Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab
Module II: Gene expression analysis and networks

- **Computational foundations:**
  - Unsupervised Learning: Expectation Maximization
  - Supervised learning: generative/discriminative models
  - Read mapping, significance testing, splice graphs
  - Folding: DP self-alignment, Context Free grammars

- **Biological applications:**
  - L6: Expression Analysis
  - L7: Transcript structure and RNA world
  - L8: Regulatory motifs and biological networks
  - L9: Epigenomics & 3D chromatin interactions
HMMs Review + Learning

Supervised + Unsupervised
Viterbi Training (max, top path)
Baum Welch Training (EM, all paths)
Combinations of marks encode epigenomic state

- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

Enhancers: • H3K4me1 • H3K27ac • DNase

Promoters: • H3K4me3 • H3K9ac • DNase

Transcribed: • H3K36me3 • H3K79me2 • H4K20me1

Repressed: • H3K9me3 • H3K27me3 • DNAmethyl

100s of known modifications, many new still emerging
Systematic mapping using ChIP-, Bisulfite-, DNase-Seq
ChIP-Seq Histone Modifications: What the raw data looks like

- Each sequence tag is 30 base pairs long
- Tags are mapped to unique positions in the ~3 billion base reference genome
- Number of reads depends on sequencing depth. Typically on the order of 10 million mapped reads.
Hashing vs. Burrows Wheeler Transform

Today: How does the BW transform actually work?

Reference genome (> 3 gigabases) → Short read

1. Extract seeds
2. Multi-seed hashing

- Position 1: ACTG CCGT AAAC TAAT
- Position 2: CTGC CGTA AACT AATG
- Position N: 

3. Index seed pairs

- Seed index (tens of gigabytes)

4. Look up each pair of seeds in index

5. Hits identify positions in genome where spaced seed pair is found

6. Confirm hits by checking “***” positions

BWT

- Burrows-Wheeler transform and indexing
  - Look up ‘suffixes’ of read
  - Hits identify positions in genome where read is found
  - Convert each hit back to genome location

Report alignment to user
Summarize multiple marks into chromatin states

ChromHMM: multi-variate hidden Markov model
HMM nomenclature for this course

- Vector $x$ = Sequence of observations
- Vector $\pi$ = Hidden path (sequence of hidden states)
- Transition matrix $A = a_{kl}$ = probability of $k \rightarrow l$ state transition
- Emission vector $E = e_k(x_i)$ = prob. of observing $x_i$ from state $k$
- Bayes’s rule: Use $P(x_i|\pi_i=k)$ to estimate $P(\pi_i=k|x_i)$

Transitions: $a_{kl} = P(\pi_i=l|\pi_{i-1}=k)$
Transition probability from state $k$ to state $l$

Emissions: $e_k(x_i) = P(x_i|p_i=k)$
Emission probability of symbol $x_i$ from state $k$
HMMs for Genome Annotation/Parsing: Viterbi Algo.

**Score annotation (path \( \pi \)) given a seq (obs \( x \)): B-P**

\[
p = \frac{1}{2} \times P(A \mid Bgdn) P(Bgdn \mid Bgdn) P(Bgdn, \ldots) P(T \mid Bgdn) P(Bgdn \mid Bgdn) \ldots P(A \mid Bgdn)
\]

\[
= \frac{1}{2} \times (0.1)^1 \times (0.4)^2 \times (0.99)^3 \times (0.01)^1 \times (0.95)^3 \times (0.05)^1
\]

\[
= 6.4 \times 10^{-10}
\]

Much less likely, due to high cost of transitions

**The Viterbi Algorithm**

- **Input:** \( x = x_1 \ldots x_N \)
- **Initialization:**
  \[ V_0(0) = 1, \quad V_k(0) = 0, \quad \text{for all} \ k > 0 \]
- **Iteration:**
  \[ V_k(i) = e_k(x_i) \times \max_j a_{jk} V_{j}(i-1) \]
- **Termination:**
  \[ P(x, \pi^*) = \max_k V_k(N) \]
- **Traceback:**
  Follow max pointers back
  Similar to aligning states to seq
- **In practice:**
  Use log scores for computation
- **Running time and space:**
  Time: \( O(K^2N) \)
  Space: \( O(KN) \)

- **Calculate**
  \[ V_k(i) = e_k(x_i) \times \max_j (V_{j}(i-1) \times a_{jk}) \]
  current max
  this emission
  max ending
  in state \( j \) at step \( i \)
  Transition
  from state \( j \)
  all possible previous states \( j \)
HMMs over all paths: Posterior Decoding

Calculate most probable label at a single position

\[ P(\text{Label}_i = B | x) \]

- Calculate most probable label, \( L^*_i \), at each position \( i \)
- Do this for all \( N \) positions gives us \( \{L^*_1, L^*_2, L^*_3, \ldots, L^*_N\} \)
- How much information have we observed? Three settings:
  - Observed nothing: Use prior information
  - Observed only character at position \( i \): Prior + emission probability
  - Observed entire sequence: Posterior decoding

### The Backward Algorithm

\[ b_k(i) = \sum_{k} b_k(i+1) \times a_{kl} \times e_l(x_{i+1}) \]

**Initialization:**
\[ b_k(n) = a_{k0}, \text{ for all } k \]

**Iteration:**
\[ b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1) \]

**Termination:**
\[ P(x) = \sum_k a_{k0} e_1(x_1) b_k(1) \]

**In practice:**
Sum of log scores is difficult
\[ \rightarrow \text{approximate } \exp(1+p+q) \]
\[ \rightarrow \text{scaling of probabilities} \]

**Running time and space:**
- Time: \( O(K^2N) \)
- Space: \( O(K) \)

### Putting it all together: Posterior decoding

**Putting it all together:**
- \( P(k) = P(\pi_i = k | x) = \frac{f_k(i)^* b_k(i)}{P(x)} \)
  - Probability that \( i^{th} \) state is \( k \), given all emissions \( x \)
- **Posterior decoding**
  - Find the most likely state at position \( i \) over all possible hidden paths given the observed sequence \( x \)
  - \( \pi^*_i = \arg\max_k P(\pi_i = k | x) \)
  - **Posterior decoding ‘path’ \( \pi^*_i \)**
    - For classification, more informative than Viterbi path \( \pi^* \)
    - More refined measure of “which hidden states” generated \( x \)
    - However, it may give an invalid sequence of states
    - Not all \( j \rightarrow k \) transitions may be possible

### Calculate total end probability recursively

**Assume we know \( b_i \) for the next time step \((i+1)\)**

\[ b_k(i) = \sum_l \left( \frac{e_l(x_{i+1})}{\text{current max}} \times a_{kl} \times b_l(i+1) \right) \]

**Sum over all possible next states**
HMM Foundations, Parsing, Decoding, Learning

1. HMM basics, evaluation, parsing, posterior decoding
   - Observations, Models, Bayes’ rule, Bayesian inference
   - Markov Chains and Hidden Markov Models
   - Calculating joint probability of one (seq, parse) \( P(x, \pi) \)
   - Viterbi algorithm: Find best parse \( \pi^* = \arg \max_{\pi} P(x, \pi) \)
   - Forward algorithm: Find total \( P(x) \), sum over all paths
   - Posterior Decoding: Most likely state \( \pi_i \) (over all paths)

2. Learning (ML training, Baum-Welch, Viterbi training)
   - Supervised: Find \( e_i(.) \) and \( a_{ij} \) given labeled sequence
   - Unsupervised: given only \( x \rightarrow \) annotation + params

3. Increasing the ‘state’ space / adding memory
   - Finding GC-rich regions vs. finding CpG islands
   - Gene structures GENSCAN, chromatin ChromHMM
### The six algorithmic settings for HMMs

**One path**

1. **Scoring x, one path**
   \[ P(x, \pi) \]
   Prob of a path, emissions

2. **Scoring x, all paths**
   \[ P(x) = \sum_{\pi} P(x, \pi) \]
   Prob of emissions, over all paths

3. **Viterbi decoding**
   \[ \pi^* = \arg\max_{\pi} P(x, \pi) \]
   Most likely path

4. **Posterior decoding**
   \[ \pi^\wedge = \{ \pi_i \mid \pi_i = \arg\max_k \sum_{\pi} P(\pi_i=k|x) \} \]
   Path containing the most likely state at any time point.

**All paths**

5. **Supervised learning, given π**
   \[ \Lambda^* = \arg\max_{\Lambda} P(x, \pi|\Lambda) \]

6. **Unsupervised learning**
   \[ \Lambda^* = \arg\max_{\Lambda} \sum_{\pi} P(x, \pi|\Lambda) \]
   Viterbi training, best path

   Baum-Welch training, over all paths
Learning: How to train an HMM

Transition probabilities
e.g. \( P(P_{i+1}|B_i) \) – the probability of entering a pathogenicity island from background DNA

Emission probabilities
i.e. the nucleotide frequencies for background DNA and pathogenicity islands
If we have a sequence that has islands marked, we can simply count:

\[
\begin{align*}
A: & \quad \frac{1}{5} \\
T: & \quad 0 \\
G: & \quad \frac{2}{5} \\
C: & \quad \frac{2}{5}
\end{align*}
\]

\[P(S|P)P(S|B)P(L_{i+1}|L_i)\]

![Diagram showing a sequence with labels and transition probabilities.](image)
How do we know how to count?

Unlabeled Data

L: P P P P P P P P P P
    start B B B B B B B B B B End

S: G C A A A A T G C

P(L_{i+1}|L_i)

<table>
<thead>
<tr>
<th>B_{i+1}</th>
<th>P_{i+1}</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_i</td>
<td>P_i</td>
<td></td>
</tr>
<tr>
<td>Start</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P(S|B)

A:
T:
G:
C:

P(S|P)

A:
T:
G:
C:
An idea:
1. Imagine we start with some parameters
2. We *could* calculate the most likely path, $P^*$, given those parameters and $S$
3. We *could* then use $P^*$ to update our parameters by maximum likelihood
4. And iterate (to convergence)
Simple case: Viterbi Training

**Initialization:**
Pick the best-guess for model parameters
(or arbitrary)

**Iteration:**
1. Perform Viterbi, to find $\pi^*$
2. Calculate $A_{kl}$, $E_k(b)$ according to $\pi^* +$ pseudocounts
3. Calculate the new parameters $a_{kl}$, $e_k(b)$

Until convergence

**Notes:**
– Convergence to local maximum guaranteed. Why?
– Does not maximize $P(x \mid \theta)$
– In general, worse performance than Baum-Welch
### One path

1. **Scoring** \( x, \text{one path} \)
   
   \[ P(x, \pi) \]
   
   Prob of a path, emissions

2. **Scoring** \( x, \text{all paths} \)
   
   \[ P(x) = \sum_\pi P(x, \pi) \]
   
   Prob of emissions, over all paths

3. **Viterbi decoding**
   
   \[ \pi^* = \arg\max_\pi P(x, \pi) \]
   
   Most likely path

4. **Posterior decoding**
   
   \[ \pi^\Lambda = \{\pi_i | \pi_i = \arg\max_k \sum_\pi P(\pi_i = k | x)\} \]
   
   Path containing the most likely state at any time point.

### All paths

5. **Supervised learning, given** \( \pi \)
   
   \[ \Lambda^* = \arg\max_\Lambda P(x, \pi | \Lambda) \]

6. **Unsupervised learning**
   
   \[ \Lambda^* = \arg\max_\Lambda \max_\pi P(x, \pi | \Lambda) \]
   
   Viterbi training, best path

#### Learning

5. **Unsupervised learning.**
   
   \[ \Lambda^* = \arg\max_\Lambda \sum_\pi P(x, \pi | \Lambda) \]
   
   Baum-Welch training, over all paths
Expectation Maximization (EM)

The basic idea is always the same:
1. Use model to estimate missing data (E step)
2. Use estimate to update model (M step)
3. Repeat until convergence

EM is a general approach for learning models (ML estimation) when there is “missing data”

1. Initialize parameters randomly
2. E Step Estimate expected probability of hidden labels, Q, given current (latest) parameters and observed (unchanging) sequence
   \[ Q = P(Labels|S, params^{t-1}) \]
3. M Step Choose new maximum likelihood parameters over probability distribution Q, given current probabilistic label assignments
   \[ params^t = \arg \max_{params} E_Q \left[ \log P(S, labels | params^{t-1}) \right] \]
4. Iterate

\[ P(S|Model) \text{ guaranteed to increase each iteration} \]
Starting with our best guess of a model M, parameters $\theta$:

Given $x = x_1 \ldots x_N$

for which the true $\pi = \pi_1 \ldots \pi_N$ is unknown,

We can get to a provably more likely parameter set $\theta$

Principle: EXPECTATION MAXIMIZATION

1. **Expected** annotations all-paths parse w/ current params *(E step)*
2. **Max**-likelihood params $A_{kl}$, $E_k$ using this all-paths parse *(M step)*
3. Repeat 1 & 2, until convergence
Max likelihood parameters | all-paths parse (M step)

(Sum over all emissions from k, at any time step i)

\[ E_k(b) = \frac{1}{P(x)} \sum_{i \mid x_i = b} \left[ f_k(i) b_k(i) \right] \]

(Sum over all \( k \rightarrow l \) transitions, at any time step i)

\[ A_{kl} = \sum_i P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \frac{\sum_i f_k(i) a_{kl} e_{i}(x_{i+1}) b_l(i+1)}{P(x \mid \theta)} \]
Max likelihood parameters | all-paths parse (M step)

Derivation:
To estimate $A_{kl}$:
At each position $i$:

Find probability transition $k \rightarrow l$ is used:

$$P(\pi_i = k, \pi_{i+1} = l \mid x) = \frac{1}{P(x)} \times P(\pi_i = k, \pi_{i+1} = l, x_1 \ldots x_N) = \frac{Q}{P(x)}$$

where $Q = P(x_1 \ldots x_i, \pi_i = k, \pi_{i+1} = l, x_{i+1} \ldots x_N) = P(\pi_i = k, x_{i+1} \ldots x_N) = P(\pi_i = k, x_{i+1} x_{i+2} \ldots x_N) f_k(i)$

So:

$$P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \frac{f_k(i) a_{kl} e_l(x_{i+1}) b_{l(i+1)}}{P(x \mid \theta)}$$

(For one such transition, at time step $i \rightarrow i+1$)
1. Scoring x, one path
\[ P(x, \pi) \]  
Prob of a path, emissions

2. Scoring x, all paths
\[ P(x) = \sum_{\pi} P(x, \pi) \]
Prob of emissions, over all paths

3. Viterbi decoding
\[ \pi^* = \arg\max_{\pi} P(x, \pi) \]
Most likely path

4. Posterior decoding
\[ \pi^\wedge = \{ \pi_i | \pi_i = \arg\max_k \sum_{\pi} P(\pi_i = k | x) \} \]
Path containing the most likely state at any time point.

