9

Chou–Fasman Prediction of the Secondary Structure of Proteins

The Chou–Fasman–Prevelige Algorithm

Peter Prevelige, Jr., and Gerald D. Fasman

I. INTRODUCTION

The Chou–Fasman algorithm for the prediction of protein secondary structure is one of the most widely used predictive schemes. This is because of its relative simplicity and its reasonably high degree of accuracy.

Peter Prevelige, Jr. • Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. *Gerald D. Fasman* • Graduate Department of Biochemistry, Brandeis University, Waltham, Massachusetts 02254.

A number of modifications of the Chou–Fasman algorithm have been developed and published (see G. D. Fasman, Chapter 6, this volume, for a review). However, in general these suffer from one of two faults: either they are completely computerized and hide much of the decision-making process from the user or they leave the user to make decisions but do not adequately describe the decision-making process used by the authors.

This chapter attempts to outline the approach that has been successfully employed by the authors over the past several years. The approach is one in which a computer program is employed to perform the arithmetic calculations and then the data reduction is performed by hand. This approach utilizes the computer to reduce the tedious calculations while at the same time allowing the individual to bring his experience and intuition to bear. The computer program itself was developed from ideas in a program originally written by Dr. George Long and Jeff Siegel in 1979.

The first section of this chapter reviews the Chou–Fasman method for prediction of protein structure. This is followed by a section that lays out the mechanics of operating the program and then by a discussion of the process of data reduction. Finally, worked examples are provided in the hope that they will make more concrete the many considerations involved in predicting a protein secondary structure.

II. REVIEW OF THE METHOD AND RATIONALE OF THE CHOU–FASMAN ALGORITHM

The Chou-Fasman algorithm is an algorithm to predict the secondary structure of proteins from their amino acid sequence. It falls into the class of the statistical approach as discussed by Fasman (Chapter 6, this volume).

The x-ray-determined structures of 15 proteins containing 2473 amino acid residues were carefully examined, and the number of occurrences of a given amino acid in the α helix, β sheet, and coil was tabulated (Table I). From this, the conformational parameters for each amino acid were calculated by considering the relative frequency of a given amino acid within a protein, its occurrence in a given type of secondary structure, and the fraction of residues occurring in that type of structure (Chou and Fasman, 1974a). This conformational parameter is essentially a measure of a given amino acid's preference to be found in α helix, β sheet, or coil. These parameters, symbolized by P_α, P_β, and P_c, respectively, presumably contain information about the physical–chemical parameters defining protein stability, such as hydrophobicity, properly weighted for their relative importance. These parameters therefore should be useful for predicting a protein's secondary structure based on the amino acid sequence.

Having computed these conformational parameters, Chou and Fasman formulated a set of empirical rules for predicting secondary structure (Chou and Fasman, 1974b). The development of these empirical rules was guided by underlying considerations of protein structure. These rules, when applied by Chou and Fasman, resulted in a 70–80% predictive accuracy. The rules were never developed as a computer algorithm and hence lack the type of rigorous definition that a computer algorithm requires. This has led to a wide variety of implementations, which have an equally wide variety of accuracies.

Chou and Fasman later extended the analysis of α helix, β sheet, and coil to include 29 proteins of known x-ray structure. This increased the total number of residues classified to 4741, or approximately double the initial number (Chou and Fasman, 1978). The most pronounced change occurred for Met. This change resulted from an underrepresentation of Met in the initial 15 proteins examined. Less pronounced changes were also seen in Asn, Asp, Ala, His, Gly, Ile, Lys, and Tyr (Table II).

Table I. Assignment of Amino Acids as Formers, Breakers, and Indifferent for Helical and β-Sheet Regions in Proteins Based on P_α and P_β Values[a]

Helical residues[b]	P_α		β-Sheet residues[c]	P_β	
$Glu^{(-)}$	1.53	H_α	Met	1.67	H_β
Ala	1.45		Val	1.65	
Leu	1.34		Ile	1.60	
$His^{(+)}$	1.24	h_α	Cys	1.30	h_β
Met	1.20		Tyr	1.29	
Gln	1.17		Phe	1.28	
Trp	1.14		Gln	1.23	
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
$Lys^{(+)}$	1.07	I_α	Trp	1.19	I_β
Ile	1.00		Ala	0.97	
$Asp^{(-)}$	0.98	i_α	$Arg^{(+)}$	0.90	i_β
Thr	0.82		Gly	0.81	
Ser	0.79		$Asp^{(-)}$	0.80	
$Arg^{(+)}$	0.79		$Lys^{(+)}$	0.74	
Cys	0.77		Ser	0.72	
Asn	0.73	b_α	$His^{(+)}$	0.71	b_β
Tyr	0.61		Asn	0.65	
Pro	0.59	B_α	Pro	0.62	
Gly	0.53		$Glu^{(-)}$	0.26	B_β

[a] Chou and Fasman (1974b).
[b] Helical assignments: H_α, strong α former; h_α, α former; I_α, weak α former; i_α, α indifferent; b_α, α breaker; B_α, strong α breaker. I_α assignments are also given to Pro and Asp (near the N-terminal helix) as well as Arg (near the C-terminal helix).
[c] β-Sheet assignments: H_β, strong β former; h_β, β former; I_β, weak β former; i_β, β indifferent; b_β, β breaker; B_β, strong β breaker; b_β assignment is also given to Trp (near the C-terminal β region).

A similar analysis, using the 29-protein data base, was performed for amino acid residues that were found in β turns (Chou and Fasman, 1977). The conformational parameter P_t was determined. In the case of turns, a significant difference was also observed in the frequency of residues in the first, second, third, and fourth positions of β turns for all residues (Table II). Some residues were found to have a dramatic positional preference, e.g., proline. Proline occurs 30% of the time in position number 2 of the β bend but less than 4% of the time in position number 3. Therefore, for the prediction of turns a method to factor in positional preference was devised (Chou and Fasman, 1979).

