Multi-label learning for identication of RNA-associated subcellular localizations

Hao Wang
Tianjin University
Yijie Ding
Suzhou University of Science and Technology
Jijun Tang
Tianjin University
Quan Zou
University of Electronic Science and Technology
Fei Guo (fguo@tju.edu.cn)
Tianjin University https://orcid.org/0000-0001-8346-0798

Research article

Keywords: RNA subcellular localization, Multi-label classi cation, Hilbert-Schmidt independence criterion, Multiple kernel learning, Web server

Posted Date: August 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-55447/v1

Version of Record: A version of this preprint was published on January 15th, 2021. See the published version at https://doi.org/10.1186/s12864-020-07347-7.

RESEARCH

Multi-label learning for identification of RNA-associated subcellular localizations

Hao Wang[1], Yijie Ding[2], Jijun Tang[4], Quan Zou[3] and Fei Guo[1*]

*Correspondence: fguo@tju.edu.cn
[1]School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University,China
Full list of author information is available at the end of the article

Abstract

Biological functions of biomolecules rely on the cellular compartments where they are located in cells. Importantly, RNAs are assigned in specific locations of a cell, enabling the cell to implement diverse biochemical processes in the way of concurrency. However, lots of existing RNA subcellular localization classifiers only solve the problem of single-label classification. In fact, a single primary RNA transcript is used to make multiple proteins. Therefore, it is of great practical significance to expand RNA subcellular localization into multi-label classification problem. In this study, we extract multi-label classification datasets about RNA-associated subcellular localizations on various types of RNAs, and then construct subcellular localization datasets on four RNA categories. In order to study Homo sapiens, we further establish human RNA subcellular localization datasets. Furthermore, we utilize different nucleotide property composition models to extract effective features to adequately represent the important information of nucleotide sequences. In the most critical part, we achieve a major challenge that is to fuse the multivariate information through multiple kernel learning based on Hilbert-Schmidt independence criterion. The optimal combined kernel can be put into an integration support vector machine model for identifying multi-label RNA subcellular localizations. To be specific, our novel method performs outstanding rather than other prediction tools on our novel benchmark datasets. Moreover, we establish user-friendly web server with the implementation of our method, which can be easily used by most experimental scientists.

Keywords: RNA subcellular localization; Multi-label classification; Hilbert-Schmidt independence criterion; Multiple kernel learning; Web server

Introduction

Biological functions of biomolecules rely on various cellular compartments. One cell can be divided into different compartments that are related to different biological processes. Thus, the cellular role of one RNA molecular could be inferred from its localization information. What's more, there has been a great deal of research on the protein subcellular localization [1–6]. Currently, the biological technology capable of whole-genome that subcellular localization has been indicated to be a fundamental regulation mode in biological cells [7].

Currently, the biological technology capable of whole-genome localization is the subcellular RNA sequencing, called SubcRNAseq, which yields high-throughput and quantitative data. Large amounts of raw subcRNAseq data have recently become available, most notably from the ENCODE consortium. A lot of research work has established the resource to make RNA localization data available to the broader scientific community. Firstly, Zhang et al. [8] built a database called RNALocate,

which collected more than 42,000 manually engineered RNA subcellular localization entries. Subsequently, Mas-Ponte et al. [9] constructed a database named LncATLAS to store the subcellular localization of lncRNA.

Considering expensive and inconvenient biological experiments [10], automatic computational tools are the highly relevant measure to speed up RNA-related studies. The computational identification of subcellular localization has been a hot topic for the last decade. In the early days, Cheng et al. [11] systematically studied the distribution of lncRNA localization in gastric cancer and revealed its relationship with gastric cancer. As a pioneer work, Feng et al. [12] developed a computational method to predict the organelle positions of non-coding RNA (ncRNAs) by collecting ncRNAs from centroids, mitochondria, and chloroplast genomes. Subsequently, Zhen et al. [13] developed lncLocator to predict the subcellular localization of long-stranded non-coding RNA. Xiao et al. [14] proposed a novel method used the sequence-to-sequence model to predict microRNA subcellular localization. Besides, Yang et al. [15] developed MiRGOFS being a GO-based functional similarity measurement for miRNA subcellular localization. Then, iLoc-mRNA [16] used binomial distribution and one-way analysis of variance to obtain the optimal nonamer composition of mRNA sequences, and applies a predictor to identify human mRNA subcellular localization.

However, most existing RNA subcellular localization classifiers only solve the problem of single-label classification. In fact, a single primary RNA transcript is used to make multiple proteins [17–19]. Therefore, it is of great practical significance to expand RNA subcellular localization into multi-label classification problem. In view of the above research, there is no multi-label RNA subcellular localization dataset available for this task. According to RNALocate database, we extract multi-label classification datasets about RNA-associated subcellular localizations on various types of RNAs, and then construct subcellular localization datasets on four RNA categories (mRNAs, lncRNAs, miRNAs and snoRNAs).

In this study, we utilize different nucleotide property composition models to adequately represent important information of nucleotide sequences. In the most critical part, we achieve a major challenge is to fuse the multivariate information through multiple kernel learning based on Hilbert-Schmidt independence criterion. The optimal combined kernel can be put into an integration support vector machine model for training a multi-label RNA subcellular localization classifier.

Material and method

In this study, we establish RNA subcellular localization datasets, and then propose an integration learning model for multi-label classification. The flowchart of our method is show in Figure S1.

Benchmark dataset

RNAs are generally divided into two categories. One is encoding RNAs, such as messenger RNAs (mRNAs), which play a very important role in transcription. Other is non-coding RNAs, including long non-coding RNA (lncRNA), microRNA (miRNA), small nucleolar RNA (snoRNA), which play an irreplaceable regulatory role in life. In order to study subcellular localization for Homo sapiens, we further

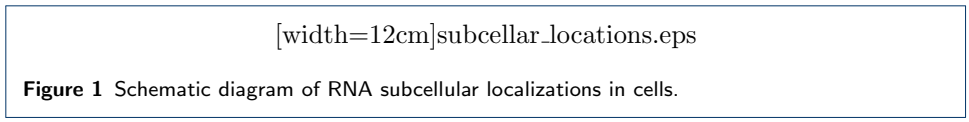

Figure 1 Schematic diagram of RNA subcellular localizations in cells.

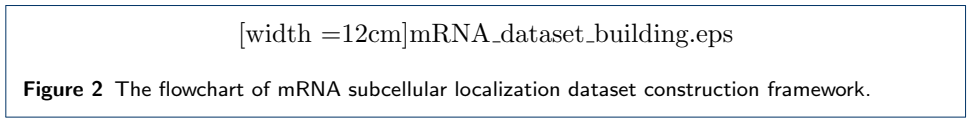

Figure 2 The flowchart of mRNA subcellular localization dataset construction framework.

establish human RNA subcellular localization datasets. Subcellular localizations of various RNAs in cells are shown in Figure 1.

We use the database of RNA subcellular localization in order to integrate, analyze and identify RNA subcellular localization for speeding up RNA structural and functional researches. The first release of RNALocate (http://www.rna-society.org/rnalocate/) contains more than 42,000 manually engineered RNA-associated subcellular localization and experimental evidence entries in more than 23100 RNA sequences, 65 organisms (e.g., homo sapiens, mus musculus, saccharomyces cerevisiae), localization of 42 subcells (e.g., cytoplasm, nucleus, endoplasmic reticulum, ribosomes), and 9 RNA categories (e.g., mRNA, microRNA, lncRNA, snoRNA). Thus, RNALocate provides a comprehensive source of subcellular localization and even insight into the function of hypothetical or new RNAs. We extract multi-label classification datasets about RNA-associated subcellular localizations on four RNA categories (mRNAs, lncRNAs, miRNAs and snoRNAs). The flowchart of mRNA subcellular localization dataset construction framework is shown in Figure 2.

RNA subcellular localization datasets

We extract four RNA subcellular localization datasets, including mRNAs, lncRNAs, miRNA and snoRNAs. The procedure for constructing RNA datasets is listed as follows.

- We download total RNA entries with curated subcellular localizations from RNAlocate, and use CD-HIT [20] to remove redundant samples with a cutoff of 80%.
- We delete samples with duplicate Gene ID and remove samples without corresponding subcellular localization labels, and then construct four RNA subcellular localization datasets.
- We count the number of samples for each category of subcellular localization labels, and then select some categories with the sample size greater than a reasonable threshold ($N/N_{max} > 1/30$).

The statistical distributions of these four RNA datasets are shown in Figure 3. Details are shown in Supplementary Table S1-S2.

Human RNA subcellular localization datasets

We also extract four Homo sapiens RNA subcellular localization datasets, including H_mRNAs, H_lncRNAs, H_miRNA and H_snoRNAs. The procedure for constructing human RNA datasets is listed as follows.

