JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 23, Number 1, 2016
© Mary Ann Liebert, Inc.
Pp. 10–20
DOI: 10.1089/cmb.2015.0134

Prediction of Ionizing Radiation Resistance in Bacteria Using a Multiple Instance Learning Model

SABEUR ARIDHI,[1,2,3] HAÏTHAM SGHAIER,[4,5] MANEL ZOGHLAMI,[1,2,3]
MONDHER MADDOURI,[3,6] and ENGELBERT MEPHU NGUIFO[1,2]

ABSTRACT

Ionizing-radiation-resistant bacteria (IRRB) are important in biotechnology. In this context, *in silico* methods of phenotypic prediction and genotype–phenotype relationship discovery are limited. In this work, we analyzed basal DNA repair proteins of most known proteome sequences of IRRB and ionizing-radiation-sensitive bacteria (IRSB) in order to learn a classifier that correctly predicts this bacterial phenotype. We formulated the problem of predicting bacterial ionizing radiation resistance (IRR) as a multiple-instance learning (MIL) problem, and we proposed a novel approach for this purpose. We provide a MIL-based prediction system that classifies a bacterium to either IRRB or IRSB. The experimental results of the proposed system are satisfactory with 91.5% of successful predictions.

Key words: bacterial ionizing radiation resistance, multiple instance learning, prediction.

1. INTRODUCTION

TO DATE, GENOMIC DATABASES INDICATE the presence of thousands of genome projects. However, limited computational works are available for the prediction of bacterial IRR (Sghaier et al., 2008, 2013; Omelchenko et al., 2005; Makarova et al., 2007; Ghosal et al., 2005) and consequently the rapid determination of useful microorganisms for several applications (bioremediation of radioactive wastes). As mentioned in a previous article (Sghaier et al., 2008), we consider IRRB as non-spore-forming bacteria that can protect their cytosolic proteins from oxidation and tolerate many DNA double-strand breaks (DSBs) after exposure to high, acute ionizing radiation (IR) (dose greater than 1 kilogray (kGy) for 90% reduction (D_{10}) in

[1] Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS)–Université Blaise Pascal (UBP), Clermont Ferrand, France.

[2] Centre National de Recherche Scientifique (CNRS), UMR 6158, LIMOS, Aubière, France.

[3] University of Tunis El Manar, Faculty of Sciences of Tunis, LIPAH, Tunis, Tunisia.

[4] National Center for Nuclear Sciences and Technology (CNSTN), Sidi Thabet Technopark, Ariana, Tunisia.

[5] Laboratory BVBGR, ISBST, University of Manouba, La Manouba, Tunisia.

[6] Taibah University, Medinah, Saudi Arabia.

This is an expanded and updated version of an earlier article published in BIOKDD '13, a workshop of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD '13) (Aridhi et al., 2013).

colony forming units [CFUs]). Partly, it seems that the shared ability of IRRB to survive the damaging effects of IR is the result of positively selected basal deoxyribonucleic acid (DNA) repair pathways (Sghaier et al., 2008) and high intracellular manganese concentration (Daly, 2012).

In this work, we study basal DNA repair proteins of IRRB and IRSB to develop a bioinformatics approach for the phenotype prediction of IRR. Thus, we consider that each studied bacterium is represented by a set of DNA repair proteins. Due to this fact, we formalize the problem of predicting IRR in bacteria as an MIL problem in which bacteria represent bags and repair proteins of each bacterium represent instances. Many MIL algorithms have been developed to solve several problems such as predicting types of protein–protein interactions (PPI) (Yamakawa et al., 2007) and drug activity prediction (Fu et al., 2012), mainly including diverse density (Maron and Pérez, 1998), citation-kNN and Bayesian-kNN (Wang and Zucker, 2000), MI-SVM (Andrews et al., 2003), and HyDR-MI (Zafra et al., 2013). Diverse density (DD) was proposed in Maron and Pérez (1998) as a general framework for solving multi-instance learning problems. The main idea of DD approach is to find a concept point in the feature space that are close to at least one instance from every positive bag and meanwhile far away from instances in negative bags. The optimal concept point is defined as the one with the maximum diversity density, which is a measure of how many different positive bags have instances near the point, and how far the negative instances are away from that point. In Wang and Zucker (2000), the minimum Hausdorff distance was used as the bag-level distance metric, defined as the shortest distance between any two instances from each bag. Using this bag-level distance, the *k-NN* algorithm predicts the label of an unseen bag. In Andrews et al. (2003), the authors proposed the algorithm *MI-SVM* to modify support vector machines. The algorithm *MI-SVM* explicitly treats the label instance labels as unobserved hidden variables subject to constraints defined by their bag labels. The goal is to maximize the usual instance margin jointly over the unknown instance labels and a linear or kernelized discriminant function. In Zafra et al. (2013), the authors proposed a feature subset selection method for MIL algorithms called HyDR-MI (hybrid dimensionality reduction method for multiple instance learning). The hybrid consists of the filter component based on an extension of the ReliefF algorithm (Zafra et al., 2012) developed for working with MIL and the wrapper component based on a genetic algorithm that optimizes the search for the best feature subset from a reduced set of features, output by the filter component.

