Volume : 2 | Issue : 6 | June 2013 • ISSN No 2277 - 8179

Research Paper

Multiple Sequence Alignment and Profile Analysis of Protein Family Utsing Hidden Markov Model

Engineeringt

KEYWORDS : Multiple Sequence Alignment, Hidden Markov Model

Navjot Kaur	Assistant Professor, at LLRIET, Moga
Rajbir Singh Cheema	Assistant Professor, at LLRIET, Moga
Harmandeep Singh	Assistant Professor, at LLRIET, Moga

ABSTRACT

Bioinformatics is the field of managing, mining and interpreting from biological sequences and structures. It act as an interface between biological and computational sciences. This scientific field deals with the management of all kind of biological information. This information is on proteins and their products, whole organisms. 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. Proteins are high molecular weight compounds composing of alpha-amino acids linked through amide formation between the carboxyl group of one acid and alpha-amino group of the next. Sequence alignment is the way of arranging the primary sequences of Proteins to identify regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences. In protein sequences alignment, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region is among lineages. It also helps to find acceptable substitution for amino acids in a protein signature.

1. INTRODUCTION

Bioinformatics is the science of managing, mining and interpreting information from biological sequences and structures. It act as an interface between biological and computational sciences. This scientific field deals with the management of all kind of biological information. This information is on proteins and their products, whole organisms. Proteins are complex organic compounds that consist of amino acids joined by peptide bonds. In the total of 20 amino acids, 9 are essential amino acids and 11 are non-essential amino acids. Proteins are essential to the structure and function of all living cells and viruses. Proteins are high molecular weight compounds composing of alpha-amino acids linked through amide formation between the carboxyl group of one acid and alpha- amino group of the next. They are components of cell membrane and perform a range of functions as enzymes, antibodies, hormones and transport molecules. An understanding of protein function is facilitated by the study of protein structure. Protein structure are far more complex than simple organic chemicals, because the size of molecules allow for many possible 3-D arrangements. Protein structure is described in terms of a hierarchy.

- Primary structure
- Secondary structure
- Tertiary structure
- Quaternary structure

Sequence alignment is the way of arranging the primary sequences of Proteins to identify regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences. In protein sequences alignment, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region is among lineages. It also helps to find acceptable substitution for amino acids in a protein signature.

Need of multiple sequences alignment, one can reveal whether there is a functional and evolutionary relationship between families of sequences. In addition, multiple sequence alignment is often used to assess sequence conservation of amino acids in protein sequences and nucleotides in DNA sequences.

2. Problem statement

There exists more known protein sequences than protein structures, structure prediction relies heavily on sequence alignment of the protein of unknown structure to other proteins of known structure. The quality of alignment does not correlate well with the level of sequence identity, and alignment algorithms do not perform as well for highly divergent sequences. Therefore, improving sequence alignments ultimately leads to better structure and function prediction methods. The work will be focused on multiple sequence alignment. The optimal alignment between two sequences will be computed. The optimal alignment is the one which gives the maximum alignment score for the input sequences. It is also possible to obtain multiple optimal alignments for the two sequences. The one usually of most interest is to find the hidden states that generated the observed output. In many cases the hidden states of the model represent something of value that is not directly observable. For a particular HMM, the Viterbi algorithm is used to find the most probable sequence of hidden states given a sequence of observed states and then the probability of the sequence given the HMM model is computed by multiplying all probabilities along the path. Forward algorithm is used to sum over all paths inductively and after that decide the most probable path and its raw score is.

3. Solution methodology

3.1 Pfam

HMMs are used extensively both for the construction of Pfam and for detecting matches to Pfam families in database sequences. HMMs are a general probabilistic modeling technique and use HMM in this study to mean a specific form of model that describes the sequence conservation in a family. This type of HMM consists of a linear chain of match, delete, and insert states. The match state contains probabilities for amino acids in a given column, whereas the transition probabilities to and from insert and delete states reflect the propensity to insert a residue or skip one at a given position. The HMM parameters can either be estimated directly from a multiple alignment or iteratively by an expectation-maximization procedure from unaligned sequences. A protein sequence can be aligned to an HMM by using dynamic programming to find its most probable path through the states.

3.2 Seed and full alignment

Construct a seed alignment for each family from a nonredundant representative set of full-length domain sequences. From the seed alignment an HMM is built, which then is used to find new members and to generate the alignment of all detected members. The process of seed alignment and member gathering is iterated as if the initial seed is unsatisfactory. The HMMs are not built from the all-member alignment because this may contain incomplete or incorrect sequences that may affect the HMM adversely. The full alignments are never edited; if they are unacceptable, either the seed alignment is improved or the method to generate the full alignment from the seed is changed.

