# Strand-Loop-Strand Motifs: Prediction of Hairpins and Diverging Turns in Proteins

**Michael Kuhn,[1] Jens Meiler,[2] and David Baker[2]***
[1]*California Institute of Technology, Pasadena, California*
[2]*Department of Biochemistry, University of Washington, Seattle, Washington*

***ABSTRACT*** **β-sheet proteins have been particularly challenging for de novo structure prediction methods, which tend to pair adjacent β-strands into β-hairpins and produce overly local topologies. To remedy this problem and facilitate de novo prediction of β-sheet protein structures, we have developed a neural network that classifies strand-loop-strand motifs by local hairpins and nonlocal diverging turns by using the amino acid sequence as input. The neural network is trained with a representative subset of the Protein Data Bank and achieves a prediction accuracy of 75.9 ± 4.4% compared to a baseline prediction rate of 59.1%. Hairpins are predicted with an accuracy of 77.3 ± 6.1%, diverging turns with an accuracy of 73.9 ± 6.0%. Incorporation of the β-hairpin/diverging turn classification into the ROSETTA de novo structure prediction method led to higher contact order models and somewhat improved tertiary structure predictions for a test set of 11 all-β-proteins and 3 αβ-proteins. The β-hairpin/diverging turn classification from amino acid sequences is available online for academic use (Meiler and Kuhn, 2003; www.jens-meiler.de/turnpred.html). Proteins 2004;54:282–288.**


**Key words: artificial neural network; protein secondary structure prediction; β-hairpin; ROSETTA; protein tertiary structure prediction; fragment replacement**

## INTRODUCTION

The output of genome-sequencing projects is far greater than the output of experimental protein structure determination. As of September 2002, the Protein Data Bank (PDB) contained 16,921 protein structures, compared to 114,033 sequence entries in SWISS-PROT.[2,3] Hence, there is a great need for accurate protein structure prediction methods. Protein structure prediction has been addressed on multiple levels, from secondary structure prediction through supersecondary structure motif recognition to the prediction of three-dimensional structures. The prediction of secondary structure with artificial neural networks has a long tradition[4–8] with the most powerful recent methods using neural networks on multiple-sequence alignments produced by PSIBLAST. Some approaches to predict supersecondary motifs have aimed at predicting a number of different motifs at once,[9–11] whereas others focus on special structures such as β-turns.[12–14]

The information gained by the prediction of supersecondary structure can be used in tertiary structure prediction (e.g., in fold recognition[15,16] and structure assembly[17]). ROSETTA,[18] one of the most successful current approaches to tertiary structure prediction,[19] generates a distribution of plausible local conformations for each segment of the chain by searching the PDB for fragments with similar local sequences. These fragments define the accessible conformational space of the sequence at a particular position. The accessible conformational space of the complete chain is searched by a Monte Carlo algorithm. At a randomly selected position of the chain, one randomly selected fragment from the fragment list at this particular position is inserted and the change in the low-resolution energy function dominated by hydrophobic burial and strand pairing is evaluated. Structure predictions are made carrying out many independent simulations and detecting broad minima on the energy landscape by cluster analysis.

ROSETTA has difficulties generating accurate models for proteins with primarily nonlocal contacts. A measure of the balance between local and nonlocal contacts is the contact order, which is defined to be the average sequence separation of the amino acids that are in contact. Indeed, populations of structural models generated by ROSETTA contain an excess of low-contact order structures.[20] Contacts between residues close in the sequence can be identified relatively rapidly and are not disrupted by most subsequent fragment insertions; thus, low-energy local interactions are formed and maintained at the expense of nonlocal energy minima. That there is a correlation between contact order and protein-folding rates: proteins with a high-contact order fold more slowly than low-contact order proteins.

A structural motif that is formed in excess in ROSETTA simulations is the β-hairpin. To minimize the energy, the optimization algorithm often pairs adjacent β-strands to form a strand-hairpin-strand motif. In ROSETTA models, about 80% of all adjacent β-strands are connected by a hairpin, whereas this ratio is close to 60% in native

Grant sponsor: Howard Hughes Medical Institute.

