Prediction of β-Turns in Proteins Using the First-Order Markov Models

Thy-Hou Lin,* Ging-Ming Wang, and Yen-Tseng Wang

Department of Life Science, National Tsing Hua University, Hsinchu, Taiwan 30043, R.O.C.

Received June 15, 2001

We present a method based on the first-order Markov models for predicting simple β-turns and loops containing multiple turns in proteins. Sequences of 338 proteins in a database are divided using the published turn criteria into the following three regions, namely, the turn, the boundary, and the nonturn ones. A transition probability matrix is constructed for either the turn or the nonturn region using the weighted transition probabilities computed for dipeptides identified from each region. There are two such matrices constructed for the boundary region since the transition probabilities for dipeptides immediately preceding or following a turn are different. The window used for scanning a protein sequence from amino (N-) to carboxyl (C-) terminal is a hexapeptide since the transition probability computed for a turn tetrapeptide is capped at both the N- and C- termini with a boundary transition probability indexed respectively from the two boundary transition matrices. A sum of the averaged product of the transition probabilities of all the hexapeptides involving each residue is computed. This is then weighted with a probability computed from assuming that all the hexapeptides are from the nonturn region to give the final prediction quantity. Both simple β-turns and loops containing multiple turns in a protein are then identified by the rising of the prediction quantity computed. The performance of the prediction scheme or the percentage (%) of correct prediction is evaluated through computation of Matthews correlation coefficients for each protein predicted. It is found that the prediction method is capable of giving prediction results with better correlation between the percent of correct prediction and the Matthews correlation coefficients for a group of test proteins as compared with those predicted using some secondary structural prediction methods. The prediction accuracy for about 40% of proteins in the database or 50% of proteins in the test set is better than 70%. Such a percentage for the test set is reduced to 30 if the structures of all the proteins in the set are treated as unknown.

INTRODUCTION

There is an increasing interest in exploring the roles of β-turns in proteins which usually serve as linkers between the secondary structure elements of the proteins. Studies of the effects of turns on the stability and folding kinetics by modifying the loop length or sequence in the turns of α-helix bundle proteins,[1−3] β-barrel proteins,[4−6] and small proteins (<100 amino acid residues)[7,8] have been performed recently. The results show that the role of turns in the stability and folding is strongly context-dependent.[8] However, other experiments using circularly permuted protein variants[9] or mutants directly created on the turn regions of a protein[10] have demonstrated the important contribution of them to the folding and stability of the protein. Reducing the entropy of the unfolded proteins using glycine-to-x or x-to-proline mutations[11,12] has been achieved since frequent occurrence of proline or glycine residues in β-turns are found. The β-turns were first recognized by Venkatachalam[13] as a sequence of tetrapeptide in which the main chain C═O (i) and the N−H ($i + 3$) group forms a hydrogen bond. Using model-building techniques, he classified the conformation types of β-turns into 6 (I, I′, II, II′, III, III′). The definition was broadened subsequently by Lewis et al.[14] to include those that were not formed by a hydrogen bond. Their definition states that the distance between $C^{\alpha}(i)$ and $C^{\alpha}(i + 3)$ is less than 7 Å and that the chain is not in a helical conformation. The number of turn types was extended from 6 to 10 (I, I′, II, II′, III, III′, IV, V, VI, VII). However, Richardson[15] had categorized seven turn types (I, I′, II, II′, VIa, VIb, IV) by using more stringent criteria to the classification of ψ,ϕ angles. Wilmot and Thornton[16] then described a new class of β-turn designated as type VIII in which the central residues ($i + 1$, $i + 2$) adopt an $\alpha_R\beta$ conformation. Therefore, the number of turn types generally accepted is now 8 (I, I′, II, II′, IV, VIa, VIb, VIII).

A number of methods have been developed for predicting β-turns since the recognition of their existence in proteins. For instances, Chou and Fasman[17] described a method based on calculating the product of derived amino acid probabilities at each of the four positions of a turn tetrapeptide. The four positions in a turn are not treated equivalent, despite the fact that the turns are considered as one homogeneous group. The location of most β-turns on the protein surface was recognized by Kuntz,[18] and this was interpreted as minima in the hydrophobicity plots which were then used for turn prediction. Cohen et al.[19] have treated turns as irregular segments in a polypeptide chain and developed techniques based on pattern recognition for prediction of turns. Their algorithms rely on the protein class to anticipate the likely spacing between turns. By analyzing a database of 59 proteins, Wilmot and Thornton[16] have found important sequence preferences within some β-turn types. They incorporated the positional trends for type I and II β-turns in the

* Corresponding author phone: 886-3-574-2759; fax: 886-3-572-1746; e-mail: thlin@life.nthu.edu.tw.

10.1021/ci0103020 CCC: $22.00 © 2002 American Chemical Society
Published on Web 12/29/2001

same prediction scheme described by Lewis et al.[14] for improving the positional prediction for β-turns. Recently, Chou et al.[20−23] described a method for prediction of simple or tight β-turns for proteins based on using the first-order Markov chain; i.e., only the first-order sequence coupling effect is taken into account. It appeared that the quality of prediction was significantly improved by the new model, implying that the residue-coupled effect along a polypeptide chain is important for the formation of β-turns.

Markov models are well-known tool for analyzing biological sequence data, and the predominant model for microbial sequence analysis is a fixed-order Markov chain.[24−27] A first-order Markov model predicts each base/residue of a DNA/protein sequence using only the previous base/residue to predict the next base/residue.[25−27] In this work, we have identified the β-turns for a database of 338 proteins using the criteria given by Wilmot and Thornton[16] and the secondary structure assignments by the method of Kabsch and Sander.[28] We divided each protein sequence to the following three regions, namely, the β-turns, the nonturn areas, and the boundary regions for residues immediately preceding or following a β-turn. However, we have found that about 37% of turns was not a simple β-turn formed by a tetrapeptide but rather a combination of different types of β-turns within a loop. Some of these multiple turns were badly distorted. However, they were also treated as the turn sequences if their corresponding φ, ϕ angles fulfilled the turn type criteria given by Wilmot and Thornton.[16] Therefore, our prediction scheme is designed for both the simple β-turns and for the loops containing multiple turns. The first-order Markov transition probability matrices were then constructed for sequences in each region. To predict simple β-turns and loops for a protein, we computed the transition probability for every hexapeptide scanning from the N- to the C-terminal of the protein. The basic scanning unit was a hexa- rather than a tetrapeptide because the transition probability of each tetrapeptide was capped at both the N- and C-ends with a boundary transition probability computed from the boundary region. A nonturn transition probability for each hexapeptide was also computed from the transition probability matrix of the nonturn region. Both of these transition probabilities were averaged to generate the corresponding transition probability for each residue in the protein. The averaged and capped transition probability for each residue was then weighted with the corresponding averaged nonturn transition probability to give the prediction quantity for each residue. Consecutive residues with rising prediction quantities computed are assigned as a simple β-turn or a loop depending on the number of residues in which the peak is spanned. The prediction accuracy for each protein was accessed through computation of the Matthews correlation coefficient.[29] We found that the percent of correct prediction for about 40% of proteins in the database was above 70. The same prediction scheme was applied to a group of 28 proteins randomly selected from the Protein Data Bank (PDB)[30] and treated as a test set, and the result was that the percent of correct prediction for 50% of these proteins was greater than 70. Such a percentage was reduced to 30 if we treated all the proteins in the test set as proteins of unknown structures.

MATERIALS AND METHODS

Proteins with redundant records in atomic coordinates were discarded from an original list chosen from the PDB.[30] We also allowed only single-chain proteins in the database. The primary list was further screened using the β-turn criteria described by Wilmot and Thornton.[16] This would eliminate some defective proteins in which some residues could not be consecutively assigned as β-turn residues. The number of proteins in the final list was 338, and they were used to form the database. There were also 28 proteins selected using the same way from the PDB, and these were treated as the test set. The PDB codes of all these proteins were listed in Table 1. Secondary structures of each of these proteins were defined using the method of Kabsch and Sander[28] implemented in the SYBYL 6.6 program.[31] As mentioned above, the β-turns for each protein were identified using the criteria that the distance between the C^{α} of residue i and the C^{α} of residue $i + 3$ was less than 7 Å, and the central residues were not helical. The β-turns identified for all the proteins were agreed with those assigned by the method of Kabsch and Sander,[28] i.e., the central residues of the identified β-turns were either the "T" or "_" residues assigned by the method. The φ, ϕ angles of each protein was computed using the Biopolymer module of SYBYL 6.6 program[31] and they were used for classification of β-turn types according to the type criteria described by Wilmot and Thornton.[16] Ideal φ, ϕ angles were allowed to vary by ±30° for classification since most of the β-turns identified were distorted. We found that except simple β-turns there were also loops containing different types of β-turns or multiple turns on the proteins. Sequences of simple β-turns or loops containing multiple turns were all treated as turn sequences. Therefore, we divided the sequence of each protein into three regions, namely, the turn, the nonturn, and the boundary regions. The boundary regions were identified for residues immediately preceding or following a turn sequence.

Sequences of each of these regions were scanned from N- to C-terminal for counting the frequency of occurrence of each dipeptide. The frequency of occurrence of each dipeptide on sequences of all the proteins in the database was also counted. The frequency of each dipeptide in each particular region was weighted by the later one. A first-order conditional probability was calculated for each dipeptide in each region. For example, for a dipeptide with sequence -*GR*-, the conditional probability $p(R/G)$ for presence of R given presence of G was calculated as follows

$$p(R/G) = \frac{f(R/G)}{f(G)} \tag{1}$$

where *f(R/G)* was the weighted frequency of the dipeptide -*GR*- found in a particular region whose first residue was *G* (Gly) and whose second residue was *R* (Arg), and *f(G)* was the total weighted frequency of those dipeptides in the same region whose first residue was *G* and the second residue was any of the 20 standard amino acid residues. We computed such a conditional probability for each dipeptide in sequences of each region from N- to C-terminal. To construct a transition probability matrix for each region, the conditional probabilities obtained for each region were sorted and then summed according to the order of a hydrophobicity scale given by Engleman et al.[32] Only one transition probability

