PAI-SAE: Predicting Adenosine To Inosine Editing Sites Based On Hybrid Features By Using Spare Auto-Encoder

Xuan Xiao[1, 2, *], Peng Wang[1, a], Zhaochun Xu[1, b], Wangren Qiu[1, 3, c] and Xinzhu Fang[1, d]

[1] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403 China
[2] Gordon Life Science Institute, Boston, Massachusetts 02478, United States of America, USA
[3] Department of Computer Science and Bond Life Science Center, University of Missouri, Columbia, MO, USA

*Corresponding author: jdzxiaoxuan@163.com; [a]1182643454@qq.com; [b]jdzxuzhaochun@163.com; [c]qiuone@163.com; [d]623359878@qq.com

Abstract. Adenosine-to-inosine RNA editing is an important post-transcriptional modification, which converts adenosines to inosines in both coding and noncoding RNA transcripts. Therefore, this modification can result in the diversification of the transcriptome. It is significant to accurately identify adenosine-to-inosine editing sites for further understanding their biological functions. Given an uncharacterized RNA sequence that contains many adenosine resides, can we identify which one of them can be converted to inosine, and which one cannot? To meet the increasingly high demand form most experimental scientists working in the area of drug development, we have developed a new predictor called PAI-SAE by hybrid features combining with dinucleotide-based auto-cross covariance (DACC), pseudo dinucleotide composition (Pse DNC) and nucleotide density, followed by a spare auto-encoder model. It has been observed via rigorous jackknife test that the predictor PAI-SAE is superior to others in this area.

1. Introduction

RNA editing is a post-transcriptional process, selectively inserting and deleting single nucleotide, or converting one nucleotide to another [1]. There are two major types of RNA editing in mammals: one is C-to-U (cytidine to uracil), the other, a much more common type, is A-to-I (adenosine to inosine) [2]. A-to-I editing usually takes place under the control of the enzyme ADARs (adenosine deaminases that act on RNA) that bind dsRNA (double-stranded RNA) structures [3]. In this catalytic process, a targeted adenosine (A) within these structures is deaminated into inosine (I), and inosine (I) can be recognized as guanosine (G), because of the similar functions to G by the cellular machinery [4]. Many biological mechanisms, such as RNA stability, localization, splicing, miRNA function and translation, are affected by the A-to-I editing event. Therefore, it is significant to accurately identify adenosine-to-inosine editing sites for further understanding their biological functions.

With the progress of RNA sequencing technology, identifying A-to-I editing sites has entered into the perspective of researchers. For example, the next-generation sequencing has been successfully used to identify hundreds of human A-to-I editing sites in non-Alu regions since 2009 [5]. And A-to-I editing sites were accurately identified in H.sapiens by transcriptome sequencing in 2012 and 2014 [4,6,7] . Following these works, A-to-I editing sites were successfully detected in M.musculus on the basis of RNA-Seq method [8].

Although great successes have been achieved in this regard, it is expensive and time-consuming to identify A-to-I editing sites by means of the standard laboratory methods. Facing the explosive growth of RNA sequences discovered in the postgenomic age, it is highly demanded to develop computational approach to help getting the information. Very recently, in a pioneering study, St Laurent et al. [9] proposed an interesting method to identify A-to-I editing sites in D. melanogaster via an iterative feedback loop of computational prediction and experimental validation. But no web-server has been provided for their method, and hence its practical application value is quite limited. For this reason, Chen et al. [10] proposed a prediction model "PAI" on the basis of support vector machine by using pseudo dinucleotide composition method to identify A-to-I editing sites in D. melanogaster in 2016. And the corresponding web-server was constructed. The next year, a predictor "iRNA-AI" [11] based on support vector machine was constructed by using the chemical properties of nucleotides and nucleotide density. In view of its importance and urgency, it is certainly worthwhile to further improve the prediction quality by introducing some novel approaches as elaborated below.

In this study, we constructed the new hybrid features by combining DACC [12], PseDNC [13,14] and nucleotide density [15] and spare auto-encoder [16,17] to develop a new predictor "PAI-SAE" to identify A-to-I RNA editing sites in D. melanogaster aimed at improving its Matthew correlation coefficient(MCC) and accuracy(ACC), the two most important and harshest metrics for predictor.

