Subject Section

Accurate Prediction of Genome-wide RNA Secondary Structure Profile Based on Extreme Gradient Boosting

Yaobin Ke[1], Jiahua Rao[1], Huiying Zhao[2], Yutong Lu[1], Nong Xiao[1*] and Yuedong Yang[1,3]*

[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510000, China

[2]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510000, China

[3]Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University) of Ministry of Education, Guangzhou, China

***To whom correspondence should be addressed.**

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: RNA secondary structure plays a vital role in fundamental cellular processes, and identification of RNA secondary structure is a key step to understand RNA functions. Recently, a few experimental methods were developed to profile genome-wide RNA secondary structure, i.e. the pairing probability of each nucleotide, through high-throughput sequencing techniques. However, these high-throughput methods have low precision and can't cover all nucleotides due to limited sequencing coverage.

Results: Here we have developed a new method for the prediction of genome-wide RNA secondary structure profile from RNA sequence based on the extreme Gradient Boosting technique. The method achieves predictions with areas under the receiver operating characteristic curve (AUC) greater than 0.9 on three different datasets, and AUC of 0.888 by an independent test on the recently released Zika virus data. These AUCs are consistently >5 % greater than the ones by the CROSS method recently developed based on a shallow neural network. Further analysis on the 1000 Genome Project data showed that our predicted unpaired probabilities are highly correlated (>0.8) with the minor allele frequencies at synonymous, non-synonymous mutations, and mutations in untranslated region, which were higher than those generated by RNAplfold. Moreover, the prediction over all human mRNA indicated a consistent result with previous observation that there is a periodic distribution of unpaired probability on codons. The accurate prediction by our method indicates that such model trained on genome-wide experimental data might be an alternative for analytical methods.

Availability: The GRASP is available for academic use at https://github.com/sysu-yanglab/GRASP.

Contact: xiaon6@mail.sysu.edu.cn or yangyd25@mail.sysu.edu.cn

Supplementary information: Supplementary data are available online.

1 Introduction

RNA plays an essential role in a wide variety of fundamental cellular processes, such as transcription, replication, protein synthesis, and regulation of gene expression(Glisovic, et al., 2008; Mortimer, et al., 2014). The structure of an RNA, including secondary structure and tertiary structure, determines its translation and other functions. Identifying secondary structure is a key groundwork to know tertiary structure and a vital premise to understand the detailed mechanism of various biological activities, such as protein-RNA interactions and translation process. Therefore, there is a critical need to identify the RNA secondary structure by an unbiased and systematic manner.

While RNA secondary structure can be obtained from a small number of RNA tertiary structures experimentally determined by low throughput techniques such as Nuclear Magnetic Resonance (NMR), X-ray Crystallography, and Cryo-electron microscopy, recently, a few experimental techniques have been developed to perform high-throughput profiling of the RNA structure by exploiting biochemical reactions. For example, Parallel Analysis of RNA Structure (PARS) distinguishes double- and single-stranded regions using catalytic activities of two enzymes, RNase V1 and S1 (able to cut double-stranded and single-stranded nucleotides respectively). This technique has been successfully applied to the yeast and the human genomes (Kertesz, et al., 2010; Wan, et al., 2014). Fragmentation sequencing (FragSeq), using nuclease P1 to generate fragments, was applied to determine single-stranded RNA regions in multiple ncRNAs(Underwood, et al., 2010). Besides, selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq) was able to measure the structures of a complex pool of RNAs(Lucks, et al., 2011) but the output was found to be sensitive to noise(Ouyang, et al., 2013). Another technique based on in vivo modification with dimethyl sulphate (DMS), which only reacts to adenine and cytosine, are of high quality. But DMS experiments are not included here because not all the nucleotide states are provided(Ding, et al., 2014; Rouskin, et al., 2014). Recently, two orthogonal high-throughput sequencing-based techniques, icSHAPE (*in vivo* click selective 2-hydroxyl acylation and profiling experiment) and PARIS (psoralen analysis of RNA interactions and structures), have been applied to the Zika virus (Zikv) for an accurate estimate of genome-wide secondary structure profile(Li, et al., 2018).

However, high-throughput genomic experiments always have high noise and are hard to cover all nucleotides on the RNA due to limited sequencing coverage(Ouyang, et al., 2013; Wang, et al., 2009). Moreover, the sequencing experiments are of heavy experimental work and high costs. Therefore, computational methods are often required. Many tools have been developed to obtain locally stable secondary structure by minimizing the free energy, such as ViennaRNA (Bernhart, et al., 2006; Lorenz, et al., 2011). Nonetheless, these methods are not accurate enough due to a lack of a precise free energy criterion(Mathews, et al., 1999), and the searching of the global minima is an NP-hard problem (Lyngso and Pedersen, 2000). To be even worse, the secondary structure with the lowest free energy is not always the actual one (Hofacker, 2014; Seetin and Mathews, 2012; Ye, et al., 2005). Recently, with the accumulation of experimental genomic data, a CROSS method was developed to predict RNA secondary structural profile by a shallow artificial neural network with only one hidden layer (Ponti, et al., 2017). The neural network is well known to have strong self-learning and non-linear fitting ability, but it is easy to fall into local optimal solution and has slow convergence with small training data (Jin-yue and Bao-ling, 2012; Roberts, 2003).

