ResearchArticle **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

ASSESSMENT OF TRANSITION AND EMISSION PROBABILITIES OF HMM FOR PROTEIN SECONDARY STRUCTURE PREDICTION

ANBARASI M*, NEHA VADNERE, TAPASVI SONI, AAKANSHA GUPTA, SALEEM DURAI M. A AND SWARNALATHA P.

School of Computing Science and Engineering, VIT University,Vellore, Tamil Nadu, India.

ABSTRACT

Protein secondary structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry. In genome analysis, secondary structure prediction can be used to predict some aspects of protein functions (classify proteins, identify domains, annotate sequences) and recognize functional motifs. In this paper, we represent protein secondary structure as a mathematical model. To extract and predict the protein secondary structure from the primary structure, we require a set of parameters. These parameters specify any constants appearing in the model and provide a mechanism for efficient and accurate use of data. The Expectation Maximization Algorithm is used to estimate the model parameters with the use of protein datasets like RS126 by using the Bayesian Probabilistic method (data set being categorical). The ultimate objective will be to obtain greater accuracy better than the previously achieved. The future work can be extended by comparing the efficiency of EM algorithm to the other algorithms for estimating the model parameters provides a scope to use these parameters for predicting secondary structure of proteins using machine learning techniques like neural networks and fuzzy logic.

KEYWORDS: Model Parameters, Expectation Maximization Algorithm, Protein Secondary Structure Prediction, Hidden Markov Model.

*Corresponding author

ANBARASI M
School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu, India.

INTRODUCTION

Proteins are the large biological molecules or the macromolecules which consist of one or more long chains of amino acid residues. Proteins are able to perform biological function, primary sequence is folded into one or more specific three-dimensional conformations, determined by a total of non-covalent interactions such as hydrogen bonding, ionic interactions, the Vander Waals forces and the hydrophobic packing. In order to recognize the functions of proteins at a molecular level, it is sometimes necessary to determine the 3D structure. Thus the proteins differ from each other primarily in the sequence of amino acids. This sequence is dictated by nucleotide sequence of their genes, and that results in folding of the protein into a specific 3D structure which determines its activity. The structural features of proteins are divided into various levels. The first level of the protein structure is called the primary structure as it refers just to the sequence of amino acids in the protein. The Polypeptide chains can sometimes fold into regular structures, which mean that the structures are same in shape for different polypeptides. These structures are called secondary structures protein[1]. In Secondary Structures, the hydrogen bonding stabilizes the repeating structures. The examples are alpha helix(H), beta sheets(E) and turns(C). Then the Tertiary Structure which presents the spatial relationship of the secondary structures to one another. The final structure to be mentioned is the Quaternary Structure which is formed by several protein molecules called protein subunits. The structural class has become one of the most important features for characterizing the overall folding type of a protein and thus plays an important role in different aspects of protein research. Knowledge of the structural class has been applied to improve the accuracy of secondary structure prediction[2]. Thus Protein structure prediction is the prediction of the two-dimensional structure of a protein from its amino acid sequence. The primary sequence of the protein consists of the full information to determine the 2D structure. Protein secondary structure prediction is an intermediate step in the prediction of 3D structure of a target protein from its primary structure. Thus the amino acid sequence of a protein can be represented as

$$\{A^{n1}\ R^{n2}\ N^{n3}\ D^{n4}\ C^{n5}\ Q^{n6}\ E^{n7}\ G^{n8}\ H^{n9}\ I^{nl0}\ L^{n11}\ K^{nl2}\ M^{n13}\ F^{nl4}\ P^{nl5}\ S^{nl6}\ T^{n17}\ W^{nl8}\ Y^{nl9}\ V^{n20}\}$$

where the letters are the one letter code of the amino acid residue (total 20 possible amino acids), and n1, n2, n3,... ,n20 represent the number of times the corresponding amino acid code repeats in the protein sequence and n=(nl+n2+ ... +n20) is the length of the protein which has to be predicted. The secondary structure of the primary sequence having length n is $\{C^{ml},H^{m2},E^{m3}\}$, where H, C, E are the different secondary structure classes and ml, m2, m3 represent the number of times the corresponding secondary structural class repeats in the secondary structure of the protein[3]. Analysis of disease likes cancer, diabetes etc required for prediction of protein structures. Hence, the need for prediction of structures of proteins arose in late 1960's and 1970's. Since then, prediction of protein secondary structure is an important research area in bioinformatics. So, a lot of research work has been done to improve accuracy of prediction using computer learning techniques. Different data mining techniques were applied initially for this purpose. In bioinformatics, the local secondary structure of proteins can be predicted if the primary structure is known. Protein secondary structure prediction consists of assigning the regions of amino acid sequences into three classes as alpha helix(H), beta sheets(E) and coil(C). Thus, the goal of secondary structure prediction is to assign an input sequence to any one of the classes. The prediction is successful or not, is determined by comparing it to DSSP algorithm applied to the crystal structure of the protein. Many data mining techniques have been applied to predict secondary structure of proteins. Some of these methods include Chau-Fasman method, k-nearest neighbour method, GOR method, Fuzzy Logic, Machine learning methods like neural networks. The best overall accuracy achieved till now is around 80%. The theoretical accuracy achieved is 90%. In this