5. Supervised learning, given \( \pi \)
\[ \Lambda^* = \arg\max_{\Lambda} P(x, \pi | \Lambda) \]

6. Unsupervised learning
\[ \Lambda^* = \arg\max_{\Lambda} \sum_{\pi} P(x, \pi | \Lambda) \]
Viterbi training, best path

Unsupervised learning
\[ \Lambda^* = \arg\max_{\Lambda} \sum_{\pi} P(x, \pi | \Lambda) \]
Baum-Welch training, over all paths
Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data

2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. Discovery and characterization of chromatin states
   - HMM Foundations, Generating, Parsing, Decoding, Learning
   - Chromatin state characterization: Functional/positional enrichment

4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence

5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
Detecting GC-rich regions: HMM architecture

Hidden (model)

Background

GC-rich promoter

Observed (world)

\[
\begin{align*}
P(A|\text{Bgd}) &= 0.25 \\
P(T|\text{Bgd}) &= 0.25 \\
P(C|\text{Bgd}) &= 0.25 \\
P(G|\text{Bgd}) &= 0.25 \\
P(A|\text{Prom}) &= 0.10 \\
P(T|\text{Prom}) &= 0.10 \\
P(C|\text{Prom}) &= 0.40 \\
P(G|\text{Prom}) &= 0.40
\end{align*}
\]

Slide credit: Serafim Batzoglou
Example: Detecting GC-rich regions: motivation

Model genome as two states:
- P: promoter
- B: background

Model different nucleotide compositions
- Background:
  \[ P(A) = P(T) = P(C) = P(G) = 25\% \]
- Promoters
  \[ P(A) = P(T) = 10\% \]
  \[ P(G) = P(C) = 40\% \]

Note: generative model, \( P(A|\text{promoter}) \) etc
Then: reverse probabilities using Baye’s rule

Model length distribution:
- Promoter: 20 bp
- Non-promoter: 100 bp

Modified from: Khuu et al PNAS 2007; NFI: Nuclear Factor I; ZDRs: Z-DNA Forming Regions
HMMs are used broadly for genome annotation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Topology / Transitions</td>
<td>2 states, different nucleotide composition</td>
<td>2 states, different conservation levels</td>
<td>2 states, different tri-nucleotide composition</td>
<td>2 states, different evolutionary signatures</td>
<td>~20 states, different composition/conservation, specific structure</td>
<td>40 states, different chromatin mark combinations</td>
</tr>
<tr>
<td>Hidden States / Annotation</td>
<td>GC-rich / AT-rich</td>
<td>Conserved / non-conserved</td>
<td>Coding exon / non-coding (intron or intergenic)</td>
<td>Coding exon / non-coding (intron or intergenic)</td>
<td>First/last/middle coding exon, UTRs, intron.4/3, intergenic, *(+/- strand)</td>
<td>Enhancer / promoter / transcribed / repressed / repetitive</td>
</tr>
<tr>
<td>Emissions / Observations</td>
<td>Nucleotides</td>
<td>Level of conservation</td>
<td>Triplets of nucleotides</td>
<td>Nucleotide triplets, conservation levels</td>
<td>Codons, nucleotides, splice sites, start/stop codons</td>
<td>Vector of chromatin mark frequencies</td>
</tr>
</tbody>
</table>
### Examples of HMMs for genome annotation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2 states, different conservation levels</td>
<td>2 states, different nucleotide composition</td>
<td>8 states, 4 each +/-, different transition probabilities</td>
<td>2 states, different tri-nucleotide composition</td>
<td>~20 states, different nucleotide composition/conservation, specific structure</td>
<td>40 states, different chromatin mark combinations</td>
<td>2 states, different evolutionary signatures</td>
</tr>
<tr>
<td>Conserved / non-conserved</td>
<td>GC-rich / AT-rich</td>
<td>CpG-rich / CpG-poor</td>
<td>Coding exon / non-coding (intron or intergenic)</td>
<td>First/last/middle coding exon, UTRs, intron 1/2/3, intergenic, *(+- strand)</td>
<td>Enhancer / promoter / transcribed / repressed / repetitive</td>
<td>Coding exon / non-coding (intron or intergenic)</td>
</tr>
<tr>
<td>Level of conservation</td>
<td>Nucleotides</td>
<td>Di-Nucleotides</td>
<td>Triplets of nucleotides</td>
<td>Codons, nucleotides, splice sites, start/stop codons</td>
<td>Vector of chromatin mark frequencies</td>
<td>64x64 matrix of codon substitution frequencies</td>
</tr>
</tbody>
</table>

L2:alignmnt L4:HMMs1 L5:HMMs2 L5:HMMs2 L5:HMMs2 L8:Epignmcs L17:CompG
Genome Annotation

Genome Sequence

> HSCKIIBE. Human gene for casein kinase II subunit beta (EC 2.7.1.37).

RNA

Transcription

Translation

Protein
Eukaryotic Gene Structure

complete mRNA

coding segment

ATG

TGA

start codon

donor site

acceptor site

donor site

acceptor site

stop codon

Model of joint distribution $P(Y,X) = P(\text{Labels},\text{Seq})$

For gene prediction, we are given $X$…

How do we select a $Y$ efficiently?
HMM architecture matters: Protein-coding genes

- Gene vs. Intergenic
- Start & Stop in/out
- UTR: 5’ and 3’ end
- Exons, Introns
- Remembering frame
  - E0,E1,E2
  - I0,I1,I2
- Sequence patterns to transition between states:
  - ATG, TAG, Acceptor/Donor, TATA, AATAAA
### Examples of HMMs for genome annotation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2 states, different conservation levels</td>
<td>2 states, different nucleotide composition</td>
<td>8 states, 4 each +/-, different transition probabilities</td>
<td>2 states, different trinucleotide composition</td>
<td>~20 states, different composition/conservation, specific structure</td>
<td>40 states, different chromatin mark combinations</td>
<td>2 states, different evolutionary signatures</td>
</tr>
<tr>
<td>Conserved / non-conserved</td>
<td>GC-rich / AT-rich</td>
<td>CpG-rich / CpG-poor</td>
<td>Coding exon / non-coding (intron or intergenic)</td>
<td>First/last/middle coding exon, UTRs, intron1/2/3, intergenic, *(+- strand)</td>
<td>Enhancer / promoter / transcribed / repressed / repetitive</td>
<td>Coding exon / non-coding (intron or intergenic)</td>
</tr>
<tr>
<td>Level of conservation</td>
<td>Nucleotides</td>
<td>Di-Nucleotides</td>
<td>Triplets of nucleotides</td>
<td>Codons, nucleotides, splice sites, start/stop codons</td>
<td>Vector of chromatin mark frequencies</td>
<td>64x64 matrix of codon substitution frequencies</td>
</tr>
<tr>
<td>L2:alignmnt</td>
<td>L4:HMMs1</td>
<td>L5:HMMs2</td>
<td>L5:HMMs2</td>
<td>L5:HMMs2</td>
<td>L8:Epignmcs</td>
<td>L17:CompG</td>
</tr>
</tbody>
</table>
Summarize multiple marks into chromatin states

ChromHMM: multi-variate hidden Markov model
Multivariate HMM for Chromatin States

- **Observed chromatin marks**: Called based on a Poisson distribution.
- **Most likely Hidden State**: 1 → 2 → 3 → 4 → 6 → 6 → 6 → 6 → 5 → 5 → 5
- **High Probability Chromatin Marks in State**:
  - **State 1**: K4me1 (0.8), K27ac (0.8)
  - **State 2**: K4me3 (0.9), K4me3 (0.8)
  - **State 3**: K4me3 (0.9)
  - **State 4**: K4me1 (0.7)
  - **State 5**: None
  - **State 6**: K36me3 (0.9)

- **200bp intervals**
- **All probabilities are learned from the data**

Ernst and Kellis
Nature Biotech 2010
Chromatin State: Emission & Transition Matrices

- Emission matrix $e_k(x_i)$
  - Multi-variate HMM
  - Emits vector of values

- Transition matrix $a_{kl}$
  - Learn spatial relationships
  - No a-priori ‘gene’ structure

Design Choice

• How to model the emission distribution
  – Model the signal directly
  – Locally binarize the data

• For $M$ input marks each state $k$ has a vector of $(p_{k1},...,p_{kM})$ of parameters for independent Bernoulli random variables which determine the emission probability for an observed combination of marks

Data Binarization

• Leads to biologically interpretable models that can be robustly learned

• Let $c_{ij}$ be the number of reads for mark $i$ mapping to bin $j$. $\lambda_i$ be the average number of reads mapping to a bin for modification $i$. The input for feature $i$ becomes ‘1’ if

$$P(X>c_{ij})<10^{-4}$$

where $X$ is a Poisson random variable with mean $\lambda_i$
### Emission Parameter Matrix $e_k(\tilde{X}_i)$

<table>
<thead>
<tr>
<th>State</th>
<th>ACGG34ac</th>
<th>ACGG23ac</th>
<th>ACGG123ac</th>
<th>ACGG1232ac</th>
<th>ACGG12321ac</th>
<th>ACGG123212ac</th>
<th>ACGG1232123ac</th>
<th>ACGG12321232ac</th>
<th>ACGG123212321ac</th>
<th>ACGG1232123212ac</th>
<th>ACGG12321232123ac</th>
<th>ACGG123212321232ac</th>
<th>ACGG1232123212321ac</th>
<th>ACGG12321232123212ac</th>
<th>ACGG123212321232123ac</th>
<th>ACGG1232123212321232ac</th>
<th>ACGG12321232123212321ac</th>
<th>ACGG123212321232123212ac</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.06</td>
<td>3.48</td>
<td>2.06</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
</tr>
<tr>
<td>2</td>
<td>3.48</td>
<td>2.05</td>
<td>3.48</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
</tr>
<tr>
<td>3</td>
<td>2.06</td>
<td>2.05</td>
<td>2.06</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
</tr>
<tr>
<td>4</td>
<td>3.48</td>
<td>2.05</td>
<td>3.48</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
</tr>
<tr>
<td>5</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
</tr>
<tr>
<td>6</td>
<td>2.06</td>
<td>2.05</td>
<td>2.06</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
</tr>
<tr>
<td>7</td>
<td>3.48</td>
<td>2.05</td>
<td>3.48</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
</tr>
<tr>
<td>8</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
</tr>
<tr>
<td>9</td>
<td>2.06</td>
<td>2.05</td>
<td>2.06</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
</tr>
<tr>
<td>10</td>
<td>3.48</td>
<td>2.05</td>
<td>3.48</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
</tr>
<tr>
<td>11</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
<td>3.48</td>
</tr>
<tr>
<td>12</td>
<td>2.06</td>
<td>2.05</td>
<td>2.06</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
<td>2.05</td>
</tr>
</tbody>
</table>

**Key points:**
- **Multi-variate HMM** emits vector of values, not just one value
- **Can emit real values (SegWay)** or binary presence /absence values (ChromHMM)
- **Use to learn mark combinations**

*Ernst and Kellis, Nature Biotech 2010*
Transition matrix $a_{kl}$

- Learns spatial relationships between neighboring states
- Reveals distinct sub-groups of states
- Reveals transitions between different groups
Example Chromatin State Annotation

- Use Baum Welch to learn hidden states and their annotations
- Learned states correspond to known functional elements
- *De novo* discovery of major types of chromatin
Model complexity matches that of genome

- Handful of repressed states capture vast majority of genome
  - Only 1% of genome split in 14 promoter states
- Modeling power well distributed where needed
Apply genome wide to classify chromatin states *de novo*

Now what? Interpret these states biologically
Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   – Overview of epigenomics, Diversity of Chromatin modifications
   – Antibodies, ChIP-Seq, data generation projects, raw data

2. Primary data processing: Read mapping, Peak calling
   – Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   – Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. Discovery and characterization of chromatin states
   – HMM Foundations, Generating, Parsing, Decoding, Learning
   – Chromatin state characterization: Functional/positional enrichment

4. Model complexity: selecting the number of states/marks
   – Selecting the number of states, selecting number of marks
   – Capturing dependencies and state-conditional mark independence

5. Learning chromatin states jointly across multiple cell types
   – Stacking vs. concatenation approach for joint multi-cell type learning
   – Defining activity profiles for linking enhancer regulatory networks
State definitions

State Enrichments

Chromatin mark frequencies for each chromatin state

Genomic and functional enrichments for each state

Promoter states

Transcribed States

Active Interg.

