Molecular Therapy
Nucleic Acids
Original Article

iRNA-3typeA: Identifying Three Types of Modification at RNA's Adenosine Sites

Wei Chen,[1,3,4] Pengmian Feng,[2] Hui Yang,[3] Hui Ding,[3] Hao Lin,[3,4] and Kuo-Chen Chou[3,4]

[1]Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China; [2]Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, School of Public Health, North China University of Science and Technology, Tangshan 063000, China; [3]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; [4]Gordon Life Science Institute, Boston, MA 02478, USA

RNA modifications are additions of chemical groups to nucleotides or their local structural changes. Knowledge about the occurrence sites of these modifications is essential for in-depth understanding of the biological functions and mechanisms and for treating some genomic diseases as well. With the avalanche of RNA sequences generated in the post-genomic age, many computational methods have been proposed for identifying various types of RNA modifications one by one. However, so far no method whatsoever has been developed for simultaneously identifying several different types of RNA modifications. To address such a challenge, we developed a predictor called "iRNA-3typeA," by which we can simultaneously identify the occurrence sites of the following three most frequently observed modifications in RNA: (1) N^1-methyladenosine (m^{1}A), (2) N^6-methyladenosine (m^{6}A), and (3) adenosine to inosine (A-to-I). It has been shown via rigorous cross-validations for the RNA sequences from *Homo sapiens* and *Mus musculus* transcriptomes that the success rates achieved by the powerful new predictor are quite high. For the convenience of broad experimental scientists, a user-friendly web server for iRNA-3typeA has been established at http://lin-group.cn/server/iRNA-3typeA/. It is anticipated that iRNA-3typeA may become a useful high throughput tool for genome analysis.

INTRODUCTION

RNA modification means the addition of chemical groups to its constitutional nucleotides or structural changes therein.[1] So far, more than 100 types of RNA modifications have been observed in cellular RNAs of all living organisms.[2] Because they are involved in a series of crucial biological activities,[3] such as mRNA splicing, mRNA nuclear processing, mRNA export, and mRNA decay,[3–6] particularly linked with human diseases, RNA modifications have drawn great attention in the scientific community.

With the development of high-throughput experimental techniques,[7–9] lots of RNA modification data have been acquired; they are very helpful for revealing the novel functions of RNA modifications. As indicated in a recent review,[10] however, most of these methods are unable to discriminate among the different RNA modifications that may simultaneously occur in the same RNA molecule. For example, the adenosine usually undergoes N^1-methyladenosine (m^{1}A), N^6-methyladenosine (m^{6}A), and adenosine to inosine (A-to-I or A → I) modifications[7] (Figure 1). Unfortunately, using the aforementioned techniques, one could not detect whether different types of RNA modifications might take place at the same time, let alone analyze their combinational biological functions.[11].

Therefore, it is urgently needed to develop computational methods to address this problem. As excellent complements to experimental techniques, computational methods have been developed to identify RNA modifications[12–18] via machine learning to train computational models based on the large data yielded from the high-throughput experiments. However, rarely are they able to simultaneously identify multiple RNA modifications.

The present study was devoted to developing a bioinformatics tool that can identify the RNA modification types for m^{1}A, m^{6}A, and A → I that may simultaneously occur on adenosine in both *Homo sapiens* and *Mus musculus* transcriptomes.

As shown in a series of recent publications,[19–31] in developing a bioinformatics tool, complying with the five-step rules yields the following advantages:[32] (1) clearer in logic deduction, (2) better illumination in stimulating other relevant tools, and (3) more usefulness in practical application.

In view of this, we elaborate the following procedures required in the five-step rules: (1) benchmark dataset, (2) sample formulation, (3) operative machine, (4) cross-validation, and (5) web server, and they are embedded into the rubrics according to the journal's format.

Received 23 February 2018; accepted 27 March 2018;
https://doi.org/10.1016/j.omtn.2018.03.012.

