ProSPs: Protein Sites Prediction Based on Sequence Fragments

Michela Quadrini[2(✉)], Massimo Cavallin[1], Sebastian Daberdaku[3], and Carlo Ferrari[1]

[1] Department of Information Engineering, University of Padova, Via Gradenigo 6/A, 35131 Padova, PD, Italy
massimo.cavallin.1@studenti.unipd.it, carlo.ferrari@unipd.it

[2] School of Science and Technology, University of Camerino, Via Madonna delle Carceri, 9, Camerino, MC, Italy
michela.quadrini@unicam.it

[3] Sorint.Tek, Sorint.LAB group, Via Giovanni Savelli 102, 35129 Padova, PD, Italy
sebastian.daberdaku@latek.it

Abstract. Identifying interacting sites of proteins is a relevant aspect for drug and vaccine design, and it provides clues for understanding the protein function. Although such a prediction is a problem extensively addressed in the literature, just a few approaches consider the protein sequence only. The use of the protein sequences is an important issue because the three-dimensional structure of proteins could be unknown. Moreover, such a structural determination experimentally is expensive and time-consuming, and it may contain errors due to experimentation. On the other hand, sequence based method suffers when the knowledge of sequence is incomplete.

In this work, we present ProSPs, a method for predicting the protein residues considering protein sequence fragments, which are obtained using sliding windows and become the samples for an unbalance binary classification problem. We use the Random Forest classifier for data training. Each amino acid is enriched using a selected subset of physicochemical and biochemical amino acid characteristics from the AAIndex1 database. We test the framework on two classes of proteins, Antibody-Antigen and Antigen-Bound Antibody, extracted from the Protein-Protein Docking Benchmark 5.0. The obtained results evaluated in terms of the area under the ROC curve (AU-ROC) on these classes outperform the sequence-based algorithms in the literature and are comparable with the ones based on three-dimensional structure.

Keywords: Random forest · Sequence method · Site interaction prediction

G. Nicosia et al. (Eds.): LOD 2021, LNCS 13163, pp. 568–580, 2022.
https://doi.org/10.1007/978-3-030-95467-3_41

1 Introduction

Proteins are versatile macromolecules consisting of one or more amino acid sequences that carry out a broad range of functions in living organisms. These biological roles are fulfilled by interacting with other molecules, including RNA, other proteins, and small ligands [3]. The interactions between two proteins, known as protein-protein interactions (PPIs), are responsible for the metabolic and signaling pathways [13]. Dysfunction or malfunction of pathways and alterations in protein interactions can cause several diseases, including neurodegenerative disorders and cancer. Therefore, the protein interaction knowledge allows us to understand how proteins perform their functional roles and design new antibacterial drugs [8]. The experimental determination of three-dimensional structures of protein complexes is labor-intensive, time-consuming, and has high costs. Therefore, efficient computational methods to predict PPIs play a fundamental role. The computational approaches can be broadly divided, according to the protein representation, into sequence-based and structure-based. The former employs information derived from the amino acid sequence alone to predict the site, while the latter considers the protein three-dimensional (3D) structure. About the sequence-based methods, the representative ones include PPiPP [1], PSIVER [14], DLPred [28], and NPS-HomPP [26]. PPiPP uses the position-specific scoring matrix (PSSM) and amino acid composition, and PSIVER takes advantage of PSSM and predicted accessibility as input for a Naive Bayes classier. DLPred uses long-short term memory to learn features such as PSSM, physical properties, and hydropathy index. To improve prediction, NPS-HomPPI infers interfacial residues from the interfacial residues of homologous interacting proteins. In the literature, structure-based methods usually perform better than sequence-based ones. About the structural-based methods, several approaches have been proposed in the literature. Some of them take advantage of the molecular surface representations for describing the structure and use Zernike descriptors or geometric invariant fingerprint (GIF) descriptors to identify possible binding sites [5,7,27]. Other methods use graph representations of proteins, such as contact maps or hierarchical representations [19]. Most of these aforementioned methods employ machine learning algorithms, including support vector machines, neural networks, Bayesian networks, naive Bayes classier, and random forests. Although structure-based approaches are generally more accurate than sequence-based ones, their applicability is limited since the structure is known or contains some errors due to experimentation. As a consequence, an improvement of sequence-based methods is necessary.

