Lecture 10: Regulatory genomics

Motif discovery, target prediction, regulatory region dissection
<table>
<thead>
<tr>
<th>Project</th>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Lec</th>
<th>Topic</th>
<th>Read*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>Thu, Sep 7</td>
<td>Introduction</td>
<td>L1</td>
<td>Intro: Biology, Algorithms, Machine Learning, Course Overview</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Sep 8</td>
<td></td>
<td>R1</td>
<td>Recitation 1: Biology and Probability Review</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Tue, Sep 12</td>
<td>Module I: Aligning and Modeling Genomes</td>
<td>L2</td>
<td>Alignment I: Dynamic Programming, Global and local alignment</td>
<td>2,3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Sep 14</td>
<td></td>
<td>L3</td>
<td>Alignment II: Database search, Rapid string matching, BLAST, BLOSUM</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Sep 15</td>
<td></td>
<td>R2</td>
<td>Recitation 2: Deriving Parameters of Alignment, Multiple Alignment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Tue, Sep 19</td>
<td>Foundations</td>
<td>L4</td>
<td>Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms</td>
<td>7,8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Sep 21</td>
<td></td>
<td>L5</td>
<td>Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Sep 22</td>
<td></td>
<td></td>
<td>No classes - student holiday</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mon, Sep 25</td>
<td>Project Intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-507</td>
<td></td>
<td></td>
<td>15,16</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Thu, Sep 28</td>
<td>Module II: Gene Expression and Epigenomics</td>
<td>L6</td>
<td>Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Sep 29</td>
<td></td>
<td>R3</td>
<td>Recitation 3: Affinity Propagation Clustering and Random Forest Classification</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Thu, Oct 3</td>
<td>Foundations</td>
<td>L7</td>
<td>Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr</td>
<td>14,15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Oct 4</td>
<td></td>
<td>L8</td>
<td>Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Oct 5</td>
<td></td>
<td>R4</td>
<td>Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Tue, Oct 6</td>
<td>Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 6</td>
<td></td>
<td>L9</td>
<td>Three-dimensional chromatin interactions: 3C, 5C, HIC, ChIA-Pet</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 7</td>
<td></td>
<td>L10</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 8</td>
<td></td>
<td>R5</td>
<td>Recitation 5: Gapped Motif Discovery, DNaseShape, PBMs, SelEx</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Thu, Oct 10</td>
<td>Foundations</td>
<td>L11</td>
<td>Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
<td>20,21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 10</td>
<td></td>
<td>L12</td>
<td>Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder</td>
<td>20,22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Oct 12</td>
<td>Project Planning: feedback 32D-507 at 4-5pm</td>
<td>R6</td>
<td>Recitation 6: Networks review, Recommendation systems, EHR, PhEwAS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 12</td>
<td></td>
<td></td>
<td>No lec, thanksgiving break</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Oct 15</td>
<td></td>
<td></td>
<td>No lec, thanksgiving break</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 15</td>
<td></td>
<td></td>
<td>No lec, thanksgiving break</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Tue, Oct 17</td>
<td>Module IV: Population Genetics and Disease Genomics</td>
<td>L13</td>
<td>Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Oct 26</td>
<td></td>
<td>L14</td>
<td>Disease Association Mapping, GWAS, organismal phenotypes</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Oct 27</td>
<td></td>
<td>R7</td>
<td>Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Tue, Nov 2</td>
<td>Foundations</td>
<td>L15</td>
<td>Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, IMWAS</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Nov 2</td>
<td></td>
<td>L16</td>
<td>Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Nov 3</td>
<td></td>
<td>R8</td>
<td>Recitation 8: Rare Variants, ExAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Nov 7</td>
<td>Module V: Comparative genomics and evolution</td>
<td>L17</td>
<td>Comparative genomics and evolutionary signatures</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Nov 10</td>
<td></td>
<td>L18</td>
<td>Genome Scale Evolution, Genome Duplication</td>
<td>4,5,7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Nov 14</td>
<td>Norecitation, Veterans Day</td>
<td></td>
<td>No recitation, Veterans Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Nov 16</td>
<td>No recitation, Veterans Day</td>
<td></td>
<td>No recitation, Veterans Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Nov 17</td>
<td>No recitation, Veterans Day</td>
<td></td>
<td>No recitation, Veterans Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Tue, Nov 28</td>
<td>Quiz</td>
<td>L21</td>
<td>Single-cell genomics: technology, analysis, microfluidics, applications, insights</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Nov 30</td>
<td></td>
<td>L22</td>
<td>Mining human phenotypes, PhEwAS, UK Biobank, meta-phenotypes + imputation</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Dec 1</td>
<td></td>
<td>R10</td>
<td>Recitation 10: Project Feedback, results, interpretation, directions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 5</td>
<td></td>
<td>L23</td>
<td>Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Dec 7</td>
<td></td>
<td>L24</td>
<td>Genome Engineering with CRISPR/Cas9 and related technologies</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri, Dec 8</td>
<td></td>
<td>R11</td>
<td>Recitation 11: Presentation Skills - Intro, discussion, Slides, Presentation skills</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 12</td>
<td></td>
<td>L25</td>
<td>Final Presentations - Part I (11am), 32-G8 reading room</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tue, Dec 12</td>
<td></td>
<td>L25</td>
<td>Final Presentations - Part I (1pm), 32-141</td>
<td></td>
</tr>
</tbody>
</table>

