Prediction of adenosine to inosine RNA editing sites

Abstract

RNA editing process like adenosine to inosine (A-to-I) editing is one of the prominent kinds of RNA editing which helps us to understand the biological functions properly. In this paper, we have applied a machine learning or data-centric approach for the prediction of A-to-I RNA editing sites. We have used simple Frequency count, pseudoKNC, cumulativeSkew, atgcRatio, gcContent and z-curve as features.We generated total 1375 features to train our model. Several algorithms (KNN, SVM, RF, Ada, Naive Bayes, Decision Tree, Logistic Regression, etc.) were applied on the benchmark dataset using a 10-fold cross-validation method. Support Vector machine obtained best result for our experiement.Our SVM model achieved an accuracy of 81.5580% with a sensitivity of 91.2%, specificity of 71.4286% and MCC of 0.6542.

1 Introduction

RNA editing is the process of modifying RNA nucleotides to change the amino acid sequence. The adenosine to inosine (A-to-I) editing is one of the prominent kinds of RNA editing which involves in various biological processes. As a result, the accurate identification of A-to-I editing site is important for researchers to understand biological functions properly.

Researchers are generating a large amount of biological data directly from biological experiments. So it is impossible to analyze these data manually. We need a convenient way to extract information from these data. This is where computational methods come into play. Researchers can use various computational methods and algorithms to find out important insights from this large amount of data automatically.

In this paper, we tried to solve the problem of finding the a-to-i RNA editing sites by using a machine learning or data-centric approach. At first, we collected the standard benchmark RNA sequence data of *D. melanogaster*. Then we generated Frequency count, pseudoKNC, cumulativeSkew, atgcRatio, gcContent and z-curve as features from the dataset. Several algorithms (KNN, SVM, RF, Ada, Naive Bayes, Decision Tree, Logistic Regression, etc.) were applied on

the benchmark dataset using a 10-fold cross-validation method. Support Vector machine obtained best result for our experiement.Our SVM model achieved an accuracy of 81.5580% with a sensitivity of 91.2%, specificity of 71.4286% and MCC of 0.6542.

2 Related Work

Computational biologists have been studying RNA editing for a long time. Machine learning has been applied in a number of studies to address the prediction of adenosine to inosine RNA editing sites [1–5].

In 2016, Chen et al [1] proposed a support vector machine based-model, called PAI, to predict A-to-I editing sites using pseudo nucleotide compositions of D. melanogaster. They used a benchmark dataset and independent dataset for measuring the performance of their model. Benchmark dataset was built based on the work of Laurent et al [6]. It contains 244 samples of RNA sequence data of the D. melanogaster. Among them, 125 items are positive samples and 119 items are negative samples. The length of each sequence is 51 with the Adenosine nucleotide in the middle denotes the modification site. The independent dataset was built by harvesting the RNA sequence data obtained from Yu and his colleagues' work [7]. It contains 300 positive samples of RNA sequence data of the D. melanogaster. It is to be noted that there are no negative samples in the independent dataset. However, each sequence is 51 nucleotides long with the Adenosine nucleotide in the middle. The researcher used the jackknife test to examine the performance of PAI to identify the A-to-I editing sites on both the dataset. PAI was applied on the benchmark dataset and it obtained an accuracy of 79.51% with a sensitivity of 85.60%, specificity of 73.11% and MCC of 0.60. PAI was also applied to the independent dataset and it was able to correctly identify 247 A-to-I editing sites out of 300. The sensitivity of the model for the independent dataset is 82.33%. Researchers can freely access The PAI web-server at http://lin-group.cn/server/PAI.

