Genomics 107 (2016) 9–15

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Non-coding RNA identification based on topology secondary structure and reading frame in organelle genome level

Cheng-Yan Wu, Qian-Zhong Li *, Zhen-Xing Feng

Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 8 December 2015
Accepted 12 December 2015
Available online 15 December 2015

Keywords:
Non-coding RNA
Spatial structure
RNA motif
Reading frame
Increment of diversity

ABSTRACT

Non-coding RNA (ncRNA) genes make transcripts as same as the encoding genes, and ncRNAs directly function as RNAs rather than serve as blueprints for proteins. As the function of ncRNA is closely related to organelle genomes, it is desirable to explore ncRNA function by confirming its provenance. In this paper, the topology secondary structure, motif and the triplets under three reading frames are considered as parameters of ncRNAs. A method of SVM combining the increment of diversity (ID) algorithm is applied to construct the classifier. When the method is applied to the ncRNA dataset less than 80% sequence identity, the overall accuracies reach 95.57%, 96.40% in the five-fold cross-validation and the jackknife test, respectively. Further, for the independent testing dataset, the average prediction success rate of our method achieved 93.24%. The higher predictive success rates indicate that our method is very helpful for distinguishing ncRNAs from various organelle genomes.

1. Introduction

Non-coding RNAs (ncRNAs) are the major products of the genome and play important roles in the regulation of genome post-transcription, cell growth, cell differentiation and proliferation. In recent years, there was a major focus on detecting RNA transcripts with no apparent protein-coding potential, due to that ncRNAs might play an important role in the regulation of vital cellular functions [1–3]. The experimental results show that some ncRNAs act as molecular switches that regulate gene expression and are related to human diseases [4–7], and ncRNAs may be the missing links in well-known oncogenic networks and tumor suppressive networks [8].

Since the function of ncRNAs correlates with their organelle genomes, the knowledge of their provenances may be very useful in understanding their role in the biological processes. *Chloroplast*, *Kinetoplast*, *Mitochondrion* and *Nuclear* are essential organelles in eukaryotic cells, they contain themselves DNA and distinct genetic systems [9,10]. Nuclear genome control most genes in the cell, while other organelle genomes related to themselves organelle's functions. Although other organelle genomes are different from nuclear genome, they have also striking similarities, such as their evolution from prokaryotes, uniparental inheritance and their dependence on nuclear genes for biogenesis. In recent years, the research on organelle genome has attracted more and more people's attentions [11–15]. And the discrimination of ncRNAs from organelle genomes by computational recognition algorithms may provide a new thought for the theoretical description of ncRNAs' function. However, the predictive investigation of ncRNA in the organelle genome has not been found, many effective researches relating to the prediction of ncRNAs mainly focused on miRNAs, siRNAs, etc. [16–21]. In this study, an ncRNA's dataset with less than 80% sequence identity (ncRNA_361) for different organelle genomes was built from NONCODE v3.0 [22]. And trinucleotides from three kinds of reading frames, the topology secondary structure parameters of ncRNAs and the motif information from ncRNA primary sequences are used to represent the features of ncRNAs. Based on these features, the ID_SVM algorithm [23–25] is proposed to classify the ncRNAs in four kinds of organelle genomes. The prediction performance is examined by using five-fold cross-validation test, the jackknife test and the independent dataset test. The good results indicated that the exacted information parameters can be the important features of ncRNAs.

2. Materials and methods

2.1. Dataset for classification experiments

The ncRNA sequences were downloaded from NONCODE v3.0, the dataset include ncRNAs residing in which genomes [22]. Only those ncRNAs with 'organelle genome' appeared in 'Keywords' were considered to construct a standard dataset. 468 ncRNA sequences were obtained, covering four kinds of organelle genomes: *Chloroplast* genome (*ch*., 67 ncRNAs), *Kinetoplast* genome (*ki*., 148 ncRNAs), *Mitochondrion* genome (*mi*., 145 ncRNAs) and *Nucleolus* genome (*nu*., 108 ncRNAs). The sequences length distribution for each kind of organelle genomes is shown in Fig. 1.

* Corresponding author.
E-mail address: qzli@imu.edu.cn (Q.-Z. Li).

