PREDICTION OF β-TURNS

PETER Y. CHOU AND GERALD D. FASMAN, *Graduate Department of Biochemistry, Brandeis University, Waltham, Massachusetts 02154 U.S.A.*

ABSTRACT An automated computer prediction of the chain reversal regions of globular proteins is described herein using the bend frequencies and β-turn conformational parameters (P_t) determined from 408 β-turns in 29 proteins calculated from x-ray atomic coordinates. The probability of bend occurrence at residue i is $p_t = f_i \times f_{i+1} \times f_{i+2} \times f_{i+3}$ with the average bend probability $<p_t> = 0.55 \times 10^{-4}$. Tetrapeptides with $p_t > 0.75 \times 10^{-4}$ ($\approx 1.5 \times <p_t>$) as well as $<P_t> > 1.00$ and $<P_\alpha> < <P_t> > <P_\beta>$ are selected by the computer as probable bends. Adjacent probable bends (i.e., 11–14, 12–15, 13–16) are compared pairwise by the computer, and the tetrapeptide with the higher p_t value is predicted as a β-turn. The percentage of bend and nonbend residues predicted correctly for 29 proteins by this computer algorithm is $\%_{t+nt} = 70\%$, whereas 78% of the β-turns were localized correctly within ± 2 residues. The average β-turn content in the 29 proteins is 32%, with helical proteins having fewer bends (17%) than β-sheet proteins (41%). Three proteins having iron-sulfur clusters were found with the highest percentages of β-turns: *Chromatium* high potential iron protein (65%), ferredoxin (57%), and rubredoxin (65%). Finally, the bend frequencies at all 12 positions from 457 β-turns in 29 proteins (Chou and Fasman, 1977) were used to test the effectiveness of predicting bends using 2, 4, 8, and 12 residues as well as different cut-off p_t values. The computer analysis showed that $1.25 <p_t>$ to be the best cut-off yielding 70% accuracy in $\%_{t+nt}$ for 4 residues and $\%_{t+nt} = 73\%$ for 12 residues in predicting the bend and nonbend regions of proteins.

INTRODUCTION

The β-turn is a structural feature of protein conformation involving four consecutive residues where the polypeptide chain folds back on itself by nearly 180°. These turns have also been called β-bends (Lewis et al., 1971, 1973), hairpin loops (Kuntz, 1972), reverse turns (Crawford et al., 1973), and 3_{10} bends (Dickerson et al., 1971). Although helical and β-sheet regions in proteins have been clearly elucidated by x-ray crystallographers, the β-turns are usually not specified as they were generally assumed to be part of the random coil or irregular regions. Although a bend frequency table based on 12 proteins was compiled by Chou and Fasman (1974) the data set relied in part on stereodiagrams inasmuch as atomic coordinates were not available. Because of the tentative nature of the data set based on relatively small statistical sampling, as well as the lack of x-ray information on the location of the chain reversal regions in most proteins, no attempt was made to predict the β-turns as was done for the helices and β-sheets in all the 19 proteins surveyed at that time. Nevertheless, guidelines for β-turn prediction were outlined using the probability of bend occurrence (p_t) in conjunction with the conformational parameters P_α, P_β, and P_t. When applied to pancreatic trypsin inhibitor, the predicted α, β, and turn regions showed close agreement to the secondary

---

Dr. Chou's present address is Department of Chemistry, Worcester Polytechnic Institute, Worcester, Mass. 01609. Dr. Fasman is the Rosenfield Professor of Biochemistry.

structure elucidated from x-ray analysis (Chou and Fasman, 1974). This method also yielded 77% accuracy in predicting the bend and nonbend regions of adenylate kinase before the x-ray structure was known (Schulz et al., 1974). Using the bend frequencies of 17 enzymes, the chain reversal regions in *lac* repressor were predicted (Chou et al., 1975) showing that 20% of the protein consisted of β-turns as compared to 37% helices and 35% β-sheets. A more refined analysis of 29 proteins revealed 408 β-turns, and these bend frequencies (Chou and Fasman, 1978) were used to predict the conformational folding in the histones (Fasman et al., 1976). A detailed documentation of 459 β-turns elucidated from the x-ray atomic coordinates of 29 proteins, together with the classification of 11 bend types based on the ϕ, ψ dihedral angles, was presented earlier (Chou and Fasman, 1977). In this paper, an automated computer method of predicting β-turns is described and applied in localizing the bend regions of 29 proteins. A predictive accuracy analysis was also made for 29 proteins based on the bend frequencies of 2, 4, 8, and 12 residues in the chain reversal region as well as using various cut-off values. It is hoped that these studies will lead to further optimization of β-turn predictions in globular proteins.

METHODS

β-Turns Elucidated from X-ray Analysis

From the x-ray atomic coordinates of 29 proteins, the C_i^α–C_{i+3}^α distances of all 4,650 tetrapeptides were computed (Chou and Fasman, 1977). Those whose distances were below 7 Å and not in a helical region were considered as β-turns. A complete listing of the 459 β-turns found from x-ray analysis, along with their tetrapeptide sequences, C_i^α–C_4^α and O_1–N_4 distances,[1] the dihedral angles of the two central bend residues, and the classification of the β-turns according to 11 bend types, was given in a previous paper (Chou and Fasman, 1977).

β-Turn Frequencies Calculated from 29 Proteins

The frequency of occurrence of the 20 amino acids in the 4 β-turn positions i, $i + 1$, $i + 2$, and $i + 3$ is given in Table I arranged in hierarchical order. For example, the number of Ala residues in the 1st, 2nd, 3rd, and 4th position of β-turns is 26, 33, 15, and 25, respectively. Dividing these by the total number of Ala residues ($n_{\text{Ala}} = 434$) gives the values $f_1 = 0.060, f_2 = 0.076, f_3 = 0.035$, and $f_4 = 0.058$ as shown in Table I. The total number of Ala residues in β-turns is $n_t = 85$, and is smaller than $n_{t_1} + n_{t_2} + n_{t_3} + n_{t_4} = 99$, because overlapping Ala residues in β-turns were not counted twice. Hence, if Ala appears as residue 12 in β-turn 9-12 and 11-14, it is counted only once in n_t. Similarly, 45 Ala residues were found in the 2nd and 3rd bend positions. The frequency of Ala residues in β-turns and middle β-turns are $f_t = 85/434 = 0.196$ and $f_{t_m} = 45/434 = 0.104$, respectively. The average frequency of all residues in β-turns and middle β-turns are $<f_t> = 1{,}400/4{,}740 = 0.295$ and $<f_{t_m}> = 764/4{,}740 = 0.161$. The β-turn and middle β-turn conformational potentials of Ala are obtained by normalization: $P_t = f_t/<f_t> = 0.196/0.295 = 0.66$ and $P_{t_m} = f_{t_m}/<f_{t_m}> = 0.104/0.161 = 0.64$. The average frequency of bend occurrence is $<f_j> = \Sigma_j/N = 408/4{,}740 = 0.086$, where $j = i, i + 1, i + 2$, or $i + 3$.

Automated Computer Prediction of β-Turns

The relative probability of β-turn occurrence at residue position i is computed from

$$p_t = f_i \times f_{i+1} \times f_{i+2} \times f_{i+3}. \quad (1)$$

[1]The notations 1, 2, 3, 4 and i, $i + 1$, $i + 2$, $i + 3$ will be used interchangeably to indicate 4 positions of the β-turn.

TABLE I
BEND FREQUENCY HIERARCHIES AND β-TURN CONFORMATIONAL POTENTIALS OF 20 AMINO ACIDS

f_i		f_{i+1}		f_{i+2}		f_{i+3}		P_t		P_{t2}	
Asn	0.161	Pro	0.301	Asn	0.191	Trp	0.167	Asn	1.56	Pro	2.04
Cys	0.149	Ser	0.139	Gly	0.190	Gly	0.152	Gly	1.56	Gly	1.63
Asp	0.147	Lys	0.115	Asp	0.179	Cys	0.128	Pro	1.52	Asp	1.61
His	0.140	Asp	0.110	Ser	0.125	Tyr	0.125	Asp	1.46	Asn	1.56
Ser	0.120	Thr	0.108	Cys	0.117	Ser	0.106	Ser	1.43	Ser	1.52
Pro	0.102	Arg	0.106	Tyr	0.114	Gln	0.098	Cys	1.19	Lys	1.13
Gly	0.102	Gln	0.098	Arg	0.099	Lys	0.095	Tyr	1.14	Tyr	1.08
Thr	0.086	Gly	0.085	His	0.093	Asn	0.091	Lys	1.01	Arg	1.05
Tyr	0.082	Asn	0.083	Glu	0.077	Arg	0.085	Gln	0.98	Thr	0.98
Trp	0.077	Met	0.082	Lys	0.072	Asp	0.081	Thr	0.96	Cys	0.92
Gln	0.074	Ala	0.076	Thr	0.065	Thr	0.079	Trp	0.96	Gln	0.84
Arg	0.070	Tyr	0.065	Phe	0.065	Leu	0.070	Arg	0.95	Glu	0.80
Met	0.068	Glu	0.060	Trp	0.064	Pro	0.068	His	0.95	His	0.77
Val	0.062	Cys	0.053	Gln	0.037	Phe	0.065	Glu	0.74	Ala	0.64
Leu	0.061	Val	0.048	Leu	0.036	Glu	0.064	Ala	0.66	Phe	0.62
Ala	0.060	His	0.047	Ala	0.035	Ala	0.058	Met	0.60	Met	0.51
Phe	0.059	Phe	0.041	Pro	0.034	Ile	0.056	Phe	0.60	Trp	0.48
Glu	0.056	Ile	0.034	Val	0.028	Met	0.055	Leu	0.59	Val	0.43
Lys	0.055	Leu	0.025	Met	0.014	His	0.054	Val	0.50	Leu	0.36
Ile	0.043	Trp	0.013	Ile	0.013	Val	0.053	Ile	0.47	Ile	0.29

f_i, f_{i+1}, f_{i+2}, and f_{i+3} are the frequencies of the 1st, 2nd, 3rd, and 4th residues in a reverse β-turn. P_t is the conformational potential of a residue in a β-turn based on all four positions of a reverse turn. P_{t2} is the conformational potential of a residue in a β-turn based on the 2nd and 3rd positions of a reverse turn. This frequency table was based on 408 β-turns in 29 proteins.

