A Data Oriented Approach to Assess the Accuracy of a Protein Secondary Structure Predictor

Saad Subair

College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh KSA
sosubair@pnu.edu.sa

ABSTRACT-Researchers in the field of protein secondary structure prediction use typical three states of secondary structures, namely: alpha helices (H) beta strands (E), and coils (C). The series of amino acids polymers linked together into adjacent chains are known as proteins. Protein secondary structure prediction is a fundamental step in determining the final structure and functions of a protein. In this work we developed a prediction machine for protein secondary structure. By investigating the amino acids benchmark data sets, it was observed that the data is grouped into two distinct states or groups almost 50% each. In this scheme, researchers classify any state which is not classified as helix or strands as coils. Hence, in this work a new way of looking to the data set is adopted. For this type of data, the Receiver Operating Characteristic (ROC) analysis is considered for analysing and interpreting the results of assessing the protein secondary structure classifier. The results revealed that ROC analysis showed similar results to that obtained using other non ROC classification methods. The ROC curves were able to discriminate the coil states from non-coil states by 72% prediction accuracy with very small standard error.

Keywords: *Protein Secondary Structure Prediction, Receiver Operating Characteristics (ROC), Area Under Curve (AUC), Binary Classification, Bioinformatics.*

المستخلص: نهج موجه نحوالبيانات لاختيار أسلوب التنبؤ ببنية البروتين الثانوية: الباحثون في مجال التنبؤ ببنية البروتين الثانوية يستخدمون ثلاث اشكال من الهياكل الثانوية، وهي: اللوالب ألفا (H) بيتا (E)، والملفوفات (C). سلسلة الأحماض الأمينية التي ترتبط معا في سلاسل مجاورة تعرف باسم البروتينات. التنبؤ ببنية البروتين الثانوية هو خطوة أساسية في تحديد هيكل ووظائف البروتين النهائية. استلهم هذا العمل من تجربة لتطوير آلة التنبؤ ببنية البروتين الثانوية. من خلال التحقيق في مجموعات بيانات الأحماض الأمينية ، لوحظ أن البيانات تنقسم إلى مجموعتين اثنين تقريبا 50٪ لكل منهما. وبالتالي، يتم تبني طريقة جديدة للنظر إلى مجموعة البيانات. لهذا النوع بحيث ان البيانات تنقسم الي مجموعتان وليست ثلاثة. واستخدم (ROC) لتحليل وتفسير النتائج. وكشفت النتائج أن التحليل ROC أظهر نتائج مماثلة لتلك التي تم الحصول عليها باستخدام أساليب التصنيف الأخرى غير ROC. وكانت ROC قادرة على التمييز بنسبة 72٪ .

INTRODUCTION

Protein has three main structures: *primary structure* which is essentially the linear amino acid sequence. *Secondary structure*s which are *alpha* helices, *beta* sheets, and coils which are formed when the sequences of primary structures tend to arrange themselves into regular conformations [1,2,3,4]. The *3D structure* and where secondary structures are elements that packed against each other in a stable configuration. The estimation of the global accuracy of a protein is usually conducted by a measure known as Q_3. The Q_3 is a measure of the overall percentage of predicted residues to the observed ones [5] and represented as: The summation of the number of residues identified in the (helix, strand, and coil) states effectively observed divided by the total number of residues. Segment Overlap measure or SOV is another measure that measures the quality of secondary structure prediction in percentage [6].

The Receiver Operating Characteristic (ROC) curve is a method for visualizing, organizing, and selecting classifiers based on their performance. ROC graphs have long been used in signal processing and detection theory to depict the trade-off between hit rates and false alarm rates of classifiers [7,8]. ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems [9]. The ROC techniques is then used extensively in biological sciences and specifically clinical medicine [10,11,12].The ability of a test to discriminate abnormal cases from normal cases is evaluated using the ROC curve analysis [10,11]. ROC curves can also be used to compare the performance of two or

more classifiers. ROC becomes popular in assessing a two-class or binary classifier and comparing many binary classifiers efficiently.

