Protein Remote Homology Detection Using Dissimilarity-Based Multiple Instance Learning

Antonelli Mensi[1], Manuele Bicego[1(✉)], Pietro Lovato[1], Marco Loog[2], and David M. J. Tax[2]

[1] University of Verona, Verona, Italy
manuele.bicego@univr.it
[2] Delft University of Technology, Delft, The Netherlands

Abstract. A challenging Pattern Recognition problem in Bioinformatics concerns the detection of a functional relation between two proteins even when they show very low sequence similarity – this is the so-called Protein Remote Homology Detection (PRHD) problem. In this paper we propose a novel approach to PRHD, which casts the problem into a Multiple Instance Learning (MIL) framework, which seems very suitable for this context. Experiments on a standard benchmark show very competitive performances, also in comparison with alternative discriminative methods.

Keywords: Protein homology · N-grams · Multiple instance learning

1 Introduction

The Protein Remote Homology Detection (PRHD) problem represents a relevant bioinformatics problem, widely studied in recent years [1,12,14]. It aims at identifying functionally or structurally-related proteins by looking at amino acid sequence similarity – where the term *remote* refers to some very challenging situations where homologous proteins exhibit very low sequence similarity. Many computational approaches have been developed to face this problem – see for example the very recent review published in [1]. In a broad sense, such approaches are divided in three main categories [1]: alignment-based methods, rank-based methods, and discriminative-based methods. Here we focus on this last category, which casts the problem in a binary classification task (homologous/not homologous), and in particular on approaches based on the Support Vector Machines (SVM) classifier – shown to reach top performances in many different benchmarks [6,14–18,20].

To apply the SVM, the typical choice is to derive a vectorial representation, so that classic kernels (such as RBF - Radial Basis Function- kernels) can be

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".

© Springer Nature Switzerland AG 2018
X. Bai et al. (Eds.): S+SSPR 2018, LNCS 11004, pp. 119–129, 2018.
https://doi.org/10.1007/978-3-319-97785-0_12

applied. In this scenario representations based on N-grams (or K-mers[1]) – short subsequences of consecutive symbols – are widely employed [15–18]. The well known Bag of Words representation is an example of such characterization [7, 15,17,18]. Here a vectorial representation is extracted consisting of the number of times the dictionary N-grams appear in the sequence. Although this leads to excellent results, the main problem of this class of approaches is that N (i.e. the length of the subsequence) is forced to remain small (such as 3). For longer N-grams, the representation becomes too large (leading to the curse of dimensionality) and too sparse (with too many zeros), thus creating problems to the SVM [4]. Actually, due to the limited length, we can not fully exploit the biological information present in longer sequences. An alternative is to devise methods which directly compute kernels on the basis of long K-mers, avoiding the explicit computation of the representation. One notable example is [11], where authors propose a K-mer based string kernel approach. In their work they showed that the best performances are obtained with K-mers of length 5.

In this paper we propose a novel approach to PRHD, which derives a novel vectorial representation for SVM-based discriminative techniques. The approach is based on the paradigm of Multiple Instance Learning (MIL – [5]), an extension of supervised learning where class labels are associated with sets (bags) of feature vectors (instances) rather than with individual feature vectors. This paradigm, which usefulness has been shown in many different contexts [2,8], has not yet been investigated in the Protein Remote Homology Detection scenario. Here we cast the PRHD problem in a MIL framework by interpreting protein sequences as bags that contain fragments of a certain length k (the instances). The classification problem is solved using a recent MIL approach based on dissimilarities between instances [3]. The MIL scenario, and in particular the dissimilarity-based approach of [3], seems to be very suitable for the PRHD problem for different reasons. First, the MIL paradigm assumes that the label of the whole bag is determined by only a small set of relevant instances [5]. This assumption is reasonable in PRHD, where the homology between two proteins is linked to the presence of a small set of highly informative fragments (such as ligand sites). Second, it does not impose any limit to the length of the K-mers, so that also biologically meaningful longer fragments can be included in the analysis. Third, the approach of [3] relies on the computation of distances between instances, which in the PRHD case can be easily defined via meaningful sequence alignment methods.

