Volume 2, No. 4, July-August 2011
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Protein Profile Analysis: an exploration with HMM

Er. Neeshu Sharma*
CSE,
RIMT MAEC
Mandi Gobindgarh, India
neeshukhn@yahoo.com

Er. Dinesh Kumar
CSE,
DAVIET
Jallandhar, India
Er.dineshk@gmail.com

Er. Reet Kamal Kaur
CSE,
RIMT MAEC
Mandi Gobindgarh, India
reetkamal1901@yahoo.co.in

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. Profile HMMs use position specific scoring for the matching & substitution of a residue and for the opening or extension of a gap. 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.

Keywords: Sequence Alignment, Profile Analysis, Hmm, Profile HMM.

I. INTRODUCTION

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.

An amino acid is any molecule that contains both an amino group and a carboxylic acid group. An amino acid residue is the residuals of an amino acid after it forms a peptide bond and loses a water molecule. Since we are interested in amino acids that form proteins, it is safe to use the terms residue and amino acid interchangeably. There are 20 different amino acids in nature that form proteins.

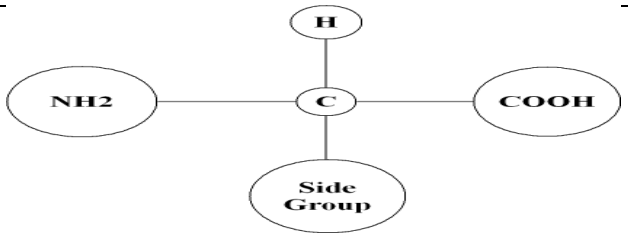

Figure 1: Structure of Amino Acid

II. PROFILE ANALYSIS

Profile Analysis: Profile analysis is a sequence comparison method for finding and aligning distantly related sequences. The comparison allows a new sequence to be aligned optimally to a family of similar sequences. The comparison uses a scoring matrix called a PAM matrix and an existing optimal alignment of two or more similar protein sequences. The group or family similar sequences are first aligned together to create a multiple sequence alignment.[16] The information in the multiple sequence alignment is then represented quantitatively as a table of position-specificsymbol comparison values and gap penalties. This table is called a profile.

The starting point for the creation of a profile is a sequence or group of aligned sequences. This probe is generally a group of functionally related proteins that have been aligned. A profile, however, can be created from a single sequence. The similarity of new sequences to an existing profile can be tested by comparing each new sequence to the profile with the same algorithm used to make optimal

© 2010, IJARCS All Rights Reserved

alignments. To understand how this is done we must first recall what alignment algorithms do. Alignment algorithms find alignments between two sequences that maximize the number of matches and minimize the number of gaps. Gaps are given penalties in the same units as the values in the scoring matrix. The best alignment is then simply defined as the alignment for which the sum of the scoring matrix values minus the gap penalties is maximal. Each row in the profile corresponds to a position in the original multiple sequence alignment. Each possible sequence symbol has a value (a column) in each row of the profile. The comparison of a sequence symbol to any row of the profile defines a specific value or "profile comparison value." The best alignments of a sequence to a profile are found by aligning the symbols of the sequence to the profile in such a way that the sum of the profile comparison values minus the gap penalties is maximal. The profile also contains gap coefficients that are specific for each position so the penalty for inserting a gap in one part of the alignment might be more or less than in another part. The position-specific gap coefficients penalize gaps in conserved regions more heavily than gaps in more variable regions.[16] The profile contains a consensus sequence for the display of alignments of other sequences to the profile. The consensus sequence character corresponds to the highest value in the row. Since the table on which the profile is based is usually the Dayhoff evolutionary distance table, the consensus residue is the residue that has the smallest evolutionary distance from all of the residues in that position of the alignment rather than simply the most frequent residue at that position. In the original approach of Dayhoff the actual estimation is restricted to only very closely related pairs of sequences. However, once a Markov model is fitted to this data, replacement frequencies characteristic for distantly related sequences can be extrapolated from the model.

