Improved Method for Prediction of Protein Backbone U-Turn Positions and Major Secondary Structural Elements Between U-Turns

Wei-Ping Hu,[1] Andrzej Kolinski,[1,2] and Jeffrey Skolnick[1]*
[1]*The Scripps Research Institute, Department of Molecular Biology, La Jolla, California*
[2]*Department of Chemistry, University of Warsaw, 02-093 Warsaw, Poland*

ABSTRACT **A new and more accurate method has been developed for predicting the backbone U-turn positions (where the chain reverses global direction) and the dominant secondary structure elements between U-turns in globular proteins. The current approach uses sequence-specific secondary structure propensities and multiple sequence information. The latter plays an important role in the enhanced success of this approach. Application to two sets (total 108) of small to medium-sized, single-domain proteins indicates that approximately 94% of the U-turn locations are correctly predicted within three residues, as are 88% of dominant secondary structure elements. These results are significantly better than our previous method (Kolinski et al., Proteins 27:290–308, 1997). The current study strongly suggests that the U-turn locations are primarily determined by local interactions. Furthermore, both global length constraints and local interactions contribute significantly to the determination of the secondary structure types between U-turns. Accurate U-turn predictions are crucial for accurate secondary structure predictions in the current method. Protein structure modeling, tertiary structure predictions, and possibly, fold recognition should benefit from the predicted structural data provided by this new method. Proteins 29:443–460, 1997.** **© 1997 Wiley-Liss, Inc.**

Key words: protein folding; turn prediction; secondary structure prediction; statistical potential; Fourier transform; local interactions

INTRODUCTION

The three-dimensional (3-D) structure of a globular protein can be viewed as being built from a series of linear, transglobular blocks connected by surface loops and turns where the backbone changes overall direction[1] (the U-turns). The blocks between these surface U-turns may consist of regular secondary structural elements, such as helices or β-strands, or irregular transglobular loops. Figure 1 uses a simple β-protein to show the above idea. One of the most important purposes of secondary structure prediction is to provide starting information for 3-D protein model building. Additionally, the positions of U-turns provide invaluable information for constructing a qualitatively correct low-resolution 3-D structure. Kolinski et al.[2] have developed a method that can predict the positions of the surface U-turns and the dominant secondary structures the protein backbone adopts in the transglobular blocks with rather high accuracy. The results obtained from this method are potentially very useful in protein model building and fold recognition and have been applied in the initial steps of a promising low-resolution protein tertiary structure prediction algorithm.[3,4]

There are, however, some practical limitations of this method. First, the model assumes that the globular proteins are spherical in shape. Although this is generally a good approximation, there are many known exceptions where the form of burial potential used in the method can be invalid. Second, this method is only applicable to relatively small, less than 100-residue, single-domain proteins. For larger proteins, it is necessary to search for other methods. Moreover, the U-turn positions predicted by the method are sometimes very diffuse, and it is difficult to objectively determine the ends of the turn regions. This might prevent accurate model building. Actually, the second limitation mentioned above reflects the inadequacy of the model in describing the structures of larger proteins. Although most secondary structure elements are still more or less linear, curved helices and strands are more common in larger proteins and, thus, the number of topological elements are less well defined. Furthermore, despite the fact that loops and turns regions are mostly exposed in larger proteins, many linear blocks are not necessarily transglobular in nature, which makes the length constraints more difficult to apply.

Grant sponsor: NIH; Grant number: GM-48835; Grant sponsor: Howard Hughes Medical Institute; Grant number: 75195-543402.

*Correspondence to: Jeffrey Skolnick, The Scripps Research Institute, Department of Molecular Biology, 10550 North Torrey Pines Road, La Jolla, CA 92037.
E-mail: skolnick@scripps.edu

Received 10 April 1997; Accepted 31 July 1997

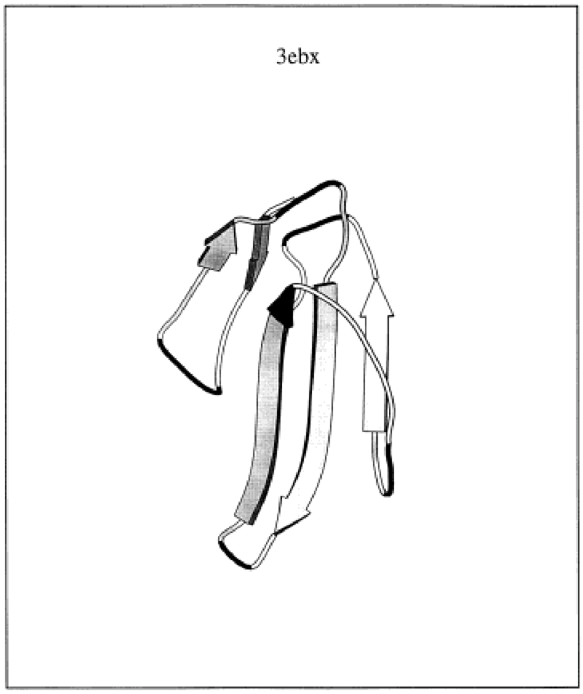

Fig. 1. Ribbon diagram of a small (62 residues) β-protein, 1 ebx (x-ray structure). This figure shows the idea that the protein structure at low resolution consists of secondary structural element blocks and U-turn regions. The black regions are the predicted U-turn positions by the current method, which correlate very well with the crystal structure. Notice that a transglobular loop crosses the main β-sheet. Even though it is not a "regular" hydrogen-bonded secondary structural element, the U-turns at both ends of this loop are correctly predicted.

In this study, we present a new method that is, on average, more accurate for small proteins than the previous method,[2] but is now also applicable to larger proteins. The new method is based on a simple statistical secondary structure propensity energy function[5] derived from experimentally determined protein structures and the use of multiple sequence information. Although the realization of the low-resolution protein model and the prediction objectives are the same, the current method differs significantly from our previous approach. First, the previous method (using only one sequence) treats the U-turns and the blocks on equal footing. It uses the sum of various energy terms to optimize the best U-turn positions and secondary structures simultaneously by fitting and orienting the structural templates. Here, U-turns are first identified by the turn propensity (defined by the statistical potential) averaged over a set of homologous sequences. Then, the dominant secondary structures of the "blocks" between U- turns are determined by the length constraints and the averaged secondary structure propensity. Here, the "dominant secondary structure" specifies the type of secondary structure conformation (helix or extended states) adopted by the majority of residues in a block. Although the previous method contains more global information, and at face value seems to be a more reasonable way to make structure predictions, we find that, in reality, the U-turn positions in protein structures seem to be primarily determined by the local interactions. Furthermore, for globular proteins to maintain roughly spherical shapes, the secondary structure types that the sequence fragments can adopt between the U-turns correlate strongly with the separations between the U-turn locations. Thus, if the U-turn positions can be accurately predicted, then in many cases the dominant secondary structure types of the blocks also can be predicted with high accuracy. The various terms in the energy function used in the previous method more realistically describe the interactions within a protein structure. However, due to the statistical nature of these energy terms, they also may introduce more "noise" into the prediction. The current method uses only one of these energy terms, and the inclusion of the multiple sequence information seems to significantly reduce the noise level. In terms of computational efficiency, the current method is very straightforward. A typical prediction with a reasonable number (less than 100) of homologous sequences takes only seconds on a modern workstation compared with the previous method that used a single sequence and took on the order of hours of CPU time. In principle, we also can use multiple sequence information in our previous method, but in practice it would be too time consuming. Further comparisons to the previous method will be made later in this article.

Our current approach bears some similarity to the turn prediction approach by Cohen et al.[6] in that both methods are designed to directly predict the turn positions in all classes of globular protein structures using local sequence properties. Their method is based on specific "sequence patterns," which involve hydrophobicity, charge, and special sequential arrangements of residues implemented in the context of a pattern-matching computer language. Some length constraints (length-dependent masking) also are applied for fine-tuning the predictions. High accuracy was achieved on their test set of proteins. However, to reach this high accuracy, the class of the hairpin structure (α/α, α/β, or β/β) has to be known beforehand, as are the complex sets of different patterns for each class of turns. On the other hand, the current method is energy based and does not require the knowledge of hairpin types or any predetermined sequence patterns. Also, the current method predicts the dominant secondary structure types of the blocks after the U-turn positions are determined, as opposed to use of the secondary structure types as input. Thus, we anticipate that the current method will be more general and easier to apply.

The outline of the rest of the article is as follows. In Materials and Methods the form of the energy function and details of implementation of the new method is presented. Then, in Results and Discussion, results on two test sets of 68 and 40 proteins, respectively, are shown, including the 38 proteins tested in the previous study, and comparison with our previous method is made. To assess the accuracy and the reason for success, predictions based on random assignment of the U-turn positions on a subset of the test proteins are made and analyzed. Finally, in the conclusion, the strengths, limitations, and possible further improvements of our current approach are discussed.

MATERIALS AND METHODS

The goal of the current study is to develop a method to accurately assign the U-turn positions in a protein sequence and the dominant secondary structure between the U-turns. First, a given amino acid sequence is scanned by using a statistical secondary structure propensity potential to identify localized regions showing a strong turn propensity. The somewhat noisy signals obtained in this way are smoothed by a local three-point averaging, by multiple sequence averaging using homologous sequences (if available), and by using a discrete Fourier transform (DFT)[7] as a high-frequency filter. Once the U-turns are identified, the secondary structure types of the blocks between the U-turn regions are decided by length constraints as well as the overall secondary structure preference of the residues within the blocks, which are also determined by the same statistical potential. A detailed description of the potential and the procedures follows.

