FUZZY PATTERN TREES FOR PROTEIN BINDING SITES PREDICTION USING WEIGHTED AVERAGING FUZZY AGGREGATION OPERATORS

Georgina Mirceva
Faculty of computer science and engineering
Ss. Cyril and Methodius University in Skopje
Skopje, R. Macedonia

Andrea Kulakov
Faculty of computer science and engineering
Ss. Cyril and Methodius University in Skopje
Skopje, R. Macedonia

ABSTRACT

The knowledge about the relationship between the protein functions and structure is very essential, since it could be used for drug design. With the high-throughput technologies the number of determined protein structures grows rapidly. However, many of these protein structures are not investigated in terms of determining their functions. Thus, the necessity for fast computational methods for annotating protein structures is evident. The functions of the protein structures could be determined by using different information. In our research we focus on annotating protein structures by detecting the protein binding sites. We have already introduced the fuzzy pattern tree induction for predicting the protein binding sites by considering the features of the amino acid residues. In this paper we introduce two additional fuzzy aggregation operators. We present some results of the evaluation of the method regarding the usage of different fuzzy aggregation operators. The results show that the prediction power of the models is increased with the inclusion of the additional fuzzy aggregation operators.

I. INTRODUCTION

Proteins are one of the main compounds in the organisms because they are involved in many processes in the living organisms. The knowledge about the relationship between their structure and functions is fundamental, and it could be used for drug design. With the technological innovations, the number of protein molecules with known structure increases fast. Also, there are experimental methods for annotating protein structures. Nevertheless, these methods are expensive and very complex. Therefore, many protein molecules are not investigated in terms of predicting their functions. Due to this, many research groups focus on developing fast and automated computation methods for annotating protein structures.

In the state of the art literature, many different methods are proposed and they take into account various information about the protein molecules. Some methods analyze the homology between protein molecules [1]. Other group of methods examines the conservation of their sequence and/or structure [2] by multiple-alignment. There are various methods that perform graph analyses on the protein-protein interaction networks [3]. Other methods [4] determine protein functions by detecting the protein binding sites, similarly as molecular biologists manually annotates protein structures. In our research we predict the possible protein binding sites.

In various methods found in the literature, different amino acid residues' features are considered for protein binding sites prediction. Accessible Surface Area (ASA) [5], depth index (DPX) [6], protrusion index (CX) [7], and hydrophobicity [8] are among the most widely used features. Since none of the amino acid residues' features does not provide enough information about the possibility that the amino acid residue would be binding site or not, therefore in the research usually a set of features is used.

Variety of methods for protein binding sites can be found in the literature [9, 10, 11, 12]. During evolution, small changes of the amino acid residues' features occur, and they should not influence the predictions. The classical classification methods are sensitive to small changes in the data, so these changes significantly influence the decisions. To overcome this problem, we already introduced the fuzzy theory for protein binding sites prediction. There are several studies about fuzzy decision trees (FDTs) [13, 14, 15, 16]. However, in FDT induction only one type of fuzzy aggregation operator is used. Later, the fuzzy pattern trees (FPTs) are introduced by Huang et al. [17] where a set of fuzzy aggregation operators could be used in the model induction.

In our previous research we already introduced the FDTs [18] and FPTs [19] for protein binding sites prediction. Besides the most basic fuzzy aggregation operators (FAOs), i.e. AND and OR, also other FAOs could be used. In this paper we introduce Weighted Average (WA) and Order Weighted Average (OWA) fuzzy aggregation operators in protein binding sites prediction. We perform comparison regarding the set of fuzzy aggregation operators that are considered.

In section 2, we present the fuzzy pattern tree-based method for predicting protein binding sites. In section 3 we provide some experimental results. Finally, in section 4 we conclude the paper and identify potential for additional improvements.