Repetit. Repress.

State definitions → State Enrichments

(see Supplementary Fig. 2 for full emission prob. matrix)
Functional enrichments enable annotation of 51 distinct states
Application of ChromHMM to 41 chromatin marks in CD4+ T-cells (Barski'07, Wang'08)

### a. Chromatin mark frequencies for each chromatin state

<table>
<thead>
<tr>
<th>State</th>
<th>Promoter states</th>
<th>Active Intergenic</th>
<th>Repressed</th>
<th>Repetitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Promoter upstream high expr; Potential enh loop.</td>
<td>Transcribed promoter; highest expr, TSS for active genes</td>
<td>Repressed promoter</td>
<td>Simple repeats (CA)n, (TG)n</td>
</tr>
<tr>
<td>2</td>
<td>Promoter upstream medium expr; Potential enh loop.</td>
<td>Transcribed promoter; highest expr, open chr, TF bind</td>
<td>Transcribed promoter; highest expr, open chr, TF bind</td>
<td>L1/LTR Repeats</td>
</tr>
<tr>
<td>3</td>
<td>Promoter upstream low expr; Potential enh loop.</td>
<td>Transcribed promoter; highest expr, downstream</td>
<td>Transcribed promoter; highest expr, open chr</td>
<td>Satellite Repeat</td>
</tr>
<tr>
<td>4</td>
<td>Transcribed promoter; highest expr, near TSS</td>
<td>Transcribed promoter; highest expr, downstream</td>
<td>Transcribed 5' proximal, med expr; Alu repeats</td>
<td>Satellite Repeat; moderate mapping bias</td>
</tr>
<tr>
<td>5</td>
<td>Transcribed promoter; high expr, near TSS</td>
<td>Transcribed 5' proximal; higher expr, open chr, TF bind</td>
<td>Transcribed less 5' proximal, med expr; open chr</td>
<td>Satellite Repeat; high mapping bias</td>
</tr>
<tr>
<td>6</td>
<td>Transcribed promoter; high expr, downstream</td>
<td>Transcribed 5' proximal, high expr</td>
<td>Transcribed less 5' proximal, lower expr; Alu repeats</td>
<td>Satellite Repeat/rRNA; extreme mapping bias</td>
</tr>
<tr>
<td>7</td>
<td>Transcribed 5' proximal; high expr</td>
<td>Transcribed 5' proximal, higher expr</td>
<td>Enhancer in Transcribed region</td>
<td>Specific Repression</td>
</tr>
<tr>
<td>8</td>
<td>Transcribed 5' proximal; med expr; Alu repeats</td>
<td>Transcribed 5' proximal, med expr</td>
<td>Spliced exons/GC Rich; open chr, TF binding</td>
<td>Specific Repression</td>
</tr>
<tr>
<td>9</td>
<td>Transcribed less 5' proximal, med expr</td>
<td>Transcribed less 5' proximal, med expr; open chr</td>
<td>Spliced exons/GC Rich</td>
<td>Specific Repression</td>
</tr>
<tr>
<td>10</td>
<td>Transcribed less 5' proximal, low expr; Alu repeats</td>
<td>Transcribed less 5' proximal, lower expr; Alu repeats</td>
<td>Spliced exons/GC Rich</td>
<td>Specific Repression</td>
</tr>
<tr>
<td>11</td>
<td>Enhancer in Transcribed region</td>
<td>Enhancer in Transcribed region</td>
<td>Spliced exons/GC Rich</td>
<td>Specific Repression</td>
</tr>
<tr>
<td>12</td>
<td>Spliced exons/GC Rich; open chr, TF binding</td>
<td>Spliced exons/GC Rich</td>
<td>Satellite Repeat</td>
<td>Specific Repression</td>
</tr>
<tr>
<td>13</td>
<td>Spliced exons/GC Rich</td>
<td>Spliced exons/GC Rich</td>
<td>Satellite Repeat</td>
<td>Specific Repression</td>
</tr>
<tr>
<td>14</td>
<td>Alu repeats</td>
<td>Alu repeats</td>
<td>Satellite Repeat</td>
<td>Specific Repression</td>
</tr>
<tr>
<td>15</td>
<td>Transcribed 5' distal; exons</td>
<td>Transcribed 5' distal; exons</td>
<td>Transcribed 5' distal; Alu repeats</td>
<td>Satellite Repeat</td>
</tr>
</tbody>
</table>
Functional properties of discovered chromatin states

State 28: 112-fold ZNF enrich

“The achievement of the repressed state by wild-type KAP1 involves decreased recruitment of RNA polymerase II, reduced levels of histone H3 K9 acteylation and H3K4 methylation, an increase in histone occupancy, enrichment of trimethyl histone H3K9, H3K36, and histone H4K20 …” MCB 2006.

Distinct types of repression:
- Chrom bands / HDAC resp
- Repeat family / composition
States show distinct mCpG, DNase, Tx, Ac profiles

TssA vs. TssBiv: diff. activity, both open, both unmethylated!
Enh vs. ReprPC: diff. activity, both intermediate DNase/Methyl
Tx: Methylated, closed, actively transcribed

⇒ Distinct modes of repression: H3K27me3 vs. DNAme vs. Het
Chromosomal ‘domains’ from chromatin state usage

- State usage → gene density, lamina, cytogenetic bands
- Quies/ZNF/het | gene rich/poor, each active/repressed
Applications to genome annotation

New protein-coding genes

In promoter (short)/low-expr states

Long intergenic non-coding RNAs/lincRNAs

Chromatin signature:

Evolutionary signature:

not protein-coding

Known coding

Evolutionary signature:

not protein-coding

New developmental enhancer regions

Assign candidate functions to intergenic SNPs from genome-wide association studies

Bing Ren, Eddy Rubin
Discovery power for promoters, transcripts

- Significantly outperforms single-marks
- Similar power to supervised learning approach
- CAGE experiments give possible upper bound
Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   – Overview of epigenomics, Diversity of Chromatin modifications
   – Antibodies, ChIP-Seq, data generation projects, raw data

2. Primary data processing: Read mapping, Peak calling
   – Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   – Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. Discovery and characterization of chromatin states
   – HMM Foundations, Generating, Parsing, Decoding, Learning
   – ChromHMM: Multi-variate HMM for chromatin state learning

4. Model complexity: selecting the number of states/marks
   – Capturing dependencies. State-conditional mark independence
   – Selecting the number of states, selecting number of marks

5. Learning chromatin states jointly across multiple cell types
   – Stacking vs. concatenation approach for joint multi-cell type learning
   – Defining activity profiles for linking enhancer regulatory networks
State-conditional mark independence

Do hidden states actually capture dependencies between marks?
Pairwise Expected vs. Observed Mark Co-Occurrence

Each point = one pair of chromatin marks
41x41 pairs plotted
X-axis: F(mark1) * F(mark2)
Y-axis: F(mark1 & mark2)
Diagonal: independence
Off-diag: dependence

 Marks become conditionally independent
 Model captures dependencies

Multi-variate HMM emits entire vector of marks at a time
Model assumes mark independence *conditional* upon state
In fact, it specifically seeks to *capture* these dependencies
Test conditional independence for each state

Promoter states

Transcribed states
Non-independence reveals cases of model violation

Active Intergenic states

- Repetitive states show more dependencies
- Conditional independence does not hold
As more states are added, dependencies captured

- With only 5 states in HMM, not enough power to distinguish different properties
  - Dependencies remain
- As model complexity increases, states learned become more precise
  - Dependencies captured
Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data

2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. Discovery and characterization of chromatin states
   - HMM Foundations, Generating, Parsing, Decoding, Learning
   - ChromHMM: Multi-variate HMM for chromatin state learning

4. Model complexity: selecting the number of states/marks
   - Capturing dependencies. State-conditional mark independence
   - Selecting the number of states, selecting number of marks

5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
Comparison of BIC Score vs. Number of States for Random and Nested Initialization

Step 1: Learn a larger model that captures ‘all’ relevant states
Step 2: Prune down model greedily eliminating least informative states
Step 3: Select arbitrary cutoff based on biological interpretation

Result: a 51-state model that captures most biology in least complexity

- Standard model selection criteria fail due to genome complexity: more states always preferred
- Instead: Start w/complex model, keep informative states, prune redundant states. Pick cutoff
Recovery of 79-state model in random vs. nested initialization

**Random Initialization**
- (states appear & disappear)

**Nested Initialization**
- (states consistently recovered)

**Nested initialization approach:**
- **First pass:** learn models of increasing complexity
- **Second pass:** form nested set of emission parameter initializations by greedily removing states from best BIC model found

**Nested models criteria:**
- Maximize sum of correlation of emission vectors with nested model
- Models learned in parallel
Functional recovery with increasing numbers of states

- Red: Maximum fold functional enrichment for corresponding biological category
- Blue: Percent of that functional category that overlaps regions annotated to this state
- Top plot: Correlation of emission parameter vector for that state to closest state
Chromatin state recovery with increasing numbers of marks

Which states are well-recovered?
Increasing numbers of marks (greedy)

Precisely what mistakes are made?
(for a given subset of 11 ENCODE marks)

State Inferred with subset of marks

State Inferred with all 41 marks

State confusion matrix with 11 ENCODE marks

Recovery of states with increasing number of marks
Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning

Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab
Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data

2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)

3. Discovery and characterization of chromatin states
   - HMM Foundations, Generating, Parsing, Decoding, Learning
   - ChromHMM: Multi-variate HMM for chromatin state learning

4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence

5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks
Epigenomic mapping across 100+ tissues/cell types

**Diverse tissues and cells**

- Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood...)
- Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood...)
- ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench...)

**Diverse epigenomic assays**

- **Histone modifications**
  - H3K4me3, H3K4me1, H3K36me3
  - H3K27me3, H3K9me3, H3K27/9ac
  - +20 more

- **Open chromatin:**
  - DNA accessibility

- **DNA methylation:**
  - WGBS, RRBS, MRE/MeDIP

- **Gene expression**
  - RNA-seq, Exon Arrays
ENCODE: Study nine marks in nine human cell lines

9 marks
- H3K4me1
- H3K4me2
- H3K4me3
- H3K27ac
- H3K9ac
- H3K27me3
- H4K20me1
- H3K36me3
- CTCF
- +WCE
- +RNA

9 human cell types
- HUVEC: Umbilical vein endothelial
- NHEK: Keratinocytes
- GM12878: Lymphoblastoid
- K562: Myelogenous leukemia
- HepG2: Liver carcinoma
- NHLF: Normal human lung fibroblast
- HMEC: Mammary epithelial cell
- HSMM: Skeletal muscle myoblasts
- H1: Embryonic

81 Chromatin Mark Tracks
(2^81 combinations)

How to learn single set of chromatin states?

Brad Bernstein ENCODE Chromatin Group

Ernst et al, Nature 2011
Roadmap 2015: 12 marks x 127 cell types

ENCODE 2019: 40 marks x 834 cell types
Solution 1: Learn independent models and cluster

### Basic approach:

- **a)** Train a k-state model in each cell type independently
- **b)** Cluster models learned independently
- **c)** Merge clusters and re-apply to each cell type

### How to cluster

- **a)** Using emission probability matrix: most similar definitions
- **b)** Using genome annotation: posterior probability decoding

<table>
<thead>
<tr>
<th>State</th>
<th>CTCF</th>
<th>H3K27me3</th>
<th>H3K36me3</th>
<th>H4K20me1</th>
<th>H3K4me1</th>
<th>H3K4me2</th>
<th>H3K4me3</th>
<th>H3K27ac</th>
<th>H3K9ac</th>
<th>WCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.2</td>
<td>72.0</td>
<td>0.2</td>
<td>9.1</td>
<td>47.9</td>
<td>77.8</td>
<td>49.5</td>
<td>1.3</td>
<td>10.2</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>11.9</td>
<td>1.9</td>
<td>6.1</td>
<td>9.0</td>
<td>52.7</td>
<td>93.7</td>
<td>55.0</td>
<td>14.1</td>
<td>44.1</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>16.4</td>
<td>1.5</td>
<td>2.4</td>
<td>5.5</td>
<td>17.0</td>
<td>92.6</td>
<td>99.0</td>
<td>95.7</td>
<td>98.1</td>
<td>1.9</td>
</tr>
<tr>
<td>4</td>
<td>11.4</td>
<td>0.6</td>
<td>14.5</td>
<td>11.3</td>
<td>96.3</td>
<td>99.3</td>
<td>75.1</td>
<td>97.2</td>
<td>85.7</td>
<td>3.7</td>
</tr>
<tr>
<td>5</td>
<td>5.3</td>
<td>0.2</td>
<td>9.5</td>
<td>2.6</td>
<td>88.1</td>
<td>56.8</td>
<td>5.3</td>
<td>84.4</td>
<td>24.3</td>
<td>1.5</td>
</tr>
<tr>
<td>6</td>
<td>6.7</td>
<td>0.9</td>
<td>1.0</td>
<td>3.2</td>
<td>58.3</td>
<td>74.7</td>
<td>8.4</td>
<td>5.8</td>
<td>5.4</td>
<td>0.8</td>
</tr>
<tr>
<td>7</td>
<td>1.6</td>
<td>0.6</td>
<td>1.6</td>
<td>1.3</td>
<td>56.5</td>
<td>2.7</td>
<td>0.4</td>
<td>5.9</td>
<td>1.6</td>
<td>0.6</td>
</tr>
<tr>
<td>8</td>
<td>91.5</td>
<td>1.8</td>
<td>0.9</td>
<td>2.8</td>
<td>6.3</td>
<td>3.3</td>
<td>0.4</td>
<td>0.5</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>4.6</td>
<td>0.3</td>
<td>43.2</td>
<td>43.1</td>
<td>36.5</td>
<td>11.5</td>
<td>1.9</td>
<td>9.1</td>
<td>3.9</td>
<td>1.3</td>
</tr>
<tr>
<td>10</td>
<td>1.2</td>
<td>0.1</td>
<td>47.2</td>
<td>2.7</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>0.3</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>11</td>
<td>0.4</td>
<td>0.1</td>
<td>2.7</td>
<td>1.7</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>12</td>
<td>0.9</td>
<td>26.8</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>13</td>
<td>0.2</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>14</td>
<td>21.9</td>
<td>27.9</td>
<td>19.1</td>
<td>41.0</td>
<td>5.7</td>
<td>4.8</td>
<td>25.9</td>
<td>5.3</td>
<td>13.1</td>
<td>37.5</td>
</tr>
<tr>
<td>15</td>
<td>85.2</td>
<td>85.1</td>
<td>91.5</td>
<td>88.4</td>
<td>76.2</td>
<td>75.9</td>
<td>90.8</td>
<td>72.7</td>
<td>85.0</td>
<td>78.3</td>
</tr>
</tbody>
</table>
Joint learning of states across multiple cell types

