Published online 18 May 2020 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* ***W239–W243***
doi: 10.1093/nar/gkaa385

mRNALoc: a novel machine-learning based *in-silico* tool to predict mRNA subcellular localization

Anjali Garg, Neelja Singhal, Ravindra Kumar and Manish Kumar *

Department of Biophysics, University of Delhi South Campus, New Delhi 110021, India

Received March 03, 2020; Revised April 14, 2020; Editorial Decision April 27, 2020; Accepted April 30, 2020

ABSTRACT

Recent evidences suggest that the localization of mRNAs near the subcellular compartment of the translated proteins is a more robust cellular tool, which optimizes protein expression, post-transcriptionally. Retention of mRNA in the nucleus can regulate the amount of protein translated from each mRNA, thus allowing a tight temporal regulation of translation or buffering of protein levels from bursty transcription. Besides, mRNA localization performs a variety of additional roles like long-distance signaling, facilitating assembly of protein complexes and coordination of developmental processes. Here, we describe a novel machine-learning based tool, mRNALoc, to predict five sub-cellular locations of eukaryotic mRNAs using cDNA/mRNA sequences. During five fold cross-validations, the maximum overall accuracy was 65.19, 75.36, 67.10, 99.70 and 73.59% for the extracellular region, endoplasmic reticulum, cytoplasm, mitochondria, and nucleus, respectively. Assessment on independent datasets revealed the prediction accuracies of 58.10, 69.23, 64.55, 96.88 and 69.35% for extracellular region, endoplasmic reticulum, cytoplasm, mitochondria, and nucleus, respectively. The corresponding values of AUC were 0.76, 0.75, 0.70, 0.98 and 0.74 for the extracellular region, endoplasmic reticulum, cytoplasm, mitochondria, and nucleus, respectively. The mRNALoc standalone software and web-server are freely available for academic use under GNU GPL at http://proteininformatics.org/mkumar/mrnaloc.

INTRODUCTION

Localization of mRNA is an evolutionarily conserved phenomenon that controls many important biological processes like cell-fate determination and polar cell growth (1). After post-transcriptional modifications, such as 5′ capping, splicing and addition of 3′ poly (A) tail, the nascently transcribed mRNA either gets localized within the nucleus or alternatively travels out of the nucleus. It has been suggested that mRNA localization has many advantages over protein localization (2–6). These are: (a) localization of mRNA to a specific location helps the cell to build a local repository of proteins at the site of function instead of transporting individual protein molecules to the site of function. This also compartmentalizes protein synthesis and forms a protein gradient within the cells, which ultimately results in local synthesis of encoded proteins at the target site; (b) mRNA localization works as a translation/co-translational regulator; (c) mRNA localization is a better energy-efficient pathway compared to protein targeting and; (d) mRNA localization aids in formation of only functional and non-harmful multi-protein complexes which aids in avoiding unnecessary protein-protein interactions that might be harmful to the cells (7,8). Not all protein synthesis occurs after mRNA localization. A large number of mRNA sequences are also transported co-translationally (9).

Five different mechanisms namely, diffusion and localized entrapment, localized degradation, localized synthesis, active transport and, polarized nuclear export are considered important for mRNA localization. However, ribonucleoprotein transport complex is the main mode by which majority of RNA is transported. Building the ribonucleoprotein complex is a sequence specific phenomenon, which is guided by a short stretch of 20–200 *cis*-acting nucleotide sequences known as 'zipcode'. It is located at the 3′ untranslated region of the mRNA sequence, although in some cases they can also be present in the 5′UTR or in the coding sequence (10,11). Proteins present in a subcellular compartment are related to the physiological and metabolic function associated with that subcellular compartment. Hence, prediction of subcellular location of mRNA might suggests the biological function of the gene from which the mRNA was transcribed. Thus, a tool that can predict the correct intracellular location of transcripts may also help in understanding how gene expression is regulated and, how cells achieve polarity.

To our knowledge, computational predictors that can predict the subcellular localization of eukaryotic mRNA are unavailable, till date. Hence, we developed a Support Vector Machine (SVM) based *in-silico* tool which can predict the eukaryotic mRNA subcellular locations on the

*To whom correspondence should be addressed. Tel: +91 11 24157263; Email: manish@south.du.ac.in

basis of primary sequence information of mRNA/cDNA. Named as mRNALoc (acronym for 'mRNA Localization), this tool is based on the experimentally validated localization data of mRNA retrieved from 'RNALocate' (12).