The proposed approach, presented in some different variants, has been tested using standard benchmarks based on the SCOP 1.53 dataset [14]. The results confirm the suitability of the proposed approach, also in comparison with alternative discriminative methods.

[1] Along the text we will refer equivalently to K-mers or N-grams.

2 General and Dissimilarity-Based MIL

In this section we introduce the general multiple instance learning paradigm, together with the approach presented in [3] that we used. Multiple Instance Learning (MIL – [5]) is concerned with problems where the objects originally are not represented by a single feature vector, but by a so-called bag. A bag is basically a *set* of feature vectors, the latter of which are also referred to as instances in this context. As opposed to the standard classification setting, a label is then assigned to the whole bag and not the individual feature vectors. This can make classification quite difficult. The basic assumption behind MIL is that a positive label of a bag indicates the presence of (at least) a positive instance inside the bag – we will see that this assumption is very suitable for our context.

Many different approaches have been proposed to solve MIL problems [2,8], here we summarize the methods proposed in [3]. These methods are based on the dissimilarity-based paradigm for classification [19], a paradigm where each object is represented by a vector of dissimilarities with respect to a set of reference objects (called prototypes). In the same spirit, in the approach of [3] each bag is encoded into a vectorial representation based on the distances between the instances of the bag and the instances of a set of prototypes.

More in detail, we are given N bags to encode and a set of L prototypes. The choice of these prototypes is crucial, but in the basic version they can also be the whole training set. Given prototype P_j containing m instances, $P_j = \{x_{j1}, ...x_{jm}\}$, we represent a bag $B_i = \{x_{i1}, ...x_{in}\}$ with n instances, by some signature extracted from the pairwise distances between all the instances of B_i and those of the prototype bag P_j. Different features can be extracted from the resulting $n \times m$ dissimilarity matrix.

1. d_{bag} feature. This feature is a scalar, and represents the average of the minimum distances between each fragment of the bag and all the fragments of the prototype.

$$d_{bag}(B_i, P_j) = \frac{1}{|B_i|} \sum_{k=1}^{|B_i|} \min_l d(x_{ik}, x_{jl})$$

   where $d(x_{ik}, x_{jl})$ represents a distance between instances of the bag.
2. d_{inst} feature. This is a vector of length m, where each component represents the minimum distance between each fragment of the prototype and all fragments of the bag.

$$d_{inst}(B_i, P_j) = \left[\min_k d(x_{ik}, x_{j1}), ..., \min_k d(x_{ik}, x_{jm})\right]$$

In the first two MIL schemes, which are called D_{bag} and D_{inst}, each bag is represented by concatenating all the d_{bag} and d_{inst} features computed with respect to all prototypes, i.e. $D_{bag}(B_i) = [d_{bag}(B_i, P_1), d_{bag}(B_i, P_2), ...d_{bag}(B_i, P_L)]$ and $D_{inst}(B_i) = [d_{inst}(B_i, P_1), d_{inst}(B_i, P_2), ...d_{inst}(B_i, P_L)]$.

These representations may have some limitations: D_{bag} may hide the most informative dissimilarities, since it is an average over all distances, not considering that only few instances are relevant. The D_{inst} method, on the contrary, considers all these dissimilarities, but the process of selection can be time consuming. Furthermore it may suffer from the curse of dimensionality. To overcome these possible limitations, the authors in [3] proposed a variant which exploits the combining classifier paradigm. The method, which we call the "ensemble" approach, is based on considering each prototype as a single subspace where a classifier is trained. Similarly to the D_{inst} method, each direction of the subspace represents the minimum distance between each instance of the prototype and all instances of the bag. The dimensionality of this subspace is therefore the number of instances of the prototype. Given L prototypes, we built L different representations, training L different classifiers. The final classifier is then found by aggregating the results of the L different classifiers via a combining function (in this sense it is an ensemble approach) – for further details please refer to [3].

