Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Original Paper

Sequence Analysis

iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC

Zhen-Dong Su[1,&], Yan Huang[2,&], Zhao-Yue Zhang[1], Ya-Wei Zhao[1], Dong Wang[1,2], Wei Chen[1,3,4,*], Kuo-Chen Chou[1,4,*] and Hao Lin[1,4,*]

[1]Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China, [2]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China, [3]Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China, [4]Gordon Life Science Institute, Boston, MA 02478, USA.

[&]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Long non-coding RNAs (lncRNAs) are a class of RNA molecules with more than 200 nucleotides. They have important functions in cell development and metabolism, such as genetic markers, genome rearrangements, chromatin modifications, cell cycle regulation, transcription and translation. Their functions are generally closely related to their localization in the cell. Therefore, knowledge about their subcellular locations can provide very useful clues or preliminary insight into their biological functions. Although biochemical experiments could determine the localization of lncRNAs in a cell, they are both time-consuming and expensive. Therefore, it is highly desirable to develop bioinformatics tools for fast and effective identification of their subcellular locations.

Results: We developed a sequence-based bioinformatics tool called “iLoc-lncRNA” to predict the subcellular locations of LncRNAs by incorporating the 8-tuple nucleotide features into the general PseKNC (Pseudo K-tuple Nucleotide Composition) via the binomial distribution approach. Rigorous jackknife tests have shown that the overall accuracy achieved by the new predictor on a stringent benchmark dataset is 86.72%, which is over 20% higher than that by the existing state-of-the-art predictor evaluated on the same tests.

Availability: A user-friendly webserver has been established at http://lin-group.cn/server/iLoc-LncRNA, by which users can easily obtain their desired results.

Contact: W Chen: chenweiimu@gmail.com, KC Chou: kcchou@gordonlifescience.org, or H Lin: hlin@uestc.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The basic unit of life is a cell. It contains many biomolecules including proteins, RNA, and DNA. To really understand the biological process inside a cell, the knowledge of the subcellular localization of protein, RNA, and DNA molecules is indispensible. In order to timely obtain the information of their subcellular localization, many bioinformatics tools for predicting the subcellular localization of proteins molecules based on their sequence information alone have been developed (see, e.g., (Cai et al., 2006; Cai et al., 2002; Cheng et al., 2017b; Cheng et al., 2018a; Cheng et al., 2017d; Chou and Cai, 2002; Chou and Cai, 2003; Chou and Shen, 2008; Chou and Shen, 2010; Chou et al., 2011; Chou et al., 2012; Lin et al., 2009; Xuao et al., 2018; Zhu et al., 2015) as well as a long list of references cited in two comprehensive reviews (Chou and Shen, 2007; Nakai, 2000). However, relatively much fewer bioinformatics tools were developed for predicting the subcellular localization of RNA molecules.

Long non-coding RNAs (lncRNAs) are a class of RNA molecules with more than 200 nucleotides and have little or no protein-coding capacity (Spizzo et al., 2012). The large-scale analysis of animal transcriptions showed that the diversity of lncRNA is far exceed that of protein-encoded mRNAs (Carninci and Hayashizaki, 2007; Carninci et al., 2005; Kapranov et al., 2007; The, 2007). lncRNA was originally thought to be a non-functional by product of RNA polymerase II transcripts that are false transcription noise (Struhl, 2007). However, more and more researches have reported that they have important biological functions. Accumulated evidences suggest that lncRNAs have important functional diversity in cell development and metabolism, including genetic markers, genome rearrangement, chromatin modification, cell cycle regulation, transcription, splicing, mRNA decay and translation (Gong and Maquat, 2011; Huarte et al., 2010; Hung et al., 2011; Kino et al., 2010; Kretz et al., 2013; Lee, 2010; Tripathi et al., 2010; Tripathi et al., 2013; Tsai et al., 2010; Xu et al., 2013a; Yap et al., 2010; Yi et al., 2013). Their abnormal expression has been shown to be associated with several types of cancer, Alzheimer’s disease, Huntington’s disease and cardiovascular diseases (Gupta et al., 2010; Johnson, 2012; Lin et al., 2007; McPherson et al., 2007; Mourtada-Maarabouni et al., 2009; Panzitt et al., 2007; Pasmant et al., 2007; Wang et al., 2010; Zhang et al., 2010; Zhao et al., 2005).

Initial studies on lncRNAs have showed that they tend to locate in the nucleus and chromatin for epigenetically regulating gene expression (Hutchinson et al., 2007; Mondal et al., 2010; Rinn et al., 2007; Tsai et al., 2010; Whitehead et al., 2009; Zhao et al., 2008). There exists a substantial population of lncRNAs in the cytoplasm (Carlevaro-Fita et al., 2016; Ulitsky and Bartel, 2013; van Heesch et al., 2014) for regulating protein translation (Schein et al., 2016; Yoon et al., 2012; Zucchelli et al., 2016), protein trafficking (Aoki et al., 2010; Kino et al., 2010), or miRNA decoys (Cesana et al., 2011). Intracellular localization of RNA is now regarded vitally important for understanding the mechanism of eukaryotic cell development and physiology (Donnelly et al., 2010; Weil et al., 2010). In prokaryotes, although there is a lack of nuclei and the coupling between transcription and translation, several studies have demonstrated that various RNA molecules are localized to specific subcellular regions in bacterial cells (Broude, 2011; Keiler, 2011). It is easily deduced that the functions of lncRNAs are closely associated with their locations in cells. Therefore, the identification of subcellular location of lncRNAs is very important.

By using the fluorescent RNA-binding MS2 protein, the first observation about mRNA in live bacterial cells showed that the RNA transcripts in most cases are near the quarter points or close to the cell center, with limited motion (Hiraga, 2000; Nevo-Dinur et al.). Valencia-Burton et al. used fluorescence protein complementation to monitor RNA localization in live prokaryotic cells and found that the lacZ mRNA, the 5S RNA and short non-coding RNA were distributed in cytoplasm, nucleoid and cell poles, respectively (Valencia-Burton et al., 2007). Although these biochemical methods provide very reliable and precise information to determine the subcellular localization of RNAs, they are both expensive and time consuming. Computational methods could overcome these disadvantages and provide high-throughput outcomes. As mentioned above, during the past three decades, many efforts have been made by focusing on the prediction of protein subcellular localization by means of bioinformatics approaches. The similarity between the distribution patterns exhibited by proteins and RNA suggests that their localization is intimately linked to each other (Nevo-Dinur et al.). This linkage suggests that the RNA subcellular localization could also be predicted by using quite similar methods.

