Computational phenotype prediction of ionizing-radiation-resistant bacteria with a multiple-instance learning model

Sabeur Aridhi[*]
CNRS, UMR 6158, LIMOS,
F-63173 Aubiere, France.
Clermont University, Blaise
Pascal University, LIMOS, BP
10448, F-63000
Clermont-Ferrand, France.
University of Tunis El Manar,
LIPAH - FST, Academic
Campus, Tunis 2092, Tunisia.
aridhi@isima.fr

Haitham Sghaier
Unit of Microbiology and
Molecular Biology, National
Center for Nuclear Sciences
and Technologies (CNSTN),
Sidi Thabet Technopark,2020
Sidi Thabet, Tunisia

Mondher Maddouri
University of Tunis El Manar,
LIPAH - FST, Academic
Campus, Tunis 2092, Tunisia.

Engelbert Mephu Nguifo
CNRS, UMR 6158, LIMOS,
F-63173 Aubiere, France.
Clermont University, Blaise
Pascal University, LIMOS, BP
10448, F-63000
Clermont-Ferrand, France
mephu@isima.fr

ABSTRACT

Ionizing-radiation-resistant bacteria (IRRB) are important in biotechnology. The use of these bacteria for the treatment of radioactive wastes is determined by their surprising capacity of adaptation to radionuclides and a variety of toxic molecules. *In silico* methods are unavailable for the purpose of phenotypic prediction and genotype-phenotype relationship discovery. We analyze basal DNA repair proteins of most known proteomes sequences of IRRB and ionizing-radiation-sensitive bacteria (IRSB) in order to learn a classifier that correctly predicts unseen bacteria. In this work, we formulate the problem of predicting IRRB as a multiple-instance learning (MIL) problem and we propose a novel approach for predicting IRRB. We use a local alignment technique to measure the similarity between protein sequences to predict ionizing-radiation-resistant bacteria. The first results are satisfactory and provide a MIL-based prediction system that predicts whether a bacterium belongs to IRRB or to IRSB. The proposed system is available online at http://home.isima.fr/irrb/.

BIOKDD '13 Chicago, IL, USA
Copyright 2013 ACM 978-1-4503-2327-7 ...$15.00.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Computational biology

General Terms

Theory

Keywords

Phenotypic prediction, ionizing-radiation-resistant bacteria, ionizing-radiation-sensitive bacteria, protein sequences, multiple-instance learning

1. INTRODUCTION

Nuclear waste contains a variety of toxic and radioactive substances. The bioremediation of these wastes with pertinent bacteria and low cost is a challenging problem [13, 16]. The use of ionizing-radiation-resistant bacteria (IRRB) for the treatment of these radioactive wastes is determined by their surprising capacity of adaptation to radionuclides and to a variety of toxic molecules. To date, genomic databases indicate the presence of thousands of genome projects. However, only a few computational works are available for the purpose of phenotypic prediction discovery that rapidly determines useful genomes for the bioremediation of radioactive wastes [16, 17].

A main idea in this context is that resistance to ionizing radiation and tolerance of desiccation are two complex phenotypes, and suggest that protection and repair mechanisms are complementary in IRRB. In addition, it seems

[*] to whom correspondence should be addressed

that the shared ability of IRRB to survive the damaging effects of ionizing radiation and desiccation is the result of basal DNA repair pathways and that basal DNA repair proteins in IRRB, unlike many of their orthologs in ionizing-radiation-sensitive bacteria (IRSB), present a strong ability to effectively repairs damage incurred to DNA.

In this work, we study the basal DNA repair protein of IRRB and IRSB to solve the problem of phenotypic prediction in IRRB. 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 phenotypic prediction in IRRB as a multiple instance learning problem (MIL). In a MIL setting, the training data is available only as set of bags of instances with labels for the bags. In our context, bacteria represent bags and repair proteins of each bacterium represent instances.

