Toward predicting protein topology: An approach to identifying β hairpins

Xavier de la Cruz*[†‡§], E. Gail Hutchinson[†¶], Adrian Shepherd[†], and Janet M. Thornton[‖]

*Institut Català per la Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys, 23, 08018 Barcelona, Spain; [†]Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, United Kingdom; [‡]Department de Bioquimica, C/Marti i Franquès, 1, 08028 Barcelona, Spain; [¶]School of Animal and Microbial Sciences, University of Reading, Whiteknights, P.O. Box 228, Reading, Berkshire RG6 6AJ, United Kingdom; and [‖]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

Communicated by Jane S. Richardson, Duke University Medical Center, Durham, NC, June 24, 2002 (received for review January 1, 2002)

Although secondary structure prediction methods have recently improved, progress from secondary to tertiary structure prediction has been limited. A promising but largely unexplored route to this goal is to predict structure motifs from secondary structure knowledge. Here we present a novel method for the recognition of β hairpins that combines secondary structure predictions and threading methods by using a database search and a neural network approach. The method successfully predicts 48 and 77%, respectively, of all of hairpin and nonhairpin β-coil-β motifs in a protein database. We find that the main contributors to motif recognition are predicted accessibility and turn propensities.

The success of the genome sequencing projects, with their pressing need for fast functional and structural annotations, has renewed interest in predicting the tertiary structure of a protein from its sequence. *Ab initio* approaches are limited by the enormous size of the conformational space and an incomplete knowledge of the interactions that contribute to protein stability (1). These difficulties have prompted the development of alternative approaches involving different simplifications of the problem. The prediction of protein secondary structure is the most widely used of them, with accuracies between 75 and 80% (2–8). These results open the way for the second part of the prediction problem: Can we derive the three-dimensional structure of a protein from the knowledge of its secondary structure?

In recent years, a number of fold recognition methods have addressed this question (9–13). Using the predicted secondary structure of a protein to query the structure database, these methods are able to retrieve a valid structural candidate in 40–60% of cases (10, 13). However, the fold recognition approach is of no use when the query protein has a novel fold. In addition, it is limited by the fact that secondary structure patterns are degenerate (10, 13) and may correspond to different three-dimensional structures.

Secondary structure predictions have also been incorporated into *ab initio* structure prediction methods (14–17). Again the results are encouraging, with several structures predicted in the low-resolution range (14, 15, 18, 19). However, such hybrid methods are still limited to small proteins because of their large computational requirements. In particular, recent studies (20) suggest that it may be difficult to predict the structure of large β proteins by using these methods, even if the native secondary structure of the protein is known. This is because the low-resolution potentials used to speed up the prediction process have to be enriched with hydrogen bonding terms, to properly model large cooperative structures like β barrels (20).

A third approach would follow the hierarchical organization of protein structures and would predict local structural motifs. These could be used in turn to derive the protein tertiary structure, as shown in Fig. 1. Supersecondary structure (SSS) elements are recurrent structural motifs consisting of two or more secondary structure units (21). The interest in their prediction lies in the fact that they are present in the majority of protein structures, particularly in the frequently occurring superfolds (22). Indeed, about 60% of residues involved in secondary structure belong to one of the three simplest structural motifs (22): β, α, and $\beta\alpha\beta$ hairpins. This, together with the fact that they have a relatively small number of possible arrangements (23), strongly suggests that the ability to correctly identify these motifs may help to simplify the structure prediction problem. Actually, identification of structural motifs has been successfully used within the context of *ab initio* protein structure predictions (24, 25).

An Approach to SSS Prediction. Here we present an approach to predicting the SSS of a protein, combining the best secondary structure predictions with threading against a database of tertiary motifs. The rationale behind our method is that in general, any given linear pattern of secondary structures can fold into different tertiary arrangements. However, some of them are very common, and our method seeks to recognize these canonical structures. For example, for the β strand-coil-β strand (βcβ) pattern, 40% of our database of 2,576 patterns fold into β hairpins.

Therefore, we developed a protocol to predict β hairpins, which are simple SSS motifs formed by two adjacent, antiparallel, hydrogen-bonded β strands (26–28). Their simplicity and ubiquity make them good targets for prediction (29–32). In general, because they capture rich three-dimensional structure information, their identification in a protein of unknown structure significantly reduces the number of possible folds available to that protein. More specifically, β hairpin predictions could be used together with low-resolution experimental data, e.g., secondary structure from NMR experiments, to extend the range of the structure determination/prediction process. In particular, correct location of β hairpins within a protein sequence can help in the identification of folds with several adjacent β strands in their secondary structures. For example, both IL-1β, a β trefoil, and the phosphotyrosine recognition domain SH2 of V-SRC, a UB roll, have four β hairpins. However, for the latter, the strands of the consecutive hairpins overlap; that is, the second strand of one hairpin becomes the first strand of the next hairpin, etc. On the contrary, for IL-1β, there is no strand overlap between adjacent hairpins. For all of the above reasons, we wondered whether it is possible, using only local sequence data, to discriminate between those βcβ patterns that form a β hairpin and those that do not.