2. Materials and Methods

2.1. Benchmark Dataset

St Laurent [9] et al. Sequenced the RNAs of the D. melanogaster to carry out genome-wide studies of adenosine-to-inosine RNA editing with single molecular sequencing in 2013. Based on experimental data, after removing redundant sequences by using CD-HIT[18], Chen et al. [10] constructed the benchmark dataset S including subset S^+ composed of 125 adenosine-to-inosine editing site sequences and subset S^- composed of 119 non-adenosine-to-inosine editing site samples. The benchmark dataset for the current study can be formulated as:

$$S = S^+ \cup S^- \tag{1}$$

Where the symbol $\cup$ represents the union of the subsets.

2.2. Feature Extraction

2.2.1. Physicochemical properties of dinucleotides. RNA is composed of four different types of nucleotides i.e. A (adenosine), C (cytidine), G (guanosine), and U (uridine). Then, sixteen dinucleotides were denoted as AA, AC, AG, AU, CA,…, UU. However, each of sixteen dinucleotides has different physicochemical property. As RNA physicochemical properties are the most intuitive features of biochemical reactions, in this paper, the eleven involved physicochemical properties were: (1). PC^1:Shift ; (2) PC^2:Slide; (3) PC^3:Rise; (4) PC^4:Tilt; (5) PC^5:Roll; (6) PC^6:Twist; (7) PC^7:Stacking energy; (8) PC^8:Enthalpy; (9) PC^9:Entropy; (10) PC^{10}:Free energy; (11) PC^{11}:Hydrophilicity. And their original values for each dinucleotide are listed in Table 1.

Table 1. The original values of the eleven physicochemical properties for each RNA dinucleotide

Code	PC^1	PC^2	PC^3	PC^4	PC^5	PC^6	PC^7	PC^8	PC^9	PC^{10}	PC^{11}
GG	-0.01	-1.78	3.32	0.30	12.1	32.0	-11.1	-12.2	-29.7	-3.26	0.17
GA	0.07	-1.70	3.38	1.30	9.40	32.0	-14.2	-13.3	-35.5	-2.35	0.10
GC	0.07	-1.39	3.22	0.00	6.10	35.0	-16.9	-14.2	-34.9	-3.42	0.26
GU	0.23	-1.43	3.24	0.80	4.80	32.0	-13.8	-10.2	-26.2	-2.24	0.27
AG	-0.04	-1.50	3.30	0.50	8.50	30.0	-14.0	-7.60	-19.2	-2.08	0.08
AA	-0.08	-1.27	3.18	-0.80	7.00	31.0	-13.7	-6.60	-18.4	-0.93	0.04
AC	0.23	-1.43	3.24	0.80	4.80	32.0	-13.8	-10.2	-26.2	-2.24	0.14
AU	-0.06	-1.36	3.24	1.10	7.10	33.0	-15.4	-5.70	-15.5	-1.10	0.14
CG	0.30	-1.89	3.30	-0.10	12.10	27.0	-15.6	-8.00	-19.4	-2.36	0.35
CA	0.11	-1.46	3.09	1.00	9.90	31.0	-14.4	-10.5	-27.0	-2.11	0.21
CC	-0.01	-1.78	3.32	0.30	8.70	32.0	-11.1	-12.2	-29.7	-3.26	0.49
CU	-0.04	-1.50	3.30	0.50	8.50	30.0	-14.0	-7.60	-19.2	-2.08	0.52
UG	0.11	-1.46	3.09	1.00	9.90	31.0	-14.4	-7.60	-19.2	-2.11	0.34
UA	-0.02	-1.45	3.26	-0.20	10.7	32.0	-16.0	-8.10	-22.6	-1.33	0.21
UC	0.07	-1.70	3.38	1.30	9.40	32.0	-14.2	-10.2	-26.2	-2.35	0.48
UU	-0.08	-1.27	3.18	-0.80	7.00	31.0	-13.7	-6.60	-18.4	-0.93	0.44

2.2.2. Dinucleotide-Based Auto-Cross Covariance. With the development of computer technology, many feature vectors that are used to represent sample sequences would be directly generated by the web server, such as Pse-in-One [19], repRNA [20], and repDNA [21], without need to go through the complex mathematical details.