ROC can be explained when you consider the results of a particular test in two populations, one population with abnormal cases, the other population with normal cases. For every possible cut- off point or criterion value you select to discriminate between the two populations. There will be some cases with the abnormal cases correctly classified as positive (true positive or TP), but some cases with the abnormal cases will be classified as negative (false negative or FN). On the other hand, some cases without the abnormal cases will be correctly classified as negative (true negative or TN), but some cases without the abnormal cases will be classified as positive (false positive or FP).

Sensitivity and Specificity are two important terms in the ROC literatures which are defined as *Sensitivity* is the probability that a test result will be positive when the abnormal case is present (true positive rate) while *Specificity* is the probability that a test result will be negative when the abnormal case is not present (true negative rate).

To measure the performance accuracy of a binary classifier, a common method is to calculate the area under the ROC curve, which is known as AUC [13].The AUC is a portion of the area of the unit square and hence its value will always be between 0 and 1. Since the random guess produces the diagonal line between (0; 0) and (1; 1), which has an area of 0.5, no practical classifier have an AUC less than 0.5 (Explained in the next section in Figure 1). Moreover, the AUC has an important statistical property that the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [14].

Many researchers argue that dichotomous (binary) classification is convenient and powerful for decision making, while it may introduces distortions [15,16]. However, the use of Receiver Operating Characteristic (ROC) curves which is mainly threshold-independent has received considerable attention in recent years.

The ROC curves or graphs are useful techniques for assessing the performance of classifiers. The ROC curves are well known in Biology and Medical decision making and they are well used in dichotomous classification. They have been increasingly adopted as a tool for analysing and visualizing many aspects of machine learning algorithms or methods. The ROC curve is a plot of the true positive rate against the false positive rate for different possible cut points of a diagnostic test.

The ROC curve illustrates the trade- off between sensitivity and specificity in the sense that any increase in sensitivity will be accompanied by a decrease in specificity. It also shows that the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test while the closer the curve comes to the diagonal of the ROC space, the less accurate the test. Further, the area under the curve (AUC) is a measure of the algorithm accuracy.

Kloczkowski *et al.* [17] argued that, regularly, proteins contain about 30% helical structure (H), about 20% strands (E), and about 50% coil (C) structure.

This means that even the most trivial prediction algorithm which assigns all residues to the coil (C) state would give approximately 50% correct prediction. This paper attempts to test the results of the prediction or classification task of a new protein secondary structure prediction method [18] while opening a discussion about the reliability of ROC curves analysis in predicting coils only states in a multi-class data set. In eight-to-three secondary structure reduction methods discussed in a previous work [19], one of the reduction methods showed that coils states composed 0.48 of the

whole data set. Several researchers in the protein secondary structure prediction reported similar ratio. Baldi *et al.* [20] reported coil only composed 0.4765 of the data set while others argued that 50% accuracy of an algorithm is not better than a random guess in protein secondary structure prediction.

MATERIALS AND METHODS

For the problem of secondary structure prediction, if we have an amino acid sequence of length n , then secondary structures corresponding to these sequences are the three states helix, strand, and coils which can be considered as $d_i=d_1$, d_2, d_n .. In the case of the dichotomy problem of two alternative classes, when predicting only one structural class, for instance: a coil versus non-coil, then, the d_i in general equals to 0 or 1 which is a binomial model of 0.5 probability for a d_i or non- d_i . In this work d_i corresponds to the coil states since it is equivalent to 0.5 of the data set. Helix and strand states together correspond to non-coil sate which is of course 0.5 of the whole dataset. So we can analyze the three class states as typically two states.

The relation between sensitivity and specificity can be expressed as:

$$Sensitivity = TP/(TP+FN) \quad (1)$$

$$Specificity = FP/(FP+TN) \quad (2)$$

where N is the total sample size which defined as:

$$N= T P + TN + F P + F N \quad (3)$$

The four numbers of the equation (i.e. *T P, TN, FP, and FN*) can be arranged into a 2 x 2 contingency or confusion matrix as shown in Table 1 to facilitate a straightforward analysis of these numbers.

The ROC curve does not provide a rule for the classification of cases. However, there are strategies that may be used to develop decision rules. Two elements are required to identify the appropriate threshold; the first is the relative cost of FP and FN errors while the second is the prevalence of positive cases. Assigning values to these costs is complex, subjective and dependent upon the context within which the classification rule will be used [10].