Recently, the eXtreme Gradient Boosting (XGBoost) technique was proposed by aggregating multiple weak learners to obtain a combined and strong learner (Chen and Guestrin, 2016). Meanwhile, as a kind of gradient boosting models, its implementation of parallel processing enables a fast model training compared to many traditional models, and can be deployed to high-performance platform for large-scale parallel computing. The technique was found to outperform other machine learning and deep learning techniques in many competitions such as Kaggle and KDDCup (Chen and Guestrin, 2016; Dhaliwal, et al., 2018), especially for datasets with sparse matrix. It has been successfully applied in many bioinformatic studies, such as miRNA-disease association(Chen, et al., 2018), protein translocation(Mendik, et al., 2019), protein-protein interactions(Basit, et al., 2018), and DNA methylation(Zou, et al., 2018).

In this study, we developed a new method for end-to-end prediction of the Genome-wide RNA Secondary Structure Profile (GRASP) from RNA sequence by using the XGBoost technique. The method achieves area under the receiver operating characteristic curve (AUC) values greater than 0.9 by cross-validations on three different datasets (high-throughput PARS yeast and human datasets, and high-quality dataset from NMR/X-ray structures), and AUC of 0.888 on an independent test of the ZIKA virus dataset. The comparison showed that our method consistently outperformed the CROSS method trained by using shallow neural networks. Moreover, the predictive power of our model was also supported by a correlation between predicted structure profile and minor allele frequencies (MAF) of genetic variants, as well as the finding that both ends of coding region have less structure.

2 Materials and Methods

Datasets

For validation of our method, we employed three training datasets (PARS-Yeast, PARS-Human, and SS-PDB) as also used in the previous study (Ponti, et al., 2017). Since their training and test sets weren't available, we followed the same steps as reported in the study to generate the datasets. In addition, an independent test set was compiled from the recently released Zika virus (Zikv) genomic data (Li, et al., 2018). Table 1 show the details of three training datasets and the independent test set.

The PARS-Yeast and PARS-Human datasets were derived from the experimentally measured RNA structural profiles probed by the PARS technique on the *S.cerevisiae* (Kertesz, et al., 2010) and *Homo sapiens* (GEO: GSE50676) (Wan, et al., 2014). The experimental data contain around 3200 and more than 35,000 non-redundant transcripts, respectively. In these two datasets, the ratio between double and single stranded frequencies was calculated as a score (PARS score) for each nucleotide. In order to obtain nucleotides with the most reliable measurements, we selected nucleotides with the highest scores on each transcript as double-stranded nucleotides (positives), and the same number with the lowest scores as the single-stranded nucleotides

Table 1. The details of the three training datasets and an independent dataset

	Transcript Sequences	Positives nucleotides	Negatives nucleotides	Total nucleotides	Note
PARS-Yeast	3196	15341	15340	30681	All transcripts from transcriptome were selected to generate dataset.
PARS-Human	36531	29454	31720	61174	28120 transcripts with experiment scores of 36531 from the human transcriptome were selected to generate dataset.
SS-PDB	202	30680	16014	46694	202 transcripts were selected by homology screening to generate dataset.
SS-ZIKV	1	1627	1618	3245	All RNA of Zika Virus was selected to generate dataset

(negatives). Around the selected nucleotides, fragments were prepared to include 18 nucleotides from its upstream and downstream, leading to a length of 37 for each sample. As fragments from different positions or transcripts might have the same sequences but different scores or labels, we kept one representative fragment set as the consistent state of secondary structure if more than 90% of the fragments were in the same state, single- or double-stranded, otherwise the fragments were all removed. Here, we selected top and bottom five nucleotides in each transcript for *S.cerevisiae* and two for *Homo sapiens* in order to keep close numbers for the two species. Finally, we obtained 15341 positives nucleotides (each represented by a fragment) and 15340 negatives for the *S.cerevisiae*, and 29454 positives and 31720 negatives for the *Homo sapiens*, namely PARS-Yeast and PARS-Human.

SS-PDB: We downloaded 1341 secondary structures of RNA from the RNAstrand (Andronescu, et al., 2008), a curated database from the three-dimensional structures of RNA that were determined by X-ray or NMR and deposited in the Protein Data Bank (PDB). By removing redundant sequences with sequence identity greater than 75% calculated by CD-hit(Li and Godzik, 2006), 202 sequences remained. The sequences include totally 46694 nucleotides, with 30, 680 positives (double stranded) and 16, 014 negatives (single stranded), namly SS-PDB.