* readings refer to chapters in compiled 2016 scribe notes, available in the materials folder on Stellar
** recitation topics will be adjusted to respond to lecture and student needs
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 $\rightarrow$ 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*

**Logo** or **motif**

- TAATTA
- CACGTG
- AGATAAAGA
- TCATTA
Disrupted motif at the heart of FTO obesity locus

Strongest association with obesity

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

Thermogenic stimuli (e.g. cold)

Browning mitochondrial thermogenesis

Restoring motif restores thermogenesis

Lean

Obese
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
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
Approaches to regulatory motif discovery

\{ 
\begin{align*}
\text{Region-based motif discovery} \quad & \\
\text{• Expectation Maximization (e.g. MEME)} & \quad \text{– Iteratively refine positions / motif profile} \\
\text{• Gibbs Sampling (e.g. AlignACE)} & \quad \text{– Iteratively sample positions / motif profile} \\
\text{• Enumeration with wildcards (e.g. Weeder)} & \quad \text{– Allows global enrichment/background score} \\
\text{• Peak-height correlation (e.g. MatrixREDUCE)} & \quad \text{– Alternative to cutoff-based approach} \\
\end{align*}
\}

\{ 
\begin{align*}
\text{Genome-wide} \quad & \\
\text{• Conservation-based discovery (e.g. MCS)} & \quad \text{– Genome-wide score, up-/down-stream bias} \\
\end{align*}
\}

\{ 
\begin{align*}
\text{In vitro / trans} \quad & \\
\text{• Protein Domains (e.g. PBMs, SELEX)} & \quad \text{– In vitro motif identification, seq-/array-based} \\
\end{align*}
\}
Motifs are not limited to DNA sequences

• Splicing Signals at the RNA level
  – Splice junctions
  – Exonic Splicing Enhancers (ESE)
  – Exonic Splicing Surpressors (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
Challenges in regulatory genomics

Regulator

TF/miRNA

Network analysis (upcoming lecture)

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

Motif

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

Sequence specificity

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

Targets

Functional instances

TFs/miRNAs: Evolutionary signatures**
miRNAs: Composition/folding

TFs/miRNAs: De novo comparative discovery**

TFs: Mass Spec (difficult)

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

TFs/miRs: Perturbation response

Network analysis

TFs: ChIP-Chip/ChIP-Seq

Challenges in regulatory genomics

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

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

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

- easy to find starting position probabilities

**Key idea:** Iterative procedure for estimating both, given uncertainty

(learning problem with hidden variables: the starting positions)
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)</th>
<th>Express clustering</th>
<th>HMM learning</th>
<th>Motif discovery</th>
<th>Update model parameters (M step)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick a best</td>
<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>
<td><strong>Greedy:</strong> Find best motif match in each sequence</td>
<td>Average of those points assigned to label</td>
</tr>
<tr>
<td>Average all</td>
<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>
<td><strong>MEME:</strong> Use all positions as a motif occurrence weighed by motif match score</td>
<td>Average of all points, weighted by membership</td>
</tr>
<tr>
<td>Sample one</td>
<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>
<td><strong>Gibbs sampling:</strong> Use one position for the motif, by sampling from the match scores</td>
<td>Average of those points assigned to label(a sample)</td>
</tr>
</tbody>
</table>

