De novo sequence-based method for ncRPI prediction using structural information

Michele Leone*
Dep. Elettronics, Information and Bioengineering
Politecnico di Milano
Milano, Italy
michele.leone@polimi.it

Marta Galvani*
Mathematics Dep.
University of Pavia
Pavia, Italy
marta.galvani@unipv.it

Marco Masseroli
Dep. Elettronics, Information and Bioengineering
Politecnico di Milano
Milano, Italy
marco.masseroli@polimi.it

***Abstract*—Improving knowledge of RNA-binding protein targets is focusing the attention towards non-coding RNAs (ncRNAs), i.e., transcripts not translated into a protein; they are associated with a wide range of biological functions through different molecular mechanisms, usually concerning the interaction with one or more protein partners. Recent studies confirmed that the alteration of ncRNA-protein interactions (ncRPIs) may be linked to various pathologies, including autoimmune and metabolic diseases, neurological and muscular disorders and cancer. Unfortunately, the limited number of structurally characterized RNA-protein complexes available does not allow to accurately establish their role in cellular processes and diseases. Experimental analyses to identify ncRNA-protein interactions are providing a large amount of valuable data, but these experiments are expensive and time consuming. For these reasons, computational approaches based on machine learning techniques appear very useful to predict ncRPIs. Yet, there are still few studies regarding the prediction of ncRPIs, especially including the use of higher-order structures, which are of vital importance for the ncRPI functions.**

In this work, a new computational method for non-coding RNA-protein interaction prediction is developed; from sequence data it derives more accurate information about the secondary structure of the molecules involved in such interactions, which it then uses in the prediction. Obtained results suggest that the use of machine learning techniques, together with considering also information on higher-order structures of ncRNAs and proteins, can be useful to better predict ncRPIs.

***Index Terms*—RNA Protein Interaction, Computational Proteomics, Machine Learning, Classification Models**

I. Introduction and Motivation

Non-coding RNAs (ncRNAs) are transcripts not involved in the translational mechanism of protein synthesis. However, the hypothesis that these molecules would play an important biological role in cell development and metabolism has begun to be supported, as in [1] and [2]. The human transcriptome is turning out to be much more than a simple series of genes that encode proteins and their isoforms deriving from alternative splicing. Indeed, the regulation of gene expression in eukaryotic organisms is also linked to many other factors, such as overlapping genes, natural antisense RNA, and the non-coding RNA [3], subdivided into two main classes according to the size of the transcript: small and long ncRNA [4], [5]. These two categories are both transcripts that generally lack significant open reading frames. However, the first class consists of polymeric molecules of no more than 200 nucleotides, while the second class includes molecules that can reach 100 kilobases in length. See Table I for a complete overview of the components of the ncRNA class.

TABLE I
Main types of ncRNA

Type	Subclass	Symbol
Small ncRNA	transfer RNA	tRNA
	microRNA	miRNA
	ribosomal RNA 5S and 5.8S	rRNA
	small interfering RNA	siRNA
	small nucleolar RNA	snoRNA
	small nuclear RNA	snRNA
Long ncRNA	ribosomal RNA 18S and 28s	rRNA
	pseudogenes	none
	long or large intergenic ncRNA	none
	long intronic ncRNA	lincRNA
	antisense RNA	aRNA

Previous studies on ribozymes have shown that particular structures within these RNAs are important for their functions [6]. Similarly, different ncRNA structures are critical for the mechanism of the steroidal RNA receptor [7]. Therefore, the prediction of stem-loop secondary structures is useful for identifying the different functions of the ncRNAs. For this purpose, there are a number of in-silico tools widely used, such as RNAfold or Mfold, which, given the nucleotide sequence in input, return the most probable least energy secondary structures, considering that the biologically correct structure is usually sub-optimal rather than the minimum energy one (Zuker's algorithm) [8]. However, a very large number of RNA secondary structures remains to be determined, as well as their interaction partners. For these cases, computational approaches are very useful, especially because experimental methods for determining ncRNA-protein interactions (ncRPIs)

*Both authors contributed equally to this research.

are expensive and time consuming.

The current number of available ncRNA-binding protein data allows the creation of computational models that can be used to predict new interactions. Most of these methods has been developed principally for the prediction of interactions between proteins and RNAs; yet, the hypothesis is that the binding proficiency could be similar, regardless of the RNA type. Therefore, in silico methods may be fundamental for the characterization of ncRNAs, whose experimental data are less abundant and often experimentally difficult to obtain.

Most of the developed algorithms use, as training features, three-dimensional protein structures linked to a fragment of RNA extracted from the Protein Data Bank [9]. From that, a series of physico-chemical characteristics (for example, orientation of hydrogen bonds, Van der Waals forces and the secondary structure) can be used to describe each protein-RNA pair. Characteristic vectors of known interactions are compared with those calculated on control sets of non-interacting pairs; statistical methods, machine learning algorithms, or ad hoc scoring systems are then used to evaluate the potential link between new interacting pairs.