*Correspondence to: David Baker, Department of Biochemistry, University of Washington, Box 357350, Seattle, WA 98195-7350. E-mail: dabaker@u.washington.edu

Received 22 May 2003; Accepted 7 July 2003

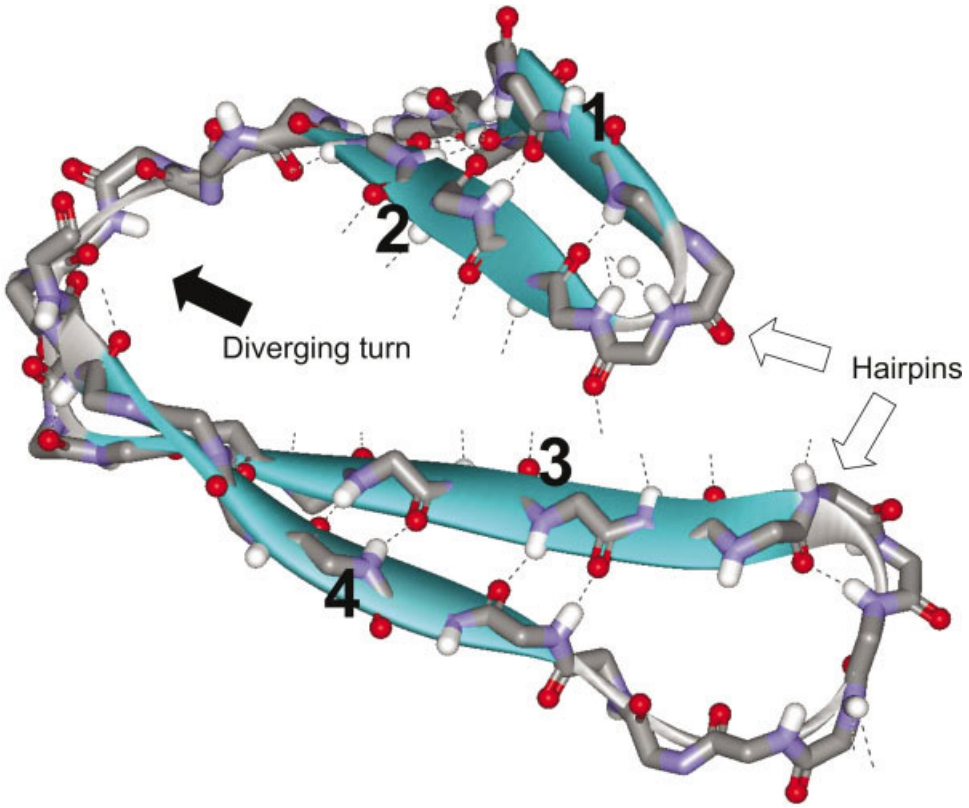


Fig. 1. Illustration of hairpins and diverging turn. The first four strands of the immunoglobulin fold 1fna_ are shown. Strands 1 and 2 and strands 3 and 4 form hairpins while the turn between strands 2 and 3 is a diverging turn.

structures. If it were possible to predict whether a hairpin is present, the formation of hairpins could be selectively disallowed. By reducing the degrees of freedom in the conformational space, an incentive for the formation of long-range contacts could be provided. This could increase the yield of higher contact order structures with potentially the correct overall topology.

Toward this end, in this article we develop a neural network-based method for predicting whether a strand loop strand motif adopts a hydrogen-bonded β-hairpin or a non-hydrogen-bonded diverging turn in which the two strands are paired with other strands (see Fig. 1). The predictions are then incorporated into the ROSETTA tertiary structure prediction method by penalizing the formation of β-hairpins in regions predicted to form diverging turns. Although the network overpredicts turns, it is quite successful in distinguishing between hairpins and diverging turns, and the predictions increase the contact order and accuracy of ROSETTA tertiary structure predictions.

## MATERIALS AND METHODS

### Data Sets and Turn Classification

A database of proteins was compiled with structures of resolution $> 2.5$ Å and $< 50\%$ sequence identity.[21] It contains 2209 proteins, with a total of ~550,000 residues. Turns in the database are detected and classified by using DSSP[22] output files. These contain secondary structure information and β-bridge partners of β-strand residues. A chain segment connecting two β-strands is considered a turn if it does not contain any β-strand or α-helix residues. If the adjacent strands are connected by one or more hydrogen bonds, the turn is classified as hairpin. Otherwise, if no β-bridges are recognized, the turn is considered to be a diverging turn. With this limitation, 5151 hairpins and 3905 diverging turns are detected in the database.

### Prediction Method

Because the sequence signals at the N- and C-termini of turns have different properties and turns have variable lengths, we decided to build two separate neural networks for predicting the state of the first residue in a turn and the state of the last residue, respectively (Fig. 2). Each ANN predicts whether the considered residue is the first/last residue of a hairpin, diverging turn, or neither.