The hidden label is:
- Cluster labels
- State path \( \pi \)
- Motif 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 < $\varepsilon$)
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 $k$

\[
M = \begin{pmatrix}
1 & 2 & 3 \\
A & 0.1 & 0.5 & 0.2 \\
C & 0.4 & 0.2 & 0.1 \\
G & 0.3 & 0.1 & 0.6 \\
T & 0.2 & 0.2 & 0.1
\end{pmatrix} \quad (\sim \text{CAG})
\]

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

\[
B = \begin{pmatrix}
A & 0.26 \\
C & 0.24 \\
G & 0.23 \\
T & 0.27
\end{pmatrix} \quad \text{(near uniform)}\]
  \text{(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 $i$.

$$Z = \begin{pmatrix}
seq1 & 0.1 & 0.1 & 0.2 & 0.6 \\
seq2 & 0.4 & 0.2 & 0.1 & 0.3 \\
seq3 & 0.3 & 0.1 & 0.5 & 0.1 \\
seq4 & 0.1 & 0.5 & 0.1 & 0.3
\end{pmatrix}$$

Some examples:

- $Z_1$: No clear winner
- $Z_2$: Two candidates
- $Z_3$: One big winner
- $Z_4$: Uniform
### Starting positions (Z_{ij}) ↔ Motif matrix M(k,c)

#### Motif matrix M(k,c)

<table>
<thead>
<tr>
<th></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>c=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=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>c=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>c=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: Z_{ij}

- **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

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

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

#### 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}
Regulatory genomics: motifs, instances, regions

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

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

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

\[ p = \begin{bmatrix} 0 & 1 & 2 & 3 \\ A & 0.25 & 0.1 & 0.5 & 0.2 \\ C & 0.25 & 0.4 & 0.2 & 0.1 \\ G & 0.25 & 0.3 & 0.1 & 0.6 \\ T & 0.25 & 0.2 & 0.2 & 0.1 \end{bmatrix} \]

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

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

\[ \vdots \]

- then normalize so that

\[ \sum_{j=1}^{L-W+1} Z_{ij} = 1 \]
Denominator: \( P(X_i) \), when motif position is known

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

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

before motif  |  motif  |  after motif

- Example:

\[
X_i = \begin{array}{cccc}
G & C & T & G \\
\end{array}, \quad B = \begin{array}{cccc}
A & 0.25 & B & 0.25 & T & 0.25 \\
C & 0.25 & & & & \\
G & 0.25 & & & & \\
T & 0.25 & & & & \\
\end{array}, \quad M = \begin{array}{ccc}
\begin{array}{ccc}
A & 1 & 2 & 3 \\
C & 0.1 & 0.5 & 0.2 \\
G & 0.4 & 0.2 & 0.1 \\
T & 0.3 & 0.1 & 0.6 \\
\end{array}
\end{array}
\]

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

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

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

before motif  

motif  

after motif

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

can be stored in a matrix  

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

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

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**

- **All methods agree**
- **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

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

pseudo-counts

total # of c’s in data set
**M-step example: Estimating M(k,c) from Z_{ij}**

\[ X_1 = \begin{array}{cccc}
A & C & A & G & C & A \\
Z_1 = & 0.1 & 0.7 & 0.1 & 0.1 \\
\end{array} \]

\[ X_2 = \begin{array}{cccc}
A & G & G & C & A & G \\
Z_2 = & 0.4 & 0.1 & 0.1 & 0.4 \\
\end{array} \]

\[ X_3 = \begin{array}{cccc}
T & C & A & G & T & C \\
Z_3 = & 0.2 & 0.6 & 0.1 & 0.1 \\
\end{array} \]

