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

Condition	*cisExpress*		
	Best 5-nt consensus	**Position**	***P*-value**
Drought	CACGT	−110 … −60	10^{-14}
Heat	CTAGA	−70 … −50	10^{-2}
Cold	CTATA	−50 … −15	10^{-34}
Roots	TCTAT	−40 … −20	10^{-21}
Seeds	CATGC	−80 … −44	10^{-9}
Nitrogen	AGGCC	−110 … −50	10^{-18}
Strength	GGCCC	−110 … −50	10^{-11}
Variability	TATAA	−50 … −10	10^{-140}
Flowers	CTATA	−40 … −20	10^{-14}
Leaves	CTTAT	−40 … −20	10^{-20}
Light	CCGCG	−110 … −90	10^{-2}

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Outdated C++ libraries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**

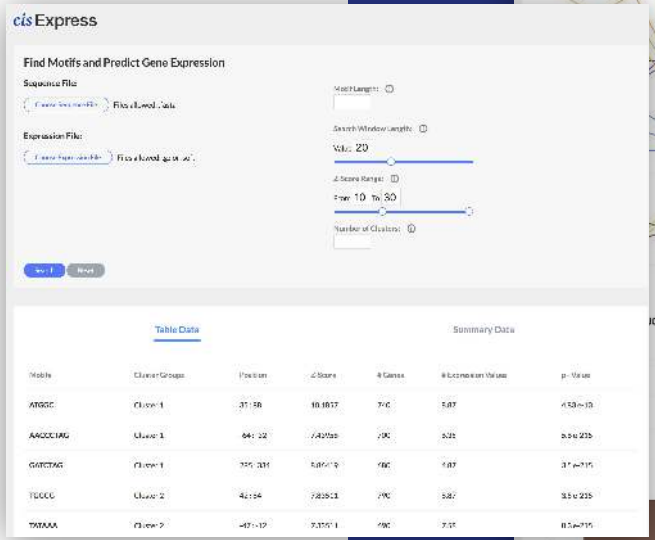
Assumptions

Running the tool on large dataset will take take a long time

Layout Implementation

[Prototype Link]

[Git]



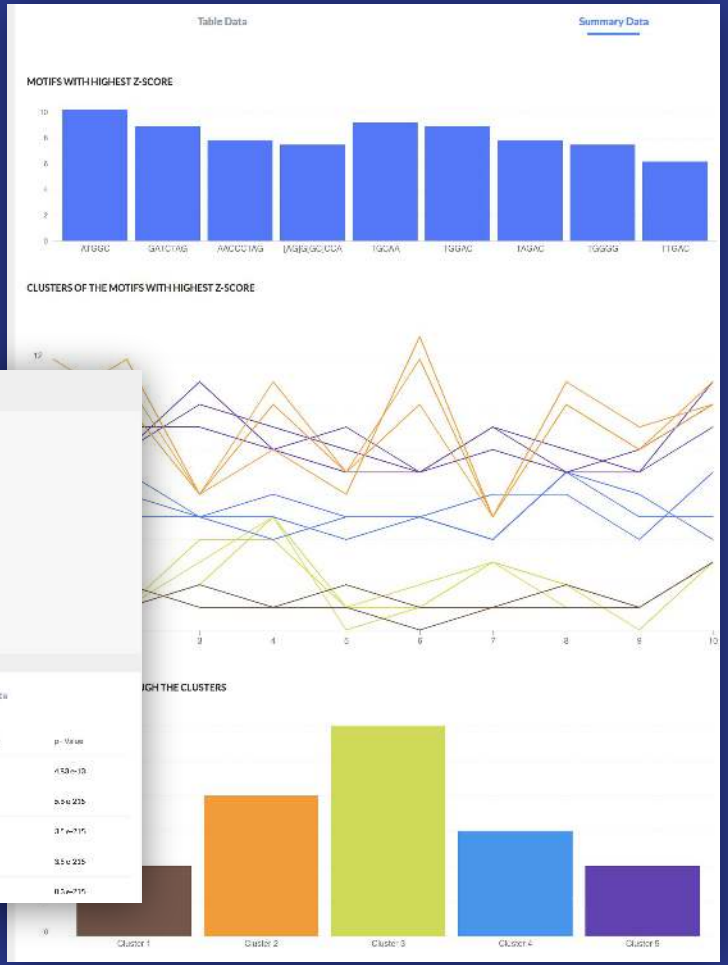
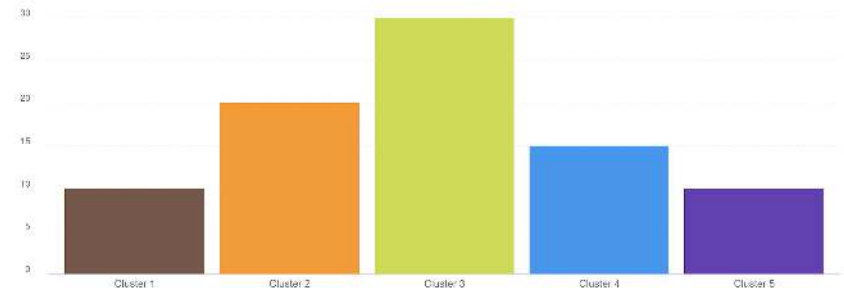


