Protein backbone dihedral angle prediction based on probabilistic models

Xin Geng, Jihong Guan
Dept. of Computer Science and Technology
Tongji University
Shanghai, China
{gengxin, jhguan}@tongji.edu.cn

Qiwen Dong, Shuigeng Zhou
Shanghai Key Lab of Intelligent Information Processing
School of Computer Science, Fudan University
Shanghai, China
{qwdong, sgzhou}@fudan.edu.cn

***Abstract*—Protein backbone dihedral angles are important descriptors of local conformation for amino acids. Protein backbone dihedral angle prediction lays the foundation for prediction of higher-order protein structure. Existing prediction methods of protein backbone angles mainly exploit traditional machine learning techniques. In this paper, we propose to use two well-known types of probabilistic models --- maximum entropy Markov models (MEMMs) and conditional random fields (CRFs) to predict the backbone dihedral angles of amino acid sequences. Experiments conducted on dataset PDB25 show that these two probabilistic models are effective in dihedral angle prediction, and CRFs outperform MEMMs.**

Keywords-Protein dihedral angle; structure prediction; maximum entropy Markov model; conditional random fields.

I. INTRODUCTION

As a kind of descriptors of the local structural conformation, the dihedral angles pairs ϕ / ψ of amino acid sequences specify the backbone conformation of the corresponding proteins. What exactly information the amino acid sequence can provide for the structure of a protein is still one of fundamentally unsolved problems in computational biology area [1]. To address this problem, an important way is to analyze backbone dihedral angles of amino acid sequences.

The dihedral angles pairs ϕ / ψ of a protein are highly correlated to its secondary structures [2]. The distribution of ϕ / ψ of α-helix and β-sheet are relatively centralized, while the distribution of the random coil structure is decentralized with some regions overlapping with that of α-helix and β-sheet. In order to define the higher-order structures of a protein, it is especially necessary to determine the dihedral angles ϕ and ψ.

By definition, the dihedral angle prediction problem is: given the Ramachandran plot and the amino acid sequence, we want to predict in what region of the Ramachandran plot the dihedral angles pair ϕ / ψ of each residue lies [2]. Ramachandran plot is a useful tool to visualize the allowed values of the dihedral angles and their population. It can be used to map the string that represents the primary sequence into a string from an alphabet of n letters, with each letter representing the region where the dihedral angle pair of each residue lies in the plot . In this work, the letters are the labels that we try to predict by using probabilistic models.

Compared with secondary structure prediction, there has been relatively little work on the prediction of the backbone dihedral angles [2]. The first method for protein backbone dihedral angle prediction is HMMSTR [3]. Based on hidden markov models (HMMs), [3] uses amino acid profile to predict the state sequences of dihedral angles. HMMSTR discretizes the backbone dihedral angles into 11 conformation states. However, it does not give the reason why the multinomial distribution is used to calculate the probability of observing a given profile at each position in a sequence.

Given the conformational regions defined in Ramachandran plot, prediction of backbone dihedral angles can be modeled as a classification problem in the space of induced dihedral angle states. R. Kuang et al. [2] devised two novel automated methods on a local structural-based sequence profile database. They applied two methods for data classification, including support vector machine (SVM) and feed-forward back-propagation artificial neutral network. O. Zimmermann and U. H. E. Hansmann [4] proposed a multi-step support vector machine procedure to divide the dihedral angles into three categories. As we can see, these methods mainly exploit traditional machine learning techniques, and the division of the Ramachandran plot into three or four conformational regions is usually related to the distribution of dihedral angles of three-state secondary structure. Although they achieved better average prediction accuracy than previous methods, their performance for the residues with random coil structure is not as good as that with α-helix and β-sheet structures.

In this study, we treat the prediction of backbone dihedral angles as the problem of labeling amino acid sequences. We use two types of probabilistic models, maximum entropy Markov models (MEMMs) [5] and conditional random fields (CRFs) [6] to predict backbone angle sequences of amino acid sequences. In comparison with HMMs, these two models can achieve better performance on many text-related tasks, including part-of-speech tagging, information extraction, etc. Here, we simply choose the sequence profiles of residues as the features used by these two models. Furthermore, different definition modes for labels of dihedral angles are taken into consideration based on different divisions of the Ramachandran plot. Our focus is on how to use the two probabilistic models to mark the amino acid sequences with dihedral angle labels.

