Article ID: WMC00632 2046-1690

Evaluation of neural network performance in the protein secondary structure prediction problem

Corresponding Author:
Dr. Eric Sakk,
Assistant Professor, Department of Computer Science, Morgan State University, 1700 E. Cold Spring Lane, 21251 - United States of America

Submitting Author:
Dr. Eric Sakk,
Assistant Professor, Department of Computer Science, Morgan State University, 1700 E. Cold Spring Lane, 21251 - United States of America

Article ID: WMC00632

Article Type: Research articles

Submitted on:26-Sep-2010, 07:22:40 AM GMT **Published on:** 26-Sep-2010, 07:24:41 AM GMT

Article URL: http://www.webmedcentral.com/article_view/632

Subject Categories:BIOINFORMATICS

Keywords: Protein secondary structure prediction, neural networks, target vector selection, classifier performance measures

How to cite the article: Sakk E , Alexander A . Evaluation of neural network performance in the protein secondary structure prediction problem . WebmedCentral BIOINFORMATICS 2010;1(9):WMC00632

Source(s) of Funding:

This publication was made possible by Grant Number G12RR017581 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH).

Evaluation of neural network performance in the protein secondary structure prediction problem

Author(s): Sakk E , Alexander A

Abstract

In the context of the protein secondary structure prediction problem, we examine the performance of various neural network architectures commonly applied in the literature. Specifically,by applying the neural network paradigm, mappings between training vectors and their desired targets are constructed. The class membership of test data and associated measures of significance are then numerically demonstrated to vary depending on the set of applied target vectors. By applying standard symbol encoding techniques, we analyze and discuss the ability of the neural network to accurately model fundamental attributes of protein secondary structure.

Introduction

The protein secondary structure prediction problem can be phrased as a supervised pattern recognition problem [1], [2] for which training data is readily available from reliable databases such as the Protein Data Bank (PDB) or CB513 [3]. Based upon training examples, amino acid subsequences derived from primary sequences are categorized according to a discrete set of classes such as alpha helix (H), beta sheet (B) or coil (C). Then, by applying some pattern recognition scheme of choice, subsequences of unknown classification are tested to predict the class to which they belong. Phrased in this way, backpropagation neural networks [4], [5], [6], [7] and variations on the neural network theme [8], [9], [10], [11], [12], [13], [14] have been applied to the secondary structure prediction problem with varied success. Furthermore, many tools currently applying hybrid methodologies such as PredictProtein [15], [16], JPRED [10], [11], [17], SCRATCH [18], [19] and PSIPRED [20], [21] rely on the neural network paradigm as part of their prediction scheme.

The purpose of this work is to revisit the application of neural networks to the protein secondary structure prediction problem. In particular, our goal is to demonstrate the variability of results generated by the neural network when different schemes are used to encode (mathematically represent) the secondary structure classes. Consider the well-studied case where three classes (H, B and C) are used to classify a given amino acid subsequence. Two equally valid and commonly applied encoding schemes for the three class problem are orthogonal encoding {H=(1,0,0), B=(0,1,0), C=(0,0,1)}and {H=(-.5,.866,0), B=(-.5,-.866,0), C=(1,0,0)} where class encodings are chosen on the vertices of an equilateral triangle. Given the same set of input training sequences, it is our observation that, for certain neural network architectures, classification results and associated performance measures can vary when two equally valid encoding schemes are employed. Such a result goes against the intuition that the secondary structure property should be independent of the encoding scheme chosen.

In this work, we extend results regarding the variability of classifier performance measures [22], [23], [24] to the protein secondary structure prediction problem. In particular, we examine the linear and backpropagation neural networks [1], [2] as template examples in order to understand how structure classification and associated confidence measures can vary. It is critical to note that the goal of this work is not to demonstrate improvements over existing techniques. The hybrid techniques outlined above have been demonstrated to outperform neural networks when used alone. Instead, given that certain models presented under this paradigm have been found to yield variable results, we focus this study on the ability of the neural network to accurately model protein secondary structure. We believe these results are relevant because they bring into discussion a body of literature that has purported to offer a viable path to the solution of the secondary structure prediction problem.

Classifier Description and Associated Performance Measures

In the supervised classification problem [1], [2], it is assumed that a training set consists of N training pairs {(X1,Y1), (X2,Y2), ..., (XN,YN)} where Xi is an input column vector and Yi is an output ('desired' response) column vector for i=1, ..., N. A given classifier is considered to be ‘trained’ when, upon presenting any input vector XK from the training set to the classifier, it

responds with the associated output YK.

