2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385263

LncRNA Subcellular Localization Signals – Are the Two Ends Equal? A Machine Learning Analysis Across Multiple Cell Lines

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

Jason R Miller
Computer Science, Mathematics, & Engineering
Shepherd University
Shepherdstown WV, USA
jrm0122@mix.wvu.edu

Gangqing Hu
Microbiology, Immunology, and Cell Biology
WVU School of Medicine
Morgantown, WV, USA
michael.hu@hsc.wvu.edu

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

***Abstract*— *In this work, we studied the question of whether the two ends of long non-coding ribonucleic acids (lncRNAs) (i.e., the 5' end and 3'end) carry similar information about subcellular localization of lncRNAs. We considered this problem from three viewpoints using machine learning models:* (1) *consideration of the classification performance of the machine learning models using features from defined regions (or segments) along the sequence,* (2) *correlation-based analysis using models built on regions/segments along the LncRNA sequence, and* (3) *analysis of the relative positions of predicted lncRNA localization motifs along the LncRNA sequence. Our results and observations suggest that the 5' region of the lncRNA sequences (the prefixes) tend to carry more localization signals when compared with the 3' region (the suffixes) of the sequences. These could have implications on how we use machine learning models for improved analysis of lncRNA subcellular localization.*

Keywords: lncRNA, localization, machine learning, motifs

I. INTRODUCTION

Most long non-coding RNA (lncRNA) genes are poorly characterized. Some clues to their cellular roles have been provided by experiments that measure whether the lncRNA transcripts are more abundant in the nucleus or cytoplasm. All lncRNA is formed in the nucleus, but much of it is exported through the nuclear membrane into the cytoplasm where, presumably, it plays some role. Subcellular localization of lncRNA is influenced by many factors, as reviewed in [1]. In many cases, localization is correlated to specific k-mers or motifs, i.e. substrings that exactly or inexactly match some pattern, found within the lncRNAs themselves. This observation has inspired many machine learning experiments designed to train models that predict localizations given the lncRNA sequences alone.

In protein-coding RNA, also called messenger RNA (mRNA), the center portion of the RNA sequence is evolutionarily constrained to list the 3-nucleotide codons that get translated into proteins. In mRNA, the initial and trailing regions are called the 5' UTR and 3' UTR, respectively. The UTR portions of mRNA often contain information beyond protein sequence, such as binding motifs and localization signals. Since lncRNA shares many characteristics with mRNA, it has been suggested that localization signals may be concentrated at the 5' end or 3' end of lncRNA [2].

In this study, we trained machine learning models to predict either nuclear or cytoplasmic subcellular localization of lncRNAs given a portion of their sequences. We compared the accuracy of models that operated only on just the prefixes, or just the suffixes, or the entire lengths of lncRNA sequences. More specifically, we performed our analysis with four machine learning models, namely, Multilayer Perceptron (MLP), Random Forest (RF), eXtreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (LGBM), first using the entire sequence (all_Seq), and then using segments from different regions of the lncRNA sequence, moving from the 5' end to the 3' end. That is, from prefix segment (5' end), second segment, middle segment, fourth segment, and suffix segment (3' end). Using all k-mer features from the sequence (all_Seq) with RF produced the highest classification accuracy of 60.45%, using 4-mers with 0 mismatch. We then evaluated the prefix, second, middle, fourth segments, and the suffix on different length segments. Using segment length 1024, with 4-mer with 0 mismatch on the prefix segment resulted in the highest segment-based results, in terms of classification accuracy. The classification accuracy seemed to decrease from the prefix (5' end) towards the suffix (3' end). The results seem to show that the prefix region (5' end) of the lncRNA sequences tended to contain more signals that are relevant to localization, when compared with the other regions, including the suffix region (3' end). This observation is further studied by performing correlation-based analysis using segments along the lncRNA sequence, and also by analyzing the relative position of ranked

localization motifs along the lncRNA sequence. A learning-based fusion using extracted prediction probabilities from machine models developed using 3 segment lengths, and 5 region segments as input to the random forest model achieved a slightly improved classification accuracy of 61.89%.

II. Background

Nearly all protein-coding RNA must localize to the cytoplasm to be functional. That is because the cytoplasm is the sole site of translation from mRNA to protein. In contrast, the nucleus is the sole site of intron splicing, and incomplete splicing is associated with nuclear retention. One study [3] looked at a splice signal called 5'SS. This signal is commonly located at 5' ends and is commonly removed by splicing. The authors found that 5'SS in the 3' ends of mRNAs does not get removed by splicing and is associated with nuclear retention. The effect was less strong, but still present, in lncRNA.