Correspondence: Wei Chen, Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China.
E-mail: chenweiimu@gmail.com
Correspondence: Hao Lin, 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.
E-mail: hlin@uestc.edu.cn

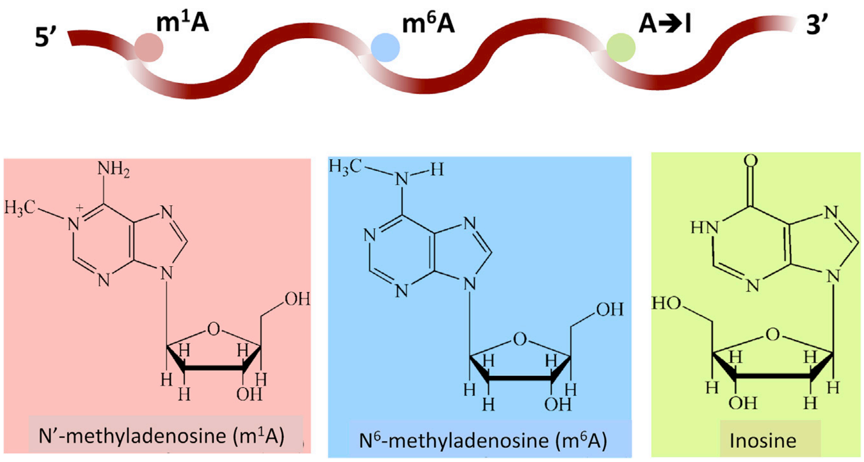


Figure 1. The Three Common Types of Modifications in RNA

(1) N^1-methyladenosine (m^{1}A), (2) N^6-methyladenosine (m^{6}A), and (3) adenosine to inosine (A-to-I).

RESULTS AND DISCUSSION

Performance Report

Listed in Table 1 are the jackknife test results obtained by the proposed predictor on the benchmark datasets (Supplemental Information S1 and Supplemental Information S2 available at http://lin-group.cn/server/iRNA3typeA/data.htm) for *H. sapiens* and *M. musculus*, respectively. As we can see from the table, the rates for both overall accuracy (Acc) and stability (MCC) are quite high for all the three different types of modifications investigated, indicating that the predictor is not only high in overall success rate but also quite stable. Therefore, the potential is quite high for iRNA-type3A to become a high-throughput tool in both basic research and drug development.

It is instructive to point out that, although the current predictor is limited in identifying m^{1}A, m^{6}A, and A→I sites for the RNA sequences from *H. sapiens* and *M. musculus*, with more experimental data available for other types of modifications and other species in future, we can easily to extend our model to cover more different types of modifications and more different species. Therefore, the current predictor is just a good start; it will be subjected to updates with the aim to continuously enhance its power and coverage scope.

Comparison with Other Classifiers

The proposed predictor iRNA-3typeA is the first predictor ever constructed for identifying the three types of RNA modifications (m^{1}A; m^{6}A; A→I) simultaneously. It is not possible to show its power via a conventional comparison since there is no other predictor whatsoever that can do the same. Nevertheless, below we can carry out a special comparison to further demonstrate its superiority.

As mentioned above, the operative machine used for iRNA-3typeA is a support vector machine (SVM) classifier. What would happen if we use other classifiers instead? Listed in Table 2 are the results when the SVM classifier was substituted with the other classifiers, respectively.

From the table, we can see the following: (1) the SVM classifier is better than J48 Tree in all the metrics rates. (2) Although the SVM classifier is a little bit lower than the BayesNet classifier and Naive Bayes classifier in identifying the m^{6}A sites for *H. sapiens*, its accuracies in identifying all the other types of modifications for both *H. sapiens* and *M. musculus* are significantly higher than those of BayesNet and Naive Bayes. All these results have further indicated that the SVM classifier is indeed a correct choice for the iRNA-3typeA predictor.

Web Server and User Guide

The last step of the five-step rules[32] is about the web server. It is indeed important because user-friendly and publicly accessible web servers represent the future direction for developing practically more useful predictors.[33] Actually, it has been demonstrated by a series of recent publications (see, e.g., Cheng et al.,[25,34–36] Liu et al.,[28] Lin et al.,[37] Jia et al.,[38,39] and Cheng and Xiao[40]) that a new prediction method with its web server available would significantly enhance its impacts.[41,42] In view of this, the web server for iRNA-3typeA has been established. Furthermore, to maximize the convenience of broad experimental scientists, a step-by-step guide is given below:

Step 1. Open the iRNA-3typeA web server at http://lin-group.cn/server/iRNA-3typeA; you will see the top page of the web server as shown in Figure 2A.

Step 2. Either type or copy/paste the query RNA sequences (in FASTA format) into the input box. Example sequences can be found by clicking on the Example button.

Step 3. Click the open circle (*H. sapiens* and *M. musculus*) to choose the species concerned, followed by clicking the Submit button. For example, if using the query RNA sequences in the Example window as the input and choosing *H. sapiens*, after submission you will see the predicted results summarized in a table (Figure 2B), clearly indicating (1) the adenosine at position 21of sequence #1 has the potential to be of the site for m^{1}A or A-to-I editing modification. (2) The adenosine at position 21 of sequence #2 has the potential to be of m^{6}A modification only. All these predicted results are fully consistent with experimental observations.