The Chou–Fasman algorithm is simple in principle. Using the conformational parameter, one finds nucleation sites within the sequence and extends them until a stretch of amino acids is encountered that is not disposed to occur in that type of structure or until a stretch is encountered that has a greater disposition for another type of structure. At that point, the structure is terminated. This process is repeated throughout the sequence until the entire sequence is predicted. The conformational parameters for coil are not employed; coil is predicted by default.

Table II. Conformational Parameters for α-Helical, β-Sheet, and β-Turn Residues in 29 Proteins[a]

	P_α			P_β			P_t		f_i		f_{i+1}		f_{i+2}		f_{i+3}	
Glu	1·51	H_α	Val	1·70	H_β	Asn	1·56	Asn	0·161	Pro	0·301	Asn	0·191	Trp	0·167
Met	1·45		Ile	1·60		Gly	1·56	Cys	0·149	Ser	0·139	Gly	0·190	Gly	0·152
Ala	1·42		Tyr	1·47		Pro	1·52	Asp	0·147	Lys	0·115	Asp	0·179	Cys	0·128
Leu	1·21		Phe	1·38		Asp	1·46	His	0·140	Asp	0·110	Ser	0·125	Tyr	0·125
Lys	1·16	h_α	Trp	1·37	h_β	Ser	1·43	Ser	0·120	Thr	0·108	Cys	0·117	Ser	0·106
Phe	1·13		Leu	1·30		Cys	1·19	Pro	0·102	Arg	0·106	Tyr	0·114	Gln	0·098
Gln	1·11		Cys	1·19		Tyr	1·14	Gly	0·102	Gln	0·098	Arg	0·099	Lys	0·095
Trp	1·08		Thr	1·19		Lys	1·01	Thr	0·086	Gly	0·085	His	0·093	Asn	0·091
Ile	1·08		Gln	1·10		Gln	0·98	Tyr	0·082	Asn	0·083	Glu	0·077	Arg	0·085
Val	1·06		Met	1·05		Thr	0·96	Trp	0·077	Met	0·082	Lys	0·072	Asp	0·081
Asp	1·01	I_α	Arg	0·93	i_β	Trp	0·96	Gln	0·074	Ala	0·076	Thr	0·065	Thr	0·079
His	1·00		Asn	0·89		Arg	0·95	Arg	0·070	Tyr	0·065	Phe	0·065	Leu	0·070
Arg	0·98	i_α	His	0·87		His	0·95	Met	0·068	Glu	0·060	Trp	0·064	Pro	0·068
Thr	0·83		Ala	0·83		Glu	0·74	Val	0·062	Cys	0·053	Gln	0·037	Phe	0·065
Ser	0·77		Ser	0·75	b_β	Ala	0·66	Leu	0·061	Val	0·048	Leu	0·036	Glu	0·064
Cys	0·70		Gly	0·75		Met	0·60	Ala	0·060	His	0·047	Ala	0·035	Ala	0·058
Tyr	0·69	b_α	Lys	0·74		Phe	0·60	Phe	0·059	Phe	0·041	Pro	0·034	Ile	0·056
Asn	0·67		Pro	0·55		Leu	0·59	Glu	0·056	Ile	0·034	Val	0·028	Met	0·055
Pro	0·57	B_α	Asp	0·54	B_β	Val	0·50	Lys	0·055	Leu	0·025	Met	0·014	His	0·054
Gly	0·57		Glu	0·37		Ile	0·47	Ile	0·043	Trp	0·013	Ile	0·013	Val	0·053

[a] P_α, P_β, and P_t are conformational parameters of helix, β sheet, and β turns, f_i, f_{i+1}, f_{i+2}, f_{i+3} are bend frequencies in the four positions of the β turn. H_α, H_β, etc., as defined previously (Chou and Fasman, 1974b). From Chou and Fasman (1977, 1978).

An abbreviated set of rules follows (Fasman, 1985).

1. A cluster of four helical residues (Hα or hα) out of six along the protein sequence will initiate a helix. The helical segment is extended in both directions until sets of tetrapeptide breakers ($\langle P_\alpha \rangle < 1.00$) are reached. Proline 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. The inner helix is defined as one omitting the three helical end residues at both the amino and carboxyl ends. Any segment that is at least six residues long with $\langle P_\alpha \rangle > 1.03$ and $\langle P_\alpha \rangle > \langle P_\beta \rangle$ is predicted as helical.

2. A cluster of three β formers or a cluster of three β formers out of five residues along the sequence will initiate a β sheet. The β sheet is propagated in both directions until terminated by a set of tetrapeptide breakers ($\langle P_\beta \rangle < 1.00$). Any segment with $\langle P_\beta \rangle > 1.05$ as well as $\langle P_\beta \rangle > \langle P_\alpha \rangle$ is predicted as β sheet.

3. The probability of a bend at residue i is calculated from $p_t = f_i \times f_{i+1} \times f_{i+2} \times f_{i+3}$. Tetrapeptides with $p_t > 0.75 \times 10^{-4}$ as well as $\langle P_t \rangle > 1.00$ and $\langle P_\alpha \rangle < \langle P_t \rangle > \langle P_\beta \rangle$ are predicted as β-turns.

4. Any segment containing overlapping α and β regions is helical if $\langle P_\alpha \rangle > \langle P_\beta \rangle$ or β sheet if $\langle P_\beta \rangle > \langle P_\alpha \rangle$.

