Pattern Recognition Letters 128 (2019) 231–236

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A dissimilarity-based multiple instance learning approach for protein remote homology detection

Antonella Mensi[a,*], Manuele Bicego[a], Pietro Lovato[a], Marco Loog[b], David M.J. Tax[b]

[a] Computer Science Department, University of Verona, Verona, Italy
[b] Delft University of Technology, Delft, the Netherlands

ARTICLE INFO

Article history:
Received 5 February 2019
Revised 26 July 2019
Accepted 26 August 2019
Available online 27 August 2019

MSC:
41A05
41A10
65D05
65D17

Keywords:
Protein Remote Homology Detection
Multiple-instance learning
Dissimilarity representation

ABSTRACT

We study the problem of Protein Remote Homology Detection, which assesses the functional similarity of two proteins. We approach this as a problem of binary multiple-instance learning (MIL) that aims to distinguish between homologous and non-homologous proteins. The particular MIL approach employed is based on the dissimilarity representation in which various schemes of combining N-gram representations are considered. This approach allows us to cope with longer N-grams, capturing a richer biological context, and results in versatile framework offering competitive performance compared to state of the art.

1. Introduction

Protein Remote Homology Detection (PRHD), is a challenging and widely studied bioinformatics problem [2,15,17]. The task is to assess whether two proteins have similar functions (i.e. are homologous). A common solution is to compare the proteins' sequences, since they determine the tertiary structure (often not available), which in turn determine the proteins' functions. The low sequence similarity often makes the task relatively complex.

Many different methods have been proposed to face the PRHD problem. Following the recent survey by Chen et al. [2], these methods can be divided in three categories: (i) alignment-based methods, which assess the homology by evaluating the pairwise alignment result of sequences; ii) ranked-based methods, that assess whether a protein belongs to a certain superfamily by looking at the most similar sequences, and iii) discriminative based methods, which are based on the use of a discriminative classifier for detecting the homology.

This study focuses on the last class of approaches: PRHD is cast as a binary classification problem that distinguishes between homologous and non-homologous proteins. More precisely, we concentrate on Support Vector Machines-based (SVM) approaches – since they obtain top performances on many benchmarks [8,17–21,26].

Since SVMs with standard kernels require as input a feature vector, a proper representation must be obtained. A very common one [18–21] is based on N-grams. An N-gram is a subsequence of consecutive symbols with fixed length N, extracted from the original sequence. This concept can be used to build a Bag of Words (BoW) representation, i.e., given a list (called a dictionary) of all possible N-grams, a vector is obtained by counting how many times each N-gram in the dictionary appears in the sequence to represent. Although rather straightforward, this technique has shown to perform very well in many studies [18–21]. Longer N-grams seem to encapsulate more significant information from a biological point of view [14]. Unfortunately, the use of BoW with N-grams with $N > 3$ is prohibitive, due to the exponential increase of the size of the representation vector, leading to the curse of dimensionality and sparsely populated vectors [6]. Alternatives to the BoW scheme are often based on a direct computation of kernels on the basis of long N-grams. A relevant example is [14], which proposes an N-gram based string kernel approach that obtains the best results using N-grams of length 5.

This paper presents a new SVM-based discriminative approach to face PRHD which overcomes the limitations of BoW-based methods. It proposes a novel representation that can employ

* Corresponding author.
E-mail address: antonella.mensi@univr.it (A. Mensi).

https://doi.org/10.1016/j.patrec.2019.08.027
0167-8655/

longer N-grams and manages a good trade-off between efficiency and accuracy.

The proposed method is based on Multiple Instance Learning (MIL), a recent learning paradigm [7] which extends classical supervised learning. The main difference with classical learning paradigms is that an object is not represented by a single feature vector, but with an unordered set of feature vectors, called instances. This set of instances, called a bag, has a unique label. This paradigm, which usefulness has been shown in many contexts [3,10], has, up to now, not been investigated in the Protein Remote Homology Detection scenario. In this paper we cast the PRHD task in a MIL framework: protein sequences are considered as bags of N-grams, i.e. subsequences, each one representing the instances. In particular, we adapted and extended a recent approach for MIL [4] which integrates the dissimilarity-based representation for Pattern Recognition, a paradigm introduced some years ago by Duin and Pekalska [23,25] to represent objects through dissimilarities. Our approach is very appropriate for the task at hand. First, the underlying MIL paradigm assumes that the label of a bag is determined by only a few of its instances [7]. This is especially true in PRHD, where the homology between two proteins is determined by the existence of few very informative subsequences (such as ligand sites). Second, the methodology is not limited by the length of the N-grams. Therefore the representation can leverage longer fragments. Finally, the proposed approach computes distances between pairs of fragments, allowing for the use of biologically meaningful and sophisticated distances.

