Supplementary Materials

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

Anjali Garg[1], Neelja Singhal[1], Ravindra Kumar[1] and Manish Kumar[1*]
[1]Department of Biophysics, University of Delhi South Campus, New Delhi, India
***Correspondence to: Manish Kumar, Department of Biophysics, University of Delhi South Campus, New Delhi, India – 110021. E-mail:** manish@south.du.ac.in

Materials and methods

Collection of subcellular location annotated dataset of mRNA

Constructing a high quality benchmark dataset is the foremost requirement to develop a reliable prediction model. For the present work we collected the mRNA sequences and their subcellular location information from RNALocate database (1). Initially a total of 28829 mRNA sequences with annotated subcellular localization were obtained. This included multi-locations mRNA also. On the basis of subcellular locations, the mRNA were classified into five subgroups namely, cytoplasmic, endoplasmic reticulum, extracellular, mitochondrial and nuclear. 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.

Redundancy removal

The sequence redundancy among mRNA might results in overestimation of prediction capability. Hence to get rid of the redundancy and to avoid homology bias in prediction, we used NCBI BLASTCLUST program to keep only sequences having alignment identity ≤ 40% over 70% or more of their full length (BLASTCLUST with "-S 40 and -L 0.7" option) (2). 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 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. The NCBI gene accession numbers of mRNA and their sequences and locations are available at mRNALoc webserver (http://proteininformatics.org/mkumar/mrnaloc/download.html).

Construction of training and independent dataset and training methodology

We divided redundancy reduced mRNA sequences of each subcellular location into two parts. The subset that contained 5/6th of total data was used as training dataset for SVM while the remaining 1/6th part was used as independent dataset to assess the performance of trained model. For training we adopted five-fold cross-validation approach during which all mRNA sequences of training dataset were randomly divided into five sets of which four sets were used for training and the remaining one set for testing. This procedure was repeated five times in such a way that each set was used once for testing.

Conversion of a nucleotide sequence into machine learning input feature

Since SVM can be trained with only fixed length input features, we converted each mRNA sequence to a fixed length numerical encodings. The most simple feature encoding is nucleotide compositions of an mRNA. However the simple nucleotide composition does not contain the sequence-order effect. To incorporate local sequence-order information K-tuple nucleotide composition (PseKNC) (3) was proposed, which has been used in development of many biological feature predictors (4-7). The value and dimension of PseKNC depends on the value of K that ranges from 2, 3, 4 and 5 for di-, tri-, tetra- and penta-nucleotide respectively. It was observed that DNA/RNA dinucleotide physical structures, including twist, tilt, roll, shift, slide and rise, significantly contribute to dealing with DNA/RNA sequences (8,9). Therefore, in this study six dinucleotide physical structures are employed to encode the pseudo 2-tuple nucleotide compositions. Additionally, 12 physicochemical properties of tri-, tetra- and pentanucleotide were also included to encode the pseudo 3, 4, 5-tuple nucleotide compositions.

Hybrid Feature Vectors (HFV)

Several studies have shown that a SVM model that is trained on a combination of more than one input features has better discrimination capability (10-13). Hence we combined and used all input features (namely PseDNC, PseTNC, PseTetraNC, and PsePentaNC) as a single hybrid input feature vector. The combined hybrid feature vector had the size 1360-D (16+64+256+1024) for an mRNA sequence.

Fragmented Sequence Encoding

In our earlier work we have observed that when a protein sequence was divided into multiple fragments and a combined input vector was made from the sequence encoding of individual fragments, the performance of predictor increased significantly (11,14,15). Hence in the present work we have also used the fragmented sequence approach to train SVM. We split each mRNA sequence into three equal parts and named them as N-terminal, M-part and C-terminal, each part individually used for feature identification. For example an mRNA sequence of length L with N number of nucleotides is divided into three chunks where first L/3rd part is used as N-terminal and second L/3rd part as M-part and last L/3rd as C-terminal. For each fragment we calculated the values of PseKNC and used the combination of PseKNC of three segments as input to the SVM. We used both single and hybrid approach of encoding as an input to SVM.

Performance Evaluation Strategies

In this study five-fold cross-validation approach was used to test the performance of SVM models after training. For training mRNA of each location was randomly divided into five approximately equal-sized non-overlapping subsets. At each parameter the SVM was trained on four subsets and tested on the remaining one subset. This procedure was repeated five times so that every time a different subset was used as the test set and each subset were used as a test set only once. The performance at each parameter was measured for each test, and the average of five test subsets was reported as the final performance.

