Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 29, S. 117-121, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 29, pp. 117-121, December 2021

Research Article

Effect of Benchmark Datasets on Protein Structure Prediction as a Concept

Nuh Azginoglu[1*]

[1*] Kayseri University, Faculty of Engineering, Architecture and Design, Departmant of Computer Engineering, Kayseri, Turkey, (ORCID: 0000-0002-4074-7366), nuhazginoglu@kayseri.edu.tr

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2021 – 21-23 October 2021)

(**DOI:** 10.31590/ejosat.1014716)

ATIF/REFERENCE: Azginoglu, N. (2021). Effect of Benchmark Datasets on Protein Structure Prediction as a Concept. *European Journal of Science and Technology*, (29), 117-121.

Abstract

Knowing the protein structures is essential in understanding the job descriptions of proteins involved in vital functions, drug design, and many more. On the other hand, protein structure prediction is an alternative bioinformatics sub-study field to shorten the process that takes a long time in the laboratory environment. Performance analyzes of the methods developed in this field are generally made on benchmark datasets. The size of the datasets directly affects the algorithm runtime. In this study, how to benchmark datasets are reflected in the results is analyzed. Within the scope of the study, two different benchmark datasets, CB513 and EVASet, and two different protein structure prediction methods, JPred and Porter, were used. The study is a source of inspiration for further studies with the idea of developing benchmark datasets that are comprehensive in terms of protein properties but contain as little data as possible in terms of data size.

Keywords: Protein structure prediction, Benchmark dataset, Concept.

Kıyaslama Veri Kümelerinin Protein Yapı Tahminine Etkisi: Bir Kavram Çalışması

Öz

Protein yapılarının bilinmesi hayati fonksiyonlarda görev alan proteinlerin görev tanımlarının anlaşılabilmesi, ilaç tasarımı ve daha birçok açıdan öneme sahiptir. Protein yapı tahmini ise laboratuvar ortamında oldukça uzun zaman alan süreci kısaltmak için alternatif bir biyoinformatik alt çalışma alanıdır. Bu alanda geliştirilen yöntemlerin performans analizleri genel itibariyle kıyaslama (benchmark) veri kümeleri üzerinden yapılmaktadır. Veri kümelerinin büyüklüğü algoritma çalışma zamanlarına doğrudan etki etmektedir. Bu çalışmada kapsamında kıyaslama veri kümelerinin sonuçlara nasıl yansıdığı analiz edilmiştir. Çalışma kapsamında iki CB513 ve EVASet olmak üzere iki farklı kıyaslama veri kümesi, JPred ve Porter olmak üzere iki farklı protein yapı tahmini yöntemi kullanılmıştır. Çalışma, protein özellikleri açısından geniş kapsamlı ancak, veri büyüklüğü anlamında olabildiğince az veri içerecek olan benchmark veri kümeleri geliştirme fikri itibariyle sonraki çalışmalar için esin kaynağı niteliğindedir.

Anahtar Kelimeler: Protein yapı tahmini, Kıyaslama veri kümesi, Kavram.

[*] Corresponding Author: nuhazginoglu@kayseri.edu.tr

1. Introduction

Proteins are the building blocks of the body and play an essential role in almost all basic functions and growth (van Goudoever et al., 2014; Aydin et al. 2019). From this point of view, knowing the structure of proteins can contribute significantly to many issues such as body defense (Krishnan, 1932), treatment and drug design (Silverman & Holladay, 2014). Protein structure is a subject that has been studied for years, and a hierarchical order consisting of four different classes (KU, 1952) has been proposed to customize the studies and conduct them more efficiently. Protein secondary structure refers to the hydrogen bonding patterns that express the state of the protein between its primary structure consisting of amino acids and its three-dimensional form in space. The way to know the tertiary structure of the protein is through the determination of the secondary structure.

Determination of protein structure in the laboratory is an uphill task that takes a long time. Thanks to the developing technology and bioinformatics studies, structure estimation studies have provided successful results in recent years. Protein secondary structure prediction is also a studied subject, and many methods have been developed specifically for the issue (Atasever et al., 2019, Azginoglu et al., 2020; Jones, 1999; Pirovano & Heringa, 2010). Many different methods have been used for protein secondary structure prediction, such as Hidden Markov Models (Asai, 1993), neural networks (Holley & Karplus, (1989), and deep learning (Spencer et al., 2014). The performance analysis of these methods is carried out on benchmark datasets, and in this respect, it has a crucial place on the subject. The benchmark dataset must represent the problem and must be of acceptable size in terms of computational cost.