Different approaches have been proposed with this aim: one of the first methods for predicting RNA-protein interactions [10] used features such as the protein localization, Gene Ontology annotations, and gene interactions, which may not be available, or be difficult to be calculated, for some RNA or proteins; in fact, the method performance on RNA binding proteins not included in the training set is very variable. The lncPro method [11], tested on the NPInter database of non-coding RNA-associated interactions [12], gave good predictive accuracy; its application to the entire human proteome revealed a massive presence of nuclear proteins, consistent with the observation that many ncRNAs reside in the nucleus [13]. The Oli suite [14] was trained with high-throughput data, including PAR-clip experiments, and utilized features that include sequence characteristics, secondary structure prediction, and the presence of structural patterns. The main problem with this method is the use of descriptors that cannot be easily applied to cases that are not present in the training set. RPISeq [15] used only information about the sequences of proteins and RNAs; Random Forest (RF) and Support Vector Machines (SVM) classifiers were trained using the 3-mer and 4-mer conjugate triad features of the amino acid and nucleotide sequences, respectively. [16] proposed an approach based on the Naive Bayes (NB) and Extended NB (ENB) classifiers, using the same datasets and characteristics very similar to those reported in [15]. The PRIPU method [17] differs from the others because it uses statistical learning methods trained exclusively with positive examples. The catRAPID [18], RPI-Pred [19] and rpiCOOL [20] methods included also structural information for the RNA-protein interaction (RPI) prediction. Yet, they derived this information from experimental analyses; thus, this approach is recognized to be weak for biomolecules that lack data experimentally obtained.

Recognizing the characteristics of structural data when not experimentally available, and using them for the prediction of interactions, is of vital importance to better understand the interaction molecular mechanisms and biological functions. In the literature there are still few studies concerning the prediction of ncRPIs that, starting from sequence information, are able to predict whether a given type of interaction takes place or not, considering also structural information not obtained through experimental results, but deriving them from sequence information. In this work, we developed a new sequence-based method that, retrieving structural information de novo from primary structure of small and long non-coding RNA, uses them to predict potential ncRNA-protein partners.

The remain of this paper is structured as follow: in Section II the whole workflow of our novel prediction method is presented, and each step is explained in details, from the data collection to the interaction prediction. Results are shown in Section III, and Section IV contains the final conclusion.

II. Step-wise workflow of the proposed method

The purpose of this work is to create an efficient and accurate method for the prediction of ncRNA-protein interactions using sequence and structural information. Figure 1 shows the workflow of this method, which includes four phases: (1) Extraction of non-coding RNA-protein pairs from the NPInter database, removal of the redundant RNA-protein pairs and selection of the interacting pairs on the basis of well-defined characteristics, with subsequent generation of a dataset containing negative interactions. (2) Extraction of sequence characteristics of both proteins and RNAs to develop the prediction method. (3) Extraction of both protein and RNA structure characteristics. (4) Prediction of RNA-protein interactions. All these phases are explained in details in the next sub-sections.

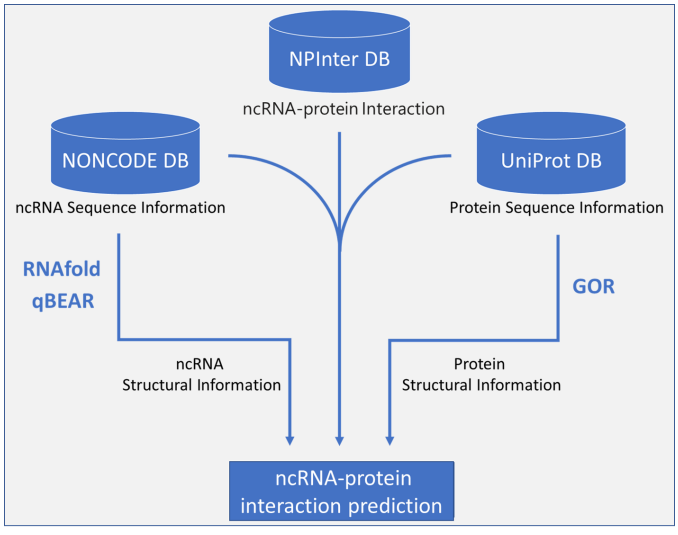

Fig. 1. Workflow for ncRNA-protein interaction prediction using secondary structures obtained starting from sequence information.

