SASYR Symposium of Applied Science for Young Researchers

4th Symposium of Applied Science for Young Researchers

PROCEEDINGS 2024

July 3 , 2024

Optimization of RNA classification using the Resonant Recognition Model

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] Universidade Tecnológica Federal do Paraná (UTFPR), Campus Cornélio Procópio, Brasil
felipebuenosouza@alunos.utfpr.edu.br, matheus.pimenta@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. The RNA sequences have different biological functions, such as mRNAs, rRNAs, tRNAs and ncRNA, to name a few. The correct identification of each class of RNA sequences is important because of the huge volume of unlabeled data available. In this context, this study proposes an approach based on the Resonant Recognition Model 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 RMM method, the results show high accuracy, indicating the potential of the proposed approach.

Keywords: RNA · RRM · RNAs classification · Feature Extraction · Bioinformatics · Pattern Recognition.

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 [6]. Adequate Bioinformatics tools and algortihms are thus becoming essential to analyze the huge amounts of biological data generated by these technologies, which requires efficiency and scalability to extract useful information from it [4].

In particular, DNA sequences are transcribed to RNA sequences, providing different classes of RNAs (transcriptome), which are important because each class of RNA can perform different biological functions in the organisms, from regulation of cells and dosage compensation, their relationship with genetic diseases and autoimmune disorders [2, 5].

This study proposes a method to extract features from RNA sequences and classify two different classes of RNA: non-coding RNA (ncRNA) and messenger RNA (mRNA). The proposed approach starts by transforming DNA 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 proposed approach reduces the dimensionality of the inputted numerical series by applying the Resonant Recognition Model (RMM) [3], selecting only the common frequencies of each RNA class. This approach proved to be efficient, achieving the expected results with lower dimensional data input.

This paper is structured in four main sections, in addition to the Introduction: Section 2 covers the method applied in this study; Section 3 describes the dataset used, the experimental methodology applied and the results obtained; finally, Section 4 lays out the conclusions and delineates future work.

2 Resonant Recognition Model

The Resonant Recognition Model (RRM) method [3] is a digital signal processing method that uses numerical series that represent amino acids sequences, in order 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, triads of nucleotides can be translated as amino acids, where each one has its EIIP value, as shown in Table 1. The values are the average energy states of the amino acid's valence electrons. In this way it is possible to convert strings sequences into numerical series, to be analyzed by digital signal processing methods. The data transformation follows the steps presented at Figure 1.

RMM extracts common frequencies between spectres through a cross-spectrum function, for two frequency spectres of different sequences. 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 Equation 1:

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

In this study, DNA sequences of different sizes were used. To define common frequencies among protein sequences, it is calculated the absolute value M_n for each coefficient of a multiple cross-spectral function, as defined by Equation 2:

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

Table 1: Electron-Ion Interaction Potential (EIIP) Values for Amino Acids [3]

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

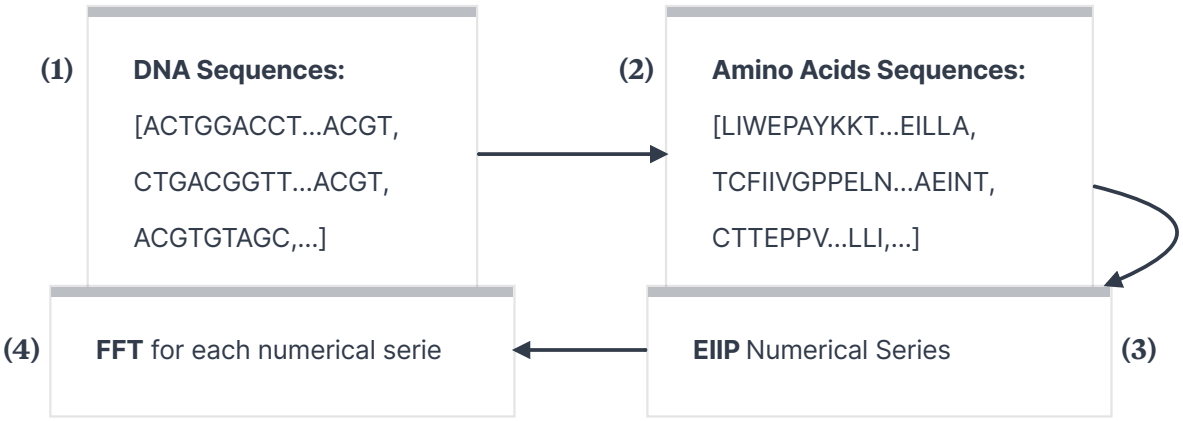

Fig. 1: From DNA sequences to Frequency Spectres: (1) Represent DNA sequences as array of strings. (2) Translate DNA strings into amino acids strings. (3) Transform each amino acid string in a numerical series using the EIIP value for each amino acid letter. (4) Use the FFT method to create a frequency spectre for each numerical series.