Looking at the studies in the literature, while there are studies on new method development, existing methods (Bouziane et al., 2015; Le et al., 2017), and comparison of prediction servers (Bujnicki et al., 2001), no studies are evaluating the protein secondary structure benchmark datasets to the best of our knowledge. Within the scope of this study, the effect of benchmark datasets on the success of the methods and the results obtained were examined. Thus, it was focused on whether it would be inconvenient to use smaller, less costly benchmark datasets instead of large datasets in computational cost.

It was determined that the data sets we used in our study did not differ much in measuring the success of the methods. However, this does not mean that the same result will be obtained when different datasets or techniques are used. If the study we have carried out as a concept study is expanded in terms of datasets, datasets that are qualitatively broad but quantitatively narrow-scoped can be developed with the results to be obtained so that running time can be reduced in parallel with computation cost.

2. Material and Method

2.1. Dataset

Two different benchmark datasets, CB513 (Cuff & Barton, 1999) and EVASet (Koh et al., 2003), were used in this study. These are challenging, and difficult datasets used to measure the performance of methods developed for secondary structure, solvent accessibility, and torsion angle estimation. We used these datasets only for secondary structure prediction. CB513 contains 513 proteins and a total of 84119 amino acids (residue), while EVASet includes 2876 proteins and a total of 584595 amino acids after excluding proteins containing less than 30 amino acids. Datasets are in the form of multiple Fasta (Pearson & Lipman, 1988) files in a single text file.

2.2. Dataset

2.2.1. JPred

JPred is a protein secondary structure prediction server using the JNet algorithm. In this study, Jpred4 (Drozdetskiy et al., 2015), the latest version of JPred, was used as a secondary structure prediction method.

We used JPred4 via web server, which has certain restrictions (http://www.compbio.dundee.ac.uk/jpred4). JPred4 does not accept submissions for more than 200 proteins at the same time. For this reason, we first divided our dataset into parts from $\leq$ 200. Therefore, it was planned to submit eighteen different jobs as (200 x 2) + (113 x 1) three for CB513 containing a total of 513 proteins, and (200 x 14) + (76 x 1) fifteen jobs for EVAset containing a total of 2876 proteins.

JPred4 does not accept proteins containing more than 800 residues of amino acids as a single input. For this reason, we divided in two the amino acid sequences of 25 proteins containing more than 800 amino acids that we detected in EVASet (The IDs of these proteins are: 1b0pA, 1bglA, 1bxrA, 1c7sA, 1clqA, 1e7uA, 1ej6A, 1ej6C, 1epwA, 1eulA, 1ffyA, 1h3nA, 1h6zA, 1hq7sA, 1hty 1jncwA, 1kjvA, 1hty5A, 1kcv2A, 1hty1, 1qb4A, 1qgkA, 2btvA.) and submitted them separately. Therefore, a total of 19 jobs, 3 for CB513 and 15+1=16 for EVASet, were submitted to the JPred4 web server with different names. The results of the proteins submitted in two fragments were combined after the estimation process.

Fasta files containing multiple proteins were submitted to the JPred4 server as advanced options with Single Sequence (Batch Mode) and Skip Searching PDB Before Prediction options. JPred accepts only a certain number of submissions per day from an e-mail address. For this reason, we need to state that submissions are made from five different e-mail accounts not to delay the works. Under these circumstances, we can say that the JPred4 estimates for the datasets we use are completed in less than half a day, and, it's a reasonable time.