To study the RNA subcellular localization, Zhang et al. constructed a database called RNALocate, which collected more than 37,700 manually curated RNA subcellular localization entries (Zhang et al., 2017). Subsequently, Mas-Ponte et al. (Mas-Ponte et al., 2017) built a database called LncATLAS to store the subcellular localization of lncRNA. Cheng et al. (Cheng and Leung, 2018) systematically investigated the distribution of lncRNA subcellular localization in gastric cancer and uncovered its association with cancer. As a pioneer work, Feng et al. (Feng et al., 2017b) developed a computational method to predict the organelle location of noncoding RNAs (ncRNAs) by collecting ncRNAs from kinetoplast, mitochondrion and chloroplast genomes. Subsequently, Zhen, et al. (Zhen et al., 2018) developed a predictor called lncLocator to predict the subcellular localization of long non-coding RNAs.

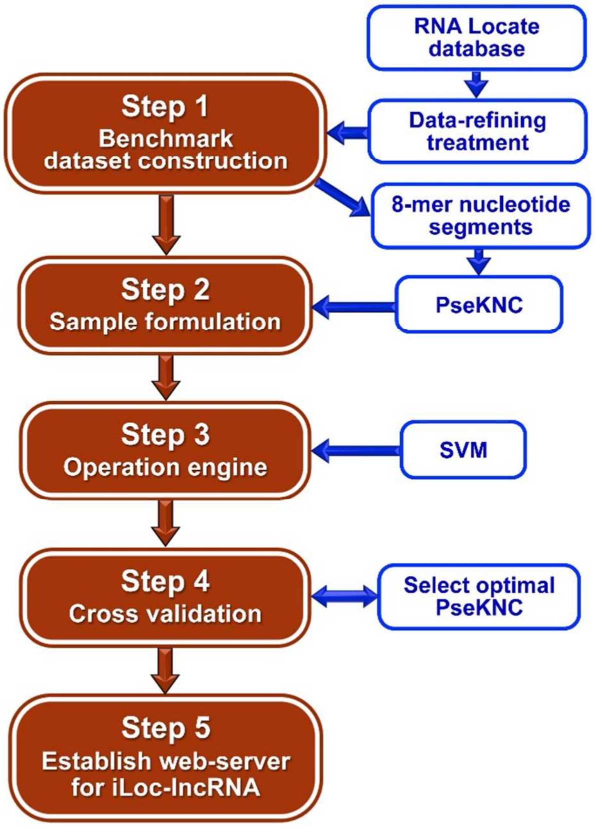

RNA Locate database
Step 1 Benchmark dataset construction
Data-refining treatment
8-mer nucleotide segments
Step 2 Sample formulation
PseKNC
Step 3 Operation engine
SVM
Step 4 Cross validation
Select optimal PseKNC
Step 5 Establish web-server for iLoc-lncRNA

Fig. 1. A flowchart to outline the 5-step rule used in this study.

In this study, we are devoted to developing a computational method to predict lncRNA subcellular localization. As demonstrated by a series of recent publications (Chen et al., 2018a; Chen et al., 2016b; Chen et al., 2017b; Feng et al., 2017a,b; Feng et al., 2018; Khan et al., 2018; Liu et al., 2017c; Liu et al., 2018b; Qiu et al., 2017a; Song et al., 2018b; Song et al., 2018c), presenting a new predictor by observing the 5-step rules (Chou, 2011) would have the following merits: (1) more transparent in logic development; (2) outcome easier to be repeated by others; (3) more inspiring; (4) large impacts.

Below, let us also follow the 5-step guidelines to present our new prediction method; i.e., (1) construct a reliable benchmark dataset to train and test model; (2) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (5) establish a user-friendly web-server for the predictor that is accessible to the public. Illustrated in **Fig.1** is an outline of the 5-steps and their detailed development.

2 Methods

2.1. Benchmark Dataset

Constructing a high quality benchmark dataset is the first prerequisite to establish a reliable model. To realize this, we collected the lncRNA samples from RNALocate (http://www.rna-society.org/rnalocate/). A total of 923 lncRNA sequences with annotated subcellular localization were obtained. Since highly similar data will cause overestimation on the prediction quality, to get rid of the redundancy and avoid bias, the CD-HIT (Li and Godzik, 2006) program was utilized to winnow those lncRNA samples that had ≥ 80% pairwise sequence identity with any other in a same subset. Finally, we obtained 655 lncRNA sequences, which are classified into four subsets, as formulated by

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \tag{1}$$

where the subset $\mathbb{S}_1$ contains 156 lncRNAs from nucleus (**Fig.2**), $\mathbb{S}_1$ contains 426 samples from cytoplasm, $\mathbb{S}_3$ contains 43 lncRNAs from ribosome, and $\mathbb{S}_4$ contains 30 lncRNAs from exosome. The symbol ∪ represents the 'union' in the set theory. For readers' convenience, the accession numbers of these lncRNA samples and their sequences are given in Supporting Information S1, which can also be directly downloaded at http://lin-group.cn/server/iLoc-LncRNA/Supp-S1.txt

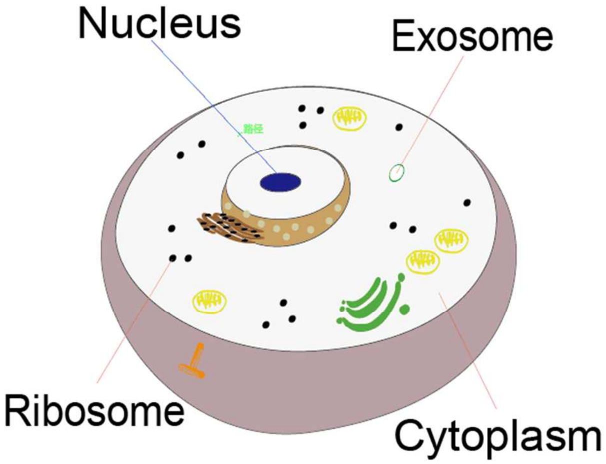


Fig. 2. A schematic drawing to show the four locations of lncRNAs in a cell.