paper, we have prepared base which is required for the prediction of protein secondary structure. We have found the transition and emission probabilities of H, E, C using parameter estimation method. These probabilities can be used further in neural network and other algorithms for prediction of protein secondary structure. We have used RS 126 data sets.

METHODOLOGY

(i) Datasets

Today most of the algorithms that are employed in protein secondary structure prediction depend on known three-dimensional structures. These algorithms are then used to impose certain parameters on unknown sequences and such methods rely on the availability of data for their predictions. The earlier algorithms for prediction of secondary structure protein reported high success but their studies were based on small quantities of data, which was mainly derived during the training sessions. For example, Lim[4] claimed 70% Q3 predictive accuracy of 25 proteins; Garnier[5] achieved 63% accuracy for 26 proteins. Using such different protein pool for training and testing of algorithms produces a challenge for objective evaluation of such type of algorithms. In Rost[6] tested a method for prediction of the proteins in which they did not use the same proteins for their algorithm. Their prediction accuracy was greater than 70%. The cross-validation methods of testing, whereby initial proteins were removed from the training pool of proteins, yield a more realistic prediction. To calculate the Transition and Emission matrices in this paper we use the dataset RS126.

(ii) Transition and Emission Probabilities

Mathematical models are required to understand real world problems in a better way. These models are captured by means of distribution model which are extracted or learned directly from data gathered about them. Every distribution model requires estimation of some parameters. These parameters provide an efficient and accurate use of data. There are different methods for estimation of parameters. Here we have used Bayesian approach for parameter estimation. Bayesian approach is helpful with uncertain data. Distribution models which use Bayesian approach are called conditional models.

Bayesian Algorithm

In Bayesian algorithm, define a distribution model, prior probability model and compute the posterior probability distribution. Consider the parameter to be estimated as 'θ' and data unit as'd', by Bayes theorem,

$$P(\theta/d) = \frac{P(\frac{d}{\theta})P(\theta)}{P(d)} \quad \text{And hence,} \quad P(\theta/d) = \frac{P(\frac{d}{\theta})P(\theta)}{\sum_{\theta'}^{e} P(\frac{d}{\theta'})P(\theta')}$$

Thus using Bayesian formula, we can find the probability of each letter in amino acid with respect to three classes H, E and C. The approach further we have used is a Hidden Markov Model.

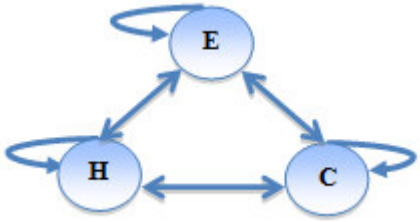
(iii) Hidden Markov Model

The hidden markov model (HMM) assumes a system to be a markov process with hidden states. In this algorithm, the states are not openly visible, so the transition and emission probabilities are considered as parameters. Every state has a probability distribution the sequence of symbols generated by an HMM give information about the sequence of states. Here, hidden refers to the state sequence through which model passes, not to the parameters of the model [7]. Hidden Markov Model has a wide and useful application in biological sequence. Thus here it is used after finding transition and emission probabilities. The HMM can be described as:

- A set of states, Q and A set of transitions, where transition probability $a_{kl} = P(\pi_i=l|\pi_{i-1} = k)$ is the probability transitioning from state k to state l for k,l ϵ Q.

An emission probability, $e_k(b) = P(x_i=b|\pi_i = k)$, for each state, k ,and each symbol, b, where $e_k(b)$ is the probability of seeing symbol b in state k. The sum of all emission probabilities at a given state must equal 1, that is, $\sum_b ek = 1$ for each state, k.