Many multiple instance learning algorithms have been developed to solve several problems such as predicting types of Protein-Protein Interactions (PPI) [19] and drug activity prediction [5], mainly including Diverse Density [9], Citation-kNN and Bayesian-kNN [18]. Diverse Density (DD) was proposed in [9] 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 [18], 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, we can predict the label of an unseen bag using the *k-NN* algorithm.

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 be represented as a set of motifs [2, 14] or as a vector in a vector space spanned by these motifs [15]. However, the use of this technique is not suitable in the context of phenotypic prediction of IRRB. 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 bag of proteins are putting 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 case.

In this paper, we propose a MIL approach for predicting IRRB using proteins implicated in basal DNA repair in IRRB. We used a local alignment technique to measure the similarity between protein sequences of the studied bacteria to predict ionizing-radiation-resistant bacteria. To the best of our knowledge, this is the first work which proposes an *in silico* approach for phenotypic prediction in IRRB.

The remainder of this paper 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 and possible future directions 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. [4] 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. [4], in a traditional setting of machine learning, an object m is represented by a feature vector (an instance) which is associated to 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.

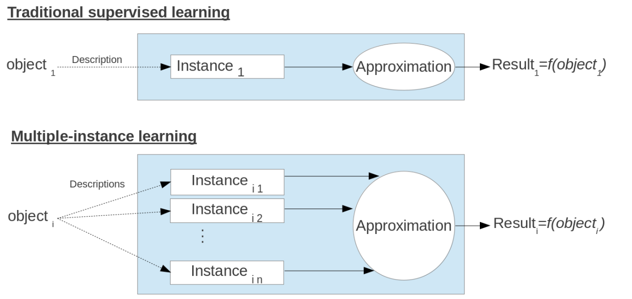

Figure 1: Differences between traditional supervised learning and multiple instance learning.