MATERIALS AND METHODS

Benchmark Datasets

The benchmark datasets for m^{1}A, m^{6}A, and A-to-I editing sites in *H. sapiens* and *M. musculus* genomes were derived from the previous

Table 1. The Success Rates Achieved by iRNA-3typeA via Jackknife Tests on the Benchmark Datasets for *H. sapiens* and *M. musculus*, Respectively

Species	Type of Modification	Sn (%)	Sp (%)	Acc (%)	MCC
H. sapiens	m^1A[a]	98.38	99.89	99.13	0.98
	m^6A[b]	81.68	99.11	90.38	0.82
	A → I[c]	86.18	95.23	90.71	0.82
M. musculus	m^1A[d]	97.46	100.00	98.73	0.97
	m^6A[e]	77.79	100.00	88.39	0.80
	A → I[f]	96.75	100.00	98.38	0.96

[a]The parameters used for SVM are $C = 8$ and $\gamma = 0.0078125$.
[b]The parameters used for SVM are $C = 128$ and $\gamma = 3.05158e\text{-}5$.
[c]The parameters used for SVM are $C = 8$ and $\gamma = 0.0078125$.
[d]The parameters used for SVM are $C = 2$ and $\gamma = 0.0078125$.
[e]The parameters used for SVM are $C = 32$ and $\gamma = 0.00012207$.
[f]The parameters used for SVM are $C = 512$ and $\gamma = 0.000488281$.

Table 2. The Comparative Results of the Proposed Predictor When Its Operating Algorithm[32] Was Replaced from SVM to Other Classifiers

Classifier	Species	Modification Type	Sn (%)	Sp (%)	Acc (%)	MCC
BayesNet[a]	*H. sapiens*	m^1A	98.81	98.85	98.83	0.98
		m^6A	82.04	100.00	91.02	0.83
		A → I	88.50	89.57	89.03	0.78
	M. musculus	m^1A	97.18	98.78	97.98	0.96
		m^6A	77.79	100.00	88.90	0.80
		A → I	96.51	99.88	98.20	0.96
Naive Bayes[a]	*H. sapiens*	m^1A	98.16	98.30	98.23	0.96
		m^6A	82.04	99.73	90.88	0.83
		A → I	89.40	87.04	88.22	0.76
	M. musculus	m^1A	96.43	97.75	97.09	0.94
		m^6A	77.79	98.62	88.22	0.78
		A → I	95.91	97.95	96.93	0.94
J48 Tree[a]	*H. sapiens*	m^1A	98.77	99.40	99.09	0.98
		m^6A	82.48	84.35	83.41	0.67
		A → I	88.18	89.04	88.60	0.77
	M. musculus	m^1A	96.71	98.68	97.70	0.95
		m^6A	83.03	82.21	82.62	0.65
		A → I	96.27	99.04	97.65	0.95
SVM[b]	*H. sapiens*	m^1A	98.46	99.89	99.18	0.98
		m^6A	80.44	100.00	90.23	0.82
		A → I	86.73	95.40	91.07	0.82
	M. musculus	m^1A	97.46	100.00	98.73	0.97
		m^6A	77.79	100.00	88.90	0.80
		A → I	97.35	100.00	98.67	0.97

All the rates below are obtained by the 10-fold cross-validations on the same benchmark datasets (Supplemental Information S1 and Supplemental Information S2 available at http://lin-group.cn/server/iRNA3typeA/data.htm).
[a]Taken from the WEKA package.[91]
[b]Proposed in this paper.

works.[12,14,43] Listed in Table 3 are the numbers of positive and negative samples for each of the benchmark datasets. It has been found by similar approaches[12,14] that the optimal length of the sequence samples in the benchmark datasets are 41nt, with the modified sites (m^1A, m^6A, or A → I editing site) at the center. For readers' convenience, the benchmark dataset thus obtained for *H. sapiens* is given in Supplemental Information S1, while that for *M. musculus* given in Supplemental Information S2; both can be downloaded from the link at http://lin-group.cn/server/iRNA3typeA/data.htm.