The input window of both neural networks contains 12 amino acids. The first network is trained to predict "beginning of a turn" if the true beginning of the turn is at the fifth position of the input window (subscript "b" in the following discussion). The second network predicts "end of turn" if the true end of the turn is at the eighth (fifth to last) position in the sequence window (subscript "e"). This ensures that turns up to the length of eight are presented completely in the input window. For example, if the

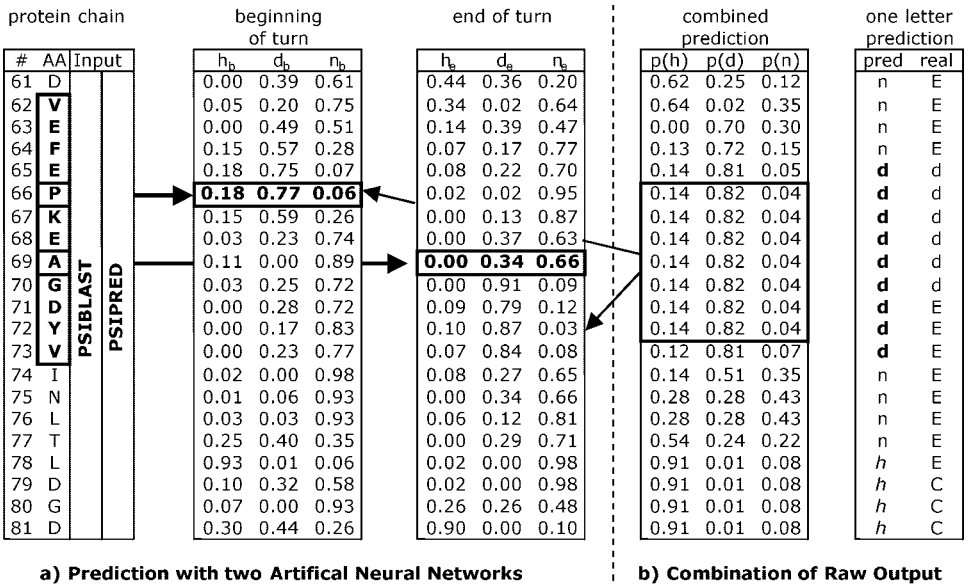


Fig. 2. Illustration of the prediction procedure (residues 61–81 of 1ksr). **a:** The box around the amino acid identifiers represents the input window that is moved over the protein chain. The PSIBLAST profile and PSIPRED secondary structure prediction are used as input for the Artificial Neural Networks that predict the beginning and end of turns. The fifth residue of the input window is predicted as potential beginning of a turn, and the eighth residue as potential end of a turn. This is done for the whole sequence. $n_b$ is the predicted probability that a residue is not the beginning of a turn, $n_e$ the probability that a residue is not the end of a turn. **b:** Raw predictions are combined to yield a three-state prediction (hairpin, diverging turn, and no turn); for details see text. On the right side, the one-letter prediction is compared with the real (super) secondary structure. The real diverging turn is correctly identified (bold type); however, a non-existing hairpin is predicted (italic type).

beginning of a two-residue hairpin is considered, the window would consist of four-strand residues, followed by the two-turn residues and six residues of the adjacent strand. If the frame is shifted so that another residue of the turn is to be predicted, the ANN is supposed to predict "no turn."

A total of 23 numerical inputs is used for each amino acid: the PSIBLAST profile (position-specific score matrices, 20 inputs, as used by Jones for secondary structure prediction)[23] and the PSIPRED secondary structure prediction (three inputs).[7] PSIPRED also uses the PSIBLAST profile as input but is trained over a large data set containing all secondary structure types and thus additionally provides valuable information for the turn prediction.

Both ANNs are of feed-forward architecture and have $12 \times 23 = 276$ inputs, 15 hidden neurons, and 3 output neurons representing the probabilities for hairpin (h), diverging turn (d), and no turn (n). The networks were trained by back-propagation[24] with a learning rate of 0.0001 and a momentum term of 0.5 with a training set of 1800 proteins. Training was stopped when the performance for a monitoring set of 200 proteins declined. The number of hidden neurons was optimized to yield the best performance in predicting the monitoring set of data.