Open the Web page by clicking the link at http://bioinformatics.hitsz.edu.cn/Pse-in-One/ and click on the serve button, you can see three different efficient tools for feature extraction including PseDAC-General, PseRAC-General and PseAAC-General and choose the second one for RNA sequences. After selecting the mode dinucleotide-based auto-cross covariance (DACC) and corresponding above-mentioned eleven physicochemical properties, you can easily obtain the desired results. Then, the necessary parameter ‘lag’ must be set. Experiments show the best results can be obtained when the value of the parameter lag is 4.

Generally, a RNA sequence R can be expressed as

$$R = R_1R_2R_3R_4R_5R_6\cdots R_L \tag{2}$$

Where L represents the length of sequence R.

Then, in accordance with the above procedure, the sample sequence R can be formulated by a 484-dimensional feature vector shown as below.

$$R = [\xi_1 \quad \xi_2 \quad \cdots \quad \xi_u \quad \cdots \quad \xi_{484}]^T \tag{3}$$

The derivation process of the Eq. (3) was described in detail in references [12,22].

2.2.3. Pseudo Dinucleotide Composition (Pse DNC). According to the references [10,23], based on the above-mentioned eleven physicochemical properties, the sample sequence R can be defined as

$$R = [v_1 \quad \cdots \quad v_{16} \quad v_{17} \quad \cdots \quad v_{16+\lambda}]^T \tag{4}$$

Where λ, the number of sequence order correlation factors, is an integer and must be smaller than $L-1$. And each component can be formulated by

$$v_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} & 1 \le k \le 16 \\ \dfrac{\omega \theta_{k-16}}{\sum_{i=1}^{16} f_i + \omega \sum_{j=1}^{\lambda} \theta_j} & 17 \le k \le 16+\lambda \end{cases} \tag{5}$$

Where ω is the weight factor; θ_j is called the j th-tier correlation factor; f_k is the normalized occurrence frequency. And the feature vector formulated by Eq. (4) can be also directly generated by the web server Pse-in-One.

Here, the last few components of the feature vector, that can show the sequence order information, are adopted to represent the RNA sequence, as shown below.

$$R = [v_{17} \quad \mathrm{L} \quad v_{16+\lambda}]^T \tag{6}$$

Experiments show that the best results can be obtained when the parameter λ and ω are set to 5 and 0.3, respectively. Then we can obtain a 5-dimensional feature vector.

2.2.4. Nucleotide Density. As described in the references [11,15], the concept of nucleotide density was proposed to reflect the frequency of a nucleotide and its distribution in a given RNA sample sequence R formulated by Eq. (2). Then the corresponding feature vector can be expressed as

$$R = [P_1 \quad P_2 \quad \mathrm{L} \quad P_i \quad \mathrm{L} \quad P_L]^T = [P_1 \quad P_2 \quad \mathrm{L} \quad P_i \quad \mathrm{L} \quad P_{51}]^T \tag{7}$$

$$P_i = \frac{\sum_{k=1}^{i} f(R_k)}{i} \tag{8}$$

Where P_i is the density of the nucleotide R_i at position i of a given RNA sample sequence with L nucleotides, and

$$f(R_k) = \begin{cases} 1 & if \quad R_k = R_i \\ 0 & otherwise \end{cases} \tag{9}$$

2.2.5. Feature Fusion. In order to increase the degree of discrimination of RNA sequences and further improve the performance of a predictive model, we can incorporate the above-mentioned three different feature extraction methods into a fusion vector to express the sample sequence formulated by Eq.(2), as shown below.

$$R = [\xi_1 \quad \xi_2 \quad \mathrm{L} \quad \xi_u \quad \mathrm{L} \quad \xi_{484} \quad v_{17} \quad \mathrm{L} \quad v_{16+5} \quad P_1 \quad \mathrm{L} \quad P_{51}]^T \tag{10}$$

2.2.6. Sparse Auto-Encoder. As a popular classifier, sparse auto-encoder has been successfully applied in bioinformatics field [24-26]. In this paper, we construct a sparse auto-encoder with two hidden layers to identify A-t-I sites. In order to achieve this more effectively, we can use the deep learning software package that can be downloaded from the website: https://github.com/rasmusbergpalm/DeepLearnToolbox. In this package, after using the SAE and NN, we can obtain the optimized results through the optimization of the parameters.

