Lecture 7: DNA sequence motif identification
Several approaches to identify regulatory elements:

1. **Combine gene expression and sequence motif occurrence**
   If a TF is activated, its target genes should have significant expression changes *(REDUCE, MDScan, Motif Regressor)*

2. **Search for enriched motifs in a pre-selected group of promoter sequences.**
   Target genes of a TF have the “same” regulatory element *(MEME, a dimer method)*

3. **Comparative genomics**
   TF binding sites are more conserved *(functional constraint)* *(Kellis, Cliften)*

4. **Methods based on cluster of motifs**
   TFs work together and their binding sites tend to be close
Several approaches to identify regulatory elements:

1. **Methods based on cluster of motifs**
   
   TFs work together and their binding sites tend to be close

   Berman, Markstein, Frith, Rajewsky, Sihna
Strategy:

1. Assemble a list of motifs
2. Identify matches of each motif
3. Find clusters of motifs and learn grammar
4. Experimental validation
Protocol (Berman et al.):

- Align binding sequences of 5 TFs (Bcd, Cad, Hb, Kr and Kni) determined by experiments
- Build 5 PSFM\s and find matches
- Find clusters (\geq 13 matches in 700 bp)
- Collapse overlapping clusters

Bicoid (Bcd)

Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>19</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>45</td>
<td>25</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>5</td>
<td>34</td>
<td>21</td>
</tr>
</tbody>
</table>

Sequence Logo

Sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Target Gene</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>catcggCTAACCTCccg</td>
<td>hb</td>
<td>(1)</td>
</tr>
<tr>
<td>tgcggCTAACCTGccc</td>
<td>hb</td>
<td>(1)</td>
</tr>
<tr>
<td>gatttggATGATCGggg</td>
<td>hb</td>
<td>(1)</td>
</tr>
<tr>
<td>gtctataacccTTATCCCCaagttact</td>
<td>hb</td>
<td>(1)</td>
</tr>
<tr>
<td>ttctgtCTATCCTGgaatgg</td>
<td>hb</td>
<td>(1)</td>
</tr>
<tr>
<td>ccgcgCTAATCCGata</td>
<td>hb</td>
<td>(1)</td>
</tr>
<tr>
<td>cccatCTAATCCCTtgcacgc</td>
<td>hb</td>
<td>(1)</td>
</tr>
<tr>
<td>catgCTAATCCTGatgactga</td>
<td>hb</td>
<td>(1)</td>
</tr>
<tr>
<td>ctgaacCTAAATCGgcct</td>
<td>kr</td>
<td>(2)</td>
</tr>
<tr>
<td>aaatTTAATCCGttcct</td>
<td>kr</td>
<td>(2)</td>
</tr>
<tr>
<td>gacaaATATCCAgcct</td>
<td>kr</td>
<td>(2)</td>
</tr>
<tr>
<td>gtctgTTATCTCgggcc</td>
<td>kr</td>
<td>(2)</td>
</tr>
<tr>
<td>ctatgtgaTTTAGCTT</td>
<td>kr</td>
<td>(2)</td>
</tr>
<tr>
<td>ttctTAAATCCGttcgtg</td>
<td>kr</td>
<td>(2)</td>
</tr>
</tbody>
</table>
Results

- Examined novel clusters with 15 binding sites (or more) per 700bp.
- Identified 28 clusters that met this criteria - these sites contain binding sites for at least two of the factors.
  - 23 fall in upstream regions
  - 3 fall in intron regions

Sensitivity = # recovered clusters/# of known clusters
Specificity = # recovered clusters/# of total identified clusters
Results

- Examined the 49 genes that could be regulated by these sites

- Ten of the 28 sites were upstream of the first intron of anterior-posterior pattern expressed genes.
  - ~35% correct predictions
The Giant (Gnt) gene

expression of endogenous giant mRNA

expression of enhancer-lacZ fusion mRNA

anterior posterior
Conclusions

- Clustering can be used to successfully determine cis-regulatory elements and can be applied to other systems.
- Clustering is more efficient when done using prior knowledge of transcription factor binding site(s).
- Computational identifications of cis-regulatory DNA regions improves when using two or more different classes of recognition sequences (motifs).
- The grammar of the cis-regulatory code is clearly more complex than simply the density of transcription factor binding sites.
Dorsal Transcription Factor

- Drosophila transcription factor involved in dorsal-ventral patterning in development.