After the JPred4 estimates are completed, an e-mail is informed, and the download link of the *tar.gz file containing the results is sent. After all the predictions were achieved, *.name files to determine the name of the relevant protein, and *.simple.html files to determine the secondary structure prediction result were used among the result folders containing many files from alignment files to structural matrices.

```
Query_name: 1qgpA
Query_length: 76

LSSHFQELSIYQDQEQRILKFLEELGEGKATTAHDLSGKLGTPKKEINRVLYSLAKKGKL
CCCCCCCCCCCCCCHHHHHHHHHHHHCCCCCCEEHHHHHHHHHCCCHHHHHHHHHHHHHCCCC
EEEEEEEEEbeEEeeEeBbEbBEEbEEEEebeBeeBBEEbEbEeEeBbEbBbeBeEEEeB


QKEAGTPPLWKIAVSD
EECCCCCCCEEECCCC
eEEEEEbbebEbeEEE
```

Figure 1. Porter Estimation Result for a Sample Protein

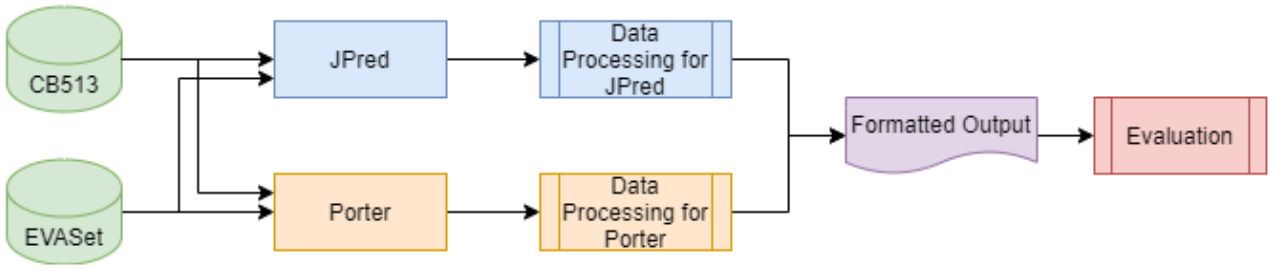

Figure 2. Experimental Setup

An automated programmatic way has been developed to select the files to be used and the deletion of the files that will not be used from the disk (it can take up approximately 200 MB of space for just a protein depending on the protein size). Python was used as the programming language, and Jupyter Notebook was used as the programming editor.

2.2.1. Porter

Porter is an ensemble of 25 different bidirectional recurrent neural networks, a protein secondary structure, and a solvent accessibility prediction server (PaleAle for solvent accessibilty prediction). In this study, Porter 4.0 (Mirabello & Pollastri, 2013), the latest version of Porter, was used for protein secondary structure prediction.

The Porter server (http://distillf.ucd.ie/porterpaleale/) accepts input in Fasta format up to 64 kilobytes (KB). For our input size not to exceed 64 KB, the datasets were submitted to the system as divided into sections of approximately 250 proteins. The CB513 ((1 x 250) + (1 x 263) = 513 proteins) is divided into two different, and the EVASet ((11 x 250) + (1 x 126) = 2876 proteins) was divided into twelve different jobs. Porter completed all the submitted jobs in about half a day, similar to JPred. Prediction result (output) of The Porter Server for a sample protein is given in Figure-1.

After the Porter predictions were completed, the results obtained for each dataset were combined. A new file was generated with two lines separated for each protein. (name of the protein and the protein secondary structure prediction query). Python programming language was used during this generating process. In its final form, the same format output was produced using both JPred and Porter prediction results.

2.3. Experimental Setup

In this study, in which two benchmark datasets and two different methods are used, firstly, the data is converted into the format accepted by the methods. Then the prediction results are obtained by giving them to the system. Outputs in different forms were converted into a standard format, and then their prediction performance was evaluated. Experiment setup and workflow are given in Figure-2.

3. Results

Average Three-state Prediction Accuracy (Q3) (Rost & Eyrich, 2001), Segment Overlap Measure (SOV) (Zemla et al., 1999), Class-specific Recall (R), and Precision (P) were used to evaluate the results of the experiments conducted in this study. The overall accuracy was calculated by taking the percentage of the value obtained by dividing the correctly predicted number of amino acids by the total number of amino acids. Q3 is one of the most popular statistical performance measures. Here three symbol denotes secondary structure labels. On the other hand, SOV aims to calculate the overlap ratio between the actual class label and the predicted class label segments.

