Teresa Guarda
Filipe Portela
Maria Fernanda Augusto (Eds.)

Communications in Computer and Information Science 2348

Advanced Research in Technologies, Information, Innovation and Sustainability

ARTIIS 2024 International Workshops
Santiago de Chile, Chile, October 21–23, 2024
Revised Selected Papers, Part I

Part 1

Springer

ARTIIS

Resonant Recognition Model as a Preprocessing Technique for RNA Classification

Felipe Bueno de Souza[1,2], Matheus Henrique Pimenta-Zanon[1], Dora Henriques[3], M. Alice Pinto[3], Carlos Balsa[2(✉)], José Rufino[2], and Fabrício Martins Lopes[1]

[1] Computer Science Department,Universidade Tecnológica Federal do Paraná (UTFPR), Alberto Carazzai, 1640, Cornélio Procpio - Paraná 86300-000, Brazil
felipebuenosouza@alunos.utfpr.edu.br, omatheuspimenta@outlook.com, fabricio@utfpr.edu.br

[2] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SUSTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
balsa@ipb.pt, rufino@ipb.pt

[3] Centro de Investigação de Montanha (CIMO), Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SUSTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
dorasmh@ipb.pt, apinto@ipb.pt

Abstract. The development of high throughput sequencing technologies, such as RNA-Seq, has enabled the generation of large volumes of biological data. Thus, it is necessary to develop computational methods to interpret this massive volume of data and contribute to knowledge discovery. RNA sequences are products of the transcription of genomic DNA sequences and represent the gene expression process that organisms use to synthesize protein or RNA molecules. These RNA sequences can be compared between organisms of the same or different species to demonstrate similar functional proteins. There are several classes of RNA sequences (mRNA, rRNA, tRNA, ncRNA, etc.), with different biological functions. The correct identification of each class of RNA sequences is important because of the huge volume of unlabelled data available. In this context, this study proposes an approach based on the Resonant Recognition Model (RRM) for feature extraction and classification regarding the ncRNA and mRNA classes. To assess the proposed approach, it was adopted the dataset from the PLEK method. Despite the reduction of the input data size achieved using the RRM model, the results show high accuracy for primary protein sequences translated from RNA sequences, signaling the potential of the proposed approach to classify RNA.

Keywords: RNA · DNA · Amino Acid · RRM · RNAs Classification · Feature Extraction · Bioinformatics · Pattern Recognition

T. Guarda et al. (Eds.): ARTIIS 2024 Workshops, CCIS 2348, pp. 3–17, 2025.
https://doi.org/10.1007/978-3-031-83435-6_1

1 Introduction

One of the current scientific challenges is the interpretation and discovery of knowledge from the large volumes of data that are nowadays generated in the most diverse fields of science. Therefore, it is important to develop efficient mathematical and computational methods to deal with such challenge.

Regarding biological data, there has recently been a significant advance in the development of high throughput sequencing technologies (Next-Generation Sequencing), making it possible to popularize the sequencing of organisms [22]. Adequate bioinformatics methods and algorithms are thus becoming essential to analyze the huge amounts of biological data generated by these technologies, which requires efficiency, scalability and interpretability of results to extract useful information from it [8,9,12].

Various sources of biological data can be considered, such as genomics, transcriptomics, proteomics and metabolomics, commonly known as "omics data", for knowledge discovery using machine learning methods [15,16,21]. While genomic (DNA) sequences are strings composed of four different nucleotides – adenine (A), thymine (T), guanine (G), cytosine (C) –, thymine is replaced by uracil (U) in the transcript (RNA) sequences [1].

DNA sequences are transcribed into RNA sequences, generating different classes of RNAs (transcriptome), and these perform different biological functions in the organisms, from regulation of cells to dosage compensation, also having relationship with genetic diseases and autoimmune disorders [14,17,20].

In order to classify these different classes of RNAs, several computational methods have been devised [3,4,10,11,13]. In this context, we propose an approach for feature extraction from RNA sequences and classification of two different classes of RNA: non-coding RNA (ncRNA) and messenger RNA (mRNA).

The approach builds on a digital signal processing method, to develop the algorithm that will extract these features. The algorithm can be characterized as a pre-processing step for the data, focusing in a better computational efficiency by means of the reduction of the amount of input data. It is thus a contribution to the development of methods that can be used to deal with the computational challenges that emerge when dealing with large volumes of biological data.

The proposed approach starts by transforming protein sequences from transcribed RNA sequences into numerical series, and analyzing their frequency spectra provided by Discrete Fourier Transform (DFT) as features, which are the input feature vector for classification. The dimensionality of the input numerical series is reduced by applying the Resonant Recognition Model (RRM) [6], selecting only the common frequencies of each RNA class. This approach proved to be efficient, achieving the expected results with lower dimensional data input, thus representing an efficient and scalable method for biological sequence analysis.

