ISSN:0975-9646

Applying Hidden Markov Model to Protein Sequence Alignment

Er. Neeshu Sharma[#1], Er. Dinesh Kumar[*2], Er. Reet Kamal Kaur[#3]

[#]CSE, PTU
[#1] RIMT-MAEC ,[#3] RIMT-MAEC
CSE, PTU
DAVIET, Jallandhar

***Abstract*----Hidden markov models is a statistical tool largely used to study protein alignments and profile analysis of a set of proteins. Finite state machines like HMM move through a series of states and produce output either when the machine has reached a particular state or when it is moving from state to another. It generates a protein sequence by emitting amino acids as it progresses through a series of states. Multiple sequence alignment is a powerful technique that is used by modern bioinformatics systems almost in all their applications. The biomedical methods and algorithms used in MSA have vast importance in solving a series of related biological problems. The well-known and widely used statistical method of characterizing the spectral properties of the residues of a genomic or proteomic pattern is the HMM approach. Profile HMMs have proved to offer a robust solution for MSA.**

***Keywords*--- HiddenMarkovModel HMM, Multiple Sequence Alignment PairwiseSequence Alignment, Alignment,Profile HMM**

1. INTRODUCTION

Sequence alignment is a way of writing one sequence on top of another where the residues in one position are supposed to have a common evolutionary origin. If the same letter occurs in both sequences then this position has been conserved in evolution. If the letters differ it is assumed that the two derive from an ancestral letter. Similar sequences may have different length, which is generally explained through insertions or deletions in sequences. Thus, a letter or a stretch of letters may be paired up with dashes in the other sequence to signify such an insertion or deletion. Since an insertion in one sequence can always be seen as a deletion in the other one frequently uses the term "indel" to represent this. There are two main areas of sequence alignment: pairwise sequence alignment and multiple sequence alignment:

1.1. Pairwise Sequence Alignment

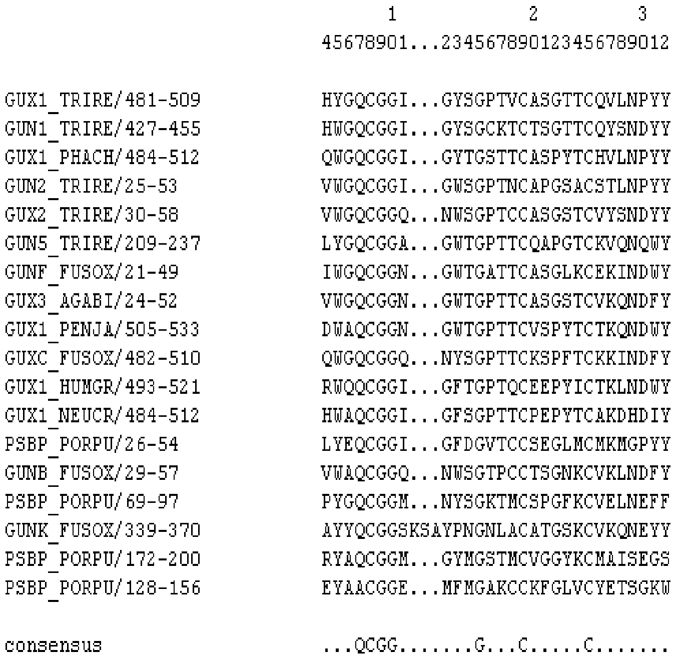
Pairwise Sequence alignment: This alignment is an arrangement of two DNA & amino acid which shows where the two sequences are similar, and where they differ. Broadly, there are three categories of methods for sequence comparison.