In Table-1 and Table-2, confusion matrixes obtained using the CB513 dataset are given. Table-1 presents the results obtained with JPred, and Table-2 presents the results obtained with the Porter method. The vertical axis represents the actual label in the tables, and the horizontal axis represents the estimation results. Table-5 and Table-6 show the results obtained using the EVASet dataset. Table-3 and Table-4 give the P, R, Q3, and SOV results of JPred and Porter methods, respectively, using the CB513 dataset. Values were calculated both on a class level and as a total. All values are given in the table, and the essential values for us to compare are the Q3 and SOV values. Table-7 and Table-8 present the results of the JPred and Porter method obtained using the EVASet dataset. Both dataset-based and method-based comparisons were made since two different datasets (Cb513 and EVASet), and two different methods (JPred and Porter) were used during the experiments.

When we make a comparison as a data set, when we look at the results obtained from the JPred method, it is seen that there is a 0.40% (78.31%-77.91%) difference in the Q3 results. Similarly, a difference of 0.40% (73.79%-73.39%) is also observed in SOV results. We can say that this difference is not statistically significant for either metric (z-score and p-value values were used at this point). For this reason, the dataset difference in CB513 and EVASet did not significantly affect the results. In the Porter method, the difference between datasets for Q3 is 0.25% (82.67%-82.42%) for CB513 and 0.59% for EVASet. (79.50%-78.91%). These values were also not statistically significant, and there was no obvious difference between the two data sets in the experimental results. When the methods are compared, it is seen that the Porter method is more successful on CB513 with a rate of 4.36% (Q3) and 5.61% (SOV), and on EVASet, 4.51% (Q3) and 5.52% (SOV) is more successful than the JPred method.

However, a detailed comparison was not made for the working time of JPred and Porter, as it was not within the scope of the study. Since the number and size of the jobs waiting in the queue cannot be known clearly, we would like to point out that such a comparison can only be made by installing these two methods on the local server.

If we evaluate the results obtained, considering that both CB513 and EVASet are difficult datasets, the fact that there is not much difference in the success rates based on the dataset shows that the dataset containing less protein is computationally preferable. In this respect, CB513 (513 protein, 84119 amino acids) among the two datasets shows that it is preferable to EVASet (2876 protein, 584595 amino acids) due to the low number of both protein and amino acids it contains in general terms. When an evaluation was made within the scope of the methods, it was seen that the Porter method gave better results in the range of about 4%-6% compared to JPred, and therefore it was preferable.

Table 1. Confusion Matrix for JPred (CB513)

		Predicted			
		H	***E***	***L***	***Total***
Actual	***H***	22476	392	6229	29097
	E	495	12833	5731	19059
	L	2557	2843	30563	35963
	Total	25528	16068	42523	84119

Table 2. Confusion Matrix for Porter (CB513)

		Predicted			
		H	***E***	***L***	***Total***
Actual	***H***	25134	249	3714	29097
	E	250	14453	4356	19059
	L	2930	3078	29955	35963
	Total	28314	17780	38025	84119

Table 3. Accuracy Measures for JPred (CB513)

	P	**R**	**SOV**
H	88.04	77.25	79.04
E	79.87	67.33	72.84
L	71.87	84.98	70.10
Total	78.31	**78.31 (Q3)**	**73.79 (SOV)**

Table 4. Accuracy Measures for Porter (CB513)

	P	**R**	**SOV**
H	88.77	86.38	86.47
E	81.29	75.83	78.99
L	78.78	83.29	74.44
Total	**82.67**	**82.67**	**79.50**

Table 5. Confusion Matrix for JPred (EVASet)

		Predicted			
		H	***E***	***L***	***Total***
Actual	***H***	165157	3247	44965	213369
	E	3449	85309	37000	125758
	L	20035	20435	204998	245468
	Total	188641	108991	286963	584595

Table 6. Confusion Matrix for Porter (EVASet)

		Predicted			
		H	***E***	***L***	***Total***
Actual	***H***	185258	2226	25885	213369
	E	1815	96947	26996	125758
	L	22912	22915	199641	245468
	Total	209985	122088	252522	584595

Table 7. Accuracy Measures for JPred (EVASet)

	P	**R**	**SOV**
H	87.55	77.40	78.93
E	78.27	67.84	72.76
L	71.44	83.51	68.98
Total	77.91	**77.91 (Q_3)**	**73.39 (SOV)**

Table 8. Accuracy Measures for Porter (EVASet)

	P	**R**	**SOV**
H	88.22	86.83	86.38
E	79.41	77.09	79.26
L	79.06	81.33	72.68
Total	82.42	**82.42 (Q_3)**	**78.91 (SOV)**

4. Conclusion and Recommendations

Within the scope of the study, the secondary structures of the proteins in the CB513 and EVASet benchmark datasets were estimated by JPred and Porter methods. The pre-processed data were pre-processed to make them suitable for the methods used, and then the results were obtained from both approaches. The outputs of these methods, which give output in different formats, have been post-processed, and a standard format has been received. Finally, the prediction successes were determined.