II. FUZZY PATTERN TREE-BASED METHOD FOR PROTEIN BINDING SITES PREDICTION

In order to predict the protein binding sites, first we extract several features of the amino acid residues'. Then, fuzzy pattern trees are induced as models for making decisions whether a given test amino acid residue would be binding site or not.

A. Extraction of the Amino Acid Residues' Features

In this research, we consider the most commonly used amino acid residues' features: Accessible Surface Area (ASA) [5], depth index (DPX) [6], protrusion index (CX) [7], and hydrophobicity [8].

The accessible surface area (ASA) is very important feature of the amino acid residues, since it gives information about the surrounding of the amino acid residue, which dictates the possibility that the residue would be involved in interaction.

ASA is usually calculated by using the rolling ball algorithm [5], where a probe sphere with a given radius is rolled around the protein molecule. We use a sphere with radius of 1.4 Å, which is the most common value. With the rolling probe algorithm, the ASA value for each atom is calculated. Since atoms are compounds of the amino acid residues, the ASA of a given amino acid residue is calculated by summing the ASA values of its atoms. Since deeply buried amino acid residues could not be involved in protein-protein interaction, we filter only the amino acid residues that are located on the protein surface. As surface residues we consider those residues for which at least 5% of their surface is accessible by the probe sphere [20].

The depth index (DPX) [6] of each atom is calculated as Euclidean distance between the atom and the nearest atom that is reached by the probe sphere. The DPX of an amino acid residue is calculated as an average value of the DPXs of its atoms.

For calculating the protrusion index (CX) [7] of an atom, first for each non-hydrogen atom we count the number of heavy atoms in its neighbouring. In this research we analyze the neighbouring atoms within 10 Å around the atom, as suggested in [7]. Then, the volume occupied by heavy atoms V_{int} is calculated by multiplying the number of heavy atoms and the average volume of an atom (we use a value of 20.1 Å according to [7]), while the rest of the volume of the sphere is considered as remaining volume V_{ext}. Finally, the protrusion index CX is calculated as ratio of the remaining and the occupied volume. The CX of an amino acid residue is calculated as an average value of the CXs of its atoms.

Hydrophobicity is related with the hydrophobic effect of the amino acids, based on which hydrophobic amino acids are mostly buried in the protein interior, while hydrophilic amino acids are more frequently found at/near the protein surface. Several hydrophobicity scales can be found in the literature. We use the most common scale proposed in [8].

B. Bottom-up Induction of Fuzzy Pattern Trees

Next, we induce models for protein binding sites prediction by using fuzzy pattern tree induction. In this research we use the bottom-up fuzzy pattern tree induction method proposed by Huang et al. [17]. For each class (binding or non-binding sites' class) we build a separate model. Test amino acid residues are classified in the class for which highest similarity is obtained.

First, we perform fuzzification of the dataset by using some fuzzy membership function (FMF). We use two straight-line fuzzy membership functions (triangular and trapezoidal), and one convex fuzzy membership function (Gaussian FMF). In fuzzification, the data set is labelled with fuzzy terms. Since we take into account four features (ASA, DPX, CX and hydrophobicity), we will induce 20 primitive trees if the number of FMFs per feature is set to 5 (N=5).

The obtained primitive trees are very simple and do not provide acceptable accuracy. Therefore, these primitive trees would be combined in order to provide model with acceptable accuracy. For this reason, we have to evaluate the primate trees, and to identify the primitive tree with highest similarity. In this research we use the Root mean squared error (RMSE) similarity measure in order to estimate the similarity between the membership values of a given fuzzy term for a given feature and the membership values of other fuzzy term for the class attribute.

The primitive trees (trees at level 0) could not achieve acceptable accuracy. Therefore, the primitive tree with highest similarity is combined (aggregated) with the other primitive trees by using some fuzzy aggregation operators, thus forming the trees at level 1. In the induction of given fuzzy pattern tree different fuzzy aggregation operators could be used in same tree based on the obtained similarity, while in fuzzy decision tree only a single fuzzy aggregation operator could be used. There are various fuzzy aggregation operators. In our previous research [19] we used the AND and OR fuzzy aggregation operators. In this paper we introduce the Weighted Average (WA) and Ordered Weighted Average (OWA) fuzzy aggregation operators for protein binding sites prediction.

