Protein classification by using four approaches for extraction of the protein ray-based descriptor

Georgina Mirceva
Faculty of computer science and engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
georgina.mirceva@finki.ukim.mk

Andrea Kulakov
Faculty of computer science and engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
andrea.kulakov@finki.ukim.mk

***Abstract*—The knowledge about the protein molecules, and how they influence the processes in the humans is very worth, because it is really needed in order to develop new drugs for diseases. In proteomics, one of the most important tasks is solving the problem of classification of protein molecules. The literature provides plethora of methods that could be used for this task. However, it is still an open issue where still there is a need for fast computational methods that would provide accurate classification of proteins. In this paper, we focus on solving this task. For that purpose, first, we extract feature vectors that hold information about the main features of the proteins. The feature vectors that are used in this study are obtained by following the procedure for extraction of our protein ray-based descriptor that we have introduced in our former studies. For that purpose, the skeleton of the protein is interpolated with predefined number of interpolation points, and then the elements of the feature vector are extracted as Euclidean distances between the interpolation points and center of mass. Besides this approach, in this study we also use three additional approaches for extraction of the feature vectors, where we focus on the change of the Euclidean distance to the center of mass between two consecutive interpolation points. After extracting feature vectors, next we apply several well-known classification methods in order to generate classification model. We present the results obtained with these four approaches used for extraction of the feature vectors.**

Keywords—protein structure, protein classification, protein ray-based descriptor

I. INTRODUCTION

Proteins are an important component of the living organisms. The understanding of protein molecules and their influence in the processes in which they participate is essential in order to be able to design new drugs for various diseases. Proteomics is an area where the focus is on discovery, analysis and understanding of the proteins. One of the most important tasks in this research area is the problem of classification of protein structures, which could help to understand these structures and to determine the functions that they make have.

The data about the protein structures that are discovered are deposited in the Protein Data Bank (PDB) [1], [2]. Nevertheless, these data are not worth if we do not determine the functions of the protein structures. It is supposed that the proteins that have common ancestor and belong to same class, also share similar functions. Therefore, the protein classification problem is among the most important tasks that should be solved regarding proteins. The literature provides plethora of methods that could be used to classify protein structures.

SCOP (Structural Classification Of Proteins) [3] is one of the most widely known methods used for protein classification. With this method, the classification of proteins is manual, thus the classification time is too long. Therefore, there is need for methods where the classification would be done in automatic manner. In the literature, also automatic and semiautomatic methods could be found. CATH (Class, Architecture, Topology and Homologous superfamily) [4] is among the semiautomatic methods, because it first tries to classify the inspected protein in automatic way, and if it is not possible for a given protein, in the next stage the protein is classified manually by human experts.

Another class of methods tries to classify protein structures by making alignment of their sequences. Needleman–Wunch [5], BLAST [6] and PSI-BLAST [7] are among the most important representatives from this category. However, the methods based on sequence alignment may not discover similarity between proteins whose structures are very similar if they do not have similar sequences. Therefore, it is better to find the similar structures by aligning the structures of the proteins, rather than making alignment of their sequences. CE [8], MAMMOTH [9] and DALI [10] are among the most broadly known methods from this category. Of course, there is also third category, where the methods perform both sequence alignment and structure alignment, as in the methods SCOPmap [11] and FastSCOP [12].

However, the methods from the three categories mentioned above, which make alignment of protein sequences or/and structures need long time to classify an inspected protein. To overcome this problem, various methods extract vectors with features for the proteins, thus later the comparison between the proteins is made by calculating the distance between their feature vectors. In this category, there are both methods that consider features of the protein sequences [13] or protein structures [14]. After extraction of the proteins' feature vectors, then a classification model could be built by using some classification method.

In this paper, we use an approach as it is used in the methods from the last category. Namely, we extract feature vectors for the proteins, and then we generate prediction model. In our previous study [15], we presented several approaches for finding similar protein structures based on the features of their tertiary structures. In [15], we focused on the task how to find similar protein structures, known as protein retrieval. In this paper, our aim is not to find the homologous proteins that are similar with the inspected proteins, but our aim here is to make decision in which class the inspected protein should be classified in.