The remaining of the paper is structured as follows: Sect. 2 covers the method proposed and applied in this study; Sect. 3 describes the dataset used, the data preparations step, the experimental methodology applied and the results obtained; finally, Sect. 4 lays out the conclusions.

2 Resonant Recognition Model

The Resonant Recognition Model (RRM) [6] is a digital signal processing method that uses numerical series that represent amino acids sequences, to extract the most discriminative information about their biological functionalities. These discrete sequences are transformed to the frequency domain by the Discrete Fourier Transformation (DFT), using the Fast Fourier Transformation (FFT) algorithm.

To get the numerical series from the DNA nucleotides string sequences, each one has its EIIP values, as shown in Table 1.

Table 1. Electron-Ion Interaction Potential (EIIP) values for nucleotides [6].

Name	Letter	EIIP Value
Adenine	A	0.1260
Cytosine	C	0.1340
Guanine	G	0.0806
Thymine	T	0.1335
Uracil	U	0.0289

Triads of nucleotides can be translated into amino acids, using other EIIP values, shown in Table 2. The values are the average energy states of the amino acid's valence electrons.

In this way it is possible to convert string sequences into numerical series, to be analyzed by digital signal processing methods. The data transformation steps are shown in Fig. 1: (1) represent DNA sequences as an array of strings; (2) translate DNA strings into amino acids strings; (3) transform each amino acid string in a numerical series using the EIIP value for each amino acid letter; (4) use the FFT method to create a frequency spectrum for each numerical series.

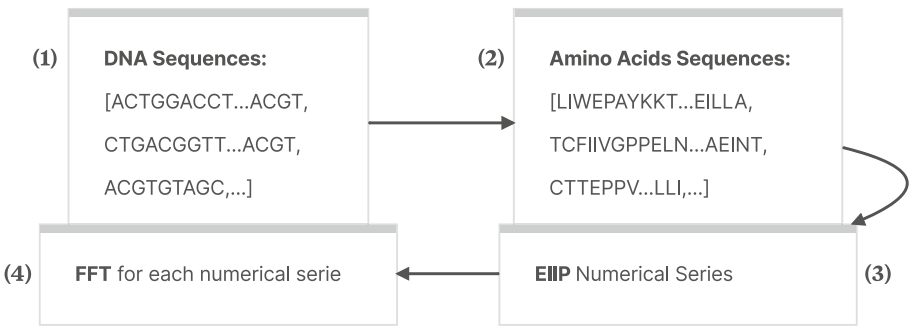

Fig. 1. From DNA sequences to Frequency Spectres.

Table 2. Electron-Ion Interaction Potential (EIIP) values for amino acids [6].

Name	Amino Acid	Letter	EIIP Value
Leucine	Leu	L	0.0000
Isoleucine	Ile	I	0.0000
Asparagine	Asn	N	0.0036
Glycine	Gly	G	0.0050
Valine	Val	V	0.0057
Glutamic Acid	Glu	E	0.0058
Proline	Pro	P	0.0198
Histidine	His	H	0.0242
Lysine	Lys	K	0.0371
Alanine	Ala	A	0.0373
Tyrosine	Tyr	Y	0.0516
Tryptophan	Trp	W	0.0548
Glutamine	Gln	Q	0.0761
Methionine	Met	M	0.0823
Serine	Ser	S	0.0829
Cysteine	Cys	C	0.0829
Thereonine	Thr	T	0.0941
Phenyalanine	Phe	F	0.0946
Arginine	Arg	R	0.0959
Aspartic Acid	Asp	D	0.1263

RRM extracts common frequencies between spectra through a cross-spectrum function, for two frequency spectra of different sequences. This procedure raises the magnitude of common frequencies, indicating prominent peaks, as represented in Fig. 2. The cross-spectrum function can be described as the multiplication of the DFT coefficient X_n from a $x(m)$ series, by the conjugate complex Y_n^* of the DFT coefficients of another series $y(m)$, as shown in Eq. (1):

$$S_n = X_n Y_n^* \quad n = 1, 2, 3, ..., N/2 \tag{1}$$

In this study, several sequences of different lengths were used. To define common frequencies among protein sequences, the absolute value M_n of each coefficient of a multiple cross-spectral function is calculated, according to:

$$|M_n| = |X1_n|.|X2_n|.|X3_n|...|XM_n| \quad n = 1, 2, 3, ..., N/2 \tag{2}$$

This multiple cross-spectrum function is called *Consensus Spectrum* for a large group of protein sequences with the same biological function [6].