ANALYSIS WORKFLOW

Data sources

In the present work, we collected the mRNA sequences and their subcellular location information from RNALocate database (version 2.0) (12). RNALocate is a manually curated database that provides complete subcellular location annotation of RNA with experimental support. Initially, a total of 28829 mRNA sequences with annotated subcellular localization were obtained. The downloaded mRNA sequences revealed their localization to both single and multiple subcellular locations. In the present study, we considered only those mRNA sequences which showed single locations. The mRNA dataset was classified in five subgroups on the basis of subcellular locations namely, cytoplasm, endoplasmic reticulum, extracellular, mitochondria and nucleus. The number of mRNA sequences in the five locations were as follows: 6964 in cytoplasm, 1998 in endoplasmic reticulum, 1131 in extracellular region, 442 in mitochondria and 6346 in nucleus.

Since, redundant mRNA sequences results in overestimation of prediction capability, hence to reduce the redundancy and to avoid homology bias in prediction, we used NCBI BLASTCLUST program to retain only sequences showing alignment identity ≤40% over 70% or more of their full length (BLASTCLUST with '-S 40 and -L 0.7' option) (13). The final non-redundant mRNA dataset contained 6376 sequences of cytoplasm, 1426 sequences of endoplasmic reticulum, 855 sequences of extracellular region, 421 sequences of mitochondria and 5831 mRNA sequences of nucleus. 5/6 part of total 40% non-redundant data was used for training the model. Remaining 1/6 data was used for the independent evaluation of the trained model. For detail about collection of dataset, redundancy removal, constructions of training and independent datasets please see supplementary material. The NCBI gene accession numbers, mRNA sequences and subcellular locations are available in the download section of mRNALoc webserver (http://proteininformatics.org/mkumar/mrnaloc/download.html).

Overview of mRNALoc

mRNALoc is a web resource to predict the subcellular localization of eukaryotic mRNA. The overall workflow of mRNALoc is shown in Figure 1. Users have to provide the mRNA sequences in a FASTA format. The submitted mRNA sequence will be converted into numerical encoding using pseudo oligonucleotide composition or pseudo K-tuple nucleotide composition (PseKNC) (14–17). On the basis of SVM prediction score, the mRNA will be predicted to localize at one of the five subcellular locations, namely cytoplasm, endoplasmic reticulum, extracellular location, mitochondria and nucleus. mRNALoc prediction is based on the five trained SVM models, each specific for one location. During prediction each model provides the prediction score for its corresponding location. The subcellular location, whose SVM model gets the maximum score, will be the predicted location. The final outcome of mRNALoc depends on the user-selected threshold. Higher thresholds would result in more specific predictions, while lower threshold would result in low specificity predictions.

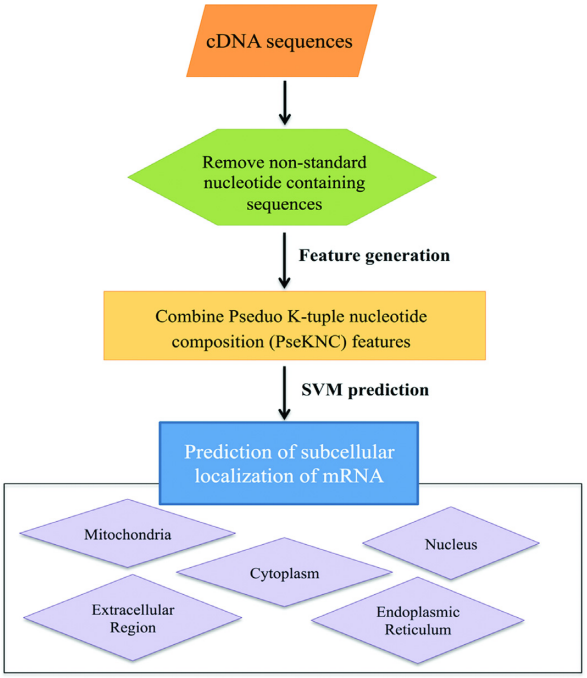

Figure 1. Overall schema of mRNALoc. mRNALoc predicts five subcellular locations *viz.*, mitochondria, cytoplasm, nucleus, endoplasmic reticulum and extracellular. Firstly, it removes the sequences from the query that has non-standard nucleotides then generates combined features from pseudo K-tuple nucleotide composition, which is further used as input for Support Vector Machine (SVM) prediction.