A 2022 study [4]measured the impact of 5'SS and two other sequence motifs known to favor nuclear retention in lncRNA: BORG, and SIRLOIN. These motifs are recognizable computationally, and presumably biochemically, by their nucleotide sequence and its implied molecular structure. The study examined the effect of these motifs on lncRNA in a particular human cell line, known as BSX, which is cultured from retinal pigment epithelium. RNA was isolated from the nucleus and from the cytoplasm of these cells under two conditions: normal, and oxidative stress. The RNA was sequenced, aligned to the human reference transcriptome, and quantified to give each transcript's cytoplasmic-over-nuclear log2 fold change. For most lncRNA, the value was near zero, indicating similar abundance in the nucleus and cytoplasm. This was true under the normal and stress conditions, but more transcripts moved to the extremes when the cells were under stress. This indicates that environmental factors influence localization. When the authors compared transcripts that do or do not harbor one of the three known nuclear localization motifs, they found that the nuclear fraction was higher for lncRNAs that contained the motifs than for those that did not. This was true under the normal and the stressed conditions. This confirms that the known factors encourage nuclear retention. Next, the authors found an effect due to multiplicity. The fold change was inversely correlated with copy count, meaning that more repeats of the signal correlated with heavier nuclear retention.

A 2019 study [5] built models to predict nuclear or cytoplasmic enrichment per RNA. The models used a combination of lncRNA features including sequence elements. The models were mildly successful, explaining 45% of the variance for mRNA genes and 34% of the variance for lncRNA genes. However, the most predictive factor was splicing efficiency, i.e. the portion of the transcripts whose introns were removed completely. This factor is measurable in cells but it may not be easily predictable from sequence alone. For this reason, one can expect low levels of accuracy in machine learning classifiers trained on the sequences alone. A 2015 study found that many lncRNAs are translated to protein in the cytoplasm [6]. This finding is counter-intuitive since lncRNA is, by definition, non-coding. However, biology is rarely binary, and the definition of lncRNA should perhaps be modified to indicate that they do not encode functional proteins. The protein products of translated lncRNA may be unstable or inactive, but in some cases they bear resemblance to functional proteins in the same or other species. Furthermore, the authors argue that some translated lncRNAs bear signs of purifying evolutionary selection, indicating that these lncRNAs contribute to overall fitness in some way. Protein translation occurs only in the cytoplasm. The authors found that localization preference for the cytoplasm correlates with lncRNA efficiency at translation to protein. This indicates that preference for the cytoplasm is distributed non-randomly among lncRNAs.

