HAL
open science

An Overview of in Silico Methods for the Prediction of Ionizing Radiation Resistance in Bacteria

Manel Zoghlami, Sabeur Aridhi, Mondher Maddouri, Engelbert Mephu Nguifo

To cite this version:

Manel Zoghlami, Sabeur Aridhi, Mondher Maddouri, Engelbert Mephu Nguifo. An Overview of in Silico Methods for the Prediction of Ionizing Radiation Resistance in Bacteria. Tamar Reeve. Ionizing Radiation: Advances in Research and Applications, Nova science publishers, pp.241-256, 2018, Physics Research and Technology Series, 978-1-53613-539-8. hal-01807944

HAL Id: hal-01807944

https://inria.hal.science/hal-01807944v1

Submitted on 20 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Chapter

AN OVERVIEW OF *in Silico* METHODS FOR THE PREDICTION OF IONIZING RADIATION RESISTANCE IN BACTERIA

Manel Zoghlami[1,2,*]***, Sabeur Aridhi***[3,]
Mondher Maddouri[4] ***and Engelbert Mephu Nguifo***[1]
[1]Clermont Auvergne University, LIMOS, Clermont-Ferrand, France
[2]University of Tunis El Manar, LIPAH, Tunis, Tunisia
[3]University of Lorraine, LORIA, Nancy, France
[4]University of Jeddah, College Of Business, Jeddah,
Kingdom of Saudi Arabia

Abstract

Ionizing-radiation-resistant bacteria (IRRB) could be used for bioremediation of radioactive wastes and in the therapeutic industry. Limited computational works are available for the prediction of bacterial ionizing radiation resistance (IRR). In this chapter, we present some works that study the causes of the high resistance of IRRB to ionizing radiation. Then we focus on presenting *in silico* approaches that use protein sequences of bacteria in order to predict if an unknown bacterium belongs to IRRB or ionizing-radiation-sensitive bacteria (IRSB). These approaches formulate the problem of predicting bacterial IRR as a multiple instance learning (MIL) problem where bacteria represent the bags and

*Corresponding Author: manel.zoghlami@gmail.com.

primary structure of basal DNA repair proteins of each bacterium represent the instances inside the bags. We also present a formulation of the problem of MIL in sequence data and explain how it could be used to solve the problem of IRR prediction in bacteria. A brief comparison of the presented approaches is provided.

Keywords: bacterial ionizing radiation resistance, multiple instance learning, phenotype prediction

1. Introduction

IRRB could be used for the treatment of mixed radioactive wastes by developing a strain to detoxify both mercury and toluene [1]. These organisms are also being engineered for *in situ* bioremediation of radioactive wastes[2]. In [3], the authors discuss the potential uses of radiation-resistant extremophiles (e.g. micro-organisms with the ability to survive in extreme environmental conditions) in biotechnology and the therapeutic industry. The major challenges of therapeutic development using extremophiles are discussed in [4].

Several *in vitro* and *in silico* works studied the causes of the high resistance of IRRB to ionizing radiation to determine peculiar features in their genomes and improve the treatment of radioactive wastes [5]. However, limited computational works are provided for the prediction of bacterial IRR [6][7]. In this work, we present some works that study the causes of the high resistance of IRRB to ionizing radiation. Then we focus on machine learning approaches that use protein sequences of bacteria to predict whether a bacterium belongs to IRRB or IRSB. These approaches formulate the problem of predicting bacterial IRR as an MIL problem where bacteria represent bags and repair proteins of each bacterium represent instances. MIL methods are a variation of machine learning methods that can be used to solve problems in which the labels are assigned to bags, i.e., a set of instances, rather than individual instances.

The remainder of this chapter is organized as follows. In Section 2, we present some computational works that aim to analyze IRRB in order to find out the causes of their resistance to ionizing radiation. In Section 3, we provide a formalization of the problem of MIL and explain how it could be used to solve the IRR prediction problem. Then we present some existing *in silico* approaches for IRR prediction in bacteria. Section 4 provides a brief comparison of the presented approaches. Concluding points make the body of Section 5.

2. Computational Works for Analyzing IRRB

Several *in vitro* and *in silico* works are proposed in order to study the causes of the high resistance of IRRB to ionizing radiation. In fact, determining the characteristics of IRRB could help to predict if a bacterium belongs to IRRB or IRSB. In this chapter, we present some of these works and show the pipeline of existent bioinformatics programs and the statistical methods they use.