The average probability of any tetrapeptide to be in the β-turn is $<p_t> = <f_j>^4 = (0.086)^4 = 0.55 \times 10^{-4}$. Two cutoff values were selected: $p_t = 1.00 \times 10^{-4}$ (a value approximately double that of the average) and $p_t = 0.75 \times 10^{-4}$ (a value that is 3/2 that of the average). All tetrapeptides with $p_t > 1.00 \times 10^{-4}$ (denoted by the symbol *) and $1.00 \times 10^{-4} < p_t > 0.75 \times 10^{-4}$ (denoted by the symbol ?) along the protein sequence were automatically listed by the computer as potential β-turns without considerations of helical or β-sheet regions. This extended list was then reduced by eliminating tetrapeptides that have either $<P_\alpha> > <P_t>$ or $<P_\beta> > <P_t>$ or $<P_t> < 1.00$. If adjacent tetrapeptides all have $p_t > 0.75 \times 10^{-4}$, they are considered pairwise, and the tetrapeptide with the higher p_t value is predicted as a β-turn. In this manner, redundant β-turns are eliminated by the automated computer prediction algorithm.

Evaluation of Predictive Accuracy

To evaluate the success of any predictive algorithm, it is necessary to compare the predicted conformational states for each residue of a protein with the observed assignments determined from x-ray crystallography. In the present case, the percentage of residues n_t predicted correctly in the β-turn conformational state t is given by:

$$\%_t = \frac{100(n_t - t_m)}{n_t}, \tag{2}$$

where n_t represents the number of β-turn residues in each protein from x-ray analysis and t_m is the number of β-turn residues missed in the prediction. The amount of overprediction is reflected in the

equation:

$$\%_{nt} = \frac{100(n_{nt} - t_o)}{n_{nt}} = \frac{100(N - n_t - t_o)}{N - n_t}, \tag{3}$$

where $\%_{nt}$ is the percentage of non-β-turn residues ($n_{nt} = N - n_t$) predicted correctly and t_o is the number of overpredicted β-turn residues. The quality of β-turn prediction is obtained by averaging $\%_t$ and $\%_{nt}$:

$$Q_t = \frac{(\%_t + \%_{nt})}{2}. \tag{4}$$

The overall accuracy of bend prediction may also be represented by:

$$\%_{t+nt} = \frac{100(N - t_m - t_o)}{N}, \tag{5}$$

where $(t_m + t_o)$ is the total number of incorrectly predicted bend and nonbend residues in the protein, and N is the total number of residues in the protein.

When there is total agreement between prediction and observation, the value of 100% is obtained for $\%_t$, $\%_{nt}$, Q_t, and $\%_{t+nt}$, whereas 0% indicates total disagreement. For a two-state system (i.e., bend or nonbend), a value of 50% for $\%_{t+nt}$ shows that the prediction is no better than random guessing. Although Eqs. 4 and 5 appear similar, it should be noted that Q_t is not a weighted average. That is, predicting proteins containing few β-turns tends to yield high Q_t values. Thus, if all 20 bend residues of a protein containing 100 residues were predicted correctly as bends ($\%_t = 100\%$) while 40 of the 80 nonbend residues were predicted correctly as nonbends ($\%_{nt} = 50\%$), the quality of bend prediction is $Q_t = 75\%$. However, the accuracy of bend prediction according to Eq. 5 yields $\%_{n+nt} = 60\%$ because there are 40 incorrectly predicted residues out of 100. If Q_t is weight-averaged,

$$\overline{Q}_t = \frac{\%_t(f_t) + \%_{nt}(f_{nt})}{N}, \tag{6}$$

the results obtained will be identical to Eq. 5. Hence, in the above example, $\overline{Q}_t = [100(0.2) + 50(0.8)]/100 = 60\%$. Similar warnings concerning the use of Q_α and Q_β when predicting proteins with low helical and β-sheet content have been given earlier (Chou and Fasman, 1974).

Another criterion that indicates how much better a given prediction is than random guessing is the correlation coefficient (Matthews, 1975). Hence, for predicting bend residues,

$$C_t = \frac{(\mathrm{p}/N - \overline{\mathrm{P}}\overline{S})}{\{\overline{\mathrm{P}}\overline{S}(1 - \overline{S})(1 - \overline{\mathrm{P}})\}^{1/2}}, \tag{7}$$

where $\mathrm{p} = (n_t - t_m)$ is the number of correctly predicted turn residues (i.e., the observed number of turn residues minus the turn residues missed in prediction) and N is the total number of residues in the protein. $\overline{\mathrm{P}}$ is the fraction of the protein predicted to be in β-turns,

$$\overline{\mathrm{P}} = (n_t - t_m + t_o)/N = (f_t)_{\text{predicted}}, \tag{8}$$

and $\overline{S}$ is the fraction of the protein observed to be in β-turns,

$$\overline{S} = n_t/N = (f_t)_{\text{x-ray}}. \tag{9}$$

Substituting Eq. 8 and Eq. 9 in Eq. 7,

$$C_t = \frac{[(n_t - t_m)/N] - [(n_t - t_m + t_o)/N](n_t/N)}{\left\{\frac{(n_t - t_m + t_o)}{N}\frac{(n_t}{N}\left(1 - \frac{n_t}{N}\right)\left[1 - \frac{(n_t - t_m + t_o)}{N}\right]\right\}1/2}. \tag{10}$$

A correlation coefficient $C = 1$ indicates perfect agreement between prediction and observation, $C = 0$ indicates that a prediction is no better than random guessing, and $C = -1$ indicates total disagreement or 0% accuracy.

RESULTS AND DISCUSSION

Computer Prediction of β-Turns

Using the bend frequencies based on 408 β-turns in 29 proteins as given in Table I, the computed β-turn profile for staphylococcal nuclease is shown in Fig. 1 with the probable β-turns listed in Table II. There are 21 tetrapeptides (denoted as * in Table II) above the cutoff point, $p_t = 1.00 \times 10^{-4}$ (indicated by solid line in Fig. 1) and 6 tetrapeptides (denoted as ?) above the lower cutoff point, $p_t = 0.75 \times 10^{-4}$ (indicated by dashed line in Fig. 1). Using the lower cutoff value, the 27 probable bends are reduced to 19 predicted bends which are shown in parentheses. It is seen that tetrapeptides 8–11 (with $<P_\alpha> = 1.06 > <P_t> = 1.05$) and 68–71 ($<P_\alpha> = 1.10 > <P_t> = 1.06$) were not predicted as bends by computer analysis because their $<P_t>$ values were lower than either $<P_\alpha>$ or $<P_\beta>$ values. Similarly, tetrapeptides 47–50, 82–85, 85–88, 94–97, 117–120, and 145–148 were not predicted as bends because adjacent tetrapeptides had higher p_t values. Hence, tetrapeptide 46–49 ($p_t = 2.88 \times 10^{-4}$, $<P_t> = 1.12$) was predicted as a β-turn rather than 47–50 ($p_t = 1.28 \times 10^{-4}$, $<P_t> = 1.28$) because of its higher bend probability p_t value despite its lower $<P_t>$ value.

