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.

To obtain print-out enter:

PRINT SUBTIL.PRN (Filename OUTPUT.PRN)
```

Figure 1. Sample of screen during program execution. User input is shown in all capital letters and bold print.

per line or as many as desired—however, each letter must be separated by a "white space." White spaces are spaces, tabs, or carriage returns. The program will accept drive and path information up to a total of 128 characters. The input file can be written using any word processor program that has a nondocument or ASCII mode.

To run the program, type PREDICT. The program will query the user for the name of the input file followed by the name of the output file (Fig. 1).

The next question asked by the program concerns the data base to use in predicting the secondary structure. There are two choices: the Chou–Fasman original data set based on 29 proteins (Chou and Fasman, 1978) or the expanded data set based on 64 proteins (P. Chou, Chapter 12, this volume).

Finally, the program will query the user for the type of output file. There are two formats, one that can be read into Lotus-123 and a plain ASCII-type file. The only difference between the two is that the Lotus version has the characters surrounded by quotation marks. This enables it to be read into Lotus using the File Import function (see Section IV).

The program will then proceed to perform the prediction and write the output file. The output file requires that the data be reduced by hand. The program does not perform this function. This is a deliberate decision. The Chou–Fasman algorithm is really a set of guidelines, and in reducing the data to a single predicted structure significant information is lost. Frequently a region will display significant propensity for more than one type of structure and be "hard to call." These regions are actually very interesting and may be sites of conformational change (see Fasman, Chapter 6, this volume). It is hoped that by being forced to be in closer contact with the raw data, the user will become aware of these potentialities.

		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	
25	N	67	89	156	b	i	98		125	*	89	2.95e-005	
26	V	106	170	50	h	H	117	*	124	*	66	1.16e-005	
27	K	116	74	101	h	b	117	*	124	*	66	4.90e-006	
28	V	106	170	50	h	H	115	*	145	*	53	7.39e-006	
29	A	142	83	66	H	i	114	*	116	*	77	3.03e-006	
30	V	106	170	50	h	H	98		114	*	96	4.00e-005	
31	I	108	160	47	h	H	85		91		123	8.99e-005	*
32	D	101	54	146	I	B	85		91		123	2.17e-004	*
33	S	77	75	143	i	b	85		91		123	1.07e-005	
34	G	57	75	156	B	b	85		91		123	6.58e-005	
35	I	108	160	47	h	H	90		91		119	6.27e-005	
36	D	101	54	146	I	B	88		72		131	1.38e-004	*
37	S	77	75	143	i	b	77		73		133	1.05e-004	*
38	S	77	75	143	i	b	83		67		134	1.55e-005	
39	H	100	87	95	I	i	94		81		113	5.28e-004	*
40	P	57	55	152	B	B	98		78		114	3.84e-005	
41	D	101	54	146	I	B	111	*	107	*	89	1.40e-005	
42	L	121	130	59	H	H	121	*	114	*	69	1.14e-005	
43	K	116	74	101	h	b	105	*	100	*	93	1.40e-005	

Figure 2. Sample of ASCII-mode output file for the Chou–Fasman algorithm. Input file, a:SUBTIL; protein name, SUBTILISIN BPN'; data base used was 29 proteins.

IV. DATA REDUCTION: THE OUTPUT FILE

A sample output file is presented in Fig. 2. The first column is the residue number, and the second column is the single-letter code for the amino acid at that position.

The next five columns are data base data; they are, in order, the P_α value, the P_β value, the P_t value, the helical assignment, and the β-sheet probability. It should be noted that since the helical and β-sheet assignments were made only for the original Chou–Fasman data set, these are the ones that are always listed. Likewise, all turn values are from the original Chou–

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.

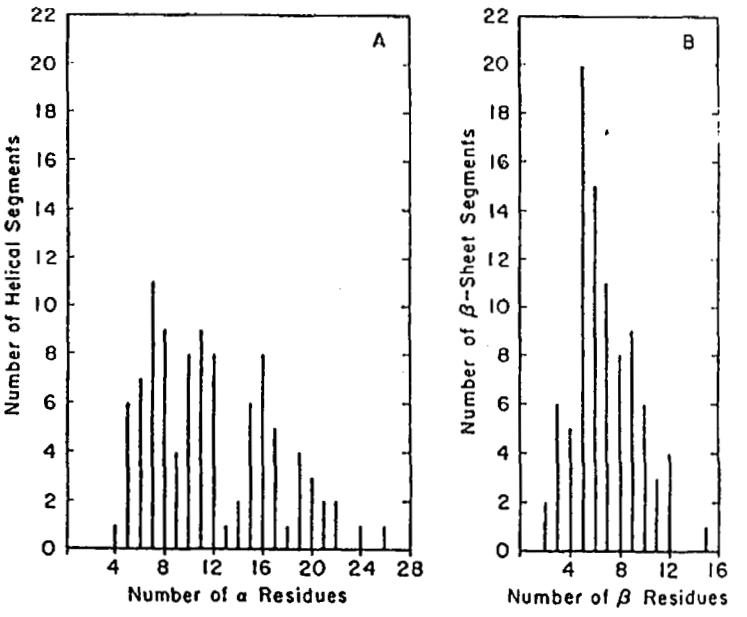

Figure 3. Distribution of the number of residues per helical segment (A) and per β-sheet segment (B) for 7 proteins determined from x-ray crystallography (Chou and Fasman, 1974b).