The predictor is called 'PAI-SAE', where 'P' stands for 'predicting', 'AI' for 'A-t-I editing sites' and 'SAE' for 'sparse auto-encoder'.

2.2.7. Prediction Quality Examination. In general, there are four conventional metrics, i.e. Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), and Matthew correlation coefficient (MCC), that are widely used to examine the performance of a predictor in the field of bioinformatics, as formulated by

$$\begin{cases} Sn = \dfrac{TP}{TP+FN} \\ Sp = \dfrac{TN}{TN+FP} \\ ACC = \dfrac{TP+TN}{TP+TN+FP+FN} \\ MCC = \dfrac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \tag{11}$$

Where TP represents the number of the A-t-I editing sample sequences correctly predicted as the A-t-I editing sample sequences; TN, the number of the non-A-t-I editing sample sequences correctly predicted as the non-A-t-I editing sample sequences; FP, the number of the non-A-t-I editing sample sequences incorrectly predicted as the A-t-I editing sample sequences; FN, the number of the A-t-I editing samples incorrectly predicted as the non-A-t-I editing samples.

2.2.8. Cross-Validation. As most scientists on biology have done, we use the validation methods to score the above-mentioned four metrics. Generally, there are three cross-validation methods, namely independent dataset test, K-fold cross-validation test and jackknife test. Although K-fold cross-validation test has more advantages in the computational time, the jackknife test can yield the unique outcome for a given benchmark dataset. Therefore, the jackknife test is adopted to examine the predictor's performance in this paper.

3. Result and Discussion

Listed in Table 2 are the rates obtained by the current PAI-SAE predictor via the jackknife test on the benchmark dataset. For facilitating comparison, listed in that tables are also the corresponding results obtained by the PAI, the existing most powerful predictor based on Supper Vector Machine for identify A-t-I editing sites in D. melanogaster.

As shown in Table 2, the scores of the four metrics used to quantitatively measure the quality of a single-label predictor, the new predictor "PAI-SAE", are higher than those of the predictor "PAI". For example, the ACC of our predictor "PAI-SAE" gains 2.46 per cent. The MCC rate has increased by 4.14 per cent, the Sn rate, by 1.60 percent, and the Sp rate, by 3.36 per cent. As pointed out in a comprehensive review, among the aforementioned four metrics, the most important are MCC and ACC. The high success rates shown in Table 2 clearly indicate that the current predictor is not only a pioneer one in this area, but also holds very high potential to become a high throughput tool for both basic research and drug development.

Table 2. The comparison of the jackknife test results on benchmark dataset.

Predictor	ACC (%)	MCC (%)	Sn (%)	Sp (%)
PAI[a]	79.51	60.00	85.60	73.11
PAI-SAE[b]	81.97	64.14	87.20	76.47

[a]The prediction method developed by Chen et al. (2016)
[b]The prediction method proposed in this paper.

4. Conclusion

Identification of adenosine-to-inosine editing sites in RNA sequences is important for the intensive study on RNA function and the development of new medicine. In this paper, a new predictor called PAI-SAE was constructed based on hybrid features combining with DACC, PseDNC and nucleotide density by using spare auto-encoder. The jackknife test results of the predictor PAI-SAE on the benchmark dataset show that our predictor is superior to others in this area. And the results were promising enough for our predictor to be used as an analytic solution to more genomic problems.