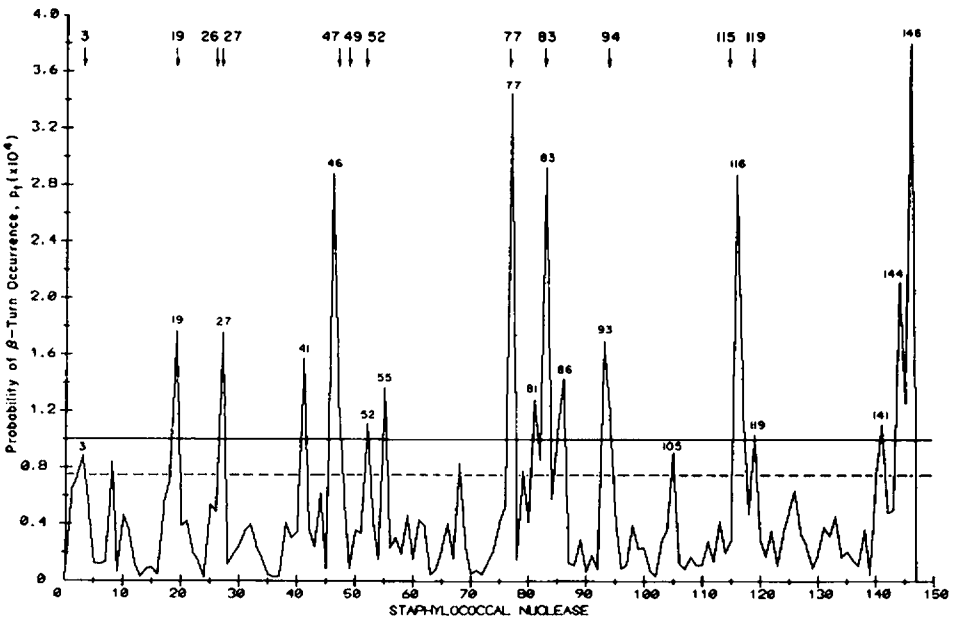

FIGURE 1 Probability of tetrapeptide β-turns in staphylococcal nuclease. The solid and dashed horizontal lines correspond to cutoff values of $p_t = 1.00 \times 10^{-4}$ (double the average bend probability, $<p_t>$) and 0.75×10^{-4} (3/2 the $<p_t>$), respectively. The numbered peaks indicate bends predicted at position i using the computer predictive algorithm depending on the cutoffs used. β-turns found by x-ray analysis are shown as arrows at the top.

TABLE II

COMPUTER PREDICTIONS OF β-TURNS IN STAPHYLOCOCCAL NUCLEASE.*
PROBABLE BENDS AND PREDICTED BENDS

β-Turn	Tetrapeptide		$p_t(\times 10^4)$	$<P_t>$	$<P_\alpha>$	$<P_\beta>$
(3–6)‡	Ser-Thr-Lys-Lys	?	0.89	1.10	0.98	0.86
8–11	His-Lys-Glu-Pro	?	0.84	1.05	1.06	0.63
(19–22)	Asp-Gly-Asp-Thr	*	1.77	1.36	0.86	0.76
(27–30)	Tyr-Lys-Gly-Gln	*	1.76	1.17	0.88	1.01
(41–44)	Thr-Pro-Glu-Thr	*	1.57	1.05	0.94	0.82
(46–49)	His-Pro-Lys-Lys	*	2.88	1.12	0.97	0.73
47–50	Pro-Lys-Lys-Gly	*	1.28	1.27	0.87	0.70
(52–55)	Glu-Lys-Tyr-Gly	*	1.12	1.11	0.98	0.83
(55–58)	Gly-Pro-Glu-Ala	*	1.37	1.12	1.02	0.63
68–71	Asn-Ala-Lys-Lys	?	0.84	1.06	1.10	0.80
(77–80)	Asn-Lys-Gly-Gln	*	3.45	1.27	0.88	0.87
(79–82)	Gly-Gln-Arg-Thr	?	0.78	1.11	0.87	0.99
(81–84)	Arg-Thr-Asp-Lys	*	1.29	1.09	1.00	0.85
82–85	Thr-Asp-Lys-Tyr	?	0.85	1.14	0.92	0.98
(83–86)	Asp-Lys-Tyr-Gly	*	2.93	1.29	0.86	0.88
85–88	Tyr-Gly-Arg-Gly	*	1.05	1.30	0.70	0.98
(86–89)	Gly-Arg-Gly-Leu	*	1.44	1.16	0.83	0.93
(93–96)	Tyr-Ala-Asp-Gly	*	1.70	1.21	0.92	0.90
94–97	Ala-Asp-Gly-Lys	*	1.19	1.17	1.04	0.72
(105–108)	Arg-Gln-Gly-Leu	?	0.91	1.01	0.97	1.02
(116–119)	Lys-Pro-Asn-Asn	*	2.88	1.41	0.77	0.77
117–120	Pro-Asn-Asn-Thr	*	1.28	1.40	0.69	0.88
(119–122)	Asn-Thr-His-Glu	*	1.03	1.05	1.00	0.83
(141–144)	Ser-Glu-Asn-Asp	*	1.11	1.30	0.99	0.64
(144–147)	Asp-Ala-Asp-Ser	*	2.12	1.26	1.05	0.67
145–148	Ala-Asp-Ser-Gly	*	1.25	1.28	0.94	0.72
(146–149)	Asp-Ser-Gly-Gln	*	3.80	1.36	0.87	0.79

*$p_t = (f_1)(f_2)(f_3)(f_4)$ is the probability of bend occurrence. $<P_t>$, $<P_\alpha>$, $<P_\beta>$ are, respectively, the average conformational potential for the tetrapeptide to be in the β-turn, α-helix, and β-sheet conformation. Tetrapeptides with $p_t > 1.00 \times 10^{-4}$ (*) or $p_t > 0.75 \times 10^{-4}$ (?) were selected as potential bends. The predicted β-turns (denoted in parentheses) all have $<P_t> > 1.00$ and $<P_\alpha> < <P_t> > <P_\beta>$ as well as $p_t > 0.75 \times 10^{-4}$.
‡(Predicted bends).

As a second example, the bend probability profile of cytochrome b_5 is plotted in Fig. 2, with the cutoff point, $p_t = 0.75 \times 10^{-4}$ represented by dashed lines. Table III shows the 16 tetrapeptides with $p_t > 1.00 \times 10^{-4}$ (denoted as *) and 3 tetrapeptides with $1.00 \times 10^{-4} < p_t > 0.75 \times 10^{-4}$. Using the lower cutoff point, the 19 probable bends are reduced to 11 predicted bends as shown in parentheses. Hence, tetrapeptide 4–7 is omitted due to $<P_\beta> = 1.35 > <P_t> = 0.95$. In addition, tetrapeptides 16–19, 19–22, 40–43, 50–53, 63–66, 82–85, and 90–93 were eliminated due to higher p_t values of adjacent tetrapeptides predicted to be β-turns. It should be noticed that 18–21 was predicted as a bend despite its lower p_t value when compared to 17–20, however, it has a higher p_t value when compared 19–22 listed in Table III. Likewise, 81–84 was predicted as a bend because it has a higher p_t value than 82–85 even though 80–83 has an even higher p_t value. That is, a sequence of adjacent probable bends beginning at residues 15–19–80 through 83 in the case for cytochrome b_5 (Table III) are compared pairwise by the computer according to their p_t values, and the

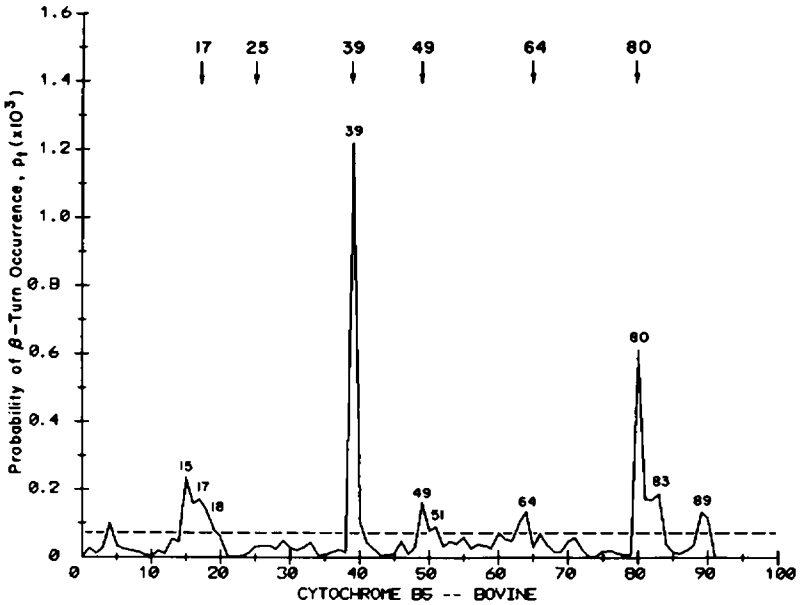

FIGURE 2 Probability of tetrapeptide β-turns in cytochrome b_5. The dashed horizontal lines correspond to a cutoff value of $p_t = 0.75 \times 10^{-4}$. The bends predicted by the automated computer analysis are numbered at the peaks with x-ray results shown as arrows.