III. OPERATION OF THE PREDICTION PROGRAM: THE INPUT FILE

The input file must be an ASCII file containing the single-letter codes for the amino acids. The format of the file does not matter—the letters can be upper or lower case, there can be one

```
> PREDICT

Enter name of input file: B:SUBTIL

Enter name of output file: B:SUBTIL.PRN

Enter the name of the protein ----> SUBTILISIN BPN'

Use database of 29 proteins (Chou-Fasman '78) ----> enter 29

Use database of 64 proteins (Chou '79) ----> enter 64

29

ASCII format enter A, or LOTUS format, enter L:

L

PROGRAM EXECUTING.
```


		Pa	Pb	Pt	a	b	⟨Pa⟩		⟨Pb⟩		⟨Pt⟩	⟨pt⟩	
1	A	142	83	66	H	i	109	*	109	*	89	3.90e-005	
2	Q	111	110	98	h	h	87		102	*	110	1.96e-005	
3	S	77	75	143	i	b	77		111	*	114	2.45e-005	
4	V	106	170	50	h	H	72		111	*	118	3.23e-004	*
5	P	57	55	152	B	B	72		111	*	118	6.68e-005	
6	Y	69	147	114	b	H	77		116	*	115	2.07e-005	
7	G	57	75	156	B	b	87		107	*	111	6.00e-005	
8	V	106	170	50	h	H	100	*	128	*	84	1.79e-005	
9	S	77	75	143	i	b	103	*	104	*	97	1.45e-005	
10	Q	111	110	98	h	h	119	*	106	*	78	1.05e-005	
11	I	108	160	47	h	H	105	*	93		91	1.18e-005	
12	K	116	74	101	h	b	114	*	73		96	8.24e-006	
13	A	142	83	66	H	i	115	*	87		85	4.42e-005	
14	P	57	55	152	B	B	105	*	88		93	1.51e-005	
15	A	142	83	66	H	i	110	*	93		90	1.48e-005	
16	L	121	130	59	H	H	102	*	100	*	98	3.51e-005	
17	H	100	87	95	I	i	86		86		123	1.09e-004	*
18	S	77	75	143	i	b	78		101	*	127	2.79e-004	*
19	Q	111	110	98	h	h	80		112	*	116	5.66e-005	

Fasman data set. All the data base values and all the output values have been multiplied by a factor of 100.

The four columns on the right are the calculated output data. The first three of these are the tetrapeptide averages calculated from the P_α, P_β, and P_t values. The asterisks indicate when the tetrapeptide average is greater than 100. The last column is the position-dependent turn calculation $\langle p_t \rangle$. This value is flagged with an asterisk when the calculated value is above 100.

An approach to data reduction is presented below in terms of the search for helical regions, the search for β-sheet regions, and the search for β-turn sites. In fact, the process is most easily performed in a single pass through the data, checking for all three types of structure, rather than three passes as might be implied by the presentation.

V. ANALYSIS OF OUTPUT

A. Search for Helical Regions

1. Helix Nucleation

Helical regions are nucleated by the presence of four sequential tetrapeptide averages with a value greater than 100. These are easily located by visually scanning for a cluster of asterisks.

2. Helix Propagation

The helix is extended towards the carboxyl terminus until the tetrapeptide average drops below 100. Thus, the following three residues are included in the helical region. At this point, the residues immediately following are examined. If they are of class H or h, they are generally included; if they are of class i, with high P_α values they are included. The $\langle P_\alpha \rangle$ for the entire region is then calculated and recalculated with the exclusion of the terminal class i residues (should any be present). If the calculated value is above the threshold value of 103, the region may be assigned as helical. In deciding whether to include a terminal residue, one should consider that the distribution of number of residues per helical segment has peaks at multiples of four residues (Fig. 3). This reflects the fact that the hydrogen-bonding scheme in an α helix has residue i hydrogen bonded to residue $i + 3$. Therefore, inclusion of unfavorable residues is more acceptable when they make the total number a multiple of four.

It will often be the case that the tetrapeptide average will drop below 100 for a residue while it is above 100 for the residues on either side. The somewhat low residue is generally included in the structural region.

3. Proline as a Helix Breaker

Proline cannot occur in the inner helix or at the C-terminal end. It can occupy the first turn in the N-terminal helix.

B. Search for β-Sheet Regions

1. β-Sheet Nucleation

β-Sheet regions are nucleated by the presence of three sequential tetrapeptide averages with a value greater than 100. These are easily located by visually scanning for a cluster of asterisks.

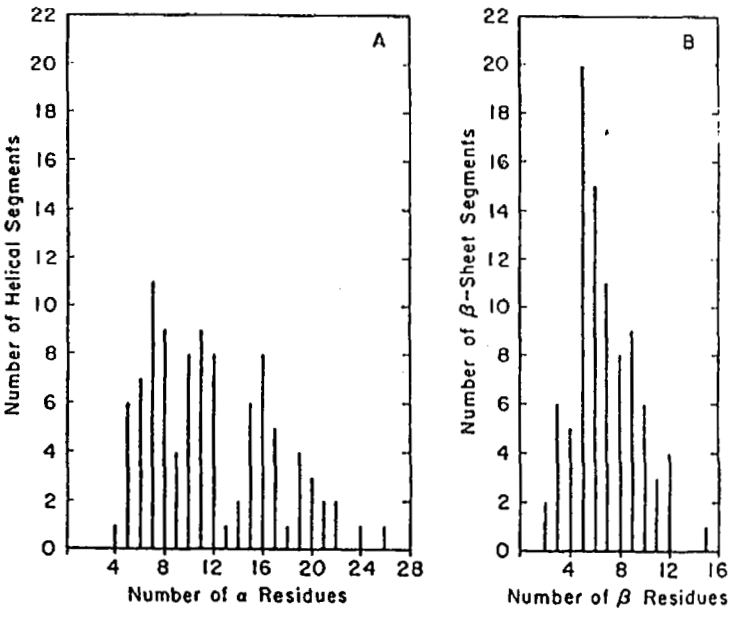


Figure 3. Distribution of the number of residues per helical segment (A) and per β-sheet segment (B) for 7 proteins determined from x-ray crystallography (Chou and Fasman, 1974b).