The protocol developed is summarized in Fig. 2 and described in detail in *Methods*. The approach is hierarchical. First, the secondary structure is predicted. All of the occurrences of the βcβ pattern in a predicted secondary structure are labeled. Each βcβ pattern is then compared with each member of a library of β hairpins with the same number of residues. Every comparison generates 14 scoring terms that are input to a neural network,

Abbreviation: SSS, supersecondary structure.

[§]To whom reprint requests should be addressed. E-mail: xavier@husky.bq.ub.es.

www.pnas.org/cgi/doi/10.1073/pnas.162376199

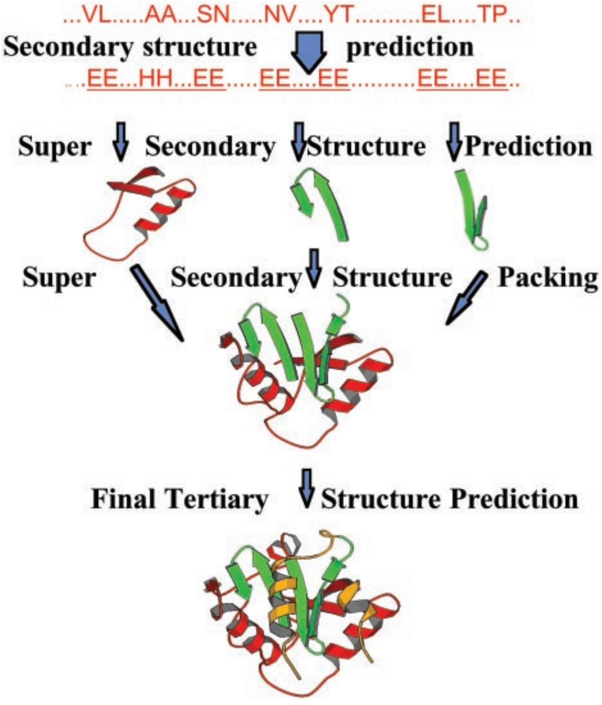
...VL.....AA...SN.....NV....YT..........EL....TP..
Secondary structure prediction
....EE...HH...EE.....EE....EE..........EE....EE..
Super Secondary Structure Prediction
Super Secondary Structure Packing
Final Tertiary Structure Prediction

Fig. 1. A hierarchical approach to protein structure prediction (see text).

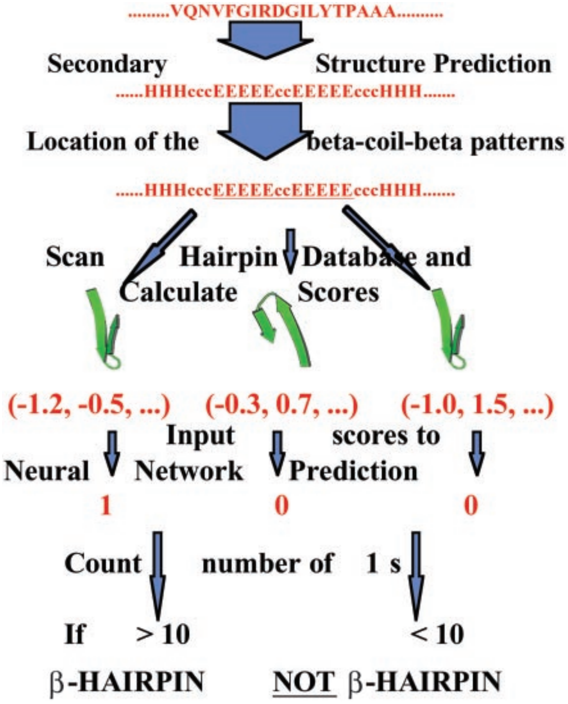
.........VQNVFGIRDGILYTPAAA..........
Secondary Structure Prediction
......HHHcccEEEEEccEEEEEcccHHH.......
Location of the beta-coil-beta patterns
......HHHcccEEEEEccEEEEEcccHHH.......
Scan Hairpin Database and Calculate Scores
(-1.2, -0.5, ...)
(-0.3, 0.7, ...)
(-1.0, 1.5, ...)
Input scores to Neural Network Prediction
1
0
0
Count number of 1 s
If > 10
< 10
β-HAIRPIN
NOT β-HAIRPIN

Fig. 2. The protocol for β hairpin prediction (see *Methods*).

which discriminates between probable hairpins (scoring 1) and nonhairpins (scoring 0). When the complete database has been scanned, if there are more than 10 matches with output 1, we predict the pattern as a β hairpin. Otherwise, the pattern is predicted as a nonhairpin.

Methods

The SSS Prediction Method. The main steps of the method are summarized in Figs. 1 and 2. This method is part of a hierarchical approach to protein structure prediction (Fig. 1) in which secondary structure is predicted in the first place. Then this information is used to predict SSS motifs. Finally, the three-dimensional structure of the protein would be obtained by packing together the predicted SSS elements and modeling the remaining protein residues (although this last part is not addressed herein).

