Algoritmos Bioinformática /Bioinformática 2023/2024

Group Assignment 2

Phylogenetics of the hemoglobin gene

The hemoglobin protein is involved in oxygen transport from the lung to the various peripheral tissues. Hemoglobins are chemically similar molecules and can be found in many different species. Differences in the sequences of these molecules can provide evidence of evolutionary relationships between species.

In this assignment, you will explore the phylogenetics evolution of hemoglobin across different species, with relation to the hemoglobin gene found in human. You will anlayze the hemoglobin subunit beta (HBB_HUMAN, UniProt P68871).

This is a more exploratory project than the previous ones, so you will have more freedom to explore different solutions and approaches! Remember that more than one approach may be correct, choose the one that makes more sense to you and discuss it.

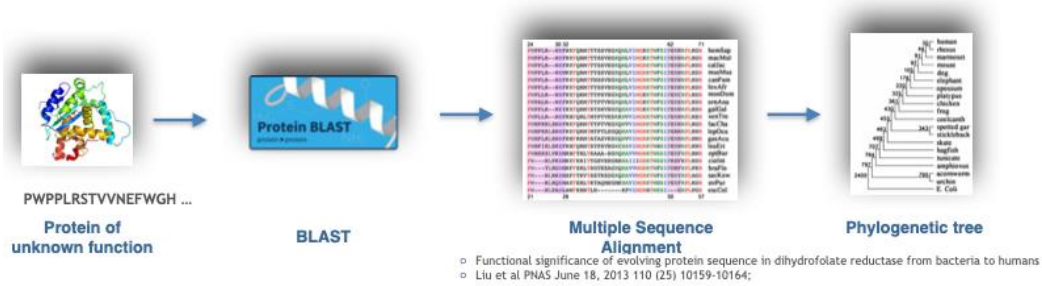

Figure 1. Pipeline to achieve a phylogenetic tree from a query sequence of unknown function or of interest to analyse.

Tools

In this work, you should explore different programming tools and packages available to handle biological sequence analysis. You can use the code provided in the classes, but the idea is to use new tools as well. In particular, you can explore the **BioPython** package and other APIs like the **Uniprot** package 'uniprot'. You can use the functionalities that best suit your goal. Use ChatGPT to obtain suggestions and examples, but beware of hallucinations; always double check!

Phylogenetic Pipeline

The goal will be to implement programmatically the different steps in the pipeline described in Figure 1 and in the classes. If you have difficulties in implementing any of the steps in the

pipeline via coding, perform the analysis manually, *i.e.*, by using the website or the tools and discuss it in the report.

1. Data Collection. Create a small function that given the identifier of the sequence (e.g. UniProt P6887) retrieves the molecular sequence and additional relevant information. You can use the UniProt database to obtain this information. Write the retrieved sequence to a file (sequence.fasta)
   Hint: Explore the UniProt python package.
   Note: the identifier number of the proteins may change. Check the appropriate id.

2. BLAST Analysis. Perform a blast of the sequence against a relevant database of protein sequences (e.g., NR protein database). Parse and filter the results and obtain the sequences **of ten species** (different from human) with the highest matching score. Note that among the high scoring matches you may have multiple sequences from the same species. You will need to retrieve a representative sequence from each species. If you can't find ten different species select the maximum number. For this step, output the 10 sequences into a fasta file (sequences_to_analyse.fasta)
   Hint: explore BioPython functionalities

3. Multiple Sequence Alignment. Using Clustal Omega to perform the MSA on the eleven sequences from step 2. Use the default parameters for protein sequence MSA.
   Hint: explore the BioPython functionalities or the Webservices (script or REST API) provided by EBI for Clustal Omega:
   https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/Clustal+Omega+Help+and+Documentation#ClustalOmegaHelpandDocumentation-WebServices

   The output of the alignment should be written into a file (alignment.txt)

4. Phylogenetic tree. Build and plot the tree. Explore different types of trees and create visualizations for them. Export to different graphical formats (e.g., png or pdf). Discuss the results.
   Hint: explore the BioPython functionalities or newer packages.

Report

For this assignment, you should submit:

- Report – a small report between 1 and 2 pages maximum, describing your approach, the main results, the difficulties, the tools, and packages used. You should provide an example of the obtained phylogenetic tree and finish with a conclusion.
- Code – submit all the developed scripts and codes. You should have a main file called *build_phylotree.py* that given a protein sequence identifier executes the pipeline by calling all the required functions. Note that this can be a unique script and all the results should be generated in the current directory. As an example, the script should be called:

```
Python build_phylotree.py P68871
```

Include a requirement.txt file with all the python modules you used and respective versions (outside the standard python modules). In case your group had issues with any of the steps of the analysis and had to use the online tools, also include the corresponding data in the submission.

Submit a **zip** file with **2 folders**: report and code as described above.

Do not forget to mention in the report all the elements involved in the assignment and any particular note on the contribution of each element to the work.