The approach has been tested using standard benchmarks based on two datasets: SCOP 1.53 [17] and SCOP 2.04 [20]. The results we obtain demonstrate the suitability of the proposed approach, comparing favorably to current State of the Art in addition.

A preliminary version of this paper was published in [22]. This manuscript extends the aforementioned paper from both a methodological and experimental point of view; a new methodology for representing the bag as a vector is proposed; moreover a novel thorough experimental evaluation on a newer benchmark is presented.

Summarizing, this paper makes four contributions with respect to the state of the art. First we employ MIL in a novel scenario, which is the challenging task of Protein Remote Homology Detection, and we obtain excellent results. Second, we extend the method by [4]. We propose a new way to extract a robust descriptor from the dissimilarity matrices, called d_{MR} and we show that this novel variant is the best choice in terms of performance, i.e. supporting the methodological assumptions we made about it. The third contribution of our work concerns the possibility to use long fragments, i.e. longer than 3, which is very difficult for BoW-based methods. This is a severe limitation since longer fragments contain more information from a biological point of view. Lastly, we perform a thorough evaluation on the effect of different choices of the parameters, exploring different ways to encode the protein sequences under the chosen methodology. We analyze and compare the obtained variants from an experimental point of view as well, to assess their suitability for PRHD.

2. General and dissimilarity-based MIL

The main concept in MIL [7] is based on a new definition of object: an object is not a collection of features but rather a set (bag) of feature vectors (instances). A label is assigned to the whole bag, and not to its single instances, differently from the standard classification paradigm. Typically, not all instances are relevant for labeling the bag. The standard MIL assumption considers a bag positive if it contains at least an instance that is positive; conversely, a bag is negative if all of its instances are. Many different types of techniques exist to solve MIL tasks [3,5,10]. In this study we use a MIL approach based on dissimilarities [4,23,25]. In the dissimilarity-based paradigm an object is encoded as a vector where its entries represent dissimilarities to other objects, called prototypes. This approach makes it particularly easy to extract vectorial representations from non-vectorial objects.

We define T as the number of bags to encode, and L the number of prototypes used to build the representation. In the most simple case, every object in the training set can be a prototype. Each bag $B_i = \{\mathbf{x}_{i1}, \dots \mathbf{x}_{in_i}\}$ is composed of n_i instances, while each prototype $P_j = \{\mathbf{x}_{j1}, \dots \mathbf{x}_{jm_j}\}$ contains m_j instances. The steps to follow to encode the bag B_i as a vector are:

1. For each bag B_i and prototype P_j compute the pairwise distances between all instances of the bag and those of the prototype.
2. Reduce the obtained matrix $d(\mathbf{x}_{ik}, \mathbf{x}_{jl})$, $(k = 1, \dots, n_i, l = 1, \dots, m_j)$ of size $n \times m$ to a compact set of features.
3. Concatenate the outputs of Step 2. for all L prototypes to obtain the final vector representing B_i.

Clearly Step 2 can be performed in various ways:

1. d^{bag}: this strategy reduces the dissimilarity matrix to a single value in the following way:

$$d_{bag}(B_i, P_j) = \frac{1}{n_i} \sum_{k=1}^{n_i} \min_l d(\mathbf{x}_{ik}, \mathbf{x}_{jl}). \tag{1}$$

It extracts the minimum distances across all the instances of the prototype, and these minimum distances are then averaged to obtain a single feature that represents the dissimilarity between a bag and a prototype.

2. $\mathbf{d}_{inst}$: this strategy produces a vector of length m_j: for each prototype instance, we extract the minimum distances across all feature vectors of B_i.

$$\mathbf{d}_{inst}(B_i, P_j) = \left[\min_k d(\mathbf{x}_{ik}, \mathbf{x}_{j1}), \dots, \min_k d(\mathbf{x}_{ik}, \mathbf{x}_{jm_j})\right]. \tag{2}$$

The final representation of a bag B_i is the concatenation of either the d_{bag} or $\mathbf{d}_{inst}$ features computed with respect to all prototypes.

$$D_{bag}(B_i) = \left[d_{bag}(B_i, P_1), d_{bag}(B_i, P_2), \dots, d_{bag}(B_i, P_L)\right], \tag{3}$$

or

$$D_{inst}(B_i) = [\mathbf{d}_{inst}(B_i, P_1), \mathbf{d}_{inst}(B_i, P_2), \dots, \mathbf{d}_{inst}(B_i, P_L)], \tag{4}$$

which have lengths L and $\sum_{k=1}^{L} m_k$, respectively.

Since D_{bag} is an average over all distances, it may hide the most informative dissimilarities. D_{inst} may highlight them, but the process is computationally more expensive, and furthermore, D_{inst} may suffer from the curse of dimensionality, when $\sum_{k=1}^{L} m_k$ is high.

