СРАВНИТЕЛЬНЫЙ АНАЛИЗ ОСОБЕННОСТЕЙ ОРГАНИЗАЦИИ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ МРНК И ДЛИННЫХ НЕКОДИРУЮЩИХ РНК

-

Волкова О.А.[1], *Кондрахин Ю.В.*[2,3], *Шарипов Р.Н.* [2,3]

[1] Федеральный исследовательский центр Институт цитологии и генетики СО РАН, Новосибирск
[2] Институт вычислительных Технологий СО РАН, Новосибирск
[3] ООО BIOSOFT.RU, Новосибирск

В настоящее время много внимания уделяется длинным некодирующим РНК (лнкРНК). Хотя лнкРНК были изначально классифицированы как некодирующие, применение технологии профилирования рибосом (Ribo-seq) позволило обнаружить транслируемые с некоторых из них функциональные пептиды. Технология Ribo-seq является мощным инструментом для экспериментального определения эффективности стадий инициации и элонгации трансляции РНК в клетке (с помощью харрингтонина, лактимидомицина и циклогексимида, соответственно). Мы разработали новую модель позиционной весовой матрицы для предсказания старта на основе данных Ribo-seq. Эта модель позволяет различать мРНК и лнкРНК человека с точностью 96%. С помощью использования данной модели для предсказания открытых рамок считывания (ОРС) в РНК, было обнаружено, что почти все лнкРНК содержат только короткие ОРС (≤300nt), чего не было обнаружено для мРНК.

Ключевые слова: мРНК; длинные некодирующие РНК; короткие ОРС; дискриминантный анализ; позиционные весовые матрицы.

COMPARATIVE ANALYSIS OF mRNA AND lncRNA SEQUENCE FEATURES

Volkova O.A.[1], *Kondrakhin Yu.V.*[2][3], *Sharipov R.N.*[2][3]

[1] The Federal Research Center Institute of Cytology and Genetics, SB RAS, Novosibirsk
[2] Institute of Computational Technologies, SB RAS, Novosibirsk
[3] BIOSOFT.RU, Ltd, Novosibirsk,

Nowadays much attention is paid to long non-coding RNA (lncRNAs). Although lncRNAs have been classified as non-coding, some functional peptides were detected using ribosome profiling technology (Ribo-seq). The Ribo-seq technology is a powerful experimental tool to characterize RNA translation in a cell with a focus on initiation (harringtonin, lactimidomycin) and elongation (cycloheximide). By exploiting Ribo-seq data, we developed a novel position weight matrix model for prediction of translation starts. This model shows 96 % accuracy of discrimination between human mRNAs and lncRNAs. When the same model was used to predict putative ORFs in RNAs, almost all lncRNAs were found to contain only small ORFs (≤300nt), which was not the case with mRNAs.

Keywords: human mRNAs; human lncRNAs; small ORFs; discriminant analysis; position weight matrix approach; IPS matrix algorithm.

1. Background

Transcription is one of important stages of realization of genetic information in cells leading to generation of rather wide spectrum of RNA types, including messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs). LncRNAs have been originally defined as non-translated RNA molecules longer than 200 nt[1] to distinguish this non-coding RNA type from other types

In different studies, functionality of lncRNAs has been explained by participation in regulation of transcription, translation, mRNA stability, miRNA generation, etc.[3] According to several experimental studies, some lncRNAs may have been translated in cells, so the task of discrimination of potentially translated and non-translated transcripts is actual.[4-6] Translated RNAs possess certain features of sequences that have been analyzed in many studies[7]. For the last decade, investigations related to identification of small open reading frames (smORFs, <300 nt) with potential for translation into small peptides (<100 aminoacids) in virtually all RNAs types have become intense.[10-19] In these studies the start of ORFs was determined, as rule, by the AUG triplet. However, in other studies, besides AUG, additional RNA features have been used to identify the start of translation initiation. Thus, existence of suitable nucleotide environment (context) for AUG was postulated to be important for translation process: adenine or guanine in position -3 and guanine at position +4.[20,21] In particular, these two RNA features were used for comparative analysis of mRNAs and lncRNAs[8]. Another RNA feature – the optimal context for initiation of translation in mammals described by consensus GCCRCCAUGG – is used for analysis of efficiency of translation initiation.[20]