3 Experiments

This section describes the experiments conducted to validate the proposed approach for RNA classification. First, it is provided a characterization of the

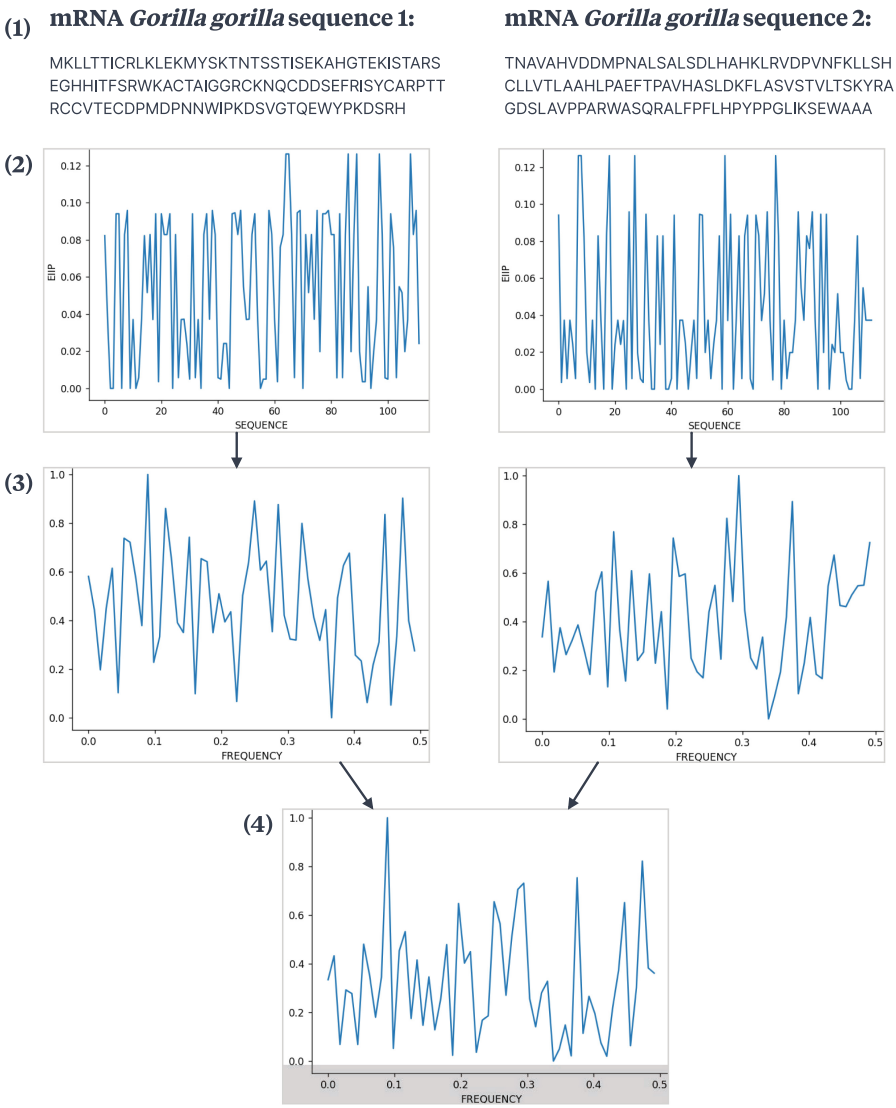

Fig. 2. An example of application of the RRM method: (1) Same size mRNA protein sequences from *Gorilla gorilla*; (2) Graphic representation of the numerical series obtained by replacing each amino acid with its EIIP value; (3) Frequency spectra obtained when applying DFT for the numerical series; (4) Frequency spectrum results obtained by the cross-spectral function. (common frequencies are represented by prominent peaks).

datasets used (Subsect. 3.1) and a description of the data preparation procedures (Subsect. 3.2); next, it is presented the methodology followed to classify the RNA classes (Subsect. 3.3); finally, a comparison is made between the results obtained in this work and a previous related one (Subsect. 3.4).

3.1 Dataset

To evaluate the RRM method, two commonly used datasets were selected: one from the PLEK tool [13] (a predictor of long non-coding RNAs and messenger RNAs based on an improved k-mer scheme), and another from the CPC2 tool [11] (a coding potential calculator based on sequence intrinsic features). These datasets were used in previous related works as [4,10], allowing a suitable comparison between methods and avoiding bias in classifications and analysis.

The PLEK tool dataset includes different numbers of sequences of two RNA classes: ncRNA (non-coding RNA) and mRNA (messenger RNA). The dataset comprises data from nine different species (see Table 3).

