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

Objective

You are familiar with hemoglobin, the oxygen-transport protein in the red blood cells of vertebrates, and also with myoglobin, an important oxygen-carrying pigment of muscle tissues. Plants have neither blood cells nor muscle tissue, but do they have globin homologs nevertheless? – You'll find out today! You will compare the sequence of the human beta-globin gene against all protein-coding genes of *Medicago truncatula* (barrel medic), which is a model legume plant with a relatively small genome of approximately 475 Mb and 39,000 protein-coding genes.

The data: query sequences and databases

☞ Make a new directory for today's exercise. You can do that either with the graphical file browser or with the terminal, whichever you prefer.

The query sequences You will have to download the human sequence from the nucleotide and protein databases at NCBI. Just as you did a couple of weeks ago, look up beta globin from human (protein ID: NP_000509, DNA ID: NM_000518). Download the protein and the DNA sequence in **fasta format**, in two **separate** text-only files: one for the protein sequence, and one file for the DNA sequence. Save it in the folder for today's exercise.

The database The *Medicago truncatula* sequences are available in fasta format on the Moodle website. Download the file, move it to a directory for today's exercise, and extract it with a right-click of your mouse. You will then see two files, medicagoDna.fasta and medicagoPep.fasta.

☞ Start the Terminal program and change into the directory (`cd`) for today's computer lab.

Formatting sequences for use as a BLAST database

The program `makeblastdb` can be used to format sequences as a BLAST database: Sequences are indexed and made computer readable, and for this a fasta file is converted into (at least) three files in binary format. Use these two commands to format the databases:

```
makeblastdb -in medicagoDna.fasta -dbtype nucl
```

```
makeblastdb -in medicagoPep.fasta -dbtype prot
```

You can type `ls` to see the newly generated files, but because they are in binary format you won't be able to read their contents.

Using the command-line BLAST

Different BLAST programs are available for different types of searches, but the minimum parameters they all need are the names of three files:
• the name of the database to be searched, given after the parameter option **-db** (e.g., "-db medicagoDNA.fasta")

- the name of the file containing the query sequence, given after the parameter option **-query**
- the name of the output file, given after the parameter option **-out**

You will not specify additional parameters, but if you want to look up which other parameters exist, you can use the command `blastp -help | less`. Remember that you can quit the program `less` by typing "quit".

Searching a protein sequence against a database of protein sequences

☞ Use the protein sequence of the human beta-globin that you just downloaded as a query for a search against the *Medicago* protein sequences, using the blastp program. Use only the three parameters listed above; they can be specified in any order.

Look at the resulting file with the linux tool `less`.

☞ What is the best hit to the human globin sequence?

☞ What are the E-value, raw score, and bit score for the best hit? How many identical and similar amino acids does the sequence with the best hit share?

☞ Can you identify all values that are required to compute the E-value? (m, n, S, S', k, λ)

☞ Do you think that the best hit is a true homolog of the human globin sequence? Why or why not?

Searching a DNA sequence against a database of DNA sequences

Use the DNA sequence of the human beta-globin that you just downloaded as a query for a search against the *Medicago* DNA sequences, using the blastn program.

☞ Execute the BLAST call: the names of the program and the input files will be different, but the options ('-query', '-db', etc) are the same as for the blastp program.

☞ Look at the resulting file with the linux tool `less`.

If there is a best hit: What is the best hit for the human globin sequence? What are the E-value, raw score, and bit score for the best hit? How many identical and similar bases does the sequence with the best hit share? Do you think that the best hit is a true homolog of the human globin sequence? Why or why not?

If no hits are found: How do you explain this result? After all, the same genes are in both files [1]. So why is a gene sequence identified as being similar on the protein level, but not on the DNA level?

Discussion

☞ Initially we asked whether plants, which have neither blood cells nor muscle tissue, do in fact have globin homologs. Based on both BLAST searches, what is your answer to this question?

☞ Which type of search (blastp vs. blastn) is the more appropriate search for the human beta globin sequence against a database of plant sequences? Justify your answer.

[1] To verify that the sequence of the best hit for the blastp search is actually in the file medicagoPep.fasta, you could search for its ID in the terminal: `grep CT025840_43 medicagoDna.fasta`.