SS-ZIKV: We downloaded the experimental scores of secondary structure profiles for Zika virus from previous study (Li, et al., 2018). As suggested by the previous study(Ponti, et al., 2017), we selected nucleotides with raw score of 0 and 1 as double- and single-stranded, respectively. Similar to the previous way for processing PARS dataset, we merged identical fragments with a consistent secondary structure state (occurring among >90% of the fragments), and removed all identical fragments if they are divergent (either state occurring <90%). Finally, we kept 1627 double-stranded fragments (positives), and 1618 single-stranded fragments (negatives), namely SS-ZIKV. For comparison with CROSS, the sequences were submitted to online server (http://service.tartaglialab.com/new_submission/cross) to obtain the predictions.

The first three training datasets were used for both cross-tests and self-tests. In the cross-test, one dataset was employed for training model, and the other two datasets were used to evaluate the performance. In the self-test, the method was separately tested on each dataset using the five-fold cross-validation. The five-fold cross validation test was conducted by randomly splitting the dataset into five subsets, where four subsets were used for training a model, and the remained was used for validation. This process repeated for five times so that each fold was tested once. All results were collected to measure the overall performance for the dataset. The SS-ZIKV was used as independent test.

Features extraction and encoding of RNA sequences

We employed a window-based strategy for features extraction of secondary structure status. For a given nucleotide, *d* nucleotides both upstream and downstream of it were selected as its features. Here, we defined a window size *l*, where $l = 2d+1$. So the window size decided the number of features to represent a nucleotide. At the beginning or end of the sequence, it was padded with the letter N if the length of upstream or downstream was less than *d*. After feature extraction, each nucleotide was encoded with the one-hot notation(Figure 1): A = (1, 0, 0, 0), C = (0, 1, 0, 0), G = (0, 0, 1, 0), U = (0, 0, 0, 1), and N = (0, 0, 0, 0). Thus, the prediction of each nucleotide has an input of $4 \times l$ matrix. By testing different window sizes, we finally chosen l = 37 for a balance of performance and training time.

Training of GRASP

GRASP was trained by using XGBoost, which is an ensemble method to generate *k* Classification and Regression Trees (CART). The input of the CART includes a vector with size of $4l$ that is flattened from the feature matrix ($4 \times l$). The training procedure of XGBoost can be outlined as follows:

1) Sort values in each feature and scan the best splitting point, the values that gives the lowest gain;
2) Select the feature with the best splitting point that optimizes the objective function；
3) Repeat the splitting in the above two steps until the maximum tree depth (set hyper-parameter) is reached;
4) Make assignment to the leaves with prediction score and prune the nodes with negative gains according to a bottom-up order;
5) Continue repeating the above steps for *k* times (*k* trees);

Table 2. The hyper-parameters used for the model training.

Hyper-parameters	Description	Value set
max_depth	Maximum depth of a tree. The larger the value, the easier to learn the local specificity but also easier to overfit.	{3, 6, 9, 12}
subsample	Subsample ratio of the training instances. Smaller value leads to faster training speed but the risk of underfit.	{0.7, 1}
colsample_bytree	A parameter for subsampling of columns.	{0.7, 1}
learning_rate	Step size shrinkage used in update to prevents overfitting. The smaller the value, the easier to underfit.	{0.05, 0.1}
reg_lambda	L2 regularization term on weights.	{0.05, 0.1, 0.5}
reg_alpha	L1 regularization term on weights.	{0.05, 0.1, 0.5}
n_estimators	The number of base learners, with the same effect as learning_rate.	{500, 1000, 2000}

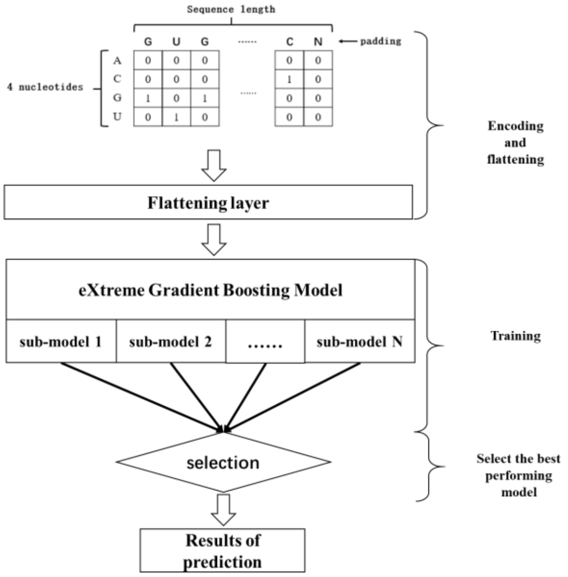

Figure 1. The flowchart of GRASP method

We used the implementation provided in the XGBoost Python library that was optimized for distributed systems. Here, we selected $l = 37$ after comparison (explain in detail in discussion section). We used grid search in Scikit-learn framework (Pedregosa, et al., 2011) to find the optimal parameters. The range of parameters set up in the training process is shown in Table 2. Moreover, the optimized XGBoost models were trained on a 16 core CPU to speed up the learning process. Parameter optimization and evaluation of the models were performed using 5-fold cross-validation.

Figure 1 shows the flowchart for the model training. First, the individual features extracted from RNA sequence were encoded and flatten. Then the models were parallelly trained by grid searching strategy with 5-fold cross-validation. The sub-model with the best AUC in validation was selected. Finally, the independent test was performed by the remained two datasets not involved in the training.

