PROTEIN SECONDARY STRUCTURE PREDICTION BASED ON THE AMINO ACIDS CONFORMATIONAL CLASSIFICATION AND NEURAL NETWORK TECHNIQUE

Guang-Zheng Zhang [1,2], *De-Shuang Huang*[2], *Hong-Qiang Wang* [1,2]

1. HeFei Institute of Intelligent Machines, Chinese Academy of Sciences
2. Department of Automation, University of Science and Technology of China
PO Box 1130, Hefei, Anhui 230031,China
gzzhang@iim.ac.cn, dshuang@iim.ac.cn

ABSTRACT

In this paper, based on the 340 protein sequences and their corresponding secondary structures got from the Protein Data Bank (*PDB*), we group the 20 different amino acids into f (Former), b (Breaker) and n (Neutral) according to their occurring frequencies in the three-state secondary structures (α-helix, β-sheets and Coil), which reflect the intrinsic preference of that amino acid for a given type of secondary structure. Then we use this information to improve the protein Secondary Structure Prediction (*SSP*) accuracy and get a better performance than the previous methods.

1. INTRODUCTION

The conformational preference of amino acids is one of the most important factors in the protein secondary structure prediction. The early studies [1] showed that some amino acids did have preference for a given type of secondary structures. But the small number of protein sequences and structures available at that time limited the further study on the problem. With more sequences and corresponding structures are determined and available on the internet, we can get a further understanding on the amino acids conformational preference and its application in protein secondary structure prediction.

In the present study, based on the 340 protein sequences (96,921 amino acids altogether) got from the Protein Data Bank (PDB), we get the conformational preference of the 20 different amino acids and use it to improve the secondary structure prediction accuracy. Section 2 introduces the statistical algorithm and the amino acids conformational classifications. In Section 3, we improve the protein secondary structure prediction accuracy by using the classification information, and get a better experimental results compared with the previous approaches. Section 4 concludes the whole paper.

The work is supported by the NSF of China.

2. EXPERIMENTAL PROCEDURES

2.1. Data Preparation

The 340 protein sequences used here are obtained from the PDB [2] and have some special features: (1) All the secondary structures are determined by X-ray diffraction technique; (2) All the protein sequences are deposited in the database before 28 May, 2003, and sorted by deposition date in descending order; (3) The counting and statistical results are based on the MATLAB version 6.5.

The occurrence number of a particular amino acid in a particular type of secondary structure reflects the intrinsic preference of that amino acid for the type of secondary structure.

Table 1. The numbers[♭] of amino acids in 3- and 8-state secondary structures in the 340 protein sequences.

3 state	H			E		C		
8 state	G	H	I	B	E	$*$	S	T
Ala	267	3428	0	63	1209	1524	541	842
Cys	35	300	0	54	37	320	100	121
Asp	242	1382	2	66	546	1517	748	979
Glu	288	2831	0	56	851	1184	572	816
Phe	134	1147	2	90	1129	700	273	319
Gly	179	980	0	85	960	1809	1315	2053
His	81	607	1	36	420	756	246	320
Ile	98	1796	0	87	1916	927	325	281
Lys	195	1972	1	73	902	1227	550	724
Leu	234	3529	0	125	2019	1486	510	706
Met	48	700	0	22	391	526	129	160
Asn	144	857	1	34	494	1040	577	837
Pro	214	641	1	60	400	2036	437	821
Gin	96	1314	0	38	571	737	318	406
Arg	163	1844	0	56	864	1055	453	555
Ser	210	1263	2	85	1039	1796	679	782
Thr	136	1219	1	93	1387	1507	532	555
Val	88	1893	4	148	2648	1321	432	357
Trp	59	485	0	11	407	241	85	131
Tye	103	1047	0	66	1013	610	253	304

♭ 3-class: $H = helix$; $E = sheets$; and $C = coil$. 8-class: H = alpha helix; G = 3-helix (3/10 helix); I =5 helix (pi helix); E = extended strand, participates in beta-ladder; B = residue in isolated beta-bridge; S = bend; T =hydrogen bonded turn; and "*".

