A Deep Learning Approach to LncRNA Subcellular Localization Using Inexact q-mers

Weijun Yi
Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV, USA
wy0003@mix.wvu.edu

Donald A. Adjeroh
Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV, USA
Donald.Adjeroh@mail.wvu.edu

***Abstract*—Long non-coding Ribonucleic Acids (lncRNAs) can be localized to different cellular components, such as the nucleus, exosome, cytoplasm, ribosome, etc. Their biological functions can be influenced by the region of the cell where they are located. Many of these lncRNAs are associated with different challenging diseases. Thus, it is crucial to study their subcellular localization. However, compared to the massive number of lncRNAs, only relatively few have annotations in terms of their subcellular localization. Conventional computational methods use *q*-mer profiles from lncRNA sequences and train machine learning models, such as support vector machines and logistic regression with the profiles. These methods focus on the exact q-mer. Given possible sequence mutations and other uncertainties in genomic sequences and their role in biological function, a consideration of these changes might improve our ability to model lncRNAs and their localization. We hypothesize that considering these changes may improve our ability to predict subcellular localization of lncRNAs. To test this hypothesis, we propose a deep learning model with inexact q-mers for the localization of lncRNAs in the cell. The proposed method can obtain a high overall accuracy of 94.7%, an average of 91.3% on a benchmark dataset, using 8-mers with mismatches. In comparison, the exact 8-mer result was 89.8%. The proposed approach outperformed existing state-of-art lncRNA localization predictors on two different datasets. Our results, therefore, support the hypothesis that deep learning models using inexact *q*-mers can improve the performance of computational lncRNA localization algorithms.**

Keywords—deep learning, subcellular localization, CNN, inexact q-mer

I. INTRODUCTION

Non-coding RNAs (ncRNAs) and protein-coding genes are two constituent parts of the human genome [1]. Usually, non-coding RNAs can be divided into small ncRNAs with lengths less than 200 nucleotides and long non-coding RNAs (lncRNAs) with lengths greater than or equal to 200 nucleotides [2]. Since lncRNAs were first discovered in the early 1990s, the family of lncRNAs has expanded rapidly. A recent study indicated that there are over 270,000 lncRNA transcripts in humans [3]. Unlike the protein-coding genes, which are functional units of heredity [4], non-coding RNAs were once deemed non-functional. They were perceived as the product of spurious transcription [5]. However, the application of high-throughput sequencing technologies [6] has shed more light on the transcriptional units. Accumulative evidence shows that ncRNAs, specifically lncRNAs, exhibit biological functions. LncRNAs have been associated with biological processing, such as chromatin modification, cell cycling, protein transcription, and translation [7], [8]. LncRNAs also play essential roles in diseases, including cancer, autism, Alzheimer's disease, and others [9]–[11]. A popular database of lncRNA-associated diseases, LncRNADisease [12] records 10,564 experimentally supported lncRNA-disease associations. With info on over 450+ unique diseases, including various cancers, nervous system disorders, etc., this underlines the critical role of lncRNAs in many complex diseases.

Similar to proteins, the function of lncRNAs has been linked with their subcellular localization in the cell [13]. Therefore, understanding the subcellular localization of lncRNAs and their dynamic changes can also help to explain the function of newly discovered lncRNAs [14]. To study the RNA subcellular localization, a database, RNALocate v2.0 [15], was constructed in 2016 and updated in 2021. 213,260 RNA subcellular localization entries validated by experimental evidence. Experimental results show RNAs can be located in the nucleus, cytoplasm, ribosome, exosome, nucleoplasm, chromatin, cytosol, endoplasmic reticulum, and plasma membrane. See [15]. The dataset contains 9,587 lncRNAs, some of them located in different components of the cell. Only 6728 unique lncRNAs were annotated. In 2017, another database, LncATLAS [16], for subcellular localization of lncRNAs was introduced by calculating the cytoplasmic/nuclear relative concentration index. 6768 lncRNAs were annotated. Compared to the large number of lncRNAs, only a few lncRNAs have been annotated.

Recent studies have analyzed the use of deep learning techniques in lncRNA identification [17]. Similarly, some work suggest that lncRNA subcellular localization can be predicted from known subcellular localization datasets using computational approaches. These studies make predictions with high accuracy by extracting shot nucleotide segments (called *q*-mers or *q*-grams) from lncRNA sequences and training machine learning models, such as Random Forest (RF), support vector machines (SVM), or deep neural network models [18]–[20].

