Problem Session (5) 2021/5/1 Yun-wei Xue

Topic: Prediction of Protein Secondary Structure

0. Introduction

0.1 Protein secondary structure concerned

α-helix:

- 3.6 amino acids per turn
- Hydrogen bond formed between every 4th residue

β-sheet:

- Consists of β-strands
- Formed by hydrogen bonds between amino acids

β-turn:

- Consists of 4 amino acid residues
- Cause a change in direction of polypeptide chain

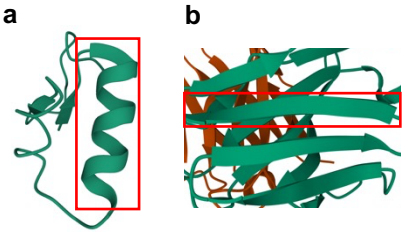


Figure 1. (a) Typical structures of α-helix (PDB ID: 3NIR, residues 6-18) (b) β-sheet (PDB ID: 1DEE, residues 19-25).

0.2 The development of methods for predicting protein secondary structure

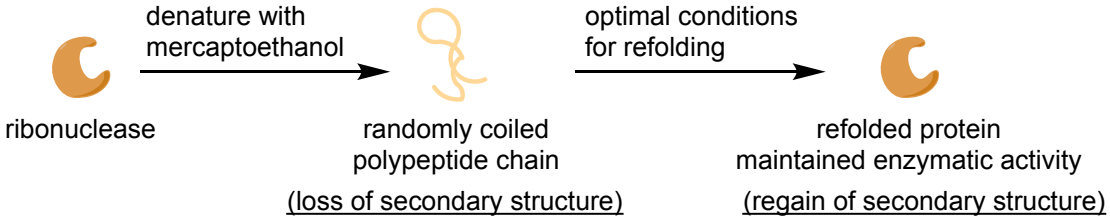


Figure 2. The regaining of enzymatic activity of ribonuclease after denaturing.

This classical experiment[1] implied that the amino acid sequence itself contains enough information for the secondary structure in a particular environment. Based on the result, methods for predicting protein secondary structure from amino acid sequence have emerged since then.

More than 40 methods have been proposed to date to predict the secondary structure of protein based on amino acid sequence. By their different approaches, they are basically categorized to three classes. The classification and representing methods are shown below:

1. Empirical statistical methods *(using parameters derived from analysis of known sequence and structure)*

(1) Chou-Fasman method[2, 3, 4] (discussed in this problem session)

(2) GOR(Garnier- Osguthorpe -Robson) method[5, 6]

2. Physical chemistry criteria

(1) Lim's method[7, 8] (stereochemical criteria)

(2) DSSP (Define Secondary Structure of Proteins)[9]

(3) STRIDE (**Str**uctural **Ide**ntification)[10]

3. Machine learning approach

(1) PSIPRED (**PSI**-blast based secondary structure **Pred**iction)[11]

(2) JPRED[12]…

1. Solution to Question 1

1.0 Assignment of amino acids with different roles in secondary structural forming

The principle of Chou-Fasman algorithm is based on the statistical analyses of number of occurrence of one single given amino acid in a specific secondary structure (α-helix, β-sheet, and β-turn). From the known sequence and structures of proteins as a data base, 20 amino acids were classified as "Former", "Breaker", or "indifferent" for relative secondary structures based on the carefully calculated conformational parameters. These conformational parameters, symbolized by P_α, P_β etc., are presumed to contain information about the physical-chemical properties of the amino acids for determining the secondary structures of proteins, such as hydrophobicity, hydrogen bond forming capacity etc.

In question 1, an amino acid sequence of Staphylococcal Nuclease with 149 residues is given. The first thing to do is to assign the amino acids with their corresponding role in the Chou-Fasman method applied. As shown below in *Figure 3*, the first amino acid A (Ala) is a strong α former as well as a β indifferent, thus assigned as "H" in column α-helix former and "i" in β-sheet former.