The goal of our study was to perform comparative analysis of protein-coding and long non-coding transcripts (RNAs). For this purpose, we used the position weight matrix (PWM) approach to obtain matrices describing the structure of translation starts. These matrices allowed to achieve 96% accuracy of discrimination between human lncRNAs and mRNAs. Obtained matrices were used also for prediction of putative ORFs in lncRNAs. It appeared to be that about 98% of ORFs were ≤300nt of length. In other words, they represented the small ORF type only.

2. Materials and Methods

2.1. RNA samples

For our analysis we compiled three RNA-samples: ‘translation start sample’, ‘mRNA sample’ and ‘lncRNA sample’. The ‘translation starts sample’ consisted of fragments of mouse mRNA (length w = 100 nt) with the canonical start codon AUG in the middle ofthe sequence. Those mRNA fragments were extracted from

mRNA sequences annotated by UCSC database (https://genome.ucsc.edu/). Start codons located in those mRNA sequences were identified on the base of results of ribosome profiling[22]. The ‘mRNA sample’ represented all protein-coding transcripts available in the Ensembl database (build 38). We extracted those transcripts only for which coding DNA sequences (CDS) were available.

The ‘lncRNA sample’ was composed in two versions: ‘lncRNA(12) sample’ and ‘lncRNA(6) sample’. For ‘lncRNA(12) sample’ all available lncRNA sequences had been extracted from Ensembl. Then, for each extracted transcript we identified the putative CDSs that began with *i*-th AUG where *i* is the order of AUG occurrence within the transcript, *i* = 1,…,12. We restricted *i* (number of putative CDSs) by 12, because our preliminary analysis of total sets of human protein-coding transcripts revealed that majority (99%) of annotated CDSs started with up to the twelfth AUG. For the second version – ‘lncRNA(6) sample’ – we restricted number of putative CDSs by 6.

2.2. IPSscore : algorithm for scoring RNA fragments and for matrix derivation

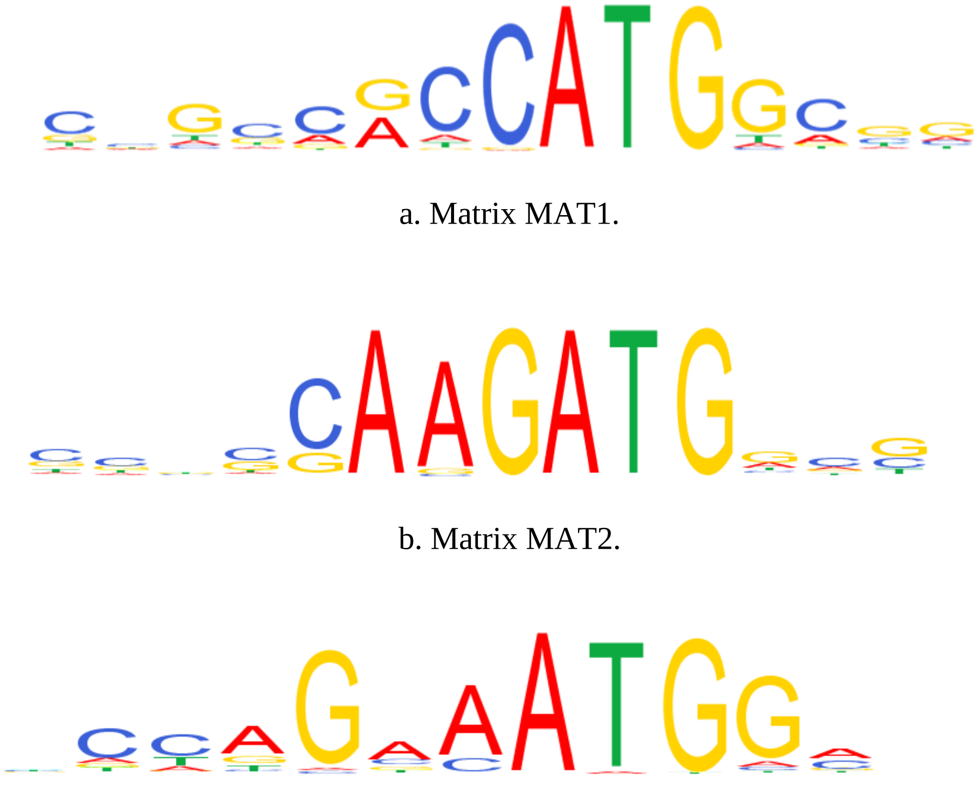
For prediction of putative translation starts, we have developed a novel method for scoring an arbitrary RNA fragment. Briefly, common additive score is calculated at the first step and then individual probability score (IPS) is evaluated. The higher IPS, the more reliable prediction of translation start. IPS depends on structure of a given PWM and nucleotide content of the sequence fragment surrounding the translation start.