2.2. Sample Formulation

Now let us consider the 2nd step of the 5-step rule (Chou, 2011); i.e., how to formulate the lncRNA sequence samples with an effective mathematical expression that can truly reflect their essential correlation with the target concerned. Given an lncRNA sequence R, its most straightforward expression is (Chen et al., 2015a)

$$\mathbf{R} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\mathrm{N}_4\mathrm{N}_5\mathrm{N}_6\mathrm{N}_7 \cdots \mathrm{N}_L \tag{2}$$

where L denotes the lncRNA's length or the number of its constituent nucleic acid residues, $\mathrm{N}_1$is the 1st residue, N_2 the 2nd residue, N_3 the 3rd residue, and so forth. Since all the existing machine-learning algorithms can only handle vectors (Chou, 2015), we have to convert an lncRNA sample from its sequential expression (Eq.2) to a vector. But a vector defined in a discrete model might completely miss all the sequence-order or pattern information. To deal with this problem, the PseKNC (Pseudo K-tuple Nucleotide Composition) was introduced (Chen et al., 2014), which is an extension of PseAAC (Pseudo Amino Acid Composition) (Chou, 2001; Chou, 2005) that can be used to deal with DNA/RNA sequences. Ever since then, the concept of PseKNC has been widely and increasingly used in many areas of computational genomics/genetics with the aim to grasp various different sequence patterns that are essential to the targets investigated (see, e.g., (Chen et al., 2013; Chen et al., 2015b; Feng et al., 2017a; Feng et al., 2018; Guo et al., 2014; Kabir and Hayat, 2016; Lin et al., 2014; Liu et al., 2018a; Liu et al., 2018b; Qiu et al., 2017b; Xiao et al., 2016; Yang et al., 2018) and a long list of references cited in a recent review papers (Chou, 2017)). According to the concept of general PseKNC (Chen et al., 2015a), any RNA sequence can be formulated as a PseKNC vector given by

$$\mathbf{R} = [\phi_1 \quad \phi_2 \quad \cdots \quad \phi_u \quad \cdots \quad \phi_\Gamma]^{\mathbf{T}} \tag{3}$$

where T is the transposing operator, the subscript Γ is an integer, and its value and the components ϕ_u $(u = 1, 2, \cdots)$ will depend on how to extract the desired features and properties from the RNA sequence. In this study, their definitions are described below.

K-tuple (or called K-mer) nucleotide composition has important biological significance (Ghandi et al., 2014) and has been widely applied in DNA/RNA regulatory element recognition (Chen et al., 2017b; Feng et al., 2018; Zhao et al., 2017; Zhu et al., 2015). Some studies on evolutionary mechanism and biological functions of 8-mers containing CG dinucleotide in yeast have shown (Jia et al., 2018) that the 8-mer distribution has a unique evolutionary mechanism. In order to characterize each lncRNA sequence as accurately as possible, the 8-mer composition was proposed to describe lncRNA samples in this study. Thus, the dimension of PseKNC in Eq.3 is

$$\Gamma = 4^K = 4^8 = 65536 \tag{4}$$

And the u-th 8-mer therein is given by

$$\phi_u = \frac{n_u}{\sum_{i=1}^{65536} n_i} = \frac{n_u}{(L-7)} \tag{5}$$

where u and L denote the numbers of the u-th 8-mer and the length of the sample sequence, respectively. Thus, the lncRNA sample can be defined in a 65536-D vector given by

$$\mathbf{R} = \left[\phi_1, \phi_2, \phi_3, \cdots, \phi_u, \cdots, \phi_{65536}\right]^{\mathbf{T}} \tag{6}$$

2.3. Feature Selection

One may notice that if the lnRNA sample is represented by a vector of 65,536 dimensions, which may cause the following three problems (Ding et al., 2012; Feng et al., 2013; Lai et al., 2017; Liu et al., 2015; Tang et al., 2016b; Wang et al., 2008; Yang et al., 2016; Zhao et al., 2016; Zhao et al., 2017; Zhu et al., 2010): (1) redundant or irrelevant noise yielding poor prediction quality; (2) over-fitting problem resulting in the model with very low generalization ability; (3) the "dimension disaster" or "curse of dimensionality". Fortunately, these problems could be improved by means of the feature selection approach. In fact, some feature selection techniques such as principal component analysis (PCA) (Du et al., 2017), analysis of variance (ANOVA) (Chen et al. 2016c; Lin et al., 2015; Tang et al., 2016a; Tang et al., 2016b; Tang et al., 2018), diffusion Maps (Yin et al., 2011), and mRMR (Minimal Redundancy Maximal Relevance) approach (Hu et al., 2011; Huang et al., 2011a; Huang et al., 2011b; Huang et al., 2012; Li et al., 2012a; Li et al., 2012b; Wang et al., 2011; Zheng et al., 2012) had been proposed to alleviate the interference from noise or irrelevant features so as to improve the prediction quality. In this study, we proposed a powerful technique based on binomial distribution (Lai et al., 2017) to winnow out the most optimal features.

In order to judge whether the occurrence of an 8-mer segment in RNA is completely random, let us define the prior probability q_j given by

$$q_j = \frac{m_j}{M} \tag{7}$$

where m_j denotes the number of a given 8-mer segment occurring in the j-th type of sample (j= 1, 2, 3, and 4 corresponding to the subcellular locations "Nucleus", "Cytoplasm", "Ribosome", and "Exosome", respectively); M is the total number of all different 8-mer segments in the four subsets.

Obviously, the probability of the i-th 8-mer occurring in the j-th type of lncRNA can be defined as

$$p\left(n_{ij}\right) = \sum_{m=n_{ij}}^{N_i} \frac{N_i!}{m!(N_i-m)!} q_j^m \left(1-q_j\right)^{N_i-m} \tag{8}$$

where N_i represents the total number of the i-th 8-mer segment in the benchmark dataset, n_{ij} represents the number of occurrences of the i-th 8-mer segment in the j-th type of lncRNA, and the sum is taken from n_{ij} to N_i. If the i-th 8-mer segment occurring in the j-th subset is not random and biologically significant, the $p\left(n_{ij}\right)$ will be very small. Thus, we may define the confidence level of this statement as CL_{ij}:

$$CL_{ij} = 1 - p\left(n_{ij}\right) \tag{9}$$

According to Eqs. (7)-(9), we ranked the 65,536 8-mer vectors in descending order based on their CL values. Because there are four kinds of subcellular locations considered in this study, there will be four CL values for each of the 8-mer segments. Finally, we assigned the largest one for the CL of the i-th 8-mer vector; i.e.,

$$CL_i = \max(CL_{i1}, CL_{i2}, CL_{i3}, CL_{i4}) \tag{10}$$

2.4. Support Vector Machine (SVM)

SVM is a machine-learning algorithm based on the statistical learning theory, which can improve the generalization ability of learning machine and minimize the risk of experience and the scope of confidence by minimizing the structural risk. Thus, a good statistic result can be usually achieved even in small sample. As a powerful supervised learning method, SVM has been widely used in bioinformatics (see, e.g., (Cai et al., 2002; Cai et al., 2003; Chen et al., 2016a; Chou and Cai, 2002; Ehsan et al., 2018; Hayat and Iqbal, 2014; Kumar et al., 2015; Lai et al., 2017; Mohabatkar et al., 2011; Zhao et al., 2017)). In this paper, the LIBSVM 3.21(Chang and Lin, 2011) was used to perform the prediction, which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Since it is suitable for non-linear classification, the radial basis function (RBF) kernel was selected as kernel function. The one-versus-one (OVO) strategy was used for multiclass classification. In order to construct the optimal model, the regularization parameter C and the kernel width parameter γ were optimized via an optimization procedure using a grid search approach, of which the search spaces for C and γ were $[2^{-5}, 2^{15}]$ and $[2^{3}, 2^{-15}]$ with step sizes of 2 and 2^{-1}, respectively.

