Lecture 11
Regulatory motif discovery
and target identification

Module III: Epigenomics and gene regulation

- Computational Foundations
  - L10: Multi-variate HMMs, IDR, peak calling
  - L11: Gibbs Sampling: between EM and Viterbi training
  - L12: Spectral algorithms, matrix operations, linear algebra

- Biological frontiers:
  - L10: Epigenomics, chromatin states, differentiation
  - L11: Regulatory motif discovery, TF binding
  - L12: Gene networks, regulatory genomics

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo

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
   - Foreground vs. background. Real vs. control motifs

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

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

ATGACTAAATCTCATTCAGAAGAA

GAL1

CCCCWCGG CCG

Gal4
Mig1
How Transcription Factors actually recognize 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

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)
PBM (Protein binding microarrays; Mukherjee, 2004)

Double stranded DNA arrays
Regulator
TF/miRNA
Targets
Network analysis (next lecture)
TFs: Homology to TFs/domains
miRNAs: Evolutionary signatures
miRNAs: Experimental cloning
TFs: Selex, DIP-Chip, Protein-Binding-Microarrays
miRNAs: Evolutionary/structural signatures
miRNAs: Experimental cloning of 5'-ends
TFs: Mass Spec (difficult)
TFs/miRNAs: De novo comparative discovery**
TFs: Enrichment in co-regulated genes/ bound regions **
TFs/miRNAs: Evolutionary signatures**
miRNAs: Composition/folding

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo

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
   - Foreground vs. background. Real vs. control motifs

Challenges in regulatory genomics

- Targets
- Functional instances
- Motif
- Sequence specificity
- Regulator
- TF/miRNA

How would you go about it?

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
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

Starting positions ↔ Motif matrix

- given aligned sequences → easy to compute profile matrix

<table>
<thead>
<tr>
<th>Sequence positions</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>

Starting positions ↔ Motif matrix

- easy to find starting position probabilities

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

Basic Iterative Approach

Given: length parameter \( W \), training set of sequences
set initial values for motif
do
  - re-estimate starting-positions from motif
  - re-estimate motif from starting-positions
until convergence (change < \( \epsilon \))
return: motif, starting-positions

Representing Motif \( M(k,c) \) and Background \( B(c) \)

- Assume motif has fixed width, \( W \)
- Motif represented by matrix of probabilities: \( M(k,c) \)
  - the probability of character \( c \) in column \( k \)

\[
M = \begin{pmatrix}
A & C & G & T \\
0.1 & 0.4 & 0.3 & 0.2 \\
0.5 & 0.2 & 0.1 & 0.6 \\
0.3 & 0.1 & 0.1 & 0.2 \\
0.1 & 0.2 & 0.2 & 0.3 \\
0.6 & 0.5 & 0.1 & 0.2 \\
0.3 & 0.1 & 0.1 & 0.2 \\
0.4 & 0.2 & 0.2 & 0.1 \\
\end{pmatrix}
\]

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

\[
B = \begin{pmatrix}
A & C & G & T \\
0.26 & 0.24 & 0.23 & 0.27 \\
\end{pmatrix}
\]

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 \( i \)

\[
Z = \begin{pmatrix}
1 & 2 & 3 & 4 \\
seq1 & 0.1 & 0.1 & 0.2 & 0.6 \\
seq2 & 0.4 & 0.2 & 0.1 & 0.3 \\
seq3 & 0.3 & 0.1 & 0.5 & 0.1 \\
seq4 & 0.1 & 0.5 & 0.1 & 0.3 \\
\end{pmatrix}
\]

Some examples:

- \( Z_1 \): no clear winner
- \( Z_2 \): two candidates
- \( Z_3 \): one big winner
- \( Z_4 \): uniform

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

- At each position, evaluate start probability by multiplying across the matrix

Starting positions (\( Z_{ij} \)) ↔ 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 \( c \)

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

Representing Motif \( M(k,c) \) and Background \( B(c) \)

- Assume motif has fixed width, \( W \)
- Motif represented by matrix of probabilities: \( M(k,c) \)
  - the probability of character \( c \) in column \( k \)

\[
M = \begin{pmatrix}
A & C & G & T \\
0.1 & 0.4 & 0.3 & 0.2 \\
0.5 & 0.2 & 0.1 & 0.6 \\
0.3 & 0.1 & 0.1 & 0.2 \\
0.1 & 0.2 & 0.2 & 0.3 \\
0.6 & 0.5 & 0.1 & 0.2 \\
0.3 & 0.1 & 0.1 & 0.2 \\
0.4 & 0.2 & 0.2 & 0.1 \\
\end{pmatrix}
\]