- We screen out samples of homo sapiens on above four RNA datasets, and construct four human RNA subcellular localization datasets.

Figure 3 The statistical distributions of four RNA subcellular localization datasets.

Figure 4 The statistical distributions of four human RNA subcellular localization datasets.

- We count the number of samples for each category, and then select some categories with the sample size greater than a reasonable threshold ($N/N_{max} > 1/12$).

The statistical distributions of these four human RNA datasets are shown in Figure 4. Details are shown in Supplementary Table S3-S4.

Nucleotide property composition representation

RNA sequence can be represented as follow: $S = (s_1, \cdots, s_l, \cdots, s_L)$, where s_l denotes the l-th ribonucleic acid and L denotes the length of S. How to formulate varied length RNA sequences as fixed length features, is the key point to effective operational problem-solving. Many studies have shown that the RNA sequence can be encoded by nucleotide property composition representation [21], which can profoundly affect the way of body behaves. Here, we encode the RNA sequence in order to better mine and explore information patterns.

k-mer nucleotide composition

For k-mer descriptor, RNAs are represented as occurrence frequencies of k neighboring nucleic acids, which has been successfully applied to human gene regulatory sequence prediction and enhancer identification. The k-mer (e.g. $k = 2$) descriptor can be calculated as follows.

$$f(t) = \frac{N(t)}{N - k + 1}, \quad t \in \{AA, AC, AG, TT\} \tag{1}$$

where $N(t)$ is the number of k-mer type t, while N is the length of a nucleotide sequence.

For $k = 1, 2, 3, 4$, there are four combinations together, each of which has 4^k distinct types of nucleotide characteristics. Therefore, we extract 340-dimensional feature vector $F_{kmer1234}$.

Only remaining 4-mer, there are 4^4 types of nucleotide characteristics. Therefore, we extract 256-dimensional feature vector F_{kmer4}.

Reverse compliment k-mer

The reverse compliment k-mer (RCKmer) is a variant of k-mer descriptor, which is not expected to be strand-specific. For instance, there are 16 types of 2-mer ('AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'CT', 'GA', 'GC', 'GG', 'GT', 'TA', 'TC', 'TG', 'TT'), 'TT' is reverse compliment with 'AA'. After removing the reverse compliment k-mer, there are only 10 distinct types of k-mer in the reverse compliment k-mer approach ('AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG','GA', 'GC', 'TA').

For 4-mer with 256 types, after removing reverse compliment 4-mer, there are 136 distinct types in the reverse compliment k-mer approach. Therefore, we extract 136-dimensional feature vector F_{RCKmer}.

Nucleic acid composition

The nucleic acid composition (NAC) encodes the frequency of each nucleic acid type in a nucleotide sequence, which is similar to 1-mer. The frequency of each natural nucleic acid ('A', 'C', 'G', 'T' or 'U') can be calculated as follows.

$$f(t) = \frac{N(t)}{N}, \qquad t \in \{A, C, G, T(U)\} \tag{2}$$

where $N(t)$ is the number of nucleic acid type t, while N is the length of a nucleotide sequence.

Therefore, we extract 4-dimensional feature vector F_{NAC}.

Di-nucleotide composition

The di-nucleotide composition (DNC) encodes the frequency of each 2-tuple of nucleic acid type in a nucleotide sequence, which is similar to 2-mer. The frequency of each 2-tuple of natural nucleic acid can be calculated as follows.

$$D(i, j) = \frac{N_{ij}}{N - 1}, \qquad i, j \in \{A, C, G, T(U)\} \tag{3}$$

where N_{ij} is the number of di-nucleotide type represented by nucleic acid types i and j.

Therefore, we extract 16-dimensional feature vector F_{DNC}.

Tri-nucleotide composition

The tri-nucleotide composition (TNC) encodes the frequency of each 3-tuple of nucleic acid type in a nucleotide sequence, which is similar to 3-mer. The frequency of each 3-tuple of natural nucleic acid can be calculated as follows.

$$D(i, j, k) = \frac{N_{ijk}}{N - 2}, \qquad i, j, k \in \{A, C, G, T(U)\} \tag{4}$$

where N_{ijk} is the number of di-nucleotide type represented by nucleic acid types i, j and k.

Therefore, we extract 64-dimensional feature vector F_{TNC}.

Composition of k-spaced nucleic acid pair

The composition of k-spaced nucleic acid pair (CKSNAP) is used to calculate the frequency of nucleic acid pairs separated by any k nucleic acids ($k = 0, 1, 2, \dots$). For each k-space, there are 16 types of nucleic acid pair composition ('A...A', 'A...C', 'A...G', 'A...T', 'C...A', 'C...C', 'C...G', 'C...T', 'G...A', 'G...C', 'G...G', 'G...T', 'T...A', 'T...C', 'T...G', 'T...T').

For $k = 0, 1, 2, 3, 4, 5$, there are six different combinations, each of which has 16 distinct types of nucleic acid pair composition. Therefore, we extract 96-dimensional feature vector F_{CKSNAP}.

Multiple kernel support vector machine classifier

We apply radial basis function (RBF) on above feature sets to construct corresponding kernels, respectively. The RBF kernel is defined as follows.

$$K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2),\ i, j = 1, 2, ..., N \tag{5}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors of samples i and j, N denotes the number of samples, and γ is the bandwidth of Gaussian kernel.

The kernel set with seven distinct kernels is denoted as follows.

$$\mathbf{K} = \{\mathbf{K}_{\text{kmer4}}, \mathbf{K}_{\text{kmer1234}}, \mathbf{K}_{\text{RCKmer}}, \mathbf{K}_{\text{NAC}}, \mathbf{K}_{\text{DNC}}, \mathbf{K}_{\text{TNC}}, \mathbf{K}_{\text{CKSNAP}}\} \tag{6}$$

Hilbert-schmidt independence criterion multiple kernel learning

We use multiple kernel learning (MKL) to figure out weights of above kernels, and then integrate them together. The optimal combinatorial kernel can be calculated as follows.

$$\mathbf{K}^* = \sum_{p=1}^{7} \beta_p \mathbf{K}^p, \quad \mathbf{K}^p \in \mathbf{R}^{N \times N} \tag{7}$$

The main purpose of hilbert-schmidt independence criterion (HSIC) [22] is to measure a difference in the distribution of two variables, which is similar to the covariance and is itself constructed according to the covariance. Let $\mathbf{X} \in \mathbf{R}^{N \times d}$ and $\mathbf{Y} \in \mathbf{R}^{N \times 1}$ be two variables from a data set of $\mathbf{Z} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, which is jointly from some probability distribution $Pr_{\mathbf{x}y}$. HSIC measures the independence between $\mathbf{x}$ and y by calculating the norm of cross-covariance operator over domain $\mathbf{X} \times \mathbf{Y}$.

Hilbert-Schmidt operator norm of $C_{\mathbf{x}y}$ is defined as follows.

$$HSIC(\mathbf{F}, \mathbf{G}, Pr_{\mathbf{x}y}) = \|C_{\mathbf{x}y}\|_{HS}^2 \tag{8}$$

Given set $\mathbf{Z}$, empirical estimate of HSIC is computed as follows.

$$\begin{aligned} HSIC(\mathbf{F}, \mathbf{G}, \mathbf{Z}) &= \frac{1}{N^2} tr(\mathbf{KU}) - \frac{2}{N^3} \mathbf{e}^T \mathbf{KUe} + \frac{1}{N^4} \mathbf{e}^T \mathbf{Kee}^T \mathbf{Ue} \\ &= \frac{1}{N^2} \left[tr(\mathbf{KU}) - \frac{1}{N} tr(\mathbf{KUee}^T) - \frac{1}{N} tr(\mathbf{UKee}^T) \right. \\ &\left. + \frac{1}{N^2} tr(\mathbf{Uee}^T \mathbf{Kee}^T) \right] \\ &= \frac{1}{N^2} tr[\mathbf{K}(\mathbf{I} - \frac{1}{N} \mathbf{ee}^T)\mathbf{U}(\mathbf{I} - \frac{1}{N} \mathbf{ee}^T)] \\ &= \frac{1}{N^2} tr(\mathbf{KHUH}) \triangleq HSIC(\mathbf{K}, \mathbf{U}) \end{aligned} \tag{9}$$

where $\mathbf{F}$ is the RKHS of feature set $\mathbf{X}$, $\mathbf{G}$ is the RKHS of label set $\mathbf{Y}$, $\mathbf{e} = (1, ..., 1)^T \in \mathbf{R}^{N \times 1}$, $\mathbf{H} = \mathbf{I} - \mathbf{ee}^T / N \in \mathbf{R}^{N \times N}$ (centering matrix), $\mathbf{K}, \mathbf{U} \in \mathbf{R}^{N \times N}$ are kernel matrices with $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{U}_{ij} = l(y_i, y_j)$, $\mathbf{I} \in \mathbf{R}^{N \times N}$ is the identity matrix. The stronger the dependence between $\mathbf{K}$ and $\mathbf{U}$, the larger the value. $\mathbf{K}$ and $\mathbf{U}$ are independent between each other, when $HSIC(\mathbf{K}, \mathbf{U}) = 0$.

