doi: 10.1093/nar/gky462

BASiNET—BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification

Eric Augusto Ito[1], Isaque Katahira[1], Fábio Fernandes da Rocha Vicente[1], Luiz Filipe Protasio Pereira[1,2] and Fabrício Martins Lopes[1,*]

[1]Department of Computer Science, Bioinformatics Graduate Program, Federal University of Technology – Paraná, Cornélio Procópio, PR 86300-000, Brazil and [2]Empresa Brasileira de Pesquisa Agropecuária, Embrapa Café, Brasília, DF 70770-901, Brazil

Received December 15, 2017; Revised May 09, 2018; Editorial Decision May 11, 2018; Accepted May 22, 2018

ABSTRACT

With the emergence of Next Generation Sequencing (NGS) technologies, a large volume of sequence data in particular *de novo* sequencing was rapidly produced at relatively low costs. In this context, computational tools are increasingly important to assist in the identification of relevant information to understand the functioning of organisms. This work introduces BASiNET, an alignment-free tool for classifying biological sequences based on the feature extraction from complex network measurements. The method initially transform the sequences and represents them as complex networks. Then it extracts topological measures and constructs a feature vector that is used to classify the sequences. The method was evaluated in the classification of coding and non-coding RNAs of 13 species and compared to the CNCI, PLEK and CPC2 methods. BASiNET outperformed all compared methods in all adopted organisms and datasets. BASiNET have classified sequences in all organisms with high accuracy and low standard deviation, showing that the method is robust and non-biased by the organism. The proposed methodology is implemented in open source in R language and freely available for download at https://cran.r-project.org/package=BASiNET.

INTRODUCTION

The advances in high-throughput sequencing (RNA-seq) have enabled a broader characterization of transcripts (1,2). In addition to the quantification of transcriptomes, these new methods have contributed to a better understanding of the genetic information contained in the RNA sequence, such as those related to non-coding RNAs (3), as well as make possible the sequencing of several species (4).

Two classes of transcripts have been extensively investigated, the mRNAs that carries information for the synthesis of proteins and, more recently, the non-coding RNAs (ncRNAs), involved mainly in epigenetic regulation. What differentiates mRNAs from ncRNAs is mainly their function. The length of the sequence is not an effective feature to classify them into mRNA or ncRNA. However, the ncRNAs are categorized into two groups according to the size of the sequence: the long non-coding RNAs (lncRNAs), with sequences >200 nucleotides and the small non-coding RNAs (sncRNAs), with sequences shorter than 200 nucleotides. (5).

These molecules are important because they act in different biological processes like transcriptional regulation (6–8), may be associated with human diseases (9) such as cancer (10,11), neurodegenerative and cardiovascular diseases (12) to cite but a few.

The sncRNAs are abundant in highly conserved organisms and are related to transcriptional gene silencing (5,13).

In this context, there is a great challenge to identify the different types of sequences in the large volume of data produced by the RNA-seq technique. Computational methods may be useful for classifying mRNA, lncRNAs and sncRNAs.

Regarding pattern recognition research field, there are well-established classification methods such as Support Vector Machines (SVM), Decision Trees, Neural Networks among others (14). One important challenge is to define suitable features from data that led to better separation between classes, i.e. to produce suitable feature space for classification (15). Therefore, the feature extraction can be crucial for the classification process and its accuracy.

It is known that ncRNAs molecules adopt 3D structures which is dependent on the nucleotides order in the sequence. Thus, to extract features from nucleotide sequences considering only its frequency may not fully capture the differences

[*]To whom correspondence should be addressed. Tel: +55 43 3520 3925; Fax: +55 43 3520 4010; Email: fabricio@utfpr.edu.br

between the coding and non-coding molecules (16). Besides, it is commonly known that information content of genomes has a very important function in the existence and development of organisms (17). In this way, even not calculating the RNA secondary structure, it is important to extract features that can capture the information content of genomes. One possible way is to consider the adjacency and frequency patterns of the nucleotides.

In this context, methods for classifying transcripts have been proposed. Among these methods are Coding Potential Calculator (CPC) (18) and CPC2 (19). CPC2 is an updated version of the CPC method, which adopts six features obtained from a transcript molecule. Three of the features are based on the Open Reading Frame prediction (ORF) and the other three are based on the alignment from UniProt proteins (20).

The features related to the ORF are the log-odds score of the prediction, the coverage of the ORF and a binary value that indicates whether the ORF starts with a start codon and ends with a stop codon. The features related to protein alignment are number of hits, a HSP (High Scoring Pairs) based score, and a frame score based on HSPs distribution from the reading frames. The CPC2 adopts some important features to be used by a SVM classifier. For example, if a sequence encodes a protein it must have an ORF and a good alignment with the corresponding protein.

However, the CPC2 method requires the existence of known protein sequences, i.e., the feature extraction is dependent on data other than the nucleotides sequence. Therefore, the method may have limitations when extracting features from *de novo* sequencing of new organisms or unknown proteins.

The PLEK (*predictor of long non-coding RNAs and messenger RNAs based on an improved k-mer scheme*) (21) is an alignment-free method, i.e. does not depend on alignment with pre-existing databases. It adopts the k-mer frequency as a feature from a sliding window with a step of one in order to count the k-mer ranging from 1 to 5 (that is, 4^k patterns for each k value). The frequency of each pattern is weighted by its size. The frequency set is used as feature vector in the SVM classifier. However, only the nucleotide frequency is considered directly. Its feature extraction does not take into account features related to the molecule structure such as the position or the adjacency between the nucleotides.