- **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**
  - \( M(1,A) = \frac{0.7 + 1}{3 + 4} = \frac{1.7}{7} = 0.24 \)
  - \( M(1,C) = \frac{1.7 + 1}{3 + 4} = \frac{2.7}{7} = 0.39 \)
  - \( M(1,G) = \frac{0.4 + 1}{3 + 4} = \frac{1.4}{7} = 0.2 \)
  - \( M(1,T) = \frac{0.2 + 1}{3 + 4} = \frac{1.2}{7} = 0.17 \)

\[ M(1, A) = \frac{Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} + 1}{Z_{1,1} + Z_{1,2} + \ldots + Z_{3,3} + Z_{3,4} + 4} \]

**Em approach:** Avg’em all

**Gibbs sampling:** Sample one

**Greedy:** Select max

\[ \begin{array}{|c|c|c|}
\hline
& 1 & 2 & 3 \\
\hline
A & 0.24 & 0.39 & 0.21 \\
\hline
C & 0.39 & 0.21 & 0.18 \\
\hline
G & 0.2 & 0.24 & 0.44 \\
\hline
T & 0.17 & 0.16 & 0.16 \\
\hline
\end{array} \]
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 \( M \))
P(Seq|Model) Landscape

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
   - 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

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>
</tr>
</thead>
<tbody>
<tr>
<td>Update assignments (E step) ➔</td>
<td>Expression clustering</td>
</tr>
<tr>
<td>Estimate hidden labels</td>
<td>Cluster labels</td>
</tr>
<tr>
<td>E step</td>
<td>K-means: Assign each point to nearest cluster</td>
</tr>
<tr>
<td>Pick a best</td>
<td><strong>Fuzzy K-means:</strong> Assign to all clusters, weighted by proximity</td>
</tr>
<tr>
<td>Average all</td>
<td><strong>N/A:</strong> Assign to a random cluster, sample by proximity</td>
</tr>
<tr>
<td>Sample one</td>
<td><strong>N/A:</strong> Assign to a random cluster, sample by proximity</td>
</tr>
</tbody>
</table>
Three examples of Greedy, Gibbs Sampling, EM

Greedy always picks maximum

Gibbs sampling picks one at random (or)

(and)

EM uses both in estimating motif

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 $\Pr(U_1...U_n)$ by iteratively sampling from
  for each $j$ $\Pr(U_j \mid U_1...U_{j-1}, U_{j+1}...U_n)$
- 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 $B$

- **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:

\[
A_j \propto \Pr(X_i \mid Z_{ij} = 1, p)
\]
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
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.
Regulatory genomics: motifs, instances, regions

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

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

Known engrailed binding site

- 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

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

Transcription factor binding

Increase power by testing conservation in many regions
## Genome-wide conservation

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Gal4</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>All intergenic regions</td>
<td>13%</td>
<td>2%</td>
</tr>
<tr>
<td>Intergenic : coding</td>
<td>13% : 3%</td>
<td>2% : 7%</td>
</tr>
<tr>
<td>Upstream : downstream</td>
<td>12:0</td>
<td>1:1</td>
</tr>
</tbody>
</table>

### Diagram

- Scer
- Spar
- Smik
- Sbay

**A signature for regulatory motifs**
Test 1: Intergenic conservation

Conserved count vs. Total count plot with a focus on the CGG-11-CCG region.
Test 2: Intergenic vs. Coding

Intergenic Conservation

C GG-11-CCG

Higher Conservation in Genes

Coding Conservation

Graph showing the comparison of intergenic vs. coding conservation with a particular focus on CGG-11-CCG and higher conservation in genes.
Test 3: Upstream vs. Downstream

Upstream Conservation vs. Downstream Conservation

Most Patterns

CGG-11-CCG

Downstream motifs?
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
   - 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

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: Known Factors

<table>
<thead>
<tr>
<th>MCS</th>
<th>Discovered motif</th>
<th>Known Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>46.8</td>
<td>GGGCGGR</td>
<td>SP-1</td>
</tr>
<tr>
<td>34.7</td>
<td>GCCATnTTg</td>
<td>YY1</td>
</tr>
<tr>
<td>32.7</td>
<td>CACGTG</td>
<td>MYC</td>
</tr>
<tr>
<td>31.2</td>
<td>GATGGGY</td>
<td>NF-Y</td>
</tr>
<tr>
<td>30.8</td>
<td>TGANrTCA</td>
<td>AP-1</td>
</tr>
<tr>
<td>29.7</td>
<td>GGGAGGGR</td>
<td>MAZ</td>
</tr>
<tr>
<td>29.5</td>
<td>TGACGTMr</td>
<td>CREB</td>
</tr>
<tr>
<td>26.0</td>
<td>CGGCCATYK</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>CATTTCCCK</td>
<td>STAT1</td>
</tr>
</tbody>
</table>