Enligthened by HSIC [23], we define optimization function as follows.

$$\max_{\boldsymbol{\beta},\mathbf{K}^*} HSIC(\mathbf{K}^*, \mathbf{U}) \tag{10a}$$

$$HSIC(\mathbf{K}^*, \mathbf{U}) = \frac{1}{N^2} tr(\mathbf{K}^*\mathbf{H}\mathbf{U}\mathbf{H}) \tag{10b}$$

$$subject\ to\ \mathbf{K}^* = \sum_{p=1}^{P} \beta_p \mathbf{K}^p, \tag{10c}$$

$$\beta_p \geq 0,\ p = 1, 2, ..., P, \tag{10d}$$

$$\sum_{p=1}^{P} \beta_p = 1 \tag{10e}$$

where $\mathbf{K}^* \in \mathbf{R}^{N\times N}$ is the optimal kernel of feature space, and $\mathbf{U} = \mathbf{y}_{train}\mathbf{y}_{train}^T \in \mathbf{R}^{N\times N}$ is ideal kernel matrix (label kernel), $\boldsymbol{\beta} \in \mathbf{R}^{P\times 1}$ is the kernel weight vector. We aim to maximize HSIC between $\mathbf{K}^*$ and $\mathbf{U}$.

Convex quadratic programming problem can be solved as follows.

$$\min_{\boldsymbol{\beta},\mathbf{K}^*} -\frac{1}{N^2} tr(\mathbf{K}^*\mathbf{H}\mathbf{U}\mathbf{H}) + \nu_1 \|\boldsymbol{\beta}\|^2 \tag{11a}$$

$$subject\ to\ \mathbf{K}^* = \sum_{p=1}^{P} \beta_p \mathbf{K}^p, \tag{11b}$$

$$\beta_p \geq 0,\ p = 1, 2, ..., P, \tag{11c}$$

$$\sum_{p=1}^{P} \beta_p = 1 \tag{11d}$$

where ν_1 is L_2 norm regularization term. The final training and testing kernels are linearly weighted by $\boldsymbol{\beta}$, respectively.

Support vector machine

Support vector Machine [24] was first proposed by Cortes and Vapnik [25]. It deals primarily with dichotomies. Given a dataset of instance-label pairs $\{\mathbf{x}_i, y_i\}$, $i = 1, 2, ..., N$, the classification decision function realized by SVM is expressed as follows.

$$f(\mathbf{x}) = sign[\sum_{i=1}^{N} y_i \alpha_i \cdot K(\mathbf{x}, \mathbf{x}_i) + b] \tag{12}$$

where $\mathbf{x}_i \in \mathbf{R}^{1\times d}$ and $y_i \in \{+1, -1\}$.

Solving the following convex Quadratic Programming (QP) problem can obtain the coefficient α_i.

$$\text{Maximize} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K(\mathbf{x}_i, \mathbf{x}_j) \tag{13a}$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C \tag{13b}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, i = 1, 2, ..., N \tag{13c}$$

where C is a regularization parameter that controls the balance between boundary and misclassification errors, and when the corresponding $\alpha_j > 0$, $\mathbf{x}_j$ is called support vector.

One-vs-rest strategy

We use an indirect strategy to solve multi-label classification problem, which can be solved by converting multi-label problem into multiple binary classification problems. The one-vs-rest strategy is to treat one class as positive samples and the rest classes as negative samples. We can build a binary classifier for each class label, thus construct a total of k binary classifiers.

Results

In this section, we compare various nucleotide representations, integration strategies and classification tools on our novel benchmark datasets.

Evaluation Measurements

Ten-fold cross-validation is a statistical technique to evaluate the performance of models in turn. Six parameters are used to analyze the performance of model [26], including Average Precision (AP), Accuracy (Acc), Coverage (Cov), Ranking Loss (L_r), Hamming Loss (L_h) and One-error (E_{one}).

$$Acc = \frac{1}{|D|} \sum_{i=1}^{|D|} \left| \frac{\hat{Y}_i \cap Y_i}{\hat{Y}_i \cup Y_i} \right| \tag{14a}$$

$$Cov = \frac{1}{|D|} \sum_{i=1}^{|D|} \max_{y_p \in Y_i} \hat{r}(y_p) - 1 \tag{14b}$$

$$AP = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{1}{|Y_i|} \sum_{y_q \in Y_i} \frac{|\{y_p | \hat{r}(y_p) \leq \hat{r}(y_q), y_p \in Y_i\}|}{\hat{r}(y_q)} \tag{14c}$$

$$L_r = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|\{(y_p, y_q) | \hat{f}(y_p) \leq \hat{f}(y_q), y_p \in Y_i, y_q \in \bar{Y}_i\}|}{|Y_i| \times |\bar{Y}_i|} \tag{14d}$$

$$L_h = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|\hat{Y}_i \Delta Y_i|}{|L|} \tag{14e}$$

$$E_{one} = \frac{1}{|D|} \sum_{i=1}^{|D|} |\arg\max \hat{f}(y_p) \notin Y_i| \tag{14f}$$

where $|D|$ represents the number of samples, $|L|$ represents the number of labels, $\hat{r}(y)$ indicates the rank of y in Y on the descending order, $\hat{f}(y)$ represents the score of y predicted by the classifier, Y represents the real label set, $\hat{Y}$ represents the prediction label set, $\bar{Y}$ denotes the complementary set of Y, Δ stands for the symmetric difference between two label sets.

For Coverage, Ranking Loss, Hamming Loss and One-error, the model can achieve the best performance with the smallest value. For Average Precision and Accuracy, the model can achieve the best performance with the largest value.

Performance of different nucleotide representations

We analyze seven different nucleotide property composition representations via 10-fold cross validation. Here, we compare single-kernel feature models on four RNA subcellular localization datasets, as shown in Table 1. It can be observed that kmer achieves best performance on mRNAs (AP:0.688) and lncRNAs (AP:0.745), NAC obtains best performance on miRNAs (AP:0.785), and DNC gains best performance on snoRNAs (AP:0.793). Details are shown in Supplementary Table S5. Also, we compare single-kernel feature models on four human RNA subcellular localization datasets, as shown in Table 2. It can be noticed that kmer achieves best performance on mRNAs (AP:0.750), lncRNAs (AP:0.753), and snoRNAs (AP:0.817), CKSNAP obtains best performance on miRNAs (AP:0.784). Details are shown in Supplementary Table S6.

Table 1 Average Precision of seven different nucleotide representations on four RNA datasets.

Models	mRNAs	lncRNAs	miRNAs	snoRNAs
$\mathbf{K}_{\text{kmer4}}$	**0.688**	**0.745**	0.782	0.782
$\mathbf{K}_{\text{kmer1234}}$	0.626	0.730	0.775	0.775
$\mathbf{K}_{\text{RCKmer}}$	0.658	0.733	0.726	0.775
$\mathbf{K}_{\text{NAC}}$	0.572	0.722	**0.785**	0.773
$\mathbf{K}_{\text{DNC}}$	0.668	0.737	0.760	**0.793**
$\mathbf{K}_{\text{TNC}}$	0.686	0.741	0.751	0.774
$\mathbf{K}_{\text{CKSNAP}}$	0.664	0.725	0.773	0.773

Table 2 Average Precision of seven different nucleotide representations on four human RNA datasets.

Models	H_mRNAs	H_lncRNAs	H_miRNAs	H_snoRNAs
$\mathbf{K}_{\text{Kmer4}}$	0.726	**0.753**	0.764	**0.817**
$\mathbf{K}_{\text{Kmer1234}}$	**0.750**	0.739	0.768	0.815
$\mathbf{K}_{\text{RCKmer}}$	0.717	0.738	0.700	0.794
$\mathbf{K}_{\text{NAC}}$	0.722	0.729	0.772	0.796
$\mathbf{K}_{\text{DNC}}$	0.736	0.726	0.740	0.808
$\mathbf{K}_{\text{TNC}}$	0.726	0.732	0.716	0.803
$\mathbf{K}_{\text{CKSNAP}}$	0.723	0.738	**0.784**	0.800

In order to further analyze characteristics, we make use of random forest (RF) to calculate the importantce score of each feature dimension. On four RNA datasets, feature scores of mRNAs have more balanced overall distribution, but feature scores of miRNAs and snoRNAs have irregular distributions, as shown in Figure 5. This phenomena is also reflected on four human RNA dataset, as shown in Figure 6. It indicates that miRNAs and snoRNAs have shorter sequences with less regular nucleotide property composition information.