The method for SSS prediction was derived for β hairpins, although it can be easily extended to other SSS motifs. It can be divided into five steps (Fig. 2). First, predict the secondary structure of the protein. Second, label all of occurrences of the $\beta c\beta$ pattern in the prediction. Third, for each pattern found, scan the hairpin database, scoring the database members by using a set of structural and sequence parameters (see below). For each database hairpin, this will give a set of 14 scores. Fourth, these scores are then processed by a neural network that will produce a discrete output, 0′ or 1′, which means that the $\beta c\beta$ pattern is unlikely or likely, respectively, to form a β hairpin. A final filter is applied: if the total number of 1′s (i.e., good matches to a hairpin in the database) is above 10, the $\beta c\beta$ pattern is predicted to be a β hairpin; otherwise, it is assigned to the nonhairpin class.

Protein Set. The protein set was obtained from the CATH (33) list of H representatives. Proteins belonging to the same homology (H) level in the CATH hierarchy have the same fold, a significant degree of sequence similarity, or common functional features. The representatives for each H level are chosen by using structure quality criteria, e.g., highest resolution for x-ray structures, etc. (C. Orengo, personal communication). The list was filtered, so that sequence identity is always lower than 35% between any pair of proteins, excluding any protein with missing residues. The latter was done because the accessibility-based scoring term depends critically on the three-dimensional environment of the residues. Missing residues could lead to incorrect residue accessibility values. Our final protein set was made of 534 proteins.

β Hairpin Database. To build the β hairpin database, we used the PROMOTIF program (34) to locate the hairpins in the 534 proteins. β hairpins having an α helix embedded in the coil fragment were discarded. A total of 1,031 β hairpins were kept. Their length varied between 6 and 52 residues, with an average of 16.9 ± 6.5.

The Ideal and Test Cases. The performance of our method was computed for two different cases, the ideal and test cases. The former was used to set an upper limit for the performance of our method in its present form, by using information of the highest quality, i.e., observed secondary structure was utilized to locate the $\beta c\beta$ patterns. It was also used, together with observed accessibility, to compute the associated scores (see below). For the test case, the information used was that available when dealing with proteins of unknown structure. That is, predicted secondary structure was used to locate the $\beta c\beta$ pattern and, together with predicted accessibility, to compute the associated scores.

Scores based on other properties were also computed. However, because they require only the structure of the database hairpin, they were computed in the same fashion in both the ideal and test cases.

Observed Secondary Structure and Accessibility. For each database hairpin, we took the observed secondary structure and accessibility from the DSSP database (ftp://ftp.embl-heidelberg.de/pub/databases/dssp), downloading the file of the corresponding protein.

PHD Predicted Secondary Structure and Accessibility. These were obtained by using the PHD software (5) (www.embl-heidelberg.de/predictprotein) with default parameters. We used the cross-validation option, so that the target protein was removed from the PHD training set before running the predictions.

Scoring Procedure. Every $\beta c\beta$ pattern found in the secondary structure prediction was compared against each hairpin of the same length in the database. The comparison was based on an ungapped residue by residue alignment between the query pattern and the database hairpin. In this alignment, the first residue in the query pattern matched the first residue in the candidate hairpin, the second matched the second, etc. To evaluate the similarity between the $\beta c\beta$ pattern and the database hairpin, we utilized 14 scores, derived from the previous alignment and based on the following properties: secondary structure (6), accessibility (1), presence of turns (1), specific pair interactions (1), nonspecific distance-based contacts (1), and four properties of the secondary structure pattern related to residue length. It is important to note that computing the scoring terms requires the use of the alignment between the $\beta c\beta$ pattern and the database hairpin, except for the four scoring terms depending on the secondary structure pattern.

To compare the secondary structures of the $\beta c\beta$ pattern and the database hairpin, we first computed three scores: one for each strand and one for the coil. These scores were computed as follows: $\Sigma_i\ \mathrm{ss}_i/n$. Where i runs from 1 to $n1$, $n2$, and $n3$, the lengths of the two β strands and the coil of the database hairpin, respectively. n will be equal to $n1$, $n2$, and $n3$, respectively. The ss_i score corresponds to the de la Cruz and Thornton potential (10), which measures the similarity between the predicted secondary structure of the $\beta c\beta$ pattern and the observed secondary structure of the database hairpin. This score is equal to:

$$\mathrm{ss}_i = \ln\{p[\mathrm{OBSss}_i/(\mathrm{PREDss}_i \cap rl)]\} - 1/3\sum_j \ln\{p[\mathrm{OBSss}_j/(\mathrm{PREDss}_i \cap rl)]\},$$

where PREDss_i is the predicted secondary structure for the ith residue in one of the secondary structure elements of the $\beta c\beta$ pattern (either one of the two strands, or the coil); rl, the reliability of the prediction, as given by PHD (5); and OBSss_i is the observed secondary structure of the corresponding residue in the database hairpin. The sum in the second right-hand term runs over the three possible secondary structure states (helix, β, and coil), and OBSss_j takes the value of each of them. The probabilities were computed by comparing the observed and predicted secondary structures for a set of proteins (10). For the ideal case, because observed secondary structure was used, PREDss_i was replaced by OBSss_i, and rl was made equal to 9, the maximum reliability index.