Traditional computational methods have focused on exact q-mers. However, given possible mutations in genomic sequences [21], and other uncertainties in biological systems, exact pattern matching may not be adequate to model problems in RNA localization. Thus, segments with inexact matches or mismatch(es) may provide equally biological information in the modeling. In this work, we are interested in whether inexact *q*-mers can impact the computational prediction of lncRNA localization based on lncRNA sequences.

978-1-6654-0126-5/21/$31.00 ©2021 IEEE

II. Background and Related Work

A. Subcellular localization

RNAs play crucial roles in cellular processes, including translating genetic information, regulating gene activity, and cellular differentiation [22]. These functions are determined by RNAs' location in the cell [14], [23]. The cell of eukaryotic organisms can be divided into functionally distinct membrane-bound compartments [23] (See Fig 1.), which are linked with different phases of biological processes [24]. To understand the function of RNA, we need to understand its subcellular localization. Experiment methods, such as FISH, which map RNAs to their subcellular localization, require knowledge of molecular chemistry and specialized instruments and techniques.

Conventionally, we divide RNAs into coding and non-coding based on their coding potential [25]. Coding RNAs encode protein. Non-coding RNAs act as cellular regulators without encoding proteins. Unlike the coding RNAs, which have been studied widely, lncRNAs are more challenging to explore, given their low expression levels [26]. Thus, using information from known datasets to predict the subcellular localization of lncRNAs has become a significant challenge. There are existing databases [16], [18], [19] which annotate lncRNAs with their subcellular localizations, such as cytoplasm, nucleus, ribosome, exosome, etc. Therefore, we can treat the prediction of subcellular localization as a classification problem. For coding RNAs, there are many predictors of protein localization, which have been developed since the 1990s [27]. Many of them take computational approaches, such as artificial neural networks, or support vector machines (SVM). However, in contrast to protein-coding RNAs, only a few methods have been proposed for predicting lncRNAs subcellular localization.

B. Prior computational approaches

Research shows that we can represent the RNA sequence using a discrete model: pseudo-k-tuple nucleotide composition (PseKNC) [28]. In the PseKNC model, q-length substrings (q-mers) are extracted from the RNA sequence. Each substring can be treated as an RNA motif that contains some biological information. Then, the RNA sequence is decomposed into a set of small-sized segments, which are typically more efficient to analyze, than long RNA sequences. Along this line of thought, Kirk et al. [29] showed that profiles based on such q-mers could be used to analyze lncRNAs subcellular localization.

General computational methods predict the localization of lncRNAs by extracting q-mer features from the lncRNA sequence. Some of them select particular nucleotide segments as features. Based on the features, they train a prediction model, such as random forest, support vector machines, or deep neural network to make a prediction. In LncLocator [18], Cao *et al.* created an annotated subcellular localization dataset of lncRNAs from RNALocate [30]. The dataset contains 612 lncRNAs localized to 5 locations in the cell, including the nucleus, cytoplasm, cytosol, ribosome, and exosome (see Table 1). They extract q-mer segments (q=4,5,6) from the lncRNA sequences. Given the low discrimination ability of very short segments, they feed 4-mer features into a stacked autoencoder model to create a high-level feature representation of the sequence. They tested their data in various scenarios. The overall accuracy was 59.8% on the five-class dataset.

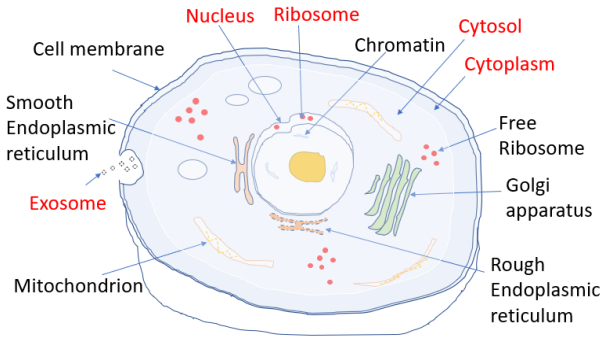

Figure 1. The structure of an animal cell. The key target lncRNA localizations in most datasets are nucleus, exosome, ribosome, and cytoplasm (indicated in red).

The iLoc-lncRNA [19] predicts lncRNAs subcellular localization by feeding octamer features into an SVM model. They build a 4-class dataset from the RNALocate database [30]. The classes correspond to the following localizations: nucleus, cytoplasm, ribosome, and exosome. There are 655 lncRNAs in the dataset (see Table 1). First, they extract 8-mer features from the lncRNA sequences. Then, because high dimensional features will produce several problems such as over-fitting and redundant noise, they selected features based on the 8-mer feature distribution probability. They finally picked 4107 8-mer features and then trained the SVM model with the extracted features. The overall accuracy was 86.72% on the 4-class dataset.