[width = 12cm]RNAs_dataset_feature_scores.eps

Figure 5 Feature importantce scores of seven characteristics on four RNA datasets.

[width = 12cm]human_RNAs_dataset_feature_scores.eps

Figure 6 Feature importantce scores of seven characteristics on four human RNA datasets.

Performance of different integration strategies

We study five different integration strategies with SVM model as base classifier via 10-fold cross validation, including binary relevance (BR) [26], ensemble classifier chain (ECC) [27], label powerest (LP) [26], multiple kernel learning with average weights (MK-AW), multiple kernel learning with Hilbert-Schmidt independence criterion (MK-HSIC).

Here, we compare five integrated SVM strategies on four RNA subcellular localization datasets, as shown in Table 3. It can be observed that MKSVM-HSIC achieves best performance on mRNAs (AP:0.703), lncRNAs (AP:0.757), miRNAs (AP:0.787), and snoRNAs (AP:0.800). Details are shown in Supplementary Table S7. Also, we compare five integrated SVM strategies on four human RNA subcellular localization datasets, as shown in Table 4. It can be observed that MK-HSIC achieves best performance on mRNAs (AP:0.755), lncRNAs (AP:0.754), miRNAs (AP:0.791), and snoRNAs (AP:0.816). Details are shown in Supplementary Table S8. Overall accuracy of our integration strategy is significantly higher than that of other four strategies. It can be found that multiple kernel learning has an obvious advantage over other general integration strategies in dealing with classification problems.

Table 3 Average Precision of five different integration strategies on four RNA datasets.

Integrations	mRNAs	lncRNAs	miRNAs	snoRNAs
SVM-BR	0.651	0.737	0.724	0.775
SVM-ECC	0.671	0.735	0.725	0.775
SVM-LP	0.652	0.738	0.712	0.775
MKSVM-AW	0.699	0.755	0.784	0.792
MKSVM-HSIC	**0.703**	**0.757**	**0.787**	**0.800**

Table 4 Average Precision of five different integration strategies on four human RNA datasets.

Integrations	H_mRNAs	H_lncRNAs	H_miRNAs	H_snoRNAs
SVM-BR	0.720	0.731	0.670	0.794
SVM-ECC	0.711	0.731	0.673	0.800
SVM-LP	0.716	0.730	0.637	0.797
MKSVM-AW[a]	0.741	0.752	0.785	0.814
MKSVM-HSIC	**0.755**	**0.754**	**0.791**	**0.816**

According to MK-HSIC strategy, we optimize all weights of effective kernels, in order to improve the correlation between optimal combined kernel and ideal kernel. All weights for seven kernels are shown in Figure 7. Details are shown in Supplementary Table S9. On miRNAs dataset, $\mathbf{K}_{\mathrm{Kmer1234}}$ has highest kernel weight, and $\mathbf{K}_{\mathrm{NAC}}$ has second highest kernel weight. On human miRNAs dataset, $\mathbf{K}_{\mathrm{NAC}}$ has highest kernel weight. On other six dataset, $\mathbf{K}_{\mathrm{DNC}}$ similarly has highest kernel weights.

[width = 12.cm]kernel_weights.eps

Figure 7 Weights for seven different kernels on various RNA datasets.

Comparison with existing classification tools

We compare the performance of different classifiers for solving multi-label classification problem via 10-fold cross validation. We use all feature sets for training SVM [28], RF [13], ML-KNN [26], extreme gradient boosting (XGBT) [29], multi-layer perceptron (MLP) [30].

Here, we compare six classification methods on four RNA subcellular localization datasets, as shown in Table 5. It can be observed that MKSVM-HSIC achieves best performance on mRNAs (AP:0.703), lncRNAs (AP:0.757) and miRNAs (AP:0.787), and XGBT obtains best performance on snoRNAs (AP:0.806). Details are shown in Supplementary Table S10. Also, we compare six classification methods on four human RNA subcellular localization datasets, as shown in Table 6. It can be noticed that MKSVM-HSIC achieves best performance on mRNAs (AP:0.755), lncRNAs (AP:0.754), miRNAs (AP:0.791), and snoRNAs (AP:0.816). Details are shown in Supplementary Table S11. As is clearly reflected by the chart, MKSVM-HSIC achieved best performance on different RNA datasets, and XGBT and RF also have good prediction results. It proves that our novel method is valid, and our new benchmark dataset is correct and meaningful.

Table 5 Average Precision of five different classifiers on four RNA datasets.

Methods	mRNAs	lncRNAs	miRNAs	snoRNAs
SVM	0.651	0.737	0.724	0.775
RF	0.640	0.753	0.728	0.776
ML-KNN	0.576	0.683	0.673	0.748
XGBT	0.701	0.751	0.785	**0.806**
MLP	0.664	0.721	0.709	0.762
MKSVM-HSIC	**0.703**	**0.757**	**0.787**	0.800

Table 6 Average Precision of five different classifiers on four human RNA datasets.

Methods	H_mRNAs	H_lncRNAs	H_miRNAs	H_snoRNAs
SVM	0.720	0.731	0.670	0.794
RF	0.724	0.732	0.728	**0.816**
ML-KNN	0.687	0.677	0.607	0.775
XGBT	**0.755**	0.745	**0.791**	0.810
MLP	0.711	0.719	0.707	0.794
MKSVM-HSIC	**0.755**	**0.754**	**0.791**	**0.816**

In order to analyze the stability, we perform T-check on MKSVM-HSIC via 10-fold cross validation. We calculate mean value and standard deviation of Average Precision, Accuracy, Coverage, Ranking Loss, Hamming Loss and One-error, as shown in Figure 8 on RNA dataset and Figure 9 on human RNA dataset. It can be seen that the variance of MKSVM-HSIC is small, so the stability and robustness of our method is very excellent. Details are shown in Supplementary Table S12.

Importantly, RNAs are assigned in specific locations of a cell, enabling the cell to implement diverse biochemical processes in the way of concurrency. To be specific, our novel method performs outstanding rather than other prediction tools on our

[width = 12cm]RNAs_datasets_box_plots.eps

Figure 8 The robustness of our novel method on four RNA datasets.

[width = 12cm]human_RNAs_datasets_box_plots.eps

Figure 9 The robustness of our novel method on four human RNA datasets.

novel benchmark datasets. Moreover, we establish user-friendly web server with the implementation of our method.

Web server

A web server is built for the new proposed method in this paper, the URL is http://lbci.tju.edu.cn/Online_services.htm, including four servers: LocmRNA, LocmiRNA, LocmiRNA and LocsnoRNA. Each one supports two prediction formats, an on-line input single sequence or an entire multiple sequence upload file. The sequence format must be $.fasta$. It will return the possibility of each label for RNA subcellular localization, and also give the suggested labels as final prediction result.

Conclusion

In this paper, we establish multi-label benchmark data sets for various RNA subcellular localizations to verify prediction tools. Furthermore, we design an integration SVM prediction model with one-vs-rest strategy to fuse a variety of nucleic acid sequence to identify RNA subcellular localization. Finally, we propose user-friendly web server with the implementation of our method, which is a useful platform for research community.

Funding

This work is supported by a grant from the National Natural Science Foundation of China (NSFC 61772362, 61902271 and 61972280) and National Key R&D Program of China (2018YFC0910405, 2017YFC0908400).

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Hao Wang conceived and designed the experiments; Yijie Ding performed the experiments and analyzed the data; Fei Guo wrote the paper; Jijun Tang and Quan Zou reviewed the manuscript. All authors have read and approved the whole manuscript.

Acknowledgements
This work is supported by a grant from the National Natural Science Foundation of China (NSFC 61772362, 61902271 and 61972280) and National Key R&D Program of China (2018YFC0910405, 2017YFC0908400).

Supplementary Files
Supplemental charts for the article are in the supplemental data file and include 12 more comprehensive tables and a flowchart.

Author details
[1]School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University,China. [2]School of Electronic and Information Engineering, Suzhou University of Science and Technology,China. [3]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China,China. [4]School of Computational Science and Engineering, University of South Carolina,U.S..