Table 1. Proteins Used in the Study

database
1ppt(1.37)[a]
1crn(1.5)
5rxn(1.2)
1ovoA(1.9)
1fxd(1.7)
1ptx(1.3)
2cab(2.0)
1ctf(1.7)
2abxA(2.5)
3icb(2.3)
1hcp
1iml
1ubq(1.8)
1kveB(1.8)
1cc5(2.5)
1srsA(3.2)
3b5c
1hip(2.0)
2bopA(1.7)
1poh(2.0)
2gn5(2.3)
2hts(1.83)
1cyo(1.5)
1molA(1.7)
1ihfB(2.5)
1who(1.9)
1beo(2.2)
1frd(1.7)
2pcy(1.8)
2hpeA(2.0)
1xer(2.0)
1tlk(2.8)
1shaA(1.5)
1onc(1.7)
1aaj(1.8)
1erw(1.8)
1qapA(2.8)
1nar(1.8)
1amp(1.8)
1ryc(1.8)
2ora(2.0)
1dhpA(2.5)
1csn(2.0)
1fnb(1.7)
1lbiA(2.7)
1gym(2.2)
8abp(1.49)
5abp(1.8)
1ecrA(2.7)
2dln(2.3)
3cpa(2.0)
5cpa(1.54)
2ctc(1.4)
4cpa(2.5)
1glg(2.0)
1dorA(2.0)
2 cmd(1.87)
1nfkA(2.3)
2exo(1.8)
1pprM(2.0)
1arp(1.9)
2pspA(1.95)
1fdd(1.9)
5fd1(1.9)
256bA(1.4)
1rei(2.0)
1acx(2.0)
1omd(1.85)
1brsA(2.0)
1cewi(2.0)
1hulA(2.4)
4cpv(1.5)
2trxA(1.68)
1jpc(2.0)
1ccr(1.5)
1tfg(1.95)
2mcm(1.5)
1dynA(2.2)
2mhr(1.7)
2rhe(1.6)
2rspA(2.0)
1rmd(2.1)
1sreA(1.78)
2cy3(1.7)
1hfh
1paz(1.55)
2gmfA(2.4)
1ppa(2.0)
2pf1(2.9)
1alc(1.7)
1pp2R(2.5)
1bpq(1.8)
1bp2(1.7)
1bsrA(1.9)
2madL(2.25)
4fgf(1.6)
3rn3(1.45)
1otgA(2.1)
1trb(2.0)
2acq(1.76)
1tadA(1.7)
4tln(2.3)
1tca(1.55)
1npc(2.0)
1pbp(1.9)
1oibA(2.4)
2pia(2.0)
2abh(1.7)
1ige
1axn(1.78)
3app(1.8)
1apte(1.8)
2apr(1.8)
1bpyA(2.2)
1bmdA(1.9)
1htrB(1.62)
1pscA(2.0)
4ape(2.1)
5ldh(2.7)
1dxy(1.9)
1prc(2.3)
4mdhA(2.5)
7api(3.0)
1ounA(2.3)
2fgf(1.77)
1azu(2.7)
1ttbA(1.7)
1rie(1.5)
2ccy(1.67)
2che(1.8)
3chy(1.66)
1cpq(1.72)
1aizA(1.8)
1rcb(2.25)
1msc(2.0)
1adl(1.6)
1icm(1.5)
1lid(1.6)
1cdmA(2.0)
1opaA(1.9)
1jacA(2.43)
1poc(2.0)
2tbd
1cof(2.3)
1eca(1.4)
1enj(1.8)
1cbs(1.8)
1jvr
1anu(2.15)
5nll(1.75)
1vhiA(2.5)
1lcl(1.8)
1stmA(1.9)
1flp(1.5)
1nhkL(1.9)
1hbiA(1.7)
2fal(1.8)
1raiD(2.5)
2hbg(1.5)
1fxl(2.0)
1aa8A(2.5)
2omf(2.4)
1xikA(1.7)
1uxy(1.8)
2liv(2.4)
1uby(2.4)
1pax(2.4)
1pedA(2.15)
1air(2.2)
2bbkH(1.75)
2mnr(1.9)
1mns(2.0)
2bltA(2.0)
1pud(1.85)
2ohxA(1.8)
2btfA(2.55)
5adh(2.9)
1kaz(1.7)
1crkA(3.0)
2sil(1.6)
2bat(2.0)
4xiaA(2.3)
1dob(2.0)
1php(1.65)
1chmA(1.9)
2mtaC(2.4)
1div(2.6)
2gdm(1.7)
1cobA(2.0)
1mbd(1.4)
1rcy(1.9)
1gds
1lh7(2.0)
1gdi(1.8)
1mniA(2.07)
2tmv(2.9)
1hfc(1.56)
1esl(2.0)
1hjrA(2.5)
1gpr(1.9)
1cyw(2.5)
1apyA(2.0)
1cpcA(1.66)
3dfr(1.7)
1klo(2.1)
1l68(1.7)
1lckA(2.5)
2cpl(1.63)
5p21(1.35)
1mhyG(2.0)
1ofv(1.7)
1wba(1.8)
1bbpA(2.0)
2fcr(1.8)
2scpA(2.0)
1gcs(2.0)
2prd(2.0)
1aoc(2.0)
1etu(2.9)
1lobA(2.0)
1ytbA(1.8)
1cau(2.3)
1csc(1.7)
1gnd(1.81)
2dkb(2.1)
1celA(1.81)
5rubB(1.7)
5rubA(1.7)
1pmi(1.7)
1ad3A(2.6)
1nhq(2.0)
1gerA(1.86)
1alkA(2.0)
1pii(2.0)
2bmhA(2.0)
1bpl(2.4)
2hpdA(2.0)
1jswA(2.8)
3grs(1.54)
1srp(2.0)
3hhrA(2.8)
2taaA(3.0)
1bmfA(2.85)
1aszA(3.0)
1btc(2.0)
8catA(2.5)
2myr
2ltnA(1.7)
1gp1A(2.0)
2stv(2.5)
1bcx(1.8)
153l(1.6)
2sas(2.4)
1fbtA(2.0)
1dsbA(2.0)
1ryt(2.1)
1etpA(2.2)
1zxq(2.2)
1vokA(2.1)
3adk(2.1)
1shcA
1ukz(1.9)
1vscA(1.9)
1bhmA(2.2)
1idsA(2.0)
2cut(1.9)
2alp(1.7)
1nox(1.59)
1ast(1.8)
2gsq(2.2)
1mngA(1.8)
1sacA(2.0)
1cfb(2.0)
1edhA(2.0)
1pipA(1.7)
1ppd(2.0)
9pap(1.65)
1lbu(1.8)
2ayh(1.6)
1dkzA(2.0)
1havA(2.0)
1nbvL(2.0)
1jud(2.5)
2brd(3.5)
1pkm(2.6)
1ddt(2.0)
1stg(1.8)
2cas(3.0)
1vnc(2.1)
1tf4A(1.9)
1gof(1.7)
1trkA(2.0)
1cdg(2.0)
1tcmA(2.2)
2tmdA(2.4)
8acn(2.0)
1alo(2.0)
1cp4(1.9)
2cstA(1.9)
1uae(1.8)
1occA(2.8)

test set
1ecd(1.4)
2fd2(1.9)
2tgi(1.8)
2pab(1.8)
1seiA(1.9)
1tgn(1.65)
1mcpH(2.7)
2ptcE(1.9)
3rp2A(1.9)
1dbs(1.8)
1lrv(2.6)
2gch(1.9)
1conA(2.0)
1lbd(2.7)
4cha(1.68)
1est(2.5)
1zymA(2.5)
1wsyA(2.2)
1nbaB(2.0)
2rslB(3.0)
4blmA(2.0)
1yasA(1.9)
1tdtA(2.2)
1mua(1.7)
4cac(2.2)
3dni(2.0)
1arb(1.2)
1btl(1.8)
1tys(1.8)
2mevl(3.0)
1nzyA(1.8)
1xjo(1.75)
2dri(1.6)
2clrA(2.0)
2sicE(1.8)
1hsaA(2.1)
1broA(2.05)
2prk(1.5)
1cnv(1.65)
2ebn(2.0)
1xsm(2.3)
1prn(1.96)
2end(1.45)
2lzm(1.7)
2gcr
1froA(2.2)
2sga(1.5)
3sgb(1.8)
4sbvA(2.8)
1ebpA(2.8)
2cna(2.0)
1timA(2.5)
1pyp(3.0)
2tbvA(2.9)
1xvaA(2.2)
1pfkA(2.4)
3gpdR(3.5)
1gotB(2.0)
3pgk(2.5)
1gal(2.3)
1kit(2.3)
1fdx
4pti(1.5)
2kauB(2.0)
5cyt(1.5)

[a] The resolution of each X-ray structure is parenthesized.

matrix was constructed for each of the turn or the nonturn region. However, there were two (b1 and b2) constructed for the boundary region since there was a difference in conditional probability for a residue being immediately preceding or following a turn sequence. Suppose there was a β-turn with sequence -X_1GREFX_2-, in which X_1 and X_2 were the residues immediately preceding or following the turn sequence. The prediction quantity $P_{q,i}$ for each residue i in the turn or not was computed as follows

$$P_{q,i} = \frac{P_{ttt,i}}{P_{nnn,i}} \tag{2}$$

where

$$P_{ttt,i} = \sum_{j=1}^{n_i} \frac{P_{tt,j}}{6} \tag{3}$$

$$P_{tt,j} = P_{b1}(G/X_1)P_t(R/G)P_t(E/R)P_t(F/E)P_{b2}(X_2/F) \tag{4}$$

Table 2. Composition of Residues of All the Simple β-Turns and Loops Identified in All the Proteins in the Database

NRT[a]	FQ[b]	F[c]	M	I	L	V	C	W	A	T	G	S	P	Y	H	Q	N	E	K	D	R
4	2599	290	143	277	517	410	131	114	758	653	1514	862	729	300	231	350	716	580	555	914	346
5	571	87	38	67	165	149	69	28	185	192	312	214	161	98	65	93	210	162	182	262	116
6	229	40	19	47	82	77	25	8	85	82	166	94	100	60	26	54	94	72	70	119	54
7	254	51	27	48	107	90	27	22	120	97	185	165	117	80	46	72	99	90	106	159	70
8	198	59	17	46	86	67	31	21	101	123	159	146	127	52	30	56	108	86	86	126	57
9	109	33	10	30	56	42	15	16	80	59	112	74	64	42	22	36	76	49	45	84	36
10	52	18	8	17	33	24	10	8	35	28	59	36	37	16	14	19	44	28	31	34	21
11	41	22	4	13	23	22	8	3	23	35	58	31	31	15	18	23	23	18	18	46	17
12	34	17	3	6	23	22	20	6	20	28	50	44	30	11	8	12	23	23	21	27	14
13	12	2	3	9	8	8	4	4	9	17	13	9	18	1	2	7	8	11	6	11	6
14	10	12	1	4	3	8	1	0	7	14	12	15	7	7	2	9	11	5	8	9	5
15	8	3	1	3	8	3	1	6	16	6	16	6	10	4	3	4	6	3	6	11	4
16	4	3	1	2	6	3	1	2	3	6	5	4	3	2	0	2	6	2	2	6	5
17	2	0	2	1	4	0	1	0	3	3	4	1	3	1	1	1	3	3	1	1	1
18	1	2	0	0	4	1	0	0	0	2	2	0	1	0	3	0	2	1	0	0	0
19	2	1	0	2	2	0	2	1	3	3	2	6	3	2	3	1	2	1	1	2	3
20	1	2	0	0	3	1	0	0	0	1	3	1	3	0	1	2	0	2	1	0	2