The Coding Non-Coding Index (CNCI) method (22) adopts codons in order to distinguishing non-coding RNAs, mainly to improve the accuracy levels with respect to the identification of lncRNAs. CNCI calculates the frequency of the 64 codons in each sequence using a sliding window, which scanned each transcript six times to generate six reading frames. For each one, the method calculates a score of the sequence. Then, the most likely coding domain sequence is identified. The six extracted features are related to the size and to the nucleotide frequency, which are used as feature vector in the SVM classifier.

The challenge of feature extraction process is to find the suitable way to obtain measurements from an object whose values are similar for objects in the same class and dissimilar to objects in different classes. In some cases, the form of the original data is not the most suitable for the direct extraction of measures and it is useful to change the representation to another feature space (14,23). In this way, the complex networks theory and its measurements have been used to represent different objects and extract more global and comprehensive features in several domains (24–30) such as interactome (31,32), cellular organization (33), three-dimensional genome organization (34) and gene networks (35–38).

This work proposes the BASiNET (BiologicAl Sequences NETwork) method based on feature extraction from complex networks. The adopted representation takes into account the global neighborhood between sequence sections of certain length, characterizing the adjacency between them through a complex network. The method does not require prior annotation of the genome, nor alignment of the sequences in database. Only the nucleotides sequence in FASTA format is required. BASiNET does not classify NGS reads directly, but assembled transcripts. The classifier is trained with assembled sequences that are longer than the NGS reads. The proposed method was evaluated and compared to the main competing methodologies considering two datasets, the first with sequences from nine species and the second with sequences from six species. In both experiments the BASiNET obtained higher accuracy than competitors methods (CNCI, PLEK and CPC2) in all evaluated species. In addition, the software that implements the BASiNET methodology is freely available at https://cran.r-project.org/package=BASiNET.

The available version of the software allows new sequences to be classified for the previously trained organisms. BASiNET implements a supervised learning algorithm. Thus, to be used with newly sequenced species, it is necessary to know the classes of a subset of transcript for the training step. Therefore, in addition, the user can perform classifier training for newly sequenced organisms since a subset of training sequences are known.

MATERIALS AND METHODS

Materials

This work adopted two datasets in order to validate the proposed method as well as to compare its results with the main competitor methods. The first dataset was obtained from PLEK (21), which presents transcripts (mRNA) and non-coding transcripts (ncRNAs) from nine species of vertebrates. The Human non-coding transcripts were obtained from the GENCODE v17 and the protein-coding transcripts from the RefSeq release 60. The mouse lncRNA were obtained from the GENCODE vM2 and the mRNA also from RefSeq release 60. The Ensembl database v72 were used to collect transcripts of the other vertebrates. The second dataset was obtained from CPC2 (19), which presents transcripts (mRNA), small non-coding transcripts (sncRNAs) and long non-coding transcripts (lncRNAs) from six species: human, mouse, zebrafish, fly, worm and the model plant *Arabidopsis thaliana*. The protein-coding sequences were obtained from RefSeq database for which proteins are annotated by Swiss-Prot. Redundant sequences were removed from the dataset. The non-coding sequences were obtained from the Ensembl v87 and Ensembl Plants v32.

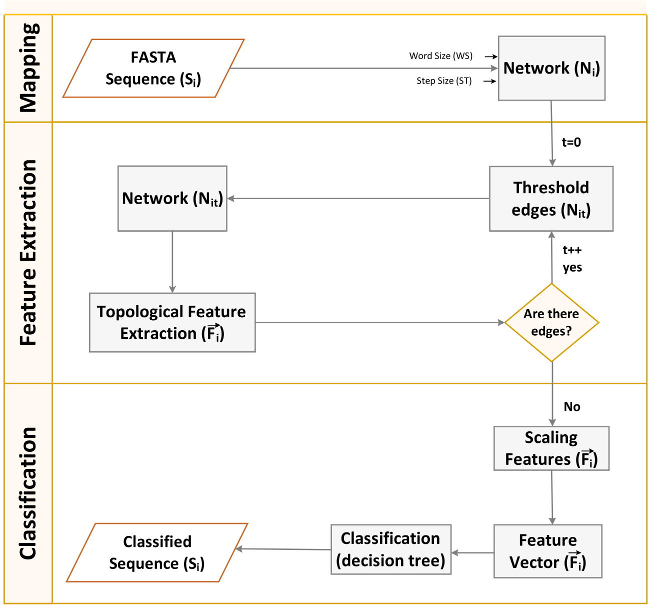
Mapping
Feature Extraction
Classification
FASTA Sequence (S_i)
Word Size (WS)
Step Size (ST)
Network (N_i)
t=0
Threshold edges (N_{it})
Network (N_{it})
t++
yes
Topological Feature Extraction ($\vec{F_i}$)
Are there edges?
No
Scaling Features ($\vec{F_i}$)
Feature Vector ($\vec{F_i}$)
Classification (decision tree)
Classified Sequence (S_i)

Figure 1. BASiNET method overview. Mapping: the sequence in FASTA format is converted in a weighted graph. Feature extraction: topological measures are computed for networks at different weight thresholds. Classification: The topological measures are used as features in order to classify the sequences.