In [6], the authors analyze four known genome sequences of IRRB in order to find out the role of positive Darwinian selection in the evolution of IRR and the tolerance of desiccation. The used pipeline contains three steps:
Step 1. Finding orthologous relationships using MultiParanoid [8].
Step 2. Aligning the sequences for each ortholog set using CLUSTAL W [9].
Step 3. Testing for positive selection using the DnaSP program [10] that requires an aligned set of orthologous sequences.
This work concludes that all basal DNA repair genes in IRRB are subject to positive selection unlike many of their orthologs in IRSB.

In [11], the authors make a comparative analysis of codon and amino acid usage patterns. The study uses 19 genomes of the phylum *Deinococcus-Thermus* and follows a six-step pipeline.
Step 1. The authors make correspondence analysis on relative synonymous codon usage and amino acid usage using the program CodonW [12]. Then they use SPSS software [13] to make correlation and variance analysis.
Step 2. In order to find out the variation in amino acid usage between radiation-sensitive and radiation-resistant genome of the studied bacteria, the STATISTICA software [14] is used and a cluster analysis on amino acid usage is performed.
Step 3. To find out a cluster of orthologous proteins, the authors use the CMG-Biotools workbench [15] and a BLAST matrix [16].
Step 4. Synonymous and non-synonymous substitution patterns in previously selected orthologous proteins are then estimated. The PAML software [17] is used in this step.
Step 5. Several indices are calculated in order to find out the factors influencing codon and amino acid usage including aromaticity, average hydrophobicity, isoelectric point and instability index.
Step 6. Finally, a COG [18] functional classification is performed in order to compare the proteins. The IMG/M system [19] is used in this step as a support for comparative analysis of metagenomes.

This study reports significant differences in synonymous codon usage bias and amino acid usage patterns between the radiation-sensitive and radiation-resistant genomes of the studied dataset.

In a recent work [20], the authors identify general patterns of microbial responses to multiple stressors in radioactive environments. They analyze three datasets including a set of bacteria isolated from soil contaminated by nuclear waste at the Hanford site (USA) [21]. The other two datasets are out of the scope of this chapter since they concern fungi and yeasts. In order to develop a filter procedure for identifying important predictor variables, machine learning and information theoretic approaches are used. The pipeline contains three main steps.

Step 1. The authors start by testing the hypothesis which supposes that the collected contaminated soil samples are statistically dependent (samples separated by 1 meter of soil depth would contain the same bacterial taxa than samples separated by 10 meters). To do so, they use Mantel tests [22] for spatial autocorrelation using the ade4 R package [23].

Step 2. In order to find out the variables which are the least important for describing the data, authors use two machine learning prediction methods: random forests [24] and random generalized linear modeling [25]. The variables that are identified as unimportant by both machine learning methods are most likely to be noise variables and are not used in the rest of the study.

Step 3. All variables are ranked in order of importance and filtered based on information theoretic approaches [26] [27]. A new model is created using the retained variables.

3. Multiple Instance Learning Approaches for Bacterial IRR Prediction

As far as we know, only three *in silico* approaches were proposed in order to predict the bacterial IRR based on a set of protein sequences. These approaches, named ABClass, ABSim and MIL naive approach [28], are three machine learning algorithms that require as input a set of proteins sequences of each bacterium and provide as output a label response: IRRB if the query bacterium is radioresistant and IRSB if not. Since these approaches are based on a MIL formalisation, we present in this section the MIL problem formulation and we highlight some well known MIL algorithms. Then, we provide a description of the above

cited approaches.

3.1. Problem Formulation

We denote Σ an *alphabet* defined as a finite set of characters or symbols. A *symbolic sequence* is defined as an ordered list of symbols [29]. The primary structure of a protein is a symbolic sequence since it is described using symbols (amino acids). Let DB be a learning database that contains a set of n labeled bags $DB = \{(B_i, Y_i), i = 1, 2 \ldots, n\}$ where $Y_i = \{-1, 1\}$ is the label of the bag B_i. Instances in B_i are sequences and are denoted by B_{ij}. Formally $B_i = \{B_{ij}, j = 1, 2 \ldots, m\}$, where m is the total number of instances in this bag. We note that according to the problem investigated in this chapter, there is a relationship between instances of different bags since there are orthologous proteins in the different bags. This relation is denoted *the across bag sequences relation* in [28]. The goal is to learn a multiple instance classifier from DB. Given a query bag $Q = \{Q_k, k = 1, 2 \ldots, q\}$, where q is the total number of instances in Q, the classifier should use sequential data in this bag and in each bag of DB in order to predict the label of Q.