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

\[
B = \begin{pmatrix}
A & C & G & T \\
0.26 & 0.24 & 0.23 & 0.27 \\
\end{pmatrix}
\]

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 \( i \)

\[
Z = \begin{pmatrix}
1 & 2 & 3 & 4 \\
seq1 & 0.1 & 0.1 & 0.2 & 0.6 \\
seq2 & 0.4 & 0.2 & 0.1 & 0.3 \\
seq3 & 0.3 & 0.1 & 0.5 & 0.1 \\
seq4 & 0.1 & 0.5 & 0.1 & 0.3 \\
\end{pmatrix}
\]

Some examples:

- \( Z_1 \): no clear winner
- \( Z_2 \): two candidates
- \( Z_3 \): one big winner
- \( Z_4 \): uniform
Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM, Gibbs), de novo
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

Calculating P($X_i$) when motif position is known

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

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

- Example:

\[
X_i = GC\ T\ T\ AG
\]

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

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

\[
\Pr(X_i \mid Z_{ij} = 1, M, B) = B(G) \times B(C) \times M(1, T) \times M(2, G) \times M(3, T) \times B(A) \times B(G) = 0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.25 \times 0.25
\]

Calculating the Z vector (using M)

- To estimate the starting positions in $Z$ at step $t$
  - 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
      \[
      Z_{ij}^{(t)} = \frac{\Pr(X_i \mid Z_{ij} = 1, M^{(t)}) \Pr(Z_{ij} = T)}{\sum_{k=1}^{L-W+1} \Pr(X_i \mid Z_{ij} = 1, M^{(t)}) \Pr(Z_{ij} = T)}
      \]
    - Assume uniform priors (motif eq likely to start at any position)

Calculating the Z vector: Example

\[
X_i = G\ C\ T\ G\ T\ A\ G
\]

\[
p = \begin{bmatrix}
  A & 0.25 \\
  C & 0.25 \\
  G & 0.25 \\
  T & 0.25
\end{bmatrix}
\]

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

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

- then normalize so that $\sum_{j=1}^{L-W+1} Z_{ij} = 1$
Aside: Simplifying $P(X_i)$

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

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

before motif  motif  after motif

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

constant for each sequence can be stored in a matrix

Motif discovery overview

1. Introduction to regulator motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
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
   - Foreground vs. background. Real vs. control motifs.

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

$$M^{(t+1)}(k,c) = \frac{n_{k,c} + d}{\sum_{c} (n_{k,c} + d)}$$

pseudo-counts

where

$$n_{c,k} = \sum_{i} \sum_{\{j \mid X_{i,j+k-1}=c\}} Z_{ij}$$

$$B^{(t+1)}(c) = \frac{n_{0,c} + d}{\sum_{c} (n_{0,c} + d)}$$

total # of c’s in data set

$$n_{0,c} = \sum_{j=1}^{W} n_{j,c}$$

The EM Algorithm

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

$$\prod_{i} Pr(X_i \mid M, B)$$

- 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.

Motif discovery overview

1. Introduction to regulator y motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo

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
   - Foreground vs. background. Real vs. control motifs.

Three examples of Greedy, Gibbs Sampling, EM

<table>
<thead>
<tr>
<th>Greedy always picks maximum</th>
<th>Gibbs sampling picks one at random (or)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_1$</td>
<td>EM uses both in estimating motif</td>
</tr>
</tbody>
</table>

| $Z_2$                       | All methods agree one big winner       |

| $Z_3$                       | Greedy ignores most of the probability |
|                            | Gibbs sampling rapidly converges to some choice |
|                            | EM averages over the entire sequence (no preference) |

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></td>
<td>Expression clustering</td>
<td>HMM learning</td>
</tr>
<tr>
<td></td>
<td>Assign each point to best label</td>
<td>K-means: Assign each point to nearest cluster</td>
</tr>
<tr>
<td></td>
<td>Pick a best point</td>
<td>Baum-Welch training: label sequence w all paths (posterior decoding)</td>
</tr>
<tr>
<td></td>
<td>Assign each point to all labels, probabilistically</td>
<td>Fuzzy K-means: Assign to all clusters, weighted by proximity</td>
</tr>
<tr>
<td></td>
<td>Average all points</td>
<td></td>
</tr>
</tbody>
</table>

| Sample one | N/A: Assign to a random cluster, sample by proximity | N/A: Sample a single label for each position, according to posterior prob | Gibbs sampling: Use one position for the motif, by sampling from the match scores | Average of those points assigned to label (a sample) |