TABLE III
COMPUTER PREDICTIONS OF β-TURNS IN CYTOCHROME b_5*.
PROBABLE BENDS AND PREDICTED BENDS

β-Turn	Tetrapeptide		p_t (× 10^4)	$<P_t>$	$<P_\alpha>$	$<P_\beta>$
4–7	Val-Lys-Tyr-Tyr	*	1.02	0.95	0.90	1.35
(15–18)‡	His-Asn-Asn-Ser	*	2.35	1.37	0.78	0.85
16–19	Asn-Asn-Ser-Lys	*	1.59	1.39	0.82	0.82
(17–20)	Asn-Ser-Lys-Ser	*	1.71	1.36	0.84	0.78
(18–21)	Ser-Lys-Ser-Thr	*	1.36	1.21	0.88	0.86
19–22	Lys-Ser-Thr-Trp	?	0.83	1.09	0.96	1.01
(39–42)	His-Pro-Gly-Gly	*	12.20	1.39	0.68	0.73
40–43	Pro-Gly-Gly-Glu	*	1.05	1.34	0.81	0.61
(49–52)	Gln-Ala-Gly-Gly	*	1.62	1.19	0.92	0.86
50–53	Ala-Gly-Gly-Asp	?	0.79	1.31	0.89	0.72
(51–54)	Gly-Gly-Asp-Ala	?	0.90	1.31	0.89	0.72
63–66	His-Ser-Thr-Asp	*	1.02	1.20	0.90	0.84
(64–67)	Ser-Thr-Asp-Ala	*	1.35	1.13	1.01	0.83
(80–83)	His-Pro-Asp-Asp	*	6.11	1.35	0.90	0.63
(81–84)	Pro-Asp-Asp-Arg	*	1.71	1.35	0.89	0.64
82–85	Asp-Asp-Arg-Ser	*	1.70	1.32	0.94	0.69
(83–86)	Asp-Arg-Ser-Lys	*	1.85	1.21	0.98	0.74
(89–92)	Lys-Pro-Ser-Glu	*	1.32	1.17	1.00	0.60
90–93	Pro-Ser-Glu-Ser	*	1.16	1.28	0.91	0.61

*Same as in Table II.
‡(Predicted bends).

TABLE IV
AUTOMATED COMPUTER PREDICTION OF β-TURNS IN 29 PROTEINS*

Adenylate kinase
6–9, 16–19, 18–21, 20–23, 30–33, 38–41, 48–51, 51–54, 85–88, 87–90, 92–95, 95–98, 119–122, 121–124, 130–133, 135–138, 140–143, 165–168, 174–177, 178–181

Carbonic anhydrase C, human
1–4, 5–8, 7–10, 9–12, 11–14, 28–31, 33–36, 40–43, 49–52, 59–62, 70–73, 71–74, 78–81, 81–84, 84–87, 85–88, 99–102, 101–104, 109–112, 124–127, 127–130, 135–138, 152–155, 167–170, 169–172, 176–179, 178–181, 183–186, 192–195, 199–202, 215–218, 217–220, 228–231, 230–233, 247–250, 249–252

Carboxypeptidase A, bovine
3–6, 5–8, 39–42, 41–44, 53–56, 54–57, 56–59, 87–90, 89–92, 91–94, 93–96, 110–113, 112–115, 120–123, 133–136, 134–137, 142–145, 144–147, 148–151, 150–153, 156–159, 157–160, 159–162, 161–164, 164–167, 166–169, 169–172, 171–174, 186–189, 196–199, 204–207, 213–216, 235–238, 238–241, 250–253, 254–257, 258–261, 260–263, 273–276, 275–278, 282–285

Chromatium Hipip
3–6, 20–23, 33–36, 37–40, 43–46, 60–63, 66–69, 72–75, 74–77, 77–80

α-Chymotrypsin, bovine
10–13, 23–26, 27–30, 35–38, 40–43, 42–45, 48–51, 62–65, 72–75, 74–77, 75–78, 89–92, 91–94, 100–103, 123–126, 126–129, 127–130, 133–136, 151–154, 163–166, 165–168, 168–171, 184–187, 188–191, 191–194, 193–196, 194–197, 202–205, 204–207, 216–219, 218–221, 220–223, 224–227

Concanavalin A, Jack Bean
10–13, 12–15, 14–17, 16–19, 19–22, 34–37, 42–45, 56–59, 67–70, 71–74, 75–78, 94–97, 96–99, 102–105, 116–119, 117–120, 119–122, 131–134, 134–137, 136–139, 142–145, 147–150, 149–152, 151–154, 160–163, 161–164, 164–167, 167–170, 169–172; 201–204, 203–206, 206–209, 216–219, 218–221, 222–225, 224–227, 225–228, 233–236

Cytochrome b_5, bovine
15–18, 17–20, 18–21, 39–42, 49–52, 51–54, 64–67, 80–83, 81–84, 83–86, 89–92

Cytochrome *c*, horse
21–24, 24–27, 26–29, 29–32, 39–42, 43–46, 48–51, 50–53, 52–55, 54–57, 70–73, 75–78

Cytochrome c_2, *Rhodospirillum rubrum*
11–14, 21–24, 24–27, 26–29, 29–32, 42–45, 43–46, 49–52, 71–74, 73–76, 80–83, 82–85, 84–87, 96–99

Elastase, porcine
17–20, 23–26, 24–27, 27–30, 36–36C, 36B–38, 40–43, 42–45, 48–51, 72–75, 74–77, 76–79, 78–81, 91–94, 95–98, 114–117, 125–128, 131–134, 132–135, 134–137, 145–149, 168–170A, 170–171, 170A–172, 176–179, 182–185, 184–187, 188A–191, 191–194, 193–196, 194–197, 203–205

Ferredoxin, *Micrococcus aerogenes*
5–8, 12–15, 15–18, 18–21, 23–26, 31–34, 33–36, 37–40, 38–41, 45–48, 49–52, 51–54

Flavodoxin, *Clostridium* MP
6–9, 8–11, 11–14, 25–28, 27–30, 29–32, 34–37, 77–80, 86–89, 87–90, 91–94, 102–105, 104–107, 125–128

Hemoglobin, glycera
19–22, 20–23, 23–26, 25–28, 27–30, 47–50, 50–53, 52–55, 75–78, 90–93, 92–95, 93–96, 104–107, 136–139

α-Hemoglobin, horse
6–9, 16–19, 22–25, 48–51, 49–52, 58–61, 62–65, 76–79, 94–97, 101–104, 113–116, 114–117, 138–141

β-Hemoglobin, horse
3–6, 22–25, 43–46, 48–51, 50–53, 55–58, 57–60, 63–66, 69–72, 78–81, 81–84, 92–95, 97–100, 99–102, 117–120, 119–122, 128–131, 143–146

Hemoglobin, lamprey
4–7, 27–30, 32–35, 87–90, 100–103, 111–114, 130–133

Hemoglobin, midge larva
12–15, 16–19, 19–22, 31–34, 43–46, 79–82, 88–91, 97–100, 112–115

Insulin, porcine (A chain)
4-7, 18–21

Insulin, porcine (B chain)
8–11, 20–23, 27–30

TABLE IV (continued)

Lactate dehydrogenase, dogfish
16–19, 17–20, 19–23, 35–38, 66–69, 68–71, 79–82, 81–85, 88–91, 99–102, 131–133, 142–145, 144–147, 146–149, 152–155, 162–165, 166–169, 183–186, 193–196, 195–198, 203–206, 217–220, 225–228, 237–240, 239–242, 269–272, 277–280, 281–284, 294–297, 308–311

Lysozyme, hen egg white
4–7, 13–16, 17–20, 19–22, 20–23, 25–28, 35–38, 37–40, 42–45, 44–47, 46–49, 47–50, 50–53, 51–54, 59–62, 64–67, 65–68, 69–72, 71–74, 72–75, 78–81, 85–88, 91–94, 99–102, 101–104, 104–107, 111–114, 113–116, 115–118, 117–120

Myogen, carp muscle
22–25, 37–40, 39–42, 51–54, 53–56, 77–80, 78–81, 89–92, 90–93, 92–95

Myoglobin, sperm whale
2–5, 23–26, 25–28, 36–39, 62–65, 116–119, 119–122, 120–123

Papain, papaya
8–11, 18–21, 20–23, 22–25, 23–26, 41–44, 44–47, 46–49, 48–51, 56–59, 57–60, 59–62, 60–63, 63–66, 67–70, 76–79, 82–85, 85–88, 95–98, 99–102, 101–104, 106–109, 114–117, 117–120, 138–141, 144–147, 151–154, 153–156, 164–167, 167–170, 173–176, 175–178, 178–181, 182–185, 183–186, 190–193, 192–195, 193–196, 195–198, 198–201, 203–206, 205–208

Ribonuclease S, bovine
12–15, 14–17, 15–18, 21–24, 22–25, 23–26, 24–27, 32–35, 36–39, 38–41, 59–62, 65–68, 69–72, 70–73, 74–77, 82–85, 86–89, 88–91, 89–92, 92–95, 110–113, 113–116

Rubredoxin, *Cl. pasteurianum*
14–17, 15–18, 17–20, 19–22, 20–23, 25–28, 27–30, 34–37, 39–42, 43–46, 45–48

Staphylococcal nuclease
3–6, 19–22, 27–30, 41–44, 46–49, 52–55, 55–58, 77–80, 79–82, 81–84, 83–86, 86–89, 93–96, 105–108, 116–119, 119–122, 141–144, 144–147, 146–149

Subtilisin BPN′
4-7, 18–21, 21–24, 23–26, 32–35, 36–39, 39–42, 44–47, 51–54, 55–58, 59–62, 60–63, 62–65, 75–78, 85–88, 98–101, 100–103, 102–105, 123–126, 125–128, 128–131, 143–146, 152–155, 155–158, 158–161, 160–163, 167–170, 169–172, 171–174, 181–184, 182–185, 193–196, 200–203, 209–212, 210–213, 212–215, 216–219, 217–220, 218–221, 224–227, 236–239, 238–241, 246–249, 257–260, 259–262, 262–265, 264–267

