CLASSIFICATION OF PROTEIN STRUCTURES BY USING FUZZY KNN CLASSIFIER AND PROTEIN VOXEL-BASED DESCRIPTOR

Prof. Dr. Mirceva G., Prof. Dr. Naumoski A., Prof. Dr. Kulakov A.
Faculty of computer science and engineering, Ss. Cyril and Methodius University in Skopje, Skopje, R. Macedonia

georginamirceva@gmail.com

Abstract: *Protein classification is among the main themes in bioinformatics, for the reason that it helps understand the protein molecules. By classifying the protein structures, the evolutionary relations between them can be discovered. The knowledge for protein structures and the functions that they might have could be used to regulate the processes in organisms, which is made by developing medications for different diseases. In the literature, plethora of methods for protein classification are offered, including manual, automatic or semiautomatic methods. The manual methods are considered as precise, but their main problem is that they are time consuming, hence by using them a large number of protein structures stay uncategorized. Therefore, the researchers intensively work on developing methods that would afford classification of protein structures in automatic way with acceptable precision. In this paper, we propose an approach for classifying protein structures. Our protein voxel-based descriptor is used to describe the features of protein structures. For classification of unclassified protein structures, we use a k nearest neighbors classifier based on fuzzy logic. For evaluation, we use knowledge for the classification of protein structures in the SCOP database. We provide some results from the evaluation of our approach. The results show that the proposed approach provide accurate classification of protein structures with reasonable speed.*

Keywords: PROTEIN STRUCTURE, PROTEIN CLASSIFICATION, PROTEIN VOXEL-BASED DESCRIPTOR, K NEAREST NEIGHBORS, FUZZY LOGIC

1. Introduction

Bioinformatics community intensively analyze protein molecules for the reason that they are essential in the organisms. The processes in the organisms can be controlled by the interactions of proteins. The knowledge gathered from the examination of proteins may be used for drug design, where the functions of the proteins in these interactions is taken into consideration. Using various types of techniques, the structures of the protein molecules have been examined. The information about protein structures is stored in the Protein Data Bank (PDB) [1], [2]. Due to the fast improvements in these techniques, the structures of proteins are determined with fast speed. However, the methods that provide classification of proteins are not able to classify them with the same speed that leads to gap in the number of proteins with determined structures and the number of proteins that are classified. Thus, there is a great necessity for developing methods for classification of protein structures.

The current literature offers various methods for classification of proteins. In the SCOP (Structural Classification Of Proteins) method [3] the decision is done in manual way, where the experts visually examine the proteins. However, the manual methods are time consuming and are not able to follow the speed of determining novel protein structures. Therefore, there are also automatic methods and semiautomatic methods. For example, the CATH (Class, Architecture, Topology and Homologous superfamily) method [4] tries to classify proteins in automatic manner first, and if the decision could not be made, then manual decision is performed.

Some methods align the sequences of the protein structures, also known as primary structures, in order to perform classification of the proteins. The most known methods in this group are Needleman–Wunch [5], BLAST [6] and PSI-BLAST [7]. However, the protein sequence is a particular sequence of amino acid residues that folds in some specific way in the three-dimensional space. In that way, two amino acid residues could be close in the space, while far in the protein sequence. Therefore, the methods based on alignment of protein sequences are not able to recognize distant homology between proteins. For that purpose, also there is a group of methods that analyze the tertiary structures of proteins, like CE [8], MAMMOTH [9] and DALI [10]. Third group of methods, like SCOPmap [11] and FastSCOP [12], combines both sequence and structure alignment.

Besides alignment of the sequences or structures of proteins, feature vectors could be extracted, and then the proteins can be compared based on the distance between their feature vectors. In the literature, there are methods that use features of the sequence [13] or structure [14] of proteins, as well as both of them. By extracting the vector of features, the protein is presented by a point in the feature space. Later, in the classification stage, the amount of information that is processed is significantly lower than by making direct alignment of proteins, so the time needed for classification is much lower. For that purpose, in our earlier study [15], we presented feature vectors that contain features that represent the protein structures. These feature vectors could be used as inputs and by using some classification method, a prediction model could be generated.

In this paper, we use the protein voxel-based descriptor presented in [15] and we apply a fuzzy *k* nearest neighbors classification method [16] to classify the unknown structures.

Here is an outline of the structure of the remaining of this paper. Section 2 provides description of the proposed approach, where the protein voxel-based descriptor [15] and fuzzy *k* nearest neighbors classification method [16] are presented. The results of the evaluation of the approach are presented and discussed in Section 3, whereas Section 4 presents the main conclusions and points out directions for further advancements of the approach.

2. The Proposed Approach

The approach used in this study has two steps. The first step performs extraction of the feature vectors of the training protein structures. In this study, we use the protein-voxel based descriptor [15]. After mapping the proteins in the feature space, next, in the second step we use the fuzzy *k* nearest neighbors classification method [16] to determine the class in which a given query protein would belong to.