Table 1: The contingency table or confusion matrix for coil states prediction

		Predicted	
Observed		C	$\overline{C}$
	C_1	TP	FN
	$\overline{C}_2$	FP	TN

C_1 Coil

$\overline{C}_2$ Not Coil

As discussed earlier, the numbers TP, TN, FP and FN depend on how the threshold is selected. In most cases, there is a trade- off between the amount of false positives and the amount of false negatives produced by the algorithm or the classifier. The ROC summarizes such results by displaying threshold values within a certain range of sensitivity or specificity. In a typical ROC curve the hit rate (sensitivity) increases with the false alarm rate (specificity).

Thus sensitivity can be defined as the probability of correctly predicting a positive example and the specificity is the probability that a positive prediction is correct. In biology and medical statistics, the word specificity is sometimes used in a different sense [20] which is beyond our discussion in this paper.

The ROC curves usually show the distribution of the number of normal and NOT normal observations arranged according to the value of a test. This distributions overlap does not distinguish normal from not normal with 100% accuracy. Further, the area of overlap indicates where the test cannot distinguish normal from not normal. In practice, a cut-point (cut score) is chosen; above which the test will be considered as abnormal and below which the test will be considered as normal. The position of the cut point will determine the number of true positive, true negatives, false positives, and false negatives. Different cut points may be chosen if we wish to minimize one of the errors types of the test

results. This curve is discussed in the next section.

The confusion matrix accuracy measures assume that data is real counts. The sensitivity of a test can be described as the proportion of true positives it detects of all the positives. All positives are the sum of (detected) true positives (TP) and (undetected) false negatives (FN). Sensitivity is therefore can be rewritten as:

$$TP/(TP+FN) \quad (4)$$

While the specificity of a test can be described as the proportion of true negatives it detects all the negatives. Thus it is a measure of how accurately it identifies negatives. All negatives are the sum of (detected) true negatives (TN) and (miss-predicted) false positives (FP). Specificity is therefore can be rewritten as:

$$TN/(TN+FP) \quad (5)$$

Finally, sensitivity and specificity represent the measures of accuracy of a certain diagnostic test or classification. In fact, the measurements have to be sensitive in order to detect differences that are important to the research question, and specific enough to show only the feature of interest. Hence, sensitivity describes how well a classification task classifies those observations in the right corresponding class (as in coils state here). Similarly, specificity describes how well a classification task classifies those observations that are not coils. Thus the definitions of sensitivity and specificity can be well depicted from equations above.

Since a typical classifier generates a variable that has values within the range 0 -1, and all of the measures described in this section depend on the numbers in the confusion matrix, these numbers are obtained by application of a threshold criterion to a continuous variable generated by the classifier. A mid value between 0 and1 which is 0.5 is the threshold applied here. Thus, a continuous variable is converted into dichotomy variable in this case. If the threshold criterion is altered, then the values in the confusion matrix will change. Often, the raw scores are available so it is relatively easy to examine the effect of changing the threshold. If we have FN errors more serious than FP errors the threshold can be adjusted to decrease the FN rate at the expense of an increased FP error rate.

The effect of the threshold on error rates can be explained by a cut-point of 0 where every case assigns as positive, while a cut-point of 1 assigns every case as negative. Therefore, as the cut-point is moved from 0 to 1 the false positive frequency falls while the false negative frequency increases. The point where these two curves cross is the point with the minimum overall error rate. Thresholds can be amended to reflect different TP and FP rates according to different objectives (This is clearly illustrated in the next section in Table 3).

RESULTS

Table 2 presents six classification methods for protein secondary structure prediction including our NN-GORV-I classifier which the core of the whole research.

Table 2: Performance of NN-GORV-I and the other five prediction methods

Prediction Method	Q_3
NN-I	64.05
GOR-IV	63.19
GOR-V	71.84
NN-II	73.58
PR OF	75.03
NN-GORV-I	79.22

The primary results in this research revealed that our classifier NN-GORV-I reached an accuracy of 79.22% using the Q_3 assessment method mentioned above and shown in Table 2.