2. β-Sheet Propagation

The sheet is extended towards the carboxyl terminus until the tetrapeptide average drop below 100. The next three residues (the four of the tetrapeptide) are included in the β-shee sequence. At this point, the residues immediately following are examined. If they are of clas H or h, they are generally included; if they are of class i with high P_β values, they ar included. The $\langle P_\beta \rangle$ for the entire region is then calculated and recalculated with the exclusio of the terminal class i residues (should any be present). If the calculated value is above th threshold value of 103, the region may be assigned as β sheet.

It will often be the case that the tetrapeptide average will drop below 100 for a residu while it is above 100 on either side. These residues are generally included in the structur region.

C. Search for β Turns

1. Definition of β Turns

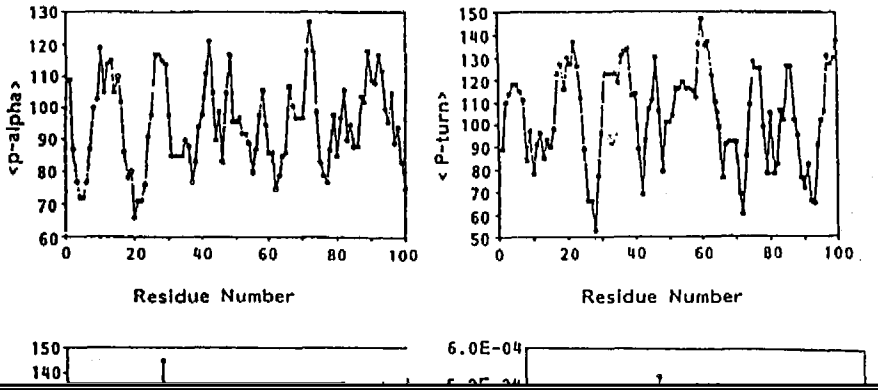
Beta turns are defined by having $\langle P_t \rangle > 100$, $\langle P_\alpha \rangle < \langle P_t \rangle > \langle P_\beta \rangle$, and $p_t > 0.75 \times 10^-$ In general, turns are assigned regardless of disruption of helical or sheet regions.

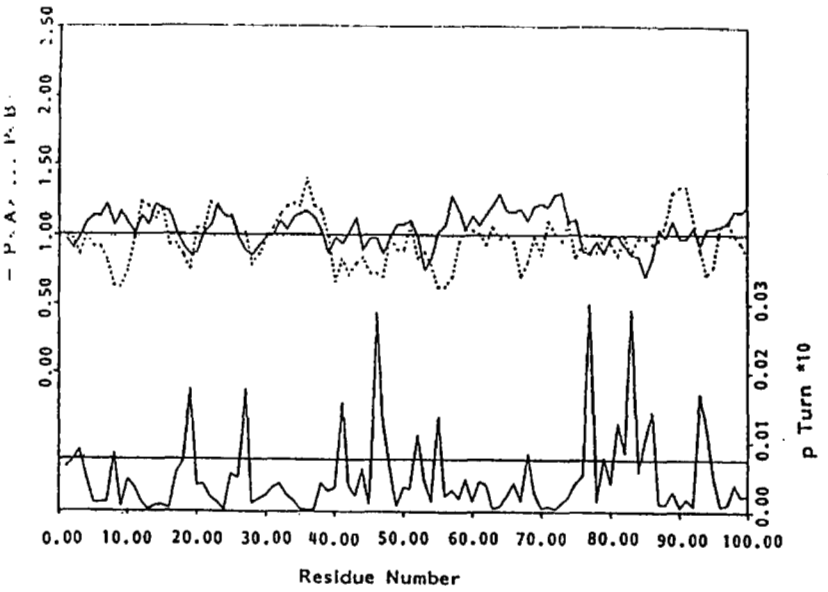
2. Resolving Overlapping Turns

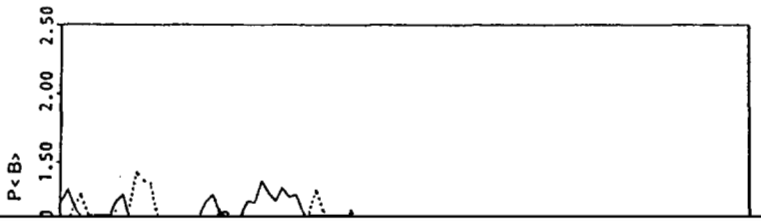
Where a series of turns is found to overlap, the assignment is made to the turn with the higher local $\langle p_t \rangle$ value. One should then go back to make sure that no potential turns have been discounted. If a turn was discounted because of overlap with another turn, which finally was not assigned, the first turn should be reconsidered.

D. Resolving Overlapping Regions

Resolving the overlapping regions is the most difficult aspect of the analysis. It is at this point that using a spreadsheet program such as Lotus-123 greatly reduces the tedium of the process. In general the overlapping regions are compared for the calculated average value. If the $\langle P_\alpha \rangle$ is greater than the $\langle P_\beta \rangle$, the region is assigned as α helical; if the situation is reversed, the region is assigned as β sheet.