Subsequently, in 2017, Chen et al [2] proposed another support vector machine based-model, called iRNA-AI, to identify the adenosine to inosine editing sites in RNA sequences of humans. They used a benchmark dataset and independent dataset for measuring the performance of their model. Benchmark dataset was derived from the DARNED database of kiran et al [8]. It contains 6000 samples of RNA sequence data of the human. Among them, 3000 items are positive samples and 3000 items are negative samples. The length of each sequence is 51 with the Adenosine nucleotide in the middle denotes the modification site. The independent dataset was also derived from the DARNED database of kiran et al [8]. It contains 3,243 positive samples and 3,243 negative samples of RNA sequence data of human. It is to be noted that none of the samples in the independent dataset occurs in the benchmark dataset and each sequence is 51 nucleotides long with the Adenosine nucleotide in the middle. The researcher used the cross-validation test to examine the performance of iRNA-AI to identify the A-to-I editing sites on a benchmark dataset. iRNA-AI was

applied on the benchmark dataset and it obtained an accuracy of 90.71% with a sensitivity of 86.18%, specificity of 95.23% and MCC of 0.82. iRNA-AI was also applied on the independent dataset and obtained an accuracy of 93.81% with a sensitivity of 84.19%, specificity of 89.36% and MCC of 0.80. Researchers can freely access The iRNA-AI web-server at http://lin-group.cn/server/iRNA-AI/.

Later on, in 2018, Chen et al [3] proposed a support vector machine based-model, called iRNA-3typeA, to Identify Three Types of Modification at RNA's Adenosine Sites. They used two benchmark datasets that contain RNA sequences from Homo sapiens and Mus musculus transcriptomes respectively. The benchmark datasets were derived from the previous work [2, 9, 10]. The benchmark dataset of Homo sapience contains three subsets. The first subset contains data(6,366 Positive samples and 6,366 Negative samples) for analyzing m1A modification, the second subset contains data(1,130 Positive samples and 1,130 Negative samples) for analyzing m6A and the third subset contains data(3,000 Positive samples and 3,000 Negative samples) for A-to-I editing. The second benchmark dataset of Mus musculus contains three subsets also. The first subset contains data(1,064 Positive samples and 1,064 Negative samples) for analyzing m1A modification, the second subset contains data(725 Positive samples and 725 Negative samples) for analyzing m6A and the third subset contains data(831 Positive samples and 831 Negative samples) for A-to-I editing. The length of each sequence is 41 with the Adenosine nucleotide in the middle. The researcher used the jackknife test to examine the performance of iRNA-3typeA to Identify three types of modification at RNA's Adenosine Sites on both the dataset. iRNA-3typeA was applied on the homo sapiens benchmark dataset and it obtained an accuracy of 99.13% with a sensitivity of 98.38%, specificity of 99.89% and MCC of 0.98 for identification of m1A modification type. It also obtained an accuracy of 90.38% with a sensitivity of 81.68%, specificity of 99.11% and MCC of 0.82 for identification of m6A modification type. For identification of a-to-I modification type, it obtained an accuracy of 90.71% with a sensitivity of 86.18%, specificity of 95.23% and MCC of 0.82. iRNA-3typeA was also applied on the M. musculus benchmark dataset and it obtained an accuracy of 98.73% with a sensitivity of 97.46%, a specificity of 100% and MCC of 0.97 for identification of m1A modification type. It also obtained an accuracy of 88.39% with a sensitivity of 77.79%, specificity of 100% and MCC of 0.80 for identification of m6A modification type. For identification of a-to-I modification type, it obtained an accuracy of 98.38% with a sensitivity of 96.75%, specificity of 100% and MCC of 0.96. Researchers can freely access the iRNA-3typeA web-server at http://lin-group.cn/server/iRNA-3typeA/

In 2018, Xiao et al [4] tried to improve the accuracy of identifying A-to-I editing sites by proposing a new support vector machine based-model, called PAI-SAE. PAI-SAE used the same dataset that was used for PAI [1]. This research introduces new hybrid features by combining dinucleotide-based auto-cross covariance (DACC), pseudo dinucleotide composition (PseDNC) and nucleotide density by using spare auto-encoder. Researchers used the cross-validation test to examine the performance of PAI-SAE to identify the A-to-I editing sites on the benchmark dataset. It obtained an accuracy of 81.97% with a sensitivity

of 87.20%, specificity of 76.47% and MCC of 64.14 for the benchmark dataset. PAI-SAE gains 2.46% higher accuracy than the PAI model.