3. Results and Discussion

3.1. Matrix derivation

In order to obtain matrices for prediction of translation starts in RNA sequences, we applied the IPSmatrix algorithm to ‘translation starts sample’ several times. The matrix induced by nnnCCRnnATGGnnn consensus of translation starts was used as initial approximation of matrix. As a result, we obtained three matrices (say, MAT1, MAT2 and MAT3) demonstrated on Figure 1. The essential difference between these matrices was observed at the four flank positions upstream the start codon AUG. As a result, these matrices were able to produce very different scores for the same translation start (see Table 1). It is interesting to note that matrix MAT1 represents the optimal context for initiation of translation that has been described previously by consensus GCCRCCAUGG[20].

In order to identify matrix with the best prediction ability, we evaluated ROC curves for each matrix and calculated the corresponding AUC values (see Table 2). Additionally we considered a union prediction model where the translation starts were predicted with the help of three matrices simultaneously. In this case it was sufficient to compute maximal IPS value max_score = max{IPS1, IPS2, IPS3} for tested RNA fragment, where IPSi is the IPS value obtained with the help of MATi matrix, i=1,2,3. What about individual matrices, one can make a confident conclusion that matrix MAT1 outperforms other matrices MAT2 and MAT3 while matrix MAT2 outperforms matrix MAT3. What about union prediction model, it outperforms each model that used single matrix MAT1 or MAT2 or MAT3. On the base of Table 2 and Figure 1 one can conclude that the total set of translation starts is heterogeneous and it is insufficient to use single matrix to predict translation starts with high accuracy.



a. Matrix MAT1.

b. Matrix MAT2.

c. Matrix MAT3.

Fig. 1. Logos for translation start matrices MAT1, MAT2 and MAT3 obtained by the IPSmatrix algorithm.

Table 1. The IPS values obtained for some translation starts with the help of matrices MAT1, MAT2 and MAT3.

Ensembl ID	MAT1	MAT2	MAT3
ENST00000599320	6.809	2.139	2.103
ENST00000452622	6.757	2.321	2.576
ENST00000576178	2.456	6.964	2.001
ENST00000433065	2.409	6.992	1.571
ENST00000568752	2.220	1.592	6.995
ENST00000507844	1.854	2.339	6.928

Table 2. The AUC values computed for individual matrices and for union prediction model.