In our work, we are interested to a specific bacteria family with high radioresistance to ionizing radiation and tolerance of desiccation. This family contains 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 a 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 ionizing-radiation-resistant bacteria. The proposed algorithm focuses on discriminating bags by the use of local alignment technique to measure the 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 M: 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: Class $R = IRRB \text{ or } IRSB$
1: for all $p_{qi} \in Q$ do
2:    for all $X_j$ do
3:       $M_{ij} \leftarrow LocalAlignment(p_{qi}, p_{X_j i})$   //$X_j = \{p_{j1}, \cdots, p_{jk}\}$ and $p_{X_j i}$ is the protein number $i$ of bacterium $X_j$
4:    end for
5: end for
6: $R \leftarrow Aggregate(M)$
7: return $R$
```

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

1. For each protein sequence p_{qi} in the query bag Q, MIL-ALIGN computes the corresponding alignment scores with each protein of bacteria in the database (line 1 to 5).
2. Store alignment scores of all protein sequences of query bacterium into a matrix M (line 3). Line i of M corresponds to a score vector of protein p_{qi} against all proteins $p_{X_j i}$ of X_j with $1 \leq j \leq n$. Element M_{ij} corresponds to the alignment score of protein p_{qi} of Q with protein $p_{X_j i}$ of bacterium X_j.
3. Apply an aggregation method to S in order to compute the final prediction result R (line 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 [8]. We initiated our analyses by retrieving orthologous proteins implicated in basal DNA repair in IRRB with fully sequenced genomes.

Table 1 presents the used IRRB and IRSB.

Table 1: Experimental set of Bacteria

ID	Bacterium	Phenotype
B1	*Acinetobacter radioresistens* SH164	IRRB
B2	*Kineococcus radiotolerans* SRS30216	
B3	*Methylobacterium radiotolerans* JCM 2831	
B4	*Deinococcus maricopensis* DSM 21211	
B5	*Gemmata obscuriglobus* UQM 2246	
B6	*Deinococcus proteolyticus* MRP	
B7	*Truepera radiovictrix* DSM 17093	
B8	*Acinetobacter radioresistens* SK82	
B9	*Escherichia coli* OP50	IRSB
B10	*Neisseria gonorrhoeae* MS11	
B11	*Neisseria gonorrhoeae* PID1	
B12	*Neisseria gonorrhoeae* DGI18	
B13	*Pseudomonas putida* S16	
B14	*Thermus thermophilus* SG0.5JP17-16	

For our experiments, we constructed a training set containing 14 bags (8 IRRB and 6 IRSB). Each bag contains at most 30 instances which correspond to proteins implicated in basal DNA repair in IRRB (see Table 2). Protein sequences were downloaded from the FTP website of the curated database SwissProt [1].

3. RESULTS AND DISCUSSION

3.1 Experimental techniques

The computations were carried out on a duo CPU 2.86 GHz PC with 2 GB memory, operating on Ubuntu Linux. In the classification process, we used the Leave-One-Out (LOO) technique [7] 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 BLAST tool [1] for computing local pairwise alignments. We implemented and tested two aggregation methods with MIL-ALIGN: the *Sum of Maximum Scores* method and the *Weighted Average of Maximum Scores* method.

Sum of Maximum Scores (SMS). For each protein in the query bacterium, we traverse the corresponding line of M 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 $total_R(M)$). Otherwise, it adds max_S to the total score of IRSB (which we call $total_S(M)$). When all the selected proteins were traversed, the SMS method compares the total scores of IRRB and IRSB. If $total_R(M)$ is greater than $total_S(M)$, classification refers IRRB. Otherwise, classification refers IRSB.

Below, we formally define the SMS method:

$$\text{SMS}(M) = \begin{cases} IRRB, & \text{if } total_R(M) \geq total_S(M), \\ IRSB, & \text{otherwise,} \end{cases}$$

where

[1] http://www.uniprot.org/downloads

Table 2: Replication, repair, and recombination proteins related to ionizing-radiation-resistant bacteria

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

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