Table 3. PLEK dataset [13].

Species	RNA Class	Number of Sequences	Average Length	σ
Gorilla gorilla	mRNA	33025	2775.39	2080.64
(Western gorilla)	ncRNA	367	291.48	88.44
Macaca mulatta	mRNA	5709	2044.99	1388.93
(Rhesus macaque)	ncRNA	359	292.86	87.92
Bos taurus	mRNA	13190	2302.28	1507.43
(Cow)	ncRNA	182	296.85	116.89
Danio rerio	mRNA	14493	2088.83	1257.14
(Zebrafish)	ncRNA	419	593.11	471.60
Mus musculus	mRNA	35765	2659.80	2269.25
(House mouse)	ncRNA	8032	530.49	929.58
Pan troglodytes	mRNA	1906	1922.76	1204.01
(Chimpanzee)	ncRNA	1164	289.70	50.45
Pongo abelii	mRNA	3401	2836.92	1195.64
(Sumatran orangutan)	ncRNA	392	290.44	86.10
Sus scrofa	mRNA	3978	1823.76	1412.85
(Boar)	ncRNA	241	381.18	247.89
Xenopus tropicalis	mRNA	8874	2294.34	1350.16
(Clawed frog)	ncRNA	279	205.24	110.53

The CPC2 dataset includes three RNA classes from six species: mRNA, and ncRNA split into long non-coding RNA (lncRNA) and small non-coding RNA

(sncRNA). ncRNAs are sequences longer than 200 nucleotides – this was the threshold to discriminate between long and small subclasses [5]. Thus, sncRNAs are too small to be applied in the proposed method, and therefore only lnRNA sequences were considered, representing the ncRNA class in Table 4.

Table 4. CPC2 dataset [11].

Species	RNA Class	Number of Sequences	Average Length	σ
Arabidopsis thaliana	mRNA	13986	1669.28	938.44
(Arabidopsis)	ncRNA	2562	351.76	194.59
Drosophila melanogaster	mRNA	3680	2852.06	2302.20
(Fruit fly)	ncRNA	2776	1016.56	1292.79
Homo sapiens	mRNA	6142	3833.17	3938.10
(Human)	ncRNA	7485	944.84	2597.15
Danio rerio	mRNA	2344	2084.25	1113.65
(Zebrafish)	ncRNA	1163	952.27	876.10
Mus musculus	mRNA	10638	2954.16	2068.44
(House mouse)	ncRNA	6460	1169.62	1139.22
Caenorhabditis elegans	mRNA	3551	1599.75	1695.39
(Worm)	ncRNA	1582	346.90	381.54

Both the PLEK and CPC2 datasets have *Mus musculus* and *Danio rerio* as common species, enabling a cross-validation between datasets, and helping to make the classification of the method as less biased as possible.

Due to the different amount of sequences, to create an unbiased prediction model, for each species, the number of sequences was limited to the minimum number of sequences between the two classes. This involved slicing the sequence's array of the class with greater amount, from index 0 to the index of the minimum value, producing a balanced number of sequences in each class of RNA.

3.2 Data Preparation

Both datasets were submitted for a pre-processing to cleanup duplicated sequences in each file, and between shared species from both datasets, using Seqkit2, an open-source toolkit for FASTA/Q file manipulation [19], which was incremented from the first version [18] with new functionalities.

To prepare the data, four commands provided by Seqkit2 were used: (i) `rmdup`, to remove duplicate sequences in a file; (ii) `common`, to find similar sequences between files with common names; (iii) `seq`, to extract those common sequences into a reference file; (iv) `grep`, to create filtered files removing the common sequences, in order to avoid over-fitting [19].

First, duplicate sequences from each dataset were removed with `rmdup`, to create cleaned files for each RNA class of each species. For shared species, such

as *Danio rerio* and *Mus musculus*, the output files had a common suffix, such as `mRNA_cleaned_*` and `ncRNA_cleaned_*`, with `*` being the name of the source dataset. With the files named this way, it was possible to use the command `common` between them. This process was implemented to handle the first part of this work, applying the method to datasets individually. Then, using `seq`, files were created with shared sequences between the common species, and `grep` was used to create filtered files without them, allowing unbiased cross-validation between datasets. The composition of these filtered files is shown in Table 5.

Table 5. Composition of the Filtered Files.