3 MIL Solution to the PRHD Problem

In our proposed approach we first cast the PRHD problem into a MIL formulation, i.e. we define bags, instances and labels. This is done in a reasonable and straightforward way: (i) each protein sequence is a bag, i.e. a collection of N-grams (instances); (ii) the fragments (N-grams) composing the protein sequence are considered the instances; (iii) finally, the label, which is attached to the set of instances, is the label of the sequence. Please note that MIL represents a natural representation for the PRHD problem: proteins typically contain a small set of meaningful fragments, which are crucial to determine the 3D structure (e.g. binding sites) and thus the function (namely the label). Clearly, the fragments can be extracted from the sequence in many different ways (random sampling, exhaustive list, and so on). Here we adopt a very simple scheme: from each sequence of length n, fragments of a fixed length k are extracted with overlap $k-1$. Each bag B_i will therefore have $n-k+1$ instances. Once cast into a MIL formulation, the PRHD problem is then input to the dissimilarity-based approach presented in the previous section. In particular, a set of prototypes $\mathcal{P} = \{P_1 \cdots P_L\}$ is chosen as a subset of the training set $\mathcal{T}$. Given a prototype P_j, for each sequence S_i we compute a dissimilarity matrix between all fragments of P_j and all the fragments of S_i (i.e. the bag B_i). As described in the previous section, from this matrix we then derive two different representations: a scalar (d_{bag}) or a set of values (d_{inst}). In the basic formulation, the dissimilarity matrices are extracted for all prototypes and concatenated to obtain the final representation of our sequence. The proposed representation can now be fed to the SVM classifier. Alternatively, the ensemble method described in the previous section can be used: the classifier is trained on d_{inst} of a single prototype, called a subspace, and then the obtained scores are combined together to obtain the final results via an ensemble classifier. Summarizing, we have three different MIL schemes: one using (D_{bag}), one using (D_{inst}), and the last using the ensemble approach (D_{ens}).

One crucial aspect of this class of approaches is the choice of the prototypes. First, the number of prototypes has to be chosen. Next, it is crucial to define the strategy with which they are chosen. Here we studied three different options:

(i) **Random choice of sequences**: the prototypes are randomly selected protein sequences of the training set.
(ii) **Informed choice of sequences**: the prototypes are chosen exploiting some a priori knowledge on the training set.
(iii) **Random fragments**: here the prototypes are not anymore objects of the training set (i.e. whole sequences), but they are built using random fragments extracted from sequences. After deciding on the number of fragments that should compose each prototype, we randomly select those fragments from the whole set of bags. Note that our proposed scheme allows to exploit long K-mers without increasing in a significant way the dimensionality. In fact, the dissimilarity matrix between bag's instances, which is at the basis of our scheme, does not depend from the length of the K-mers, but only the the number. This permits to exploit longer fragments with respect to classic N-grams methods, which may contain more important biological information, such as that related to folding.

4 Experiments

The proposed approach has been tested on the standard benchmark dataset[2], based on the SCOP 1.53 [14]. Even if quite old and not complete, this represents a standard dataset for protein remote homology detection, permitting to compare most of the methods introduced in this field [6,14–18,20]. Following the standard protocol introduced in [14], the PRHD problem has been cast in a set of 54 binary classification problems, each one involving a specific protein family. As done in some recent studies [15–17], before extracting N-grams we re-wrote each protein sequence using information extracted from the corresponding profile, determined by following the recent [16], which employed a public implementation of the PsiFreq program[3].