A. Interaction data collection and negative interaction set creation

NPInter was chosen as a reference database, as it integrates experimentally verified interactions between non-coding RNAs and other biomolecules. ncRNAs were annotated with the NONCODE [21] ID, or with the miRBase [22] ID, while proteins were annotated with the ID of UniProt [23], RefSeq [24], or UniGene [25] depending on the type of molecule (protein, RNA encoding protein, or DNA, respectively). Once the IDs of the interacting pairs were obtained, the related sequences were extracted. In particular, the ncRNA sequences were obtained from the NONCODE website (http://www.noncode.org/), while those of the proteins were taken from the UniProt one (https://www.uniprot.org/).

Three datasets were adopted to test the method proposed in this paper: RPI2241 [15], widely used in the field of predicting ncRNA-protein interaction [11], [15], [26], and two datasets that we generated, named RPI4756 and RPI1020. RPI2241 is composed of non-redundant experimentally-validated ncRNA-protein interaction pairs extracted from the three-dimensional structures of RNA-protein complexes within the Protein Data Bank (PDB) [9]. RPI4756 and RPI1020 are composed of non-redundant RNA-protein interactions, extracted from the NPInter database, regarding respectively long non-coding RNA and small non-coding RNA according to their RNA sequence length. The summary information of these datasets is listed in Table II.

TABLE II
CHARACTERISTICS OF THE THREE CONSIDERED DATASETS

Dataset	Interactions	RNAs	Proteins	Negatives
RPI2241	2,077	801	1,912	3,221
RPI4756	4,756	1,160	70	5,554
RPI1020	1,020	165	237	1,824

To predict the probability of an RNA-protein interaction, it is necessary to generate, for both long and small non-coding RNA classes, a balanced dataset containing both *interacting* and *non-interacting* RNA-protein pairs. To generate the *non-interacting* pairs, for each RNA class we randomly paired RNAs and proteins by selecting them among those in the class dataset, and then removing generated pairs similar to those known to interact. In other words, if RNA A interacts with protein B and the latter one is similar to another protein C, then we cannot generate a non-interacting RNA-protein pair involving A and C.

Two RNAs (or proteins) sequences are marked as similar if their similarity is more than 30%. We used the Needleman-Wunsch algorithm [27] to find the similarity between two sequences. This algorithm is based on the pairwise sequence alignment, a common method used to identify the regions of similarity between two DNA/RNA or protein sequences. It is a dynamic programming method that divides a large problem into a series of smaller ones and, after solving optimally the smaller problems, it uses their solutions to construct an optimal solution for the original problem. The Percent Sequence Identity of each pair of sequences is then evaluated from the pairwise alignment, by dividing the number of identical positions in the two aligned sequences by the number of aligned positions and then multiplying it by 100.

B. Sequence information encoding

For representing RNA sequences the standard 4-letter alphabet was used. Proteins were instead coded with the conjunct triad feature (CTF) representation, previously used in [28]. In this representation, the 20 amino acids are classified into 7 groups according to their dipole moments and the volume of their side chains: {A, G, V}, {I, L, F, P}, {Y, M, T, S}, {H, N, Q, W}, {R, K}, {D, E}, and {C}. Each protein sequence was then represented using the derived reduced 7-letter alphabet.

C. Structural information extraction and enconding

RNA functions are strictly related to how RNAs fold into the space, which can be simplified in their secondary structure. This is simpler to compute than the 3D fold, for which there are no reliable algorithms yet, especially for medium and large size RNAs, probably due to the low number of available crystallographic or NMR structures of RNA molecules. Many RNA-binding proteins can recognize, in their RNA partner, some structures for which they have affinity; these structures can be approximated by the RNA secondary structure. So, entering a description of the RNA secondary structure in an RNA-protein interaction predictor can improve prediction performances by adding important information about the phenomenon.

The structure of an RNA can be considered as a combination of the different structural elements that compose it, each of which contributes independently to the free energy of the overall structure [29]. The different possible structural elements in an RNA are the following:

- Single-stranded regions;
- Stem: contiguous pairs of complementary nucleotides;
- Hairpin loop: a loop at the end of a stem;
- Bulge loop: interruption of a stem in one side;
- Internal loop: interruption of a stem on both sides;
- Multi-branched loop: interruption between three or more divergent stems.

Formally, the secondary structure of an RNA molecule can be defined as the set of base pairs $s_{i,j}$ between the nucleotides i and j, always with $i < j$. Classical algorithms, such as [8] and [30], for predicting the secondary structure of RNAs from their primary structure are not able to handle these particular pairing. For this reason, in this work two different tools are used to represent the information of the secondary structure: RNAfold and BEAR.

RNAfold is one of the basic program of the Vienna RNA package, which predicts the minimum free energy (MFE) of the secondary structure of individual sequences using the dynamic programming algorithm of Zuker and Stiegler [31]. In addition to the MFE prediction, the probabilities of pairing between the sequence bases are calculated using the algorithm introduced in [32]. In this way, it is possible to represent

the information of the secondary structure of the sequence in a simple vector consisting of points and brackets, where each character represents a nucleotide. The open parentheses indicate that the nucleotide is coupled to another nucleotide before it. Closed parentheses indicate that a base is coupled to another base after it. The points, on the other hand, indicate an unmatched base.

