Prof. Dr. Daniel Huson
Anna Gorska
Zentrum für Bioinformatik
Fachbereich Informatik
Mathematisch-Naturwissenschaftliche Fakultät

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Advanced Java for Bioinformatics — WS 2015/16

Due: 11 November 2015

1 Clustering Viewer (10 points)

A word about CD-HIT

`CD-HIT`[1] is a popular program used for clustering of biological sequences. First, it reversely sorts the sequences by their length. Next it takes the longest sequence and creates the first cluster. The first sequence to be assigned to the cluster is its representative. The algorithm goes through the list and for each sequence decides whether to include it into the current cluster. Sequence is included if its similarity to the representative exceeds a user-specified threshold. When the sequence is included into the cluster it is removed from the list. Subsequent clusters are created analogously: the longest sequence of the current list is taken as the representative, the list is scanned and the sequences are taken out. Program finishes when the list is empty.

[1] Li, W., \& Godzik, A. (2006).
Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.
Bioinformatics (Oxford, England), 22(13), 1658,Äì9. doi:10.1093/bioinformatics/btl158

Task

Your task is to write an `Clustering Viewer` using a `TreeTableView`. Representative sequences serve as root nodes. When collapsed only statistics about the representative sequence is presented.

Read in files with sequences (`.fasta`) and clustering (`.clsr`). Get the `.clsr` filename using `FileChooser`. The `.fasta` file should be found automatically, expect it to be in the same location, with .fasta extension and without percentage annotation. The reason is there might be more then one clustering per single `fasta` file, e.g. with various thresholds. Possible file name pair: `staph_aur_aur_16S.fasta, staph_aur_aur_16S_90.clsr`. Inform the user if you can't read or find files.

Your viewer should have columns: SequenceId, Strain, Sequence Length, Sequence Similarity (for the representative it should be 100%). The sequences you can find in the fasta file, and the clustering in the `.clstr`.

The exemplary fasta file contains 16S rRNA sequences from *Staphylococcus Aureus* downloaded from `Silva` database. Sequence headers look like that:

```
>AB680132.1.1477 Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Staphylococcus;
Staphylococcus aureus subsp. aureus
>CP003979.483035.484865 Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Staphylococcus;
Staphylococcus aureus subsp. aureus CN1
```

Fragment of the CD-HIT output:

```
>Cluster 0
0       1477nt, >AB680132.1.1477... at +/100.00%
1       1477nt, >AB681291.1.1477... at +/100.00%
..
31      1233nt, >CP003045.2150855.21... at +/100.00%
32      1831nt, >CP003979.483035.484... *
```

As you see the CD-HIT output truncates sequence IDs. Therefore, you should also truncate IDs until the first dot, therefore from the example: AB680132 and CP003979. Strain is the last part of the path, for the example: *Staphylococcus aureus subsp. aureus* and *Staphylococcus aureus subsp. aureus CN1*. Sequence length can be taken from the clustering file or computed using a sequence in the fasta file. The cluster representative sequence is the one with * instead of similarity percentage. There should be exactly one in each cluster.

Not to restrain your imagination there are no screen-shots this time. Try to style your `TreeTableView` a little bit, so it is nice to use.