Thermolysin
3–6, 9–12, 10–13, 14–17, 16–19, 19–22, 31–34, 33–36, 35–38, 43–46, 50–53, 52–55, 57–60, 65–68, 72–75, 80–83, 82–85, 88–91, 92–95, 94–97, 105–108, 107–110, 109–112, 115–118, 122–125, 124–127, 133–136, 149–152, 157–160, 159–162, 179–182, 181–184, 183–186, 194–197, 198–201, 205–208, 207–210, 209–212, 211–214, 213–216, 214–217, 216–219, 218–221, 221–224, 225–228, 226–229, 231–234, 233–236, 245–248, 248–251, 259–262, 261–264, 276–279, 278–281, 280–283, 295–298, 298–301

Trypsin inhibitor, bovine pancreatic
1–4, 8–11, 10–13, 12–15, 35–38, 41–44, 43–46, 53–56, 55–58

*Cutoff value = $1.50 <p_t> = 0.75 \times 10^{-4}$ based on 4 residues.

tetrapeptide with the lower p_t value is eliminated while the higher p_t tetrapeptide is predicted as a β-turn. A complete listing of the β-turns predicted for 29 proteins by the automated computer prediction algorithm is given in Table IV.

Bend Residues and β-Turns Correctly Predicted and Localized

The computer bend predictions in Table IV were compared with the 457 β-turns elucidated from x-ray analysis (see Table II of Chou and Fasman, 1977). A summary of the percent accuracy of the automated bend predictions in 29 proteins using the cutoff $p_t > 0.75 \times 10^{-4}$ and $<P_\alpha> < <P_t> > <P_\beta>$ is given in Table V. The accuracy of β-turn residues

TABLE V

PERCENT ACCURACY OF AUTOMATED COMPUTER PREDICTION OF β-TURNS IN 29 PROTEINS*

Protein	N	n_t	t_o	t_m	$\%_t$	$\%_{nt}$	Q_t	C_t
Adenylate kinase	194	58	29	17	70.7	78.7	74.7	0.47
Carbonic anhydrase C, human	259	76	65	20	73.7	64.5	69.1	0.35
Carboxypeptidase A, bovine	307	109	52	39	64.2	73.7	69.0	0.37
Chromatium Hipip	85	55	3	21	61.8	90.0	75.9	0.50
α-chymotrypsin, bovine	241	113	39	48	57.5	70.5	64.0	0.28
Concanavalin A, Jack Bean	237	93	44	23	75.3	69.4	72.4	0.44
Cytochrome b_5, bovine	93	24	12	4	83.3	82.6	83.0	0.61
Cytochrome *c*, horse	104	27	21	10	63.0	72.7	67.8	0.32
Cytochrome c_2, *R. rubrum*	112	41	21	16	61.0	70.4	65.7	0.31
Elastase, porcine	240	92	34	30	67.4	77.0	72.2	0.44
Ferredoxin, *M. aerogenes*	54	31	13	5	83.9	43.5	63.7	0.30
Flavodoxin, *Clostridium* MP	138	49	27	28	42.9	69.7	56.3	0.13
Hemoglobin, glycera	147	20	31	11	45.0	75.6	60.3	0.16
A-hemoglobin, horse	141	24	32	9	62.5	72.6	67.6	0.28
B-hemoglobin, horse	146	20	49	2	90.0	61.1	75.6	0.35
Hemoglobin, lamprey	148	27	17	16	40.7	86.0	63.3	0.26
Hemoglobin, midge larva	136	29	19	13	55.2	82.2	68.7	0.35
Insulin, porcine (A chain)	21	8	7	7	12.5	46.2	29.3	−0.41
Insulin, porcine (B chain)	30	8	5	1	87.5	77.3	82.4	0.58
Lactate dehydrogenase, dogfish	329	112	51	56	50.0	76.5	63.2	0.27
Lysozyme, hen egg white	129	61	37	13	78.7	45.6	62.1	0.26
Myogen, carp muscle	108	24	16	11	54.2	81.0	67.6	0.33
Myoglobin sperm whale	153	26	21	21	19.2	83.5	51.3	0.03
Papain, papaya	212	57	71	4	93.0	54.2	73.6	0.42
Ribonuclease S, bovine	124	36	28	2	94.4	68.2	81.3	0.57
Rubredoxin, *Cl. pasteurianum*	54	35	3	8	77.1	84.2	80.7	0.59
Staphylococcal nuclease	149	42	27	5	88.1	74.8	81.4	0.57
Subtilisin BPN′	275	95	69	17	82.1	61.7	71.9	0.42
Thermolysin	316	116	78	23	80.2	61.0	70.6	0.40
Trypsin inhibitor, Bovine pancreatic	58	16	18	6	62.5	57.1	59.8	0.18
Summary	4,740	1,524	939	486	68.1	70.8	69.5	0.37

The symbols N and n_t represent, respectively, the total number of residues and the number of β-turn residues in each protein. t_o and t_m are, respectively, the number of β-turn residues overpredicted and missed in prediction. $\%_t = 100\,(n_t - t_m)/n_t$ and $\%_{nt} = 100\,(N - n_t - t_o)/(N - n_t)$ are, respectively, the percent of β-turn residues and the percent of non-β-turn residues predicted correctly. The quality of bend prediction is given by the average of $\%_t$ and $\%_{nt}$, $Q_t = (\%_t + \%_{nt})/2$. C_t is the correlation coefficient between predicted and observed β-turns as given by Eq. 10. $C_t = 1.00$ indicates a perfect prediction; $C_t = 0$ indicates a prediction no better than random (i.e., $Q_t = 50\%$); $C_t = -1.0$ indicates total disagreement (i.e., $Q_t = 0\%$). The figures in the first 4 columns of the Summary row were used to compute the overall percent accuracy for the 29 proteins.

*The probability of β-turn occurrence is computed from $p_t = (f_1)(f_2)(f_3)(f_4)$ based on the frequency of residues at the 1st, 2nd, 3rd, and 4th position in the β-turns of 29 proteins from Table I. Tetrapeptides with $p_t > 0.75 \times 10^{-4}$ and $<P_\alpha> < <P_t> > <P_\beta>$ were predicted as β-turns.

predicted correctly is $\%_t = 68\%$, whereas the accuracy of nonbend residues predicted as nonbends is $\%_{nt} = 71\%$. The quality of bend prediction is $Q_t = 70\%$, and the β-turn correlation coefficient is $C_t = 0.37$. The automated computer analysis using the cutoff point $p_t > 1.00 \times 10^{-4}$ gave the following results for β-turn prediction: $\%_t = 59\%$, $\%_{nt} = 78\%$, $Q_t = 68\%$, and $C_t = 0.36$. Hence, the lower cutoff value ($p_t = 3/2 <p_t> = 0.75 \times 10^{-4}$) predicts more bend residues correctly, whereas the higher cutoff ($p_t = 2 <p_t> = 1.00 \times 10^{-4}$) predicts more nonbend residues correctly. Although the Q_t values show 68% and 70% for the higher and lower p_t cutoffs, respectively, the overall bend predictive accuracy according to Eq. 5 is slightly better using the higher cutoff $\%_{t+nt} = 72\%$ as compared to the lower cutoff, $\%_{t+nt} = 70\%$. The β-turn correlation coefficient C_t for both methods are almost identical.

The effect of using different cutoff points in bend predictions may be seen in Fig. 1 for staphylococcal nuclease. The 12 β-turns found by x-ray analysis are shown in arrows whereas the high probability peaks represent the predicted β-turns. When the cutoff value $p_t = 2 <p_t>$ (solid line) is lowered to $3/2 <p_t>$ (dashed lines), an additional β-turn (tetrapeptide 3–6) is correctly identified. Because there are 42 bend residues in staphylococcal nuclease, this accounts for a dramatic 10% increase in $\%_t$ (78.6% → 88.1%). On the other hand, there is only a 4% decrease in $\%_{nt}$ (78.5 → 74.8%), even though extra bends (8–11, 68–71, 79–82, 82–85, and 105–108) were observed above the lower cutoff. However, bends 8–11 and 68–71 were not predicted by computer analysis because their $<P_t>$ values were lower than $<P_\alpha>$ or $<P_\beta>$ values (Table II). Tetrapeptide 82–85 was not predicted as a bend due to higher p_t values of its adjacent neighbors 81–84 and 83–86. Although 79–82 was included as a bend in Table II, no overprediction resulted because these residues were already included in the predicted β-turns at 77–80 and 81–84. Hence, the lower cutoff resulted in only four overpredicted β-turn residues at 105–108 with $Q_t = 81.4\%$, $C_t = 0.57$ as compared to $Q_t = 78.5\%$, $C_t = 0.53$ for the higher cutoff.