Protein Voxel-Based Descriptor

The protein voxel-based descriptor contains features of the primary, secondary and tertiary structure of the protein. Regarding the tertiary structure, we use the voxel descriptor [17] that is originally proposed for comparing 3D objects. Concerning the primary and secondary structure, the features are extracted as in [18].

The extraction of the protein voxel-based descriptor is done in the following way. First, a mesh model of the protein structure is generated, by making triangulation of the atoms of the protein that are treated as spheres and with triangulation they are presented by a given number of triangles. In order to obtain feature vector that is invariant to translation, the protein is translated so that its center of

mass is in the center of the coordinate system. With the aim to obtain feature vector that is invariant to scaling, next we perform scaling of the mesh model thus the most distant vertex of the triangles in the mesh model is at a distance equal to 1 from the center of mass.

After obtaining the mesh model, next, a voxelization is made. With this step, we transform the continual into discrete space. For that purpose, first, discretization is made, where the continuous three-dimensional space is divided into equal cubes named voxels. Then, for each of the voxels, we calculate the ratio of the area of the mesh that is in the inspected voxel. For that purpose, the triangles are divided into p_j^2 triangles with a surface $\delta = S_j / p_j^2$, where S_j denotes the area of the triangle T_j that is currently divided. If a given triangle T_j has vertices in one voxel, then p_j is set to 1. Otherwise, it is calculated as

$$p_j = \left| \sqrt{p_{\min} \frac{S_j}{S}} \right|, \quad (1)$$

where S is the total surface of the triangles and p_{min} defines the quality of the approximation. In this study, we use p_{min} = 32000 as in [17]. For each voxel, the surface that is placed in the voxel is incremented for δ.

With the previous step, as output, we obtain three-dimensional matrix that could be used as feature vector. However, the number of features contained in this matrix could be significantly reduced. Moreover, this three-dimensional matrix as a feature vector is not invariant to rotation. Therefore, in the next step, we apply 3D Discrete Fourier Transform thus obtaining feature vector that is invariant to rotation. In this way, the new version of the feature vector is also a three-dimensional matrix.

Next, the indices are shifted so the voxel in the center has indices (0, 0, 0). Because there is a symmetry between the elements of the obtained feature vector, therefore only the non-symmetrical features are considered that corresponds to the values of the voxels with indices that satisfy $1 \le |p| + |q| + |s| \le N/2$, where ($p$, q, s) are the indices of the voxel and N is the number of slices for one coordinate used in the discretization of the space. Further, the features are divided by the feature that correspond to the voxel with indices (0, 0, 0). More details about the extraction of the geometrical features contained in the voxel descriptor can be found in [17] and [15].

Besides the features of the tertiary structure of the proteins, we also consider several features of their primary and secondary structure. In this study, we use the features used in [18]. Regarding primary structure, we consider the ratio of each amino acid and the ratio of the hydrophobic amino acids in the protein. From the features of the secondary structure, we consider the ratios of the types of helices, as well as the number of occurrences of each type of secondary structure element (helix, sheet and turn). More details about these features can be found in [18] and [15].

Fuzzy k Nearest Neighbors Classifier

For classifying a given query sample (protein chain in this case), first its protein voxel-based descriptor is extracted. Then, this query sample is compared with the training samples and its class is determined by using the fuzzy k nearest neighbors (Fuzzy KNN) classification method [16].

The Fuzzy KNN method is inspired from the well known k nearest neighbors (KNN) classification method [19], but adjusted for sets in fuzzy logic. KNN could be used to perform simple majority voting of the nearest neighbors, or the neighbors may have different weights in the voting in order to give higher significance to the votes of the closer neighbors. For the second approach, the distance between the examined sample and the nearest neighbor could be used in order to calculate the weight of the vote of that neighbor.

Let assume we are using k nearest neighbors for making decisions. The Fuzzy KNN method first identifies the k nearest neighbors for the inspected sample q, which are denoted as *NN* (nearest neighbors). For that purpose, the similarity between a given training sample x and the query sample q is calculated as $S(x,q)=1/D(x,q)^2$, where $D(x,q)$ denotes the distance between x and q. Then, the examined sample q is classified by maximizing

$$\frac{\sum_{x \in NN} S(x,q) M_c(x)}{\sum_{x \in NN} S(x,q)}, \quad (2)$$

where c is the examined class, while $M_c(x)$ denotes the membership function for that class. The membership function could be crisp, defined as

$$M_c(x) = \begin{cases} 1, & x \in C \\ 0, & x \notin C \end{cases}, \quad (3)$$

where C is a set of the samples in class c. In this study, since we are using an approach based on fuzzy set theory, therefore instead of using a crisp function we are using the gradual function presented in [16]. The membership function that is used is defined as

$$M_c(x) = \begin{cases} 0.51 + 0.49 \frac{n_C}{k}, & x \in C \\ 0.49 \frac{n_C}{k}, & x \notin C \end{cases}, \quad (4)$$

where $n_C = |C|$ is the size of the set C.