In 2019, Ahmad et al [5] proposed a new enhanced support vector machine based-model, called EPAI-NC to predict A-to-I RNA editing sites using nucleotide compositions. EPAI-NC used the same dataset that was used for PAI [1]. This model used two sequence based features (composition of l-mers and n-gapped l-mers) for training and evaluation of the model. 10 fold cross-validation method was applied on the benchmark dataset and EPAI-NC achieved an accuracy of 93.90% with a sensitivity of 96.80%, specificity of 90.80% and MCC of 87.90. EPAI-NC correctly identified 253 A-to-I RNA editing sites out of 300 for the independent dataset. Researchers can freely access the iRNA-3typeA web-server at http://epai-nc.info.

3 Materials and methods

This section provides the details of materials and methods in this paper.Here we have used the famous five step rule by Kuo-Chen Chou.Firstly we extracted five features(Simple Frequency, pseudoKNC, cumulativeSkew, atgcRatio,gcContent and z-curve) from the benchmark dataset. The benchmark dataset cotains the RNA sequence data of *D. melanogaster*. Then we used SVM to train the model and the best result was saved to present in this paper. The block diagram of this paper is given in Fig: 1 :

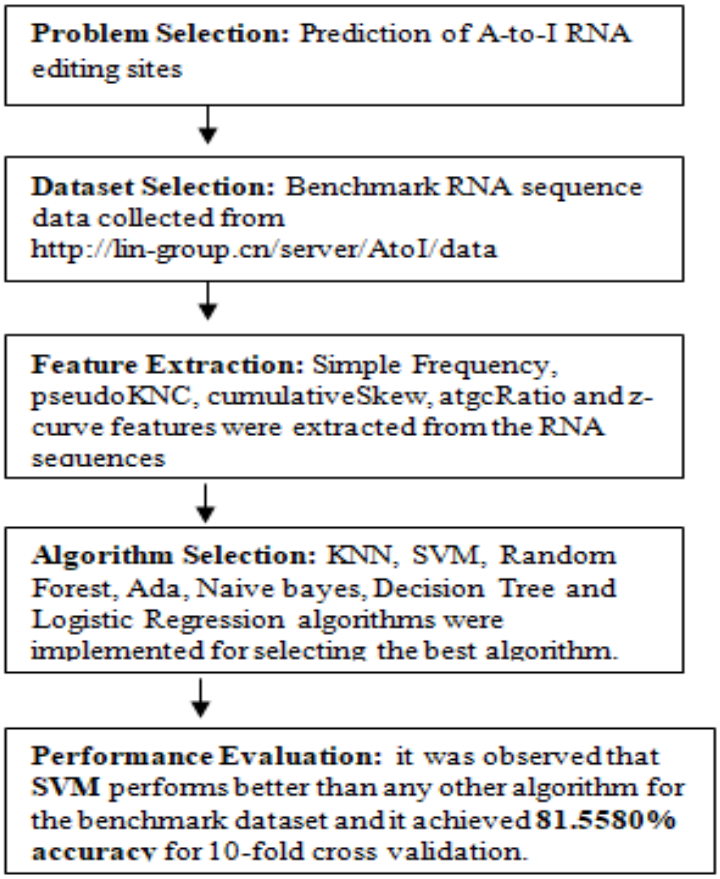

Figure 1: block diagram of working procedure

3.1 Dataset

In this paper, we have used two datasets (benchmark dataset and independent dataset) for our experiment. Benchmark dataset contains 244 samples of RNA sequence data of the D. melanogaster. Among them, 125 items are positive samples and 119 items are negative samples. The length of each sequence is 51 with the Adenosine nucleotide in the middle denotes the modification site. On the other hand, the independent dataset contains 300 positive samples of RNA sequence data of the D. melanogaster and each sequence is 51 nucleotides long with the Adenosine nucleotide in the middle. It is important to note that there are no negative samples in the independent dataset. All the Datasets have been collected from the PAI experiment [1].

