ISSN:0975-9646

Bayesian Approach to Prediction of Protein Secondary Structure

Asmita A. Yendralwar, Swapnali L.Waghmare , Rajlaxmi M. Biyani, Satish S. Kumbhar

College of Engineering,Pune

***Abstract*-Protein secondary structure prediction is an important problem in bioinformatics and has many applications. In this research we investigate the Bayesian approach for secondary structure prediction of protein. Accuracy improvement in protein secondary structure prediction is focus of our study. We used Three-state-per residue accuracy (Q3) measure for comparative study between Bayesian method and different approaches like Hidden semi Markov Model (HSMM), Dynamic Bayesian Network (DBN), Hybrid model of Support Vector Machine (SVM) and Bayesian Segmentation Model.**

Keywords: Protein secondary structure prediction, Bayesian Method, Q3 measure, HMM, DBN, SVM.

INTRODUCTION

Tertiary structure of a protein can be predicted by some experimental methods like X-ray crystallography, NMR, etc. But as there is a lot of progress in protein engineering and design, these experimental methods are quite slow and expensive. So there is growing interest developing software to do this task. Since direct prediction of tertiary structure of protein is a critical task, prediction of secondary structure of protein is an important intermediate step to predict tertiary structure[8].

How to make an accurate prediction of protein secondary structure is unsolved problem in structural bioinformatics. The accuracy of secondary structure prediction of protein is improving towards 88% which is theoretically estimated limit [5]. One common way to predict secondary structure is to classify them into A three-state classification i.e. helix, sheet and coil, which is often used for secondary structure.

Many methods have been proposed for secondary structure prediction and can be categorized into three groups. The first one contains the computational methods which make use of parameters obtained from analysis of known sequences. Second group includes methods that are based on stereo-chemical criteria. And the third group consists of machine learning algorithm. Accuracy of third one is much better than first two. Over the years, many machine learning methods like Neural Network, SVM have been implemented.

BAYESIAN METHOD

Mathematically, Bayes' theorem gives the relationship between the probabilities of A and B, P(A) and P(B), and the conditional probabilities of A given B and B given A, P(A|B) and P(B|A). In its most common form, it is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The basic idea of the Bayesian method to predict secondary structure of protein is to present relationship of amino acid sequence and the sequence of secondary structure based on prior and likelihood of secondary structural segments.

Consider following notations:

R = $(r_1, r_2, ...,r_n)$to be the sequence of n amino acid residues, where r_i denotes the i^{th} residue.

t = $(t_1, t_2, ...,t_n)$to be the sequence of secondary structural type corresponding to respective residues where $t_i \in$ {H, E, C}.

S = $(s_1, s_2, ...,s_m)$to be the sequence of m positions denoting the end of each individual secondary structural segment

T = $(T_1, T_2, ..., T_m)$ to be the sequence of secondary structural types corresponding to respective segments where $T_i \in \{H,E,C\}$, For all i ∈ 1,2 ,...,m.

CCCCEEEEECCEEEECCCHHHHHHH can be represented as

m = 5

S = (4,9,11,15,18).

T = (C,E,C,E,C,H)

This model specifies the probabilistic dependencies between sequence and structure elements. In this approach, the intra segment independence is not assumed.

Problem of secondary structure prediction is the problem of maximizing the posterior probability of a structure given its primary sequence [5]. Therefore, for given primary sequence, R, the vector (m, S, T) should be found with maximum posterior probability P(m, S, T|R). According to Bayes theorem it can be expressed as:

$$P(R|m,S,T) = \prod_{j=1}^{m} P(R_{[S_{J-1}+1:S_J]}|S,T)$$

Where the *j*th term is the likelihood of amino acids in *j*th segment i.e. subsequence of R starting at position $S_{j-1}+1$ and ending at S_j. To completely specify the joint distribution P(m ,S,T,R) the prior probability distribution is also need to be provided.

$$P(m,S,T) = P(m)\prod_{j=1}^{m} P(T_j|T_{j-1})P(S_j|S_{j-1},T_j)$$

Here from the formula, we can see that each segment type depends on its nearest neighbors only. P(m) is taken to be an improper uniform. The probability of transition of a segment with secondary structure T_{j-1} to a segment having secondary structure T_j . Value of this term are estimated

from a representative set of 2482 unrelated proteins[3]. The last term in the equation gives the length distribution of an uniform secondary structure segment [5][1].

Though we assumed that the amino acid residues at different secondary structure to be independent, those within the same segment are allowed to depend on the neighboring residues[3]. To reflect this dependencies P(R| m,S,T) is modeled as :
For alpha helix:

$$P(R_{[S_{J-1}+1:\,S_J]}|S_{J-1},S_j,H) = \prod_{i=S_{j-1}+1}^{S_{j-1}+l_N^H} P^H_{N_{i-S_{j-1}}}(R_i|R_{[S_{J-1}+1:\,i-1]}) \times \prod_{i=S_{j-1}+l_N^H+1}^{S_j-l_C^H} P^H_I(R_i|R_{[S_{J-1}+1:\,i-1]}) \times \prod_{i=S_j-l_C^H+1}^{S_j} P^H_{C_{S_j-i+1}}(R_i|R_{[S_{J-1}+1:\,i-1]})$$

Where Here l_N^H indicates the length of the helix N-cap model, *Ni* , *Ci* indicate the *i*th position from the N- and C-termini respectively; and *I* indicates an internal (non cap) position[1].
Here first product term models the distribution of amino acids at each of the first l_N^HN-terminal and the last term for the C-terminal positions whereas the
middle term models all internal positions as identically distributed but dependent.