In this study, we perform protein classification by using the protein ray-based descriptor [15]. In [15], we showed that the protein ray-based descriptor is very accurate for finding similar proteins, even though it is very simple. The extraction of the protein ray-based descriptor starts with interpolation of the protein skeleton. Then, the elements of the feature vector are extracted as Euclidean distances between the obtained interpolation points and the center of mass. Besides this approach, in this paper additionally we consider three additional approaches where we analyze how these Euclidean distances change as we traverse the skeleton of the protein from one interpolation point to its consecutive interpolation point. After extraction of the feature vectors, then we apply several classification methods in order to build classification model for making class decisions. In this paper we use the following classification methods: C4.5 [16], Naive Bayes [17], Bayesian Network [18], k-nearest neighbors (knn) [19] and Support Vector Machines (SVM) [20], [21].

The remaining of this paper is structured in this way. In Section 2, we give description of the original protein ray-based descriptor. Besides this approach, where the Euclidean distances between the interpolation points and center of mass are used as features, we also present three other approaches where the features of the feature vectors are calculated based on the difference between two consecutive elements of the protein ray-based descriptor. Section 3 is focused on the evaluation of the approaches presented in this study. This section provides results obtained by using the four approaches for feature vector extraction combined with various classification methods for model generation. The conclusions are presented in Section 4, which also contains several ideas for future work.

II. Protein Classification by Using the Protein Ray-Based Descriptor

In this study, the classification of protein structures is made by using feature vectors, which contain information about the geometrical features of the proteins. For that purpose, first, for each training protein structure a corresponding feature vector is generated. In the second stage, prediction model is built by using some classification method, where the elements of the feature vectors correspond to the attributes in the data set that is used for training the model. Once the prediction model is generated, next, we can make decisions in which class a given query protein belongs to.

Protein molecules contain one or several chains. The ground true data that would be used in this study contain information about the classes of the protein chains, therefore we need to obtain a feature vector for each protein chain that would be used in the study. In this way, the samples in the data set correspond to the individual protein chains of the protein molecules.

As it was mentioned before, in this paper we use four approaches for extraction of the protein ray-based descriptor. These approaches are very close to each other, the difference is just in the last step where the final values for the elements of the feature vector are calculated. The first approach corresponds to the original version of our protein ray-based descriptor that was presented in our previous study [15]. The remaining three approaches focus on the difference between the consecutive elements in the feature vector. First, we give description of the first approach, as it is presented in [15], and then we give description about the differences introduced with the other three approaches.

A. Protein Ray-Based Descriptor

In the extraction of the protein ray-based descriptor [15], we consider only the C_α atoms, which form the protein's skeleton. We take into account the information how the C_α atoms of the protein are positioned in the 3D space, thus forming its 3D model. First, this model is scaled thus obtaining model where the Euclidean distance between the most distant C_α atom and the center of mass is 1. In this way, scale invariance is provided.

Different protein chains have different number of C_α atoms, thus we are not able to extract some feature vector directly, by considering all these atoms, because in that way we will obtain feature vectors with different lengths. Therefore, we interpolate the skeleton of the protein backbone with interpolation points, where the number of interpolation points is predefined and is equal for each protein chain. The skeleton of the protein chain could be seen as a curve in the 3D space, where the consecutive C_α atoms are connected by this curve. The idea of the interpolation is to find predefined number of points that would be good representatives of this curve. In [15], we considered two different ways how to make interpolation of the protein skeleton. In this study, we use uniform interpolation, which showed as better choice according to the results presented in [15].

Next, we give description how the interpolation of the skeleton is made. For that purpose, we calculate the length of the protein's skeleton by summing up the Euclidean distances between each pair of two consecutive C_α atoms. Then, we need to find interpolation points that are placed over the skeleton and form segments over the skeleton with same length. In this study, we interpolate the skeleton with N=64 interpolation points, thus we obtain vectors with 64 features.