The predictor thus constructed is called "iLoc-lncRNA" where "iLoc" stands for "identify or predict subcellular localization', and "lncRNA" for "long non-coding RNAs".

2.5. Performance Evaluation

The classification performance for the subcellular localization of lncRNA was evaluated using sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC) and overall accuracy (OA) (Chen et al., 2007), which are formulated as (Cheng et al., 2018a; Cheng et al., 2018b; Cheng et al., 2017d; Feng et al., 2013; Liu et al., 2018b; Xiao et al., 2017)

$$\begin{cases} \mathrm{Sn}(i) = 1 - \frac{N_{-}^{+}(i)}{N^{+}(i)} & 0 \le \mathrm{Sn}(i) \le 1 \\ \mathrm{Sp}(i) = 1 - \frac{N_{+}^{-}(i)}{N^{-}(i)} & 0 \le \mathrm{Sp}(i) \le 1 \\ \mathrm{MCC}(i) = \frac{1-\left(\frac{N_{-}^{+}(i)}{N^{+}(i)}+\frac{N_{+}^{-}(i)}{N^{-}(i)}\right)}{\sqrt{\left(1+\frac{N_{+}^{-}(i)-N_{-}^{+}(i)}{N^{+}(i)}\right)\left(1+\frac{N_{-}^{+}(i)-N_{+}^{-}(i)}{N^{-}(i)}\right)}} & -1 \le \mathrm{MCC}(i) \le 1 \\ \mathrm{OA} = \frac{1}{\delta}\sum_{i=1}^{\zeta}\left[N^{+}(i) - N_{-}^{+}(i)\right] & 0 \le \mathrm{OA} \le 1 \end{cases} \tag{11}$$

where $N^{+}(i)$ is the total number of lncRNA samples in the i-th subset, $N_{-}^{+}(i)$ is the number of the samples in $N^{+}(i)$ that are incorrectly predicted to be of other locations; $N^{-}(i)$ is the total number of lncRNA samples in any location but not the i-th location, whereas $N_{+}^{-}(i)$ is the number of the samples in $N^{-}(i)$ that are incorrectly predicted to be of the i-th location; ζ is the total number of the concerned, and δ is the number of the total samples in the benchmark dataset.

It is instructive to point out, however, the set of metrics of Eq.11 is valid for the single-label systems in which each sample has one and only one label or just belongs to one attribute. For the multi-label systems where a sample may simultaneously belong to several different attributes, whose existence has become increasingly frequent in system biology (Cheng et al., 2017a; Cheng et al., 2017b; Cheng et al., 2017c; Cheng et al., 2018a; Cheng et al., 2017d; Xiao et al., 2017), system medicine (Cheng et al., 2017e; Cheng et al., 2017f) and biomedicine (Qiu et al., 2016b), a completely different set of metrics as defined in (Chou, 2013) is needed."

3 Results and discussion

3.1 Prediction accuracy

As described in Section 2.2, each LncRNA sample was formulated as a 65,536-D PseKNC vector (Eq.6). By examining the performance of iLoc-lncRNA predictor via the 5-fold cross-validation on the benchmark dataset, we observed that the overall accuracy is 69.77% when C=2^{9} and γ=2^{-15}. Although high-dimensional feature vector may contain more information of the LncRNA sample, it may unavoidably include a lot of noise as well, which could reduce the predictor's accuracy. Moreover, it is time-consuming to train the model using a high-dimensional vector. Therefore, to construct a more accurate predictor, it is necessary to exclude noise from the high-dimensional feature vectors. To realize this, the binomial distribution approach as given in Eqs.7-10 can serve to do so. By investigating the performance of iLoc-lncRNA predictor with the CL being 99.99%, we found that the corresponding model could improve the accuracy from 69.77% to 72.06%. Even though, it is still far from our satisfaction because the number of these 8-mer segments was so small that many important information might be lost. Therefore, it is crucially important to choose the optimal number of features to build a robust and efficient predictive model.

We used the IFS strategy to build the optimal feature subsets. At first, the feature subset started from the 8-mer-vector with the maximum CL value in the ranked feature set. Then, a new feature subset was produced when the second 8-mer with the second biggest CL value was added. This process was repeated from the highest CL value to the lowest CL value until all candidate 8-mer vectors were added. Thus, a total of

65,536 feature subsets were collected and the same number of SVM-based models were built accordingly. Their prediction capabilities were investigated by using the 5-fold cross-validation test. The most optimal feature set was obtained when the overall accuracy reaches its maximum. The corresponding IFS curve was plotted in a 3-D Cartesian coordinate system with feature dimension as its X-coordinate, 1-*CL* as its Y-coordinate and overall accuracy as its Z-coordinate (**Fig. 3**). It can be seen that the overall accuracy was reaching its maximum of 86.11% when the *CL* was selected as 99.19%, with the number of 8-mers features being 4107. In other words, when $\Gamma = 4107$ for the PseKNC of Eq.3, the model would perform the best. The 4,107 vector components thus obtained for each of the protein samples in the benchmark dataset are given in Supporting Information S2, which can also be directly downloaded at http://lin-group.cn/server/iLoc-LncRNA/Supp-S2.txt.

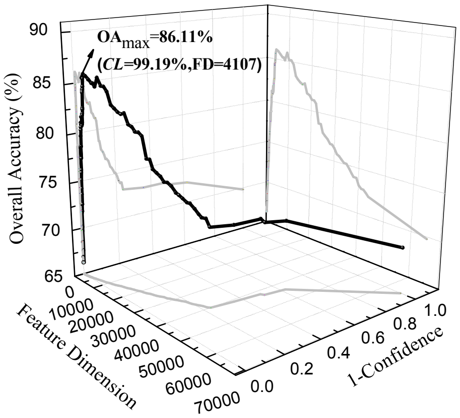


Fig. 3. A plot showing the IFS procedure in a 3-D space. When the dimension of Eq.3 was $\Gamma = 4107$ a peak of 86.11% was reached by 5-fold cross-validation.