Acknowledgments

This work was financially supported by the National Nature Science Foundation of China (No. 31560316, 31760315, 61300139, 61761023), Natural Science Foundation of Jiangxi Province, China (No. 20171ACB20023, 20171BAB202020, 20142BAB207013), the Department of Education of JiangXi Province (GJJ160866, GJJ160909, GJJ160910), China Postdoctoral Science Foundation Funded Project (Project No.2017M612949).

References

[1] B. Zinshteyn and K. Nishikura, Adenosine-to-inosine RNA editing, Wiley Interdiscip. Rev. Syst. Biol. Med. 1 (2009) 202-209.

[2] K. Licht and M. F. Jantsch, Rapid and dynamic transcriptome regulation by RNA editing and RNA modifications, J. Cell Biol. 213 (2016) 15-22.

[3] P. Barraud and F. H. Allain, ADAR proteins: double-stranded RNA and Z-DNA binding domains, Curr. Top. Microbiol. Immunol. 353 (2012) 35-60.

[4] G. Ramaswami, et al., Accurate identification of human Alu and non-Alu RNA editing sites, Nat. Methods. 9 (2012) 579-581.

[5] J. B. Li, et al., Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing, Science. 324 (2009) 1210-1213.

[6] J. H. Bahn, et al., Accurate identification of A-to-I RNA editing in human by transcriptome sequencing, Genome Res. 22 (2012) 142-150.

[7] M. Sakurai, et al., A biochemical landscape of A-to-I RNA editing in the human brain transcriptome, Genome Res. 24 (2014) 522-534.

[8] S. Alon, et al., The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing, eLife. 4 (2015).

[9] G. St Laurent, et al., Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila, Nat. Struct. Mol. Biol. 20 (2013) 1333-1339.

[10] W. Chen, P. Feng, H. Ding and H. Lin, PAI: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions, Sci. Rep. 6 (2016) 35123.

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

[12] Z. Liu, et al., pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties, Anal. Biochem. 497 (2016) 60-67.

[13] W. Chen, P. M. Feng, H. Lin and K. C. Chou, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, Biomed Res Int. 2014 (2014) 623149.

[14] W. R. Qiu, S. Y. Jiang, Z. C. Xu, X. Xiao and K. C. Chou, iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition, Oncotarget. 8 (2017) 41178-41188.

[15] W. Chen, H. Tang, J. Ye, H. Lin and K. C. Chou, iRNA-PseU: Identifying RNA pseudouridine sites, Molecular therapy Nucleic acids. 5 (2016) e332.

[16] B. A. Olshausen and D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381 (1996) 607-609.

[17] F. Gu, F. Florez-Revuelta, D. Monekosso and P. Remagnino, Marginalised Stacked Denoising Autoencoders for Robust Representation of Real-Time Multi-View Action Recognition,

Sensors 15 (2015) 17209-17231.

[18] L. Fu, B. Niu, Z. Zhu, S. Wu and W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150-3152.

[19] B. Liu, et al., Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65-71.

[20] B. Liu, F. Liu, L. Fang, X. Wang and K. C. Chou, repRNA: a web server for generating various feature vectors of RNA sequences, Mol. Genet. Genomics 291 (2016) 473-481.

[21] B. Liu, F. Liu, L. Fang, X. Wang and K. C. Chou, repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects, Bioinformatics 31 (2015) 1307-1309.

[22] X. Xiao, P. Wang and K. C. Chou, iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix, PLoS One 7 (2012) e30869.

[23] W. Chen, T. Y. Lei, D. C. Jin, H. Lin and K. C. Chou, PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53-60.

[24] W. Xu, L. Zhang and Y. Lu, SD-MSAEs: Promoter recognition in human genome based on deep feature extraction. Journal of biomedical informatics 61 (2016) 55-62.

[25] S. P. Nguyen, Y. Shang and D. Xu, DL-PRO: A Novel Deep Learning Method for Protein Model Quality Assessment, Proc Int Jt Conf Neural Netw. 2014 (2014) 2071-2078.

[26] Z. C. Xu, P. Wang, W. R. Qiu and Xiao, X. iSS-PC: Identifying Splicing Sites via Physical-Chemical Properties Using Deep Sparse Auto-Encoder, Sci. Rep. 7 (2017) 8222.