3.2 Features

After selecting the benchmark dataset we have used the following feature extraction methods for generating features for our algorithm:

Frequency feature: RNA sequence contains A, C, G and U. we simply counted the frequency of A, C, G and U in the sequence and used it as a feature for our model. So the total number of features in this category is four.

pseudoKNC feature: pseudoKNC was calculated from the composition of A,C,G and U. When k=1, feature structure will be X. When k=2, feature structure will be X, and XX. When k=3, feature structure will be X, XX, and XXX. In this paper we used the value for k is 5. So the total number of features in this category is $4 + 4^2 + 4^3 + 4^4 + 4^5 = 1364$

cumulativeSkew feature: Due to deamination process there is a difference of the count of G and U in forward and reverse strands. The forward strand often have more G and U. The cumulative skew is defined formally as:

$$GC - Skew = \frac{\sum G - \sum C}{\sum G + \sum C}; AT - Skew = \frac{\sum A - \sum U}{\sum A + \sum U}; \tag{1}$$

So the total number of features in this category is two.

atgcRatio feature: Single feature will be generated in this category. The equation is given below:

$$AT/GCRatio = \frac{\sum A + \sum U}{\sum G + \sum C} \tag{2}$$

gcContent feature: Single feature will be generated in this category. The equation is given below:

$$gcContent = \frac{\sum G + \sum C}{\sum A + \sum C + \sum G + \sum U} * 100\% \tag{3}$$

z-curve feature: Z-curve theory is often used in sequence analysis. It has got three components in three axis.They are defined as: $x - axis = (\sum A + \sum G) - (\sum C + \sum U); y - axis = (\sum A + \sum C) - (\sum G + \sum U); z - axis = (\sum A + \sum U) - (\sum G + \sum C)$; So the total number of features in this category is three.

3.3 Algorithms

Here We have used several algorithms for selecting the best algorithm for our experiment. We used KNN,SVM,Random Forest, Ada, Naive bayes, Decision Tree and Logistic Regression algorithms and found that SVM is performing better for our experiment. SVM was applied on the benchmark dataset and it obtained an accuracy of 81.5580% with a sensitivity of 91.2%, specifcity of 71.4286% and MCC of 0.6542. We used 'rbf' kernel type with regularization parameter value 15 for SVM and enabled the probability estimates. We executed our training with anaconda platform on an Intel Core i5 processor. Python 3.7 version was used for our experiment.

3.4 Performance Evaluation

To measure the performance of machine learning algorithms we used a number of performance metrics:Accuracy(Acc), Sensitivity (S_n), Specificity (S_p), Mathew's Correlation Coefficient (MCC), Receiver Operating Characteristic (ROC), Precision-Recall curves and F1-Score. These performance metrics are dependent on the following things:

True Positives (TP): It indicates the number of correctly identified positive instances.

True Negatives (TN): It denotes the number of correctly identified negative instances.

False Positives (FP): It is the number of incorrectly identified positive samples.

False Negatives (FN): It is the number of incorrectly identified negative samples.

Based on the above things the other measures are defined as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$S_n = \frac{TP}{TP + FN} \tag{5}$$

$$S_p = \frac{TN}{TN + FP} \tag{6}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

Here, Accuracy is the ratio of correctly identified and all instances.Precision is the ratio of truly positive and all positive samples. Recall is also known as sensitivity. It is the ratio of correctly identified true positive and all true positive samples. F1-Score is defined as the following:

$$F1 - Score = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{8}$$

ROC is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. All these metrics accept MCC has got values in the range [0,1]. Here 1 means a best classifier and 0 is the worst classifier. MCC has got the values in the range of [-1,1]. Performance of the different classifiers on the benchmark dataset are displayed in the Table 1.