Evaluation Metrics

The performance of the model was measured by the area under the receiver-operating characteristic curve (AUC), accuracy (ACC), precision, recall value and F1-Score score. The relevant formulas for these measurements are shown as below:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F1 - score = \frac{(2 * precision * recall)}{precision + recall}$$

, where TP is true positives, the number of paired nucleotides that are predicted to be paired. Similarly, TN, FP and FN are the numbers of true negative, false positive and false negative, respectively.

The 1000-Genome dataset

The 1000 Genomes Phase 3 VCF file was downloaded from Ensembl annotated by ANNOVAR (Wang, et al., 2010), which leads to single nucleotide variations (SNVs) along with their minor allele frequencies (MAF), including 223,693 cases in 5′ untranslated regions (UTR), 899,976 in 3′ UTR, 16,847 in stop-gain region, 704,643 nonsynonymous and 427,077 synonymous regions. MAF is the frequency of the least common allele in a population. For each category, SNVs were sorted and equally separated into 50 bins according to predicted score for secondary structure (or predicted ASA, accessible surface area, by RNA-snap) at their mutation positions. Log (Predicted values) and log (MAF) were averaged, as well as Pearson’s correlation coefficients were calculated based on the average values(Yang, et al., 2017).

Human genome data

We downloaded 89,732 transcripts sequences in the human genome from Gencode version v26, which referred to Ensembl v88. Genes without 5′ UTR, coding sequence (CDS), or 3′ UTR were removed, which led to 60876 transcripts from 18527 genes.

Comparison to RNAplfold

For genome-scale studies, we compared the GRASP to secondary structure profile prediction software RNAplfold from the package ViennaRNA 2.1.9 (Lorenz et al. 2011). By using command "RNAplfold -u 1 -W 37" with a window size of 37, we obtained the unpaired probability for each nucleotide, which could be converted to paired probability by using "1 – unpaired probability".

3 Results

3.1 Prediction of RNA secondary structure profile

As shown in Table 3, our method achieved AUC values of 0.942, 0.967, and 0.901 by the five-fold cross-validation on the PARS-Human, PARS-Yeast, and SS-PDB, respectively. At a threshold of 0.5, the respective accuracy values are 0.871, 0.901, and 0.844 for the prediction of secondary structure profile (nucleotides to be paired or not). The balanced measures by F1-score are all above 0.85, where the precision and recall values are all above 0.86.

In order to make stricter tests, we performed the cross-tests between three datasets, where model was trained on one dataset, and tested on the other two datasets. As shown in Table 4, the model trained by PARS-Human achieved an AUC essentially the same as the one achieved on the PARS-Yeast by 5-fold cross validation (0.94 vs 0.94). Meanwhile, a close AUC value was also achieved on the model trained by PARS-Yeast (0.93 vs 0.97). The similar performance by self-tests and cross-tests on the PARS-Human and PARS-Yeast demonstrated the robustness of our method on different genomes. Differently, when these two models are applied to the SS-PDB dataset, the models trained by PARS-Human and PARS-Yeast achieved close but significantly lower AUC values (0.65 and 0.63). This is likely due to the difference in two experimental techniques. On the other hand, the analytical method, RNAplfold achieved lower AUC values with 0.76, 0.86, and 0.67 for PARS-Human, PARS-Yeast, and SS-PDB, respectively (Table 4). Notably, for PARS-Human, when we selected top and bottom 5 nucleotides instead of two as mentioned in the dataset, the results were similar (Table S3, Table S4).

When compared to the reported results by the CROSS, a method trained by a shallow neural network, our method performs consistently better in all cross-tests datasets: the AUCs achieved by our method are 4.4~10% better than those by CROSS with an average of 6.7% (Table S2). The differences in datasets should have little impact, as the improvements are essentially the same as the one (5.7%) in the independent test on the SS-ZIKV dataset by using CROSS's online server (see section 3.3). Our method achieved slightly lower AUCs in the self-tests (the results of cross-validation).

Table 3. The performance of GRASP on three training datasets by the 5-fold cross validation.

	PARS-Human	PARS-Yeast	SS-PDB
AUC	0.942±0.001	0.967±0.002	0.901±0.004
ACC	0.871±0.002	0.901±0.003	0.844±0.005
Precision	0.874±0.004	0.905±0.006	0.848±0.005
Recall	0.854±0.005	0.906±0.004	0.929±0.003
F1-score	0.864±0.002	0.906±0.003	0.887±0.004

Table 4. Comparisons of performances on three datasets by GRASP and RNAplfold

	GRASP			RNAplfold
train test (all)	PARS-Human	PARS-Yeast	SS-PDB	-------
PARS-Human	0.94	0.93	0.75	0.76
PARS-Yeast	0.94	0.97	0.76	0.86
PDB	0.65	0.63	0.90	0.67