Species	Dataset	RNA class	Number of Sequences	Average Length
Danio rerio	PLEK	mRNA	12263	2078.8
		ncRNA	253	503
	CPC2	mRNA	115	1890
		ncRNA	996	989.2
Mus musculus	PLEK	mRNA	26545	2574
		ncRNA	5833	264.8
	CPC2	mRNA	1435	3250.8
		ncRNA	4251	1135.1

3.3 Methodology

The aim of this study was to develop a binary classification to classify sequences between ncRNAs and mRNAs, focusing on reducing the dimensionality of the input sequences. For that, common frequencies from mRNA and ncRNA were extracted using miscellaneous sequences from the same class. This process was carried out individually for each one of the nine target species listed in Table 3.

To perform the spectrum analysis, the frequencies were first distributed in a histogram of $N = 512$ bins, where frequencies range from 0.0 to 0.5. Note that 0.5 is the maximum frequency value of the spectrum, because the mean distance of amino acids in a peptide chain is considered equidistant; therefore, the distance between points in a numerical series is set arbitrarily with a value of $d = 1$, making the maximum frequency to be $F = 1/2d = 0.5$ [6].

Numerical series provided by nucleotides EIIP values have a different spectrum resolution, as the distance between the nucleotides is not equidistant. The distance between nucleotides is $d = 0.89$, making the maximum frequency in the DNA and RNA spectra to be $F = 1/2d = 0.56$ [7].

The distribution of values in a histogram ensures that sequences of different sizes are analyzed equivalently. First, two histograms were assembled, one for mRNA and one for ncRNA. These histograms have their coefficient values multiplied to form a spectrum of common frequencies by considering normalized values between 0 and 1, extracting which are the discriminatory frequencies for each class. The frequencies with a magnitude lower than 0.1 were filtered and considered noise. Figure 3 shows the process by which the histogram is built.

Then, to validate the selected frequencies as discriminatory parts, a classification was performed by considering histograms representing the DFTs as feature vectors, ensuring that sequences of different sizes were analyzed equally, since larger sequences have more frequency points than smaller ones. Thus, different frequency points were allocated to the frequencies closest to the histogram, and a modulus of these values was performed when several values are assigned to a single frequency. This process can be visualized in Fig. 4. The result is a collection of histograms that can be interpreted as discrete sequences of values.

The discrete sequences were classified with their full size, and then, for each sequence, only the most discriminating frequency indices extracted through RRM in the first stage were selected, thus reducing the size of the input sequences and analyzing the performance of the two classifications.

The adopted classification algorithm was a Decision Tree, using the optimized version of the CART algorithm [2], with 10-fold cross-validation to assess the training performance of the classification model in the individual evaluation of each dataset.

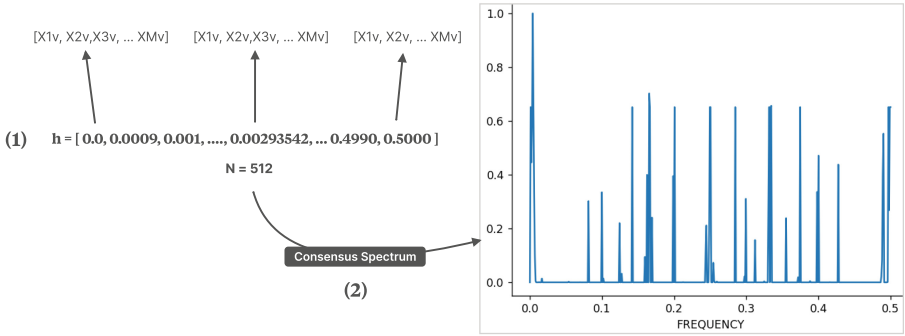

Fig. 3. Building the frequencies histogram based on the multiple cross-spectrum function *Consensus Spectrum*: (1) For each spectrum in the class, DFT sequences were iterated to assign each value to the respective frequency in the histogram – because sequences have different dimensions, this process created a proportional histogram for the cross-spectral function; (2) With the histogram assembled, Eq. (2) is applied, to create the spectrum signal with the peaks of common frequencies.

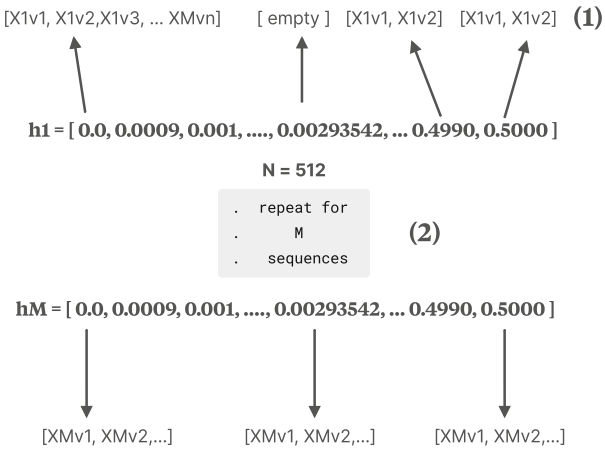

Fig. 4. Construction of the frequency histograms to analyze all the sequence spectra as a same dimension signal of $N = 512$ frequency intervals from 0.0 to 0.5: (1) The DFT sequence was iterated to assign each value to the respective frequency in the frequency histogram; (2) The first step was repeated for each one of the DFT sequences, creating M histograms that will be analyzed as a spectrum signal.