The classification performance for the subcellular location of mRNA was evaluated using sensitivity (S_n), specificity (S_p), Matthew's correlation coefficient (MCC) and accuracy (ACC). All of these parameters were derived from the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). These parameters are formulated as:

$$S_n = TP \div (TP + FN) * 100 \qquad (1)$$

$$S_p = TN \div (TN + FN) * 100 \qquad (2)$$

$$ACC = (TP + TN) \div (TP + FP + TN + FN) * 100 \qquad (3)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (4)$$

In eq. 1..4 for a particular location X, TP was the number of mRNA sequence correctly predicted to be present in X location, TN represent mRNA which were correctly predicted to be not present in X location, FP is the number of sequences belong to location Y but wrongly predicted as location X and FN is the mRNA sequences which actually belong to location X but predicted as Y location.

To describe the performance of models across the entire range of SVM decision values, receiver operating characteristics (ROC) curve (16) and area under ROC curve (AUC) were used. ROC curves showed the true positive rate as a function of the false positive rate. The area under the ROC curve (the AUC score) is a way to transform the information provided by ROC curve to a single scalar value representing expected performance. Random prediction by a SVM model will have ROC curve at the diagonal line with AUC score of 0.5, while a perfect predictor will produce a ROC curve along the left and top boundary of the square and will have AUC score one.

Results and Discussion

Performance on Complete Sequence Information

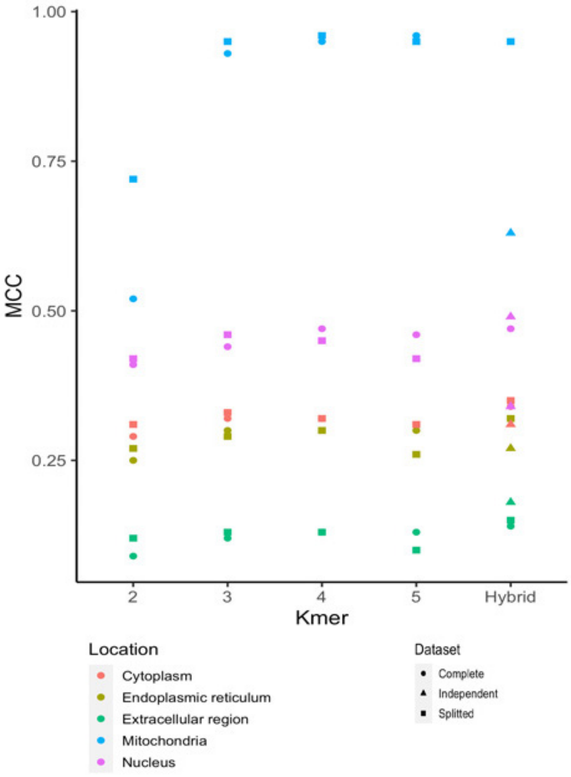
In all five locations firstly we used PseDNC (K=2) as input. Subsequently the value of K increased to five. We noticed a significant improvement in performance when value of K increased from 2 to 4 and after K=4 no increase in performance was observed.

Performance on Fragmented Sequence Information

Similar to the complete sequence information we observed a considerable improvement in performance when value of K increased from 2 to 4 and after K=4 no increase in performance was observed. However we didn't notice any improvement in performance when fragmented PseKNC was used as input instead of complete sequence information.

All Features Helped, but the Combination Performed Best and Most Robust

When all five forms of PseKNC were merged together as a single input, we noticed a significant improvement in performance in case of both complete and fragmented sequence input. In most locations an increase of 2% was observed with hybrid inputs both in case of complete and splitted PseKNC (**Supplementary Figure 1**).



Supplementary Figure 1. Performance achieved during five-fold cross-validation with different values of K-mer.

Receiver Operating Characteristics (ROC) Plot and Area Under ROC Curve (AUC) analysis

Use of overall accuracy to estimate the performance of a predictor developed on an imbalanced dataset might provide an unrealistic assessment of a classifier's performance because even a random one-sided prediction of majority class members may create an impression of a highly accurate predictor, which is infact based on an one-sided random predictor. In the present work, to avoid the influence of majority data types on prediction efficiency calculations, we gave equal weightage to both sensitivity and specificity. Another way of objective estimation of classifier accuracy is receiver operating characteristic (ROC) plot, which is a very popular way to measure overall performance of a classifier. ROC plot demonstrates the trade-off between sensitivity and specificity at different thresholds and is generated by plotting 'True positive rate' *vs.* 'False positive rate'. Further, the area under the ROC curve (AUC) is also frequently used to estimate the performance of a classifier. As shown in Supplementary figures S2 and Supplementary table 1, the ROC plot and their corresponding AUC values also supports the conclusion that SVM model developed using hybrid composition as input has performed better in comparison to the ones which were developed using a single input. Also since performance of both complete sequence and splitted composition based SVM modules are same, in all further studies we used hybrid full sequence composition based SVM modules.