[a] Number of residues in a turn (a simple β-turn containing four residues) or a loop. [b] Frequency of the simple β-turns or the loops identified in the database. [c] Amino acid residues are arranged in order of a hydophobicity scale given by Engleman et al.[32]

and

$$P_{nnn,i} = \sum_{j=1}^{n_i} \frac{P_{nn,j}}{6} \tag{5}$$

$$P_{nn,j} = P_n(G/X_1)P_n(R/G)P_n(E/R)P_n(F/E)P_n(X_2/F) \tag{6}$$

where $P_t(\bullet)$, $P_{b1}(\bullet)$, $P_{b2}(\bullet)$, and $P_n(\bullet)$ were the transition probabilities indexed from the transition probability matrices constructed for the turn, the boundary (b1 and b2), and the nonturn regions, respectively, according to the given dipeptide sequence in each parenthesis. $P_{tt,j}$ and $P_{nn,j}$ were the product of these transition probabilities for the hexapeptide j being a turn or a nonturn one. Each of these two quantities was averaged over each residue in the hexapeptide j. Since every hexapeptide in the sequence of a protein was scanned from N- to C-terminal, each residue i in the protein would engage n_i hexapeptides scanned. Therefore, $P_{ttt,i}$ and $P_{nnn,i}$ were computed by summing the averaged transition probabilities of all the hexapeptides involving residue i. Finally, $P_{ttt,i}$ was weighted with $P_{nnn,i}$ to give the prediction quantity $P_{q,i}$ for residue i. The prediction quantity was computed for each residue i scanned from N- to C-terminal for each protein. Residues that were identified as β-turns were labeled as "2" otherwise they were labeled as "1" residues. Prediction quantities computed for all the two residues were collected and sorted in order of magnitude for each protein. We employed the Matthews correlation coefficient[29] C_M defined as follows for assessing the prediction accuracy for each protein

$$C_M = \frac{r\bar{r} - t\bar{t}}{\sqrt{(r + t)(r + \bar{t})(\bar{r} + t)(\bar{r} + \bar{t})}} \tag{7}$$

where r is the number of correctly predicted residues of 2s, $\bar{r}$ is the number of correctly predicted residues of 1s, t is the number of residues of 2s incorrectly predicted, and $\bar{t}$ is the number of residues of 1s incorrectly predicted. Each of the sorted prediction quantities was treated as a cutoff for computation of C_M. An all 2 tetrapeptide was considered as a peak and correctly predicted if the prediction quantity of any of its two central residues computed was greater than a cutoff. However, for an all 1 tetrapeptide to be considered as an incorrectly predicted β-turn, all of its four prediction quantities computed must be greater than a cutoff. The sorted prediction quantity that gave rise to the largest C_M computed was used as the cutoff to count the percent of correct prediction for each protein. For a protein of unknown structure, the prediction quantities computed were sorted, and each sorted prediction quantity was treated as a cutoff for finding the maximum number of peaks existing in the prediction quantities. The definition of a peak or correct prediction used was similar to that described above for the case of proteins of known structures. The entire prediction scheme was written in FORTRAN77 and implemented on a Silicon Graphics O_2 computer with a R12000 processor. The coil or turn regions of several proteins selected from the test set were also predicted using the secondary structure prediction methods of Qian-Sejnowski (QS),[33] Garnier-Osguthorpe-Robson (GOR),[34] and Maxfield-Scheraga (MS)[35] implemented in the SYBYL 6.6 program.[31]

RESULTS

The total number of turns including simple β-turns and loops containing multiple turns identified for all the 338 proteins of the database (Table 1) is 4127. The largest loop containing multiple turns identified is in protein 1trkA for residues 89−108. The sequence of the loop consists of at least six simple β-turns, namely, I (89, 90), II (91, 92), II (94, 95), VIb (99, 100), VIb (103, 104), and II (107, 108); according to the classification criteria described by Wilmot and Thornton.[16] The number of residues involved in forming a multiple turn is varying from 5 to 20 for the database studied (Table 2). The frequency of occurrence of these multiple turns is decreasing when the complexity of them is increasing. The composition of residues of all the turns are also listed in Table 2 in an order of decreasing hydrophobicity as according to the hydrophobicity scale given by Engleman et al.[32] Table 2 also shows the composition of residues of all the simple β-turns which makes about 67% of the total turns identified. Apparently, the frequency of residues in decreasing order in the turns are Gly, Asp, Ser, Ala, Pro, Asn, Thr, Lys, Glu, Leu, Val, Arg, Gln, Tyr, Phe, Ile, His, Cys, Met, and Trp. There are also some differences

Table 3. Analysis of β-turn Types for Proteins of the Database Containing 10-Residue Loops

protein	residues	I	I’	II	II’	IV	VIa	VIb	VIII
1ad3A	235−244		*					*	
1aoc	8−17		*	*					
1aszA	209−218						*		**
1celA	263−272	**		**			*		
1cnv	50−59								
1cnv	130−139			*			*	*	
1crkA	319−328	*							*
1dbs	119−128	**					*		
1dorA	127−136			*				*	
1eca	35−44	**		*				*	
1ecrA	77−86	*					*		*
1edhA	39−48			*					*
1esl	94−103			**			*		
1est	210−219	*		*					*
1etpA	156−165	*							*
1flp	41−50	**	*						
1gds	86−95								
1glg	195−204	***							
1gnd	308−317			*			*	*	
1idsA	11−20	**			*				*
1jswA	84−93								
1lbd	153−162	*		*					*
1mbd	40−49	***							
1nar	22−31	*						*	*
1onc	21−30	**	*				*		
1pipA	2−11	*						*	*
1pkm	84−93	**					*		
1prc	56−65	**						*	
1prc	136−145		*				*	*	
1trkA	49−58	**						*	
1vnc	173−182	**	*				*		
2cas	326−335	*							*
2cstA	121−130		*				*		
2cy3	48−57	**	*						
2cy3	74−83	**					*		
2dln	145−154	*	*	*					
2fcr	57−66	*	*				*		
2mtaC	37−46	*		*			*		
2omf	25−34	**			*				
2ora	59−68	**							*
2prk	79−88	**							*
2tbd	120−129								
2tmdA	64−73	*					*	*	
2tmdA	694−703	*							**
2tmdA	712−721			**			*		
5rubA	2−11	*						*	
5rubA	372−381	**						*	
5rubB	2−11	**						*	
8acn	43−52	**					*		
8catA	163−172	**					*		
8catA	289−298						*	*	
9pap	2−11	*						*	*

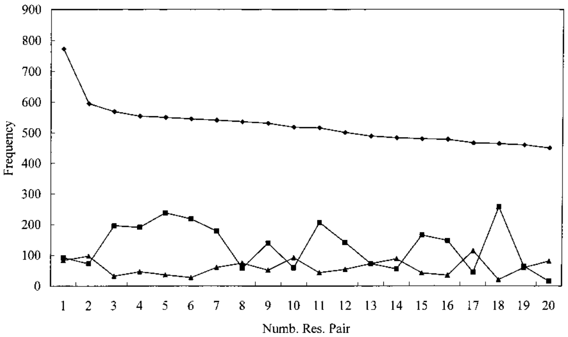


Figure 1. The sorted frequencies of the 20 most frequent dipeptides identified for all the proteins in the database. The dipeptides in the order of frequency are Ala-Ala, Ala-Leu, Ala-Gly, Gly-Ala, Ser-Gly, Gly-Gly, Gly-Val, Leu-Leu, Leu-Gly, Val-Ala, Gly-Ser, Gly-Leu, Leu-Ala, Ala-Val, Gly-Thr, Thr-Gly, Val-Leu, Asp-Gly, Leu-Lys, and Val-Val. The corresponding frequencies of these dipeptides in the turn, and the nonturn regions of the protein database are also presented. The diamonds represent frequencies of the dipeptides in all the proteins, the squares represent frequencies of the dipeptides in the turn region, and the triangles represent frequencies of the dipeptides in the nonturn region.

in frequency counted for residues in the simple β-turns or in the loops containing multiple turns (Table 2). The major difference is in those counted for residue Ala, Pro, Asn, Thr, Glu, Lys, Leu, Val, Gln, and Arg. The comparison reveals that there are no stringent requirements for sequences either being simple β-turns or loops. We do not intend to separate residues assigned in a loop to different individual β-turns since all the residues in the multiple turns of the loop are consecutive and are fulfilling the same turn criteria. The composition of different types of β-turns in these loops is complicated. A detail analysis on types of β-turns appearing on a loop containing 10 residues or seven tetrapeptides is presented in Table 3. The frequency of occurrence of these loops is 52 (Table 2). The number of turns appearing on these loops varies from 0 to 4 (Table 3). The order of abundance of each type of β-turns on these loops is I (58), VIa (20), II (19), VIb (16), VIII (17), I′ (7), II′ (2), and IV (0). This would agree with those analyzed by Wilmot and Thornton[16] for 59 proteins in which they show that the most abundant types of simple β-turns are I, II, and VIII. Table 3 also shows that type I or VIII or VIa β-turns are likely to associate with a type I turn to form a loop; while type II is more likely to engage with type VIa or II, and type I′ is likely to join with either type VIb or VIa. There is no apparent type assigned for loops (86−95) and (120−129) of protein 1gds and 2tbd, while there are four assigned for loops (35−44), (21−30) and (173−182) of protein 1eca, 1onc, and 1vnc, respectively (Table 3). Residue Arg, Ala, Phe, Leu, Thr, Ser, and Glu are counted with high positional frequency as being central residues of type I β-turn among these loops (Table 3). This is slightly different from that observed by Wilmot and Thornton[16] for simple β-turns of 59 proteins in which the highly frequent residues counted are Asp, Pro, Ser, Glu, Thr, Arg, and Asn.

