Hidden Markov Models - Assignment 2

Franco Chiesa Docampo and Luis Felipe Aycardi

Bioinformatics - LGBIO2010

1 Introduction

Protein families are groups of evolutionary-related proteins that present related functions and similarities in sequences or structures. When different proteins share a common ancestor, they are likely to show significant similarity in their sequences. Therefore, identifying this similarity between a protein sequence and a protein family is a used to determine if the protein belongs to the family.

However, there are different ways to represent protein families, and as consequence, there are different ways to search for similarities between sequences and protein families. Some of them are (i) performing pairwise alignments between a given protein sequence and typical members of a known protein family, (ii) a blast algorithm with the most similar sequences and (iii) the alignment of the same protein sequence to a profile Hidden Markov Model (pHMM) representing the family. This algorithms are going to be tested to asses whether the given narwhal's sequence, AEB0, is a 7-transmembrane G-protein coupled receptors (abbreviated 7- TM receptors or GPCRs) or not.

2 Pairwise Alignments

While performing pairwise alignments four key features have to be taken into account: gaps, substitutions, matches and the resulting score of the alignment. The higher the amount of gaps, substitutions and strength of the mismatches between amino-acids, the lower the score will be and therefore the worse the alignment is going to be.
A global alignment (Needleman–Wunsch algorithm) would not be recommendable because it would force too many gaps when aligning the sequences since in this case they both differ in length. Thus the score would be too low and the alignment sub-optimal. This type of algorithm is employed when there is a strong reason to believe that both sequences are related along their entire length and that they are approximately the same size, this is not the case in our study.
Since the aligned sequences differ on length, we expect some gaps at the beginning and/or end of the alignment. A semi-global algorithm does not penalize gaps at the beginning and/or end of such alignment. This technique is useful in cases where we believe that both sequences are related along the entire length of the region where they overlap. Using semi-global technique looks like a good alignment strategy to adopt.
Global and semi-global alignments are used when it is expected that the two sequences are related from end to end. In contrast, local alignment is employed in situations where the two aligned sequences share one or more local regions that are related, but are not related globally from end to end. In other words, the Smith–Waterman algorithm finds the segments between two sequences that have similarities while the Needleman–Wunsch algorithm aligns two complete sequences. This is why We think that local alignment could also be a suitable algorithm for our case study since it is very probable that the AEB0 has suffered several mutations along time (noise through mutations), thus creating substitutions between sequences and cancelling the possibility of a perfect matching alignment.
To put it differently, the Smith-Waterman algorithm could be used in this case to perform the best possible alignments between subsequences of x and y. By

doing this way, we could identify regions of similarity within long sequences that are often widely differ. According to Smith-Waterman criteria, the most biologically significant regions in these protein sequences are the ones that align very well. If we assume this, consequently we will have as well biologically less-significant regions that do not align very well. Another problem with local alignments is that they are more difficult to calculate and require more computing power because of the fact that the algorithm has to identify multiple regions of similarity between sequence.
In order to decide between the Smith-Waterman and the semi-global alignment algorithm we think we have to look deeper into the data and analyze the results we can get from it.
Regarding the PAM300 matrix, we consider it as a good scoring tool for comparing two specific proteins with suspected homology because such matrix indicates the likelihood of an amino-acid being replaced by another one. Such likelihood is represented through a series of 300 point accepted mutations during a specified evolutionary interval, rather than two amino-acids being aligned due to chance. This matrix allows us to study this problem in an evolutionary context, and that is biologically relevant for the alignment problem presented here. That being said, we consider that we could apply a different PAM matrix such as for example a PAM200 (as suggested by Stephen Altschul in 1991). Regarding the gap opening and gap extension penalties, we will leave them as suggested by the authors of this assignment. The reason we do this is because we looked into the NCBI website and followed its recommendations for gap open and gap extension penalty scores.

Next we are going to present the results of a pairwise alignment between the AEB0 sequence and the 5HT1A HUMAN sequence in table 1. We are going to employ semi-global (”local-global” inside the ”pairwiseAlignment” R function) and local alignment techniques.

Scores could suggest there is some sort of similarity between the aligned sequences. After performing this two alignments, question arises about how likely were this particular observations. Maybe the hits that happened in this alignments occurred by chance.

Table 1: Local and semi-global alignments.

Results	Smith-Waterman	Semi-Global
Score	150	127
Size	240	270
P. of Id.	20.83	17.04
Matches	50	46
Mismatches	121	135
Gaps	69	89
Non - Gaps	171	181

This sets the ground for doing a statistical significance test. The statistical significance of the scores are assessed by the P-value. The term ‘p-value’ of an alignment designates the probability of an alignment with this score (or a higher one) occurring by chance alone. We choose as a significant p-value a $p \leq 0.05$ (equal or less than 5 percent of being wrong). We present the p-values of both algorithm approaches in table 2.