Subsequently, the per residue predictions $h_b(i)$, $d_b(i)$, $n_b(i)$ and $h_e(i)$, $d_e(i)$, $n_e(i)$ are combined to derive turn probabilities p(x, i…j) for turn type x beginning at residue i and end residue j:

$$p(n, i\ldots j) = [n_b(i) + n_e(j)]/2$$

$$p(h, i\ldots j) = [1 - p(n, i\ldots j)] \cdot [h_b(i) + h_e(j)]/[h_b(i) + h_e(j) + d_b(i) + d_e(j)]$$

$$p(d, i\ldots j) = [1 - p(n, i\ldots j)] \cdot [d_b(i) + d_e(j)]/[h_b(i) + h_e(j) + d_b(i) + d_e(j)]$$

For all possible turn lengths of up to eight residues in the protein sequence, these probabilities are calculated by averaging the results of the prediction for the first and last residue of the turn (36 possibilities for each residue). This averaged probability is assigned to all residues in the turn. For each residue, the highest turn probability assignment of those 36 possibilities is kept as the final output of the prediction method (Fig. 2).

In a final step, a one-letter code is derived by assigning the small letters *h* or *d* to amino acids with predicted hairpin or diverging turn probabilities $> 0.75$. All other amino acids remain in the state *n*. Note that not the one-letter code but the three-state probabilities are used for all further calculations.

Other approaches, such as the prediction of all turn residues instead of just the first and last residue and training specialized networks for specific turn lengths, were tested. Isolated results were promising; however, it was not possible to combine predictions for the various turn lengths. It also proved difficult to predict the length of turns with a single neural network.

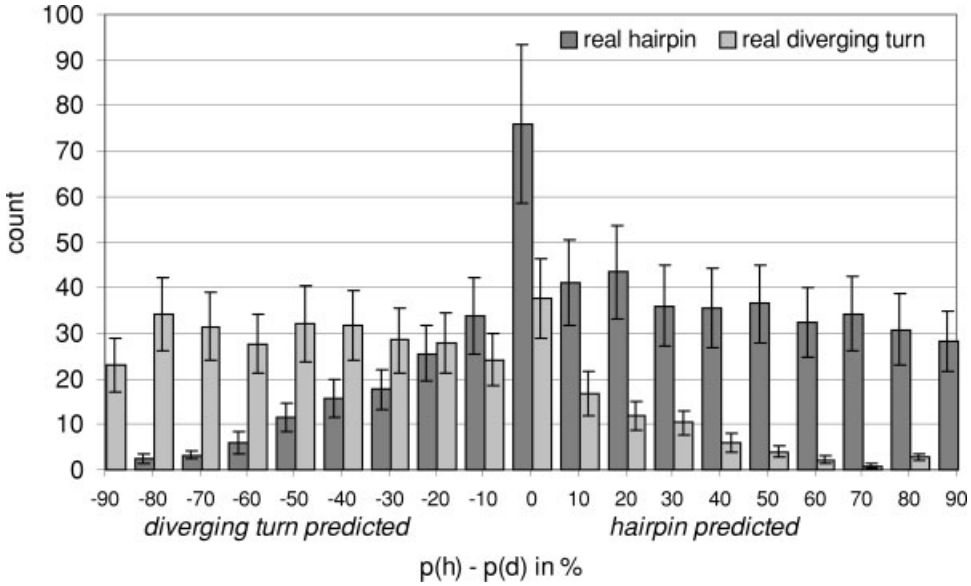


Fig. 3. Histogram of prediction confidence showing the relationship between the confidence of the prediction and the prediction accuracy. The *x* axis is the hairpin probability minus the diverging turn probability. It is evident that the greater the bias of the prediction toward one of the two turn types, the smaller the chance of misclassification. Error bars are the standard deviation across 10 evaluations of the independent data set with differently trained prediction networks. The large peak corresponding to a difference of zero results both from ambiguous turn sequences and real turns that were not detected.

### Assessment of Prediction Performance

The accuracy of the prediction was assessed by cross-validation: The complete data set (2000 proteins) was split into 10 parts, and each of these was used to monitor training over the remaining 1800 proteins. All ANN were tested against the monitoring set they had been trained with and against an independent data set containing 209 proteins. The prediction networks used for tertiary structure prediction were trained with the whole training data set, using the independent data set as monitoring set.

The classification of turns was evaluated by averaging the prediction results over all turn residues. In this fashion, all real turns were classified as either hairpin or diverging turn. Two performance measures were calculated: accuracy and confidence. Accuracy ($Q_a$) is defined as the fraction of real turns that have been correctly classified. Confidence ($Q_c$) is the fraction of predicted turns that are correct predictions. In the example of hairpin accuracy and confidence, if c is the number of correctly classified hairpins, d the number of hairpins that have been classified as diverging turn, and h the number of diverging turns that have been predicted as hairpin, $Q_a = c/(c + d)$ and $Q_c = c/(c + h)$.