Method

In this work, the complex networks are represented by undirected weighted graphs. Formally, a graph G is defined as a set of vertices V (or nodes) and edges E (or links): $G = \{V, E\}$. The edges are represented by a pair (i, j) which corresponds to a link from vertex i to vertex j. In the case of undirected graphs, an edge (i, j) indicates that there is a link between i and j independent of the direction. For each edge is also associated a weight $w(i, j) \in \mathbb{N}$. Thus, a graph is represented by an adjacency matrix, A, which can be obtained from the application of a threshold function, $\Theta(W, t)$. Where, W is the weight matrix and each position $w_{i,j} = w(i, j)$. $t \in \mathbb{N}$ is the threshold value. The function $a_{i,j} = 1$ if $w_{i,j} > t$ and $a_{i,j} = 0$, otherwise. Where $a_{i,j}$ is a position of the adjacency matrix. In this way, from the adjacency matrix A, several measurements of network characterization can be obtained.

An overview of the method is presented in Figure 1. The method consists of three steps: (i) mapping, (ii) feature extraction and (iii) classification. Figure 1 presents the overview of the BASiNET method.

In a general way, the Mapping step consists of (i) input of the sequence in FASTA format and (ii) create the network representation from the sequence considering the parameters Word Size (WS) and Step Size (ST). The Feature Extraction step includes: (i) application of thresholds in each of the networks and consequent reduction of the number of edges and (ii) the extraction of network topological measurements in each threshold. A topological measurement is the computation of some feature of the graph. For example, the average degree (number of edges) of the vertices. The Classification step is performed after the feature extraction of all sequences and respective networks considered in the experiment. Each step will be explained in detail bellow. As a result, it is built an undirected weighted network. The weight of the edges represent the frequency that each word (vertex) was identified as a neighbor from the other word. Thus, the topology of the created network represents the organization of all adjacencies of all words in the sequence.

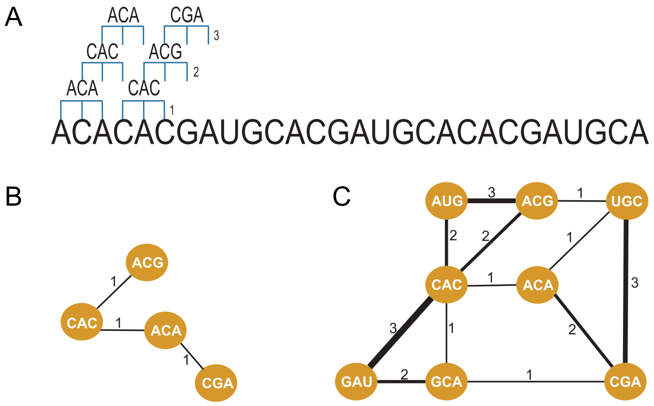

Figure 2. Mapping the sequence to an undirected weighted network. (**A**) The first three iterations of the algorithm with $WS = 3$ and $ST = 1$. Iteration 1: the vertex ACA is linked to vertex CAC, which are adjacent words in the sequence with this ST. The window slides with $ST = 1$. Iteration 2:the vertex CAC is linked to ACG. The window slides with $ST = 1$; The vertex ACA is linked to CGA. (**B**) Building the network from the first three illustrative iterations in sub-figure (a). At each interaction, two adjacent words occur again, +1 is added to the weight of the edge. For example, the words GAU and CAC occurred three times as neighbors. (**C**) The result network after all iterations.

Mapping. The Mapping step (Figure 2) is performed as follows. For each transcript are considered their nucleotides (A, C, U, G) in FASTA format. There are two parameters: (i) WS which is related to the number of nucleotides considered to compose a network vertex, for example, a WS = 3 means that will be considered words of 3 adjacent nucleotides and (ii) ST, which refers to the length of the step to the next neighboring word. Figure 2 presents an illustrative example of the Mapping step for the RNA sequence 'ACAC ACGAUGCACGAUGCACACGAUGCA' adopting the parameters $WS = 3$ and $ST = 1$ applied in this work. The complete mapping example is available at supplementary file 1.

Feature extraction. The second step, Feature Extraction is performed in order to consider different resolutions of a network in each iteration. The weight of an edge represents the frequency of the adjacency between the pair of vertices. Since some adjacent sequences can be more frequent than others, the method apply a threshold to the weight of the edges to capture the adjacencies at different frequencies. Thus, the method starts considering all the network edges and after each iteration considering only the most persistent edges (patterns). Figure 3 presents an illustrative example of the Feature Extraction step.

