Analytical Biochemistry 561–562 (2018) 59–65

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

# iRNA(m6A)-PseDNC: Identifying $N^6$-methyladenosine sites using pseudo dinucleotide composition

Wei Chen[a,b,d,*], Hui Ding[c], Xu Zhou[a], Hao Lin[c,d,**], Kuo-Chen Chou[c,d,***]

[a] School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, 063000, China
[b] Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, 611730, China
[c] 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
[d] Gordon Life Science Institute, Boston, MA, 02478, USA

ARTICLE INFO

*Keywords:*
$N^6$-methyladenosine
Pseudo nucleotide composition
RNA modification
Support vector machine
5-step rules

ABSTRACT

As a prevalent post-transcriptional modification, $N^6$-methyladenosine ($m^6A$) plays key roles in a series of biological processes. Although experimental technologies have been developed and applied to identify $m^6A$ sites, they are still cost-ineffective for transcriptome-wide detections of $m^6A$. As good complements to the experimental techniques, some computational methods have been proposed to identify $m^6A$ sites. However, their performance remains unsatisfactory. In this study, we firstly proposed an Euclidean distance based method to construct a high quality benchmark dataset. By encoding the RNA sequences using pseudo nucleotide composition, a new predictor called iRNA(m6A)-PseDNC was developed to identify $m^6A$ sites in the *Saccharomyces cerevisiae* genome. It has been demonstrated by the 10-fold cross validation test that the performance of iRNA(m6A)-PseDNC is superior to the existing methods. Meanwhile, for the convenience of most experimental scientists, established at the site http://lin-group.cn/server/iRNA(m6A)-PseDNC.php is its web-server, by which users can easily get their desired results without need to go through the detailed mathematics. It is anticipated that iRNA(m6A)-PseDNC will become a useful high throughput tool for identifying $m^6A$ sites in the *S. cerevisiae* genome.

## 1. Introduction

Among the ~150 kinds of chemical modifications identified in cellular RNAs, $N^6$-methyladenosine ($m^6A$) is the most prevalent one in mRNA and noncoding RNA [1]. Since it was first detected in 1970s [2], $m^6A$ has been observed in a wide range of eukaryotes. As indicated by recent evidences, $m^6A$ plays fundamental regulatory roles in a series of biological processes, such as affecting mRNA splicing and stability, translation, stem cell pluripotency, as well as immune response [3–7]. Therefore, the transcriptome-wide annotation of $m^6A$ site will be helpful to understanding its biological functions.

In 2012, the high-throughput sequencing technique termed MeRIP-Seq and $m^6A$-seq [8,9], were proposed to detect transcriptome wide $m^6A$ sites in *S. cerevisiae*, Mus musculus, and Homo sapiens. Since this technique relies solely on immunoprecipitation of fragmented RNA, the resolution of MeRIP-Seq and $m^6A$-seq is not satisfactory [5]. In 2015, Linder et al. proposed the miCLIP technique [10], which provides the single-nucleotide-resolution profile of $m^6A$ sites in human transcriptome. Although these experimental techniques promote the research progresses on $m^6A$ modifications, they are still costly and time consuming in performing transcriptome wide analysis.

During the last several years, many computational methods have been developed for identifying $m^6A$ sites in the *S. cerevisiae* genome. Based on the $m^6A$-seq data, Schwartz et al. proposed the first computational model for identifying $m^6A$ site in the *S. cerevisiae* genome [11]. Inspired by their work, Chen et al. developed two bioinformatics tools

[*] Corresponding author. School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, 063000, China.

[**] Corresponding author. 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.