Table 2. Transition numbers and percentages of 20 different amino acids (N.=Number, P.=Percentage)

Amino	Transition HE		Transition HC		Transition EC		Transition EH		Transition CE		Transition CH	
Acid	N.	P.	N.	P.	N.	P.	N.	P.	N.	P.	N.	P.
Ala	0	0	321	11.53	220	5.49	6	5.00	250	6.10	118	4.39
Cys	2	10.53	33	1.19	65	1.62	0	0	48	1.17	34	1.26
Asp	1	5.26	77	2.77	235	5.87	14	11.67	360	8.78	383	14.23
Glu	2	10.53	244	8.77	184	4.60	2	1.67	183	4.46	112	4.16
Phe	3	15.79	121	4.35	220	5.49	3	2.50	117	2.85	59	2.19
Gly	1	5.27	63	2.26	152	3.80	8	6.67	545	13.29	289	10.74
His	0	0	65	2.34	99	2.47	5	4.17	101	2.46	69	2.56
Ile	0	0	130	4.67	356	8.89	4	3.33	169	4.12	41	1.52
Lys	0	0	229	8.23	188	4.70	6	5.00	243	5.93	87	3.23
Leu	0	0	416	14.95	372	9.29	2	1.67	200	4.88	118	4.39
Met	1	5.27	75	2.69	56	1.40	0	0	77	1.88	36	0.97
Asn	1	5.27	106	3.81	207	5.17	7	5.83	241	5.88	213	7.92
Pro	0	0	1	0.04	176	4.40	11	9.17	337	8.22	188	6.70
Gin	1	5.27	122	4.38	118	2.95	2	1.67	141	3.44	59	2.19
Arg	3	15.79	174	6.25	163	4.07	6	5.00	206	5.02	81	3.01
Ser	0	0	167	6.00	228	5.69	14	11.67	228	5.56	385	14.31
Thr	0	0	157	5.64	243	6.07	19	15.83	241	5.88	290	10.78
Val	2	10.53	120	4.31	474	11.84	3	2.50	260	6.34	54	2.01
Trp	0	0	39	1.40	88	2.20	5	4.17	38	0.93	24	0.89
Tyr	2	10.53	123	4.42	160	4.00	3	2.50	115	2.80	61	2.27
SUM	19	100.0	2783	100.0	4004	100.0	120	100.0	4100	100.0	2691	100.0

Table 1 shows the numbers of all the 20 amino acids for the corresponding secondary structures (the 3-state and 8-state). Here we assume that the amino acid conformational preference only depends on the type of amino acid and the type of secondary structure, but not on the position of the residue in the sequence.

2.2. Statistical Algorithm

In this section, we will consider the problem of counting the breaker numbers of the 20 different amino acids for their corresponding secondary structures. As for the ith protein sequence, its primary structure is $PS_i = \{P_i^1, P_i^2, \cdots, P_i^{L_i}\}$ and the corresponding secondary structure is $SS_i = \{S_i^1, S_i^2, \cdots, S_i^{L_i}\}$ (L_i is the sequence length and $i \in \{1, 2, \cdots, 340\}$). if the jth amino acid, PS_i^j, secondary structure, SS_i^j, is α-helix and the $j+1$th residue secondary structure is β-sheet, which indicates that the secondary structure change from α-helix to β-sheet at the j position, we denote it as *Transition HE*. In this way, we can get all the numbers of the *Transition* for the 20 different amino acids (as shown in Table 2).