3.2 Performance of consensus model

Since three training datasets represent different genomes or experimental techniques, it is interesting to know the performance by combining all datasets. We randomly selected 90% samples from each dataset, and put them together to train a consensus model. The remaining 10% of the three datasets were used as independent test sets. As shown in the Table 5, the consensus model achieved close to the highest AUC values among three independents. The average AUC is 0.927, significantly higher than the 0.846, 0.840, and 0.812 by models trained only on the PARS-Human, PARS-Yeast and SS-PDB training sets, respectively. Besides, it is important to note that, for the test set of SS-PDB data, the consensus model outperforms the models only trained on human or yeast data with increasements of AUC from less than 0.65 to nearly 0.90. These improvements indicate that the consensus model could eliminate the difference between two experimental techniques and is better for general prediction of RNA secondary structures. Moreover, we also assessed the performances on tRNA molecules. By randomly selecting 30% (6 chains) from the totally 21 tRNA chains in the SS-PDB as the test set, we re-trained the consensus model with the same hyper-parameters as above using the remaining data in the training set. The re-trained consensus model achieved an AUC of 0.887 on the 6 tRNA chains, higher than the 0.823 and 0.759 by RNAplfold and CROSS, respectively.

3.3 Independent test on SS-ZIKV

We further tested our consensus model on the recently released secondary structure of the Zika virus measured by the icSHAPE technique(Li, et al., 2018). Though the training of our consensus model didn't include dataset by such technique, the model made a prediction with an AUC of 0.888 (Figure 2) on the SS-ZIKV dataset. By comparison, though the CROSS (global) method has included two

Table 5. Comparison of AUC values on the independent tests consisting of 10% samples by models trained with 90% samples of each dataset or their combination (Consensus).

train test (10%)	Consensus model	PARS-Human	PARS -Yeast	SS-PDB
PARS-Human	0.940	0.945	0.925	0.754
PARS-Yeast	0.960	0.945	0.967	0.773
SS-PDB	0.886	0.648	0.628	0.908
Average	**0.927**	0.846	0.840	0.812

datasets by SHAPE and icSHAPE experimental techniques in their model training, their final model reached an AUC of 0.840 that is 5.7% lower than our method. The difference of two AUC values was significant (P-value<1E-6) according to the statistical test(Hanley and McNeil, 1982; Lowry). The RNAplfold achieved the lowest AUC of 0.799 that is 11% lower than the one by our method.

3.4 Relation of predicted secondary structure with the MAF of genetic variants

To demonstrate the biological significance of our predicted secondary structure profiles, we examined whether the predicted paired states of nucleotides were related with the minor allele frequencies (MAF) of genetic variants observed from the 1000 Genomes Project for healthy individuals(Huang, et al., 2012).

As shown in Fig 3A, the unpaired probabilities (i.e. 1 - paired probability) predicted by GRASP showed high correlations with MAF for most types of mutations, with the highest Pearson's correlation coefficient (PCC) of 0.901 from synonymous mutations. This is probably because synonymous mutations that don't change expressed proteins affect biological functions mainly through the change of RNA secondary structure. The correlation is especially clear for the paired states (left region in Fig 3B) as mutations of paired nucleotides mostly destroy the paired states and likely cause diseases. Relatively, the points close to 0 (unpaired states) are less related as mutations of unpaired nucleotides are

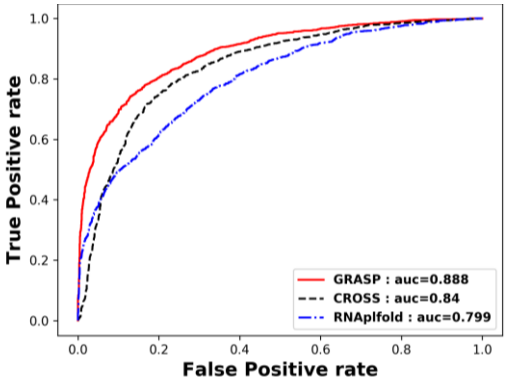

Figure 2. Receiver Operating Characteristic (ROC) curves reveal test performances of three models on Zikv RNA genomic prediction. GRASP consensus model performs the best among all models.