Subsequently, the rigorous jackknife tests was used on the same benchmark dataset to examine the performance of the new proposed predictor iLoc-lncRNA when $\Gamma = 4107$ for the PseKNC of Eq.3. The final outcomes thus obtained by the iLoc-lncRNA predictor for Sn, Sp, MCC, and OA (cf. Eq.11) are listed in Table 1, where for facilitating comparison with the corresponding results by lncLocatior (Zhen et al., 2018), the state-of-the-art predictor for the same purpose, the re-estimated results are also given.

Table 1. A comparison of the proposed predictor with the existing predictor.

Location	iLoc-lncRNA[a]				lncLocator[b]			
	Sn[c] (%)	Sp[c] (%)	MCC[c]	OA[c] (%)	Sn[c] (%)	Sp[c] (%)	MCC[c]	OA[c] (%)
Nucleus	77.56	97.59	0.796	86.72	38.15	92.17	0.357	66.50
Cytoplasm	99.06	67.68	0.742		88.01	36.36	0.288	
Ribosome	46.51	99.83	0.652		7.00	97.53	0.070	
Exosome	16.67	1.00	0.400		4.00	97.27	0.015	

[a] Proposed predictor in this paper.

[b] The existing state-of-the-art predictor (Zhen et al., 2018).

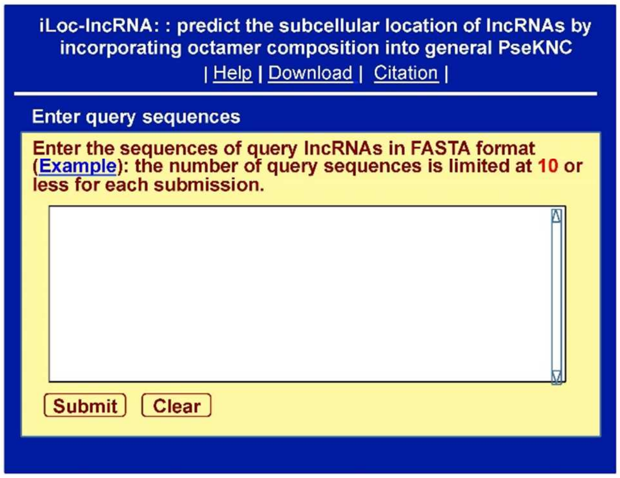
[c] See Eq.11 for the definition of metrics.

As we can see from the table, the proposed iLoc-lncRNA is remarkably superior to the lncLocator (Zhen et al., 2018) from the measurement by each of the four metrics in Eq.11. Particularly, the overall accuracy achieved by the proposed predictor is over 20% high than the existing state-of-the-art predictor, implying that the powerful new predictor will become a high throughput tool widely used in both basic research and drug development.

3.2 Web-server and user guide

As pointed out in (Chou and Shen, 2009), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors. Actually, user-friendly web-servers as given in a series of recent publications (Chen et al., 2017a; Chen et al., 2018b; Jia et al., 2015; Jia et al., 2016a; Jia et al., 2016b; Liang et al., 2017; Liu et al., 2017a; Liu et al., 2018a; Liu et al., 2017b; Liu et al., 2016; Qiu et al., 2016a; Qiu et al., 2016c; Song et al., 2018a; Song et al., 2018c; Wang et al., 2018; Wang et al., 2017; Xu et al., 2013b; Xu et al., 2014; Yang et al., 2018) will substantially increase the impacts of the bioinformatics tools because they can be easily used by broad experimental scientists (Chou, 2017). In view of this, a user-friendly and public accessible web-server for the new iLoc-lncRNA predictor has also been established. Moreover, to maximize users' convenience, a step-by-step guide is given below.

Step 1. Open the web server at http://lin-group.cn/server/iLoc-LncRNA and you will see the top page of iLoc-LncRNA shown on your computer screen (**Fig.4**).



iLoc-lncRNA: : predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC

Enter query sequences

Enter the sequences of query lncRNAs in FASTA format (Example): the number of query sequences is limited at 10 or less for each submission.

Fig. 4. A semi-screenshot for the top page of the iLoc-LncRNA webserver at http://lin-group.cn/server/iLoc-LncRNA.

Step 2. Either type or copy/paste the query RNA sequences into the input box at the center of Figure 4. The input sequences should be in the FASTA format. And click on the Submit button to see the predicted result. For example, if using the four query RNA sequences in the Example window as the input, after clicking the Submit button, you will see the following shown on the screen of your computer. (1) The first query LncRNA is predicted to locate in Nucleus. (2) The second query LncRNA in cytoplasm. (3) The third query LncRNA in ribosome. (4) The fourth query LncRNA in exosome. All these results are perfectly consistent with experimental observations.

Step 3. Click the Download button to get the Supporting Information mentioned in this paper.

Step 4. Click on the Citation button to find the relevant papers that play the key roles in developing the iLoc-LncRNA predictor.

Step 5. Click on the Help button to view the relevant instructions and the caveat when using it.

4 Conclusion

In this paper, a binomial distribution-based feature selection technique was introduced to reduce the feature dimension for avoiding the over-fitting problem, excluding the redundant information, reducing computational complexity, and improving accuracy as well as generalization ability of the model. In fact, some traditional feature selection techniques such as the ANOVA have been used to optimize features. However, these techniques are usually suitable for the data obeying normal distribution. For high dimension k-mer composition, the features obey binomial distribution. Thus, we may use binomial distribution to perform feature selection.

The proposed predictor "iLoc-lncRNA" is superior to the existing state-of-the-art predictor in identifying the subcellular localization of lncRNAs, as clearly indicated by the compelling data listed in Table 1. The powerful predictor will undoubtedly become a high throughput bioinformatics tool for in-depth studying various cellular biological processes including genetic markers, genome rearrangements, chromatin modifications, cell cycle regulation, transcription and translation. It has not escaped our notice that the novel approach presented here may also be used to deal with many other biological systems.

Acknowledgements

The authors wish to thank the three anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this study.

Funding

This work was supported by the National Nature Scientific Foundation of China (61772119, 31771471), the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015Z006, ZYGX2016J125, ZYGX2016J118), Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244), the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028).

Conflict of Interest: none declared.

References

Aoki, K., *et al.* (2010) A thymus-specific noncoding RNA, Thy-ncR1, is a cytoplasmic riboregulator of MFAP4 mRNA in immature T-cell lines, *BMC Mol. Bio.* **11**, 99.

Broude, N.E. (2011) Analysis of RNA localization and metabolism in single live bacterial cells: achievements and challenges, *Mol. Microbiol.* **80**, 1137-1147.