The ROC curves provide an efficient way to display the relationship between sensitivity and specificity and the cut- off point for positive and negative tests [22, 23]. The ROC curves describe the performance of a test used to discriminate between normal and abnormal cases based on a variable measured on a continuous scale.

The area under the ROC function (AUC) is usually taken to be an important index because it provides a single measure of overall accuracy that is not dependent upon a particular threshold [14,16].

With reference to Figure 1, the results show that the value of the AUC is between 0.5 and 1.0. If the value is 0.5, as in the diagonal line on the plot, the scores for two groups do not differ. A score of 1.0 indicates no overlap in the distributions of the group scores.

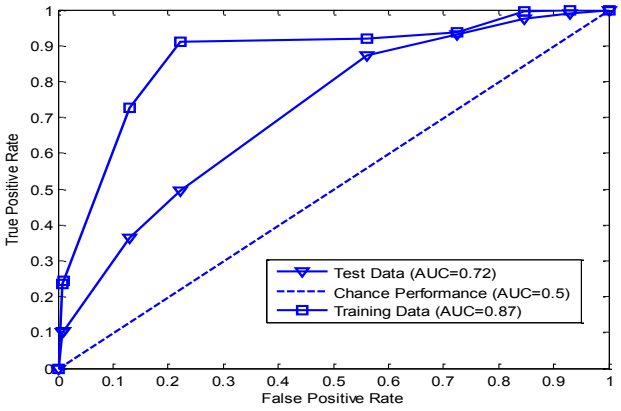

Figure 1: Area under curve (AUC) for training data, test data, and chance performance or random guess

Typically, values of the AUC will not achieve these limits. A value of 0.87 for the AUC means that for 87% of the time a random selection from the positive group will have a score greater than a random selection from the negative class. Usually the AUC for the training data is higher than that for the test data as shown in Figure 1.

This is expected since most classification methods will perform best on the data used to generate the classification rule which is the training data set, and less on the test data set. Researchers argued that some caution is necessary when using ROC methods with biological data since biological cases may not be directly equivalent to the original definition. In particular, the original ROC model assumes that the group allocation is absolutely reliable and each signal is homogeneously presented and processed [24].

In this work, the coil states consist 48% of the data when we use one of the reduction methods of the Define Secondary Structure of Proteins or DSSP definition [17,25,26]. It can be seen clearly that the coils states constitutes approximately 0.5 of the data set. The ROC analysis is applied here to discriminate between coils and non-coils states.

Nine cut scores of 10772 secondary structures outputs sample predicted by the new secondary structure prediction method under consideration [21]. The true positive (TP) row represents the situation that coils states predicted by the prediction method as coils while the false positive (FP) represents the situation that NOT coils states predicted by the prediction method as coils.

As discussed in a previous work [18], the total number of residues in the data base used in training and testing the algorithms is more than 80000 residues. The test sample used in this experiment was chosen from 10772 secondary structure predicted states for its appropriate cut scores and convenience in calculations and representation.

According to their respective cut scores, the true positive rate (TPR) which is the sensitivity of the test and the false positive rate which is (1- specificity) of the test are shown in Table 3 that shows the respective area for each cut score.

Table 3: The cut scores, true positive rate (TPR), false positive rate (FPR), and area under ROC (AUC) for the coil state only prediction

Cut Score	TPR	FPR	Area
1	1.0000	1.0000	0.0710
2	0.9895	0.9287	0.0805
3	0.9752	0.8467	0.1161
4	0.9310	0.7249	0.1471
5	0.8722	0.5618	0.2320
6	0.4949	0.2224	0.0399
7	0.3630	0.1293	0.0279
8	0.1043	0.0097	0.0002
9	0.0998	0.0073	0.0004
10	0.0000	0.0000	0.0000
AUC	-	-	***0.7151***
SE	-	-	*0.0057*

The summation of the nine scores areas represents the area under the curve (AUC). This area under the curve measures the prediction accuracy. The AUC of this test as shown in the table is 0.7151 with standard error (SE) of 0.0057 as calculated from the nine cut scores.