978-1-4244-4713-8/10/$25.00 ©2010 IEEE

II. MATERIALS AND METHODS

A. Datase

All the protein sequences we use for experiments are non-redundant protein structures extracted from the PDB25 dataset, which is a subset of the PDB database [7]. There is less than 25% sequence identity between any two proteins and any protein has a resolution better than 2.5 Å. The structures with missing atoms and chain breaks are also excluded. The backbone dihedral angles ϕ and ψ of each protein sequence, ranging from -180° to 180°, are calculated by using DSSP program [8]. After processing by DSSP, every amino acid is marked by the label in terms of the conformation state of its dihedral angles pair ϕ / ψ in the Ramachandran plot.

B. Labels of protein backbone dihedral angles

We define the labels of protein backbone dihedral angles according to the major conformational states on the Ramachandran plot. Here, we use the partition of the Ramachandran plot presented by Shortle [9]. Fifteen labels ranging from 1 to 15 are derived from 15 regions of the Ramachandran plot. The subdivisions used in this paper are described in Fig. 1.

We also design other 4 label definition modes based on the subdivisions used by propensities defined by Shortle [9], and the coarse-grained classes are defined with combination of some regions of 15 discrete representations. Other regions used in our paper and the labels with its corresponding regions are shown in Fig. 2 and Table. I, respectively. Omitting the N- and C-termini, the backbone dihedral angles of each amino acid are mapped to one of the regions in the plot, so in every category of labels it is sure that there is only one label for an amino acid.

C. Predicting protein backbone dihedral angles based on MEMMs

As a discriminative model, MEMM has more advantages for labeling sequential data than traditional generative models (e.g. HMMs). MEMM shows excellent performance when it is used to solve conditional problems in which the observations are given in terms of many overlapping features or the set of all possible observations is not reasonable enumerable. In the following, we will present how to utilize MEMMs to predict protein backbone dihedral angles.

For each amino acid in a protein sequence, we employ its PSI-BLAST log-odds score profile [10] as the representation of observations. From the datasets used for training and testing, the amino acid at each position of the sequences is converted to a string, which comprises the log-odds score profile of every amino acid within the window centered by this position. This string is marked by the label corresponding to the amino acid at the central position. Not all the strings have the same length and the string will be shorter if blank positions in window segments are neglected. An example can be seen in Fig. 3 with a window size of five. As MEMMs calculate state transition probabilities based on maximum entropy, the position of window center is an important factor for parameter estimation. Every protein sequence should be neither cut apart nor integrated with other sequences.

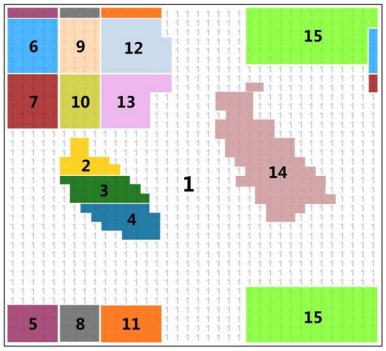

Figure 1. The Ramachandran plot is partitioned into 10 degree by 10 degree grids and 15 regions are used in this paper