3.2. Multiple Instance Learning: An Overview

In a traditional setting of supervised learning task, the training set is composed of feature vectors (instances), where each feature vector has a label. In MIL task, we learn a classifier based on a training set of bags, where each bag contains multiple feature vectors and it is the bag that carries a label [30]. We do not know the labels of the instances inside the bags. The task of MIL investigated in this chapter is to learn a classifier from the training set that correctly predicts unseen bags.

Several MIL algorithms have been proposed. A review of MIL approaches with a comparative study could be found in [30] and [31]. A common assumption in MIL field is that a positive bag contains at least one positive instance, while in a negative bag all of the instances are negative. This assumption is called *the standard multiple instance assumption*. Several MIL algorithms adopt this asymmetrical assumption including MI-SVM [32] and Diverse Density [33]. MI-SVM is an adaptation of support vector machines (SVM) to the MIL problem. The key point of Diverse Density algorithm is to find a concept point in the feature space that is close to at least one instance from every positive bag and meanwhile far away from instances in negative bags. Other methods

adopt the so called *collective assumption* which states that all instances in a bag contribute to define its label. This assumption could be suitable to the problem investigated in this chapter since we do not know which set of proteins helps us to classify a bacterium as IRRB. Some algorithms make a positive instance selection. In a recent work [34], the authors present the MILKDE algorithm which identifies the most representative instances in each positive bag based on a likelihood computation.

According to [30], the MIL methods could be categorized according to how the information existent in the MI data is exploited. Three categories could be defined depending on the adopted paradigm:

- Methods following the instance-space paradigm assume that the discriminative information is present at the instance-level. We consider the characteristics of individual instances in the learning process without looking at more global characteristics of the whole bag. The simplest algorithm in this category is the SIL algorithm [35] which trains a standard supervised classifier then simply uses the sum as aggregation rule to obtain the bag-level classifier.

- Methods following the bag-space paradigm treat each bag as a whole entity. A global bag-level information is used to make the discriminative decision instead of aggregating instance-level decisions. A commonly used approach is to define a distance function in order to compare bags. In [36], the authors present an extension of the classical KNN algorithm [37] called the Citation-kNN algorithm. It classifies a bag based on the labels of its neighbors (references and citers).

- Methods following the embedded-space paradigm map each bag to a feature vector which summarizes the relevant information about the whole bag. The bag-level information is extracted explicitly through the definition of a mapping function, while it is done implicitly in the bag-space paradigm. In [38], the authors propose an MIL algorithm that computes the dissimilarities of a bag to other bags in the training set and uses these dissimilarities as a feature representation.

3.3. A Naive MIL Approach for IRR Prediction

In [28], the authors present a naive MIL approach for sequence data with across bag relationships. Since the bacterial IRR prediction problem could be formal-

ized as an MIL problem and there is a relationship between protein sequences across bags (orthologous proteins), then we can apply the naive approach in our case. This approach contains two steps: a preprocessing step and a learning step.

Step 1. The preprocessing step transforms the set of sequences into an attribute-value matrix where each row corresponds to a sequence and each column corresponds to an attribute. When we deal with sequence data, the most used technique to transform data into an attribute-value format is to extract motifs that serve as attributes. We note that finding a uniform description of all instances using a set of motifs is not always an easy task. Since the naive approach takes into account the across bag relationships between instances, the preprocessing step extracts motifs from each set of related instances. The union of these extracted motifs is then used as an attribute set to construct the descriptive matrix. The presence or the absence of a motif in a sequence is respectively denoted by 1 or 0. It is worthwhile to mention that only a subset of the used attributes is representative for each processed sequence. Therefore, we may have a big sparse matrix when trying to present the whole sequence data using an attribute value format.

Step 2. The second step consists in applying an existing MIL classifier.

3.4. *ABClass*: Across Bag Sequences Classification Approach

In order to avoid the use of one large vector of features to describe sequence data, the ABClass approach [28] that takes into account the across bag relationships between instances is proposed. It contains the following steps.