3.3 Seed alignment construction

The initial members of a seed are collected from one of several sources: Database, NCBI, BLAST results. Families are chosen on

Volume : 2 | Issue : 6 | June 2013 • ISSN No 2277 - 8179

an ad hoc basis, with a bias toward families with many members. If the source provided a complete alignment of the seed members, this is used, but usually an alignment has to be built and compared with known salient features such as active site residues or structurally important residues. Alignment method HMM training is used to produces the best alignment. In a few cases manual editing of the seed alignment is necessary. Any sequence that is suspected to contain an error such as truncation, frame shift, or incorrect splicing is not included in the seed alignment to avoid adding noise to the HMM. This is important because up to 5% of the sequences in Swissprot may contain such errors.

3.4 HMM construction
From each seed alignment an HMM is built by using the Viterbi and Forward algorithms. To avoid over fitting and to make the HMM more general, amino acid frequency priors are normally derived according to an ad hoc pseudocount method using the BLOSUM substitution matrix.

3.5 Full alignment construction
Each HMM thus constructed is then compared with all the loaded sequences. For the families where the initial seed alignment is derived, the new HMM is constructed by using algorithm that constrains the known structural alignment, allowing only the sequences of unknown structure to be realigned. By extracting all matching sequence fragments and aligning them to the HMM, a full alignment is created. The method used for constructing the full alignment and the score cutoffs used are recorded for each family.

3.6 Format
The Pfam format for the alignments is for each sequence segment name/start-end followed by the padded sequence on one line. The name is given from NCBI and the start and end are the coordinates of the first and last residues of the sequence segment, and accession number is added to the end of each sequence line. Each family in Pfam-A has a permanent referenceable accession number (Pfxxxxx), an ID name, and a definition line.

3.7 Incremental updating
Pfam is designed with easy updating in mind. If new sequences are released, they are compared with the existing models and if they score above the cutoff they are automatically added to the full alignment. Normally the seed alignment is not altered, except for the updating of corrected seed sequences. However, if new sequences give rise to problems, the seeds may have to be improved to become more specific for the respective families.

4. Algorithms
4.1 Viterbi
The Viterbi model tries to first find out the most probable path and then the probability of the sequence given the HMM model is computed by multiplying all probabilities along the path. What the algorithm basically does is that at every stage in its journey, it tries to figure out the most probable path leading from a particular amino acid's emission to another's. It selects the most probable path to go from one amino acid emission to another amino acid emission after having compared the probability scores corresponding of each path. The algorithm does the above every time it wants to go from one amino acid to another. The resultant path that stretches along the whole sequence of amino acids is said to be the most probable path. The algorithm employs a matrix. The columns of the matrix are indexed by the states in the model, and the rows are indexed by the sequence.

4.2 Forward
In the forward algorithm, the strategy is to sum over all paths inductively and after that decide which the most probable path is and what its raw score is. Therefore, what this algorithm does is that at every stage it tries to sum over the probabilities of all the paths leading from one amino acid emission to another amino emission instead of taking the max as the Viterbi algorithm did. It does this recursively for each leap from one amino acid emission to another. Therefore, at the end of the day, we end up with the summed over probability of generating the required sequence of amino acids.

4.3 Smith-waterman
The Smith–Waterman algorithm is a well-known algorithm for performing local which is used for determining similar regions between two nucleotide or protein sequences. Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.

5.1 Work flow Diagram

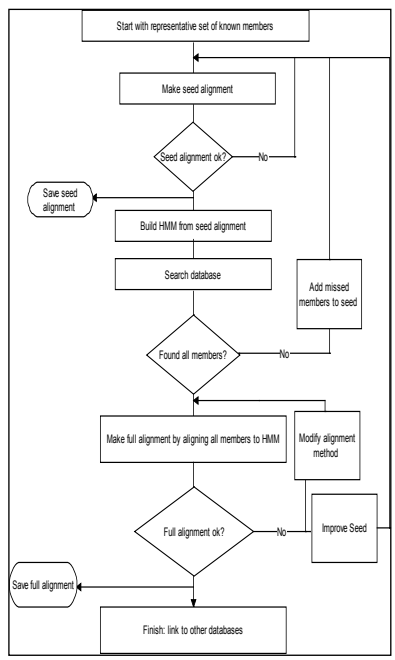


Figure 3.1: The procedure to construct the alignments atnd HMM for a Pfam family. Initial seed alignments are taken from NCBI. 'ok' we mean that known conserved features are correctly aligned.

