A NEW NEURAL NETWORK FOR Β-TURN PREDICTION: THE EFFECT OF SITE-SPECIFIC AMINO ACID PREFERENCE

ZHONG-RU XIE and MING-JING HWANG

Institute of Bioinformatics, National Yang-Ming University,
Institute of Biomedical Sciences, Academia Sinica,
Taipei, Taiwan

Abstract

The prediction of β-turn, despite the observation that one out of four residues in protein belongs to this structure element, has attracted considerably less attention comparing to secondary structure predictions. Neural network machine learning is a popular approach to address such a problem of structural bioinformatics. In this paper, we describe a new neural network model for β-turn prediction that accounts for site-specific amino acid preference, a property ignored in previous training models. We showed that the statistics of amino acid preference at specific sites within and around a β-turn is rather significant, and incorporation of this property helps improve the network performance. Furthermore, by contrasting with a previous model, we revealed a deficiency of not incorporating this site-specific property in previous models.

Introduction

β-turn

Prediction of protein secondary structure is an intermediate step in the prediction of its tertiary structure. Most secondary structure prediction methods predict only three states — α-helix, β-sheet and coil [1]. However, in addition to these three repetitive structural states, tight turn is a significant element frequently occurring in protein structures. Based on the number of their constituent amino acid residues, tight turns are categorized as δ-, γ-, β-, α- and π- turns [1]. Of these five tight turns, the occurrence of β-turn is the most frequent, constituting approximately 25% to 30% of the residues in globular proteins [2]; in contrast, the second most frequently occurring tight turn, γ-turn, takes up only 3.4% of the total residues [3]. β-turn formation is also an important stage in protein folding [4], and because β-turns usually occur on solvent-exposed surfaces, they often participate in molecular recognition processes in the interactions between peptide substrates and receptors [5].

Despite that β-turn is a common and critical structure element, and that a great number of secondary structure prediction methods have been developed, β-turn prediction algorithms are surprisingly few. Most of the β-turn prediction methods are early statistical approaches, which achieve limited accuracy [1]. As accurate β-turn prediction would increase the accuracy and reliability of secondary structure prediction, which in turn would contribute to improve the prediction of tertiary structure and the identification of structural motifs such as β-hairpin, there is a need to explore more sophisticated β-turn prediction algorithms.

β-turn Prediction

The widely accepted definition for β-turn is: A β-turn comprises four consecutive residues where the distance between Cα(i) and Cα(i+3) is less than 7 Å, and the tetrapeptide is not in a helical conformation [1]. Based on these criteria, a number of β-turn prediction algorithms have been developed. They can be categorized as: 1) Site-Independent Model, 2) 1–4 and 2–3 Residue-Correlation Model, 3) Sequence-Coupled Model, and 4) Others [2].

Because a β-turn is consisted of four consecutive amino acid residues, the prediction for β-turn can be performed based on the probabilities of the 20 amino acid residues occurring at each of the 4 oligopeptide subsites.

The Site-Independent Model is a simple prediction method that multiplies the probability of each kind of the 20 amino acids occurring at each of the four subsites. Different from the Site-Independent Model, both the 1–4 and 2–3 Residue-Correlation Model and the Sequence-Coupled Model do not consider the occurrences of the 4 residues as completely independent incidents. The 1–4 and 2–3 Residue-Correlation Model is based on the observation that when a tetrapeptide folds into a β-turn, the interaction between 1st and 4th as well as between 2nd and 3rd residues becomes remarkable. Particularly, a hydrogen bond may form between the backbone carbonyl oxygen of the 1st residue and the backbone amino hydrogen of the 4th residue. The Sequence-Coupled Model also incorporates conditional probabilities. However, it is a residue-coupled model that calculates the conditional probabilities of 1–2, 2–3 and 3–4 residues.

As β-turn prediction has only two outcomes — β-turn and non-β-turn, the former should take up ~25% of the occurrences according to what is observed in protein structures — it is not sufficient to evaluate the performance of a prediction algorithm based only on prediction accuracy, which could be misleading when, for example, a method is biased to give more non-β-turn prediction outcomes.