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.

When one region of structure is wholely contained within another and the lengths are similar, the procedure is relatively straightforward. If, for example, there is a region of β sheet contained within a longer region of potential helix, the average value for the contained β-sheet region is calculated and compared to the average helical value for the same residues. If the average β-sheet value is greater, then the region is assigned as β sheet; otherwise it is α helix.

If one region of structure is significantly longer than another, it is handled as described above. Should the contained region have a larger average value, the assignment is made for the contained region. The user should then go back and reevaluate the remaining region to see if it will be stable on its own.

A problem arises when the contained region has a larger average value and the leftover residues are not long enough to form a stable structure. Should a small region of stable structure disrupt a longer region? Although generally, a small locally stable region is not allowed to disrupt a longer region of secondary structure, if the difference in average value is large it may do so.

VI. GRAPHIC DISPLAY

It is frequently useful to look at the tetrapeptide averages for α helix, β sheet, and β turns in a graphic format. If the Lotus output file has been generated and the file imported, this is a simple procedure using the graph option (see Fig. 4).

A program to display simultaneously the three tetrapeptide averages for the α helix, β sheet, and β turns, called PREVIEW, has been written by R. Lee for use on an IBM-PC. Such a plot is displayed in Fig. 5 for staphylococcal nuclease, which is discussed later in this chapter. As an aid in portraying the prediction of secondary structure, a diagramatic plot can be made of the secondary structure, as was demonstrated in the original paper on the prediction of secondary structure (Chou and Fasman, 1974b). A diagramatic representation of the bovine pancreatic trypsin inhibitor is shown in Fig. 6 (Chou and Fasman, 1974b).

VII. PORTABILITY

This program is written in the C programming language. The source code is reproduced in Appendix 1. The data set is contained in an Include file called Protein.Dat. The Include file

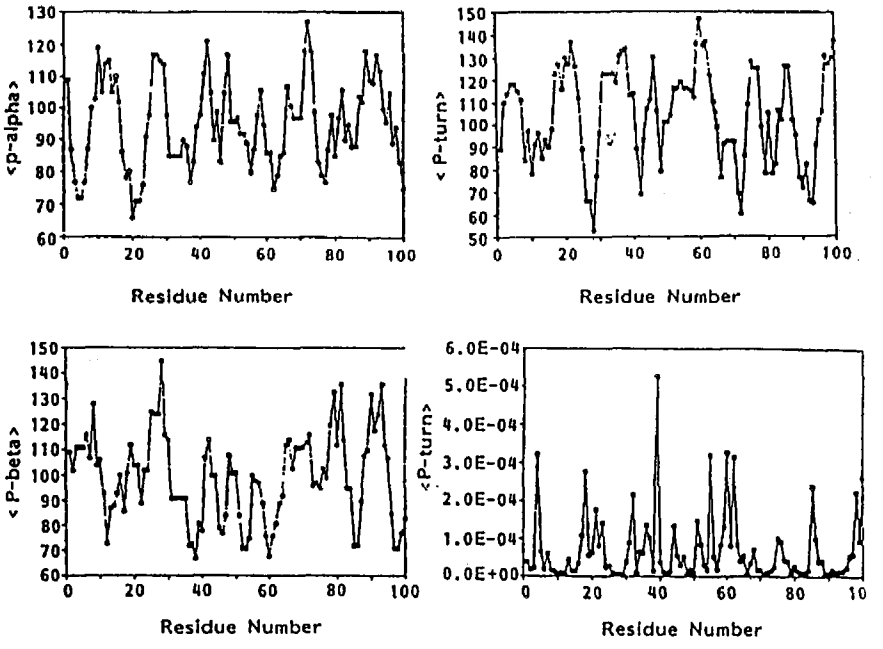

Figure 4. Graphic display of tetrapeptide averages produced using Lotus.

program must be available for compilation. The program uses only standard C functions and thus should be readily portable from machine to machine.

VIII. LOTUS HINTS

The Lotus-compatible version of the output file can be read into Lotus using the File Import function. The file should be imported as "numbers." Once imported it is useful to reformat the file so that it is displayed conveniently on the screen. In order to do this, select the Worksheet function, followed by the Global function. Then use the Column-Width feature and set the column width to 4. This takes care of all but the position-dependent turn data, which will be displayed as all zeros. This can be set as follows: select Range, then Format, and finally select Scientific with two decimal places. The range will now appear as all asterisks. Adjust the column width by selecting Worksheet and Column-Width: set the column width to 9.

In order to obtain the average of the P_α or P_β values for a range, use the @AVG function. A convenient place to locate the calculation is in the adjacent column where the asterisks are located. It is also possible to obtain a running average by fixing the starting point, using the average function, and copying the function through the range.