Sample Formulation

An RNA sample with 41 nt is usually sequentially formulated by

$$\mathbf{R} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\cdots\mathrm{N}_i\cdots\mathrm{N}_{41}, \qquad \text{(Equation 1)}$$

where

$$\mathrm{N}_i \in \{\ \mathrm{A(adenine)},\quad \mathrm{C(cytosine)},\quad \mathrm{G(guanine)},\quad \mathrm{U(uracil)}\ \} \qquad \text{(Equation 2)}$$

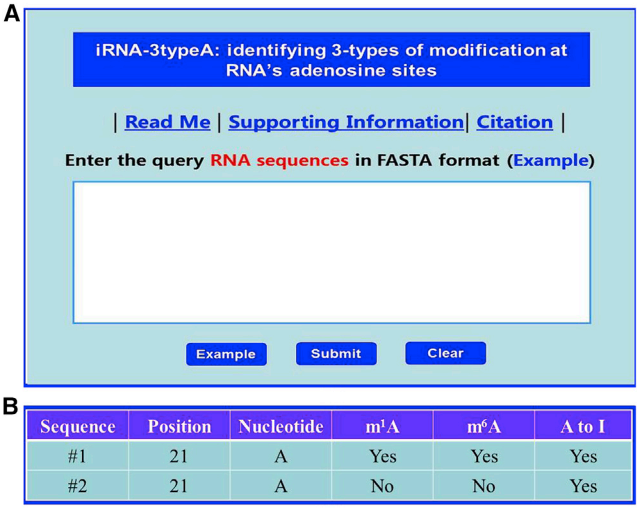
denotes the nucleotide at the i-th sequence position, and $\in$ is the a symbol in the set theory meaning "member of."

To enable the existing machine-learning algorithms handle the RNA sample,[41] the first thing we need to do is to convert its sequential formulation into a vector. But a vector in a discrete framework might totally miss all the sequence-order information or pattern feature. To deal with this problem, the PseAAC (pseudo amino acid composition) was introduced.[44] Ever since the concept of PseAAC was proposed, it has been swiftly penetrated into many biomedicine and drug development areas[45,46] and nearly all the areas of computational proteomics (see, e.g.,Esmaeili et al.,[47] Mohabatkar et al.,[48] Nanni et al.,[49] Pacharawongsakda and Theeramunkong,[50] Mondal and Pai,[51] Ahman et al.,[52] Kabir and Hayat,[53] Yu et al.,[54] Zhang and Duan,[55] Muthu Krishnan,[56] and a long list of references cited in two review papers[42,57]). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, this idea has been extended to deal with DNA/RNA sequences[21,28,37,58–60] in computational genomics via PseKNC (pseudo K-tuple nucleotide composition).[61,62] According to Chen et al.[63], the general form of PseKNC can be formulated as

$$\mathbf{R} = [\phi_1\phi_2\cdots\phi_u\cdots\phi_\Gamma]^{\mathrm{T}}, \qquad \text{(Equation 3)}$$

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

The four bases (A, C, G, and U) of RNA have different chemical properties and structures.[64,65] Therefore, based on their different chemical properties and structures,[64,65] A, C, G, and U can be represented by (1, 1, 1), (0, 0, 1), (1, 0, 0), and (0, 1, 0), respectively.[20,27] For instance, the RNA sequence with six nucleotides "GUGCAG" can be expressed



iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites

| Read Me | Supporting Information| Citation |

Enter the query RNA sequences in FASTA format (Example)

Sequence	Position	Nucleotide	m¹A	m⁶A	A to I
#1	21	A	Yes	Yes	Yes
#2	21	A	No	No	Yes

Figure 2. The Semi-screenshot for the Top Page of the iRNA-3typeA Web Server and the Prediction Result of the Two Example Query Sequences

The Semi-screenshot for the top page of the iRNA-3typeA Web Server (top panel) and the Prediction Result of the two example query sequences (bottom panel).

by the vector of $(3 \times 6) = 18$ components; i.e., [1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0].

Moreover, to incorporate into Equation 3 the sequence-coupled information[66] for the nucleotides around the modification sites, we adopt the lingering density as defined below

$$D_i = \frac{1}{\|L\|_i} \sum_{j=1}^{\ell} f(\mathrm{N}_j), \qquad \text{(Equation 4)}$$

where D_i is the density of the nucleotide N_i at the site i of a RNA sequence, $\|L_i\|$ the length of the sliding substring concerned; ℓ denotes each of the site locations counted in the substring, and

$$f(\mathrm{N}_j) = \begin{cases} 1, & \text{if } \mathrm{N}_j = \text{the nucleotide concerned} \\ 0, & \text{otherwise} \end{cases} \qquad \text{(Equation 5)}$$

For example, the RNA sequence "GUGCAG" can be represented by the vector [1, 0.5, 0.66, 0.25, 0.2, 0.5].