The detailed pseudo-code is as follows:

input: the 340 protein sequences $PS_i = \{P_i^1, P_i^2, \cdots, P_i^{L_i}\}$ and their corresponding secondary structures $SS_i = \{S_i^1, S_i^2, \cdots, S_i^{L_i}\}$.

output: the numbers of breakers for the 3-class secondary structures

algorithm:

for $i = 1$ to 340
  for $j = 1$ to L_i
    if $S_i^j = H$ and $S_i^{j+1} = E$
      if $p_i^j = A$ then $TArray(1,1) = TArray(1,1) + 1$
      elseif $p_i^j = C$ then $TArray(2,1) = TArray(2,1) + 1$
      $\cdots\cdots$ 20 amino acids altogether
      elseif $p_i^j = Y$ then $TArray(20,1) = TArray(20,1) + 1$
      end
    elseif $S_i^j = H$ and $S_i^{j+1} = C \cdots\cdots$
    elseif $S_i^j = E$ and $S_i^{j+1} = H \cdots\cdots$
    elseif $S_i^j = E$ and $S_i^{j+1} = C \cdots\cdots$
    elseif $S_i^j = C$ and $S_i^{j+1} = H \cdots\cdots$
    elseif $S_i^j = C$ and $S_i^{j+1} = E \cdots\cdots$
    end
  end
  NofTHE $= \sum_{k=1}^{20}\{TArray(k,1)\}$
  NofTHC $= \sum_{k=1}^{20}\{TArray(k,2)\}$
  NofTEH $= \sum_{k=1}^{20}\{TArray(k,3)\}$
  NofTEC $= \sum_{k=1}^{20}\{TArray(k,4)\}$
  NofTCH $= \sum_{k=1}^{20}\{TArray(k,5)\}$
  NofTCE $= \sum_{k=1}^{20}\{TArray(k,6)\}$
end

where the $TArray$ stands for the *Transition Array* (20×6). The 20 rows represent the 20 different amino acids respectively, and the 6 columns give the numbers of *Transition HE, Transition HC, Transition EH, Transition EC, Transition CH* and *Transition CE*. The NofTHE represents *the Numbers of Transition HE*, and so do the NofTHC, NofTEH,

Table 3. Conformational Classification[♭♭] of the 20 different amino acids.

Amino	H Breaker		H Former		E Breaker		E Former		C Breaker		C Former		Amino Acids
Acid	N.	P.	N.	P.	N.	P.	N.	P.	N.	P.	N.	P.	Classification
Ala	321	**11.460**	124	4.41	226	5.48	250	**6.07**	368	5.42	541	**7.97**	*bff*
Glu	246	**8.78**	114	4.06	186	4.51	185	4.49	295	4.34	428	6.31	*bnn*
Lys	229	**8.17**	93	3.31	194	4.70	243	5.90	330	4.86	417	6.14	*bff*
Leu	416	**14.85**	120	4.27	374	**9.07**	200	4.86	318	4.68	788	**11.61**	*bbf*
Arg	177	**6.32**	87	3.10	169	4.10	209	**5.07**	287	4.23	337	4.97	*ffn*
Asp	78	2.78	397	**14.12**	249	6.04	361	**8.76**	743	**10.94**	312	4.60	*ffb*
Gly	64	2.28	297	**10.57**	160	3.88	546	**13.26**	834	**12.28**	215	3.17	*ffb*
Asn	107	3.82	220	**7.83**	214	5.19	242	**5.88**	454	**6.69**	313	4.61	*ffb*
Pro	1	0.04	199	**7.08**	188	4.56	337	**8.18**	525	**7.73**	178	2.62	*ffb*
Sir	167	5.96	399	**14.19**	242	**5.87**	228	5.54	613	**9.03**	395	5.82	*fbb*
Thr	157	5.60	309	**10.99**	262	**6.35**	241	5.85	531	**7.82**	400	5.89	*fbb*
Phe	124	4.43	62	2.21	223	**5.41**	120	2.91	176	2.29	341	**5.02**	*nbf*
Ile	130	4.64	45	1.61	360	**8.73**	169	4.10	210	3.09	486	**7.16**	*nbf*
Val	122	4.35	57	2.03	477	**11.57**	262	6.36	314	4.62	594	**8.75**	*nbf*
Cys	35	1.25	34	1.21	65	1.58	50	1.21	82	1.21	98	1.44	*nnn*
His	65	2.32	74	2.63	104	2.52	101	2.45	170	2.50	164	2.42	*nnn*
Met	76	2.71	26	0.92	56	1.36	78	1.89	103	1.52	131	1.93	*nnn*
Gin	123	4.39	61	2.17	120	2.91	142	3.45	200	2.95	240	3.54	*nnn*
Trp	39	1.39	29	1.03	93	2.26	28	0.68	62	0.97	127	1.87	*nnn*
Tye	125	4.46	64	2.28	163	3.95	117	2.84	176	2.59	283	4.17	*nnn*
SUM	2802	100.0	2811	100.0	4124	100.0	4119	100.0	6791	100.0	6787	100.0	