- Segment methods compare all windows (overlapping segments of a predetermined length (e.g., 10 amino acids)) from one sequence to all segments from the other. This is the approach used in dot plots. [18]

- Optimal global alignment methods allow the best overall score for the comparison of the two sequences to be obtained, including a consideration of gaps. Global: All positions are aligned
CA--GATTCGAAT!
CGCCGATT---AT!
- Optimal local alignment algorithms seek to identify the best local similarities between two sequences but, unlike segment methods, include explicit consideration of gaps. Local: A (contiguous) subset of positions are aligned ..GATT.....!
....GATT.. ![18]. Based on differences between the two sequences, one can calculate the "cost" of aligning the two sequences by using replacements, deletions and insertions, and assign a similarity score. [18]

1.2 Multiple Sequence Alignment:

Multiple sequence alignment [18] aims to find similarities between many sequences where all similar sequences can be compared in one single figure or table. The basic idea is that the sequences are aligned on top of each other, so that a coordinate system is set up, where each row is the sequence for one protein, and each column is the 'same' position in each sequence. Each column corresponds to a specific residue in the 'prototypical' protein. However both pair wise and multiple sequence alignment algorithms use substitution matrices to score the sequence alignment. But MSA is hard and less tractable than Pair wise Sequence Alignment. MSA where all similar sequences can be compared in one single figure or table. The basic idea is that the sequences are aligned on top of each other, so that a coordinate system is set up, where each row is the sequence for one protein, and each column is the 'same' position in each sequence. Each column corresponds to a specific residue in the 'prototypical' protein.



```
                                          1            2         3
                                    45678901...23456789012345678901 2

GUX1_TRIRE/481-509                  HYGQCGGI...GYSGPTVCASGTTCQVLNPYY
GUN1_TRIRE/427-455                  HWGQCGGI...GYSGCKTCTSGTTCQYSNDYY
GUX1_PHACH/484-512                  QWGQCGGI...GYTGSTTCASPYTCHVLNPYY
GUN2_TRIRE/25-53                    VWGQCGGI...GWSGPTNCAPGSACSTLNPYY
GUX2_TRIRE/30-58                    VWGQCGGQ...NWSGPTCCASGSTCVYSNDYY
GUN5_TRIRE/209-237                  LYGQCGGA...GWTGPTTCQAPGTCKVQNQWY
GUNF_FUSOX/21-49                    IWGQCGGN...GWTGATTCASGLKCEKINDWY
GUX3_AGABI/24-52                    VWGQCGGN...GWTGPTTCASGSTCVKQNDFY
GUX1_PENJA/505-533                  DWAQCGGN...GWTGPTTCVSPYTCTKQNDWY
GUXC_FUSOX/482-510                  QWGQCGGQ...NYSGPTTCKSPFTCKKINDFY
GUX1_HUMGR/493-521                  RWQQCGGI...GFTGPTQCEEPYICTKLNDWY
GUX1_NEUCR/484-512                  HWAQCGGI...GFSGPTTCPEPYTCAKDHDIY
PSBP_PORPU/26-54                    LYEQCGGI...GFDGVTCCSEGLMCMKMGPYY
GUNB_FUSOX/29-57                    VWAQCGGQ...NWSGTPCCTSGNKCVKLNDFY
PSBP_PORPU/69-97                    PYGQCGGM...NYSGKTMCSPGFKCVELNEFF
GUNK_FUSOX/339-370                  AYYQCGGSKSAYPNGNLACATGSKCVKQNEYY
PSBP_PORPU/172-200                  RYAQCGGM...GYMGSTMCVGGYKCMAISEGS
PSBP_PORPU/128-156                  EYAACGGE...MFMGAKCCKFGLVCYETSGKW

consensus                           ...QCGG......G...C.....C.......
```

Fig 1.1: An example of Sequence Alignment

The multiple alignment of these sequences is taken from Pfam, shown below is the so-called seed alignment,, containing the sequences the Pfam curators have used to define the family. This is just a part of the complete alignment file; some comments have been removed. For each sequence, the SWISS-PROT identifier and the position in the parent protein is given on the left. The top line shows the position numbers using the 1CBH 3D structure scheme. The bottom line shows the consensus, which we define here as the same amino-acid residue type in 14 or more sequences (out of 18). Please note that this definition of consensus is just one of many possible [18].

1.3 Major Approaches to MSA

Dynamic Programming: The dynamic-programming approach computes an optimal alignment for a given score function, assuming that the score function is decomposable. Because of its high running time it's not typically used in practice

Progressive alignments: This approach repeatedly aligns two sequences, two alignments, or a sequence with an alignment.

Several heuristics have been proposed to compute the order in which two sequences or aligned.