Gudenas et al. [20] built a two-class dataset from the ENCODE project. First, they quantified the lncRNA transcript differences between nuclear and cytosolic, applying $\log_2$ fold-change threshold to allocate 8678 lncRNAs to cytosolic and nuclear, 4380 for cytosolic, and 4298 for nuclear. They then extracted q-mer features (q=2,3,4,5) from the lncRNA sequences. Next, they added RNA-protein binding motifs to the feature map, and passed these to a deep neural network. They obtained an accuracy of 72.4%. Fan et al., in lncLocPred [31], built a four-class dataset from the RNALocate database [30]. The database contains 396 lncRNAs. They use this dataset as an independent dataset and dataset in iLoc-lncRNA [19] as the benchmark dataset. First, they collect features using q-mers (q=5,6,8), triplet, and PseDNC. They then trained a logistic regression model using the selected features.

C. Our approach

This paper examines the impact of inexact q-mer profiles on the prediction performance on multi-label lncRNA subcellular localization. In this paper, both exact and inexact q-mer profiles are extracted from the lncRNA sequences to build feature maps, and then a 1D convolutional neural network (1D CNN) model is trained. To compare the performance of this method with the existing state-of-the-art techniques, we use the datasets from LncLocator [18] with 5-components and iLoc-lncRNA [19]

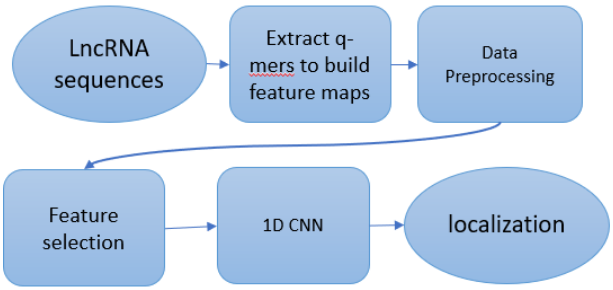

Figure 2. Workflow of our approach. See text.

with four components. We also test the pre-trained features from iLoc-lncRNA.

III. METHODS

Figure 2 shows the workflow in our methodology. We first extract *q*-mer profiles from the lncRNA subcellular localization dataset to build a feature map and then apply data preprocessing. We finally feed our preprocessed data into a 1D CNN model to do classification.

A. Dataset

The datasets from lncLocator [18], iLoc-lncRNA [19] are tested. Both obtained their lncRNA sequences from RNALocate [30]. Four subcellular localizations (classes) are retained in iLoc-lncRNA dataset (as our benchmark dataset) and five in lncLocator, including the nucleus, cytoplasm, cytosol, ribosome, and exosome (See Table 1). We test our method on these two datasets. Because of the time limitation, we didn't try the dataset in lncLocPred. The datasets have annotated each lncRNA sequence with a particular localization. We treat the task as a supervised classification problem.

B. Feature representation

LncRNA is transcribed from DNA. *LncRNA* consists of *a* string of nucleotides bases. These bases are adenine (A), guanine (G), uracil (U), and cytosine (C). The sequence of lncRNA can be represented as: $S = S_1S_2...S_i...S_n$, with $S_i \in$ {A, G, C, U}. Here n is the length of the sequence, and S_i is i^{th} nucleotide base, $1 \leqslant i \leqslant n$.

1) q-mer profile

The *q*-mer is a substring of a sequence with length *q*. A possible q-mer will be a q-length substring with one of the A, C, G, U symbols for a lncRNA sequence. There are 4^q possible different *q*-mers in one lncRNA sequence. We build the feature map with the *q*-mer profile, which captures the probability distribution

Table 1. Characterizing two lncRNA subcellular localization datasets. Both are derived from RNALocate [30].

location	number of lncRNA	
	LncLocator	iLoc-lncRNA
Nucleus	152	156
Cytoplasm	301	426
Cytosol	91	---
Ribosome	43	43
Exosome	25	30
Total	612	655

Table 2. (8,1) mismatch q-gram, for example, Q=AGCUAGUA. See text.