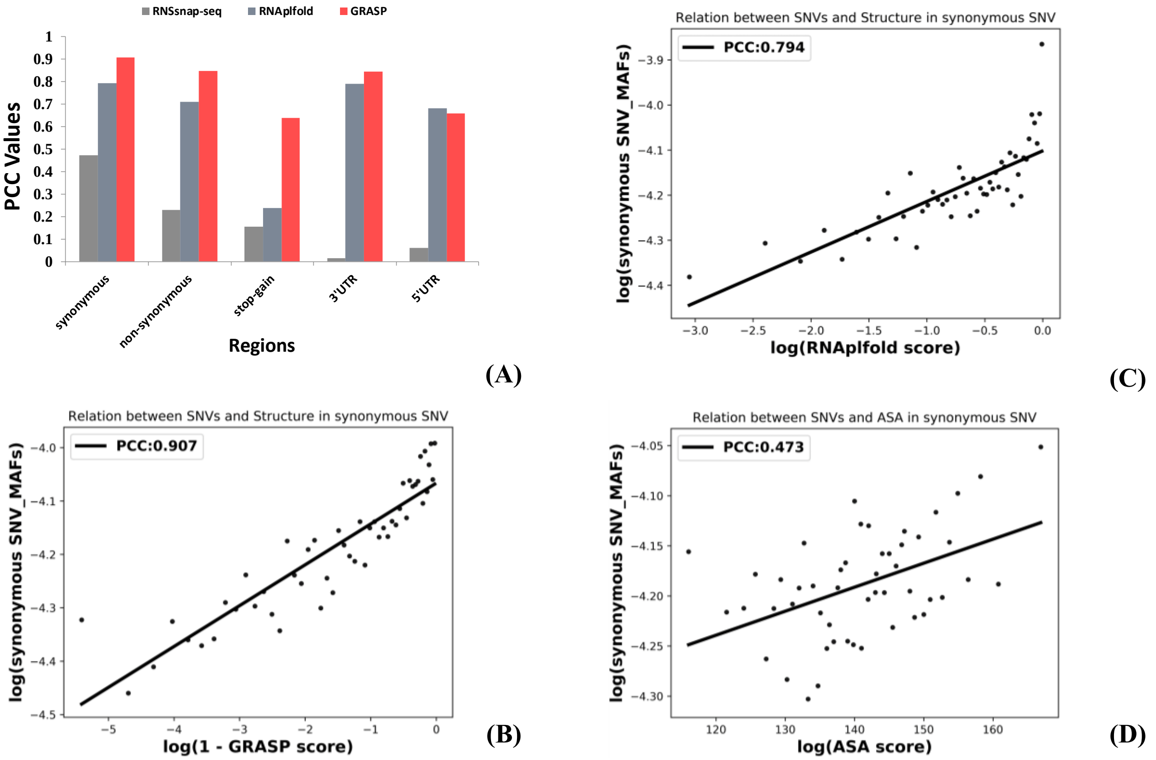

Figure 3. Positive association between minor allele frequencies (MAF) of genetic variants and predicted secondary structures or ASA at mutation sites. (A) Pearson's correlation coefficients between the average MAF of single-nucleotide variations and unpaired probabilities by RNAplfold, unpaired probabilities by GRASP, and ASA(accessible surface area) by RNAsnap-seq for synonymous, nonsynonymous, and stop-gain mutations at the coding region, mutations at 3′ and 5′ untranslated regions, respectively. For synonymous mutations at the coding region, the relation of the average MAF from the 1000 Genomes Project with the average of (B) the unpaired probability predicted by GRASP, (C) the unpaired probability by RNAplfold, and (D) predicted ASA by RNAsnap-seq. The average was calculated over bins sorted by predicted values. PCC values are as labelled.

less probable to change the paired states. This was confirmed by the PCC of -0.925 when using the unpaired probabilities through mutated sequences. When using the changes of the unpaired probabilities, there is still a strong negative correlation, but the correlation is slightly weaker (PCC=-0.853). This is consistent with our previous findings that the prediction of overall states is still limited to reflecting the differences by mutations(Chen, et al., 2019; Yang, et al., 2017). The predictions by RNAplfold achieved a PCC of 0.794 that's greater than the PCC of 0.473 with the predicted ASA(accessible surface area) from the RNAsnap-seq, consistent with the previous study (Yang, et al., 2017). This ranking order is consistent with most types of mutations, non-synonymous mutations, stop-gain mutations, and mutations occurring in the 3'UTR (untranslated region). (Figure 3 and Figure S1-S3 in supplemental file).

3.5 Predicted secondary structure plays as a key signal in translation

To further explore the potential function of the secondary structure in translation process, we explored distribution of paired probability in coding area of mRNAs. As shown in Figure 4, the paired probability predicted by GRASP reveals a three-nucleotide periodicity across coding regions. In each codon, the first nucleotide is always least structured and the second one is more structured than the other two, similar to previously observation(Kertesz, et al., 2010; Wan, et al., 2014). This vibration frequency indicates there is an unambiguous definition of codon boundaries during translation process. The results of RNAplfold also shows periodicity, but not so obvious. Moreover, it was shown that there was a sudden drop and then fast rise in paired probability near the start codons as well as stop codons. The curve of GRASP forms a deeper bottom than that of RNAplfold (Figure 4). Namely, both the starting site and ending site tend to be unpaired. This is consistent with the previous finding that over 80% of the start codon are free of secondary structure by analyzing mRNAs of prokaryotic and eukaryotic(Ganoza and Louis, 1994), as well as free energy preference at the 5' and 3' ends of siRNA(Shabalina, et al., 2006). This kind of enrichment of unpaired nucleotides can help to start the protein translation process. Additionally, the transition of secondary structure over the boundary is likely an important signal for a correct recognition of coding regions.