The above cited algorithms use an attribute-value format to represent their data. A most used approach to represent protein sequences in an attribute-value format is to extract motifs that can serve as attributes. Appropriately chosen sequence motifs may reduce noise in the data and indicate active regions of the protein. A protein can then be represented as a set of motifs (Ben-Hur and Brutlag, 2003; Saidi et al., 2012) or as a vector in a vector space spanned by these motifs (Saidi et al., 2010). However, the use of this technique is not suitable in the context of phenotypic prediction of bacterial IRR. This is due to the fact that the set of proteins of each bag must be represented (in the attribute-value format) with the same set of attributes, which is possible only if all extracted motifs from the different bags of proteins were used together as a unique set of motifs. As the different bags of proteins are processed disjointly, it is necessary to design a novel approach for such cases.

In this article, we propose an MIL approach for predicting bacterial IRR using proteins implicated in basal DNA repair. For this purpose, we used a local alignment technique to measure the similarity between protein sequences of the studied bacteria. To the best of our knowledge, this is the first work that proposes an *in silico* approach for phenotypic prediction of bacterial IRR.

The remainder of this article is organized as follows. Section 2 presents the materials and methods used in our study. In section 3, we describe our experimental techniques and we discuss the obtained results. Concluding points make the body of section 4.

2. MATERIALS AND METHODS

2.1. *Terminology and problem formulation*

The task of multiple instance learning (MIL) was coined by Dietterich et al. (1997) when they were investigating the problem of drug activity prediction. In multiple-instance learning, the training set is composed of n labeled bags. Each bag in the training set contains k instances and have a bag label $y_i \in \{-1, +1\}$. We notice that instances of each bag have labels $y_{ij} \in \{-1, +1\}$, but these values are not known during training. The most common assumption in this field is that a bag is labeled positive if at least one of its instances is positive, which can be expressed as follows:

$$y_i = \max_j (y_{ij}). \tag{1}$$

The task of MIL is to learn a classifier from the training set that correctly predicts unseen bags. Although MIL is quite similar to traditional supervised learning, the main difference between the two approaches can be found in the class labels provided by the data. According to the specification given by Dietterich et al. (1997), in a traditional setting of machine learning, an object m is represented by a feature vector (an instance), which is associated with a label. However, in a multiple instance setting, each object m may have k various instances denoted $m_1, m_2, \cdots, m_k$. The difference between the traditional setting of machine learning and the multiple instance learning setting can be represented clearly in Figure 1, where the difference between the input objects is shown.

In our work, we are interested in the prediction of the phenotype of IRR in a family composed of a set of bacteria. Let $DB = \{X_1, \ldots, X_n\}$ be a bacteria database. Each bacterium in the database is represented by a set of proteins $X_i = \{p_{i1}, \cdots, p_{ik}\}$ and belongs to a class label y_i with $y_i = \{IRRB, IRSB\}$. The problem of phenotypic prediction of IRRB can be seen as an MIL problem in which bacteria represent bags, and basal DNA repair proteins of each bacterium represent instances.

The problem investigated in this work is to learn a multiple-instance classifier in this setting. Given a query bacterium $Q = \{p_1, \cdots, p_k\}$, the classifier must use primary structures of basal DNA repair proteins in Q and in each bag of DB to predict the label of Q.

2.2. *MIL-ALIGN algorithm*

Based on the formalization, we propose the MIL-ALIGN algorithm allowing to predict IRRB. The proposed algorithm focuses on discriminating bags by the use of local alignment technique to measure the