In this work, we introduce ProSPs, a method for predicting the protein interaction sites taking into account protein sequence fragments. Considering sequence fragments is relevant when the entire sequence of proteins is unknown. Such fragments, obtained using sliding windows approach over the whole sequence or the known part of it, become the samples for an unbalance binary classification problem. To determine whether a single residue is part of the complex or not, we used a Machine Learning approach using a Random Forest as a classifier [23]. Although methods based on Random Forests achieve good

results with unbalanced data [23], we also employed Random Sampling and a classifier combination approach, which further improved predictions made from unbalanced data. Such predictions are performed on the central residues of sliding windows, which are extracted from the entire protein sequence. Their length is computed using a normalized version of the metric introduced by Sikic *et al.* [23]. In other words, we select the length of the sliding windows considering the minimum difference of normalized entropy. To better represent the data, each amino acid is equipped with eight high-quality amino acid indices of physicochemical and biochemical properties extracted from the AAindex1 database [12]. We tested the framework on two classes of proteins, Antibody-Antigen and Antigen-Bound Antibody, extracted from the Protein-Protein Docking Benchmark 5.0 [25] supposing that their 3D structures are unknown. We selected these classes since antibodies recognize and bind several antigens. Such characteristic makes them the most valuable category of biopharmaceuticals for both diagnostic and therapeutic applications. To evaluate the model performance on the two data sets, we consider only the area under the receiver operating characteristic curve (AU-ROC) since the recognition of PPIs interface sites is an imbalanced classification problem. This aspect can lead to classifiers that tend to label all the samples as belonging to the majority class, thus trivially obtaining a high accuracy measure. The obtained results in terms of AU-ROC on the data set outperform the sequence-based algorithms in the literature. Moreover, they are comparable with the ones based on three-dimensional structures.

The paper is organized as follows. In Sect. 2, we describe the dataset entries, which consist of sliding windows of a predetermined number of residues In Sect. 3, we describe the used dataset and the results obtained with the model, described in Sect. 2.3. The paper ends with some conclusions and future perspective, Sect. 4.

2 Materials and Methods

2.1 Dataset Entry

Each dataset sample consists of sliding windows of a predetermined number of residues extracted from the entire length of the antibody chains sequence. The sliding window length can influence the classification of results.

Window Length Selection. To determine the sliding window length, we proposed a method based on entropy, similar to one proposed by Sikic *et al.* [23] and applied in Sriwastava *et al.* [24]. Our approach takes advantage of the normalized entropy difference between the occurrence of a particular number of interacting residues within a window length of N residues and the uniform occurrence distribution. To carry out this calculation, we define the interacting residues for all proteins in the datasets, and we compute the number of interacting residues using different length sliding windows. Finally, we consider only the windows

having a central interacting residue. The normalized window entropy formula is

$$\frac{-\sum_{i=1}^{N} p_i \cdot \log_2 p_i - \log_2 N}{N} \tag{1}$$

where N is the length of a window, p_i is the frequency appearance of i interacting residues in a window of N residues, and $\log_2 N$ is the window entropy when the interacting residues number is distributed uniformly. In other words, the value $\log_2 N$ represents the maximum possible entropy of the window.