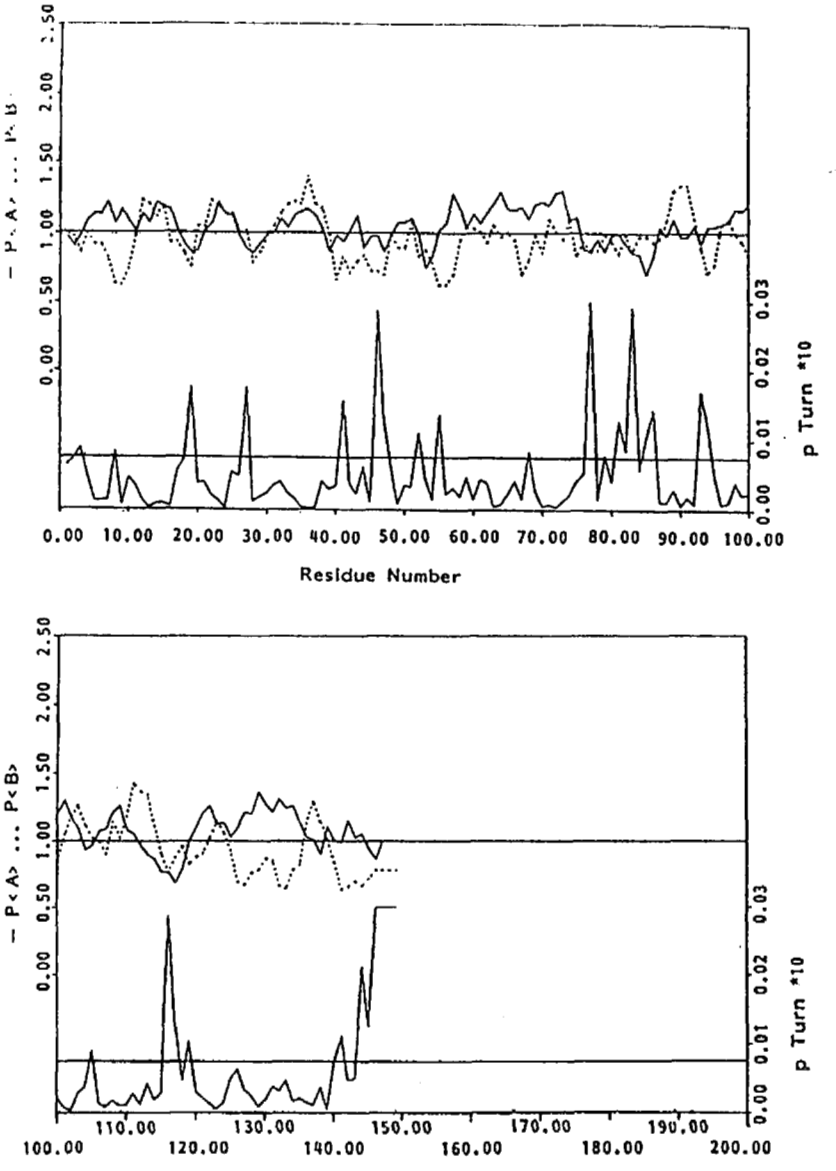

Figure 5. Predicted conformational profile of staphylococcal nuclease. The average helical and β-sheet potential of tetrapeptides i to $i + 3$ based on single-residue information. Data base is 64 proteins (see Chou, Chapter 12, this volume). $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ are shown, respectively, by the solid line and dotted lines. The β-turn probability profile is shown as a solid line at the bottom of the profile. The horizontal line corresponds to the cutoff value of 0.75×10^{-3}.

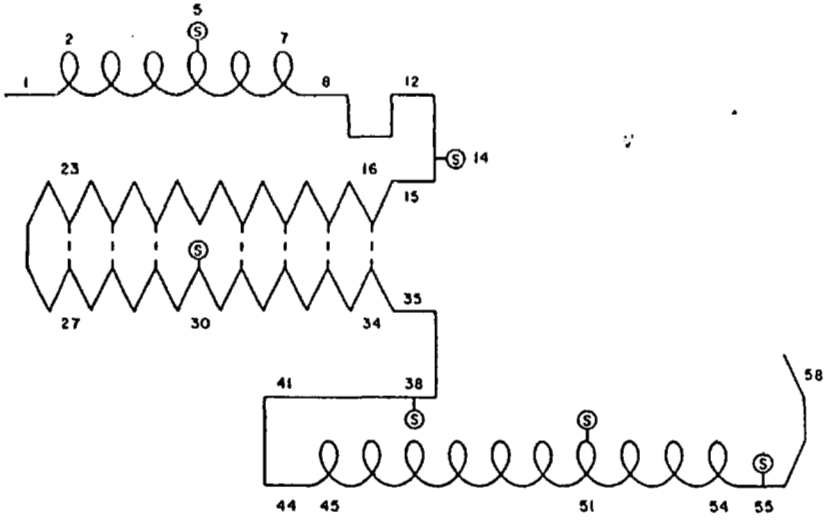

Figure 6. Schematic diagram of helical, β-sheet, and reverse β-turn regions predicted in pancreatic trypsin inhibitor. Residues are represented in their respective conformational states: helical, β-sheet, and coil. Chain reversals denote β-turn tetrapeptides. Hydrogen bonding between antiparallel β sheets 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

The prediction of the secondary structure of staphylococcal nuclease is compared to the x-ray diffraction results of Huber *et al.* (1971) in Table III. All regions of α and β structure were correctly predicted: four α-helical regions and two β-sheet regions. Within each region the numbers of overpredicted residues and missed residues are also listed in Table III. A total of 44 residues were either overpredicted or missed. Therefore, a total of 71% (subtracting both missed and overpredicted residues) of the residues were correctly predicted. The point to be emphasized is that all secondary structural regions were correctly predicted, and thus a fairly accurate picture of the secondary structure was obtained. The predicted β turns are also listed in Table III but cannot be compared, as they are not identified as such in the original paper (Huber *et al.*, 1971).