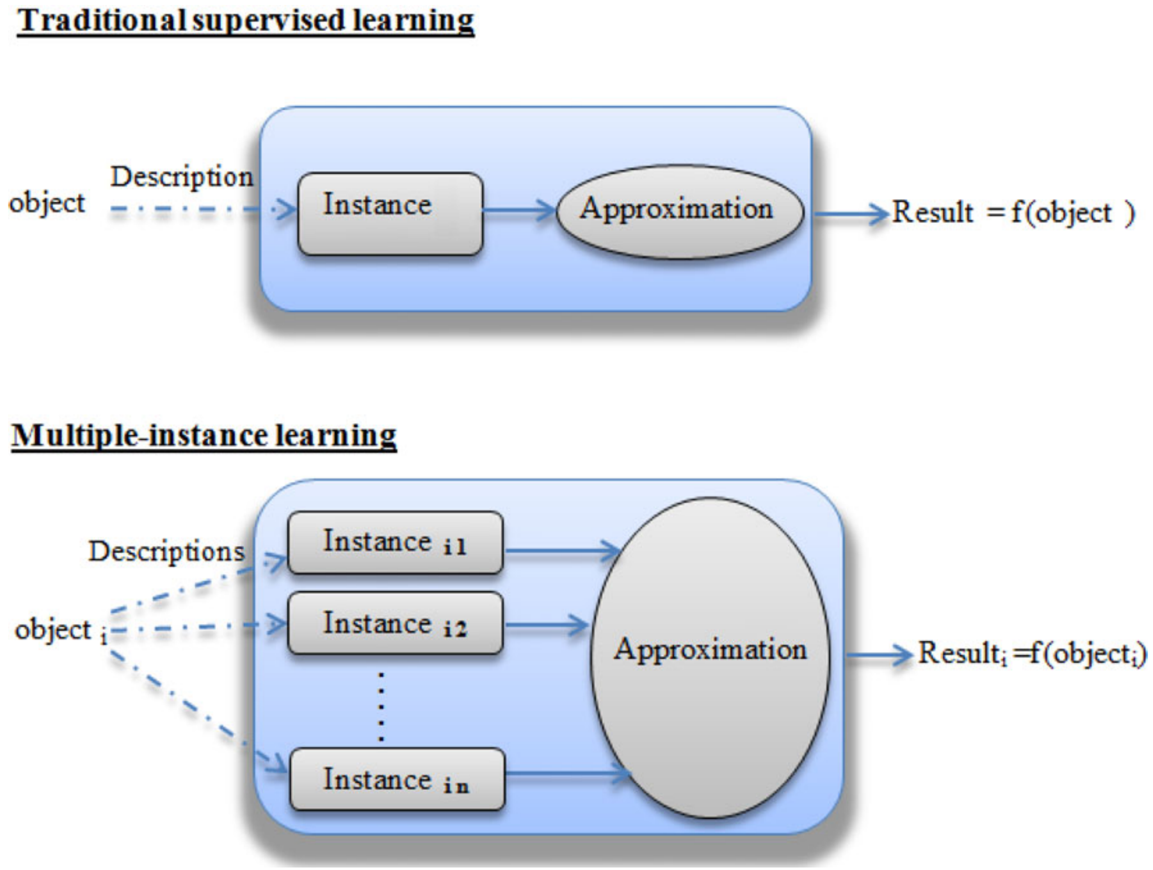

FIG. 1. Differences between traditional supervised learning and multiple instance learning.

similarity between each protein sequence in the query bag and corresponding protein sequence in the different bags of the learning database.

In MIL-ALIGN algorithm we use the following variables for input data and for accumulating data during the execution of the algorithm:

- the variable Q: corresponds to the query bag (the query bacterium), which is a vector of protein sequences.
- the variable DB: corresponds to the bacteria database.
- the variable S: corresponds to a matrix used to store alignment score vectors.

Algorithm 1: MIL-ALIGN

```
Require: Learning database $DB = \{(X_1, y_1), \cdots, (X_n, y_n)\}$, Query $Q = \{p_{q1}, \cdots, p_{qk}\}$
Ensure: Prediction result $R$
1:  for all $p_{qi} \in Q$ do
2:      for all $X_j$ do
3:          $S_{ij} \leftarrow LocalAlignment(p_{qi}, p_{X_j i})$   $//X_j = \{p_{j1}, \cdots, p_{jk}\}$ and $p_{X_{ji}}$ is the protein number $i$ of bacterium $X_j$
4:      end for
5:  end for
6:  $R \leftarrow Aggregate(S)$
7:  return $R$
```

Informally, the algorithm works as follows (see Algorithm 1):

1. For each protein sequence p_i in the query bag Q, MIL-ALIGN computes the corresponding alignment scores (line 1 to 5).
2. Group alignment scores of all protein sequences of query bacterium into a matrix S (line 3). Line i of S corresponds to a score vector of protein p_i against all proteins $p_{X_{ji}}$ of X_j with $1 \leq j \leq n$. Element S_{ij} corresponds to the alignment score of protein p_{q^i} of Q with protein $p_{X_{ji}}$ of bacterium X_j.
3. Apply an aggregation method to S in order to compute the final prediction result R (line 6 to 7). A query bacterium is predicted as IRRB (respectively IRSB) if the aggregation result of similarity scores of its proteins against associated proteins in the learning database is IRRB (respectively IRSB).

2.3. *Experimental environment*

Information on complete and ongoing IRRB genome sequencing projects was obtained from the GOLD database (Liolios et al., 2008). We initiated our analyses by retrieving orthologous proteins implicated in basal DNA repair in IRRB and IRSB with sequenced genomes. Proteins of the bacterium *Deinococcus radiodurans* (B7) were downloaded from the UniProt website. PrfectBLAST tool (Santiago-Sotelo and Ramirez-Prado, 2012) was used to identify orthologous proteins. Proteomes of other bacteria were downloaded from the NCBI FTP website.

For our experiments, we constructed a database containing 28 bags (14 IRRB and 14 IRSB). Table 1 presents the used IRRB and IRSB. Each bacterium contains 25 to 31 instances that correspond to proteins implicated in basal DNA repair in IRRB (see Table 2).