3.4 Results

To evaluate the proposed method, both datasets were treated individually in order to compare the results with those of the original methods implemented on each dataset. Then, two cross-validations were performed between the shared species to evaluate the behavior of the method and the classifier: training the model with the PLEK dataset and validating it with the CPC2 filtered files (created at data preparation, as shown in Sect. 3.2), and training the model with the CPC2 dataset and validating it with PLEK filtered files.

Table 6 shows similar results to those obtained with the PLEK tool, with a slight improvement when considering mRNAs, however with similar accuracies overall. Comparing the results between sequences with size $N = 512$, and sequences with reduced sizes, it is noticeable that the peak values from common frequencies can be interpreted as feature coefficients to be analyzed when the approach is to classify sequences. Therefore, by reducing the size of the input data, the computational complexity was reduced, achieving better performance.

Table 7 presents the comparison of the original CPC2 results with those obtained using its dataset, and also with those produced when using the reduced inputted data generated with RRM. For this dataset, the achieved accuracies were lower than CPC2 results. Despite this, the accuracies achieved are still considerable, and the dimensionality reduction did not implied any relevant changes on their values, thus indicating the impressive dimensional reduction of the feature space for the input sequences. The reduced number of discriminative extracted features shows that the proposed method could be a valuable contribution as a pre-processing step for other classification methods in similar cases.

Table 6. Classification results for mRNA and ncRNA classes for each specie from PLEK dataset [13].

Species	RNA class	PLEK accuracy	N-512 DFT histogram accuracy	N-size sequences frequency peaks accuracy	N
Gorilla gorilla	mRNA	83.8%	91%	94%	41
	ncRNA	99.7%	93%	96%	
Specie Average		**91.7%**	**92%**	**95%**	
Macaca mulatta	mRNA	85%	92%	98%	38
	ncRNA	100%	93%	94%	
Specie Average		**92.5%**	**92%**	**96%**	
Bos taurus	mRNA	94.8%	98%	100%	38
	ncRNA	99.5%	98%	100%	
Specie Average		**97.1%**	**98%**	**100%**	
Danio rerio	mRNA	91.3%	78%	76%	3
	ncRNA	90.9%	82%	77%	
Specie Average		**91.1%**	**80%**	**76%**	
Mus musculus	mRNA	88.1%	79%	81%	2
	ncRNA	89.9%	80%	79%	
Specie Average		**89%**	**80%**	**80%**	
Pan troglodytes	mRNA	87.1%	94%	99%	42
	ncRNA	99.9%	97%	95%	
Specie Average		**92.5%**	**95%**	**97%**	
Pongo abelii	mRNA	98%	98%	98%	40
	ncRNA	100%	97%	96%	
Specie Average		**99%**	**98%**	**97%**	
Sus scrofa	mRNA	85.1%	88%	85%	8
	ncRNA	98.3%	88%	86%	
Specie Average		**91.7%**	**88%**	**86%**	
Xenopus tropicalis	mRNA	94.5%	96%	98%	90
	ncRNA	100%	97%	97%	
Specie Average		**97.2%**	**97%**	**98%**	
Average accuracy per class	mRNA	89.7%	90.4%	**92.1%**	
	ncRNA	**97.6%**	91.7%	91.1%	

Table 7. Classification results for mRNA and ncRNA classes for each specie from CPC2.

Species	RNA class	CPC2 accuracy	N-512 DFT histogram accuracy	N-size sequences frequency peaks accuracy	N
Arabidopsis thaliana	mRNA	99.7%	93%	92%	1
	ncRNA	95.3%	93%	89%	
Specie Average		**97.5%**	**93%**	**90%**	
Drosophila melanogaster	mRNA	94.6%	75%	71%	2
	ncRNA	91.9%	69%	66%	
Specie Average		**93.2%**	**72%**	**68%**	
Homo sapiens	mRNA	95.9%	80%	80%	510
	ncRNA	92.8%	76%	78%	
Specie Average		**94.3%**	**78%**	**79%**	
Danio rerio	mRNA	95.5%	73%	68%	1
	ncRNA	88.1%	74%	64%	
Specie Average		**91.8%**	**74%**	**66%**	
Mus musculus	mRNA	93.9%	72%	66%	2
	ncRNA	95.0%	69%	65%	
Specie Average		**94.4%**	**70%**	**66%**	
Caenorhabditis elegans	mRNA	96.5%	90%	88%	1
	ncRNA	99.9%	87%	86%	
Specie Average		**98.2%**	**89%**	**87%**	
Average accuracy per class	mRNA	**96%**	80.5%	77.5%	
	ncRNA	**93.8%**	78%	74.7%	

Table 8. Classification results for mRNA and ncRNA classes cross validation between datasets training with CPC2 dataset and validating with PLEK dataset.