2.2 Features

In this work, we consider some physico-chemical and biochemical properties of amino acids that are published in the AAindex [12]. AAindex is a database containing numerical indices that represent various physico-chemical and biochemical properties of residues and residue pairs published in the literature. Each amino acid index is a set of 20 numerical values representing any of the different physico-chemical and biological properties of each amino acid: the AAindex1 section of the database is a collection of 566 such indices. Using a consensus fuzzy clustering method on all available indices in the AAindex1, Saha *et al.* [22] identified three high-quality subsets (HQIs) of all available indices, namely HQI8, HQI24, and HQI40. In this work, we use the features of the HQI8 amino acid index set, reported in Table 1, that are identified as the medoids (centers) of 8 clusters obtained by using the correlation coefficient between indices as a distance measure.

Table 1. HQI8 indices.

Entry name	Description
BLAM930101	Alpha helix propensity of position 44 in T4 lysozyme
BIOV880101	Information value for accessibility; average fraction 35%
MAXF760101	Normalized frequency of alpha-helix
TSAJ990101	Volumes including the crystallographic waters using the ProtOr
NAKH920108	AA composition of MEM of multi-spanning proteins
CEDJ970104	Composition of amino acids in intracellular proteins (percent)
LIFS790101	Conformational preference for all beta-strands
MIYS990104	Optimized relative partition energies - method C

2.3 Dataset Entries Definition

Each entries of the dataset is defined on a sliding window of N residues. It is a vector that consists of $N \cdot k + 2$ elements, where k is the number of selected features. The input vector shows the following scheme:

- *Features* represent the first $N \cdot k$ elements of the input vector that assigns the k selected features for each amino acids of the window.
- *interface* indicate whether the window is interfacing with a protein or not, 1 or -1 respectively. A window is defined as interfacing if its central residue interacts with another one of the other proteins.
- *group* identifies the protein that contains the sequence.

Since each residue for each chain is the possible center of an interface window, the number of windows belonging to the "interface" and "non-interface" classes in the dataset reflects exactly the number of residues previously indicated as interacting and non-interacting. There is a one-to-one correspondence between each amino acid in the antibody sequence and the windows.

2.4 Dataset Imbalance Reducing

The prediction of PPI sites can be considered a classification problem, whose objective is to assign a label, either 1 (interface) or 0 (no-interface), to each residue. This problem is extremely imbalanced. This imbalance makes it difficult for a classifier to learn significant patterns with particular reference to the samples belonging to the minority class. Therefore, a random subsampling method is used to reduce the dataset imbalance. The "Non- Interface" samples of the training set were randomly undersampled to reduce the class imbalance ratio.

2.5 Random Forest

Random Forest is an ensemble model for classification and regression. The model operates by constructing a multitude of decision trees at training time and outputting the class that is the mode (for classification problem) or mean/average prediction (for regression) of the classes output by individual trees. Developed by Breiman [4], the model combines the bagging approach with the random selection of features, introduced independently by Ho [9,10] and Amit and Geman [2], to ensure that the decision trees of the forest are uncorrelated from each other. In bagging, the decision trees depend on trees, which are created from a different bootstrap sample of the training dataset. A bootstrap sample is a sample of the training dataset with replacement, i.e., each sample may appear more than once in the sample. In details, let S be the training set containing m samples, the bagging procedure will initially realize B replicated datasets extracting by uniform sampling with replacement of m' samples from the entire dataset S. In each dataset S^i with $i \in \{1, 2, \ldots, B\}$ will therefore be possible to find samples of S repeated several times, while some may not be selected at all. The replicated datasets permit to train of decision trees, which will then make up the Random Forest so that each tree will only see different portions of the original dataset during training. This bagging approach is combined with the random selection of features, that uses only different random subsets of the entire feature space to train each tree in the random forest. This means that some features used to

train a single tree may not be used to train other trees belonging to the forest. Typically, for a classification problem with p features, $\sqrt{p}$ features are used in each split.