There are three categories of fuzzy aggregations operators, i.e. t-norm, t-conorms, and averaging operators. In this paper we introduce Weighted Averaging (WA) and Ordered Weighted Averaging (OWA) [21] fuzzy aggregation operators. With the WA operator of dimension n we perform mapping $R^n \rightarrow R$ by

$$WA(a_1, a_2, ..., a_n) = \sum_{j=1}^{n} \omega_j a_j, \qquad \sum_{i=1}^{n} \omega_i = 1$$

where $w = (w_1, w_2, \ldots, w_n)^T$, $w_i \in [0, 1]$, $1 \leq i \leq n$, is an associated n-dimensional vector with weights. By the OWA operator the mapping is done by

$$OWA(a_1, a_2, ..., a_n) = \sum_{j=1}^{n} \omega_j (f_j(a_1, a_2, ..., a_n)), \qquad \sum_{i=1}^{n} \omega_i = 1$$

where $f_j(a_1, a_2, \ldots, a_n)$ returns the j-th largest element of the set $\{a_1, a_2, \ldots, a_n\}$. The main difference between these fuzzy aggregation operators is that OWA does not have specific weights associated for the elements; rather the weights are associated with the ordered position of the elements.

After obtaining the trees at level 1, we identify the tree with highest similarity by using the RMSE similarity measure. This tree is further combined with the trees that do not have highest similarity at that level. This procedure is repeated until some predefined stop criterion is satisfied. Different criteria could be used for termination. For example, the induction of the tree could stop when the trees obtained at given level obtains lower similarity than the tree with highest similarity at the previous level. Other stop criterion could be to terminate the fuzzy pattern tree induction when the tree obtains some predefined depth. In our research we use the second criterion and induce trees with depth 5. We have made analysis by setting deeper depth (depth 10), but the induced models didn't reached the maximum allowed depth.

We can obtain two types of models, i.e. simple models and general models. In the induction of simple models the tree with highest similarity at a given level could be aggregated only with primitive trees (trees at level 0), while in the induction of general models the tree with highest similarity could be aggregated with the trees from all levels. In this research we induce simple models.

III. EVALUATION

Next, we will evaluate the method. As a standard of truth we use a part of the BIND database [22]. The BIND database contains information about the protein binding sites. We have limited computational power, and we are not able to consider the entire BIND database in the analysis. Therefore, from the BIND database we consider only the information about the most representative protein chains. As representative protein chains we consider the chains that do not have more than 40% similarities in their sequences by using the criterion given in [23]. After building the representative data set, next we divide the data into training and test set. From the representative data set, we filter the protein chains that do not have more than 20% sequence similarities between themselves by using the same criterion. In this way we form the test data set, while the remaining protein chains are taken in the training data set. The training data set contains 1062 protein chains with 365862 amino acid residues, while the test data set contains 1858 protein chains with 608434 amino acid residues. Next, we filter the amino acid residues located at the protein surface by considering only the residues for which at least 5% of their surface could be reached by the probe sphere. After this filtering, we have 284168 amino acids in the training data set, and 484637 amino acids in the test data set. From the information stored in the BIND database we can conclude that the data set is not balanced. Namely, from the amino acids in the data set only 10% are binding sites and even 90% are not binding sites. In order to prevent building models that are biased towards the dominant class (the non-binding sites class in this case), we balance the training data set. This balancing is done only on the training data set, while the test data set remains unbalanced. Therefore, in the evaluation we must use some measure that is appropriate for unbalanced data sets. In this research we use the Area under ROC curve (AUC-ROC) measure to evaluate the prediction models. AUC-ROC obtains values in the interval [0,1]. The higher the value is, the more accurate the prediction model is.