[♭♭] The 3 letters *b*, *f* & *n* stand for *Breaker*, *Former* and *Neutral* respectively. The first letter corresponds to α-Helix, the secondary the β-Sheets and the last the Coil. E.g., the classification of Ala is *bff*, this indicates that it is α-helix *Breaker*, β-sheets *Former* and Coil *Former* at the same time.

NofTEC, NofTCH and NofTCE correspondingly. The 6 percentages are determined by the equation 1.

$$P_{i,j} = \frac{TArray(i,j)}{\sum_{k=1}^{20}\{TArray(k,j)\}} \tag{1}$$

where $i \in (1, 2, \cdots, 20)$ and $j \in (1, 2, \cdots, 6)$.

2.3. Conformational Classification

From Table 2, we can get 6 different numbers and percentages of *Transition*. As to the *Transition HE*, we can consider it as the breaking of H, which indicates that the secondary structure α-helix is broke at this position, and at the same time we can also consider it as the beginning of E, which indicates that the β-sheet begin at the position of *Transition HE*. So the number of breaking and beginning of different secondary structure are got by the following equation:

$$\left.\begin{array}{l} \text{NofHBe=NofTEH+NofTCH} \\ \text{NofHBr=NofTHE+NofTHC} \\ \text{NofEBe=NofTHE+NofTCE} \\ \text{NofEBr=NofTEH+NofTEC} \\ \text{NofCBe=NofTEC+NofTHC} \\ \text{NofCBr=NofTCH+NofTCE} \end{array}\right\} \tag{2}$$

where $NofHBe$ represents the *Number of α-Helix Beginning*, $NofHBr$ represents the *Number of Helix Breaking*, and so do the $NofEBe$, $NofEBr$, $NofCBe$ and $NofCBr$ correspondingly. The Table 3 shows the detailed percentages of Breaking and Beginning, which are used to define the preference of a particular amino acid for a particular secondary structure.

To a given type secondary structure, we can get the two percentages: the *Breaking* percentage and the *Beginning* percentage. If the difference of the two percentages is less than 1%, we consider the amino acid as **Neutral** to the type of secondary structure (denoted by n); else if the *Beginning* percentage is 1% greater than the *Breaking* percentage, we think that the amino acid is a **Former** of the type secondary structure (denoted by f); if the *Breaking* percentage is 1% greater than *Beginning* percentage, we consider it as a **Breaker** of the secondary structure (denoted by b). The detailed amino acids conformational classifications are shown in Table 3.

3. ITS APPLICATION IN SSP AND SIMULATION RESULTS

The secondary structure prediction network and algorithm used in this study is similar to the paper [3] and [4]. The only difference between our approach and the references is: we use amino acids conformational classification information to reconstruct the input vectors by increasing dimension from 22 to 25, the last three dimensions indicate the residue (amino acid) conformational classification, and get a better results relative to the previous methods.

3.1. Training Error Convergent Rate by Different Sliding Window Widths

The relationship between the training error convergent rate of the network and the different window widths is shown in Fig.1. Here, we tested the network with the widths form 1 to 11, with an interval of 2, corresponding to the sub-figures of A, B, $\cdots$, F of Fig.1. The results show that the training error with the window widths of 1 and 3 were almost not convergent, while the ones with the widths greater than 7 can converge effectively (the sub-graph D with the number of 7 converged at about 1820 iterations, E with 9 at about 1680 iterations, F with 11 at 4100 iterations).

