ISSN 2250-0987
Neeshu Sharma et al, UNIASCIT, Vol 2 (1), 2012, 141-146

Investigating HMM in Protein Profile Analysis

Er. Neeshu Sharma[1], Er. Dinesh Kumar[2]

[1] *RIMT-Maharaja Agarssen Engg. College, Mandi Gobindgarh*

[2] *DAVIET ,Jallandhar*

***Abstract*----** HMM has found its application in almost every field. Applying HMM to biological sequences has its own advantages. HMM's being more systematic and specific, yield a result better than consensus techniques. In this research work we have applied HMM to profile analysis. With our technique it was found that HMM outperformed Dot plots, local alignment and global alignment techniques considerably.

Keywords:, Profile Analysis, HMM, Profile HMM, Dot plots, Local alignment, Global alignment.

1. INTRODUCTION

Bioinformatics, the application of computational techniques to analyze the information associated with biomolecules on a large-scale, has now firmly established itself as a discipline in molecular biology, and encompasses a wide range of subject areas from structural biology, genomics to gene expression studies Bioinformatics is the application of computer technology to the management of biological information. It is the analysis of biological information using computers and statistical techniques; the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. Bioinformatics is more of a tool than a discipline, the tool for analysis of Biological Data.

The present role of bioinformatics is to aid biologists in gathering and processing genomic data to study protein function.

Proteins are complex organic compounds that consist of amino acids joined by peptide bonds. Proteins are essential to the structure and function of all living cells and viruses. Many proteins function as enzymes or form subunits of enzymes. Some proteins play structural or mechanical roles. Some proteins function in immune response and the storage and transport of various ligands. Proteins serve as nutrients as well; they provide the organism with the amino acids that are not synthesized by that organism. Proteins are amongst the most actively studied molecules in biochemistry and they were

discovered by the Swedish scientist, Jons Jakob Berzelius in 1838.

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

The HMM method has been traditionally used in signal processing, speech recognition, and, more recently, bioinformatics. It may generally be used in pattern recognition problems, anywhere there may be a model producing a sequence of observations. In bioinformatics, it has been used in sequence alignment, in silico gene detection, structure prediction, data-mining literature, and so on. Difficulties with the HMM method include the need for accurate, applicable, and sufficiently sized training sets of data. As for the example of gene detection, in order to accurately predict genes in the human genome, many genes in the genome must be accurately known.

2. HMM FRAMEWORK

This model makes use of alignment as the profiling method and uses conceptual alignment as the type of alignment. Alignment can be considered the most important unsupervised learning problem. The diagram shows the general architecture of an instantiated HMM

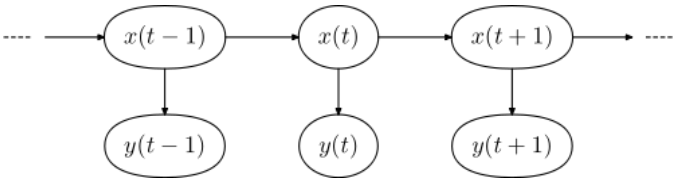

Figure 1.1: Architecture of Hidden Markov Model

Each oval shape represents a random variable that can adopt any of a number of values. The random variable x(t) is the hidden state at time t (with the model from the above diagram, x(t) ε { x1, x2, x3 }). The random variable y(t) is the observation at time t (with y(t) ε { y1, y2, y3, y4 }). The arrows in the diagram (often called a trellis diagram) denote conditional dependencies [30].

From the diagram, it is clear that the conditional probability distribution of the hidden variable x(t) at time t, given the values of the hidden variable x at all times, depends only on the value of the hidden variable x(t − 1): the values at time t − 2 and before have no influence. This is called the Markov property. Similarly, the value of the observed variable y(t) only depends on the value of the hidden variable x(t) (both at time t).

In the standard type of hidden Markov model considered here, the state space of the hidden variables is discrete, while the observations themselves can either be discrete (typically generated from a categorical distribution) or continuous (typically from a Gaussian distribution). The parameters of a hidden Markov model are of two types, transition probabilities and emission probabilities (also known as output probabilities). The transition probabilities control the way the hidden state at time t is chosen given the hidden state at time t – 1.