To solve this, Cheplygina et al. [4] proposes a variant which combines their respective strengths: maintain a low dimensionality, like D_{bag}, while still capturing the information retained in the dissimilarities, as D_{inst} does. To achieve this, the combining classifier paradigm is exploited [13]. This ensemble approach, called D_{ens}, treats each prototype as an independent subspace where a single classifier is trained and tested. The final classifier is then obtained by combining all these classifiers (one for each subspace). As in D_{inst}, the directions of each subspace correspond to the minimum distance between each instance of the given prototype and all instances of the bag. Thus, the dimensionality of the subspace, and thus of the final vector representing the original bag, reduces to the number of instances of a single prototype. Summarizing, after choosing L prototypes, we build L different representations and train L different classifiers. The final classifier is given by the aggregation of the results of the L classifiers independently trained,

using a specific combination function (this justifies the terminology of ensemble approach – for additional details please refer to Cheplygina et al. [4]).

3. MIL solution to the PRHD problem

The PRHD problem is cast into a MIL formulation by: (i) decomposing a protein sequence into N-grams, (ii) using the fragments (N-grams) as the instances, (iii) considering the protein sequence as the bag, and (iv) attaching the label of the whole sequence to the bag of instances.

The fragments can be extracted in several different ways (random sampling, exhaustive list, and so on). Here we use a simple strategy: given a sequence of length n, we extract all subsequences of length N, which is fixed, with maximum overlap, i.e. $N-1$. In this way, each protein sequence with length n is represented by a bag B_i containing $n-N+1$ N-grams. Then, we use the obtained bags as input to a dissimilarity-based approach, as described in Section 2. First, we define the set $\mathcal{P}=\{P_1\cdots P_L\}$ of prototypes (here we define it on the basis of $\mathcal{T}$, the original training set or on the basis of $\mathcal{T}'$ which contains a bag for each feature vector present in $\mathcal{T}$). Second, each prototype P_j is encoded as a MIL bag in the same way B_i was encoded. Subsequently, a matrix of dissimilarities is created for each pair (B_i, P_j). Finally, we extract a feature vector which summarizes the dissimilarity between the bag and the prototype.

Next to d_{bag} and $\mathbf{d}_{inst}$ presented in the previous section, we propose a novel approach, which we called $\mathbf{d}_{MR}$, where *MR* stands for multiresolution. First note that a fragment in a prototype may be more or less useful for the representation, depending on the sequence which is represented. This may affect the $\mathbf{d}_{inst}$ feature vector, which considers the distance from all fragments of the prototype. To solve this problem, we consider distances from *groups of fragments*, so that every sequence can use, in some sense, the *most suitable fragment* among a group of possible choices. In the PHRD domain, such a group can be defined as a set of q consecutive N-grams in a given sequence:

$$\mathbf{d}^q_{MR}(B_i,P_j)=\Big[min\Big(\min_k d(x_{ik},x_{j1}),\ldots,\min_k d(x_{ik},x_{jq})\Big),\ldots, min\Big(\min_k d(x_{ik},x_{jq+1}),\ldots,\min_k d(x_{ik},x_{j2q})\Big),\ldots\Big]. \quad (5)$$

When $q=2$ consecutive N-grams are used in a group, the dimensionality of the representation is halved, therefore reducing its *resolution*. In this way we are partially recovering the problem of destroying the structure of the sequence which is typical of N-gram based approaches.

Two observations must be made. First, in case $q=1$, the MR approach corresponds to $\mathbf{d}_{inst}$. Second, whenever the number of fragments is not a multiple of q, we must deal with extraction of the last feature.[1]

Finally, to obtain the vector representing B_i, two alternatives are possible. Either we concatenate the feature vectors obtained with the different prototypes (D_{inst}, D_{bag} or D_{MR}), or we use the D_{ens} method, which trains a classifier on each prototype independently. As to the application of D_{ens}, we can use the approach described in Section 2 (which we will refer to as $D_{ens}(Inst)$) based on the a posteriori combination of classifiers trained on different prototypes, or we can extend this by also combining different resolutions on the same prototype (referred to as $D_{ens}(MR)$). We will demonstrate in the experimental part that this strategy achieves state-of-the-art results. Note that L is the same as in the other approaches, even if the total number of subspaces will be higher (which depends on q).

[1] In this paper we keep the last set of fragments smaller than q.

One of the most crucial steps in dissimilarity-based approaches is the choice of the prototypes, both in terms of the number and in terms of the selection strategy [24,25]. Here we considered three different options.

i) **Random selection of sequences**: prototypes are randomly selected from protein sequences of the training set. In general, choosing prototypes randomly leads to good results [24,25].
ii) **Informed selection of sequences**: the selection of the prototypes depends on some a priori knowledge of the dataset. For example, we can select as prototype a sequence which is the most "central" in a given family (i.e. the sequence whose distance to all other sequences of the family is minimum).
iii) **Random fragments**: prototypes are randomly selected from fragments of the entire set of training bags, i.e. they do not correspond to whole protein sequences. The number of fragments must be fixed. This permits to have more diverse prototypes.

