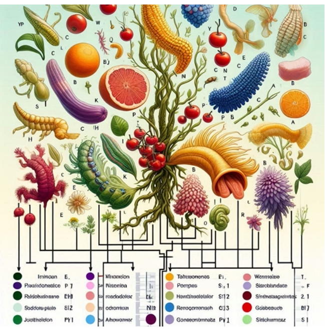

Algoritmos Bioinformática /Bioinformática
Group Assignment 1
2024/205

Phylogenetics of a Taste 2 Receptor Member gene

The **Taste 2 Receptor Member 38 (T2R38)** protein controls the ability to taste glucosinolates, a family of bitter-tasting compounds found in some plant families. These receptors are present across multiple species, with variation of this gene being related to dietary preferences.

In this assignment, you will explore the phylogenetics of T2R38 across different species, with a focus on its relationship to the T2R38 gene in humans, which will be provided as a **nucleotide** sequence.

You will have freedom to explore different solutions. Remember that more than one approach may be correct, choose the one that makes more sense to you and discuss it, justifying its adoption. You will be evaluated on the developed approach, the corresponding choices of algorithms and respective parameters.

Goal of the project

The main goal of this project is to implement a tool to automate **Phylogenetic Analysis,** based on the algorithm discussed in class**. Sequence retrieval** and **analysis algorithms** required to this analysis should be applied programmatically.

This tool should be able to take as input two arguments:

- **Target sequence**: either a DNA or a Protein sequence. However, the phylogenetic analysis will always be performed based on the corresponding **protein sequence.** Note that in case the sequence provided is a DNA sequence, you should begin by identifying the correct reading frame. Hint: In the case of a DNA input sequence, you can show the corresponding protein sequence for the user to confirm its selection.
- The **number of species to include in the analysis** (try n=10 to keep a manageable running time). The species correspond to the different species for which your input sequence has homologs. You should select *n* different in a decreasing order of similarity (from the BLAST analysis).

In this work, you should explore different programming tools and packages available to handle biological sequence analysis. You should primarily use the code provided in the classes, but other tools can be used to complement the analysis. In particular, you can explore the *BioPython* package. If you have difficulties in implementing any of the steps in the pipeline via coding, perform the analysis manually, *i.e.*, by using a website or a tool, then use the corresponding generated file to keep going with the pipeline. Briefly discuss these issues in the report.

The implemented workflow should include at least the following steps:

1. Pre-processing of the input sequence (verification of sequence validity, transcription or translation if necessary, etc …) as the target sequence of the analysis
2. Search for similar sequences in relevant databases (select the appropriate BLAST search tools, adapt the parameters, etc …)
3. Filter the BLAST output to retrieve the relevant sequences.
4. Process the set of selected sequences and perform the MSA.
5. Create a phylogenetic tree from the MSA.
6. Integrate all the information in a document (webpage, Markdown, ….)
7. Other steps considered necessary for the analysis.

Instructions for Submission

For this assignment, you should submit in Moodle within the corresponding timeline:

Code – submit all the developed scripts and codes.

Include a main script named *phylo_analysis.py* that:

1. Takes a sequence file (DNA or protein) and the number of species to analyze as input.

2. Executes the entire pipeline, generating results in the current directory.

3. As an example, an analysis performed on the file *sequence.fasta* with *n=10* species (inclusive of the species from the target sequences)

4. The script should be run as follows:

```
Python phylo_analysis.py sequence.fasta 10
```

5. You are also encouraged to provide any intermediary files as part of their submission.

Report – a small report of **4 pages maximum**, including references, with a motivation, approach (main include a figure with a scheme of the pipeline), the main results, a discussion on the implementation challenges, the tools, and packages used. You should provide the obtained phylogenetic tree and discuss your results based on evidence provided by it. To generate this report with the template for Application Notes from the journal Bioinformatics, that can be downloaded in a word or latex format from here:

https://academic.oup.com/bioinformatics/pages/instructions_for_authors

Word:

https://static.primary.prod.gcms.the-infra.com/static/site/bioinformatics/document/cabios-word-temp.zip?node=9b946b36faff1c196dba&version=541801:1bc4a5f5521ada3b11e7

Latex:

https://www.overleaf.com/latex/templates/oup-general-template/ybpypwncdxyb

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

Report: pdf file with the name of the group

Code: Python files. Do not submit notebooks.

Do not forget to mention in the report all the elements involved in the assignment and any particular note if the contribution of each element to the work significantly deviates from the contribution of the remaining elements. Each group should be composed of three elements.