For example the table value for a profile that is 25 amino acids will have 25 rows of 20 scores, each score in row for matching one of the amino acids in length is to be searched each 25 amino acids 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 section will be the most similar to the profile.

The profile method differs in two major respects from methods of sequence comparison in common use:

a. Any number of known sequences can be used to construct the profile, allowing more information to be used in the testing of the target than is possible with pairwise alignment methods.
b. The profile includes the penalties for insertion or deletion at each position, which allow one to include the probe secondary structure in the testing scheme.

A. Role of Profile Analysis:

Typical scenarios of a profiling approach become relevant, particularly, in the cases of the first two groups, where researchers commonly wish to combine information derived from several sources about a single query or target sequence. For example, users might use the sequence alignment and search tool BLAST to identify homologs of their gene of interest in other species, and then use these results to locate a solved protein structure for one of the homologs. Similarly, they might also want to know the likely secondary structure of the mRNA encoding the gene of interest, or whether a company sells a DNA Construct containing the gene. Sequence profiling tools serve to automate and integrate the process of seeking such disparate information by rendering the process of searching several different external databases transparent to the user.

Advantages of sequence profiling tools include the ability to use multiple of these specialized tools in a single query and present the output with a common interface, the ability to direct the output of one set of tools or database searches into the input of another, and the capacity to disseminate hosting and compilation obligations to a network of research groups and institutions rather than a single centralized repository.

B. Techniques for Profile Analysis:

a. Protein Microarrays:

Protein microarrays consist of antibodies, proteins, protein fragments, peptides or carbohydrate elements that are immobilized in a grid-like pattern on a glass surface. The arrayed molecules are then used to screen and assess protein interaction patterns with samples containing distinct proteins.[17]

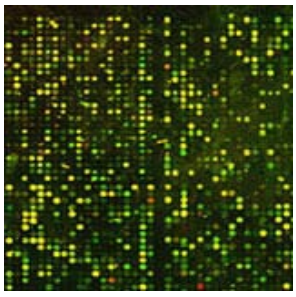

Figure 2: Protein Microarrays

These microarrays are used to identify protein-protein interactions, to identify the substrates of proteins or to identify the targets of biologically active small molecules. And with this growth comes a need for bioinformatics tools to analyze the microarrays.

b. Protein Amino Acid Sequences:

The analysis of amino acid sequences, or primary structure, of proteins provides the foundation for many other types of protein studies. The primary structure ultimately determines how proteins fold into functional 3D structures. Primary structure is used in multiple sequence alignment studies to determine the evolutionary relationships between proteins, and to determine relationships between structure and function in related proteins.

Figure 3: Protein Amino Acid Sequences

© 2010, IJARCS All Rights Reserved

c. Protein-Ligand Docking:

In drug discovery and development, the manner in which small-molecule compounds bind or dock with proteins is of the utmost importance. Proteins are often the main targets for new drugs. And many drug compounds are small molecules that are designed to bind preferentially to specific proteins. Because of this need to design small molecules for protein docking, many bioinformatics tools exist for the analysis of protein-ligand interactions. These tools often fall in the category of computational chemistry. At the atomic scales in which compounds dock with proteins, the interactions are biochemical and biophysical in nature [17]

d. Protein Folds:

Although there is no universal agreement on how to define protein folds, one simple characterization of folds is "an arrangement of secondary structures into a unique tertiary structure." That is, protein amino acid sequences arrange themselves in recognizable, identifiable, 3D structures. Some of these structures are so common in many different proteins that they are given special names, i.e. Rossmann folds, TIM barrels, etc.[17]

III. 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. a set of states **X**
b. 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.
c. States of **X** are "hidden" states.
d. an alphabet **Σ** of symbols emitted in states of **X**, a set of emission probabilities **E,** an **X** x **Σ** matrix
e. ei(b) ≡ P(b | Xi). The probability that b is emitted in state i. (Emissions are sometimes called observations.)[1]

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.