Cai, Y.D., *et al.* (2006) Using LogitBoost classifier to predict protein structural classes, *J. Theor. Biol.* **238**, 172-176.

Cai, Y.D., *et al.* (2002) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect, *J. Cell. Biochem.* **84**, 343-348.

Cai, Y.D., *et al.* (2003) Support vector machines for predicting membrane protein types by using functional domain composition, *Biophys. J.* **84**, 3257-3263.

Carlevaro-Fita, J., *et al.* (2016) Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells, *RNA*. **22**, 867-882.

Carninci, P. and Hayashizaki, Y. (2007) Noncoding RNA transcription beyond annotated genes, *Curr. Opin. Genet. Dev.* **17**, 139-144.

Carninci, P., *et al.* (2005) The Transcriptional Landscape of the Mammalian Genome, *Science*. **309**, 1559.

Cesana, M., *et al.* (2011) A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA, *Cell*. **147**, 358-369.

Chang, C.C. and Lin, C.J. (2011) LIBSVM: A Library for Support Vector Machines, *ACM Trans. Intell. Syst. Technol.* **2**, 27.

Chen, J., *et al.* (2016a) dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation, *Sc. Rep.* **6**, 32333.

Chen, X.X., *et al.* (2016c) Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. *BioMed Res. Int.* 2016: 1654623.

Chen, W., *et al.* (2017a) iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, *Oncotarget*. **8**, 4208-4217.

Chen, W., *et al.* (2018a) iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites, *Mol. Ther. Nucleic Acids*. doi:10.1016/j.omtn.2018.03.012.

Chen, W., *et al.* (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition *Nucleic Acids Res.* **41**, e68.

Chen, W., *et al.* (2014) PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition, *Anal. Biochem.* **456**, 53-60.

Chen, W., *et al.* (2015a) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, *Mol BioSyst*. **11**, 2620-2634.

Chen, W., *et al.* (2016b) iRNA-PseU: Identifying RNA pseudouridine sites *Mol. Ther. Nucleic Acids*. **5**, e332.

Chen, W., *et al.* (2017b) iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, *Bioinformatics*. **33**, 3518-3523.

Chen, W., *et al.* (2015b) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions, *Bioinformatics*. **31**, 119-120.

Chen, Z., *et al.* (2018b) iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, *Bioinformatics*. doi: 10.1093/bioinformatics/bty140/4924718.

Cheng, L. and Leung, K. (2018) Quantification of non-coding RNA target localization diversity and its application in cancers, *J. Mol. Cell. Biol.* **10**, 130-138

Cheng, X., *et al.* (2017a) pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC, *Genomics*. doi:10.1016/j.ygeno.2017.10.002.

Cheng, X., *et al.* (2017b) pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC, *Mol. BioSyst.* **13**, 1722-1727.

Cheng, X., *et al.* (2017c) pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, *Gene (Erratum: ibid., 2018, Vol.644, 156-156)*. **628**, 315-321.

Cheng, X., *et al.* (2018a) pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, *Genomics*. **110**, 50-58.

Cheng, X., *et al.* (2018b) pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information, *Bioinformatics*. **34**, 1448-1456.

Cheng, X., *et al.* (2017d) pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, *Bioinformatics*. **33**, 3524-3531.

Cheng, X., *et al.* (2017e) iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, *Oncotarget*. **8**, 58494-58503.

Cheng, X., *et al.* (2017f) iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, *Bioinformatics (Corrigendum, ibid., 2017, Vol.33, 2610)*. **33**, 341-346.

Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition, *PROTEINS (Erratum: ibid., 2001, Vol.44, 60)*. **43**, 246-255.

Chou, K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, *Bioinformatics*. **21**, 10-19.

Chou, K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review), *J. Theor. Biol.* **273**, 236-247.

Chou, K.C. (2013) Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems, *Mol. BioSyst.* **9**, 1092-1100.

Chou, K.C. (2015) Impacts of bioinformatics to medicinal chemistry, *Med. Chem.* **11**, 218-234.

Chou, K.C. (2017) An unprecedented revolution in medicinal chemistry driven by the progress of biological science, *Curr. Top. Med. Chem.* **17**, 2337-2358.

Chou, K.C. and Cai, Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location, *J. Biol. Chem.* **277**, 45765-45769.

Chou, K.C. and Cai, Y.D. (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology, *Biochem. Biophys. Res. Commun.* **311**, 743-747.

Chou, K.C. and Shen, H.B. (2007) Review: Recent progresses in protein subcellular location prediction, *Anal. Biochem.* **370**, 1-16.

Chou, K.C. and Shen, H.B. (2008) Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms, *Nat. Protoc.* **3**, 153-162.

Chou, K.C. and Shen, H.B. (2009) Recent advances in developing web-servers for predicting protein attributes, *Natural Science*. **1**, 63-92

Chou, K.C. and Shen, H.B. (2010) Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms, *Natural Science*. **2**, 1090-1103.

Chou, K.C.*, et al.* (2011) iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins, *PLoS One*. **6**, e18258.

Chou, K.C.*, et al.* (2012) iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, *Mol. BioSyst.* **8**, 629-641.

Ding, C.*, et al.* (2012) Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions, *J. Proteomics*. **77**, 321-328.

Donnelly, C.J.*, et al.* (2010) Subcellular Communication Through RNA Transport and Localized Protein Synthesis, *Traffic*. **11**, 1498-1505.

Du, Q.*, et al.* (2017) 2L-PCA: a two-level principal component analyzer for quantitative drug design and its applications, *Oncotarget*. **8**, 70564-70578.

Ehsan, A.*, et al.* (2018) A Novel Modeling in Mathematical Biology for Classification of Signal Peptides, *Sci. Rep.* **8**, 1039.

Feng, P.*, et al.* (2017a) iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, *Mol. Ther. Nucleic Acids*. **7**, 155-163.

Feng, P.*, et al.* (2018) iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, *Genomics*. doi:10.1016/j.ygeno.2018.01.005.

Feng, P.*, et al.* (2017b) Predicting the Organelle Location of Noncoding RNAs Using Pseudo Nucleotide Compositions, *Interdiscip Sci*. **9**, 540-544.

Feng, P.M.*, et al.* (2013) iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, *Anal. Biochem.* **442**, 118-125.

Ghandi, M.*, et al.* (2014) Robust k-mer frequency estimation using gapped k-mers, *J. Math. Biol.* **69**, 469-500.

Gong, C. and Maquat, L.E. (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements, *Nature*. **470**, 284-288.

Guo, S.H.*, et al.* (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, *Bioinformatics*. **30**, 1522-1529.

Gupta, R.A.*, et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis, *Nature*. **464**, 1071-1076.