Table 2: Local and semi-global p-values.

Results	Smith-Waterman	Semi-Global
p-value	0.292	0.228

Even though both alignments show a good score, their respective p-values demonstrate that these alignments are not statistical significant. Which means that their optimal alignment was due to randomness rather than to biological similarity. Therefore we conclude that the AEB0 protein sequence is not related to the 5HT1A HUMAN sequence. By observing both p-values, we can estimate that the semi-global algorithm presents better results because its p-value is closer to 0.05 than the local one. Thus the alignments produced by the semi-global approach are less random and more significant.

Next we will proceed to 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. Then we will report the proportion of members presenting sequence similarity with the Putative Monodon GPCR sequence. Finally we will present a conclusion about the membership of the AEB0 sequence to the 7-TM receptors family based on experiments of this section. In the code it can be appreciated that we located the "score pwa[[k]]/nchar pwa[[k]]" inside the rows of the "sc pstot". The bigger the percentile of the score, the more far away it is from the center zone of the normal distribution of scores (in the matrix 64x1000). Thus the less random and more specific the score of the alignment is. Resulting in a significant score. Results will be presented in table 3.

Table 3: Percentage of members from 7-TM receptors family presenting sequence similarity with AEB0 sequence.

Results	Smith-Waterman	Semi-Global
Percentage	48%	64%

When utilizing the Smith-Waterman algorithm, 48 percent of family members present sequence similarity to the AEB0 sequence. When using the semi-global alignment, 64 percent of family members present sequence similarity to the AEB0 sequence.

As a concluding remark, we determine that there is no strong evidence supporting the theory that the AEB0 sequence belongs to the 7-TM receptors family because the general similarity to this group (of 64 sequences) is low.

3 Blast

However, the previous approach only considers a set of 64 sequences which may not represent accurately and broadly the family made of thousands of different proteins. Therefore, the analysis and conclusions could be biased.

For that, we identified the closest 1000 sequences (in terms of similarity and from the "non-redundant" database) to AEB0 and determined whether they are considered as G-protein coupled receptors or not by comparing the hit definition. If it contained the "G-protein coupled receptor" string, the hit was considered as a GPCR.

Then, we repeated the procedure with the 10 first members of the family as query, expecting a big degree of similarity. The proportion of GPCRs in the closest sequences (1000) to the AEB0 and the known GPCR are presented in Table 4.

Table 4: Proportion of hits considered as GPCR.

Query	Proportion of hits
AEB0	483/1000
Receptors 1	6/1000
Receptors 2	0/1000
Receptors 3	10/1000
Receptors 4	0/1000
Receptors 5	2/1000
Receptors 6	6/1000
Receptors 7	1/1000
Receptors 8	0/1000
Receptors 9	213/1000
Receptors 10	3/1000

As it can be seen, the proportion of hits that consider the sequence AEB0 as member of the family GPCR is low (less than 50 percent). The label is not shared in the majority of the 1000 most similar sequences. Since it does not even reached half of the targets, is tempting to say that the sequence does not belong to the family of proteins. On the other hand, the proportions found for comparing the sequences of the family with their respective hits are even lower than the previous one. This would imply that the sequences (receptors) are not part of the 7-TM family, which is opposite to the assumption we had since the beginning of the work. Therefore, it is clear that deciding whether a sequence belongs or not to the family can not be only judged by the selected characteristic. There are multiple other parameters we are not taking into account when determining if it belongs, among them:

- Hsp bit-score
- Hsp score
- Hsp evalue
- Statistics db-num
- Statistics db-len
- Statistics hsp-len
- Statistics eff-space
- Statistics kappa
- Statistics lambda
- Statistics entropy

Therefore, we consider insufficient the proposed criteria (of searching for the ”G-protein coupled receptor” string) in order to determine the proportion of hits considered as GPCR. We believe that the mentioned fact stands for all sequences (AEB0 and the 10 first sequences of the 64 references).

4 pHMMs

The last phase includes then, determine whether the AEB0 sequence belongs to the 7-TM receptors family, following the pHMM representation from the Pf online database. As a profile HMM can have various structures, based on the type of alignments and sequences, the structure of the pHMM of the GPCRs family can be analyzed.

After doing a pHMM analysis, it was clear that the transitions between the silent states (I) and the delete states (D) were not part of the structures of the family. Additionally, no transitions were found in the other sense either (from D to I). The result was a structure as the one depicted in Figure 1.