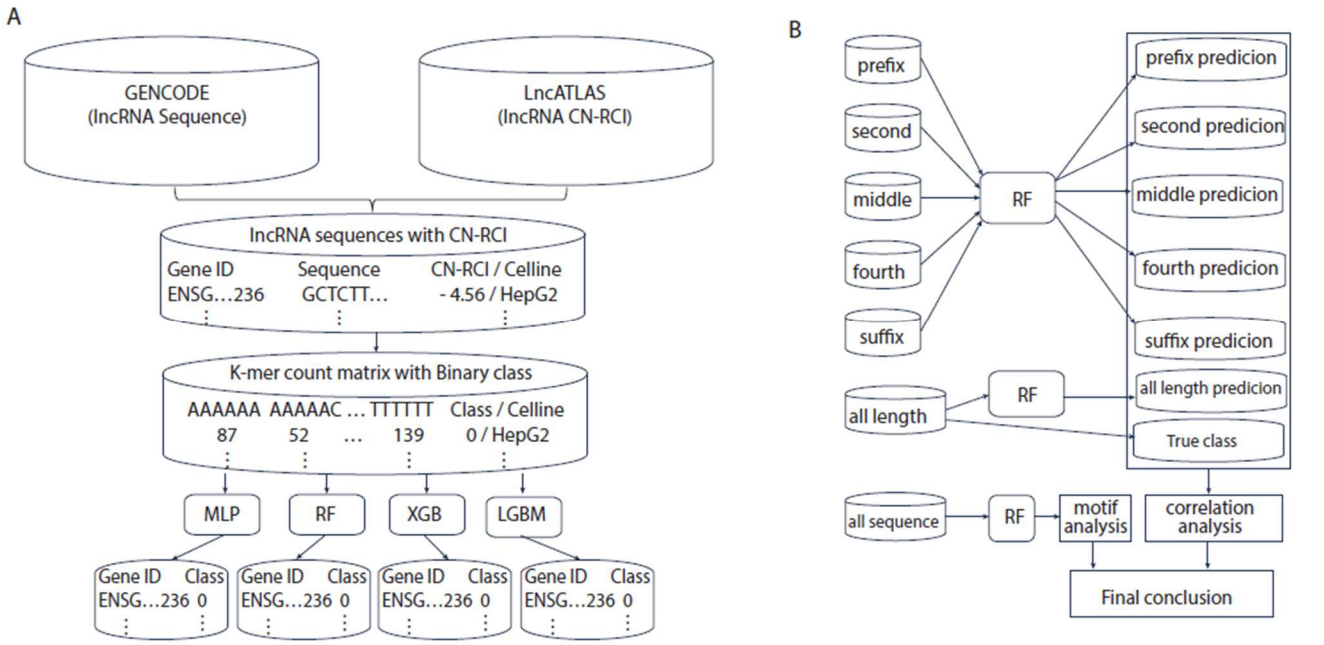

Figure 1. Workflow for the reported analysis. (A) classification using k-mers; (B) analysis using correlation, and motif location.

In a 2022 study, Ron and Ulitsky [7] measured the nuclear vs cytoplasmic enrichment of certain sequences associated with either compartment. Though most of the findings concerned another type of RNA (circRNA), the authors reported some associations between lncRNA sequence and localization. For example, the authors found that nuclear localization was associated with many short sequences taken from the MALAT1 lncRNA, which is known to have nuclear function.

The foregoing suggests that there could be some value in further studying lncRNA sequences, especially on the question of whether the 5' and 3' ends of a lncRNA could hold important information that may be relevant to localization. Clearly, an improved understanding of the role of these lncRNA ends in localization could be exploited to build improved computational methods for predicting subcellular localization for lncRNAs, based primarily on information from their sequences.

III. METHODS

Figure 1 shows the basic workflow in our analysis. Panel A shows our baseline model. We build our datasets first, and then create baselines of the machine learning models. Panel B shows the workflow of following analysis. Using the baseline models, we identify a suitable machine learning model to be used to analyze the different regions of the lncRNA, using varying segment lengths on the sequence. We performed lncRNA localization classification using the models. Further, based on the predicted class and the true classes of the lncRNA sequences, we performed correlation analysis, using results from the different regions of the lncRNA sequence. Finally, we performed further analysis by considering the relative positions of predicted k-mer localization sequence motifs, as obtained from the best performing machine learning model.

A. Dataset

We downloaded lncRNA subcellular localization data from lncATLAS [8]. The lncATLAS considered two localizations in the cells: nuclear and the cytoplasmic, which is the fluid inside a cell but outside the cell's nucleus. They computed the cytoplasmic-nuclear relative concentration index (CN-RCI) values for the genes from 15 cell lines. We obtained the gene ensemble ID and a CN-RCI value for each gene from the dataset. We then download lncRNA sequence data and annotation file (.gff file) from GENCODE (https://www.gencodegenes.org/human/) release 42. The GENCODE dataset includes repeated genes with duplicate transcripts. We combined the two datasets according to their gene ensemble ID and retained transcripts that had at least one CN-RCI value for the cell lines. We used the longest transcript for a given gene, and removed the duplicated genes. We separate the dataset into training and test sets with a ratio of 4:1. We obtained 4662 genes for the training set. The length of the lncRNA transcripts varied from 72 to 205,000. To reduce the computation workload, we only analyzed genes with sequence lengths between 200 and 5000 in the training set. Thus, we obtained 4607 genes (transcripts) for the training set.

We used segments of lengths 256, 512, and 1024 respectively, from the sequences to check whether the prefixes (5' ends), and suffixes (3'ends) contain similar amounts of localization information. To make results from the different lengths more comparable, we performed our analysis only on sequences with lengths greater than or equal to 1027. This reduced our data to 2834 sequences in total. We extract subsequences with lengths 256, 512, and 1024 respectively, from 5 positions in each given lncRNA sequence. These 5 positions thus defined 5 segments, namely, the prefix segment, the second segment, the middle, the fourth segment, and the suffix segment. See Figure 2.

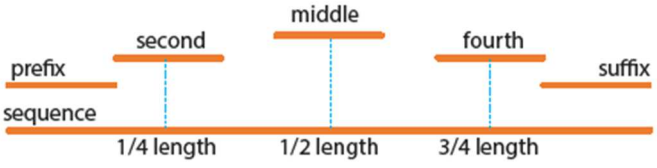

Figure 2. The 5 segments in the sequence.

