Software Impacts 12 (2022) 100268

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

ORFhunteR: An accurate approach to the automatic identification and annotation of open reading frames in human mRNA molecules

Vasily V. Grinev [a,*,1], Mikalai M. Yatskou [b,1], Victor V. Skakun [b], Maryna K. Chepeleva [b,c], Petr V. Nazarov [c,*]

[a] Department of Genetics, Belarusian State University, 4 Nezavisimosti Avenue, Minsk 220030, Belarus
[b] Department of Systems Analysis and Computer Modelling, Belarusian State University, 4 Nezavisimosti Avenue, Minsk 220030, Belarus
[c] Department of Cancer Research, Luxembourg Institute of Health, 1AB rue Thomas Edison, L-1445 Strassen, Luxembourg

ARTICLE INFO

Keywords:
Open reading frame
Classification
Gene prediction
Transcriptome

ABSTRACT

The coding potential of RNA molecules can be estimated using algorithms that find open reading frames (ORFs). However, previously developed algorithms show limited performance. We developed a computational approach dedicated to the automatic identification of ORFs in a large set of human mRNA molecules. It is based on the vectorization of nucleotide sequences followed by classification using a random forest. The predictive model was validated on human mRNA molecules from the NCBI RefSeq and Ensembl databases and demonstrated almost 95% accuracy in detecting true ORFs. Our method is implemented into a powerful R/Bioconductor package ORFhunteR.

Code metadata

Current code version	v1.0.0
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2022-6
Permanent link to reproducible capsule	https://codeocean.com/capsule/9151601/tree/v1
Legal code license	MIT
Code versioning system used	git
Software code languages, tools, and services used	C++, R
Compilation requirements, operating environments, and dependencies	R 4.0.2 or later, rtools40 or later. Dependencies (all are installed automatically during package installation): Biostrings, rtracklayer, Peptides, BSgenome.Hsapiens.UCSC.hg38, data.table, stringr, randomForest, xfun, XML, Rcpp.
If available, link to developer documentation/manual	http://bioconductor.org/packages/release/bioc/manuals/ORFhunteR/man/ORFhunteR.pdf
Support email for questions	grinev_vv@bsu.by

1. Introduction

High-throughput technologies allow capturing sequence information about whole transcriptomes with reasonable cost and time. Several high-performance computational approaches have been developed to restore the structure of full-length RNA molecules (or transcripts) from short RNA-Seq reads and to get qualitative and quantitative characteristics of these molecules [1,2]. One of the most important properties of transcripts is their coding potential, which can be assessed by various algorithms. Many methods, including one implemented in ORFfinder

Abbreviations: CPF, category-position-frequency; ECDF, empirical cumulative distribution of frequency (probability); lncRNA, long non-coding RNA; mRNA, messenger RNA(protein-coding RNA); ORF, open reading frame

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding authors.
E-mail addresses: grinev_vv@bsu.by (V.V. Grinev), petr.nazarov@lih.lu (P.V. Nazarov).

[1] Equal contribution.

https://doi.org/10.1016/j.simpa.2022.100268
Received 21 January 2022; Received in revised form 17 February 2022; Accepted 1 March 2022