We also studied prototypes from another point of view: we designed two schemes called Same for All (SfA) and Different for All (DfA). The former corresponds to using the same set $\mathcal{P}$ of prototypes for all classification problems, while the latter instead employs a specific set of prototypes for each classification problem.

Note that our proposed approach allows to exploit long N-grams, i.e. $N>3$, since long fragments do not cause an exponential increase of the dimensionality (as it does in BoW-based techniques). Actually the dimensionality of the dissimilarity matrix and consequently that of the final vector representing the bag does not depend on the length of the N-grams, but only on the number of fragments and/or prototypes.

4. Experiments

All the variants of the proposed approach have been tested on the standard Protein Remote Homology Detection benchmark dataset[2], based on the SCOP 1.53 [17]. SCOP 1.53 is often used [8,17–21,26] and allows for a direct comparison with other discriminative methods. However it is somewhat dated and incomplete and to confirm the suitability of our method we also tested some of the best variants of our approach on SCOP 2.04,[3] a newer and more accurate dataset. It is processed in the same way as SCOP 1.53 and made available by the authors of Lovato et al. [20]. In all cases, we followed the protocol from [17] that casts the PRHD problem as several binary classification problems, one per protein family. For SCOP 1.53 we have 54 families, for the other dataset 89.

The N-grams extraction, as done in other studies [18–20], is performed after encoding each protein sequence using information extracted from the corresponding profile, a representation which takes into consideration evolutionary information derived from a multiple sequence alignment [1]. After processing the sequences, we extract N-grams and build the MIL representation. From the MIL representation, we extract feature vectors, as described in Section 3. The vectors are the inputs to a SVM classifier. As done in other works [9,18–21,26], we use the public GIST implementation,[4] leaving most of the parameters to their default value, except for the kernel function which was set to the radial basis function. To measure the accuracy, we employ the ROC50 score [11] which is a standard accuracy measure for the PRHD context that takes into consideration unbalanced datasets. ROC50 is the area under the Receiver Operating Characteristic curve – up to the first 50 false

[2] Available at http://noble.gs.washington.edu/proj/svm-pairwise/.
[3] Available at http://profs.sci.univr.it/~bicego/code/scop2.04.zip.
[4] Downloadable from http://www.chibi.ubc.ca/gist / [17].

Table 1
The different variants of the proposed approach.

Variant	MIL	Prot. Sel.	Detail
1.$\mathbf{D}_{bag}$-Info	D_{bag}	Informed	SfA: most central positive train seq per family, DfA: positive train seqs of the family
2.$\mathbf{D}_{inst}$-Info	D_{inst}	Informed	SfA: most central positive train seq of dataset, DfA: most central positive train seq of the family
3.$\mathbf{D}_{inst}$-RndFrag	D_{inst}	Rand Frag	fixed number of fragments randomly selected among all positive training fragments (SfA) or family-specific ones (DfA)
4.$\mathbf{D}_{ens}$(Inst)-RndSeq-Mean	$D_{ens}(Inst)$	Rand Seq	Prototype: random sequence from all positive training seqs (SfA) or among those belonging to the family (DfA), combination: mean.
5.$\mathbf{D}_{ens}$(Inst)-RndSeq-Max	$D_{ens}(Inst)$	Rand Seq	Equal to $\mathbf{D}_{ens}$(Inst)-RndSeq-Mean, combination: maximum
6.$\mathbf{D}_{ens}$(Inst)-RndFrag-Mean	$D_{ens}(Inst)$	Rand Frag	Prototypes created as in $\mathbf{D}_{inst}$-RndFrag, combination: mean
7.$\mathbf{D}_{ens}$(Inst)-RndFrag-Max	$D_{ens}(Inst)$	Rand Frag	Equal to $\mathbf{D}_{ens}$(Inst)-RndFrag-Mean, combination: maximum
8.$\mathbf{D}_{ens}$(MR)-RndFrag-Mean	$D_{ens}(MR)$	Rand Frag	Prototypes as in $\mathbf{D}_{inst}$-RndFrag, combination: mean, $q = 1,2,3,4$
9.$\mathbf{D}_{ens}$(MR)-RndFrag-Max	$D_{ens}(MR)$	Rand Frag	Equal to $\mathbf{D}_{ens}$(MR)-RndFrag-Mean, combination: maximum

positives. The score ranges from 0 to 1, where 0 indicates that the classifier is unable to distinguish the two classes and 1 indicates a perfect separation [16].

The experiments were repeated for different lengths of the N-grams, which are $N \in \{2, 3, 4, 5, 6, 9, 12\}$. The distance between the fragments is computed by using the Jukes-Cantor distance [12], a biological distance based on the Hamming metric. No alignment was carried out prior to the distance computation. We consider different variants of the proposed approach, attempting to capture the most relevant combinations of the basic schemes introduced in Sections 2 and 3 and of the way the prototypes are chosen. The variants investigated are described in Table 1.