The total number of dipeptides from N- to C-terminal of all the proteins in the database is 400 since there are 20 standard amino acid residues. The total frequency of each dipeptide appears in the database is 82 119, while that appears in the turn region is 19 698. Residues immediately preceding or following a turn region are treated as the boundary region. Therefore each turn region is capped with two boundary dipeptides with one at the N- and the other at the C-terminal. It appears that Gly is the most frequent residue found in the turn region since each of the 13 most frequent dipeptides identified in the turn region containing a Gly, and they are Asp-Gly (257), Ser-Gly (238), Gly-Gly (218), Pro-Gly (209), Gly-Ser (207), Ala-Gly (196), Asn-Gly (195), Gly-Ala (191), Gly-Asp (181), Gly-Val (178), Gly-Thr (165), Gly-Lys (162), and Glu-Gly (160), respectively. A comparison of frequency of the 20 most frequent dipeptides on sequences of all the proteins in the database and on those of the corresponding turn and the nonturn regions is presented in Figure 1. The plot reveals the preference of an abundant dipeptide being in the turn or the nonturn region. For example, there is a strong preference for both the Asp-Gly and Ser-Gly dipeptides being in the turn region since their appearances in the region are more

Table 4. Transition Probability Matrix of the Turn or Loop Regions

	F	M	I	L	V	C	W	A	T	G	S	P	Y	H	Q	N	E	K	D	R
F	0.030	0.059	0.029	0.050	0.029	0.054	0.031	0.049	0.036	0.082	0.069	0.105	0.029	0.049	0.045	0.047	0.049	0.055	0.041	0.059
M	0.049	0.009	0.024	0.043	0.036	0.030	0.035	0.029	0.057	0.100	0.081	0.078	0.038	0.057	0.063	0.058	0.049	0.055	0.064	0.045
I	0.039	0.025	0.031	0.016	0.034	0.043	0.060	0.026	0.043	0.095	0.049	0.119	0.028	0.062	0.046	0.080	0.038	0.049	0.074	0.046
L	0.046	0.031	0.020	0.035	0.024	0.069	0.055	0.048	0.049	0.086	0.055	0.077	0.046	0.047	0.038	0.070	0.062	0.046	0.054	0.043
V	0.043	0.039	0.022	0.037	0.013	0.080	0.027	0.044	0.059	0.085	0.054	0.067	0.037	0.050	0.033	0.090	0.055	0.058	0.082	0.027
C	0.020	0.052	0.020	0.055	0.034	0.009	0.015	0.069	0.079	0.057	0.059	0.104	0.033	0.046	0.065	0.055	0.046	0.066	0.053	0.062
W	0.024	0.064	0.020	0.046	0.008	0.079	0.014	0.021	0.044	0.075	0.092	0.097	0.039	0.011	0.059	0.101	0.016	0.052	0.056	0.082
A	0.037	0.046	0.015	0.033	0.031	0.030	0.039	0.033	0.058	0.095	0.065	0.074	0.048	0.023	0.057	0.083	0.064	0.040	0.086	0.043
T	0.043	0.022	0.020	0.036	0.035	0.027	0.031	0.043	0.055	0.066	0.065	0.058	0.039	0.057	0.072	0.084	0.049	0.067	0.081	0.052
G	0.039	0.042	0.028	0.041	0.048	0.050	0.047	0.050	0.050	0.058	0.059	0.046	0.042	0.046	0.065	0.054	0.060	0.056	0.066	0.052
S	0.032	0.033	0.021	0.026	0.032	0.049	0.026	0.035	0.065	0.071	0.064	0.065	0.043	0.054	0.053	0.073	0.063	0.061	0.070	0.064
P	0.050	0.014	0.023	0.031	0.026	0.034	0.041	0.054	0.057	0.080	0.061	0.050	0.055	0.065	0.053	0.084	0.053	0.049	0.074	0.050
Y	0.046	0.041	0.013	0.024	0.030	0.049	0.056	0.043	0.042	0.094	0.057	0.101	0.044	0.050	0.046	0.064	0.038	0.060	0.037	0.067
H	0.032	0.031	0.048	0.039	0.037	0.029	0.033	0.046	0.044	0.071	0.066	0.132	0.048	0.015	0.054	0.051	0.054	0.056	0.061	0.054
Q	0.038	0.040	0.021	0.035	0.037	0.036	0.038	0.039	0.075	0.093	0.059	0.079	0.031	0.047	0.049	0.098	0.047	0.046	0.066	0.027
N	0.030	0.044	0.033	0.038	0.032	0.032	0.039	0.048	0.052	0.091	0.068	0.080	0.045	0.043	0.046	0.057	0.058	0.046	0.061	0.057
E	0.018	0.017	0.023	0.031	0.032	0.029	0.064	0.038	0.050	0.094	0.068	0.069	0.056	0.046	0.036	0.087	0.053	0.063	0.087	0.041
K	0.032	0.013	0.018	0.034	0.035	0.065	0.025	0.035	0.046	0.092	0.057	0.072	0.052	0.094	0.038	0.085	0.037	0.052	0.080	0.038
D	0.031	0.038	0.027	0.032	0.031	0.045	0.042	0.040	0.056	0.075	0.067	0.071	0.040	0.047	0.057	0.077	0.055	0.058	0.066	0.044
R	0.043	0.047	0.019	0.031	0.038	0.062	0.007	0.031	0.061	0.079	0.054	0.067.	0.051	0.058	0.055	0.077	0.059	0.042	0.068	0.051

frequent than those in the nonturn one. The frequency of each dipeptide in the boundary region can be calculated by deduction of the frequency of the dipeptide in the turn region from the total frequency. However, residues at the N- or C-terminal of a turn region are repeatedly counted as being a turn or a boundary dipeptide. The transition probability matrices of the turn, the boundary, and the nonturn region for the residues arranged in order of a hydrophobic scale given by Engleman et al.[32] are presented in Tables 4−6, respectively. The three largest transition probabilities computed among all the transition probabilities of the turn region (Table 4) are for dipeptide Met (column)-His (row) (0.132), Met-Ile (0.119), and Cys-Trp (0.101), respectively. Note that these dipeptides are absent from Figure 1 in which frequencies of the 20 most frequent dipeptides of the database are compared. This is because each transition probability is computed from weighted rather unweighted frequencies. There are two transition matrices constructed for residues either immediately preceding (b1) or following (b2) a turn region (Table 5 (parts a and b). Transition probabilities computed for dipeptide Cys-Ala, Cys-Thr, Cys-Gly, Cys-Asn, Asn-Gly, Thr-Gly, His-Gly, Trp-Thr, Trp-Gly, Trp-Pro, Trp-Lys, and Trp-Asp in the b1 matrix and those computed for dipeptide Lys-Cys, Ala-Cys, Gly-Cys, Ser-Cys, Pro-Cys, Glu-Cys, Arg-Cys, Gly-Phe, Gly-Met, and Thr-Trp in the b2 matrix are zero. Most of these dipeptides involve a Cys or Trp residue implies that the probability for a turn sequence being initiated or terminated by these twos is rather rare. Transition probabilities in the transition matrix of the nonturn region (Table 6) are contrasting somewhat with those in the transition matrix of the turn one (Table 6). For example, the smallest transition probabilities in the transition matrix of the nonturn region computed are for dipeptide Asp-Asp (0.013), Gly-Ala (0.013), and Gly-Asp (0.013), while the corresponding values in the turn transition matrix (Table 6) computed for them are 0.066, 0.095, and 0.075, respectively.

The window used for scanning a protein from N- to C-terminal for prediction of β-turns is a hexapeptide in which the dipeptides at both ends are treated as boundaries while those between are treated as turn. The probability for a hexapeptide being a β-turn is computed as a product of the transition probabilities of the boundary dipeptide at the N-terminal, the three dipeptides in the central, and the other boundary dipeptide at the C-terminal and they are indexed from the b1 (Table 5a), the turn (Table 4), and the b2 matrix (Table 5b), respectively. By assuming all the five dipeptides are from the nonturn region, a product of the five nonturn transition probabilities indexed from the nonturn transition probability matrix (Table 6) is also computed. Both of these twos are averaged with 6 since there are six residues in each hexapeptide. Since each protein is scanned by a hexapeptide from N- to C-terminal, each residue in the protein will engage in a certain number of hexapeptides. Both of the averaged probabilities are summed for the number of hexapeptides involving the residue. The final prediction quantity for each residue is computed by weighting the summed averaged probability for the residue being in the turn region with that being in the nonturn one. The prediction quantity computed, the protein sequence, the assignment of secondary structures by the method of Kabsch and Sander,[28] and the label for β-turn regions defined by the criteria of Wilmot and Thornton[16] for protein 1ebpA of the test set are presented in Figure 2. The magnitude of the prediction quantity has been scaled into a range of 1−20 in the plot. There are some regions for which the prediction quantity computed are rising, and these are recognized as the suspicious simple β-turns or loops containing multiple turns (Figure 2). The percent of correct prediction and the Matthews correlation coefficient C_M computed for the protein is 78 and 0.76, respectively. Apparently, most of the predicted β-turns agree with those assigned by the Kabsch and Sander[28] method or identified by the Wilmot and Thornton[16] criteria except those for residues 3−4, 7, 14, 22−26, 54−55, 78−82, 113−116, and 198 which are incorrectly predicted (Figure 2). The cutoff used in determining the percent of correct prediction varies from protein to protein since the percent of correct prediction is determined using the largest C_M computed. The percent of correct prediction and the corresponding Matthews correlation coefficients computed and sorted in an order of the number of residue of all the database proteins and of all the test set proteins are presented in Figure 3a,b, respectively. Both of these plots show that the percent of correct prediction and the corresponding Matthews correlation coefficients computed decrease slightly with the increase of the number of residues in the proteins. Protein 1ppt, 1crn, 1ovoA, 1ptx, 1ctf, 1kveB, 1poh, 1frd, 1erw, 1acx, 1hulA, 2mhr, 1ppa,

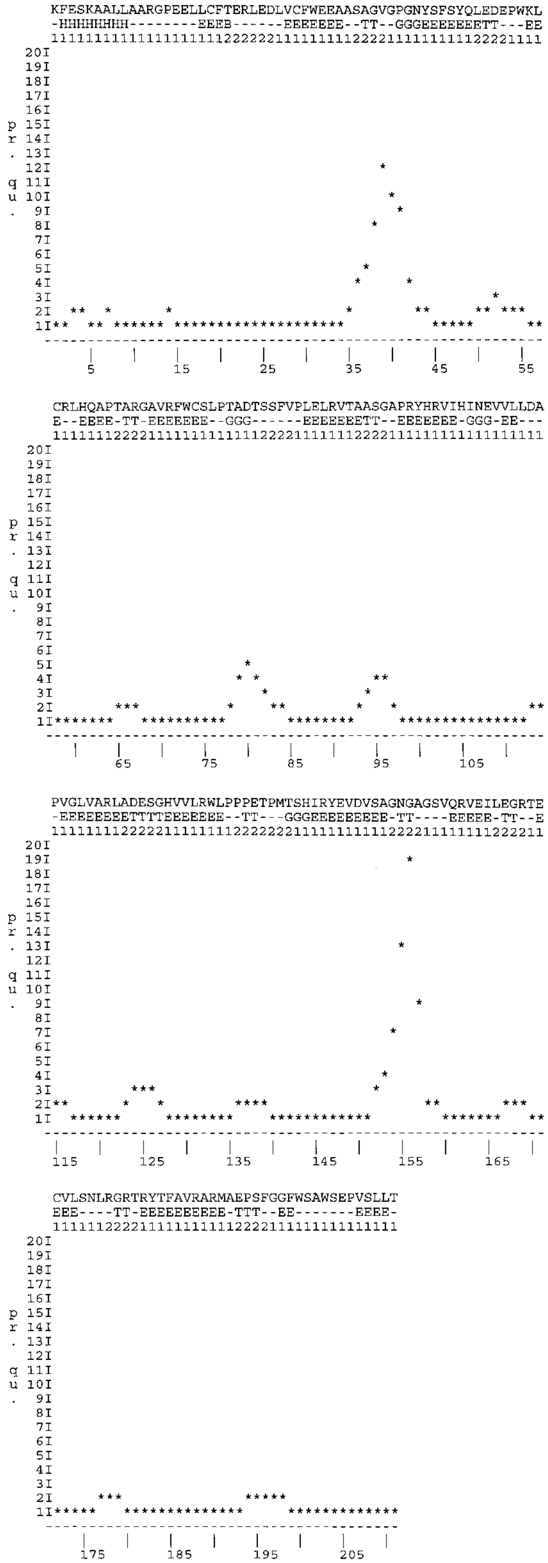
Table 5. Transition Probability Matrices b1 and b2 of the Boundary Regions