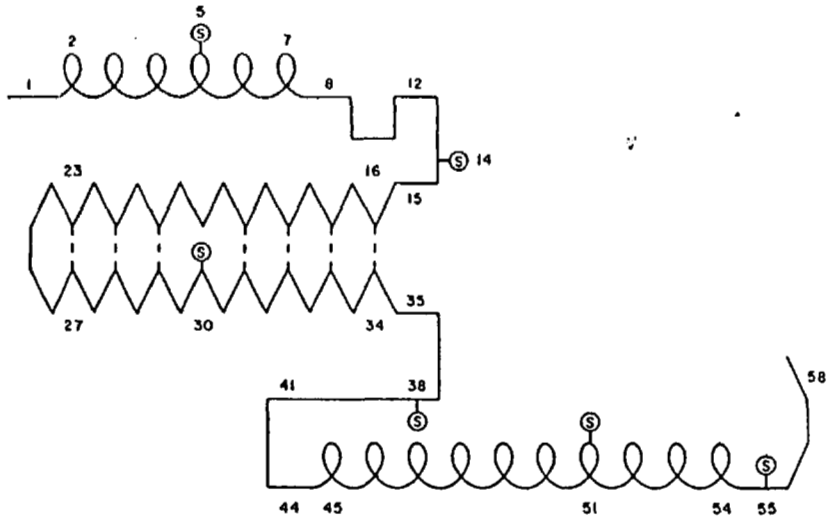


Figure 6. Schematic diagram of helical, β-sheet, and reverse β-turn regions predicted in pancreatic trypsin inhibitor. Residues are represented in their respective conformational states: helical, β-sheet, and coil. Chain reversals denote β-turn tetrapeptides. Hydrogen bonding between antiparallel β sheets is represented by dashed lines. Conformational boundary residues are numbered, as well as the six Cys residues indicated by S. It should be noted that in the present scale each helical loop represents a single helical residue and not a single turn consisting of 3.6 residues (Chou and Fasman, 1974b).

IX. ACCURACY OF PREDICTION

Table III. The Conformational Prediction[a] of Staphylococcal Nuclease Compared to X-Ray Results[b]

		Pa	Pb	Pt	a	b	⟨Pa⟩	⟨Pb⟩	⟨Pt⟩	⟨pt⟩
1	A	142	83	66	H	i	96	99	100	6.40e-005

54	Y	69	147	114	b	H	83	78	124	1.52e-005	
55	G	57	75	156	B	b	101 *	62	112	1.37e-004	*
56	P	57	55	152	B	B	106 *	62	108	2.27e-005	
57	E	151	37	74	H	B	128 *	69	87	3.09e-005	
58	A	142	83	66	H	i	118 *	94	83	1.90e-005	
59	S	77	75	143	i	b	103 *	103 *	91	4.68e-005	
60	A	142	83	66	H	i	113 *	103 *	80	1.52e-005	
61	F	113	138	60	h	h	107 *	101 *	89	4.36e-005	
62	T	83	119	96	i	h	115 *	93	89	3.92e-005	
63	K	116	74	101	h	b	120 *	105 *	78	4.69e-006	
64	K	116	74	101	h	b	129 *	96	71	8.08e-006	
65	M	145	105	60	H	h	117 *	100 *	85	2.29e-005	
111	V	106	170	50	h	H	105 *	142 *	70	2.85e-005	
112	A	142	83	66	H	i	96	136 *	86	1.36e-005	
113	Y	69	147	114	b	H	90	134 *	94	4.26e-005	
114	V	106	170	50	h	H	87	111 *	104	1.97e-005	
115	Y	69	147	114	b	H	77	91	130	2.92e-005	
116	K	116	74	101	h	b	76	76	141	2.88e-004	*
117	P	57	55	152	B	B	68	88	140	1.28e-004	*
118	N	67	89	156	b	i	79	96	125	4.69e-005	
119	N	67	89	156	b	i	100 *	83	105	1.03e-004	*
120	T	83	119	96	i	h	111 *	88	90	3.05e-005	
121	H	100	87	95	I	i	120 *	91	81	2.18e-005	
122	E	151	37	74	H	B	126 *	101 *	72	1.38e-005	

97–104 α helix. Since $\langle P_t \rangle > \langle P_\alpha \rangle$ at residue 105, the turn assignment was made for residues 105–108.
105–108 β turn.
109–115 β sheet. Residues 109–115 were assigned as β sheet because the most stable helix of six residues that could be made (residues 109–114) has $\langle P_\alpha \rangle = 1.13$, significantly less than the value obtained for $\langle P_\beta \rangle = 1.21$.
116–119 β turn.
120–143 α helix: $\langle P_\alpha \rangle = 1.13$.
136–140] Residues 136–140 have $\langle P_\beta \rangle = 1.08$ versus $\langle P_\alpha \rangle = 1.01$ and could be assigned as a β sheet, but as it is a short region within a much longer α region, it is not assigned.
141–144 β turn. This turn precludes the turn possible at residues 141–144. The turn at 144–147 has a higher $\langle p_t \rangle$.
146–149 β turn.

XI. PREDICTION OF SUBTILISIN

This is an analysis of the prediction for subtilisin. The output file is reproduced in Fig. 8.

Residue — Structural assignment

1–3 β sheet.
4–7 β turn.

		Pa	Pb	Pt	a	b	⟨Pa⟩	⟨Pb⟩	⟨Pt⟩	⟨pt⟩	
1	A	142	83	66	H	i	109 *	109 *	89	3.90e-005	
2	Q	111	110	98	h	h	87	102 *	110	1.96e-005	
3	S	77	75	143	i	b	77	111 *	114	2.45e-005	
4	V	106	170	50	h	H	72	111 *	118	3.23e-004	*
5	P	57	55	152	B	B	72	111 *	118	6.68e-005	
6	Y	69	147	114	b	H	77	116 *	115	2.07e-005	
7	G	57	75	156	B	b	87	107 *	111	6.00e-005	
8	V	106	170	50	h	H	100 *	128 *	84	1.79e-005	
9	S	77	75	143	i	b	103 *	104 *	97	1.45e-005	
10	Q	111	110	98	h	h	119 *	106 *	78	1.05e-005	
11	I	108	160	47	h	H	105 *	93	91	1.18e-005	
12	K	116	74	101	h	b	114 *	73	96	8.24e-006	
13	A	142	83	66	H	i	115 *	87	85	4.42e-005	
14	P	57	55	152	B	B	105 *	88	93	1.51e-005	
15	A	142	83	66	H	i	110 *	93	90	1.48e-005	
16	L	121	130	59	H	H	102 *	100 *	98	3.51e-005	
17	H	100	87	95	I	i	86	86	123	1.09e-004	*
18	S	77	75	143	i	b	78	101 *	127	2.79e-004	*
19	Q	111	110	98	h	h	80	112 *	116	5.66e-005	
20	G	57	75	156	B	b	66	104 *	130	6.55e-005	
21	Y	69	147	114	b	H	71	104 *	127	1.78e-004	*
22	T	83	119	96	i	h	71	89	137	8.32e-005	*
23	G	57	75	156	B	b	76	102 *	126	1.44e-004	*
24	S	77	75	143	i	b	91	102 *	112	2.65e-005	