http://dx.doi.org/10.1016/j.ygeno.2015.12.002
0888-7543/

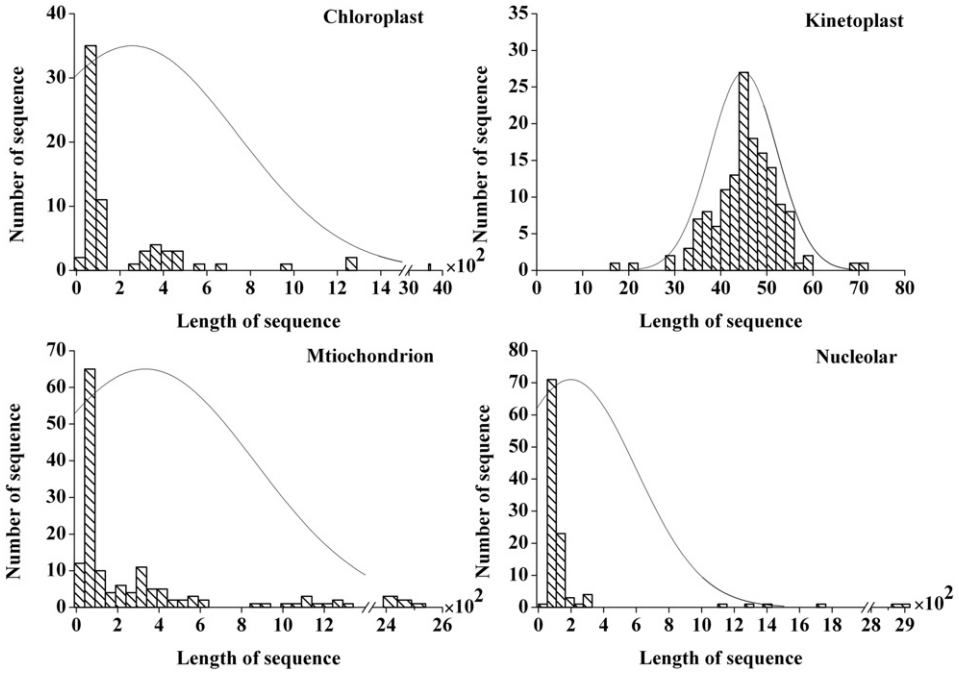

Fig. 1. The length distribution of four types of organelle genomes in standard dataset.

Fig. 1 shows that the distribution of sequences length (denoted as L) from *Chloroplast* genome and *Mitochondrion* genome are very similar. The number of ncRNA sequences with $L \le 200$ in *Chloroplast* genome and *Mitochondrion* genome is 48 and 90, respectively, which are about 65% of the total number for each genome. Compared with other three genomes, the length distribution of *Kinetoplast* genome sequences is relatively concentrated (15 bp–75 bp). Moreover, the numbers of ncRNAs for four kinds of organelle genomes decreased with the increasing of sequence length.

In this paper, based on the 468 ncRNA sequences, a benchmark dataset was constructed with the cutoff threshold of 80% by using a culling program CD-HIT [26], and named ncRNA_361. These sequences are classified into the four organelle genomes: (1) 57 *Chloroplast* genome, (2) 126 *Kinetoplast* genome, (3) 106 *Mitochondrion* genome and (4) 72 *Nucleolus* genome. The names of sequences in the ncRNA_361 dataset are given in Supplementary material S1.

In order to further estimate the effectiveness of the selected information feature and prediction method, 10% sequences (37 sequences) of each genome in ncRNA_361 are randomly selected as independent testing dataset (named as the ncRNA_Ident dataset). The other 324 sequences are named as the ncRNA_324 dataset. The names of sequences in the ncRNA_Ident dataset are given in Supplementary material S2.

2.2. Schemes of information parameters

2.2.1. Triplets under three reading frames

For a given sequence, according to the first nucleotide of triplets relative to the initiation site of the sequence, there are three kinds of reading frames (denoted by W1, W2, and W3). Take one ncRNA sequence as an example, AGAAUAAUAAAAUAAAAAAUAAAGAGUAGUAAUUUUUAUUAUCCCAUUCGGAUGGAUUAU.

In the first reading frame (W1), triplets are AGA, AUA, AUA, AAA, etc. In the second reading frame (W2), triplets are GAA, UAA, UAA, AAU, etc. In the third reading frame (W3), triplets are AAU, AAU, AAA, AUA, etc., respectively.

It has been shown previously that the structure of ncRNA may provide insights into biological functions. In order to use this ncRNA structural information as a feature set for machine-learning techniques, the RNA-fold software [27] is used to predict the secondary structure of ncRNA sequences. The ncRNA secondary structures are shown as the sequence constituted by two kinds of symbols: brackets and dots. That is to say that each nucleotide only has two statuses: paired denoted by brackets ("(“or")") and unpaired denoted by dots (“.”), in which we do not distinguish “(” and “)” and use “(” for both situations. Thus, the structure-sequence mode is formed by the ncRNA secondary structure constituted by brackets and dots and the primary sequence. Combining with three types of reading frames of the primary sequence, the feature vectors are selected from the structure-sequence. In each kind of reading frame, triplets along the structure-sequence can be denoted as “U(((”, “A((.”, etc., here U or A presents the type of the middle nucleotide among the triplets [28,29].

So, for one ncRNA sequence, the 32 (4×2^3) structure-sequence triplet features are given by one reading frame, denoted by T_η:

$$T_\eta = \left[t_1^A, t_2^A, \cdots, t_8^A, \cdots, t_i^j, \cdots, t_1^U, \cdots, t_7^U, t_8^U\right] \qquad \eta \in \{W1, W2, W3\}$$
$$j = \begin{cases} A & i = 1,2,\cdots,8 \\ C & i = 1,2,\cdots,8 \\ G & i = 1,2,\cdots,8 \\ U & i = 1,2,\cdots,8 \end{cases} \tag{1}$$

where t_i^j is the number of the i-th triplets with the j-th middle nucleotide from the structure-sequence mode under one reading frame. $T_\eta(\eta \in \{W1, W2, W3\})$ denotes the structure-sequence triplets from the first, second or third reading frame. For the three reading frames, the structure-sequence triplet features can be denoted by T_{tri}:

$$T_{tri} = T_{W1} \cup T_{W2} \cup T_{W3} \tag{2}$$

2.2.2. The topology secondary structure parameters

The structure-sequence features only denote that the bases are paired or unpaired in ncRNA primary sequence level. In other words, these features describe ncRNA secondary structure to a single dimension.

Owing to the fact that the structure of an RNA is determined by the complex pattern of base–base interactions, including base paired secondary structures and long-range tertiary interactions [30,31]. And RNA chain frequently folds back on itself to form local structures: base-paired segments and various loops. Therefore, in this paper, the