The whole procedure was repeated by using a Chou and Fassman-like potential (35), to give an additional three scores. These scores are used to measure the likelihood that the $\beta c\beta$ pattern adopts the secondary structure of the database hairpin. They are equal to: $\mathrm{ss}(\mathrm{aa}_i, \mathrm{ss}_i) = f(\mathrm{aa}_i, \mathrm{ss}_i)/\langle f(\mathrm{ss})\rangle$, where aa_i is the residue type found at position i in the query $\beta c\beta$ pattern, ss_i is the secondary structure (helix, β, or coil) of the ith residue in a database hairpin. $\mathrm{ss}(\mathrm{aa}_i, \mathrm{ss}_i)$ is the score of assigning to residue i of the query $\beta c\beta$ pattern the secondary structure ss_i of the ith residue in the database hairpin; $f(\mathrm{aa}_i, \mathrm{ss}_i)$ is the probability of finding a residue of type aa_i in secondary structure ss_i; and $\langle f(\mathrm{ss})\rangle$ is the average of $f(\mathrm{aa}_i, \mathrm{ss}_i)$ over the 20 amino acid types. Both $f(\mathrm{aa}_i, \mathrm{ss}_i)$ and $\langle f(\mathrm{ss})\rangle$ are computed using a database of known protein structures.

Thus a final six scores were generated to measure how well the predicted secondary structure of the query sequence matches the secondary structure of the database hairpin.

The overall similarity between the predicted accessibility of the $\beta c\beta$ pattern and the observed accessibility of the database hairpin was measured by using the following score: $\Sigma_i\ \mathrm{ac}_i/n$. Where n is the total length of the $\beta c\beta$ pattern and ac_i is the value of the de la Cruz and Thornton accessibility potential (10). This potential is equal to:

$$\ln\{p[\mathrm{OBSac}_i/(\mathrm{PREDac}_i \cap rl)]\} - 1/3\sum_j \ln\{p[\mathrm{OBSac}_j/(\mathrm{PREDac}_i \cap rl)]\},$$

where PREDac_i is the predicted accessibility for the ith residue of the $\beta c\beta$ pattern, rl the reliability of the prediction, as given by PHD (5). OBSac_i is the observed accessibility for the ith residue of the database hairpin. The sum in the second right-hand term runs over the three possible accessibility states (exposed, buried, and half-buried), and OBSac_j takes the value of each of them. The probabilities were computed by comparing the observed and predicted accessibilities for a set of proteins (10). For the ideal case, we followed the same procedure as for the secondary structure (see above).

To measure the likelihood that the putative query loop sequence will adopt the turn conformation of the database hairpin, we used a turn score. This score was generated by using the Hutchinson and Thornton potential (36). To this end, we utilized the turn information from the database hairpin. For example, if the database hairpin had a type I turn starting at position 5, the turn score was computed, adding the turn propensities of residues at positions 5–8 in the query sequence. The final score was then divided by four. For hairpins with multiple turns, the final score was obtained by averaging over the number of turns.

The two β strands in a β hairpin are arranged in an antiparallel fashion, displaying a specific pattern of residue pairings between them. It has been recently observed that not all of the residue pairs are equally allowed in antiparallel pairings (37). To evaluate the likelihood that the sequence of the $\beta c\beta$ pattern is consistent with the pairing observed for the database hairpin, we used the Hutchinson *et al.* potential (37). This potential is a pair-specific term that measures the empirical probability that two amino acids will form a "pair" in a β ladder [for a proper definition of pairing, see Hutchinson *et al.* (37)]. To compute this score, we used the information of the pairing pattern of the database hairpin. For example, if residues i and j formed a pair in the matched database hairpin, the score would be equal to the propensity of residues i and j in the $\beta c\beta$ pattern to form such a pair. For hairpins with multiple pairs, the final score was taken as the average.

To evaluate whether the $\beta c\beta$ pattern is likely to be "stable" if adopting a hairpin structure, we utilized a coarse-grained potential: the shell potential by Park *et al.* (38). This potential is a secondary structure unspecific distance-based potential, computed for all of the residue pairs at a $C\beta$-$C\beta$ distance lower than 7 Å ($C\alpha$ atoms are used for Gly residues) and an interresidue distance larger than 1. Scoring a given $\beta c\beta$ pattern required threading the sequence onto the structure of the database hairpin and then computing the overall value of the potential. Because it is based on the use of $C\alpha$ and $C\beta$ atoms, there was no need to model the side chains.

Finally, to measure whether the secondary structure pattern of the $\beta c\beta$ pattern was "hairpin-like," we used four simple parameters: the lengths of the two predicted β strands, that of the predicted coil, and the absolute value of the length difference between both predicted strands.

The Neural Network. A feed-forward network (39) with one input layer, one hidden layer, and one output layer was used. A total of 14 parameters were input to the network—the 14 scoring terms described above. The hidden layer had two units, and the output layer had one unit. The network output was either 1 (a positive β hairpin prediction) or 0.