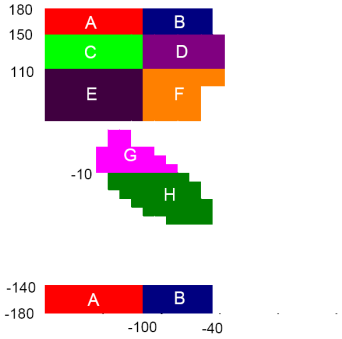

Figure 2. Some regions used for 4 coarse-grained classes

TABLE I. THE DEFINITION OF FOUR TYPES OF LABELS

I(4)	II(6)	III(9)	IV(12)
a=2+3+4	a=5+6+7+8+9+10	a=5+6+8+9	a=A
b=5+6+7+8+9+10+11+12+13	b=11+12+13	b=7+10	b=C
c=14	c=2+3+4	c=12	c=E
d=1+15	d=14	d=13	d=B
	e=15	e=G	e=D
	f=1	f=H	f=F
		g=14	g=2
		h=15	h=3
		i=1	i=4
			j=14
			k=15
			l=1

```
amino acid sequence:     A G  E D  V G  A  P  P
label sequence(Type Ⅱ):   13 8 15 2 14 11 11 11
log-odds score profile:  A C D E F G H I K L M N P Q R S T V W Y
                       A 5-1-2-1-3 0-2-2-1-2-1-2-1-1-2 1 0 0 -3-2
                       G -1-3-2-3 4 5-2-2-2-2-2-2-3-3-3-1-2-3-1 0
                       E -2-4 5 4-4-2-1-4 0-4-3 0-2 1-1-1-2-3-4-3
                       D 2-3 3-1-4-2-2-3-1-3-3-3-1 6 -1-2-1-1-2-4-3
                       V -1-2-3-2-1-4-3 1-2 3 1-3-3 2-2-2-1 3-3-2
                       G -1-3-2-2-4 5-2-3 -1-3-3-1-3-1 2-1 2-3-3-3
                       A 2-2 2-1-3-2-3-1-2-2-2-2 5-2-3-1-1 1-4-3
                       P -2-3-3-2-2-3 5-2-2 2-1-26-2-2-2-2-2-4-1
                       P -2-4-2-2-5-3-3-4-2-4-3-3 8-2-3-2-2-3-5-4
string representation(5):
                     #0A-1 #0C-3 ⋯ #1A-2#1C-4 ⋯ #2W-4#2Y-3    13
                      #-1A-1 #-1C-3 ⋯ #0A-2#0C-4⋯ #2W-4#2Y-3    8
                     #-2A-1#-2C-3 ⋯ #-1A-2#-2C-4 ⋯ #2W-3#2Y-2 15
                     #-2A-2#-2C-4 ⋯ #-1A-1#-2C-3 ⋯ #2W-4#2Y-3   2
                     #-2A2#-2C-3 ⋯ #-1A-1#-2C-2 ⋯ #2W-3#2Y-3   14
                     #-2A-1#-2C-2 ⋯ #-1A-1#-2C-3 ⋯ #2W-5#2Y-4 11
```

Figure 3. An example of the feature representation

D. *Predicting protein backbone dihedral angles sequences based on CRFs*

CRFs, also discriminative models, can be understood as the sequence version of maximum entropy models. They offer several advantages over MEMMs with a drawback of label bias problem caused by the per-state normalization of transition scores [6]. Since comparison between these two probabilistic models will be conducted in our experiments, we use similar features to represent amino acids so that the comparison is unbiased.

The same parameters for training used by these two models are listed in Table. II. For model learning, we calculate the state transition probability based on first-order Morkov models. The features appearing less than five times and the context predicates appearing less than once are not considered and the iterative scaling algorithms for the models' parameters estimation take 50 iteration. We use the publicly available FlexCRFs-0.3 package [11] in our experiments and the results are shown in next section.

III. RESULTS AND DISCUSSION

Two hundred protein sequences are randomly selected from the dataset PDB25 for experiments. Structures of all the proteins in the dataset are known. Because of the lack of backbone atoms, the N- and C-terminal residues are removed from our training and testing datasets. There are totally 45246 residues. Five-fold cross validation experiments are conducted with the MEMMs and CRFs models.

A. *Performance Comparison between MEMMs and CRFs*

We first compare the best prediction accuracies of MEMM and CRFs, and the results of prediction accuracies for the five categories of labels definition are shown in Table. III.

Obviously, in most cases CRFs outperforms MEMM by about 4~7%, while MEMM gets a little better prediction by less than 1% when dihedral angles are divided into 4 classes. These results are consistent with the fact that for most tasks of labeling sequence data, CRFs is a powerful tool that has been widely used in natural language processing area previously. Numbers in the brackets of column 1 and column 2 in Table. III are the size of the window we set when the best accuracy is achieved. Some details of the performance of 4 different window lengths will be demonstrated and discussed in the next section.

From the results, we can roughly infer that the coarser the class granularity is, the more accurate the prediction is. As Table. III indicates, when the backbone dihedral angles are grouped into 4 classes and 12 classes, the optimal prediction accuracies are greater than 67% and less than 39% for the two models, respectively. However, there is an exception that the accuracy of Type V with 15 classes is higher than Type IV with 12 classes. Hence, we may argue that the division of Ramachandran plot into 15 classes is better than into 12 classes.