	F	M	I	L	V	C	W	A	T	G	S	P	Y	H	Q	N	E	K	D	R
a. Transition Probability Matrix b1																				
F	0.038	0.047	0.041	0.062	0.042	0.049	0.009	0.054	0.056	0.079	0.058	0.049	0.054	0.032	0.053	0.053	0.063	0.045	0.065	0.051
M	0.048	0.046	0.043	0.057	0.056	0.029	0.008	0.065	0.064	0.064	0.055	0.063	0.034	0.065	0.042	0.075	0.049	0.043	0.042	0.051
I	0.066	0.035	0.047	0.060	0.049	0.022	0.024	0.066	0.050	0.059	0.062	0.054	0.056	0.027	0.046	0.045	0.068	0.062	0.057	0.046
L	0.046	0.045	0.037	0.037	0.055	0.013	0.043	0.062	0.055	0.061	0.064	0.060	0.069	0.043	0.050	0.044	0.075	0.033	0.055	0.054
V	0.044	0.020	0.047	0.078	0.037	0.024	0.022	0.060	0.053	0.062	0.066	0.070	0.058	0.030	0.040	0.064	0.057	0.040	0.072	0.056
C	0.071	0.027	0.041	0.062	0.049	0.038	0.060	0.054	0.063	0.058	0.038	0.036	0.038	0.047	0.061	0.040	0.055	0.052	0.073	0.039
W	0.043	0.034	0.046	0.058	0.063	0.054	0.031	0.086	0.062	0.055	0.050	0.064	0.034	0.037	0.038	0.039	0.064	0.062	0.056	0.025
A	0.066	0.019	0.072	0.065	0.080	0.000	0.033	0.055	0.059	0.055	0.044	0.070	0.046	0.061	0.038	0.062	0.053	0.041	0.067	0.016
T	0.070	0.019	0.052	0.095	0.045	0.000	0.000	0.090	0.065	0.068	0.045	0.089	0.039	0.048	0.025	0.047	0.030	0.024	0.082	0.065
G	0.047	0.054	0.019	0.030	0.041	0.000	0.000	0.134	0.000	0.149	0.086	0.116	0.026	0.000	0.058	0.000	0.066	0.071	0.075	0.028
S	0.057	0.031	0.068	0.067	0.112	0.017	0.037	0.076	0.047	0.057	0.034	0.070	0.030	0.016	0.024	0.061	0.032	0.049	0.083	0.034
P	0.042	0.046	0.054	0.078	0.034	0.033	0.000	0.069	0.037	0.059	0.072	0.101	0.028	0.065	0.012	0.020	0.079	0.016	0.093	0.063
Y	0.040	0.042	0.043	0.054	0.058	0.017	0.014	0.068	0.073	0.058	0.044	0.061	0.049	0.034	0.049	0.066	0.054	0.057	0.065	0.052
H	0.070	0.043	0.046	0.072	0.058	0.016	0.037	0.081	0.068	0.042	0.048	0.032	0.026	0.033	0.051	0.061	0.051	0.056	0.056	0.054
Q	0.048	0.032	0.032	0.061	0.066	0.034	0.012	0.088	0.033	0.059	0.059	0.051	0.064	0.026	0.059	0.050	0.071	0.051	0.041	0.063
N	0.055	0.000	0.026	0.050	0.083	0.000	0.049	0.059	0.065	0.067	0.075	0.059	0.022	0.043	0.089	0.067	0.019	0.050	0.066	0.056
E	0.073	0.033	0.030	0.067	0.037	0.035	0.015	0.077	0.058	0.071	0.083	0.074	0.021	0.023	0.034	0.054	0.067	0.067	0.054	0.029
K	0.095	0.000	0.015	0.083	0.041	0.095	0.000	0.064	0.036	0.056	0.070	0.042	0.059	0.064	0.080	0.022	0.032	0.032	0.093	0.022
D	0.034	0.091	0.049	0.069	0.056	0.069	0.000	0.060	0.035	0.039	0.033	0.066	0.049	0.071	0.054	0.047	0.044	0.047	0.057	0.031
R	0.056	0.064	0.036	0.056	0.063	0.011	0.015	0.067	0.053	0.085	0.054	0.046	0.061	0.053	0.031	0.034	0.055	0.052	0.058	0.052
b. Transition Probability Matrix b2																				
F	0.070	0.124	0.053	0.006	0.032	0.069	0.177	0.018	0.014	0.000	0.025	0.022	0.084	0.138	0.035	0.027	0.020	0.038	0.022	0.027
M	0.054	0.095	0.151	0.038	0.022	0.099	0.190	0.015	0.017	0.000	0.000	0.020	0.152	0.042	0.038	0.000	0.033	0.012	0.007	0.011
I	0.078	0.221	0.094	0.021	0.047	0.100	0.126	0.016	0.018	0.002	0.017	0.023	0.044	0.065	0.024	0.010	0.017	0.021	0.011	0.046
L	0.080	0.136	0.086	0.019	0.023	0.100	0.146	0.016	0.022	0.004	0.014	0.021	0.077	0.094	0.039	0.012	0.021	0.019	0.020	0.050
V	0.072	0.149	0.074	0.028	0.032	0.109	0.132	0.016	0.015	0.002	0.018	0.014	0.069	0.078	0.060	0.014	0.029	0.023	0.009	0.058
C	0.024	0.215	0.048	0.000	0.065	0.133	0.277	0.000	0.006	0.000	0.000	0.000	0.078	0.068	0.043	0.009	0.000	0.027	0.007	0.000
W	0.112	0.118	0.102	0.034	0.065	0.055	0.118	0.025	0.000	0.009	0.021	0.019	0.087	0.032	0.060	0.016	0.011	0.022	0.021	0.072
A	0.070	0.164	0.089	0.033	0.027	0.073	0.141	0.014	0.015	0.004	0.011	0.020	0.069	0.104	0.051	0.016	0.022	0.018	0.007	0.053
T	0.067	0.152	0.072	0.031	0.027	0.104	0.111	0.024	0.026	0.004	0.022	0.019	0.071	0.099	0.040	0.014	0.028	0.029	0.017	0.044
G	0.073	0.119	0.104	0.031	0.034	0.075	0.162	0.023	0.028	0.008	0.019	0.023	0.083	0.067	0.035	0.015	0.027	0.014	0.008	0.055
S	0.067	0.169	0.095	0.021	0.025	0.055	0.171	0.028	0.026	0.003	0.015	0.016	0.075	0.064	0.050	0.010	0.031	0.016	0.024	0.037
P	0.070	0.142	0.070	0.027	0.030	0.086	0.144	0.020	0.030	0.002	0.017	0.020	0.075	0.072	0.057	0.016	0.021	0.023	0.015	0.062
Y	0.103	0.129	0.101	0.023	0.039	0.121	0.151	0.013	0.012	0.004	0.015	0.006	0.098	0.060	0.050	0.009	0.015	0.028	0.005	0.019
H	0.076	0.140	0.079	0.022	0.031	0.089	0.119	0.024	0.012	0.004	0.006	0.005	0.092	0.118	0.022	0.009	0.014	0.043	0.029	0.066
Q	0.059	0.104	0.098	0.022	0.038	0.154	0.144	0.008	0.009	0.002	0.009	0.013	0.042	0.108	0.045	0.013	0.025	0.021	0.023	0.066
N	0.080	0.139	0.061	0.022	0.036	0.144	0.105	0.029	0.023	0.003	0.013	0.023	0.069	0.088	0.048	0.013	0.026	0.016	0.018	0.046
E	0.077	0.139	0.100	0.017	0.039	0.114	0.159	0.018	0.022	0.001	0.009	0.018	0.051	0.066	0.064	0.006	0.025	0.022	0.003	0.049
K	0.055	0.179	0.072	0.018	0.027	0.073	0.187	0.014	0.025	0.001	0.012	0.016	0.077	0.041	0.063	0.027	0.023	0.016	0.011	0.062
D	0.082	0.114	0.094	0.033	0.036	0.094	0.140	0.015	0.019	0.006	0.021	0.011	0.060	0.108	0.049	0.011	0.019	0.033	0.005	0.049
R	0.069	0.097	0.067	0.034	0.028	0.104	0.183	0.034	0.009	0.002	0.017	0.017	0.066	0.092	0.045	0.010	0.023	0.025	0.007	0.071