4.1. SCOP 1.53: results and analyses

For SCOP 1.53, the ROC50 scores averaged over the 54 families are reported in Table 2 for each variant. Since the most suitable N may be different depending on the family under analysis (due to some intrinsic biological properties), we selected the highest score for each classification problem among the different choices, i.e. lengths, of N-grams. An additional analysis which highlights the preferred lengths for each experiment is presented later on in this section.

From Table 2, we can see that the most basic variant, $\mathbf{D}_{bag}$-Info performs almost as well as the ensemble variants. This surprising result – each prototype is represented by a single value– suggests that the representation is already very informative. Second, we notice that in general the *SfA* variant, i.e. the one for which the prototypes are equal across all families, is better than *DfA* where prototypes are indeed family-specific. Clearly this is not due to the fact that identical prototypes are used, because each classification problem is solved independently. Instead, the reason may be that the prototypes are generated starting from a bigger training set, not family-dependent, which contains more variability. This permits to have a potentially richer representation of the sequences. Interestingly, the results obtained with the D_{inst} technique show that prototypes with randomly selected fragments reach better performances than prototypes that were chosen in an informed way. This result is also supported by literature [25], where it has been shown that with big datasets random selection is a good technique for choosing the prototypes. Furthermore, in most cases the average combining rule to combine the scores leads to better results than the maximum combining rule, which is in line with some other studies in the field of combining classifiers [13,27]. In general, the $\mathbf{D}_{ens}(MR)$ scheme is the one which performs best. This result is very important as it confirms that the usage of an ensemble approach, as stated in [4], is able to overcome the limitations of the D_{bag} and D_{inst} representations. It also shows that using the same prototype in different ways, i.e. by extracting features at multiple resolutions via the D_{MR} approach, is better than using $D_{ens}(Inst)$ and thus it is better suited to exploit the information contained in the fragments.

Table 2
ROC50 accuracies of the different variants of the proposed approach.

Variant	ROC50 (SfA)	ROC50 (DfA)
$\mathbf{D}_{bag}$-Info	0.863	0.711
$\mathbf{D}_{inst}$-Info	0.820	0.781
$\mathbf{D}_{inst}$-RndFrag	0.867	0.862
$\mathbf{D}_{ens}$(Inst)-RndSeq-Mean	0.878	0.792
$\mathbf{D}_{ens}$(Inst)-RndSeq-Max	0.819	0.781
$\mathbf{D}_{ens}$(Inst)-RndFrag-Mean	0.886	0.859
$\mathbf{D}_{ens}$(Inst)-RndFrag-Max	0.860	0.840
$\mathbf{D}_{ens}$(MR)-RndFrag-Mean	0.890	0.828
$\mathbf{D}_{ens}$(MR)-RndFrag-Max	0.844	0.793

To understand what the influence of the number of prototypes L is, we carried out an additional set of experiments using the two best performing techniques, i.e. the variants $\mathbf{D}_{ens}(Inst)$-RndFrag-Mean (SfA) and $\mathbf{D}_{ens}(MR)$-RndFrag-Mean (SfA). In Table 3 (a) we report the ROC50 for different values if L. Performances seem to remain more or less stable when more than seven prototypes are used for both variants. This suggests that the approach is robust against variations in L, provided that this number exceeds a minimum (7 in this case). Another notable point is that the $\mathbf{D}_{ens}(MR)$ method performs better than the other ensemble-based approach in most cases.

To understand whether using long fragments leads to better classification, we analyzed the distribution of the best N. We partitioned the lengths in short $N \in \{2, 3\}$ and long $N > 3$. This analysis has been performed in two different ways. First we consider the best results, in terms of ROC50 scores, obtained for each family. Fig. 1 (a) shows a bar plot: for each variant the length of the darker bar indicates the average number of families that prefer a long N-gram, where the average was computed between the SfA and DfA versions of the same variant; analogously the length of the lighter bars indicates those families that prefer a short fragment. For the majority of families the best results are obtained when using longer N-grams. The second analysis focuses on the variant $\mathbf{D}_{ens}(Inst)$-RndFrag-Mean (SfA). For each family, fixed a number of prototypes, we choose two ROC50 scores: the first is the best score between those obtained with short N-grams, i.e. when $N \in 2$, 3, the other is chosen as the best score obtained when $N > 3$. Indeed the plot in Fig. 1(b) depicts how the averaged ROC50 score varies as the number of prototypes increases for short and long N-grams. The length of the vertical bars describes the standard error of the given population. From this analysis we see that the averaged scores are higher when dealing with longer N-grams. We also observe that on average the score tends to increase slightly when increasing the number of prototypes, but reaches a plateau eventually. This confirms what we stated in the previous analysis about the robustness with respect to the number of prototypes. All

Table 3
ROC50 results of the variants $\mathbf{D}_{ens}$(Inst)-RndFrag-Mean (SfA), $\mathbf{D}_{ens}$(MR)-RndFrag-Mean (SfA): (a) with varying number of prototypes and (b) with varying N-grams.