Step 1. A preprocessing step identifies orthologous protein sequences. Ideally, each protein has an orthologous sequence in each bag. This is defined as an across bag dependency. This relationship between sequences of different bags will be used in the learning step. The learning dataset is divided into sets of orthologous sequences. We note that a protein may not have any ortholog in a bag.

Step 2. A set of motifs is extracted from each set of orthologous sequences. In the experimental tests, DMS [39] is used as a motif extraction method. DMS allows building motifs that can discriminate a family of sequences from other ones. It first identifies motifs in the protein sequences. Then it filters them in order to keep only the discriminative and minimal ones.

Step 3. Extracted motifs are used to encode instances in order to create a dis-

criminative model. Each set of related instances (i.e. orthologous proteins) is represented by its own motifs vector. This reduces the number of attributes that are not representative for the processed protein sequence. The WEKA [40] data mining tool is used in order to apply existing well known classifiers to generate models. The following three classifiers have been used: SMO [41] [42], J48 [43] and Naive Bayes [44].
Step 4. To predict the IRR of an unlabeled bacterium, the extracted motifs are used to represent its instances. Each protein sequence is represented using a vector. Then it is compared with the corresponding model (already generated from its orthologous proteins) to provide a partial prediction result.
Step 5. Finally, an aggregation step is applied on the partial results in order to compute the final prediction result.

3.5. *ABSim*: Across Bag Sequences Similarity Approach

In [28], the authors present ABSim, an algorithm that discriminates bags by measuring the similarity between each instance sequence in the query bag and corresponding related sequences in the different bags of the learning database. ABSim is an extension of a previously proposed algorithm named MIL-Align [45]. The difference between these two algorithms is that ABSim is more general since MIL-Align was originally proposed to deal with biological data while ABSim aims to deal with MIL in sequence data in general e.g. textual data. When applied to the problem of bacterial IRR prediction, ABSim uses the local alignment score as similarity measure to compare protein sequences. The algorithm works as follows:
Step 1. For each protein sequence in the query bag, the algorithm computes the corresponding alignment scores between this protein and its orthologs in other bags.
Step 2. Alignment scores are grouped into a matrix. Each line corresponds to a score vector of a protein against all its orthologs in other bacteria.
Step 3. An aggregation method is applied to the matrix in order to compute the final prediction result. Two aggregation methods named SMS and WAMS are proposed by the authors.

3.6. Experimental Results of MIL Approaches

The MIL naive approach, ABClass and ABSim are used in [28] in order to resolve the problem of phenotype prediction of bacterial IRR. Bacteria rep-

resent the bags and primary structure of basal DNA repair proteins represent the sequences. An unknown bacterium is affiliated to either IRRB or IRSB. The used dataset is described in [45]. It consists of 28 bacteria (14 IRRB and 14 IRSB). Each bacterium/bag contains 25 to 31 instances that correspond to proteins implicated in basal DNA repair in IRRB. Proteins of the bacterium *Deinococcus radiodurans* were downloaded from the UniProt website. Proteomes of other bacteria were downloaded from the NCBI FTP website. ABSim and ABClass tools can be downloaded at the following link: http://homepages.loria.fr/SAridhi/software/MIL/.

It is worthwhile to mention that ABClass is tested under several different settings (classifiers and motif extraction settings) and that the obtained accuracy results depend on the used settings. The over all accuracy results reported in [28] are very close but a slightly better accuracy rates are provided either by ABClass or by ABSim according to the used settings. Using both approaches, two bacteria (*M. radiotolerans* and *B. abortus*) generate the lowest rates of successful prediction compared to the rate of the other bacteria. These results may help to understand some characteristics of the studied data. A probable biological explanation is provided in [45] and notes that this could be explained by the increased rate of sequence evolution in endosymbiotic bacteria [46].

4. Comparison

Table 4.1 provides a short comparison of the previously presented approaches. The three works [6] [20] and [11] are interested in analyzing different characteristics of IRRB. The works in [6] and [11] are based on a pipeline that uses existent bioinformatics programs in order to perform their studies. The study in [7] is based on a filter of machine learning and information theoretic approaches. Although these works are important to understand the causes of the high resistance of IRRB to ionizing radiation, they do not provide a tool to predict if an unknown bacterium belongs to IRRB or IRSB.