Secondary Structure Propensity (r14) Potential

This short-range sequence-specific secondary structure propensity potential has been discussed previously[5] and has been used in lattice simulations of protein folding[3–5] and in threading.[8] It is based on the statistics of the occurrence of particular consecutive triplets of Cα-Cα vectors in the database of known protein 3-D structures. For easier derivation and practical use, the three-vector descriptor is mapped onto the "chiral" distance between the ends of vectors,

$$e_s = e_s(a_i, a_{i+1}, r^{2*}_{\mathrm{i}-1,i+2})$$

$$r^{2*}_{\mathrm{i}-1,i+2} = r^{2*}_{i-1,i+2}\ \mathrm{sign}\ ((v_{i-1} \times v_i) \cdot v_{i+1}) \qquad (1)$$

where a_i is the amino acid residue type at position i, v_i is the vector from Cα_i to Cα_{i+1}, and $r^{2*}_{i-1,i+2}$ is the "chiral" square of distance from Cα_{i-1} to Cα_{i+1} and is positive if the three vectors are arranged in a right-handed conformation and negative if they are in a left-handed conformation. The observed values (in Å^2) of $r^{2*}_{i-1,i+2}$ in experimentally determined structures of proteins are grouped into six coarse-grained bins that correspond to different structural classes:

$-128 < r^{2*}_{i-1,i+2} < 85$	bin = 1	extended β
$-85 < r^{2*}_{i-1,i+2} < 38$	bin = 2	loops and turns (left-handed)
$-38 < r^{2*}_{i-1,i+2} < 0$		
$0 < r^{2*}_{i-1,i+2} < 37$	bin = 4	helix (right-handed)
$37 < r^{2*}_{i-1,i+2} < 83$	bin = 5	loops and turns (right-handed)
$83 < r^{2*}_{i-1,i+2} < 135$	bin = 6	extended β

The generally used form is thus

$$e_{r14} = e_{r_{i-1,\, i+2}}(a_i, a_{i+1}, \mathrm{bin}) \qquad (2)$$

where the bin number signifies the range of "chiral" distance squared and the structural class. The term *r14* indicates that the energy is a function of the distance between two Cα atoms separated by three residues in the sequence. A total of 234 high-resolution protein structures were used in the potential derivation. The entire set of 2,400 (20 × 20 × 6) energy parameters may be obtained from the authors upon request. It is noted that the structures corresponding to specific r14 bins, which are based only on local Cα-Cα distances, do not always correlate with the hydrogen bond-based secondary structure classification.[9] In particular, bin 1 and bin 6 can represent extended loop regions as well as hydrogen-bonded β-strands; bin 2 and bin 5 often represent turn regions connecting different secondary structure elements. Bin 5 also can include distorted helices. Table I shows a comparison of the r14 definition with that of the DSSP[9] assignments for three representative proteins.

Turn Position Determination

We define a "turn state energetic preference," E'_T, at each residue as

$$E'_T(i) = \min\ [e_{r14}(a_i, a_{i+1}, 2), e_{r14}(a_i, a_{i+1}, 5)]. \qquad (3)$$

Similarly, the "extended state energetic preference," E'_E, and the "helix energetic preference," E'_H, are defined as follows:

$$E'_E(i) = \min\ [e_{r14}(a_i, a_{i+1}, 1), e_{r14}(a_i, a_{i+1}, 6)]. \qquad (4)$$

$$E'_H(i) = e_{r14}(a_i, a_{i+1}, 4). \qquad (5)$$

TABLE I. Comparison Between the r14[†] and DSSP Assignments

	Protein: 1pra
	10 20 30 40 50 60
Sequence	SISSRVKSKR IQLGLNQAEL AQKVGTTQQS IEQLENGKTK RPRFLPELAS ALGVSVDWLLNGT
r14	.HHHHHHHHH HTTETHHHHH HHTTETHHHH HHHHTTTETE TTTHHHHHHH TTETHHHHHEE..
DSSP	HHHHHHHHHH HHT HHHHH HHHTS HHHH HHHHHT SS TTHHHHHHH HT HHHHHS
	Protein: 1cyo
	10 20 30 40 50 60
Sequence	LKCHNTQLPF IYKTCPEGKN LCFKATLKKF PLKFPVKRGC ADNCPKNSAL LKYVCCSTDKCN
r14	.EETTTTTTE EEEETTTTTE EEEEETTTTT EETETEETEE TTETTETTTE EEEEETTTTE..
DSSP	EE S SSS EE TT EEEEEEEETTS SS EEEEE ESS TT EEEEEEESSTT
	Protein: 1ego
	10 20 30 40 50 60 70 80
Sequence	MQTVIFGRSG CPYCVRAKDL AEKLSNERDD FQYQYVDIRA EGITKEDLQQ KAGKPVETVPQIF VDQQHIG GYTDFAAWVKENLDA
r14	.EEETETTTT HHHHHHHHHH HHHHHHTTTE EEEEETHHHT EETHHHHHHH TTETTETTEEEET TTTETTT THHHHHHHHHHHE..
DSSP	EEEEE TT STHHHHHHHH HHHHHHHHSS EEEEE HHH HT SHHHHH HT S S EEEETTEEEESSHHHHHHHHHHHH

[†] Bins 1 and 6 are assigned as "E," bins 2 and 5 are assigned as "T," and bins 3 and 4 are assigned as "H."

Because a left-handed helix is very rare and the statistics for this bin are not very good, we do not use bin number 3 in calculating E'_H. To avoid local fluctuations, these energies are averaged over three consecutive residue positions:

$$E_T(i) = ⅓[E'_T(i - 1) + E'_T(i) + E'_T(i + 1)] \quad (6)$$

$$E_E(i) = ⅓[E'_E(i - 1) + E'_E(i) + E'_E(i + 1)] \quad (7)$$

$$E_H(i) = ⅓[E'_H(i - 1) + E'_H(i) + E'_H(i + 1)]. \quad (8)$$

We then define

$$E^*_T(i) = E_T(i) - \min\,[E_E(i), E_H(i)] \quad (9)$$

which is a measure of the turn propensity relative to the most favorable helix or extended type of structure. If one now plots the E^*_T along the sequence, distinct low-energy minima are usually apparent. These regions correspond to the predicted turns. However, the E^*_T vs. residue position curve is usually noisy, with closely spaced minima. For easier automatic assignment of the turn positions, we filter out the "high-frequency" component of the E^*_T curve by using a DFT and inverse Fourier transform. The $E_T(i)$ is first transformed into the "frequency domain" by

$$F\,(u) = \frac{1}{N}\sum_i E^*_T(i)\exp\,(-j2\pi ui/N) \quad (10)$$

where N is the number of data points and j is $\sqrt{-1}$. We then define a cut-off frequency as

$$u_{cut} = 1/n_{\min} \quad (11)$$

where $n_{\min}$ is the minimum length (in residues) of a β-type block as defined in the previous study:[2]

$$n_{\min} = 1.8\ S_0/3.4 \quad (12)$$

$$S_0 = 2.2\ m^{0.38} \quad (13)$$

where S_0 is the estimated radius of gyration[10] in angstroms, 1.8 S_0 is the estimated diameter of the hydrophobic core, 3.4 Å is the axial translation of an extended state per residue, and m is the number of residues of the protein. We modify $F(u)$ as follows:

$$F^*(u) = F(u), \quad \text{if } u < u_{\text{cut}}$$

$$F^*(u) = 0, \quad \text{otherwise.} \quad (14)$$

We then perform the inverse Fourier transform

$$E^f_T(i) = \Sigma F^*(u)\exp\,(j2\pi ui/N). \quad (15)$$

Usually, the resulting $E^f_T(i)$ is a smooth curve with only a few minima, or valleys. A typical example is shown in Figure 2. A valley is defined at residue i_p

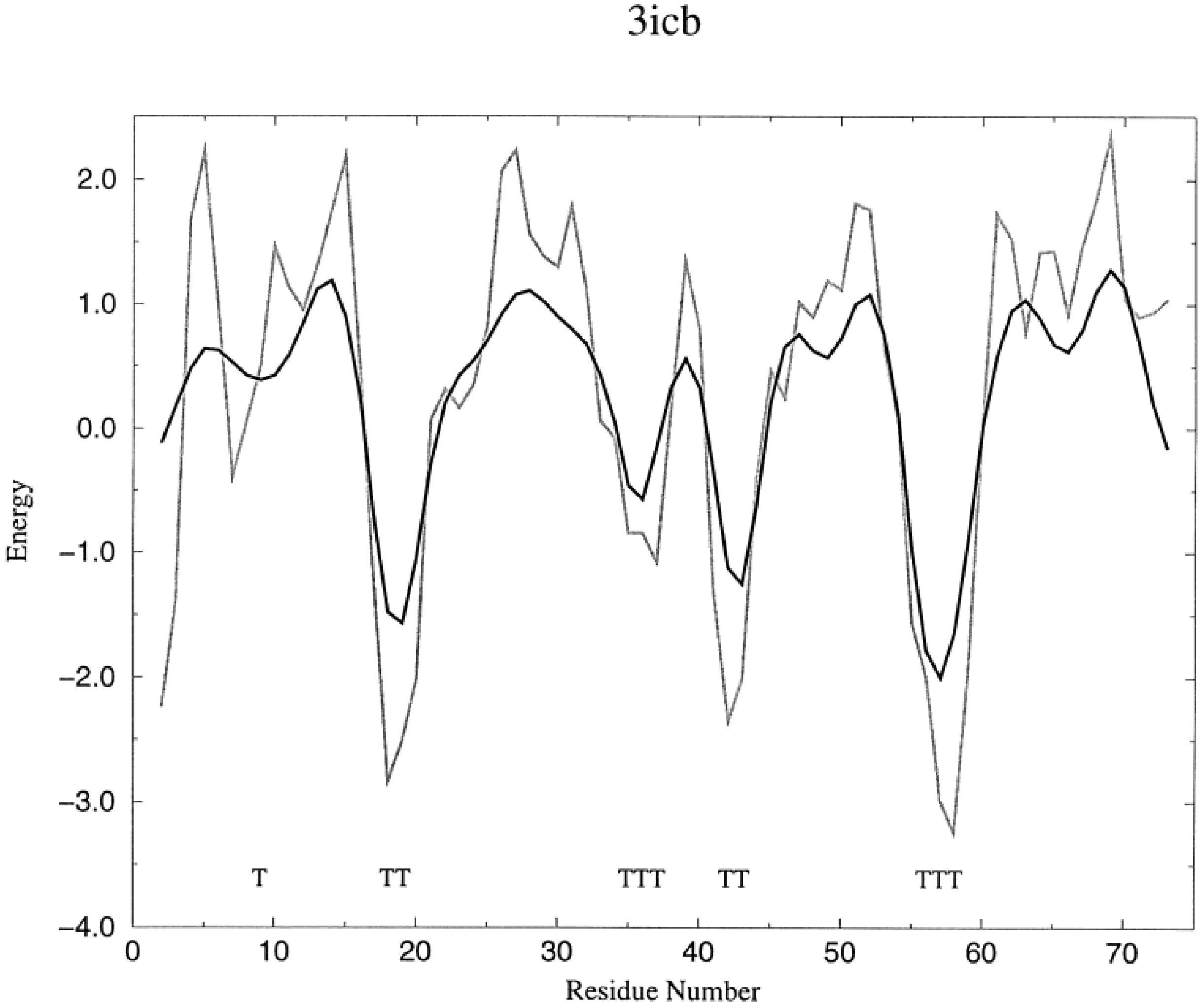

Fig. 2. Energy plot of the E_T^* [Eq. (9)] (gray, thinner line) and the E_T^f [Eq. (15)] (black) curves of a typical protein, 3icb. The Fourier high-frequency filter smooths the local fluctuations of the turn propensity. The predicted U-turn regions are marked with "T" below the curves.