Nr. Prototypes	1	2	3	4	5	7	10	15	20	30	40	50
$\mathbf{D}_{ens}$(Inst)-RndFrag-Mean	0.854	0.859	0.862	0.870	0.869	0.895	0.886	0.885	0.885	0.885	0.886	0.885
$\mathbf{D}_{ens}$(MR)-RndFrag-Mean	0.866	0.870	0.867	0.862	0.873	0.890	0.890	0.888	0.890	0.881	0.875	0.876
					(a)							

N-gram	2GRAMS	3GRAMS	4GRAMS	5GRAMS	6GRAMS	9GRAMS	12GRAMS
$\mathbf{D}_{ens}$(Inst)-RndFrag-Mean	0.784	0.845	0.861	0.842	0.856	0.865	0.872
$\mathbf{D}_{ens}$(MR)-RndFragMean	0.789	0.857	0.879	0.862	0.863	0.879	0.897
					(b)		

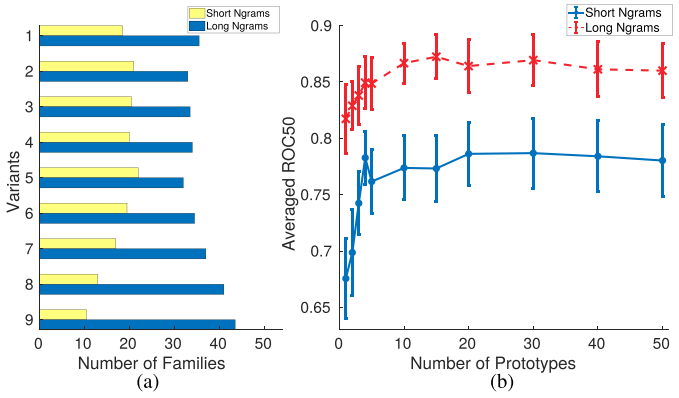

Fig. 1. Analysis of preferred N-gram length on SCOP 1.53. (a) The distribution of the best length over all approaches. The number on the left of each bar indicates the variant with respect to Table 1. (b) The ROC50 performance as a function of the number of prototypes in the $\mathbf{D}_{ens}$(Inst)-RndFrag-Mean (SfA) variant (best viewed in colors).

in all, it seems that using longer N-grams allows to reach better performances. This result is important as it confirms the biological hypothesis that longer subsequences are more informative, allowing a richer representation. A richer representation allows, in turn, a better discrimination and therefore better performances.

To further establish the suitability of longer subsequences in the PRHD context, we perform an additional analysis. It simply consists of fixing the length of the N-gram after which, for each family, we choose the number of prototypes that performs the best in terms of ROC50. We carried out this analysis for the two best variants $\mathbf{D}_{ens}$(Inst)-RndFrag-Mean (SfA) and $\mathbf{D}_{ens}$(MR)-RndFrag-Mean (SfA). Table 3(b) presents the ROC50 score averaged across all families. As can been seen the score increases with larger N in almost all cases, once more confirming the significance of longer fragments over short ones. In addition, we see that the $\mathbf{D}_{ens}$(MR)-RndFrag-Mean (SfA) approach outperforms the other one for all N-grams.

Table 4
ROC50 accuracies on SCOP 2.04 of the best variants of the proposed approach.

Variant	ROC50 (SfA)
$\mathbf{D}_{ens}(Inst)$-RndSeq-Mean	0.934
$\mathbf{D}_{ens}(Inst)$-RndSeq-Max	0.923
$\mathbf{D}_{ens}(Inst)$-RndFrag-Mean	0.930
$\mathbf{D}_{ens}(Inst)$-RndFrag-Max	0.916
$\mathbf{D}_{ens}(MR)$-RndFrag-Mean	0.948
$\mathbf{D}_{ens}(MR)$-RndFrag-Max	0.892

4.2. SCOP 2.04: results and analyses

Results are presented on SCOP 2.04 for some of the best variants. We decided on all variants based on D_{ens} in their SfA version. The ROC50 scores are again averaged and extracted in the same way as described in Section 4.1. Table 4 reports on the performance with $L = 10$. Clearly, the proposed framework performs satisfactorily also for this newer dataset. These experiments further reinforce that averaging the scores is a better choice when combining classifiers. In addition we see that the ensemble variant based on D_{MR} outperforms the one based solely on d_{inst} also for this newer dataset and so again combining classifers seems more suitable for the task at hand.