[***] Corresponding author. 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 addresses:* chenweimimu@gmail.com (W. Chen), hding@uestc.edu.cn (H. Ding), sxzhouxu@126.com (X. Zhou), hlin@uestc.edu.cn (H. Lin), kcchou@gordonlifescience.org (K.-C. Chou).

https://doi.org/10.1016/j.ab.2018.09.002
Received 26 July 2018; Received in revised form 31 August 2018; Accepted 3 September 2018
Available online 08 September 2018

termed iRNA-Methyl and m$^6$Apred to identify m$^6$A sites, respectively [12,13]. Later on, by encoding RNA sequences using both sequence information and the RNA secondary structures, Zhou et al. developed a random forest based method to predict m$^6$A sites in *S. cerevisiae* [14]. In order to improve the accuracy of computationally identifying m$^6$A site in the *S. cerevisiae* genome, Chen et al. developed RAM-ESVM method that was built by using ensemble classifiers [15]. More recently, by encoding RNA sequences using multi-interval nucleotide pair position specificity, Xing et al. proposed a sequence-based predictor called RAM-NPPS for identifying m$^6$A sites [16].

Although these computational methods yielded encouraging results for computationally identifying m$^6$A sites in the *S. cerevisiae* [17], further improvement is needed. Particularly, most of those methods were trained by the dataset constructed by Chen et al. [12]. In that dataset, the negative samples were randomly selected from a huge amount of candidates, and hence unavoidably had some arbitrariness. Accordingly, their reliability might be questioned [18]. The present study was initiated in an attempt to develop a new and more powerful method to identify m$^6$A sites by refining the benchmark dataset.

As done in a series of recent reports [12,19–36], here we are to propose the new predictor also according to the Chou's 5-sep rules [37] because doing so will make the entire process more logic and transparent.

## 2. Materials and methods

### 2.1. Benchmark dataset

The first step in the 5-step rules [37] is how to construct or select a valid benchmark dataset to train and test the predictor. By following the same procedures as reported in Ref. [12], we first obtained 1307 positive samples and 33,280 negative samples. It was observed via preliminary trials that the optimal sequence length is 51 nt. But instead of using the random selection method to reduce the number of negative samples as reported in Ref. [12], in this study we adopted the subset-balancing treatment according to the Euclidean distance [38], as elaborated below.

First of all, for the reason that will be discussed later, we calculated the average value of each of the 22 features [12] over the 33,280 negative samples; i.e.,

$$\bar{d}_i = \frac{1}{33280}\sum_{j=1}^{33280} d_i^{\,j} \qquad (i = 1,2,\cdots,22) \tag{1}$$

where $d_i^{\,j}$ is the value of the $i$-th feature for the $j$-th negative sample. Thus, the center of the negative samples can be defined as

$$\overline{\mathbf{R}} = [\overline{d_1} \ \ \overline{d_2} \ \ \cdots \ \ \overline{d_{16}} \ \ \overline{d_{17}} \ \ \cdots \ \ \overline{d_{22}}]^{\mathbf{T}} \tag{2}$$

Subsequently, we calculated the Euclidean distance ($D^j$) between the $j$-th negative sample and the center $\overline{\mathbf{R}}$; i.e.,

$$D^j = \sqrt{\sum_{i=1}^{22} (d_i^{\,j} - \bar{d}_i)^2} \qquad (j = 1,2,\cdots,33280) \tag{3}$$

According to their Euclidean distance values, the 33,280 negative samples were sorted in an ascending order, and the top ranked 1307 negative samples (i.e., they were closest to the center) were picked up to form the refined negative subset.

For reader's convenience, the detailed sequences in the original 1307 positive samples and the refined 1307 negative samples are given in Supporting Information S1 and S2, which can be directly downloaded via the link at http://lin-group.cn/server/iRNA(m6A)-PseDNC.php.

### 2.2. Sample formulation

The second step in the 5-step rules [37] is how to formulate the biological sequence samples with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms can only handle vector but not sequence samples, as elucidated in a comprehensive review [39]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition [40] or PseAAC [41] was proposed. Ever since then, it has been widely used in nearly all the areas of computational proteomics (see, e.g. [42–83], as well as a long list of references cited in Ref. [84]). Because it has been widely and increasingly used, recently three powerful open access soft-wares, called 'PseAAC-Builder' [85], 'propy' [86], and 'PseAAC-General' [87], were established: the former two are for generating various modes of special PseAAC [88]; while the 3rd one for those of general PseAAC [37], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode, "Gene Ontology" mode, and "Sequential Evolution" or "PSSM" mode. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, its idea and approach were extended to PseKNC (Pseudo K-tuple Nucleotide Composition) to generate various feature vectors for DNA/RNA sequences [89] that have proved very successful as well [25,90–95].