References

1. Chou KC, Shen HB. Large-scale plant protein subcellular location prediction. Journal of Cellular Biochemistry. 2006;100(3):665–678.
2. Chou KC, Shen HB. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. Biochemical Biophysical Research Communications. 2006;347(1):0–157.
3. Shen HB, Chou KC. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Engineering Design Selection Peds. 2007;20(11):561–567.
4. Shen HB, Yang J, Chou KC. Methodology development for predicting subcellular localization and other attributes of proteins. Expert Review of Proteomics. 2007;4(4):453–463.
5. Shen HB, Yang J, Chou KC. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids. 2007;33(1):57–67.
6. Shen HB, Chou KC. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. Analytical Biochemistry. 2009;394(2):269–274.
7. Ayers D. Long Non-Coding RNAs: Novel Emergent Biomarkers for Cancer Diagnostics. Journal of Cancer Research Treatment. 2013;1(2):31–35.
8. Zhang T, Tan P, Wang L, Jin N, Li Y, Zhang L, et al. RNALocate: a resource for RNA subcellular localizations. Nucleic acids research. 2016;45(D1):D135–D138.
9. Mas-Ponte D, Carlevaro-Fita J, Palumbo E, Pulido TH, Guigo R, Johnson R. LncATLAS database for subcellular localization of long noncoding RNAs. Rna. 2017;23(7):1080–1087.
10. Chou KC, Shen HB. Recent progress in protein subcellular location prediction. Analytical Biochemistry. 2007;370(1):1–16.
11. Cheng L, Leung KS. Quantification of non-coding RNA target localization diversity and its application in cancers. Journal of molecular cell biology. 2018;10(2):130–138.
12. Feng P, Zhang J, Tang H, Chen W, Lin H. Predicting the organelle location of noncoding RNAs using pseudo nucleotide compositions. Interdisciplinary Sciences: Computational Life Sciences. 2017;9(4):540–544.
13. Cao Z, Pan X, Yang Y, Huang Y, Shen HB. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. Bioinformatics. 2018 02;34(13):2185–2194. Available from: https://doi.org/10.1093/bioinformatics/bty085.
14. Xiao Y, Cai J, Yang Y, Zhao H, Shen H. Prediction of MicroRNA Subcellular Localization by Using a Sequence-to-Sequence Model. In: 2018 IEEE International Conference on Data Mining (ICDM). IEEE; 2018. p. 1332–1337.
15. Yang Y, Fu X, Qu W, Xiao Y, Shen HB. MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA–disease association. Bioinformatics. 2018;34(20):3547–3556.
16. Zhao-Yue Z, Yu-He Y, Hui D, Dong W, Wei C, Hao L. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. Briefings in Bioinformatics. 2020;.
17. Shen HB, Chou KC. Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites. Journal of Biomolecular Structure Dynamics. 2010;28(2):175–186.
18. Shen HB, Chou KC. Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochemical Biophysical Research Communications. 2007;355(4):0–1011.
19. Ying-Ying X, Fan Y, Hong-Bin S. Incorporating organelle correlations into semi-supervised learning for protein subcellular localization prediction. Bioinformatics. 2016;(14):14.
20. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010 01;26(5):680–682. Available from: https://doi.org/10.1093/bioinformatics/btq003.
21. Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform. 2019;10.
22. Gretton A, Bousquet O, Smola A, Schölkopf B. Measuring Statistical Dependence with Hilbert-Schmidt Norms. vol. 3734; 2005. .
23. Yamada M, Jitkrittum W, Sigal L, et al. High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso. Neural Computation. 2013;26(1):185–207.
24. Ding Y, Tang J, Guo F. Identification of drug-target interactions via multiple information integration. Information Sciences. 2017;418-419:546 – 560.
25. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273–297.
26. Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering. 2013;26(8):1819–1837.
27. Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. Machine learning. 2011;85(3):333.
28. Su ZD, Huang Y, Zhang ZY, Zhao YW, Wang D, Chen W, et al. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. 2018 06;34(24):4196–4204. Available from: https://doi.org/10.1093/bioinformatics/bty508.
29. Chen T, He T, Benesty M, Khotilovich V, Tang Y. Xgboost: extreme gradient boosting. R package version 04-2. 2015;p. 1–4.
30. Oh C, Zak SH, Mirzaei H, Buck C, Regnier FE, Zhang X. Neural network prediction of peptide separation in strong anion exchange chromatography. Bioinformatics. 2007;23(1):114–118.

Figures

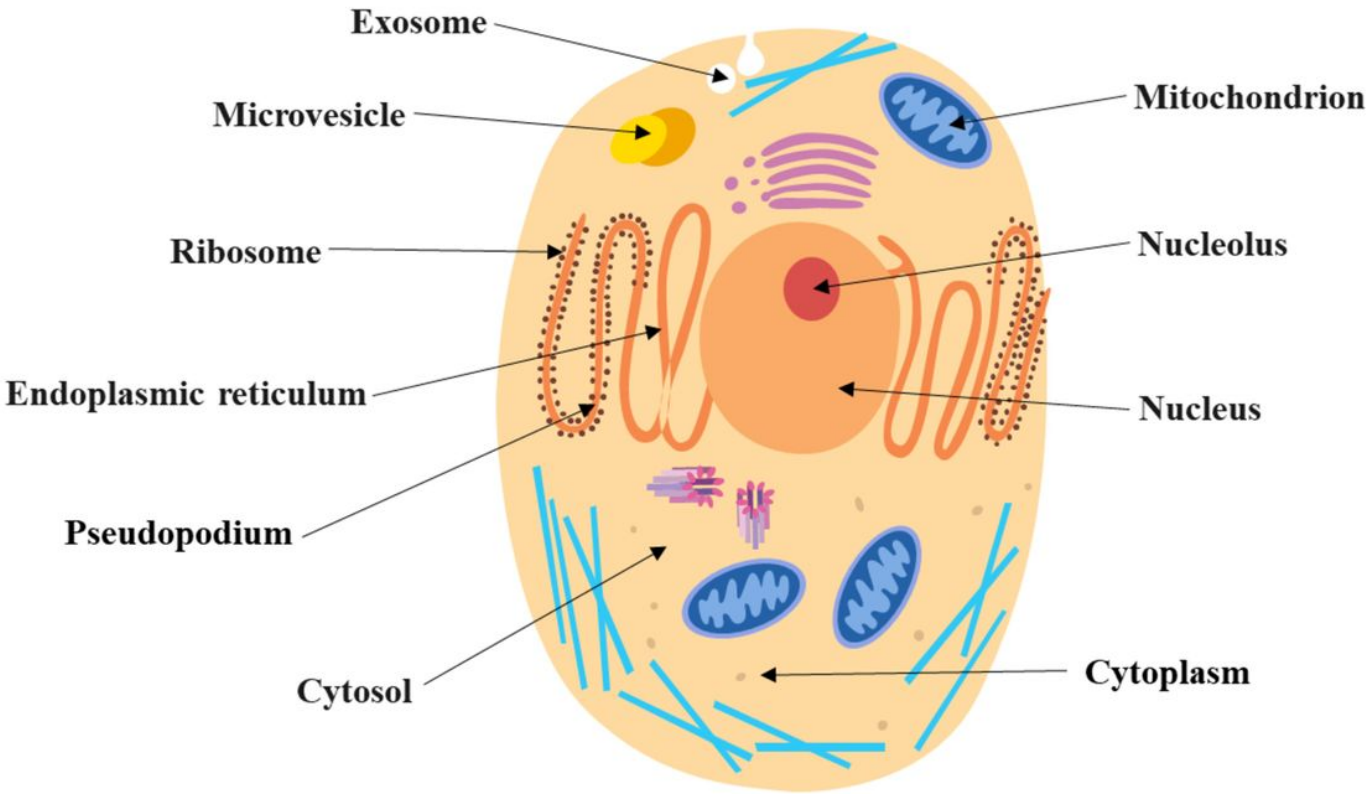

Figure 1

Schematic diagram of RNA subcellular localizations in cells.