BEAR (Brand nEw Alphabet for RNAs) is a new encoding that allows the information of the secondary structure of a sequence to be stored within a single string of characters, by unequivocally linking the information of the secondary structure to each nucleotide in the sequence. In BEAR coding, each letter represents information about the length and the type of the structure it belongs. Unlike the dot-bracket notation, the assignment of each nucleotide to the element of the secondary structure to which it belongs allows distinguishing nucleotides described with the same symbol (a point or a bracket) but belonging to a different element of the secondary structure. A BEAR representation of the RNA secondary structure is a string, having the same length as the associated RNA sequence, but encoding, at the same time, more structural information than the standard point-bracket notation [33]. For example, BEAR encoding specifies whether the unpaired nucleotides belong to a hairpin loop or to a bulge. There are two versions of the alphabet used by BEAR, the complete one and the one reduced to 18 characters, called qBEAR. We adopted the latter one, converting the RNA sequences by using the encoding of the qBEAR alphabet. Table III reports the employed conversion rules from the BEAR to the qBEAR alphabet.

TABLE III
CONVERSION TABLE FROM BEAR TO QBEAR ALPHABET

BEAR	qBEAR	Description
a b c d e	Z	short stem
f g h i	A	medium stem
=	Q	long stem
j k l m n o p q r	X	short loop
s t u v w x y z	S	medium loop
^	W	long loop
! " # $ % 2 3 4 5 6	C	short internal loop
& ′ ( ) 7 8 9	D	medium internal loop
+ >	E	long internal loop
[]	B	bulge
{ }	G	bulge branch
:	T	branch
A B C D E	V	short stem branch
F G H I	F	medium stem branch
J	R	long stem branch
K L M N Y Z ∼ ?	N	short internal loop branch
O P Q R S _ | / \	H	medium internal loop branch
T U W Y Z @	Y	long internal loop branch

For the protein secondary structure prediction, we used the GOR (Garnier-Osguthorpe-Robson) V algorithm [34], based on the probability parameters derived from protein tertiary structures obtained by X-ray crystallography. This method considers not only the propensities of individual amino acids to form particular secondary structures, but also the conditional probability of an amino acid to form a secondary structure given that its immediate neighbours have already formed that structure [35]. In addition, this procedure is able to predict whether an amino acid belongs to an alpha helix, a beta sheet, or a coil, considering 17-amino-acid sequence windows. This approach was proven to be about 73.5% accurate; it is one of the most accurate methods for secondary structure assignment that only use a single sequence. A GOR V representation of the protein secondary structure is a string, having the same length as the associated protein sequence, but converted in a new 3-letter alphabet {H, E, C} depending on the amino acid belonging respectively to an helix, strand, or coil.

D. Feature extraction and representation

Our RNA-protein interaction prediction model is built considering both sequence and structural information of RNAs and proteins. Both primary and secondary structures are complex elements to be directly considered as features in a binary classification model; so, for this purpose a new representation for them is needed.

From the sequence information it is clear that the interaction between an RNA and a protein does not involve just an amino acid and a nucleotide, but it involves a set of neighbouring nucleotides and a set of neighbouring amino acids. This consideration justifies the idea of treating the primary structure of each RNA and each protein as the combination of small portions of the sequence, i.e., the combination of different *k-mers*. Previous works, such as [15] and [36], considered as feature space for the primary structure of a sequence the frequency matrix of all possible *k-mers* of the sequence. As we would like to identify the most influencing *k-mers* in the interaction prediction, in this work we built and considered the *term frequency-inverse document frequency (tf-idf)* matrix. The *tf-idf* value is widely used in the information retrieval to understand how important a word is to a document in a collection, or corpus, of documents. In our case, a word is a *k-mer*, a document is an RNA (or protein) sequence, and the corpus is the sequence dataset under analysis. Thus, in our *tf-idf* matrix each row is an RNA (or protein), each column represents a k-mer, and each matrix element is their *tf-idf* value.

The *tf-idf* is a numerical statistic composed of the Term Frequency (*tf*) and the Inverse Document Frequency (*idf*), respectively expressed as:

$$tf_{i,j} = \frac{n_{i,j}}{|d_j|}$$

$$idf_i = log \frac{|D|}{|\{d : i \in d\}|}$$

where, in our case, $n_{i,j}$ and $|d_j|$ are respectively the number of times a *k-mer* i occurs and the number of all *k-mers* in the observed RNA (or protein) j, $|D|$ is the total number of RNAs (or proteins) in the dataset under analysis, and the denominator of idf_i is the number of RNAs (or proteins) in the dataset that contain the observed *k-mer* i. The *tf-idf* value is then evaluated as the product between the tf and idf values.