As shown in Supporting Information S1 or S2, each of the RNA samples in this study has the form of

$$\mathbf{R} = \mathrm{N_1N_2N_3\cdots N_\mathit{i}\cdots N_{51}} \tag{4}$$

where $\mathrm{N_1}$ represents the 1st nucleotide, $\mathrm{N_2}$ the 2nd nucleotide, and so forth. They can be any of the four nucleotides A, C, G or U. According to PseKNC [89], the RNA sequence can be formulated as a discrete vector

$$\mathbf{R} = [d_1 \ \ d_2 \ \ \cdots \ \ d_{16} \ \ d_{16+1} \ \ \cdots \ \ d_{16+\lambda}]^{\mathrm{T}} \tag{5}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 16) \\[2ex] \dfrac{w\theta_{u-16}}{\sum_{i=1}^{16} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (16 < u \le 16+\lambda) \end{cases} \tag{6}$$

In Eq. (6), $f_u$ ($u = 1,2, \ \cdots, 16$) is the normalized occurrence frequency of the $u$-th non-overlapping dinucleotides in the RNA sequence. $\lambda$ can be viewed as the number of the total pseudo components used to reflect the long-range or global sequence effect, and $w$ is the weight factor. $\theta_j$ is the $j$-th tier correlation factor that reflects the sequence order correlation between all the $j$-th most contiguous dinucleotide along a RNA sequence as formulated by

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1} \mathrm{C}_{i,\,i+j} \qquad (j = 1,2,\cdots,\lambda;\ \lambda< L) \tag{7}$$

where $\mathrm{C}_{i,\,i+j}$ is the correlation function and is defined by

$$\mathrm{C}_{i,\,i+j} = \frac{1}{\mu}\sum_{g=1}^{\mu} [P_g(\mathrm{D}_i) - P_g(\mathrm{D}_{i+j})]^2 \tag{8}$$

where $\mu = 3$ is the number of RNA physicochemical properties considered, $P_g(\mathrm{D}_i)$ is the numerical value of the $g$-th ($g = 1, 2, 3, ..., u$) RNA local structural property for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$ at position $i$ and $P_g(\mathrm{D}_{i+j})$ the corresponding value for the dinucleotide $\mathrm{R}_{i+j}\mathrm{R}_{i+j+1}$ at position $i + j$.

Listed in Table 1 are the three physicochemical properties obtained by standard conversion for the 16 different dinucleotides in RNA. The concrete procedures for how to convert the original 16 physicochemical properties to their standard ones have been elaborated in Ref. [12], and hence there is no need to repeat here.

**Table 1**
The three physicochemical properties after standard conversion for the 16 different dinucleotides in RNA.

| Dinucleotide | Enthalpy (Ka/mol) | Entropy (eU) | Free energy (Ka/mol) |
|---|---|---|---|
| GG | −1.08 | −0.88 | −1.45 |
| GA | −1.50 | −1.82 | −0.28 |
| GC | −1.85 | −1.72 | −1.66 |
| GU | −0.30 | −0.32 | −0.14 |
| AG | 0.71 | 0.82 | 0.07 |
| AA | 1.10 | 0.95 | 1.56 |
| AC | −0.30 | −0.32 | −0.14 |
| AU | 1.44 | 1.42 | 1.34 |
| CG | 0.55 | 0.79 | −0.29 |
| CA | −0.42 | −0.57 | 0.03 |
| CC | −1.08 | −0.88 | −1.45 |
| CU | 0.71 | 0.82 | 0.07 |
| UG | 0.71 | 0.82 | 0.03 |
| UA | 0.51 | 0.27 | 1.04 |
| UC | −0.30 | −0.32 | −0.28 |
| UU | 1.10 | 0.95 | 1.56 |

In order to avoid over-fitting or the "high-dimension disaster" problem [96], the search of the two parameters in Eq. (6), namely λ and $w$, was in the following ranges [3,6] and [0, 1] with the steps of 1 and 0.1, respectively. It was found that the optimal values for λ and $w$ are 6 and 0.9, respectively.