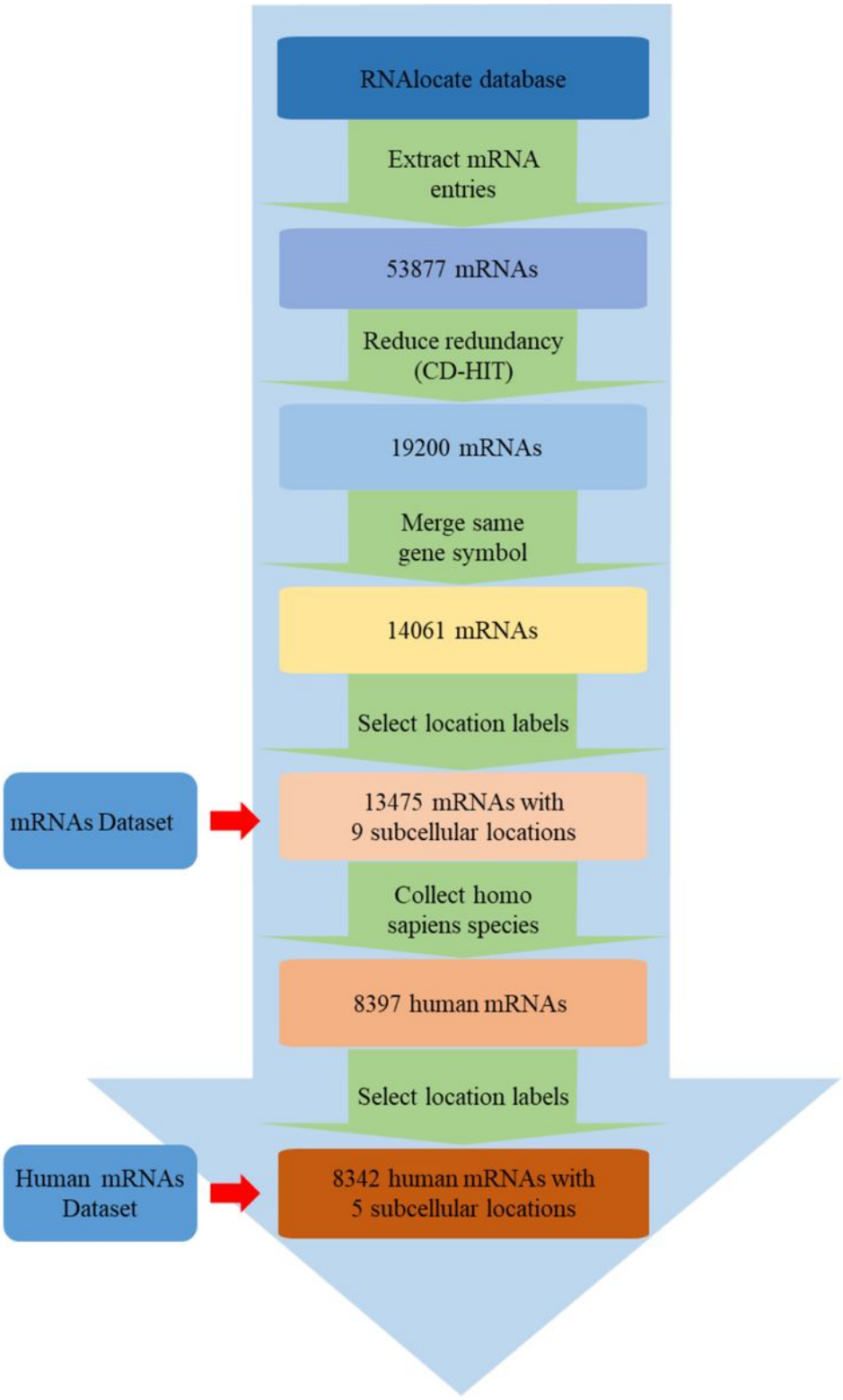

Figure 2

The owchart of mRNA subcellular localization dataset construction framework.

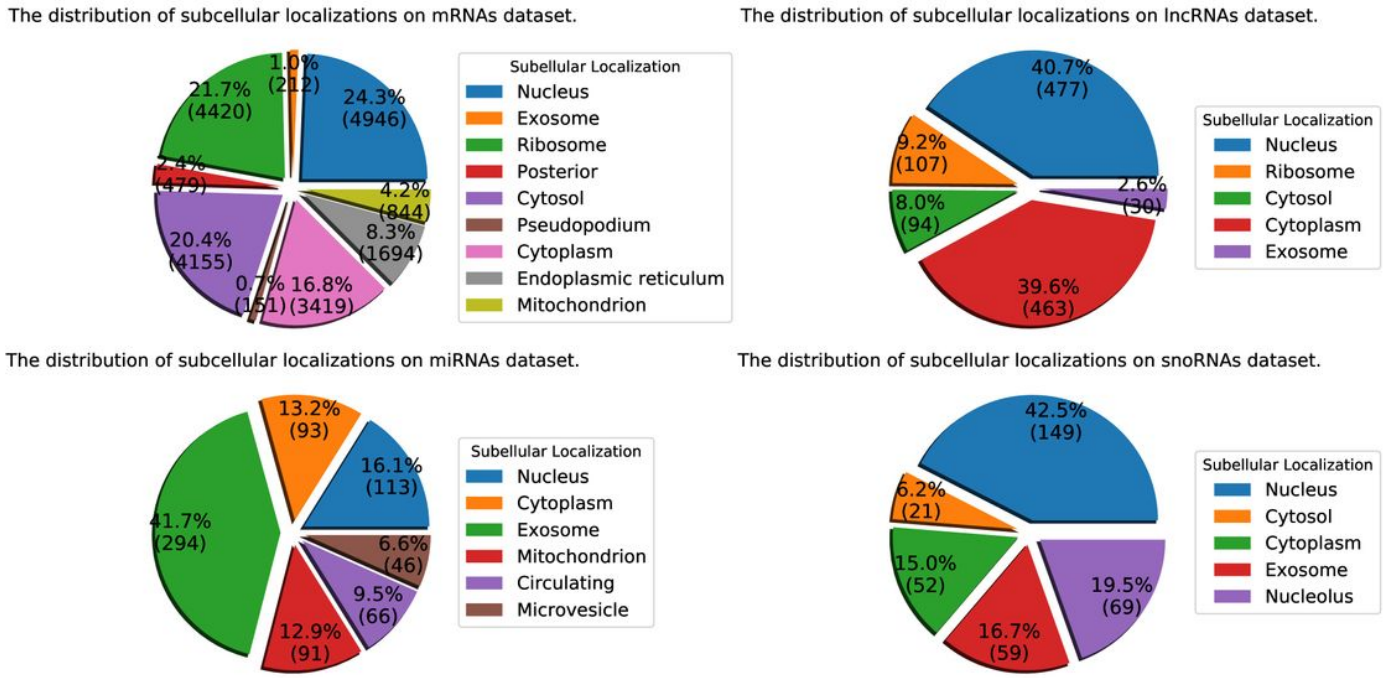

Figure 3

The statistical distributions of four RNA subcellular localization datasets.

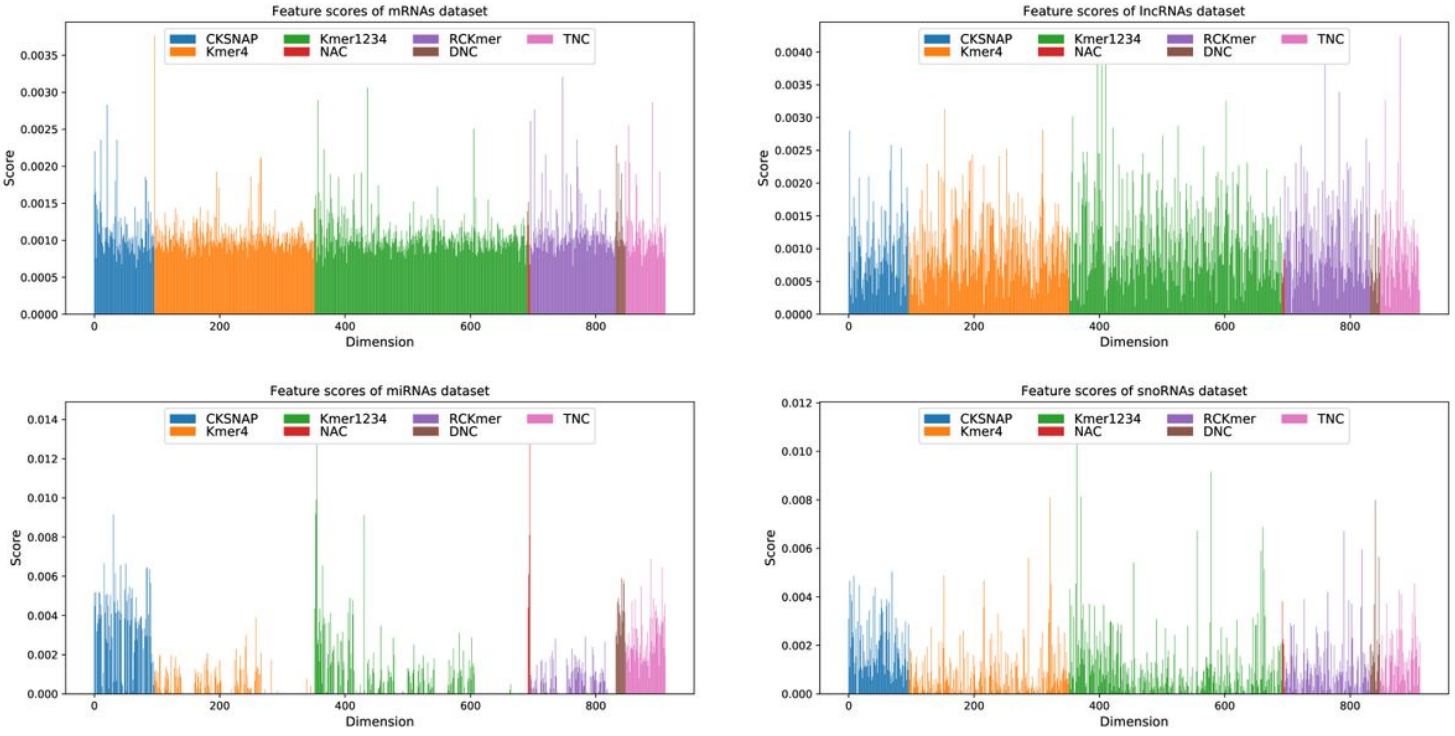

Figure 4

The statistical distributions of four human RNA subcellular localization datasets.