Figure 2 shows the ROC curve travels above the diagonal line and below the top left corner of the graph indicting that the area of this curve is above null guess 0.5 and of course below the perfect prediction 1.0.

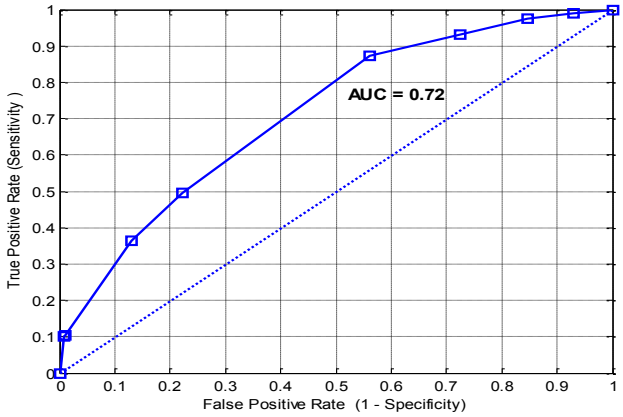

Figure 2: The area under ROC (AUC) for the prediction algorithm considering coil only classification.

The computed AUC as shown in the figure and described in Table 3 is 0.72 and the standard error is 0.0057. This proves that the secondary structure prediction algorithm is able to discriminate the coils states from non-coils with 72% prediction accuracy with a very minor experimental or standard error. Although there is a loss in the entropy in this procedure due to the 0.48 probability of the coils states in the database instead of 0.5, this result is in-line with what has been reported by Subair and Deris [18] using Q3 and SOV measures. Further this result shows a comparative agreement with the correlation coefficients reported by the same authors [18].

In this work, the adoption of the receiver operating characteristics (ROC) analysis aims to determine the discriminative ability of the prediction algorithm to distinguish the coil states only since they constitute approximately 0.5 of the data. This test might be controversial since it is conducted on a three-class classifier and not a binary classifier. The nature of the data set that constitutes the three classes of secondary structure made the data set divided into two classes for the coil states that constitute half of the data set. The ROC analysis test arrived at a conclusion that the prediction algorithm was able to distinguish between two classes (coils/not coils) at 72% of the times.

CONCLUSIONS

The protein secondary structure coils states are classified using the receiver operating characteristics ROC curve and analysis. The trade- off between the true positive rate (sensitivity) and the false positive rate was plotted in an ROC curve and the area under the curve (AUC) was estimated and found that the new prediction algorithm was able to correctly classify 72% of the coils states. Although this accuracy is less than the accuracy discussed in the previous work, when using other evaluation measures like Q3 and SOV [17], the results proved that ROC classification and analysis is reliable in the case of protein data. It is not unusual to find that the accuracy of ROC analysis here is less than the accuracy obtained by the Q3 and SOV measure since there is loss in the entropy of the TP, FP, TN, and FN numbers as discussed earlier in the methodology section. In addition, describing the data set as coils and not coils in its discrete binary meaning is not accurately satisfied in this case.

ACKNOWLEDGMENTS

The author would like to thank Princess Nourah bint Abdulrahman University, Riyadh KSA for supporting this research and other related publications.

REFERENCES

[1] Heilig, R., Eckenberg, R., Petit J. L., Fonknechten ,N, Da Silva et al (2003). The DNA Sequence and Analysis of Human Chromosome 14. Nature. Vol. 421, No. 6923, PP:601-607.

[2] Subair, S (2012) Protein Secondary Structure Prediction: Using Artificial Neural Networks and Information Theory. Lambert Publishing, Germany. ISBN-10: 3847330667 | ISBN-13: 978-3847330660

[3] Pauling, L. and Corey, R. B. (1951). Configurations of Polypeptide Chains With Favoured Orientations Around Single Bonds: Two New Pleated Sheets. Proc. Natl. Acad. Sci. USA. Vol. 37, PP: 729-740.

[4] Kendrew., J. C. Dickerson RE, Strandberg BE, Hart RG, and Davies D.R. (1960). Structure of Myoglobin. Nature. Vol. 185, PP: 422-427.

[5] Schulz, G. E. and Schirmer, R. H. (1979). Principles of Proteins Strcuture. Springer-Verlag., New York., 1979.