These results are an improvement over the original prediction of staphylococcal nuclease (Chou and Fasman, 1974b), and the improvement can be attributed to the larger data base (64 versus 29 proteins) used to obtain the conformational parameters P_α and P_β.

X. PREDICTION OF STAPHYLOCOCCAL NUCLEASE

This is an analysis of the prediction for staphylococcal nuclease. The output file is reproduced in Fig. 7.

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

Predicted	X ray	Residues overpredicted	Residues missed
7–18 α	12–19 α	5	1
23–26 β	21–27 β	0	3
31–40 β	30–36 β	4	1
59–76 α	54–67 α	10	8
97–104 α	99–106 α	2	2
120–140 α	122–134 α	8	0
Total		29	15
β turns[c]			
3–6			
19–22			
27–30			
41–44			
46–49			
55–58			
77–80			
83–86			
93–96			
105–108			
116–119			
141–144			
146–149			

[a] Based on the predictive analysis of Fig. 7.
[b] Huber *et al.* (1971).
[c] Percentage correctly predicted = (149 − 44)/149 × 100 = 71%.

Residue	Structural assignment
3–6	β turn: $\langle P_t \rangle > \langle P_\alpha \rangle$ and $\langle P_t \rangle > \langle P_\beta \rangle$; $\langle p_t \rangle >$ threshold.
7–18	α helix: $\langle P_\alpha \rangle = 1.13$; starts at residue 7; $\langle P_\alpha \rangle > \langle P_t \rangle$ for the region of 12–15; $\langle P_\alpha \rangle = 1.14$ and $\langle P_\beta \rangle = 1.18$; however, this is short for a β-sheet region, so it is not assigned as β. The proline is at position 5 of the helix. Therefore, it might cause a bend in the helix.
19–22	β turn.
23–26	β sheet: $\langle P_\beta \rangle = 1.20$ and $\langle P_\alpha \rangle = 1.19$.
27–30	β turn.
31–40	β sheet: $\langle P_\beta \rangle = 1.12$ and $\langle P_\alpha \rangle = 1.06$.
41–44	β turn.
46–49	β turn. The possible turn at residues 47–50, $\langle p_t \rangle = 1.28 \times 10^{-4}$, is precluded by the stronger turn at 46–49, $\langle p_t \rangle = 2.88 \times 10^{-4}$.
55–58	β turn. The presence of this turn precludes the one that is possible at residues 52–55.
59–76	α helix. The possible turn at residues 68–71 does not have $\langle P_t \rangle > \langle P_\alpha \rangle$. There is no region of β-sheet potential that is greater than α helix throughout this region. Notice that the region has been extended to include the helix-forming residue at 76, $\langle P_\alpha \rangle = 1.17$.

		Pa	Pb	Pt	a	b	⟨Pa⟩		⟨Pb⟩		⟨Pt⟩	⟨pt⟩	
1	A	142	83	66	H	i	96		99		100	6.40e-005	
2	T	83	119	96	i	h	89		96		109	7.38e-005	
3	S	77	75	143	i	b	98		85		110	8.86e-005	*
4	T	83	119	96	i	h	109	*	99		89	4.98e-005	
5	K	116	74	101	h	b	113	*	91		89	1.23e-005	
6	K	116	74	101	h	b	113	*	91		89	1.21e-005	
7	L	121	130	59	H	H	122	*	82		82	1.32e-005	
8	H	100	87	95	I	i	106	*	63		105	8.43e-005	*
9	K	116	74	101	h	b	116	*	62		98	6.51e-006	
10	E	151	37	74	H	B	108	*	73		97	4.66e-005	
11	P	57	55	152	B	B	100	*	96		93	3.53e-005	
12	A	142	83	66	H	i	113	*	123	*	67	1.31e-005	
13	T	83	119	96	i	h	107	*	120	*	75	2.66e-006	
14	L	121	130	59	H	H	121	*	111	*	68	8.66e-006	
15	I	108	160	47	h	H	118	*	119	*	65	9.69e-006	
16	K	116	74	101	h	b	116	*	92		90	4.40e-006	
17	A	142	83	66	H	i	102	*	93		103	5.55e-005	
18	I	108	160	47	h	H	91		85		123	7.28e-005	
19	D	101	54	146	I	B	85		75		136	1.77e-004	*
20	G	57	75	156	B	b	86		104	*	112	3.87e-005	
21	D	101	54	146	I	B	101	*	104	*	98	4.22e-005	
22	T	83	119	96	i	h	106	*	123	*	76	2.08e-005	
23	V	106	170	50	h	H	122	*	119	*	67	1.41e-005	
24	K	116	74	101	h	b	112	*	114	*	83	2.41e-006	
25	L	121	130	59	H	H	112	*	114	*	83	5.42e-005	