Profiling: This method computes a profile for a set of sequences, and the profile is then used to align the sequences. Although deterministic profiling methods were proposed earlier, probabilistic methods based on hidden Markov Models are most commonly used aligned.[4]

2. HIDDEN MARKOV MODEL (HMM)

Hidden Markov models are sophisticated and flexible statistical tool for the study of protein models. Using HMMs to analyze proteins is part of a new scientific field called bioinformatics, based on the relationship between computer science, statistics and molecular biology. Hidden Markov models (HMMs) offer a more systematic approach to estimating model parameters. The HMM is a dynamic kind of statistical profile. Like an ordinary profile, it is built by analyzing the distribution of amino acids in a training set of related proteins. However, an HMM has a more complex topology than a profile. It can be visualized as a finite state machine. Finite state machines typically move through a series of states and produce some kind of output either when the machine has reached a particular state or when it is moving from state to state. A markov model is a statistical model that stepwise goes through some kind of change. Markov model is characterized by the property that the change is dependent only on the current state. HMMs are hidden because only the symbols emitted by system are observable, not the underlying walks between states[15]. HMMs are the Legos of computational sequence analysis.A Hidden Markov Model M is defined by

- a set of states **X**
- a set **A** of transition probabilities between the states, an |**X**| x |**X**| matrix. aij ≡ P(Xj | Xi) The probability of going from state i to state j.
- States of **X** are "hidden" states.

- an alphabet **Σ** of symbols emitted in states of **X**, a set of emission probabilities **E,** an **X** x **Σ** matrix
- ei(b) ≡ P(b | Xi). The probability that b is emitted in state i. (Emissions are sometimes called observations.)

It is important to note that in most cases of HMM use in bioinformatics a fictitious inversion occurs between causes and effects when dealing with emissions. For example, one can synthesize a (known) polymer sequence that can have different (unknown) features along the sequence. In an HMM one must choose as emissions the monomers of the sequence, because they are the only known data, and as internal states the features to be estimated. In this way, one hypothesizes that the sequence is the effect and the features are the cause, while obviously the reverse is true. An excellent case is provided by the polypeptides, for which it is just the amino acid sequence that causes the secondary structures, while in an HMM the amino acids are assumed as emissions and the secondary structures are assumed as internal states. States "emit" certain symbols according to these probabilities.

Example of HMM [9].

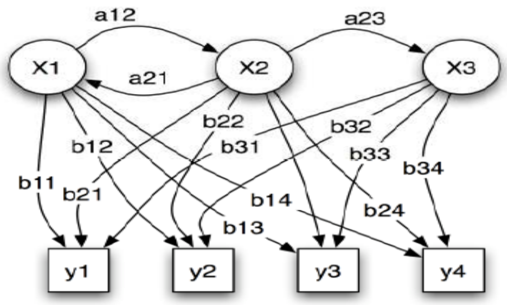


Fig 1.2: Hidden Markov Model

Probabilistic parameters of a hidden Markov model given in the above example.
x — states
y — possible observations
a — state transition probabilities
b — output probabilities

2.1 Major Applications of HMM in Bioinformatics

The HMMs are in general well suited for natural language processing, and have been initially employed in speech-recognition and later in optical character recognition, and melody classification. In bioinformatics, many algorithms based on HMMs have been applied to biological sequence analysis, as gene finding and protein family characterization. A detailed description of all applications would be, in our opinion, outside the scope and the size of a normal survey paper. Nevertheless, in order to give a feeling of how the models described in the first part are implemented in real-life bioinformatics problems, we shall describe in more detail, in what follows, a single application, i.e. the use, for multiple sequence alignment, of the profile HMM, which is a powerful, simple, and very popular algorithm, especially suited to this purpose.[13]

2.2 Profile HMM

Profile HMMs use position specific scoring for the matching & substitution of a residue and for the opening or extension of a gap. Profile hidden Markov models (HMMs) have several advantages over standard profiles. Profile HMMs have a formal probabilistic basis and have a consistent theory behind gap and insertion scores, in contrast to standard profile methods which use heuristic methods. HMMs apply a statistical method to estimate the true frequency of a residue at a given position in the alignment from its observed frequency while standard profiles use the observed frequency itself to assign the score for that residue. This means that a profile HMM derived from only 10 to 20 aligned sequences can be of equivalent quality to a standard profile created from 40 to 50 aligned sequences. [14] In general, producing good profile HMMs requires less skill and manual intervention than producing good standard profiles. A profile HMM has several types of probabilities associated with it. One type is the transition probability -- the probability of transitioning from one state to another. In a simple ungapped model, the probability of a transition from one match state to the next match state is 1.0 and the path through the model is strictly linear, moving from the match state of node n to the match state of node n+1.