Supplementary Table 1. Estimation of the performance metrics for mRNA location identification under different combination of K-mer features. Values indicated in italics and bold font show the highest prediction score of training, and only in bold font show the performance on independent dataset.

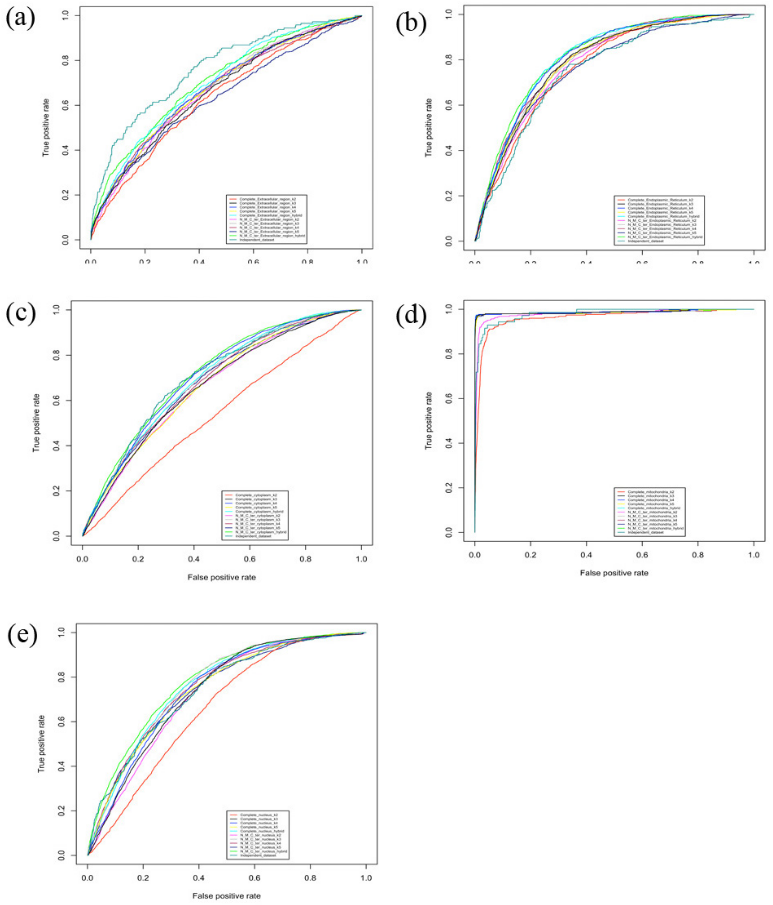
	Sen (%)	Spe (%)	ACC (%)	MCC	THR	AUC	K-mer	# of features	Location
Complete sequence	58.17	61.08	60.91	0.09	-0.2	0.63	2	16	**Extracellular region**
	60.00	63.93	63.70	0.12	-0.4	0.66	3	64	
	61.27	65.27	65.04	0.13	-0.4	0.67	4	256	
	62.96	63.36	63.34	0.13	-0.6	0.67	5	1024	
	62.67	***65.34***	***65.19***	***0.14***	***-0.2***	***0.69***	***2+3+4+5***	1360	
	70.30	70.41	70.40	0.25	0.4	0.77	2	16	**Endoplasmic reticulum**
	72.49	73.53	73.43	0.30	0.1	0.79	3	64	
	74.26	74.86	74.80	0.32	0.2	0.81	4	256	
	72.40	73.68	73.56	0.30	0	0.79	5	1024	
	74.09	***75.49***	***75.36***	***0.32***	***0.4***	***0.81***	***2+3+4+5***	1360	
	62.26	66.53	64.71	0.29	0.5	0.54	2	16	**Cytoplasm**
	66.05	66.35	66.22	0.32	0.4	0.66	3	64	
	65.37	66.80	66.19	0.32	-0.4	0.70	4	256	
	65.56	65.38	65.46	0.31	0.3	0.66	5	1024	
	66.69	***67.41***	***67.10***	***0.34***	***0.4***	***0.69***	***2+3+4+5***	1360	
	91.43	93.68	93.62	0.52	-0.3	0.96	2	16	**Mitochondria**
	96.00	99.69	99.59	0.93	-0.1	0.98	3	64	
	97.14	99.77	99.70	0.95	-0.3	0.98	4	256	
	95.14	99.89	99.75	0.96	-0.2	0.98	5	1024	
	96.28	***99.79***	***99.70***	***0.95***	***0.1***	***0.98***	***2+3+4+5***	1360	
	69.56	71.83	70.94	0.41	-0.3	0.66	2	16	**Nucleus**
	72.62	72.49	72.55	0.44	0.4	0.73	3	64	
	75.08	72.27	73.37	0.47	0.3	0.74	4	256	
	74.07	72.35	73.02	0.46	-0.3	0.74	5	1024	
	74.17	***73.22***	***73.59***	***0.47***	***0.4***	***0.76***	***2+3+4+5***	1360	
Splitted sequence	60.42	63.90	63.70	0.12	-0.1	0.66	2	48	**Extracellular region**
	62.39	64.31	64.21	0.13	-0.3	0.67	3	192	
	61.69	64.09	63.95	0.13	-0.5	0.67	4	768	
	59.01	61.86	61.70	0.10	-1	0.63	5	3072	
	64.37	***65.42***	***65.36***	***0.15***	***-0.3***	***0.70***	***2+3+4+***	***4080***	