Solution 2: Stacking
- Learns each combination of activity as a separate state
- Ex: ES-specific enhancers: enhancer marks in ES, no marks in other cell types

Solution 3: Concatenation
- Requires that profiled marks are the same (or treat as missing data)
- Ensures common state definitions across cell types
Joint learning with different subsets of marks (Solution 3)

Option (a) Treat missing tracks as missing data
- EM framework allows for unspecified data points
- As long as pairwise relationship observed in some cell type

Option (b) Chromatin mark imputation
- Explicitly predict max-likelihood chromatin track for missing data
- Less powerful if ultimate goal is chromatin state learning
ENCODE: Study nine marks in nine human cell lines

<table>
<thead>
<tr>
<th>9 marks</th>
<th>9 human cell types</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me1</td>
<td>HUVEC: Umbilical vein endothelial</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>NHEK: Keratinocytes</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>GM12878: Lymphoblastoid</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>K562: Myelogenous leukemia</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>HepG2: Liver carcinoma</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>NHLF: Normal human lung fibroblast</td>
</tr>
<tr>
<td>H4K20me1</td>
<td>HMEC: Mammary epithelial cell</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>HSMM: Skeletal muscle myoblasts</td>
</tr>
<tr>
<td>CTCF</td>
<td>H1: Embryonic</td>
</tr>
<tr>
<td>+WCE</td>
<td></td>
</tr>
<tr>
<td>+RNA</td>
<td></td>
</tr>
</tbody>
</table>

81 Chromatin Mark Tracks (2^81 combinations)

Concatenation approach:
- Learned jointly across cell types
- State definitions are common
- State locations are dynamic

Brad Bernstein ENCODE Chromatin Group

<table>
<thead>
<tr>
<th>Chromatin States</th>
<th>CTCF</th>
<th>H3K27me3</th>
<th>H3K36me3</th>
<th>H4K20me1</th>
<th>H3K4me1</th>
<th>H3K4me2</th>
<th>H3K4me3</th>
<th>H3K27ac</th>
<th>H3K9ac</th>
<th>WCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>17</td>
<td>93</td>
<td>99</td>
<td>96</td>
<td>98</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>2</td>
<td>6</td>
<td>9</td>
<td>53</td>
<td>94</td>
<td>95</td>
<td>14</td>
<td>44</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>72</td>
<td>0</td>
<td>9</td>
<td>48</td>
<td>78</td>
<td>49</td>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>1</td>
<td>15</td>
<td>11</td>
<td>96</td>
<td>99</td>
<td>75</td>
<td>97</td>
<td>86</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>10</td>
<td>3</td>
<td>88</td>
<td>57</td>
<td>57</td>
<td>84</td>
<td>25</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>58</td>
<td>75</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>56</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>92</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>0</td>
<td>43</td>
<td>43</td>
<td>37</td>
<td>11</td>
<td>2</td>
<td>9</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
<td>47</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>27</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>22</td>
<td>28</td>
<td>19</td>
<td>41</td>
<td>6</td>
<td>5</td>
<td>26</td>
<td>5</td>
<td>13</td>
<td>37</td>
</tr>
<tr>
<td>15</td>
<td>85</td>
<td>85</td>
<td>91</td>
<td>88</td>
<td>76</td>
<td>77</td>
<td>81</td>
<td>73</td>
<td>85</td>
<td>78</td>
</tr>
</tbody>
</table>

Chromatin Mark Observation Frequency (%): (fold)

- (+/-kbp TSS) (Conserved non-exon) DNase (K562) C-Myc (K562) NF-kB (GM12878) Transcript (NHLF)

Candidate state annotation:
- Active Promoter
- Weak Promoter
- Inactive/poised Promoter
- Strong enhancer
- Strong enhancer
- Weak/poised enhancer
- Weak/poised enhancer
- Insulator
- Transcriptional transition
- Transcriptional elongation
- Weak transcribed
- Polycomb-repressed
- Heterochrom.; low signal
- Repetitive/CNV
- Repetitive/CNV

Functional enrichments (fold)
Chromatin states dynamics across nine cell types

- Single annotation track for each cell type
- Summarize cell-type activity at a glance
- Can study 9-cell activity pattern across...
Epigenomic mapping across 100+ tissues/cell types

**Diverse tissues and cells**
- Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood...)
- Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood...)
- ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench...)

**Diverse epigenomic assays**
- Histone modifications
  - H3K4me3, H3K4me1, H3K36me3
  - H3K27me3, H3K9me3, H3K27/9ac
  - +20 more
- Open chromatin:
  - DNA accessibility
- DNA methylation:
  - WGBS, RRBS, MRE/MeDIP
- Gene expression
  - RNA-seq, Exon Arrays

Anshul Kundaje
Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning

- Enhancer-gene linking: Correlation, Hi-C, eQTLs
- TF motif discovery: Enrichment, EM, Gibbs Sampling
- Deep learning convolution CNNs for motif discovery
- Global motif discovery: Comparative Genomics
- Motif Instance Identification: Branch Length Score
- Regulatory region dissection: MPRA, HiDRA

GWAS hit

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab
Link enhancers to their target genes
3 lines of evidence:

**Physical**

Hi-C: Physical proximity in 3D

**Functional**

Enhancer-gene activity correlation

**Genetic**

eQTL evidence: SNP effect on expression

Complementary evidence at physical, functional, genetic level
Correlation-based links of enhancer networks

Regulators $\rightarrow$ Enhancers $\rightarrow$ Target genes
Chromatin state annotations across 127 epigenomes

Reveal epigenomic variability: enh/prom/tx/repr/het

Anshul Kundaje
2.3M enhancer regions $\Leftrightarrow$ only ~200 activity patterns
Introducing multi-cell activity profiles

Gene expression
Chromatin States
Active TF motif enrichment
TF regulator expression
Dip-aligned motif biases

HUVEC
NHEK
GM12878
K562
HepG2
NHLF
HMEC
HSMM
H1

Link enhancers to target genes

ON  Active enhancer  Motif enrichment  TF On  Motif aligned
OFF  Repressed  Motif depletion  TF Off  Flat profile
Activity-based linking of enhancers to target genes

Finding correct target of enhancer in divergently transcribed genes

Compute correlations between gene expression levels and enhancer associated histone modification signals
Visualizing 10,000s predicted enhancer-gene links

• Overlapping regulatory units, both few and many
• Both upstream and downstream elements linked
• Enhancers correlate with sequence constraint
Chromatin dynamics: linking enhancer networks

TFs → enhancers → target genes
Introducing multi-cell activity profiles

<table>
<thead>
<tr>
<th>Gene expression</th>
<th>Chromatin States</th>
<th>Active TF motif enrichment</th>
<th>TF regulator expression</th>
<th>Dip-aligned motif biases</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUVEC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHEK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM12878</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K562</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HepG2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHLF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMEC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSMM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Link TFs to target enhancers
Predict activators vs. repressors

- **ON**
- **OFF**
- **Active enhancer**
- **Repressed**
- **Motif enrichment**
- **Motif depletion**
- **TF On**
- **TF Off**
- **Motif aligned**
- **Flat profile**
Coordinated activity reveals activators/repressors

Enhancer activity

Activity signatures for each TF

- Enhancer networks: Regulator $\rightarrow$ enhancer $\rightarrow$ target gene
Regulatory motifs predicted to drive enhancer modules

- Activator and repressor motifs consistent with tissues

Pouya Kheradpour
Causal motifs supported by dips & enhancer assays

Dip evidence of TF binding (nucleosome displacement)

Enhancer activity halved by single-motif disruption

Motifs bound by TF, contribute to enhancers
Lecture 5: Regulatory circuitry
Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs

TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA
1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix ↔ positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint $(M, Z_{ij})$ distribution
   - Sampling motif positions based on the $Z$ vector
   - More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score $\Rightarrow$ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays
Regulatory motif discovery

• Regulatory motifs
  – Genes are turned on / off in response to changing environments
  – No direct addressing: subroutines (genes) contain sequence tags (motifs)
  – Specialized proteins (transcription factors) recognize these tags

• What makes motif discovery hard?
  – Motifs are short (6-8 bp), sometimes degenerate
  – Can contain any set of nucleotides (no ATG or other rules)
  – Act at variable distances upstream (or downstream) of target gene
The regulatory code: All about regulatory motifs

- **The parts list:** ~20-30k genes
  - Protein-coding genes, RNA genes (tRNA, microRNA, snRNA)
- **The circuitry:** constructs controlling gene usage
  - Enhancers, promoters, splicing, post-transcriptional motifs
- **The regulatory code, complications:**
  - Combinatorial coding of ‘unique tags’
    - Data-centric encoding of addresses
  - Overlaid with ‘memory’ marks
    - Large-scale on/off states
  - Modulation of the large-scale coding
    - Post-transcriptional and post-translational information
- **Today:** discovering motifs in co-regulated promoters and *de novo* motif discovery & target identification
TFs use DNA-binding domains to recognize specific DNA sequences in the genome.

DNA-binding domain of *Engrailed*
Disrupted motif at the heart of FTO obesity locus

**Strongest association with obesity**

C-to-T disruption of AT-rich regulatory motif

**Restoring motif restores thermogenesis**

- **Browning mitochondrial thermogenesis**
  - UCP1
  - PGC1α
  - PRDM16

**THERMOGENIC STIMULI (E.G. COLD)**

- **ARID5B**
- **IRX3**
- **IRX5**

- **AATATT motif**

**LEAN**

- White adipocytes

**OBESE**

- Lipid storage
Regulator structure ↔ recognized motifs

- Proteins ‘feel’ DNA
  - Read chemical properties of bases
  - Do NOT open DNA (no base complementarity)

- 3D Topology dictates specificity
  - Fully constrained positions:  
    - every atom matters
  - “Ambiguous / degenerate” positions
    - loosely contacted

- Other types of recognition
  - MicroRNAs: complementarity
  - Nucleosomes: GC content
  - RNAs: structure/seqn combination
Motifs summarize TF sequence specificity

- Summarize information
- Integrate many positions
- Measure of information
- Distinguish motif vs. motif instance
- Assumptions:
  - Independence
  - Fixed spacing

### Target genes bound by ABF1 regulator

<table>
<thead>
<tr>
<th>Target genes bound by ABF1 regulator</th>
<th>Coordinates</th>
<th>Genome sequence at bound site</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACS1 acetyl CoA synthetase</td>
<td>-491 -479</td>
<td>ATCATTTCTGGACG</td>
</tr>
<tr>
<td>ACS1 acetyl CoA synthetase</td>
<td>-433 -421</td>
<td>ATCATCTCGGACG</td>
</tr>
<tr>
<td>ACS1 acetyl CoA synthetase</td>
<td>-311 -299</td>
<td>ATCATTTGCCACG</td>
</tr>
<tr>
<td>CHA1 catabolic L-serine dehydratase</td>
<td>-280 -254</td>
<td>A</td>
</tr>
<tr>
<td>ENO2 Enolase</td>
<td>-470 -461</td>
<td>ggcgttat</td>
</tr>
<tr>
<td>HMR silencer</td>
<td>-256 -283</td>
<td>ATCAATAC</td>
</tr>
<tr>
<td>LPD1 lipoamide dehydrogenase</td>
<td>-288 -300</td>
<td>gat</td>
</tr>
<tr>
<td>LPD1 lipoamide dehydrogenase</td>
<td>-301 -313</td>
<td>gat</td>
</tr>
<tr>
<td>PGK phosphoglycerate kinase</td>
<td>-523 -496</td>
<td>CAAAACAA</td>
</tr>
<tr>
<td>RPC160 RNA pol III/C 160 kDa subunit</td>
<td>-385 -349</td>
<td>ATCACTATATACG</td>
</tr>
<tr>
<td>RPC40 RNA pol III/C 40 kDa subunit</td>
<td>-137 -116</td>
<td>GTCACTATAAACG</td>
</tr>
<tr>
<td>rpl2 ribosomal protein L2</td>
<td>-185 -167</td>
<td>TAAT</td>
</tr>
<tr>
<td>SPR3 CDC3/10/11/12 family homolog</td>
<td>-315 -303</td>
<td>ATCACTAAATACG</td>
</tr>
<tr>
<td>YPT1 TUB2</td>
<td>-193 -172</td>
<td>CCTAG</td>
</tr>
</tbody>
</table>