3. RESULTS AND DISCUSSION

3.1. *Experimental techniques*

Computations were carried out on an i7 CPU 2.49 GHz PC with 6 GB memory, operating on Linux Ubuntu. In the classification process, we used the leave-one-out (LOO) technique (Han et al., 2011) also known as *jack-knife test*. For each dataset (comprising n bags), only one bag is kept for the test and the remaining part is used for the training. This action is repeated n times. In our context, the leave-one-out is considered to be the most objective test technique compared to the other ones (i.e., hold-out, n-cross-validation), as our training set contains a small number of bacteria.

For our tests, we used the basic local alignment search tool (BLAST) (Altschul et al., 1990) for computing local alignments. We implemented two aggregation methods to be used with MIL-ALIGN: the *sum of maximum scores* method and the *weighted average of maximum scores* method.

Table 1. IRRB and IRSB Learning Set

Phenotype	*ID*	*Bacterium*	*Phylogenetic group*	*D_{10} (kGy)*[a]
IRRB	B1	*Chroococcidiopsis thermalis* PCC 7203	Cyanobacteria	4[b] (Billi et al., 2002)
	B2	*Deinococcus deserti* VCD115	*Deinococcus-Thermus*	>7.5 (Slade and Radman, 2011)
	B3	*Deinococcus geothermalis* DSM 11300	*Deinococcus-Thermus*	10–16 (Slade and Radman, 2011)
	B4	*Deinococcus gobiensis* I 0	*Deinococcus-Thermus*	12.7 (Slade and Radman, 2011)
	B5	*Deinococcus maricopensis* DSM 21211	*Deinococcus-Thermus*	~11 (Rainey et al., 2005)
	B6	*Deinococcus proteolyticus* MRP	*Deinococcus-Thermus*	>15 (Brooks and Murray, 1981)
	B7	*Deinococcus radiodurans* R1	*Deinococcus-Thermus*	10 (Ito et al., 1983)
	B8	*Geodermatophilus obscurus* DSM 43160	Actinobacteria	9 (Gtari et al., 2012)
	B9	*Kineococcus radiotolerans* SRS30216	Actinobacteria	2 (Phillips et al., 2002)
	B10	*Kocuria rhizophila* DC2201	Actinobacteria	2[c] (Rainey et al., 1997; Brooks and Murray, 1981)
	B11	*Methylobacterium radiotolerans* JCM 2831	Proteobacteria	1 (Green and Bousfield, 1983; Ito and Iizuka, 1971)
	B12	*Modestobacter marinus*	Actinobacteria	6 (Gtari et al., 2012)
	B13	*Rubrobacter xylanophilus* DSM 9941	Actinobacteria	5.5 (Ferreira et al., 1999)
	B14	*Truepera radiovictrix* DSM 17093	*Deinococcus-Thermus*	>5 Albuquerque et al., 2005)
IRSB	B15	*Brucella abortus* S19	Proteobacteria	0.34 (Federighi and Tholozan, 2001)
	B16	*Escherichia coli* B REL606	Proteobacteria	0.7 (Daly et al., 2004)
	B17	*Escherichia coli* str. K-12 substr. DH10B	Proteobacteria	0.7 (Daly et al., 2004)
	B18	*Neisseria gonorrhoeae* FA 1090	Proteobacteria	0.07–0.125 (Daly et al., 2004)
	B19	*Neisseria gonorrhoeae* TCDC NG08107	Proteobacteria	0.07–0.125 (Daly et al., 2004)
	B20	*Pseudomonas putida* S16	Proteobacteria	0.25 (Daly et al., 2004)
	B21	*Shewanella oneidensis* MR-1	Proteobacteria	0.07 (Daly et al., 2004)
	B22	*Shigella dysenteriae*1617	Proteobacteria	0.22 (Federighi and Tholozan, 2001)
	B23	*Thermus thermophilus* HB27	*Deinococcus-Thermus*	0.8 (Federighi and Tholozan, 2001)
	B24	*Thermus thermophilus* HB8	*Deinococcus-Thermus*	0.8[d] (Federighi and Tholozan, 2001)
	B25	*Thermus thermophilus* JL-18	*Deinococcus-Thermus*	0.8[d] (Federighi and Tholozan, 2001)
	B26	*Thermus thermophilus* SG0.5JP17-16	*Deinococcus-Thermus*	0.8[d] (Federighi and Tholozan, 2001)
	B27	*Vibrio parahaemolyticus* RIMD 2210633	Proteobacteria	0.03–0.06 (Federighi and Tholozan, 2001)
	B28	*Yersinia enterocolitica* 8081	Proteobacteria	0.1–0.21 (Federighi and Tholozan, 2001)

[a] D_{10}: Dose for 90% reduction in colony forming units (CFUs); for IRRB, it is greater than 1 kGy.

[b] *Chroococcidiopsis* spp.

[c] *Kocuria rosea.*

[d] *T. thermophilus* HB27.

IRRB, ionizing-radiation-resistant bacteria; IRSB, ionizing-radiation-sensitive bacteria.