A. Advantages of Hidden Markov Model:

Statistical Grounding

a. Statisticians are comfortable with the theory behind hidden Markov models
b. Freedom to manipulate the training and verification processes
c. Mathematical / theoretical analysis of the results and processes
d. HMMs are still very powerful modeling tools – far more powerful than many statistical methods Modularity
e. HMMs can be combined into larger HMMs

Transparency of the Model

f. Assuming an architecture with a good design
g. People can read the model and make sense of it
h. The model itself can help increase understanding

Incorporation of Prior Knowledge

i. Incorporate prior knowledge into the architecture
j. Initialize the model close to something believed to be correct

Use prior knowledge to constrain training process Example of HMM [1].

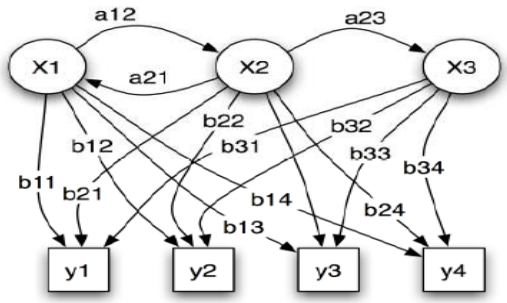

Figure 4: Hidden Markov Model

Probabilistic parameters of a hidden Markov model given in the above example.

x — states
y — possible observations
a — state transition probabilities
b — output probabilities

B. 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

© 2010, IJARCS All Rights Reserved

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]

C. 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.[1].

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.

D. 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:

a. By aligning sequences to the model, one can construct multiple alignments.
b. The model itself can offer insight into the characteristics of the family when one examines the structure and probabilities of the trained HMM.
c. 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.

IV. 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 pair wise methods. Following are the steps followed in this research work:

A. 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	a_{11}	a_{12}	a_{13}	…	…	a_{1l}
2	a_{21}	a_{122}	a_{23}	…	…	a_{2l}
3	a_{31}					
-						
-						
N	a_{N1}	a_{N2}	a_{N3}	…	…	a_{Nl}

where aij denotes the amino acid from the ith sequence at the jth position.

© 2010, IJARCS All Rights Reserved

B. 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

$$P_{ij} = f_{ij}/b_i$$

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	P_{L1}	P_{L2}	P_{L3}	…	…		P_{Ll}
V	P_{V1}	P_{V2}	P_{V3}	…	…		P_{Vl}
F	P_{F1}						
.							
.							
.							

And we use this table to compute:

$$Score_{ij} = \log(P_{ij})$$

C. Test new sequences against the profile:

To use the profile to score a new sequence, we do the following:

a) Slide a window of width l over the new sequence.
b) The score of the window equals the sum of the scores of each position in the window.
c) 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.

V. 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:

A. Real user interface.
B. Provision to include other sequences (i.e. with different accession numbers and their supported files) automatically.
C. Provision to access the data from a database.
D. Provision for choice of alignment technique
E. Provision to incorporate various input formats

VI. REFERENCES

[1] Sharma N.,Kumar D., Kaur Reet. (2011) "Applying Hidden markov model to sequence alignment", Vol 2 (3),pages 1031-1035

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

© 2010, IJARCS All Rights Reserved

[14] Wong, L., Chua, H., 17]W.R. Taylor, and C.A. Orengo, "Protein structure alignment", J. Mol. Biol., vol. 208, 1989, pp. 1-22.
[15] Li, Z., Liu, G. and Sung, W. (2008) "Graph – Based Protein Function Prediction", Genome Informatics, vol. 16(1), pp. 17-23.
[16] http://www.avatar.se/molbioinfo2001/multali.html
[17] http://www.b-eye-network.com/view/1127
[18] http://www.caspur.it/~castri/bioinformatica/gcghelp/profileanalysis.html

© 2010, IJARCS All Rights Reserved