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 traverse the corresponding line of M 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 $max_R(M)$) and the maximum score among the scores of alignments against IRSB (which we call $max_S(M)$). It then compares these scores. If the $max_R(M)$ is greater than $max_S(M)$, it adds $max_R(M)$ 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 $max_S(M)$ 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 traversed, we compare the average of total scores of IRRB (which we called $avg_R(M)$) and the average of total scores of IRSB (which we called $avg_S(M)$). If $avg_R(M)$ is greater than $avg_S(M)$, prediction refers IRRB. Otherwise, classification refers IRSB.

Below, we formally define the WAMS method:

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

where

- $avg_R(M) = total_R(M)/num_M$, and
- $avg_S(M) = total_S(M)/num_M$,

and

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

where w_i is the weight of the protein p_i.

3.2 Results

In order to simulate traditional setting of machine learning in the context of predicting IRRB, we conducted a set of experiments with MIL-ALIGN by selecting just one protein for each bacterium in the training set. Each experiment consists of aggregating alignment scores between a protein sequence of a query bacterium and the corresponding protein sequences of each bacterium in the learning database. We present in Table 3 classification 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 only 22 experiments (with only 22 proteins). This is due to the fact that experiments on proteins which are not expressed at least for

Table 3: Classification results with the traditional setting of machine learning

Protein	Dataset		Accuracy (%)	Sensitivity (%)	Specificity (%)
	IRRB	IRSB			
DNA primase	8 (B1 B2 B3 B4 B5 B6 B7 B8)	6 (B9 B10 B11 B12 B13 B14)	**85.7**	87.5	83.3
Replicative DNA helicase	8 (B1 B2 B3 B4 B5 B6 B7 B8)	6 (B9 B10 B11 B12 B13 B14)	78.5	85.7	71.4
DNA topoisomerase I	8 (B1 B2 B3 B4 B5 B6 B7 B8)	6 (B9 B10 B11 B12 B13 B14)	78.5	85.7	71.4
DNA gyrase, subunit A	8 (B1 B2 B3 B4 B5 B6 B7 B8)	6 (B9 B10 B11 B12 B13 B14)	71.4	75	66.6
Endonuclease III	8 (B1 B2 B3 B4 B5 B6 B7 B8)	6 (B9 B10 B11 B12 B13 B14)	71.4	70	75
Formamidopyrimidine-DNA glycosylase	8 (B1 B2 B3 B4 B5 B6 B7 B8)	6 (B9 B10 B11 B12 B13 B14)	71.4	75	66.6
RecA Protein	8 (B1 B2 B3 B4 B5 B6 B7 B8)	6 (B9 B10 B11 B12 B13 B14)	64.2	66.6	60
DNA polymerase III, α subunit	8 (B1 B2 B3 B4 B5 B6 B7 B8)	6 (B9 B10 B11 B12 B13 B14)	57	66.6	55.5
Excinuclease ABC, subunit A	8 (B1 B2 B3 B4 B5 B6 B7 B8)	4 (B9 B10 B11 B12 B13 B14)	75	87.5	60
DNA helicase RecG	5 (B1 B4 B6 B7 B8)	6 (B9 B10 B11 B12 B13 B14)	**90.9**	83.3	100
Excinuclease ABC, subunit C	6 (B1 B2 B3 B5 B7 B8)	5 (B9 B10 B11 B12 B13)	**81.8**	100	71.4
Transcription-repair coupling factor	6 (B1 B2 B3 B5 B7 B8)	5 (B9 B10 B11 B12 B14)	72.7	71.4	75
DNA polymerase III, τ/γ subunit	6 (B2 B3 B4 B5 B6 B7)	5 (B9 B10 B11 B13 B14)	72.7	80	66.6
DNA gyrase, subunit B	5 (B1 B2 B3 B5 B8)	6 (B9 B10 B11 B12 B13 B14)	63.6	60	66.6
Holliday junction resolvase	4 (B1 B2 B3 B4 B6)	6 (B9 B10 B11 B12 B13 B14)	70	66.6	71.4
DNA polymerase III, ϵ subunit	6 (B1 B2 B3 B4 B6 B8)	3 (B9 B13 B14)	77.7	83.3	66.6
Excinuclease ABC, subunit B	6 (B1 B2 B3 B5 B7 B8)	3 (B9 B12 B13)	44.4	66.6	33.3
RecF protein	5 (B1 B2 B4 B6 B7)	3 (B9 B13 B14)	75	80	66.6
A/G-specific adenine glycosylase	7 (B1 B3 B4 B5 B6 B8)	1 (B13)	75	85.7	0
Single-stranded DNA-binding protein	6 (B1 B4 B5 B6 B7 B8)	2 (B9 B13)	50	66.6	0
Ribonuclease HII	2 (B1 B8)	5 (B9 B10 B11 B12 B13)	**85.7**	66.6	100
DNA-directed DNA polymerase	4 (B2 B3 B5 B6)	1 (B13)	60	75	0

one IRRB bacterium and for one IRSB bacterium were not conducted. Results in Table 3 show that the use of our algorithm with just one instance for each bag in the learning database allow good accuracy values especially with some specific proteins. However, almost all results were generated without the whole set of bacteria. In fact, when a protein is not expressed in a specific bacterium, we do not use the bacterium in the learning database. For example, the protein *DNA helicase RecG* is expressed for only 11 bacteria (5 IRRB and 6 IRSB) from the set of 14 bacteria of the training set (see Table 1).

