Enriched Bag of Words for Protein Remote Homology Detection

Andrea Cucci, Pietro Lovato, and Manuele Bicego(✉)

Dipartimento di Informatica - Ca' Vignal 2, Università degli Studi di Verona,
Strada le Grazie 15, 37134 Verona, Italy
manuele.bicego@univr.it

Abstract. One of the most challenging Pattern Recognition problems in Bioinformatics is to detect if two proteins that show very low sequence similarity are functionally or structurally related – this is the so-called Protein Remote Homology Detection (PRHD) problem. Even if in this context approaches based on the "Bag of Words" (BoW) paradigm showed high potential, there is still room for further refinements, especially by considering the peculiar application context. In this paper we proposed a modified BoW representation for PRHD, which enriches the classic BoW with information derived from the evolutionary history of mutations each protein is subjected to. An experimental comparison on a standard benchmark demonstrates the feasibility of the proposed technique.

Keywords: Bag of words · N-grams · Sequence classification

1 Introduction

In recent years, several Pattern Recognition problems have been successfully faced by approaches based on the "Bag of Words" (BoW) representation [21]. This representation is particularly appropriate when the pattern is characterized (or assumed to be characterized) by the repetition of basic, "constituting" elements called words. By assuming that all possible words are stored in a dictionary, the BoW vector for one particular object is obtained by *counting* the number of times each element of the dictionary occurs in the object. One of the main advantages of this representation is that it can represent in a vector space many types of objects, even ones that are non-vectorial in nature (like documents, strings, sequences), for which less computational tools are available. The success of this paradigm has been demonstrated in many fields [2–4,21]: in particular, in the bioinformatics context, different BoW approaches [6,14–16] have been proposed in recent years – with the name of N-gram methods – to face the so-called Protein Remote Homology Detection (PRHD) problem [1,10,12]. This represents a central problem in bioinformatics aimed 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

A. Robles-Kelly et al. (Eds.): S+SSPR 2016, LNCS 10029, pp. 463–473, 2016.
DOI: 10.1007/978-3-319-49055-7_41

homologous proteins exhibit very low sequence similarity. In this context, the BoW paradigm is instantiated by considering as words the so-called N-grams, i.e. short sequences of aminoacids of fixed length (N), extracted from the aminoacidic sequence – in the basic formulation [6] – or even from evolutionary representations, i.e. the profiles [14,15].

In this context, approaches based on the BoW representation achieved state of the art prediction performances. Yet, the potentialities of this representation have not been completely exploited, but can be enriched by using some peculiarities of the specific application scenario. In particular, to solve the PRHD task it is needed to capture the homology between proteins, linked to evolutionary aspects, such as insertions, deletions and mutations incurred between the two sequences. Let us concentrate to this last operation, which represents the case when an aminoacid in the sequence is substituted with another aminoacid during evolution. Biologically, there are mutations which are very likely to happen (due to the similar chemical-physical characteristics of the aminoacids), whereas some others are less likely. A good representation for PRHD should capture this aspect; the BoW approach, in its original formulation for PRHD, does not permit to model this aspect[1]: if there is a mutation, we simply count for a *different* word, independently from the fact that the mutation is highly probable or not to happen in nature. However, the BoW paradigm can be extended to cope with this aspect, and this represents the main goal of this paper. More in detail, here we propose a BoW approach to PRHD which modifies the process of counting words, in order to take into account the evolutionary relations between words. The idea is straightforward: in the classical setting, when we observe a word w, we increment its counter by 1. Here we propose to extend this process and to increment also the counters of words which are "biologically likely" mutations of the word w. More specifically, we propose to increment the counter of all other words w' by a value which is directly proportional to the probability of mutation of w in w'. This information is estimated from the so-called substitution matrices (the most famous example being the BLOSUM [9]), employed in sequence-alignment approaches, which quantitatively measure how likely it is, in nature, to observe particular mutations. In this sense, the BoW vector is *enriched* by evolutionary information derived from the specific application scenario.