Thus, by using both nucleotide chemical properties and the lingering density (cf. Equation 4), each nucleotide can be defined by four variables. Accordingly, the RNA sequence of Equation 1 can be defined by a vector with $(41 \times 6) = 164$ components; namely $\Gamma = 164$ for Equation 3 now.

Table 3. A Breakdown of the Benchmark Dataset

		Number of Samples		
Species	Attribute	m^1A	m^6A	A→I
H. sapiens	positive	6,366	1,130	3,000
	negative	6,366	1,130	3,000
M. musculus	positive	1,064	725	831
	negative	1,064	725	831

Operative Machine

In this study, the SVM was chosen as the operative machine. The SVM has been widely used in computational genomics and proteomics (see, e.g., Ehsan et al.,[26] Feng et al.,[20,27,67–69] Chen et al.,[70–72] Lin et al.,[73] Lai et al.,[74] Zhao et al.,[75] and Yang et al.[76]). The implementation of the SVM was conducted by using the LibSVM package 3.18 available at https://www.csie.ntu.edu.tw/~cjlin/libsvm/. The radial basis kernel function (RBF) was used to obtain the classification hyperplane, and the grid search method was applied to optimize the regularization parameter C and kernel parameter γ.

The predictor obtained via the above procedures is called "iRNA-3typeA," where "i" stands for "identify," and "3typeA" means RNA's "three types of modifications at adenosine sites." Illustrated in Figure 3 is a flowchart to show the process of how the iRNA-3typeA predictor is working.

Cross-Validation

To evaluate the quality of a new predictor, we need to consider the following two problems. What metrics should be used to quantitatively display its performance? And what concrete procedure should be followed to derive the metrics' values?

(1) A set of four metrics. In literature, the following four conventional metrics are generally used to evaluate a predictor's quality:[77] (1) Acc, (2) MCC, (3) sensitivity (Sn), and (4) specificity (Sp). But the conventional expressions copied directly from math books are lacking in inductivity and hard to understand for most biological scientists. Fortunately, by using the symbols introduced by Chou in studying signal peptides,[78] the four metrics can be converted to a set of intuitive ones[58,79] as given below:

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\[2ex] \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\[2ex] \mathrm{Acc} = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\[2ex] \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases}, \qquad \text{(Equation 6)}$$

where N^+ represents the total number of positive samples investigated, while N_-^+ is the number of positive samples incorrectly predicted to be negative, and N^- represents the total number of negative

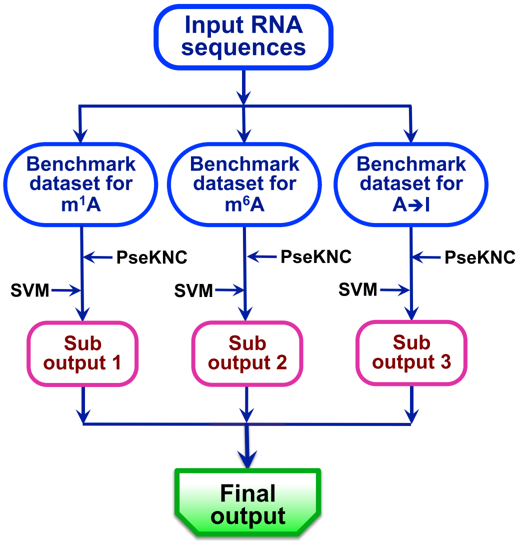


Figure 3. A Flowchart to Show How the iRNA-3typeA Predictor Is Working

samples investigated, while N_{+}^{-} the number of the negative samples incorrectly predicted to be positive. With the set of formulations in Equation 6, the meanings of Sn, Sp, Acc, and MCC have become much more intuitive and easier to understand, as discussed in a series of recent studies in various biological areas (see, e.g., Liu et al.,[21,24,28,60] Ehsan et al.,[26] Feng et al.,[20,27] Song et al.,[31] Lin et al.,[37] and Xu et al.[80,81]).