### Scoring Function

To score structural models generated with the ROSETTA program, hairpins that are predicted by ROSETTA are penalized with the goal of reducing the excess of wrongly predicted hairpins. First, β-strands separated by less than nine turn residues are identified by using secondary structure information from the fragments used to generate the structure. If a conformation contains a hairpin in a region predicted to be diverging turn, a penalty is added to the standard ROSETTA score.[25] The probability that the move that created the hairpin will be accepted is thereby reduced. The penalty is proportional to the length of the shorter strand and the difference between the two turn probabilities. With $l_1$ and $l_2$ the lengths of the adjacent strands and p(h) and p(d) the predicted probabilities for hairpin/diverging turn, the penalty is $\min(l_1, l_2) \cdot \max(0, p(d) - p(h))$. Using the difference between the probabilities for diverging turn and hairpin is based on the fact that a larger difference corresponds to a greater confidence (see Fig. 3). On the other hand, if a hairpin is predicted as hairpin, no score is given, so as not to encourage the formation of hairpins.

### Evaluation of the Hairpin Penalty

Fourteen proteins were used to explore the possibility of using the hairpin/diverging turn prediction in protein structure prediction. These proteins are all-β- and αβ-proteins out of a data set that is used to benchmark ROSETTA. Care was taken to remove close homologues (>50% sequence identity) from the set of proteins used for training the neural networks. There are 11 all-β- and three αβ-proteins, with sequence lengths between 58 and 121.

Two sets of 10,000 structural models were generated for each protein, using ROSETTA both with and without the hairpin penalty. In preliminary tests, the weight of the score was varied over a wide range to find an optimal balance relative to the standard energy function.

The 10,000 generated models are split up into 10 sets of models containing 1000 structures. For each of these 10 sets, first-percentile root-mean-square deviation (RMSD)

**TABLE I. Prediction Results by Residue**

| | Monitoring data set prediction (%) | | Independent data set prediction (%) | |
|---|---|---|---|---|
| Structure | Turn | No turn | Turn | No turn |
| Turn | 6.4 ± 0.7 | 0.4 ± 0.1 | 7.1 ± 0.2 | 0.5 ± 0.0 |
| No turn | 32.2 ± 2.0 | 61.0 ± 4.8 | 34.7 ± 1.3 | 57.7 ± 0.8 |
| | Hairpin | Diverging turn | Hairpin | Diverging turn |
| Hairpin | 40.0 ± 5.2 | 15.1 ± 1.5 | 41.8 ± 0.7 | 17.0 ± 0.6 |
| Diverging turn | 9.0 ± 1.4 | 36.0 ± 2.9 | 8.7 ± 0.9 | 32.5 ± 0.9 |

**TABLE II. Prediction Results by Turn**

| | Monitoring data set prediction (%) | | Independent data set prediction (%) | |
|---|---|---|---|---|
| Structure | Hairpin | Diverging turn | Hairpin | Diverging turn |
| Accuracy | 78.2 ± 6.5 | 74.4 ± 6.9 | 77.3 ± 6.1 | 73.9 ± 6.0 |
| Confidence | 79.9 ± 6.7 | 72.4 ± 6.7 | 81.1 ± 6.4 | 69.3 ± 5.6 |

to native and the fraction of structural models within 5% of native contact order were determined as quality measures. Because the creation of structural models with the Monte Carlo method is a random process, variations in the prediction results are expected. Calculating the standard deviation over different prediction results makes it possible to judge whether the observed changes of the results are significant.

## RESULTS

### Prediction of Turns

Predictions were made for turns of one to eight residues, which covers ~75% of all strand-loop-strand patterns. Both the length and type of turns were determined by using DSSP.[22] Thus, a turn is regarded as a hairpin if DSSP recognizes residues in the adjacent β-strands as being connected by β-bridges. Otherwise, the turn is considered as diverging turn.

Ideally, the prediction of turns could be used to detect the location of turns and to classify the detected turns. As indicated in Table I, only the second aim was achieved. Given any real turn, it is correctly classified as hairpin or diverging turn with 75.9 ± 4.4% probability (Table II). This is well above the "baseline" of 59.1%, the fraction of turns that would be correctly "classified" if all turns were predicted as hairpins. Hairpins are predicted with an accuracy of 77.3 ± 6.1% and diverging turns with 73.9 ± 6.0%. Figure 3 shows the trade-off between confidence and the fraction of turns that are predicted with that confidence.