### Position Weight Matrix (PWM)

<table>
<thead>
<tr>
<th>Position</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>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>56</td>
<td>4</td>
<td>4</td>
<td>81</td>
<td>4</td>
<td>23</td>
<td>15</td>
<td>27</td>
<td>31</td>
<td>89</td>
<td>23</td>
<td>4</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>32</td>
<td>4</td>
<td>4</td>
<td>12</td>
<td>4</td>
<td>31</td>
<td>23</td>
<td>4</td>
<td>19</td>
<td>23</td>
<td>4</td>
<td>89</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>4</td>
<td>89</td>
<td>4</td>
<td>58</td>
<td>12</td>
<td>23</td>
<td>19</td>
<td>19</td>
<td>23</td>
<td>4</td>
<td>69</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>89</td>
<td>4</td>
<td>4</td>
<td>35</td>
<td>35</td>
<td>39</td>
<td>50</td>
<td>31</td>
<td>23</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

### Motif Logo

#### Consensus

<table>
<thead>
<tr>
<th>R</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>Y</th>
<th>N</th>
<th>N</th>
<th>H</th>
<th>N</th>
<th>N</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>R</th>
</tr>
</thead>
</table>
Motifs are not limited to DNA sequences

- Splicing Signals at the RNA level
  - Splice junctions
  - Exonic Splicing Enhancers (ESE)
  - Exonic Splicing Suppressors (ESS)

- Domains and epitopes at the Protein level
  - Glycosylation sites
  - Kinase targets
  - Targetting signals
  - MHC binding specificities

- Recurring patterns at the physiological level
  - Expression patterns during the cell cycle
  - Heart beat patterns predicting cardiac arrest
    - Final project in previous year, now used in Boston hospitals!
  - Any probabilistic recurring pattern
Approaches to regulatory motif discovery

Region-based motif discovery

- Expectation Maximization (e.g. MEME)
  - Iteratively refine positions / motif profile
- Gibbs Sampling (e.g. AlignACE)
  - Iteratively sample positions / motif profile
- Enumeration with wildcards (e.g. Weeder)
  - Allows global enrichment/background score
- Peak-height correlation (e.g. MatrixREDUCE)
  - Alternative to cutoff-based approach

Genome-wide

- Conservation-based discovery (e.g. MCS)
  - Genome-wide score, up-/down-stream bias

In vitro / trans

- Protein Domains (e.g. PBM, SELEX)
  - In vitro motif identification, seq-/array-based
Experimental factor-centric discovery of motifs

SELEX (Systematic Evolution of Ligands by Exponential Enrichment; Klug & Famulok, 1994).

DIP-Chip (DNA-immunoprecipitation with microarray detection; Liu et al., 2005)

PBMs (Protein binding microarrays; Mukherjee, 2004) Double stranded DNA arrays
Challenges in regulatory genomics

TFs: Homology to TFs/domains
miRNAs: Evolutionary signatures
miRNAs: Experimental cloning

TFs: Mass Spec (difficult)

TFs: Selex, DIP-Chip, Protein-Binding-Microarrays
miRNAs: Evolutionary/structural signatures
siRNAs: Experimental cloning of 5’-ends

TFs/miRNAs: De novo comparative discovery**

Network analysis (upcoming lecture)

Targets

Functional instances

Evolutionary footprints
DNase, TF/Chrom ChIP, Chromatin ‘dips’, MPRA

TFs: Enrichment in co-regulated genes/bound regions **

Regulator

TF/miRNA

Motif

Sequence specificity

* = Covered in today’s lecture
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix ⇔ positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint $(M, Z_{ij})$ distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score $\Rightarrow$ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays
Enrichment-based discovery methods

Given a set of co-regulated/functionally related genes, find common motifs in their promoter regions

- Align the promoters to each other using local alignment
- Use expert knowledge for what motifs should look like
- Find ‘median’ string by enumeration (motif/sample driven)
- Start with conserved blocks in the upstream regions
Starting positions $\Leftrightarrow$ Motif matrix

- given **aligned** sequences $\Rightarrow$ easy to compute profile matrix

**Key idea:** Iterative procedure for estimating both, given uncertainty
(learning problem with hidden variables: the starting positions)

![Shared motif diagram with sequence positions and profile matrix]

<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>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.6</td>
<td>0.1</td>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>G</td>
<td>0.2</td>
<td>0.2</td>
<td>0.6</td>
<td>0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>
### Three options for assigning points, and their parallels across K-means, HMMs, Motifs

<table>
<thead>
<tr>
<th>Update rule</th>
<th>Update assignments (E step) ➔ Update model parameters (M step) ➔ max likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick a best</td>
<td>Assign each point to best label Assign each point to nearest cluster <strong>K-means</strong>: Assign each point to nearest cluster</td>
</tr>
<tr>
<td>Assign each point to all labels, probabilistically</td>
<td><strong>Fuzzy K-means</strong>: Assign to all clusters, weighted by proximity</td>
</tr>
<tr>
<td>Pick one label at random, based on their relative probability</td>
<td>N/A: Assign to a random cluster, sample by proximity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm implementing E step in each of the three settings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expression clustering</strong></td>
</tr>
<tr>
<td>The hidden label is:</td>
</tr>
</tbody>
</table>

- **Expression clustering**: Cluster labels
- **HMM learning**: State path $\pi$
- **Motif discovery**: Motif positions
Basic Iterative Approach

Given: length parameter \( W \), training set of sequences
set initial values for \textit{motif}
do
\[ \rightarrow \text{re-estimate starting-positions from motif} \]
\[ \rightarrow \text{re-estimate motif from starting-positions} \]
until convergence (change < \( \varepsilon \))
return: \textit{motif}, \textit{starting-positions}
Representing Motif $M(k,c)$ and Background $B(c)$

- Assume motif has fixed width, $W$
- Motif represented by matrix of probabilities: $M(k,c)$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>0.4</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>G</td>
<td>0.3</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>T</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

$M = \begin{bmatrix}
A & 0.1 & 0.5 & 0.2 \\
C & 0.4 & 0.2 & 0.1 \\
G & 0.3 & 0.1 & 0.6 \\
T & 0.2 & 0.2 & 0.1 \\
\end{bmatrix}$

($\sim$CAG)

- Background represented by $B(c)$, frequency of each base

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.27</td>
<td></td>
</tr>
</tbody>
</table>

$B = \begin{bmatrix}
A & 0.26 \\
C & 0.24 \\
G & 0.23 \\
T & 0.27 \\
\end{bmatrix}$

(near uniform)

(see also: di-nucleotide etc)
Representing the starting position probabilities ($Z_{ij}$)

- The element $Z_{ij}$ of the matrix $Z$ represents the probability that the motif starts in position $j$ in sequence $i$.

\[
Z = \begin{bmatrix}
\text{seq1} & 0.1 & 0.1 & 0.2 & 0.6 \\
\text{seq2} & 0.4 & 0.2 & 0.1 & 0.3 \\
\text{seq3} & 0.3 & 0.1 & 0.5 & 0.1 \\
\text{seq4} & 0.1 & 0.5 & 0.1 & 0.3 \\
\end{bmatrix}
\]

Some examples:

- $Z_1$: no clear winner
- $Z_2$: two candidates
- $Z_3$: one big winner
- $Z_4$: uniform
Starting positions \((Z_{ij}) \leftrightarrow \text{Motif matrix } M(k,c)\)

- **\(Z_{ij}\):** Probability that on sequence \(i\), motif start at position \(j\)
- **\(M(k,c)\):** Probability that \(k\)th character of motif is letter \(c\)

### Motif: \(M(k,c)\)

<table>
<thead>
<tr>
<th>(c=A)</th>
<th>(k=1)</th>
<th>(k=2)</th>
<th>(k=3)</th>
<th>(k=4)</th>
<th>(k=5)</th>
<th>(k=6)</th>
<th>(k=7)</th>
<th>(k=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.1)</td>
<td>(0.3)</td>
<td>(0.1)</td>
<td>(0.2)</td>
<td>(0.2)</td>
<td>(0.4)</td>
<td>(0.3)</td>
<td>(0.1)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c=C)</th>
<th>(k=1)</th>
<th>(k=2)</th>
<th>(k=3)</th>
<th>(k=4)</th>
<th>(k=5)</th>
<th>(k=6)</th>
<th>(k=7)</th>
<th>(k=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.5)</td>
<td>(0.2)</td>
<td>(0.1)</td>
<td>(0.1)</td>
<td>(0.6)</td>
<td>(0.1)</td>
<td>(0.2)</td>
<td>(0.7)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c=G)</th>
<th>(k=1)</th>
<th>(k=2)</th>
<th>(k=3)</th>
<th>(k=4)</th>
<th>(k=5)</th>
<th>(k=6)</th>
<th>(k=7)</th>
<th>(k=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.2)</td>
<td>(0.2)</td>
<td>(0.6)</td>
<td>(0.5)</td>
<td>(0.1)</td>
<td>(0.2)</td>
<td>(0.2)</td>
<td>(0.1)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c=T)</th>
<th>(k=1)</th>
<th>(k=2)</th>
<th>(k=3)</th>
<th>(k=4)</th>
<th>(k=5)</th>
<th>(k=6)</th>
<th>(k=7)</th>
<th>(k=8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.2)</td>
<td>(0.3)</td>
<td>(0.2)</td>
<td>(0.1)</td>
<td>(0.3)</td>
<td>(0.3)</td>
<td>(0.3)</td>
<td>(0.1)</td>
<td></td>
</tr>
</tbody>
</table>

### Three variations for re-computing motif \(M(k,c)\) from \(Z_{ij}\) matrix

- **Expectation maximization** → All starts weighted by \(Z_{ij}\) prob distribution
- **Gibbs sampling** → Single start for each seq \(X_i\) by sampling \(Z_{ij}\)
- **Greedy approach** → Best start for each seq \(X_i\) by maximum \(Z_{ij}\)

### Computing \(Z_{ij}\) matrix from \(M(k,c)\) is straightforward

- At each position, evaluate start probability by multiplying across the matrix
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix ⇔ positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint $(M, Z_{ij})$ distribution
   - Sampling motif positions based on the $Z$ vector
   - More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score $\Rightarrow$ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays
E-step:
Estimate $Z_{ij}$ positions from matrix

Starting positions: $Z_{ij}$  
Motif: $M(k,c)$
Calculating the Z vector (using M)

- To estimate the starting positions in Z at step t
  \[ Z_{ij}^{(t)} = \Pr(Z_{ij} = 1 \mid X_i, M^{(t)}) = \frac{\Pr(X_i \mid Z_{ij} = 1, M^{(t)}) \Pr(Z_{ij} = 1)}{\Pr(X_i)} \] (Bayes’ rule)

- At iteration t, calculate \( Z_{ij}^{(t)} \) based on \( M^{(t)} \)
  - We just saw how to calculate \( \Pr(X_i \mid Z_{ij}=1,M^{(t)}) \)
  - To obtain total probability \( \Pr(X_i) \), sum over all starting positions

\[
Z_{ij}^{(t)} = \frac{\Pr(X_i \mid Z_{ij} = 1, M^{(t)}) \Pr(Z_{ij} = 1)}{\sum_{k=1}^{L-W+1} \Pr(X_i \mid Z_{ik} = 1, M^{(t)}) \Pr(Z_{ik} = 1)}
\]

- Assume uniform priors (motif eq likely to start at any position)
Calculating the Z vector: Example

\[
X_i = \begin{bmatrix}
G & C & T & G \\
T & A & G
\end{bmatrix}
\]

\[
p = \begin{bmatrix}
A & 0.25 & 0.1 & 0.5 & 0.2 \\
C & 0.25 & 0.4 & 0.2 & 0.1 \\
G & 0.25 & 0.3 & 0.1 & 0.6 \\
T & 0.25 & 0.2 & 0.2 & 0.1
\end{bmatrix}
\]

\[
Z_{i1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25
\]

\[
Z_{i2} = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25
\]

\[
\sum_{j=1}^{L-W+1} Z_{ij} = 1
\]

• then normalize so that
Denominator: \( P(X_i) \), when motif position is known

- Probability of training sequence \( X_i \), given hypothesized start position \( j \)

\[
\Pr(X_i \mid Z_{ij} = 1, M, B) = \prod_{k=1}^{j-1} B(X_{i,k}) \prod_{k=j}^{j+W-1} M(k-j+1, X_{i,k}) \prod_{k=j+W}^{L} B(X_{i,k})
\]

\( \text{before motif} \quad \text{motif} \quad \text{after motif} \)

- Example:

\[
X_i = \begin{array}{cccc}
G & C & T & G & T & A & G \\
\end{array}
\]

\[
B = \begin{array}{cccc}
A & 0.25 & & \\
C & 0.25 & & \\
G & 0.25 & & \\
T & 0.25 & & \\
\end{array}
\]

\[
M = \begin{array}{ccc}
1 & 2 & 3 \\
A & 0.1 & 0.5 & 0.2 \\
C & 0.4 & 0.2 & 0.1 \\
G & 0.3 & 0.1 & 0.6 \\
T & 0.2 & 0.2 & 0.1 \\
\end{array}
\]

\[
\Pr(X_i \mid Z_{i3} = 1, M, B) = \frac{B(G) \times B(C) \times M(1,T) \times M(2,G) \times M(3,T) \times B(A) \times B(G)}{0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25}
\]
Aside: Simplifying $P(X_i)$