There are also emissions probabilities associated with each match state, based on the probability of a given residue existing at that position in the alignment. For example, for a fairly well conserved column in a protein alignment, the emissions probability for the most common amino acid may be 0.81, while for each of the other 19 amino acids it may be 0.01. If you follow a path through the model to generate a sequence consistent with the model, the probability of any sequence that is generated depends on the transition and emissions probabilities at each node. In order to model real sequences, we also need to consider the possibility that gaps might occur when a model is aligned to a sequence. Two types of gaps may arise. The first type occurs when the sequence contains a region that is not present in the model (an insertion in the sequence). The second type occurs when there is a region in the model that is not present in the sequence (a deletion in the sequence). To handle these cases, each node in

the profile HMM must now have three states: the match state, an insert state, and a delete state. The model also needs more types of transition probabilities: match>match, match->insert, match->delete, insert- >match, etc.[15]
Aligning a sequence to a profile HMM is done by a dynamic programming algorithm that finds the most probable path that the sequence may take through the model, using the transition and emissions probabilities to score each possible path.

2.2.1 Purpose of Profile HMM

Profile HMMs are statistical tools that can model the commonalities of the amino acid sequences for a family of proteins. Considered to be more expressive than a standard consensus sequence or a regular expression, profile HMMs allow position dependent insertion and deletion penalties, as well as the option to use a separate distribution for inserted portions of the amino acid sequence. Once a model is trained on a number of amino acid sequences from a given family or group, it is most commonly used for three purposes:

1. By aligning sequences to the model, one can construct multiple alignments.
2. The model itself can offer insight into the characteristics of the family when one examines the structure and probabilities of the trained HMM.
3. The model can be used to score how well a new protein sequence fits the family motif. For example, one could train a model on a number of proteins in a family, and then match sequences in a database to that model in order to try to find other family members. This technique is also used to infer protein structure and function.

3. PRESENT WORK

Profile analysis has long been a useful tool in finding and aligning distantly related sequences and in identifying known sequence domains in new sequences. Basically, a profile is a description of the consensus of a multiple sequence alignment. It uses a position-specific scoring system to capture information about the degree of conservation at various positions in the multiple alignments. This makes it a much more sensitive and specific method for database searching than pairwise methods, such as those used by BLAST or FastA, that use position-independent scoring [14]. The common pairwise comparison methods are usually not sensitive and specific enough for analyzing distantly related sequences. In contrast, Hidden Markov Model (HMM) profiles provide a better alternative to relate a query sequence to a statistical description of a family of sequences. HMM profiles use a position-specific scoring system to capture information about the degree of conservation at various positions in the multiple alignments of these sequences. HMM profile analysis can be used for multiple sequence alignment, for database searching, to analyze sequence composition and pattern segmentation, and to predict protein structure and locate genes by predicting open reading frames. This research work shows how HMM profiles are used to characterize protein families.
The following steps were followed:

1. Accessing Pfam databases
2. Profile HMM Alignment
3. Looking for similarity with sequence comparison
4. Exploring Profile HMM Alignment Options

4. CONCLUSION AND FUTURE WORK

Currently, one very promising approach for protein family related analysis of amino acid sequences is the application of so-called Profile Hidden Markov Models (Profile HMMs) as probabilistic target family models. Given a training set of protein data, discrete HMMs are estimated. These models are then evaluated for unknown query sequences which are aligned to the explicit protein family models. Such explicit target family models are favorable for sequence analysis since family specific data is incorporated into the analysis. One of the main purposes of developing profile HMMs is to use them to detect potential membership in a family. We can use either the Viterbi algorithm to get the most probable alignment or the forward algorithm to calculate the full probability of the sequence summed over all possible paths.
The research can be extended to:

1. Real user interface.
2. Provision to include other sequences (i.e. with different accession numbers and their supported files) automatically.
3. Provision to access the data from a database.
4. Provision for choice of alignment technique
5. Provision to incorporate various input formats

REFERENCES

[1] Can, T., Wang, Y. (2004) *“Automated Protein Classification using Consensus Decision”,* Bioinformatics, vol. 12, pp. 317-327.