Not all the sequences (lncRNAs) have CN-RCI values in every cell line. For each cell line, we will drop the sequences that do not have a value in the cell line. In this work, we considered 3 cell lines, A549, MCF.7, and SK.N.SH for our first two analyses (classification, and correlation-based analysis). We set the class threshold to 0. Thus, for a lncRNA with a CN-RCI value greater or equal to 0, we set the class label to 1 (i.e., more likely to be cytoplasmic). Otherwise, we set the class label to 0 (i.e., more likely to be nuclear). For each cell line, we perform down-sampling to build a balanced dataset, with equal numbers for the two label categories. Thus, we obtain balanced datasets having A549 with 968 genes, MCF.7 with 796 genes, and SK.N.SH with 734 genes.

B. Feature representation

LncRNA is transcribed from DNA. LncRNA consists of a string of nucleotide bases, namely, adenine (A), guanine (G), uracil (U), and cytosine (C). For our feature space, we use only sequence-based features, using simple k-mers (k-length substrings) from the sequence. We used k=4, and k=6, and used the k-mer counts as our feature representation for each lncRNA sequence. The k-mer counts are easy to compute, though more efficient algorithms are available using suffix trees and suffix arrays [13]. Using the k-mer counts, we generate the k-mer profile for a given sequence or segment. The profiles are normalized by the sequence length, or by the segment length as needed to capture the probability of occurrence of each given k-mer in the sequence, or segment.

C. Machine Learning Models

In this work, we studied lncRNA subcellular localization as a classification problem, using four machine learning models, namely, the random forest (RF) classifier from Scikit [9], XGBoost [10], light gradient boosting machine classifier (LGBM) [11], and multilayer perceptron classifier (MLP). For the first three models, we use the default parameters. We build 3 layers (64,32,2) MLP model with TensorFlow [12].

Our baseline models are built on the k-mer profiles obtained using the full-length sequence (all_len), normalized by gene's sequence length. We than evaluated the respective classification performance on the 5 different regions of the sequence, k-mer

profiles normalized by region length. For each model, we analyzed 3 cell lines, namely, A549, MCF.7, and SK.N.SH. We used 5-fold cross validation on the balanced training sets. Based on the predicted localization labels, we performed class correlation analysis. We then performed motif location analysis, by considering the distribution of motifs along the sequence, using the relative position of predicted lncRNA motifs along the sequence.

D. Evaluation

In this work, we performed our classification study using a balanced dataset, obtained via under-sampling. We performed 5-fold cross-validation twice on this data, and recorded the average results from the two runs. We report average of the classification accuracy over the 3 cell lines.

IV. Results

A. Classification performance

Table 1 and Table 2 show the classification results for the baseline models (using k-mer profiles from all the sequence), using the four machine learning models (MLP, RF, XGB, and LGBM). We have included results for exact k-mers (_0miss) and inexact k-mers (as introduced in [14]) with k-mismatches (_1miss, _2miss, 3_miss). The results show that the RF model with default parameters resulted in the highest accuracy of 60.45% on 4-mer with 0 mismatch, and 60.25% on 6-mer with 1 mismatch. In subsequent analysis, we will focus random forest model (RF) with 4mer with 0 mismatch and 6mer with 1 mismatch. In subsequent analysis, we will focus on the random forest (RF) model, using 4mers with 0 mismatch, and 6mers with 1 mismatch.

We tested the RF model on segments from different regions of the lncRNA sequences, using different segment lengths. Results are shown in Tables 3 and 4. The highest score was 59.3%, using prefixes and segment length 1024. Expectedly, the results show that the classification accuracy increases with increasing segment length -- from 58.73% with length 256 to 59.3% with length 1024. This tendency of improved classification accuracy with increasing segment length was also generally observed for each of the five regional segments. Considering the 5 segments (or regions) at a given sequence length, the accuracy was observed to generally decrease from the prefix (5' end) to the suffix (3' end). This indicates that localization signals are stronger or more likely to be found in the prefix (5' end) of the lncRNA sequence, when compared with the suffix (3' end). This was a surprising observation, thus we set out to perform further analysis (see below).

We also further tested the performance using a learning-based fusion scheme. We extracted the predicted probabilities from the 15 models (i.e., 3 segment lengths, 5 regions), and used these as the input features to the machine learning models. Once again, the random forest (RF) model produced the best results, giving an accuracy of 61.89% on 4mers, 0 mismatch, and 60.67% on 6mers, 0 mismatch. See results in Table 5.

Table 1. Classification results using the baseline models on 4-mers.