The naive MIL approach, ABClass and ABSim provide as output a prediction label for an unknown bacterium (IRRB or IRSB). The naive MIL approach uses standard MIL classifiers after making a preprocessing step which extracted motifs and adapt the data to the required format. This preprocessing step is time consuming and could lead to a huge matrix. ABCLass takes advantage of the across bag relationships between sequences in order to reduce the number of attributes that are not representative for each sequence during the encoding

Table 4.1. Comparison of the presented approaches on IRR analyses\prediction in bacteria.

Approach	Task	Method
Sghaier et al., 2008 [6]	Analysis of IRRB	Pipeline
Banerjee et al., 2014 [11]	Analysis of IRRB	Pipeline
Shuryak and Dadachova, 2016 [20]	Analysis of IRRB	Information theory and machine learning
MIL Naive Approach [28]	IRR Prediction	MIL classifiers and motif extraction
ABClass [28]	IRR Prediction	MIL model and optimized use of motifs
ABSim [28]	IRR Prediction	MIL model and alignment score

step. This relationship is also used during the learning step when generating partial models for each set of related sequences. ABSim does not use motifs to represent data since no encoding step is needed. The local alignment score is used to perform the prediction. This makes ABSim faster and easier to use than ABClass unless we already have the representative motifs for each set of orthologous proteins or if we think that the extraction of motifs will not be an expensive task (according to the data size, the used motifs extractor and the extraction settings e.g. required motifs length). As mentioned in the previous section, a slightly better accuracy result could be provided either by ABClass or by ABSim according to the used settings.

5. Conclusion

Prediction of IRR in bacteria is a challenging task. Several works were interested in finding out the causes of the resistance of some bacteria to IRR. These works generally use a pipeline of existing bioinformatics programs. Other works provide machine learning algorithms in order to predict the bacterial IRR. They are based on an MIL formalization. In this chapter, we studied existing works on prediction of IRR in bacteria and we presented a comparison of the studied approaches. Based on our study, we mention that the used settings in the preprocessing step and the learning step influence the choice of the approach to use.

References

[1] Brim H, McFarlan SC, Fredrickson JK, Minton KW, Zhai M, Wackett LP, et al. Engineering Deinococcus radiodurans for metal remediation in radioactive mixed waste environments. *Nature biotechnology.* 2000;18(1):85–90.

[2] Brim H, Venkateswaran A, Kostandarithes HM, Fredrickson JK, Daly MJ. Engineering Deinococcus geothermalis for bioremediation of high-temperature radioactive waste environments. *Applied and environmental microbiology.* 2003;69(8):4575–4582.

[3] Gabani P, Singh OV. Radiation-resistant extremophiles and their potential in biotechnology and therapeutics. *Applied microbiology and biotechnology*. 2013;97(3):993–1004.

[4] Singh O, Gabani P. Extremophiles: radiation resistance microbial reserves and therapeutic implications. *Journal of applied microbiology.* 2011;110(4):851–861.

[5] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology.* 1970;48(3):443–453.

[6] Sghaier H, Ghedira K, Benkahla A, Barkallah I. Basal DNA repair machinery is subject to positive selection in ionizing-radiation-resistant bacteria. *BMC genomics*. 2008;9(1):297.

[7] Makarova KS, Omelchenko MV, Gaidamakova EK, Matrosova VY, Vasilenko A, Zhai M, et al. *Deinococcus geothermalis: the pool of extreme radiation resistance genes shrinks.* PLoS One. 2007;2(9):e955.

[8] Alexeyenko A, Tamas I, Liu G, Sonnhammer EL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*. 2006;22(14):e9–e15.

[9] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*. 1994;22(22):4673–4680.

[10] Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. 2009;25(11):1451–1452. Available from: http://www.ub.edu/dnasp/.

[11] Banerjee R, Roy A, Mukhopadhyay S. Genomic and proteomic signatures of radiation and thermophilic adaptation in the Deinococcus-Thermus genomes. *International Journal of Pharmacy and Pharmaceutical Sciences.* 2014;6:287–300.

[12] Peden J. CodonW software; 1997. Available from: http://codonw.sourceforge.net/.

[13] SPSS Inc. SPSS software for windows (version 15.0); 2007.

[14] StatSoft Inc. Statistica 6 software; 2002. Available from: http://www.statsoft.com/Products/STATISTICA-Features.

[15] Vesth T, Lagesen K, Acar Ö, Ussery D. *CMG-biotools, a free workbench for basic comparative microbial genomics*. PLoS One. 2013;8(4):e60120.

[16] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*. 1990;215(3):403–410.