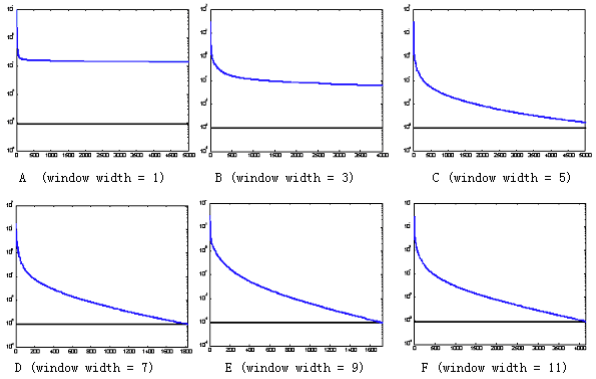

Fig. 1. Training error convergent rates vs. different widths

3.2. Accuracies by Different Sliding Window Widths

Table 4 gives the accuracies by different sliding windows widths . It can be found that the accuracy by single residue (the width is 1) information is very low, only about 55.31%. On the other hand, the prediction accuracy based on the surrounding residues (central around the residue to be predicted) information got about 20% improvements. From Table 4, we can see that the prediction by the width of 7 got the best performance with the accuracy of about 74.86%. In other words, the 7 sequential residues centered around the predicted amino acid have the most important role in determining the secondary structure.

Table 4. Accuracies vs. different sliding windows widths

Width	Training set $Q_3(\%)$	Testing set $Q_3(\%)$	Q_H (%)	Q_E (%)	Q_C (%)
1	68.32	54.73	60.83	38.25	58.47
3	93.18	69.24	69.80	50.31	76.44
5	98.64	72.23	70.65	58.33	79.86
7	**99.43**	**74.86**	**72.74**	**61.54**	**83.97**
9	100.0	72.86	70.32	58.84	79.21
11	100.0	72.44	68.31	62.02	81.32

3.3. Comparing with Other Methods

Several approaches, such as Chou & Fasman method, GOR method, PhD method and etc., have been applied in the protein secondary structure prediction successfully. Table 5 shows the detailed accuracies of different methods and indicate that our approach, conformational classification method, has a better accuracy of 74.28%.

Table 5. The accuracies of different methods

method	accuracy (Q_3)
Chou-Fasman	57%
Garnier, Osguthorpe and Robson	66%
Rost & Sander	68-72%
Conformational Classification	**74.86%**

4. CONCLUSIONS

In this paper we use amino acids conformational classification to improve the SSP problem and get a better accuracy. The experimental results indicate that the conformational preference has a promise future in the SSP problem. Future works will include tertiary structure prediction by conformational classification.

5. REFERENCES

[1] Michael Levitt, "Confromational Preferences of Amino Acids in Globular Proteins", *Biochemistry*, 17:4278-4285(1978).

[2] H.M. Berman, J. Westbrook, Z. Feng and etc., "The Protein Data Bank", *Nucleic Acids Research*, 28 pp, 235-242(2000).

[3] John-Marc Chandonia and Martin Karplus, "New methods for Accurate Prediction of Protein Secondary Structure", *PROTEINS: Structure, Function and Genetics*, 35:293-306(1999).

[4] V. DI Francesco, J. Garnier and P.J. Munson, "Improving protein secondary structure prediction with aligned homologous sequences", *Protein Science*, 5:106-113(1996).

[5] D.S.Huang and S.D.Ma, "Linear and nonlinear feedforward neural network classifiers: A comprehensive understanding", *Journal of Intelligent Systems*, Vol.9, No.1, 1-38, (1999).

[6] Gasteiger E., Gattiker A. and etc., "ExPASy: the proteomics server for in-depth protein knowledge and analysis", *Nucleic Acids Res*, 31:3784-3788(2003).