An example transcriptional regulatory cascade
Here, controlling *Salmonella* bacteria multidrug resistance

RamR represses the *ramA* gene, which encodes the activator protein for the *acrAB* drug efflux pump genes.
Historically, DNA and RNA binding sites were defined biochemically (DNase footprinting, gel shift assays, etc.)

Now, many binding motifs are discovered bioinformatically

Isolate different nucleic acid segments bound by copies of the protein (e.g. all sites bound across a genome) → Sequence → Search computationally for recurring **motifs**
Transcription factor regulatory networks can be highly complex, e.g. as for embryonic stem cell regulators.

**MOTIFS**

<table>
<thead>
<tr>
<th>TF</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEM13</td>
<td>CCCATTGTTTCG</td>
</tr>
<tr>
<td>HEM13</td>
<td>TTCCTGTTTCG</td>
</tr>
<tr>
<td>HEM13</td>
<td>TCAATTGGTTAG</td>
</tr>
<tr>
<td>ANB1</td>
<td>CTATTGGTTGC</td>
</tr>
<tr>
<td>ANB1</td>
<td>TCCATTGGTTC</td>
</tr>
<tr>
<td>ANB1</td>
<td>CCTATTGGTTC</td>
</tr>
<tr>
<td>ANB1</td>
<td>TCCATTGGTTC</td>
</tr>
<tr>
<td>ROX1</td>
<td>CCAATTGTTCG</td>
</tr>
</tbody>
</table>

Binding sites of the transcription factor ROX1

**consensus**

**frequencies**

**scaled by information content**

\[
I_{\log}(i) = -\sum \frac{f_{bi}}{f_{bi}} \log \frac{f_{bi}}{P_p}
\]

freq of nuc b in genome
So, here’s the challenge:

Given a set of DNA sequences that contain a motif (e.g., promoters of co-expressed genes), how do we discover it computationally?

Could we just count all instances of each $k$-mer?

Why or why not?

$\Rightarrow$ promoters and DNA binding sites are not well conserved
How does motif discovery work?

Sites in target sequences

<table>
<thead>
<tr>
<th>Motif model</th>
</tr>
</thead>
<tbody>
<tr>
<td>AATCAAGTT</td>
</tr>
<tr>
<td>AGGCTGAAAACAAAAGTTTTCGAGTA</td>
</tr>
<tr>
<td>GAGATAACGCCGTGATAGAATCTCATTGCC</td>
</tr>
<tr>
<td>ATATTTCCGAGCTGAGCAGACCCGGTTTGG</td>
</tr>
<tr>
<td>GAACGCCAACAGGACGTTTCACATGAA</td>
</tr>
</tbody>
</table>

Assign sites to motif

Update the motif model

Assign sites to motif

Update the motif model

Assign sites to motif

Update the motif model

etc.
How does motif discovery work?

Motif finding often uses expectation-maximization i.e. alternating between building/updating a motif model and assigning sequences to that motif model.

Searches the space of possible motifs for optimal solutions without testing everything.

Most common approach = Gibbs sampling

We will consider N sequences, each with a motif of length w:

- \( A_k \) = position in seq k of motif
- \( q_{ij} \) = probability of finding nucleotide (or aa) j at position i in motif
  - i ranges from 1 to w
  - j ranges across the nucleotides (or aa)
- \( p_j \) = background probability of finding nucleotide (or aa) j
Start by choosing \( w \) and randomly positioning each motif:

\[ A_k = \text{position in seq } k \text{ of motif} \]

\[ q_{ij} = \text{probability of finding nucleotide (or aa) } j \text{ at position } i \text{ in motif} \]

\[ i \text{ ranges from 1 to } w \]

\[ j \text{ ranges across the nucleotides (or aa)} \]

\[ p_j = \text{background probability of finding nucleotide (or aa) } j \]

NOTE: You won’t give any information at all about what or where the motif should be!

Predictive update step: Randomly choose one sequence, calculate \( q_{ij} \) and \( p_j \) from N-1 remaining sequences

\[ q_{ij} = \text{probability of finding nucleotide (or aa) } j \text{ at position } i \text{ in motif} \]

\[ i \text{ ranges from 1 to } w \]

\[ j \text{ ranges across the nucleotides (or aa)} \]

\[ p_j = \text{background probability of finding nucleotide (or aa) } j \]
Stochastic sampling step: For withheld sequence, slide motif down sequence & calculate agreement with model

Withheld sequence

\[
\frac{\Pi(q_{ij})^{c_{xij}}}{\Pi(p_{ij})^{c_{xij}}}
\]

Position in sequence

(see the paper for details)

Here's the cool part. DON'T just choose the maximum. INSTEAD, select a new \( A_k \) position proportional to this odds ratio.

Then, choose a new sequence to withhold, and repeat everything.
Over many iterations, this magically converges to the most enriched motifs. Note, it’s stochastic:

Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation

- Measure mRNA abundances using DNA microarrays
- Search for motifs in promoters of glucose vs galactose controlled genes

Discovered motifs

Known motif

Galactose upstream activation sequence

"AlignAce"
Measure mRNA abundances using DNA microarrays

Search for motifs in promoters of heat-induced and repressed genes

Discovered motifs

Known motif

Cell cycle activation motif, histone activator

If you need them, we now know the binding motifs for 100's of transcription factors at 1000's of distinct sites in the human genome, including many new motifs.

e.g., http://compbio.mit.edu/encode-motifs/

<table>
<thead>
<tr>
<th>Novel1</th>
<th>Novel2</th>
<th>Novel3</th>
<th>Novel4</th>
<th>Novel5</th>
<th>Novel6</th>
<th>Novel7</th>
<th>Novel8</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1_disc1</td>
<td>EGR1_disc4</td>
<td>SP2_disc3</td>
<td>TATA_disc5</td>
<td>E2F_disc6</td>
<td>AGGCGG_A</td>
<td>TATA_disc7</td>
<td>AGAGAA_A</td>
</tr>
<tr>
<td>CHD2_disc1</td>
<td>TCCCGCA_A</td>
<td>AG_CTGGTT</td>
<td>TRIM28_disc2</td>
<td>SMARC_disc2</td>
<td>AGGCGG_A</td>
<td>TRIM28_disc2</td>
<td>AGAGAA_A</td>
</tr>
<tr>
<td>ETS_disc3</td>
<td>TCCCGCA_A</td>
<td>TCF12_disc3</td>
<td>RFX5_disc3</td>
<td>ZNF143_disc1</td>
<td>TATA_disc5</td>
<td>RFX5_disc3</td>
<td>E2F_disc6</td>
</tr>
<tr>
<td>ETS_disc6</td>
<td>TCCCGCA_A</td>
<td>ZBTB7A_disc2</td>
<td>TATA_disc5</td>
<td>ZNF143_disc2</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>AAATAAA</td>
</tr>
<tr>
<td>NRC1_disc3</td>
<td>SETDB1_disc1</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
</tr>
<tr>
<td>ZBTB33_disc1</td>
<td>STX5_disc1</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
</tr>
<tr>
<td>ZBTB33_disc2</td>
<td>STX5_disc2</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
</tr>
<tr>
<td>ZBTB33_disc3</td>
<td>STX5_disc3</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
</tr>
<tr>
<td>ZBTB33_disc4</td>
<td>STX5_disc4</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
<td>TATA_disc5</td>
</tr>
</tbody>
</table>
Here’s a good place to start if you want to do this practically: http://meme-suite.org/