(2) Jackknife test. Now the next problem is how to test the values of these metrics in an objective way. As is well known, the independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test are the three cross-validation methods widely used for testing a prediction method.[82] Of the three test methods, however, the jackknife test is deemed the least arbitrary and most objective one.[32] Accordingly, the jackknife test has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors (see, e.g., Ahmad et al.,[52,83] Lin et al.,[84] Tang et al.,[85] Tripathi and Pandey,[86] and Dao et al.[87]). In view of this, the jackknife test was also adopted in the current study to examine the proposed predictor. During the jackknife test, each sample in the benchmark dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the one being identified. One more advantage of using the jackknife test is that there is no need to artificially separate the benchmark dataset into two subsets, one for training the model and one for testing it. This is because the outcome obtained by the jackknife test is actually a combination from many different independent dataset tests.[88–90]

AUTHOR CONTRIBUTIONS

W.C. and H.L. designed the study; P.F., H.Y., and H.D. conducted the experiments; W.C., H.L., and K.-C.C. analyzed the results; W.C., H.L., and K.-C.C. wrote the paper.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors wish to thank the three anonymous reviewers, whose constructive comments were very helpful for further strengthening the presentation of this paper. This work was supported by the Natural Science Foundation of China (No. 31771471 and 61772119), the 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), and the Applied Basic Research Program of Sichuan Province (No. 2015JY0100).

REFERENCES

1. Gilbert, W.V., Bell, T.A., and Schaening, C. (2016). Messenger RNA modifications: form, distribution, and function. Science *352*, 1408–1412.

2. Machnicka, M.A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K.M., et al. (2013). MODOMICS: a database of RNA modification pathways—2013 update. Nucleic Acids Res. *41*, D262–D267.

3. Roundtree, I.A., Evans, M.E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. Cell *169*, 1187–1200.

4. Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., Yi, C., Lindahl, T., Pan, T., Yang, Y.G., and He, C. (2011). N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat. Chem. Biol. *7*, 885–887.

5. Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120.

6. Zhao, B.S., Roundtree, I.A., and He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. Nat. Rev. Mol. Cell Biol. *18*, 31–42.

7. Li, X., Xiong, X., Wang, K., Wang, L., Shu, X., Ma, S., and Yi, C. (2016). Transcriptome-wide mapping reveals reversible and dynamic N(1)-methyladenosine methylome. Nat. Chem. Biol. *12*, 311–316.

8. Chen, K., Lu, Z., Wang, X., Fu, Y., Luo, G.Z., Liu, N., Han, D., Dominissini, D., Dai, Q., Pan, T., and He, C. (2015). High-resolution N(6) -methyladenosine (m(6) A) map using photo-crosslinking-assisted m(6) A sequencing. Angew. Chem. Int. Ed. Engl. *54*, 1587–1590.

9. Helm, M., and Motorin, Y. (2017). Detecting RNA modifications in the epitranscriptome: predict and validate. Nat. Rev. Genet. *18*, 275–291.

10. Esteller, M., and Pandolfi, P.P. (2017). The epitranscriptome of noncoding RNAs in cancer. Cancer Discov. *7*, 359–368.

11. Nachtergaele, S., and He, C. (2017). The emerging biology of RNA post-transcriptional modifications. RNA Biol. *14*, 156–163.

12. Chen, W., Tang, H., and Lin, H. (2017). MethyRNA: a web server for identification of N^6-methyladenosine sites. J. Biomol. Struct. Dyn. *35*, 683–687.

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

14. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.C. (2017). iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget *8*, 4208–4217.

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

16. Qiu, W.R., Jiang, S.Y., Sun, B.Q., Xiao, X., Cheng, X., and Chou, K.C. (2017). iRNA-2methyl: identify RNA 2′-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. Med. Chem. *13*, 734–743.

17. Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K.C. (2016). iRNA-PseU: identifying RNA pseudouridine sites. Mol. Ther. Nucleic Acids *5*, e332.

18. Feng, P., Ding, H., Chen, W., and Lin, H. (2016). Identifying RNA 5-methylcytosine sites via pseudo nucleotide compositions. Mol. Biosyst. *12*, 3307–3311.

19. Cheng, X., Xiao, X., and Chou, K.C. (2017). pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. Mol. Biosyst. *13*, 1722–1727.

20. Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., and Chou, K.C. (2017). 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.

21. Liu, B., Wang, S., Long, R., and Chou, K.C. (2017). iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics *33*, 35–41.

22. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., Jia, J.H., and Chou, K.C. (2017). iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics, Published online November 16, 2017. https://doi.org/10.1016/j.ygeno.2017.10.008.

23. Xiao, X., Cheng, X., Su, S., and Nao, Q. (2017). pLoc-mGpos: Incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. Nat. Sci. *9*, 331–349.

24. Liu, L.M., Xu, Y., and Chou, K.C. (2017). iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Med. Chem. *13*, 552–559.

25. Cheng, X., Xiao, X., and Chou, K.C. (2018). pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics *110*, 50–58.

26. Ehsan, A., Mahmood, K., Khan, Y.D., Khan, S.A., and Chou, K.C. (2018). A novel modeling in mathematical biology for classification of aignal peptides. Sci. Rep. *8*, 1039.

27. Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.C. (2018). iDNA6mA-PseKNC: identifying DNA N^6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics, Published online January 31, 2018. https://doi.org/10.1016/j.ygeno.2018.01.005.

28. Liu, B., Yang, F., Huang, D.S., and Chou, K.C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics *34*, 33–40.

29. Song, J., Li, F., Takemoto, K., Haffari, G., Akutsu, T., Chou, K.C., and Webb, G.I. (2018). 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.

30. Yang, H., Qiu, W.R., Liu, G., Guo, F.B., and Lin, H. (2018). iRSpot-Pse6NC: identifying recombination spots in *Saccharomyces cerevisiae* by incorporating hexamer composition into general PseKNC. Int. J. Biol. Sci. https://doi.org/10.7150/ijbs.246.

31. Song, J., Wang, Y., Li, F., Akutsu, T., and Rawlings, N.D. (2018). iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief. Bioinform. https://doi.org/10.1093/bib/bbx180.

32. Chou, K.C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. *273*, 236–247.

33. Shen, H.B. (2009). Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. *1*, 63–92.

34. Cheng, X., Zhao, S.G., Lin, W.Z., Xiao, X., and Chou, K.C. (2017). pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics *33*, 3524–3531.

35. Cheng, X., Xiao, X., and Chou, K.C. (2017). pLoc-mGneg: predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics, Published online October 6, 2017. https://doi.org/10.1016/j.ygeno.2017.10.002.

36. Cheng, X., Xiao, X., and Chou, K.C. (2017). pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics, Published online November 2, 2017. https://doi.org/10.1093/bioinformatics/btx711.

37. Lin, H., Deng, E.Z., Ding, H., Chen, W., and Chou, K.C. (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.

38. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (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.

39. Jia, J., Zhang, L., Liu, Z., Xiao, X., and Chou, K.C. (2016). pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics *32*, 3133–3141.

40. Cheng, X., and Xiao, X. (2017). pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. Gene *628*, 315–321.

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

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

43. Chen, W., Feng, P., Tang, H., Ding, H., and Lin, H. (2016). RAMPred: identifying the N(1)-methyladenosine sites in eukaryotic transcriptomes. Sci. Rep. *6*, 31080.

44. Chou, K.C. (2001). Prediction of protein cellular attributes using pseudo amino acid composition. Proteins *43*, 246–255.

45. Zhong, W.Z., and Zhou, S.F. (2014). Molecular science for drug development and biomedicine. Int. J. Mol. Sci. *15*, 20072–20078.

46. Zhou, G.P., and Zhong, W.Z. (2016). Perspectives in medicinal chemistry. Curr. Top. Med. Chem. *16*, 381–382.

47. Esmaeili, M., Mohabatkar, H., and Mohsenzadeh, S. (2010). Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J. Theor. Biol. *263*, 203–209.

48. Mohabatkar, H., Mohammad Beigi, M., and Esmaeili, A. (2011). Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. J. Theor. Biol. *281*, 18–23.

49. Nanni, L., Lumini, A., Gupta, D., and Garg, A. (2012). Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans. Comput. Biol. Bioinformatics *9*, 467–475.

50. Pacharawongsakda, E., and Theeramunkong, T. (2013). Predict subcellular locations of singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of Chou's PseAAC. IEEE Trans. Nanobioscience *12*, 311–320.

51. Mondal, S., and Pai, P.P. (2014). Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J. Theor. Biol. *356*, 30–35.

52. Ahmad, S., Kabir, M., and Hayat, M. (2015). Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into Chou's general PseAAC. Comput. Methods Programs Biomed. *122*, 165–174.

53. 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.

54. Yu, B., Li, S., Qiu, W.Y., Chen, C., Chen, R.X., Wang, L., Wang, M.H., and Zhang, Y. (2017). Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. Oncotarget *8*, 107640–107665.

55. Zhang, S., and Duan, X. (2018). Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J. Theor. Biol. *437*, 239–250.

56. Muthu Krishnan, S. (2018). Using Chou's general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains. J. Theor. Biol. *445*, 62–74.

57. Chou, K.C. (2009). Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr. Proteomics *6*, 262–274.

58. Chen, W., Feng, P.M., Lin, H., and Chou, K.C. (2013). iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. *41*, e68.

59. Qiu, W.R., Xiao, X., and Chou, K.C. (2014). iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int. J. Mol. Sci. *15*, 1746–1766.