The Random Forest classifier has several hyper-parameters that can be tuned:

- number of decision trees inside the forest
- maximum depth of each decision tree
- minimal impurity of a node for it to be converted into a leaf
- maximum number of attributes used per tree training
- minimal impurity decrease of resulting subdatasets for a node to be created

3 Experiments

3.1 Dataset

The Protein–Protein Docking Benchmark 5.0 (DB5) [25], the standard benchmark dataset for assessing docking and interface prediction methods, is the dataset in this work. The benchmark consists of 230 non-redundant, high-quality structures of protein-protein complexes, selected from the Protein Data Bank (PDB). PDB organizes the complexes according to the functional eight different classes: Antibody-Antigen (A), Antigen- Bound Antibody (AB), Enzyme-Inhibitor (EI), Enzyme-Substrate (ES), Enzyme complex with a regulatory or accessory chain, Others, G-protein containing (OG), Others, Receptor containing (OR), and Others, miscellaneous (OX). This study considers only complexes of classes A and AB. For each class, we separated the receptor proteins from the ligand ones. To easily compare our approach with other methods in the literature, we split the data into training and test sets, as shown in Table 2 and proposed in [6]. The residues of a given protein is labeled as part of the PPI interface if they had at least one heavy (non-hydrogen) atom within 5Å from any heavy atom of the other protein (the same threshold used in [6]).

Table 2. The table gives the PDB code and chain ID of each protein used in this study (the PDB code in parentheses identifies the corresponding bound complex in the DB5 database).

Dataset	Training set	Test set
$\mathbf{A}_r$	1AY1.HL (1BGX), 1BVL.BA (1BVK), 2FAT.HL (2FD6), 2I24.N (2I25), 3EO0.AB (3EO1), 3G6A.LH (3G6D), 3HMW.LH (3HMX), 3L7E.LH (3L5W), 3MXV.LH (3MXW), 3V6F.AB (3V6Z), 4GXV.HL (4GXU)	1FGN.LH (1AHW), 1DQQ.CD (1DQJ), 1QBL.HL (1WEJ), 1GIG.LH (2VIS), 2VXU.HL (2VXT), 3RVT.CD (3RVW), 4G5Z.HL (4G6J)
$\mathbf{A}_l$	1TAQ.A (1BGX), 3LZT (1BVK), 1A43 (1E6J), 1YWH.A (2FD6), 1IK0.A (3G6D), 1F45.AB (3HMX), 3M1N.A (3MXW), 3F5V.A (3RVW), 3KXS.F (3V6Z), 1DOL.A (4DN4), 4I1B.A (4G6J), 1RUZ.HIJKLM (4GXU)	1TFH.A (1AHW), 1HRC (1WEJ), 2VIU.ACE (2VIS), 1J0S.A (2VXT), 1QM1.A (2W9E), 1TGJ.AB (3EO1), 3F74.A (3EOA), 2FK0.ABCDEF (4FQI)
$\mathbf{AB}_r$	1BJ1.HL (1BJ1), 1FSK.BC (1FSK), 1I9R.HL (1I9R), 1K4C.AB (1K4C), 1KXQ.H (1KXQ), 2JEL.HL (2JEL), 1QFW.HL (9QFW)	1IQD.AB (1IQD), 1NCA.HL (1NCA), 1NSN.HL (1NSN), 1QFW.IM (1QFW), 2HMI.CD (2HMI)
$\mathbf{AB}_l$	2VPF.GH (1BJ1), 1BV1 (1FSK), 1D7P.M (1IQD), 7NN9 (1NCA), 1HRP.AB (1QFW), 1S6P.AB (2HMI), 1POH (2JEL)	1ALY.ABC (1I9R), 1JVM.ABCD (1K4C), 1PPI (1KXQ), 1KDC (1NSN)

3.2 Implementation and Results

In our work, we use a Python implementation of the Random Forest [4] classifier provided by 0.22.2 version of scikit-learn package [15]. The scheme of our approach is shown in Fig. 1, while the code used in this manuscript are available from the corresponding author upon reasonable request.