(Here the former residues are highlighted with blue and green background respectively for easier identification.)

residue	amino acid	3 letter abbr.	α-helix former	β-sheet former
1	A	Ala	H	i
2	T	Thr	i	h
3	S	Ser	i	b
4	T	Thr	i	h
5	K	Lys	h	b
6	K	Lys	h	b
7	L	Leu	H	h
8	H	His	I	i
9	K	Lys	h	b
10	E	Glu	H	B

Figure 3. The assignment of roles for amino acids 1-10 in the given sequence.
(*For identification, the elements concerning α-helix is colored in blue and the elements concerning β-sheet is colored in green, respectively*)

After the assignment of the whole sequence, the clusters of α-helix formers or β-sheet formers can be easily observed which roughly represent the regions showing higher probability of forming helical or sheet secondary structures. As an example, residues 5 to 10 (blue rectangle shown in *Figure 3*) is a cluster of α-helical formers with 5 formers out of 6 residues (which meets the requirement of 4 out 6 residues are helical formers) and is considered to be a region with high probability of forming α-helix structure. However, clusters observed alone is not sufficient enough to determine a specific secondary structure region. To be more precise, the conformational parameters P_x, their average values $\langle P_x \rangle$, as well as other factors including the existence of Proline, the emerge of β-turn structures, are taken into consideration when applied with Chou-Fasman algorithm.

1.1 α-Helix assignment

**The assignment of a-helix and b-sheet structures follows a process of initiation, extension, and termination.*

residue	amino acid	3 letter abbr.	P_α	P_β	a	b	<P_α>	<P_β>
12	A	Ala	1.42	0.83	H	i	1.17	1.16
13	T	Thr	0.83	1.19	i	h		
14	L	Leu	1.21	1.3	H	h		
15	I	Ile	1.08	1.6	h	H		
16	K	Lys	1.16	0.74	h	b		
17	A	Ala	1.42	0.83	H	i		
18	I	Ile	1.08	1.6	h	H		

Figure 4. Residues 12 to 18 as a typical α former cluster to initiate α-helix.

1.1.1 Initiation: Following the rule described in the question "*A cluster of four helical residues (H_α or h_α) out of six along the protein sequence will initiate α helix.*", residues 12 to 18 are considered to meet the requirement to initiate α-helix. Meanwhile, among the given residues 12 to 18, this sequence also meets the requirement to initiate β-sheet since there is a cluster of β formers (residue 13 to 15) as well. In this case, the average values of respective P_α or P_β (symbolized by $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$) of the given sequence are calculated to determine which secondary structure to assign. As a result, in this method, the $\langle P_\alpha \rangle$ value (1.17) is larger than $\langle P_\beta \rangle$ value (1.16). So, this segment of residues 12 to 18 is assigned to initiate α-helix.

residue	amino acid	3 letter abbr.	P_α	a	<P_α>	assignment
5	K	Lys	1.16	h	1.1325	
6	K	Lys	1.16	h	1.1325	
7	L	Leu	1.21	H	1.225	
8	H	His	1	I	1.065	termination
9	K	Lys	1.16	h	1.17	α-helix extension
10	E	Glu	1.53	H	1.0875	
11	P	Pro	0.57	B	1.0075	
12	A	Ala	1.42	H	1.135	α-helix initiation
13	T	Thr	0.83	i	1.07	
14	L	Leu	1.21	H	1.2175	
15	I	Ile	1.08	h	1.185	
16	K	Lys	1.16	h	1.1675	
17	A	Ala	1.42	H	1.02	
18	I	Ile	1.08	h	0.9175	
19	D	Asp	1.01	I	0.855	termination
20	G	Gly	0.57	B	0.8675	
21	D	Asp	1.01	I	1.015	
22	T	Thr	0.83	i	1.065	

Figure 5. Residues 9 to 18 assigned as α-helix.

1.1.2 extension-termination: The assigned α-helix initiation (residues 12 - 18) extends in both directions until a termination is found. In the direction towards N-terminus, Pro at residue 11 was found. Following the rule described in the question "Proline (Pro) cannot occur in the inner helix or at the C-terminal helical end but can occur within the last three residues at the N-terminal end." Thus, the α-helical extension ended at residue 9 for the existence of Pro at residue 11. On the other hand, towards the C-terminus, the tetrapeptide (residue 19 to 22) came up with a $\langle P_\alpha \rangle$ value below 1.00 which also terminates the extension of the α-helix. Thus, the segment of residues 9 to 18 is assigned with α-helix.

