LGBIO2010 – Bioinformatics

Assignment 2

Hidden Markov Models

1 Protein families

Protein families are groups of evolutionarily-related proteins that present related functions and similarities in sequences or structures. Proteins that share a common ancestor are likely to show significant sequence similarity. Identifying such similarity between a protein sequence and a protein family is, therefore, a common approach to determine if the protein is a family member. However, there are different ways to represent protein families. As a consequence, there are different ways to search for similarities between sequences and protein families.

As a first approach, you will perform pairwise alignments between a given protein sequence and typical members of a known protein family to decide whether the protein belongs to the family or not. Herein the family is represented by its members. As a comparison, you will make the same decision based on the alignment of the same protein sequence to a *profile Hidden Markov Model* (pHMM) representing the family.

Your analysis will focus on the *7-transmembrane G-protein coupled receptors* (abbreviated 7-TM receptors or GPCRs) family and a protein sequence of the narwhal. The 7-TM receptors form a large protein family that includes hormones, neurotransmitters, and light receptors. GPCRs are involved in many diseases and are also the target of approximately 40% of all modern drugs, including two of the top ten global best-selling drugs[1]. While all of these receptors have common patterns in their sequences, there is still a large amount of sequence variation between members.

The narwhal (*Monodon monoceros*) is a toothed whale (*Odontoceti*) whose most prominent characteristic is the possession of a "*tusk*". The *tusk* is a helical canine tooth protruding from the left side of the upper jaw through the lip. While most males have a canine tooth, about 15 percent of females grow it. About one in 500 males has two *tusks*, occurring when the right canine also grows through the lip. There is only one known case of a female growing a second *tusk*.

Discovered canine tooths were believed to be horns of unicorns. Subsequent observations associate *tusks* with weapons and with tools. Modern studies determined that the canine tooth is an innervated sensory organ connecting seawater stimuli of the environment with narwhal's brain.

A thorough study of the *Monodon monoceros*' genome should help understand the evolution of such a tooth and the possible influence of sexual selection on the observed sexual dimorphism. As part of this study, you'll have to determine whether the given narwhal's sequence, AEB0, is a GPCR or not.

[1] Overington J.P., Al-Lazikani B., Hopkins A.L. (Dec 2006). "How many drug targets are there?". *Nature Reviews. Drug Discovery* 5 (12): 993-6.

2 Pairwise alignments

You will first address the problem of assigning a given protein sequence to a known family by mean of pairwise sequence alignments. You will develop a strategy to test if two proteins have significant sequence similarity. You will use this strategy to compare the given protein sequence to various family members.

We want to determine whether the *AEB0* sequence (*Putative* Monodon *GPCR*, stored in *AEB0.fasta*) belongs to the 7-TM receptors family (see characteristics of the family on the *Pfam*[a] database, accession number `PF00001.18`). The family will be represented by the sequence of some of its typical members. Sequences of the members to consider are stored in the *receptors.fasta* file.

2.1. You'll have to align the *AEB0* sequence to sequences of the GPCRs family. Global, semi-global, and local pairwise alignments of sequences have pros and cons. Which alignment should you promote among the three alternatives? Base your answer on the characteristics of the 7-TM receptors family and the expectations of each alignment type. Explain why you choose one type and why you discard the two others.

We propose the PAM300 matrix as substitution matrix, penalties of 11 for gap opening and 1 for gap extension as default parameters. You are free to adapt these parameters. Report the values that you used and motivate your choice. If you decide to modify the parameter's values, explain why the chosen values are more suitable than the default ones.

2.2. Perform a pairwise alignment of the *AEB0* sequence and the *5HT1A_HUMAN* sequence (the first sequence in *receptors.fasta* file). Report your results (at least the alignment score, the alignment size, the percentage of identity, the number of match (or identity), the number of mismatch (or similarity) and the number of gaps).

2.3. Test the statistical significance of the alignment and present your approach. Report your results and conclusions.

2.4. You aligned a single member of the family to the *AEB0* sequence. Perform pairwise alignment of the *Putative* Monodon *GPCR* sequence to each of the 64 members of the family and test the statistical significance of each alignment.

Report the proportion of members presenting sequence similarity with the Putative *Monodon* GPCR sequence.