1	A	G	C	U	A	G	U	A
2	*	G	C	U	A	G	U	A
3	A	*	C	U	A	G	U	A
4	A	G	*	U	A	G	U	A
5	A	G	C	*	A	G	U	A
6	A	G	C	U	*	G	U	A
7	A	G	C	U	A	*	U	A
8	A	G	C	U	A	G	*	A
9	A	G	C	U	A	G	U	*

(or frequency of occurrence) of each given possible *q*-mer. For the *q*-mer profile, typically, each row corresponds to one lncRNA, and each column corresponds to one of the possible *q*-mers. Each cell is a feature value that represents the frequency of the *q*-mer in the given sequence. We compute the feature value by running a *q*-length window with stride one across the sequence. If the segment is in the sequence, then the frequency of the segment is set to its feature value. Otherwise, the feature value is 0. Based on this, we define the feature map (FM) of a lncRNA sequence as FM(S) = $\{Q_i: f_i\}$, $1 \leqslant i \leqslant N\}$, Here S is the sequence, Q_i is the i^{th} *q*-mer, f_i is the corresponding feature value, and N is the number of possible unique *q*-mers in the sequence. For example, we can compute the 3-mer feature map for the sequence **S=*AGCUAGUA***. First, we find all the 3-mer combinations of A, G, C, and U. Then, we map the frequency of each 3-mer. Finally, we get the feature map: FM(**S**)={AAA:0, AAG:0, …, AGC:1, …, AGU:1, …, UUU:0}.

2) Inexact q-mer profiles.

In this work, we introduce the idea of inexact *q*-mer profile. We focus on the *q*-mers with *k*-mismatch(es), also called the (*q, k*)-mismatch kernel, which provides the idea of mismatching in biological interest [32], [33]. Given a *q*-mer, we compute the frequency of other matching *q*-mers, where a match is allowed to admit at most *k*-mismatches, here $k < q$. Each matching *q*-mer is still required to have the same length of *q*, just like the given *q*-mer. Thus, for a given *q*-mer, say *Q*, the result of (*q*, *k*)-mismatch is thus a set of q-mers, such that each feature (*q*-mer) in the collection has the same length as *Q*, and there are at least *q-k* base(s) that have an exact match with bases in the given *q*-mer, *Q*. For example, for the *q*-mer sequence **Q=*AGCUAGUA***, *the* (8, 1)-mismatches are shown in Table 2. We use the hamming distance to measure the mismatch. Row 1 is the original sequence. From row 2 to 9, each row denotes a mismatch which happened at a different location of the original sequence. The asterisk indicates one of three bases that is other than the original base, respectively. Thus, for each row, there are 3 possible mismatch *q*-mers. Hence, in this example, there are 24 mismatch *q*-mers. We then set the frequency value of 8-mer, ***AGCUAGUA*,** with 25 (24 mismatches + 1 match). This work uses a naïve method to compute the (*q, k*)-mismatch feature map. There exist efficient data structures using suffix trees and suffix arrays[34] to compute the feature map.

3) Data preprocessing

The feature maps of the lncRNA sequences in the two datasets are counts of the q-mers. First, we normalize the counts according to the length of lncRNA sequences, respectively. We then split the dataset into training and testing sets with a ratio 4 to 1. We then do z-score normalization on the training and test sets. The formula is as follows: $z_i = (x_i - \mu)/\sigma$. Here z_i is the score of i-th q-mer feature, x_i is the count of q-mer, μ is the mean q-mer count of all of the lncRNA sequences, and σ is the standard deviation.

4) Feature selection

The dimension of the feature map is 4^q. It grows exponentially with q. A high-dimensional feature map will typically be noisy, which will reduce the prediction accuracy, often leading to over-fitting [35]. Further, this can also pose a significant computational challenge, especially as q increases. We test our q-mer approach with q=3, 4, …, 8, including some q-mer combinations. For 7-mer and 8-mer, we have 16384 and 65536 features, respectively. Thus, we need to reduce the size of the feature map. We applied the χ^2 test feature selection method from Scikit-learn [36] to get a feature rank and then select the optimal subset. We started with a subset with the first feature in the rank and added eight features into the subset each time. We tested the performance of the model and took the subgroup with the highest accuracy. For 8-mers, we tried the 4107 pre-trained features reported in iLoc-lncRNA[19].

C. Deep learning architecture

The convolutional neural network is a class of deep neural networks that employ a mathematical operation called convolution in at least one of its layers [35]. With convolution, a new feature map from the input feature is detected. Unlike 2D CNN, which broadly operates on 2-dimension data such as images and videos, 1D CNN is designed to work on one-dimensional signals such as time series digital signal processing (DSP). This is often used in time domain analysis and frequency domain analysis.

The feature map of the lncRNA dataset has two attributes: 1) The feature is with a fixed length, and 2) only the feature frequency is considered. Furthermore, the location of each q-mer feature is ignored. Thus, a 1D CNN model is suitable for this scenario. In this paper, we build the 1D CNN model with Keras [37]. Figure 3 shows the proposed architecture for the 1D CNN model.