In order to study the incorrectly classified bacteria with the traditional setting of machine learning, we computed for each bacterium in the learning database, the percentage of experiments that fail to correctly classify the bacterium (see Table 4).

As shown in Table 4, some bacteria present high rates of failed predictions. This means that we fail to correctly predict the phenotype of those bacteria with most proteins. On the other hand, the results illustrated in Table 4 may help to understand some characteristics of the studied bacteria. For example, the *Thermus thermophilus SG0.5JP17-16* bacterium presents a high rate of failed predictions (83.33 %). It mean that in most cases, *Thermus thermophilus SG0.5JP17-16* is predicted as IRRB. This result shows that *Thermus thermophilus* SG0.5JP17-16 might allow a strong ability for DNA protection and repair mechanisms and confirm the *in vitro* results presented in [10], [12] and [11].

In order to study the importance of considering the problem of classifying ionizing-radiation-resistant bacteria as a multiple instance learning problem, we present in Table 5 the experimental results of MIL-ALIGN using a set of 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 a high values of sensitivity and specificity especially with the WAMS aggregation method. This can be explained by a good choice of proteins to represents the studied bacteria. For example, with the combination of DNA primase (P8), DNA helicase RecG (P24) and A/G-specific adenine glycosylase (P28) and with the WAMS aggregation method, we have **92.8 %** of accuracy, 88.8 % of sensitivity and 100 % of specificity. We do not exceed these values in all the cases presented in Table 3. This result can be explained by the complementarity between DNA primase (P8), DNA helicase RecG (P24) and A/G-specific adenine glycosylase (P28). In fact, DNA primase (P8) and DNA helicase RecG (P24) present good accuracies in a traditional supervised learning setting (see Table 3) and A/G-specific adenine glycosylase (P28) presents the ability to correctly classify bacteria that are incorrectly classified with DNA primase (P8) and DNA helicase RecG (P24).

Table 5 suggests that ionizing resistant radiation is better reflected in three biological processes : (i) synthesis by the DNA primase (P8) of small RNA primers for the Okazaki fragments on both template strands at replication forks during chromosomal DNA synthesis; (ii) maintaining genomic stability and integrity by controlling recombination events, and repairing DNA damage by the DNA helicase RecG (P24); and (iii) repair of G-A mispairs and oxidatively damaged form of guanine by MutY (P28).

The high values of specificity presented by MIL-ALIGN show the ability of MIL-ALIGN to identify negative bags (IRSB).

4. CONCLUSION

In this paper, we addressed the issue of classifying ionizing-radiation-resistant bacteria (IRRB). 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 first results of MIL-ALIGN are satisfactory.

In the future work, we will study the performance of the proposed approach to improve its efficiency. Also, we will study the use of a priori knowledge to improve the efficiency

Table 4: Percentage of failed classifications

Phenotype	Bacterium	Rate of failed predictions (%)
IRRB	Acinetobacter radioresistens SH164	15
	Kineococcus radiotolerans SRS30216	33.33
	Methylobacterium radiotolerans JCM 2831	**77.77**
	Deinococcus maricopensis DSM 21211	0
	Gemmata obscuriglobus UQM 2246	47.05
	Deinococcus proteolyticus MRP	5.88
	Truepera radiovictrix DSM 17093	27.77
	Acinetobacter radioresistens SK82	11.11
IRSB	*Escherichia coli* OP50	20
	Neisseria gonorrhoeae MS11	6.25
	Neisseria gonorrhoeae PID1	0
	Neisseria gonorrhoeae DGI18	0
	Pseudomonas putida S16	47.61
	Thermus thermophilus SG0.5JP17-16	**83.33**

Table 5: Experimental results of MIL-ALIGN with leave-one-out-based evaluation technique

Used proteins	Aggregation method	Accuracy (%)	Sensitivity (%)	Specificity (%)
All proteins	SMS	71.39	75	66.6
	WAMS	78.5	72.7	100
DNA Polymerase proteins	SMS	71.39	75	66.6
	WAMS	78.5	77.7	80
Replication complex proteins	SMS	71.39	75	66.6