where

$$E_T^f(i_p - 1) > E_T^f(i_p) < E_T^f(i_p + 1). \qquad (16)$$

The "major" valleys (or turn signals) are then selected as the centers of the predicted turn positions. The criteria for a "major" valley at residue i_p are 1) the E_T^* value at one of the five residues centered at i_p (from $i_p - 2$ to $i_p + 2$) must be less then zero, i.e., the turn is preferred over the other two types of secondary structures, and 2) the E_T^* value at i_p must be less than a threshold $E_{\min}$, which is defined in the current study as

$$E_{\min} = \max\,(-0.5, \overline{E}_p/3) \qquad (17)$$

where $\overline{E}_p$ is the average value of E_T^* in all the valleys. The width of a valley, which is here made equivalent to the range of a turn, is found by extending the turn region from the minimum position (i_p) toward both sides until

$$E_T^f(i_p - n_L) \geq E_T^f(i_p) + 0.3|E_T^f(i_p)|, \qquad (18)$$

$$E_T^f(i_p + n_R) \geq E_T^f(i_p) + 0.3|E_T^f(i_p)| \qquad (19)$$

where n_L and n_p are positive integers. The region from residue $i_p - n_L$ to $i_p + n_R$ is then assigned as a U-turn region. The parameters −0.5 and 3 in Eq. (17) and 0.3 in Eqs. (18)–(19) were empirically determined values from the predictions on the first 10 proteins listed in Table II. Although we find that our prediction is not very sensitive to these empirical parameters, this makes the first 10 proteins not part of the "genuine" test set. As will be discussed in the

TABLE IIA. PDB Names, Sizes (in residues), and Types (numbers of α-helices and β-strands) of the 68 Test Proteins in Test Set I

Name	Size	Type	Name	Size	Type	Name	Size	Type
1gb1	56	1α4β	2utg	70	4α	2aza A	129	1α8β
proA	46	3α	1ctf	68	3α3β	3fis A	73	4α
1fas	61	5β	1crn	46	2α2β	3icb	75	4α
1pou	71	4α	1msh	72	1α3β	3wrp	101	6α
1tlk	103	8β	1ftz	70	3α	1bel	120	3α6β
1ris	97	2α4β	1cis	66	1α4β	1cnp A	90	4α
1lpt	90	4α	1tin	69	1α3β	1ert	105	4α5β
1ten	89	7β	1cvo	62	5β	1fmb	104	2α8β
1mjc	69	5β	1adr	76	5α	1iba	78	3α3β
1gps	47	1α3β	1hme	77	3α	1kte	105	5α4β
1tfi	50	4β	1vna	65	1α4β	1ncm	99	8β
1tpm	50	5β	2ait	74	6β	1nin	105	8β
Alcc	51	3α	1cod	62	5β	1pbk	116	1α6β
1pra	63	5α	1cb1	78	4α	1upj	99	8β
1c5a	66	4α	1aca	86	4α	1rge	96	1α6β
1trf	76	4α	1aaj	105	9β	1acx	108	8β
1lea	72	3α2β	1aap A	56	2α2β	1azu	126	1α8β
2ptl	78	1α4β	1aba	87	3α4β	1bds	43	3β
1hdn	85	3α4β	1bov A	69	1α5β	1cbh	36	3β
1bta	89	4α3β	1cd8	114	9β	3ebx	62	5β
1ego	85	3α4β	1ifb	131	2α10β	6hir	49	4β
1svq	94	2α5β	1mba	146	8α	3rnt	104	1α6β
1ubq	76	1α5β	1pcy	99	7β			

TABLE IIB. PDB Names, Sizes (in residues), and Types (numbers of α-helices and β-strands) of the 40 Test Proteins in Test Set II

Name	Size	Type	Name	Size	Type	Name	Size	Type
1bfm A	69	3α	1ihf B	94	3α3β	1pyt A	94	3α3β
1bmg	84	8β	1ihw A	52	3α3β	1sap	66	1α4β
1cew I	108	2α5β	1mol A	94	1α6β	1smp I	100	1α8β
1ctj	89	5α	1myl B	40	2α	1tif	76	2α4β
1ecm A	91	3α	1ntx	60	5β	1tig	88	2α4β
1erw	51	4α5β	1orc	64	3α3β	1tii D	98	2α6β
1fim	102	2α4β	1otf A	59	1α2β	1vih	71	3α3β
1fip	73	4α	1pdg A	87	6β	1wap A	68	7β
1ftt	68	3α	1pfs A	78	6β	1ytf C	46	6β
1grx	85	3α4β	1pht	83	1α5β	2bop A	85	2α4β
1gua B	76	1α5β	1pog	62	3α	2crt	60	5β
1hcn A	85	6β	1poh	85	2α4β	2hpe A	99	1α8β
1hsm	79	3α	1prt D	98	1α7β			
1hst A	74	3α2β	1ptf	87	3α4β			

next section, the accuracy obtained for these 10 proteins is not very different from that of the other test proteins.

Secondary Structure Determination

After the U-turn regions have been determined, the sequence fragments between these turns are used to determine the dominant secondary structure of the "blocks," which is either a helix or extended state. It is noted in this study that if a block is shorter than four residues, it is simply assigned as a loop or extended region. The dominant secondary structures are first determined by length constraints, i.e., if the blocks are linear and transglobular in nature, then a very long block can be only a helix and a short block very likely an extended state. We define

$$\mathrm{MinH} = \mathrm{Min}\,(10,\ 1.8\ S_0/2.3 - 1) \tag{20}$$

$$\mathrm{MaxE} = \mathrm{Max}\,(11,\ 2.5\ S_0/3.4 + 3) \tag{21}$$

where the number 2.3, which corresponds to the average length in angstroms of a residue that is 60% helical and 40% extended in nature, in Eq. (20) is empirically determined to give a reasonable lower limit for helix sizes in small proteins. (A topological helix block may contain several residues in the extended state. Thus, if 1.5 Å is the axial translation of a helical residue, as used in Eq. (20), the resulting value of MinH would be too large.) The "2.5 S_0" in Eq. (21) is the estimated diameter of the entire globule. Any blocks that are longer than MaxE residues are first assigned as helices, and those shorter than MinH as extended states.

Supposing that the block begins at residue ibeg and ends at residue iend, we calculate the following energies:

$$\overline{E}_E = \sum_{i=ibeg}^{iend} E_E(i) \tag{22}$$

$$\overline{E}_H = \sum_{i=ibeg}^{iend} E_H(i) \tag{23}$$

The secondary structures of those blocks assigned by the length constraints can be overridden when the energies calculated above indicate differently. Specifically, when the block size is longer than MaxE and less than MaxE+3 residues, but $\overline{E}_E < \overline{E}_H$, and $|(\overline{E}_H - \overline{E}_E)/\overline{E}_H| \geq 0.25$, then the block is assigned as an extended state instead of a helix. Similarly, when the block size is shorter than MinH and longer than max(5, MinH-3) residues, $\overline{E}_H < \overline{E}_E$, and $|(\overline{E}_E - \overline{E}_H)/\overline{E}_E| \geq 0.25$, then the block is assigned as a helix instead of an extended state.

For those blocks having lengths between MinH and MaxE, the secondary structures are determined primarily by $\overline{E}_E$ and $\overline{E}_H$. In particular, if $\overline{E}_E < \overline{E}_H$, then the secondary structure of the block is assigned as an extended state; otherwise, it is a helical block. If $\overline{E}_E = \overline{E}_H$, then the block is assigned as a helix if its length is longer than 10 residues; otherwise, it is assigned as an extended state. If the block size is less than four residues, it is assigned as a loop or turn region regardless of the energy values.

Multiple-Sequence Averaging

It is generally assumed that proteins with homologous sequences adopt very similar structures.[11] On the basis of this assumption, it is desirable to use a set of aligned homologous sequences (if any) to improve the signal-to-noise ratio in the prediction. In the current study, all the homologous sequences are weighted the same. If we suppose that there are *ns* homologous sequences available for prediction, the multiple sequence averaging replaces Eqs. (3)–(5) by

$$E'_T(i) = \frac{1}{ns}\sum_{j=1}^{ns} \min\,[e_{r14}(a_{i,j}, a_{i+1,j}, 2), e_{r14}(a_{i,j}, a_{i+1,j,} 5)] \tag{24}$$

$$E'_E(i) = \frac{1}{ns}\sum_{j=1}^{ns} \min\,[e_{r14}(a_{i,j}, a_{i+1,j}, 1), e_{r14}(a_{i,j}, a_{i+1,j,} 6)] \tag{25}$$

$$E'_H(i) = \frac{1}{ns}\sum_{j=1}^{ns} e_{r14}(a_{i,j}, a_{i+1,j}, 4) \tag{26}$$

where a_{ij} is the ith residue in the jth sequence. The gap positions are skipped and do not contribute to the averaging. Thus, *ns* in Eqs. (24)–(26) is defined as the actual number of sequences that do not have a gap at residue *i* in the multiple sequence alignment. Figure 3 shows the E'_T curves with and without multiple sequence averaging for a typical protein. The U-turn and secondary structure predictions can then proceed as outlined above without modification. As discussed in the next section, use of multiple sequence averaging, in general, significantly improves the prediction. In the current study, multiple sequence alignment information is obtained from the PHD[12,13] program, but this information also can be obtained from a variety of other sequence search programs. A flow chart illustrating the entire prediction procedure is depicted in Figure 4.

Test Proteins

Two test sets of proteins with known structures are used in the current study. The PDB[14] names of these proteins are listed in Table II. Approximately 65% have more than 10 "quality" homologous sequences whose sequence identity ranges between 85 and 35%. The first test set consists of 68 proteins, including the 38 proteins tested in the previous study and 30 new ones. The structures of 17 of these 68 proteins were used in the derivation of the r14 potential, forming a self-consistent subset for validating the method. This point is discussed further in the next section. To test further the predictive power of the current method, predictions on a second test set of 40 proteins also are performed. None of the 40 protein structures is used in the derivation of the r14 potential or any other empirical parameters in the current method.