Once determined, the MIL representations are then employed to train a SVM classifier. As done in many previous works [7,15–18,20], we used the public GIST implementation[4], setting the kernel type to radial basis, and keeping the remaining parameters to their default values. Detection accuracies are measured using the ROC50 score [9]. This score, specifically designed for the PRHD context, improves the classic Area under the ROC curve. In particular, it represents the area under the ROC50 curve (with a value ranging from 0 to 1), which plots true positives as a function of false positives – up to the first 50 false positives. A score of 1 indicates perfect separation of positives from negatives, whereas a score of 0 indicates that none of the top 50 sequences selected by the algorithm were positives [13].

[2] Available at http://noble.gs.washington.edu/proj/svm-pairwise/.
[3] Available at http://bioinformatics.hitsz.edu.cn/main/~binliu/remote.
[4] Downloadable from http://www.chibi.ubc.ca/gist/ [14].

For the proposed approach, we repeated the experiment for $k = \{2, 3, 4, 5, 6, 9, 12\}$. The distance between the K-mers was computed using the classic Jukes-Cantor distance, based on the Hamming distance. Please note that this is a basic distance between sequences, which does not imply any alignment. It can be expected that performances may improve even more when more advanced sequence comparison methods are used, for instance methods that allow for the comparison of K-mers of different lengths. We tested different variants of the proposed approach, trying to cover the most interesting combinations of the basic scheme ((D_{bag}), (D_{inst}), and (D_{ens})) and the way prototypes are chosen. For all variants we investigated two possible options, which derive from the fact that the benchmark contains 54 classification problems. In particular, in the first version (called SfA – Same for All) the prototypes were kept identical among all 54 problems. In the second version (called DfA - Different for All) a different set of prototypes is used for each family. In particular the following variants have been investigated:

(i) **$\mathbf{D}_{bag}$-Info**. In this variant, we used the D_{bag} information to build the representation, choosing the prototypes in an informed way. In the SfA version, we used 54 prototypes, equal for all families: each prototype is the most central sequence of the positive training set of each family, that is the one with lowest distance to all other sequences. In the DfA version, for each family we used as prototypes all the sequences in the positive part of training set.

(ii) **$\mathbf{D}_{inst}$-Info**. In this variant we used the D_{inst} information to build the representation. Due to the high dimensionality of this representation, we choose to employ a single prototype, chosen in an informed way. In particular, in the SfA version, the prototype was chosen as the most central sequence among all positive training sequences of the 54 families. In the DfA version, for each family the prototype was chosen as the most central sequence among the positive training sequences of the considered family.

(iii) **$\mathbf{D}_{inst}$-RndFrag**. In this variant we used again the D_{inst} information to build the representation, employing again one prototype. However the prototype was chosen using random fragments. In the SfA version, the fragments are extracted from the set composed by the fragments of all the positive training sequences of all families. The cardinality of the prototype P is the ratio between the total number of fragments of the just mentioned bag and the total number of positive training sequences. In the DfA version, for each family the random fragments are chosen among the set composed by the fragments of all the positive training sequences of the considered family. The cardinality of each prototype P is the ratio between the total number of fragments of the just mentioned bag and the number of positive training sequences.

(iv) **$\mathbf{D}_{ens}$-RndSeq-Mean**. In this variant we used the ensemble MIL scheme to build the representation, using random sequences as prototypes. In particular, in the SfA version, we randomly chose 10 prototypes from the set of all positive training sequences of the 54 problems. Then we extract the

D_{inst} representation for each prototype, training a different SVM for each of them. Once computed the SVM scores, a "mean" combiner function is used to get the final score (i.e. the mean of all scores). In the DfA version, the 10 prototypes were different for each classification problem. In particular, for each family we selected 10 prototypes from the set of positive training sequences of that family. A study on the performances by using a different number of prototypes is reported later.

(v) **$\mathbf{D}_{ens}$-RndSeq-Max**. This is identical to the **$\mathbf{D}_{ens}$-RndSeq-Mean** except that the combiner was a "max" combiner (i.e. the max among the scores).