The network was trained following the procedure described in Shepherd *et al.* (40) presenting the network with a number of inputs, together with their associated target outputs. For nonhairpins, the network was presented with all of the scores resulting from threading the query sequence onto all of the same-sized database hairpins. For β hairpins, only one set of scores was presented to the network, that resulting from self comparison of the sequence with its corresponding β hairpin. This training protocol introduced a bias toward better predictions for nonhairpins. However, it was followed because β hairpins can show large variations for the selected properties (accessibility, etc.). These differences may be comparable to those found between hairpins and nonhairpins and thus could lead to a poorer training of the network.

The neural network performance was tested by using the full prediction procedure. For the true prediction test, no information whatsoever on the observed hairpin structures or $\beta c\beta$ patterns was used. In addition, after training, the performance in the test case was not used to alter the training procedure. The network weights were optimized by using scaled conjugate gradients with 50 iterations (40). The results cited in the present work were obtained by using a 5-fold crossvalidation procedure. The set of 534 proteins was randomly divided into 5 subsets. The five different combinations of four subsets were used to train the network. For each combination, the network was tested on the excluded subset. The results for the test sets were then averaged to provide the results included in Table 1. There was an average 108 β hairpins in the test sets.

A final filter was used to limit false positives. A minimum of 10 neural network outputs equal to 1′ was required for a β hairpin prediction. This number was derived after testing the effect of different values (0, 5, 10, 15) on the averaged performances of the training sets. This filter may result in no β hairpin predictions for the less frequent hairpins.

It must be noted that in the ideal case (when observed secondary structure and accessibility of the query hairpin were used for the scoring process), self-comparison was allowed. That is, the hairpin used to query the database was kept in the database, which was not true for the test case. That is, when testing our method, the β hairpin corresponding to a given $\beta c\beta$ pattern was effectively purged from the database.

Performance Measures. The performance of our procedure was evaluated by using three different parameters: *Qp*, *Qo*, and *S*. Percentage of correct predictions (*Qp*): $Qp = (100\cdot cp)/(cp + ip)$, where *cp* and *ip* were the number of correct and incorrect predictions, respectively. We computed *Qp* independently for hairpin and nonhairpin predictions, using all of the predictions together. In the ideal case, a prediction was taken as correct when: (*i*) the SSS of the query $\beta c\beta$ pattern was correctly identified as hairpin or nonhairpin; and (*ii*) the beginning of the SSS motif was within ±2 residues from the actual location of the pattern. The latter was done to take into account inaccuracies in secondary structure assignment methods (4). It was extended to ±4 residues in the true prediction case to allow for the greater inaccuracies in secondary structure predictions (41). However, ±2 and ±3 shifts gave similar results (results not shown).

Coverage (*Qo*): $Qo = (100\cdot cp)/(cp + np)$, where *cp* is the same as before, and *np* is the number of nonpredicted instances, e.g., when computing *Qo* for the hairpins, np is the number of nonpredicted hairpins.

Finally, we used the normalized performance relative to a purely random prediction method, *S* (40). *S* is defined in such a way as to eliminate overprediction advantages in the nonrandom method. Let *p* and *n* be the number of correct hairpin and nonhairpin predictions, and o and u the number of incorrect hairpin and nonhairpin predictions. For the specific hairpin and nonhairpin cases, *S* is defined as:

$$S(\text{hairpin}) = [p\cdot t - (p + o)\cdot(p + u)]/[t^2 - (p + o)\cdot(p + u)]$$

$$S(\text{nonhairpin}) = [n\cdot t - (n + o)\cdot(n + u)]/[t^2 - (n + o)\cdot(n + u)]$$

where $t = p + n + o + u$.

Table 1. Overall prediction performance

	Ideal case		Test case	
	β hairpin	Non-β hairpin	β hairpin	Non-β hairpin
*Qp**	55.9 ± 2.1	73.6 ± 3.7	47.7 ± 3.9	77.4 ± 2.7
Qo†	64.2 ± 8.3	65.8 ± 6.7	30.1 ± 7.9	87.6 ± 4.5
S‡	34.3 ± 8.3	26.3 ± 3.6	15.9 ± 5.5	28.5 ± 6.3

*Percentage of correct predictions (see *Methods*). Average values are shown, followed by the standard deviations. Both were computed from the results of the cross-validation procedure (see *Methods*).

†Coverage (see *Methods*).

‡Performance relative to random (see *Methods*).

The values of these parameters were computed for those cases for which the $\beta c\beta$ pattern was present, because our prediction method focuses on the ability to assign the proper SSS motif to a given secondary structure pattern.

Results

To test our approach, we distinguish two cases (see *Methods*): the ideal case and the true prediction case. In the former, observed secondary structure and accessibility were utilized, together with other properties, to compute the different scores used by the neural network to produce the hairpin/nonhairpin prediction. This provides a first and simple check of the novel recognition procedure and also sets an upper limit for its performance. The difference for the true prediction case is that predicted, instead of observed, secondary structure and accessibility were used when scoring the match between the query $\beta c\beta$ pattern and the database hairpins.