5.2 Work flow Diagram

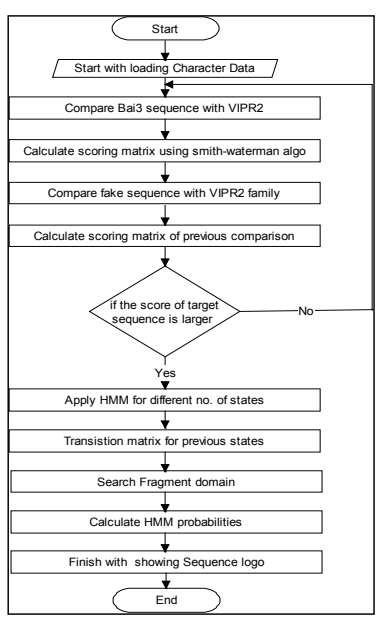
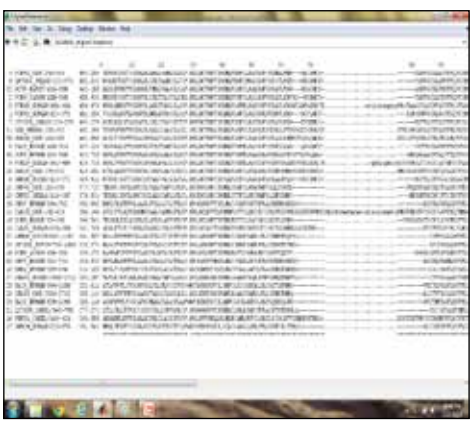


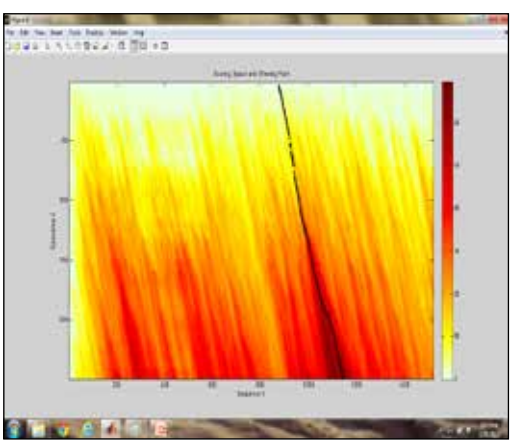
Figure 3.2: Calculate HMM probability and Transition matrix of aligned sequences

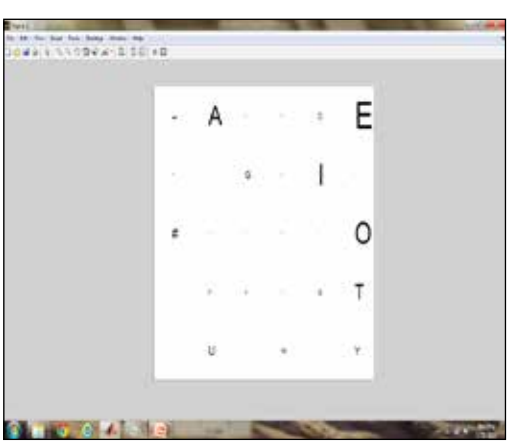
6. Conclusion

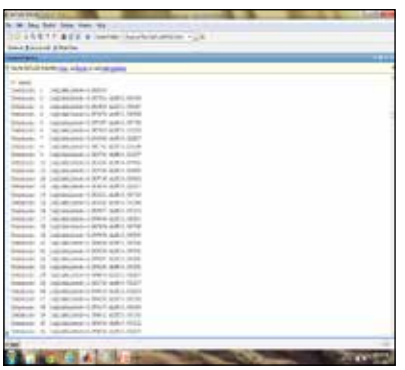
Currently, one very promising approach for protein family related analysis of amino acid sequences is the application of so-called Profile Hidden Markov Models (Profile HMMs) as probabilistic target family models. Firstly, A seed alignment is constructed for each family. From that Seed alignment, an HMM is constructed. One of the main purposes of developing Profile HMM is to find new members and to generate the alignment of all generated members. The process of seed alignment and member gathering is iterated. Each HMM thus constructed is then compared with all sequences and full alignment is created. Normally the seed alignment is not altered, except for the updating of corrected seed sequences. However, if new sequences give rise to problems, such as strong cross-reaction between families, the seed may have to be improved to become more specific for the respective families. Hidden Markov Model profile use a position specific scoring system to capture the related information in the multiple sequence alignment.