(vi) **$\mathbf{D}_{ens}$-RndFrag-Mean**. This variant is similar to **$\mathbf{D}_{ens}$-RndSeq-Mean**, except that the prototypes are built using Random Fragments. Prototypes, for both SfA and DfA versions are determined as described in the **$\mathbf{D}_{inst}$-RndFrag** variant. In this version we used the "mean" combiner.

(vii) **$\mathbf{D}_{ens}$-RndFrag-Max**. This is identical to the **$\mathbf{D}_{ens}$-RndFrag-Mean** except that we used the "Max" combiner.

For each experiment we selected the best result among the different lengths of N-grams (which can be reasonably different depending on the specific family addressed). A further analysis on the preferred length has been reported later in the section. ROC50 values, averaged over the 54 families, are reported in Table 1, for the different variants. From the table we make different observations. First, it is interesting to note that the most basic variant of our scheme, namely the **$\mathbf{D}_{bag}$-Info**, is performing very well, at the same level of the most complicated variants. This suggests that the extracted information, even in its basic form, is already very informative. Second, it seems evident that choosing the same set of prototypes for all families permits to reach better performances in almost all cases. Actually we are convinced that the crucial point is not that the prototypes are the same for all classification problem (each classification problem is solved independently), but rather that this set is chosen among the whole set of sequences rather than the single training set of a given family. This permits to have a more variable set of prototypes which permits to get a richer representation. Interestingly, the informed choice of the prototypes does not improve in a substantial way the performances. As a final observation, it is important

Table 1. ROC50 accuracies of the different variants of the proposed approach.

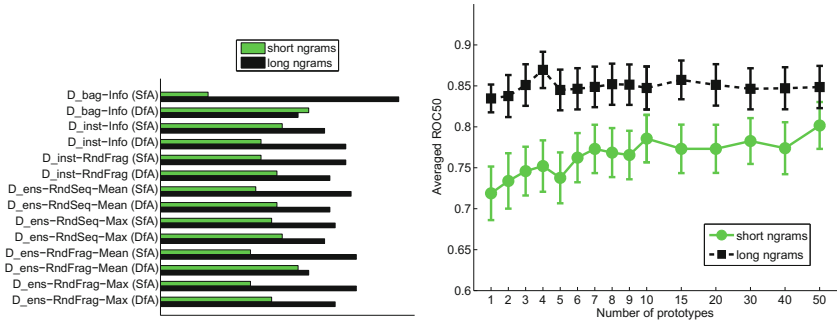
Variant	MIL scheme	Prot. Sel.	ROC50 (SfA)	ROC50 (DfA)
$\mathbf{D}_{bag}$-Info	D_{bag}	Informed	0.863	0.711
$\mathbf{D}_{inst}$-Info	D_{inst}	Informed	0.820	0.781
$\mathbf{D}_{inst}$-RndFrag	D_{inst}	Rand Frag	0.867	0.862
$\mathbf{D}_{ens}$-RndSeq-Mean	D_{ens}	Rand Seq	0.878	0.792
$\mathbf{D}_{ens}$-RndSeq-Max	D_{ens}	Rand Seq	0.819	0.781
$\mathbf{D}_{ens}$-RndFrag-Mean	D_{ens}	Rand Frag	0.882	0.847
$\mathbf{D}_{ens}$-RndFrag-Max	D_{ens}	Rand Frag	0.837	0.878

Table 2. Results of the variant $\mathbf{D}_{ens}$**-RndFrag-Mean** (SfA) with varying number of prototypes.

Nr. prototypes	**1**	**2**	**3**	**4**	**5**	**7**	**10**	**15**	**20**	**30**	**40**	**50**
ROC 50	0.867	0.872	0.886	0.892	0.880	0.882	0.882	0.874	0.879	0.868	0.870	0.880

to note that when combining the classifiers in the $\mathbf{D}_{ens}$ class of approaches the best result is obtained with the mean rule (in line with other studies in classifiers combination [10]).