60. Liu, B., Yang, F., and Chou, K.C. (2017). 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. Mol. Ther. Nucleic Acids *7*, 267–277.

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

62. Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.C. (2015). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics *31*, 119–120.

63. Chen, W., Lin, H., and Chou, K.C. (2015). Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol. Biosyst. *11*, 2620–2634.

64. Chen, W., Feng, P., Tang, H., Ding, H., and Lin, H. (2016). Identifying 2′-O-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions. Genomics *107*, 255–258.

65. Li, W.C., Deng, E.Z., Ding, H., Chen, W., and Lin, H. (2015). iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. Chemometr. Intell. Lab. Syst. *141*, 100–106.

66. Chou, K.C. (1993). A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J. Biol. Chem. *268*, 16938–16948.

67. Feng, P.M., Chen, W., Lin, H., and Chou, K.C. (2013). iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal. Biochem. *442*, 118–125.

68. Feng, P.M., Lin, H., and Chen, W. (2013). Identification of antioxidants from sequence information using naïve Bayes. Comput. Math. Methods Med. *2013*, 567529.

69. Feng, P.M., Ding, H., Chen, W., and Lin, H. (2013). Naïve Bayes classifier with feature selection to identify phage virion proteins. Comput. Math. Methods Med. *2013*, 530696.

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

71. Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics *33*, 3518–3523.

72. Chen, X.X., Tang, H., Li, W.C., Wu, H., Chen, W., Ding, H., and Lin, H. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. BioMed Res. Int. *2016*, 1654623.

73. Lin, H., Liang, Z.Y., Tang, H., and Chen, W. (2017). Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM Trans. Comput. Biol. Bioinformatics, Published online February 8, 2017. https://doi.org/10.1109/TCBB.2017.2666141.

74. Lai, H.Y., Chen, X.X., Chen, W., Tang, H., and Lin, H. (2017). Sequence-based predictive modeling to identify cancerlectins. Oncotarget *8*, 28169–28175.

75. Zhao, Y.W., Lai, H.Y., Tang, H., Chen, W., and Lin, H. (2016). Prediction of phosphothreonine sites in human proteins by fusing different features. Sci. Rep. *6*, 34817.

76. Yang, H., Tang, H., Chen, X.X., Zhang, C.J., Zhu, P.P., Ding, H., Chen, W., and Lin, H. (2016). Identification of secretory proteins in *Mycobacterium tuberculosis* using pseudo amino acid composition. BioMed Res. Int. *2016*, 5413903.

77. Chen, J., Liu, H., Yang, J., and Chou, K.C. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids *33*, 423–428.

78. Chou, K.C. (2001). Prediction of signal peptides using scaled window. Peptides *22*, 1973–1979.

79. Xu, Y., Ding, J., Wu, L.Y., and Chou, K.C. (2013). iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE *8*, e55844.

80. Xu, Y., Shao, X.J., Wu, L.Y., Deng, N.Y., and Chou, K.C. (2013). iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ *1*, e171.

81. Xu, Y., Wang, Z., Li, C., and Chou, K.C. (2017). iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. Med. Chem. *13*, 544–551.

82. Chou, K.C., and Zhang, C.T. (1995). Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. *30*, 275–349.

83. Ahmad, K., Waris, M., and Hayat, M. (2016). Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition. J. Membr. Biol. *249*, 293–304.

84. Lin, H., Liu, W.X., He, J., Liu, X.H., Ding, H., and Chen, W. (2015). Predicting cancerlectins by the optimal g-gap dipeptides. Sci. Rep. *5*, 16964.

85. Tang, H., Zou, P., Zhang, C., Chen, R., Chen, W., and Lin, H. (2016). Identification of apolipoprotein using feature selection technique. Sci. Rep. *6*, 30441.

86. Tripathi, P., and Pandey, P.N. (2017). A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition. J. Theor. Biol. *424*, 49–54.

87. Dao, F.Y., Yang, H., Su, Z.D., Yang, W., Wu, Y., Hui, D., Chen, W., Tang, H., and Lin, H. (2017). Recent advances in conotoxin classification by using machine learning methods. Molecules *22*, e1057.

88. Chou, K.C., and Shen, H.B. (2007). Recent progress in protein subcellular location prediction. Anal. Biochem. *370*, 1–16.

89. 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.

90. Shen, H.B. (2010). Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Sci. *2*, 1090–1103.

91. Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I.H. (2004). Data mining in bioinformatics using Weka. Bioinformatics *20*, 2479–2481.