(*i*) The Ideal Case. In this case, we utilized the observed secondary structure and accessibility to compute the values of the following scores: six secondary structure-based scores, the accessibility score, the lengths of the strands and coil, and finally the absolute value of the difference between both strands sizes. For the remaining terms, i.e., the turn term and both the specific and unspecific contact terms, we used the three-dimensional structure of the database hairpin to derive the associated scores. For example, the unspecific pairing term was computed using the interresidue Cβ-Cβ distances for the database hairpin.

The results obtained (Table 1) show that, on average, 55.9% (±2.1) and 73.6% (±3.7) of the β hairpin and non-β hairpin predictions, respectively, are correct. Despite being smaller, the β hairpin success rate still represents a 34.3% (±8.3) improvement over a random method (Table 1). The coverage was essentially the same for both hairpins and nonhairpins. These results show that if the secondary structure assignment is accurate, hairpin recognition works reasonably well.

We evaluated whether there is a dependence of the prediction rates on the hairpin lengths. To this end, we computed the average *Qp* for all those $\beta c\beta$ patterns (hairpins and nonhairpins together) for which there were at least 15 cases of the given length per crossvalidation set and at least three crossvalidation sets. The resulting *Qp* values varied between 50 and 68%. No clear relationship was observed between *Qp* and pattern length (Fig. 3).

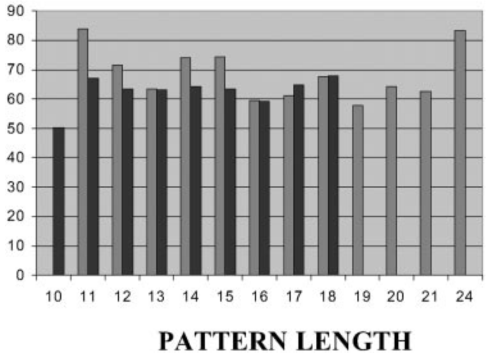

Fig. 3. Length (abscissae) dependence of the overall (β hairpins + nonhairpins) prediction rate of the method (ordinates). Dark gray, ideal case; light gray, test case.

(*ii*) The Test Case. We next tested our procedure by using only sequence information. For example, when studying a given target protein, only predicted secondary structure was used to locate the $\beta c \beta$ patterns, etc. Only 542 of 1,031 database hairpins were found. The remaining 589 hairpins could not be located, because their $\beta c \beta$ patterns were missing from the secondary structure predictions. However, the number of nonhairpins was almost the same, going from 1,545 to 1,495, because spurious $\beta c \beta$ patterns are created by incorrect secondary structure predictions. The prediction problem then becomes more difficult, as the proportion of hairpins relative to nonhairpins decreased from 0.67 to 0.36. The following assessment of hairpin prediction refers only to the 2,037 $\beta c \beta$ patterns found in the secondary structure predictions of the query proteins.

The prediction rate shows a decrease in the hairpin predictions, 47.7% (±3.9) correct predictions, when compared with that of the ideal case, 55.9% (±2.1) correct predictions. The coverage is also smaller, with 30.1% (±7.9) identified hairpins vs. 64.2% (±8.3) in the ideal case. However, the normalized improvement over random, 15.9% (±5.5), is still significant.

The prediction rate for the nonhairpins is comparable to that of the ideal case, 77.4% (±2.7) correct predictions vs. 73.6% (±3.7) in the ideal case, which is also true for the normalized performance relative to random: 28.5% (±6.3) vs. 26.3% (±3.6) in the ideal case.

The length dependence was also studied in this case (Fig. 3). We found that *Qp* varies within a somewhat broader range, 61–84%, but no clear trend could be found. In addition, there are no clear differences between the ideal and test cases.

Discussion

The procedure for SSS prediction is based on the decomposition of the problem in two separate parts: secondary structure prediction and SSS motif identification. This strategy has proved fruitful when used in other structure prediction problems, like turn prediction (40). Our approach is entirely different from the few SSS prediction methods described until now (29–31), which try to predict SSS mainly from sequence. In our case, we first use secondary structure predictions to locate possible SSS motifs. Then, a recognition procedure is used to identify the corresponding motif. By following this decomposition scheme, we can address the first half of the problem by using already available high-performance secondary structure prediction methods. We are then able to concentrate on the identification process.

When evaluating this approach in the test case, we observe that only about half the hairpins were predicted as $\beta c \beta$ patterns in stage 1. For the remaining hairpins, the corresponding $\beta c \beta$ pattern did not exist in the secondary structure prediction. This becomes the major obstacle to our method, emphasizing the importance of improving secondary structure prediction accuracy. In stage 2, we can see that the prediction rates for hairpins and nonhairpins, 47.7% (±3.9) and 77.4% (±2.7) correct predictions, respectively, are better than random (Table 1), which shows that local sequence contains information that helps to determine the local fold, beyond secondary structure. This observation concurs with the work by Wodak and coworkers (32), who found that a local signal helps to determine the conformation of $\alpha\alpha$ turns. Testing our method in the ideal case led to better results in the sense of more balanced predictions: 55.9% (±2.1) and 73.6% (±3.7) β hairpin and nonhairpin successful predictions vs. 47.7% (±3.9) and 77.4% (±2.7) for the test case. However, despite this clear improvement, the fact that the correct prediction rate still was below 100% stresses the relevance of global context in determining local structure.