1.2 β-sheet assignment

residue	amino acid	3 letter abbr.	P_α	P_β	a	b	$<P_\alpha>$	$<P_\beta>$
32	M	Met	1.45	1.05	H	h	1.14	1.27
33	T	Thr	0.83	1.19	i	h		
34	F	Phe	1.13	1.38	h	h		
35	R	Arg	0.98	0.93	i	i		
36	L	Leu	1.21	1.3	H	h		
37	L	Leu	1.21	1.3	H	h		
38	L	Leu	1.21	1.3	H	h		
39	V	Val	1.06	1.7	h	H		

Figure 6. Residues 32 to 39 as a typical β former cluster to initiate β-helix.

1.2.1 Initiation: The assignment of β-sheet also follows the initiation-extension-termination process similar to the α-helix assignment described in section 1.1. Take residues 32 to 39 as an example. Among this sequence, 7 out of 8 residues are assigned as β formers which is qualified to initiate a β-sheet structure. Since the sequence itself is also rich in α formers at the same time, $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ were calculated to be 1.14 and 1.27 respectively in which the value $\langle P_\beta \rangle$ stands out. As a result, residues 32 to 39 are assigned to initiate a β-sheet structure.

residue	amino acid	3 letter abbr.	P_β	a	$<P_\beta>$	assignment
27	Y	Tyr	1.47	H	1.015	
28	K	Lys	0.74	b	0.785	
29	G	Gly	0.75	b	0.8625	
30	Q	Gln	1.1	h	0.9725	termination
31	P	Pro	0.55	B	1.0425	β-sheet extension
32	M	Met	1.05	h	1.1375	
33	T	Thr	1.19	h	1.2	
34	F	Phe	1.38	h	1.2275	
35	R	Arg	0.93	i	1.2075	β-sheet initiation
36	L	Leu	1.3	h	1.4	
37	L	Leu	1.3	h	1.21	
38	L	Leu	1.3	h	1.1825	
39	V	Val	1.7	H	0.995	
40	D	Asp	0.54	B	0.6625	termination
41	T	Thr	1.19	h	0.825	
42	P	Pro	0.55	B	0.7125	
43	E	Glu	0.37	B	0.7925	

Figure 7. Residues 31 to 39 assigned as β-sheet.

1.2.2 Extension-termination: Likely, the assigned β-sheet initiation extends in both directions towards N- and C- termini. In both directions, tetrapeptides with low $\langle P_\beta \rangle$ values occurred at residue 30 and residue 40, respectively. The β-sheet structure initiated at residues 32 to 39 terminated at residue 30 and residue 40. As a result, the segment of residues 31 to 39 is assigned as β-sheet structure.

1.3 β-turn assignment

Due to the different structural features of β-turn comparing to α-helix and β-sheet, the assignment of β-turn relies on two different parameters: P_t and p_t. The capitalized P_t stands for the total occurrence of one residue in the β-turns; while the lower case p_t means the relative probability that a tetrapeptide (four residues) will form a β-turn. Other than the initiation-extension-termination process, the assignment of β-turn mainly consists of localization through parameter p_t and comparison between P_t and P_α or P_β.

residue	amino acid	3 letter abbr.	p_t
27	Y	Tyr	0.000176
28	K	Lys	1.18E-05
29	G	Gly	1.87E-05
30	Q	Gln	2.46E-05

Figure 8. Residues 27 to 30 with a p_t value larger than 0.75×10^{-4}.