In order to see how critical the number of prototypes L is, we performed another set of experiments using the best performing technique, i.e. the variant $\mathbf{D}_{ens}$**-RndFrag-Mean** (SfA). We varied the number of prototypes from 1 to 50, and the corresponding accuracies are reported in Table 2. It appears that performances do not vary too much when more than 3 prototypes are used. This suggests that the approach is robust against variations in L, provided that this number exceeds a minimum (3 in this case). Another interesting aspect to be analysed concerns the length of the K-mers. As already mentioned, in our experiments we computed results by varying the length k of the fragments, selecting, for each family, the length leading to the best accuracy. It seems interesting to observe the distribution of such best k, in order to discover if the MIL approach prefers short or long N-grams. To do that, for each variant, we count how many times the best result is obtained with *short N-grams* (N-grams of length 2 or 3) or with *long N-grams* (N larger than 3). Such analysis is reported in Fig. 1(a). In all cases except the $\mathbf{D}_{bag}$**-Info**(DfA) variant, longer fragments give better results. Furthermore, in Fig. 1(b) the accuracies obtained by $\mathbf{D}_{ens}$**-RndFrag-Mean** (SfA) are shown for an increasing number of prototypes (results of Table 2), divided in two cases: method with *short N-grams* and



(a) (b)

Fig. 1. Analysis of preferred N-gram length: (a) the distribution of the best length over all approaches and (b) the ROC50 performance as a function of the number of prototypes.

method with *long N-grams*. The results with *long N-grams* are better and seem to be more independent from the number of prototypes (whereas with *short N-grams* there seems to be an increasing behaviour). All these findings confirm our intuition that exploiting longer fragments can be beneficial for facing the Protein Remote Homology Detection problem.

4.1 Comparison with the State of the Art

In Table 3 we compared the proposed scheme with alternative approaches present in the literature. The SCOP 1.53 dataset, even if being old, has been widely used as benchmark for many different approaches. We reported in the table comparative results taken from the very recent [17], which are related to both Bag of Words approaches as well as more complicated alternatives. We can see that the proposed approach is very competitive, well comparing with alternatives. In particular, the proposed approach is better than almost all methods presented in the table, with the exception of the very complex Soft PLSA approach [17]: this recent method, however, starts from a larger set of information – the complete profile of each protein together with evolutionary probabilities – whereas our approach only uses the most probable profile (for more information, interested readers are referred to [17]).

Table 3. Comparison with state of the art. For the proposed approach we reported the best obtained result, i.e. the result for **$\mathbf{D}_{ens}$-RndFrag-Mean** (SfA) with 4 prototypes – see Table 2.

N-grams based approaches			*Other approaches*		
Method	**Year**	**ROC50**	**Method**	**Year**	**ROC50**
BoW-row-2gram	2017	0.772 [17]	SVM-pairwise	2014	0.787 [16]
Soft BoW	2017	0.844 [17]	SVM-LA	2014	0.752 [16]
Soft PLSA	2017	0.917 [17]	HHSearch	2017	0.801 [17]
SVM-N-gram	2014	0.589 [16]	Profile (5,7.5)	2005	0.796 [11]
SVM-N-gram-LSA	2008	0.628 [15]	PSI-BLAST	2007	0.330 [6]
SVM-Top-N-gram (n = 2)	2008	0.713 [15]	SVM-Bprofile-LSA	2007	0.698 [6]
SVM-Top-N-gram-combine	2008	0.763 [15]	SVM-Pattern-LSA	2008	0.626 [15]
SVM-N-gram-p1	2014	0.726 [16]	SVM-Motif-LSA	2008	0.628 [15]
SVM-N-gram-KTA	2014	0.731 [16]	SVM-LA-p1	2014	0.888 [16]
ROC50 of the proposed approach: 0.892					

5 Conclusions

In this paper we presented a Multiple Instance Learning approach for Protein Remote Homology detection. The proposed scheme casts the PRHD problem into the MIL paradigm by considering protein sequences as bags of N-grams, i.e. short fragments of the sequence. A dissimilarity-based approach is then used to face the MIL problem, based on the matrix of pairwise distances of fragments of a given protein and fragments of a set of prototypes. An empirical evaluation on standard datasets confirms the suitability of the proposed framework. Future directions include analysis of richer dissimilaritites as well as the selection of biologically relevant prototypes (e.g. binding sites).