The proposed approach has been thoroughly evaluated using the standard SCOP[2] 1.53 superfamily benchmark [12], representing the most widely employed dataset to test PRHD approaches. Obtained results demonstrate that the proposed approach reaches satisfactory results in relation to other N-gram based techniques, as well as in comparison to a broader spectrum of approaches proposed in the recent literature.

The rest of the paper is organized as follows: in Sect. 2 we summarize the classic Bag of Words approaches for Protein Remote Homology Detection, whereas in Sect. 3 we present the proposed approach. The experimental

[1] Actually, in computer vision, some approaches dealing with weights have been proposed – e.g. see [17].

[2] http://scop.berkeley.edu/ [7].

evaluation is described in Sect. 4; finally, in Sect. 5, conclusions are drawn and future perspectives are envisaged.

2 BoW Approaches for PRHD

In this section we summarize how a BoW representation can be extracted from a biological sequence – this scheme being at the basis of different PRHD systems [6,14–16]. First, we introduce how "words" and "dictionary" are defined in this context. We consider as words *sequence N-grams*: a N-gram of a sequence $S = s_1 \dots s_L$ is defined as a subsequence of N consecutive symbols $g_l = s_l \dots s_{l+N-1}$. Once fixed the length N, we can define a dictionary $\mathbb{D}$ as the set of all possible subsequences of length N built using the alphabet $\mathcal{A}$ (the four symbols A, T, C, G in case of nucleotides, or 20 symbols in case of aminoacids). Therefore the dictionary $\mathbb{D}$ contains $W = \mathcal{A}^N$ words.

Given a sequence S, its Bag of Words representation $BoW(S)$ is obtained by counting how many times each word (N-gram) $v_i \in \mathbb{D}$ occurs in S. Let us introduce more formally the counting process, mainly to fix the notation used to present the proposed approach. In the first step all the N-grams $g_1, ...g_G$ present in the sequence S are extracted (where G depends on the length L of the sequence and on the degree of overlap with which the N-grams are extracted from the sequence). Then, each g_i is represented via a vector $\mathbf{w}_i$,

$$g_i \longrightarrow \mathbf{w}_i = [0, 0, \cdots, 1, \cdots 0] \tag{1}$$

This W-dimensional vector encodes the fact that g_i corresponds to the j-th word v_j of the dictionary $\mathbb{D}$ via the "1-of-W" scheme: in the $\mathbf{w}_i$ vector all the elements are zero, except one, which is 1; the position of the non zero element is the position in the dictionary $\mathbb{D}$ of the N-gram g_i extracted from the sequence. Given such representation, the Bag of Words representation of S is obtained by summing element-wise all the vectors $\mathbf{w}_1, ..., \mathbf{w}_G$:

$$BoW(S) = \mathbf{w}_1 + \mathbf{w}_2 + \cdots + \mathbf{w}_G \tag{2}$$

See the left part of Fig. 1 for a schematic sketch of the BoW scheme.

This representation has been successfully employed in the case of Protein Remote Homology Detection, typically as direct input to discriminative classifier such as Support Vector Machines [14,15], or after the employment of more sophisticated models, such as topic models [16]. In all these approaches, the BoW representation has been extracted from different kinds of biological sequences: raw sequences (as in [6]), evolutionary representations of the biological sequences – called *profiles* (as in [14,15]), or even in combination with the corresponding 3D structures (as in [16]).