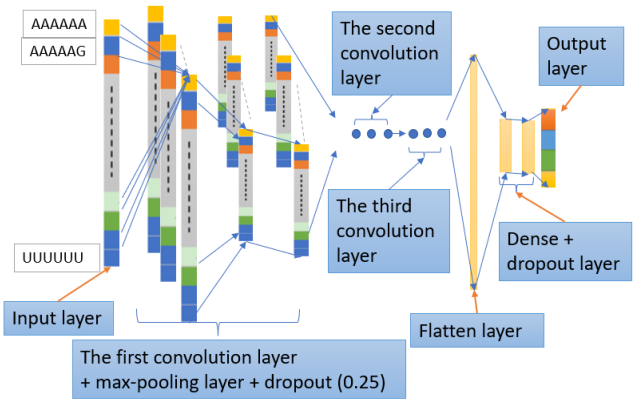

Figure 3. 1D CNN architecture used in this work. We were using 6-mer profile as an example. The second and the third convolution layer blocks have the same structure as the first one, that is, a convolution layer followed by a maxpooling layer with 0.25 dropout.

D. Evaluation

We use overall accuracy (OA), sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC) and F1-score to evaluate the performance of the 1D CNN model, which were computed with the equation OA = (TP + TN) / (TP + TN + FP + FN), Sn = TP / (TP + FN), Sp = TN / (TN + FP), MCC = (TP × TN – FP × FN) /*sqrt* $((TP+FP)(TP+FN)(TN+FP)(TN+FN))$, Here TP is the true positive, the number of positive samples we predict correctly. TN is the true negative, the number of negative samples we predict correctly predicted. FP is false positive, the number of negative samples we incorrectly predict as positive. FN is the false negative, the number of positive samples predicted as negative, and *sqrt* is square root function.

IV. Experiments & Results

We test our 1D CNN model on the 5-class dataset from the lncLocator [18] and the 4-class dataset from the iLoc [19]. We tried some q-mer combinations and q-mer (q=3, 4, 5, 6, 7, 8) with various mismatches. Given the randomness of the CNN model [38], we did experiments ten times for different scenarios on the two datasets. Finally, we calculated the average of the performance. The results showed that our model performed better on the 4-class iLoc dataset than the 5-class lncLocator dataset and that q-mer with mismatch(es) could improve the classification performance.

A. Results with exact q-mers

The overall accuracy using exact q-mers on the two datasets is shown in Table 3. The table shows that our model performed better on the 4-class (iLoc-lncRNA) dataset, when compared with the 5-class (lncLocator) dataset. With increasing q, the overall accuracy on the iLoc-lncRNA dataset rose from around 64% (for 3-mer) to 89.85% (for 8-mer), and from about 53% (for 3-mer) to 71.46% (for 8-mer) on the lncLocator dataset. It indicates that the longer the segments might provide more discriminative features for determining the lncRNA subcellular localization. Combining different q-mers using fusion did not seem to improve the result (results not shown).

Table 3. Overall accuracy using exact q-mers.

	3	4 mer	5 mer	6 mer	7 mer	8 mer
iLoc-lncRNA	64.35	64.89	64.27	65.5	68.78	89.85
lncLocator	53.33	54.72	56.1	54.96	53.58	71.46

B. Results using q-mers with k-mismatch

We experimented on q-mers (q=3, 4, 5, 6, 7, 8) with k mismatches ($0 \le k \le q$-1) on the two datasets. Table 4 shows the overall accuracy (mean) of using q-mer with k mismatch(es) on the lncLocator [18] dataset. The table shows the overall accuracy of the prediction in different scenarios. For example,

we can have the utmost seven mismatches in 8-mer. When the number of mismatches is greater than 2, there is a slight improvement in the overall accuracy. The overall accuracy increases from 71.5% to 71.9%, and the highest is 72.2%. However, for smaller q-mers, e.g., q=3, 4, and 5, increasing the number of mismatches did not necessarily lead to increased accuracy and perhaps more noise in the model. This may point to the need for a more detailed study of the interplay between q and k in this (q, k)-mismatch model.

Table 4. Results of q-mers with mismatch(es) on 5-class lncLocator dataset.