Accuracy Assessment Method

The measure of the prediction accuracy, which is the same as used previously,[2] is as follows. If there is overlap between the predicted and the actual U-turn regions, the errors in U-turn positions are measured as the additional overlap (if any) between the predicted turn regions and the regular secondary structure regions (based on DSSP assignment) in experimentally determined protein structures. If there is no overlap between the predicted and the actual U-turn regions, then the distance between the C-

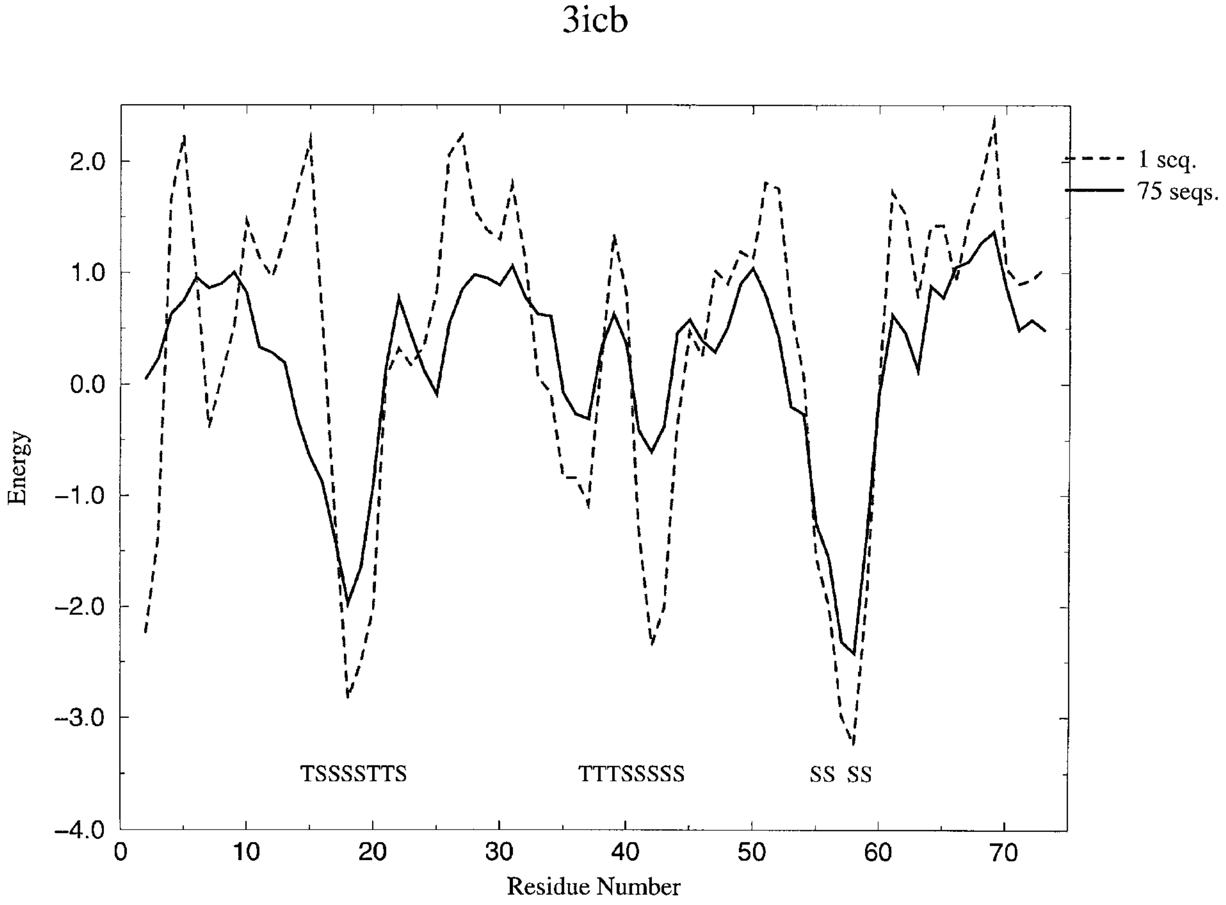

Fig. 3. Energy plot of the $E_{\ddagger}$ of a typical protein, 3icb, with a single sequence (dashed curve) and with multiple-sequence averaging (solid curve). The DSSP assignment of the turn regions are marked below the curves. The shapes of both curves are usually similar; however, the multiple-sequence curve usually contains less false U-turn signals. In the current case, the single-sequence curve contains a false signal at residue 7.

terminal ends of the predicted and the actual turn regions is used. The number of over and under predicted turns also are examined. The accuracy in secondary structure block assignment is measured as the ratio of the number of correctly predicted dominant secondary structures of the blocks to the total number of regular linear secondary structure elements, as classified by DSSP, excluding very short secondary structural elements (less than three β-residues or less than four helical residues in a block). A secondary structure element in an experimentally determined protein structure is counted as correctly predicted if (1) at least half of the predicted block region overlaps with a block region in the actual protein structure and (2) in the experimentally determined protein structure, the predominant secondary structure type in the block region (which is predicted by the current method) is the same as the prediction. Only one predicted block region can be assigned to a particular block region in an experimentally determined protein structure. For example, if the current method predicts residues 21–35 to be a block between two U-turns and the predicted dominant secondary structure type is helical, and in the experimentally determined protein structure, residues 23–34 are a block region and 10 of them are assigned as helical by DSSP, then we say the prediction is correct. However, if in the experimentally determined protein structure, a block runs from residues 30–40 and is helical, then the predicted block does not correspond to the actual block because less than half of the predicted block region overlaps with the actual block region. Thus, the predicted helical block from residues 23–34 does not count as a correct prediction.

For the second set of test proteins, an additional prediction based on the PHD three-state assignment instead of our r14 energy function has been performed for comparison purposes. The U-turn regions are taken from the sequence regions between the assigned secondary structure elements (more than one consecutive “H” or “E” assignment). The prediction also is assessed by the same method mentioned above.

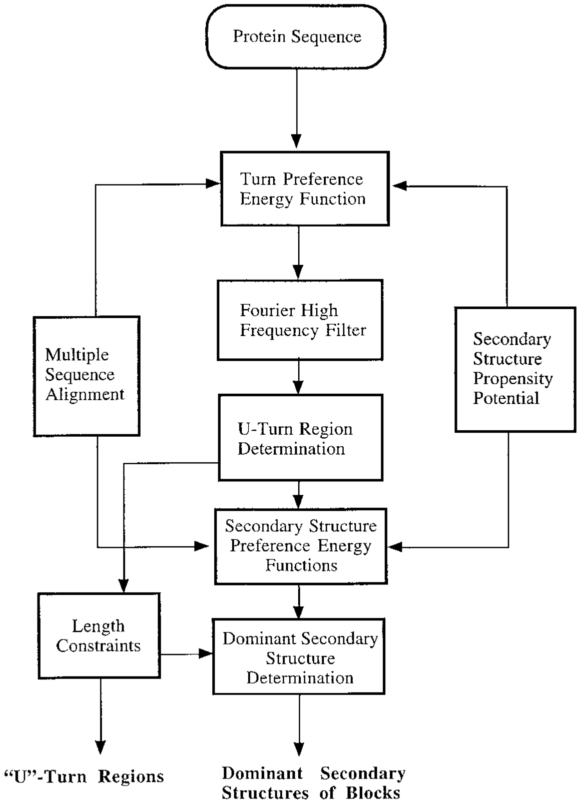
Protein Sequence
Turn Preference Energy Function
Fourier High Frequency Filter
Multiple Sequence Alignment
Secondary Structure Propensity Potential
U-Turn Region Determination
Secondary Structure Preference Energy Functions
Length Constraints
Dominant Secondary Structure Determination
"U"-Turn Regions
Dominant Secondary Structures of Blocks

Fig. 4. Flow chart of the procedure used in the current approach.

RESULTS AND DISCUSSION

Overall Accuracy

The first set of 68 test proteins is divided into three overlapping sets: 1) the 38 proteins tested in the previous method, 2) the 17 proteins whose structures were used in the derivation of the r14 potential, and 3) the 51 proteins whose structures were not used in the derivation of the potential. Table IIIA shows the overall prediction accuracy and the comparison with the previous method. On average, the inclusion of the multiple sequence alignment information (if available) improves the results significantly. In the following discussion, unless otherwise stated, we refer only to the multiple sequence averaged results. It can be seen from Table IIIA that, for the same set of 38 proteins, the turn positions are predicted much more accurately than in our previous method, with an average error improving from more than two residues to less than one residue. Approximately 94% of the U-turns are predicted within three residues, whereas our previous method achieved only 80% accuracy. (This should not be confused with the 95% accuracy reported in our previous study,[2] which defined a correct prediction when the predicted U-turn regions have any overlap with the actual U-turn regions. Although our current method gives approximately the same accuracy by that definition, it usually predicts narrower and more well-defined U-turn regions.) The prediction of dominant secondary structure also improves significantly from 82 to 88% when multiple sequence averaging is used. Due to the nature of the new method, it is more likely to overpredict turns; however, the new method is only slightly worse in this aspect compared with the previous method, and there are fewer underprediction errors in our new method. A summary of current predictions with multiple sequences and comparison with our previous method for the 38 proteins is given in Table IV in the same format as Table II of Ref. 2. (The prediction details of other proteins in the current study and the prediction program can be found on our World Wide Web server.[15])

As seen in Table IIIA, the prediction accuracy is consistent across all three sets of proteins, suggesting that there is very little "memory" in the statistical potential. The slightly lower accuracy for the 17 proteins used in the r14 potential derivation is probably also due to the insufficient statistics within this small set. The overall accuracy for the entire 68–protein set is very encouraging: 94% of the U-turn regions are predicted within three residues (84% are within 1 residue) with an average error of less than one residue; 89% of the identity of dominant secondary structures in blocks are correctly predicted. Another good way to measure the prediction accuracy is to calculate the Matthews coefficient.[16] In the current study, the overall prediction gives a Matthews coefficient of 0.92 for the U-turn regions and a value of 0.79 for the predicted identity of dominant secondary structure. It is noted that we found no homologous sequences for four of the test proteins and that 24 of the test proteins have less than 10 quality homologous sequences. However, if the 24 test proteins are excluded in the test set, the accuracy increases only slightly, e.g., the overall accuracy of the dominant secondary structures becomes 90%. All other accuracy measures are almost identical to the values in Table IIIA and, thus, are not shown in this study. This suggests that, even though multiple sequence averaging is very helpful for the predictions, perhaps only a few quality homologous sequences are sufficient to yield much better accuracy. As mentioned in the previous section, the first 10 proteins in the test set are used in the determination of some empirical parameters [Eqs. (17)–(19)]. Table IIIA also lists the prediction accuracy obtained for these proteins. It can be seen that they are not particularly higher than those of the other proteins and are very close to the average accuracy.