More specifically, the initial iteration considers all the identified network edges and the adopted network measurements are extracted from the network, i.e. the threshold is

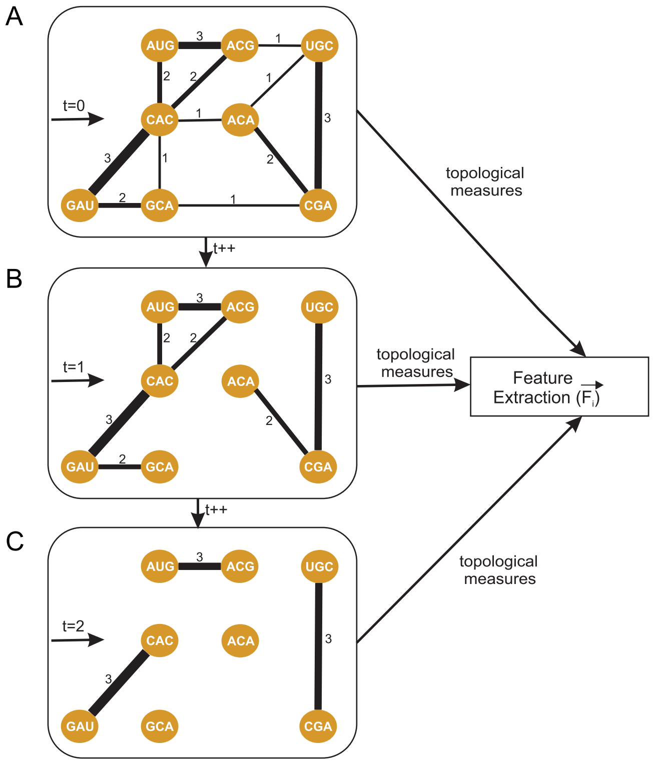

Figure 3. Overview of the Feature Extraction step. (**A**) $t = 0$ considers all the identified edges and the topological measures are extracted and appended in the respective feature vector, one for each sequence. (**B**) $t = 1$ and (**C**) $t = 2$ the threshold is applied removing the edges lower than the threshold and the topological measures are extracted and appended in the respective feature vector, one for each sequence. $t = 3$ produce a network without edges and the process is terminated.

not applied, $t = 0$. The threshold is incremented (t ++) in the next iteration, thus removing the edges with weight equal to one and again the adopted network measurements are extracted from the resulting network. The iterations are repeated until no more edges left in the network.

As a result, a feature vector is built in order to characterize the network considering the different levels (thresholds) of frequencies of adjacent nucleotides.

Regarding the network measures, the complex networks can be characterized by their topological measurements, which are essential in many network investigations, including representation, characterization, classification and modeling (27–29,31,33,35,39–43). Thus, 10 topological measures commonly used for the network characterization were adopted: assortativity (ASS), average degree (DEG), maximum degree (MAX), minimum degree (MIN), average betweenness centrality (BET), clustering coefficient (CC), average short path length (ASPL), average standard deviation (SD), frequency of motifs with size 3 (MT3) and frequency of motifs with size 4 (MT4) (25,26,28).

Classification. The final step is the Classification from the feature vector, one for each input sequence. In order to avoid the influence of the different sequences length and different scales from network measures, a Min-Max rescale procedure is applied to the feature vectors. Consider a Feature Vector $\vec{F} = f_1, f_2, \ldots, f_m$, the Min-Max Normalization maps a value f_k to fn_k in the range [0, 1] defined as $fn_k = (f_k - f_{min})/(f_{max} - f_{min})$. Where f_{max} is the maximum value, f_{min} is the minimal value for the topological measure, f_k is the original measure value and fn_k is the respective rescaled value. As a result, all the topological measures are defined into the interval [0, 1] and than the decision tree algorithm (23) is performed in order to classify the sequences with 10-fold cross-validation.

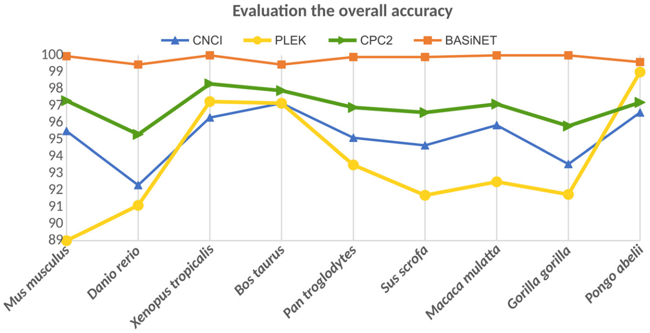

Figure 4. Overall classification accuracy using BASiNET compared to the CNCI, PLEK and CPC2 methods in the first experiment.

RESULTS

To evaluate BASiNET accuracy for classification of mRNAs and ncRNAs, prediction results were compare with CNCI (22), PLEK (21) and CPC2 (19) competitor methods using cross-species data from nine vertebrates species. Table 1 presents the obtained results in the first experiment by considering the proposed method and the following competitors.

The results indicate the separability of the classes since BASiNET reached higher accuracy levels than other methods for the classification of mRNAs in all observed species. It is also possible to observe that BASiNET achieve the average accuracy of mRNAs 7.46% higher than CNCI; 10.25% higher than PLEK and 5.24% higher than CPC2. Regarding the identification of ncRNAs, BASiNET achieve the average accuracy of 1.7%, 2.1%, 0.5% higher than CNCI, PLEK and CPC2, respectively. In addition, BASiNET obtained higher average results and lower standard deviation in both coding and non-coding mRNA classification. Figure 4 shows the overall accuracy obtained for mRNA and ncRNA prediction in the first experiment.

BASiNET was also used to predict the three transcript classes: mRNA, lncRNA and sncRNA using dataset from six species.The second experiment was developed in order to evaluate the proposed method as well as to compare its results with competitor methods considering a cross-species prediction with six species and three transcript classes: mRNA, lncRNA and sncRNA. The adopted dataset in the second experiment was obtained in (19). Table 2 presents the second experiment results by considering the proposed method and the same competitors methods adopted in the first experiment.