3 The Proposed Approach

The main idea behind the proposed approach stems from the observation that the classic Bag of Words scheme for Protein Remote Homology Detection is

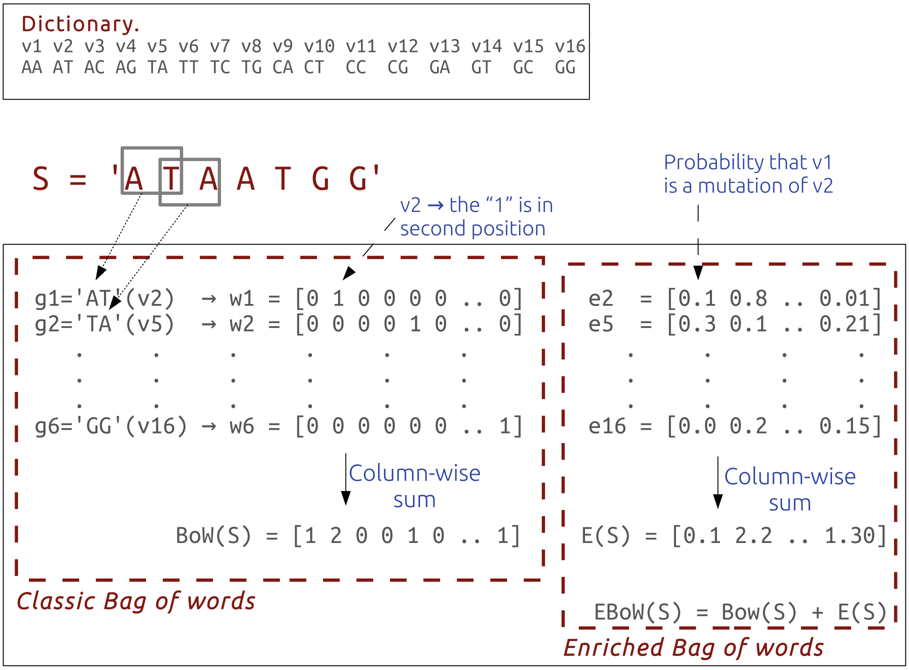
Dictionary.
v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v14 v15 v16
AA AT AC AG TA TT TC TG CA CT CC CG GA GT GC GG
S = 'A T A A T G G'
v2 → the "1" is in second position
Probability that v1 is a mutation of v2
g1='AT'(v2) → w1 = [0 1 0 0 0 0 .. 0]
g2='TA'(v5) → w2 = [0 0 0 0 1 0 .. 0]
g6='GG'(v16) → w6 = [0 0 0 0 0 0 .. 1]
Column-wise sum
BoW(S) = [1 2 0 0 1 0 .. 1]
Classic Bag of words
e2 = [0.1 0.8 .. 0.01]
e5 = [0.3 0.1 .. 0.21]
e16 = [0.0 0.2 .. 0.15]
Column-wise sum
E(S) = [0.1 2.2 .. 1.30]
EBoW(S) = Bow(S) + E(S)
Enriched Bag of words

Fig. 1. Sketch of the Bag of Words representation and the proposed Enriched Bag of Words approach. We are considering nucleotidic sequences (therefore the alphabet contains the symbols A,T,C,G); in the specific case, the sequence length is 7 ($L = 7$), and we used N-grams of length 2, with overlap = 1. This means that the number of N-grams extracted from the sequence is 6 ($G = 6$).

not able to encode evolutionary relations which can exist between words: if a substitution occurs (i.e. an aminoacid is replaced by another during evolution), the classic BoW simply counts for another word in the dictionary, independently from how likely is this substitution: actually, in nature, there are definitely different probabilities of mutation between aminoacids (which compose the words in the BoW), which depend on the family, the chemical properties or the structural features. To cope with this aspect, we propose a scheme based on the following idea: if an N-gram g_i in the sequence corresponds to the word v_j, we increment the count of v_j by 1 (as in the classic BoW), but we also increment the counters of the words which are "biologically likely" mutations of the word v_j: such increments are clearly directly proportional to the probability of being mutation of v_j.

More formally, a given N-gram g_i, corresponding to the j-th word of the dictionary, is represented by $\mathbf{w}'_i$, defined as:

$$g_i \longrightarrow \mathbf{w}' = \mathbf{w}_i + \mathbf{e}_j \tag{3}$$

where $\mathbf{w}_i$ is defined as in Eq. (1) (all zeros and a 1 in position j), and

$$\mathbf{e}_j = [e_{j1}, e_{j2}, \cdots, e_{jW}] \tag{4}$$

is the *enrichment vector*, a vector of length W which, in every position k, indicates how much probable is that the k-th word of the dictionary is a mutation of v_j. This vector permits to explicitly encode the biological a priori knowledge on the relation which occurs between the words of the dictionary.