We examined the AUC-ROC when different type of fuzzy membership function is used. In this research we considered the triangular, trapezoidal and Gaussian fuzzy membership functions. Also we made experiments by using different number of fuzzy membership functions per amino acid residues' feature (N=3, 4 and 5). In this research we use the RMSE similarity measure. In the aggregation we aggregate the tree with highest similarity among the trees at the last level with the primitive trees, thus obtaining simple models. As a stop criterion we use the maximal allowed depth criterion, and we set the maximal depth to 5.

In Table 1 we present the AUC-ROC values by using only the AND and OR fuzzy aggregation operators, while Table 2 presents the results when AND, OR, WA and OWA fuzzy aggregation operators are used. In our previous research [19] we already examined the influence of the type of FMF and the number of fuzzy membership functions used per feature by using the AND and OR fuzzy aggregation operators. In this paper we perform similar analysis where additionally WA and OWA fuzzy aggregation operators are used together with the AND and OR operators.

Table 1: The results for AUC-ROC obtained by using AND and OR fuzzy aggregation operators.

FMF	N=3	N=4	N=5
Triangular	0,5564	**0,5644**	0,5459
Trapezoidal	0,5386	0,5568	**0,5649**
Gaussian	0,5564	**0,5644**	0,5386

Table 2: The results for AUC-ROC obtained by using AND, OR, WA and OWA fuzzy aggregation operators.

FMF	N=3	N=4	N=5
Triangular	0,5670	**0,5744**	0,5727
Trapezoidal	0,5431	0,5580	**0,5662**
Gaussian	0,5667	0,5708	**0,5722**

In this case, when AND, OR, WA and OWA operators are used, the AUC-ROC values are increased in all experiments. The highest AUC-ROC of 0,5744 is obtained by using 4 triangular FMFs. The increase of AUC-ROC is highest (increase of 0,0336) for 5 Gaussian FMFs, and then the next highest increase (increase of 0.0268) is for 5 triangular FMFs.

IV. CONCLUSION

Our research presented in this paper aims to provide efficient method for protein binding sites prediction that could be used for protein annotation. First, we extracted several amino acid residues' features, and then we induced models for making decision whether a given amino acid residue is binding site or not by using a bottom-up fuzzy pattern tree induction. In our previous research papers we already introduced this method and we made experimental analysis by using only the AND and OR fuzzy aggregation operators. In this paper additionally we introduced the WA and OWA fuzzy aggregation operators in order to increase the prediction power of the method.

By considering the WA and OWA fuzzy aggregation operators together with the AND and OR operators, the AUC-ROC values are increased in all experiments. The highest AUC-ROC is obtained by using 4 triangular FMFs.

As future work, we identified several directions for additional improvements. In this research we used only triangular, trapezoidal and Gaussian FMFs. Besides them, we can also induce models by using the bell, log-normal, sigmoid (+1), sigmoid (-1) and other FMFs. Regarding similarity measure, besides the most simple measure used in our research, i.e. RMSE measure, we can use other similarity measures like Jaccard, Cosine etc. In this paper we introduced the WA and OWA fuzzy aggregation operators for protein binding sites prediction and we showed that by using them the prediction models are enhanced. Besides these fuzzy aggregation operators, also other types of operators could be considered, like Yager, Sklar etc.