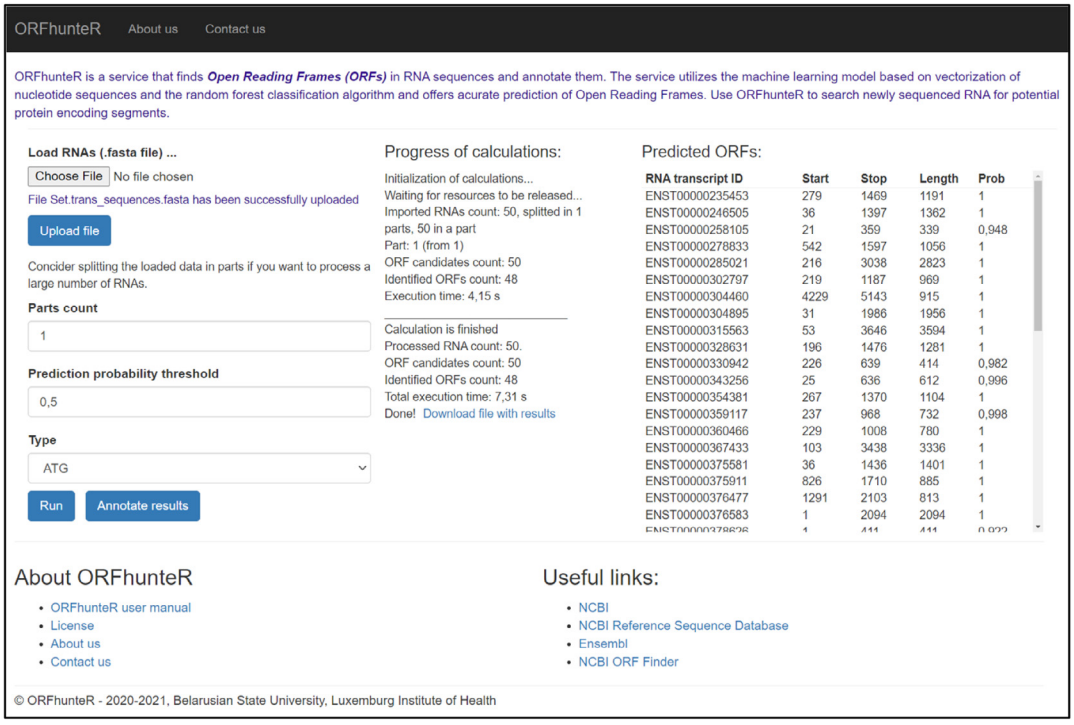

Fig. 1. An interface of the online tool (https://orfhunter.bsu.by/), implementing ORFhunteR. The sample dataset from the corresponding GitHub repository was used here.

at NCBI [3], are based on the selection of the longest open reading frame (ORF) from possible candidates. Other computational techniques, for example, used for meta-genomics analysis [4,5], propose vectorization of sequence features of ORF-candidates, which is an efficient conversion of nucleotide sequence into a vector of features. However, these algorithmic approaches have several limitations. First, they do not allow making a reasonable choice among ORFs in the case when multiple candidates are presented. Second, they do not provide automated computational tools for high-throughput analysis pipelines and large datasets. Third, they show low accuracy in predicting ORFs and require significant computing power. Finally, they lack integration with software for the analysis of structural and functional characteristics of RNA molecules. Here we present a computational approach and its implementation, an ORFhunteR – R/Bioconductor package, aimed at automatic determination of true ORFs in mRNA molecules. The proposed method is based on the vectorization of nucleotide sequences followed by a random-forest classification. Our package also provides automatic annotation of the identified ORFs. In addition, a user-friendly version of ORFhunteR, based on a pre-trained model, is implemented as a web application (https://orfhunter.bsu.by/, Fig. 1). The approach was validated on two large public datasets (NCBI RefSeq, Ensembl) and a pre-trained model is provided together with the package.

2. Methods

The proposed computational approach for the automatic identification of the true ORFs integrates algorithms for vectorization [6,7] and random forest-based classification [8]. Our pipeline includes the following five steps: (i) building of a set of reference ORFs, (ii) vectorization of reference ORFs into sequence features, (iii) training of the classification model, (iv) identification of the true ORFs in a set of mRNA molecules and (v, optionally) annotation of the identified ORFs (Fig. 2a).

2.1. Feature extraction: vectorization of sequences

Nucleotide sequences are vectorized into 104 features. The first **84 features** represent frequencies of mono-, di-, and trinucleotides. The frequencies are calculated by the standard algorithms of R/Bioconductor package *Biostrings*. Next, **6 features** based on nucleotide correlation factors are included [7]. The length of a candidate ORF and its logarithm are both used, to preserve sensitivity to variability in short and long ORFs (**2 features**). Finally, **12 features** of the category-position-frequency (CPF) model [6] represent the local frequency-based entropy values of sequences (see Methods in [9]). Vectorization is implemented in C++ and R programming languages using R/Bioconductor and CRAN packages. This provides a significant performance improvement compared to pure R (C++ with *Rpp* package increased the performance of the analysis by almost 100-fold).