							5		
	71.22	71.19	71.19	0.27	-0.1	0.78	2	48	**Endoplasmic reticulum**
	72.15	72.93	72.85	0.29	0.2	0.79	3	192	
	73.58	73.59	73.59	0.30	-0.1	0.80	4	768	
	69.11	71.87	71.60	0.26	-0.3	0.76	5	3072	
	73.16	***75.42***	***75.21***	***0.32***	***0.3***	***0.82***	***2+3+4+5***	***4080***	
	67.82	63.90	65.58	0.31	0.4	0.66	2	48	**Cytoplasm**
	66.93	66.30	66.57	0.33	0.4	0.69	3	192	
	66.67	65.65	66.08	0.32	-0.4	0.68	4	768	
	66.86	63.96	65.20	0.31	0.2	0.68	5	3072	
	67.02	***68.38***	***67.80***	***0.35***	***0.4***	***0.71***	***2+3+4+5***	***4080***	
	93.14	97.82	97.69	0.72	-0.5	0.98	2	48	**Mitochondria**
	96.00	99.81	99.70	0.95	-0.1	0.98	3	192	
	96.00	99.88	99.78	0.96	-0.1	0.98	4	768	
	95.71	99.86	99.74	0.95	-0.3	0.98	5	3072	
	95.43	***99.82***	***99.70***	***0.95***	***0.1***	***0.98***	***2+3+4+5***	***4080***	
	72.48	70.81	71.47	0.42	-0.3	0.72	2	48	**Nucleus**
	74.17	72.31	73.04	0.46	-0.3	0.76	3	192	
	71.70	73.88	73.03	0.45	0.3	0.75	4	768	
	71.45	70.79	71.05	0.42	-0.3	0.74	5	3072	
	75.26	***74.19***	***74.61***	***0.49***	***0.3***	***0.78***	***2+3+4+5***	***4080***	
Independent dataset	**81.38**	**56.67**	**58.1**	**0.18**	**-0.2**	**0.76**	**2+3+4+5**	**1360**	**Extracellular region**
	75.10	**68.6**	**69.23**	**0.27**	**0.4**	**0.75**	**2+3+4+5**	**1360**	**Endoplasmic reticulum**
	73.26	**58.06**	**64.55**	**0.31**	**0.4**	**0.70**	**2+3+4+5**	**1360**	**Cytoplasm**
	87.32	**97.16**	**96.88**	**0.63**	**0.1**	**0.98**	**2+3+4+5**	**1360**	**Mitochondria**
	50.20	**81.62**	**69.35**	**0.34**	**0.4**	**0.74**	**2+3+4+5**	**1360**	**Nucleus**

Sen: Sensitivity, Spe: Specificity, ACC: Accuracy, MCC: Mathew's correlation coefficient, THR: Threshold, and AUC: Area under ROC curve

Benchmarks on independent mRNA datasets

We also evaluated the performance of mRNALoc hybrid full sequence composition based SVM models on an independent dataset. The data was 1/6th part of the original data collected initially and it was not was used to train the SVM. On same prediction parameters, it gave prediction result with sensitivity, specificity, accuracy and MCC values 75.10, 68.60, 69.23 and 0.27 for endoplasmic reticulum, 81.38, 56.67, 58.1 and 0.18 for extracellular region, 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.