Table:
motifs, position, z-score

Table Data Summary Data

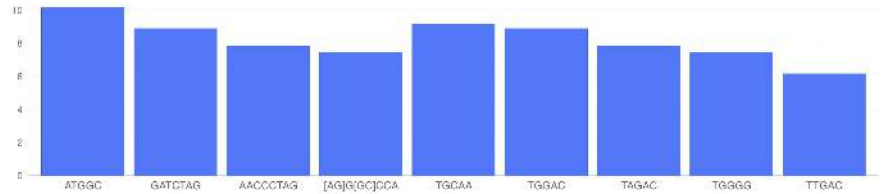
Motifs	Cluster Groups	Position	Z-Score	# Genes	# Expression Values	p-Value
ATGGC	Cluster 1	35:68	10.1857	740	8.87	4.83e-15
AACCCTAG	Cluster 1	-64:-57	7.43958	700	6.35	5.6e-215
GATCTAG	Cluster 1	285:334	8.86419	680	4.87	3.5e-213
TGGCG	Cluster 2	42:84	7.83511	790	8.87	3.5e-213
TATAAA	Cluster 2	-47:-12	7.33511	690	7.58	9.3e-213
[AG]G[GC]CCA	Cluster 2	-94:-79	6.13991	660	6.57	9.3e-213
TGCAA	Cluster 3	40:65	6.13511	720	8.87	4.7e-215
TGGAC	Cluster 3	42:84	5.13511	690	7.58	4.7e-215
TGGAG	Cluster 3	56:82	4.13511	500	6.47	4.58e-215

COMPARE MOTIFS THROUGH THE CLUSTERS

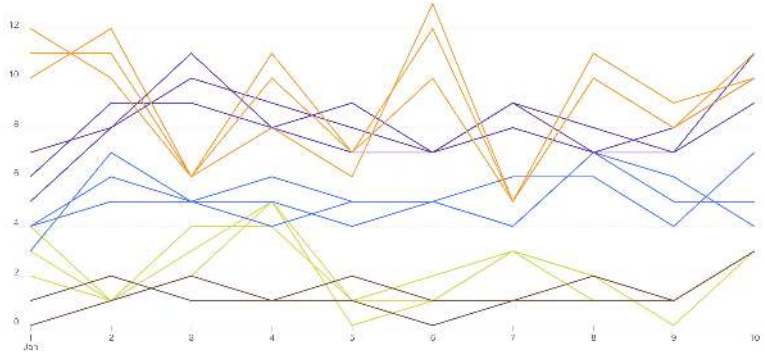


Summary:
motifs with highest z-score,
clusters of the motifs with highest z-score
compare motifs through the clusters

MOTIFS WITH HIGHEST Z-SCORE



CLUSTERS OF THE MOTIFS WITH HIGHEST Z-SCORE



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

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

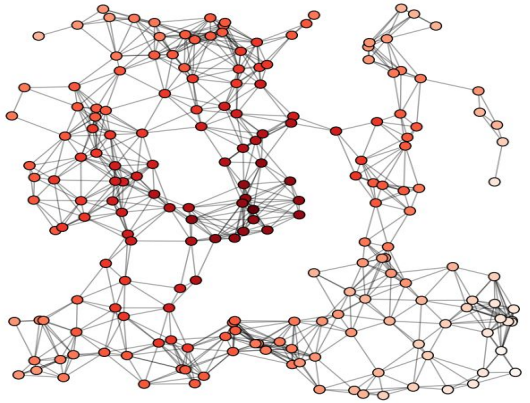
Merging similar motifs

AGGCC

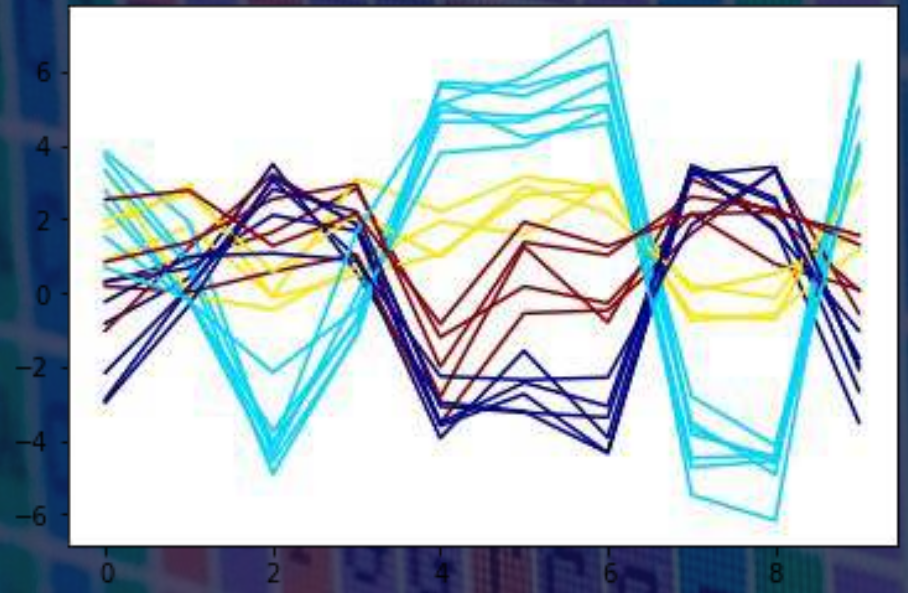
GCCCA

→ [AG]G[CG]CCA

Clustering



Time series



Hidden Markov Models

separate HMMs for each gene cluster

the set of HMMs as a discriminant model for unknown gene function prediction