2.1 Mathematical Perspective

Hidden Markov models are widely used in science, engineering and many other areas (speech recognition, optical character recognition, machine translation, bioinformatics, computer vision, finance and economics, and in social science).

Definition: The Hidden Markov Model (HMM) is a variant of a *finite state machine* having a set of hidden *states*, Q, an output *alphabet* (observations), O, transition probabilities, A, output (emission) probabilities, B, and initial state probabilities, Π. The current state is not observable. Instead, each state produces an output with a certain probability (B). Usually the states, Q, and outputs, O, are understood, so an HMM is said to be a triple, (A, B, Π).

Formal Definition:

Hidden states Q = { qi }, i = 1, . . . , N .

Transition probabilities A = {aij = P(qj at t +1 | qi at t)}, where P(a | b) is the conditional probability of a given b, t = 1, . . . , T is time, and qi in Q. Informally, A is the probability that the next state is qj given that the current state is qi.

Observations (symbols) O = { ok }, k = 1, . . . , M .

Emission probabilities B = { bik = bi(ok) = P(ok | qi) }, where ok in O. Informally, B is the probability that the output is ok given that the current state is qi.

Initial state probabilities Π = {pi = P(qi at t = 1)}.

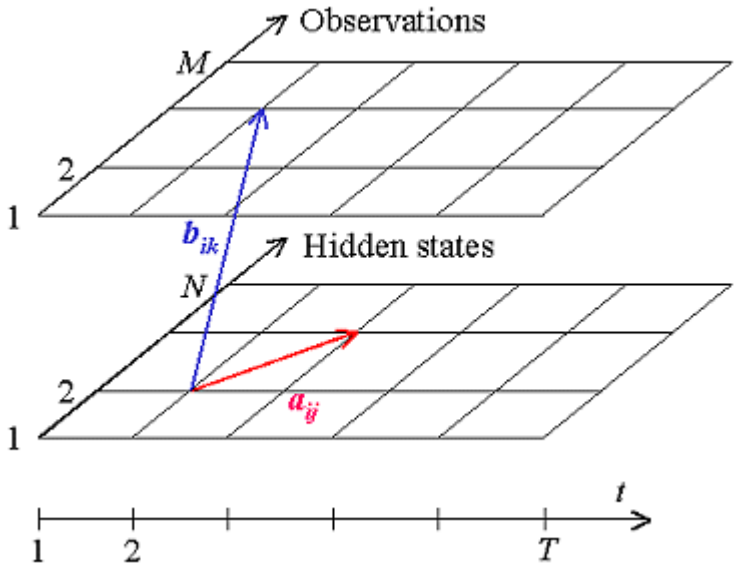

Fig 1.2: Representing HMM States

The model is characterized by the complete set of parameters: Λ = {A, B, Π }.

3. CRITERIA OF PROFILE ANALYSIS

Profile analysis is a key tool in bioinformatics. 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 alignment 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 demonstration shows how HMM profiles are used to characterize protein families [31].
The qualitative and quantitative characterization of protein abundance profiles over a series of time points or a set of environmental conditions is becoming increasingly

important. Using isobaric mass tagging experiments, mass spectrometry-based quantitative proteomics deliver accurate peptide abundance profiles for relative quantitation. Associated data analysis workflows need to provide tailored statistical treatment that (i) takes the correlation structure of the normalized peptide abundance profiles into account and (ii) allows inference of protein-level similarity. We introduce a suitable distance measure for relative abundance profiles, derive a statistical test for equality and propose a protein-level representation of peptide-level measurements. This yields a workflow that delivers a similarity ranking of protein abundance profiles with respect to a defined reference. All procedures have in common that they operate based on the true correlation structure that underlies the measurements. This optimizes power and delivers more intuitive and efficient results than existing methods that do not take these circumstances into account. 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 pair wise method Following are the steps followed in this research work:
Align the sequences in the family: Initially, we will assume that there are no gaps in the alignment. We look at the alignment of N sequences of l positions as follows:

Table 1: Alignment of sequences

Sequence	Position					
	1	2	3	4	…	l
1	a11	a12	a13	…	…	a1l
2	a21	a122	a23	…	…	a2l
3	a31					
-						
-						
N	aN1	aN2	aN3	…	…	aNl

where aij denotes the amino acid from the ith sequence at the jth position.
Use the alignment to create a profile: We build the profile as follows. We compute:
fij = % of column j that is amino acid i
bi = % of background which is amino acid i
The background" can be computed, for example, from a large sequence database, or from a genome, or from some particular protein family.
Now compute the 20 x l array Pij , where
Pij = fij/bi
Intuitively, Pij is the "propensity" for amino acid i in the j position in the alignment.
This gives us the following table:

Table 2: Alignment to compute the Profile

Sequence	Position						
	1	2	3	4	5	…	L
L	PL1	PL2	PL3	…	…		PLl
V	PV1	PV2	PV3	…	…		PVl
F	PF1						
.							
.							
.							

And we use this table to compute:
Scoreij = log(Pij)
Test new sequences against the profile: To use the profile to score a new sequence, we do the following:
Slide a window of width l over the new sequence.
The score of the window equals the sum of the scores of each position in the window.
If the score of the window is higher than the cut off, which is determined empirically, we can conclude that the window is a member of the family. In addition, the higher the score, the more confident the prediction.

4.METHODOLOGY OF WORK

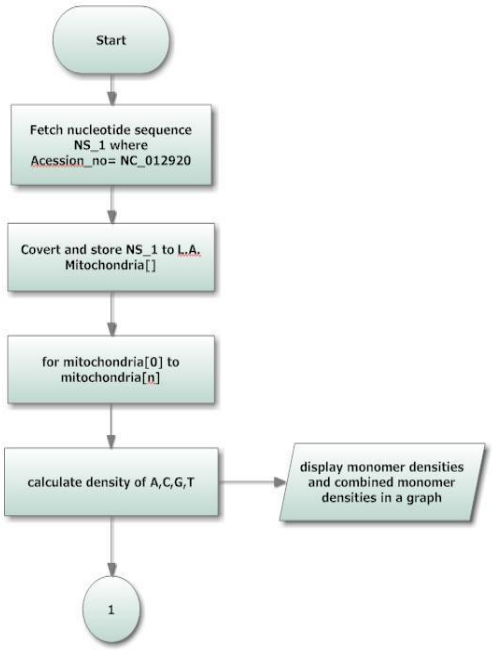

Fig 1.3: Flowchart for accessing a sequence and plotting Monomer and combined Monomer densities