- Probability of training sequence $X_i$, given hypothesized start position $j$

$$\Pr(X_i \mid Z_{ij} = 1, M, B) = \prod_{k=1}^{j-1} B(X_{i,k}) \prod_{k=j}^{j+W-1} M(k - j + 1, X_{i,k}) \prod_{k=j+W}^{L} B(X_{i,k})$$

\[= \prod_{k=j}^{j+W-1} \frac{M(k - j + 1, X_{i,k})}{B(X_{i,k})} \prod_{k=1}^{L} B(X_{i,k}) \]

- Before motif
- Motif
- After motif

Can be stored in a matrix

Constant for each sequence
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix $\Leftrightarrow$ positions
   – E step: Estimate motif positions $Z_{ij}$ from motif matrix
   – M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint ($M, Z_{ij}$) distribution
   – Sampling motif positions based on the $Z$ vector
   – More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   – Genome-wide conservation scores, motif extension
   – Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   – Phylogenies, Branch length score $\Rightarrow$ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   – 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   – HiDRA: random ATAC fragmentation + self-reporter assays
M-step:
Max-likelihood motif from $Z_{ij}$ positions

Starting positions: $Z_{ij}$
Motif: $M(k,c)$
Three examples for **Greedy, Gibbs Sampling, EM**

- **Greedy always picks maximum**
- **Gibbs sampling picks one at random** (or)
- **EM uses both in estimating motif**

**Z₁**

- All methods agree

**Z₂**

- One big winner

**Z₃**

- Greedy ignores most of the probability
- Gibbs sampling rapidly converges to some choice
- EM averages over the entire sequence (slow/no convergence)
The M-step: Estimating the motif $M$

- recall $M(k, c)$ represents the probability of character $c$ in position $k$; $B(c)$ stores values for the background.

\[
M^{(t+1)}(k, c) = \frac{n_{k,c} + d}{\sum_c (n_{k,c} + d)}
\]

where \[
n_{c,k} = \sum_i \sum_{\{j|X_{i,j+k-1}=c\}} Z_{ij}
\]

\[
B^{(t+1)}(c) = \frac{n_{0,c} + d}{\sum_c (n_{0,c} + d)}
\]

where \[
n_{0,c} = n_c - \sum_{j=1}^W n_{j,c}
\]
M-step example: Estimating $M(k,c)$ from $Z_{ij}$

\[
\begin{align*}
X_1 &= \begin{array}{cccc}
A & C & A & G \\
Z_1 &= 0.1 & 0.7 & 0.1 & 0.1
\end{array} \\
X_2 &= \begin{array}{cccc}
A & G & G & C & A & G \\
Z_2 &= 0.4 & 0.1 & 0.1 & 0.4
\end{array} \\
X_3 &= \begin{array}{cccc}
T & C & A & G \\
Z_3 &= 0.2 & 0.6 & 0.1 & 0.1
\end{array}
\]

- **EM**: sum over full probability
  - $n_{1,A} = 0.1 + 0.1 + 0.4 + 0.1 = 0.7$
  - $n_{1,C} = 0.7 + 0.4 + 0.6 = 1.7$
  - $n_{1,G} = 0.1 + 0.1 + 0.1 + 0.1 = 0.4$
  - $n_{1,T} = 0.2 = 0.2$
  - Total: $T = 0.7 + 1.7 + 0.4 + 0.2 = 3.0$

- **Normalize and add pseudo-counts**
  - $M(1,A) = \frac{0.7 + 1}{T + 4} = \frac{1.7}{7} = 0.24$
  - $M(1,C) = \frac{1.7 + 1}{T + 4} = \frac{2.7}{7} = 0.39$
  - $M(1,G) = \frac{0.4 + 1}{T + 4} = \frac{1.4}{7} = 0.2$
  - $M(1,T) = \frac{0.2 + 1}{T + 4} = \frac{1.2}{7} = 0.17$

\[
M(k,c) = \begin{array}{ccc}
A & 0.24 & 0.39 & 0.21 \\
C & 0.39 & 0.21 & 0.18 \\
G & 0.2 & 0.24 & 0.44 \\
T & 0.17 & 0.16 & 0.16
\end{array}
\]

- **Em approach**: Avg’em all
- **Gibbs sampling**: Sample one
- **Greedy**: Select max
The EM Algorithm

- EM converges to a local maximum in the likelihood of the data given the model:

\[ \prod_i \Pr(X_i | M, B) \]

- Deterministic iterations max direction of ascent
- Usually converges in a small number of iterations
- Sensitive to initial starting point (i.e. values in $M$)
EM searches for parameters to increase $P(\text{seqs}|\text{parameters})$

Useful to think of $P(\text{seqs}|\text{parameters})$ as a function of parameters.

EM starts at an initial set of parameters.

And then “climbs uphill” until it reaches a local maximum.

Where EM starts can make a big difference.
One solution: Search from Many Different Starts

To minimize the effects of local maxima, you should search multiple times from different starting points

MEME uses this idea:

Start at many points
Run for one iteration
Choose starting point that got the “highest” and continue
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix ↔ positions
   – E step: Estimate motif positions $Z_{ij}$ from motif matrix
   – M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint $(M, Z_{ij})$ distribution
   – Sampling motif positions based on the Z vector
   – More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   – Genome-wide conservation scores, motif extension
   – Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   – Phylogenies, Branch length score $\Rightarrow$ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   – 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   – HiDRA: random ATAC fragmentation + self-reporter assays
## Three options for assigning points, and their parallels across K-means, HMMs, Motifs

<table>
<thead>
<tr>
<th>Update rule</th>
<th>Algorithm implementing E step in each of the three settings</th>
<th>Update model parameters (M step) ➔ max likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>The hidden label is:</td>
<td>Expression clustering</td>
<td>HMM learning</td>
</tr>
<tr>
<td>Assign each point to best label</td>
<td><strong>K-means:</strong> Assign each point to nearest cluster</td>
<td><strong>Viterbi training:</strong> label sequence with best path</td>
</tr>
<tr>
<td>Assign each point to all labels, probabilistically</td>
<td><strong>Fuzzy K-means:</strong> Assign to all clusters, weighted by proximity</td>
<td><strong>Baum-Welch training:</strong> label sequence w all paths (posterior decoding)</td>
</tr>
<tr>
<td>Pick one label at random, based on their relative probability</td>
<td><strong>N/A:</strong> Assign to a random cluster, sample by proximity</td>
<td><strong>N/A:</strong> Sample a single label for each position, according to posterior prob.</td>
</tr>
</tbody>
</table>
Three examples of Greedy, Gibbs Sampling, EM

- Greedy always picks maximum
- Gibbs sampling picks one at random (or)
- EM uses both in estimating motif

- All methods agree

- Greedy ignores most of the probability
- Gibbs sampling rapidly converges to some choice
- EM averages over the entire sequence (no preference)
Gibbs Sampling

• A general procedure for sampling from the joint distribution of a set of random variables \( \Pr(U_1 \ldots U_n) \) by iteratively sampling from for each \( j \) \( \Pr(U_j | U_1 \ldots U_{j-1}, U_{j+1} \ldots U_n) \)

• Useful when it’s hard to explicitly express means, stdevs, covariances across the multiple dimensions

• Useful for supervised, unsupervised, semi-supervised learning
  – Specify variables that are known, sample over all other variables

• Approximate:
  – Joint distribution: the samples drawn
  – Marginal distributions: examine samples for subset of variables
  – Expected value: average over samples

• Example of Markov-Chain Monte Carlo (MCMC)
  – The sample approximates an unknown distribution
  – Stationary distribution of sample (only start counting after burn-in)
  – Assume independence of samples (only consider every 100)

• Special case of Metropolis-Hastings
  – In its basic implementation of sampling step
  – But it’s a more general sampling framework
Gibbs Sampling for motif discovery

- First application to motif finding: Lawrence et al 1993
  - Can view as a stochastic analog of EM for motif discovery task
  - Less susceptible to local minima than EM
- EM maintains distribution $Z_i$ over the starting points for each seq
- Gibbs sampling selects specific starting point $a_i$ for each seq
  - but keeps resampling these starting points

Given: length parameter $W$, training set of sequences

Choose random positions for $a$

Do

Pick a sequence $X_i$

Estimate $p$ given current motif positions $a$ (update step)
  - (using all sequences but $X_i$)
  - Sample a new motif position $a_i$ for $X_i$ (sampling step)

Until convergence

Return: $p$, $a$
Popular implementation: AlignACE, BioProspector

AlignACE: first statistical motif finder
BioProspector: improved version of AlignACE

Both use basic Gibbs Sampling algorithm:

1. **Initialization:**
   a. Select random locations in sequences $X_1, \ldots, X_N$
   b. Compute an initial model $M$ from these locations

2. **Sampling Iterations:**
   a. Remove one sequence $X_i$
   b. Recalculate model
   c. Pick a new location of motif in $X_i$ according to probability
      the location is a motif occurrence

In practice, run algorithm from multiple random initializations:

1. Initialize
2. Run until convergence
3. Repeat 1,2 several times, report common motifs
Gibbs Sampling (AlignACE)

• **Given:**
  - \(X_1, \ldots, X_N,\)
  - motif length \(W,\)
  - background \(B,\)

\[
\sum_{i=1}^{N} \sum_{k=1}^{W} \log \frac{M(k, X_{i,a_i+k})}{B(X_{i,a_i+k})}
\]

• **Find:**
  - Model \(M\)
  - Locations \(a_1, \ldots, a_N\) in \(X_1, \ldots, X_N\)

Maximizing log-odds likelihood ratio
This is the same as the EM objective (notice log and notation change)
Gibbs Sampling (AlignACE)

Predictive Update:

• Select a sequence $x_i$
• Remove $x_i$, recompute model:

$$M(k, c) = \frac{d + \sum_{s \neq i} (X_{s, a_s + k} = c)}{(N - 1) + 4d}$$

where $d$ is a pseudocount to avoid 0s
Sampling New Motif Positions

- for each possible starting position, \( a_i=j \), compute a weight

\[
A_j = \prod_{k=j}^{j+w-1} \frac{M(k - j + 1, X_{i,k})}{B(X_{i,k})}
\]

- randomly select a new starting position \( a_i \) according to these weights (normalizing across the sequence, again like with MEME)

- Note, this is equivalent to using the likelihood from MEME because:

\[
A_j \propto \Pr(X_i \mid Z_{ij} = 1, p)
\]

![Prob](image)
Advantages / Disadvantages

Advantages:
• Easier to implement
• Less dependent on initial parameters
• More versatile, easier to enhance with heuristics

Disadvantages:
• More dependent on all sequences to exhibit the motif
• Less systematic search of initial parameter space
Gibbs Sampling and Climbing

Because gibbs sampling does always choose the best new location it can move to another place not directly uphill.

In theory, Gibbs Sampling less likely to get stuck at a local maxima.
Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab
Human Vision ⇔ many layers of abstraction ⇔ Deep learning

Goodfellow 2016
Key idea: **Representation learning**

In deep learning, the two tasks are **coupled**:

- the classification task “drives” the **feature extraction**
- Extremely powerful and general paradigm

→ Be creative! The field is still at its infancy!
→ New application domains (e.g. beyond images) can have **structure** that current architectures **do not capture/exploit**
→ Genomics/biology/neuroscience can help drive development of **new architectures**
## Key design principles of CNNs (+brain counterparts)

<table>
<thead>
<tr>
<th>Property</th>
<th>Human Visual System Property</th>
<th>Deep Learning CNN Building Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locality</td>
<td>Low-level neurons respond to local patches (receptive field)</td>
<td><strong>Local</strong> computation of convolutional filters (not a fully-connected network)</td>
</tr>
<tr>
<td>Filters</td>
<td>Specialized neurons carry out low-level detection operation</td>
<td>Low-level <strong>filters</strong> carry out the same operation throughout the network</td>
</tr>
<tr>
<td>Layers / abstraction</td>
<td>Layers of neurons learn increasingly abstract ‘concepts’</td>
<td><strong>Layers</strong> of hidden units, abstract concepts learned from simpler parts / building blocks</td>
</tr>
<tr>
<td>Threshold</td>
<td>Neurons fire after cross activation threshold ➞ non-linearity</td>
<td><strong>Activation</strong> functions introduce non-linearities ➞ expand universe of functions</td>
</tr>
<tr>
<td>Pooling</td>
<td>Higher-level neurons invariant to exact position, sum/max of prev.</td>
<td><strong>Max/Avg pooling</strong> layers: positional invariance reduced # parameters, speed up compute</td>
</tr>
<tr>
<td>Multimodal</td>
<td>Different neurons extract different features of image</td>
<td><strong>Multiple filters</strong> applied simultaneously, each captures different aspects of original image</td>
</tr>
<tr>
<td>Sampling Density</td>
<td>Central vision sampled densely by photoreceptors than periphery</td>
<td>Adjust <strong>stride</strong> of filter application to denser (slower) vs. sparser (faster) sampling</td>
</tr>
<tr>
<td>Saturation</td>
<td>Neurons ‘tired’ after activation, signal quiets down</td>
<td>Limiting weight of individual hidden units, dropout learning, <strong>regularization</strong></td>
</tr>
<tr>
<td>Learn/Reinforcement</td>
<td>Useful connections strengthened over time</td>
<td><strong>Back-propagation</strong>, adjusting weights across the hierarchy</td>
</tr>
<tr>
<td>Feed-forward edges</td>
<td>Neurons with long connections from lower levels to higher ones</td>
<td>Residual networks (<strong>ResNets</strong>) feed lower-level signal, avoid vanishing gradients</td>
</tr>
</tbody>
</table>
Predictive model of regulatory DNA

Transcription factor ChIP-seq data OR chromatin accessibility (DNase-seq / ATAC-seq data)

DNA sequences ($S_i$)

Class = +1 (20.2)
Class = +1 (10.6)
Class = +1 (15.8)
Class = 0 (0.3)
Class = 0 (1.2)
Class = 0 (3.5)

Arvey et al. 2012
Ghandi et al. 2014
Setty et al. 2015
Convolutional neural network (CNN) with DNA sequence inputs

- 100s conv. filters (motif detectors)
- Combinations of motifs (motif syntax)

- Binary or continuous output

Is TF bound? / TF binding signal

Automatically learns predictive patterns from raw DNA sequence to maximize prediction accuracy of binding labels

One-hot encoded input: DNA sequence represented as ones and zeros

Alipanahi et al. 2015 (DeepBind)
Kelley et al. 2016 (Basset)
Zhou et al. 2015 (DeepSEA)
High-resolution ‘shapes’ and ‘spans’ of TF and chromatin profiles capture exquisite information about protein-DNA contacts.