After finding the interpolation points, the feature vector could be extracted. When we introduced the protein ray-based descriptor, we were inspired from the ray descriptor [22] that is used for making retrieval of similar 3D objects. In the ray descriptor, for a given 3D object a mash model is obtained, and then the feature vector is extracting by calculating the distances between the center of mass and the vertices in the mash model. Similarly, here we extract the elements of the feature vector by calculating the Euclidean distances between the interpolation points that were found in the previous step and the center of mass. By calculating the feature vector's elements in this way, we provide feature vector that is invariant to both translation and rotation.

B. Four Approaches for Extraction of the Protein Ray-Based Descriptor

The description presented in the previous sub-section corresponds to the first approach used for extraction of the protein ray-based descriptor. With this approach, the interpolation points are examined individually, because each element of the feature vector focuses on one of the interpolation points. If we try the visualize this approach, the idea of the protein ray-based descriptor is to present how the skeleton of the protein goes towards the center of mass or goes towards the protein surface. If we assume that there are concentric spheres in the space where the protein is placed, then, with the protein ray-based descriptor we describe how the skeleton of the protein passes from one concentric sphere to another as we traverse along the backbone. This is

illustrated in our previous paper [23], where we build HMM (Hidden Markov Model) for classification of proteins.

In this study, we came with the same idea that we used in [23], where we tried several different ways in order to represent the Euclidean distances that are obtained. With the first approach, which is actually the original version of the protein ray-based descriptor [15], the elements of the feature vector $f_{Eucl} = [f_1, f_2, \ldots, f_N]$ are calculated as Euclidean distances between the interpolation points and center of mass. The i-th elements of the feature vector is $f_i = D_i$, for i=1, 2,…, N, where D_i denotes the Euclidean distance between the i-th interpolation point and the center of mass. In the second approach, we analyze the difference between two consecutive interpolation points, therefore the feature vector is calculated as $f_{diff} = [diff_1, diff_2, \ldots, diff_{N-1}]$, where $diff_i = f_i - f_{i+1} = D_i - D_{i+1}$, i=1, 2,…, N–1. With the third approach, the feature vector is calculated as $f_{abs} = [abs_1, abs_2, \ldots, abs_{N-1}]$, where $abs_i = |diff_i|$, i=1, 2,…, N–1. The fourth approach just considers whether the difference between two consecutive Euclidean distances rises or declines, without considering the amount of the increase or decrease. Thus the feature vector with the fourth approach is calculated as $f_{sign} = [sign_1, sign_2, \ldots, sign_{N-1}]$, where $sign_i = sign(diff_i)$, i=1, 2,…, N–1, where the function $sign(x)$ is defined as

$$sign(x) = \begin{cases} 1, \text{if } x > 0 \\ 0, \text{if } x = 0. \\ -1, \text{if } x < 0 \end{cases} \quad (1)$$

C. Classification Methods

After extraction of the feature vectors for all training samples, next, we generate a model that would be later used for making decisions about the classes of the test samples. The elements of the feature vectors correspond to the descriptive attributes that are used to describe the samples, the protein chains in this case.

In this paper we use several classification methods that are commonly used for solving classification task, i.e.: C4.5 [16], Naive Bayes [17], Bayesian Network [18], k-nearest neighbors (knn) [19] and Support Vector Machines (SVM) [20], [21]. For knn, we use k=1.

III. Experimental Results

The standard of truth used in this study is obtained from part of the knowledge from the SCOP database [3], which holds knowledge how the protein chains are classified by human experts. With the SCOP method [3], the classification is in hierarchical manner, where the SCOP domain level is considered as most important for protein classification. Therefore, in this study the classes correspond to the different SCOP domains at domain level. We formed a data set that holds 6145 protein chains from 150 SCOP domains. The protein chains are approximately uniformly distributed in these SCOP domains. This means that we have 6145 samples and 150 classes. This data set is divided in a ratio 90:10 into training and test data set, thus obtaining 5531 samples for training the models and 614 samples for testing the models. The prediction power of the obtained classification models is estimated using the classification accuracy evaluation measure. Additionally, we also present the results obtained for the AUC-ROC (Area under the ROC curve) evaluation measure.