### Table: Mammalian motifs

<table>
<thead>
<tr>
<th>MCS</th>
<th>Discovered motif</th>
<th>Known Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>70/174</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table: Fly motifs

<table>
<thead>
<tr>
<th>MCS</th>
<th>Discovered motif</th>
<th>Known Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>35/145</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

1. Most motifs avoided in ubiquitously expressed genes

2. Functional clusters emerge
Regulatory genomics: motifs, instances, regions

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

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/assembling 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
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
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

Original motif

Genome sequence

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 enrichment 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

\[
\frac{\text{# in foreground}}{\text{size of background}} \div \frac{\text{# in background}}{\text{size of foreground}}
\]

\[
\frac{\text{# in foreground}}{\text{# in background}} \div \frac{\text{# control in foreground}}{\text{# control in background}}
\]

Binomial confidence interval

Use this

0.0
fraction
1.0

Foreground (e.g. TF bound):

Background (e.g. Intergenic):

Fraction of motif instances can be compared to fraction of control motif instances in foreground
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
Confidence selects for functional instances

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

miRNA motifs

- BLS (12 species)
- Full cons. melanogaster group
- Full cons. in D. Pseudoobscura
- Full cons. Sophophora species

TF motifs

- BLS (12 species)
- Full cons. melanogaster group
- Full cons. in D. Pseudoobscura
- Full cons. Sophophora species
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 ≥50% of regions with a motif

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

Mammals

- CTCF
- HNF6
- HNF4A
- HNF1A
- Stat1

- Wei, et al., Cell (2006)
- p53
- RELA
- Myc


Flies

- Zeitlinger, et al., Genes & Devel (2007)
- Twist
- Snail

- Sandmann, et al., Genes & Devel (2007)
- Mef2
- CrebA

- Abrams and Andrew, Devel Cell (2005)

Barski, et al., Cell (2007)


Sandmann, et al., Devel Cell (2005)
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
More enrichment when binding conserved

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.

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 $P < 10^{-6}$)
Regulatory genomics: motifs, instances, regions

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 \textit{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. \textit{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: 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**

Network analysis (next lecture)

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

Targets

Functional instances

TFs: ChIP-Chip/ChIP-Seq
TFs/miRs: Perturbation response

TFs/miRNAs: Evolutionary signatures**
miRNAs: Composition/folding

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

* = 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
Problem: Not all annotated enhancers are real

Luciferase assays
- PCR
- Chromatin mark
- Predicted enhancer
- min
- luc2
- Activity
- pGL4.23

Massively-parallel assays
- Array-based synthesis / library construction
- Transfect
- Predicted enhancer
- minP
- luc2
- Barcode
- pGL4.23

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

Slow, tedious, time-consuming

10,000+ elements at a time

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 \( \Rightarrow \) insert upstream
- Couple each with a barcode \( \Rightarrow \) insert downstream
- Make 10,000s of elements \( \Rightarrow \) plasmids, transfection
- High-throughput test in diff. cell types \( \Rightarrow \) 10k measurements

Application: Test 10,000 variants in 1 experiment

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

CRE

INFB

**CRE**

Cryptic CREB(1) CREB(2) CREB(3) CREB(4)

**INFB**

ATF-2 IRF-3(1) IRF-3(2) p50

c-Jun IRF-7(1) IRF-7(2) RelA
Systematic motif disruption in 2000 regions for 5 activators and 2 repressors in 2 human cell lines

Kheradpour, Kellis et al. Genome Research, 2013
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
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
   - 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

6. De novo dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. **Position offset matters.**
   - 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

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?
Tile differences allow motif discovery

- Increased resolution allows testing of only 30-bp intervals
- *De novo* discovered motifs match known motifs
- Discovery distinguishes activating vs. repressive factors