55	T	83	119	96	i	h	80	100 *	116	3.21e-004 *
56	P	57	55	152	B	B	87	98	116	5.39e-005
57	N	67	89	156	b	i	98	97	115	1.98e-005
58	F	113	138	60	h	h	106 *	89	112	8.38e-005 *
59	Q	111	110	98	h	h	95	76	136	1.33e-004 *

112	E	151	37	74	H	B	127 *	104 *	70	1.43e-006
113	W	108	137	96	h	h	125 *	115 *	68	4.41e-006
114	A	142	83	66	H	i	114 *	103 *	83	6.50e-006
115	I	108	160	47	h	H	96	105 *	106	5.68e-005

169	G	57	75	156	B	b	74		87		130	9.09e-005 *
170	K	116	74	101	h	b	79		87		127	1.29e-005
171	Y	69	147	114	b	H	77		111	*	114	1.64e-004 *
172	P	57	55	152	B	B	87		115	*	98	2.22e-005
173	S	77	75	143	i	b	108	*	122	*	76	4.34e-006
174	V	106	170	50	h	H	115	*	145	*	53	3.91e-006
175	I	108	160	47	h	H	103	*	122	*	79	1.39e-005
176	A	142	83	66	H	i	111	*	102	*	84	3.17e-005
177	V	106	170	50	h	H	102	*	124	*	80	9.78e-006
178	G	57	75	156	B	b	90		100	*	107	3.30e-005
179	A	142	83	66	H	i	93		104	*	107	4.98e-005
180	V	106	170	50	h	H	86		102	*	115	9.56e-005 *
181	G	57	75	156	B	b	77		96		131	7.62e-005 *
182	N	67	89	156	b	i	77		96		131	3.21e-004 *
183	K	116	74	101	h	b	96		94		109	3.94e-005
184	Y	69	147	114	b	H	84		113	*	112	3.05e-005
185	G	57	75	156	B	b	83		98		123	8.04e-005 *
186	A	142	83	66	H	i	83		98		123	1.13e-004 *
187	Y	69	147	114	b	H	69		107	*	130	1.02e-004 *
188	N	67	89	156	b	i	71		89		137	9.43e-005 *
189	G	57	75	156	B	b	90		93		113	7.57e-005 *
190	T	83	119	96	i	h	111	*	95		91	9.71e-006
191	S	77	75	143	i	b	110	*	84		103	3.65e-005
192	M	145	105	60	H	h	105	*	79		105	4.39e-005
193	A	142	83	66	H	i	94		75		114	1.53e-005
194	S	77	75	143	i	b	85		96		110	1.78e-004 *
195	P	57	55	152	B	B	101	*	98		90	7.79e-006
196	H	100	87	95	I	i	101	*	103	*	91	3.58e-005
197	V	106	170	50	h	H	111	*	102	*	84	5.19e-005
198	A	142	83	66	H	i	120	*	81		88	1.04e-005
199	G	57	75	156	B	b	120	*	81		88	1.57e-005
200	A	142	83	66	H	i	136	*	94		64	1.12e-005
201	A	142	83	66	H	i	128	*	114	*	59	9.19e-006
202	A	142	83	66	H	i	123	*	125	*	57	1.36e-006
203	L	121	130	59	H	H	106	*	123	*	77	7.91e-006
204	I	108	160	47	h	H	105	*	109	*	87	1.28e-005
205	L	121	130	59	H	H	103	*	91		99	3.30e-005
206	S	77	75	143	i	b	87		72		122	8.73e-005 *
207	K	116	74	101	h	b	85		76		126	8.00e-006
208	H	100	87	95	I	i	83		92		124	1.34e-003 *
209	P	57	55	152	B	B	78		100	*	125	4.28e-005
210	N	67	89	156	b	i	81		108	*	126	1.24e-005
211	W	108	137	96	h	h	85		116	*	111	1.25e-004 *
212	T	83	119	96	i	h	86		109	*	111	4.55e-005
213	N	67	89	156	b	i	91		122	*	100	3.41e-005
214	T	83	119	96	i	h	99		123	*	84	2.01e-005
215	Q	111	110	98	h	h	98		112	*	96	3.73e-005
216	V	106	170	50	h	H	89		103	*	107	8.71e-005 *
217	R	98	93	95	i	i	93		93		110	8.51e-005 *
218	S	77	75	143	i	b	96		97		110	5.88e-005
219	S	77	75	143	i	b	94		101	*	114	1.01e-005
220	L	121	130	59	H	H	95		112	*	102	9.02e-005 *
221	Q	111	110	98	h	h	86		109	*	111	3.15e-005
222	N	67	89	156	b	i	79		111	*	111	8.93e-005 *
223	T	83	119	96	i	h	91		107	*	97	5.74e-005
224	T	83	119	96	i	h	100	*	110	*	88	4.68e-005
225	T	83	119	96	i	h	94		99		103	5.41e-005

Figure 8 (cont.)