Table 2 presents the obtained accuracy considering six organisms and three classes of transcripts for each specie,

Table 1. Accuracy for the classification using BASiNET compared to the CNCI, PLEK and CPC2 methods in the first experiment

Species	Class of RNA	Transcripts	CNCI	PLEK	CPC2	BASiNET
Mus musculus	mRNA	26062	93.9	88.1	94.7	100.0
	ncRNA	2963	97.1	89.9	99.9	99.9
Danio rerio	mRNA	14493	95.3	91.3	96.6	100.0
	ncRNA	419	89.3	90.9	94.0	98.9
Xenopus tropicalis	mRNA	8874	92.9	94.5	96.5	100.0
	ncRNA	279	99.7	100.0	100.0	100.0
Bos taurus	mRNA	13190	94.3	94.8	95.9	100.0
	ncRNA	182	100.0	99.5	100.0	98.9
Pan troglodytes	mRNA	1906	90.2	87.1	93.9	100.0
	ncRNA	1166	100.0	99.9	100.0	99.8
Sus scrofa	mRNA	3978	93.4	85.1	94.9	99.9
	ncRNA	241	95.9	98.3	98.3	99.6
Macaca mulatta	mRNA	5709	92.0	85.0	94.2	100.0
	ncRNA	359	99.7	100.0	100.0	100.0
Gorilla gorilla	mRNA	33025	87.4	83.8	91.6	100.0
	ncRNA	367	99.7	99.7	100.0	100.0
Pongo abelii	mRNA	3401	93.4	98.0	94.4	100.0
	ncRNA	392	99.8	100.0	100.0	99.2
Average	mRNA	—	92.53	89.74	94.75	99.99
	ncRNA	—	97.91	97.58	99.13	99.59
Overall average	**mRNA and ncRNA**	**—**	**95.22**	**93.66**	**96.94**	**99.79**
Standard deviation	mRNA	—	2.27	4.81	1.45	0.03
	ncRNA	—	3.35	3.88	1.89	0.44

in comparison with other methods. These results indicate that BASiNET achieve better results for mRNA classification than competitors for all adopted species presenting improvements of 9.67%, 19.7%, 3.6% higher than CNCI, PLEK and CPC2, respectively. Regarding the ncRNAs (lncRNAs and sncRNAs), it is observed that even though BASiNET have reached lower individual accuracy than mRNA classification, the accuracies were improved in 2.74%, 3.28% and 2.53% than CNCI, PLEK and CPC2, respectively and present lower standard deviation in both coding and non-coding identification. The achieved results reinforces the adequacy of the method and its robustness, presenting lower variation. Only for sncRNAs the BASiNET presents 99.7% of accuracy, slightly lower than PLEK and CPC2, which presented 100% of correct results for the three class of transcript prediction. Figure 5 shows the overall accuracy obtained for the three class RNA.

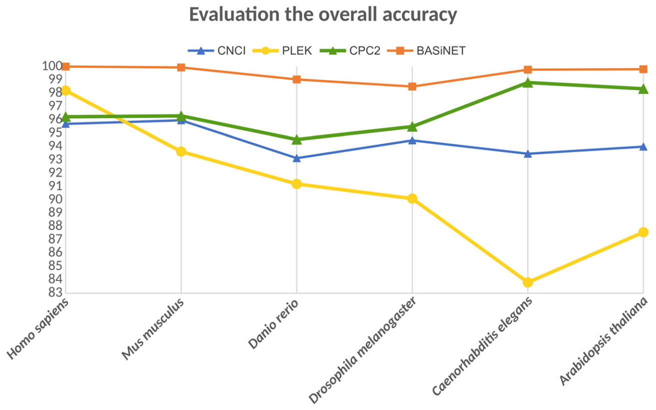

Figure 5. Overall classification accuracy using BASiNET compared to the CNCI, PLEK and CPC2 methods in three class of transcript prediction

DISCUSSION

In order to identify which adopted complex network measurements were most relevant in the sequence classification, the generated decision trees allow the identification of the frequency that each measure was applied in the RNA species classification. In the first analyses, performed with the PLEK dataset, six topological measures were selected for classification in order of decreasing relevance: ASPL, BET, DEG, ASS, MAX and MIN. Figure 6 presents the frequencies of each topological measure selected for decision tree classification.

Regarding the second experiment, performed with the CPC2 dataset, all the ten adopted topological measures were selected by the decision trees for the RNA classification. Again the ASPL and BET topological measures were selected as the main features. Figure 7 presents the frequencies at which each topological measure was selected as a feature for decision tree classification.