Species	RNA class	N-512 DFT histogram accuracy	N-size sequences frequency peaks accuracy	N
Danio rerio	mRNA	87%	89%	5
	ncRNA	77%	75%	
Specie Average		**82%**	**82%**	
Mus musculus	mRNA	90%	88%	2
	ncRNA	64%	65%	
Specie Average		**77%**	**76.5%**	
Average accuracy per class	mRNA	88.5%	88.5%	
	ncRNA	70.5%	70%	

Table 9. Classification results for mRNA and ncRNA classes cross validation between datasets training with PLEK dataset and validating with CPC2 dataset.

Species	RNA class	N-512 DFT histogram accuracy	N-size sequences frequency peaks accuracy	N
Danio rerio	mRNA	61%	64%	7
	ncRNA	70%	69%	
Specie Average		**65.5%**	**66.5%**	
Mus musculus	mRNA	55%	56%	2
	ncRNA	61%	65%	
Specie Average		**58%**	**60.5%**	
Average accuracy per class	mRNA	58%	60%	
	ncRNA	65.5%	67%	

To better evaluate the proposed method, it was performed a cross-validation between each of the dataset's shared species, *Danio Rerio* and *Mus musculus*, training the decision tree with one dataset and validating it with the other one. The first validation was training the model entirely with the CPC2 dataset, and then using the trained model to classify the filtered PLEK data, so that all the sequences used in validation are different from those used during training, avoiding possible overfitting of the classifier. This same process was carried out in the second cross-validation, this time training with the complete PLEK data and validating with the filtered CPC2 data. The results are in Tables 8 and 9.

Comparing the two tables, it is noticeable that the CPC2 dataset is more robust. The accuracies achieved with the model trained with this dataset were considerably higher than those with the PLEK dataset, showing the model's better ability to generalize. This shows how important data quality is for classification. The proposed method consistently produced suitable results, as all the tests maintained similar accuracies and even improved when considered only the peak values of the spectra for reduction the dimensionality of the feature space.

4 Conclusion

Biological sequences may be analysed in many different ways. Moreover, an increasingly amount of biological data is becoming available. Thus, computational solutions that increase the performance of the analysis, scalability and interpretability of the results are extremely relevant. These solutions may have the potential of dispensing with the need to resort to high-performance computational systems to handle this analysis, increasing applicability and usability for a wide range of users, beyond researchers that have access to supercomputers.

This study presents an approach based on RRM as a functional method for dimensionality reduction for analyzing the frequency spectra of primary protein

sequences. Using a simple decision tree as a classifying algorithm, high accuracies were achieved, even when reducing the input data dimensionality. Therefore, this approach points to a viable and scalable solution to biologic data analysis, reducing the complexity of the input data of a classifier, but avoiding a loss in the quality of this data and, as a result, in the classification of RNA sequences.

Comparing the accuracies between the proposed approach with the lower dimensionality generated, we did not obtain a percentage increase or a significant worsening either, maintaining close accuracies values. However, these values were achieved using a decision tree classifier which returns reliable values, as it allows us to easily understand how predictions are made, bringing transparency, explainability and confidence to predictive models. Thus, we can conclude the RRM method is a feasible pre-processing step for classification models.

As an additional contribution, the proposed approach is open and freely accessible at https://github.com/SALIPE/RRM-PLEK-APPLICATION, contributing to the replicability of the study and further contributions.

Acknowledgments. This work was supported by national funds through the Fundação Araucária (Grant number 035/2019, 138/2021 and NAPI - Bioinformática), CNPq 440412/2022-6 and 408312/2023-8), FCT/MCTES (PIDDAC): CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDB/05757/2020); CIMO, UIDB/00690/2020 (DOI: 10.54499/UIDB/00690/2020) and UIDP/00690/2020 (DOI: 10.54499/UIDP/00690/2020); and SusTEC, LA/P/0007/2020 (DOI: 10.54499/ LA/P/0007/2020).