	MLP_0 miss	MLP_1 miss	MLP_2 miss	RF_0 miss	RF_1 miss	RF_2 miss	XGB_0 miss	XGB_1 miss	XGB_2 miss	LGBM_0 miss	LGBM_1 miss	LGBM_2 miss
A549	58.52	60.22	58.36	63.27	62.44	59.5	61.77	61.36	60.12	62.09	62.55	60.32
MCF.7	55.09	56.72	57.29	57.92	57.85	55.02	55.9	58.04	53.2	57.09	58.42	54.96
SK.N.SH	56.27	56.13	54.22	60.15	57.77	55.52	58.18	56.27	56.13	57.29	57.97	56.47
mean	56.63	57.69	56.62	60.45	59.35	56.68	58.62	58.56	56.48	58.82	59.65	57.25

Table 2. Classification results using the baseline models on 6-mers.

	MLP_0 miss	MLP_1 miss	MLP_2 miss	MLP_3 miss	RF_0 miss	RF_1 miss	RF_2 miss	RF_3 miss	XGB_0 miss	XGB_1 miss	XGB_2 miss	XGB_3 miss	LGBM_0 miss	LGBM_1 miss	LGBM_2 miss	LGBM_3 miss
A549	59.55	60.79	60.63	61.46	60.48	63.58	62.44	62.34	59.14	61.36	61.67	60.64	60.53	64	62.29	62.34
MCF.7	55.59	57.29	57.92	56.91	58.11	58.23	59.11	55.77	56.97	57.85	56.28	54.52	59.04	56.78	58.73	56.72
SK.N.SH	57.22	57.91	58.38	58.86	56.81	58.93	57.29	56.74	58.59	57.02	54.09	58.04	58.04	57.29	55.59	56.47
mean	57.45	58.66	58.98	59.08	58.47	60.25	59.61	58.28	58.23	58.74	57.35	57.73	59.2	59.36	58.87	58.51

Table 3. Classification performance using 4mer with 0 mismatch on segments from different sequence regions.

Segment length	Prefix	Second	Middle	Fourth	Suffix
256	58.73	56.34	55.57	53.92	52.82
512	58.53	57.5	56.7	56.29	55.39
1024	59.3	58.33	59.13	57.66	55.16

Table4. Classification performance using 6mer with 1 mismatch on segments from different sequence regions..

Segment length	Prefix	Second	Middle	Fourth	Suffix
256	57.85	53.7	55.44	52.58	53.45
512	55.94	55.95	56.32	55.62	55.04
1024	58.17	57.47	58.31	56.91	55.98

Table5. Classification performance using learning-based fusion.

15 features	4mer0miss	4mer1miss	6mer0miss	6mer1miss	6mer2miss
RF	61.89	61.28	60.67	59.7	60.62

Table 6. Correlation analysis using 4-mers. The correlation coefficient between true class and different model predicted classes.

	Class	all_len	256_pfx	256_second	256_mid	256_fourth	256_sfx	512_pfx	512_second	512_mid	512_fourth	512_sfx	1024_pfx	1024_second	1024_mid	1024_fourth	1024_sfx
Class	1	0.209	0.175	0.127	0.112	0.078	0.056	0.171	0.151	0.134	0.126	0.108	0.186	0.176	0.183	0.154	0.103
all_len	0.209	1	0.247	0.273	0.259	0.232	0.125	0.309	0.379	0.371	0.303	0.255	0.423	0.281	0.453	0.372	0.371
256_pfx	0.175	0.247	1	0.123	0.027	0.014	0.03	0.474	0.165	0.053	0.046	0.032	0.347	0.091	0.112	0.039	0.076
256_second	0.127	0.273	0.123	1	0.145	0.112	0.078	0.17	0.382	0.203	0.146	0.092	0.268	0.232	0.243	0.152	0.155
256_mid	0.112	0.259	0.027	0.145	1	0.166	0.074	0.062	0.176	0.43	0.21	0.139	0.168	0.127	0.346	0.218	0.23
256_fourth	0.078	0.232	0.014	0.112	0.166	1	0.102	0.054	0.125	0.207	0.377	0.187	0.091	0.09	0.244	0.307	0.288
256_sfx	0.056	0.125	0.03	0.078	0.074	0.102	1	0.05	0.078	0.101	0.147	0.252	0.067	0.055	0.118	0.161	0.198
512_pfx	0.171	0.309	0.474	0.17	0.062	0.054	0.05	1	0.234	0.097	0.068	0.065	0.448	0.135	0.139	0.098	0.113
512_second	0.151	0.379	0.165	0.382	0.176	0.125	0.078	0.234	1	0.241	0.16	0.107	0.367	0.328	0.314	0.2	0.165
512_mid	0.134	0.371	0.053	0.203	0.43	0.207	0.101	0.097	0.241	1	0.26	0.196	0.225	0.209	0.475	0.306	0.288
512_fourth	0.126	0.303	0.046	0.146	0.21	0.377	0.147	0.068	0.16	0.26	1	0.282	0.157	0.127	0.282	0.418	0.372
512_sfx	0.108	0.255	0.032	0.092	0.139	0.187	0.252	0.065	0.107	0.196	0.282	1	0.108	0.09	0.228	0.33	0.352
1024_pfx	0.186	0.423	0.347	0.268	0.168	0.091	0.067	0.448	0.367	0.225	0.157	0.108	1	0.25	0.307	0.179	0.182
1024_second	0.176	0.281	0.091	0.232	0.127	0.09	0.055	0.135	0.328	0.209	0.127	0.09	0.25	1	0.252	0.215	0.125
1024_mid	0.183	0.453	0.112	0.243	0.346	0.244	0.118	0.139	0.314	0.475	0.282	0.228	0.307	0.252	1	0.343	0.323
1024_fourth	0.154	0.372	0.039	0.152	0.218	0.307	0.161	0.098	0.2	0.306	0.418	0.33	0.179	0.215	0.343	1	0.447
1024_sfx	0.103	0.371	0.076	0.155	0.23	0.288	0.198	0.113	0.165	0.288	0.372	0.352	0.182	0.125	0.323	0.447	1