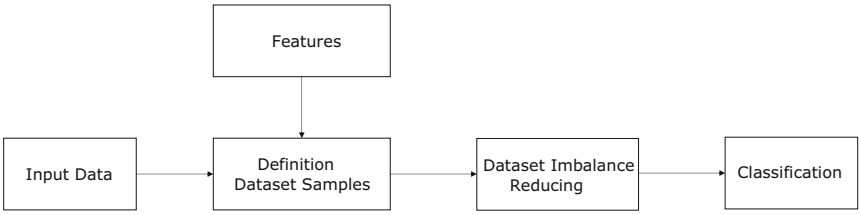
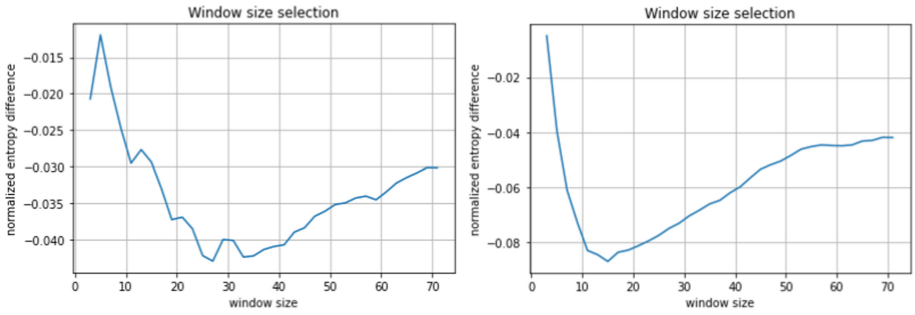


Fig. 1. The scheme of our approach

The first step in creating the dataset samples is to identify the interaction residues. Considering the PDB files of the protein-protein complexes taken from DB5 as an input, the tool extracts the sequence and the interaction residues taking advantage of Biopython package. The second step of the dataset samples creation is the identification of the number of residues to build the sliding windows. To carry out this calculation, only the windows with the central residue classified as interacting (label equal to 1) were considered. In the case of windows with the interacting central residues less than $\lfloor \frac{N}{2} \rfloor$ residues away from the edge of the chain, to keep the size of the windows fixed, a padding of 0 was used to cover the positions that would have been cut. To find the best value of N that minimizes the Eq. 1, a range of possible values from 3 to 71 with a step of 2 were tested. The results of this analysis of classes A and AB are in Figs. 2 and 3. Therefore, the dataset entries are windows composed of 28 residues for A_l class, 14 residues for A_r class, 28 residues for AB_l class, and 22 residues for AB_r class. In the entries definition, we need to introduce some padded residues when we consider the first and $N-1$ last amino acids. The padded residues are conceived as fictitious elements equipped with unnatural features. In particular, these values were obtained by increasing the maximum possible value of considered index of HIQ8 set increased by 1.

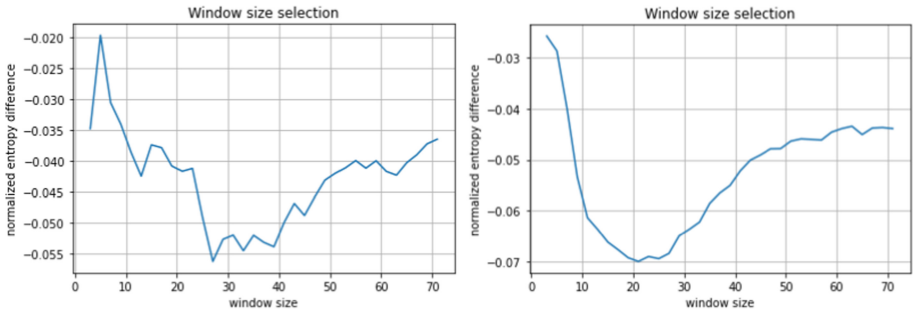
The tool reduces the dataset imbalance by a random subsampling method. In particular, it uses the RandomUnderSampler algorithm of Imbalanced-learn, an open source library relying on scikit-learn, obtaining a ratio of 60-40 of the number of samples in the "Interface" (minority) class over the number of samples in the "No-Interface" (majority) class after resampling. Finally, we tune the hyper-parameters of the Random Forest Algorithm, which is implemented with RandomForestClassifier from the sklearn-ensemble package of version 0.22.2 of Scikit-Learn.