Given this correction, the *Enriched* Bag of Words representation is obtained by following the same scheme of Eq. (2), i.e. by the columwise summation of all $\mathbf{w}'_i$:

$$EBoW(S) = \mathbf{w}'_1 + \mathbf{w}'_2 + \cdots + \mathbf{w}'_G \tag{5}$$

Rearranging the summation, we obtain:

$$EBoW(S) = BoW(S) + E(S) \tag{6}$$

where $E(S)$ represents the “enrichment” (or correction) made to the Bag of Words representation of sequence S.

3.1 Computing the Enrichment Vectors

The enrichment vectors are obtained by starting from the so-called substitution matrices (the most famous one called BLOSUM [9]), matrices which encode the biological knowledge related to mutations. This matrix is typically used to perform sequence alignments, and a given entry (i, j) encodes the rate at which the aminoacid i is likely to mutate into the aminoacid j (the higher, the more likely it is). Intuitively, this matrix has the highest values on the diagonal; for off-diagonal elements, the matrix should reflect the fact that there are some mutations that are highly improbable, due for example to physical or chemical properties of aminoacids.

To define our enrichment vectors, we start from the approach used to derive the BLOSUM matrix, proposed in [9]. In that paper, the matrix has been built by starting from *blocks* of related sequences[3]. From these blocks (more than 2000 blocks have been used in the original paper of [9]), the *expected probability of occurrence* of each pair of symbols can be computed, and this represents our starting information. For example, in the case of the nucleotidic sequences (i.e. the alphabet is composed by ‘A’, ‘T’, ‘C’, ‘G’), this matrix is

$$\mathbf{M}_1 = \begin{bmatrix} P(A \rightarrow A) & P(A \rightarrow T) & P(A \rightarrow C) & P(A \rightarrow G) \\ P(T \rightarrow A) & \cdots & \cdots & P(T \rightarrow G) \\ & & \ddots & \\ P(G \rightarrow A) & \cdots & \cdots & P(G \rightarrow G) \end{bmatrix}$$

[3] In biological terms, related sequences are the ones which belong to the same evolutionary family – namely, they share the same biological function.

where $P(x \rightarrow y)$ indicates the probability that the nucleotide 'x' mutates into the nucleotide 'y'. If we use as words 1-grams, namely the entries of the alphabet $\mathcal{A}$, then we can directly employ this matrix to derive the required enrichment vectors. For $N > 1$, however, we should provide a larger matrix, containing the mutation probability for each pair of N-grams. Here we define this probability via the multiplication of the probabilities of pairs of symbols; for example, in the case of 2-grams, we have

$$P(xy \rightarrow kj) = P(x \rightarrow k)P(y \rightarrow j)$$

We are aware that by employing this simple scheme we are assuming that the symbols inside the N-gram are probabilistically independent: however this simplifying assumption is accepted and employed in many applications dealing with biological sequences – e.g. for multiple sequence alignment [18]. In formula, the mutation matrix $\mathbf{M}_N$ for N-grams of length N is obtained inductively by employing the Kronecker tensor product "$\otimes$":

$$\mathbf{M}_N = \mathbf{M}_{N-1} \otimes \mathbf{M}_1 \tag{7}$$

Finally, the matrix of enrichment vectors for N-grams of length N $\mathbf{E}_N = [\mathbf{e}_1; \mathbf{e}_2; ...; \mathbf{e}_W]$ is obtained by normalizing the mutation matrix in order to have a reasonable range.

$$\mathbf{E}_N = \frac{\mathbf{M}_N}{\max_{i,j} \mathbf{M}_N} \tag{8}$$

4 Experimental Evaluation

The experimental evaluation is based on a famous benchmark[4] widely employed to assess the detection capabilities of many protein remote homology detection systems [12], extracted from SCOP version 1.53 and containing 4352 sequences from 54 different families. The protein remote homology detection task is cast into a binary classification problem: to simulate remote homology, 54 different subsets are created: in each of this, an entire target family is left out as positive testing set. Positive training sequences are selected from other families belonging to the same superfamily (i.e. sharing remote homology), whereas negative examples are taken from other super-families. Please note that class labels are very unbalanced, with a vast majority of objects belonging to the negative class (on average the positive class (train + test) is composed by 49 sequences, whereas the negative one is made by 4267).

