Applied Bioinformatics DD2404 - Reducing noise in protein multialignments

Hanna Hassan, hanhas@kth.se

2014.12.17

Project chosen: Reducing noise in protein multialignments

We have started with mapping out how the first part of the project, reducing the noise, should be approached in code.

The first step we took was to make the program be able to read the input file. We have also made a dictionary containing all the amino acids which we have thought to be a necessity to be able to apply the noise conditions. A counting function has been made (count_character) to count occurences of the different amino acids in the input file's aligned columns. Code which takes out the first column in an input sequence and put it into a string has also been written.

The coming steps in the noise-reduction part is to write code for the noise conditions in the same definition and make them True if the condition is met.

This route was decided upon to make it easier to remove columns, in a later function, with our true/false trial all the input columns will go through.

2014.12.18

The noise-reduction steps have been written and are working. The columns in the multialignments are removed if the column fulfills:

- there are more than 50% indels,
- at least 50% of amino acids are unique,
- no amino acid appears more than twice.

Control experiments which were tested made us take care of the cases: the file not being able to open correctly, if all the columns are noisy (in other words all of them have been removed) and if the input file is empty.

The result we get now is after the file is read, the noisy columns are removed and the sequences without noisy columns are written out. Some corrections need to be made before heading on to the next part of the project, inferring trees.

2014.12.19

Today, some corrections were made regarding how the output of the program looks. The empty lists from the lasts session have been fixed. Names of the sequences have also been added and the output is now written in a FASTA-format.

A control for if the column contains a character which is not an amino acid was written.

Small fixes may be needed to work on next time before inferring trees for the alignments.

2015.06.15

Made small fixes on the functions, making sure they work with different control-files and that error messages are displayed.

2015.12.18
Complications, project cannot be finished due to FastPhylo and DendroPy not being able to be installed on the school computers. Need to wait for this to be fixed.

2016.09.24
Problems with DendroPy and FastPhylo still exist, started writing on the introduction part of the report.

2016.10.07
Tried the professors suggestions of changing download directory. FastPhylo and DenroPy packages work and can now be installed on the school computers. Trees can now be inferred and the program can be finished

2016.11.04
Compared results when BIONJ was used to FNJ. Settled for using the FNJ option.

2016.11.10
Started and writing introduction and materials and methods section of the report.

2016.11.20
Read through what has been written and wrote controls.

2016.12.03
Added small fixes to the report such as finishing up the controls section and abstract section.

2016.12.04
The project was finished. Code and report was double-checked and necessary files were uploaded to GitHub.