MAT1	MAT2	MAT3	Union prediction model
0.719	0.690	0.625	0.771

3.2. Comparative analysis of protein-coding and long non-coding RNAs

Protein-coding and long non-coding RNA samples (mRNA sample and lncRNA sample, respectively) have been formed as described in Materials and Methods. Thus, in case of the mRNA sample we used annotated transcripts, while for the lncRNA sample putative structural elements (CDSs and UTRs) were determined *in silico*. Fisher's discriminant model[9] was used for comparison of protein-coding RNAs and lncRNAs. For discrimination we have selected the following eight RNA features: lg-value of length of full transcript (lg(full transcript length)), lg-value of length of CDS (lg(CDS length)), lg-value of length of 5'-UTR (lg(5'-UTR length)), lg-value of length of 3'-UTR (lg(3'-UTR length)), IPS of translation start calculated by union prediction model (max_score) and three IPSs of translation start calculated with the help of individual matrices (IPS(MAT1), IPS(MAT2) and IPS(MAT3)).

The quality of discrimination was measured by true classification rates (TCRs). Table 3 contains these numerical characteristics computed for the mRNA and lncRNA samples. On the base of these characteristics, we made conclusion that selected RNA features discriminated protein-coding and non-coding RNAs with high accuracy, because the majority of transcripts (at least, 96%) were classified properly. It is important to note that this conclusion is invariant with respect to choice of version of the lncRNA sample. In other words, the process of construction of the lncRNA sample had no impact on conclusion. It is also worthwhile to note that in previous study[8] only 80.1% transcripts were

classified properly with the help of the same Fisher's discriminant analysis and another set of 14 RNA features.

Table 3. True classification rates for discrimination between lncRNAs and protein-coding RNAs.

Dataset	**TCRs**	
	lncRNA sample = lncRNA(6)	**lncRNA sample = lncRNA(12)**
mRNA sample	0.954	0.949
lncRNA sample	0.964	0.972
Union of mRNA and lincRNA samples	0.960	0.965

Table 4. Mean values of the RNA features and significance of differences for the matched RNA samples.

RNA feature	**Mean value for mRNA sample**	**Mean value for lncRNA(6) sample**	**Wilcoxon test statistic (normal approximation)**	**p-value**
lg(CDS length)	2.915	1.686	311.563	$<10^{-35}$
lg(full transcript length)	3.213	2.895	164.716	$<10^{-35}$
max_score	3.342	2.782	95.492	$<10^{-35}$
IPS(MAT1)	2.465	1.827	88.135	$<10^{-35}$
IPS(MAT2)	2.461	2.102	59.261	$<10^{-35}$
lg(3'-UTR length)	2.732	2.629	36.127	$<10^{-35}$
IPS(MAT3)	2.279	2.099	30.465	$<10^{-35}$
lg(5'-UTR length)	2.235	2.180	24.775	<10-35

Using Wilcoxon rank sum test we have assessed the contribution of the mRNA features to discrimination between protein-coding RNAs and lncRNAs. Table 4 demonstrates the mean values of the RNA features in the considered samples as well as Wilcoxon test statistics (normal approximation) and p-values characterizing significance of differences between the matched RNA samples. It is easy to see that all considered features had significant influence on discrimination. According to values of Wilcoxon test statistics, the most significant feature was lg(CDS length) while the less significant feature was lg(5'-UTR length). According to Table 4, protein-coding RNAs were characterized by higher values of all considered IPSs of translation starts in comparison to lncRNAs. Finally, full transcripts, CDSs, 5'-UTRs and 3'-UTRs were longer in mRNAs than in lncRNAs (that was expected). In particular, Figure 2 also demonstrates the difference between mRNAs and lncRNAs with the help of corresponding densities of lg(CDS length) and max_score.