References

1. Alberts, B., et al.: Molecular Biology of the Cell: Seventh International Student Edition with Registration Card. WW Norton & Company (2022)
2. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and Regression Trees. Taylor & Francis (1984). https://books.google.pt/books?id=JwQx-WOmSyQC
3. Breve, M.M., Lopes, F.M.: A simplified complex network-based approach to MRNA and NCRNA transcript classification. In: Setubal, J.C., Silva, W.M. (eds.) Advances in Bioinformatics and Computational Biology, pp. 192–203. Springer, Cham (2020)
4. Breve, M.M., Pimenta-Zanon, M.H., Lopes, F.M.: Basinetentropy: an alignment-free method for classification of biological sequences through complex networks and entropy maximization (2022)
5. Brosnan, C.A., Voinnet, O.: The long and the short of noncoding RNAS. Current Opin. Cell Biol. **21**(3), 416–425 (2009). https://doi.org/10.1016/j.ceb.2009.04.001, nucleus and gene expression
6. Cosic, I.: Macromolecular bioactivity: is it resonant interaction between macromolecules? - theory and applications. IEEE Trans. Bio-med. Eng. **41**, 1101–14 (1995). https://doi.org/10.1109/10.335859
7. Cosic, I.: The resonant recognition model of macromolecular bioactivity. Birkhäuser Basel (2012). https://doi.org/10.1007/978-3-0348-7475-5

8. de Holanda Maia, M.R., Plastino, A., Freitas, A., de Magalhaes, J.P.: Interpretable ensembles of classifiers for uncertain data with bioinformatics applications. IEEE/ACM Trans. Comput. Biol. Bioinf. **20**(3), 1829–1841 (2022)
9. Iqbal, N., Kumar, P.: From data science to bioscience: emerging era of bioinformatics applications, tools and challenges. Procedia Comput. Sci. **218**, 1516–1528 (2023)
10. Ito, E.A., Katahira, I., Vicente, F.F.d.R., Pereira, L.F.P., Lopes, F.M.: BASiNET - BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification. Nucleic Acids Res. **46**(16), e96–e96 (2018). https://doi.org/10.1093/nar/gky462
11. Kang, Y.J., et al.: CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. **45**(W1), W12–W16 (2017).https://doi.org/10.1093/nar/gkx428
12. Karim, M.R., et al.: Explainable AI for bioinformatics: methods, tools and applications. Briefings Bioinform. **24**(5), bbad236 (2023). https://doi.org/10.1093/bib/bbad236
13. Li, A., Zhang, J., Zhou, Z.: PLEK: a tool for predicting long non-coding RNAS and messenger RNAS based on an improved k-MER scheme. BMC Bioinform. **15**(1), 311 (2014). https://doi.org/10.1186/1471-2105-15-311
14. Mattick, J.S., et al.: Long non-coding RNAS: definitions, functions, challenges and recommendations. Nat. Rev. Mol. Cell Biol. **24**(6), 430–447 (2023)
15. Picard, M., Scott-Boyer, M.P., Bodein, A., Périn, O., Droit, A.: Integration strategies of multi-omics data for machine learning analysis. Comput. Struct. Biotechnol. J. **19**, 3735–3746 (2021). https://doi.org/10.1016/j.csbj.2021.06.030
16. Reel, P.S., Reel, S., Pearson, E., Trucco, E., Jefferson, E.: Using machine learning approaches for multi-omics data analysis: a review. Biotechnol. Adv. **49**, 107739 (2021). https://doi.org/10.1016/j.biotechadv.2021.107739
17. Santosh, B., Varshney, A., Yadava, P.K.: Non-coding RNAS: biological functions and applications. Cell Biochem. Funct. **33**(1), 14–22 (2015)
18. Shen, W., Le, S., Li, Y., Hu, F.: Seqkit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. PLOS ONE **11**(10), 1–10 (2016). https://doi.org/10.1371/journal.pone.0163962
19. Shen, W., Sipos, B., Zhao, L.: Seqkit2: a swiss army knife for sequence and alignment processing. iMeta n/a(n/a), e191. https://doi.org/10.1002/imt2.191
20. Statello, L., Guo, C.J., Chen, L.L., Huarte, M.: Gene regulation by long non-coding RNAS and its biological functions. Nat. Rev. Mol. Cell Biol. **22**(2), 96–118 (2021)
21. Vicente, F.F., Lopes, F.M., Hashimoto, R.F., Cesar, R.M.: Assessing the gain of biological data integration in gene networks inference. BMC Genomics **13**, 1–12 (2012)
22. Villaseñor-Altamirano, A.B., Balderas-Martínez, Y.I., Medina-Rivera, A.: Review of gene expression using microarray and RNA-seq. In: Rigor and Reproducibility in Genetics and Genomics, pp. 159–187. Elsevier (2024)