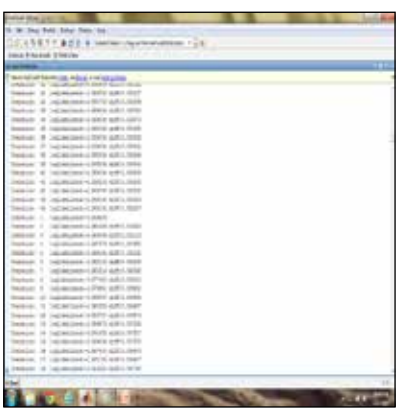
7. Result and Discussion

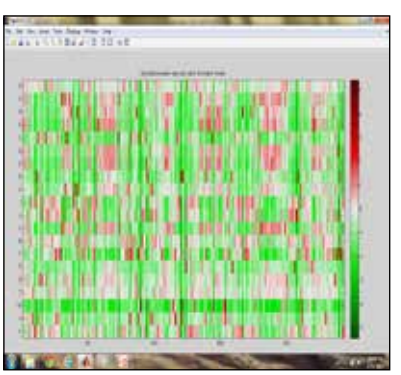


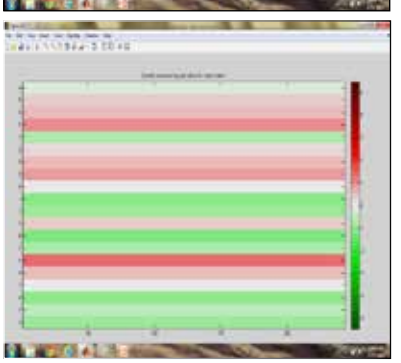




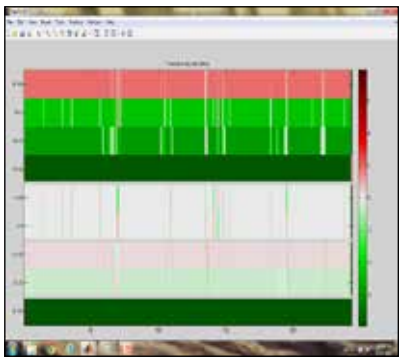


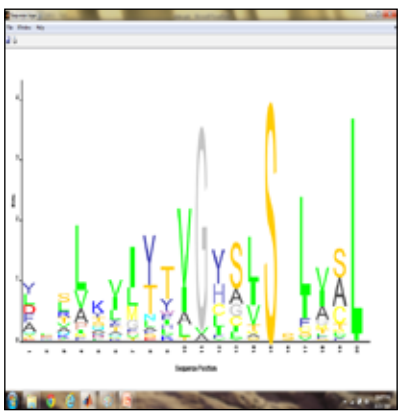






Volume : 2 | Issue : 6 | June 2013 • ISSN No 2277 - 8179





8. Future Scope

The Project can be extended to:

- Remove the drawback that transitions and emissions that don't appear in the training dataset would acquire zero probability (would never be allowed)

Solution: add pseudo counts to the observed frequencies

- Provision to include other sequences (i.e. with different accession numbers and their supported files) automatically.
- Provision to access the data from a database.
- Provision for choice of alignment technique.
- Provision to incorporate various input formats.

REFERENCE

1. Adam M. Szalkowski and Maria Anisimova (2011) "Markov Models of Amino Acid Substitution to Study Proteins with Intrinsically Disordered Regions" vol. 6, Issue 5. | 2. Albayrak, et al. (2010) "Clustering of protein families into functional subtypes using Relative Complexity Measure with reduced amino acid alphabets" Biological Sciences and Bioengineering, Sabanci University, Orhanli, Tuzla, Istanbul, Turkey, pp. 1-10. | 3. Er. Neeshu Sharma et al (2011) "HMM's interpolation of protiens for profile analysis" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), vol. 1.3, pp. 1-10. | 4. Krane, D. and Raymer, M. (2006) "Fundamental Concepts of Bioinformatics", Pearson Education Publishers. | 5. Rastogi, S. C., Mendiratta, N. and Rastogi, P. (2005) "Bioinformatics Methods and Applications", third edition, PHI publication, pp.1-350. | 6. R.C. Edgar, K. Sjölander (2003) "Simultaneous Sequence Alignment and Tree Construction Using Hidden Markov Models" vol. 16, pp. 20-25. | 7. Sharma N.,Kumar D., Kaur Reet. (2011) "Applying Hidden markov model to sequence alignment", vol 2 (3), pp. 1031-1035. | 8. Sierk, et al. (2010) "Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments", Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908 USA, vol. 146, pp. 669-680 |