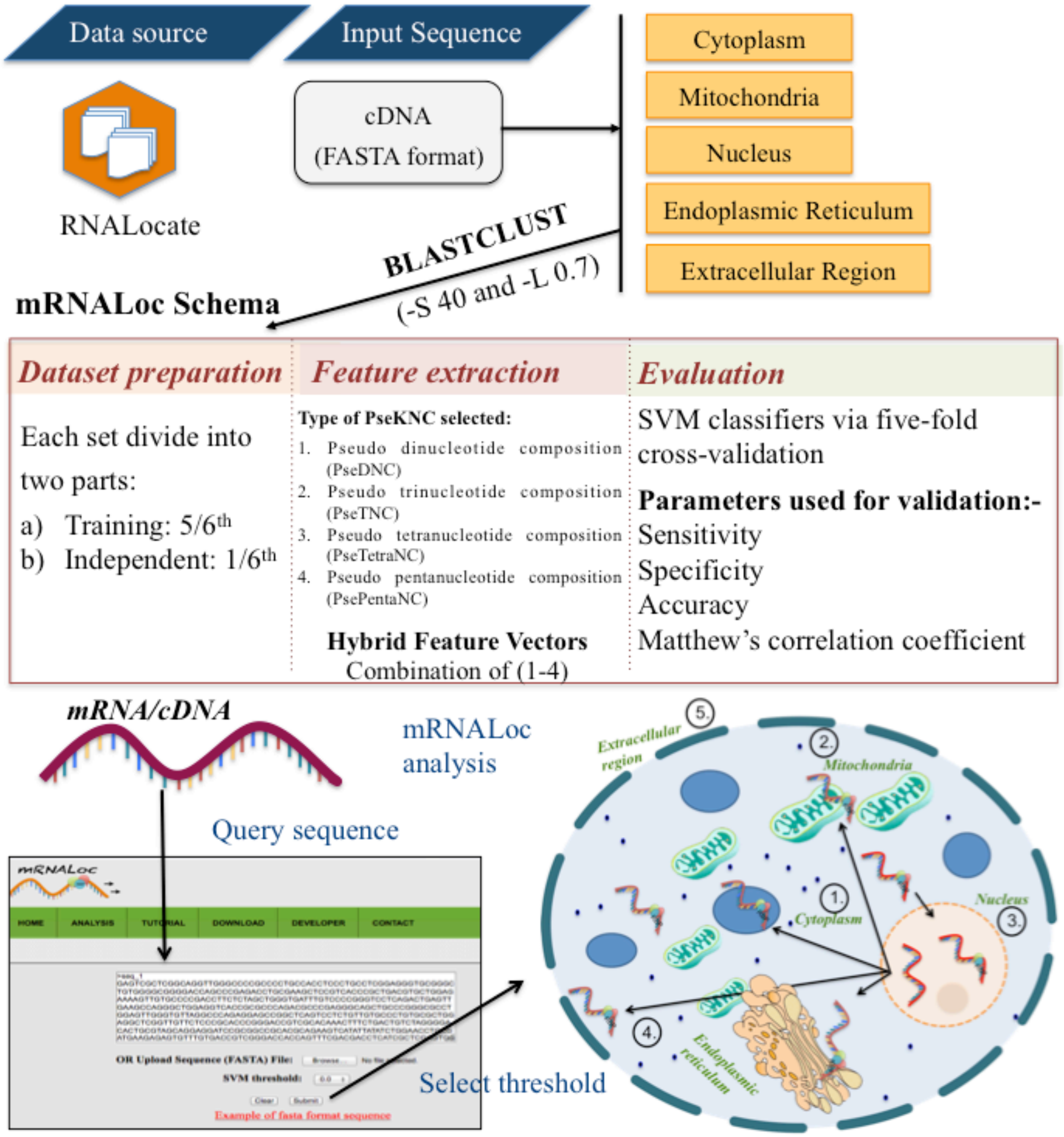


Supplementary Figure 2. The ROC curves for (a) extracellular region (b) endoplasmic reticulum (c) cytoplasm (d) mitochondria (e) nucleus. A ROC cure plots the true positive rate (i.e. sensitivity) against the false positive rate. For each location performance on complete and splitted sequence is shown at different Kmer values. The performance on independent data is also shown.