When the input training data can be partitioned into M distinct classes, a set E = {e1, e2, …, eM} of target column vectors are chosen to encode (i.e. mathematically represent) each class. In this case, all output vectors Yi in the training set must come from the set of target vectors E. For example, as pointed out in the introduction, the three secondary structure classes {H,B,C} can be encoded using a set of three target vectors. Since there appear to be any number of equally valid choices to represent the three classes, it is interesting to evaluate the neural network classification approach by examining various performance measures to be outlined in this section. For this study, we apply and analyze two types of neural networks as the supervised classification method: the linear neural network [2], [22], [23] and the backpropagation neural network [1], [2].

After a given classifier is trained, when presented with an input vector, x, of unknown classification, it will respond with an output answer f(x). The class membership C(x) is then determined by choosing the target vector from the set E that is closest to the classifier answer f(x). When characterizing the performance of a pattern classifier, one often presents a set S of test vectors and analyzes the associated output. In addition to determining the class membership C(x), it is also possible to rank the distance between a specific target vector eJ (representing class J) and the classifier response to the set of vectors in S. In this case a distance criterion ρJ(x) similar to that of the class membership can be applied [22]. [23]. The ranking ρJ(x) can be thought of as a distance-based measure of confidence indicating the statistical significance of a class membership prediction. In other words, the higher the ranking of an input vector, x, with respect to class J, the smaller the distance of the classifier answer, f(x), to the target vector eJ. Given a classifier and two distinct choices for the set of target vectors E1 and E2, our goal is to investigate if the above performance measures remain invariant. This study will therefore evaluate neural network performance by using existing secondary structure training data to determine if, given a set of test sequences S, the equalities C1(x) = C2(x) and ρ 1J(x) = ρ2J(x) hold true for any input vector x in the test set S.

Results

In this section, we numerically demonstrate that, when different target vector encodings are applied, performance measures outlined above, in certain cases, are observed to vary widely. To do this, we apply the neural network paradigm to one hundred protein sequences extracted from the CB513 database [3] available through the JPRED secondary structure prediction engine [17]. A moving window of length 17 is applied to each protein sequence where, in order to avoid protein terminal effects, the first and last 50 amino acids are omitted from the analysis. The secondary structure classification of the central residue is then assigned to each window of 17 amino acids. For the one hundred sequences analyzed, a total of 12000 windows of length 17 were extracted. To encode the input amino acid sequences of length 17, in a manner similar to [6], [8], we employ sparse orthogonal encoding [26, Ch. 6] which maps symbols from a given sequence alphabet onto a set of orthogonal vectors. In particular, for an alphabet containing K symbols, a unique K dimensional unit vector is assigned to each symbol; furthermore, the kth unit vector contains a one at the kth position and is zero at all other positions. Hence, if all training sequences and unknown test sequences are of uniform length L, an encoded input vector will be of dimension n where n = LK. In our case, K = 20 and L = 17; hence, the dimension of any given input vector is n = 340. The secondary structure classification associated with a given input vector is then encoded using either E1 = {e1H=(1,0,0), e1B =(0,1,0), e1C =(0,0,1)} or E2 ={ e2H=(-.5,.866,0), e2B =(-.5,-.866,0), e2C =(1,0,0)} as the set of target vectors. Both the linear and the backpropagation network have been tested first by training using E1 and then comparing classifier performance with their counterparts trained using E2. In all numerical experiments, MATLAB has been used for simulating and testing these networks. Multiple cross validation trials are required in order to prevent potential dependency of the evaluated accuracy on the particular training or test sets chosen [6], [8]. In this work, we apply a hold-n-out strategy similar to that of [7] using 85% of the 12000 encoded sequences as training data (i.e. N=10200) and 15% as test data to validate the classification results.

Recognition rates for both the linear and backpropagation rates using either set of target vector encodings were approximately 65% which is typical of this genre of classifiers that have applied similar encoding methodologies [4], [5], [6], [7]. Although these aggregate values remain consistent, we now present data demonstrating that, while class membership and ranking remain invariant for the linear network, these measures of performance vary considerably for the backpropagation network which was trained with m = 17 seventeen hidden nodes and a mean squared training error less than .2.

Ranking results for a representative test for linear and backpropagation networks are presented in Tables 1 and 3. Class membership data are presented in Tables 2 and 4. Specifically, Tables 1 and 3 list the indices of test vectors generating the top 20 network responses that are closest to the helix target vector. Hence, ρ1H(x) and ρ2H(x) represent the distance of the ith test vector network response from the helix class target vectors e1H and e2H. Out of 1800 vectors tested, the distances referred to in Tables 1 and 3 were ranked from 1-20. In Tables 2 and 4, for each class, the total number of vectors classified using E1 are analyzed to examine the total number that retained their classification using E2.