	0 miss	1 miss	2 miss	3 miss	4 miss	5 miss	6 miss	7 miss
8 mer	71.5	69.9	71.9	70.9	71.8	71.7	70.7	72.2
7 mer	53.6	52.1	54.8	55.5	55.5	54.2	55.4	
6 mer	54.7	55.0	54.6	53.4	52.0	56.6		
5 mer	56.1	55.0	53.6	52.4	54.8			
4 mer	54.7	51.8	53.1	50.9				
3 mer	53.3	53.1	51.5					

Table 5 shows the corresponding results for on the iLoc dataset. We see the overall accuracy is improved with $k > 3$ using 8-mers. For example, the highest score is 91.3%, 1.4% higher than the exact 8-mer. There are also significant improvements in using q= 5, 6, 7 with the k-mismatches. Thus, we can conclude that q-mer with mismatches performs better than the exact q-mer on this dataset.

Table 5. Results of q-mers with k-mismatch(es) on the 4-class iLoc dataset.

	0 miss	1 miss	2 miss	3 miss	4 miss	5 miss	6 miss	7 miss
8 mer	89.9	88.1	89.2	90.3	90.8	90.1	90.1	91.3
7 mer	68.8	70.2	71.1	70.6	71.4	71.0	70.5	
6 mer	65.5	65.9	67.3	66.3	65.3	65.2		
5 mer	64.3	65.1	65.5	66.3	64.9			
4 mer	64.9	63.8	64.4	63.3				
3 mer	64.4	64.4	65.0					

C. Comparison with existing state-of-the-art predictors

We compare with the two popular state-of-the-art lncRNA subcellular localization methods. Table 6 shows the performance difference between our proposed method with inexact q-mers, and the method of lncLocator [18], using the 5-class lncLocator dataset. The proposed q-mer method with k-mismatches showed improvement over previous results in [18]. Table 7 shows a similar performance comparison on the 4-class iLoc dataset [19]. We compare our approach using 8-mers with 7 mismatches against the localization method proposed in [19]. The average accuracy of our approach is 91.3%±0.026, which is 4.58% higher than iLoc-lncRNA [19].

Table 6. Comparative results on lncLocator dataset [18].

		Nucleus	Cytoplasm	Ribosome	Exosome	Cytosol	OA
Our method	Sn	0.729	0.855	0.722	0.32	0.378	72.2
	Sp	0.905	0.752	0.979	0.993	0.936	
	MCC	0.637	0.612	0.716	0.386	0.35	
lncLocator	Sn	0.3815	0.8801	0.07	0.04		66.5
	Sp	0.9217	0.3636	0.9753	0.9727		
	MCC	0.357	0.288	0.07	0.015		

From the results above, we can see a general tendency that the accuracy increases with increasing q. The results also show the power of the 1D-CNN model and architecture we used in this work. The deep learning model did an excellent job on subcellular localization. In (q, k) mismatch model, we can improve the prediction accuracy. When we take 8-mer with 7 mismatches, we can get an average of 91.3% and a maximum accuracy of 94.7% on the iLoc dataset. It appears that the inexact q-mer may be capturing some crucial biological signals that are important for lncRNA localization in the cell.

Table 7. Comparative results on iLoc dataset [19].

		Nucleus	Cytoplasm	Ribosome	Exosome	OA
our method	Sn	0.865	0.959	0.933	0.483	91.3
	Sp	0.966	0.852	0.996	0.994	
	MCC	0.84	0.828	0.936	0.584	
iLoc	Sn	0.7756	0.9906	0.4651	0.1667	86.72
	Sp	0.9759	0.6768	0.9983	1	
	MCC	0.796	0.742	0.652	0.4	

V. Conclusion & Discussion

LncRNAs can exist in different regions of the cell and show some crucial biological functions that may relate to diseases. Therefore, understanding their subcellular localization becomes an urgent task. However, compared to the vast lncRNA family, people have annotated only very few of them with their subcellular localization. It is possible to annotate the lncRNAs subcellular localization using computational methods based on the existing lncRNA atlas.

The conventional computational methods annotate the lncRNA subcellular localization by extracting q-mer profiles from the lncRNA sequence. Then, they train Machine Learning models with q-mer profiles and get some good results. Given the gene mutation, there may be some changes in the lncRNA sequence, and these changes exhibit various biological functions which can relate to certain diseases. We hypothesized that these changes may affect how we perform subcellular localization.

In this paper, to test this hypothesis, we train a 1D CNN model with q-mer profile. To compare the performances, we try q-mer with various mismatches. The results show an upward trend in overall accuracy when the number of mismatches increased. It turns out that the mismatch on q-mer profile can improve the prediction performance. The proposed approach surpasses the state-of-the-art methods in predicting subcellular localization of lncRNAs.