Table 2. Prediction performance for β hairpins by using only independent scoring terms (see *Discussion*)

	Ideal		Test	
	SS*	AC¶	SS*	AC¶
Qp†	51.2 ± 2.7	60.5 ± 3.0	57.3 ± 8.2	32.5 ± 3.3
Qo‡	17.4 ± 3.3	76.0 ± 3.0	6.7 ± 2.9	28.7 ± 3.6
S§	4.8 ± 4.9	53.0 ± 1.3	3.8 ± 2.0	6.3 ± 3.5

*Results for our procedure when using only the six secondary structure-based terms as input to the NN (see *Methods*).

†,‡,§Same as in Table 1.

¶Same as before, but using the accessibility-based term (see *Methods*).

From the prediction point of view, the results for the test case show a clear improvement over random predictions (Table 1) and open the question of what features of the scoring function led to recognition. To answer this question we trained six separate neural networks by using a subset of scores for each network: (*i*) six secondary structure-based terms; (*ii*) accessibility; (*iii*) turn; (*iv*) specific pair interactions; (*v*) nonspecific distance-based contacts; and (*vi*) four properties of the secondary structure pattern. The training and prediction protocols were repeated as before for each of the networks. For the ideal case, we found that only secondary structure and accessibility-related scores, when considered in isolation, display substantial β hairpin recognition rates, 51.2% (±2.7) and 60.5% (±3.0), respectively, together with high coverages (Table 2). However, for the accessibility score only, the results of our procedure are better than random predictions (Table 2). For the test case, secondary structure scores again show the highest hairpin prediction rate, 57.3% (±8.2). However, as before, these prediction results are close to random predictions (Table 2). Interestingly, in this case, the turn term has the second highest recognition rate, 34% (±4.6), 8.8% (±4.4) better than random, suggesting that turn propensities may compensate for incorrect secondary structure predictions, by giving a lower score to residue stretches incorrectly predicted as coil. This observation is supported by the fact that for the ideal case, the turn term alone did not show any prediction power. In the case of the accessibility score, the prediction rate dropped to 32.5% (±3.3), although it was still better than random predictions.

The results for the accessibility score, both in the ideal and in the test case, indicate that the accessibility pattern of the hairpin is a powerful contributor to recognition. This recognition power probably appears because the accessibility pattern reflects the three-dimensional environment of the hairpin and thus the long-range interactions involved in stabilizing the hairpin. From the prediction point of view, that recognition happens in the test case, where no hairpins from homologues to the target protein were present, shows that hairpins with similar environments may be found in the database. The latter indicates that our procedure can be applied to proteins with no relatives in the protein database.

We see that results obtained when using the different scoring terms independently (Table 2) are always worse than those obtained when using all of them together (Table 1). This fact suggests that an important contribution to the success of our method comes from the correlations between scoring terms identified by the neural network.

Finally, it is interesting to note that no clear trend was observed for prediction rates in relation to the length of the $\beta c\beta$ patterns (Fig. 3). This result was obtained for those pattern lengths, 10–24, for which there was a similar number ($\approx$100) of hairpins in the database. Considering that hairpins of different sizes may be very different, the latter suggests that our scoring scheme is able to correctly identify some common characteristics between patterns, independent of their lengths. Thus only marginal benefits would be obtained from improving our method by using size-dependent neural networks. This independence of size is relevant in the sense that the prediction problem is mostly reduced to a SSS motif recognition problem only.

Future Improvements

The final goal of SSS predictions is to pave the way for tertiary structure predictions (Fig. 1). In this sense, our results support what we believe is a promising approach to the problem. The previous analysis suggests that future improvements may come from three different directions: better secondary structure and accessibility predictions, improvements in the prediction procedure, and growth in the database. The former will increase the rate of recovered hairpin $\beta c\beta$ patterns in the first step of our prediction protocol.

The prediction procedure can be improved by replacing the 10 1′s filter (see *Methods*) by a second neural network. The latter would be used to assess the likelihood that a given $\beta c\beta$ pattern corresponds to a nonhairpin, following a procedure analogous to that presented here. $\beta c\beta$ patterns giving a weak β hairpin signal and a strong nonhairpin signal could be predicted as nonhairpins more reliably.

Finally, improvements may come from enlarging the hairpin database (10). Increasing the database size will provide a better sampling of the hairpin space, thus increasing the number of cases for which a correct hit can be found. Adding hairpins from different representatives of the same protein family, with low intermotif sequence identity, is a possible way to expand our database.

We are at present exploring these options as well as the extension of the method to other SSS motifs.