Hayat, M. and Iqbal, N. (2014) Discriminating protein structure classes by incorporating Pseudo Average Chemical Shift to Chou's general PseAAC and Support Vector Machine, *Comput. Methods. Programs. Biomed*. **116**, 184-192.

Hiraga, S. (2000) Dynamic Localization of Bacterial and Plasmid Chromosomes, *Annu. Rev. Genet.* **34**, 21-59.

Hu, L.*, et al.* (2011) Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties *PLoS ONE*. **6**, e14556.

Huang, T.*, et al.* (2011a) Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property, *PLoS ONE*. **6**, e25297.

Huang, T.*, et al.* (2011b) Predicting Transcriptional Activity of Multiple Site p53 Mutants Based on Hybrid Properties, *PLoS One*. **6**, e22940.

Huang, T.*, et al.* (2012) Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma, *PLoS One*. **7**, e34460.

Huarte, M.*, et al.* (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response, *Cell*. **142**, 409-419.

Hung, T.*, et al.* (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters, *Nat. Genet.* **43**, 621-629.

Hutchinson, J.N.*, et al.* (2007) A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains, *BMC Genomics*. **8**, 39.

Jia, J.*, et al.* (2015) iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC, *J. Theor. Biol.* **377**, 47-56.

Jia, J.*, et al.* (2016a) iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC, *Oncotarget*. **7**, 34558-34570.

Jia, J.*, et al.* (2016b) pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, *J. Theor. Biol.* **394**, 223-230.

Jia, Y.*, et al.* (2018) Spectrum structures and biological functions of 8-mers in the human genome, *Genomics*. doi: 10.1016/j.ygeno.2018.03.006.

Johnson, R. (2012) Long non-coding RNAs in Huntington's disease neurodegeneration, *Neurobiol. Dis.* **46**, 245-254.

Kabir, M. and Hayat, M. (2016) iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples, *Mol. Genet. Genomics* **291**, 285-296.

Kapranov, P.*, et al.* (2007) RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription, *Science*. **316**, 1484.

Keiler, K.C. (2011) RNA localization in bacteria, *Curr. Opin. Microbiol.* **14**, 155-159.

Khan, Y.D.*, et al.* (2018) iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC, *Anal. Biochem.* **550**, 109-116.

Kino, T.*, et al.* (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor, *Sci. Signal*. **3**, ra8.

Kretz, M.*, et al.* (2013) Control of somatic tissue differentiation by the long non-coding RNA TINCR, *Nature*. **493**, 231-235.

Kumar, R.*, et al.* (2015) Prediction of beta-lactamase and its class by Chou's pseudo amino acid composition and support vector machine, *J. Theor. Biol.* **365**, 96-103.

Lai, H.Y.*, et al.* (2017) Sequence-based predictive modeling to identify cancerlectins, *Oncotarget*. **8**, 28169-28175.

Lee, J.T. (2010) The X as model for RNA's niche in epigenomic regulation, *Cold Spring Harbor perspectives in biology*. **2**, a003749.

Li, B.Q.*, et al.* (2012a) Prediction of Protein Domain with mRMR Feature Selection and Analysis, *PLoS One*. **7**, e39308.

Li, B.Q.*, et al.* (2012b) Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network, *PLoS One*. **7**, e33393.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*. **22**, 1658-1659.

Liang, Z.Y.*, et al.* (2017) Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics*. **33**, 467-469.

Lin, H.*, et al.* (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, *Nucleic Acids Res.* **42**, 12961-12972.

Lin, H.*, et al.* (2015) Predicting cancerlectins by the optimal g-gap dipeptides, *Sci. Rep.* **5**, 16964.

Lin, H.*, et al.* (2009) Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition, *Acta Biotheor*. **57**, 321-330.

Lin, R.*, et al.* (2007) A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas, *Oncogene*. **26**, 851-858.

Liu, B.*, et al.* (2015) Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, *J. Theor. Biol.* **85**, 153-159.

Liu, B.*, et al.* (2017a) iRSpot-EL: identify recombination spots with an ensemble learning approach, *Bioinformatics*. **33**, 35-41.

Liu, B.*, et al.* (2018a) iRO-3wPseKNC: Identify DNA replication origins by three-window-based PseKNC, *Bioinformatics*. doi: 10.1093/bioinformatics/bty312/4978052.

Liu, B.*, et al.* (2017b) Pse-Analysis: a python package for DNA/RNA and protein/peptide sequence analysis based on pseudo components and kernel methods, *Oncotarget*. **8**, 13338-13343.

Liu, B.*, et al.* (2017c) 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function, *Mol. Ther. Nucleic Acids*. **7**, 267-277.

Liu, B.*, et al.* (2018b) iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, *Bioinformatics*. **34**, 33-40.

Liu, Z.*, et al.* (2016) pRNAm-PC: Predicting N-methyladenosine sites in RNA sequences via physical-chemical properties, *Anal. Biochem.* **497**, 60-67.

Mas-Ponte, D.*, et al.* (2017) LncATLAS database for subcellular localization of long noncoding RNAs, *RNA*. **23**, 1080-1087.

McPherson, R.*, et al.* (2007) A common allele on chromosome 9 associated with coronary heart disease, *Science*. **316**, 1488-1491.

Mohabatkar, H.*, et al.* (2011) Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo amino acid composition and support vector machine, *J. Theor. Biol.* **281**, 18-23.

Mondal, T.*, et al.* (2010) Characterization of the RNA content of chromatin, *Genome Res*. **20**, 899-907.

Mourtada-Maarabouni, M.*, et al.* (2009) GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer, *Oncogene*. **28**, 195-

208.

Nakai, K. (2000) Protein sorting signals and prediction of subcellular localization, *Adv. Protein Chem*. **54**, 277-344.

Nevo-Dinur, K., *et al.* Subcellular localization of RNA and proteins in prokaryotes, *Trends Genet*. **28**, 314-322.

Panzitt, K., *et al.* (2007) Characterization of HULC, a Novel Gene With Striking Up-Regulation in Hepatocellular Carcinoma, as Noncoding RNA, *Gastroenterology*. **132**, 330-342.

Pasmant, E., *et al.* (2007) Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF, *Cancer Res*. **67**, 3963-3969.

Qiu, W.R., *et al.* (2017a) Identify and analysis crotonylation sites in histone by using support vector machines, *Artif. Intell. Med*. **83**, 75-81.

Qiu, W.R., *et al.* (2017b) iRNA-2methyl: identify RNA 2′-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier, *Med. Chem*. **13**, 734-743.

Qiu, W.R., *et al.* (2016a) iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC, *Oncotarget*. **7**, 44310-44321.