We acknowledge some potential limitations in this work. First, the datasets used are relatively small. Only hundreds of lncRNAs are contained in these datasets. From a small dataset, it is hard to extract sufficient information to predict new unannotated lncRNAs. Second, the dataset is unbalanced. With unbalanced datasets, a model may perform well at predicting the majority classes while doing poorly in minority classes. More specific attention to this data imbalance problem could improve the results further. Finally, with the potential exponential increase in the feature space as q increases, computational challenges abound, both with respect to time and space. These issues make a case for potential future direction using this idea of inexact q-grams, especially given the improved comparative performance over state-of-the-art.

ACKNOWLEDGMENT

This work was supported in part by grants from the US National Science Foundation (Award #1747788, and #1920920).

REFERENCES

[1] J. Brosius, "The Fragmented Gene," *Annals of the New York Academy of Sciences*, vol. 1178, no. 1, pp. 186–193, Oct. 2009, doi: 10.1111/j.1749-6632.2009.05004.x.

[2] C. Han Li and Y. Chen, "Small and Long Non-Coding RNAs: Novel Targets in Perspective Cancer Therapy," *Current Genomics*, vol. 16, no. 5, pp. 319–326, Oct. 2015.

[3] L. Ma *et al.*, "LncBook: a curated knowledgebase of human long non-coding RNAs," *Nucleic Acids Res*, vol. 47, no. Database issue, pp. D128–D134, Jan. 2019, doi: 10.1093/nar/gky960.

[4] "What is a gene?: MedlinePlus Genetics." https://medlineplus.gov/genetics/understanding/basics/gene/ (accessed Aug. 29, 2021).

[5] A. F. Palazzo and E. S. Lee, "Non-coding RNA: what is functional and what is junk?," *Front Genet*, vol. 6, p. 2, Jan. 2015, doi: 10.3389/fgene.2015.00002.

[6] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-Throughput Sequencing Technologies," *Molecular Cell*, vol. 58, no. 4, pp. 586–597, May 2015, doi: 10.1016/j.molcel.2015.05.004.

[7] J. Zhu, H. Fu, Y. Wu, and X. Zheng, "Function of lncRNAs and approaches to lncRNA-protein interactions," *Sci China Life Sci*, vol. 56, no. 10, pp. 876–885, Oct. 2013, doi: 10.1007/s11427-013-4553-6.

[8] L. Ma, V. B. Bajic, and Z. Zhang, "On the classification of long non-coding RNAs," *RNA Biology*, vol. 10, no. 6, pp. 924–933, Jun. 2013, doi: 10.4161/rna.24604.

[9] Y. Fang and M. J. Fullwood, "Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer," *Genomics, Proteomics & Bioinformatics*, vol. 14, no. 1, pp. 42–54, Feb. 2016, doi: 10.1016/j.gpb.2015.09.006.

[10] N. N. Parikshak *et al.*, "Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism," *Nature*, vol. 540, no. 7633, pp. 423–427, Dec. 2016, doi: 10.1038/nature20612.

[11] Q. Luo and Y. Chen, "Long noncoding RNAs and Alzheimer's disease," *Clin Interv Aging*, vol. 11, pp. 867–872, Jun. 2016, doi: 10.2147/CIA.S107037.

[12] Z. Bao, Z. Yang, Z. Huang, Y. Zhou, Q. Cui, and D. Dong, "LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1034–D1037, Jan. 2019, doi: 10.1093/nar/gky905.

[13] L.-L. Chen, "Linking Long Noncoding RNA Localization and Function," *Trends in Biochemical Sciences*, vol. 41, no. 9, pp. 761–772, Sep. 2016, doi: 10.1016/j.tibs.2016.07.003.

[14] J. Carlevaro-Fita and R. Johnson, "Global Positioning System: Understanding Long Noncoding RNAs through Subcellular Localization," *Molecular Cell*, vol. 73, no. 5, pp. 869–883, Mar. 2019, doi: 10.1016/j.molcel.2019.02.008.

[15] T. Cui *et al.*, "RNALocate v2.0: an updated resource for RNA subcellular localization with increased coverage and annotation," *Nucleic Acids Research*, no. gkab825, Sep. 2021, doi: 10.1093/nar/gkab825.

[16] D. Mas-Ponte, J. Carlevaro-Fita, E. Palumbo, T. H. Pulido, R. Guigo, and R. Johnson, "LncATLAS database for subcellular localization of long noncoding RNAs," *RNA*, vol. 23, no. 7, pp. 1080–1087, Jul. 2017, doi: 10.1261/rna.060814.117.

[17] J. Miller J and D.A. Adjeroh, "Exploring neural network models for LncRNA sequence identification", Proceedings, IEEE International Conference on Bioinformatics and Biomedicine (BIBM'20), South Korea, Dec. 2020.