Protein-DNA binding expt.

DNA accessibility experiments

https://doi.org/10.3109/10409238.2015.1051505
Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery

Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix ↔ positions
   – E step: Estimate motif positions $Z_{ij}$ from motif matrix
   – M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint $(M,Z_{ij})$ distribution
   – Sampling motif positions based on the $Z$ vector
   – More likely to find global maximum, easy to implement

4. Evolutionary signatures for *de novo* motif discovery
   – Genome-wide conservation scores, motif extension
   – Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   – Phylogenies, Branch length score $\Rightarrow$ Confidence score

6. *De novo* dissection of regulatory regions in high-resolution
   – 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   – HiDRA: random ATAC fragmentation + self-reporter assays
Motivation for *de novo* genome-wide motif discovery

- Both TF and region centric approaches are not comprehensive and are biased
- TF centric approaches generally require transcription factor (or antibody to factor)
  - Lots of time and money
  - Also have computational challenges
- *De novo* discovery using conservation is unbiased but can’t match motif to factor and require multiple genomes
Evolutionary signatures for regulatory motifs

Starting by looking at known motif instances

Individual motif instances are preferentially conserved

Can we just take conservation islands and call them motifs?

- No. Many conservation islands are due to chance or perhaps due to non-motif conservation

Kellis el al, Nature 2003
Xie et al. Nature 2005
Stark et al, Nature 2007
Conservation islands overlap known motifs

Increase power by testing conservation in many regions
Evaluate conservation within:

1. All intergenic regions
   - Gal4: 13%
   - Controls: 2%

2. Intergenic : coding
   - Gal4: 13% : 3%
   - Controls: 2% : 7%

3. Upstream : downstream
   - Gal4: 12:0
   - Controls: 1:1

A signature for regulatory motifs
Test 1: Intergenic conservation

Total count vs. Conserved count

CGG-11-CCG
Test 2: Intergenic vs. Coding

Higher Conservation in Genes

CGG-11-CCG
Test 3: Upstream vs. Downstream

CGG-11-CCG

Downstream motifs?

Most Patterns

Upstream Conservation

Downstream Conservation
Conservation for TF motif discovery

1. Enumerate motif seeds

- Six non-degenerate characters with variable size gap in the middle

2. Score seed motifs

- Use a conservation ratio corrected for composition and small counts to rank seed motifs

3. Expand seed motifs

- Use expanded nucleotide IUPAC alphabet to fill unspecified bases around seed using hill climbing

4. Cluster to remove redundancy

- Using sequence similarity

Kellis, Nature 2003
Learning motif degeneracy using evolution

- Record frequency with which one sequence is “replaced” by another in evolution
- Use this to find clusters of k-mers that correspond to a single motif

Tanay, Genome Research 2004
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix ↔ positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint (M, $Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score $\rightarrow$ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays
Validation of the discovered motifs

• Because genome-wide motif discovery is *de novo*, we can use functional datasets for validation
  – Enrichment in co-regulated genes
  – Overlap with TF binding experiments
  – Enrichment in genes from the same complex
  – Positional biases with respect to transcription start
  – Upstream vs. downstream / inter vs. intra-genic bias
  – Similarity to known transcription factor motifs

• Each of these metrics can also be used for discovery
  – In general, split metrics into discovery vs. validation
  – As long as they are *independent*!
  – Strategies that combine them all lose ability to validate
    • Directed experimental validation approaches are then needed
Similarity to known motifs

- If discovered motifs are real, we expect them to match motifs in large databases of known motifs.
- We find this (significantly higher than with random motifs).
- Why not perfect agreement?
  - Many known motifs are not conserved.
  - Known motifs are biased; may have missed real motifs.

<table>
<thead>
<tr>
<th>MCS</th>
<th>Discovered motif</th>
<th>Known Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>46.8</td>
<td>GGGCGGGR</td>
<td>SP-1</td>
</tr>
<tr>
<td>34.7</td>
<td>GCCATTTTg</td>
<td>YY1</td>
</tr>
<tr>
<td>32.7</td>
<td>CACGTG</td>
<td>MYC</td>
</tr>
<tr>
<td>31.2</td>
<td>GATGGG</td>
<td>NF-Y</td>
</tr>
<tr>
<td>30.8</td>
<td>TGATnTCA</td>
<td>AP-1</td>
</tr>
<tr>
<td>29.7</td>
<td>GGGAGGRR</td>
<td>MAZ</td>
</tr>
<tr>
<td>29.5</td>
<td>TGACGTMR</td>
<td>CREB</td>
</tr>
<tr>
<td>26.0</td>
<td>CCGCCATYK</td>
<td>Nf-MUE1</td>
</tr>
<tr>
<td>25.0</td>
<td>TGACCTTG</td>
<td>Err</td>
</tr>
<tr>
<td>22.6</td>
<td>CCGGAARY</td>
<td>Elk-1</td>
</tr>
<tr>
<td>19.8</td>
<td>SCGGAAGY</td>
<td>Gabp</td>
</tr>
<tr>
<td>17.9</td>
<td>CAATTTCCK</td>
<td>Stat1</td>
</tr>
</tbody>
</table>

70/174 mammalian motifs

<table>
<thead>
<tr>
<th>MCS</th>
<th>Discovered motif</th>
<th>Known Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>65.6</td>
<td>CTAATTAAA</td>
<td>en</td>
</tr>
<tr>
<td>57.3</td>
<td>TTKCAATTAA</td>
<td>Repo</td>
</tr>
<tr>
<td>54.9</td>
<td>WATTRATTK</td>
<td>Ara</td>
</tr>
<tr>
<td>54.4</td>
<td>AAATTTATG</td>
<td>Prd</td>
</tr>
<tr>
<td>51</td>
<td>GCAATAAA</td>
<td>Vvl</td>
</tr>
<tr>
<td>46.7</td>
<td>DTAAATTRYN</td>
<td>Ubx</td>
</tr>
<tr>
<td>45.7</td>
<td>TGATTAAT</td>
<td>Ap</td>
</tr>
<tr>
<td>43.1</td>
<td>YMATTAAAAA</td>
<td>Abd-A</td>
</tr>
<tr>
<td>41.2</td>
<td>AAACNNGTT</td>
<td>Ftz</td>
</tr>
<tr>
<td>40</td>
<td>RATTKAATT</td>
<td>Br-Z3</td>
</tr>
<tr>
<td>39.5</td>
<td>GCACGTGT</td>
<td></td>
</tr>
<tr>
<td>38.8</td>
<td>AACASCTG</td>
<td></td>
</tr>
</tbody>
</table>

35/145 fly motifs

Stark, Nature 2007
Xie, Nature 2005
Positional bias of motif matches

• Motifs are involved in initiation of transcription
  → Motif matches biased versus TSS
    – 10% of fly motifs
    – 34% of mammalian motifs
  → Depletion of TF motifs in coding sequence
    – 57% of fly motifs
  → Clustering of motif matches
    – 19% of fly motifs
Motifs have functional enrichments

For both fly (top) and mammals (bottom), motifs are enriched in genes expressed in specific tissues.

Reveals modules of cooperating motifs.
Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix $\leftrightarrow$ positions
   – E step: Estimate motif positions $Z_{ij}$ from motif matrix
   – M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint $(M, Z_{ij})$ distribution
   – Sampling motif positions based on the $Z$ vector
   – More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   – Genome-wide conservation scores, motif extension
   – Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   – Phylogenies, Branch length score $\rightarrow$ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   – 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   – HiDRA: random ATAC fragmentation + self-reporter assays
Motif instance identification

How do we determine the functional binding sites of regulators?

Kheradpour, Stark, Roy, Kellis, Genome Research 2007
Experimental target identification: ChIP-chip/seq

Limitations:
- Antibody availability
- Restricted to specific stages/tissues
- Biological functionality of most binding sites unknown
- Resolution can be limited (can’t usually identify the precise base pairs)

Ren et al., 2000; Iyer et al., 2001 (ChIP-chip)
Robertson et al., 2007 (ChIP-seq)
Computational target identification

• Single genome approaches using motif clustering (e.g. Berman 2002; Schroeder 2004; Philippakis 2006)
  – Requires set of specific factors that act together
  – Miss instances of motifs that may occur alone

• Multi-genome approaches (phylogentic footprinting) (e.g. Moses 2004; Blanchette and Tompa 2002; Etwiller 2005; Lewis 2003)
  – Tend to either require absolute conservation or have a strict model of evolution
Challenges in target identification

- **Simple case**
  - Instance fully conserved in orthologous position near genes

- **Motif turn-around/movement**
  - Motif instance is not found in orthologous place due to birth/death or alignment errors

- **Distal/missing matches**
  - Due to sequencing/assemble errors or turnover
  - Distal instances can be difficult to assign to gene
Computing Branch Length Score (BLS)

BLS = 2.23sps (78%)

Allows for:
1. Mutations permitted by motif degeneracy
2. Misalignment/movement of motifs within window (up to hundreds of nucleotides)
3. Missing motif in dense species tree
1. Evaluate chance likelihood of a given score
   • Sequence could also be conserved due to overlap with un-annotated element (e.g. non-coding RNA)

2. Account for differences in motif composition and length
   • For example, short motif more likely to be conserved by chance
1. Use motif-specific shuffled control motifs determine the expected number of instances at each BLS by chance alone or due to non-motif conservation

2. Compute Confidence Score as fraction of instances over noise at a given BLS (=1 – false discovery rate)
Producing control motifs

When evaluating the conservation, enrichment, etc, of motifs, it is useful to have a set of “control motifs”

1. Produce 100 shuffles of our original motif
2. Filter motifs, requiring they match the genome with about (+/- 20%) of our original motif
3. Sort potential control motifs based on their similarity to other known motifs
4. Cluster potential control motifs and take at most one from each cluster, in increasing order of similarity to known motifs
Computing enrichments: background vs. foreground

- **Background vs. foreground**
  - co-regulated promoters vs. all genes
  - Bound by TF vs. other intergenic regions
- **Enrichment**: *fraction of motif instances in foreground* vs. *fraction of bases in foreground*

- **Correct for composition/conservation level**: compute enrichmt w/control motifs
  - Fraction of motif instances can be compared to *fraction of control motif instances in foreground*
  - A hypergeometric p-value can be computed (similar to $\chi^2$, but better for small numbers)

- **Fractions can be made more conservative using a binomial confidence interval**
1. Confidence selects for transcription factor motif instances in promoters and miRNA motifs in 3’ UTRs
Validation of discovered motif instances

Use independent experimental evidence
Look for functional biases / enrichments
1. Confidence selects for transcription factor motif instances in promoters and miRNA motifs in 3’ UTRs
2. miRNA motifs are found preferentially on the plus strand, whereas no such preference is found for TF motifs
Increased sensitivity using BLS
ChIP-Seq and ChIP-Chip technologies allow for identifying binding sites of a motif experimentally

- Conserved CTCF motif instances highly enriched in ChIP-Seq sites
- High enrichment does not require low sensitivity
- Many motif instances are verified

ChIP data from Barski, et al., Cell (2007)
Enrichment found for many factors

**Mammals**

- CTCF
- HNF6
- HNF4A
- HNF1A
- Stat1

**Flies**

- Twist
- Snail

Bar charts showing enrichment in ChIP regions with 0% and 50% confidence levels.
Enrichment increases in conserved bound regions

1. ChIP bound regions may not be conserved
2. For CTCF we also have binding data in mouse
3. Enrichment in intersection is dramatically higher

Human: Barski, et al., Cell (2007)
Mouse: Bernstein, unpublished
1. ChIP bound regions may not be conserved
2. For CTCF we also have binding data in mouse
3. Enrichment in intersection is dramatically higher
4. Trend persists for other factors where we have multi-species ChIP data
1. Motifs at 60% confidence and ChIP have similar enrichments (depletion for the repressor Snail) in the functional promoters

2. Enrichments persist even when you look at non-overlapping subsets

3. Intersection of two regions has strongest signal

4. Evolutionary and experimental evidence is complementary
   - ChIP includes species specific regions and differentiate tissues
   - Conserved instances include binding sites not seen in tissues surveyed

Fly regulatory network at 60% confidence

Several connections confirmed by literature (directly or indirectly)

Global view of instances allows us to make network level observations:

- 46% of targets were co-expressed with their factor in at least one tissue ($P < 2 \times 10^{-3}$)
- TFs were more targeted by TFs ($P < 10^{-20}$) and by miRNAs ($P < 5 \times 10^{-5}$)
- TF in-degree associated with miRNA in-degree (high-high: $P < 10^{-4}$; low-low $P < 10^{-6}$)
Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab
1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix $\leftrightarrow$ positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint $(M, Z_{ij})$ distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score $\Rightarrow$ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays
Challenges in regulatory genomics

TFs: Homology to TFs/domains
miRNAs: Evolutionary signatures
miRNAs: Experimental cloning

TFs: Mass Spec (difficult)

miRNAs: Evolutionary/structural signatures
miRNAs: Experimental cloning of 5’-ends

TFs/miRNAs: De novo comparative discovery**

Network analysis (next lecture)

Evolutionary footprints
DNase, TF/Chrom ChIP, Chromatin ‘dips’, MPRA

TFs: Enrichment in co-regulated genes/bound regions **

TFs: Perturbation response
TFs/miRs: Evolutionary signatures**
miRNAs: Composition/folding

Targets

TFs: ChIP-Chip/ChIP-Seq

TFs: Selex, DIP-Chip, Protein-Binding-Microarrays

Regulator

TF/miRNA

Motif

Sequence specificity

* = Covered in today’s lecture
From Identification to Large-Scale Confirmation and Dissection of Candidate Regulatory Regions

ENCODE, Roadmap Epigenomics, et al: Histone marks, TF binding, DNase, FAIRE, ...