Looking at the formulas, we can observe that the *tf-idf* value increases proportionally to the number of times a *k-mer* appears in the observed sequence and is offset by the frequency of the *k-mer* in the sequences of the dataset. This helps to adjust for the fact that some *k-mers* can be more frequent in general and not specifically for a particular interaction.

Concerning the secondary structure, we analysed differently RNAs and proteins. As explained in the previous Section, for the RNA secondary structure we used the qBEAR representation, obtaining, for each RNA, a character string having the same length as the associated RNA sequence and composed of 18 different characters. We then evaluated the *tf-idf* value for each sub-structure in the sequence, considering the T value as separator, since it represents a non-paired nucleotide. In this way it is possible to represent the information about the molecule three dimensional space in a mono-dimensional vector.

The protein secondary structure is instead represented using the GOR V representation, converted into the 3-letter alphabet: H, E, C, as explained in the previous Section. Thus, each protein structure is represented as a string of repeated letters. To analyse them, we identified subsequences by splitting each string where a changing of letter is present (for example, the string EEEHHC is splitted as: EEE HH C); then, for each subsequence we evaluated its *tf-idf* value, obtaining the corresponding matrix of all the considered sequences.

E. Binary prediction model

The problem of RNA-protein interaction prediction can be viewed as a binary classification problem, where the two possible states are: *interaction* or *non-interaction*. The assumption under this approach, in predicting the interaction between proteins and RNAs, is that the composition of residues of the two interacting molecules must follow rules that determine whether the interaction is possible.

Supervised learning techniques can be used to solve the problem of identifying the probability of interaction between RNAs and proteins. Many different methods have been developed with the aim of solving binary classification tasks. Previous work, such as [15], [26] and [37], demonstrated that, for closely related studies, Support Vector Machine (SVM) and Random Forest (RF) are the best performing algorithms. Support Vector Machine [38] can model non-linear boundaries by determining a linear boundary in a larger and transformed version of the original feature space. Random Forest, introduced by [39] and extended by [40], uses a bagging procedure combining different decision trees. This method deals with over-fitting problems of a single decision tree, decreasing the variance of the model.

In this study, besides SVM and RF, we also applied the Extreme Gradient Boosted (XGBoost) method [41]; it is an implementation of gradient boosted decision trees, designed to speed up the process and improve performance. XGBoost is based on the Gradient Boosting Machine [42], a learning algorithm that uses a boosting procedure where many weak learners are combined to obtain better results. In this method the distribution of the training set in each tree is based on the performance of previously created classifiers, and a function of the performance of a classifier is used as a weight for voting. To control better over-fitting, in the XGBoost method a more regularized model formalization is introduced.

All these techniques have two main advantages: first, they can deal with an high number of features, using regularization techniques (such as in SVM and XGBoost), or selecting iteratively a random subset of features (such as in RF); second they are model-free, meaning that no a priori analytical formulation of dependence between covariates and outcome is hypothesized, allowing to model both linear and non-linear dependencies, often outperforming classical methods [43].

The model comparison is derived under a 10-fold cross validation procedure, i.e., training and validation sets are iteratively selected with a cross-validation procedure and performances are averaged, as described in [44]; in this way, we can ensure robustness of results.

When evaluating the performance of a binary prediction, it is important to balance false positives and false negatives. Several measures have been proposed in the literature to this aim. We compared the different model results by measuring the Area Under the ROC Curve (AUC) and the H-measure. The ROC curve is based on the different values of *Specificity* and *Sensitivity*, when varying between 0 and 1 the threshold for classifying an observation to one class or to the other one, and observing the assigned probability. The AUC value gives a measure, ranging between 0 and 1, of the discrimination power of a binary model. An alternative quantity to evaluate model performance is the H-measure, introduced by [45] as a new, more coherent measure, after proving that the AUC value has the weakness of needing an interpretation that uses a cost distribution that depends on the model.

To compare our results with previous works, such as [15] and [16], RNA-protein pairs were classified as interacting or non-interacting by setting the classification threshold at 0.5, and accuracy, precision, recall and F1-measure values were evaluated.

III. Results

The collected datasets used for our experiments consisted of four different variables: the two primary and secondary structures of the interacting RNAs and proteins, where the structural information are derived as explained in Section II. Negative samples, i.e., non-interacting RNA-protein pairs, were built as explained in Section II. Briefly, for each dataset a number of random pairs of RNAs and proteins were generated randomly by picking RNAs and proteins among those in the dataset; negative samples were then cleaned by deleting pairs similar to interacting ones.