Predictions were also performed on the second set of 40 test protein structures. Table IIIB summarizes the prediction accuracy. As mentioned in the previous section, additional predictions based solely on the output of the PHD program also are analyzed. It can be seen that the accuracy obtained with our new

TABLE IIIA. Prediction Accuracy for Test Set I and Comparison With Previous Method

	Previous test set (38 proteins)	Structures used in the E_{r14} derivation (17 proteins)	Structures not used in the E_{r14} derivation (51 proteins)	First 10 structures (10 proteins)	The entire test set I (68 proteins)
Average U-turn errors in residues per U-turn	(2.2)[†] 1.0[‡] 0.7[§]	1.2 1.1	0.8 0.6	1.0 0.6	0.9 0.7
Predicted U-turns region with errors[2] three residues (%)	(80) 92 95	90 90	94 96	92 96	92 94
Overpredicted U-turns[¶] (%)	(5.4) 8.2 6.9	11.9 11.9	5.9 5.2	8.2 10.0	7.3 6.7
Underpredicted U-turns[††] (%)	(4.0) 2.2 1.1	1.7 0.9	1.4 0.7	2.0 0.0	1.5 0.7
Matthews Coefficient for U-turn prediction	(0.73) 0.84 0.88	0.87 0.88	0.92 0.94	0.90 0.91	0.91 0.92
Dominant secondary structure (%)	(82) 76 88	76 88	83 90	80 90	81 89
Matthews Coefficient for dominant secondary structure prediction	(0.75) 0.60 0.79	0.69 0.76	0.70 0.80	0.60 0.81	0.65 0.79

[†] The numbers in the parentheses are the results from the previous study.[2]
[‡] Results using a single sequence.
[§] Results using multiple sequence averaging (if no homologous sequences available, a single sequence is used).
[¶] Percentage of the mispredicted U-turn region that divides a secondary structure element.
[††] Percentage of the U-turn regions that are completely missed by the prediction.

method is very similar to that of the first test set. The method based on the PHD assignment does not work as well as our current method of predicting U-turn positions and dominant secondary structures, and the PHD method tends to underpredict (or miss) the U-turns. This method, however, seldom overpredicts the U-turns. This opens up the possibility of a combined approach with our current algorithm, which is discussed further in this section. In this test set, the PHD method never misassigns a helix to a β-strand and, thus, has a higher Matthews coefficient for dominant secondary structure prediction. The seemingly high standard deviations in Table IIIB are caused by the discrete nature of the prediction, i.e., the error in U-turn prediction is measured in multiples of residues, and any imperfection in dominant secondary structure prediction of a protein will cause more than 10% error because all proteins in the test set have less than 10 secondary structure elements.

Figures 5–7 present typical energy plots for α, β, and αβ proteins with multiple sequence averaging. As seen in these plots, the E_T^f curves clearly identify the turn regions, whereas the E_H and E_E curves show the relative helical and extended-state propensities between the turn regions. Figure 8 illustrates a typical example of an overpredicted turn in the middle of a helix and a case where the length constraints override the secondary structure propensity.

Turn Predictions

It is actually a little surprising that a very simple local energy function can be used to determine turn positions and secondary structure types with such high accuracy. It has been shown that the loop or turn regions are easier to locate because they usually contain a high percentage of hydrophilic and charged residues or certain residue patterns.[17–23] Such patterns are implicit in the statistical potential, and the current approach does not require any explicitly predefined patterns. Furthermore, the predictive power of the current approach is reinforced by multiple sequence information under the assumptions that homologous sequences have very similar structures. The high accuracy of the U-turn prediction from this study suggests, once again, that the turn positions are determined predominantly by the local interactions[19] and do not heavily depend on tertiary interactions.

One drawback of the current method is that because it looks for the sequence regions showing strong turn propensity, it has a tendency to overpredict U-turns, as seen in Table III. Apparently, a turn

TABLE IIIB. Prediction Accuracy for Test Set II

	Entire test set II (40 proteins)
Average U-turn errors in residues per U-turn	0.7[†] ± 1.2[¶] 0.7[‡] ± 1.2 1.5[§] ± 1.8
Predicted U-turns region with errors[2] 3 residues (%)	91 95 79
Overpredicted U-turns[¶] (%)	7.2 6.3 2.5
Underpredicted U-turns (%)	5.8 2.2 10.1
Matthews Coefficient for U-turn prediction	0.87 0.92 0.87
Dominant secondary structure (%)	83 ± 16[†] 87 ± 13 80 ± 20
Matthews Coefficient for dominant secondary structure prediction	0.76 0.78 0.89

[†] Results obtained using single sequence.
[‡] Results obtained using multiple sequences.
[§] Results extracted from output of the PHD program.
[¶] Standard deviation.

position can be either a U-turn or a local distortion of the polypeptide chain (e.g., a β-bulge), the latter of which is less useful and can even be confusing for low-resolution modeling. A natural question arises in that because the statistical potential by definition does not distinguish between a U-turn and a local turn, how does the current method distinguish between these two types of turns? In fact, it is sometimes very difficult to determine whether a predicted turn is located within a secondary structure block or is a global U-turn position. However, we estimate that less than 20% of the β-strands in experimentally determined protein structures contain a β-bulge. Also, the Fourier high-frequency filtering smooths out local fluctuations in turn propensity. As a result, the local turn propensity does not significantly interfere with the global U-turn recognition, although it cannot be eliminated completely. As seen in Table III, the ability to predict the positions of the U-turn regions with a much higher accuracy more than compensates for the slightly higher overprediction rate when compared with our original approach.

Our observation shows that the current method achieves slightly higher accuracy for β-proteins. For the 28 all-β proteins in test sets 1 and 2, more than 98% of the U-turns are correctly predicted with 96% of them within two residues in position, which is compared with the 72% obtained by Wilmot and Thornton.[19] Their method, however, does provide additional information on the types of β-turns.

Dominant Secondary Structure Prediction

In our current study, the dominant secondary structure types of approximately 60% of the blocks are determined by length constraints, with an accuracy higher than 90%. This suggests that the length constraint is one of the dominant factors determining the secondary structure types no matter what types of residues the block actually contains. The dominant secondary structure types of those medium-sized blocks are predicted by using the statistical potential as outlined in the previous section and have a prediction accuracy of approximately 85%. Because this is much better than random, it suggests that the local interactions also play an important role in secondary structure determination. Obviously, given that tertiary interactions are, in practice, also important, it is surprising how well one can do when these contributions are neglected. Nevertheless, there are errors in the approach; we next turn to a more detailed examination of their nature.

Sources of Error

The current method tends to overpredict turns. Often, the errors in dominant secondary structure prediction are associated with the errors in U-turn predictions. For example, one of the primary sources of error is a turn predicted to be in the middle of a helix. For a medium-sized helix, this splits the block into two short elements. Then, on the basis of the length constraints, one or both of the elements are predicted to be in the extended state. Another common source of error is that a long and usually curved β-strand, which is usually hydrogen-bonded to two different β-sheets, is predicted as a helix based on length constraints. Also, sometimes the N- or C-terminal irregular loops and turns are predicted as helices. The above types of errors contribute to 60–70% of the wrong predictions of the dominant secondary structures. The rest of the errors may come from the inability to use local interaction energies to distinguish between secondary structure types, perhaps due to tertiary interactions that override the local preferences of the backbone conformations.

Partial remedies exist for the first three types of errors described in the previous paragraph. First, if a sharp turn signal appears between two short helix-favored regions, it is a good indication that it is a false turn signal, or it may correspond to a kink in the middle of the helix. This might be eliminated by use of a residue-based secondary structure prediction method. Second, if a long block (which is predicted as a helix by length constraints) shows a strong extended state character from the energy functions and the prediction based on a single sequence shows a turn in the middle of the block, it is

TABLE IV. Comparison[†] With Previous Method[2] for the Same Set of 38 Proteins

	PDB name	U-turn prediction accuracy	Errors of U-turn locations	Secondary structure block prediction accuracy	Comments on wrong assignment of current algorithm
1	1gbl	4/4[‡] 4/4 (1 over)	0-2/2-3-0 1-1-0-0	5/5 5/5	One extra turn predicted in the second strand[§]
2	proA	2/2 2/2 (1 over)	2-3 0-2	3/3 2/3	One extra turn predicted in the second helix
3	1fas	5/5 5/5	2-1-0-2-0 2-0-0-0-0	5/5 5/5	—
4	1pou	3/3 3/3	4-3-2-0 2-0-0-0	4/4 4/4	—
5	1tlk	7/7 7/7	5-3-2/1-1-2-4-7 0-0-0-0-0-0-0	7/8 8/8	—
6	1ris	5/5 5/5	3-5-6-3/4-4 0-1-1-2-0	5/6 6/6	—
7	1lpt	4/4 4/4 (2 over)	0-5-2-0 0-3-0-1	4/4 3/4	One helix predicted as a β-strand
8	1ten	6/7 7/7	1/1-3-0-3-2-2/1 0-1-0-0-1-0-0	7/8 8/8	—
9	1mjc	5/5 5/5	1-2-0-2-0 0-1-0-0-0	6/6 5/6	Last β-strand predicted as a helix
10	1gps	4/3 3/3 (1 over)	0-1-2-1 1-0-2	4/4 3/4	One helix predicted as a β-strand
11	1tfi	4/3 3/3	0-0-0-1 0-0-0	4/4 3/4	One β-strand predicted as a helix
12	1tpm	4/4 4/4	0-0-4-1 1-0-1-0	5/5 5/5	—
13	Alcc	3/2 2/2	2-0-8 0-0/2	2/3 2/3	One short helix predicted as a β-strand
14	1pra	4/4 4/4	7-3-0-2 0-0-0-0	3/5 5/5	—
15	1c5a	3/3 3/3 (1 over)	4-0-5 0-4-1	3/4 4/4	—
16	1trf	4/3 3/3	2-3-3-0 0-0-0/1	3/4 4/4	—
17	1lea	4/4 4/4 (1 over)	2-2-4-0 0-0-1-0	4/5 5/5	One short β-strand inserted
18	2ptl	6/5 5/5 (1 over)	1-3-3-4-3-2 0-2-2-0-0	5/5 3/5	Two β-strands predicted as helices
19	1hdn	5/6 5/5 (1 over)	0-3-4-1-2/2 0-3-0-0-1	6/7 6/7	One short β-strand missing
20	1bta	6/6 6/6	2-0-2-3-0-1 0-0-3-0-0-0	7/7 6/7	One β-strand predicted as a helix
21	1ego	8/7 7/7	0-7-2-2-0-2-0-7 0-3-1-0-0-1-2	6/7 7/7	—
22	1svq	6/6 6/6	1-0-1-0-3-0 1-0-1-0-0-0	6/7 7/7	—
23	1ubq	5/6 6/6	3-0-3-1-4 0-0-1-0-0-0	4/6 5/6	One β-strand predicted as a helix
24	2utg	3/3 3/3 (1 over)	3-2-7 1-0-0	4/4 4/4	One short helix inserted
25	1ctf	4/5 5/5	0-3-0-1/1 0-0-0-0-0	5/6 4/6	One β-strand predicted as a helix, one helix predicted as a β-strand