This multiple cross-spectrum function is called *Consensus Spectrum* for a large group of protein sequences with the same biological function [3].

3 Experiments

This sections is organized as follows: first, it is provided a characterization of the dataset used (Subsection 3.1); then, it is presented the methodology applied to classify the RNA classes (Subsection 3.2); finally, a comparison is made between the results obtained in this work and a previous related one (Subsection 3.3).

3.1 Dataset

To evaluate the RMM method, a dataset from the PLEK site (predictor of long non-coding RNAs and messenger RNAs based on an improved k-mer scheme) [5] was used. This dataset includes different numbers of sequences of two RNA classes: ncRNA and mRNA. The dataset comprises data from 9 different organisms species (see Table 2).

Table 2: PLEK dataset [5]

Species	RNA Class	Number of Sequences	Sequences Mean Size	σ
Gorilla gorilla	mRNA	33025	2775.4	2080.6
(Western gorilla)	ncRNA	367	291.5	88.4
Macaca mulatta	mRNA	5709	2044.9	1388.9
(Rhesus macaque)	ncRNA	359	292.8	87.9
Bos taurus	mRNA	13190	2302.3	1507.4
(Cow)	ncRNA	182	296.8	116.9
Danio rerio	mRNA	14493	2088.8	1257.1
(Zebrafish)	ncRNA	419	593.1	471.6
Mus musculus	mRNA	35765	2659.8	2269.2
(House mouse)	ncRNA	8032	530.5	929.6
Pan troglodytes	mRNA	1906	1922.7	1204
(Chimpanzee)	ncRNA	1164	289.7	50.4
Pongo abelii	mRNA	3401	2836.9	1195.6
(Sumatran orangutan)	ncRNA	392	290.4	86.1
Sus scrofa	mRNA	3978	1823.7	1412.8
(Boar)	ncRNA	241	381.2	247.9
Xenopus tropicalis	mRNA	8874	2294.3	1350.1
(Clawed frog)	ncRNA	279	205.2	110.5

Because of 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, slicing the sequence's array of the class with greater amount from index 0 to index of the minimum value, producing a balanced number of sequences in each class of RNA.

3.2 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 2.

To perform the spectre 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 spectre, 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 $F = d/2 = 0.5$ [3].

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 2 shows the process by which the histogram is built.

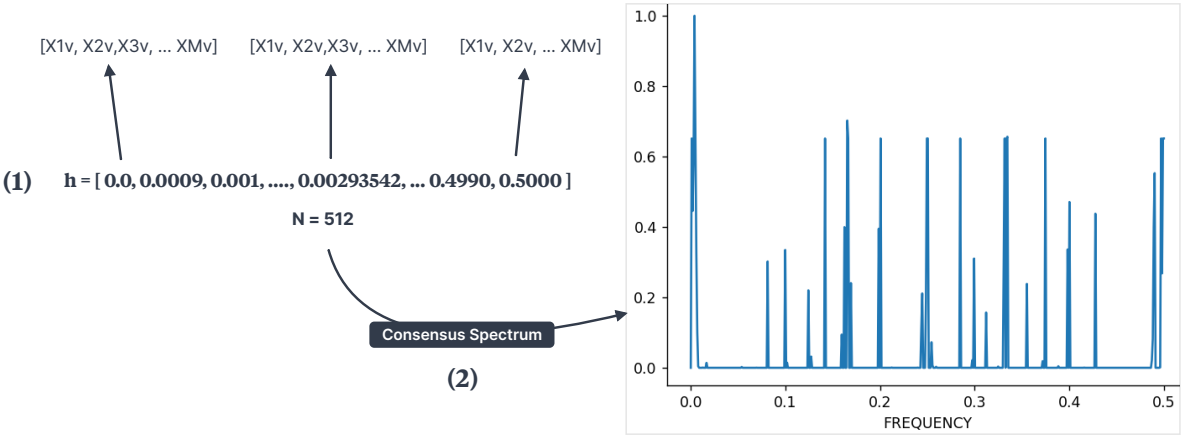

Fig. 2: Building the frequencies histogram based on the multiple cross-spectrum function *Consensus Spectrum*: (1) For each spectrum in the class, DFT sequences were iterated to assign each value to the respective frequency in the histogram; Because sequences have different dimensions, this process created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

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 Figure 3. The result is a collection of histograms that can be interpreted as discrete sequences of values.