Therefore, the four parameters commonly used to measure the performance of β-turn prediction algorithms are: 1) Qtotal (Qt): total prediction accuracy, 2) Qpredicted (Qp): percentage of correct positive prediction, 3) Qobserved (Qo): sensitivity, and 4) MCC: Matthews Correlation Coefficient, which accounts for both over- and under-predictions. They are defined in the equations given below, where “p” denotes the number of correctly predicted β-turn residues, “n” the number of correctly predicted non-β-turn residues, “o” the number of incorrectly predicted β-turn residues (false positives), “u” the number of incorrectly predicted non-β-turn residues (false negatives), and “t” the total number of residues predicted. “Qpredicted” and “Qobserved” are the proportion of false positive prediction results and that of false negative results, respectively. The MCC value is an overall evaluation parameter, which is dimensionless.

MCC has a theoretical value between 0 (for random prediction) and 1 (for perfect prediction).

$$Q_{total} = \left(\frac{p+n}{t}\right) \times 100$$

$$\mathrm{MCC} = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}}$$

$$Q_{predicted} = \left(\frac{p}{p+o}\right) \times 100$$

$$Q_{observed} = \left(\frac{p}{p+u}\right) \times 100 \qquad (1)$$

Machine Learning Approaches

Most of the recent algorithms that generally outperform earlier statistical approaches in the prediction of protein structure states have been developed via machine learning, neural networks and support vector machines (SVM) being most notable. Neural network algorithms usually use a segment of peptide sequence as the basis for prediction, where it automatically looks for subtle correlations between the input amino acids and their structural preference via a back-propagation training process. In these approaches, each of the segment residues is transformed into 20 (or 21) nodes of numeric data, which are then used as 20 (or 21) numerical values for the input nodes (or neurons) of the neural network. During the training process, the correlations between each set of the input nodes and output data are automatically adjusted to be in line with the relationship between the structure and the preference of amino acids.

In 2003, Kaur and Raghava proposed a neural network method for the prediction of β-turns utilizing multiple sequence alignments [6]. They constructed two serial feed-forward back-propagation networks, both of which have an input window of 9 residues wide (21 nodes in each residue) and a single hidden layer of 10 units (nodes). The first layer, a sequence-to-structure network, is trained with the multiple sequence alignment in the form of PSI-BLAST [7]-generated position-specific scoring matrices. The preliminary predictions from the first network along with PSIPRED [8]-predicted secondary structure states are then used as input to the second, structure-to-structure network to refine the predictions. They achieved a MCC value of 0.37 using multiple sequence alignment on the first layer and 0.43 overall using the first-layer results plus secondary structure prediction on the second layer. Their results are among the best reported in the literature for β-turn predictions.

However, in Kaur and Raghava's network, the group of 20 nodes, representing the 20 kinds of amino acids, for the central residue of the peptide segment is adjusted to merely fit the general correlations between the structure and the amino acid preference; site-specific amino acid preference is not taken into account. Here we show that a statistical analysis on the occurrence of the 20 amino acids at each of the four sites of the β-turn, and of its adjacent sites also, revealed marked site-specific preference, and incorporation of this preference improved network performance.

Materials and Methods

The Data Set

The data set in this study is consisted of 426 non-redundant protein structures as originally established by Guruprasad and Rajkumar (2000) [3]. Selected from Protein Data Bank [9], the data set was obtained using the program PDB_SELECT [10] such that no two chains of the selected representative proteins have > 25% sequence identity. All the structures selected are determined by X-ray crystallography at 2.0 Å resolution or better. Each chain contains at least one β-turn, and the β-turn assignment is based on the annotation of PDBsum [11].

Previous Neural Network Training Methods vs. Site-specific Amino Acid Preference Based Training Method

A back-propagation training procedure is used to optimize the weights of the neural network. During training, the network response at the output layer is compared to a supplied set of known answers (training targets). The errors are computed and back-propagated through the network in an attempt to improve the network response. The nodal weight factors are then adjusted by the amounts determined by the training algorithm. The iterative procedure of processing the inputs through the network, computing the errors and back-propagating the errors to adjust the weights constitutes the learning process.