TABLE IV. (Continued)

	PDB name	U-turn prediction accuracy	Errors of U-turn locations	Secondary structure block prediction accuracy	Comments on wrong assignment of current algorithm
26	1crn	4/4	7-0-2-0	3/4	Two helices predicted as β-strands
		(2 over)	0-0-2-0	2/4	
27	1msh	4/4	0-0-4-0	2/4	One helix predicted as a β-strand
		4/4	0-0-0-0	3/4	
28	1ftz	4/4	4-3-3-0	3/3	One helix predicted as a β-strand
		4/4	1-0-2-3	2/3	
29	1cis	4/5	6-0-0-3	0/5	—
		5/5	0-0-0-0-0	5/5	
30	1tin	6/5	0-4-3-0-3-0	4/4	—
		5/5	0-0-0-0-0	4/4	
31	1cvo	5/5	0-3-1/2-4-0	4/5	—
		5/5	0-0-0-2-0	5/5	
32	1adr	4/4	4-3-0-4	3/5	One helix predicted as a β-strand, one helix missing
		4/4	3-1-1-0	3/5	
		(1 over)			
33	1hme	2/3	2-4	3/3	One helix predicted as a β-strand
		3/3	1-2	2/3	
		(1 over)			
34	1vna	4/4	4-2-4-0	2/5	One helix predicted to be within a turn region
		4/4	0-0-1-0	4/5	
35	2ait	6/6	2/2-1/1-4-0-3-1	5/6	One β-strand predicted as a helix, one β-strand missing
		6/6	1-2-1-0-1-0	4/6	
36	1cod	4/4	4-3-4-1	3/5	—
		4/4	1-0-0-0	5/5	
37	1cbl	3/3	3-0-0	4/4	—
		3/3	0-0-0	4/4	
38	1aca	4/4	1-1-0-4/4	3/4	—
		4/4	0-0-0-4	4/4	

[†] See Table II in Ref. 2 and text for the definition of various accuracy measures.

[‡] In columns 3–5, the first line in every row refers to the prediction by our previous method, the second line refers to the prediction by our current method with multiple sequences.

[§] Column 6 refers only to the current prediction scheme; see Table II in Ref. 2 for comments on predictions from the previous method.

usually an indication of a long, curved β-strand. Finally, sometimes a short to medium-sized helix is predicted at the termini of the sequence, and the energy curves [Eqs. (9), (25)–(26)] show very noisy patterns. Furthermore, the helix energy function [Eq. (26)] does not have a usual minimum close to the center of the block and does not convincingly dominate the other two (turn and extended state) curves, making it possible that it is just a piece of structure consisting of irregular turns and loops. However, it is very difficult to unbiasly implement these observations into the algorithm. We have tried to make the prediction as objective and automatic as possible to eliminate subjective human factors. However, after obtaining the prediction from the algorithm, it is sometimes helpful and informative to plot and study the energy curves themselves.

Size Considerations

In our previous method, we were restricted to small proteins because of the burial potential model used. Here, the burial energy is not considered, and the statistical potential does not depend as heavily on the shape or size of the protein. To account for the nonlinear and nontransglobular nature of the blocks, the length constraints [Eqs. (20)–(21)] used in the secondary structure assignment are not always size dependent, e.g., the minimum block size is always set to four residues. Thus, it is anticipated that the current method can be applied to larger proteins. Nineteen of the 108 test proteins in the current study are larger than 100 residues, compared with only 1 in the 38 test proteins in the previous study.[2] Encouragingly, the prediction accuracy of the dominant secondary structures of these proteins is 88%, very close to the average. It should be pointed out, however, that in larger proteins, the blocks are not always transglobular in nature, and the shape of the protein may be very nonspherical. Thus, the size constraints are harder to apply, i.e., it is difficult to estimate the $n_{\min}$, MaxE, and MinH in larger proteins. Also, on average, there are more tertiary interactions in larger proteins, but it is not clear whether its importance relative to the local interactions in determining the secondary structures increases with protein size. In general, for very large,

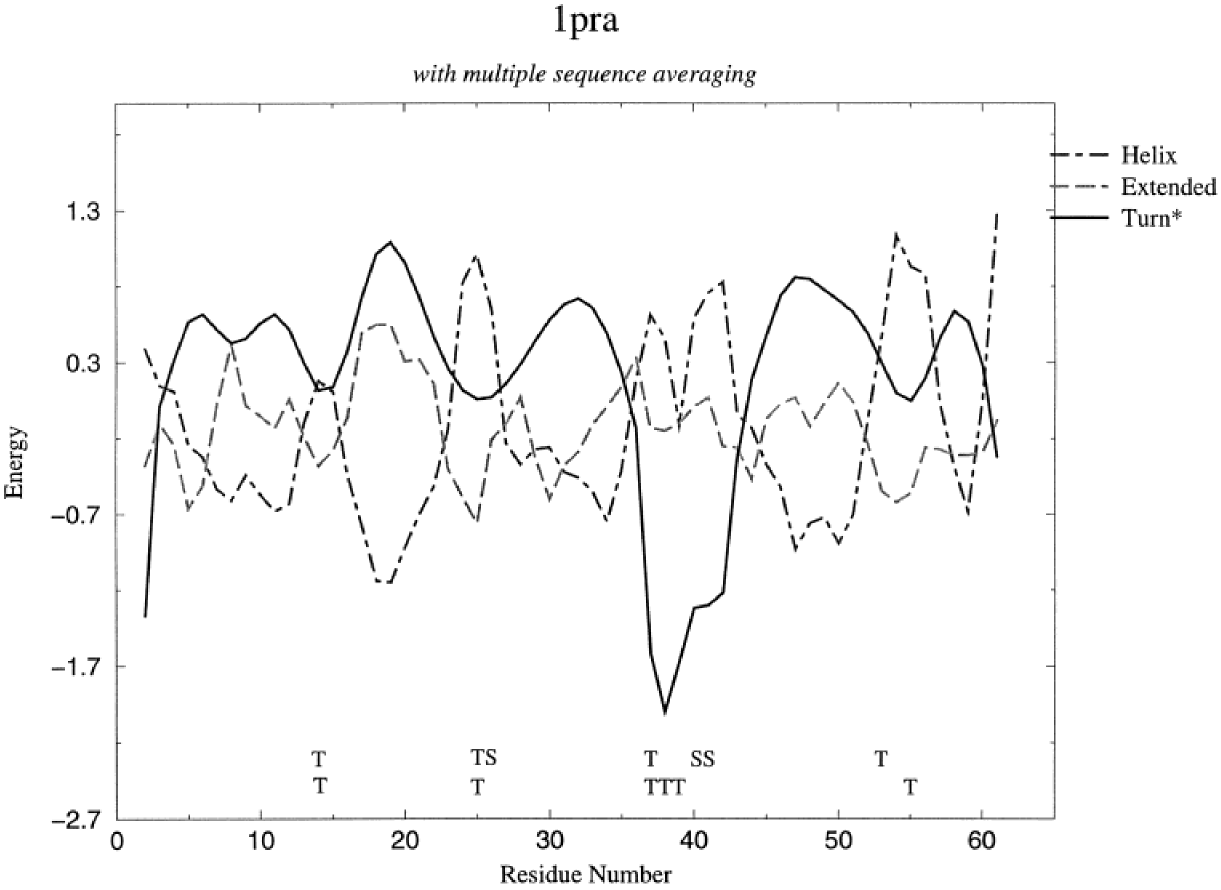

Fig. 5. Energy plot (with multiple-sequence averaging) of a typical α-protein, 1 pra. The dot-dashed, dashed, and solid curves correspond to E_H [Eq. (8)], E_E [Eq. (7)], and E_T^f [Eq. (15)], respectively. (For clarity, the $E_T^{\ddagger}$ curve is not shown, but it should be noted here that both $E_T^{\ddagger}$ and E_T^f curves are used in determining the U-turn positions.) The four major turn signals (minima) of the solid curve separate the sequence into five blocks. Below the curves, both the DSSP assignment of the turn regions ("T" or "S" on the first line) and the predicted U-turn positions ("T" on the second line) are marked. The helical energies are lower than the extended energies in most of the block regions, and thus the five helices are clearly resolved. The predicted U-turn regions correlate very well with the DSSP assignment.

single-domain proteins, the ability to use local interactions to predict secondary structure types can be expected to be less successful due to the contribution of tertiary interactions and less obvious length constraints. However, we speculate that with carefully chosen length constraints, it is possible to extend the current method to apply to multiple domain proteins. These issues need further investigation in the future.