2.2. Sequence classification and ORF identification

Classification of the sequences was done by a classical random forest from the R-package *randomForest* (500 trees). In parallel, it assesses the significance of the features by the Gini index. Discovery data were separated into 75% training and 25% validation sub-sets, with the latter, used to estimate the accuracy of the classifier.

2.3. ORFhunteR pipeline usage

The analysis pipeline includes several major steps, each with specific functions (Fig. 2b). It starts with the loading of sequences from fasta-, gtf- or gff-file using *loadTrExper*. In the uploaded transcripts, ORF candidates are identified using the functions *codonStartStop* and *findORFs*. These ORF candidates are vectorized into sequence features by *vectorizeORFs* in conjunction with the C-based functions *getBaoMetrics* and *getCorrelationFactors*. The vectorized ORF candidates are classified into true ORFs and pseudo-ORFs by function *predictORFs*. If necessary, the nucleotide sequence of the true ORFs can be obtained using the function *getSeqORFs*. Finally, identified true ORFs can be annotated by the function *annotateORFs*, together with the functions *findPTCs* and *translateORFs* (provides transcript ID, length of 5'UTRs, type of start codon, start coordinate of ORF, stop coordinate of ORF, type of stop codon, PTC status of stop codon, length of ORF, length of 3'UTRs, molecular weight of *in silico* translated protein, isoelectric point of a protein sequence and potential protein interaction index). Herewith, the function *findPTCs* identifies premature termination codons in transcripts of interest while function *translateORFs* translates ORFs to proteins.

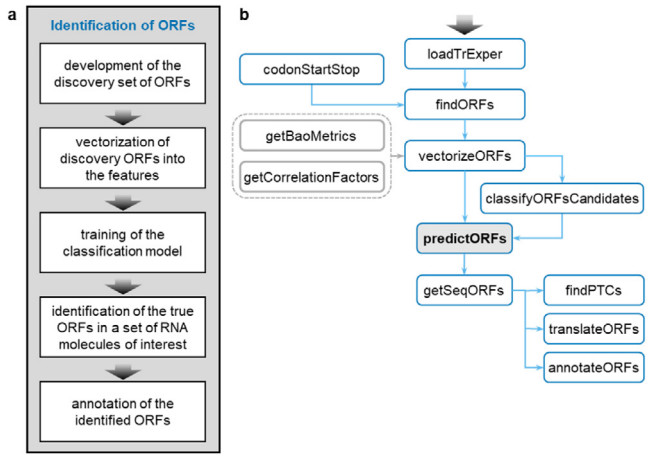

Fig. 2. An overview of ORFhunteR pipeline (a) and dependencies of the main functions (b).

2.4. Data and method validation setup

The approach was applied to two large datasets. A **discovery dataset** included 128161 well-annotated mRNA molecules of protein-coding genes and 4235 long non-coding RNA (lncRNA) molecules from the manually curated NCBI RefSeq database (release 109, GRCh38.p12 reference assembly of the human genome). Coordinates and extracted sequences of highly confident true ORFs in mRNA molecules were collected, resulting in 113085 records in total. Additionally, we calculated coordinates and extracted 108800 sequences of pseudo-ORFs from lncRNA molecules. Similar to real ORFs, pseudo-ORFs begin with ATG start codon and end in-frame with one of the stop codons, but are not translated into proteins. These two sets of ORFs were combined into a single well-balanced reference or training set of true ORFs and pseudo-ORFs (imbalance index of 1.04) and used in random forest classification to construct the trained model to be applied further for the ORF identification in discovery and test datasets. As a **test dataset**, mRNA and lncRNA sequences from Ensembl (release 97, GRCh38.p12 reference assembly of the human genome) were used. To avoid artifacts, we excluded: (i) mitochondrial transcripts, (ii) 5' incomplete transcripts, containing canonical stop codon but lacking a start codon inside the sequence, (iii) 3' incomplete transcripts containing canonical start codon ATG but lacking a stop codon inside the sequence, (iv) both 5' and 3' incomplete transcripts lacking start and stop codons inside the sequence, (v) and transcripts with non-canonical start codons CTG, GTG or TTG. We combined filtered Ensembl mRNAs (56765 records in total) and lncRNAs (74980 records in total) into a single test set of RNA molecules.