TRAINING OF PREDICTIVE SVM MODELS

The performance mRNALoc for all subcellular locations during five-fold cross-validation mode of training is shown in Table 1. Using the combined input of PseKNC ($K = 2$, 3, 4 and 5) we found 65.19, 75.36, 67.10, 99.70 and 73.59% accuracy of prediction for mRNA whose subcellular locations were extracellular region, endoplasmic reticulum, cytoplasm, mitochondria and nucleus, respectively. When evaluated on an independent dataset, mRNALoc did prediction with sensitivity, specificity, accuracy and MCC values of 81.38, 56.67, 58.10 and 0.18 for extracellular region, 75.10, 68.60, 69.23 and 0.27 for endoplasmic reticulum, 73.26, 58.06, 64.55 and 0.31 for cytoplasm, 87.32, 97.16, 96.88 and 0.63 for mitochondria and 50.20, 81.62, 69.35 and 0.34 for nucleus, respectively (Supplementary Figures S1 and S2, Supplementary Table S1).

COMPARISON WITH EXISTING MRNA SUBCELLULAR LOCALIZATION PREDICTION METHODS

Though, the role of mRNA localization is unambiguously established in cellular physiology, attempts to build *in-silico*

Table 1. The performance metrics for mRNA subcellular localization under hybrid *K*-mer feature (2+3+4+5), and performance of the SVM based classifiers (mRNALoc) on independent data

Location	Sen (%)	Spe (%)	ACC (%)	MCC	THR	AUC
Training dataset						
Extracellular region	62.67	65.34	65.19	0.14	−0.20	0.69
Endoplasmic reticulum	74.09	75.49	75.36	0.32	0.40	0.81
Cytoplasm	66.69	67.41	67.10	0.34	0.40	0.69
Mitochondria	96.28	99.79	99.70	0.95	0.10	0.98
Nucleus	74.17	73.22	73.59	0.47	0.40	0.76
Independent dataset						
Extracellular region	81.38	56.67	58.10	0.18	−0.20	0.76
Endoplasmic reticulum	75.10	68.60	69.23	0.27	0.40	0.75
Cytoplasm	73.26	58.06	64.55	0.31	0.40	0.70
Mitochondria	87.32	97.16	96.88	0.63	0.10	0.98
Nucleus	50.20	81.62	69.35	0.34	0.40	0.74

Sen: sensitivity, Spe: specificity, ACC: accuracy, MCC: Mathews correlation coefficient, THR: threshold, and AUC: area under ROC curve.

Table 2. Comparative evaluation of mRNALoc and iLoc-mRNA. In extracellular region and mitochondria no human mRNA was present, hence these two locations were not included in the evaluation

		mRNALoc		iLoc-mRNA	
Location	Number of human mRNA sequences	True positive	False negative	True positive	False negative
Cytoplasm	50	35	15	18	32
Endoplasmic reticulum	50	34	16	37	13
Extracellular region	0	0	0	0	0
Mitochondria	0	0	0	0	0
Nucleus	50	33	17	13	37

tools to predict the subcellular localizations of mRNA are negligible in comparison to protein subcellular localization prediction tools. Recently, Yan *et al.* proposed a deep-learning based method, named as RNATracker (18), to predict the subcellular localization of mRNA using data from CeFra-Seq (19) and APEX-RIP (3). Using the data from RNALocate, a human mRNA subcellular localization method iLoc-mRNA was also developed (20).