The results show that smaller datasets with less computational cost can be sufficient for performance evaluation, particularly for JPred and Porter. However, we can say that the characteristics of the dataset and the method are also crucial points. This study aims to give a different perspective on benchmark datasets for protein structure prediction studies.

5. Acknowledge

This study was supported as Project Number: FHD-2021-1045 by Kayseri University Scientific Research Projects Unit. We thank Kayseri University Scientific Research Projects unit for their contributions.

References

Asai, K., Hayamizu, S., & Handa, K. I. (1993). Prediction of protein secondary structure by the hidden Markov model. Bioinformatics, 9(2), 141-146.

Atasever, S., Azgınoglu, N., Erbay, H., & Aydın, Z. (2021). 3-State Protein Secondary Structure Prediction based on SCOPe Classes. Brazilian Archives of Biology and Technology, 64.

Aydin, Z., Azginoglu, N., Bilgin, H. I., & Celik, M. (2019). Developing structural profile matrices for protein secondary structure and solvent accessibility prediction. Bioinformatics, 35(20), 4004-4010.

Azginoglu, N., Aydin, Z., & Celik, M. (2020). Structural profile matrices for predicting structural properties of proteins. Journal of Bioinformatics and Computational Biology, 18(04), 2050022.

Bouziane, H., Messabih, B., & Chouarfia, A. (2015). Effect of simple ensemble methods on protein secondary structure prediction. Soft Computing, 19(6), 1663-1678.

Bujnicki, J. M., Elofsson, A., Fischer, D., & Rychlewski, L. (2001). LiveBench-1: Continuous benchmarking of protein structure prediction servers. Protein Science, 10(2), 352-361.

Cuff, J. A., & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics, 34(4), 508-519.

Drozdetskiy, A., Cole, C., Procter, J., & Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. Nucleic acids research, 43(W1), W389-W394.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology, 292(2), 195-202.

Holley, L. H., & Karplus, M. (1989). Protein secondary structure prediction with a neural network. Proceedings of the National Academy of Sciences, 86(1), 152-156.

Koh, I. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Eswar, N., ... & Rost, B. (2003). EVA: evaluation of protein structure prediction servers. Nucleic Acids Research, 31(13), 3311-3315.

Krishnan, K. V. (1932). The Defence Mechanism of the Human Body. The Indian medical gazette, 67(11), 637.

KU, L. L. (1952). Lane medical lectures: proteins and enzymes.

Mirabello, C., & Pollastri, G. (2013). Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics, 29(16), 2056-2058.

Le, Q., Sievers, F., & Higgins, D. G. (2017). Protein multiple sequence alignment benchmarking through secondary structure prediction. Bioinformatics, 33(9), 1331-1337.

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences, 85(8), 2444-2448.

Pirovano, W., & Heringa, J. (2010). Protein secondary structure prediction. Data Mining Techniques for the Life Sciences, 327-348.

Rost, B., & Eyrich, V. A. (2001). EVA: large-scale analysis of secondary structure prediction. Proteins: Structure, Function, and Bioinformatics, 45(S5), 192-199.

Silverman, R. B., & Holladay, M. W. (2014). The organic chemistry of drug design and drug action. Academic press.

Spencer, M., Eickholt, J., & Cheng, J. (2014). A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM transactions on computational biology and bioinformatics, 12(1), 103-112.

Van Goudoever, J. B., Vlaardingerbroek, H., van den Akker, C. H., de Groof, F., & van der Schoor, S. R. (2014). Amino acids and proteins. Nutritional Care of Preterm Infants, 110, 49-63.

Zemla, A., Venclovas, Č., Fidelis, K., & Rost, B. (1999). A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins: Structure, Function, and Bioinformatics, 34(2), 220-223.