B. Correlation analysis

Table 6 shows the correlation coefficients between the classification labels produced using the machine learning model (RF in this case) on the different regions/segments (using 4-mers), and the true class labels. The results include correlation results between regions using the same lengths, and also different regions with different lengths. The correlation analysis between true class and the predicted classes from different models seems to suggest the same observations as the direct classification results, namely, (1) the prefix, second, and middle segments have relatively higher correlation with the true class, compared with fourth, and suffix. For length 1024 segments, the prefix, second prefix, and middle had correlation coefficient of 0.186, 0.176, and 0.183 respectively, which then decreases to 0.154 and 0.103 for the fourth segment, and the suffix, respectively. Similar observations were made for segment lengths 256 and 512. (2) The longer the segment length, the higher the correlation with the true class. We also can observe that, as expected, nearby regions had more correlation, when compared with distal regions.

C. Motif based analysis

To identify the key motifs that may be important in subcellular localization, we utilized the RF classifier to generate feature ranks for each cell line. For this analysis, we used 14 cell lines in total, namely, A549, MCF.7, HT1080, GM12878, SK.MEL.5, HeLa.S3, SK.N.DZ, SK.N.SH. IMR.90, K562, HepG2, HUVEC, NCI.H460. and NHEK. (We did not use H1.hESC, the stem-cell cell line, as it appeared to have very different characteristics when compared with the others). Specifically, we focused on 6-mer with 1 mismatch, and generated the RF feature importance ranking for inexact k-mers. For each cell line, we identified the top 200 k-mers (the motifs). The results showed there were 644 motifs shared between at least two cell lines. Some top-ranked motifs in one cell line were also observed in the top-ranked motifs for other cell lines, while certain top-ranked motifs tended to appear in few cell lines.

We then checked the CN-RCI distribution per cell line that linked to the top ranked motifs. For each cell line, we divided the genes into two groups: **genes with the motif**, or **genes without the motif**, and then computed the average RCI value for the genes within the two groups. Finally, we compute the average of differences across all the cell lines. See Tables 7. For example, the motif "*AATAAA"* appeared in the top 200 motifs for 9 cell lines, as listed in Table 7. In cell line A549, the average RCI value from genes with the motif is -0.315, without the motif is -0.677. The difference between the average for those **with** and **without** the motif is 0.362. Similarly, we compute the difference for all the other cell lines. Finally, for this example motif, we obtain the mean difference across all the 9 cell lines where it appeared as 0.317. See Table 7.

We set our localization thresholds as ±0.3. Thus, motifs with mean difference greater than or equal to 0.3 are deemed more likely to be located in the cytoplasm. The analysis above identified 157 such "cytoplasmic motifs". Similarly, motifs with mean difference less than -0.3, are deemed more likely to be in the nucleus. We identified 45 such "nuclear-leaning motifs". We then analyzed the relative position(s) of a given motif along the lncRNA sequences where the motif occurred. We first find the

Table 7. The average of RCI value with and without a given motif, and the mean difference across the cell lines. Example using the motif "AATAAA".