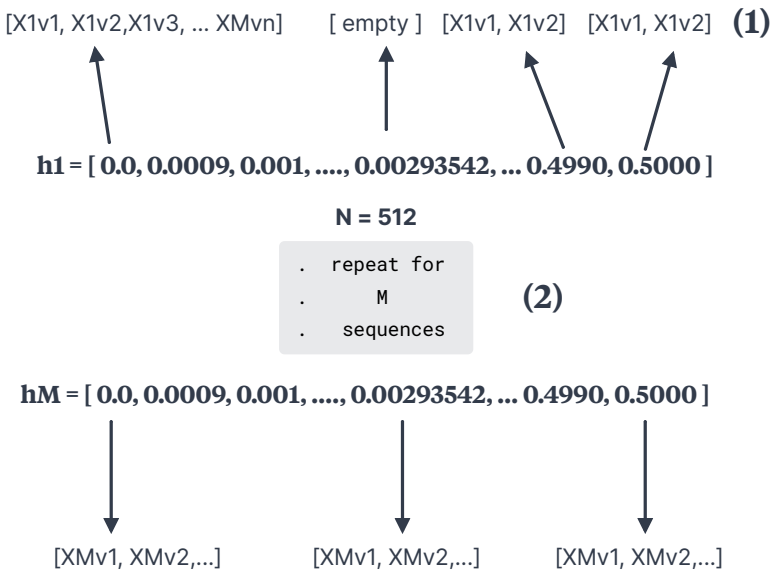

Fig. 3: Construction of the frequency histograms to analyze all the sequences spectra as a same dimension signal of $N = 512$ frequency intervals from 0.0 to 0.5: (1) The DFT sequence was iterated to assign each value to the respective frequency in the frequency histogram. (2) The first step was repeated for each one of the DFT sequences, creating M histograms that will be analyzed as a spectrum signal.

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 [1], with 10-fold cross-validation to assess the training performance of the classification model.

3.3 Results

The results of the classifications for each species can be seen in Table 3.

Table 3 shows similar results to those obtained with the PLEK tool [5], some better than others, but overall with similar accuracies. Comparing the results between sequences with size $N = 512$, and sequences with reduced sizes, it is possible to see that the peak values from common frequency can be interpreted as characteristics coefficients to be analyzed when the approach is to classify sequences. Therefore, by reducing the size of the input data, it is possible to achieve better performance and computational efficiency.

4 Conclusions

Biological sequences may be analysed in many different ways. Moreover, an increasingly amount of biological data is becoming available. Thus, computational solutions to increase the performance of the analysis of such are very important.

Table 3: Classification results for mRNA and ncRNA classes for each specie from PLEK dataset [5]

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%	
Macaca mulatta	mRNA	85%	92%	98%	38
	ncRNA	100%	93%	94%	
Bos taurus	mRNA	94.8%	98%	100%	38
	ncRNA	99.5%	98%	100%	
Danio rerio	mRNA	91.3%	78%	76%	3
	ncRNA	90.9%	82%	77%	
Mus musculus	mRNA	88.1%	79%	81%	2
	ncRNA	89.9%	80%	79%	
Pan troglodytes	mRNA	87.1%	94%	99%	42
	ncRNA	99.9%	97%	95%	
Pongo abelii	mRNA	98%	98%	98%	40
	ncRNA	100%	97%	96%	
Sus scrofa	mRNA	85.1%	88%	85%	8
	ncRNA	98.3%	88%	86%	
Xenopus tropicalis	mRNA	94.5%	96%	98%	90
	ncRNA	100%	97%	97%	

This study presents an approach based on RRM as a functional method for dimensionality reduction for analyzing the frequency spectra of DNA sequences. Here, using a simple decision tree as a classifying algorithm, high accuracies were achieved, even when reducing the input data dimensionality.

For future work, it will be considered the development of a method to map which parts of the original sequences generate the common frequency peaks that were used for dimensionality reduction in this work.

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. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and Regression Trees. Taylor & Francis (1984), https://books.google.pt/books?id=JwQx-WOmSyQC
2. 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)
3. Cosic, I.: Macromolecular bioactivity: Is it resonant interaction between macromolecules? - theory and applications. IEEE transactions on bio-medical engineering **41**, 1101–14 (01 1995). https://doi.org/10.1109/10.335859
4. Iqbal, N., Kumar, P.: From data science to bioscience: emerging era of bioinformatics applications, tools and challenges. Procedia Computer Science **218**, 1516–1528 (2023)

5. 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 Bioinformatics **15**(1), 311 (2014). https://doi.org/10.1186/1471-2105-15-311
6. 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)