Table 2. Accuracy for the classification using BASiNET compared to the CNCI, PLEK and CPC2 methods in the second experiment

Species	Class of RNA	Transcripts	CNCI	PLEK	CPC2	BASiNET
Homo sapiens	mRNA	6142	91.4	97.0	95.9	100.0
	lncRNA	7485	99.2	97.6	92.8	100.0
	sncRNA	4534	96.5	100.0	100.0	100.0
Mus musculus	mRNA	10638	91.9	89.2	93.9	100.0
	lncRNA	6460	96.8	91.7	95.0	99.9
	sncRNA	5791	99.2	100.0	100.0	99.9
Danio rerio	mRNA	2344	95.9	94.4	95.5	99.5
	lncRNA	1163	99.5	79.2	88.1	98.9
	sncRNA	365	84.0	100.0	100.0	98.7
Drosophila melanogaster	mRNA	3680	94.8	82.8	94.6	98.5
	lncRNA	2776	99.1	87.5	91.9	97.3
	sncRNA	780	89.5	100.0	100.0	99.7
Caenorhabditis elegans	mRNA	3551	82.9	53.0	96.5	100.0
	lncRNA	1582	99.3	98.4	99.9	99.4
	sncRNA	7888	98.2	100.0	100.0	99.9
Arabidopsis thaliana	mRNA	13986	82.8	63.1	99.7	99.7
	lncRNA	2562	99.7	99.6	95.3	99.7
	sncRNA	1291	99.5	100.0	100.0	100.0
Overall average	**mRNA**	—	**89.95**	**79.92**	**96.02**	**99.62**
	lncRNA and sncRNA	—	**96.71**	**96.17**	**96.92**	**99.45**
Standard deviation	mRNA	—	5.76	17.92	2.03	0.58
	lncRNA and sncRNA	—	4.91	6.67	4.18	0.81

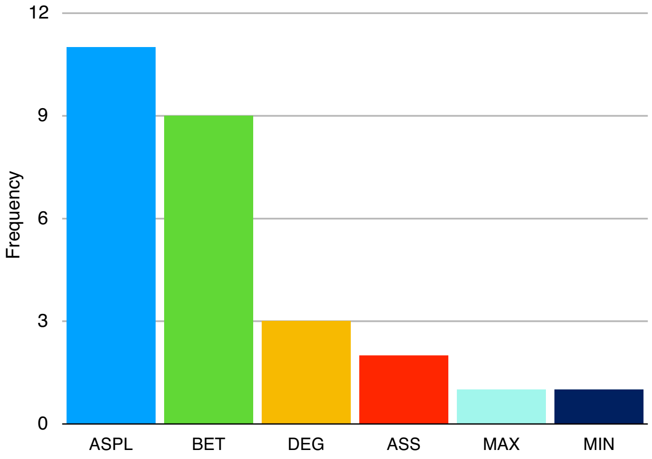

Figure 6. Histogram of the topological measures used in decision trees considering the first experiment.

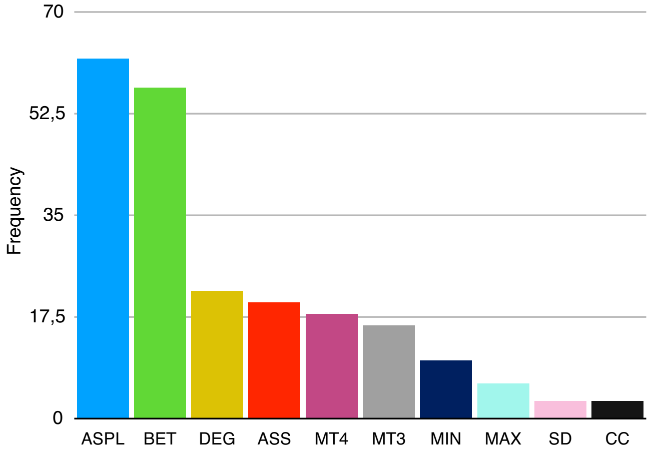

Figure 7. Histogram of the topological measures used in decision trees considering the second experiment.

The ASPL and BET features were the most frequent in all the experiments, indicating their relevance to the classification. Therefore, it is important to observe what they mean and better understand its importance in the classification. A path between a pair of vertices (*i*, *j*) is a sequence of edges that connect *i* and *j*. A Shortest Path Length between *i* and *j* is one of the paths with minimum length. The ASPL is the average of the shortest paths between all the vertex pairs in the network. Let $s_{(i,j)}$ the shortest path value between *i* and

j. The ASPL is defined as:

$$\mathrm{ASPL} = \frac{1}{N(N-1)} \sum_{i \neq j} s_{(i,j)} \quad (1)$$

The BET quantifies the relevance of a vertex in relation to all the paths of the network. By computing all the shortest paths between a pair of vertices (i, j), is obtained for a vertex $v \neq (i, j)$ the number of paths that pass through it. Thus, BET (betweenness) quantifies the proportion of paths passing through v in relation to all the paths between i and j, defined as:

$$B_v = \frac{\sum_{i,j} Q(i, j, v)}{Q(i, j)}, i \neq j \quad (2)$$

where, $Q(i, j, v)$ is the number of shortest paths between i and j that pass through v. $Q(i, j)$ is the number of minimum paths between i and j. The BET is defined as:

$$\mathrm{BET} = \frac{1}{N} \sum B_v \quad (3)$$

where N is the number of network vertices.

The more paths pass through a vertex, the greater will be your betweenness. A vertex that acts as a link between two or more groups of few vertices connected together must have a high betweenness since the paths between the different groups must pass through that vertex. In the proposed method a vertex corresponds to word ('piece') of the RNA sequence and an edge corresponds to the adjacency between the words.