- Transcription can be inhibited or induced by a TF, Dorsal, depending on the promoter. Also, transcription induction is concentration dependent.

Markstein et al., 2002, 99, 2, 763-768
A. Frequency of clusters of 2, 3 and 4 Di binding sites in windows of 1000 bp vs. 400 bp

Di sites searched: GGGWWWCCM and GGGWDWWWCCM

B. Statistical analysis of observed clustering

<table>
<thead>
<tr>
<th>Window size</th>
<th>Cluster ≥ 2 sites</th>
<th>Cluster ≥ 3 sites</th>
<th>Cluster ≥ 4 sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exp</td>
<td>obs</td>
<td>σ</td>
</tr>
<tr>
<td>400 bp</td>
<td>266</td>
<td>327</td>
<td>3.7</td>
</tr>
<tr>
<td>1000 bp</td>
<td>651</td>
<td>676</td>
<td>1.0</td>
</tr>
</tbody>
</table>

- Use a degenerate Dorsal consensus sequences to scan entire Drosophila genome.
C. Distribution of DI sites at Sog, Ady, and Phm

Results
sog expression pattern

described as follows:

- **endogenous**
- **sog-LacZ 6-kb intron 1**
- **sog-LacZ 293-bp intron 1**
Ady and Phm expression patterns
Results

- Computational searches successfully identified genes that are activated at high (Phm), intermediate (Ady), and low (Sog) levels of Dorsal.

- At least 33% are known, or indicated, to be regulated by dorsal (5/15).
Limitations:

1. Require good prior knowledge (combination of motifs, PSFMs)

2. Weak sites may be ignored.

3. Window size matters

4. Need annotation of the genome (promoter region)

5. No grammar is learned
Summary:

1. Assemble a list of motifs
2. Identify matches of each motif
3. Find clusters of motifs
4. Experimental validation
Hidden Markov Model based methods (Ahab, Stubb)

1. Weak sites are correctly considered.

2. Take motif correlation into account (orientation and spacing)

3. Include phylogenetic information


Frith et al.,
Bioinformatics, 17,10, 878, 2001
## Identification of regulatory elements:

<table>
<thead>
<tr>
<th>Method</th>
<th>Inputs</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>REDUCE</td>
<td>Expression/ChIP-chip + promoter</td>
<td>Consensus motif</td>
</tr>
<tr>
<td>MDScan</td>
<td>ChIP-chip + promoter</td>
<td>PSFM</td>
</tr>
<tr>
<td>Motif Regressor</td>
<td>Expression/ChIP-chip + promoter</td>
<td>PSFM</td>
</tr>
<tr>
<td>MEME</td>
<td>Pre-selected sequences</td>
<td>PSFM</td>
</tr>
<tr>
<td>A dimer method</td>
<td>promoter</td>
<td>PSFM</td>
</tr>
<tr>
<td>Kellis, Cliften, FastCompare</td>
<td>Promoters of orthologs</td>
<td>Consensus motif</td>
</tr>
<tr>
<td>Markstein, Berman</td>
<td>Motifs + promoters</td>
<td>cis-regulatory modules</td>
</tr>
<tr>
<td>Ahab, Stubb</td>
<td>Motifs + (orthologous) promoters</td>
<td>cis-regulatory modules (grammar)</td>
</tr>
</tbody>
</table>
Critical issues in understanding transcriptional regulation:

Define the source code (deployment of regulatory elements in the promoter region):

- Identify regulatory elements (also called binding site or binding motif of transcription factors).
- Determine the target genes of transcription factors (TFs).
- Determine the regulatory logic delivered by combinatorial regulation of TFs.

How the source code is executed depends on:

- Which TFs are activated.
- The mode of co-regulation of TFs, e.g. competition, synergy, …

Understand transcriptional regulatory networks:

- Characterize the topologies of networks
- Identify network motifs and relate them to biological functions
- Study the dynamics of networks
Readings:

   Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved
   in pattern formation in the Drosophila genome

   Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the
   Drosophila embryo.