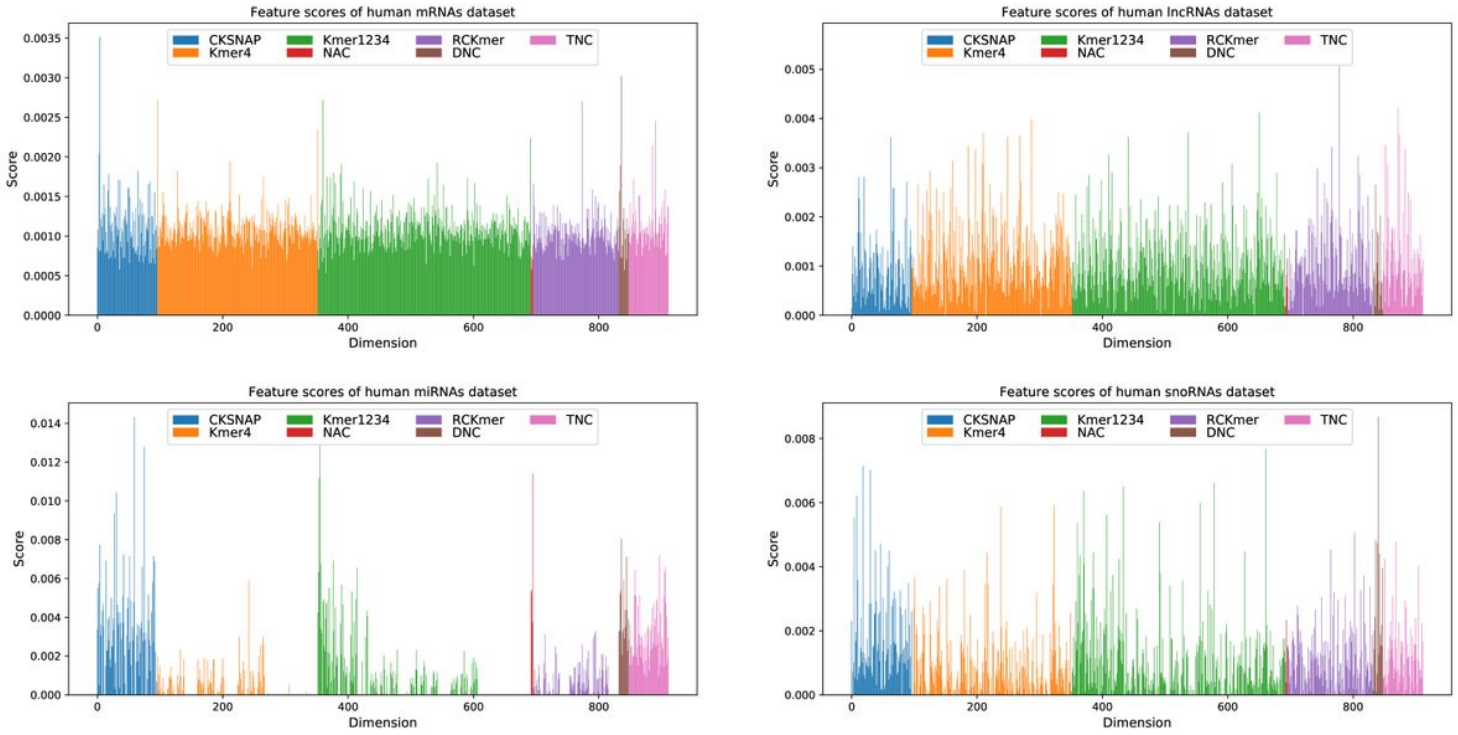

Figure 5

Feature importantce scores of seven characteristics on four RNA datasets.

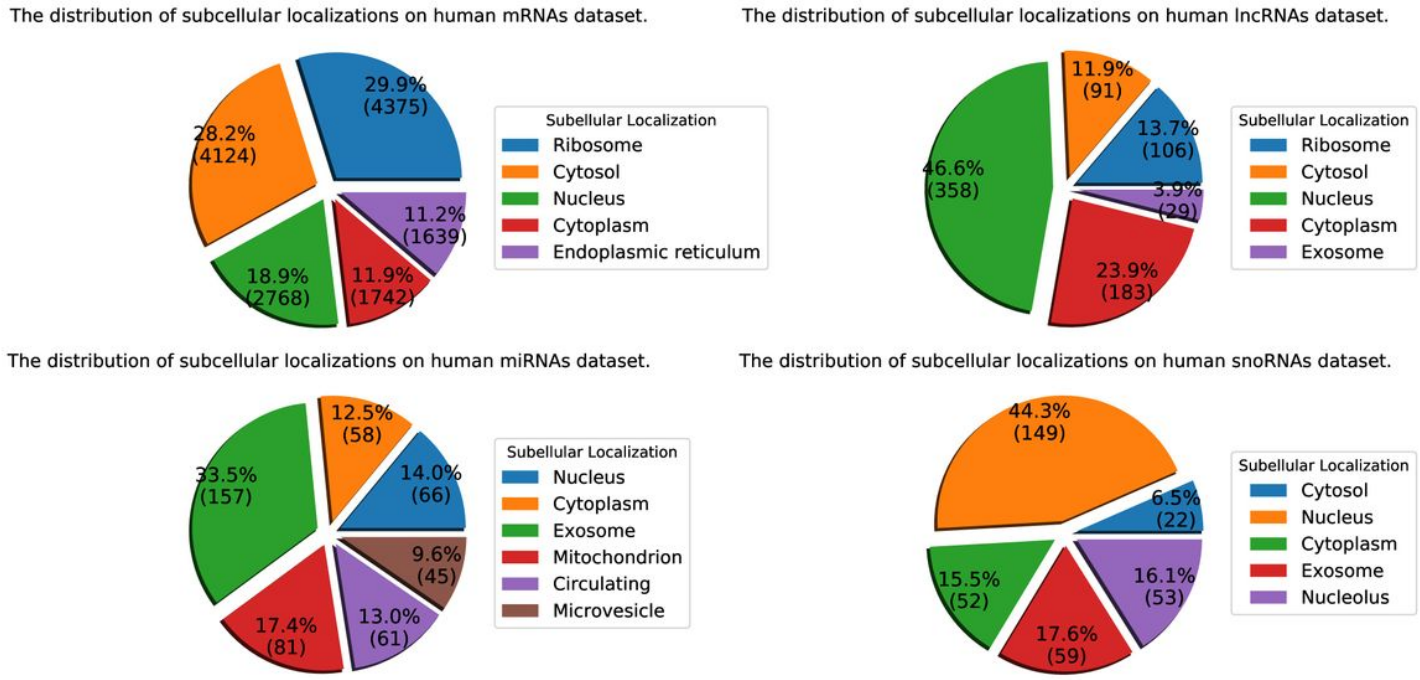

Figure 6

Feature importantce scores of seven characteristics on four human RNA datasets.