As in many previous works [5,6,13–15,19], classification is performed using SVM via the public GIST implementation[5], setting the kernel type to radial basis, and keeping the remaining parameters to their default values. Detection accuracies are measured using the receiver operating characteristic (ROC) score and the ROC50 score [8]. In both cases, the larger the value the better the

[4] Available at http://noble.gs.washington.edu/proj/svm-pairwise/.

[5] Downloadable from http://www.chibi.ubc.ca/gist/ [12].

detection. In particular, the former represents the usual area under the ROC curve, whereas the latter measures the area under the ROC curve 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.

4.1 Results and Discussion

The proposed approach has been compared with the corresponding classic Bag of Words representation in different experimental conditions. In particular, we tested the improvement obtained by the enrichment on BoW representations defined from the raw sequence and from its evolutionary representation (the profile), using different N-grams (1-gram, 2-gram, 3-gram). For what concerns the proposed method, we employed different variants of the BLOSUM matrices. Roughly speaking, a different number after the name "BLOSUM" indicates a more or less strict definition of "similar sequences" in the construction of blocks (see Sect. 3.1).

ROC and ROC50 scores, averaged over all the families of the dataset, are shown in Table 1. To assess statistical significance of our results and demonstrate

Table 1. ROC (top) and ROC50 (bottom) scores. "EBoWX" indicates that the enrichment vectors have been obtained by using the BLOSUMX matrix. In bold we put results for which the p-value of the statistical test is less than 0.05.

Sequence based

	1-grams	2-grams	3-grams
BoW	0.8601	0.8709	0.8117
EBoW-45	0.8644	**0.8998**	**0.9131**
EBoW-50	0.8638	**0.8996**	**0.9139**
EBoW-62	0.8647	**0.8990**	**0.9114**
EBoW-80	0.8653	**0.8968**	**0.9061**
EBoW-90	0.8652	**0.8950**	**0.9016**

Profile based

	1-grams	2-grams	3-grams
BoW	0.9070	0.9290	0.8876
EBoW-45	0.9054	**0.9458**	**0.9494**
EBoW-50	0.9048	**0.9453**	**0.9494**
EBoW-62	0.9042	**0.9453**	**0.9466**
EBoW-80	0.9046	**0.9440**	**0.9413**
EBoW-90	0.9051	**0.9427**	**0.9384**

(ROC)

Sequence based

	1-grams	2-grams	3-grams
BoW	0.6054	0.6331	0.5848
EBoW-45	0.6256	**0.6925**	**0.7175**
EBoW-50	0.6270	**0.6888**	**0.7216**
EBoW-62	0.6274	0.6763	**0.7052**
EBoW-80	0.6325	**0.6894**	**0.6776**
EBoW-90	0.6301	**0.6886**	**0.6737**

Profile based

	1-grams	2-grams	3-grams
BoW	0.6928	0.7741	0.7220
EBoW-45	**0.6552**	**0.7914**	0.7832
EBoW-50	0.6670	**0.7830**	**0.7944**
EBoW-62	0.6719	0.7863	0.7931
EBoW-80	0.6826	0.7829	**0.8007**
EBoW-90	0.6731	0.7774	**0.8151**