226	K	116	74	101	h	b	98		83		115	2.12e-005
227	L	121	130	59	H	H	89		83		126	9.84e-005 *
228	G	57	75	156	B	b	87		85		126	9.12e-005 *
229	D	101	54	146	I	B	90		103	*	115	1.66e-004 *
230	S	77	75	143	i	b	82		126	*	107	7.01e-005
231	F	113	138	60	h	h	77		126	*	111	6.65e-005
232	Y	69	147	114	b	H	77		110	*	121	9.62e-005 *
233	Y	69	147	114	b	H	74		92		131	7.63e-005 *
234	G	57	75	156	B	b	87		88		118	1.56e-004 *
235	K	116	74	101	h	b	100	*	109	*	90	9.42e-006
236	G	57	75	156	B	b	88		113	*	104	3.02e-006
237	L	121	130	59	H	H	100	*	137	*	78	2.10e-005
238	I	108	160	47	h	H	98		132	*	87	9.79e-006
239	N	67	89	156	b	i	106	*	113	*	92	1.66e-005
240	V	106	170	50	h	H	125	*	111	*	70	1.23e-005
241	Q	111	110	98	h	h	134	*	89		74	1.14e-005
242	A	142	83	66	H	i	134	*	89		74	1.56e-005
243	A	142	83	66	H	i	0		0		0	0.00e+000
244	A	142	83	66	H	i	0		0		0	0.00e+000
245	Q	111	110	98	h	h	0		0		0	0.00e+000
246	^@	0	0	0	^@	^@	0		0		0	0.00e+000

Figure 8 (cont.)

128–131 β turn.
132–142 α helix.
143–146 β turn.
147–153 β sheet, $\langle P_\beta \rangle = 1.33$, $\langle P_\alpha \rangle = 1.21$.
154–157 β turn.
158–161 β turn.
162–166 β sheet.
167–170 β turn.
171–181 β sheet.
182–185 β turn.
186–189 β turn.
190–193 α helix.
194–197 β turn.
198–207 α helix.
208–211 β turn.
212–227 β sheet.
228–231 β turn.
234–237 β turn.
238–245 α helix, $\langle P_\alpha \rangle = 1.19$, $\langle P_\beta \rangle = 1.06$.

XII. APPENDIXES

Appendix 1: C Language Source Code for Program PREDICT

```
#include <stdio.h>
#include <ctype.h>
#define MAXLENGTH 1000
struct protein_data {
    int code;
    int p_alpha;
    int p_beta;
    int p_turn;
    float bend[4];
    int alpha_class;
    int beta_class;
    int p4_alpha;
    int p4_beta;
    int p4_turn;
    float turn_prod;
} sequence[MAXLENGTH];


#include <protein.dat>

char infile[128], outfile[128], prot_name[64];
main()
{
    int c,length=1;
    char ans[6];
    FILE *fopen(), *fpi, *fpo;
    int fclose(FILE *fpo);
        printf("\n\n     Chou-Fasman-Prevelige Algorithm\n\n\n\n\n");
        printf("\n\n   (C) copyright 1988, Peter Prevelige, all rights reserved\n\n\n");
    do {
        printf("Enter name of input file:  ");
        scanf("%s",infile);
        if ((fpi = fopen(infile, "r")) == NULL)
          printf("No such file exists.\n");
        }while (fpi == NULL);

    do {
        printf("Enter name of output file:  ");
        scanf("%s",outfile);
        if ((fpo = fopen(outfile, "w")) == NULL)
          printf("Cannot open the file for output.\nWill output to the screen.\n");
        }while (fpi == NULL);

        printf("Enter name of protein:   ");
        scanf("%s",prot_name);


    while ((c=getc(fpi)) != EOF)
        if (!isspace(c)) sequence[length++].code = toupper(c);


    get_probability(length);

    tetra_ave(length) ;

    print_it(length,fpo);
```



```
    fclose(fpo);
}


char d_base[5];

get_probability(length)
    int length;
{
    int i,j,k,dbase;

    do{
        printf("Use database of 29 proteins (Chou & Fasman '78)---> enter 29\n");
        printf("Use database of 64 proteins    (Chou '79)---> enter 64\n");

        dbase = 2;

        scanf("%s",d_base);

        if(d_base[0]=='2')dbase=0;

        if(d_base[0]=='6')dbase=1;

    } while (dbase > 1);


    for (i=1,j=0;i<length;i++){
        while((sequence[i].code != data[j].c) && j < 20) j++;
        if (j == 20) (printf("Illegal data point #%d is%d\n",i,sequence[i].code);exit(1););
        sequence[i].p_alpha = data[j].p_a[dbase];
        sequence[i].p_beta  = data[j].p_b[dbase];
        sequence[i].p_turn  = data[j].p_t;
        for(k=0;k<=3;k++) sequence[i].bend[k]  = data[j].b[k];
        sequence[i].alpha_class = data[j].a_class[dbase];
        sequence[i].beta_class  = data[j].b_class[dbase];
        j=0;
        }

}


#define TURN_PRODUCT 0.75E-4


tetra_ave(length)         /* calculates tetrapeptide averages of protein */
int length;
{
    int i=1, j=0;
    int asum, bsum, tsum;
    float tprod;

    for (i=1; i < length - 3; i++){
        asum=bsum=tsum=0;
        tprod=1;
        for (j=0; j<=3; j++){
          asum += sequence[i+j].p_alpha;
          bsum += sequence[i+j].p_beta;
          tsum += sequence[i+j].p_turn;
```

```
        tprod*= sequence[i+j].bend[j];
    }
    sequence[i].p4_alpha = asum/4;
    sequence[i].p4_beta  = bsum/4;
    sequence[i].p4_turn  = tsum/4;
    sequence[i].turn_prod= tprod;
  }
}
#define ALPHA_CUT 100
#define BETA_CUT 100
#define TURN_CUT 0.75e-4

print_it(length,fpo)

FILE *fpo;
int length;


    int i;
    char format, *forma, *forml, *form;

    forma="%4d %c  %3d %3d %3d  %c  %c   %3d %c  %3d %c  %3d  %.2c %c\n";
    forml="%4d \"%c\" %3d %3d %3d  \"%c\"  \"%c\"   %3d \"%c\"  %3d \"%c\"  %3d
%.2c \"%c\"\n";

    format=getchar();
    do (printf("ASCII format enter A, or LOTUS format enter L:\n " );
    format=toupper(getchar());)

    while ((format != 'A') && (format != 'L'));
        fprintf(fpo," Chou-Fasman-Algorithm\n\n");
        fprintf(fpo,"Input file: %s\n",infile);
        fprintf(fpo,"Protein name: %s\n",prot_name);
        fprintf(fpo,"Database used was %s proteins\n\n",d_base);
        fprintf(fpo,"      Pa Pb Pt  a  b   <Pa>  <Pb>  <Pt>  <pt>\n");
        fprintf(fpo,"      ------------------------------------------\n");

    if (format == 'L') form=forml;
    else form=forma;

    for(i=1;i<=length;i++)
        fprintf(fpo,form,i,
            sequence[i].code,
            sequence[i].p_alpha,
            sequence[i].p_beta,
            sequence[i].p_turn,
            sequence[i].alpha_class,
            sequence[i].beta_class,
            sequence[i].p4_alpha, (sequence[i].p4_alpha >= ALPHA_CUT) ? '*' : ' ',
            sequence[i].p4_beta, (sequence[i].p4_beta >= BETA_CUT) ? '*' : ' ',
            sequence[i].p4_turn,
            sequence[i].turn_prod, (sequence[i].turn_prod >= TURN_CUT) ? '*' : ' ')
            ;
```