We made experiments by using the four approaches for extraction of the protein ray-based descriptor in combination with the five classification methods listed before. The results for the classification accuracy are given in Table 1, while the results for AUC-ROC are given in Table 2. The bolded values correspond to the best results obtained with each of the classification methods.

As it can be seen from the results, by using the first approach, where the Euclidean distances between the interpolation points and center of mass are used as feature vectors' elements, the best results are obtained. Then, the second approach follows, where we analyze how the Euclidean distances changes from one element of the vector to another. Or described in other way, we analyze how the protein backbone goes towards its center or its surface. With the last two approaches, we misplace some of the information that is considered with the second approach. With the third approach, we evade the evidence whether the difference between two consecutive Euclidean distances is increased or decreased, while with the fourth approach we keep that information, but we lose the information about the amount of the increase or decrease. Although with the fourth approach we obtain feature vector that requires less memory to be kept, in general it showed as better choice than the third approach.

The first approach corresponds to the absolute representation in [23], the second approach corresponds to the relative representation in [23], while the fourth approach is almost as the binary representation in [23]. If we compare the results from [23] and the results from this study, it is evident that with the HMM used in [23] the best results are obtained with relative representation (corresponds to the second approach from this study), while in this study the best results are obtained with the absolute representation (corresponds to the first approach from this study). The reason for that is the type of classification method that is used. With HMM, the model is defined by a final number of states, and the next state depends on one or several previous states. This type of model is appropriate if we want to analyze sequences, in our case the sequence corresponds to the transition of the skeleton from one concentric sphere to another. In this study, we use other type of classification methods, where we do not have states that are dependent from the previous states, and therefore the results are different, as expected.

Regarding the classification methods, we can conclude that, in general, with Bayesian Network best results are obtained, then knn, SVM and Naïve Bayes follow, while C4.5 showed as the worst choice in this case.

TABLE I. The Results for Classification Accuracy by Using the Four Approaches for Extraction of the Protein Ray-Based Descriptor in Combination with the Five Classification Methods

Classification Method	Approach for extraction of the protein-ray based descriptor			
	Eucl	***diff***	***abs***	***sign***
C4.5	**92.997**	91.042	88.111	89.739
Naive Bayes	**94.625**	92.182	90.717	90.717
Bayesian Network	**96.417**	95.603	94.788	93.322
knn	**98.534**	97.883	97.231	97.394
SVM	**97.557**	96.743	96.254	97.231
AVERAGE	**96.026**	94.691	93.420	93.681

TABLE II. THE RESULTS FOR AUC-ROC BY USING THE FOUR APPROACHES FOR EXTRACTION OF THE PROTEIN RAY-BASED DESCRIPTOR IN COMBINATION WITH THE FIVE CLASSIFICATION METHODS

Classification Method	**Approach for extraction of the protein-ray based descriptor**			
	Eucl	***diff***	***abs***	***sign***
C4.5	**0.971**	0.964	0.947	0.958
Naive Bayes	**0.996**	**0.996**	0.994	0.995
Bayesian Network	**1.000**	**1.000**	0.999	0.999
knn	**0.993**	0.990	0.987	0.992
SVM	0.995	0.994	**0.996**	0.990
AVERAGE	**0.991**	0.989	0.985	0.987

IV. CONCLUSION AND FUTURE WORK

This study focused on solving a problem for protein classification based on the features of the tertiary structure of proteins. For that purpose, first we generated feature vectors for the training samples, and then we applied five well-known classification methods for generating classification models. In this paper, we used our protein ray-based descriptor as a feature vector. However, besides its original version (denoted as the first approach), where the Euclidean distances between the interpolation points and center of mass are used as features, we also considered three additional approaches that focus on the changes of the values of the previously extracted features. With the second approach, we present how the values of the features increase or decrease as we navigate along the skeleton of the protein, while the third and fourth approach consider only the amount of the change or the direction of the change, respectively.