Though, both RNATracker and iLoc-mRNA are based on two different mRNA subcellular localization datasets and, were developed using two different approaches, mRNALoc has several advantages over both RNATracker and iLoc-mRNA. For example, (a) localization data produced by CeFra-Seq/APEX-RIP are inherently noisy and sometimes inaccurate also (18). The mRNALoc was developed from datasets retrieved from RNALocate (12), which contains manually curated mRNA subcellular localization information with experimental evidences. (b) The RNATracker among all the isoforms, considered only the longest isoform while, mRNALoc did not made any such distinction. (c) Redundant mRNA sequences were not removed from RNATracker and in iLoc-mRNA the redundancy threshold was 80%. While in mRNALoc, we used 40% non-redundant mRNA sequences to train the predictor. This may be the reason underlying high MCC and AUC for RNATracker and iLoc-mRNA. (d) Both, RNATracker and iLoc-mRNA were developed using only localization data of human mRNA. On the contrary, mRNALoc is a general-purpose eukaryotic mRNA subcellular localization prediction tool, which is applicable to all eukaryotes. (e) RNATracker also excluded low expressed genes, but mRNALoc made no such distinction (Supplementary Table S2).

We also conducted one-to-one comparison of performance of iLoc-mRNA and mRNALoc. As RNATracker required gene expression and coordination files for prediction, it was not possible to include it in the evaluation. For comparison we used the independent dataset of mRNALoc. Since, iLoc-mRNA is specifically designed for human mRNA subcellular localization prediction, we used 50 human mRNA sequences of independent dataset of mRNALoc. The number of human mRNA in different locations and prediction result of mRNALoc and iLoc-mRNA is shown in Table 2. In extracellular region and mitochondria, we didn't find human mRNA sequences in mRNALoc independent dataset hence, these locations were not included in the evaluation.

As shown in Table 2, for cytoplasm and nucleus the performance of mRNALoc was better than iLoc-mRNA but, in endoplasmic reticulum the performance of iLoc-mRNA was better than mRNALoc. It is also pertinent to mention that in iLoc-mRNA prediction were made for one of the following locations namely, cytosol/cytoplasm, ribosome, endoplasmic reticulum, and nucleus/exosome/dendrite/mitochondrion. We feel that combining nucleus, exosome, dendrite, and mitochondria as a single location is not appropriate as these are diverse subcellular locations which should not be merged in a single category.

DESCRIPTION OF THE WEBSERVER

Implementation of mRNALoc

The web server is hosted on a Linux system. The back-end pipeline is implemented in the Perl language. The webserver has an intuitive interface and 'how-to' guide to help the user.