[2] Cheng BY, Carbonell JG, Klein-Seetharaman J.(2005)*“Protein classification based on text document classification techniques.”*Journal of Bioinformatics, Volume 212, Pages 67-70.

[3] Devos, D. and Valencia, A. (2000) *“Practical Limits of Function Prediction”*, Protein Design Group, National Centre for Biotechnology, CNB-CSIC Madrid, E-28049, Spain, pp. 134-170.

[4] Erik L. L. Sonnhammer, Sean R. Eddy, Ewan Birney, Alex Bateman and Richard Durbin (1998) *"Pfam: multiple sequence alignments and HMM-profiles of protein domains",* Nucleic Acids Research vol. 26, No.1, pp. 320-322.

[5] Georgina Mirceva1 and Danco Davcev (2009) *"HMM based approach for classifying protein structures"* International Journal of Bio- Science and Bio- Technolog, vol. 1, no.1, pp. 37-46.

[6] N. von Öhsen, I. Sommer, R. Zimmer (2003) *"Profile-Profile Alignment: A Powerful Tool for Protein Structure Prediction"* Pacific Symposium on Biocomputing, Vol 8, pp 252-263.

[7] Park, C.Y., Park, S.H., Kim, D.H., Park, S.H. and Hwang, C.J. (2004) *"A new protein Classification method using dynamic classifier",* Bioinformatics, vol. 9, pp 32-35.

[8] Herbert Popp, Mona Singh and Johnson parker (2002) *"Topics in Computational Molecular Biology"* Lecture notes in bio computing, pp.1-11.

[9] Raninder Kaur, Shavinder Kaur, Reet Kamal Kaur and Amandeep Kaur (2010) *"Characterization of Parathyroid Hormone using HMM Framework"* International Journal of Computer Applications, vol. 1, no. 16, pp. 65-68.

[10] Tamales, J., Ouzounis, C., Casari, G., Sander, C. and Valencia, A. (1998) *"EUCLID: Automatic classification of proteins in functional classes by their database annotations"*, Bioinformatics, pp. 542-543.

[11] T. Plötz, and G.A. Fink, *"Pattern recognition methods for advanced stochastic protein sequence analysis using HMMs",* Pattern Recognition, vol. 39, 2006, pp. 2267-2280.

[12] Thakoor N, Gao J, Jung S.(2007) *"Hidden Markov model-based weighted likelihood discriminant for 2-D shape classification."* Online journal at Springerlink.com

[13] Tolga Can, Orhan C¸ amoglu, Ambuj K. Singh, Yuan-Fang Wang (2004) *"Automated Protein Classification Using Consensus Decision"* Journal of Molecular Biology, Volume 348, Issue 4, Pages 66-68.

[14] Usman Roshan and Dennis R. Livesay (2006) *"Probalign: multiple sequence alignment using partition function posterior probabilities"* Bioinformatics, Vol. 22, No. 22, pp 2715-2721.

[15] Valeria De Fonzo, Filippo Aluffi-Pentini and Valerio Parisi. (2009) *"Hidden Markov Models in Bioinformatics",* Current Bioinformatics, 2007, Vol. 2, No. 1, pp. 49-61.

[16] Wong, L., Chua, H., 17]W.R. Taylor, and C.A. Orengo, *"Protein structure alignment",* J. Mol. Biol., vol. 208, 1989, pp. 1-22.

[17] Li, Z., Liu, G. and Sung, W. (2008) *"Graph – Based Protein Function Prediction",* Genome Informatics, vol. 16(1), pp. 17-23.

[18] http://www.avatar.se/molbioinfo2001/multali.html