Comparison With Residue-Based Secondary Structure Prediction Methods

The current method identifies the turns with high accuracy and gives a low-resolution prediction of the distance-based dominant secondary structure types of the blocks between turns. In contrast, residue-based methods[12,13] can predict the hydrogen bond-based secondary structure fairly accurately at the individual residue level and can refine the block-based secondary structure prediction. We have examined a combined approach in which the U-turn positions are determined by using the current method, and within each block the dominant secondary structures are predicted by using the residue-based assignment of the PHD program. Such a hybrid approach[4] (using our previous method and the PHD program) has been used recently as the initial steps in a protein tertiary structure prediction algorithm. The resulting accuracy of the current combined approach is approximately 90% in dominant secondary structure prediction, very similar to our current block-based approach. Despite its high accuracy of secondary structure prediction at the residue level, the PHD method sometimes completely misses a β-strand or short helix. Thus, this hybrid method does not necessarily improve the prediction. From the above discussion, it is apparent that for the purpose of low-resolution 3-D protein model building, our current method, perhaps refined by the high-accuracy, residue-based approach to

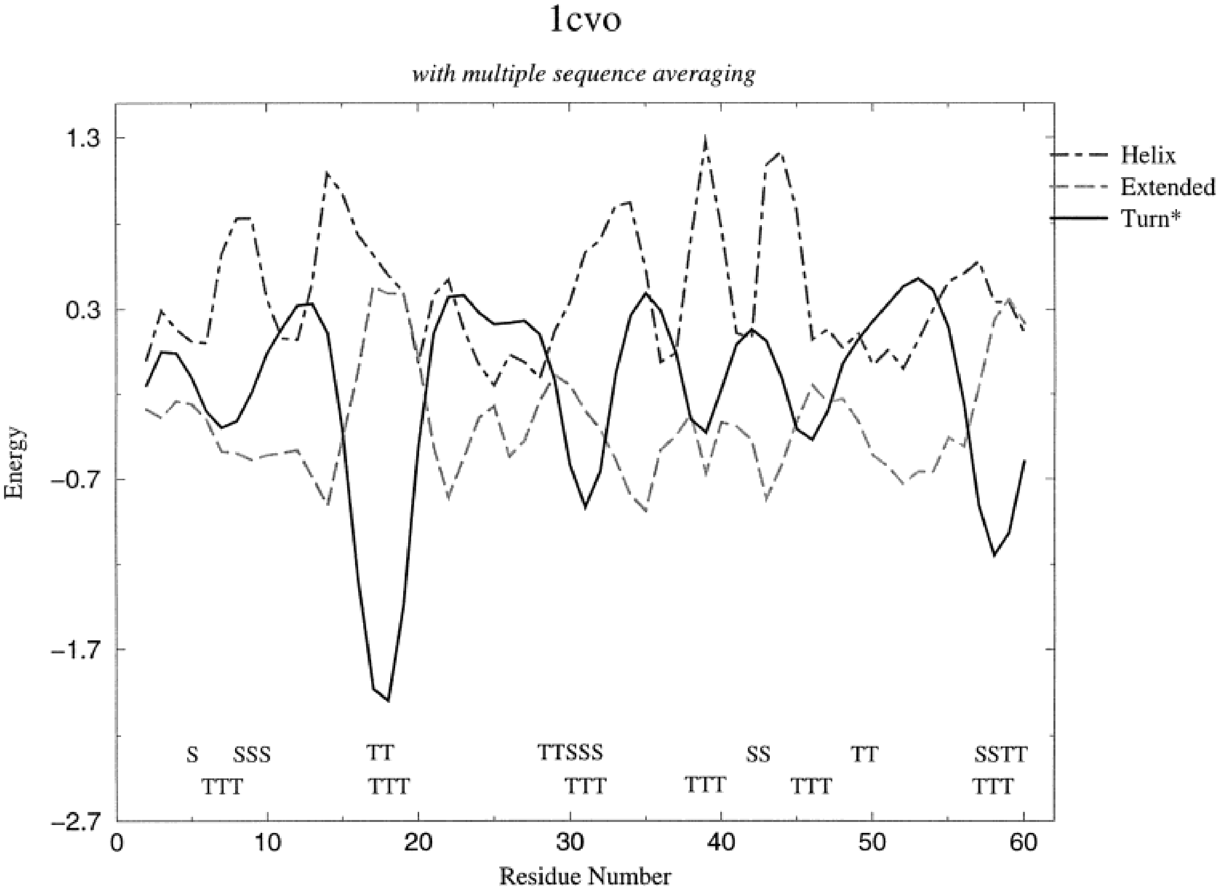

Fig. 6. Energy plot (with multiple-sequence averaging) of a typical β-protein, 1 cvo. The labeling scheme is the same as in Figure 5. The five major turn signals (minima) of the solid curve (excluding the one at the C-terminal, which is only used as an N-cap for the last block) separate the sequence into six blocks. The extended energies are lower than the helical energies in all of the block regions, and thus, five β-strands and an extended loop (residues 40–45) are clearly resolved. Except in the extended loop region, the predicted U-turn regions correlate very well with the DSSP assignment.

correct for turn overprediction, is superior to using only a residue-based, three-state prediction assignment.

Reasons for High Accuracy of Secondary Structure Prediction

As mentioned above, the accuracy of the dominant secondary structure prediction depends heavily on the accuracy of the U-turn region prediction. However, it would be interesting to know how much extra accuracy one gains compared with a random prediction of the turn regions. That is, is the enhanced accuracy due only to the averaging window for the blocks and nothing more? To examine the effects of good U turn predictions on the accuracy of dominant secondary structure prediction, we randomize the U-turn prediction part of the current method. The implementation of this random prediction protocol is as follows. (1) The total number of U turns is randomly chosen from m/25 to m/8 where m is number of residues. (2) The size of these U-turn regions are randomly chosen from 1 to 8 residues. (3) The U-turns are randomly distributed in the sequence with the constraint that the size of the block between the U-turns should be between 4 and 25 residues. On the basis of the randomly assigned turn regions, the dominant secondary structure between these U-turns is then predicted in the same way as described in Materials and Methods with multiple sequence averaging. Twenty such “random” predictions are performed and averaged for a given sequence. The results for a subset of 8 proteins (2α, 2β, and 4αβ) are shown in Table V. The average turn position error is 2.7 residues (70% of predictions are within 3 residues), and the prediction accuracy of the dominant secondary structure of blocks is 49%, close to a random prediction of either helical or extended blocks. However, there are significant differences among different sequences. It seems that the helical proteins have a higher accuracy of the secondary structure predictions, but a lower accuracy in the U-turn region prediction, presumably due to their longer block sizes. β proteins have smaller errors for the location of the U-turn regions because of their

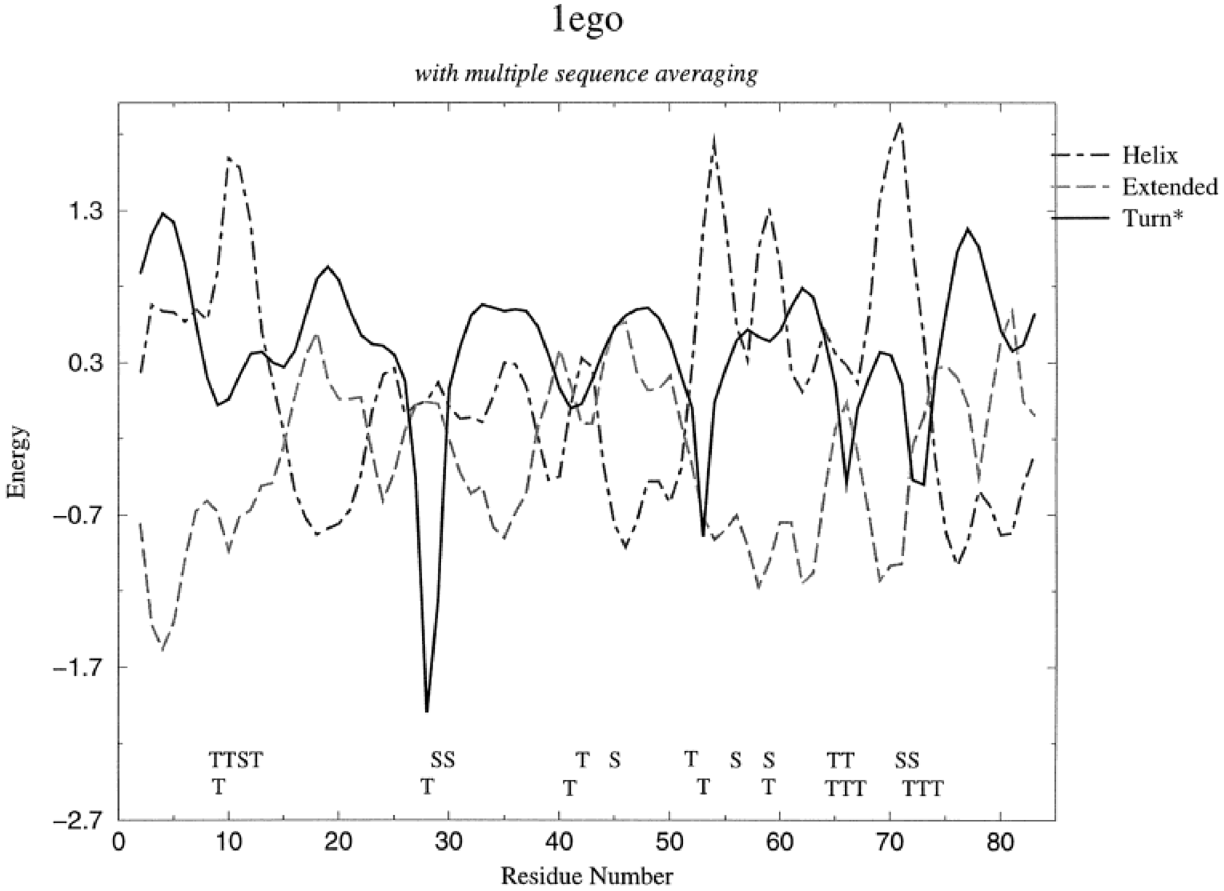

Fig. 7. Energy plot (with multiple-sequence averaging) of a typical αβ protein, 1 ego. The labeling scheme is the same as in Figure 5. The six major turn signals (minima) of the solid curve and a less prominent signal at residue 59 separate the sequence into eight blocks. The three helical regions (the second, fourth, and the last blocks) are clearly resolved. Residues 54–58 are in an extended loop region, and the other four blocks correspond to β-strands. The predicted U-turn regions correlate well with the DSSP assignment.

shorter block sizes. However, the high accuracy in secondary structure prediction does not necessarily follow because in actual β-proteins, the number of U-turns are at the high end of the allowed values. Thus, the chance of underpredicting U-turns is high in random predictions, with blocks often assigned to be helical due to the size constraints. α-β proteins show medium accuracy in the U-turn prediction and lower accuracy in the dominant secondary structure prediction. Speaking overall, the quality of the prediction based on the random assignment of the U-turn regions is significantly inferior to the current approach. It is thus obvious that good prediction of the U-turns is essential for good prediction of dominant secondary structures in the blocks.