B. *Prediction accuracies for different window sizes*

Betancourt and Skolnick [12] found that the backbone dihedral angle distribution is closely related to the identity and conformation of adjacent residues. Keskin et al. [13] obtained the results from the PDB database that the native-state dihedral angles are strongly determined by the specific amino acid sequence of the protein. So we use the amino acids' profiles as the features, which contain evolutionary information of amino acids, within the window of a certain length. The experimental results for window sizes of 5, 7, 9 and 11 are shown in Fig. 4.

TABLE II. THE PARAMETERS OF THE TWO MODELS

Parameter	Value
rare threshold for features	5
rare threshold for context predicates	1
number of training iterations	50
initial values for weights	0
σ^2	100
ε	0.0001

TABLE III. THE BEST PERFORMANCE OF THE TWO MODELS

Type	MEMM	CRFs
I	68.57(7)	67.95(7)
II	53.61(7)	59.37(9)
III	47.97(9)	52.81(5)
IV	35.42(7)	39.33(9)
V	36.35(7)	39.95(7)

For both the two models and all the five types of label definition, there is an accuracy peak around the window size of 7 or 9. This is in accordance with the results of Ohkubo and Brooks III [14] that non-nearest-neighbor interaction exists within five residues of polypeptides (>10 residues). We can also draw another conclusion from Fig. 4 that for a fixed window length and the same label definition type, the average prediction accuracy of CRFs is more or less better than that of MEMM, and sometimes it is even better than that of MEMM with coarser-grained labels.

C. *Prediction accuracies for different classes*

Table. IV compares the prediction accuracy of each class for different label definition types. Although the overall performance of CRFs is better than that of MEMM, the prediction accuracy achieved by MEMM is impressive sometimes, concretely, classes of b and c of type II and class c of type III. The reason why both the two models show bad performance for some classes, for example, class a of type II and III, class k of type IV, is probably that the number of residues belonging to these classes is relatively small. There

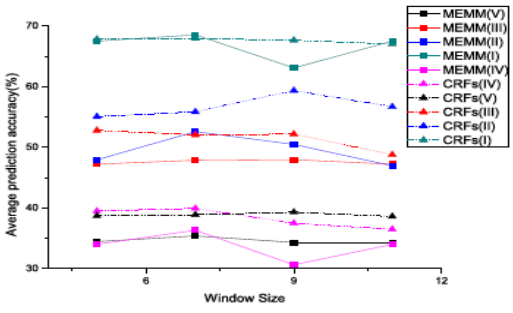

Figure 4. The average prediction accuracies for 4 different window sizes. Dot lines represent CRFs results and solid lines are MEMM results. The results of label definition Type I to V are shown in dark cyan, blue, red, megenta and black, respectively

are fewer than 300 residues for class a of type II for instance, and it appears that MEMM is more susceptible to the change of the number of residues. Furthermore, the secondary structures of the residues in the classes with low prediction accuracy in each label type are mostly random coil, which may be caused by the simple features used in the two methods.

D. *Performance Comparison with other methods*

R. Kuang et al. [2] grouped dihedral angles into four states, which is a little different from our label definition Type I. The first method they proposed, combining artificial neutral network (ANN) with a local structure-based sequence profile database, uses nine consecutive residues to predict the backbone conformational state of central residue. The second is the SVM method, in which the feature representation of nine-residue amino acid segments is the profile with the secondary structure predicted by PSI-PRED. The overall results obtained by their two approaches are better than existing methods then and the SVM method is better than the ANN method. One interesting negative result is that their prediction accuracy on the smallest class is quite low, only 0.3%. Our effort here, is not to improve the prediction accuracy, however the MEMM and CRFs methods still outperform other methods [2] for some classes and 22.28% residues labeled by d can be correctly predicted in our work. Compared with SVM, MEMMs and CRFs maybe more suitable for the prediction of dihedral angles of residues whose regions on Ramachandran plot are relatively variable.

Actually, we also conducted experiments by using second-order models, in which the state transition probability are calculated based on second-order Morkov models and a slight improvement of about 1% can be obtained in each case.