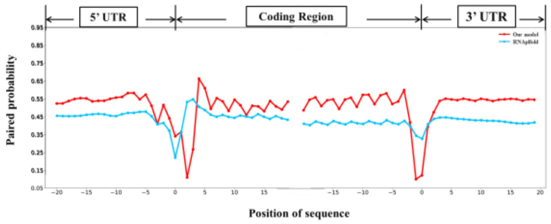

Figure 4. GRASP shows clearer patterns at different regions in transcripts than RNAplfold. It can be seen apparently that the average paired probability at coding sequence (CDS) is periodical distribution in unit of a codon, which is distinct from the pattern at untranslated regions (UTR). The result of GRASP shows a clearer vibration frequency than RNAplfold. At the two ends of CDS, GRASP also predicts a more significant and deeper bottom than RNAplfold. The results demonstrate the better biological utility of our model.

4 Discussion

In this study, we have developed a new method GRASP to predict RNA secondary structure from sequence based on XGBoost. To train the model, we used sequence information around a given nucleotide. We found that a window size of 37 nucleotides provided the best performance, as shown in Table S1 and Figure S4. For example, when model trained by PARS-Human was applied to PARS-Yeast, the AUC values increased from 0.932 to 0.944 when the window size augments to 37, and also increased from 0.913 to 0.926 when model trained by PARS-Yeast was applied to PARS-Human. Taking the average AUCs of cross-tests as concerned(Figure S4), among three datasets, the values increased significantly when window size increased from 13 to 37, but the growth stopped and a decrease trend appeared after 37. Ideally, the window size should cover the entire sequence of an RNA chain so that a machine-learning method can learn potential interactions between all nucleotides (local or nonlocal interactions). However, the growth in the number of features is easy to cause over-training due to limited number of training samples, and will also significantly increase the computational costs during model training and prediction. As a balance performance and computational costs, we chosen a window size of 37.

We observed that the cross-tests between two PARS datasets achieved AUCs above 0.9, close to their respective self-test performances. Nonetheless, models trained by these datasets had a much lower performance on the SS-PDB dataset with AUCs around 0.65. Similarly, the model trained by SS-PDB did not perform well on predictions of two PARS datasets with AUCs around 0.75, much lower than the self-test result on SS-PDB. The divergences might result from the different techniques to produce the datasets. SS-PDB dataset was derived from 3D structure determined by X-ray or NMR, reflecting an in-vitro structural states, whereas PARS measured the paired or unpaired states of nucleotides by their reactions with chemical reactants. As a compromise, our consensus model trained on both types of data achieved the best performance for all tests. Though the consensus model only included experimental data by PARS technique and PDB data, it achieved the best results on the independent test set of the Zika virus RNA genomics measured by icSHAPE technique, indicating the robustness of our model.

GRASP was further validated by using 2.2 million genetic variants found in the 1000 Genomes Project. Previous studies assumed that the higher populated genetic variants on average, the less association with diseases(Hu and Ng, 2012; Zhao, et al., 2013). This expectation effectively supported for the predicted disease susceptibility of genetic variants(Yang, et al., 2017). Therefore, if mutation-induced disruption of functional RNA structures is one of the potential trigger of disorders(Halvorsen, et al., 2010), it is expected that predicted probability scores of structures at mutation sites would have a positive correlations with average MAF values, similar to predicted RNA solvent accessibility(Zhao, et al., 2013).The expectation was proved by strong correlation between unpaired nucleotides and higher allele frequencies by a PCC > 0.8 in the mutation sites located in coding region, 3′ UTR. The relative weak correlation for 5'UTR and stop-gain mutations by both

GRASP and RNAplfold suggests that these kinds of mutations had different functional mechanism in genomes. Moreover, it could be found that the PCCs between MAF and GRASP were stronger than the relations between MAF and RNAplfold, which suggests that the machine-based method by training on high-throughput experimental data was better to characterize the genomes than analytical methods fitted for small data. With the decreasing costs in sequencing technique, more and more genomic data will be obtained, and machine learning models can be constructed to remove experimental noise and to extend into regions with low or no experimental coverage. Besides, the stronger correlation to MAF than predicted ASA also suggests that mutations in paired regions are more disruptive than those in buried RNA regions (the smaller predicted ASA, the more buried).

The positive results by correlating with MAF of genetic variants in 1000-Genomes Project encouraged us to make genome-scale application of GRASP to more than 18000 genes for identifying potential mechanism under translation process. It was found that the average paired probability shows periodical distribution in the codons of CDS. The 2nd nucleotides in codon always had higher paired probability in average than the 1st and 3rd nucleotides. This periodic fluctuation didn’t appear in the 3’UTR or 5’UTR. Moreover, near the start codons and end codons, the cliff-like curves of paired probability indicate that both ends of coding region are less structured, which is consistent with the need to interact with the ribosome for translation. In order to prevent the possible artifact caused by the selective window size, we also ran RNAplfold with the default window size (70), and found that the patterns remained highly similar (Figure S5). These results further suggested that our predictions help to unearth more biologically meaningful results, which might support downstream analysis such as protein function, protein–ligand interactions(Chen, et al., 2019) .

Our method has been trained based on XGBoost, which supports parallel computing that is advantageous for post-deployment on the super computer for large-scale calculation and public use. And the overall performance by the method indicates that fitting to the high throughput experimental data might be a substitution for analytical methods.