Sum of maximum scores (SMS). For each protein in the query bacterium, we scan the corresponding line of S, which contains the obtained scores against all other bacteria of the training database. The SMS method selects the maximum score among the alignments scores against IRRB (which we call max_R) and the maximum score among the scores of alignments against IRSB (which we call max_S). It then compares these scores. If max_R is greater than max_S, it adds max_R to the total score of IRRB [which we call $total_R(S)$]. Otherwise, it adds max_S to the total score of IRSB [which we call $total_S(S)$]. When all selected proteins were processed, the SMS method compares total scores of IRRB and IRSB. If $total_R(S)$ is greater than $total_S(S)$, the prediction output is IRRB. Otherwise, the prediction output is IRSB.

Table 2. Replication, Repair, and Recombination Proteins

ID	*Protein*	*Function*
P1	Hypothetical DNA polymerase	DNA polymerase
P2	DNA polymerase III, α subunit	
P3	DNA-directed DNA polymerase	
P4	DNA polymerase III, τ/γ subunit	
P5	Single-stranded DNA-binding protein	Replication complex
P6	Replicative DNA helicase	
P7	DNA primase	
P8	DNA gyrase, subunit B	
P9	DNA topoisomerase I	
P10	DNA gyrase, subunit A	
P11	Smf proteins	Other DNA-associated proteins
P12	Endonuclease III	
P13	Holliday junction resolvase	
P14	Formamidopyrimidine-DNA glycosylase	
P15	Holliday junction DNA helicase	
P16	RecF protein	
P17	DNA repair protein radA	
P18	Holliday junction binding protein	
P19	Excinuclease ABC, subunit C	
P20	DNA repair protein RecN	
P21	Transcription-repair coupling factor	
P22	Excinuclease ABC, subunit A	
P23	DNA helicase II	
P24	DNA helicase RecG	
P25	Exonuclease SbcD, putative	
P26	Exonuclease SbcC	
P27	Ribonuclease HII	
P28	Excinuclease ABC, subunit B	
P29	A/G-specific adenine glycosylase	
P30	RecA protein	
P31	DNA-3-methyladenine glycosidase II, putative	

Below, we formally define the SMS method:

$$\mathrm{SMS}(S) = \begin{cases} \mathit{IRRB}, & \text{if } \mathit{total}_R(S) \geq \mathit{total}_S(S), \\ \mathit{IRSB}, & \text{otherwise,} \end{cases}$$

where

- $\mathit{total}_R(S) = \sum_{i=1}^{n} \max_{1 \leq j \leq k} S_{ij}$ such that $y_j = \mathit{IRRB}$, and
- $\mathit{total}_S(S) = \sum_{i=1}^{n} \max_{1 \leq j \leq k} S_{ij}$ such that $y_j = \mathit{IRSB}$.

Weighted average of maximum scores (WAMS). With the WAMS method, each protein p_i has a given weight w_i. For each protein in the query bacterium, we scan the corresponding line of S, which contains the obtained scores against all other bacteria of the training database. The WAMS method selects the maximum score among the scores of alignments against IRRB [which we call $\mathit{max}_R(S)$] and the maximum score among the scores of alignments against IRSB [which we call $\mathit{max}_S(S)$]. It then compares these scores. If the $\mathit{max}_R(S)$ is greater than $\mathit{max}_S(S)$, it adds $\mathit{max}_R(S)$ multiplied by the weight of the protein to the total score of IRRB and it increments the number of IRRB having a max score. Otherwise, it adds $\mathit{max}_S(S)$ multiplied by the weight of the protein to the total score of IRSB and it increments the number of IRSB having a max score. When all the selected proteins were processed, we compare the average of total scores of IRRB [which we called $\mathit{avg}_R(S)$] and the average of total scores of IRSB [which we called $\mathit{avg}_S(S)$]. If $\mathit{avg}_R(S)$ is greater than $\mathit{avg}_S(S)$, the prediction output is IRRB. Otherwise, the prediction output is IRSB.

Below, we formally define the WAMS method:

$$\mathrm{WAMS}(S) = \begin{cases} IRRB, & \text{if } avg_R(S) \geq avg_S(S), \\ IRSB, & \text{otherwise,} \end{cases}$$

where

- $avg_R(S) = total_R(S)/num_R$, and
- $avg_S(S) = total_S(S)/num_S$,

and

- $total_R(S) = \sum_{i=1}^{n} \max_{1 \leq j \leq k} \; S_{ij} \cdot w_i$ such that $y_j = IRRB$, and
- $total_S(S) = \sum_{i=1}^{n} \max_{1 \leq j \leq k} \; S_{ij} \cdot w_i$ such that $y_j = IRSB$,

where w_i is the weight of the protein p_i.