We carried out the same analyses we performed on SCOP 1.53 on the length of the N-grams. Fig. 2(a) shows a bar plot analogous to Fig. 1(a), with the exception that on this dataset the analysis was performed only on the SfA variants; the conclusion is similar, since for most families the best results are obtained when using longer N-grams. Fig. 2(b) shows a plot built in the same way as Fig. 1(b): when using longer N-grams better performances are achieved, independently of the number of prototypes.

4.3. Comparison with the state of the art

In Table 5, a comparisons with other methods on SCOP 1.53 is presented. We can see that the proposed approach is very com-

Table 5
SCOP 1.53, comparison with state of the art. For the proposed approach we reported the best obtained result, i.e. the result for $\mathbf{D}_{ens}(Inst)$-RndFrag-Mean (SfA) with 4 prototypes – see Table 3 (a).

N-grams based approaches			Other approaches		
Method	Year	ROC50	Method	Year	ROC50
BoW-row-2gram	2017	0.772 [20]	SVM-pairwise	2014	0.787 [19]
Soft BoW	2017	0.844 [20]	SVM-LA	2014	0.752 [19]
Soft PLSA	2017	0.917 [20]	HHSearch	2017	0.801 [20]
SVM-N-gram	2014	0.589 [19]	Profile (5,7.5)	2005	0.796 [14]
SVM-N-gram-LSA	2008	0.628 [18]	PSI-BLAST	2007	0.330 [8]
SVM-Top-N-gram (n = 2)	2008	0.713 [18]	SVM-Bprofile-LSA	2007	0.698 [8]
SVM-Top-N-gram-combine	2008	0.763 [18]	SVM-Pattern-LSA	2008	0.626 [18]
SVM-N-gram-p1	2014	0.726 [19]	SVM-Motif-LSA	2008	0.628 [18]
SVM-N-gram-KTA	2014	0.731 [19]	SVM-LA-p1	2014	0.888 [19]
ROC50 of the proposed approach:		0.897			

Table 6
SCOP 2.04, comparison with state of the art. For the proposed approach we reported the best obtained result, i.e. the result for $\mathbf{D}_{ens}(MR)$-RndFrag-Mean (SfA) with 10 prototypes – see Table 4.

Approach	ROC50
BoW row(1,2)-gram	0.864 [20]
softBoW,prod-col 2gram	0.899 [20]
softPLSA,prod-col(1,2)-gram	0.942 [20]
softPLSA,prod-row(1,2)-gram	0.944 [20]
	ROC50 of the proposed approach: 0.948

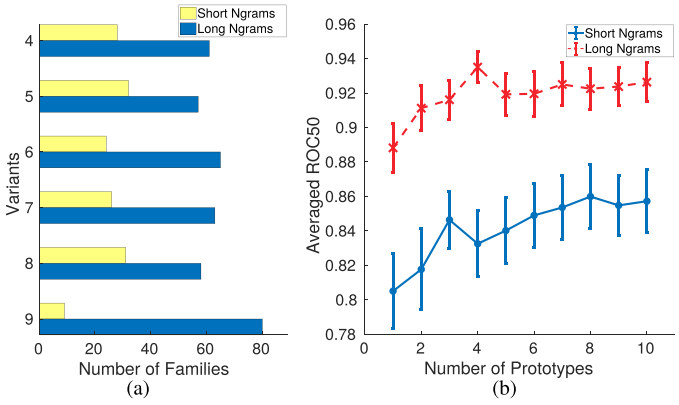

Fig. 2. Analysis of preferred N-gram length on SCOP 2.04. (a) The distribution of the best length over all approaches. The number on the left of each bar indicates the variant with respect to Table 1. (b) The ROC50 performance as a function of the number of prototypes (best viewed in colors).

petitive, well comparing with alternatives. Our ROC50 best score, obtained with the variant $\mathbf{D}_{ens}(MR)$-RndFrag-Mean, performs better than most approaches from the current literature. The only method that outperforms ours is SoftPLSA, which is a very complex approach based on a richer representation that exploits more evolutionary information (see [20] for more details). However, on the more accurate SCOP 2.04, our proposed approach outperforms also SoftPLSA, confirming to be a flexible and accurate alternative solution for the PRHD (see Table 6).

5. Conclusions

This paper presents a new approach to solve the Protein Remote Homology Detection problem, based on the Multiple Instance Learning paradigm combined with a dissimilarity-based representation and tailoring the methodology proposed by [4]. We designed various specific approaches and performed an extensive comparison on different datasets to test the robustness and suitability of the method; the experimental part also showed that the usage of longer fragments allows to reach better performances, confirming their informativeness.

Declaration of Competing Interest

None.

Acknowledgments

M. Bicego and P. Lovato were partially supported by the University of Verona through the program “Bando di Ateneo per la Ricerca di Base 2015”.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2019.08.027.