Previous neural network methods for structure-state prediction of proteins (e.g. secondary structure prediction and turn prediction) stipulate that the structure of a residue is dependent upon its adjacent amino acid sequences. According to most of these methods, patterns are presented as windows of a certain number (n) of residues, in which a prediction is made for the central residue (ith residue) [6, 8] or a residue in a specific position of the window [12], as shown in Figure 1A. In this way, the group of 20 nodes for the central residue is adjusted to merely fit the general correlations between the structure state of this residue and the amino acid preference deduced for each site on this structure fragment. As the central residue is the point of focus, these methods generally do not care if the adjacent groups of nodes do not fit a certain structure state. In other words, a residue may be predicted as a β-turn residue even if its neighboring residues are not. In addition, site-specific amino acid preference is not considered.

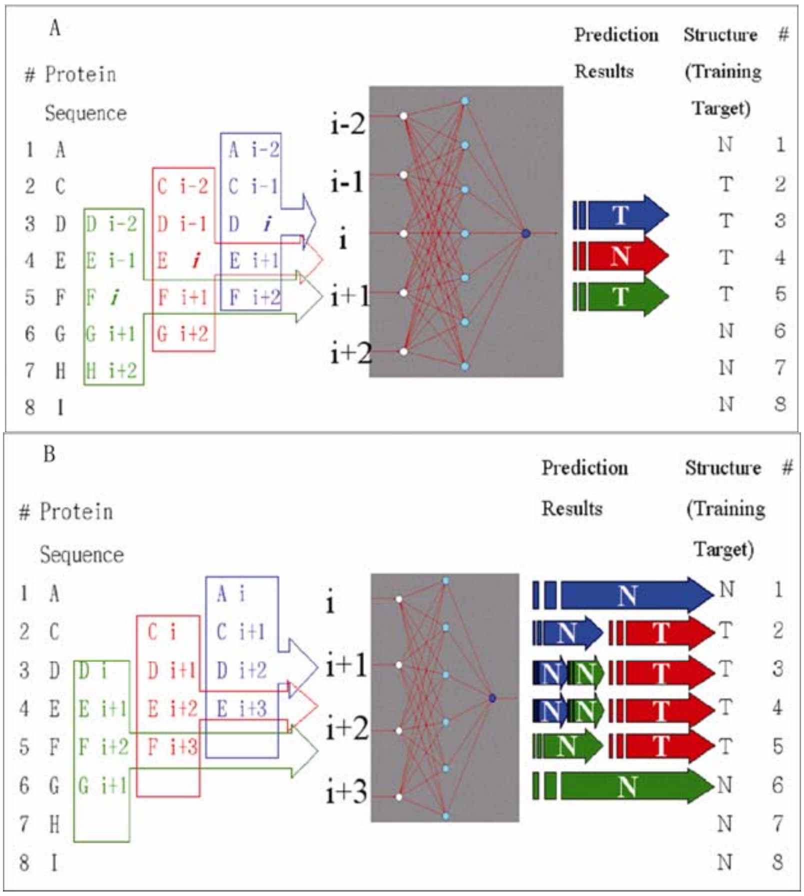

Figure 1. Illustrations of the differences between two neural network training methods: (A) A demonstration of the previous methods: The sequence ACDEFGHI on the left indicates a fragment of consecutive 8 amino acid residues in an input protein sequence. The sequence NTTTTNNN on the right indicates the known structure state of the corresponding residues: N for non-β-turn residue and T for β-turn residue. Note that T always occurs 4 or more times consecutively, as a β-turn contains exactly 4 consecutive residues. In this illustration, the neural network contains 5 groups (i-2, i-1, i, i+1, i+2; actual implementation has more) of input nodes (white dots in the gray rectangle), with each group comprising 20+1 units (such as probabilities of 20 amino acids derived from multiple sequence alignment plus an additional piece of information) to code for the residues, and 1 output node (the dark blue dot in the gray rectangle) for β-turn prediction of site i. The blue, red and green boxed and arrowed rectangles indicate 3 input data, and the "T" or "N" labeled blue, red and green arrows are the corresponding prediction results of the 3 input data. (B) A demonstration of the network model used in the present study: The network contains 4 groups of input nodes (9 residues actually used), i, i+1, i+2, i+3, and 1 output node, the result of which is either a "T" or a "N" assigned to all of the 4 predicted sites. The outputs are indicated by arrows using the same color of their corresponding inputs. Note that if a residue receives a "T" prediction from any of the input, it will be assigned as T (β-turn residue).