[17] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution.* 2007;24(8):1586–1591.

[18] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research.* 2001;29(1):22–28.

[19] Markowitz VM, Chen IMA, Chu K, Szeto E, Palaniappan K, Grechkin Y, et al. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic acids research*. 2011;40(1):123–129.

[20] Shuryak I, Dadachova E. *Quantitative modeling of microbial population responses to chronic irradiation combined with other stressors.* PloS one. 2016;11(1):e0147696.

[21] Fredrickson JK, Zachara JM, Balkwill DL, Kennedy D, Shu-mei WL, Kostandarithes HM, et al. Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the Hanford Site, Washington State. *Applied and environmental microbiology.* 2004;70(7):4230–4241.

[22] Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer research.* 1967;27(2 Part 1):209–220.

[23] Dray S, Dufour AB, et al. The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software.* 2007;22(4):1–20.

[24] Breiman L. Random forests. *Machine learning*. 2001;45(1):5–32.

[25] Song L, Langfelder P, Horvath S. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC bioinformatics.* 2013;14(1):5.

[26] Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. *Springer Science & Business Media*; 2003.

[27] Burnham KP, Anderson DR, Huyvaert KP. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*. 2011;65(1):23–35.

[28] Zoghlami M, Aridhi S, Sghaier H, Maddouri M, Nguifo EM. *A multiple instance learning approach for sequence data with across bag dependencies.* CoRR. 2016;abs/1602.00163. Available from: http://arxiv.org/abs/1602.00163.

[29] Xing Z, Pei J, Keogh E. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter.* 2010;12(1):40–48.

[30] Amores J. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*. 2013;201:81–105.

[31] Alpaydın E, Cheplygina V, Loog M, Tax DM. Single-vs. multiple-instance classification. *Pattern Recognition.* 2015;48(9):2831–2838.

[32] Andrews S, Tsochantaridis I, Hofmann T. Support Vector Machines for Multiple-Instance Learning. In: Thrun, Obermayer K, editors. *Advances in Neural Information Processing Systems.* Cambridge, MA: MIT Press; 2003. p. 561–568.

[33] Maron O, Pérez TL. A Framework for Multiple-Instance Learning. In: Jordan MI, Kearns MJ, Solla SA, editors. *Advances in Neural Information Processing Systems*. vol. 10. Cambridge, MA: The MIT Press; 1998. p. 570–576.

[34] Faria AW, Coelho FGF, Silva A, Rocha H, Almeida G, Lemos AP, et al. MILKDE: A new approach for multiple instance learning based on positive instance selection and kernel density estimation. *Engineering Applications of Artificial Intelligence.* 2017;59:196–204.

[35] Bunescu RC, Mooney RJ. Multiple instance learning for sparse positive bags. In: *Proc. 24th international conference on Machine learning*. ACM; 2007. p. 105–112.

[36] Wang J. Solving the multiple-instance problem: A lazy learning approach. In: *Proc. 17th International conference on Machine Learning*. Morgan Kaufmann; 2000. p. 1119–1125.

[37] Shakhnarovich G, Darrell T, Indyk P. Nearest-neighbor methods in learning and vision. *IEEE transactions on neural networks*. 2008;19(2):377.

[38] Cheplygina V, Tax DM, Loog M. Multiple instance learning with bag dissimilarities. *Pattern Recognition*. 2015;48(1):264–275.

[39] Maddouri M, Elloumi M. Encoding of Primary Structures of Biological Macromolecules Within a Data Mining Perspective. *Journal of Computer Science and Technology*. 2004;19(1):78–88.

[40] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*. 2009;11(1):10–18.

[41] Platt J. Fast training of support vector machines using sequential minimal optimization. In: *Advances in kernel methods.* MIT Press; 1999. p. 185–208.

[42] Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural computation.* 2001;13(3):637–649.

[43] Quinlan JR. C4. 5: programs for machine learning. Morgan Kaufmann Publishers; 1993.

[44] John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. In: *Proc. 11th conference on uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc.; 1995. p. 338–345.

[45] Aridhi S, Sghaier H, Zoghlami M, Maddouri M, Nguifo EM. Prediction of Ionizing Radiation Resistance in Bacteria Using a Multiple Instance Learning Model. *Journal of Computational Biology.* 2016;23(1):10–20.

[46] Woolfit M, Bromham L. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Molecular Biology and Evolution*. 2003;20(9):1545–1555.

MA