26	M	145	105	60	H	h	96		100	*	107	4.84e-005	
27	Y	69	147	114	b	H	88		101	*	117	1.76e-004	*
28	K	116	74	101	h	b	85		78		126	1.18e-005	
29	G	57	75	156	B	b	92		86		116	1.87e-005	
30	Q	111	110	98	h	h	99		97		101	2.46e-005	
31	P	57	55	152	B	B	99		104	*	92	3.53e-005	
32	M	145	105	60	H	h	109	*	113	*	77	4.06e-005	
33	T	83	119	96	i	h	103	*	120	*	77	2.44e-005	
34	F	113	138	60	h	h	113	*	122	*	68	1.58e-005	
35	R	98	93	95	i	i	115	*	120	*	68	4.41e-006	
36	L	121	130	59	H	H	117	*	140	*	56	2.91e-006	
37	L	121	130	59	H	H	112	*	121	*	78	3.46e-006	
38	L	121	130	59	H	H	102	*	118	*	87	4.14e-005	
39	V	106	170	50	h	H	86		99		111	3.01e-005	
40	D	101	54	146	I	B	98		66		117	3.45e-005	
41	T	83	119	96	i	h	93		82		104	1.57e-004	*
42	P	57	55	152	B	B	101	*	71		105	3.78e-005	
43	E	151	37	74	H	B	112	*	79		91	2.35e-005	
44	T	83	119	96	i	h	89		83		111	6.25e-005	
45	K	116	74	101	h	b	97		72		112	8.35e-006	
46	H	100	87	95	I	i	97		72		112	2.88e-004	*
47	P	57	55	152	B	B	86		69		127	1.28e-004	*
48	K	116	74	101	h	b	98		98		102	6.37e-005	
49	K	116	74	101	h	b	107	*	89		95	8.38e-006	
50	G	57	75	156	B	b	107	*	89		95	3.58e-005	
51	V	106	170	50	h	H	110	*	107	*	84	3.35e-005	
52	E	151	37	74	H	B	98		83		111	1.12e-004	*
53	K	116	74	101	h	b	74		87		130	4.62e-005	

Figure 7. Output file for staphylococcal nuclease. Note that a represents α, and b β. Chou–Fasman algorithm: input file, a:NUCLEASE; protein name, STAPH; data base used was 29 proteins.

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
66	V	106	170	50	h	H	116	*	94		86	4.12e-005
67	E	151	37	74	H	B	119	*	70		99	1.55e-005
68	N	67	89	156	b	i	110	*	80		106	8.37e-005 *
69	A	142	83	66	H	i	120	*	97		78	2.78e-005
70	K	116	74	101	h	b	122	*	86		80	5.26e-006
71	K	116	74	101	h	b	120	*	110	*	68	7.63e-006
72	I	108	160	47	h	H	129	*	101	*	61	4.62e-006
73	E	151	37	74	H	B	130	*	95		64	1.35e-005
74	V	106	170	50	h	H	109	*	108	*	85	2.20e-005
75	E	151	37	74	H	B	111	*	84		97	4.17e-005
76	F	113	138	60	h	h	88		94		118	5.36e-005
77	N	67	89	156	b	i	87		87		127	3.45e-004 *
78	K	116	74	101	h	b	95		88		112	1.47e-005
79	G	57	75	156	B	b	87		99		111	7.82e-005 *
80	Q	111	110	98	h	h	98		94		108	4.13e-005
81	R	98	93	95	i	i	99		85		109	1.29e-004 *
82	T	83	119	96	i	h	92		98		114	8.51e-005 *
83	D	101	54	146	I	B	85		87		129	2.93e-004 *
84	K	116	74	101	h	b	85		97		116	5.77e-005
85	Y	69	147	114	b	H	70		97		130	1.05e-004 *
86	G	57	75	156	B	b	83		93		116	1.44e-004 *
87	R	98	93	95	i	i	104	*	95		94	1.24e-005
88	G	57	75	156	B	b	97		108	*	98	1.12e-005
89	L	121	130	59	H	H	110	*	130	*	71	2.96e-005
90	A	142	83	66	H	i	97		134	*	85	6.34e-006
91	Y	69	147	114	b	H	97		134	*	85	1.84e-005
92	I	108	160	47	h	H	105	*	111	*	93	7.92e-006
93	Y	69	147	114	b	H	92		89		120	1.70e-004 *
94	A	142	83	66	H	i	104	*	71		117	1.19e-004 *
95	D	101	54	146	I	B	104	*	77		115	4.95e-005
96	G	57	75	156	B	b	106	*	106	*	91	8.70e-006
97	K	116	74	101	h	b	108	*	109	*	91	1.15e-005
98	M	145	105	60	H	h	117	*	100	*	85	3.99e-005
99	V	106	170	50	h	H	116	*	94		86	2.30e-005
100	N	67	89	156	b	i	120	*	84		88	2.37e-005
101	E	151	37	74	H	B	130	*	105	*	62	8.12e-006
102	A	142	83	66	H	i	116	*	119	*	67	3.57e-006
103	L	121	130	59	H	H	109	*	125	*	75	2.84e-005
104	V	106	170	50	h	H	93		112	*	99	3.70e-005
105	R	98	93	95	i	i	96		102	*	102	9.12e-005 *
106	Q	111	110	98	h	h	107	*	99		94	1.31e-005
107	G	57	75	156	B	b	109	*	90		95	8.48e-006
108	L	121	130	59	H	H	121	*	114	*	69	1.77e-005
109	A	142	83	66	H	i	126	*	102	*	70	1.12e-005
110	K	116	74	101	h	b	108	*	118	*	82	1.15e-005

Figure 7 (cont.)