In this study, we proposed a new model to produce a training process in which the weights of each group of the nodes are adjusted to fit the preference patterns on each site of the β-turn and of the neighboring residues as well. As shown in Figure 1B, if the (i)th amino acid residue of the input window occurs, as in the case of the target (i.e. true answer), exactly on the 1st site of the β-turn, while the (i+1)th residue occurs on the 2nd site, and so on, the neural network will perform a positive training. When the input window shifts, e.g. the (i)th residue occurs on the 2nd site of the β-turn, and the (i+1)th residue on the 3rd site, and so on, the neural network will perform a negative training. As a result, each group of the nodes will be trained to fit the preference patterns on specific sites within and around the β-turn.

Neural Network Architecture

Besides the implementation to account for site-specific preference, our network architecture follows that of Kaur and Raghava [6]. Briefly, two serial feed-forward back-propagation networks with a single hidden layer were used. The number of hidden nodes was optimized and the two networks used were a sequence-to-structure network in the first layer and a structure-to-structure network in the second layer. The first network had the input window containing information of 9 residues and 24 nodes in the single hidden layer (these numbers of residues and nodes produced best performance among several combinations tested). The input to the first network was a multiple alignment profile. The target output was a single continuous number, which was converted to a binary number — one for β-turn and zero for non-β-turn. The window was shifted residue by residue through the protein chain, yielding N patterns for a chain with N residues. The prediction results obtained from the first layer network along with the secondary structure prediction results from PSIPRED were used as input to the second layer. Specifically, besides the first layer output, each of the 9 residues of the 2nd network input window was given reliability indices of the three secondary structure states (helix, strand and coil).

Results

Statistics of Amino Acid Preference at Specific Sites of β-turn

In this study, the occurrence probability of the 20 kinds of amino acids contained in the non-redundant dataset of 426 proteins on sites within and in the vicinity of a β-turn (sites i to i+3 corresponding to the 1st to 4th residue of the β-turn, and sites i-3 to i-1 and i+4 to i+6 corresponding to the three residues preceding and following the β-turn) and their occurrence probability in the whole dataset were calculated. The one-sample test for binomial proportion [13] was performed on the occurrence probability of the 20 kinds of amino acids on these sites. Table 1 shows the z-value results. In this table, a z value > 2 or < -2 indicates the occurrence frequency of a certain amino acid at a certain site is significantly higher or lower than its occurrence frequency in the dataset. The larger the absolute z-value, the more significant the difference is. As may be seen from Table 1, different sites, particularly the four sites of β-turn, have very different preference patterns for different kinds of amino acids. For example, both the 1st (i) and 2nd (i+1) site have a strong preference for proline, whereas the 3rd (i+2) site does not and in fact selects against it. In contrast, glycine appears to be significantly preferred at the3rd (i+2) and 4th (i+3) site, but not at the 2nd (i+1) site. There are many other notable preference patterns. Thus, the amino acid preference patterns on different specific sites indeed differ significantly. This provides a basis for the new neural network training strategy, which allows neural network to more precisely adjust the weights of each group of the input nodes to fit the preference patterns on the specific sites of β-turn in the training process.

Table 1. z values of amino acid preference on the sites within (site i to i+3) and around a β-turn produced by one-sample test for binomial proportion. Those discussed in the text are highlighted.