3.2. Results

In order to simulate the traditional setting of machine learning in the context of prediction of IRR in bacteria, we conducted a set of experiments with MIL-ALIGN by selecting just one protein for each bacterium in the learning set. Each experiment consists of aggregating alignment scores between a protein

Table 3. Learning Results with the Traditional Setting of Machine Learning

Protein ID	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*
P1	85.7	100	77.7
P2	89.2	92.3	86.6
P3	82.1	90.9	76.4
P4	89.2	92.3	86.6
P5	89.2	92.3	86.6
P6	89.2	92.3	86.6
P7	89.2	92.3	86.6
P8	78.5	83.3	75
P9	89.2	92.3	86.6
P10	89.2	92.3	86.6
P11	89.2	92.3	86.6
P12	89.2	92.3	86.6
P13	78.5	90	72.2
P14	89.2	92.3	86.6
P15	85.7	91.6	81.2
P16	89.2	92.3	86.6
P17	85.7	91.6	81.2
P18	85.7	91.6	81.2
P19	89.2	92.3	86.6
P20	85.7	91.6	81.2
P21	85.7	91.6	81.2
P22	89.2	92.3	86.6
P23	89.2	92.3	86.6
P24	89.2	92.3	86.6
P25	85.7	91.6	81.2
P26	82.1	90.9	76.4
P27	82.1	100	73.6
P28	89.2	92.3	86.6
P29	78.5	90	72.2
P30	89.2	92.3	86.6
P31	78.5	78.5	78.5

Table 4. Experimental Results of MIL-ALIGN with LOO-Based Evaluation Technique

Used proteins	*Aggregation method*	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*
All proteins	SMS	92.8	92.8	92.8
	WAMS	89.2	92.3	86.6
DNA polymerase proteins	SMS	89.2	92.3	86.6
	WAMS	89.2	92.3	86.6
Replication complex proteins	SMS	92.8	92.8	92.8
	WAMS	92.8	92.8	92.8
Other DNA-associated proteins	SMS	92.8	92.8	92.8
	WAMS	92.8	92.8	92.8

SMS, sum of maximum scores; WAMS, weighted average of maximum scores.

sequence of a query bacterium and the corresponding protein sequences of each bacterium in the learning database. We present in Table 3 learning results with the traditional setting of machine learning. The LOO-based evaluation technique was used to generate the presented results.

As shown in Table 3, we conducted 31 experiments (with 31 proteins). Results show that the use of our algorithm with just one instance for each bag in the learning database allows good accuracy values.

In order to study the importance of considering the problem of predicting bacterial IRR as a multiple instance learning problem, we present in Table 4 the experimental results of MIL-ALIGN using a set of

Table 5. Percentage of Successful Predictions Using MIL

Phenotype	*Bacterium ID*	*Successful predictions (%)*
IRRB	B1	100
	B2	100
	B3	100
	B4	100
	B5	100
	B6	100
	B7	100
	B8	100
	B9	100
	B10	100
	B11	0
	B12	100
	B13	100
	B14	62.5[a]
IRSB	B15	0
	B16	100
	B17	100
	B18	100
	B19	100
	B20	100
	B21	100
	B22	100
	B23	100
	B24	100
	B25	100
	B26	100
	B27	100
	B28	100

[a] Successfully classified bacterium using three settings: (1) all proteins with SMS aggregation method; (2) replication complex proteins with SMS and WAMS aggregation methods; and (3) other DNA-associated proteins with SMS and WAMS aggregation methods.

proteins to represent the studied bacteria. For each set of proteins and for each aggregation method, we present the accuracy, the sensitivity, and the specificity of MIL-ALIGN. We notice that the WAMS aggregation method was used with equally weighted proteins. We used the LOO-based evaluation technique to generate the presented results.

We notice that the use of the whole set of proteins to represent the studied bacteria allows good accuracy accompanied by high values of sensitivity and specificity. This can be explained by the pertinent choice of basal DNA repair proteins to predict the phenotype of IRR. The high values of specificity presented by MIL-ALIGN indicate the ability of this algorithm to identify negative bags (IRSB). Using all proteins, we have 92.8% accuracy and specificity. We do not exceed these values in all the cases of mono-instance learning presented in Table 3. As shown in Table 4, the SMS aggregation method allows better results than the WAMS aggregation method using the whole set of proteins to represent the studied bacteria. Using the other subsets of proteins (DNA polymerase, replication complex, and other DNA-associated proteins) to represent the bacteria, SMS and WAMS present the same results.

In order to study the correctly classified bacteria with the MIL, we computed for each bacterium in the learning database the percentage of experiments that succeed to classify the bacterium (see Table 5).

As shown in Table 5, more than 89% of tested bacteria show successful predictions of 100%. This means that we succeed to correctly predict the IRR phenotype of those bacteria. On the other hand, the results illustrated in Table 5 may help to understand some characteristics of the studied bacteria. In particular, the IRRB *M. radiotolerans* (B11) and the IRSB *B. abortus* (B15) present a high rate of failed predictions. It means that in most cases, *M. radiotolerans* is predicted as IRSB and *B. abortus* is predicted as IRRB; the former is an intracellular parasite (Halling et al., 2005) and the latter is an endosymbiont of most plant species (Fedorov et al., 2013). A probable explanation for these two failed predictions is the increased rate of sequence evolution in endosymbiotic bacteria (Woolfit and Bromham, 2003). As our training set is composed mainly of members of the phylum *Deinococcus-Thermus*; expectedly, the *Deinococcus* bacteria (B2-B7) present a very low rate of failed predictions.