<table>
<thead>
<tr>
<th>Motif Name</th>
<th>HepG2 Activating</th>
<th>HepG2 Repressing</th>
<th>K562 Activating</th>
<th>K562 Repressing</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATA_known14</td>
<td>1.0</td>
<td>1.0</td>
<td>6.4</td>
<td>1.0</td>
</tr>
<tr>
<td>LMO2_2</td>
<td>1.0</td>
<td>1.0</td>
<td>4.4</td>
<td>1.0</td>
</tr>
<tr>
<td>TAL1_disc1</td>
<td>1.0</td>
<td>1.0</td>
<td>3.0</td>
<td>1.0</td>
</tr>
<tr>
<td>CPHX_1</td>
<td>1.0</td>
<td>1.0</td>
<td>2.7</td>
<td>1.0</td>
</tr>
<tr>
<td>JDP2_2</td>
<td>1.0</td>
<td>1.0</td>
<td>2.4</td>
<td>1.0</td>
</tr>
<tr>
<td>NFE2L2_3</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>HNF4_known9</td>
<td>4.6</td>
<td>0.3</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>NR2F6_2</td>
<td>3.7</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>HNF1B_4</td>
<td>3.5</td>
<td>0.6</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>RXRA_known10</td>
<td>3.4</td>
<td>0.9</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>PPARA_4</td>
<td>3.1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>HNF1A_4</td>
<td>2.8</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>HNF1_4</td>
<td>2.6</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>TCF7L2_disc1</td>
<td>2.4</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>AP1_known4</td>
<td>2.3</td>
<td>1.0</td>
<td>1.6</td>
<td>1.0</td>
</tr>
<tr>
<td>SMARC_disc1</td>
<td>2.2</td>
<td>1.0</td>
<td>1.1</td>
<td>1.0</td>
</tr>
<tr>
<td>CEBPB_known1</td>
<td>2.1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>TLX2_2</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>FOXJ2_4</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>EGR1_disc2</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>RFX2_3</td>
<td>1.0</td>
<td>3.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>RFX5_known9</td>
<td>1.0</td>
<td>2.9</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Regulatory genomics: motifs, instances, regions

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 \textit{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. \textit{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 elmts

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

Elements tested in:
- h1esc
- HepG2
- K562
- Huvec

Elements selected in:
- h1esc
- HepG2
- K562
- Huvec

Array 1 (7,860 regions)
Array 2 (7,860 regions)

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

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

\[A_{r,t,k} \sim N(\mu_{a_r}, \sigma^2_{a_r}) \quad k = 1, \ldots, K\]

Observation is mean of true signals

\[
\mu_{a_r} = \frac{1}{|M_r|} \sum_{m \in M_r} m
\]

Measurements normally distributed

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

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

\[
\sigma^2_{m_{r,t}} = \frac{1}{|M_r|} \sum_{m \in M_r} (m - \frac{1}{|M_r|} \sum_{m \in M_r} m)^2
\]
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

HNF4

TP53

RFX5

HNF1

E2F

HNF4
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
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?)
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 \( \Rightarrow \) 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 input DNA library recapitulates DNase/ATAC-Seq

Fig 1c from Buenrostro et al. Nature Methods 2013

Preferential selection of putative regulatory elements
HiDRA input DNA library: long, active, densely-covered regions

HiDRA DNA library captures more active elements

Fragments: 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

- 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

High-resolution genome-wide functional dissection of transcriptional regulatory regions in human

Xinchen Wang\textsuperscript{1,2,3}, Liang He\textsuperscript{2,3}, Sarah M. Goggin\textsuperscript{2}, Alham Saadat\textsuperscript{2}, Li Wang\textsuperscript{2}, Melina Claussnitzer\textsuperscript{2,4,*}, Manolis Kellis\textsuperscript{2,3,*}

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

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

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
Recitation tomorrow: *in vitro* motif identification

- PBMs: Protein binding microarrays
- SELEX: Selection-based motif identification
- De Bruijn graphs to generate PBM probes
- From k-mers to motifs
- Gapped motifs
- Degenerate motifs and DNA bending (DNA shape)
- Relaxing independence assumptions in PWMs

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

### PBMs (Protein binding microarrays; Mukherjee, 2004)

Double stranded DNA arrays

- From k-mers to motifs
- Gapped motifs
- Degenerate motifs and DNA bending (DNA shape)
- Relaxing independence assumptions in PWMs