CAGT: Clustering AGregation Tool v0.1

Max Libbrecht
August 4, 2010

Introduction

CAGT is a tool for visualizing and analyzing large set of signal profiles. It was developed to view the histone profile neighborhoods around transcription factors. It uses a clustering algorithm to split the data into similar sets, then uses an assortment of graphical tools to display

Installation

To run it, you'll need rpy2 and pycluster. You can get rpy here:

`http://rpy.sourceforge.net/rpy2.html`

and pycluster here:

`http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm`

Other than that, all the other libraries used ship with python.

To run, type

```
$ python cagt.py
```

This will cluster all the profiles and put a number of figures in `cagt/output/`. See below for information on configuring parameters.

Parameters

The code takes a number of parameters, which are in the `parameters.py` file in the top level folder (`cagt/`). The function of the parameters are described in the file.

Also, the code automatically saves its progress so you don't have to start over if you want to restart it. If you want to use the old clustering, find the lines in `histone_clustering` that calls `k_cluster`, and comment out the line where `k_cluster` is called, and the line that pickles it The line after that unpickles the old result. You can do the same with the line that calls `get_histone_data`. The lines in question are well-marked in the code.

File Format

CAGT expects a data file in the path specified by the `filename` parameter in `parameters.py`. The data file should have the following (ASCII) format:

```
value<TAB>value<TAB>...<TAB>value<NEWLINE>
value<TAB>value<TAB>...<TAB>value<NEWLINE>
```

There must be the same number of values per line, but there may be any number of lines. Each line is taken to be a separate profile, with the values being the signal at each location.