RamR represses the *ramA* gene, which encodes the activator protein for the *acrAB* drug efflux pump genes.

**An example transcriptional regulatory cascade**

Here, controlling *Salmonella* bacteria multidrug resistance
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>CCCATTGTTCTC</td>
</tr>
<tr>
<td>HEM13</td>
<td>TTTCTGTTTCTC</td>
</tr>
<tr>
<td>HEM13</td>
<td>TCAATTGTTTAG</td>
</tr>
<tr>
<td>ANB1</td>
<td>CTCATTGTTCTC</td>
</tr>
<tr>
<td>ANB1</td>
<td>TCAATTGTTCTC</td>
</tr>
<tr>
<td>ANB1</td>
<td>CCTATTGTTTCT</td>
</tr>
<tr>
<td>ROX1</td>
<td>CCAATTGTTTTC</td>
</tr>
</tbody>
</table>

**Consensus**

**frequencies**

scaled by information content

\[
I_{\text{seq}}(i) = -\sum_{b} f_{b,i} \log_{2} \frac{f_{b,i}}{P_b} \\
\text{freq of nuc b in genome}
\]
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?

→ promoters and DNA binding sites are not well conserved
How does motif discovery work?

Sites in target sequences

Motif model

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:

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

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

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

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

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

Odds ratio of agreement with model vs. background

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

(see the paper for details)

Here’s the cool part: DON’T just choose the maximum. INSTEAD, select a new \( A_k \) position \textit{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:

3 runs on the same data

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

Galactose vs. glucose

Measure mRNA abundances using DNA microarrays

Search for motifs in promoters of glucose vs galactose controlled genes

Discover motifs

Known motif

Galactose upstream activation sequence

"AlignAce"
Heat shock vs. 30 °C

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/
Here’s a good place to start if you want to do this practically:  http://meme-suite.org/

Note: online MEME suite can sometimes be quite laggy. GibbsCluster is a good alternative for peptide motifs:  https://services.healthtech.dtu.dk/service.php?GibbsCluster-2.0

Both can also be installed on your own computer