4 Experimental Analysis

All the classifiers are probabilistic in the nature of their output and depends on the threshold. The default threshold is 0.50. By changing the threshold might change the nature and performance metrics of the classifier. AUC or

Algorithm	Accuracy	auROC	auPR	F1_Score	MCC	Recall	Sensitivity	Specificity
KNN	49.6159%	0.6253	0.6279	0.0451	0.0401	0.0244	2.4000%	99.1597%
SVM	81.5580%	0.8660	0.8639	0.8356	0.6542	0.9122	91.2000%	71.4286%
RF	80.2203%	0.8228	0.7943	0.8120	0.6144	0.8404	84.0000%	76.4706%
Ada	74.1594%	0.7852	0.7830	0.7531	0.4934	0.7776	77.6000%	70.5882%
NB	65.9928%	0.6549	0.6173	0.7160	0.3347	0.8295	83.2000%	47.8992%
DT	69.6580%	0.6975	0.6569	0.6929	0.4018	0.6731	67.2000%	72.2689%
LR	78.0138%	0.8630	0.8663	0.7905	0.5702	0.8199	81.6000%	73.9496%

Table 1: Performance of different classifiers on the benchmark Dataset

area under curve for ROC is one of the measures that is not dependent on the thresholds. So,we generated the AUC or area under curve for ROC to analyze the performance of the classifiers. This curve was generated by the metrics described in the performance evaluation section. Figure 2 shows the AUC or area under curve for ROC.

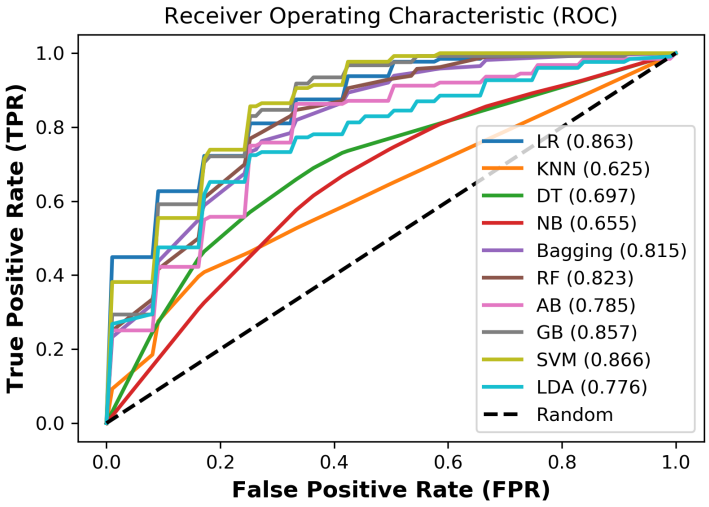

Figure 2: ROC curve

It is evident that SVM performed better for the benchmark dataset as compared to the other classifiers and achieved an accuracy of 81.5580% with a sensitivity of 91.2%, specifcity of 71.4286% and MCC of 0.6542. In order to understand the performance easily we also generated a bloxpot graph based on the accuracy of all the classifiers which is shown in figure: 3.