Table 6. Transition Probability Matrix of the Nonturn Regions

	F	M	I	L	V	C	W	A	T	G	S	P	Y	H	Q	N	E	K	D	R
F	0.058	0.082	0.050	0.038	0.041	0.072	0.068	0.038	0.042	0.043	0.039	0.036	0.064	0.064	0.049	0.044	0.048	0.039	0.045	0.039
M	0.055	0.064	0.062	0.051	0.047	0.054	0.067	0.049	0.047	0.043	0.038	0.041	0.059	0.068	0.039	0.056	0.042	0.033	0.033	0.051
I	0.065	0.089	0.061	0.044	0.047	0.059	0.068	0.045	0.037	0.035	0.043	0.039	0.064	0.047	0.042	0.033	0.049	0.048	0.033	0.051
L	0.069	0.103	0.068	0.031	0.038	0.064	0.098	0.033	0.036	0.021	0.037	0.031	0.075	0.073	0.045	0.025	0.044	0.029	0.028	0.054
V	0.062	0.105	0.064	0.049	0.036	0.069	0.093	0.036	0.029	0.021	0.033	0.035	0.066	0.061	0.052	0.032	0.040	0.029	0.025	0.062
C	0.053	0.092	0.057	0.051	0.052	0.072	0.104	0.033	0.043	0.036	0.029	0.026	0.050	0.061	0.052	0.028	0.041	0.042	0.044	0.037
W	0.061	0.061	0.048	0.050	0.056	0.057	0.071	0.064	0.052	0.041	0.040	0.052	0.037	0.045	0.047	0.034	0.055	0.049	0.044	0.039
A	0.065	0.127	0.081	0.038	0.042	0.060	0.099	0.025	0.036	0.013	0.023	0.030	0.065	0.093	0.049	0.023	0.032	0.023	0.025	0.053
T	0.068	0.118	0.074	0.041	0.040	0.067	0.094	0.043	0.036	0.019	0.028	0.034	0.068	0.076	0.039	0.025	0.029	0.025	0.033	0.044
G	0.067	0.110	0.098	0.035	0.036	0.069	0.140	0.027	0.029	0.016	0.027	0.032	0.075	0.063	0.038	0.019	0.026	0.021	0.016	0.055
S	0.061	0.129	0.084	0.036	0.051	0.045	0.123	0.038	0.026	0.018	0.022	0.025	0.075	0.061	0.039	0.024	0.037	0.029	0.035	0.042
P	0.065	0.126	0.067	0.034	0.033	0.075	0.112	0.029	0.027	0.015	0.030	0.038	0.058	0.066	0.047	0.021	0.037	0.026	0.029	0.063
Y	0.062	0.081	0.074	0.043	0.048	0.052	0.071	0.044	0.043	0.031	0.026	0.036	0.059	0.059	0.055	0.043	0.041	0.047	0.038	0.049
H	0.076	0.076	0.057	0.047	0.047	0.053	0.069	0.055	0.047	0.026	0.031	0.022	0.049	0.071	0.048	0.039	0.046	0.041	0.042	0.060
Q	0.055	0.072	0.075	0.044	0.055	0.075	0.089	0.044	0.022	0.025	0.028	0.030	0.054	0.068	0.047	0.029	0.048	0.040	0.034	0.066
N	0.070	0.105	0.062	0.033	0.047	0.094	0.097	0.031	0.030	0.015	0.029	0.030	0.065	0.087	0.057	0.020	0.031	0.030	0.024	0.043
E	0.067	0.107	0.079	0.037	0.043	0.091	0.097	0.039	0.030	0.020	0.032	0.032	0.047	0.058	0.054	0.026	0.034	0.039	0.021	0.050
K	0.070	0.117	0.064	0.036	0.037	0.072	0.118	0.034	0.026	0.018	0.029	0.025	0.064	0.053	0.074	0.024	0.030	0.028	0.033	0.048
D	0.073	0.108	0.088	0.039	0.045	0.079	0.108	0.024	0.025	0.013	0.029	0.025	0.057	0.092	0.054	0.020	0.026	0.036	0.013	0.048
R	0.065	0.085	0.056	0.045	0.048	0.065	0.107	0.044	0.029	0.033	0.033	0.029	0.069	0.064	0.043	0.022	0.038	0.040	0.032	0.055

1bpq, 1bp2, 1otgA, 1ttbA, 2ccy, 1icm, 1opaA, 1cbs, 1nhkL, 1mbd, 1mngA, and 1jud are predicted with a 100% of correct prediction (Figure 3a). However, these are smaller proteins consisting of smaller number of residues, namely, from 36 to 220 (Figure 3a). The 28 proteins in the test set are selected randomly from the PDB.[30] By using a multiple sequence alignment program given in the GCG package,[36] it is found that the sequence homology between the 10 proteins with 100% correct prediction predicted in the data set and all the proteins studied is low (Figure 4). The best percent of correct prediction computed for all the test set proteins is also 100 and which is for protein 1fdx of 54 residues (Figure 3b).



```
     KFESKAALLAARGPEELLCFTERLEDLVCFWEEAASAGVGPGNYSFSYQLEDEPWKL
     -HHHHHHHH--------EEEB------EEEEEEE--TT--GGGEEEEEEETT---EE
     111111111111111111122222211111111122221111111111122221111
  20I
  19I
  18I
  17I
  16I
p 15I
r 14I
. 13I
  12I                                      *
q 11I
u 10I                                       *
.  9I                                        *
   8I                                     *
   7I
   6I
   5I                                    *
   4I                                   *     *
   3I                                                   *
   2I  **  *      *                    *       **     ** ***
   1I**  ** ****** ********************          *****      **
    ----------------------------------------------------------
         |    |    |    |    |    |    |    |    |    |    |
         5         15        25        35        45        55

     CRLHQAPTARGAVRFWCSLPTADTSSFVPLELRVTAASGAPRYHRVIHINEVVLLDA
     E--EEEE-TT-EEEEEEE--GGG------EEEEEEETT--EEEEEEE-GGG-EE---
     111111122221111111111112222111111122221111111111111111111
  20I
  19I
  18I
  17I
  16I
p 15I
r 14I
. 13I
  12I
q 11I
u 10I
.  9I
   8I
   7I
   6I
   5I                      *
   4I                     * *             **
   3I                        *           *
   2I       ***          *    **        *   *               **
   1I*******   **********       ********     ***************
    ----------------------------------------------------------
       |    |    |    |    |    |    |    |    |    |    |
            65        75        85        95        105

     PVGLVARLADESGHVVLRWLPPPETPMTSHIRYEVDVSAGNGAGSVQRVEILEGRTE
     -EEEEEEEETTTTEEEEEEE--TT---GGGEEEEEEEEE-TT----EEEEE-TT--E
     111111112222211111111222222211111111111222211111111222211
  20I
  19I                                         *
  18I
  17I
  16I
p 15I
r 14I
. 13I                                        *
  12I
q 11I
u 10I
.  9I                                          *
   8I
   7I                                       *
   6I
   5I
   4I                                      *
   3I         ***                         *
   2I**      *   *        ****                  **       ***
   1I  ******     ********    ************        *******   **
    ----------------------------------------------------------
     |    |    |    |    |    |    |    |    |    |    |    |
     115       125       135       145       155       165

     CVLSNLRGRTRYTFAVRARMAEPSFGGFWSAWSEPVSLLT
     EEE----TT-EEEEEEEEEE-TTT--EE-------EEEE-
     1111112222111111111111222211111111111111
  20I
  19I
  18I
  17I
  16I
p 15I
r 14I
. 13I
  12I
q 11I
u 10I
.  9I
   8I
   7I
   6I
   5I
   4I
   3I
   2I     ***              *****
   1I*****   **************     *************
    -----------------------------------------
        |    |    |    |    |    |    |    |
        175       185       195       205
```

Figure 2. Prediction of the β-turns for protein 1ebpA of 211 residues of the test set. The prediction quantity computed for each residue is scaled into a range of 1−20. The percent of correct prediction and the corresponding Matthews correlation coefficient computed are 78 and 0.76, respectively. The sequence of the protein, the secondary structures assigned by the method of Kabsch and Sander,[28] and the turn residues identified by the criteria of Wilmot and Thornton[16] are all presented in the plot. Residues identified by the Wilmot and Thornton[16] criteria as β-turns are represented with digit 2 otherwise they are represented with digit 1.

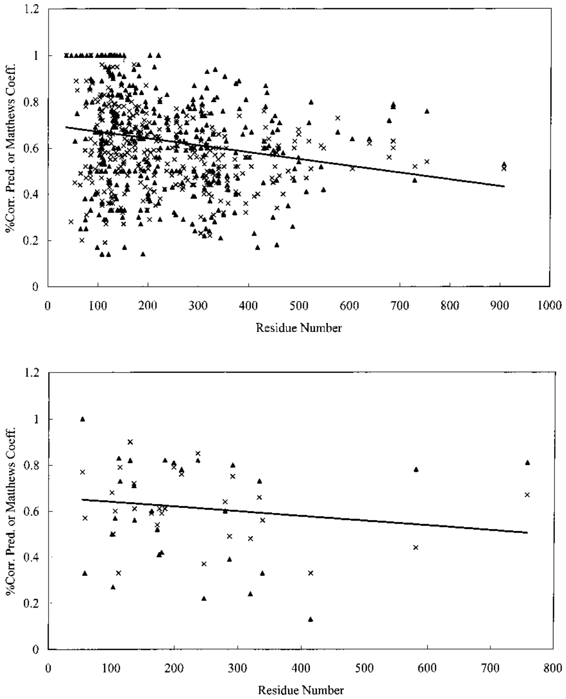


Figure 3. (a) The percent of correct prediction and the corresponding Matthews correlation coefficient computed for all the proteins of the database are plotted against the number of residues of each protein. The triangles represent the percent of correct prediction, while the crosses represent the Matthews correlation coefficient computed for each protein. (b) The percent of correct prediction and the corresponding Matthews correlation coefficient computed for all the proteins of the database are plotted against the number of residues of each protein. The triangles represent the percent of correct prediction, while the crosses represent the Matthews correlation coefficient computed for each protein.

These two plots also show that most of the marginal prediction results are confined to small or medium size proteins. For example, the percent of correct prediction computed for protein 2cab of 65 residues, 1iml of 76 residues, 2gmfA of 121 residues, 1bcx of 185 residues, 1ryt of 190 residues, 1xsm of 288 residues, 1pprM of 312 residues, 1ige of 322 residues, 2liv of 344 residues, 1uae of 418 residues, 1bp1 of 456 residues, and 1bmfA of 487 residues of the database are 25, 25, 14, 27, 14, 27, 22, 24, 21, 17, 18, and 26, respectively, and that computed for protein 5cyt of 103 residues, 1pfkA of 320 residues, and 3pgk of 415 residues of the test set is 27, 24, and 13, respectively. It is also found that 50% of all the test set proteins or 40% of all the database proteins having a percent of correct prediction greater than 70. This implies the prediction accuracy is rather unaffected by the fact that whether the predicted protein is in the database. A comparison of prediction results for 15 proteins selected from the test set by the first-order Markov models (Mk) and by several secondary structure prediction methods implemented in the SYBYL 6.6 program[31] is presented in Table 7 and in Figure 5 for proteins 1ebpA and 1seiA, respectively. The proteins in Table 7 are listed in an order of the percent of correct prediction by the Mk method. The percent of correct prediction and the corresponding Matthews correlation coefficients are computed only for the coil or turn regions predicted by these secondary structure prediction methods. Although the best percent of correct prediction is given by the QS method,[33] the corresponding Matthews correlation coefficients computed by the method are worse than those computed by the other ones (Table 7 and Figure 5). This