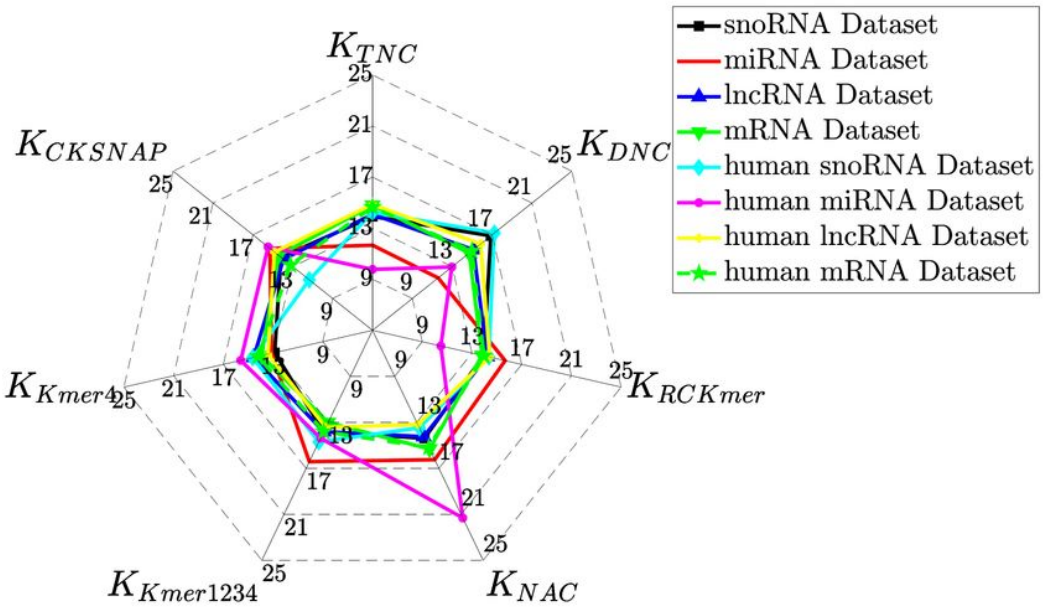

Figure 7

Weights for seven di_erent kernels on various RNA datasets.

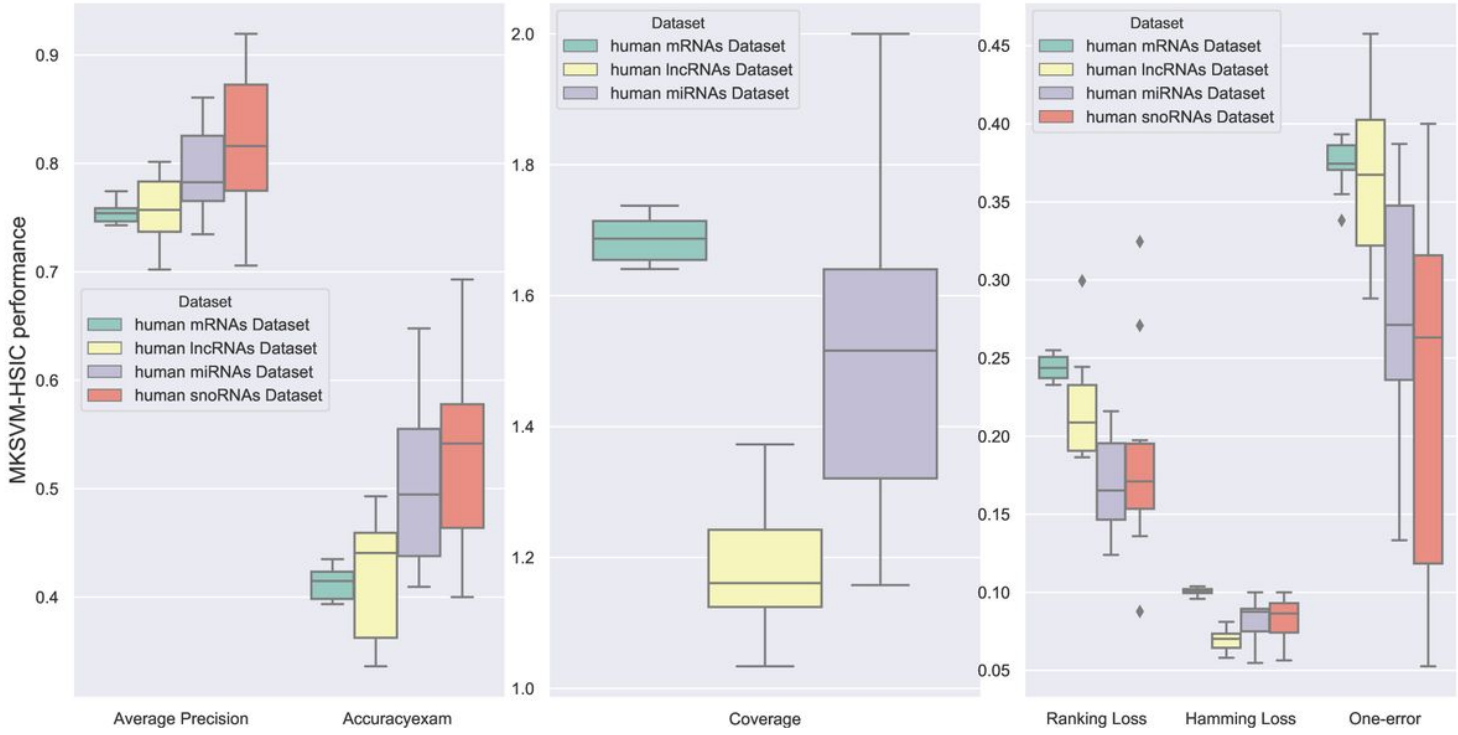

Figure 8

The robustness of our novel method on four RNA datasets.

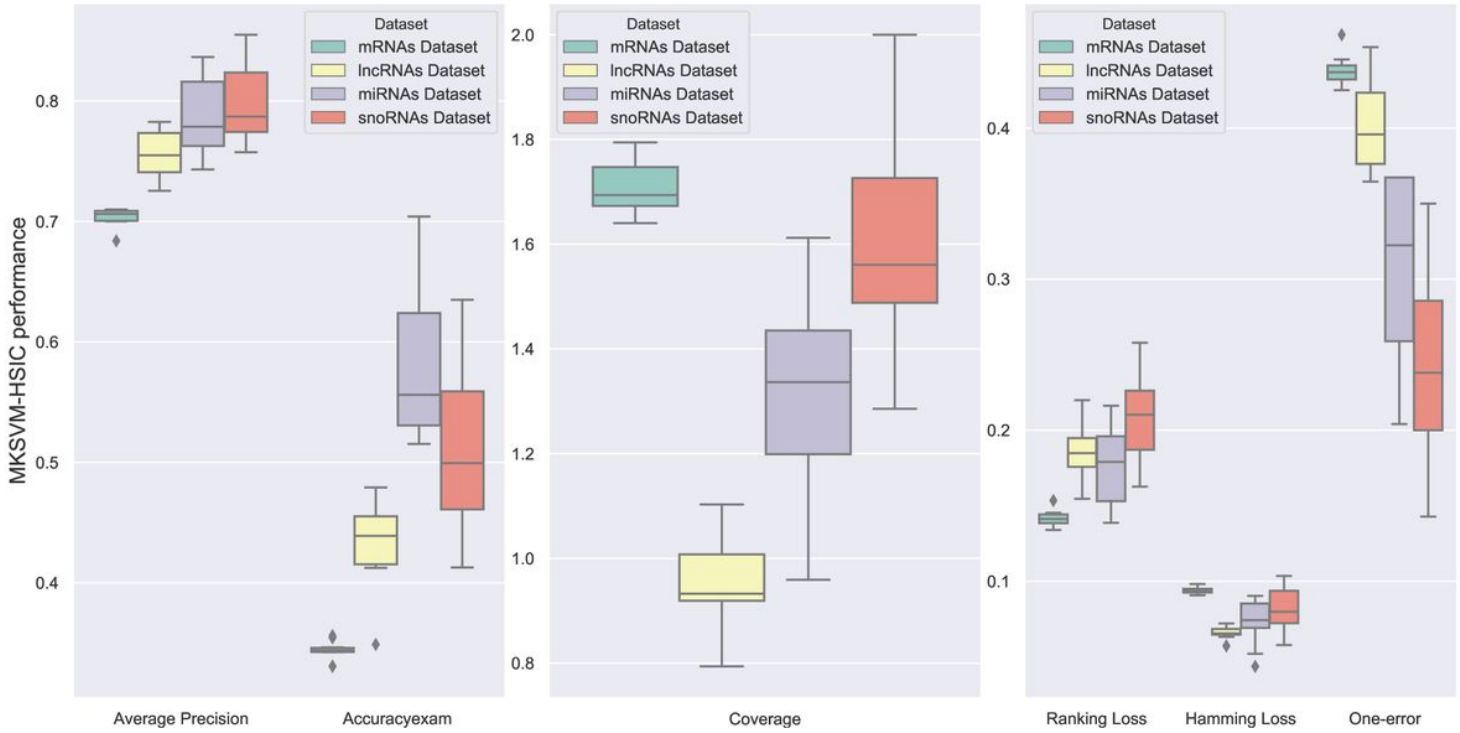

Figure 9

The robustness of our novel method on four human RNA datasets.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata.pdf