In the case of cytochrome b_5, there are six nonoverlapping bends from x-ray analysis represented by arrows in Fig. 2. The high probability bend peaks show that five of the β-turns were exactly localized, whereas the type III′ bend at 25–28 was missed completely in the prediction. Using the cutoff, $p_t = 0.75 \times 10^{-4}$ (shown in dashed lines of Fig. 2), the automated computer prediction gave $\%_t = 83\%$, $\%_{nt} = 83\%$, $Q_t = 83\%$, and $C_t = 0.61$. A slight improvement is obtained by selecting the higher cutoff $p_t = 1.00 \times 10^{-4}$ yielding $\%_t = 83\%$, $\%_{nt} = 87\%$, $Q_t = 85\%$, and $C_t = 0.66$. It is interesting to note that His 39 and His 63, which bind to the heme iron at the top of the crevasse of cytochrome b_5 (Mathews et al., 1972), were previously assigned to the random coil state. A survey of His at 12 positions around the β-turns of 29 proteins (Chou and Fasman, 1977) showed that is has the highest bend potential at the ith position ($P_{t_1} = 1.53$) and just before the bend at the $(t - 1)$ position ($P_{t-1} = 1.61$). Hence it is not surprising to find His residues 39 and 63 as well as 26 and 80 located at chain reversal regions in cytochrome b_5. In particular, the tetrapeptide 39–42, His-Pro-Gly-Gly, in cytochrome b_5 with $p_t = 1.22 \times 10^{-3}$ (see Fig. 2) has such a high probability of bend occurrence that its p_t value is 22 times that of the average $<p_t>$. The only other β-turn in the 29 proteins surveyed (Chou and Fasman, 1977) that had a higher p_t value was subtilisin BPN′ 238–241, His-Pro-Asn-Trp, with $p_t = 1.34 \times 10^{-3}$. It is surprising that

both of these highly probable bends have His-Pro at the ith and $i + 1$ positions and that both are type I bends.

The number of β-turns and the number of β-turn residues, n_t in each of the 29 proteins are listed in Table VI according to the x-ray analysis (Table II of Chou & Fasman, 1977). With the exception of cytochrome c_2, the β-turn content in the heme proteins are all less than the average found in proteins, $< f_t > = 0.32$. Helical proteins such as the six globins have less β-turns (17%) than average, whereas β-sheet proteins (concanavalin A, α-chymotrypsin, and elastase) have more β-turns (41%) than the average. It is interesting to note that the three proteins with highest β-turn content—*Chromatium* HiPIP (65%), ferredoxin (57%), and rubredoxin (65%) all have iron-sulfur clusters. The geometrical constraint of the iron atom coordination to the four Cys residues from distant part of the protein sequence gives rise to greater chain-reversals in the folding of these proteins. The fact that Cys occurs with greater than average frequency at 11 of the 12 bend positions was mentioned earlier (Chou and Fasman, 1977).

An examination of the list of 459 β-turns from x-ray studies in Table II of Chou and Fasman (1977) shows that 264 bends were predicted exactly based on probabilities alone (* indicates $p_t > 1.00 \times 10^{-4}$ and ? indicates $1.00 \times 10^{-4} < p_t > 0.75 \times 10^{-4}$). This is represented in the last row as 58% under the ± 0 column of correctly localized β-turns in Table VI. Similarly 71% of the bends were localized within one residue (i.e., x-ray bend 36–39, predicted 37–40) as shown in the ± 1 column, and 78% of the bends were correctly localized within two residues as shown in the ± 2 column. Of the 421 bends classified into 11 bend types (Table I of Chou and Fasman, 1977) 96 bends (23%) were not localized within ± 2 residues. A further analysis was made to see whether the probability bend prediction missed certain bend types (within ± 2 residues) more often than others. The results are as follows (bends missed/total bends = fraction of bends missed) for the following bend types: I (47/176 = 0.27), I′ (2/13 = 0.15), II (11/64 = 0.17), II′ (5/20 = 0.25), III (16/77 = 0.21), III′ (4/13 = 0.31), IV (7/35 = 0.20), V (0/3 = 0), V′ (2/4 = 0.50), VI (0/8 = 0), and VII (2/8 = 0.25). Although the statistical sampling of the different bend types is too small in most cases for reliable correlation, it is interesting nevertheless to find type I′ and II bends more frequently localized (85 and 83%, respectively) than average (77%). This is due predominantly to the high frequency of Gly at the third position of type I′ and II bends which results in high p_t values. It is also worthy to note that the correct localization of bend type I in chain reversal regions was less than the average (73% cf. 77%) despite the fact that more type I bends (176/421 = 0.42) were used in calculating the β-turn potentials than other bend types. It is clear from the above analysis that more statistical sampling of β-turns classified in various bend types are required before chain reversal regions can be accurately predicted according to their φ, ψ dihedral angles.

β-Turn Predictions Using Environmental Bend Positional Potentials

Although tetrapeptide sequences have been used extensively to predict chain reversal regions (Lewis et al., 1971, 1973; Crawford et al., 1973; Chou and Fasman 1974), it was pointed out that only the inner two residues contain the bonds whose orientations actually define a reverse turn (Robson and Pain, 1974). Of the 17 β-turns observed in *Chromatium* HiPIP, Carter et al. (1974) found that the 4-residue and 2-residue method predicted 10 and 16 of the bends,

TABLE VI

NUMBER OF β-TURNS CORRECTLY LOCALIZED IN 29 PROTEINS USING THE CUTOFF VALUE $p_t > 0.75 \times 10^{-4}$ THAT CORRESPONDS TO ½ THE AVERAGE PROBABILITY OF BEND OCCURRENCE

Protein	No. of β-turns	n_t	f_t	Correctly localized β-turns		
				±0	±1	±2
1 Adenylate kinase	17	58	0.30	10	13	14
2 Carbonic anhydrase C	21*	77*	0.29	11	17	19
3 Carboxypeptidase A	36	109	0.36	20	25	28
4 *Chromatium* HiPIP	17	55	0.65	9	11	12
5 α-Chymotrypsin, bovine	32	113	0.47	16	18	20
6 Concanavalin A, Jack Bean	28	93	0.39	16	22	23
7 Cytochrome b_5, bovine	6	24	0.26	5	5	5
8 Cytochrome *c*, horse	7	27	0.26	4	4	4
9 Cytochrome c_2, *R. rubrum*	11	41	0.37	6	6	6
10 Elastase, porcine	27	92	0.38	15	16	16
11 Ferredoxin, *M. aerogenes*	10	31	0.57	8	9	10
12 Flavodoxin, *Clostridium* MP	17	49	0.36	4	6	7
13 Hemoglobin, glycera	5	20	0.14	3	3	4
14 α-Hemoglobin, horse	6	24	0.17	4	5	5
15 β-Hemoglobin, horse	6	20	0.14	5	6	6
16 Hemoglobin, lamprey	7	27	0.18	4	5	5
17 Hemoglobin, midge larva	10*	32*	0.21	5	6	7
18 Insulin, porcine	4	16	0.31	3	3	3
19 Lactate dehydrogenase	33	112	0.34	12	17	22
20 Lysozyme, egg white	20	61	0.47	13	17	19
21 Myogen, carp muscle	6	24	0.22	4	4	4
22 Myoglobin, sperm whale	9	26	0.17	4	5	6
23 Papain, papaya	18	57	0.27	12	15	17
24 Ribonuclease S, bovine	11	36	0.29	8	10	11
25 Rubredoxin, *Cl. pasteurianum*	11	35	0.65	7	8	9
26 Staphylococcal nuclease	12	42	0.28	9	11	11
27 Subtilisin BPN'	29	95	0.35	20	25	27
28 Thermolysin	39	116	0.37	25	32	35
29 Trypsin inhibitor, bovine	4	16	0.28	2	2	3
Total	459*	1528*	0.32	264	326	358
β-Turns correctly localized, %				58	71	78

*Two additional β-turns were located from x-ray analysis (carbonic anhydrase C 27–30 and midge larva hemoglobin 72–75) after the computer bend prediction, increasing the total bends from 457 to 459. The number of β-turns observed from x-ray analysis in the 29 proteins are given in Table II of Chou and Fasman (1977). n_t is the number of bend residues in each protein where overlapping β-turn residues were not counted twice. f_t is the fraction of β-turn residues in each protein. The number of bends that were localized exactly (i.e., x-ray bend 17–20, predicted 17–20) is given in the ±0 column. (These correctly predicted bends are indicated by * and ? in Table II of Chou and Fasman, 1977.) The number of bends localized within one residue (i.e., x-ray bend 36–39, predicted 37–40) is given in the ±1 column. The number of bends localized within two residues (i.e., x-ray bend 74–77, predicted 72–75) is given in the ±2 column.

respectively. These results suggest that the middle bend residues may be more influential than those found at positions 1 and 4 in dictating the locus of a reverse turn. On the other hand, an environmental analysis of β-turn neighboring residues from the x-ray studies of 29 proteins indicates that the chain reversal regions are stabilized by β-β, α-α, and α-β interactions (Chou and Fasman, 1977). Because the β-turn potential at the 4 positions before and after the bend also showed dramatic positional preferences for the 20 amino acids, the use of these neighboring bend protentials may yield better β-turn predictions than the conventional 4-residue approach.