→ identification of candidate regulatory regions

Next challenge: confirm/dissect 10,000s of regions!

• Test thousands of candidate regulatory regions at once
• Identify regulatory positions at or near nucleotide level resolution independent of sequence motifs
• Distinguish activating vs. repressive nucleotides

ENCODEx Project Consortium et al, 2012
Problem: Not all annotated enhancers are real

Luciferase assays

- PCR
- Predicted enhancer
- luc2
- Activity
- pGL4.23

Massively-parallel assays

- Array-based synthesis / library construction
- Transfect
- Predicted enhancer
- luc2
- pGL4.23
- Barcode

Slow, tedious, time-consuming

10,000+ elements at a time

2659 in vivo tested elements
1444 elements with enhancer activity
Visel et al. NAR 2007

Difference between endogenous epigenomic signatures (e.g. H3K27ac) vs. being able to actually drive expression of a reporter gene (take DNA sequence segment out of context)
Enabling Technology: Massively Parallel Reporter Assay (MPRA)

- Synthesize many enhancer versions → insert upstream
- Couple each with a barcode → insert downstream
- Make 10,000s of elements → plasmids, transfection
- High-throughput test in diff. cell types → 10k measurements

Can we achieve (1) large scale application (2) nucleotide level resolution, and (3) direction of effect, all without knowing motifs or precise 145bp to test?
High-resolution tiling dissection of individual regulatory regions

Challenge: hundreds of constructs needed for each region
Can test thousands of regions jointly?
Systematic motif disruption in 2000 regions for 5 activators and 2 repressors in 2 human cell lines

Kheradpour, Kellis et al. Genome Research, 2013

54000+ measurements (x2 cells, 2x repl)
What to perturb: Guided by computational predictions

- Chromatin mark-based cell line specific enhancers
- Oct4 predicted activator of embryonic stem cells
- Gfi1 predicted repressor K562/GM12878 cells

Coordinated activity reveals activators/repressors
HNF1 and HNF4 are predicted activators of HepG2 enhancers

- **Model:** Disruption of the motif site would abolish enhancer state
Example activator: conserved HNF4 motif match

WT expression specific to HepG2

Motif match disruptions reduce expression to background

Non-disruptive changes maintain expression

Random changes depend on effect to motif match

Random changes depend on effect to motif match
Results hold across 2000+ enhancers

- Scramble abolishes reporter expression
- Neutral mutations show no change
- Increasing mutations show more expression
- Repressor mutations → expression increase
- Motif context matters
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix ↔ positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint $(M, Z_{ij})$ distribution
   - Sampling motif positions based on the $Z$ vector
   - More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score $\rightarrow$ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays
Effect of enhancer position on reporter activity

Centers of selected regions show strongest activity

Tile regions at 30-bp offsets

Chromatin dips in matched cell show strongest activity
An offset of 30-bp can make a big difference

Replicates of same tile are highly consistent

Consecutive tiles can differ greatly
Consecutive tile diffs due to motif inclusion/exclusion

- Inclusion/exclusion of 30-bp intervals
  - Akin to systematic disruption
  - Increase resolution from tile (145bp) to offset (30bp)

Applications:
- Use to discover motifs?
- Further increase resolution?
• Increased resolution allows testing of only 30-bp intervals
• *De novo* discovered motifs match known motifs
• Discovery distinguishes activating vs. repressive factors
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix\(\leftrightarrow\)positions
   – E step: Estimate motif positions \(Z_{ij}\) from motif matrix
   – M step: Find max-likelihood motif from all positions \(Z_{ij}\)

3. Gibbs Sampling: Sample from joint \((M, Z_{ij})\) distribution
   – Sampling motif positions based on the \(Z\) vector
   – More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   – Genome-wide conservation scores, motif extension
   – Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   – Phylogenies, Branch length score \(\Rightarrow\) Confidence score

6. De novo dissection of regulatory regions in high-resolution
   – 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   – HiDRA: random ATAC fragmentation + self-reporter assays
Experimental design for high-resolution tiling

- Tile 295-bp regions (vs. 385)
- Tile @ 5-bp offsets (vs @30)
- Center on DNase peaks (vs dips)
- Single barcode / offset (vs 24)
- 244K spot array (vs 54K)
- 25 chromatin states (vs. Enh)
- Both minP and SV40 promoters
Chromatin state vs. reporter activity of DNase elements

Select 15,720 DNase elements across all 25 chromatin states
Regions selected in 4 cell types, tiled in HepG2,K562

<table>
<thead>
<tr>
<th>Elements tested in:</th>
<th>HepG2</th>
<th>K562</th>
</tr>
</thead>
<tbody>
<tr>
<td>minP</td>
<td>Rep1</td>
<td>Rep1</td>
</tr>
<tr>
<td>Rep2</td>
<td>Rep1</td>
<td>Rep2</td>
</tr>
<tr>
<td>SV40</td>
<td>Rep1</td>
<td>Rep1</td>
</tr>
<tr>
<td>h1esc</td>
<td>Rep2</td>
<td>Rep2</td>
</tr>
<tr>
<td>HepG2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K562</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huvec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>h1esc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HepG2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K562</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huvec</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

15,720 regions x 31 offsets x 2 promoters x 2 reps x 2 cell lines
Computational inference model

Measure $M_1..M_{31}$ $\Rightarrow$ infer $A_1..A_{59}$

Joint samples from multivariate normal

$$\begin{bmatrix} A_{r,t} \\ M_{r,t} \end{bmatrix} \sim N(\mu_{x_{r,t}}, \Sigma_{x_{r,t}})$$

True activity normally distributed

$$A_{r,t,k} \sim N(\mu_a, \sigma^2_a) \quad k = 1, ..., K$$

Observation is mean of true signals

$$\mu_a_r = \frac{1}{|M_r|} \sum_{m \in M_r} m$$

Measurements normally distributed

$$M_{r,t,j} \sim N(\mu_{m_{r,t,j}}, \sigma^2_{m_r})$$

Mean set to empirical mean

$$\mu_{m_{r,t,j}} = \frac{1}{N} \sum_{i=0}^{N-1} A_{r,t,j+i}$$

Variance set to empirical variance

$$\sigma^2_{m_r} = \frac{1}{|M_r|} \sum_{m \in M_r} (m - \frac{1}{|M_r|} \sum_{m \in M_r} m)^2$$
Examples of tiling data deconvolution

Detect activating/repressive elements at high resolution
Deconvolved regulatory signal vs. activator motif

Other CENTIPEDE motifs active in HepG2

GABPA (MA0062.2)

60 sites containing GABPA HepG2 motifs predicted by CENTIPEDE

CENTIPEDE predictions (Pique-Regi, et al, 2011)
46 sites containing NRSF HepG2 motifs predicted by CENTIPEDE

CENTIPEDE predictions (Pique-Regi, et al, 2011)
Aggregate Motif Score Highly Correlated between K562 and HepG2

Comparing to ~1900 motifs - both known and discovered on ENCODE TF ChIP-seq data (Kheradpour and Kellis, 2014) with >= 20 instances overlapping testing regions
Top Activating and Repressive Motifs Revealed

Motif discovered in multiple ENCODE data sets. Associated TF(s) uncertain. Associated with high conservation and gene expression (Xie et al, 2005; Pique-Regi, et al 2011)
Cell Type Specific Motifs Revealed

HepG2 Regulation Score

K562 Regulation Score

GATA

HNF1

E2F

RFX5

HNF4

TP53

HepG2 Regulation Score

K562 Regulation Score
Inferred positions match regulatory motifs

Predicted Activation and Repressive Bases Strongly Enrich for Predicted Binding Sites in HepG2 + K562
Active/repressed positions are evol. conserved

Strongest enrichment for repressive positions
Slight depletion at strongest activating positions
ERV1 repeat elements can drive activity

Strongest activating nucleotides match ERV1 repeats (by contrast, LINE elements strongly depleted)

Enable rapid evolution of gene-regulatory networks
DNase elements in different chromatin states differ in their activity levels.

Promoter, Enhancer regions highly activating.
ReprPC regions highly repressive.

HepG2 matched
Accessible regions drive stronger activity

For both activating and repressive positions
Discovery of repressors that act in active regions

- REST acts as a repressor in repressive regions (as expected)
- But RFX5 acts as a repressor only in active regions (modulator?)
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix \( \leftrightarrow \) positions
   - E step: Estimate motif positions \( Z_{ij} \) from motif matrix
   - M step: Find max-likelihood motif from all positions \( Z_{ij} \)

3. Gibbs Sampling: Sample from joint \((M, Z_{ij})\) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score \( \rightarrow \) Confidence score

6. De novo dissection of regulatory regions in high-resolution
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays
HiDRA: Longer probes + Hi-res dissection + 7M tests

ATAC selection ➔ No synthesis ➔ 7M tests
3'UTR incorp. ➔ Self-transcribe ➔ No barcode
Dense, random start/end ➔ Region tiling

High-resolution inference of driver nucleotides
➔ Exploit differences between neighboring fragments
➔ Driver nucleotides match motifs, evolut. conservation
HiDRA enables testing of larger fragments

![HiDRA size distribution graph](image)

- Single nucleosome
- Dimer

**HiDRA size distribution**

- Sharpr-MPRA tiled region size (295 nt)
- Median HiDRA tiled region length (1328 nt)

- Median HiDRA fragment size (337 nt)
- MPRA fragment size (145 nt)

Frequency (HiDRA fragments, blue)

Frequency (tiled regions, green)
HiDRA input DNA library recapitulates DNase/ATAC-Seq

Preferential selection of putative regulatory elements

Fig 1c from Buenrostro et al. Nature Methods 2013
HiDRA input DNA library: long, active, densely-covered regions

HiDRA DNA library captures more active elements

**b**

HiDRA size distribution

- Single nucleosome
- Dimer
- Shapr-MPRA tiled region size (295nt)
- Median HiDRA tiled region length (1328nt)
- 5X

- Median HiDRA fragment size (337nt)
- MPRA fragment size (145nt)

**C**

Active enhancer state

128X more regions with 10+ HiDRA fragments vs. random sampling

Active TSS state

209X

ATAC-seq peaks

219X

Number of unique fragments covering regions

- HiDRA fragments
- Randomly sampled fragments

**Fragment**s: 99% are 169-477 nt (median: 337nt)
**Regions**: 99% are 513-4,036 nt (median: 1,328nt)

Up to 200-fold higher coverage for putative regulatory elements
HiDRA captures known enhancers, known motifs

High sensitivity / high specificity vs. Luciferase assays

Quantitative read-out

Capture known motifs
Sharpr2 algorithm infers high-resolution driver nucleotides

- Exploit differences between neighboring fragments
- Driver nucleotides match motifs, evolutionary conservation

- Enrichment: $P<10^{-73}$
HiDRA high-resolution drivers help dissect GWAS loci

Pinpoint causal GWAS variants

- General method to dissect non-coding variation
- Applicable to millions of genomic regions simultaneously

Pinpoint causal GWAS variants
HiDRA activity differences between risk and non-risk alleles

Allele-specific activity for IBD-associated rs2382817
HiDRA summary

• 7M fragments tested in one experiment
• Longer fragments (~350nt on average)
• High reproducibility, 0.95 for higher-activity elmt
• Up to 200-fold enrichment for regulatory regions
• High-resolution dissection of driver nts
• Captures known motifs, conserved nucleotides
• Pinpoints driver SNPs in GWAS loci
• Reveals diffs between risk and non-risk alleles
• General tool for testing regulatory regions

doi.org/10.1101/193136
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation

2. Expectation maximization: Motif matrix ↔ positions
   – E step: Estimate motif positions $Z_{ij}$ from motif matrix
   – M step: Find max-likelihood motif from all positions $Z_{ij}$

3. Gibbs Sampling: Sample from joint $(M, Z_{ij})$ distribution
   – Sampling motif positions based on the $Z$ vector
   – More likely to find global maximum, easy to implement

4. Evolutionary signatures for de novo motif discovery
   – Genome-wide conservation scores, motif extension
   – Validation of discovered motifs: functional datasets

5. Evolutionary signatures for instance identification
   – Phylogenies, Branch length score ➔ Confidence score

6. De novo dissection of regulatory regions in high-resolution
   – 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   – HiDRA: random ATAC fragmentation + self-reporter assays