In this way, a vertex with high betweenness could correspond to a codon that occurred frequently. However, sparsely dispersed across the sequence between little connected vertex groups. This codon would be adjacent to several groups. One possibility to be investigated is that this measure may map periodic patterns in general. Therefore, the betweenness could characterize patterns features that distinguish coding and non-coding RNAs.

Regarding the ASPL, a high shortest path length between a pair of vertices, that is, between a pair of sequence words, indicates that there are many words between them. More simply, these words are distant in the sequence. For example, a minimum path equal to 10 between the words GAU and UGC means that there are at least nine other words between them. On the other hand, a low value of ASPL indicates that these words occur in near positions in the sequence. In this way, this metric potentially characterizes relations between nonadjacent words, capturing the global structure of distant words.

Regarding the threshold, it was adopted in order to generate networks with different resolutions of edges frequencies, initially considering all the identified edges and than maintaining only the most frequents after each iteration. In other words, a higher threshold can capture repetitions. Thus, this criterion associated with the adopted measurements of complex networks possibly maps more meaningful structures of the biological sequences.

The BASiNET presented superior accuracy than other methods in the classification of all organisms in all adopted datasets. All other methods use specific and known biological characteristics which are important for the biological sequences. However, there may be relationships between 'pieces' of sequences that are not directly identifiable by other features extraction techniques. The results indicate that the adopted feature extraction possibly map the pattern of the adjacency and frequency of nucleotides that distinguishes the two classes of molecules.

The proposed method identifies adjacent sections of genomic sequences, counts their frequencies and constructs a weighted graph. This is, it maps how many times different adjacent sections are repeated with different frequencies. However, instead of mapping one type of repetition at a time, the graph represents a global map of the all frequencies of all adjacent sections across the entire sequence. To summarize and quantify this mapping, measurements of complex networks are applied to the graph.To identify the adjacency patterns at different frequency levels, the threshold is applied. Then, to summarize and quantify this mapping, measurements of complex networks are applied to the graphs. Thus, the method is able to extract this type of global feature from genomic sequences and (possibly) when this type of feature distinguishes two classes of sequences, the method must accurately classify RNAs.

For example, simple sequence repeats (SSR) are known to be more abundant in non-coding DNA than in protein coding sequences (44). Possibly, the graphs of these two classes of sequences have different weights at the edges and different topologies due to the different frequency and adjacency patterns of repeats.

In addition, it is known that RNA structure patterns are related to the sequence. For example, there is a relationship between the structure of loops and the conservation of RNA sequence (45). It is possible that the method captures the adjacency and frequency related to these sections of the sequences.

Thus, BASiNET does not requires previous alignment and works well when longer sequences are available.

The computation time of BASiNET is suitable for the classification of thousands of sequences. The supplementary file 2 presents the computational time for all adopted species in order to make clear that BASiNET can be used from a personal computer.

CONCLUSION

The classification between mRNA and long and small ncRNAs sequences is challenging in face of the large amount of new data produced by high-performance sequencing, in particular, by the *de novo* sequencing data. This work presented a feature extraction method for biological sequences (RNAs) classification based on complex networks and its topological measures. In the proposed method the sequences are mapped and represented by means of a complex networks. Besides the representation, the method makes the characterization by calculating topological measures of the network. The measurements form a feature vector that is used to classify the sequences. The method was applied in two datasets presented in PLEK and CPC2 methods. The BASiNET results were compared with the CNCI, PLEK and CPC2 methods. The accuracy results from the comparison of BASiNET with other methods, when applied in the two datasets, showed that BASiNET outperformed the others in all data sets, including mRNAs and ncRNAs. In addition, the results indicate that the representation of biological sequences as complex networks is able to extract features more adequate to its classification than those adopted by the other methods.

Finally, the BASiNET method was implemented in open source (R language) and the program is freely available for download at https://cran.r-project.org/package=BASiNET.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr. André Yoshiaki Kashiwabara for his valuable suggestions.

FUNDING

Conselho Nacional de Desenvolvimento Científico e Tecnológico [406099/2016-2 to F.M.L., Research Fellowship to L.F.P.P., Scholarship PIBITI-CNPq to E.A.T.]; Federal University of Technology – Paraná; Fundação Araucária. Funding for open access charge: Federal University of Technology – Paraná.
Conflict of interest statement. None declared.