GRASP is now freely available for academic use at GitHub: https://github.com/sysu-yanglab/GRASP.

Funding

This work has been supported by the National Key R&D Program of China (2018YFC0910500), National Natural Science Foundation of China (U1611261, 61772566, and 81801132), Guangdong Frontier & Key Tech Innovation Pro-gram (2018B010109006, 2019B020228001) and Introducing Innovative and Entrepreneurial Teams (2016ZT06D211).

Conflict of Interest: none declared.

References

Basit, A.H., *et al.* Training host-pathogen protein-protein interaction predictors. *J Bioinform Comput Biol* 2018:1850014.

Bernhart, S.H., Hofacker, I.L. and Stadler, P.F. Local RNA base pairing probabilities in large sequences. *Bioinformatics* 2006;22(5):614-615.

Chen, K., *et al.* Predicting the change of exon splicing caused by genetic variant using support vector regression. *Hum Mutat* 2019;40(9):1235-1242.

Chen, P., *et al.* DLIGAND2: an improved knowledge-based energy function for protein–ligand interactions using the distance-scaled, finite, ideal-gas reference state. 2019;11(1):52.

Chen, T. and Guestrin, C. XGBoost:A Scalable Tree Boosting System. In, *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. p. 785-794.

Chen, X., *et al.* EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. *Cell Death Dis* 2018;9(1):3.

Dhaliwal, S.S., Nahid, A.A. and Abbas, R. Effective Intrusion Detection System Using XGBoost. *Information* 2018;9(7):149.

Ding, Y., *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 2014;505(7485):696-700.

Ganoza, M.C. and Louis, B.G. Potential secondary structure at the translational start domain of eukaryotic and prokaryotic mRNAs. *Biochimie* 1994;76(5):428-439.

Glisovic, T., *et al.* RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 2008;582(14):1977-1986.

Hanley, J.A. and McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.

Hofacker, I.L. Energy-directed RNA structure prediction. *Methods Mol Biol* 2014;1097:71-84.

Hu, J. and Ng, P.C. Predicting the effects of frameshifting indels. *Genome Biol* 2012;13(2).

Huang, J., *et al.* 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur J Hum Genet* 2012;20(7):801-805.

Jin-yue, L. and Bao-ling, Z. Application of BP neural network based on GA in function fitting. In, *Proceedings of 2012 2nd International Conference on Computer Science and Network Technology*. 2012. p. 875-878.

Kertesz, M., *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* 2010;467(7311):103-107.

Li, P., *et al.* Integrative Analysis of Zika Virus Genome RNA Structure Reveals Critical Determinants of Viral Infectivity. *Cell Host Microbe* 2018;24(6):875-886 e875.

Li, W. and Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22(13):1658-1659.

Lorenz, R., *et al.* ViennaRNA Package 2.0. *Algorithm Mol Biol* 2011;6.

Lowry, R. VassarStats: Website for Statistical Computation.

Lucks, J.B., *et al.* Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci U S A* 2011;108(27):11063-11068.

Lyngso, R.B. and Pedersen, C.N. RNA pseudoknot prediction in energy-based models. *J Comput Biol* 2000;7(3-4):409-427.

Mathews, D.H., *et al.* Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 1999;288(5):911-940.

Mendik, P., *et al.* Translocatome: a novel resource for the analysis of protein translocation between cellular organelles. *Nucleic Acids Res* 2019;47(D1):D495-D505.

Mortimer, S.A., Kidwell, M.A. and Doudna, J.A. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* 2014;15(7):469-479.

Ouyang, Z., Snyder, M.P. and Chang, H.Y. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res* 2013;23(2):377-387.

Pedregosa, F., *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;12:2825-2830.

Ponti, R.D., *et al.* A high-throughput approach to profile RNA structure. *Nucleic Acids Res* 2017;45(5).

Roberts, P.D. Two-dimensional analysis of a gradient method in function space optimal control algorithm. In, *42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475)*. 2003. p. 1212-1217 Vol.1212.

Rouskin, S., *et al.* Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* 2014;505(7485):701-705.

Seetin, M.G. and Mathews, D.H. RNA structure prediction: an overview of methods. *Methods Mol Biol* 2012;905:99-122.

Shabalina, S.A., Spiridonov, A.N. and Ogurtsov, A.Y. Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics* 2006;7(1):65.

Underwood, J.G., *et al.* FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods* 2010;7(12):995-1001.

Wan, Y., *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 2014;505(7485):706-709.

Wang, Z., Gerstein, M. and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10(1):57-63.

Yang, Y.D., *et al.* Genome-scale characterization of RNA tertiary structures and their functional impact by RNA solvent accessibility prediction. *RNA* 2017;23(1):14-22.

Ye, D., Yu, C.C. and Lawrence, C.E. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *Rna-a Publication of the Rna Society* 2005;11(8):1157-1166.

Zhao, H.Y., *et al.* DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biol* 2013;14(3).

Zou, L.S., *et al.* BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC Genomics* 2018;19(1):390.