 2011]) and Newburger and Bulyk, 2009) ranges from 0.40 to 0.66 and median peak width varies from different characteristics: the number of peaks varies from a few to several hundred (Yale and UCD (see Supplementary Table S1)). We made a motif dis-
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 \( C^{20} \) 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 A), 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

### Table A2

<table>
<thead>
<tr>
<th>Consensus</th>
<th>scores</th>
<th>ratio</th>
<th>fg.frac</th>
<th>bg.frac</th>
<th>logo</th>
<th>DB match</th>
<th>DB Evalue</th>
<th>DB logo</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNVCAGATGGNN</td>
<td>69</td>
<td>5.12</td>
<td>0.68</td>
<td>0.14</td>
<td></td>
<td>Tcfe2a</td>
<td>1.5e-09</td>
<td></td>
</tr>
<tr>
<td>NNCAGCTGSNN</td>
<td>42</td>
<td>1.57</td>
<td>0.92</td>
<td>0.73</td>
<td></td>
<td>Ascl2</td>
<td>9.2e-08</td>
<td></td>
</tr>
<tr>
<td>NNATCAATNN</td>
<td>22</td>
<td>3.34</td>
<td>0.13</td>
<td>0.04</td>
<td></td>
<td>Pbx1</td>
<td>9.1e-05</td>
<td></td>
</tr>
<tr>
<td>NNRCACTGNN</td>
<td>-68</td>
<td>0.35</td>
<td>0.47</td>
<td>0.90</td>
<td></td>
<td>Ascl2</td>
<td>2.4e-08</td>
<td></td>
</tr>
<tr>
<td>NNTNASTCANN</td>
<td>-21</td>
<td>0.51</td>
<td>0.17</td>
<td>0.27</td>
<td></td>
<td>AP1</td>
<td>1.4e-05</td>
<td></td>
</tr>
<tr>
<td>NNRCCACANN</td>
<td>-19</td>
<td>0.66</td>
<td>0.29</td>
<td>0.41</td>
<td></td>
<td>RUNX1</td>
<td>7.1e-07</td>
<td></td>
</tr>
<tr>
<td>NNNHNAATADNN</td>
<td>-18</td>
<td>0.53</td>
<td>0.12</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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. 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 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, pancrease 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

**Fig. 4.** Predicting the specificity of similar bHLH TFs. (A1) Predicted PWMs for MyoD (left) and NeuroD2 (right), (A2) Discriminative motifs based on direct comparison of MyoD sites (foreground) with NeuroD2 sites (background), suggesting MyoD and NeuroD2 preferred ebox and cofactor motifs. (B1 and B2) Comparison of MyoD and MSC. Similar to A1 and A2. (B3) Gel shift demonstrating that MSC/E12 heterodimer binds strongly at CCAGCTGG and MyoD/E12 binds weakly, whereas GCAGCTGC binds strongly to both MyoD/E12 and MSC/E12. MSC homodimer also binds stronger at CCAGCTGG than GCAGCTGC.
change significantly with number of bootstraps performed or on bootstraps. The distribution of standard deviation does not.

computed the standard deviation of scores for each variant based on replicates, sample sizes and motif enrichment levels. We com-

CAGCTG example shown earlier in text and tested all extension wide peaks, but keeping peaks with reasonable width is import-

proportion of true foreground. (E) When motif enrichment is low, motif portion of true foreground.

all CAGCTG to different sample sizes. (D) The standard deviation of motif scores for ground in shuffled foreground, (C) Distribution of all

ent symbols correspond to different sample sizes. (C) Distribution of all

-mer scores. (A) Score distribution of for all 6mers. CAGCTG ebox, marked by

Fig. 6.

bootstrap iterations, total sample size 500 and only 20% of true

strapping mean and variance. Upper panel: 160 bootstrap iterations, total motif scores. The 95% confidence intervals are plotted based on boot-

Y

Y

Y

Y

Y

motifs. For example, direct TFBS should contain a clear DNaseI

Further, downstream analysis can be used to validate predicted motifs. For example, direct TFBS should contain a clear DNaseI

good background for comparison with adjustment to other bias.

some generic features common to most TFBS and can serve as a

given cell type can be viewed as the union of all TFBS that are

test for potential biases as we find them, each time making the

associated motifs are truly specific to the given TFs, we can iteratively
tend to have enrichment of AP1 sites. To determine if the asso-

other GC-rich motifs, and TFs with most of sites in distal regions

moter regions are likely to be associated with ETS, SP1 and

category. For example, TFs that bind predominantly in pro-

profiles, we identified some common motifs for TFs in a certain

affects binding of TFs. By examination of a large set of ChIP-Seq

cess for this method is proper choice of background, which might

whether the predicted motifs are truly involved in the biological

Discriminative motif analysis is a powerful tool to address

unclear if these TFs collectively determine the accessibility of the

tend to colocalize on a common set of accessible regions. It is

ChIP-Seq and DNaseI hypersensitivity studies suggest that TFs

patterns to be enriched in a given dataset. In addition, many

some positioning signals and so forth. Genomic sequences in

promoters, which usually show enrichment of TFBS, also con-

tain different characteristics from other parts of genome such as

.random, which presents complicated higher order structure such

is important to understand the nature of these motifs and why

effectively use the motif prediction results to guide further study, it

besides scalability concerns, they are likely to produce large num-

datasets. High-throughput datasets present different challenges:

signal to noise ratio due to lack of power from small input

The main challenge of traditional motif discovery is to increase

stronger motifs, as measured by the PWM scores, particularly

Bailey and Machanick, 2012

e et al.

2012

Fig. 6 D).
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.