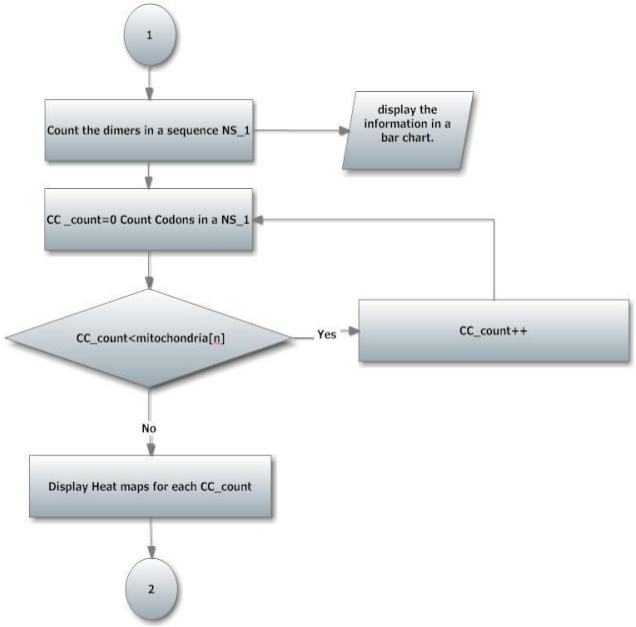

Fig 1.4: Counting Dimmers And plotting Heat maps

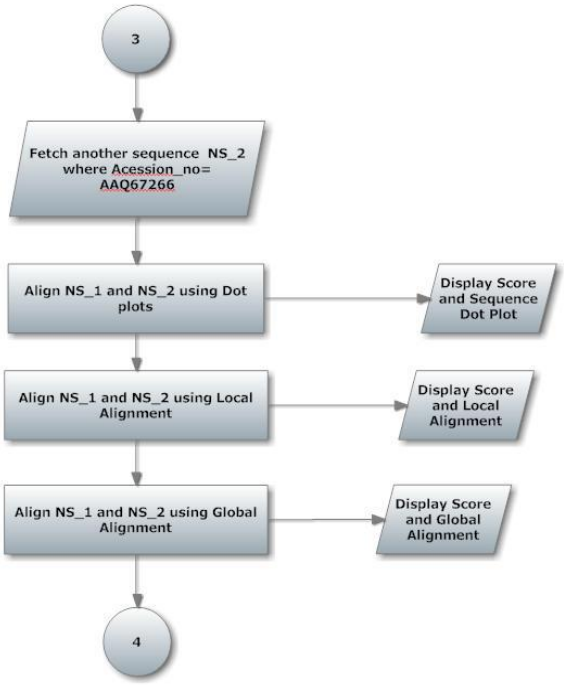

Fig 1.6: Aligning the sequences using Dot plots, Local alignment and Global Alignment

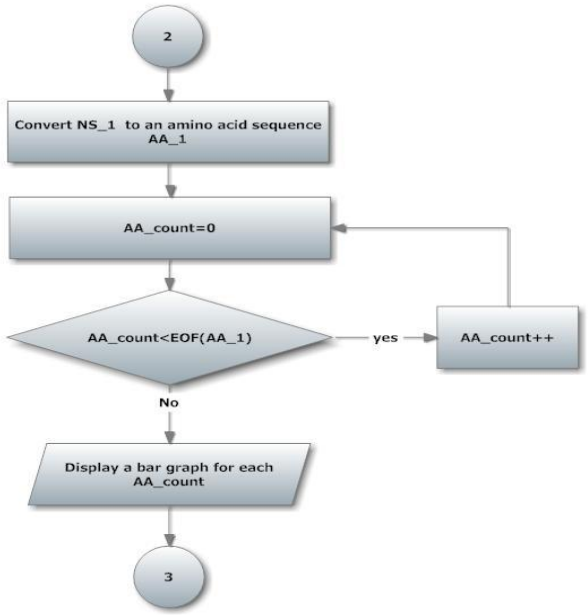

Fig 1.5: Converting to amino acid sequence and representing each amino acid count