Observe that, for the linear network, indices for the top 20 ranked vectors remain invariant indicating ranking invariance; in addition, no change in class membership is observed. On the other hand, Tables 3 and 4 clearly indicate a lack of consistency when considering the ranking and class membership of test vectors. A particularly troubling observation is that very few vectors ranked in the top 20 with respect to e1H were ranked in the top 20 with respect to e2H. Furthermore, Table 4 indicates that the class membership of a substantial number of test vectors changed when an alternative set of target vectors were employed. The data also indicates that the greatest change in class membership took place for alpha helical sequences; thus, implying that there is substantial disagreement over the modeling of this secondary structure element by the backpropagation network due to a simple transformation of the target vectors.

Conclusions

In this work we have addressed the ability of the neural network to accurately model protein secondary structure. Specifically, we have presented numerical data for analyzing how secondary structure classification and confidence measures vary depending on the type of neural network architecture and target vector encoding scheme employed. By applying methods similar to those encountered in the literature [4], [5], [6], [7], we have demonstrated that classifier performance measures can vary considerably. Intuitively, given a sensible encoding scheme, one would desire that the properties determining protein secondary structure would be independent of the target vectors chosen. However, while the linear network does demonstrate this invariance, the backpropagation network does not. It is of interest that performance measures such as the classification and ranking can vary due to a simple transformation of the target vectors using the backpropagation network. The ranking can be thought of as a distance based measure of confidence indicating the statistical significance of a class membership prediction. If invariance cannot be guaranteed, associated performance measures run the risk of being unreliable.

The examples presented above are important because they consider network architectures and sets of target vectors commonly applied in the literature. The purpose of the neural network is to create a reliable mapping between input and output training data and, ideally, extract parameters from the model that contribute to the understanding of protein secondary structure. If the properties of such a mapping are inconsistent, then further research is still required in order to understand the ability of the neural network to accurately model fundamental attributes of protein secondary structure.

Acknowledgements

This publication was made possible by Grant Number G12RR017581 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH).

References

[1] C. Bishop, Pattern Recognition and Machine Learning. Springer, 2007.

[2] M. T. Hagan, H. B. Demuth, and M. H. Beale, Neural Network Design. PWS Publishing, 1996.

[3] J. Cuff and G. Barton, “Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction,” Proteins: Structure, Function, and Genetics, vol. 40, pp. 502–511, 2000.

[4] N. Qian and T. J. Sejnowski, “Predicting the secondary structure of globular proteins using neural network models,” Journal of Molecular Biology, vol. 202, pp. 865–884, 1988.

[5] L. H. Holley and M. Karplus, “Protein secondary structure prediction with a neural network,” Proc. Natl Acad. Sci. USA, vol. 86, pp. 152–156, 1989.

[6] J. Chandonia and M. Karplus, “Neural networks for secondary structure and structural class predictions,” Protein Science, vol. 4, pp. 275–285, 1995.

[7] J. Chandonia and M. Karplus, “The importance of larger data sets for protein secondary structure prediction with neural networks,” Protein Science, vol.