References

1. Chen, J., Guo, M., Wang, X., Liu, B.: A comprehensive review and comparison of different computational methods for protein remote homology detection. Brief. Bioinf. **19**, 1–14 (2016)
2. Chen, Y., Bi, J., Wang, J.Z.: MILES: multiple-instance learning via embedded instance selection. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 1931–1947 (2006)
3. Cheplygina, V., Tax, D., Loog, M.: Dissimilarity-based ensembles for multiple instance learning. IEEE Trans. Neural Netw. Learn. Syst. **27**(6), 1379–1391 (2016)
4. Cucci, A., Lovato, P., Bicego, M.: Enriched bag of words for protein remote homology detection. In: Robles-Kelly, A., Loog, M., Biggio, B., Escolano, F., Wilson, R. (eds.) S+SSPR 2016. LNCS, vol. 10029, pp. 463–473. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49055-7_41
5. Dietterich, T., Lathrop, R., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1–2), 31–71 (1997)
6. Dong, Q., Lin, L., Wang, X.: Protein remote homology detection based on binary profiles. In: Hochreiter, S., Wagner, R. (eds.) BIRD 2007. LNCS, vol. 4414, pp. 212–223. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71233-6_17
7. Dong, Q., Wang, X., Lin, L.: Application of latent semantic analysis to protein remote homology detection. Bioinformatics **22**(3), 285–290 (2006)
8. Fung, G., Dundar, M., Krishnapuram, B., Rao, R.: Multiple instance learning for computer aided diagnosis. Proc. Adv. Neural Inf. Process. Syst. **19**, 425–432 (2007)
9. Gribskov, M., Robinson, N.: Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput. Chem. **20**(1), 25–33 (1996)
10. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
11. Kuang, R., Wang, K., Wang, K., Siddiqi, M., Freund, Y., Leslie, C.: Profile-based string kernels for remote homology detection and motif extraction. J. Bioinf. Comput. Biol. **3**(03), 527–550 (2005)
12. Kuksa, P.P., Pavlovic, V.: Efficient evaluation of large sequence kernels. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 759–767. ACM (2012)
13. Leslie, C., Eskin, E., Noble, W.: The spectrum kernel: a string kernel for SVM protein classification. In: PSB, pp. 566–575 (2002)

14. Liao, L., Noble, W.: Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J. Comput. Biol. **10**(6), 857–868 (2003)
15. Liu, B., Wang, X., Lin, L., Dong, Q., Wang, X.: A discriminative method for protein remote homology detection and fold recognition combining top-n-grams and latent semantic analysis. BMC Bioinf. **9**(1), 510 (2008). https://doi.org/10.1186/1471-2105-9-510
16. Liu, B., et al.: Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics **30**(4), 472–479 (2014)
17. Lovato, P., Cristani, M., Bicego, M.: Soft Ngram representation and modeling for protein remote homology detection. IEEE/ACM Trans. Comput. Biol. Bioinf. **14**(6), 1482–1488 (2017)
18. Lovato, P., Giorgetti, A., Bicego, M.: A multimodal approach for protein remote homology detection. IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB) **12**(5), 1193–1198 (2015)
19. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition: Foundations and Applications, Machine Perception and Artificial Intelligence, vol. 64. World Scientific, Singapore (2005)
20. Rangwala, H., Karypis, G.: Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics **21**(23), 4239–4247 (2005)