cisExpress

Online tool to find promoter motifs

Participants:
Dasha Balashova, Elena Polyakova

Scientific Supervisor:
Tatiana V. Tatarinova

INSTITUTE OF BIOINFORMATICS
15.12.2018
# Arabidopsis thaliana

<table>
<thead>
<tr>
<th>Condition</th>
<th>cisExpress</th>
<th>Position</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best 5-nt consensus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drought</td>
<td>CACGT</td>
<td>-110 ... -60</td>
<td>$10^{-14}$</td>
</tr>
<tr>
<td>Heat</td>
<td>CTAGA</td>
<td>-70 ... -50</td>
<td>$10^{-2}$</td>
</tr>
<tr>
<td>Cold</td>
<td>CTATA</td>
<td>-50 ... -15</td>
<td>$10^{-34}$</td>
</tr>
<tr>
<td>Roots</td>
<td>TCTAT</td>
<td>-40 ... -20</td>
<td>$10^{-21}$</td>
</tr>
<tr>
<td>Seeds</td>
<td>CATGC</td>
<td>-80 ... -44</td>
<td>$10^{-9}$</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>AGGCC</td>
<td>-110 ... -50</td>
<td>$10^{-18}$</td>
</tr>
<tr>
<td>Strength</td>
<td>GGCCC</td>
<td>-110 ... -50</td>
<td>$10^{-11}$</td>
</tr>
<tr>
<td>Variability</td>
<td>TATAA</td>
<td>-50 ... -10</td>
<td>$10^{-140}$</td>
</tr>
<tr>
<td>Flowers</td>
<td>CTATA</td>
<td>-40 ... -20</td>
<td>$10^{-14}$</td>
</tr>
<tr>
<td>Leaves</td>
<td>CTTAT</td>
<td>-40 ... -20</td>
<td>$10^{-20}$</td>
</tr>
<tr>
<td>Light</td>
<td>CCGCG</td>
<td>-110 ... -90</td>
<td>$10^{-2}$</td>
</tr>
</tbody>
</table>
Challenges

Main Challenges

- cisExpress.org doesn’t load. Hard to see UI implementation.
- Outdated C++ libraries

Proposed Solution

- Redesign and develop UI prototype according to described requirements.
- Develop algorithms on Python

Assumptions

- Running the tool on large dataset will take take a long time
Layout
Implementation

Prototype Link

Git
# Table: motifs, position, z-score

<table>
<thead>
<tr>
<th>Motifs</th>
<th>Cluster Groups</th>
<th>Position</th>
<th>Z-Score</th>
<th># Genes</th>
<th>Expression Values</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGSC</td>
<td>Cluster 1</td>
<td>56-89</td>
<td>30.5357</td>
<td>740</td>
<td>8.87</td>
<td>4.83e-53</td>
</tr>
<tr>
<td>AAKCTAG</td>
<td>Cluster 1</td>
<td>-64-02</td>
<td>7.09288</td>
<td>700</td>
<td>6.35</td>
<td>5.8e-233</td>
</tr>
<tr>
<td>GATCTAG</td>
<td>Cluster 1</td>
<td>293-334</td>
<td>8.64519</td>
<td>680</td>
<td>4.87</td>
<td>3.2e-233</td>
</tr>
<tr>
<td>TGSCG</td>
<td>Cluster 2</td>
<td>40-84</td>
<td>7.08511</td>
<td>790</td>
<td>8.87</td>
<td>3.2e-233</td>
</tr>
<tr>
<td>TATAAA</td>
<td>Cluster 2</td>
<td>-47-12</td>
<td>7.32011</td>
<td>690</td>
<td>7.58</td>
<td>8.3e-235</td>
</tr>
<tr>
<td>JAG2G2CECA</td>
<td>Cluster 2</td>
<td>-94-73</td>
<td>6.58881</td>
<td>680</td>
<td>6.67</td>
<td>8.3e-235</td>
</tr>
<tr>
<td>TSCAA</td>
<td>Cluster 3</td>
<td>40-85</td>
<td>6.35811</td>
<td>790</td>
<td>8.87</td>
<td>4.7e-235</td>
</tr>
<tr>
<td>TIGAC</td>
<td>Cluster 3</td>
<td>40-84</td>
<td>6.35811</td>
<td>690</td>
<td>7.56</td>
<td>4.7e-235</td>
</tr>
<tr>
<td>TISAG</td>
<td>Cluster 3</td>
<td>30-82</td>
<td>6.35811</td>
<td>500</td>
<td>6.67</td>
<td>4.98e-235</td>
</tr>
</tbody>
</table>

# Summary: motifs with highest z-score, clusters of the motifs with highest z-score compare motifs through the clusters

**Motifs with Highest Z-Score**

- ATGSC
- GATCTAG
- AAKCTAG
- JAG2G2CECA
- TSCCA
- TIGAC
- TISAG
- TGGCA
- TAGAC
- TGGCA
- TTGAC

**Clusters of the Motifs with Highest Z-Score**

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Algorithm

Assumptions

Function of promoter motifs is position-specific.

Gene expression data provides reasonable measurements of transcript abundance and reflect promoter activity.

It can be in form of microarray or RNA-seq experiments

Stages

Finding “seed” motifs.

Optimizing the motifs obtained by the first part of the method.
Algorithm

Initial data processing

\[ Z_{\text{score}}(w, k) = \frac{d_{\text{with}}(w, k) - d_{\text{without}}(w, k)}{\sqrt{\frac{\text{Stdev}^2_{\text{with}}(w, k)}{n_{\text{with}}(w, k)} + \frac{\text{Stdev}^2_{\text{without}}(w, k)}{n_{\text{without}}(w, k)}}} \]

Merging similar motifs

\[ \text{AGGCC} \rightarrow \text{[AG][CG]CCA} \]

Clustering
Initial data processing

Hidden Markov Models

- separate HMMs for each gene cluster
- the set of HMMs as a discriminant model for unknown gene function prediction