(For identification, the elements concerning β-turn are colored by light yellow)

1.3.1 Localization of β-turn: Here, take residues 27 to 30 as a typical example for the assignment of β-turn. The f_i value of Tyr, f_{i+1} value of Lys, f_{i+2} value of Gly, and f_{i+3} value of Gln are 0.082, 0.115, 0.190, and 0.098 respectively. The p_t value of residue 27 to 30 is thus calculated through the formula: $p_t = f_i \times f_{i+1} \times f_{i+2} \times f_{i+3}$, which gives out the calculated p_t value: 0.000176 larger than the determined cutoff value 0.75×10^{-4}. This result shows that this tetrapeptide has potential to form a β-turn based on the frequency analysis.

residue	amino acid	P_α	P_β	P_t	a	b	$\langle P_\alpha \rangle$	$\langle P_\beta \rangle$	$\langle P_t \rangle$	p_t	assignment
27	Y	0.69	1.47	1.14	b	H	0.8825	1.015	1.1725	0.000176	β-turn
28	K	1.16	0.74	1.01	h	b	0.8525	0.785	1.2675	1.18E-05	
29	G	0.57	0.75	1.56	B	b	0.925	0.8625	1.165	1.87E-05	
30	Q	1.11	1.1	0.98	h	h	0.99	0.9725	1.015	2.46E-05	

Figure 9. Residues 27 – 30 assigned as β-turn.

1.3.2 Determination of β-turn: The p_t value only represents the higher probability of β-turn occurrence of the tetrapeptide. It does not rule out the possibility that this tetrapeptide actually localizes inside α-helix or β-sheet. In this case, the value of $\langle P_t \rangle$ is calculated and compared with the corresponding $\langle P_\alpha \rangle$ or $\langle P_\beta \rangle$. As marked by red rectangle in *Figure 9*, $\langle P_t \rangle = 1.1725$ is apparently larger than $\langle P_\alpha \rangle$ or $\langle P_\beta \rangle$, thus determining the segment of residues 27 to 30 assigned as β-turn.

1.4 Concise answer to question 1

After the assignment process of all sequence with the method described above, the answer to question 1 is shown as below (for the details, please refer to the attached Excel file):

Table 1. The assignment of Staphylococcal Nuclease structure using Chou-Fasman algorithm

Predicted results			
residues	secondary structure		
9-18	α-helix	β-turn:	3-6
23-26	β-sheet		19-22
31-39	β-sheet		27-30
59-76	α-helix		41-44
88-92	β-sheet		46-49
97-104	α-helix		55-58
109-115	β-sheet		77-80
120-137	α-helix		83-86
			93-96
			105-108
			116-119
			141-144
			146-149