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
  \( \rightarrow \) 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 \( B \)

- **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)

```
\begin{align*}
\sum_{i=1}^{N} \sum_{k=1}^{W} \log \frac{M(k, X_{i,a_i+k})}{B(X_{i,a_i+k})}
\end{align*}
```

Sampling New Motif Positions

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

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

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

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

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

Advantages / Disadvantages

- Very similar to EM

**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
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

Conservation islands overlap known motifs

In theory, Gibbs Sampling less likely to get stuck a local maxima

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 a local maxima

Gibbs Sampling less likely to get stuck a local maxima

Evolutionary signatures for regulatory motifs

- Start 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

Evolutionary signatures for de novo motif discovery

- Genome-wide conservation scores, motif extension
- Validation of discovered motifs: functional datasets

Evolutionary signatures for instance identification

- Phylogenies, Branch length score ➔ Confidence score
- Foreground vs. background. Real vs. control motifs.

Genome-wide conservation

Evaluate conservation within:

(1) All intergenic regions  13%  2%
(2) Intergenic : coding  13% : 3%  2% : 7%
(3) Upstream : downstream  12:0  1:1

A signature for regulatory motifs
Test 1: Intergenic conservation

Test 2: Intergenic vs. Coding

Test 3: Upstream vs. Downstream

Conservation for TF motif discovery

1. Enumerate motif seeds
   \[
   \begin{array}{c|c|c}
   T & G & C \\ \hline
   \text{gap} & \text{TAG} \\ \end{array}
   \]
   - 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
   \[
   \begin{array}{c|c|c|c|c}
   S & R & T & G & C & Y \\ \hline
   \text{gap} & \text{WTAGR} \\ \end{array}
   \]
   - 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

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM, Gibbs), de novo

2. Expectation maximization: Motif matrix \( Z_{ij} \)
   - 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 \emph{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
   - Foreground vs. background. Real vs. control motifs.
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

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

Motif instance identification

How do we determine the functional binding sites of regulators?

Kheradpour, Stark, Roy, Kellis, Genome Research 2007
### 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 (phylogenetic 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/assembly errors or turnover
  - Distal instances can be difficult to assign to gene

### Computing Branch Length Score (BLS)

**CTCF**

\[
\text{BLS} = 2.23\text{sps (78%)}
\]

- 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

### Branch Length Score → Confidence

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

### 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

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)
**Computing enrichments: background vs. foreground**

- **Background vs. forgeround**
  - 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 enrichment with 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 χ², but better for small numbers)
- Fractions can be made more conservative using a binomial confidence interval

---

**Validation of discovered motif instances**

Use independent experimental evidence
Look for functional biases / enrichments

---

**Increased sensitivity using BLS**

<table>
<thead>
<tr>
<th>MicroRNA motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'UTR</td>
</tr>
<tr>
<td><img src="image1.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

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

---

**Intersection with CTCF ChIP-Seq regions**

- **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

<table>
<thead>
<tr>
<th><img src="image5.png" alt="Graph" /></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

**1. Confidence selects for transcription factor motif instances in promoters and miRNA motifs in 3' UTRs**

---

**Increased sensitivity using BLS**

<table>
<thead>
<tr>
<th><img src="image7.png" alt="Graph" /></th>
</tr>
</thead>
</table>

**1. Confidence selects for functional instances**
Enrichment found for many factors

Barski, et al., Cell (2007)
Lim, et al., Molecular Cell (2007)
Wei, et al., Cell (2006)

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

Comparing ChIP to Conservation

TFs: 67 of 83 (81%) 46k instances
miRNAs: 49 of 67 (86%) 4k instances

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 x 10^-3)
- TFs were more targeted by TFs (P < 10^-10) and by miRNAs (P < 5 x 10^-9)
- TF in-degree associated with miRNA in-degree (high-high: P < 10^-7; low-low P < 10^-4)

Fly regulatory network at 60% confidence

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
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
   - Foreground vs. background. Real vs. control motifs.