AATAAA	A549	MCF.7	HT1080	NHEK	GM12878	HeLa.S3	SK.N.SH	K562	HepG2	Mean_diff
With	-0.315	-1.430	-0.349	-0.993	-1.018	-1.334	-1.160	-0.799	-1.185	-0.954
Without	-0.677	-1.519	-0.710	-1.330	-1.145	-1.968	-1.499	-1.067	-1.524	-1.271
Difference	0.362	0.089	0.361	0.337	0.128	0.633	0.338	0.268	0.339	0.317

position of the motif in the sequence, say p_i, and then obtain the relative position by dividing p_i with sequence length n. This normalizes the position to a range of 0 to 1, which then allows us to compare relative positions across sequence with different lengths. We performed the analysis for three groups of motifs, namely, cytoplasmic motifs (157 motifs), nuclear-leaning motifs (45 motifs), and both types of motifs combined (202 motifs). We used our training set of 3913 genes, (H1.hESC removed), where the motifs occurred. For each gene, we computed the relative positions of occurrence for each of the 202 motifs (where one occurred). We then generate the distribution of the relative positions of the 202 motifs across all genes and all cell lines where they occurred. The results are shown in Figure 3.

The results show that the nuclear-leaning motifs (with mean difference < -0.3) tend to appear more in the first half of the sequence. The density goes up sharply and then goes down from prefix (5' end) to the middle and then all the way to the suffix (3' end). The localization signals seem to be located more towards the prefix region (5' end), which is consistent with our previous observation, based on classification performance and correlation analysis. However, for the cytoplasmic motifs, there was a significant surge in the motif density as we come closer to the lncRNA ends – both the prefix (5' region) and suffix (3' region). This seems to suggest that, for lncRNAs that localize in the cytoplasm, the suffix region (3' end) could also hold significant localization signals, perhaps even more so than the middle region. This observation may imply possibly significant differences between nuclear-leaning and cytoplasmic lncRNAs. This certainly requires further analysis using more data, and perhaps via some biological experimental studies. Using the whole set of motifs, the trend also aligned with our earlier observations, that is, generally more signals (or more localization motifs in this case) as we move from the 5' region to the 3' region.

V. Discussion and Conclusion

Using the analysis on top-ranked motifs from random-forest feature importance ranking, we identified 202 motifs, 157 cytoplasmic-leaning, and 45 nuclear-leaning. The analysis of the relative location of these motifs along the sequence seem to support the observation that the two ends of the lncRNA sequence may not be contributing equally in terms of the strength or amount of subcellular localization signals they may contain. Lubelsky and Ulitsky [7] identified 49 tiles that contain the SINE-derived nuclear-RNA-localization element (SIRLOIN) best matching sequences in NucLibA library and 81 tiles in NucLibB library, which are associated with nuclear enrichment. By querying the motifs in these SIRLOIN regions, 31 of the 45 nuclear motifs and 55 of the 157 cytoplasmic motifs were found in genes that localize in the nuclear region.

Overall, our classification results using machine learning models on the segments reveal that the prefix (5' end) of the sequence tends to contain more localization signals than the suffix (3' end). The correlation analysis and further motif positional analysis provide further support to this observation. The case for cytoplasmic genes is not clear cut, and requires further analysis. If the results from our computational studies can be verified in the wet lab, this could have significant implications in the analysis of lncRNA sequences.

Towards such improved analysis, we made an initial attempt to use our observations from this work to develop improved machine learning models for lncRNA subcellular localization. For instance, using the above identified 45 nuclear-leaning and 157 cytoplasmic-leaning motifs based on 6mers with 1 mismatch, we developed a RF models using the 5 regional segments, and 3 sequence lengths. We obtained the highest classification rate of 62.2% using just the prefix segments, with segment length 1024. Further, we applied learning-based score fusion using the predicted classification probabilities from the 15 models. Essentially, here, we used a machine-learning model to automatically assign appropriate weights to the results from different regional segments at a