5, pp. 768–774, 1996.
[8] B. Rost and C. Sander, “Improved prediction of protein secondary structure by use of sequence profiles and neural networks,” Proc. Natl Acad. Sci. USA, vol. 90, pp. 7558–7562, 1993.
[9] B. Rost and C. Sander, “Prediction of protein secondary structure at better than 70% accuracy,” J Mol Biol, vol. 232, pp. 584–599, 1993.
[10] J. Cuff, M. Clamp, A. Siddiqui, M. Finlay, and G. Barton, “JPred: a consensus secondary structure prediction server,” Bioinformatics, vol. 14, pp. 892–893, 1998.
[11] J. Cuff and G. Barton, “Evaluation and improvement of multiple sequence methods for protein secondary structure prediction,” Proteins: Structure, Function, and Genetics, vol. 34, pp. 508–519, 1999.
[12] S. Hua and Z. Sun, “A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach,” Journal of Molecular Biology, vol. 308, pp. 397–407, 2001.
[13] L. Wang, J. Liu, and H. Zhou, “A comparison of two machine learning methods for protein secondary structure prediction,” Proceedings of 2004 International Conference on Machine Learning and Cybernetics, vol. 5, pp. 2730–2735, 2004.
[14] G.-Z. Zhang, D. Huang, Y. Zhu, and Y. Li, “Improving protein secondary structure prediction by using the residue conformational classes,” Pattern Recognition Letters, vol. 26, pp. 2346–2352, 2005.
[15] B. Rost, “PHD: Predicting one-dimensional protein structure by profile based neural networks,” Methods in Enzymology, vol. 288, pp. 525–539, 1996.
[16] B. Rost, G. Yachdav, and J. Liu, “The PredictProtein server,” Nucleic Acids Research, vol. 32, pp. W321–W326, 2004.
[17] C. Cole, J. D. Barber, and G. J. Barton, “The Jpred 3 secondary structure prediction server,” Nucleic Acids Research, vol. 36, pp. W197–W201, 2008.
[18] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, “Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles,” Proteins, vol. 47, pp. 228–235, 2002.
[19] J. Cheng, A. Randall, M. Sweredoski, and P. Baldi, “SCRATCH: a protein structure and structural feature prediction server,” Nucleic Acids Research, vol. 33, pp. W72–W76, 2005.
[20] D. Jones, “Protein secondary structure prediction based on position-specific scoring matrices,” J. Mol. Biol., vol. 292, pp. 195–202, 1999.
[21] K. Bryson, L. McGuffin, R. Marsden, J. Ward, J. Sodhi, and D. Jones, “Protein structure prediction servers at university college london,” Nucleic Acids Research, vol. 33, pp. W36–W38, 2005.
[22] A. Alexander, “Non-invariance of pattern recognition measures using neural networks as applied to the protein secondary structure prediction problem,” Morgan State University (M.S. Thesis), 2008.
[23] E. Sakk, D. Schneider, C. Myers, and S. Cartinhour, “On the selection of target vectors for a class of supervised pattern recognizers,” IEEE Transactions on Neural Networks, vol. 20, pp. 745–757, 2009.
[24] E. Sakk, “A pseudoinverse invariance property derived from the linear least squares estimation problem,” Technical Report, Department of Computer Science, Morgan State University, August 2008.
[25] G. Golub and C. Van Loan, Matrix Computations. Johns Hopkins University Press, 1989.
[26] P. Baldi and S. Brunak, Bioinformatics: The Machine Learning Approach. MIT Press, 1998.

Illustrations

Illustration 1

Table 1

Table 1: Ranking results for the linear network.

i_{E1}	$\rho^1_H(x)$	i_{E2}	$\rho^2_H(x)$
1205	0.0780	1205	0.1170
42	0.0867	42	0.1300
1031	0.0976	1031	0.1464
1773	0.1113	1773	0.1670
598	0.1238	598	0.1857
1761	0.1267	1761	0.1900
862	0.1354	862	0.2031
1073	0.1409	1073	0.2114
277	0.1459	277	0.2188
115	0.1540	115	0.2309
1505	0.1821	1505	0.2731
392	0.1839	392	0.2759
1421	0.1904	1421	0.2856
147	0.2001	147	0.3001
990	0.2044	990	0.3066
1457	0.2127	1457	0.3191
1288	0.2150	1288	0.3225
352	0.2160	352	0.3239
1232	0.2198	1232	0.3297
280	0.2311	280	0.3466

Illustration 2

Table 2

Table 2: Class membership results for the linear network.

Class	E^1	E^2	*% change*
H	202	202	0
E	621	621	0
C	977	977	0

Illustration 3

Table 3

Table 3: Ranking results for the back propagation network.

$^{i}E1$	$\rho^{1}H(x)$	$^{i}E2$	$\rho^{2}H(x)$
817	0.0107	926	0.0101
887	0.0231	1604	0.0130
264	0.0405	887	0.0209
1183	0.0711	1145	0.0214
684	0.0727	461	0.0232
623	0.0874	583	0.0329
911	0.0891	1086	0.0339
1382	0.0917	1382	0.0478
1610	0.0939	413	0.0489
551	0.1060	225	0.0608
1042	0.1150	438	0.0609
924	0.1322	911	0.0613
727	0.1339	207	0.0774
438	0.1356	559	0.0885
577	0.1363	481	0.0945
896	0.1500	1548	0.0947
175	0.1513	962	0.0968
1138	0.1549	85	0.1012
583	0.1581	195	0.1111
559	0.1655	9	0.1167

Illustration 4

Table 4

Table 4: Class membership results for the back propagation network.

Class	E^1	E^2	*% change*
H	225	142	36.9
E	581	476	18.1
C	994	878	11.7

Disclaimer