The probability of bend occurrence at residue i obtained from the frequencies of 2, 4, 8, and 12 residues is:

$$\begin{aligned} p_{t_m} &= (f_{i+1})(f_{i+2}) \\ p_t &= (f_i)(f_{i+1})(f_{i+2})(f_{i+3}) \\ p_{t_8} &= (f_{t-2})(f_{t-1})(p_t)(f_{t+1})(f_{t+2}) \\ p_{t_{12}} &= (f_{t-4})(f_{t-3})(p_{t_8})(f_{t+3})(f_{t+4}) \end{aligned} \tag{11}$$

Because the average frequency of a bend residue at position j in proteins is $<f_t> = 457/4{,}740 = 0.096$, the average probability of bend occurrence calculated from 2, 4, 8, and 12 residues is, respectively: $<p_{t_m}> = 0.92 \times 10^{-12}$, $<p_t> = 0.85 \times 10^{-4}$, $<p_{t_8}> = 0.72 \times 10^{-8}$, and $<p_{t_{12}}> = 0.61 \times 10^{-12}$.

Cutoffs at 1.00, 1.25, 1.50, 1.75, and 2.00 times these average probabilities of bend occurrence were analyzed by automatic computer prediction (Table VII). It is seen that higher $\%_t$ and lower $\%_{nt}$ values are obtained as the bend probability cutoff value is raised. The overall accuracy of bend prediction, $\%_{t+nt}$ (Eq. 5), which is equivalent to the weight-averaged quality of bend prediction, $\overline{Q}_t$ (Eq. 6), shows a slight gain in percent accuracy with higher cutoff values. At the average bend probability corresponding to cutoff 1.00, the $\%_{t+nt}$ value increases as more residues in the neighborhood of chain reversal are considered in the predictive analysis: 2 residues (68%), 4 residues (69.1%), 8 residues (70.1%), 12 residues (71.6%). The cutoff 1.25 $<p_t>$ was selected as most suitable for 4 residue bend prediction because it was the only cutoff that yielded 70% accuracy according to $\%_t$, $\%_{nt}$, Q_t, and $\%_{t+nt}$. It also yielded the best Q_t and C_t values for 2, 8, and 12 residues bend analysis.

The accuracy analysis of the automated computer bend prediction in 29 proteins using the cutoff 1.25 based on 4 and 12 residues is given in Table VIII. It is seen that prediction by 4 residues yielded better results in $\%_t$ for 14 proteins and identical $\%_t$ values for 7 proteins. In contrast, prediction by 12 residues gave better results in $\%_{nt}$ for 23 of the 29 proteins. That is, the use of environmental bend frequencies (Table V in Chou and Fasman, 1977) will correctly identify more nonbend residues as nonbends. The overall bend and nonbend predictive accuracy, $\%_{t+nt}$, is greater in 21 of the 29 proteins using the 12-residue method as compared to the 4-residue method. When these proteins were classified in groups, it is seen that 8 of the 9 predominantly helical heme proteins and all 3 of the predominantly β-sheet proteins (concanavalin A, α-chymotrypsin, elastase) have better $\%_{t+nt}$ values when 12 residues are used in the automatic bend prediction. However, in the three predominantly β-turn proteins (*Chromatium* HiPIP, ferredoxin, rubredoxin), the 4-residue bend prediction gave better

TABLE VII

PREDICTIVE ACCURACY OF β-TURNS FOR 29 PROTEINS BASED ON ENVIRONMENTAL ANALYSIS AND USING VARIOUS CUTOFF VALUES*

Residues‡	Cutoff*	p_t‡	t_o	t_m	$\%_t$	$\%_{nt}$	Q_t	C_t	$\%_{t+nt}$
2	1.00	0.92×10^{-2}	1,077	442	71.0	66.6	68.8	0.35	68.0
2	1.25	1.15×10^{-2}	904	505	66.9	71.9	69.4	0.37	70.3
2	1.50	1.38×10^{-2}	816	547	64.1	74.7	69.4	0.37	71.2
2	1.75	1.61×10^{-2}	686	628	58.8	78.7	68.7	0.37	72.3
2	2.00	1.84×10^{-2}	543	772	49.3	83.1	66.2	0.34	72.3
4	1.00	0.85×10^{-4}	1,069	398	73.9	66.8	70.3	0.38	69.1
4	1.25	1.06×10^{-4}	961	451	70.4	70.2	70.3	0.38	70.2
4	1.50	1.27×10^{-4}	828	529	65.3	74.3	69.8	0.38	71.4
4	1.75	1.49×10^{-4}	714	626	58.9	77.8	68.4	0.36	71.7
4	2.00	1.70×10^{-4}	605	666	56.3	81.2	68.8	0.38	73.2
8	1.00	0.72×10^{-8}	987	429	71.9	69.3	70.6	0.39	70.1
8	1.25	0.90×10^{-8}	853	477	68.7	73.5	71.1	0.40	71.9
8	1.50	1.08×10^{-8}	779	535	64.9	75.8	70.4	0.39	72.3
8	1.75	1.26×10^{-8}	714	586	61.5	77.8	69.7	0.39	72.6
8	2.00	1.44×10^{-8}	660	663	56.5	79.5	68.0	0.36	72.1
12	1.00	0.61×10^{-12}	916	430	71.8	71.6	71.7	0.41	71.6
12	1.25	0.76×10^{-12}	800	474	68.9	75.2	72.0	0.42	73.1
12	1.50	0.91×10^{-12}	709	541	64.5	78.0	71.2	0.41	73.6
12	1.75	1.06×10^{-12}	624	584	61.7	80.6	71.2	0.42	74.5
12	2.00	1.21×10^{-12}	563	648	57.5	82.5	70.0	0.41	74.5

The symbols t_o, t_m, $\%_t$, $\%_{nt}$, Q_t, and C_t are the same as in Table V. $\%_{t+nt} = 100(N-t_o-t_m)/N$ is the percentage of residues predicted correctly as either bends or nonbends. $\%_{t+nt}$ is the weighted average of $\%_t$ and $\%_{nt}$.

*Based on the bend frequencies of 457 β-turns (Chou and Fasman, 1977). The cutoff value 1.00 corresponds to the average probability of bend occurrence $<p_t>$, whereas 1.50 corresponds to $\frac{1}{2} \times <p_t>$ and 2.00 to $2 \times <p_t>$ in reference to the respective residues used in the bend analysis.

‡Probability of bend occurrence based on environmental analysis were calculated as follows: $p_{t_m} = (f_{i+1})(f_{i+2})$ for 2 residues; $p_t = (f_i)(f_{i+1})(f_{i+2})(f_{i+3})$ for 4 residues, $p_{t_8} = (f_{i-2})(f_{i-1})(p_t)(f_{i+1})(f_{i+2})$ for 8 residues, $p_{t_{12}} = (f_{i-4})(f_{i-3})(p_{t_8})(f_{i+3})(f_{i+4})$ for 12 residues.

results in $\%_{t+nt}$ for ferredoxin (67 cf. 65%) and rubredoxin (80 cf. 69%) and similar $\%_{t+nt}$ for *Chromatium* HiPIP (72 cf. 73%).

Recently, Tanaka and Scheraga (1976) developed a statistical mechanical treatment of protein conformation wherein chain reversal regions were predicted in 23 proteins after the predictions of helical and β-sheet regions. Their percent bend localized (± 2 residues) = 219/372 = 59% may be compared to the 78% correctly localized bends with ± 2 residues from the automatic computer predictions as shown in Table VI. Lenstra (1977) has compared the accuracy of bend predictions in 25 proteins according to the computer methods of Nagano (1977) and Argos et al. (1976), showing values of $C_t = 0.40$ and 0.44, respectively. These results compare favorably with the automatic computer bend predictions in 29 proteins according to 12 residues with $C_t = 0.42$ as shown in Table VIII. Although the automatic bend prediction using 4 residues has a smaller bend correlation coefficient $C_t = 0.38$, its $\%_t = 70.4\%$ is slightly better than $\%_t = 68.9\%$ from the 12-residue method. It should be remembered, however, that the automatic computer bend predictions presented herein were made without considerations of predicted helices and β-sheets. Hence, if potential bends were

TABLE VIII

PERCENT ACCURACY OF AUTOMATED COMPUTER PREDICTION OF β-TURNS IN 29 PROTEINS BASED ON 4 RESIDUES AND 12 RESIDUES USING CUTOFF VALUE 1.25 × AVERAGE BEND PROBABILITY*