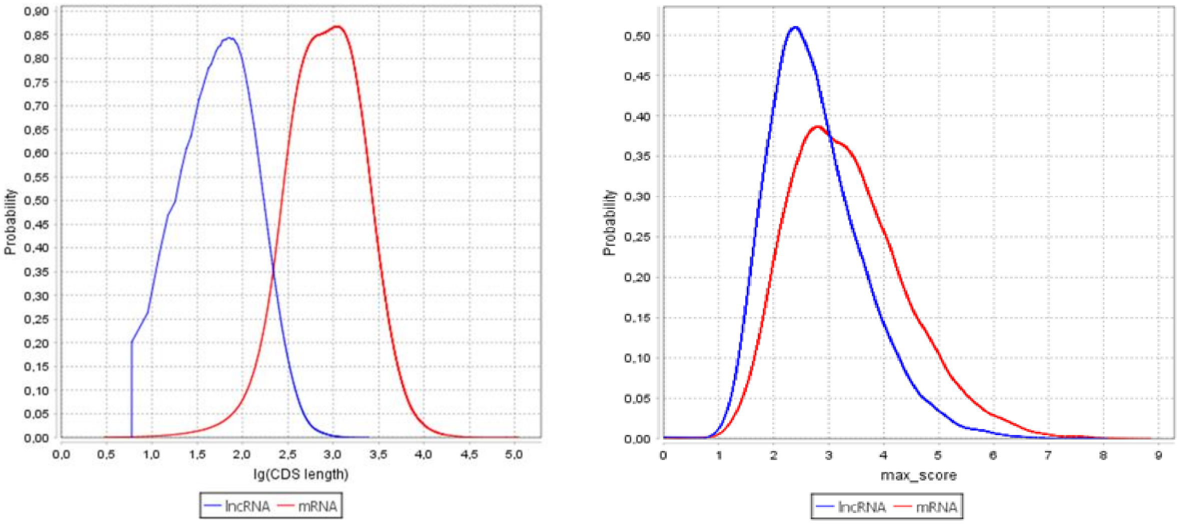


Fig. 2. Densities of lg(CDS length) and max_score evaluated on the mRNA and lncRNA samples.

3.3. Prediction of potentially translated regions within lncRNA transcripts

Because it had been declared in some reports that a part of sequences corresponding to non-coding RNAs has been observed before being occupied by ribosomes, we checked translation potential of lncRNAs by using our matrices. Initially, we predicted the putative translation starts in all sequences from the lncRNA sample. For this purpose, we applied the union prediction model where three derived matrices MAT1, MAT2 and MAT3 were used simultaneously. In order to get high reliability of prediction, we determined the optimal, but high IPS-threshold =4.0. As a result, 15,131 translation starts were predicted. Then we identified putative CDS for each predicted start of translation. It appeared to be that majority of the identified CDSs were short. Thus, 88.3% of the identified CDSs were less than 150 nt of length, while the maximal length for 97.8% of identified CDSs achieved 300 nt. Obtained results were in a good agreement with our expectation that lncRNA CDSs should be rather short.

4. Conclusions

On the base of our results we made the following conclusions:

1. A novel method, IPSmatrix, was developed for derivation of matrices describing the structure of translation starts. Three different matrices were obtained by application of IPSmatrix to the starts of translation extracted from Ribo-seq data.[13] We revealed that these matrices were complementary to each other, i.e. to achieve maximal accuracy of prediction of translation starts in mRNAs they had to be exploited simultaneously.

2. The IPSmatrix method allowed to get new RNA features – the scores obtained by application of three matrices mentioned above – that helped to achieve almost 100% accuracy of discrimination between mRNAs and lncRNAs.

3. The same three matrices were considered as suitable for prediction of putative CDSs in lncRNAs. This conclusion was based on the following results:

- application of the matrices to known mRNAs yielded, often, long CDSs;
- application of the matrices to lncRNAs yielded short putative CDSs that were classified, mostly, into small ORFs – the object of massive modern investigations.

4. Acknowledgements

This work was supported by the Russian Foundation for Basic Research (14-04-01284).