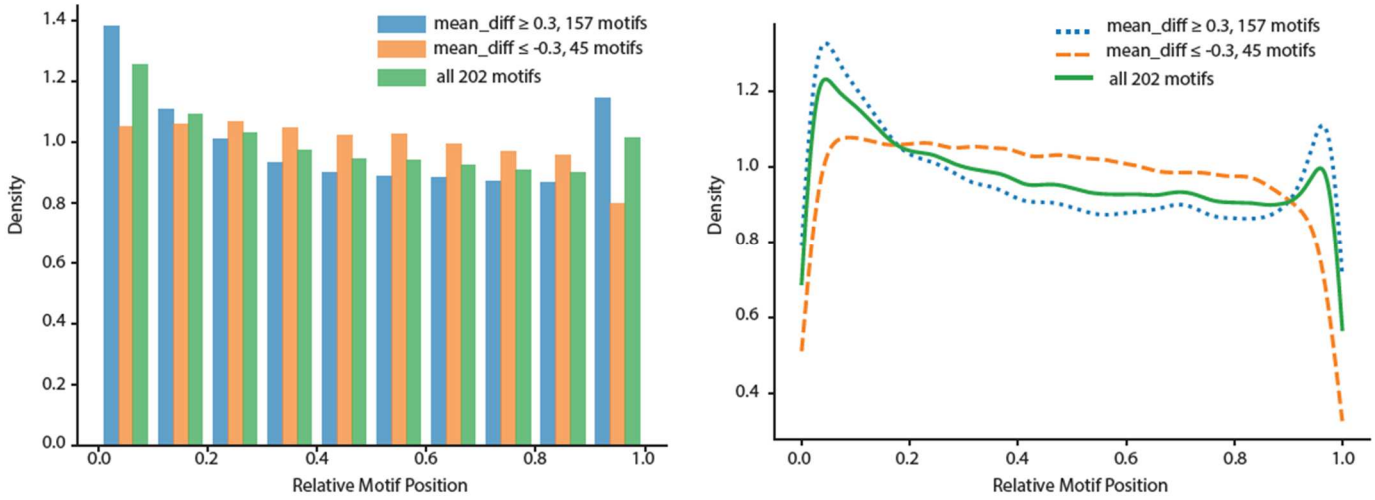

Figure 3. Distribution of relative positions for three groups of localization motifs. Results are shown for all 202 motifs, cytoplasmic motifs (with mean CN-RCI difference >= 0.3), and nuclear-leaning motifs (with mean CN-RCI difference < -0.3).

given segment length, and combined these for the final sub-cellular localization decision. This produced the best overall average classification accuracy of 61.89%, using 4mers with 0 mismatch (Table 5). These results point to the potential to build on the reported observations in this work to further improve the localization models using more sophisticated machine learning techniques, such as deep learning models and architectures.

We can point out some limitations of the current study. The segment-based analysis is essentially performance based. The ideal scenario would be to simply use segment windows with 1 position overlaps across the length of the entire lncRNA sequence, rather than just the prefix, middle, and suffix. Though this could be very computationally intensive, it will provide a clearer picture on whether localization signals have any association with certain regions in the sequence. Further, the motif position analysis was based mainly on predicted motifs from a machine learning model. It should be possible to perform a similar analysis using already characterized and biologically validated lncRNA localization motifs.

ACKNOWLEDGMENT

This work is supported in part by the US National Science Foundation (NSF), Award #s: 1747788, 1920920, 2125872.

REFERENCES

[1] C.-J. Guo, G. Xu, and L.-L. Chen, "Mechanisms of Long Noncoding RNA Nuclear Retention," *Trends Biochem Sci*, vol. 45, no. 11, pp. 947–960, Nov. 2020, doi: 10.1016/j.tibs.2020.07.001.

[2] A. F. Palazzo and E. S. Lee, "Sequence Determinants for Nuclear Retention and Cytoplasmic Export of mRNAs and lncRNAs," *Frontiers in Genetics*, vol. 9, 2018, Available: https://www.frontiersin.org/articles/10.3389/fgene.2018.00440

[3] E. S. Lee, A. Akef, K. Mahadevan, and A. F. Palazzo, "The consensus 5' splice site motif inhibits mRNA nuclear export," *PLoS One*, vol. 10, no. 3, p. e0122743, 2015, doi: 10.1371/journal.pone.0122743.

[4] T. J. Kaczynski, E. D. Au, and M. H. Farkas, "Exploring the lncRNA localization landscape within the retinal pigment epithelium under normal and stress conditions," *BMC Genomics*, vol. 23, no. 1, p. 539, Jul. 2022, doi: 10.1186/s12864-022-08777-1.

[5] B. Zuckerman and I. Ulitsky, "Predictive models of subcellular localization of long RNAs," *RNA*, vol. 25, no. 5, pp. 557–572, May 2019, doi: 10.1261/rna.068288.118.

[6] Z. Ji, R. Song, A. Regev, and K. Struhl, "Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins," *Elife*, vol. 4, p. e08890, Dec. 2015, doi: 10.7554/eLife.08890.

[7] M. Ron and I. Ulitsky, "Context-specific effects of sequence elements on subcellular localization of linear and circular RNAs," *Nat Commun*, vol. 13, no. 1, p. 2481, May 2022, doi: 10.1038/s41467-022-30183-0.

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

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

[10] "XGBoost Documentation — xgboost 2.0.1 documentation." Accessed: Nov. 03, 2023. [Online]. https://xgboost.readthedocs.io/en/stable/index.html

[11] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems".

[13] Adjeroh, Donald, Timothy Bell, and Amar Mukherjee. The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching. Springer Science & Business Media, 2008.

[14] W. Yi, and D. A. Adjeroh. "A deep learning approach to lncRNA subcellular localization using inexact q-mers," In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2128-2133. IEEE, 2021.