Figure 1
The structure of a Hidden Markov Model



If we know the transition and emission probabilities then only we can use HMM. So we need some algorithm to find out transition and emission probabilities. In this paper, we have used Expectation Maximization (EM) algorithm for estimating transition and emission probability parameters.

(iv) Estimation Maximization Algorithm

One of the popular iteration refinement algorithms for finding the parameter estimates is Estimation Maximization Algorithm. In this paper, it can be used to assign any amino acid letter (one out of 20 amino acids) to one of the classes- H,E or C based on their probabilities. The EM algorithm starts by guessing the initial parameters. This algorithm consists of two steps:

1) Expectation step
2) Maximization step

Expectation step

It assigns each object to one of the classes. Here we have object as 20 amino acids and classes are three classes of secondary structure namely H, E and C.This set of amino acids can be represented as

X = {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,Y,V,W} and classes C={H, E, C}

$$P(x_i = C_k) = \frac{P(Ck)\ P(xi|\ Ck)}{P(xi)}$$

This step calculates the probability of membership of object x_i.

Maximization Step

It uses the probability estimate obtained from the expectation step to refine the model parameters.

Example $$m_k = \frac{1}{n}\sum_{i=1}^{n} \frac{xi\ \ P(xi \epsilon Ck)}{\sum_j P(xi \epsilon Cj)}$$

This maximization step is the maximization of the likelihood of the distributions of data.

IMPLEMENTATION

In this paper, we have used RS126 dataset for defining basic parameters for EM algorithm. RS 126 data set consists of 126 files which contain an observed sequence of amino acids for proteins and their corresponding DSSP and STRIDE sequences to be used in experiments. Direct methods like X-ray analysis would be very difficult when it comes to analysis of large number of amino acid sequences of many proteins. Hence there must be some other method which is less expensive in both time and complexity[8]. Our aim is to find out the secondary structure like alpha helix, beta sheets and coils from given sequence of amino acids in protein. There are some methods for defining secondary structures, they are DSSP, STRIDE, DEFINE etc. We have used STRIDE method for our training data. STRIDE means Structural Identification. As in DSSP, hydrogen bonding is the criteria for defining the structures; in STRIDE dihedral angle is also the criteria. We took amino acid sequences and their corresponding STRIDE sequences for our training data. We have enumerated 20 amino acids in decimal codes as follows:

Table 1
STRIDE Classes and their codes used in our Algorithm

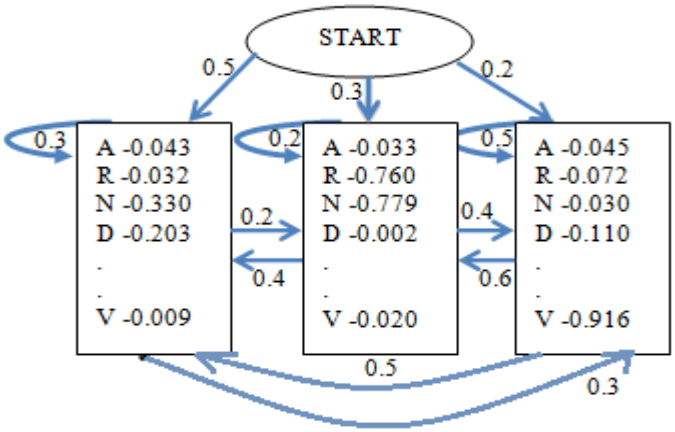
STRIDE Classes	Secondary Structure	Code
H	Alpha Helix	0
E	Beta Sheets	1
C	Coil	2

Table 2
List of amino acids in proteins and the coding's

Amino Acids	1-letter symbol	Code
Alanine	A	0
Arginine	R	1
Asparagine	N	2
Asparagate	D	3
Cysteine	C	4
Glutamate	E	5
Glutamine	Q	6
Glycine	G	7
Histidine	H	8
Isolecine	I	9
Leucine	L	10
Lysine	K	11
Methionine	M	12
Phenylalannine	F	13
Proline	P	14
Serine	S	15
Threonine	T	16
Tryptophan	W	17
Theonine	Y	18
Valine	V	19

The amino acid sequences have been extracted from the RS126 file, enumerated and their values stored in a single file. In the same way, all STRIDE sequences are extracted from all 126 files, enumerated and stored in another file. These two files are useful for getting the initial transition and emission probability values so that training data can be created. All pairs of transitions between H-H, H-E, H-C, E-H, E-E, E-C, C-H, C-E, C-C etc., are counted. In the same way the amino acid sequence pairs are counted. The transition probabilities are obtained as the number of occurring of each pair divided by the total occurrences of all pairs. All amino acid residues in corresponding H, E, and C are counted in the same way to get the emission probability.