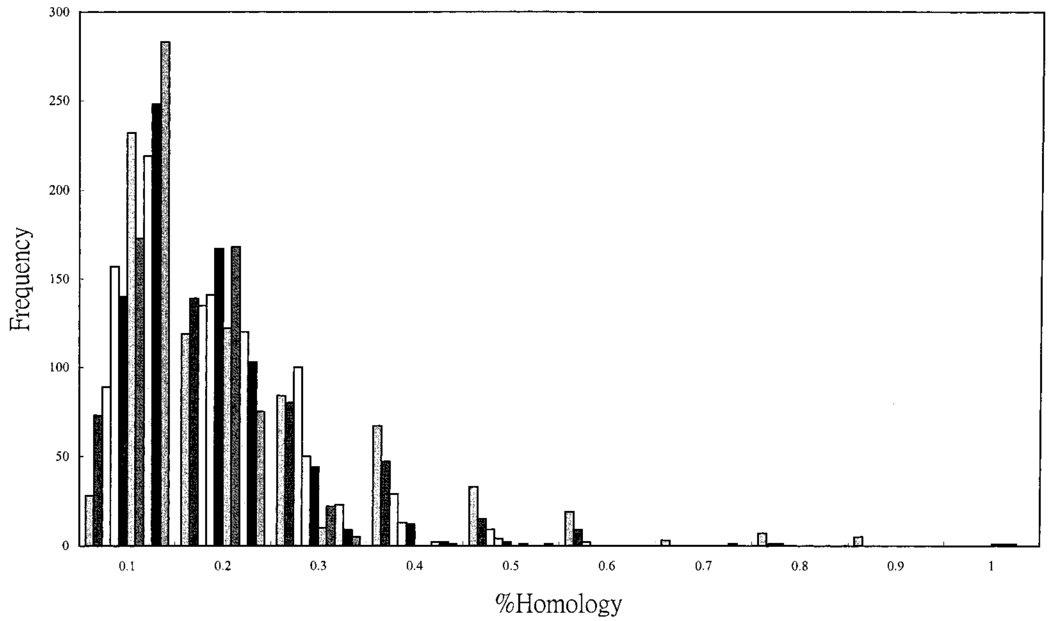


Figure 4. Distribution of percent of homology in sequences for 10 proteins with 100% correct prediction by the Mk method. A pairwise sequence alignment for each of these proteins against the total proteins (338+28) studied is performed using the GCG[36] package. The 10 proteins used in the sequence alignment are 1ppt (36), 1crn (46), 1ovoA (56), 1kveB (77), 1poh (85), 1frd (98), 1erw (105), 1acx (107), 1hulA (108), and 2mhr (114), respectively. The number of residues in each protein is parenthesized.

Table 7. Comparison of Prediction Results for the β-Turns or Loop (Coil) Regions for Several Proteins Selected from the Test Set by the First-Order Markov Models (Mk and Mk2) with Those by the Qian-Sejnowski (QS),[33] Garnier-Osguthorpe-Robson (GOR),[34] and Maxfield-Scheraga (MS)[35] Methods

		Mk		Mk2		QS		GOR		MS	
protein	NR[a]	%[b]	Matt[c]	%[b]	Matt[c]	%[b]	Matt[c]	%[b]	Matt[c]	%[b]	Matt[c]
1fdx	54	100	0.77	100	0.50	100	0.58	71	0.50	100	0.58
2tgi	112	83	0.33	79	0.26	100	0.24	100	0.45	83	0.21
1seiA	130	82	0.90	100	0.76	82	0.37	55	0.66	64	0.48
3sgb	185	82	0.61	94	0.53	94	0.58	88	0.82	94	0.70
2cna	237	82	0.85	62	0.71	96	0.58	93	0.85	93	0.82
4sbvA	199	81	0.79	89	0.54	89	0.71	74	0.85	96	0.87
1kit	757	81	0.67	90	0.58	95	0.54	86	0.72	81	0.62
1xvaA	292	80	0.75	84	0.61	87	0.49	80	0.75	83	0.54
1ebpA	211	78	0.76	100	0.66	83	0.47	72	0.70	72	0.51
1gal	581	78	0.44	87	0.63	90	0.38	71	0.53	78	0.45
2pab	114	73	0.79	82	0.79	91	0.46	55	0.50	82	0.51
3gpdR	334	73	0.66	55	0.53	96	0.67	58	0.70	83	0.74
1ecd	136	71	0.72	86	0.50	79	0.32	50	0.59	79	0.61
2lzm	164	60	0.59	68	0.42	80	0.24	80	0.50	100	0.43
1pyp	280	60	0.64	92	0.47	100	0.77	59	0.67	76	0.68
		78 ± 9[d]	0.68[e]	84 ± 3	0.57	90 ± 7	0.49	73 ± 15	0.65	84 ± 10	0.58

[a] The number of residue in each protein. [b] Percentage of correct prediction by a particular method. [c] Matthews correlation coefficient computed for the method. [d] Mean ± standard deviation computed for a particular method. [e] Mean computed for a particular method.

indicates that more residues are incorrectly predicted as coils by the QS method[33] (see Figure 5 for protein 1ebpA and 1seiA). This is also true for some proteins predicted by the GOR[34] or the MS[35] methods (Figure 5) even the percent of correct prediction obtained by the twos for these proteins are better than that by the Mk one. Apparently, for proteins of known structures, the Mk method gives prediction results in which the percent of correct prediction is better correlated with the Matthews correlation coefficients computed. For the same proteins treated as proteins of unknown structures, the percent of correct prediction, and the Matthews correlation coefficients computed by the Mk method are designated as the Mk2 prediction results and are also listed in Table 7 or presented in Figure 5 for protein 1ebpA and 1seiA, respectively. The global statistics of the percent of correct prediction of the Mk2 prediction results are similar to those given by the MS[35] or better than those by the GOR[34] but are slightly worse than those by the QS[33] method (Table 7 and Figure 5). However, the Matthews correlation coefficients of the Mk2 prediction results computed are better than those by the QS[33] method (Table 7).

DISCUSSION

The prediction scheme presented here can be used to predict simple β-turns and loops containing multiple turns as well since both of these structural elements existing in proteins. A loop region is usually composed of several simple β-turns of different turn types. Unlike most of the simple β-turns which serve to alter the direction of a protein chain nearly 180°, some of the loops are simply linkers between

```
                              1ebpA
     KFESKAALLAARGPEELLCFTERLEDLVCFWEEAASAGVGPGNYSFSYQLEDEPWKLCRLHQAPTARGAVRF
     -HHHHHHH--------EEEB------EEEEEEE--TT--GGGEEEEEEEETT---EEE--EEEE-TT-EEEE
     111111111111111111112222221111111122221111111111122221111111111122221111
Mk   111111111111111111111111111111111122222222221111122222211111111222221111
Mk2  111111111111111111111111111111111112222222221111122221111111111111111111
QS   222111111122222111111111111111112222222222222222222222111122222222222111
GOR  111111111122111111111111111111111122222222221111222121111122222222221112
MS   222221111122211111111111111111111122222222222211121111111112222222221111

     WCSLPTADTSSFVPLELRVTAASGAPRYHRVIHINEVVLLDAPVGLVARLADESGHVVLRWLPPPETPMTSH
     EEE--GGG------EEEEEEETT--EEEEEEE-GGG-EE----EEEEEEEETTTTEEEEEEE--TT---GGG
     111111111222211111112222111111111111111111111111112222211111111222222211
Mk   111112222222111111112222211111111111111122221111112222211111111222211111
Mk2  111112222222111111111222221111111111111122221111112222211111111112211111
QS   112222222222221111112222222211111111112222221111112222111122222222222222
GOR  222222222221111111111222222111111111111111111111111111211112222222222122
MS   111222222222211111111122222111111111111112111111111112111122222222222211

     IRYEVDVSAGNGAGSVQRVEILEGRTECVLSNLRGRTRYTFAVRARMAEPSFGGFWSAWSEPVSLLT
     EEEEEEEEE-TT----EEEEE-TT--EEEE----TT-EEEEEEEEEE-TTT--EE-------EEEE-
     1111111112222111111122221111111122221111111111112222111111111111111
Mk   1111111222222221111112222211111222222111111111112222221111111111111
Mk2  1111111222222221111111111111111122221111111111111222211111111111111
QS   1111111222222222111112222111222222211111111111222222222222222222222
GOR  1111111222222221111112211111111222222111111111222222222222222222211
MS   2111111222222222111112222111222222222211111112222222222122222222222


                              1seiA
     VMTDPIADMLTAIRNANMVRHEKLEVPASKIKREIAEILKREGFIRDYEYIEDNKQGILRIFLKYGPNERVI
     ---HHHHHHHHHHHHHHHTT--EEEEE--HHHHHHHHHHHHTT-EEEEEEEE----EEEEEEE-B----B--
     111111111111111111111111111111111111111111111111111222221111111222222222
Mk   111111111111111111111111111111111111111111111111111222222111111222222221
Mk2  111111111111122221111111122222111111111111111111111122221111111222222221
QS   222222111111111111111111222222211111111122222222222222222111112222222111
GOR  111111111111111111111111111111111111111112222221111111111111111222222111
MS   222222211111111111111111112111111111111121111111111111121111112222222111

     TGLKRISKPGLRVYVKAHEVPRVLNGLGIAILSTSQGVLTDKEARQKGTGGEIIAYVI
     --EEE--BTTB--EE-GGG-----TT--EEEEEETTEEEEHHHHHHHT--EEEEEEE-
     2111112222111111111111122221111122221111111111111111111111
Mk   1111222222211111111112222221112222222221111111111111111111
Mk2  1122122222221111111111122222111122222222222221222222111111
QS   1122222222211111222221122221112222221222222222222222111112
GOR  1111222222111111111111122211111222222221111112222222111111
MS   1221122222211112121211112222111122222211111222222222112222
```

Figure 5. A comparison of prediction results for β-turns of protein 1ebpA and 1seiA of the test set by the first-order Markov models (Mk and Mk2) and by the Qian-Sejnowski (QS),[33] the Garnier-Osguthorpe-Robson (GOR),[34] and the Maxfield-Scheraga (MS)[35] secondary structure prediction methods implemented in the SYBYL 6.6 program.[31] The protein sequences, the secondary structure assignments by the method of Kabsch and Sander,[28] and the turn residues identified by the criteria of Wilmot and Thornton[16] are also presented on the first, the second, and the third line, respectively. Residues identified or predicted as simple β-turns or loops (coils) are designated with digit 2, otherwise they are designated with digit 1. The Mk2 prediction results are obtained by treating each protein as a protein of unknown structure.