To obtain an analyzable dataset, the procedure explained in Section II C was employed, building the *tf-idf* matrix of all the considered sequences. On the basis of the results obtained in preliminary tests comparing the frequencies of the normalized k-mers in the RNA sequences for different values of k, we decided to encode the RNA sequence information using a

vector of 256 (4*4*4*4) elements; each position of the vector represents the *tf-idf* value of the corresponding 4-mer contained in the sequence of the considered RNA (for example, AAUG, CGAU, CCGG). Each protein sequence, converted into the 7-letter reduced alphabet, was then represented using the *tf-idf* values of the 3-mers that it contains, obtaining a vector of 343 (7*7*7) elements. Therefore, concerning the sequence information, using this approach, each RNA-protein pair was represented as a vector of 599 elements, in which the first 256 elements represent the RNA sequence and the subsequent 343 ones the protein sequence.

Concerning the secondary structure, for RNA sequences the *tf-idf* value for each sub-structure was evaluated considering the value T as separator. For protein structure information, protein sequences were splitted as explained in previous Section, and then the *tf-idf* value was evaluated for each sub-structure. Using this procedure, the final datasets contained 11,134, 1,035 and 1,188 features, respectively for the datasets RPI4756, RPI1020 and RPI2241.

Three prediction models, i.e., Random Forest, Support Vector Machine and Extreme Gradient Boosting, were comparatively evaluated using a 10-fold cross validation procedure, to obtain more stable results. It is important to underline that these three different predictive methods are employed in the phase of RNA-protein interaction prediction of the workflow in Figure 1, while the other parts of the algorithm remain unchanged.

Model parameters were selected after validation procedure. Random Forest was trained using 1,000 trees and a number of random variables to be picked at each iteration equal to $\sqrt{\frac{NumberofFeatures}{3}}$. A polynomial Kernel of degree 5 was used for Support Vector Machine, with cost C of constraint violation for the regularization term equal to 1, and γ parameter equal to 1. The Extreme Grandient Boosting was trained with 250 boosting iterations for the dataset RPI4756 and 150 boosting iterations for the datasets RPI1020 and RPI2241, and with learning rate equal to 0.3.

Model performances were evaluated and compared using AUC and H-measure values; results are reported in Table IV.

TABLE IV
MODEL PERFORMANCE COMPARISON

Dataset	Model	AUC	H-measure	Acc[1]	Prec	F1
RPI4756	RF	0.86	0.43	0.78	0.82	0.70
	SVM	0.86	0.44	0.78	0.77	0.76
	XGBoost	0.85	0.41	0.76	0.75	0.74
RPI1020	RF	0.89	0.54	0.84	0.91	0.74
	SVM	0.88	0.52	0.81	0.79	0.72
	XGBoost	0.90	0.55	0.86	0.82	0.77
RPI2241	RF	0.86	0.44	0.80	0.82	0.67
	SVM	0.77	0.30	0.71	0.75	0.53
	XGBoost	0.85	0.43	0.80	0.76	0.71

Observing AUC and H-measure values, Table IV shows that for the dataset RPI4756 Random Forest and Support Vector Machine performances are better than the one of the XGBoost model, while XGBoost has higher performance for dataset RPI1020. However, performing the De Long test [46] and considering its p-value to compare AUC and H-measure values, it can be seen that for both datasets all the model performances are not statistically different. Conversely, for the dataset RPI2241 SVM performance is lower than the other two method ones.

Concerning Accuracy and other measures obtained setting the classification threshold at 0.5 and classifying the RNA-protein pairs as interacting or non-interacting, the methodology employed in this work is competitive with respect to previous works. In particular, we compared our model with two state-of-the-art methods, [15] and [16], applied to the dataset RPI2241. Both methods take as inputs primary sequences of RNAs and proteins. RPISeq, introduced in [15], combines k-mer features and Random Forest classifier; whereas, the method [16] predicts ncRNA-protein interactions by extracting sequence-based features to represent each RNA-protein pair and using them to train a naive Bayes model. Results comparison among the method introduced in this work and the two previous works on the same dataset RPI2241 are shown in Table V.

TABLE V
MODEL PERFORMANCE COMPARISON ON DATASET RPI2241

Model	AUC	H-measure	Acc[2]	Prec	F1
RF	0.86	0.44	0.80	0.82	0.67
SVM	0.77	0.30	0.71	0.75	0.53
XGBoost	0.85	0.43	0.80	0.76	0.71
Random Forest [15]	-	-	0.90	0.89	0.90
Naive Bayes [16]	-	-	0.76	0.72	-

Looking at the results at hand, our performances are competitive with the state-of-the-art, although they are not the best ones on the dataset RPI2241, which consists of RPI involving rRNA or ribosomal proteins; this suggests that more data is needed to improve our predictions in the case of such molecules. As for the two datasets consisting of non-coding RNAs, RPI4756 and RPI1020, better performances are observed in the second one, consisting of small non-coding RNA; this suggests that the approach needs improved features to describe information of higher structures regarding very long molecules. However, the main advantage of our work is that, while remaining competitive, additional information is introduced regarding the secondary structure of RNA and proteins. As explained in the previous section, this new information is important as it represents the de novo secondary structure; thus, it allows avoiding long and costly experiments and enables the representation of RNA and protein interactions with a higher level of complexity than considering only nucleotides and amino acid sequences.