ACKNOWLEDGEMENTS

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] A. E. Todd, C. A. Orengo and J. M. Thornton, "Evolution of function in protein superfamilies, from a structural perspective," *J. Mol. Biol.*, vol. 307, no. 4, pp. 1113–1143, 2001.
[2] A. R. Panchenko, F. Kondrashov and S. Bryant, "Prediction of functional sites by analysis of sequence and structure conservation," *Protein Science*, vol. 13, no. 4, pp. 884–892, 2004.
[3] M. Kirac, G. Ozsoyoglul and J. Yang, "Annotating proteins by mining protein interaction networks," *Bioinformatics*, vol. 22, no. 14, pp. e260–e270, 2006.
[4] N. Tuncbag, G. Kar, O. Keskin, A. Gursoy and R. Nussinov, "A survey of available tools and web servers for analysis of protein-protein interactions and interfaces," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 217–232, 2009.
[5] A. Shrake and J. A. Rupley, "Environment and exposure to solvent of protein atoms," *Lysozyme and insulin, J. Mol. Biol.*, vol. 79, no. 2, pp. 351–371, 1973.
[6] A. Pintar, O. Carugo and S. Pongor, "DPX: for the analysis of the protein core," *Bioinformatics*, vol. 19, no. 2, pp. 313–314, 2003.
[7] A. Pintar, O. Carugo and S. Pongor, "CX, an algorithm that identifies protruding atoms in proteins," *Bioinformatics*, vol. 18, no. 7, pp. 980–984, 2002.
[8] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–132, 1982.
[9] A. S. Aytuna, A. Gursoy and O. Keskin, "Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces," *Bioinformatics*, vol. 21, no. 12, pp. 2850–2855, 2005.
[10] H. Neuvirth, R. Raz and G. Schreiber, "ProMate: a structure based prediction program to identify the location of protein-protein binding sites," *J. Mol. Biol.*, vol. 338, no. 1, pp. 181–199, 2004.
[11] J. R. Bradford and D. R. Westhead, "Improved prediction of protein-protein binding sites using a support vector machines approach," *Bioinformatics*, vol. 21, no. 8, pp. 1487–1494, 2005.
[12] S. Jones and J. M. Thornton, "Prediction of protein-protein interaction sites using patch analysis," *J. Mol. Biol.*, vol. 272, no. 1, pp. 133–143, 1997.
[13] C. Z. Janikow, "Fuzzy decision trees: issues and methods," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 28, no. 1, pp. 1–14, 1998.
[14] C. Olaru and L. Wehenkel, "A complete fuzzy decision tree technique," *Fuzzy Sets and Systems*, vol. 138, no. 2, pp. 221–254, 2003.
[15] A. Suárez and J. F. Lutsko, "Globally optimal fuzzy decision trees for classification and regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1297–1311, 1999.
[16] X. Wang, B. Chen, G. Olan and F. Ye, "On the optimization of fuzzy decision trees," *Fuzzy Sets and Systems*, vol. 112, no. 1, pp. 117–125, 2000.
[17] Z. H. Huang, T. D. Gedeon and M. Nikravesh, "Pattern trees induction: a new machine learning method," *IEEE Transaction on Fuzzy Systems*, vol. 16, no. 3, pp. 958–970, 2008.
[18] G. Mirceva, A. Naumoski, V. Stojkovik, D. Temelkovski and D. Davcev, "Method for Protein Active Sites Detection Based on Fuzzy Decision Trees," Database Theory and Application / Bio-Science and Bio-Technology, DTA/BSBT 2011, CCIS 258, pp. 143-150, 2011, Springer-Verlag Berlin Heidelberg 2011.
[19] G. Mirceva and A. Kulakov, "Fuzzy pattern trees for predicting the protein binding sites," 9th Conference for Informatics and Information Technology (CIIT 2012), Bitola, Macedonia, 2012.
[20] C. Chothia, "The Nature of the Accessible and Buried Surfaces in Proteins," J. Mol. Biol., vol. 105, no. 1, pp. 1–12, 1976.
[21] R. R. Yager, "On ordered weighted averaging aggregation operators in multicritera decison making," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.
[22] G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson and C. W. Hogue, "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 242–245, 2001.
[23] J. –M. Chandonia, G. Hon, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt and S. E. Brenner, "The ASTRAL Compendium in 2004," *Nucleic Acids Res.*, vol. 32, pp. D189–D192, 2004.