References

[1] S. Altschul, T. Madden, A.A. Schaffer, et al., Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic Acid Res. 25 (17) (1997) 3389–3402.
[2] J. Chen, M. Guo, X. Wang, B. Liu, A comprehensive review and comparison of different computational methods for protein remote homology detection, Briefings Bioinf. (2016) 1–14.
[3] Y. Chen, J. Bi, J.Z. Wang, Miles: multiple-instance learning via embedded instance selection, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 1931–1947.
[4] V. Cheplygina, D.M.J. Tax, M. Loog, Dissimilarity-based ensembles for multiple instance learning, IEEE Trans. Neural Netw. Learn.Syst. 27 (6) (2016).
[5] V. Cheplygina, D.M.J. Tax, M. Loog, On classification with bags, groups and sets, Pattern Recognit. Lett. 59 (2015) 11–17.
[6] A. Cucci, P. Lovato, M. Bicego, Enriched bag of words for protein remote homology detection, in: Proc. Int. Workshop on Statistical Techniques in Pattern Recognition (S+SSPR2016), 2016, pp. 463–473.
[7] T. Dietterich, R. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artif. Intell. 89 (1–2) (1997) 31–71.
[8] Q. Dong, L. Lin, X. Wang, Protein remote homology detection based on binary profiles, Bioinf. Res. Dev. (2007) 212–223.
[9] Q. Dong, X. Wang, L. Lin, Application of latent semantic analysis to protein remote homology detection., Bioinformatics 22 (3) (2006) 285–290.
[10] G. Fung, M. Dundar, B. Krishnapuram, R. Rao, Multiple instance learning for computer aided diagnosis, in: Proc. Adv. Neural Inf. Process. Syst., 19, 2007, pp. 425–432.
[11] M. Gribskov, N. Robinson, Use of receiver operating characteristic (roc) analysis to evaluate sequence matching., Comput. Chem. 20 (1) (1996) 25–33.
[12] T.H. Jukes, C.R. Cantor, Chapter 24 - evolution of protein molecules, in: H. MUNRO (Ed.), Mammalian Protein Metabolism, Academic Press, 1969, pp. 21–132.
[13] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach.Intell. 20 (3) (1998) 226–239.
[14] R. Kuang, K. Wang, K. Wang, M. Siddiqi, Y. Freund, C. Leslie, Profile-based string kernels for remote homology detection and motif extraction, J. Bioinf. Comput.Biol. 3 (03) (2005) 527–550.
[15] P.P. Kuksa, V. Pavlovic, Efficient evaluation of large sequence kernels, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 759–767.
[16] C. Leslie, E. Eskin, W. Noble, The spectrum kernel: a string kernel for svm protein classification, in: PSB, 2002, pp. 566–575.
[17] L. Liao, W. Noble, Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships, J. Comput. Biol. 10 (6) (2003) 857–868.
[18] B. Liu, X. Wang, L. Lin, Q. Dong, X. Wang, A discriminative method for protein remote homology detection and fold recognition combining top-n-grams and latent semantic analysis, BMC Bioinf. 9 (1) (2008) 510.
[19] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Q. Dong, K. Chou, Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, Bioinformatics 30(4) (2014) 472–479.
[20] P. Lovato, M. Cristani, M. Bicego, Soft ngram representation and modeling for protein remote homology detection, IEEE/ACM Trans. Comput. Biol.Bioinf. 14 (6) (2017) 1482–1488.
[21] P. Lovato, A. Giorgetti, M. Bicego, A multimodal approach for protein remote homology detection, IEEE/ACM Trans. Comput. Biol.Bioinf. (TCBB) 12 (5) (2015) 1193–1198.
[22] A. Mensi, M. Bicego, P. Lovato, M. Loog, D.M.J. Tax, Protein remote homology detection using dissimilarity-based multiple instance learning, in: X. Bai, E.R. Hancock, T.K. Ho, R.C. Wilson, B. Biggio, A. Robles-Kelly (Eds.), Structural, Syntactic, and Statistical Pattern Recognition, Springer International Publishing, Cham, 2018, pp. 119–129.
[23] E. Pekalska, R.P.W. Duin, Dissimilarity representations allow for building good classifiers, Pattern Recognit. Lett. 23 (8) (2002) 943–956.
[24] E. Pekalska, R.P.W. Duin, P. Paclík, Prototype selection for dissimilarity-based classifiers, Pattern Recognit. 39 (2) (2006) 189–208.
[25] E. Pekalska, R.P.W. Duin, The Dissimilarity Representation for Pattern Recognition: Foundations and Applications, Machine Perception and Artificial Intelligence, 64, World Scientific, Singapore, 2005.
[26] H. Rangwala, G. Karypis, Profile-based direct kernels for remote homology detection and fold recognition, Bioinformatics 21 (23) (2005) 4239–4247.
[27] D.M.J. Tax, M. Van Breukelen, R.P.W. Duin, J. Kittler, Combining multiple classifiers by averaging or by multiplying? Pattern Recognit. 33 (9) (2000) 1475–1485.