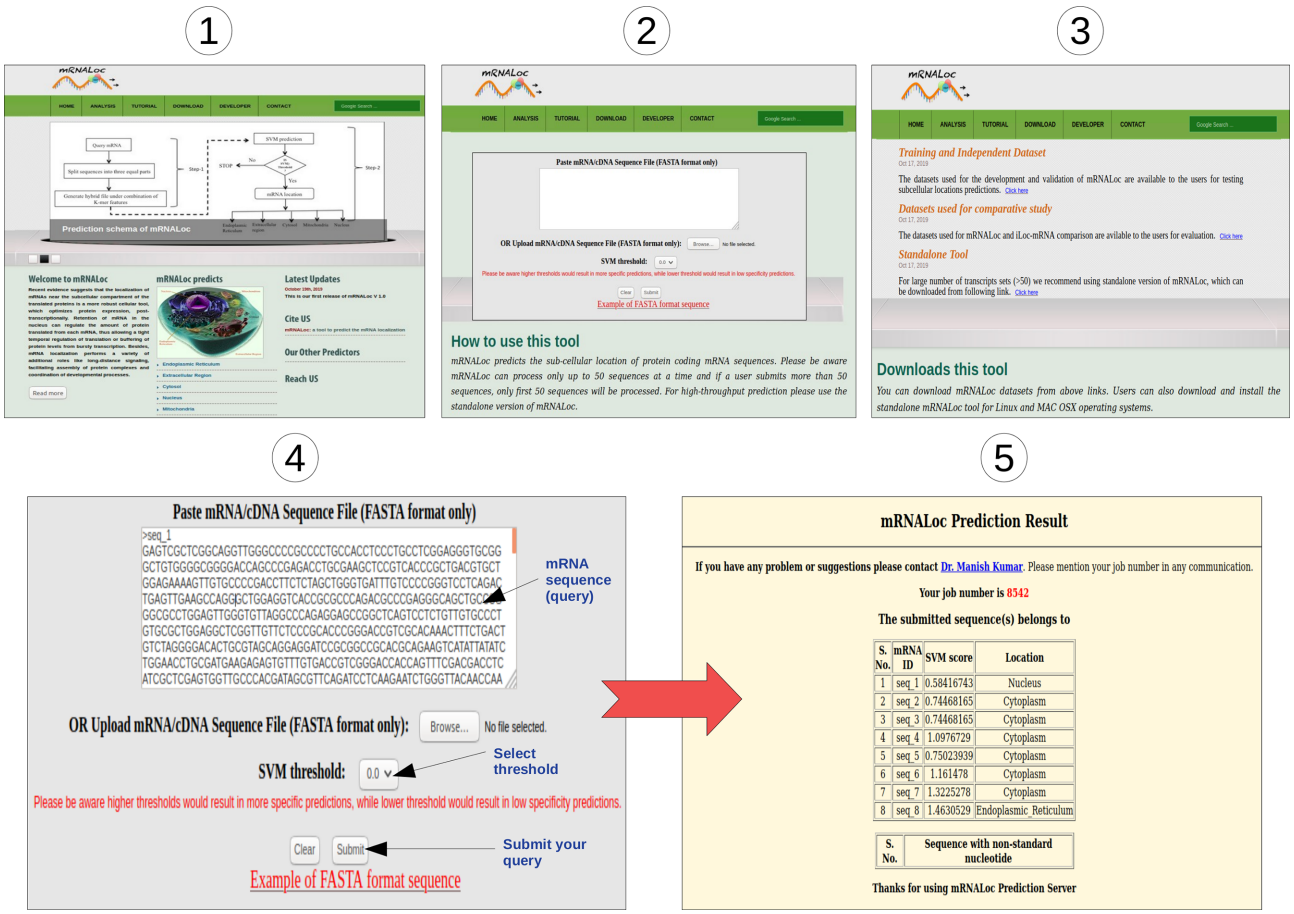

Figure 2. Screenshots of mRNALoc webserver.

Each mRNA query sequence must be at least 100 bp long and contains only valid characters, namely 'A', 'C', 'G' and 'T/U'. Sequences having non-standard nucleotides will be omitted from the prediction pipeline (Figure 2).

The output of mRNALoc

The output of mRNALoc is presented in a tabular format. It contains the highest scores obtained from the five SVM models and the location to which the mRNA is assigned. A maximum of fifty sequences can be processed by mRNALoc webserver in one go. Hence, for genome scale prediction a standalone version will be required (Figure 2 and Supplementary Figure S3).

CONCLUSIONS AND FUTURE PROSPECTS

The annotation of subcellular localization has been addressed mainly at the protein level. Many *in silico* tools were developed to predict protein subcellular location using machine-learning techniques. It has been unequivocally established that both mRNA and protein localization play an equal role in protein translocation. In future versions of mRNALoc we would like to overcome some of the limitations of the present tool. The first and foremost is that our tool is currently limited by the accuracy of the RNALocate datasets. Though, RNALocate contain data from 65 organisms, most of the data is enriched with the common model organisms like, *Homo sapiens*, *Mus musculus*, and *Saccharomyces cerevisiae* etc. Moreover, considering at the biological level, instead of cytosol, mitochondria or extracellular locations, axons, dendrites, dendritic spines, or anterior/posterior vs dorsal/ventral locations are more relevant. Another, limitation is that due to lesser availability of plant mRNA localization data compared to other domains of life, mRNALoc performance might be compromised (21). The performance of a machine-learning method depends on the data on which it is trained. We believe that with development of new and better RNA localization finding techniques, information about RNA localization in plants would also be available in the near future and future versions of mRNALoc would then support prediction of plant mRNA sequences, also. We admit that mRNALoc is in an early stage of development and training on additional datasets is needed to further improve our tool. Further prediction of mRNA localization will also help in predicting the novel zipcodes which may guide researchers to cast new hypothesis for unraveling the finer details of mechanism of mRNA-protein complex formation which is actually responsible for mRNA location. Though, the current version of mRNALoc supports prediction of only eukaryotic mRNA, the future versions of mRNALoc would definitely include data from other organisms and locations.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Indian Council of Medical Research (ICMR)-JRF scheme [3/1/3 J.R.F.-2016/LS/HRD-(32262) to A.G.]; CSIR Senior Research Associateship (Scientists' Pool Scheme) [9089A/2019-Pool to N.S.]. Funding for open access charge: Indian Council of Medical Research.
Conflict of interest statement. None declared.