(ROC50)

that increments in ROC/ROC50 scores gained with the proposed approach are not due to mere chance, we performed a Wilcoxon signed-rank test, reporting in the tables in bold the results for which the corresponding p-value is less than 0.05 (i.e. bold numbers indicate a statistically significant difference). From the tables different observations can be derived. In general, it can be seen that the proposed enrichment is almost always beneficial for 2-grams and 3-grams, with some really important improvements – for example, with 3-grams and BoW based on sequences, the ROC (ROC50) score improves from 0.81 to 0.91 (from 0.58 to 0.72), this representing a remarkable result. This is more evident by looking at the ROC scores. For what concerns the different BLOSUM employed, no differences can be observed in the ROC scores; however, considering the ROC50 scores, it seems evident that this choice has an impact. Unfortunately, a general rule can not be derived: for some configurations a stricter BLOSUM is better, for others the other way around holds. In general, we can say that for 2-grams and 3-grams there is always a configuration for which a statistically significant improvement can be obtained.

For 1-grams such improvement is not so evident (ROC50 results also highlight one case when the biological enrichment results in a worst performance). For what concerns 2- and 3-grams, it seems evident that the proposed enrichment permits to derive a better representation for classification. We think that this is due to a twofold beneficial effect that the approach produces on the representation: from one hand, we are injecting useful information which permits to recover from the uncertainty present in the counting process – a peculiarity of this application. From the other hand, the proposed approach permits to reduce the huge sparsity of the Bag of Words vectors within this application. In fact, within the SCOP datasets the sequences have an average length of 200, thus resulting in around 200 N-grams (if we consider the maximum possible overlap); when using 3-grams, the Bag of Words vector has 8000 entries (the size of the dictionary, 20^3) to be filled with around 200 ones; this implies that most of the entries are zero (this problem is less severe with 2-grams). Even if good classification methods able to deal with sparse representations exist, in this specific case a SVM with the rbf kernel has been used, for fair comparison with state of the art, thus this sparsity problem may have an impact. To provide some empirical support to our intuition, we performed two experiments, focusing on BoW computed from profiles. In the first, we select as Enrichment Matrix a random probability matrix – this solution would in principle alleviate the sparsity problem, but it does not injects any evolutionary information. In the second, we removed from the Enriched BoW low values so that the number of zero-value entries is the same as in the standard Bag of Words representation – this solution only injects evolutionary information, without solving the sparsity problem. ROC values are shown in Table 2: the accuracies obtained in the 3-grams case suggest that there is a beneficial effect both in only the reduction of the sparsity (BoW plus random Enrichment) and in the truncated injection of relevant information (Truncated Effect): however the proposed approach, which combines both effect, obtain the best effect. This is not so evident by looking at the results with 2-grams, where only the complete approach permits to improve the accuracies.

Table 2. Properties of the Enriched BoW representation in the PRHD.

	Classic BoW	BoW + random **E**	Truncated EBoW	Proposed EBoW
2-grams	0.9290	0.9086	0.9213	0.9458
3-grams	0.8876	0.9140	0.9042	0.9494

As a final analysis, we reported in Table 3 some comparative results with other approaches of the literature applied to the SCOP 1.53 benchmark. When compared to other techniques that are based on Bag of Words, the proposed approach behaves very well, outperforming all the alternative techniques; looking at the global picture, the table shows very promising results, also in comparison with other approaches, where satisfactory performances are reached both using the ROC and the ROC50 evaluation measures. Please note that the results can be further improved, for example by deriving the enrichment vectors from more recent and accurate substitution matrices or by tuning the impact of the enrichment (for example by putting a weight α in Eq. (3)).

Table 3. Average ROC scores for the 54 families in the SCOP 1.53 superfamily benchmark for different methods.