CONCLUSION

In the continuing effort to improve our original surface U-turn and block secondary structure prediction algorithm, we have developed a new method that uses a statistical secondary structure propensity, evolutionary information, and some data-processing techniques to produce higher prediction accuracy and efficiency. In the current study of two test sets (total 108) of proteins, approximately 94% of the U-turns are predicted within three residues of that observed in experimentally determined protein structures, with the average error per U-turn of approximately one residue. The prediction accuracy of the dominant secondary structure in the blocks is approximately 88%. Multiple sequence averaging significantly improves the prediction over the single sequence case. The prediction accuracy is very similar whether or not the proteins of interest are part of the training set used to build the potential. Also, the results are very similar for smaller (less than 90 residues) and somewhat larger proteins (less than 150 residues). The success of the current approach indicates that local interactions play important roles in determining the U-turn position and the secondary structure types. Recently, the importance of local interactions in protein folding also has been described in the work of Unger and Moult,[24] Aurora et al.,[25] and Srinivasan and Rose.[26]

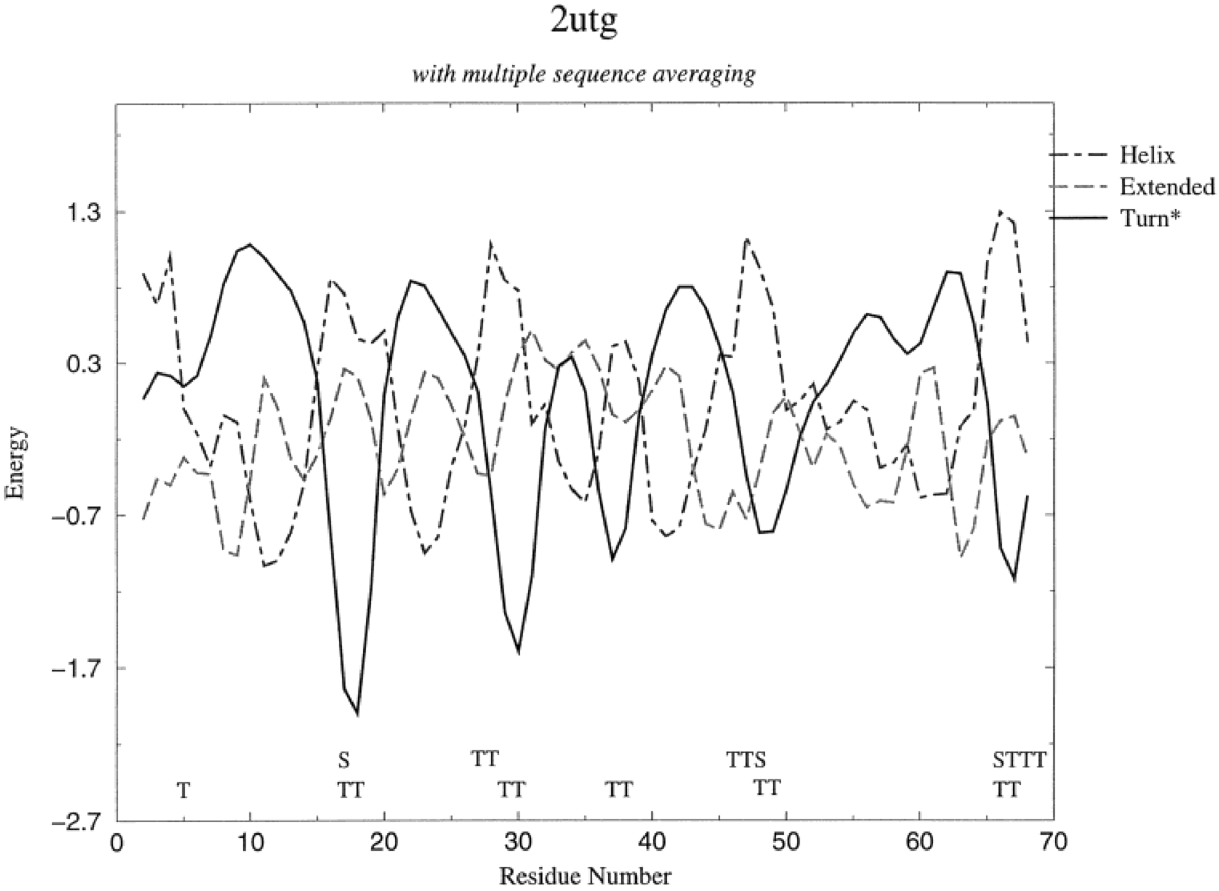

Fig. 8. Energy plot (with multiple-sequence averaging) of an α protein, 2utg (chain A). The labeling scheme is the same as in Figure 5. An overpredicted turn (at residues 37–38) is located in the middle of the third helix, which separates two predicted short helix-preferring regions. The last block is correctly predicted to be a helix by length constraints even though the dashed curve (corresponding to the extended conformations) is lower in most of the region.

Compared with our previous method,[2] both provide comparable accuracy when using a single sequence, but the new algorithm is much more efficient. However, the much simpler implementation of the current method offers the advantage of using information from multiple sequence alignment, which significantly improves the accuracy with a negligible increase in computational cost. The current method has been applied to slightly larger proteins, and it is possible to extend the approach to multiple domain proteins. For the same set of 38 proteins, the current approach that uses multiple sequence information is more accurate than our previous method in all aspects except for slightly more overpredictions of U-turn regions, especially within long helices.

The current prediction scheme is complementary to standard residue-based secondary structure prediction protocols[12–13] because the former concentrates explicitly on U-turn prediction and gives good length-based secondary structure predictions, whereas the latter provides accurate hydrogen bond-based predictions of secondary structure types. In some cases, a hybrid approach could provide higher

TABLE V. Accuracy of Random Predictions[†] on 8 Proteins[‡]

PDB name	Turn[§]	Secondary structure[¶] (%)
1pou	4.1	64
1cnp	3.1	64
1tpm	1.5	63
1aaj	1.8	39
1gb1	2.6	51
1bov A	2.9	39
1ctf	3.7	36
1ubq	1.8	35
Average	2.7	49

[†] Random predictions are performed 20 times for each protein and averaged; see text.

[‡] The first two are α-proteins, the next two are β-proteins, and the rest are αβ proteins.

[§] Average U-turn errors in residues per U-turn.

[¶] Percentage of correctly predicted dominant secondary structure.

accuracy and more information for practical use. Such accurately predicted structural data are crucial for 3-D protein model building and are anticipated to improve the success rate of a newly developed tertiary structure prediction method.[4] The results obtained from the current method also may serve as an additional filter for fold recognition methods. Possible further improvements of our approach include (1) applying more sophisticated pattern recognition techniques to the energy curves to discard the overpredicted turns and to determine more accurately the dominant secondary structure types in the borderline cases; (2) modifying the size constraints to extend the ability to treat multiple domain proteins; and (3) more effective use of homologous protein sequences.

ACKNOWLEDGMENTS

We thank Drs. Adam Godzik, Baohong Zhang, and Li Zhang for helpful discussions. This work was supported in part by NIH grant GM-48835. The research of Andrzej Kolinski was supported in part by an International Research Scholars grant from the Howard Hughes Medical Institute (75195-543402).

REFERENCES

1. Richardson, J. The anatomy and taxonomy of protein structure. Adv. Prot. Chem. 34:167–339, 1981.
2. Kolinski, A., Skolnick, J., Godzik, A., Hu, W.P. A method for the prediction of surface "U"-turns and transglobular connections in small proteins. Proteins 27:290–308, 1997.
3. Skolnick, J., Kolinski, A., Ortiz, A.R. MONSSTER: A method for folding globular proteins with a small number of distance restraints. J. Mol. Biol. 265:217–241, 1997.
4. Ortiz, A.R., Hu, W.P., Kolinski, A., Skolnick, J. A method for prediction of the low resolution tertiary structure of small proteins. J. Mol. Graph., in press, 1997.
5. Kolinski, A., Milik, M., Skolnick, J. A reduced model of short range interactions in polypeptide chains. J. Chem. Phys. 103:4312–4323, 1995.
6. Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J. Turn prediction in proteins using a pattern-matching approach. Biochemistry 25:266–275, 1986.
7. Gonzalez, R.C., Wintz, P. "Digital Image Processing." 2nd. edit. Menlo Park, CA: Addison Wesley, 1987:61–109.
8. Skolnick, J., Jaroszewski, L., Kolinski, A., Godzik, A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci. 6:676–688, 1997.
9. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
10. Kolinski, A., Skolnick, J. Monte Carlo simulation of protein folding. I. Lattice model and interaction scheme. Proteins 18:338–352, 1994.
11. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–58, 1991.
12. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
13. Rost, B., Sander, C. Progress of 1D protein structure prediction at last. Proteins 23:295–300, 1996.
14. Bernstein, F.C., Koetzle, T.F., William, G.J.B., et al. The Protein Data Bank: A computer-based archival file for macromolecule structures. J. Mol. Biol. 112:535–542, 1977.
15. The prediction program and related information can be found in our group World-Wide Web server at "http://moray3.scripps.edu/newll/."
16. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405:442–451, 1975.
17. Rose, G.D. Prediction of chain turns in globular proteins on a hydrophobic basis. Nature 272:589–590, 1978.
18. Rose, G.D., Gierasch, L.M., Smith, J.A. Turns in peptides and proteins. Adv. Protein Chem. 37:1–109, 1985.
19. Wilmot, C.M., Thornton, J.M. Analysis and prediction of the different types of β-turns in proteins. J. Mol. Biol. 203:221–232, 1988.
20. Rooman, M.J., Wodak, S.J., Thornton, J.M. Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. Protein Eng. 3:23–27, 1989.
21. Hutchinson, E.G., Thornton, J.M. A revised set of potentials for β-turn formation in proteins. Protein Sci. 3:2207–2216, 1994.
22. Ohage, E.C., Graml, W., Walter, N.M., Steinbacher, S., Steipe, B. β-turn propensities as paradigms for the analysis of structural motif to engineer protein stability. Protein Sci. 6:233–241, 1997.
23. Chou, P.Y., Fasman, G.D. Empirical predictions of protein conformation. Annu. Rev. Biochem. 47:251–276, 1978.
24. Unger, R., Moult, J. Local interactions dominate folding in a simple protein model. J. Mol. Biol. 259:988–994, 1996.
25. Aurora, R., Creamer, T.P., Srinivasan, R., Rose, G.D. Local interactions in protein folding: Lessons from the alpha-helix. J. Biol. Chem. 272:1413–1416, 1997.
26. Srinivasan, R., Rose, G.D. LINUS: A hierarchic procedure to predict the fold of a protein. Proteins 22:81–99, 1995.