Despite the good distinction rate between turns, too many turns are predicted. This bias was introduced purposely to ensure that all real turns are reliably classified. During tertiary structure prediction, all turn regions present in the native structure need to be scored correctly. For this reason, it is important to detect as many native turns as possible. This is achieved at the cost of predicting too many turns. However, this overprediction has little negative effect on tertiary structure prediction: predicted turns are solely used to penalize modeled hairpins that are in the wrong place. Thus, the only effect of overpredicting turns is to discourage the formation of hairpins in places where the native structure has no turn at all.

The rather high-sequence identity level of 50% within the protein database used for training the neural networks does not bias the prediction toward higher success rates as is shown by repeating the calculations with a sequence identity cutoff of 25%. Besides the somewhat larger standard deviation (as expected for smaller data sets), the overall success rate remains with 74.6 ± 5.3% unchanged within the limits of the standard deviation.

### Protein Structure Prediction

Based on the prediction of hairpins and diverging turns, a scoring function was developed and used during protein structure prediction with ROSETTA. Although the prediction classifies turns with high accuracy, the location of turns is unclear. Structural models for the same protein often differ in secondary structure, both in location and type of the secondary structure segments. It proved to be disadvantageous to reward formation of turns in regions they are predicted, because this provided an incentive for the formation of turns. Structural models generated with this full scoring indeed contained an excess of turns: extended loop regions with ambiguous secondary structure prediction were often turned into β-strands. For this reason, the scoring was limited to hairpins. This finding addresses a major shortcoming of ROSETTA: the overprediction of hairpins. Every hairpin in the given structural model receives a penalty if a diverging turn is predicted for that location. The score is proportional to the confidence of the prediction and the length of the shorter β-strand (see Materials and Methods).

Table III shows a comparison between models generated by using standard ROSETTA and ROSETTA with the hairpin penalty. The RMSD to native in the test set of 11 all-β-proteins is improved for five proteins (1aboA, 1fna, 1gvp, 1ksr, and 2ncm, improvements from 0.5 to 2.2 Å), is unchanged for four proteins (1c9oA, 1danT, 1tuc, and 1tul, with deviations up to 0.3 Å), and becomes slightly worse for two proteins (1vie and 1who) with poor hairpin predictions. In αβ-proteins, a slight (but still significant) improvement is observed for 2sak and 2tgi. 4ubpB does not change significantly. Even though this test set is far from comprehensive, the improvements in the predictions are encouraging.

The new method penalizes the formation of hairpins in specific parts of the chain. This would be expected to reduce the excess of overly local structures consisting primarily of hairpins. Indeed, with the hairpin penalty, higher contact order (CO) models are generated than with standard ROSETTA (Fig. 4). As shown in Table III, a significant increase in the fraction of models within 5% of the native CO is observed in all but two cases.