2.5. Present a reasoned conclusion about the membership of the *AEB0* sequence to the 7-TM receptors family based on experiments of this section.

[a] `http://pfam.xfam.org/`

Hints

- The **`read.fasta()`** and **`getSequence()`** functions from the **`seqinr`** R package respectively read FASTA files and return the corresponding amino acid sequences.
- The **`Biostrings`** R package has a convenient **`pairwiseAlignment()`** function to per-

form sequence alignments. It also provides **`aligned()`** **`score()`**, **`pattern()`**, **`nmatch()`** and other functions to extract information from alignments. Details are provided in the documentation of these functions.

- A substitution matrix can be loaded as data frame with the **`read.table()`** function. The **`as.matrix()`** function will conveniently attempts to turn it (if given as argument) into a matrix. Substitution matrices can be downloaded from: `ftp://ftp.ncbi.nih.gov/blast/matrices/`

3 Blast

We could criticize the representativeness of the 64 sequences used in the previous section. Does the selected subset of sequences represent accurately and broadly the family made of thousands of different proteins? As an example, you could repeat the protocol in previous section on another subset of GPCR sequences and achieve different results from those previously obtained. What if the choice of representative sequences is biased and your conclusions are the result only of this bias?

You'll address this bias through an alternative approach. You'll first identify the closest sequences (in terms of similarity) to AEB0 and determine whether they are considered as G-protein coupled receptors. The process will be repeated on the typical members of the family. You'll then compare the proportion of GPCRs in the closests sequences of the known GPCR and on the AEB0 sequence.

3.1. We propose the following approach:

- Perform a *Blast* on a *query* sequence
- Retrieve the 1,000 best hits
- For each hit:
  - If the *hit definition* contains the `"G-protein coupled receptor"` string, the hit is considered as a GPCR
- Report the proportion of hits considered as GPCR

3.2. Report the results of this approach using the amino acid sequence AEB0 as query. Explain your results and conclude on the membership of the sequence family.

3.3. Report the results of this approach using, as query, each of the ten first sequences from the *receptors.fasta* file. As the later are GPCRs, you would expect most of their hits to be GPCRs too[a]. You would notice, however, that a tiny fraction of hits contains the "*G-protein coupled receptor*" string in their definition. Explore *Blast*'s output and explain the difference between expectation and observation.

3.4. Does question 3.3. change your conclusion about the membership of the AEB0 sequence to the 7-TM receptors family? Motivate your answer.

[a] Remember that *Blast* identifies the sequences with maximal similarity and significant sequence similarity indicates that sequences belong to the same family.

Hints

- The **annotate** R package has convenient **blastSequences()** function to perform blast searches. We suggest the following parameters values: `database="nr"` and `filter="mL"`.
- The **blastSequences()** function returns an XML object. If you want to access to a particular tag, let say `<TAG>value</TAG>`, the following command lists the tag and value of each hit: `x["//TAG"]`, where `x` is the result from **blastSequences()**.
- Function **grepl()** is convenient to test if a particular string is contained in another string.

4 Profile Hidden Markov Models

In this part of the project, you'll align a sequence to a pHMM. The aim is still to determine the membership of the given protein sequence to the known family. The *Pfam*[2] database is a large collection of protein families. This database provides many details on each family including profile HMMs. You will characterize the pHMM of the protein family of interest and align it with the given protein sequence. This pHMM is built with thousands of sequences including all the typical family members used in the first part of the project.

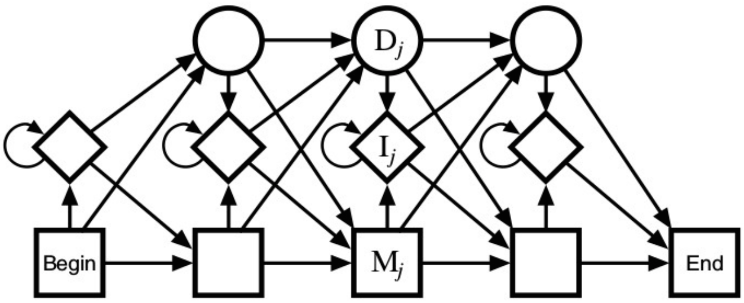

Figure 1: Example of a structure of a profile Hidden Markov Model.