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
123	Q	111	110	98	h	h	112	*	115	*	77	5.66e-006
124	L	121	130	59	H	H	114	*	106	*	78	1.43e-005
125	L	121	130	59	H	H	103	*	93		99	4.93e-005
126	R	98	93	95	i	i	110	*	69		103	6.44e-005
127	K	116	74	101	h	b	121	*	67		96	3.41e-005
128	S	77	75	143	i	b	120	*	76		95	2.47e-005
129	E	151	37	74	H	B	136	*	78		76	9.13e-006
130	A	142	83	66	H	i	127	*	87		82	1.96e-005
131	Q	111	110	98	h	h	121	*	85		91	3.85e-005
132	A	142	83	66	H	i	131	*	67		85	3.18e-005
133	K	116	74	101	h	b	124	*	64		94	4.63e-005
134	K	116	74	101	h	b	126	*	78		83	1.66e-005
135	E	151	37	74	H	B	113	*	82		97	2.11e-005
136	K	116	74	101	h	b	103	*	113	*	90	1.47e-005
137	L	121	130	59	H	H	101	*	129	*	89	1.10e-005
138	N	67	89	156	b	i	90		115	*	110	3.71e-005
139	I	108	160	47	h	H	111	*	102	*	90	4.47e-006
140	W	108	137	96	h	h	100	*	84		117	7.50e-005
141	S	77	75	143	i	b	99		63		129	1.11e-004 *
142	E	151	37	74	H	B	115	*	65		110	4.83e-005
143	N	67	89	156	b	i	102	*	70		128	5.02e-005
144	D	101	54	146	I	B	105	*	66		125	2.12e-004 *
145	A	142	83	66	H	i	94		71		127	1.25e-004 *
146	D	101	54	146	I	B	86		78		135	3.80e-004 *
147	S	77	75	143	i	b	0		0		0	0.00e+000
148	G	57	75	156	B	b	0		0		0	0.00e+000
149	Q	111	110	98	h	h	0		0		0	0.00e+000
150	^@	0	. 0	0	^@	^@	0		0		0	0.00e+000

Figure 7 (cont.)

77–80 β turn. This has been assigned as a turn because it has the largest local value of $\langle p_t \rangle$.

83–86 β turn. This assignment has been made for the same reason as above, the strongest local value of $\langle p_t \rangle$.

88–92 β sheet: $\langle P_\beta \rangle = 1.19$, $\langle P_\alpha \rangle = 0.97$.

93–96 β turn.

[97–115] The region from residue 97 to residue 115 has several possible predictions. This may be taken to mean that there is the possibility of conformational changes taking place (see Fasman, Chapter 6, this volume). A β-turn assignment could possibly be made at residues 105–108. The $\langle P_t \rangle$ is equal to the $\langle P_\beta \rangle$ there. If this turn is not predicted, then either α helix or β sheet could run through to about residue 112 or 115, respectively. For residues 97–104, $\langle P_\alpha \rangle = 1.19$ and $\langle P_\beta \rangle = 1.07$, so the assignment is made to α helix.

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.
8–10	β sheet: $\langle P_\beta \rangle = 1.18$, $\langle P_\alpha \rangle = 0.98$.
11–17	α helix. Note inclusion of indifferent residue, $\langle P_\alpha \rangle = 1.12$.
18–21	β turn.
23–26	β turn.
[11–26]*	α helix. Note these two β turns could be included in the α-helical region.
27–31	β sheet. Note inclusion of residue 31 because it is of class h, and the value $\langle P_\beta \rangle$ is 1.31.
32–35	β turn.
36–39	β turn.
44–47	β turn.
51–54	β turn.
55–58	β turn.
60–63	β turn. Highest local value of $\langle p_t \rangle$.
[32–63]	α helix + 3_{10} helix.
65–74	β sheet; $\langle P_\beta \rangle = 1.06$. $\langle P_\alpha \rangle = 1.02$ for residues 66–74. This region is difficult to assign and is a candidate for a region displaying an inducible change in secondary structure.
75–78	β turn.
79–84	β sheet, $\langle P_\beta \rangle = 1.30$, $\langle P_\alpha \rangle = 0.93$.
85–88	β turn.
89–96	β sheet, $\langle P_\beta \rangle = 1.23$. There is helical potential from residues 89–96; $\langle P_\alpha \rangle = 1.15$.
100–103	β turn.
104–122	β sheet, $\langle P_\beta \rangle = 1.14$, $\langle P_\beta \rangle = 1.03$. There may be a β turn located at residues 118–121.
123–126	β turn.