Residue\Site	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6
A	-1.72	-4.05	-6.24	-5.31	-2.45	-12.78	-3.34	-7.63	-6.09	-2.10
C	1.52	1.95	4.32	5.13	-4.20	-2.86	2.15	-1.05	2.49	3.21
D	-4.41	-3.09	4.50	14.30	5.20	21.81	-0.81	-0.67	-2.55	-4.03
E	-3.81	-2.94	-5.68	-5.40	5.43	-3.35	-4.70	-1.91	-1.75	-2.28
F	3.26	3.30	2.50	-0.36	-5.68	-3.68	-1.69	-1.33	1.01	4.67
G	-0.93	0.70	-0.91	1.91	-0.56	35.21	18.80	-0.61	-2.12	-1.77
H	0.29	3.25	2.24	2.11	-0.11	2.98	0.52	-0.73	0.69	0.50
I	3.16	2.08	0.99	-7.39	-8.59	-12.67	-6.40	-4.13	2.38	2.51
K	-2.28	-1.51	-2.19	-3.98	6.81	-3.43	2.64	6.58	-0.20	-3.42
L	2.29	-2.38	-1.06	-5.57	-10.80	-13.54	-6.34	-6.36	0.71	-0.89
M	-0.88	-0.09	-0.14	-2.69	-6.79	-6.45	-2.19	-3.12	-0.96	-2.52
N	-0.26	-1.80	1.49	9.12	2.36	24.72	1.13	2.28	-1.44	-2.11
P	-2.03	2.13	-0.57	12.78	33.19	-7.18	2.96	15.41	6.96	1.17
Q	-1.28	-2.16	-3.64	-4.87	-2.84	-3.33	-0.49	-1.43	-3.08	-2.84
R	-1.08	-1.21	1.15	-5.82	-0.57	-3.50	-0.45	1.89	-1.59	-2.72
S	-0.23	-1.02	-1.00	6.20	4.66	2.98	0.50	3.29	-1.32	0.10
T	0.73	0.64	1.41	2.02	-2.40	-1.20	3.63	6.37	0.78	3.80
V	4.18	4.73	1.92	-7.14	-8.63	-13.23	-4.91	-3.51	4.56	5.31
W	1.32	1.57	3.26	-1.12	-1.80	-1.89	-0.64	-0.86	2.97	1.66
Y	3.80	4.48	4.09	0.01	-4.25	-2.56	-1.31	-1.13	2.55	4.51
No. of Res.	7042	7072	7101	7129	7129	7129	7129	7079	7040	7015

Prediction Using Multiple Sequence Alignment in the First Layer

Our first-layer network was trained using input of multiple sequence alignment profiles generated from PSI-BLAST [12], as was done in the study of Kaur and Raghava [6]. The main difference is the new neural network model we used to fit site-specific amino acid preference, as described above. We performed a seven-fold cross validation, and the results, in comparison with those of BetaTPred2 (the current version of Kaur and Raghava's program for predicting β-turn [6]), were presented in Table 2. As may be seen, our results were significantly better. Specifically, our network achieved an MCC value of 0.402, which is significantly higher ($p < 10e-8$) than that (0.37) of the first layer network of BetaTPred2. The values of Qtotal and Qpredicted were also improved, though at the cost of slightly degraded Qobserved. These data indicate that the proportion of false positive prediction results has been significantly decreased with our model. In other words, the probability of correct prediction is significantly increased.

Table 2. Comparisons of results from the first layer between this study and that of Kaur and Raghava (BetaPred2) [6]. SD: standard deviation.

	BetaTPred2 [6]		This study	
	Average	SD	Average	SD
MCC	0.37	0.01	0.402	0.01
Qt	73.5	1.5	74.9	1.9
Qp	47.2	1.9	53.2	2.4
Qo	64.3	2.2	62.6	6.3

Prediction Using First Layer Output Plus Secondary Structure Information in the Second Layer

Again, following the procedures of Kaur and Raghava [6], our second layer was trained with the first layer output and the secondary structure prediction results from PSIPRED [10]. Cross-validation results shown in Table 3 yielded an MCC value of 0.443, which is just a bit higher than that (0.43) of BetaTPred2. Similar to the results of the first layer (Table 2), we improved on Qtotal and Qpredicted, but not Qobserved.

Table 3. Comparisons of results from the second layer between this study and that of Kaur and Raghava (BetaPred2) [6]. SD: standard deviation.