From observing the resulting pHMM in figure 1b), we conclude that this model is not taking into account transitions between insertion and deletion states. We believe that the purpose of this is to avoid excessive insertions and deletions while performing the multiple alignments. Therefore avoiding excessive ”gapping” and protecting the alignment between sequences that may be distant or not. We prefer the model proposed in figure 1b) because of the recently explained reasons.

???????

The next step is to perform a Viterbi algorithm to match the sequence of interest with the model. However, the standard algorithm does not designed for pHMM and some modifications are required to implement insert and delete states. The steps are first do the computation of the Viterbi recurrence with log’s, then including a background model to produce a log-odds score and finally repeat the procedure with the forward recurrence.

With an adapted Viterbi algorithm, we could align the AEB0 sequence to the pHMM representing the 7-TM receptors family. The results are the following:

- Probability: 6.69457893161665e-302

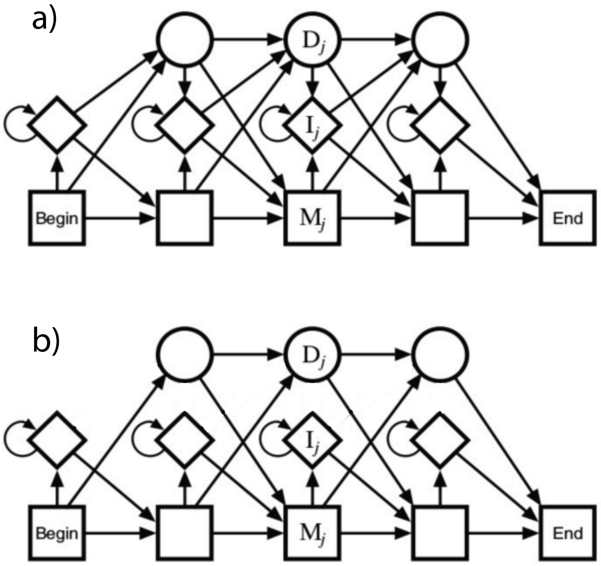

Figure 1: Profile Hidden Markov Model. a) Structure provided as example. b) Structure found for the GPCRs family.

- Score: -693.479399999999
- Alignment: I_268

In this case, the score is the product of the multiple logarithmic summation of value that represent the most likely path. Therefore, it can not be simply compared with the scores previously obtained with the pairwise alignments, its unique for this kind of algorithms.

To asses if it is statistical significant, 10 permutations of the AEB0 sequence was done. Then we computed the percentile of the first Viterbi alignment between the AEB0 and the pHMM of the 7-TM and we obtained the p-value. Our threshold for determining if the alignment was statistical significant was of 0.05. The result is the following: p-value ≈ 0. This means that the alignment result is statistically significant. In other words, this result is not random.

5 Summary

The most important conclusion is that mapping a sequence like one we have to a family of this nature is not straightforward process. According to the Pfam description “G-protein-coupled receptors, GPCRs, constitute a vast protein family that encompasses a wide range of functions. They show considerable diversity at the sequence level, on the basis of which they can be separated into distinct groups. GPCRs are usually described as “superfamily” because they embrace a group of families for which there are indications of evolutionary relationship, but between which there is no statistically significant similarity in sequence”.

This means that the variability among all the sequences in the family is big and the methods tested are too specific in a sense. The more robust test is, however, the implementation of a pHMM to compare the query sequence, as it takes advantage of a general model that represents the whole family. As the task was trying to relate the sequence to a group of such variability, the one that worked with more general parameters was the pHMM. Other techniques could be useful when comparing to smaller groups (as it was shown in the pairwise alignment) or to find sequences with a high degree of similarity (as obtained with blastSequence).

6 References

- http://www.bioinfo.org.cn/lectures/index-35.html
- http://www.cs.cmu.edu/durand/03-711/2015/Lectures/PW_sequence_alignment _2015.pdf
- Pairwise Sequence Alignments - Patrick Aboyoun, Gentleman Lab, Fred Hutchinson Cancer Research Center Seattle, WA. January 3, 2019
- https://www.ncbi.nlm.nih.gov/books/NBK279684/
- https://academic.oup.com/bib/article/7/1/2/262762
- https://www.statsdirect.com/help/basics/p_values.htm
- https://www.ebi.ac.uk/training/online/course/pfam-database-creating-protein-families/what-are-profile-hidden-markov-models-hmms
- http://pfam.xfam.org/family/PF00001.18cite_note-PUB00004961-2