REFERENCE

1. Kloc,M., Zearfoss,N.R. and Etkin,L.D. (2002) Mechanisms of subcellular mRNA localization. *Cell*, **108**, 533–544.
2. Lecuyer,E., Yoshida,H., Parthasarathy,N., Alm,C., Babak,T., Cerovina,T., Hughes,T.R., Tomancak,P. and Krause,H.M. (2007) Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, **131**, 174–187.
3. Kaewsapsak,P., Shechner,D.M., Mallard,W., Rinn,J.L. and Ting,A.Y. (2017) Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *eLife*, **6**, e29224.
4. Medioni,C., Mowry,K. and Besse,F. (2012) Principles and roles of mRNA localization in animal development. *Development*, **139**, 3263–3276.
5. Wang,E.T., Taliaferro,J.M., Lee,J.A., Sudhakaran,I.P., Rossoll,W., Gross,C., Moss,K.R. and Bassell,G.J. (2016) Dysregulation of mRNA localization and translation in genetic disease. *J. Neurosci.*, **36**, 11418–11426.
6. Hughes,S.C. and Simmonds,A.J. (2019) Drosophila mRNA localization during later Development: Past, Present, and Future. *Front. Genet.*, **10**, 135.
7. Di Liegro,C.M., Schiera,G. and Di Liegro,I. (2014) Regulation of mRNA transport, localization and translation in the nervous system of mammals (Review). *Int. J. Mol. Med.*, **33**, 747–762.
8. Vandepoele,K., Simillion,C. and Van de Peer,Y. (2002) Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. *Trends Genet.*, **18**, 606–608.
9. Weis,B.L., Schleiff,E. and Zerges,W. (2013) Protein targeting to subcellular organelles via mRNA localization. *Biochim. Biophys. Acta*, **1833**, 260–273.
10. Heasman,J., Wessely,O., Langland,R., Craig,E.J. and Kessler,D.S. (2001) Vegetal localization of maternal mRNAs is disrupted by VegT depletion. *Dev. Biol.*, **240**, 377–386.
11. Kloc,M. and Etkin,L.D. (1994) Delocalization of Vg1 mRNA from the vegetal cortex in Xenopus oocytes after destruction of Xlsirt RNA. *Science (New York, N.Y.)*, **265**, 1101–1103.
12. Zhang,T., Tan,P., Wang,L., Jin,N., Li,Y., Zhang,L., Yang,H., Hu,Z., Zhang,L., Hu,C. *et al.* (2017) RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.*, **45**, D135–D138.
13. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
14. Liu,B., Wang,S., Long,R. and Chou,K.C. (2017) iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics*, **33**, 35–41.
15. Liu,B., Liu,F., Fang,L., Wang,X. and Chou,K.C. (2015) repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, **31**, 1307–1309.
16. Chen,W., Feng,P., Ding,H., Lin,H. and Chou,K.C. (2015) iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*, **490**, 26–33.
17. Liu,B., Fang,L., Long,R., Lan,X. and Chou,K.C. (2016) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, **32**, 362–369.
18. Yan,Z., Lecuyer,E. and Blanchette,M. (2019) Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics*, **35**, i333–i342.
19. Benoit Bouvrette,L.P., Cody,N.A.L., Bergalet,J., Lefebvre,F.A., Diot,C., Wang,X., Blanchette,M. and Lecuyer,E. (2018) CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells. *RNA*, **24**, 98–113.
20. Zhang,Z.Y., Yang,Y.H., Ding,H., Wang,D., Chen,W. and Lin,H. (2020) Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief. Bioinformatics*, doi:10.1093/bib/bbz177.
21. Tian,L., Chou,H.L., Fukuda,M., Kumamaru,T. and Okita,T.W. (2020) mRNA localization in plant cells. *Plant Physiol.*, **182**, 97–109.