For evaluation, we used a part of the knowledge from the SCOP database, as a ground truth. The results showed that with the classification methods used in this study, it is best to use the original version of the protein ray-based descriptor (the first approach). Next, the second approach follows because it preserves both the amount and the direction of the change. The fourth approach although requires less memory than the third approach, in general it showed as better than the third approach, meaning that it is more important on which places there are changes of the direction in which the skeleton moves (towards the center or towards the surface), while the amount of this change is a little bit less important than the direction.

Continuing the research for solving the protein classification problem, we plan to extend our studies in several directions. We believe that the choice of the feature vector is the most important factor, thus if we have better attributes, we can later make more accurate models. Therefore, we will continue our hunt for other feature vectors and other features that are the most important and relevant for this task. Besides geometrical features, also some features of the primary and secondary structure of the proteins could be considered. Of course, our effort will be also put on looking for the most appropriate classification method, that would lead to most accurate models.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of computer science and engineering at the "Ss. Cyril and Methodius University in Skopje", Skopje, Macedonia.

REFERENCES

[1] 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 Res., vol. 28, no. 1, pp. 235–242, January 2000.

[2] RCSB Protein Data Bank, http://www.rcsb.org, 2019.

[3] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "Scop: a structural classification of proteins database for the investigation of sequences and structures," J. Mol. Biol., vol. 247, no. 4, pp. 536–540, April 1995.

[4] C. A. Orengo, A. D. Michie, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH – a hierarchic classification of protein domain structures," Structure, vol. 5, no. 8, pp. 1093–1108, August 1997.

[5] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J. Mol. Biol., vol. 48, no. 3, pp. 443–453, March 1970.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," J. Mol. Biol., vol. 215, no. 3, pp. 403–410, October 1990.

[7] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Res., vol. 25, no. 17, pp. 3389–3402, September 1997.

[8] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," Protein Eng., vol. 11, no. 9, pp. 739–747, September 1998.

[9] A. R. Ortiz, C. E. Strauss, and O. Olmea, "Mammoth: an automated method for model comparison," Protein Sci., vol. 11, no. 11, pp. 2606–2621, November 2002.

[10] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," J. Mol. Biol., vol. 233, no. 1, pp. 123–138, September 1993.

[11] S. Cheek, Y. Qi, S. S. Krishna, L. N. Kinch, and N. V. Grishin, "SCOPmap: automated assignment of protein structures to evolutionary superfamilies," BMC Bioinformatics, vol. 5, pp. 197–221, December 2004.

[12] C. H. Tung and J. M. Yang, "fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies," Nucleic Acids Res., vol. 35, W438–W443, July 2007.

[13] K. Marsolo, S. Parthasarathy, and C. Ding, "A multi-level approach to SCOP fold recognition," IEEE Symposium on Bioinformatics and Bioeng., pp. 57–64, October 2005.

[14] P. H. Chi, Efficient protein tertiary structure retrievals and classifications using content based comparison algorithms, PhD thesis, University of Missouri-Columbia, 2007.

[15] G. Mirceva, I. Cingovska, Z. Dimov, and D. Davcev, "Efficient approaches for retrieving protein tertiary structures," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 9, no. 4, pp. 1166–1179, July/August 2012.

[16] R. Quinlan, C4.5: Programs for Machine Learning, 1st ed. San Mateo, CA, USA: Morgan Kaufmann Publishers, 1993.

[17] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," In: Besnard, P., Hanks, S. (Eds.), Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, USA, pp. 338–345, 1995.

[18] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Mach. Learn., vol. 29, no. 2–3, pp. 131–163, November/December 1997.

[19] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," Mach. Learn., vol. 6, no. 1, pp. 37–66, January 1991.

[20] V. Vapnik, The Nature of Statistical Learning Theory, 2nd ed., Springer, New York, 1999.

[21] C. J. C. Burges, "A tutorial on support vector machine for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121–167, 1998.

[22] D. V. Vranic, 3D Model Retrieval, Ph.D. Thesis, University of Leipzig, 2004.

[23] G. Mirceva, M. Mirchev, and D. Davcev, "Hidden Markov Models for classifying protein secondary and tertiary structures," Journal of Convergence, vol. 1, no. 1, pp. 57–64, 2010.