IV. CONCLUSION

A large number of cellular processes are mediated by RNA-protein interactions; therefore, the study of RPIs is useful

[1]Acc: accuracy; Prec: precision; F1: F1-measure.

[2]Acc: accuracy; Prec: precision; F1: F1-measure.

for understanding their mechanism. In recent years, high-throughput sequencing methods have led to the discovery of numerous ncRNAs, which are able to regulate gene expression by interacting with protein partners. Studying the correct interaction partners is crucial to understand their function. Unfortunately, experimental methods for identifying RPIs are expensive and time-consuming. In this scenario, computational approaches to predict RPIs are very useful.

In this work, a new method is developed to predict RNA-protein interactions. Obtained results suggest that starting from sequence information it is possible to construct a good predictor by adding information on the secondary structure, avoiding the use of expensive and time-consuming experiments. The advantage of this procedure is also that, while still predicting well the RNA-protein interactions, new information are added allowing better understanding of the phenomenon under study. In particular, these predictions can be used to: (i) identify RNA or protein target partners, and (ii) create RNA-protein interaction networks. Thus, such an instrument could find interesting applications in the analysis of pathological variations that alter these interaction networks.

REFERENCES

[1] J. Wang, J. Zhang, H. Zheng, J. Li, D. Liu, H. Li, R. Samudrala, J. Yu, G.K. WoNg, "Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs", Nature, vol. 431, pp. 757, 2004.

[2] J.E. Wilusz, H. Sunwoo, D.L. Spector, "Long noncoding RNAs: functional surprises from the RNA world", Genes Dev., vol. 23, pp. 1494-1504, 2009.

[3] R. Jaenisch, A. Bird, "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals", Nat. Genet., vol. 33, pp. 245-254, 2003.

[4] C. A. Brosnan, O. Voinnet, "The long and the short of noncoding RNAs", Curr. Opin. Cell. Biol., vol. 21, pp. 416-425, 2009.

[5] F. F. Costa, "Non-coding RNAs: new players in eukaryotic biology", Gene, vol. 357, no.2, pp. 83-94, 2005.

[6] N. Toor, K. S. Keating, A. M. Pyle, "Structural insights into RNA splicing", Curr. Opin. Struct. Biol., vol. 19, pp. 260-266, 2009.

[7] R. B. Lanz, B. Razani, A. D. Goldberg, et al., "Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA)", Proc. Natl. Acad. Sci. USA, vol. 99, pp. 16081-16086, 2002.

[8] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction", Nucleic. Acids Res., vol. 31, pp. 3406-3415, 2003

[9] H.M. Berman, J. Westbrook, Z. Feng, et al., "The Protein Data Bank", Nucleic Acids Res, vol. 28, no.1, pp. 235-242, 2000.

[10] V. Pancaldi, J. Bhler, "In silico characterization and prediction of global protein-mRNA interactions in yeast", Nucleic Acids Res, vol. 39, no.14, pp. 5826-5836, 2011.

[11] Q.Lu, S. Ren, M. Lu, et al., "Computational prediction of associations between long non-coding RNAs and proteins", BMC Genomics, vol. 14, pp. 651, 2013.

[12] Y. Hao, W. Wu, H. Li, J. Yuan, J. Luo, Y. Zhao, R. Chen, "NPInter v3.0: an upgraded database of noncoding RNA-associated interactions", Database (Oxford), 2016.

[13] S. Djebali, C.A. Davis, A. Merkel, et al., "Landscape of transcription in human cells", Nature, vol. 489, no. 7414, pp. 101-108, 2012.

[14] C.M. Livi, E. Blanzieri, "Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures", BMC Bioinformatics, vol. 15, pp. 123, 2014.

[15] U.K. Muppirala, V.G. Honavar, D. Dobbs, "Predicting RNA-protein interactions using only sequence information", BMC Bioinformatics, vol. 12, pp. 489, 2011.

[16] Y. Wang, X. Chen, Z.P. Liu, et al., "De novo prediction of RNA-protein interactions from sequence information", Mol Biosyst, vol. 9, no. 1, pp. 133-142, 2013.

[17] Z. Cheng, S. Zhou, J. Guan, "Computationally predicting protein-RNA interactions using only positive and unlabeled examples", J Bioinform. Comput. Biol., vol. 13, no. 3, pp. 1541005, 2015.