Among the available hyper-parameters, we tune



Ligands Receptos

Fig. 2. Normalized Entropy difference for different windows sizes for ligands (left) and receptors (right) of class A



Ligands Receptos

Fig. 3. Normalized Entropy difference for different windows sizes for ligands (left) and receptors (right) of class AB

- the number of decision trees in the forest, n estimators (tested value: 30)
- the criterion used to evaluate the imputiry of a split, criterion (tested Value: "gini", "entropy")
- the maximum depth of each tree, max features (tested values: 2^i, for $i \in \{2, \ldots, 6\}$)
- the maximum number of features that can be used to create a splitting rule (tested values: "sqrt", "log2")
  For example for a total of n features when "sqrt" is used for this parameter only $\sqrt{n}$ features are considered when performing each split
- the minimum number of samples in a leaf to consider further splitting (tested Values: 2^i, for $i \in \{1, \ldots, 6\}$)
- the minimum number of samples required for a leaf (tested Values: 2^i, for $i \in \{1, \ldots, 6\}$)
- bootstrap that indicates whether to use bootstrap when creating each tree during the fitting phase (tested values: True, False)

To search for the optimal combination of hyper-parameter values, the RandomizedSearchCV function of the sklearn.model selection package from Scikit-Learn was used. This method allows to tune hyper-parameters using randomly selected combinations of values from the set provided as input while also fitting the model using cross-validation. As a cross-validation technique, GroupKFold function of the sklearn.model selection package from Scikit-Learn is used. The scoring used to train Random Forest is F1 Weighted, derived from the following formula of the F1 score:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

F_1 weighted computes the F_1 score for each class separately, attributing a different weight based on the support, which is the number of true instances of each label. This metric is chosen as it takes into account the dataset's imbalance, giving more importance to the minority class. Furthermore, this score is a more expressive indicator of the classifier's real prediction capabilities compared to other metrics commonly used such as AU ROC, as it highlights the model's ability to effectively predict samples belonging to the "interface" class. Table 3 illustrates the values selected for the hyper-parameters of the final model of each class (A_r, A_l, AB_r, AB_l) following Randomized Search using Group K Fold Cross-Validation with K = 10 folds and F1 weighted as scoring function.

Table 3. Hyperparameters values chosen from Randomized Search for ligands and receptors of classes A and AB

Hyperparameters	A_r	A_l	AB_r	AB_l
n estimators	115	52	94	73
Criterion	gini	entropy	gini	entropy
Max features	log2	$log2$	log2	log2
Max depth	16	16	8	4
Min samples split	4	32	8	4
Min samples leaf	2	8	2	4
Bootstrap	False	False	True	False
F1 score	0.918	0.626	0.825	0.632

We applied our framework on ligands and receptors of classes A and AB. The performance results, evaluated in terms of AU-ROC, for the proposed methodology on the test set are presented in Tables 4 and 5.

The experiments were trained using 32 parallel threads on a HPC Server with eight 12-Core Intel Xeon Gold 5118 CPUs @2.30 GHz and 1.5 TB RAM running Fedora Linux 25.

Thanks to the appropriate division of molecules, we can compare our results with the ones obtained in [6]. The proposed methodology was also compared with

Table 4. Classification results (AU-ROC) on the test set for the proteins of class A of DB5.

	Receptors		Ligands	
	Bound	Unbound	Bound	Unbound
ProSps	0.962	0.962	0.615	0.615
Other Methods				
GCN Method with Contact Map	0.953	0.952	0.737	0.760
GCN Method with Hierarchical Representation 10	0.963	0.962	0.729	0.755
Daberdaku *et al.*	0.954	0.942	0.589	0.595
SPPIDER	0.773	0.754	0.630	0.575
NPS-HomPPI	0.796	0.780	0.610	0.626
PrISE	0.770	0.758	0.622	0.569

Table 5. Classification results (AU-ROC) on the test set for the proteins of class AB of DB5.