2. Evaluation of the obtained result

2.1 Comparison between predicted and experimental results

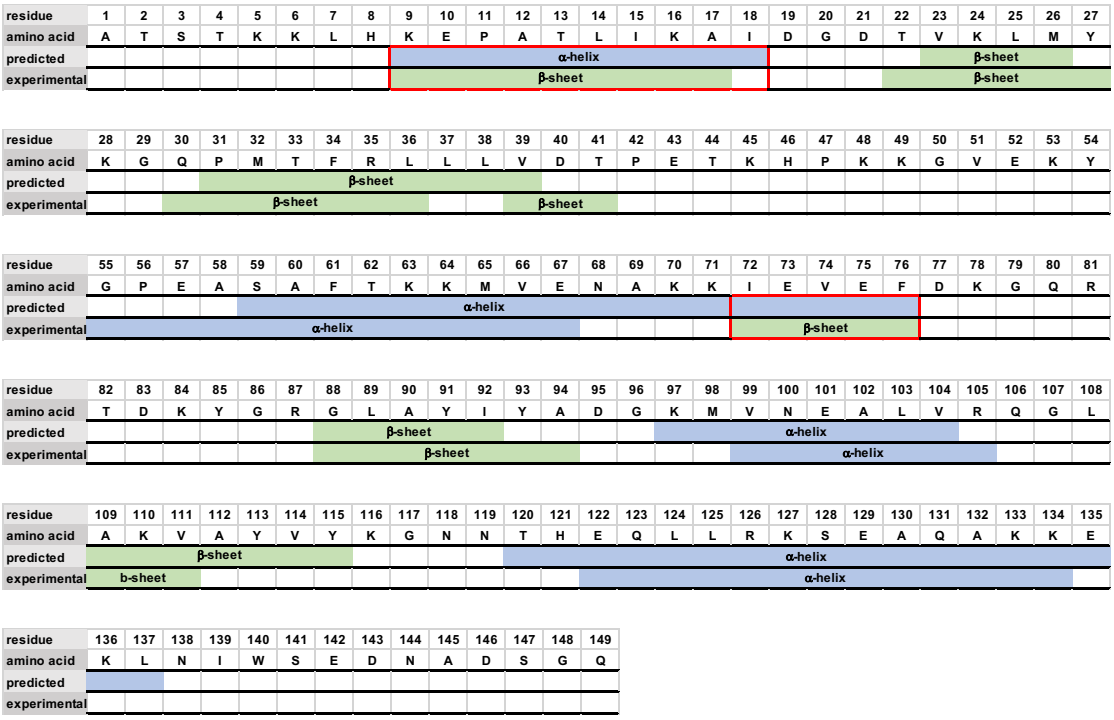


Figure 10. Schematic display of difference between the predicted and experimental results.

As the specific data about β-turn is not modeled in the PDB results, a comparison was conducted between the α-helix and β-sheet structures. Shown in *Figure 10*, the distribution of α-helix and β-sheet are roughly in accordance with each other, except for the highlighted parts with red rectangle. There apparently exists a tendency to over predict α-helix as well as under predict the β-sheet structure when Chou-Fasman algorithm is applied.

Predicted results		Experimental results (PDB ID: 1SNP)		Residues overpredicted	Residues missed
9-18	α-helix	9-17	β-sheet	1	8
23-26	β-sheet	22-27	β-sheet	0	2
31-39	β-sheet	30-36	β-sheet	2	1
		39-41	β-sheet	0	2
59-76	α-helix	55-67	α-helix	4	4
		72-76	β-sheet	0	5
88-92	β-sheet	88-94	β-sheet	0	2
97-104	α-helix	99-105	α-helix	2	1
109-115	β-sheet	109-111	β-sheet	4	0
120-137	α-helix	122-134	α-helix	5	0
			Total error	18	25
			accuracy	71%	

Figure 11. The comparison between the predicted and experimental results.

To evaluate the prediction method applied this time, the accuracy was calculated as: $accuracy = \frac{n(residues)-n(total\ error)}{n(residues)} \times 100\%$. As a result, a moderate accuracy of 71% is given for Staphylococcal Nuclease this time. (The actual accuracy of Chou-Fasman algorithm was determined to be quite low, only 50 – 60% proved by a large number of proteins.)

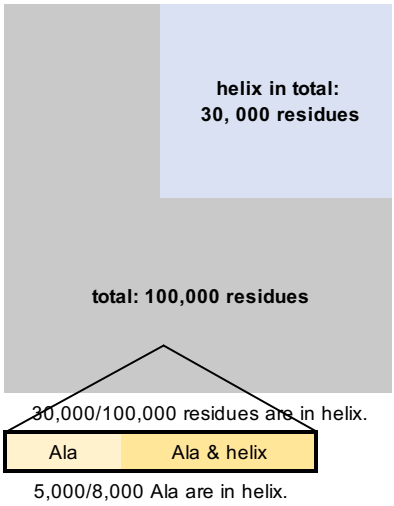
2.2 Main problems of the Chou-Fasman algorithm

2.2.1 Size of data base

As an empirical method based on the statistical analysis of known sequence and structure, the size of data base is absolutely important for the accuracy of the prediction result. However, limited by the efficiency and computing resource at that time, the data size utilized for this method is actually rather small. The original data base from which the conformation parameters are calculated, contained only 15 proteins, with 2473 residues.[2] At the same year, the data base size approximately doubled to 29 proteins, with 4741 residues.[3] Finally, 64 proteins, with 11,445 residues [4] were taken into the account when this method was last updated. Even though, the size is still too small to compare with recent prediction method through neural network and machine learning which is literally trained on PDB (Protein Data Bank) and fed with countless data.[13] The small sized data base has largely limited the accuracy of the Chou-Fasman algorithm.