[18] Z. Cao, X. Pan, Y. Yang, Y. Huang, and H.-B. Shen, "The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier," *Bioinformatics*, vol. 34, no. 13, pp. 2185–2194, Jul. 2018, doi: 10.1093/bioinformatics/bty085.

[19] Z.-D. Su *et al.*, "iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC," *Bioinformatics*, vol. 34, no. 24, pp. 4196–4204, Dec. 2018, doi: 10.1093/bioinformatics/bty508.

[20] B. L. Gudenas and L. Wang, "Prediction of LncRNA Subcellular Localization with Deep Learning from Sequence Features," *Sci Rep*, vol. 8, no. 1, p. 16385, Nov. 2018, doi: 10.1038/s41598-018-34708-w.

[21] R. Karki, D. Pandya, R. C. Elston, and C. Ferlini, "Defining 'mutation' and 'polymorphism' in the era of personal genomics," *BMC Med Genomics*, vol. 8, p. 37, Jul. 2015, doi: 10.1186/s12920-015-0115-z.

[22] S. Clancy, "RNA functions," vol. 1(1), p. 102, 2008.

[23] P. M. Macdonald, "mRNA localization: assembly of transport complexes and their incorporation into particles," *Curr Opin Genet Dev*, vol. 21, no. 4, pp. 407–413, Aug. 2011, doi: 10.1016/j.gde.2011.04.005.

[24] J. Sprenger, J. Lynn Fink, S. Karunaratne, K. Hanson, N. A. Hamilton, and R. D. Teasdale, "LOCATE: a mammalian protein subcellular localization database," *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D230-233, Jan. 2008, doi: 10.1093/nar/gkm950.

[25] J. Li and C. Liu, "Coding or Noncoding, the Converging Concepts of RNAs," *Front Genet*, vol. 10, p. 496, 2019, doi: 10.3389/fgene.2019.00496.

[26] J. Seiler, M. Breinig, M. Caudron-Herger, M. Polycarpou-Schwarz, M. Boutros, and S. Diederichs, "The lncRNA VELUCT strongly regulates viability of lung cancer cells despite its extremely low abundance," *Nucleic Acids Res*, vol. 45, no. 9, pp. 5458–5469, May 2017, doi: 10.1093/nar/gkx076.

[27] P. Dönnes and A. Höglund, "Predicting protein subcellular localization: past, present, and future," *Genomics Proteomics Bioinformatics*, vol. 2, no. 4, pp. 209–215, Nov. 2004, doi: 10.1016/s1672-0229(04)02027-3.

[28] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, and K.-C. Chou, "PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition," *Analytical Biochemistry*, vol. 456, pp. 53–60, Jul. 2014, doi: 10.1016/j.ab.2014.04.001.

[29] J. M. Kirk *et al.*, "Functional classification of long non-coding RNAs by k-mer content," *Nat Genet*, vol. 50, no. 10, pp. 1474–1482, Oct. 2018, doi: 10.1038/s41588-018-0207-8.

[30] T. Zhang *et al.*, "RNALocate: a resource for RNA subcellular localizations," *Nucleic Acids Research*, vol. 45, no. D1, pp. D135–D138, Jan. 2017, doi: 10.1093/nar/gkw728.

[31] Y. Fan, M. Chen, and Q. Zhu, "lncLocPred: Predicting LncRNA Subcellular Localization Using Multiple Sequence Feature Information," *IEEE Access*, vol. 8, pp. 124702–124711, 2020, doi: 10.1109/ACCESS.2020.3007317.

[32] C. Leslie, J. Weston, E. Eskin, and W. S. Noble, "Mismatch String Kernels for SVM Protein Classification," p. 8.

[33] C. Leslie, E. Eskin, and W. S. Noble, "THE SPECTRUM KERNEL: A STRING KERNEL FOR SVM PROTEIN CLASSIFICATION," in *Biocomputing 2002*, Kauai, Hawaii, USA, Dec. 2001, pp. 564–575. doi: 10.1142/9789812799623_0053.

[34] D. Adjeroh, T. Bell, and A. Mukherjee, *The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching*. 2008, p. 351. doi: 10.1007/978-0-387-78909-5.

[35] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[36] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[37] T. O'Malley *et al.*, "KerasTuner." 2019. [Online]. Available: https://github.com/keras-team/keras-tuner

[38] S. Scardapane and D. Wang, "Randomness in neural networks: an overview," *WIREs Data Mining and Knowledge Discovery*, vol. 7, no. 2, p. e1200, 2017, doi: 10.1002/widm.1200.