REFERENCES

1. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
2. Goodwin,S., McPherson,J. D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.
3. Ozsolak,F. and Milos,P.M. (2010) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
4. da Fonseca,R.R., Albrechtsen,A., Themudo,G.E., Ramos-Madrigal,J., Sibbesen,J.A., Maretty,L., Zepeda-Mendoza,M.L., Campos,P.F., Heller,R. and Pereira,R.J. (2016) Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Mar. Genomics*, **30**, 3–13.
5. Wang,K.C. and Chang,H.Y. (2011) Molecular mechanisms of long noncoding RNAs. *Molecular cell*, **43**, 904–914.
6. Wang,X., Arai,S., Song,X., Reichart,D., Du,K., Pascual,G., Tempst,P., Rosenfeld,M.G., Glass,C.K. and Kurokawa,R. (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature*, **454**, 126–130.
7. Snustad,D.P. (2011) *Principles of Genetics*, John Wiley And Sons, Inc., NY, p. 1999.
8. Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.
9. Wapinski,O. and Chang,H.Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol.*, **21**, 354–361.
10. Spizzo,R., Almeida,M.I., Colombatti,A. and Calin,G.A. (2012) Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene*, **31**, 4577–4587.
11. Zhao,Y., Li,H., Fang,S., Kang,Y., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q., Chen,R. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**, D203–D208.
12. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2012) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.
13. Lee,R.C. and Ambros,V. (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.
14. Bishop,C.M. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus.
15. Saeys,Y., Inza,I. and Larranaga,P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.
16. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*, Cambridge University Press.
17. Abante,J., Ghaffari,N., Johnson,C.D. and Datta,A. (2017) HiMMe: using genetic patterns as a proxy for genome assembly reliability assessment. *BMC Genomics*, **18**, 694.
18. Kong,L., Zhang,Y., Ye,Z.-Q., Liu,X.-Q., Zhao,S.-Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
19. Kang,Y.-J., Yang,D.-C., Kong,L., Hou,M., Meng,Y.-Q., Wei,L. and Gao,G. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, **45**, W12–W16.
20. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**,D158–D169.
21. Li,A., Zhang,J. and Zhou,Z. (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, **15**, 311.
22. Sun,L., Luo,H., Bu,D., Zhao,G., Yu,K., Zhang,C., Liu,Y., Chen,R. and Zhao,Y. (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.*, **41**, e166.
23. Theodoridis,S. and Koutroumbas,K. (2008) *Pattern Recognition*, 4th edn, Academic Press, USA.
24. Vazquez,A., Dobrin,R., Sergi,D., Eckmann,J.-P., Oltvai,Z.N. and Barabasi,A.-L. (2004) The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 17940–17945.
25. Boccaletti,S., Latora,V., Moreno,Y., Chavez,M. and Hwang,D.-U. (2006) Complex networks: Structure and dynamics. *Phys. Rep.*, **424**, 175–308.
26. Milo,R., Shen-Orr,S., Itzkovitz,S., Kashtan,N., Chklovskii,D. and Alon,U. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.
27. Newman,M.E.J. (2003) The Structure and Function of Complex Networks. *SIAM Review*, **45**, 167–256.
28. Costa,L.d.F., Rodrigues,F.A., Travieso,G. and Villas Boas,P.R. (2007) Characterization of complex networks: a survey of measurements. *Adv. Phys.*, **56**, 167–242.
29. Backes,A.R., Casanova,D. and Bruno,O.M. (2013) Texture analysis and classification: a complex network-based approach. *Inform. Sci.*, **219**, 168–180.
30. Conque,B.M.M., Kashiwabara,A.Y. and Lopes,F.M. (2016) A feature extraction approach based on complex networks for genomic sequences recognition. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. pp. 1803–1807.
31. Barabási,A.-L., Gulbahce,N. and Loscalzo,J. (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.
32. Pavlopoulos,G.A., Secrier,M., Moschopoulos,C.N., Soldatos,T.G., Kossida,S., Aerts,J., Schneider,R. and Bagos,P.G. (2011) Using graph theory to analyze biological networks. *BioData Mining*, **4**, 10.
33. Barabási,A.-L. and Oltvai,Z. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
34. Kruse,K., Sewitz,S. and Babu,M.M. (2012) A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. *Nucleic Acids Res.*, **41**, 701–710.
35. Costa,L.d.F., Rodrigues,F.A. and Cristino,A.S. (2008) Complex networks: the key to systems biology. *Genet. Mol. Biol.*, **31**, 591–601.
36. Lopes,F.M., Cesar Jr,R.M. and Costa,L.D.F. (2011a) Gene expression complex networks: synthesis, identification, and analysis. *J. Comput. Biol.*, **18**, 1353–1367.
37. Lopes,F.M., Martins,D.C., Barrera,J. and Cesar,R.M. (2014) A feature selection technique for inference of graphs from their known topological properties: Revealing scale-free gene regulatory networks. *Inform. Sci.*, **272**, 1–15.
38. Vicente,F.F.R. and Lopes,F.M. (2014) SFFS-WS: A feature selection algorithm exploring the small-world properties of GNs. In *Pattern*

Recognition in Bioinformatics, Proceedings Springer Berlin / Heidelberg 9th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB), Stockholm, Sweden Vol. 8626 of Lecture Notes in Computer Science, pp. 60–71.

39. Albert,R. and Barabási,A.-L. (2002) Statistical mechanics of complex networks. *Rev. Modern Phys.*, **74**, 47–97.
40. Albert,R. (2005) Scale-free networks in cell biology. *J. Cell Sci.*, **118**, 4947–4957.
41. Aittokallio,T. and Schwikowski,B. (2006) Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.*, **7**, 243–255.
42. Khanin,R. and Wit,E. (2006) How scale-free are biological networks. *J. Comput. Biol.*, **13**, 810–818.
43. Barabási,A.-L. (2009) Scale-free networks: a decade and Beyond. *Science*, **325**, 412–413.
44. Dokholyan,N.V., Buldyrev,S.V., Havlin,S. and Stanley,H. (2000) Distributions of dimeric tandem repeats in non-coding and coding DNA sequences. *J. Theoret. Biol.*, **202**, 273–282.
45. Schudoma,C., May,P., Nikiforova,V. and Walther,D. (2010) Sequence-structure relationships in RNA loops: establishing the basis for loop homology modeling. *Nucleic Acids Res.*, **38**, 970–980.