3. Results of validation

We started by applying the method to the discovery dataset (NCBI RefSeq), where the accuracy in detecting true ORFs on the validation sub-set reached 98.3%. We identified some CPFs (h_{SS}, h_{MK}, h_{KK}, see [6]) and the length as the most important features for the determination of ORFs.

In the prevention of overfitting, we tested the trained model on independent Ensembl RNA data. On this dataset, the approach allowed identifying the true ORFs with an accuracy of 94.9%. In fact, 91.9% of ORFs that were identified in Ensembl human mRNA molecules demonstrated a probability of 0.9 or higher to be coding. At the same time, probability values (to be coding ORF) for ORFs from various lncRNAs strongly differed (Fig. 3). All mRNAs are properly classified, while only a minor part of lncRNAs could be misclassified due to evolutionary-caused structural similarities with coding mRNA.

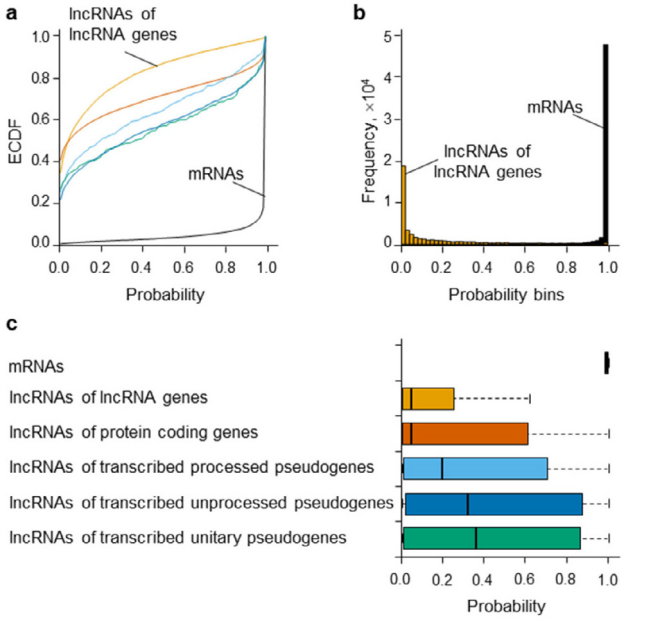

Fig. 3. Distribution of probability values for identification of pseudo-ORFs and true ORFs in different human RNA molecules. (a) Empirical cumulative distribution of frequency (ECDF) values for ORFs that were identified in protein-coding mRNAs and various non-coding RNAs. (b) Frequency of probability values for ORFs that were identified in mRNAs and lncRNAs encoded by lncRNA genes. (c) Boxplot demonstrating the distribution of probability values for ORFs that were identified in mRNAs and lncRNAs different gene biotypes.

4. Potential and existing impact

Our package automatically detects open reading frames in large collections of human mRNA molecules with high accuracy. It provides a systematic, automated, and high-throughput approach to sequencing data analysis. Moreover, it automatically annotates identified open reading frames. These properties distinguish our tool from alternative solutions. Our software is applicable for solving fundamental tasks linked to genomic aberrations in cancers as was recently reported in high-impact journals [10–12]. It also can be applied to more practical tasks: the differential diagnostics of human diseases and the development of predictive models for disease progression and clinical outcomes. An additional asset of the tool is linked to its potential applicability for studies related to personalized medicine.