	WAMS	78.5	77.7	80
Other DNA-associated proteins	SMS	78.5	85.7	71.4
	WAMS	78.5	72.7	100
P8 P24 P28	SMS	**85.7**	87.7	83.3
	WAMS	**92.8**	88.8	100
P6 P7 P8 P24 P28	SMS	**85.7**	87.7	83.3
	WAMS	**92.8**	88.8	100

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 to the high resistance of IRRB to the ionizing radiations and desiccation. In fact, many antioxidant enzymes may play important roles in scavenging free radicals caused by irradiation [6]. Finally, we will study possible extensions of our approach with other learning models [3].

Acknowledgement

This work was supported by the French-Tunisian project: General Directorate of Scientific Research in Tunisia (DGRST)/National Center for Scientific Research in France (CNRS) [IRRB 11/R 14-09]. We would like to thank Jan-Thierry Wegener for English proofreading. We also would like to thank Valerian Callens for technical help.

5. REFERENCES

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, Oct. 1990.

[2] A. Ben-Hur and D. L. Brutlag. Remote homology detection: a motif based approach. In *ISMB (Supplement of Bioinformatics)*, pages 26–33, 2003.

[3] W. Dhifli, R. Saidi, and E. M. Nguifo. Mining representative unsubstituted graph patterns using prior similarity matrix. *CoRR*, abs/1303.2054, 2013.

[4] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, Jan. 1997.

[5] G. Fu, X. Nan, H. Liu, R. Y. Patel, P. R. Daga, Y. Chen, D. Wilkins, and R. J. Doerksen. Implementation of multiple-instance learning in drug activity prediction. *BMC Bioinformatics*, 13(S-15):S3, 2012.

[6] N. Gao, B.-G. Ma, Y.-S. Zhang, Q. Song, L.-L. Chen, and H.-Y. Zhang. Gene Expression Analysis of Four Radiation-resistant Bacteria. *Genomics Insights*, 2:11–22, 06 2009.

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

[8] K. Liolios, K. Mavromatis, N. Tavernarakis, and N. C. Kyrpides. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 36(suppl 1):D475–D479, Jan. 2008.

[9] O. Maron and T. L. Pérez. A Framework for Multiple-Instance Learning. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, pages 570–576, Cambridge, MA, 1998. The MIT Press.

[10] H. Nishida and M. Nishiyama. Evolution of lysine biosynthesis in the phylum deinococcus-thermus. *Int J Evol Biol*, 2012, 2012.

[11] N. Ohtani, M. Tomita, and M. Itaya. An extreme thermophile, Thermus thermophilus, is a polyploid bacterium. *J Bacteriol*, 192(20):5499–505, 2010.

[12] M. Omelchenko, Y. Wolf, E. Gaidamakova, V. Matrosova, A. Vasilenko, M. Zhai, M. Daly, E. Koonin, and K. Makarova. Comparative genomics of Thermus thermophilus and Deinococcus radiodurans : divergent routes of adaptation to thermophily and radiation resistance. *BMC Evolutionary Biology*, 5(1):1–22, 2005.

[13] M. V. Omelchenko, Y. I. Wolf, E. K. Gaidamakova, V. Y. Matrosova, A. Vasilenko, M. Zhai, M. J. Daly, E. V. Koonin, and K. S. Makarova. Comparative genomics of Thermus thermophilus and Deinococcus radiodurans: divergent routes of adaptation to thermophily and radiation resistance. *BMC Evol Biol*, 5:57, 2005.

[14] R. Saidi, S. Aridhi, E. Mephu Nguifo, and M. Maddouri. Feature extraction in protein sequences classification: a new stability measure. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '12, pages 683–689, New York, NY, USA, 2012. ACM.

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

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

[17] H. Sghaier, S. Thorvaldsen, and N. Saied. There are more small amino acids and fewer aromatic rings in proteins of ionizing radiation-resistant bacteria. *Annals of Microbiology*, pages 1–9, 2013.

[18] J. Wang and J.-D. Zucker. Solving the Multiple-Instance Problem: A Lazy Learning Approach. In *Proc. 17th International Conf. on Machine Learning*, pages 1119–1125. Morgan Kaufmann, 2000.

[19] H. Yamakawa, K. Maruhashi, and Y. Nakao. Predicting Types of Protein-Protein Interactions Using a Multiple-Instance Learning Model. In T. Washio, K. Satoh, H. Takeda, and A. Inokuchi, editors, *New Frontiers in Artificial Intelligence*, volume 4384 of *Lecture Notes in Computer Science*, pages 42–53. Springer Berlin Heidelberg, 2007.