Protein	N	n_t	4 Residues							12 Residues						
			t_o	t_m	$\%_t$	$\%_m$	Q_t	C_t	$\%_{t+m}$	t_o	t_m	$\%_t$	$\%_m$	Q_t	C_t	$\%_{t+m}$
Adenylate kinase, porcine	194	58	25	17	70.7	81.6	76.2	0.51	78.4	18	22	62.1	86.8	74.4	0.50	79.4
Carbonic anhydrase C, human	259	76	68	12	84.2	62.8	73.5	0.43	69.1	57	11	85.5	68.9	77.2	0.50	73.7
Carboxypeptidase A, bovine	307	109	51	37	66.1	74.2	70.1	0.39	71.3	54	35	67.9	72.7	70.3	0.39	71.0
Chromatium HiPIP	85	55	3	21	61.8	90.0	75.9	0.50	71.8	2	21	61.8	93.3	77.6	0.53	72.9
α-Chymotrypsin, bovine	241	113	46	47	58.4	65.2	61.8	0.24	61.4	37	39	65.5	72.0	68.7	0.38	68.5
Concanavalin A, Jack Bean	237	93	43	23	75.3	70.1	72.7	0.44	72.2	34	27	71.0	76.4	73.7	0.47	74.3
Cytochrome b_5, bovine	93	24	11	4	83.3	84.1	83.7	0.63	83.9	8	4	83.3	88.4	85.9	0.68	87.1
Cytochrome c, horse	104	27	21	10	63.0	72.7	67.8	0.32	70.2	19	8	70.4	75.3	72.8	0.42	74.0
Cytochrome c_2, *R. rubrum*	112	41	19	17	58.5	73.2	65.9	0.31	67.9	21	18	56.1	70.4	63.3	0.26	65.2
Elastase, porcine	240	92	42	23	75.0	71.6	73.3	0.45	72.9	37	24	73.9	75.0	74.5	0.48	74.6
Ferredoxin, *M. aerogenes*	54	31	13	5	83.9	43.5	63.7	0.30	66.7	12	7	77.4	47.8	62.6	0.26	64.8
Flavodoxin, *Clostridium* MP	138	49	26	24	51.0	70.8	60.9	0.22	63.8	29	24	51.0	67.4	59.2	0.18	61.6
Hemoglobin, glycera	147	20	35	6	70.0	72.4	71.2	0.31	72.1	28	3	85.0	78.0	81.5	0.47	78.9
α-Hemoglobin, horse	141	24	34	12	50.0	70.9	60.5	0.17	67.4	21	16	33.3	82.1	57.7	0.14	73.8
β-Hemoglobin, horse	146	20	53	2	90.0	57.9	74.0	0.33	62.3	34	1	95.0	73.0	84.0	0.49	76.0
Hemoglobin, lamprey	148	27	21	12	55.6	82.6	69.1	0.34	77.7	15	16	40.7	87.6	64.2	0.29	79.1
Hemoglobin, midge larva	136	29	15	13	55.2	86.0	70.6	0.40	79.4	9	15	48.3	91.6	69.9	0.44	82.4
Insulin, porcine (A chain)	21	8	7	7	12.5	46.2	29.3	−0.41	33.3	7	5	37.5	46.2	41.8	−0.16	42.9
Insulin, porcine (B chain)	30	8	5	1	87.5	77.3	82.4	0.58	80.0	6	0	100.0	72.7	86.4	0.64	80.0
Lactate dehydrogenase, dogfish	329	112	53	51	54.5	75.6	65.0	0.30	67.8	43	59	47.3	80.2	63.8	0.29	69.0
Lysozyme, hen egg white	129	61	37	13	78.7	45.6	62.1	0.26	61.2	30	16	73.8	55.9	64.8	0.30	64.3
Myogen, carp muscle	108	24	20	12	50.0	76.2	63.1	0.24	70.4	19	14	41.7	77.4	59.5	0.18	69.4
Myoglobin, sperm whale	153	26	25	21	19.2	80.3	49.8	−0.00	69.9	18	21	19.2	85.8	52.5	0.05	74.5
Papain, papaya	212	57	72	8	86.0	53.5	69.8	0.35	62.3	54	8	86.0	65.2	75.6	0.45	70.8
Ribonuclease S, bovine	124	36	28	2	94.4	68.2	81.3	0.57	75.8	26	6	83.3	70.5	76.9	0.49	74.2
Rubredoxin, *Cl. pasteurianum*	54	35	3	8	77.1	84.2	80.7	0.59	79.6	10	7	80.0	47.4	63.7	0.29	68.5
Staphylococcal nuclease	149	42	27	5	88.1	74.8	81.4	0.57	78.5	21	7	83.3	80.4	81.9	0.59	81.2
Subtilisin BPN′	275	95	67	13	86.3	62.8	74.5	0.47	70.9	60	13	86.3	66.7	76.5	0.50	73.5
Thermolysin	316	116	77	19	83.6	61.5	72.6	0.44	69.6	56	19	83.6	72.0	77.8	0.54	76.3
Trypsin inhibitor, bovine pancreatic	58	16	14	6	62.5	66.7	64.6	0.26	65.5	15	8	50.0	64.3	57.1	0.13	62.1
Summary	4,740	1,524	961	451	70.4	70.2	70.3	0.38	70.2	800	474	68.9	75.2	72.0	0.42	73.1

*For 4 residues, the cutoff is $1.25 \langle p_t \rangle = 1.06 \times 10^{-4}$. For 12 residues, the cutoff is $1.25 \langle p_{t12} \rangle = 0.76 \times 10^{-12}$. Symbols are same as in footnotes of Tables V and VII.

analyzed after or in conjunction with the prediction of α- and β-regions as in the case of proteinase inhibitors and various ribonucleases shown in the following paper (Chou and Fasman, 1979) improvements in predicting the chain reversal regions in proteins can be expected by using only four residues.

We appreciate the help of Miss Janet Kolodner and Mr. Tim Hickey for assistance in computer programming.

This research was generously supported in part by grants from the U. S. Public Health Service (GM 17533), National Science Foundation (PCM76-21856), and the American Cancer Society (NP-92E). This is publication No. 1249 from the Graduate Department of Biochemistry, Brandeis University, Waltham, Mass. 02154.

Received for publication 16 February 1978 and in revised form 10 January 1979.

REFERENCES

ARGOS, P., J. SCHWARZ, and J. SCHWARZ. 1976. An assessment of protein secondary structure prediction methods based on amino acid sequence. *Biochim. Biophys. Acta.* **439**:261.

CARTER, C. W., JR., J. KRAUT, S. T. FREER, N-H. XUONG, R. A. ALDEN, and R. G. BARTSCH. 1974. Two-angstrom crystal structure of oxidized *Chromatium* high potential iron protein. *J. Biol. Chem.* **249**:4212.

CHOU, P. Y., A. J. ADLER and G. D. FASMAN. 1975. Conformational prediction and circular dichroism studies on the *lac* repressor. *J. Mol. Biol.* **96**:29.

CHOU, P. Y., and G. D. FASMAN. 1974. Prediction of protein conformation. *Biochemistry.* **13**:222.

CHOU, P. Y., and G. D. FASMAN. 1977. β-Turns in proteins. *J. Mol. Biol.* **115**:135.

CHOU, P. Y., and G. D. FASMAN. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas. Mol. Biol.* **47**:45.

CHOU, P. Y., and G. D. FASMAN. 1979. Conservation of chain reversal regions in proteins. *Biophys. J.* **26**:385.

CRAWFORD, J. L., W. N. LIPSCOMB, and C. G. SCHELLMAN. 1973. The reverse turn as a polypeptide conformation in globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **70**:538.

DICKERSON, R. E., T. TAKANO, D. EISENBERG, O. B. KALLAI, L. SAMSON, A. COOPER, and E. MARGOLIASH. 1971. Ferricytochrome c—I. General features of the horse and bonito proteins at 2.8 Å resolution. *J. Biol. Chem.* **246**:1511.

FASMAN, G. D., P. Y. CHOU, and A. J. ADLER. 1976. Prediction of the conformation of the histones. *Biophys. J.* **16**:1201.

KUNTZ, I. D. 1972. Protein folding. *J. Am. Chem. Soc.* **94**:4009.

LENSTRA, J. A. 1977. Evaluation of secondary structure predictions in proteins. *Biochim. Biophys. Acta.* **491**:333.

LEWIS, P. N., F. A. MOMANY, and H. A. SCHERAGA. 1971. Folding of polypeptide chains in proteins: a proposed mechanism for folding. *Proc. Natl. Acad. Sci. U.S.A.* **68**:2293.

LEWIS, P. N., F. A. MOMANY, and H. A. SCHERAGA. 1973. Chain reservals in proteins. *Biochim. Biophys. Acta.* **303**:211.

MATHEWS, F. S., M. LEVINE, and P. ARGOS. 1972. Three-dimensional Fourier synthesis of calf liver cytochrome b_5 at 2.8 Å resolution. *J. Mol. Biol.* **64**:449.

MATTHEWS, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **405**:442.

NAGANO, K. 1977. Triplet information in helix prediction applied to the analysis of super-secondary structures. *J. Mol. Biol.* **109**:251.

ROBSON, B., and R. H. PAIN. 1974. Analysis of the code relating sequence to conformation in globular proteins. *Biochem. J.* **141**:899.

SCHULZ, G. E., C. D. BARRY, J. FRIEDMAN, P. Y. CHOU, G. D. FASMAN, A. V. FINKELSTEIN, V. I. LIM, O. B. PTITSYN, E. A. KABAT, T. T. WU, M. LEVITT, B. ROBSON, and K. NAGANO. 1974. Comparison of predicted and experimentally determined secondary structure of adenyl kinase. *Nature (Lond.).* **250**:140.

TANAKA, S., and H. A. SCHERAGA. 1976. Statistical mechanical treatment of protein conformation. 4. A four-state model for specific-sequence copolymers of amino acids. *Macromolecules.* **9**:812.