Bioinformatics, computer lab IX
University of Potsdam WS 2019/2020, S. Hartmann

Taxonomic assignment of sequence fragments

You have used the command-line version of BLAST before, today you'll use the online version. This is a good opportunity to review the material covered in the BLAST lecture, including how the statistical significance of the result is computed and how BLAST results are interpreted.

A second objective is of course to assign (if possible) a taxonomic group to a sequence based on a BLAST result.

The data

Two sequences that were obtained from a project on the human microbiome have been made available on the Moodle course website.

1 NCBI BLAST

Start on the main NCBI website (`http://www.ncbi.nlm.nih.gov/`) and navigate to the BLAST site as discussed in class. Alternatively, you can directly start from the main BLAST site at `http://blast.ncbi.nlm.nih.gov/`.

☞ Under the heading 'Web BLAST', you can choose the BLAST program you want run: click on 'nucleotide blast'.

☞ Use the first DNA sequence as a query: Paste it into the online form as discussed during lecture.

☞ Specify the database for the search: under "Choose Search Set", "Database", click on "rRNA/ITS databases" and then select "16S ribosomal RNA sequences (Bacteria and Archaea)".

☞ Leave all other settings unchanged and execute the search by clicking on the button "BLAST".

☞ Look at the list of database hits and at some of the alignments. This looks a little different than the command-line result, but it serves the same purpose.

☞ In the summary table of the results, you see columns you also saw when you used the command-line BLAST: database sequence ID of the hit, an alignment score, and the E-value. In the online version you also see a column "Query cover(age)". This is the percentage of the query that is covered in the alignment with the database sequence. Can (should?) this information be used to evaluate the quality of a hit?

☞ Look at the first (best) few alignments between the query sequence and database sequences: how long are the alignments? How long are the database sequences, and which regions are

covered in the alignment? (How long is the query sequence, and which region is covered in the alignment?)

☞ From which prokaryotes are the best database matches to the query sequence? Is it easy to decide from which lineage or species/strain the sample sequence was derived?

☞ In the "Taxonomy" tab you'll find a list, sorted by bit score, of the taxonomic lineages and organisms to which the database matches belong. Are they all from the same lineage? Again: Is it easy to decide from which lineage or species/strain the sample sequence was derived?

☞ Back to the header: under "Other reports", follow the link "Distance tree of results". This shows a tree that is based on pairwise sequence comparisons and not on a multiple sequence alignment. Nevertheless, it gives you a first idea of the rRNA sequences (and species) to which the query sequence might be closely related.

☞ To make the tree more compact, some of the clades (groups) have been collapsed into green triangles: you can expand these clades to show all sequences/species that they contain by clicking on (or mousing over) a triangle and then selecting "expand" in the pop-up menu.

☞ Based on the information presented here: from which organism might the sequence come? How confident are you about this result? What kind of follow-up analyses would you suggest?

☞ Repeat the analysis for the second sequence.

☞ Two general approaches for identifying sequences from a metagenomics study were covered in class. Which one of these did you use during this exercise?

The first sample sequence is most likely derived from *Clostridioides difficile* (also known as *Peptoclostridium difficile* or *Clostridium difficile*), and the second sample sequence from *Lactobacillus delbrueckii*, possibly subspecies *bulgaricus*.

☞ Do a quick web search to find out more about these bacteria: to which group do they belong? what are their functions in the human gastrointestinal tract? Are they indicative of a healthy person?

2 Optional: Microbes of our daily lives

There are many projects about the human microbiome; here are links to the websites of two large projects: `http://commonfund.nih.gov/hmp/` and `http://hmpdacc.org/`. There is a lot of data and information here, more than we have time for.

A team of scientists in North Carolina (USA) has begun to "explore the biodiversity in our daily lives": they inventory microbes (and other life) in places like homes, pets, foreheads, and belly buttons. Check out their project websites if you like: `http://www.yourwildlife.org/our-projects/`

Many more microbiome projects have been initiated worldwide, and the links provided here describe just a few of them. If you want to learn more, do a search with relevant keywords on PubMed or Google Scholar, if you like.