[2] http://pfam.xfam.org/

We want to determine whether the *AEB0* sequence (*Putative* Monodon *GPCR*, stored in *AEB0.fasta*) belongs to the 7-TM receptors family. The protein family is represented by the pHMM (contained in the *7tm_1.hmm* file) obtained from the Pfam (accession number `PF00001.18`) online database. The *pHMM.R* file provides methods to read pHMMs and to align observations to pHMMs. Part of the work has been prepared for you.

4.1. Profile HMM can have various structures: comment on the differences between the pHMM of the GPCRs family and the one presented in figure 1. Present biological motivations to prefer one pHMM over the other.

4.2. The **`viterbi_algorithm()`** function from the **`pHMM.R`** file is an R implementation of the Viterbi algorithm to align a sequence of observations to a pHMM. It identifies the most likely path and computes the associated probability. The implementation does not, however, reports the list of hidden states that the most likely path goes through. Adapt this function such that it also reports the list of hidden states.

4.3. The standard Viterbi algorithm defined in pages 21 and 22 of the slides '*05_HMMs.pdf*' is not designed for pHMM. What are the key modifications of the Viterbi algorithm required for a working implementation handling insert states and delete states? Describe the algorithmic steps needed to deal with a pHMM. You may want to look at the slides '*06_Multiple_Alignments_+_profile_HMMs.pdf*'.

4.4. Align the *AEB0* sequence to the pHMM representing the 7-TM receptors family and report your results. Explain how the score should be interpreted. Can you compare it to the score you obtained with the pairwise alignments?

4.5. Use your adapted **`viterbi_algorithm()`** function to report the alignment of the *AEB0* sequence. The alignment is reported as a string which is the concatenation of:

- `X` if the state is a match, where `X` is the symbol of the emitted amino acid,
- `-` if the state corresponds to a deletion,
- `+X` if the state is an insertion, where `X` is the symbol of the emitted amino acid.

As an example, the alignment of the first 5 amino acids of the sequence, starting with 1 match, 1 insertion, 5 deletion and 3 matches, could be: `"F+L-----CFR"`

4.6. Test the statistical significance of the alignment score and explain your approach. Report your results and conclusions.

Hints

- The **`source()`** function is convenient to load R code from external files, such as *pHMM.R*. Please check also *A very brief introduction to R*, available from Moodle to get more information in this regard.
- The **`viterbi()`** function from the *pHMM.R* file performs a Viterbi alignment. It requires a pHMM and a sequence to align. Note that **`viterbi()`** function performs the same job as the **`viterbi_algorithm()`** except that results are produced faster.
- The sequence must be a vector of characters. A sequence `s` can be transformed into a vector

of characters using the **`strsplit()`** function.

- The pHMM is obtained from the **`readHMMFile()`** function which takes as input a file containing the pHMM in a specific format. To have such a file from an HMMER3 file (such as the *7tm_1.hmm* file), you'll need the **`convertHMMER3()`** function.

5 Summary of your analysis

You addressed the same problem of assigning a given protein to a known family with three alternative approaches. Doing so gives different types of results even if they are based on a similar principle: aligning sequences in a pairwise fashion or aligning a sequence to a representative model of a known family.

Provide a short summary of your results and a short comparison between the pairwise and pHMM alignments you performed. Indicate which approach you would prefer considering their respective strengths and weaknesses.

Results

Upload on Moodle a **zip archive file**, for this assignment, containing

- a PDF report including your answers to all questions in a frame in the present document. Your report should mention any software resource you have been reusing (typically publicly available R packages).
- the R code you have been writing for this assignment.

Submitting the information requested above for grading your work implies that each member of your group is accepting the *anti-plagarism* policy summarized below.

I hereby certify that the results and code that I will submit for this project is coming from my own work and that of my teammate (if any in my group). The submitted works will not be (even partial) copy/paste from information found on the internet or from the work of other groups.

I also certify that I will not distribute any answer or code related to these projects, in person or on any repository (github, bitbucket, Facebook groups, ...) accessible to anybody beyond my teammate, even after the deadlines.

Any violation of the above statements will be considered as cheating and will be reported as such to the President of the Jury.