[18] M. Bellucci, F. Agostini, M. Masin, et al., "Predicting protein associations with long noncoding RNAs", Nat. Methods, vol. 8, no. 6, pp. 444-445, 2011.

[19] V. Suresh, L. Liu, D. Adjeroh, et al., "RPI-Pred: predicting ncRNAprotein interaction using sequence and structural information", Nucleic Acids Res, vol. 43, no. 3, pp. 1370-1379, 2015.

[20] M. Akbaripour-Elahabad et al. "rpiCOOL: A tool for In Silico RNAprotein interaction detection using random forest". J. Theor. Biol., vol. 402, pp. 1-8, 2016.

[21] Y. Zhao, H. Li, S. Fang, et al. "NONCODE 2016: an informative and valuable data source of long non-coding RNAs". Nucleic Acids Res, vol. 44, no. D1, pp. D203-D208, 2015

[22] A. Kozomara, M. Birgaoanu, S. Griffiths-Jones, "miRBase: from microRNA sequences to function". Nucleic Acids Res. vol. 47, pp. D155-D162, 2018.

[23] The UniProt Consortium UniProt: the universal protein knowledgebase. Nucleic Acids Res. vol. 45, pp. D158-D169, 2017.

[24] K.D. Pruitt, T. Tatusova, D.R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". Nucleic Acids Res. vol. 35, pp. D61-D65, 2007.

[25] J.U. Pontius, L. Wagner, G.D., Schuler, "UniGene: A unified view of the transcriptome", NCBI Handbook, pp. 1-12, 2003.

[26] Y. Wang, J. Wang, Z. Yang, N. Deng, "Sequence-based protein-protein interaction prediction via support vector machine". Journal of Syst. Sci. Complex, vol. 23, pp. 1012-1023, 2010.

[27] S.B. Needleman, C.D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology. vol. 48, no.3, pp. 443-453, 1970.

[28] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, "Predicting protein-protein interactions based only on sequences information", Proc. Natl. Acad. Sci. USA, vol. 104, pp. 4337-4341, 2007.

[29] P. Schuster, W. Fontana, et al., "From sequences to shapes and back: a case study in RNA secondary structures", Proc. Royal Society London B., vol. 255, pp. 279-284, 1994.

[30] R. Nussinov, G. Pieczenik, J.R. Griggs, and D.J. Kleitman, "Algorithms for Loop Matchings", SIAM Journal on Applied Mathematics, vol. 35, no. 1, pp. 68-82, 1978.

[31] M. Zuker, P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information". Nucleic Acids Res. vol. 9, pp. 133-148, 1981.

[32] J.S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure", Biopolymers, vol. 29 pp. 1105-1119, 1990.

[33] E. Mattei, G. Ausiello, F. Ferr, M. Helmer-Citterich, "A novel approach to represent and compare RNA secondary structures", Nucleic Acids Res; vol. 42, no. 10, pp. 6146-6157, 2014.

[34] A. Kloczkowski, K.L. Ting, R.L. Jernigan, J. Garnier, "Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence", Proteins Structure Function and Bioinformatics vol. 49, no.2, pp. 154-166, 2002.

[35] J. Garnier, D.J. Osguthorpe, B. Robson, "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins", J. Mol. Biol., vol. 120, pp. 97-120, 1978.

[36] D. Adjeroh, M. Allaga, J. Tan, J. Lin, Y. Jiang, A. Abbasi and X. Zhou, "Feature-Based and String-Based Models for Predicting RNA-Protein Interaction", Molecules, vol. 23, pp. 697, 2018.

[37] Z.P. Liu, L.Y. Wu, X.S. Zhang, L. Chen, "Prediction of protein-RNA binding sites by a random forest method with combined features", Bioinformatics, vol. 26, pp. 1616-1622, 2010.

[38] V.N. Vapnik and A.Y. Chervonenkis, "Theory of pattern recognition: Statistical problems of learning", Nauka, 1974

[39] T.K. Ho, "Random Decision Forests", Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, vol. 1416, pp. 278-282, 1995.

[40] Breiman L., "Random forests", Journal of Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[41] T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System", Proceeding of the 22nd ACM SIGKDD International Conference, pp. 1-10, 2016.

[42] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm", Thirteenth International Conference in Machine Learning, pp. 148-156, 1996.
[43] C.M. Bishop, "Pattern Recognition and Machine Learning", Chap 1,3. Springer, 2006.
[44] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning", Springer, 2001, Chap 7.
[45] D.J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve", Journal of Machine Learning, vol. 77, pp. 103-123, 2009.
[46] E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson, "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristics Curves: A Nonparametric Approach", Biometrics, vol. 44, pp. 837-845, 1988.