*The β-turn, ϕ,ψ angles are the same as those found in the 3_{10} helix. Therefore, several consecutive β turns can be added onto an α-helical region. Frequently terminal residues of an α helix have the 3_{10} geometry.

		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	
25	N	67	89	156	b	i	98	125 *	89	2.95e-005	
26	V	106	170	50	h	H	117 *	124 *	66	1.16e-005	
27	K	116	74	101	h	b	117 *	124 *	66	4.90e-006	
28	V	106	170	50	h	H	115 *	145 *	53	7.39e-006	
29	A	142	83	66	H	i	114 *	116 *	77	3.03e-006	
30	V	106	170	50	h	H	98	114 *	96	4.00e-005	
31	I	108	160	47	h	H	85	91	123	8.99e-005	*
32	D	101	54	146	I	B	85	91	123	2.17e-004	*
33	S	77	75	143	i	b	85	91	123	1.07e-005	
34	G	57	75	156	B	b	85	91	123	6.58e-005	
35	I	108	160	47	h	H	90	91	119	6.27e-005	
36	D	101	54	146	I	B	88	72	131	1.38e-004	*
37	S	77	75	143	i	b	77	73	133	1.05e-004	*
38	S	77	75	143	i	b	83	67	134	1.55e-005	
39	H	100	87	95	I	i	94	81	113	5.28e-004	*
40	P	57	55	152	B	B	98	78	114	3.84e-005	
41	D	101	54	146	I	B	111 *	107 *	89	1.40e-005	
42	L	121	130	59	H	H	121 *	114 *	69	1.14e-005	
43	K	116	74	101	h	b	105 *	100 *	93	1.40e-005	
44	V	106	170	50	h	H	90	100 *	107	1.36e-004	*
45	A	142	83	66	H	i	99	79	111	5.62e-005	
46	G	57	75	156	B	b	83	77	130	3.22e-005	
47	G	57	75	156	B	b	105 *	84	106	5.33e-005	
48	A	142	83	66	H	i	117 *	108 *	79	6.19e-006	
49	S	77	75	143	i	b	96	101 *	101	1.87e-005	
50	M	145	105	60	H	h	96	101 *	101	1.18e-005	
51	V	106	170	50	h	H	97	84	104	1.49e-004	*
52	P	57	55	152	B	B	92	71	116	8.62e-005	*
53	S	77	75	143	i	b	92	71	116	3.18e-005	
54	E	151	37	74	H	B	89	75	119	1.87e-005	

Figure 8. Output file for subtilisin. Note that a represents α, and b β. Chou–Fasman algorithm; input file a:SUBTIL; protein name, SUBTILISIN; data base used was 29 proteins.

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 *
60	D	101	54	146	I	B	86		68		147	3.27e-004 *
61	D	101	54	146	I	B	86		76		135	8.24e-005 *
62	N	67	89	156	b	i	75		81		137	3.16e-004 *
63	S	77	75	143	i	b	79		89		122	8.47e-005 *
64	H	100	87	95	I	i	85		92		110	4.18e-005
65	G	57	75	156	B	b	86		112	*	99	5.43e-005
66	T	83	119	96	i	h	107	*	114	*	76	6.56e-006
67	H	100	87	95	I	i	101	*	103	*	91	3.58e-005
68	V	106	170	50	h	H	97		111	*	92	7.07e-005
69	A	142	83	66	H	i	97		111	*	92	1.76e-005
70	G	57	75	156	B	b	97		111	*	92	1.79e-005
71	T	83	119	96	i	h	118	*	113	*	69	8.38e-006
72	V	106	170	50	h	H	127	*	116	*	60	1.15e-005
73	A	142	83	66	H	i	118	*	96		86	1.49e-005
74	A	142	83	66	H	i	99		97		109	2.61e-005
75	L	121	130	59	H	H	83		95		128	1.03e-004 *
76	N	67	89	156	b	i	79		103	*	125	9.35e-005 *
77	N	67	89	156	b	i	77		99		125	4.42e-005
78	S	77	75	143	i	b	87		120	*	99	4.11e-005
79	I	108	160	47	h	H	98		133	*	78	7.16e-006
80	G	57	75	156	B	b	85		112	*	105	2.68e-005
81	V	106	170	50	h	H	97		136	*	78	1.56e-005
82	L	121	130	59	H	H	106	*	114	*	82	8.42e-006
83	G	57	75	156	B	b	90		95		106	1.17e-005
84	V	106	170	50	h	H	95		95		102	1.70e-005
85	A	142	83	66	H	i	88		72		126	2.39e-004 *
86	P	57	55	152	B	B	88		72		126	1.03e-004 *
87	S	77	75	143	i	b	104	*	90		102	4.09e-005
88	S	77	75	143	i	b	102	*	108	*	95	4.10e-005
89	A	142	83	66	H	i	118	*	110	*	76	9.92e-006
90	L	121	130	59	H	H	109	*	132	*	72	7.36e-006
91	Y	69	147	114	b	H	108	*	118	*	82	1.66e-005
92	A	142	83	66	H	i	117	*	124	*	66	1.10e-005
93	V	106	170	50	h	H	112	*	136	*	65	1.40e-005
94	K	116	74	101	h	b	100	*	112	*	91	1.44e-005
95	V	106	170	50	h	H	96		107	*	102	2.39e-005
96	L	121	130	59	H	H	105	*	85		106	5.38e-005
97	G	57	75	156	B	b	89		71		131	5.97e-005
98	D	101	54	146	I	B	94		71		127	2.25e-004 *
99	A	142	83	66	H	i	83		77		130	9.69e-005 *
100	G	57	75	156	B	b	75		83		138	2.64e-004 *
101	S	77	75	143	i	b	78		101	*	127	4.72e-005
102	G	57	75	156	B	b	78		101	*	127	1.21e-004 *
103	Q	111	110	98	h	h	91		117	*	112	1.00e-004 *
104	Y	69	147	114	b	H	90		129	*	100	4.09e-005
105	S	77	75	143	i	b	100	*	133	*	83	1.14e-006
106	W	108	137	96	h	h	97		136	*	86	3.10e-006
107	I	108	160	47	h	H	85		121	*	101	4.24e-005
108	I	108	160	47	h	H	85		121	*	101	3.80e-005
109	N	67	89	156	b	i	95		90		108	1.14e-005
110	G	57	75	156	B	b	106	*	102	*	93	4.46e-005
111	I	108	160	47	h	H	127	*	104	*	70	9.58e-006

Figure 8 (cont.)