	Receptors		Ligands	
	Bound	Unbound	Bound	Unbound
ProSps	0.841	0.841	0.702	0.702
Other Methods				
GCN Method with Contact Map Å	0.904	0.899	0.711	0.778
GCN Method with Hierarchical Representation	0.905	0.903	0.749	0.800
Daberdaku *et al.*	0.813	0.840	0.599	0.729
SPPIDER	0.757	0.783	0.573	0.556
NPS-HomPPI	0.701	0.698	0.675	0.713
PrISE	0.776	0.789	0.683	0.649

two state-of-the-art homology-based PPI interface prediction algorithms: NPS-HomPPI [26] and PrISE [11], and with the well-known structure-based approach SPPIDER [16,17]. The results obtained with ProSPs and evaluated in terms of the AU-ROC on ligands and receptors of A and AB classes outperform HomPPI, a competitor predictors sequence-based, except for ligand of AB class. Moreover, they are comparable with the other results obtained with approaches based on the on three- dimensional structure (GCN Method with Contact Map, GCN Method with Hierarchical Representation, Daberdaku et al., SPIDER, PrISE).

4 Conclusions and Future Work

In this work, we have faced the protein interfaces prediction considering fragments of the amino acid sequence. We have proposed ProSPs, a method based

on the sliding windows approach and the Random Forest technique as a residue classifier. Such a method considers the minimum difference of normalized entropy to select the length of the sliding windows. Such a sliding windows approach is a fundamental aspect when only parts of proteins are known since it allows us to consider only fragments of the amino acid sequence. We tested the ProSPs on two classes of proteins, Antibody-Antigen and Antigen-Bound Antibody, extracted from the Protein-Protein Docking Benchmark 5.0. The obtained results evaluated in terms of the AU-ROC on these classes outperform HomPPI, a sequence-based competitor. They are comparable with the ones based on three-dimensional structure (GCN Method with Contact Map, GCN Method with Hierarchical Representation, Daberdaku et al., SPIDER, PrISE). As future work, we plan to apply our framework to all classes of DB5. Moreover, we intend to investigate the role of the length of the sliding windows. Therefore, we want to consider other methods to determine the number. Feature selection is another fundamental aspect to investigate since it represents a crucial step to represent the data. Moreover, our approach achieves better classification results for receptors than ligands, so we plan to evaluate different sets of features for the various protein classes. Motivated by the obtained results, we intend to extend the framework for predicting interacting sites in Protein-RNA interaction complexes. Moreover, we want also to consider the whole 3D structure of proteins considering structural features, such as the protein secondary structure by further exploring the RNA-based methodology introduced in [18,20,21].

Funding. MQ is supported by the "GNCS - INdAM". CF has been partially supported by the University of Padua project BIRD189710/18 "Reliable identification of the PPI interface in multiunit protein complexes".

References

1. Ahmad, S., Mizuguchi, K.: Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. PLoS ONE **6**(12), e29104 (2011)
2. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural Comput. **9**(7), 1545–1588 (1997)
3. Berggård, T., Linse, S., James, P.: Methods for the detection and analysis of protein-protein interactions. Proteomics **7**(16), 2833–2842 (2007)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Daberdaku, S.: Structure-based antibody paratope prediction with 3D zernike descriptors and SVM. In: Raposo, M., Ribeiro, P., Sério, S., Staiano, A., Ciaramella, A. (eds.) CIBB 2018. LNCS, vol. 11925, pp. 27–49. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-34585-3_4
6. Daberdaku, S., Ferrari, C.: Exploring the potential of 3D Zernike descriptors and SVM for protein-protein interface prediction. BMC Bioinform. **19**(1), 35 (2018)
7. Daberdaku, S., Ferrari, C.: Antibody interface prediction with 3D Zernike descriptors and SVM. Bioinformatics **35**(11), 1870–1876 (2019)
8. Fry, D.C.: Protein-protein interactions as targets for small molecule drug discovery. Peptide Sci. Original Res. Biomolecules **84**(6), 535–552 (2006)
9. Ho, T.K.: Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282. IEEE (1995)

10. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998)
11. Jordan, R.A., Yasser, E.M., Dobbs, D., Honavar, V.: Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinform. **13**(1), 41 (2012)
12. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M.: Aaindex: amino acid index database, progress report 2008. Nucleic Acids Res. **36**(suppl_1), D202–D205 (2007)
13. Keskin, O., Tuncbag, N., Gursoy, A.: Predicting protein-protein interactions from the molecular to the proteome level. Chem. Rev. **116**(8), 4884–4909 (2016)
14. Murakami, Y., Mizuguchi, K.: Applying the naïve bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. Bioinformatics **26**(15), 1841–1848 (2010)
15. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011). http://jmlr.org/papers/v12/pedregosa11a.html
16. Porollo, A., Meller, J.: Prediction-based fingerprints of protein-protein interactions. Proteins: Struct. Funct. Bioinform. **66**(3), 630–645 (2007)
17. Porollo, A., Meller, J., Cai, W., Hong, H.: Computational methods for prediction of protein-protein interaction sites. Protein-Protein Interact. Comput. Exp. Tools **472**, 3–26 (2012)
18. Quadrini, M., Culmone, R., Merelli, E.: Topological classification of RNA structures via intersection graph. In: Martín-Vide, C., Neruda, R., Vega-Rodríguez, M.A. (eds.) TPNC 2017. LNCS, vol. 10687, pp. 203–215. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71069-3_16
19. Quadrini, M., Daberdaku, S., Ferrari, C.: Hierarchical representation and graph convolutional networks for the prediction of protein–protein interaction sites. In: Nicosia, G., Ojha, V., La Malfa, E., Jansen, G., Sciacca, V., Pardalos, P., Giuffrida, G., Umeton, R. (eds.) LOD 2020. LNCS, vol. 12566, pp. 409–420. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-64580-9_34
20. Quadrini., M., Merelli., E., Piergallini., R.: Loop grammars to identify RNA structural patterns. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS, pp. 302–309. SciTePress (2019)
21. Quadrini, M., Tesei, L., Merelli, E.: ASPRAlign: a tool for the alignment of RNA secondary structures with arbitrary pseudoknots. Bioinformatics **36**(11), 3578–3579 (2020)
22. Saha, I., Maulik, U., Bandyopadhyay, S., Plewczynski, D.: Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino Acids **43**(2), 583–594 (2012)
23. Šikić, M., Tomić, S., Vlahoviček, K.: Prediction of protein-protein interaction sites in sequences and 3d structures by random forests. PLoS Comput. Biol. **5**(1), e1000278 (2009)
24. Sriwastava, B.K., Basu, S., Maulik, U.: Predicting protein-protein interaction sites with a novel membership based fuzzy SVM classifier. IEEE/ACM Trans. Comput. Biol. Bioinf. **12**(6), 1394–1404 (2015)
25. Vreven, T., et al.: Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J. Mol. Biol. **427**(19), 3031–3041 (2015)
26. Xue, L.C., Dobbs, D., Honavar, V.: Homppi: a class of sequence homology based protein-protein interface prediction methods. BMC Bioinform. **12**(1), 244 (2011)

27. Yin, S., Proctor, E.A., Lugovskoy, A.A., Dokholyan, N.V.: Fast screening of protein surfaces using geometric invariant fingerprints. Proc. Natl. Acad. Sci. **106**(39), 16622–16626 (2009)
28. Zhang, B., Li, J., Quan, L., Chen, Y., Lü, Q.: Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. Neurocomputing **357**, 86–100 (2019)