2.2.2 Concept of the Chou-Fasman algorithm

The bottom logic of the Chou-Fasman algorithm is based on the frequency of one single particular amino acid in an assigned secondary structure. A simple example is given below:



Assume that a data set with total 100,000 residues, 3/10 of them are in helical conformation. Inside the data set, Ala are observed 8,000 times and among the 8,000 Ala, 5,000 Ala are in helix. In this case:

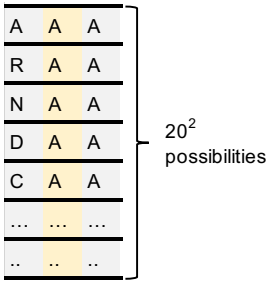
$$\frac{f(helix)}{f(total)} = \frac{30{,}000}{100{,}000}; \frac{f(Ala,\ helix)}{f(Ala)} = \frac{5{,}000}{8{,}000}$$

$$p(Ala, helix) = \left[\frac{f(Ala,helix)}{f(Ala)}\right] / \left[\frac{f(helix)}{f(total)}\right] = 2.08$$

As described, the propensity of Ala as helix in this data set is calculated to be 2.08, implying the potential of Ala to form helical structure is about 2-fold larger than average amino acid.

Figure 12. Propensity of Ala as helix in an assumed data set

The concept of the Chou-Fasman algorithm is based on the statistical analysis of frequency of each amino acid in each secondary structure but with more rigorous method than displayed above. However, this idea ignored the context of a particular residue, but only focus on the propensity of one amino acid. An idea to improve this concept is to take the context of the residue into consideration. For example, neighboring residues of one amino acid can be treated as one unit for analysis instead of one single amino acid alone. In this case, 20^2 possibilities for each amino acid could be proposed. Then calculate the propensity of the amino acid in the triple as helix or sheet structure. It will be like: instead of $p(A, helix)$, $p[(A_{i-1}A_iA_{i+1}), helix]$ will be calculated for prediction. This idea of considering neighboring residues was utilized in the GOR method[5, 6] for protein secondary structure prediction in which actually 8 residues instead of 1 residue in each direction were applied.



3. Comments

Other than the two problems discussed above, there still exists short comings for the Chou-Fasman method when considering long range interaction, multichain, resulting its poor accuracy and a tendency to under predict β-sheet structure. For the time being, although Chou-Fasman is still used for protein secondary prediction because of its easiness of conduction and fast speed, the mainstream of protein prediction method is relying more on the machine learning and homologue in data bank.

Here an overview of protein structure prediction is depicted below:

		Overall Accruracy[14]
1st generation	Chou-Fasman GOR	**50-60%**
2nd generation	DSSP STRIDE	**55-70%**
3rd generation	PSIPRED JPRED	**70-80%**

Reference:

1. Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H. *P. Natl. Acad. Sci. USA* **1961**, *47*, 1309.
2. Chou, P. Y.; Fasman, G. D. *Biochemistry* **1974**, *13*, 211.
3. Chou, P. Y.; Fasman, G. D. *Biochemistry* **1974**, *13*, 222.
4. Fasman, G. D. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Springer Verlag: New York, US, **1989**; pp 391–416.
5. Garnier, J.; Osguthorpe, D. J.; Robson, B. *J. Mol. Biol.* **1978**, *120*, 97.
6. Garnier, J.; Gibrat, J.; Robson, B. *Method Enzymol.* **1996**, 540.
7. Lim, V. I. *J. Mol. Biol.* **1974**, *88*, 857.
8. Lim, V. I. *J. Mol. Biol* **1974**, *88*, 873.
9. Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.
10. Frishman, D.; Argos, P. *Proteins* **1995**, *23*, 566.
11. D. T. Jones *J. Mol. Biol* **1999**, *292*, 195.
12. Drozdetskiy, A.; Cole, C.; Procter, J.; Barton, G. J. *Nucleic Acids Res.* **2015**, *43*, W389.
13. Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. *Nature* **2020**, *577*, 706.
14. Montgomerie, S.; Sundararaj, S.; Gallin, W.; Wishart, D.; *BMC Bioinformatics* **2006**, *7*, 301.