IV. CONCLUSION

This paper introduces two probabilistic models, MEMM and CRFs to predict protein backbone dihedral angles. Experimental results show that the performance of the two models with simple feature representation is acceptable, and the optimal window size is seven or nine, which is consistent with previous work on non-local interactions of residues. The two methods obtain better prediction accuracy for the residues with non-regular secondary structure than existing methods. As for the future work, we will try to employ more structure features of residues and investigate more effective feature representations to improve the prediction accuracy.

TABLE IV. AVERAGE PREDICTION ACCURACY OF EVERY CLASS

I	II	III	IV	V
a70.33/71.44	a .28/53.58	a 0/0	a 0/0	1 14.92/6.68
b67.15/67.52	b52.58/57.44	b56.41/48.68	b21.48/26.71	2 22.06/28.00
c 22.76/4.71	c43.84/51.09	c50.27/56.36	c 6.48/36.42	3 47.57/39.78
d 22.28/6.63	d 26.92/5.37	d 16.88/3.57	d 5.51/16.12	4 12.50/0.32
	e 2.44/1.76	e42.98/45.19	e 8.36/42.91	5 23.29/22.65
	f 15.27/6.44	f 8.59/41.06	f 13.39/3.09	6 18.75/0.59
		g 29.44/5.93	g 6.23/11.27	7 8.70/1.03
		h 3.77/1.74	h 6.35/30.15	8 38.80/42.16
		i18.88/4.17	i33.46/38.78	9 5.04/1.19
			j29.80/5.04	10 33.33/0
			k1.47/0.88	1133.90/34.29
			l5.12/6.46	12 0/0.79
				1336.43/28.97
				14 28.33/4.23
				15 2.27/2.75

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC) (No. 60873040 and No. 60873070), National Basic Research Program of China (No. 2010CB126604), Shanghai Leading Academic Discipline Project (No.B114) and China Postdoctoral Science Foundation (No. 20080440572 and and No. 200902201). Jihong Guan was also supported by the Program for New Century Excellent Talents in University of China (NCET-06-0376).

REFERENCES

[1] M. Socolich, S. W. Lockless1, W. P. Russ, H. Lee, K. H. Gardner and R. Ranganathan, "Evolutionary information for specifying a protein fold," Nature, Vol. 437, 2005, pp.512-518

[2] R. Kuang, C. S. Leslie, and A. Yang, "Protein backbone angle prediction with machine learning approaches," Bioinformatics, Vol. 20(10), 2004, pp. 1612-1621.

[3] C. Bystroff, V. Thorsson and D. Baker, "HMMSTR: a Hidden Markov Model for Local Sequence-Structure Correlations in Proteins," J. Mol. Biol. Vol. 301, 2000, pp. 173-190.

[4] O. Zimmermann and U. H. E. Hansmann, "Support vector machines for prediction of dihedral angle regions," Bioinformatics, Vol. 22, 2006, pp. 3009-3015.

[5] A. McCallum, D. Freitag, F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," Proc.ICML, 2000, pp. 591-598.

[6] J. Lafferty, A. McCallum, F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc.ICML, 2001, pp. 282-289.

[7] U. HOBOHM, M. SCHARF, R. SCHNEIDER and C. SANDER, "Selection of representative protein data sets," Protein Sci, Vol. 1, 1992, pp. 409-417.

[8] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," Biopolymers,Vol. 22, 1983, pp. 2577-2637..

[9] D. Shortle, "Composites of local structure propensities: Evidence for local encoding of long-range structure," Protein Sci, Vol. 11, 2002, pp. 18-26,.

[10] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Research, Vol. 25(17) , 1997, pp. 3389-3402.

[11] http://www.jaist.ac.jp/~hieuxuan/flexcrfs/flexcrfs.html

[12] M. R. Betancourt and J. Skolnick, "Local Propensities and Statistical Potentials of Backbone Dihedral Angles in Proteins", J. Mol. Biol. Vol. 342, 2004, pp. 635-649

[13] O. Keskin, D. Yuret, A. Gursoy, M. Turkay, and B. Erman, "Relationships Between Amino Acid Sequence and Backbone Torsion Angle Preferences," Proteins, Vol. 55, 2004, pp.992-998.

[14] Y. Z. Ohkubo and C. L. Brooks III, "Exploring Flory's isolated-pair hypothesis: Statistical mechanics of helix-coil transitions in polyalanine and the C-peptide from RNase A", PNAS, Vol. 100(24), 2003, pp. 13916-13921.