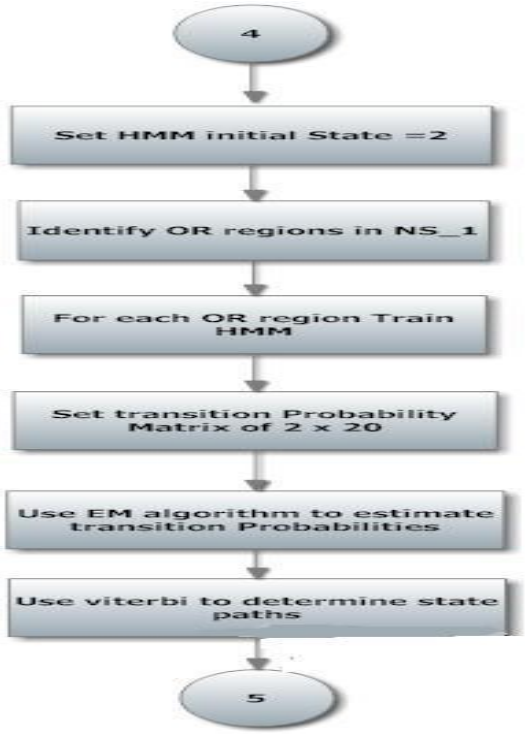

Fig 1.7: Creating and training HMM

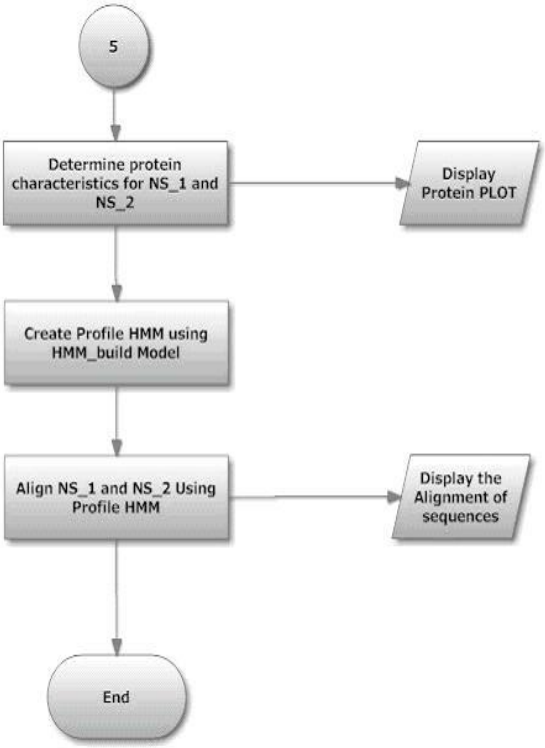

Fig 1.8: Plotting protein Properties, Identifying Profiles and Aligning using HMM

5. CONCLUSION AND FUTURE WORK

Profiles are found by performing the global msa of a group of sequences and then removing the more highly conserved regions in the alignment into a smaller msa. A scoring matrix for the msa, called a profile, is then made. The profile is composed of columns much like a mini-msa and may include matches, mismatches, insertions, and deletions. Once produced, the profile is used to search a target sequence for possible matches to the profile using the scores in the table to evaluate the likelihood at each position. For example, the table value for a profile that is 25 amino acids long will have 25 rows of 20 scores, each score in a row for matching one of the amino acids at the corresponding position in the profile. If a sequence 100 amino acids in length is to be searched, each 25- amino-acid-long stretch of sequence will be examined, 1–25, 2–26, ... 76–100. The first 25-amino-acid-long stretch will be evaluated using the profile scores for the amino acids in that sequence, then the next 25-long stretch, and so on. The highest-scoring sections will be the most similar to the profile.

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] Andrew E. Teschendorff1, Ali Naderi1, Nuno L. Barbosa-Morais, and Carlos Caldas(2006) " *Protein homology detection by HMM–HMM comparison",* Department of Oncology , Vol. 22, pp.2269–2275.

[2] *Anoop Kumar and Lenore Cowen(2010) "Recognition of beta-structural motifs using hidden Markov models trained with simulated evolution", Department of Computer Science,Vol 26,pp.i287-i293.*

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

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

[5] Christopher Tamas and Richard Hughey (1998)"Reduced space hidden Markov model training", Department of computer engineering, Jack Vol. 14 ,pp.401-406.

[6] Dariusz Mrozek†, Bożena Małysiak-Mrozek(2010) "An Improved Method for Protein

[7] Devos, D. and Valencia, A. (2000) "Practical Limits of Function Prediction", Protein Design Group, National Centre for Biotechnology, CNB-CSIC Madrid, EBaskin School engineering, University of California, USA, Spain, pp. 134-170.

[8] 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.

[9] 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.

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

[11] Johanne Söding(2005) *"Protein homology detection by HMM and HMM comparison"* Department of Protein Evolution, Vol. 21 , pp. 951–960

[12] Jia Song, Chunmei Liu, Yinglei Song, Junfeng Qu, and Gurdeep S. Hura (2010*)" Alignment of multiple proteins with an ensemble of Hidden Markov Models"* ,Data mining in Bioinformatics, Vol 4,pp.60-71.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

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

Author Biographies

Er. Neeshu Sharma Neeshu sharma was born on September 17, 1984 at kurukshetra, India. She completed her B.Tech in Computer science from Kurukshetra University in the year 2005 and is pursuing her M.Tech from DAVIET college Jallandhar.

Er. Dinesh Kumar has completed his B.Tech and M.tech in Computer Sciences and is cirrenctly pursuing his P.hD. he had guided 7 M.tech research Thesis and has active research publications in the field of Machine Learning & Natural Language Processing, Computer Networks, Data Structures.