Method	ROC	ROC50	Reference
Enriched BoW (3-gram)	0.949	0.815	This paper
Bag of words based methods			
SVM-N-gram	0.826	0.589	[6]
SVM-N-gram-LSA	0.878	0.628	[6]
SVM-Top-N-gram (n = 1)	0.907	0.696	[14]
SVM-Top-N-gram (n = 2)	0.923	0.713	[14]
SVM-Top-N-gram-combine	0.933	0.767	[14]
SVM-N-gram-p1	0.887	0.726	[15]
SVM-N-gram-KTA	0.892	0.731	[15]
Other methods			
SVM-pairwise	0.908	0.787	[15]
SVM-LA	0.925	0.752	[20]
Profile *(5,7.5)*	0.980	0.794	[11]
SVM-Pattern-LSA	0.879	0.626	[6]
SVM-Motif-LSA	0.860	0.628	[6]
PSI-BLAST	0.676	0.330	[5]
SVM-Bprofile-LSA	0.921	0.698	[5]
SVM-PDT-profile ($\beta = 8$, n = 2)	0.950	0.740	[13]
SVM-LA-p1	0.958	0.888	[15]

5 Conclusions

In this paper we proposed an enriched BoW representation for Protein Remote Homology Detection, which injects evolutionary information into the counting process, thus resulting in a richer and biologically relevant representation. The proposed scheme has been tested on a standard benchmark, obtaining very promising results. As a future work, we plan to investigate the suitability of the proposed scheme in other domains, such as text processing. Clearly, in this latter case, the main challenge is to define how words are related through similarities in meaning.

References

1. Altschul, S.F., Madden, T.L., Schffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res. **25**(17), 3389–3402 (1997)
2. Bicego, M., Lovato, P., Perina, A., Fasoli, M., Delledonne, M., Pezzotti, M., Polverari, A., Murino, V.: Investigating topic models' capabilities in expression microarray data classification. IEEE/ACM Trans. Comput. Biol. Bioinform. **9**(6), 1831–1836 (2012)
3. Brelstaff, G., Bicego, M., Culeddu, N., Chessa, M.: Bag of peaks: interpretation of nmr spectrometry. Bioinformatics **25**(2), 258–264 (2009)
4. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
5. 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). doi:10.1007/978-3-540-71233-6_17
6. Dong, Q., Wang, X., Lin, L.: Application of latent semantic analysis to protein remote homology detection. Bioinformatics **22**(3), 285–290 (2006)
7. Fox, N.K., Brenner, S.E., Chandonia, J.: SCOPe: structural classification of proteins - extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. **42**(Database–Issue), 304–309 (2014)
8. Gribskov, M., Robinson, N.L.: Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput. Chem. **20**(1), 25–33 (1996)
9. Henikoff, S., Henikoff, J.: Amino acid substitution matrices from protein blocks. PNAS **89**(22), 10915–10919 (1992)
10. Karplus, K., Barrett, C., Hughey, R.: Hidden Markov models for detecting remote protein homologies. Bioinformatics **14**, 846–856 (1998)
11. Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., Leslie, C.: Profile-based string kernels for remote homology detection and motif extraction. J. Bioinform. Comput. Biol. **3**(03), 527–550 (2005)
12. Liao, L., Noble, W.S.: Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J. Comput. Biol. **10**(6), 857–868 (2003)
13. Liu, B., Wang, X., Chen, Q., Dong, Q., Lan, X.: Using amino acid physicochemical distance transformation for fast protein remote homology detection. PLoS ONE **7**(9), e46633 (2012)

14. 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)
15. Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., Dong, Q., Chou, K.C.: Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics **30**(4), 472–479 (2014)
16. Lovato, P., Giorgetti, A., Bicego, M.: A multimodal approach for protein remote homology detection. IEEE/ACM Trans. Comput. Biol. Bioinform. **12**(5), 1193–1198 (2015)
17. Marszaek, M., Schmid, C.: Spatial weighting for bag-of-features. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2118–2125 (2006)
18. Pevsner, J.: Bioinformatics and Functional Genomics. Wiley, Hoboken (2003)
19. Rangwala, H., Karypis, G.: Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics **21**(23), 4239–4247 (2005)
20. Saigo, H., Vert, J.P., Ueda, N., Akutsu, T.: Protein homology detection using string alignment kernels. Bioinformatics **20**(11), 1682–1689 (2004)
21. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill Inc., New York (1986)