secondary structures. While a simple β-turn is formed by a tetrapeptide, a loop is usually composed of seven to 20 residues or even more. There are some differences between our method presented here and that published by Chou et al.[20−23] First, the window used in our method for sliding the sequence of a protein is a hexa- rather than a tetrapeptide since both the conditional probabilities of the two boundary dipeptides are counted. Second, a prediction profile can be drawn for each protein predicted by our method since the prediction quantity computed is numerically consecutive for the protein. Third, while Chou's works[20−23] are focused on the prediction of different turn types for tight turns, our method is aimed at the prediction for both simple β-turns and loops containing multiple turns in proteins. It is known that adding a loop to two flanking elements of secondary structures is the most tedious work in building a protein model. In general, if any homology of known structure has a loop of the same length in the corresponding region as the model, this fragment is a good choice, since it is likely to be subject to environmental constraints similar to those found in the model. If the loop is likely to belong to a defined class of loops, it may be preferable to select a loop of that class from a known structure. Otherwise, one has to find loop fragments whose geometry is compatible with the geometry of the conserved regions flanking the loop in the model and try to use geometric and sequence information to discriminate among candidate loop fragments. This method of using fragments from known structures to construct loop regions of a model has been described by Jones and Thirup[37] and elegantly elaborated by Claessens et al.[38]

Because sequence determines the secondary structure of segments of a protein chain, several strategies have been attempted for predicting secondary structure, but they rarely achieve more than 65% correctly predicted residue positions.[39] The prediction accuracy for secondary structures may be enhanced through analysis on the structural classes[40] (all α, all β, α/β, or $\alpha+\beta$) or alignment of protein sequences against that of a target protein of known structure.[41] While identification of β-turns on proteins can be fairly achieved using a hydrophobicity profile,[18] they are literally treated as "coils" or undefined regions by most of the secondary structure prediction methods. By assuming that formation of a β-turn depends only on the interaction between nearest neighbors, it is feasible to apply the first-order Markov models to predict the existence of it on a protein.[20−23] In the first-order Markov models, the output of a state depends on the state immediately previous; i.e., a residue is dependent on the previous residue.[42] Thus for 20 standard residues, we compute 400 transition probabilities: *P(F/F), P(F/M), P(F/I), ..., P(R/R)*. These transition probabilities are indexed from transition matrices constructed for sequences being the β-turns, the boundary, or the nonturn regions. There are only three transition probabilities needed to be indexed at a time if the window used for scanning a protein sequence from N- to C-terminal is a β-turn tetrapeptide. To enhance the prediction accuracy, these transition probabilities are capped at both ends with two transition probabilities indexed from the two boundary transition matrices. For example, the Matthews correlation coefficient computed for protein 1seiA with or without capping is 0.90 and 0.66, respectively. Since the dipeptides from the nonturn region account for 57% of the total of the database, the final prediction quantity must be weighted with that computed for the residue being in a nonturn region. The Matthews correlation coefficient computed for the unweighted prediction quantity for protein 1seiA is reduced to 0.75 and is caused by doubling of the number of incorrectly predicted residues. Since the total number of dipeptides increases linearly with the number of proteins, the prediction accuracy depends strongly on the size of a database. The number of proteins used in the database is properly adequate since the overall prediction result of the test set is slightly better than that of the database. The Matthews correlation coefficient employed will reflect the degree of incorrect prediction, and it will also change with the cutoff used to define regions of the incorrect prediction. However, the percent of correct prediction of each protein is directly related to the number of correctly predicted 2 residues. The prediction result is dictated by the sequence but not by the size of a protein since most of the bad prediction results are predicted for small rather than for large proteins. Treatment of sequences using the first-order Markov models appears to be appropriate since formation of β-turns requires only interactions between nearest neighbors.[20−23] Since no prior knowledge is added into the training sequences, the first-order Markov models we used here are simple ones. We are exploring ways to enhance the prediction accuracy of the current method by using more sophisticated priors by switching from the alphabet of the primary sequences to a different representation based on the chemical or physical properties of the amino acids in the sequences. Higher order Markov models such as the second-order ones may be also used to strengthen the performance. To achieve this, a selection mechanism based on computation of the highest posterior probability[43] must be added into the models, and the number of proteins used in the database must be also increased substantially to generate all the 8000 tripeptides.

ACKNOWLEDGMENT

This work is supported in part by a grant NSC89-2113-M007-055 from the National Science Council, ROC.

REFERENCES AND NOTES

(1) Brunet, A. P.; Huang, E. S.; Huffine, M. E.; Loeb, J. E.; Robert, J. W.; Hecht, M. H. The role of turns in the structure of an alpha-helical protein. *Nature* **1993**, *364*, 355−358.

(2) Predki, P. F.; Agrawal, V.; Brunger, A. T.; Regan, L. Amino acid substitutions in a surface turn modulate protein stability. *Nat. Struct. Biol*. **1996**, *3*, 54−58.

(3) Nagi, A. D.; Anderson, K. S.; Regan, L. Using loop length variants to dissect the folding pathway of a four-helix-bundle protein. *J. Mol. Biol*. **1999**, *286*, 257−265.

(4) Ybe, J.; Hecht, M. Sequence replacements in the central beta-turn of plastocyanin. *Protein Sci*. **1996**, *5*, 814−824.

(5) Martinez, J. C.; Pisabarro, M. T.; Serrano, L. Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol*. **1998**, *5*, 721−729.

(6) Kim, K.; Frieden, C. Turn scanning by site-directed mutagenesis: Application to the protein folding problem using the intestinal fatty acid binding protein. *Protein Sci*. **1998**, *7*, 1821−1828.

(7) Zhou, H.; Hoess, R. H.; DeGrado, W. F. In vitro evolution of thermodynamically stable turns. *Nat. Struct. Biol*. **1996**, *3*, 446−451.

(8) Gu, H. D.; Kim, D.; Baker, D. Contrasting roles for symmetrically disposed beta-turns in the folding of a small protein. *J. Mol. Biol*. **1997**, *274*, 588−596.

(9) Garrett, J. B.; Mullins, L. S.; Raushel, F. M. Are turns required for the folding of ribonuclease T1. *Protein Sci*. **1996**, *5*, 204−211.

(10) Takano, K.; Yamagata, Y.; Yutani, K. Role of amino acid residues at turns in the conformational stability and folding of human lysozyme. *Biochemistry* **2000**, *39*, 8655−8665.

(11) Hecht, M. H.; Sturtevant, J. M.; Sauer, R. T. Stabilization of λ repressor against thermal denaturation by site-directed Gly-Ala changes in α-helix 3. *Proteins* **1986**, *1*, 43−46.

(12) Nicholson, H.; Tronrud, D. E.; Becktel, W. J.; Matthews, B. W. Analysis of the effectiveness of proline substitutions and glycine replacements in increasing the stability of phage T4 lysozyme. *Biopolymers* **1992**, *32*, 1431−1441.

(13) Venkatachalam, C. M. Stereochemical criteria for polypeptides and proteins. V. conformation of a system of three linked peptide units. *Biopolymers* **1968**, *6*, 1425−1436.

(14) Lewis, P. N.; Momany, F. A.; Scheraga, H. A. Chain reversals in proteins. *Biochim. Biophys. Acta* **1973**, *303*, 211−229.

(15) Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem*. **1981**, *34*, 167−339.

(16) Wilmot, C. M.; Thornton, J. M. Analysis and prediction of the different types of β-turn in proteins. *J. Mol. Biol.* **1988**, *203*, 221−232.

(17) Chou, P. Y.; Fasman, G. D. Prediction of protein conformation. *Biochemistry* **1974**, *13*, 222−245.

(18) Kuntz, I. D. Protein folding. *J. Am. Chem. Soc*. **1972**, *94*, 4009−4012.

(19) Cohen, F. E.; Abarbanel, R. M.; Kuntz, I. D.; Fletterick, R. J. Turn prediction in proteins using a pattern-matching approach. *Biochemistry* **1986**, *25*, 266−275.

(20) Chou, K. C. Prediction of β-turns. *J. Peptide Res*. **1997**, *49*, 120−144.

(21) Zhang, C. T.; Chou, K. C. Prediction of β-turns in proteins by 1−4 and 2−3 correlation model. *Biopolymers* **1997**, *41*, 673−702.

(22) Cai, Y. D.; Yu, H.; Chou, K. C. Prediction of β-turns. *J. Protein Chem*. **1998**, *17*, 363−376.

(23) Chou, K. C. Prediction of tight turns and their types in proteins. *Anal. Biochem*. **2000**, *286*, 1−16.

(24) Krogh, A.; Brown, M.; Mian, S.; Sjölander, K.; Haussler, D. Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol*. **1994**, *235*, 1501−1531.

(25) Borodovsky, M.; McIninch, J. D.; Koonin, E. V.; Rudd, K. E.; Médigue, C.; Danchin, A. Detection of new genes in bacterial genome using Markov models for three gene classes. *Nucleic Acids Res*. **1995**, *23*, 3554−3562.

(26) Salzberg, S. L.; Delcher, A. L.; Kasif, S.; White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*. **1998**, *26*, 544−548.

(27) Audic, S.; Claverie, J. M. Self-identification of protein-coding regions in microbial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 10026−10031.

(28) Kabsch, W.; Sander, C. Dictionary of protein secondary structures: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

(29) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.

(30) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(31) SYBYL 6.6; The Tripos Associates: St. Louis, MO, 1999.

(32) Engelman, D. M.; Steitz, T. A.; Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* **1986**, *15*, 321−353.

(33) Qian, N.; Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol*. **1988**, *202*, 865−884.

(34) Garnier, J. R.; Osguthorpe, D. J.; Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **1978**, *120*, 97−120.

(35) Maxfield, F. R.; Scheraga, H. A. Improvements in the prediction of protein backbone topography by reduction of statistical errors. *Biochemistry* **1979**, *18*, 697−704.

(36) Wisconsin Package Version 10.2; Genetics Computer Group (GCG): Madison, WI.

(37) Jones, T. A.; Thirup, S. Using known substructures in protein model building and crystallography. *EMBO J.* **1986**, *5*, 819−822.

(38) Claessens, M.; Van Gutsem, E.; Lasters, I.; Wodak, S. Modeling the polypeptide backbone with “spare parts” from known protein structures. *Protein Eng*. **1989**, *2*, 335−345.

(39) Fasman, G. D. Protein conformational prediction. *Trends Biochem. Sci*. **1989**, *14*, 259−299.

(40) Levitt, M,; Chothia, C. Structural patterns in globular proteins. *Nature* **1976**, *261*, 552−558.

(41) Niermann, T.; Kirschner, K. Use of homologous sequences to improve protein secondary structure prediction. *Methodol. Enzymol.* **1991**, *202*, 45−59.

(42) Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257−285.

(43) Baldi, P. On the convergence of a clustering algorithm for protein coding regions in microbial genomes. *Bioinfomatics* **2000**, *16*, 367−371.

CI0103020