3. Results and Discussion

For evaluation, we used 6145 protein chains that are classified in 150 different SCOP domains. The protein chains correspond to the samples in the set, while the domains are the output classes (the possible values for the target attribute). The information about the classification of the protein chains in SCOP domains is obtained from the SCOP database [3]. The distribution of the protein chains used in this study is approximately uniform. This set is divided into training set (90% of the chains) and test set (10% of the chains). In this way, the training set that is obtained has 5531 chains, while the remaining 614 chains form the test set. As evaluation measure, the classification accuracy is used, which gives evidence about the percent of the test samples that are classified correctly. The experimental results are presented on Fig. 1. We made experiments by using different values for the number of nearest neighbors that are considered for making predictions (k = 1, 2, 3, 4, 5, 10, 15, 20). In order to give better picture of the benefit of using fuzzy sets instead of classical sets, we made experiments by applying the classical KNN classification method and the Fuzzy KNN classification method that is based on fuzzy logic.

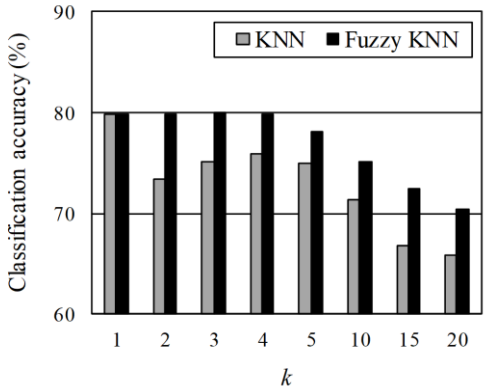

Fig. 1 *Classification accuracy achieved by KNN and Fuzzy KNN classification methods by using different number of nearest neighbors k.*

As it can be seen from the results, Fuzzy KNN provides better results than the classical KNN classification method. By using Fuzzy KNN, it is best to use between k=1 and k=4 nearest neighbors, and then by increasing the number of nearest neighbors that are considered for making decisions, the classification accuracy declines. Regarding KNN, the best result is obtained by using k=1 nearest neighbor. By using k=2 and k=3, lower accuracy is obtained than by using k=4. Then, by increasing the number of nearest neighbors (for k>4), the classification accuracy declines. The highest classification accuracy of 79.97% is achieved with Fuzzy KNN classifier by considering k=3 nearest neighbors.

The obtained results for the proposed approach are comparable to the results obtained with the existing approaches. Also, this approach provides classification of proteins with reasonable speed. Namely, the time needed for classification of all 614 test protein chains used in this study is in range of minutes, which is much better than the time needed with manual methods.

4. Conclusion

In this study, we proposed a novel approach that could be used to make decisions about the classification of protein chains into SCOP domains. For each training protein chain, its protein voxel-based descriptor is extracted, which is a feature vector that contains features about the primary, secondary and tertiary structure of the protein. The classification of an unknown test protein chain is done in two steps. First, its protein voxel-based descriptor is extracted, and then by applying the Fuzzy KNN classifier the class of the unknown protein chain is determined.

For evaluation, we used information about the classification of the protein chains in SCOP domains. The results show that it is best to use up to k=4 nearest neighbors, while by further growth of the number of nearest neighbors the results are getting worse. The Fuzzy KNN classifier was compared with the classical KNN method, and the results indicate that Fuzzy KNN is better.

As future work, we plan to extend this study in several directions. Regarding the feature vector, besides the protein voxel-based descriptor, we also plan to use some of the other feature vectors that we already used in our previous studies where these descriptors were used for retrieving similar protein structures. Concerning the classification method, we may also apply some other distance and similarity measures for estimating the similarity between two samples. Besides the Fuzzy KNN classification method, we plan to apply other classification methods, including methods based on classical set theory as well as fuzzy set theory.

Acknowledgment

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University in Skopje", Skopje, R. 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, 2000.

[2] RCSB Protein Data Bank, http://www.rcsb.org, 2018.

[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, 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, 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, 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, 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, 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, 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, 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, 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, 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, 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, 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, 2012.

[16] J. M. Keller, M. R. Gray, and J. R. Givens, "A fuzzy k-nearest neighbor algorithm," IEEE Transactions on Systems, Man, and Cybernetics, vol. 15, no. 4, pp. 580–585, 1985.

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

[18] P. Daras, D. Zarpalas, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Three-Dimensional Shape-Structure Comparison Method for Protein Classification," IEEE/ACM Trans. Comput. Biol. Bioinform, vol. 3, no. 3, pp. 193–207, 2006.

[19] D. Aha, D. Kibler, "Instance-based learning algorithms," Machine Learning, vol. 6, no. 1, pp. 37–66, 1991.