	BetaTPred2 [6]		This study	
	Average	SD	Average	SD
MCC	0.43	0.01	0.443	0.01
Qt	75.5	1.7	76.4	2.3
Qp	49.8	2.0	55.6	3.5
Qo	72.3	2.6	66.6	7.5

Discussion

In this study, we have developed a new neural network model to account for site-specific amino acid preference for β-turn predictions. We showed that site-specific preference is statistically significant and when incorporated in the neural network training can improve the network performance. In fact, ignoring site-specific preference may be a source of errors for previous models such as that of Kaur and Raghava [6]. For example, as shown in Table 1, Cysteine frequently occurs but Lysine rarely occurs on the 1st site of β-turn (z values 5.13 and -3.98), whereas on the 2nd site, the occurrence preference for the two amino acids is reversed (z values -4.20 and 6.81). In the training process of previous models, the (i)th group of neurons must fit all of the amino acids preferred on four sites simultaneously. If the residue of the 1st site of β-turn is the input to the (i)th group of neurons, the neuron weight of Cysteine will be increased and that of Lysine will be decreased. However, if the residue of the 2nd site of β-turn is the input to the (i)th group of neurons, the neuron weights of Cysteine and Lysine will be adjusted in the opposite way. This extreme example indicates possible interference of training data subsets using previous models. As the weights of a particular group of neurons are not adjusted to fit the amino acid preference on specific sites, but are merely updated as a general pattern to fit most of the preference, the prediction power would be compromised. This is corroborated by the observation that our main improvement (for the first layer) was achieved by increasing the value of Qp (Table 2), or reducing the false positive rate. Additionally, because only one residue is predicted in each prediction process using the previous models, the prediction results of consecutive residues in a sequence taken together are likely to conflict with each other; with the site-specific model (Figure 1B), contradictory adjacent predictions are eliminated.

The less-than-expected improvement by the second layer (MCC from 0.402 to 0.443), as opposed to that (MCC from 0.37 to 0.43) of Kaur and Raghava’s model (Table 3 vs. Table 2), revealed a possible role of the second layer in previous network models. Many secondary structure prediction methods use two serial neural networks for prediction, where even if the second layer network does not involve other data except for the initial prediction results from the first layer, significantly greater improvement from the first layer is still achieved [8, 14]. Our study suggests that the function of the second layer network in these models is likely to reconcile or filter the initially disaccord results, whereas in our site-specific model, this is already achieved to a large extent in the first layer.

Tight turns are usually classified as coil in secondary structure assignment. However, its structural and functional significance is no less than that of α-helix or β-sheet, and could play a prominent role in the prediction of tertiary structures. Indeed, despite that the accuracy of secondary structure prediction methods has exceeded 75% [14], that for terminals of α-helix and β-strand has not yet reached a satisfactory level. Accurate tight turn predictions could remedy this problem as they could complement nicely with existing secondary structure predictions. This study demonstrated the merit of incorporating site-specific amino acid preference for β-turn prediction and provided insight into a deficiency of previous models. The same idea should be applicable to other structure-state predictions with beneficial results.

References

1. K. C. Chou. REVIEW: Prediction of tight turns and their types in proteins. *Analytical Biochemistry*, 286:1–16, 2000.
2. H. Kaur and G. P. S. Raghava. An evaluation of β-turn prediction methods. *Bioinformatics*, 18:1508–1514, 2002.
3. K. Guruprasad and S. Rajkumar. β- and γ-turns in preteins revisited: A new set of amino acid turn-type dependent positional preferences and potentials. *J. Biosci.,* 25:143–156, 2000.
4. K. Takano, Y. Yamagata and K. Yutani. Role of amino acid residues at turns in the conformational stability and folding of human lysozyme. *Biochemistry*, 39:8655–8665, 2000.
5. G. D. Rose, L. M. Gierasch and J. A. Smith. Turns in peptides and proteins. *Adv. Protein Chem.,* 37:100–109, 1985.
6. H. Kaur and G. P. S. Raghava. Prediction of β-turns in proteins from multiple alignment using neural network. *Protein Science.,* 12:627–634, 2003.
7. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25:3389–3402, 1997.
8. D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol., 292:195–202, 1999.
9. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research.,* 28:235–242, 2000.
10. U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci.,* 3:522–524, 1994.
11. R. A. Laskowski. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research,* 29:221–222, 2001.
12. M. Kuhn, J. Meiler and D. Baker. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. *PROTEINS: Structure, Function, and Bioinformatics*, 54:282–288, 2004.
13. B. Rosner. *Fundamentals of Biostatistics.* (5th ed.). Boston: Harvard University Press.
14. B. Rost. Review: Protein Secondary Structure Prediction Continues to Rise. *Journal of Structural Biology.* 134:204–218, 2001