[6] Rost., B. (2001). Review: Protein Secondary Structure Prediction Continues To Rise. J. Struct. Biol. Vol. 134, PP: 204–218.

[7] Egan, J. P. (1975). Signal Detection Theory and ROC Analysis. Series in Cognitition and Perceptron. New York: Academic Press.

[8] De Carvalho, V. I.; Jara, A.; Hanson, T. E. and de Carvalho, M. (2013). Bayesian nonparametric ROC regression modeling, Bayesian Analysis, Vol. 8, PP: 623–646.

[9] Swets., J. (1988). Measuring the Accuracy of Diagnostic Systems. Science. Vol. 240, PP: 1285-1293.

[10] Zweig, G. and Campbell. C. C. (1993). Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. Clinical Chemistry. Vol. 39, No. 4, PP: 561-577.

[11] Macskassy, S., Provost, F., & Rosset, S. (2005). ROC Confidence Bands: An Empirical Evaluation. Proceedings of the 22nd International Conference on Machine Learning (ICML). Bonn, Germany.

[12] Devlin, S.A.; Thomas, E. G. and Emerson, S. S. (2013). Robustness of approaches to ROC curve modeling under misspecification of the underlying probability model, Communications in Statistics—Theory and Methods, Vol. 42, PP: 3655–3664.

[13] Bradley, A. P. (1997). The Use of the Area under the ROC Curve in The Evaluation of Machine Learning Algorithms. Pattern Recognition. Vol. 30, No.7, PP: 1145-1159.

[14] Hand, D. J. and Till, R. J. (2001). A Simple Generalisation of the Area under the ROC Curve For Multiple Class Classification Problems, Machine Learning. Vol. 45, PP: 171-186.

[15] Fielding, A. H. and Bell, J. F. (1997). A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. Environ. Conserv. Vol. 24, PP: 38–49.

[16] Hand, D. J. (1997). Construction and Assignment of Classification Rules. NY: John Wiley and Sons.

[17] Kloczkowski, A., Ting, K. L., Jernigan, R. L. and Garnier, J. (2002). Combining the GOR V Algorithm with Evolutionary Information for Protein Secondary Structure Prediction from Amino Acid Sequence. Proteins: Structure. Function and Genetics. Vol. 49, PP: 154-166.

[18] Subair, S (2012) Protein Secondary Structure Prediction: Using Artificial Neural Networks and Information Theory. Lambert Publishing, Germany. ISBN-10: 3847330667 | ISBN-13: 978-3847330660

[19] Subair, S. O. and Deris, S. (2007). Predicting Protein Secondary Structure Using Artificial Neural Networks and Information Theory. In: Application of Agents and Intelligent Information Technologies. Edited by Vijayan Sugumaran PP: 337-362. Idea-Group Publishing. USA

[20] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. and Nielsen, H. (2000). Assessing The Accuracy of Prediction Algorithms For Classification: An Overview. Bioinformatics. Vol. 16, PP: 412-424.

[21] Mirabello, C.; Pollastri, G. Porter, Pale Ale (2013): High-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics, Vol. 29, PP: 2056–2058

[22] Obuchowski, N. (2000). Sample Size Tables For Receiver Operating Characteristic Studies. AJR Am J Roentgenol. Vol. 175, PP: 603-608.

[23] Hughes, G. and Bhattacharya, B. (2013). Symmetry properties of binormal and bi-gamma receiver operating characteristic curves are described by Kullback–Leibler divergences, Entropy, 15, 1342–1356.

[24] Hanley, J. A. and Mcneil, B. J. (1983). The Meaning and Use of the Area Under The Receiver Operating Characteristic (ROC) Curve. Radiology. Vol. 148, PP: 839-43.

[25] Kryshtafovych, A.; Fidelis, K.; Moult, J. (2013) CASP10 results compared to those of previous CASP experiments. Proteins, Vol. 82, PP: 164–174.

[26] Lusci, A.; Pollastri, G.; Baldi, P.(2013) Deep Architectures And Deep Learning In Chemoinformatics The Prediction of Aqueous Solubility For Drug Like Molecules. J. Chem. Inf. Model., Vol. 53, PP: 1563–1575.