4. CONCLUSION

In this article, we addressed the issue of prediction of bacterial IRR phenotype. We have considered that this problem is a multiple-instance learning problem in which bacteria represent bags and repair proteins of each bacterium represent instances. We have formulated the studied problem and described our proposed algorithm MIL-ALIGN for phenotype prediction in the case of IRRB. By running experiments on a real dataset, we have shown that experimental results of MIL-ALIGN are satisfactory with 91.5% of successful predictions.

In the future work, we will study the performance of the proposed approach to improve its efficiency, particularly for endosymbiont bacteria. Also, we will study the use of *a priori* knowledge to improve the efficiency of our algorithm. This *a priori* knowledge can be used to assign weights to proteins during the learning step of our approach. A notable interest will be dedicated to the study of other proteins that can be involved with the high resistance of IRRB to the IR and desiccation, two positively correlated phenotypes.

ACKNOWLEDGMENTS

This work was partially supported by the French-Tunisian project: General Directorate of Scientific Research in Tunisia (DGRST)/National Center for Scientific Research in France (CNRS) [IRRB11/R-14-09], by the French Region of Auvergne, and by the Fédération de Recherche en Environment [UBP/CNRS-FR-3467].

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

REFERENCES

Albuquerque, L., Simoes, C., Nobre, M.F., et al. 2005. *Truepera radiovictrix* gen. nov., sp. nov., a new radiation resistant species and the proposal of *Trueperaceae* fam. nov. *FEMS Microbiol. Lett.* 247, 161–169.

Altschul, S.F., Gish, W., Miller, W., et al. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Andrews, S., Tsochantaridis, I., and Hofmann, T. 2003. Support vector machines for 11 multiple-instance learning, 561–568. *In Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA.

Aridhi, S., Sghaier, H., Maddouri, M., et al. 2013. Computational phenotype prediction of ionizing-radiation-resistant bacteria with a multiple-instance learning model, 18–24. *In Proceedings of the 12th International Workshop on Data Mining in Bioinformatics (BioKDD13).* ACM, New York, NY.

Ben-Hur, A., and Brutlag, D. 2003. Remote homology detection: A motif based approach. *Bioinformatics*. 19, 26–33.

Billi, D., Friedmann, E.I., Hofer, K.G., et al. 2002. Ionizing-radiation resistance in the desiccation-tolerant cyanobacterium *Chroococcidiopsis*. *Appl. Environ. Microbiol.* 66, 1489–1492.

Brooks, B.W., and Murray, R.G.E. 1981. Nomenclature for *Micrococcus radiodurans* and other radiation-resistant cocci: *Deinococcaceae* fam. nov. and *Deinococcus* gen. nov., including five species. *Int. J. Syst. Bacteriol.* 31, 353–360.

Daly, M.J. 2012. Death by protein damage in irradiated cells. *DNA Repair.* 11, 12–21.

Daly, M.J., Gaidamakova, E.K., Matrosova, V.Y., et al. 2004. Accumulation of Mn(II) in *Deinococcus radiodurans* facilitates gamma-radiation resistance. *Science.* 306, 1025–1028.

Dietterich, T.G., Lathrop, R.H., and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* 89, 31–71.

Federighi, M., and Tholozan, J.-L. 2001. Traitements ionisants et hautes pressions des aliments. *Economica.* 247, 161–169.

Fedorov, D.N., Ekimova, G.A., Doronina, N.V., et al. 2013. 1-Aminocyclopropane-1- carboxylate (ACC) deaminases from M*ethylobacterium radiotolerans* and *Methylobacterium nodulans* with higher specificity for ACC. *FEMS Microbiol. Lett.* 343, 70–76.

Ferreira, A.C., Nobre, M.F., Moore, E., et al. 1999. Characterization and radiation resistance of new isolates of *Rubrobacter radiotolerans* and *Rubrobacter xylanophilus*, Extremophiles. *Int. J. Syst. Bacteriol.* 3, 235–238.

Fu, G., Nan, X., Liu, H., et al. 2012. Implementation of multiple-instance learning in drug activity prediction. *BMC Bioinformatics.* 13, S3.

Ghosal, D., Omelchenko, M.V., Gaidamakova, E.K., et al. 2005. How radiation kills cells: Survival of *Deinococcus radiodurans* and *Shewanella oneidensis* under oxidative stress. *FEMS Microbiol. Rev.* 29, 361–375.

Green, P.N., and Bousfield, I.J., 1983. Emendation of *Methylobacterium* (Patt, Cole, and Hanson 1976); *Methylobacterium rhodinum* (Heumann 1962) comb. nov. corrig.; *Methylobacterium radiotolerans* (Ito and Iizuka 1971) comb. nov. corrig.; and *Methylobacterium mesophilicum* (Austin and Goodfellow 1979) comb. nov. *Int. J. Syst. Bacteriol.* 33, 875–877.

Gtari, M., Essoussi, I., Maaoui, R., et al. 2012. Contrasted resistance of stone-dwelling *Geodermatophilaceae* species to stresses known to give rise to reactive oxygen species. *FEMS Microbiol. Ecol.* 80, 566–577.