Webserver and user guide

For the convenience of usage by the experimental scientists, we developed a publicly accessible webserver **mRNALoc** using the trained SVM models. A step-by-step instruction on how to use this tool is given as follows.

Step 1.: Open the HOME page of mRNALoc at http://proteininformatics.org/mku/mrnaloc/home.html. At the left side of this page a brief introduction about the mRNA localization and mRNALoc can be found.



Supplementary Figure 3. The complete workflow of mRNALoc development including webserver schema.

Step 2.: On ANALYSIS page either type or copy/paste or upload the mRNA sequences. The input sequences must be in the FASTA format. The mRNALoc webserver predict only 50 sequences at a time. In case more than 50 sequences will be submitted, only first 50 sequences will be processed. User can change the SVM threshold value from drop down menu (default value is 0). The nature of prediction depends on the SVM threshold. For example selection of higher threshold will result in high specificity while lower threshold would result in high specificity. High specificity means a high number of false negatives and low number of true positives. High sensitivity means a high number of false positives and low number of true negatives.

Note: From DOWNLOAD page users can download benchmark datasets that were used in training and testing of mRNALoc predictor. Further standalone version of tool can also be downloaded from this page. In standalone version there is no limit of number of number of sequences.

Cautions. Each input mRNA query sequences must be 100 bp or longer and only contains valid characters: ‘A’, ‘C’, ‘G’ and ‘T/U’. If a sequence has non-standard nucleotides, the sequence will be removed from prediction pipeline.

Comparison with existing mRNA subcellular localization prediction methods

Supplementary Table 2. Brief comparison of the advantages and limitations among mRNALoc, RNATracker and iLoc-mRNA.

Feature	**mRNALoc**	**RNATracker**	**iLoc-mRNA**
Benchmark data source	RNALocate	CeFra-Seq/APEX-RIP	RNALocate
Data redundancy threshold	40%	80%	80%
Types of mRNA sequences incorporated in the study	All types of genes and isoforms	Only highly expressed and longest isoforms	Not mentioned
Tool used to develop prediction Model	SVM	Deep neural network	SVM
Can work on	All Eukaryotes	Human only	Human only
Input data	mRNA/cDNA	Sequencing read and transcript coordination file	mRNA/cDNA
Webserver	Yes	No	Yes
Standalone availability	Yes	Yes	Yes

Reference

1. 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 research*, **45**, D135-D138.
2. McGinnis, S. and Madden, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, **32**, W20-25.
3. Chen, W., Lei, T.Y., Jin, D.C., Lin, H. and Chou, K.C. (2014) PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal Biochem*, **456**, 53-60.
4. 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.
5. 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.
6. Chen, W., Feng, P., Ding, H., Lin, H. and Chou, K.C. (2015) iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Analytical biochemistry*, **490**, 26-33.
7. 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.
8. Liu, B., Yang, F., Huang, D.S. and Chou, K.C. (2018) iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics*, **34**, 33-40.
9. Liu, B., Yang, F. and Chou, K.C. (2017) 2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function. *Mol Ther Nucleic Acids*, **7**, 267-277.
10. Tahir, M., Hayat, M. and Khan, S.A. (2018) A Two-Layer Computational Model for Discrimination of Enhancer and Their Types Using Hybrid Features Pace of Pseudo K-Tuple Nucleotide Composition. *Arab J Sci Eng* **43**, 6719–6727.
11. Kumar, M., Verma, R. and Raghava, G.P. (2006) Prediction of mitochondrial proteins using support vector machine and hidden Markov model. *The Journal of biological chemistry*, **281**, 5357-5363.
12. Mishra, N.K., Kumar, M. and Raghava, G.P. (2007) Support vector machine based prediction of glutathione S-transferase proteins. *Protein Pept Lett*, **14**, 575-580.
13. Kumar, M. and Raghava, G.P. (2009) Prediction of nuclear proteins using SVM and HMM models. *BMC Bioinformatics*, **10**, 22.
14. Kumar, R., Kumari, B. and Kumar, M. (2018) Proteome-wide prediction and annotation of mitochondrial and sub-mitochondrial proteins by incorporating domain information. *Mitochondrion*, **42**, 11-22.
15. Kumar, R., Jain, S., Kumari, B. and Kumar, M. (2014) Protein sub-nuclear localization prediction using SVM and Pfam domain information. *PLoS One*, **9**, e98345.
16. Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940-3941.