In silico cis-analysis

promoter analysis

- Promoters and cis-elements
- Searching for patterns
- Searching redundant patterns
What is a promoter? and why care about it?

- Activator
- Repressor
- RNA pol II
- Mediator
- TFIID
- TFIIA
- TFIIIB
- TFII E
- TFII H
- TATA
- GENE
- CDS
A little about these cis-elements

- Transcription Factors (TF) often bind to them
- They are normally from 5-10 bp long
- They are often placed in the region from TSS and upstream
- Often they come in clusters *i.e.* must be placed in some ‘syntax’ to be functional
- We assume, they are shared by the promoters in a regulon
- They are not always conserved 100%
A cis-element
as it might be seen by a TF
Sound hard to find?

it is hard
Some tricks makes it possible

Regulon (cluster)

Comparing promoters from a regulon to all other promoters
  *i.e.* using a negative set (all other promoters in species X)
  This allows us to use hypergeometric statistics

Ranked list of genes

The distribution of sequences with a given pattern along a rank
  *i.e.* is the pattern overrepresented in promoters with low (or high)
  p-values in a microarray experiment
  This allows us to use Kolmogorov-Smirnoff
But the elements may not always be 100% conserved

If we only had a weight matrice to describe our pattern we could find less conserved patterns.

With a Gibbs sampler we can work backwards:
  - We give it promoters and it builds a weight matrice.

**Motif model (residue frequency x 100):**

<table>
<thead>
<tr>
<th>POS</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>89</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td>92</td>
<td>.</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
<td>94</td>
<td>.</td>
<td>1.2</td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td>92</td>
<td>.</td>
<td>.</td>
<td>2.0</td>
</tr>
<tr>
<td>5</td>
<td>94</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1.2</td>
</tr>
<tr>
<td>6</td>
<td>94</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1.2</td>
</tr>
</tbody>
</table>
Let’s try it
Saco-patterns and Gibbs sampler

Go to the demo server, linked from the course page
  • Try the ‘saco_patterns’ program
  • Use the 17 arabidopsis promoters as input
  • Try different pattern lengths

Which of the two statistical options is meaningful to use?
Do you find anything that you have seen before?

Try to use the Gibbs sampler too…
  • Looking for 6bp words
  • Try different settings

Do you find the same pattern every time you run the gibbs sampler?
Do you find anything that you have seen before?