References

[1] Pelechano V, Steinmetz LM, Gene regulation by antisense transcription, *Nat Rev Genet* **14(12)**:880-893, 2013

[2] Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, Lizio M, Kawaji H, Kasukawa T, Itoh M, Burroughs AM, Noma S, Djebali S, Alam T, Medvedeva YA, Testa AC, Lipovich L, Yip CW, Abugessaisa I, Mendez M, Hasegawa A, Tang D, Lassmann T, Heutink P, Babina M, Wells CA, Kojima S, Nakamura Y, Suzuki H, Daub CO, de Hoon MJ, Arner E, Hayashizaki Y, Carninci P, Forrest AR, An atlas of human long non-coding RNAs with accurate 5' ends, *Nature* **543**:199–204, 2017.

[3] Bunch H, Gene regulation of mammalian long non-coding RNA, *Mol Genet Genomics* doi: 10.1007/s00438-017-1370-9, 2017.

[4] Banfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE Jr, Kundaje A, Gunawardena HP, Yu Y, Xie L, Krajewski K, Strahl BD, Chen X, Bickel P, Giddings MC, Brown JB, Lipovich L, Long noncoding RNAs are rarely translated in two human cell lines, *Genome Res* **22(9)**:1646-1657, 2012.

[5] Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, Giraldez AJ, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation, *EMBO J* **33(9)**:981-993, 2014.

[6] Juntawong P, Girke T, Bazin J, Bailey-Serres J, Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis, *Proc Natl Acad Sci USA* **111(1)**:E203-212, 2014.

[7] Hinnebusch AG, Ivanov IP, Sonenberg N, Translational control by 5'-untranslated regions of eukaryotic mRNAs, *Science* **352(6292)**:1413-1416, 2016.

[8] Volkova OA, Kondrakhin YV, Yevshin IS, Valeev TF, Sharipov RN, Assessment of translational importance of mammalian mRNA sequence features based on Ribo-Seq and mRNA-Seq data, *J Bioinform Comput Biol* **14(2)**, 2016.

[9] Mardia KV, Kent JT, Bibby JM, Multivariate analysis, Academic Press, 1979.

[10] Dinger ME, Pang KC, Mercer TR, Mattick JS, Differentiating protein-coding and noncoding RNA: challenges and ambiguities, *PLOS Comput. Biol* **4**:e1000176, 2008.

[11] Firth AE, Brown CM, Detecting overlapping coding sequences in virus genomes, *BMC Bioinform* **7**:75, 2006.

[12] Hsu PY, Calviello L, Wu HL, Li FW, Rothfels CJ, Ohler U, Benfey PN, Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis, *Proc Natl Acad Sci USA* **113**:E7126–135, 2016.

[13] Ingolia NT, Lareau LF, Weissman JS, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes, *Cell* **147(4)**:789-802, 2011.

[14] Ji Z, Song R, Regev A, Struhl K, Many lncRNAs, 5 ? UTRs, and pseudogenes are translated and some are likely to express functional proteins, *eLife* **4**:e08890, 2015.

[15] Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP, Hundreds of putatively functional small open reading frames in Drosophila, *Genome Biol* **12**:R118, 2011.

[16] Pauli A, Valen E, Schier AF, Identifying (non-)coding RNAs and small peptides: challenges and opportunities, *BioEssays* **37**:103–112, 2015.

[17] Raj A, Wang SH, Shim H, Harpak A, Li YI, Engelmann B, Stephens M, Gilad Y, Pritchard JK, Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling, *eLife* **5**:e13328, 2016.

[18] Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, Coller J, Baker KE, Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae, *Cell Rep* **7**:1858–1866, 2014.

[19] Plaza S, Menschaert G3, Payre F, In Search of Lost Small Peptides, *Annu Rev Cell Dev Biol* **33**:391-416, 2017.

[20] Kozak M, Regulation of translation via mRNA structure in prokaryotes and eukaryotes, *Gene* **21(361)**:13–37, 2005.

[21] Pisarev AV, Hellen CU, Pestova TV, Recycling of eukaryotic posttermination ribosomal complexes, *Cell* **131(2)**:286-299, 2007.

[22] Wingender E, The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation, *Brief Bioinform* **9**:326-332, 2008.