Qiu, W.R., *et al.* (2016b) iPTM-mLys: identifying multiple lysine PTM sites and their different types, *Bioinformatics*. **32**, 3116-3123.

Qiu, W.R., *et al.* (2016c) iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier, *Oncotarget*. **7**, 51270-51283.

Rinn, J.L., *et al.* (2007) Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Non-Coding RNAs, *Cell*. **129**, 1311-1323.

Schein, A., *et al.* (2016) Identification of antisense long noncoding RNAs that function as SINEUPs in human cells, *Sci. Rep*. **6**, 33605.

Song, J., *et al.* (2018a) PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy, *Bioinformatics*. **34**, 684-687.

Song, J., *et al.* (2018b) PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural and network features in a machine learning framework, *J. Theor. Biol*. **443**, 125-137.

Song, J., *et al.* (2018c) iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, *Briefings in Bioinformatics*. doi: 10.1093/bib/bby028.

Spizzo, R., *et al.* (2012) Long non-coding RNAs and cancer: a new frontier of translational research?, *Oncogene*. **31**, 4577-4587.

Struhl, K. (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II, *Nat. Struct.Mol. Biol*. **14**, 103.

Tang, H., *et al.* (2016a) Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique, *Mol Biosyst*. **12**, 1269-1275.

Tang, H., *et al.* (2016b) Prediction of cell-penetrating peptides with feature selection techniques, *Biochem. Biophys. Res. Commun*. **477**, 150-154.

Tang, H., *et al.* (2018) HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci*. **14**, 957-964.

The, E.P.C. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*. **447**, 799-816.

Tripathi, V., *et al.* (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation, *Mol. Cell*. **39**, 925-938.

Tripathi, V., *et al.* (2013) Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB, *PLoS Genet*. **9**, e1003368.

Tsai, M.C., *et al.* (2010) Long noncoding RNA as modular scaffold of histone modification complexes, *Science*. **329**, 689-693.

Ulitsky, I. and Bartel, D.P. (2013) lincRNAs: genomics, evolution, and mechanisms, *Cell*. **154**, 26-46.

Valencia-Burton, M., *et al.* (2007) RNA visualization in live bacterial cells using fluorescent protein complementation, *Nat. Methods*. **4**, 421.

van Heesch, S., *et al.* (2014) Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes, *Genome Biol*. **15**, R6-R6.

Wang, J., *et al.* (2010) CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer, *Nucleic Acids Res*. **38**, 5366-5383.

Wang, J., *et al.* (2018) Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors, *Bioinformatics*. 10.1093/bioinformatics/bty155.

Wang, J., *et al.* (2017) POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles, *Bioinformatics*. **33**, 2756-2758.

Wang, P., *et al.* (2011) Prediction of antimicrobial peptides based on sequence alignment and feature selection methods, *PLoS One*. **6**, e18476.

Wang, T., *et al.* (2008) Predicting membrane protein types by the LLDA algorithm, *Protein Pept. Lett*. **15**, 915-921.

Weil, T.T., *et al.* (2010) Making the message clear: visualizing mRNA localization, *Trends Cell Biol*. **20**, 380-390.

Whitehead, J., *et al.* (2009) Regulation of the mammalian epigenome by long noncoding RNAs, *Biochim. Biophys. Acta*. **1790**, 936-947.

Xiao, X., *et al.* (2017) pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins, *Natural Science*. **9**, 331-349.

Xiao, X., *et al.* (2016) iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition, *Oncotarget*. **7**, 34180-34189.

Xu, D., *et al.* (2013a) Long noncoding RNAs associated with liver regeneration 1 accelerates hepatocyte proliferation during liver regeneration by activating Wnt/beta-catenin signaling, *Hepatology*. **58**, 739-751.

Xu, Y., *et al.* (2013b) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, *PeerJ*. **1**, e171.

Xu, Y., *et al.* (2014) iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, *PLoS One*. **9**, e105018.

Xuao, X., *et al.* (2018) pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC, *Genomics*. doi:10.1016/j.ygeno.2018.05.017.

Yang, H., *et al.* (2018) iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci*. **14**, 883-891.

Yang, H., *et al.* (2016) Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition, *Biomed. Res. Int*. **2016**, 5413903.

Yap, K.L., *et al.* (2010) Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a, *Molecular cell*. **38**, 662-674.

Yi, F., *et al.* (2013) RNA-seq identified a super-long intergenic transcript functioning in adipogenesis, *RNA Biol*. **10**, 991-1001.

Yin, J., *et al.* (2011) Conotoxin superfamily prediction using diffusion maps dimensionality reduction and subspace classifier, *Curr. Protein Pept. Sci*. **12**, 580-588.

Yoon, J.H., *et al.* (2012) LincRNA-p21 suppresses target mRNA translation, *Mol. Cell*. **47**, 648-655.

Zhang, T., *et al.* (2017) RNALocate: a resource for RNA subcellular localizations, *Nucleic Acids Res*. **45**, D135-D138.

Zhang, X., *et al.* (2010) Maternally Expressed Gene 3 (MEG3) Noncoding Ribonucleic Acid: Isoform Structure, Expression, and Functions, *Endocrinology*. **151**, 939-947.

Zhao, J., *et al.* (2005) Hypermethylation of the promoter region is associated with the loss of MEG3 gene expression in human pituitary tumors, *J. Clin. Endocrinol. Metab*. **90**, 2179-2186.

Zhao, J., *et al.* (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome, *Science*. **322**, 750-756.

Zhao, Y., *et al.* (2016) Prediction of phosphothreonine sites in human proteins by fusing different features, *Sci. Rep*. **6**, 34817.

Zhao, Y.W., *et al.* (2017) IonchanPred 2.0: A Tool to Predict Ion Channels and Their Types, *Int. J. Mol. Sci*. **18**, 1838.

Zhen, C., *et al.* (2018) The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier, *Bioinformatics*. doi: 10.1093/bioinformatics/bty085.

Zheng, L.L., *et al.* (2012) A comparison of computational methods for identifying virulence factors, *PLoS One*. **7**, e42517.

Zhu, L., *et al.* (2010) Improving the accuracy of predicting disulfide connectivity by feature selection, *J. Comput. Chem*. **31**, 1478-1485.

Zhu, P.P., *et al.* (2015) Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition, *Mol. BioSyst*. **11**, 558-563.

Zucchelli, S., *et al.* (2016) Engineering Translation in Mammalian Cell Factories to Increase Protein Yield: The Unexpected Use of Long Non-Coding SINEUP RNAs, *Comput.Struct. Biotechnol. J*. **14**, 404-410.