**TABLE III. Effect of Hairpin Penalty on ROSETTA Simulations**

| Protein | | | | ANN Predicted/Real | | Results of structure prediction with ROSETTA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | RMSD to native (Å)[a] | | Models with native CO[b] | |
| PDB | N | %α | %β | hairpin | diverging turn | Plain ROSETTA | With hairpin score | Plain ROSETTA | With hairpin score |
| 1aboA | 58 | 5% | 50% | 2/2 | 1/1 | 7.9 ± 0.3 | 5.7 ± 0.3 | 17.6 ± 0.8% | 24.4 ± 1.6% |
| 1c9oA | 66 | 5% | 62% | 3/3 | 0/0 | 3.8 ± 0.1 | 3.9 ± 0.1 | 27.1 ± 1.7% | 27.4 ± 1.1% |
| 1danT | 75 | 5% | 57% | 2/2 | 2/2 | 8.8 ± 0.2 | 8.9 ± 0.2 | 13.2 ± 1.4% | 13.3 ± 1.0% |
| 1fna_ | 91 | 0% | 46% | 1/2 | 2/2 | 6.2 ± 0.3 | 5.5 ± 0.3 | 8.6 ± 0.8% | 13.8 ± 1.1% |
| 1gvp_ | 87 | 7% | 46% | 2/3 | 3/3 | 9.7 ± 0.2 | 9.0 ± 0.3 | 20.6 ± 1.4% | 27.6 ± 1.2% |
| 1ksr_ | 92 | 0% | 41% | 1/1 | 2/2 | 8.3 ± 0.1 | 7.8 ± 0.2 | 16.0 ± 1.1% | 21.7 ± 1.2% |
| 1tuc_ | 61 | 5% | 44% | 2/3 | 0/0 | 5.6 ± 0.2 | 5.4 ± 0.2 | 33.4 ± 1.3% | 37.4 ± 1.6% |
| 1tul_ | 102 | 7% | 51% | 1/1 | 4/5 | 10.0 ± 0.2 | 9.7 ± 0.3 | 0.5 ± 0.2% | 1.4 ± 0.5% |
| 1vie_ | 60 | 5% | 43% | 0/3 | 1/1 | 7.1 ± 0.2 | 7.4 ± 0.2 | 8.9 ± 0.7% | 19.2 ± 1.1% |
| 1who_ | 94 | 0% | 50% | 2/4 | 1/4 | 7.7 ± 0.2 | 8.1 ± 0.2 | 3.5 ± 0.4% | 6.7 ± 0.8% |
| 2ncm_ | 96 | 0% | 55% | 3/3 | 5/5 | 8.5 ± 0.5 | 7.5 ± 0.3 | 0.2 ± 0.1% | 2.6 ± 0.7% |
| 2sak_ | 121 | 10% | 45% | 2/2 | 3/3 | 13.2 ± 0.2 | 12.5 ± 0.2 | 0.7 ± 0.3% | 1.9 ± 0.3% |
| 2tgi_ | 112 | 21% | 41% | 1/1 | 4/4 | 13.3 ± 0.1 | 12.5 ± 0.2 | 8.8 ± 0.9% | 12.9 ± 1.0% |
| 4ubpB | 103 | 17% | 32% | 0/1 | 4/5 | 11.1 ± 0.3 | 11.1 ± 0.2 | 7.8 ± 1.5% | 10.5 ± 0.8% |

[a]RMSD to native Cα position of the 10th best model out of 1000 generated.
[b]Fraction of models that have a contact order with in a range of ±5% of the contact order of the native fold.

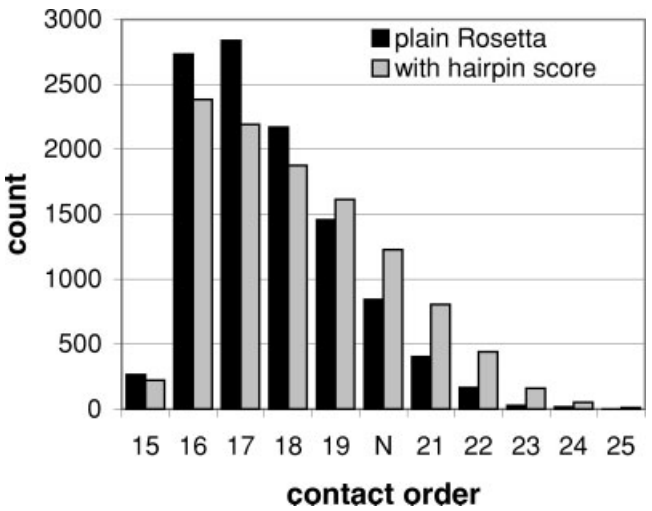


Fig. 4. Distribution of contact order in 10,000 generated structural models for 1abo_. The caption "N" on the abscissa corresponds to the native contact order of 20.3. The hairpin penalty enhances sampling of structures with higher contact order.

## DISCUSSION

It is unclear to what extent our method for predicting whether a strand-loop-strand segment will adopt a β-hairpin or a diverging turn can be improved. The major limitation is its strictly local nature. Local properties clearly have a large influence on the nature of turns, as the reasonable level of success of our method shows. Nonetheless, a number of turns will have "flexible" sequences that can form both hairpins and diverging turns. Indeed, some examples of sequence homology between hairpins and diverging turns have been observed in the database, and the turn prediction is ambiguous for a significant fraction of turns. Here, the global folding of the rest of the protein has a tremendous influence on local structure. Because the formation of secondary and tertiary structure during protein folding is interdependent,[26] it will never be possible to accurately predict secondary and supersecondary structure without predicting tertiary structure at the same time.

Recently, de la Cruz et al.[27] published another approach for the identification of hairpins. To identify hairpins, strand-loop-strand patterns are identified by secondary structure prediction with PHD.[6] These potential hairpins are matched against a database of complete strand-loop-strand patterns known to form hairpins. By using a scoring scheme and an artificial neural network, the number of strong matches is calculated. The strand-loop-strand pattern is identified as hairpin if this number is above a certain threshold.