ACCURACY OF BAYESIAN METHOD

Accuracy of Bayesian method is checked by 3 state per residue accuracy percentile :

$Q_3 = (N_c/N) *100$

Here Nc is the total number of correctly predicted residues and N is the total number of residues. The same equation can be used for each of secondary structure type, Q_H, Q_E, Q_C.

$Q_i = (N_{ci}/N_i) * 100$

Accuracy of BSPSS is checked against EVA set of "sequence unique" proteins derived from the PDB database [5].

Table I
Accuracy of BSPSS

Q3	Q_H	Q_E	Q_C
62.207	52.634	24.215	81.903

We can see that accuracy of BSPSS separately is not much good as compare to other methods which are discussed below.

IMPROVEMENT OVER BAYESIAN METHOD

1 ***Semi Markov Model***

Accuracy of IPSSP is checked against EVA set [5] :

Table II
Accuracy of IPSSP

Q3	Q_H	Q_E	Q_C
63.88	55.669	32.754	80.303

Overall accuracy of this method is higher than BSPSS. This improvement is done by introducing three residue dependency models (both probabilistic and heuristic) incorporating the statistically significant amino acid correlation patterns at structural segment borders, allowing dependencies to positions outside the segments to relax the condition of segment independence and introducing an iterative training strategy to refine estimates of model parameters.

2. ***Hybrid SVM and Bayesian approach:***

SVM is helpful for input of Bayesian method. Using Bayesian decision after the SVM can increase the prediction accuracy. SVM can easily handle non-vector inputs, such as variable length sequences or graphs. SVM map an input sample to a high dimensional feature space and try to find an optimal hyper plane that minimizes the recognition error for the training data using the nonlinear transformation function.
Results for this model [4]:

Table III
Accuracy of SVM-Bayesian classifier

E/~E	82.2
H/~H	83.3
C/~C	74.2
H/E	77.5
E/C	78.3
C/H	81.8

3. ***Markov Chain Monte Carlo algorithm:***

In Bayesian approach, there is assumption of intra segment independence which is clearly violated in the case of protein sequences, due to the nonlocal forces involved in protein folding. For example, beta-sheets consist of beta-strands linked by backbone hydrogen bonds beta-sheets are thus a major structural motif which involves interactions of sequentially distant segments to form a stable native fold [1].
By using MCMC model, We can improve accuracy of beta sheet prediction [2].

Table IV
Accuracy of MCMC

Q3	Q_E
65.1	59.6

4. ***Dynamic Bayes network:***

A dynamic Bayesian network approach to protein secondary structure prediction uses a multivariate Gaussian distribution, and simultaneously takes into account the dependency between the profile and secondary structure and the dependency between profiles of neighboring residues. Further improvement is achieved by combining the DBN with an NN (a method called DBNN). It shows better Q3 accuracy than many popular methods and is competitive to the current state-of-the-arts.

Accuracy of DBNN is tested on SD576 [7]:

Table V
Accuracy of DBNN

Q3
78.8

CONCLUSION

In summary, Bayesian segmentation is a probabilistic model of protein sequence/structure relationships in terms of structural segments. It formulates secondary structure prediction as a general Bayesian inference problem. But it suffers from relatively low accuracy because of assumptions like inter segment independence. With methods like HSMM, SVM and Bayesian hybrid model, Dynamic Bayesian Network analysts have been trying to overcome its limitations. And the accuracy of protein secondary structure prediction has been improving steadily towards the 88% estimated theoretical limit [5].

REFERENCES

[1] Bayesian Protein Structure Prediction Scott C. Schmidler, Jun S. Liu, Douglas L. Brutlag

[2] Stochastic Segment Interaction Models for Biological Sequence Analysis Scott C. Schmidler , Jun S. Liu , Douglas L. Brutlag

[3] Protein secondary structure prediction for a single-sequence using hidden semi-Markov models Zafer Aydin, Yucel Altunbasak, and Mark Borodovsky

[4] Protein Secondary Structure Prediction Using SVM with Bayesian Method Wen Yuan Liu, Shui Xing Wang, Bao Wen Wang ,Jia Xin Yu

[5] Protein secondary structure prediction for a single-sequence using hidden semi-Markov models Zafer Aydin1, Yucel Altunbasak1 and Mark Borodovsky

[6] PREDICTION OF PROTEIN SECONDARY STRUCTURE USING BAYESIAN METHOD AND SUPPORT VECTOR MACHINES Minh Ngoc Nguyen, Jagath C. Rajapakse.

[7] A dynamic Bayesian network approach to protein secondary structure prediction Xin-Qiu Yao, Huaiqiu Zhu and Zhen-Su She

[8]Machine Learning Methods for Protein Structure Prediction Jianlin Cheng, Allison N. Tegge, Member, IEEE, and Pierre Baldi, Senior Member, IEEE