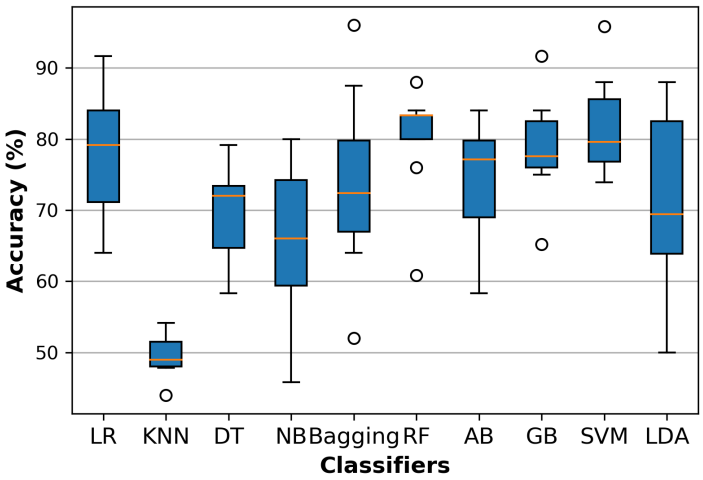

Figure 3: Boxplot

5 Conclusion

In this paper, we presented a model for prediction of adenosine to inosine RNA editing sites. We have used sequenced based features based on the frequency of A,C,G,U,pseudoKNC,cumulativeSkew,atgcRatio,gcContent and z-curve. We did not try other sequenced based features like composition of l-mers and n-gapped l-mers to train our model. At the same time we also did not use any feature selection method for which we believe that our model is not performing as good as the EPAI-NC [1, 5] model. In future, we wish to develop a generic website and database for other types of RNA editing and use several types of nucleotide composition features for our model.

References

[1] Wei Chen, Pengmian Feng, Hui Ding, and Hao Lin. Pai: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions. *Scientific reports*, 6:35123, 2016.

[2] Wei Chen, Pengmian Feng, Hui Yang, Hui Ding, Hao Lin, and Kuo-Chen Chou. irna-ai: identifying the adenosine to inosine editing sites in rna sequences. *Oncotarget*, 8(3):4208, 2017.

[3] Wei Chen, Pengmian Feng, Hui Yang, Hui Ding, Hao Lin, and Kuo-Chen Chou. irna-3typea: identifying three types of modification at rna's adenosine sites. *Molecular Therapy-Nucleic Acids*, 11:468–474, 2018.

[4] Xuan Xiao, Peng Wang, Zhaochun Xu, Wangren Qiu, and Xinzhu Fang. Pai-sae: Predicting adenosine to inosine editing sites based on hybrid features by using spare auto-encoder. In *IOP Conference Series: Earth and Environmental Science*, volume 170, page 052018. IOP Publishing, 2018.

[5] Ahsan Ahmad and Swakkhar Shatabda. Epai-nc: Enhanced prediction of adenosine to inosine rna editing sites using nucleotide compositions. *Analytical biochemistry*, 569:16–21, 2019.

[6] Georges St Laurent, Michael R Tackett, Sergey Nechkin, Dmitry Shtokalo, Denis Antonets, Yiannis A Savva, Rachel Maloney, Philipp Kapranov, Charles E Lawrence, and Robert A Reenan. Genome-wide analysis of a-to-i rna editing by single-molecule sequencing in drosophila. *Nature structural & molecular biology*, 20(11):1333, 2013.

[7] Yao Yu, Hongxia Zhou, Yimeng Kong, Bohu Pan, Longxian Chen, Hongbing Wang, Pei Hao, and Xuan Li. The landscape of a-to-i rna editome is shaped by both positive and purifying selection. *PLoS genetics*, 12(7):e1006191, 2016.

[8] Anmol Kiran and Pavel V Baranov. Darned: a database of rna editing in humans. *Bioinformatics*, 26(14):1772–1776, 2010.

[9] Wei Chen, Hua Tang, and Hao Lin. Methyrna: a web server for identification of n6-methyladenosine sites. *Journal of Biomolecular Structure and Dynamics*, 35(3):683–687, 2017.

[10] Wei Chen, Pengmian Feng, Hua Tang, Hui Ding, and Hao Lin. Rampred: Identifying the n 1-methyladenosine sites in eukaryotic transcriptomes. *Scientific reports*, 6:31080, 2016.