Figure 2
Transition and Emission probability



After getting the Transition and Emission probability matrices from the C program, these matrices are used as inputs into the HMM algorithm. This is done to train the data as a confirmation to the accuracy of the Transition and Emission probability matrices created. In order to obtain Estimated Transition and Emission probability matrices, these were given as input to the *hmmgenerate* and *hmmtrain* MATLAB functions.

CONCLUSION

In this paper, we have calculated transition and emission probabilities in terms of parameters for Hidden Markov Model. This has formulated transition and emission probabilities as matrices using RS126 dataset. The algorithm was designed to formulate the matrices rather than being given as ready input. Data sets are tested with the algorithm, Expectation Maximization (EM).The correctness of matrices is further proved by applying HMM to estimate the parameters, and then generate the sequence, thereby checking the accuracy. For a given generated sequence, the Transition and Emission Matrices generated by our C program yielded an accuracy of 79.77%. Further, as a future work, the paper aimed to use these parameters for predicting secondary structure of proteins using machine learning techniques like neural networks and fuzzy logic for greater accuracy better than the previously achieved.

REFERENCES

1. Zikrija Avdagic, Elvir Purisevic, Emir Buza, Zlatan Coralic, Neural NetworkAlgorithm for Prediction of Secondary Protein Structure, Acta Inform Med.17(2): 67-70, (2009).
2. M. Gromiha, S. Selvaraj, Protein secondary structure prediction in differen Structural classes, Protein Engineering, 11(4),249–251, (1998)
3. Lichao Zhang , Xiqiang Zhao, Liang Kong , Shuxia Liu, A novel predictor for protein structural class based on integrated information of the secondary structure sequence , Biochimie, 103:131-6,(2014).
4. Lim IV. Algorithms for prediction of α helices and β structural regions in globular proteins. J Mol Biology, 88, 873-94, (1974).
5. J. Garnier, D. Osguthorpe, and B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, J. Mol. Biol., 120, 97–120, (1978).
6. B. Rost, C. Sander, “Prediction of protein secondary structure at better than 70% accuracy,” *J. Mol. Biol.*, 232, 584–599, (1993).
7. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques,

second edition,Morgan kaufmann publishers, ISBN 1-55860-901-6, (2006).
8. Andey Krishnaji, Allam Appa Rao, An Improved Hybrid Neuro Fuzzy Genetic System (I-HNFGS) for Protein Secondary Structure Prediction from Amino Acid Sequence, International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1218-1223, (2013)
9. Jacek Blazewicz, Peter L. Hammar, And Piotr Lukaslak, Predicting Secondary Structures of Proteins, IEEE Engineering in Medicine and Biology Magazine, 24(3):88-94, (2005).
10. Ashish Ghosh, Bijnan Parai, Protein secondary structure prediction using distance based classifiers, International Journal of Approximate Reasoning, 47(1), 37-44,(2008).
11. Saejoon Kim, Protein β-turn prediction using nearest-neighbor method,Bioinformatics, Oxford University Press, 20(1), 40-44,(2003).
12. Lee, Jennifer L. Leopold1, Cyriac Kandoth1 and Ronald, Protein Secondary Structure Prediction Using RT-RICO: A Rule-Based Approach Leong, The Open Bioinformatics Journal, 4, 17-30, (2010).
13. Julian Lee, Measures for the Assessment of Fuzzy Predictions of Protein Secondary Structure, Proteins: Structure, Function, and Bioinformatics,65(2), 453-462,(2006).
14. P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, Exploiting the past and the future in protein secondary structure prediction, Bioinformatics, 15, 937–946, (1999).
15. R. King and M. Sternberg, Identification and application of the concepts important for accurate and reliable protein secondary structure prediction, Prot. Sci., 5, 2298–2310, (1996).
16. D. Frishman and P. Argos, Seventy-five percent accuracy in protein secondary structure prediction, Proteins, 27, 329–335, (1997).
17. B. Rost, C. Sander, and R. Schneider, Redefining the goals of protein secondary structure prediction, J. Mol. Biol., 235, 13–26, (1994).
18. Swapnil G Sanmukh, Waman N Paunikar, Tarun K Ghosh, Tapan Chakrabarti, Structural & Functional Prediction Of Hypothetical Proteins In Bacteriophages Against Halophilic Bacteria- An In Silico Approach, International Journal of Pharma and Bio Sciences, 2(2), 61-70,(2011).