The approach of de la Cruz et al. is not limited to turns of a certain length because it uses a scoring scheme against a database of hairpins. For this reason, it is not restricted to 75% of all strand-loop-strand patterns, as our approach is. By using secondary structure prediction to identify potential hairpins, de la Cruz et al. miss about 50% of all hairpins. Using the native secondary structure to identify potential hairpins, they were able to correctly identify 64.2 ± 8.3% of all hairpins and 65.8 ± 6.7% of all diverging turns. The method presented in this article correctly classifies 77.3 ± 6.1% of all hairpins and 73.9 ± 6.0% of all diverging turns up to length eight, without relying on information about the native secondary structure. The database method's prediction performance drops to 30.1 ± 7.9% for hairpins when the predicted secondary structure is used. The accuracy of the prediction method presented in this article is higher, which is at least partly a result of the limitation in turn length: shorter turns can be expected to show more significant sequence signals. The higher level of confidence with the current method makes it more straightforward to apply it tertiary structure prediction.

## CONCLUSION

The results presented in this article show that it is possible to distinguish between hairpin and diverging turns with 75.9 ± 4.4% accuracy. The accuracy of predicting ends of strands reached by current secondary structure prediction methods is not high enough to reliably identify strand-loop-strand motifs de novo. Our method cannot alleviate this problem and focuses on the classification hairpin/diverging turn. Based on this classification, a score was introduced into the de novo protein structure prediction algorithm ROSETTA that selectively disallows local structures. In turn, higher contact order structures are sampled. This modification was shown to improve predictions made by the ROSETTA algorithm in 50% of the 14 all-β and αβ topologies analyzed.

## ACKNOWLEDGMENTS

M.K. thanks the California Institute of Technology for support with a Summer Undergraduate Research Fellowship. J.M. acknowledges the support of a Human Frontier Science Program fellowship.

## REFERENCES

1. Meiler J, Kuhn M. TURNPRED: Prediction of hairpins and diverging turns in proteins. www.jens-meiler.de/turnpred.html, 2003.
2. PDB. PDB current holdings (September 24, 2002). http://www.rcsb.org/pdb/holdings.html#holdings, 2002.
3. SWISS-PROT. SWISS-PROT Protein Knogwledgebase Release 40.28 Statistics. http://www.expasy.org/sprot/relnotes/relstat.html, 2002.
4. Holley LH, Karplus M. Protein secondary structure prediction with a neural network. Proc Natl Acad Sci USA 1989;86:152–6.
5. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 1993;90:7558–7562.
6. Rost B. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 1996;266(Computer Methods for Macromolecular Sequence Analysis):525–539.
7. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
8. Meiler J, Müller M, Zeidler A, Schmäschke F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. J Mol Model 2001;2001:360–369.
9. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190.
10. Burke DF, Deane CM. Improved protein loop prediction from sequence alone. Protein Eng 2001;14:473–478.
11. Sun Z, Rao X, Peng L, Xu D. Prediction of protein supersecondary structures based on the artificial neural network method. Protein Eng 1997;10:763–769.
12. Chou KC. Prediction of beta-turns. J Pept Res 1997;49:120–144.
13. Shepherd AJ, Gorse D, Thornton JM. Prediction of the location and type of beta-turns in proteins using neural networks. Protein Sci 1999;8:1045–1055.
14. Cai YD, Liu XJ, Xu XB, Chou KC. Support vector machines for the classification and prediction of beta-turn types. J Pept Sci 2002;8:297–301.
15. de la Cruz X, Thornton JM. Factors limiting the performance of prediction-based fold recognition methods. Protein Sci 1999;8:750–759.
16. Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol 1997;270:471–480.
17. Jones DT. Predicting novel protein folds by using FRAGFOLD. Proteins 2001;Suppl 5:127–132.
18. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
19. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins 2001;45 Suppl 5:119–126.
20. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci 2002;11:1937–1944.
21. Dunbrack RL. Culling the PDB by resolution and sequence identity. http://www.fccc.edu/research/labs/dunbrack/culledpdb.html, 2002.
22. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
23. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
24. Meiler J. SMART: training of artifical neural networks. www.jens-meiler.de, 2002.
25. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
26. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci 2003;100:12105–12110.
27. de la Cruz X, Hutchinson EG, Shepherd A, Thornton JM. Toward predicting protein topology: an approach to identifying beta hairpins. Proc Natl Acad Sci USA 2002;99:11157–11162.