5. Limitations and future development

The ORFhunteR software package has several limitations: it depends on third-party R-libraries, has a slow initialization, and has non-optimal processing of large data files. An automatic update of model files should also be implemented. These disadvantages will be eliminated by building specialized C/C++ libraries, optimizing the codes for working with big data [13], and enabling an automatic update of prediction models. In the future, we are planning to implement classification models to identify open reading frames starting from alternative start codons CTG, GTG, and TTG. In addition, it is planned to develop predictive models based on mixed data from NCBI RefSeq and Ensembl/GENCODE databases. Moreover, the list of annotations will be significantly expanded.

6. Conclusions

The efficient computational approach for the identification of unknown ORFs in mRNA molecules was developed and integrated into the corresponding R/Bioconductor package ORFhunteR. It is based on vectorization of the sequence features of ORFs candidates and

predicting the most evident by a random forest classifier. Our numerical tests resulted in the accuracy of ORFs identification of 98.3% and 94.9% on the verification and validation datasets.

Finally, it is important to mention that the developed approach has three advantages over the competing modern strategies [4,14]: (i) it requires less computing resources and works much faster than neural nets; (ii) it is less prone to overfitting and uses a limited set of vectorized features (unlike the statistical approaches utilizing thousands of features); (iii) random forest classifiers show much better interpretability compared to deep learning or boosting models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

VVG, MMY, VVS, and MKC were supported by the Ministry of Education of the Republic of Belarus, grant GPSR "Convergenciya–2020" N3.08.3 (registration number 20190531). PVN and MKC were supported by the Luxembourg National Research Fund (C17/BM/11664971/DEMICS).

References

[1] E.R. Mardis, DNA sequencing technologies: 2006–2016, Nat. Protoc. 12 (2) (2017) 213–218.

[2] J.A. Reuter, D.V. Spacek, M.P. Snyder, High-throughput sequencing technologies, Mol. Cell 58 (4) (2015) 586–597.

[3] E.W. Sayers, et al., Database resources of the national center for Biotechnology information, Nucleic Acids Res. 47 (D1) (2019) D23–D28.

[4] A. Al-Ajlan, A. El Allali, CNN-MGP: Convolutional neural networks for metagenomics gene prediction, Interdiscip. Sci. 11 (4) (2019) 628–635.

[5] K.J. Hoff, et al., Gene prediction in metagenomic fragments: a large scale machine learning approach, BMC Bioinformatics 9 (2008) 217.

[6] J. Bao, R. Yuan, Z. Bao, An improved alignment-free model for DNA sequence similarity metric, BMC Bioinformatics 15 (2014) 321.

[7] R. Mao, et al., Comparative analyses between retained introns and constitutively spliced introns in Arabidopsis thaliana using random forest and support vector machine, PLoS One 9 (8) (2014) e104049.

[8] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[9] V.V. Grinev, et al., ORFhunteR: an accurate approach for the automatic identification and annotation of open reading frames in human mRNA molecules, 2021, bioRxiv.

[10] A. Radzisheuskaya, et al., Complex-dependent histone acetyltransferase activity of KAT8 determines its role in transcription and cellular homeostasis, Mol. Cell 81 (8) (2021) 1749–1765, e8.

[11] V.V. Grinev, et al., RUNX1/RUNX1T1 mediates alternative splicing and reorganises the transcriptional landscape in leukemia, Nature Commun. 12 (1) (2021) 520.

[12] R. Tirtakusuma, et al., Epigenetic regulator genes direct the fate of multipotent progenitor cell of origin in lineage switched MLL/AF4 leukaemia, 2021, bioRxiv.

[13] M.M. Yatskou, V.V. Apanasovich, Computational platform FluorSimStudio for processing kinetic curves of fluorescence decay using simulation modeling and data mining algorithms, J. Appl. Spectrosc. 88 (3) (2021) 571–579.

[14] J. Wen, et al., A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network, BMC Bioinformatics 20 (1) (2019) 469.