Halling, S.M., Peterson-Burch, B.D., Bricker, B.J., et al. 2005. Completion of the genome sequence of *Brucella abortus* and comparison to the highly similar genomes of *Brucella melitensis* and *Brucella suis*. *Am. Soc. Microbiol.* 187, 2715–2726.

Han, J., Kamber, M., and Pei, J. 2011. *Data Mining: Concepts and Techniques*, 3rd edition. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Ito, H., and Iizuka, H. 1971. Taxonomic studies on a radio-resistant *Pseudomonas*. Part XII. Studies on the microorganisms of cereal grain. *Agric. Biol. Chem.* 35, 1566–1571.

Ito, H., Watanabe, H., Takeshia, M., et al. 1983. Isolation and identification of radiationresistant cocci belonging to the genus *Deinococcus* from sewage sludges and animal feeds. *Agric. Biol. Chem.* 47, 1239–1247.

Liolios, K., Mavromatis, K., Tavernarakis, N., et al. 2008. The Genomes On Line Database (GOLD) in 2007: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 36, D475–D479.

Makarova, K.S., Omelchenko, M.V., Gaidamakova, E.K., et al. 2007. *Deinococcus geothermalis*: The pool of extreme radiation resistance genes shrinks. *PLoS ONE.* 2, e955.

Maron, O., and Péerez, T.L. 1998. A framework for multiple-instance learning, 570–576. *In Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA.

Omelchenko, M.V., Wolf, Y.I., Gaidamakova, E.K., et al. 2005. Comparative genomics of *Thermus thermophilus* and *Deinococcus radiodurans*: Divergent routes of adaptation to thermophily and radiation resistance. *BMC Evol. Biol.* 5, 57.

Phillips, R.W., Wiegel, J., Berry, C.J., et al. 2002. *Kineococcus radiotolerans* sp. nov., a radiation-resistant, gram-positive bacterium. *Int. J. Syst. Evol. Microbiol.* 52, 933–938.

Rainey, F.A., Nobre, M.F., Schumann, P., et al. 1997. Phylogenetic diversity of the deinococci as determined by 16S ribosomal DNA sequence comparison. *Int. J. Syst. Bacteriol.* 47, 510–514.

Rainey, F.A., Ray, K., Ferreira, M., et al. 2005. Extensive diversity of ionizing-radiation-resistant bacteria recovered from Sonoran Desert soil and description of nine new species of the genus *Deinococcus* obtained from a single soil sample. *Appl. Environ. Microbiol.* 71, 5225–5235.

Saidi, R., Aridhi, S., Mephu Nguifo, E., et al. 2012. Feature extraction in protein sequences classiffication: A new stability measure, 683–689. *In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB 12)*. ACM, New York, NY.

Saidi, R., Maddouri, M., and Mephu Nguifo, E. 2010. Protein sequences classification by means of feature extraction with substitution matrices. *BMC Bioinformatics.* 11, 175.

Santiago-Sotelo, P., and Ramirez-Prado, J.H. 2012. prfectBLAST: A platform-independent portable front end for the command terminal BLAST+ stand-alone suite. *BioTechniques*. 53, 299–300.

Sghaier, H., Ghedira, K., Benkahla, A., et al. 2008. Basal DNA repair machinery is subject to positive selection in ionizing-radiation-resistant bacteria. *BMC Genomics*. 9, 297.

Sghaier, H., Thorvaldsen, S., and Malek Saied, N. 2013. There are more small amino acids and fewer aromatic rings in proteins of ionizing radiation-resistant bacteria. *Ann. Microbiol*. 63, 1483–1491.

Slade, D., and Radman, M. 2011. Oxidative stress resistance in *Deinococcus radiodurans*. *Microbiol. Mol. Biol. Rev*. 75, 133–191.

Wang, J., and Zucker, J.D. 2000. Solving the multiple-instance problem: A lazy learning approach, 1119–1126. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Woolfit, M., and Bromham, L. 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol. Biol. Evol.* 20, 1545–1555.

Yamakawa, H., Maruhashi, K., and Nakao, Y. 2007. Predicting types of proteinprotein interactions using a multiple-instance learning model, 42–53. *New Frontiers in Artificial Intelligence*. Springer, Berlin.

Zafra, A., Pechenizkiy, M., and Ventura, S. 2012. ReliefF-MI: An extension of ReliefF to multiple instance learning. *Neurocomputing*. 75, 210–218.

Zafra, A., Pechenizkiy, M., and Ventura, S. 2013. Hydr-mi: A hybrid algorithm to reduce dimensionality in multiple instance learning. *Inf. Sci.* 222, 282–301.

Address correspondence to:
Prof. Engelbert Mephu Nguifo
LIMOS, CNRS UMR 6158
Campus Universitaire des Cézeaux
1 rue de la Chebarde
63178 Aubiere Cedex, France

E-mail: mephu@isima.fr *and* sabeur.aridhi@gmail.com