X.d.l.C. acknowledges the support of a "Human Frontier Science Program" fellowship. He also acknowledges Dr. B. Rost for the loan of a copy of the program PHD. E.G.H. was funded by Biotechnology and Biological Sciences Research Council (BBSRC) Grant 31/9708971. This is a publication from the Bloomsbury Center for Structural Biology, funded by the BBSRC Grant 31/JRI07365 for computing.

1. Dill, K. A. (1990) *Biochemistry* **29,** 7133–7155.
2. Petersen, T. N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G. P. & Lund, O. (2000) *Proteins* **41,** 17–20.
3. Jones, D. T. (1999) *J. Mol. Biol.* **292,** 195–202.
4. Cuff, J. A. & Barton, G. J. (1999) *Proteins* **34,** 508–519.
5. Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232,** 584–599.
6. King, R. D. & Sternberg, M. J. (1996) *Protein Sci.* **5,** 2298–2310.
7. Salamov, A. A. & Solovyev, V. V. (1997) *J. Mol. Biol.* **268,** 31–36.
8. Frishman, D. & Argos, P. (1997) *Proteins* **27,** 329–335.
9. Simons, K. T., Strauss, C. & Baker, D. (2001) *J. Mol. Biol.* **306,** 1191–1199.
10. de la Cruz, X. & Thornton, J. M. (1999) *Protein Sci.* **8,** 750–759.
11. Jones, D. T., Tress, M., Bryson, K. & Hadley, C. (1999) *Proteins* **37,** 104–111.
12. Panchenko, A. R., Marchler-Bauer, A. & Bryant, S. H. (1999) *Proteins* **37,** 133–140.
13. Rost, B., Schneider, R. & Sander, C. (1997) *J. Mol. Biol.* **270,** 471–480.
14. Ortiz, A. R., Kolinski, A. & Skolnick, J. (1998) *J. Mol. Biol.* **277,** 419–448.
15. Dandekar, T. & Argos, P. (1996) *J. Mol. Biol.* **256,** 645–660.
16. Abagyan, R. & Totrov, M. (1994) *J. Mol. Biol.* **235,** 983–1002.
17. Kang, H. S., Kurochkina, N. A. & Lee, B. (1993) *J. Mol. Biol.* **229,** 448–460.
18. Bowie, J. U. & Eisenberg, D. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 4436–4440.
19. Monge, A., Friesner, R. A. & Honig, B. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5027–5029.
20. Eyrich, A. V., Standley, D. M. & Friesner, R. A. (1999) *J. Mol. Biol.* **288,** 725–742.
21. Rao, S. T. & Rossmann, M. G. (1973) *J. Mol. Biol.* **76,** 241–256.
22. Salem, G. M., Hutchinson, E. G., Orengo, C. A. & Thornton, J. M. (1999) *J. Mol. Biol.* **287,** 969–981.
23. Taylor, W. R. & Thornton, J. M. (1983) *Nature (London)* **301,** 540–542.
24. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997) *J. Mol. Biol.* **268,** 209–225.
25. Jones, D. T. (2001) *Proteins* **45,** 127–132.
26. Sibanda, B. L. & Thornton, J. M. (1985) *Nature (London)* **316,** 170–174.
27. Milner-White, E. J. & Poet, R. (1986) *Biochem. J.* **240,** 289–292.
28. Efimov, A. V. (1993) *Prog. Biophys. Mol. Biol.* **60,** 201–239.
29. Godzik, A. & Skolnick, J. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 12098–12102.
30. Jenny, T. F., Gerloff, D. L., Cohen, M. A. & Benner, S. A. (1995) *Proteins* **21,** 1–10.
31. Sun, Z., Rao, X., Peng, L. & Xu, D. (1997) *Protein Eng.* **10,** 763–769.
32. Wintjens, R. T., Rooman, M. J. & Wodak, S. J. (1996) *J. Mol. Biol.* **255,** 235–253.
33. Pearl, F. M. G., Lee, D., Bray, J. E., Sillitoe, I., Todd, A. E., Harrison, A. P., Thornton, J. M. & Orengo, C. A. (2000) *Nucleic Acids Res.* **1,** 277–282.
34. Hutchinson, E. G. & Thornton, J. M. (1996) *Protein Sci.* **5,** 212–220.
35. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13,** 211–222.
36. Hutchinson, E. G. & Thornton, J. M. (1994) *Protein Sci.* **3,** 2207–2216.
37. Hutchinson, E. G., Sessions, R. B., Thornton, J. M. & Woolfson, D. N. (1998) *Protein Sci.* **7,** 2287–2300.
38. Park, B. H., Huang, E. S. & Levitt, M. (1997) *J. Mol. Biol.* **266,** 831–846.
39. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, eds. Rumelhart, D. E. & McClelland, J. L. (MIT Press, Cambridge, MA), pp. 318–362.
40. Shepherd, A. J., Gorse, D. & Thornton, J. M. (1999) *Protein Sci.* **8,** 1045–1055.
41. Rost, B., Sander, C. & Schneider, R. (1994) *J. Mol. Biol.* **235,** 13–26.