Appendix 2: Include File "Protein.Dat"

```
struct p_data {
      int c;
      int p_a[2];
      int p_b[2];
      int p_t;
      float b[4];
      int a_class[2];
      int b_class[2];
      int p4_a;
      int p4_b;
      int p4_t;
};
struct p_data data[] ={
      'A',142,139, 83, 79, 66,0.060,0.076,0.035,0.058,'H','H','I','i',0,0
      'R', 98,100, 93, 94, 95,0.070,0.106,0.099,0.085,'i','h','I','i',0,0
      'N', 67, 78, 89, 66,156,0.161,0.083,0.191,0.091,'b','i','i','b',0,0
      'D',101,106, 54, 66,146,0.147,0.110,0.179,0.081,'I','h','B','b',0,0
      'C', 70, 95,119,107,119,0.149,0.053,0.117,0.128,'i','i','h','h',0,0
      'Q',111,112,110,100, 98,0.074,0.098,0.037,0.098,'h','h','h','I',0,0
      'E',151,144, 37, 51, 74,0.056,0.060,0.077,0.064,'H','H','B','b',0,0
      'G', 57, 64, 75, 87,156,0.102,0.085,0.190,0.152,'B','B','b','i',0,0
      'H',100,112, 87, 83, 95,0.140,0.047,0.093,0.054,'I','h','i','i',0,0
      'I',108, 99,160,157, 47,0.043,0.034,0.013,0.056,'h','i','H','H',0,0
      'L',121,130,130,117, 59,0.061,0.025,0.036,0.070,'H','H','H','h',0,0
      'K',116,121, 74, 73,101,0.055,0.115,0.072,0.095,'h','h','b','b',0,0,
      'M',145,132,105,101, 60,0.068,0.082,0.014,0.055,'H','H','h','I',0,0,
      'F',113,111,138,123, 60,0.059,0.041,0.065,0.065,'h','h','h','h',0,0,
      'P', 57, 55, 55, 62,152,0.102,0.301,0.034,0.068,'B','B','B','B',0,0,
      'S', 77, 72, 75, 94,143,0.120,0.139,0.125,0.106,'i','b','b','i',0,0,
      'T', 83, 78,119,133, 96,0.086,0.108,0.065,0.079,'i','i','h','h',0,0,
      'W',108,103,137,124, 96,0.077,0.013,0.064,0.167,'h','I','h','h',0,0,
      'Y', 69, 73,147,131,114,0.082,0.065,0.114,0.125,'b','b','H','h',0,0,
      'V',106, 97,170,164, 50,0.062,0.048,0.028,0.053,'h','i','H','H',0,0,
};
```

XIII. REFERENCES

Chou, P. Y., and Fasman, G. D., 1974a, Conformational parameters for amino acids in helical, β-sheet a random coil, regions calculated from proteins, *Biochemistry* 13:211–222.

Chou, P. Y., and Fasman, G. D., 1974b, Prediction of protein conformation, *Biochemistry* 13:222–245.

Chou, P. Y., and Fasman, G. D., 1977, β-Turns in proteins, *J. Mol. Biol.* 115:135–175.

Chou, P. Y., and Fasman, G. D., 1978, Prediction of the secondary structure of proteins from their amino ac sequence, *Adv. Enzymol.* 47:45–148.

Chou, P. Y., and Fasman, G. D., 1979, Prediction of β-turns, *Biophys. J.* 26:367–384.

Fasman, G. D., 1985, A critique of the utility of the prediction of protein secondary structure, *J. Biosci.* 8:15 23.

Huber, R., Kulka, D., Rühlmann, A., and Steigmann, W., 1971, Pancreatic trypsin inhibitor (Kunitz). Part Structure and function, *Cold Spring Harbor Symp. Quant. Biol.* 36:141–150.