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
116	A	142	83	66	H	i	105	*	91		109	5.23e-005
117	N	67	89	156	b	i	95		84		129	1.52e-005
118	N	67	89	156	b	i	104	*	104	*	103	1.25e-004 *
119	M	145	105	60	H	h	115	*	122	*	75	1.17e-005
120	D	101	54	146	I	B	95		118	*	99	8.35e-006
121	V	106	170	50	h	H	106	*	131	*	78	2.21e-005
122	I	108	160	47	h	H	99		107	*	101	5.30e-006
123	N	67	89	156	b	i	102	*	99		104	1.16e-004 *
124	M	145	105	60	H	h	100	*	96		104	5.17e-005
125	S	77	75	143	i	b	78		88		128	8.66e-005 *
126	L	121	130	59	H	H	73		83		130	6.70e-005
127	G	57	75	156	B	b	62		70		151	3.12e-005
128	G	57	75	156	B	b	62		70		151	5.83e-004 *
129	P	57	55	152	B	B	67		70		148	2.86e-004 *
130	S	77	75	143	i	b	88		77		127	7.39e-005
131	G	57	75	156	B	b	104	*	79		107	2.88e-005
132	S	77	75	143	i	b	120	*	92		83	2.23e-005
133	A	142	83	66	H	i	130	*	92		73	1.56e-005
134	A	142	83	66	H	i	130	*	92		73	6.26e-006
135	L	121	130	59	H	H	130	*	92		73	1.42e-005
136	K	116	74	101	h	b	126	*	102	*	70	7.75e-006
137	A	142	83	66	H	i	122	*	97		82	1.03e-005
138	A	142	83	66	H	i	116	*	95		90	4.90e-005
139	V	106	170	50	h	H	116	*	95		90	2.85e-005
140	D	101	54	146	I	B	116	*	95		90	3.14e-005
141	K	116	74	101	h	b	126	*	102	*	70	6.79e-006
142	A	142	83	66	H	i	116	*	102	*	81	1.07e-005
143	V	106	170	50	h	H	95		100	*	103	8.95e-005 *
144	A	142	83	66	H	i	95		100	*	103	8.40e-005 *
145	S	77	75	143	i	b	86		122	*	99	1.51e-005
146	G	57	75	156	B	b	93		146	*	76	7.27e-006
147	V	106	170	50	h	H	106	*	170	*	50	4.42e-006
148	V	106	170	50	h	H	115	*	148	*	54	4.83e-006
149	V	106	170	50	h	H	124	*	126	*	58	6.04e-006
150	V	106	170	50	h	H	133	*	104	*	62	9.57e-006
151	A	142	83	66	H	i	120	*	81		88	2.43e-005
152	A	142	83	66	H	i	102	*	82		111	7.88e-005 *
153	A	142	83	66	H	i	104	*	71		113	6.23e-005
154	G	57	75	156	B	b	83		69		135	9.91e-005 *
155	N	67	89	156	b	i	88		69		132	1.95e-004 *
156	E	151	37	74	H	B	92		76		117	4.70e-005
157	G	57	75	156	B	b	68		86		137	1.40e-004 *
158	S	77	75	143	i	b	73		86		134	2.61e-004 *
159	T	83	119	96	i	h	73		86		134	9.69e-005 *
160	G	57	75	156	B	b	72		75		146	1.88e-004 *
161	S	77	75	143	i	b	78		86		131	1.65e-004 *
162	S	77	75	143	i	b	85		109	*	108	5.75e-005
163	S	77	75	143	i	b	80		109	*	111	5.52e-005
164	T	83	119	96	i	h	78		127	*	104	9.80e-005 *
165	V	106	170	50	h	H	72		111	*	118	4.09e-005
166	G	57	75	156	B	b	60		88		144	3.43e-005
167	Y	69	147	114	b	H	74		87		130	4.46e-004 *
168	P	57	55	152	B	B	74		87		130	7.80e-005 *

Figure 8 (cont.)

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.