Accordingly, the RNA sequence sample can be formulated by a (16 + 6)=22-D (dimensional) vector as given below

$$\mathbf{R} = [d_1 \quad d_2 \quad \cdots \quad d_{16} \quad d_{17} \quad \cdots \quad d_{22}]^{\mathrm{T}} \tag{9}$$

where the first 16 components are used to incorporate the short-range or local sequence order information of the RNA sample, while the remaining 6 components used to incorporate its long-range or global sequence order information.

### 2.3. Operation engine: support vector machine

The third step in the 5-step rules [37] is how to introduce or develop a powerful algorithm (or engine) to operate the prediction. Support vector machine (SVM) is a powerful and popular method for pattern recognition and has been widely used in the realm of bioinformatics [31,97–103]. The basic idea of SVM is to transform the input data into a high dimensional feature space and then determine the optimal separating hyperplane. In the current study, the LibSVM package 3.18 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used to implement SVM. Owing to its effectiveness and speed in training process, the radial basis kernel function (RBF) was used to obtain the classification hyperplane in the current study. For the SVM operation engine, a grid search approach was employed to optimize the regularization parameter $C$ and kernel parameter γ by using the 5-fold cross validation test in the following ranges [$2^{-5}$, $2^{15}$] and [$2^{-15}$, $2^{-5}$] with the steps of 2 and $2^{-1}$, respectively.

The predictor thus obtained is called "iRNA(m6A)-PseDNC", where "i" stands for "identify", "RNA(m6A)" for "RNA N6-methyladenosine site", and "PseDNC" for "via Pseudo Dinucleotide Composition".

### 2.4. Cross-validation

The fourth step in the 5-step rules [37] is how to properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor. To address this problem, we need to consider the two issues. (1) What metrics should be used to quantitatively measure the predictor's quality? (2) What test method should be utilized to score the metrics?

#### 2.4.1. A set of four intuitive metrics

To measure the quality of a predictor, four metrics [104] are often used in literature; they are (1) overall accuracy or Acc, (2) Mathew's correlation coefficient or MCC, (3) sensitivity or Sn, and (4) specificity or Sp. But their conventional formulations directly copied from math books are difficult to understand for most experimental scientists, particularly the one for MCC. Fortunately, by using the symbols introduced by Chou [105,106] in studying the signal peptide cleavage sites, a set of intuitive metrics were derived [107–109], as given below

$$\begin{cases} \mathrm{Sn}=1 - \dfrac{N_-^+}{N^+} & 0 \ \le \mathrm{Sn} \le 1 \\ \mathrm{Sp}=1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc}= \ \Lambda= 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC}= \dfrac{1-\left(\dfrac{N_-^+}{N^+}+\dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1+\dfrac{N_+^- - N_-^+}{N^+}\right)\left(1+\dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \tag{10}$$

where represents the total number of positive samples investigated, while $N_-^+$ is the number of positive samples incorrectly predicted to be of negative one; $N^-$ the total number of negative samples investigated, while $N_+^-$ the number of the negative samples incorrectly predicted to be of positive one. Because of its merit in intuitiveness, the set of metrics (Eq. (10)) has been increasingly and widely used in computational biology (see, e.g., [12,20–33] [56,59] [91–95] [110–144]). It is instructive to point out, however, that both the original four metrics [104] in math books and the intuitive ones in Eq. (7) are valid only for the single-label systems (where each sample belongs to one and only one class). For the multi-label systems (where a sample may simultaneously belong to several classes), whose existence has become more frequent in system biology [145–151], system medicine [152,153] and biomedicine [154], a completely different set of metrics as defined in Ref. [155] is absolutely needed."

#### 2.4.2. Cross-validation

In statistical prediction, the following three cross-validation methods are often used to check the performance of a predictor: independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test [38]. Of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in Ref. [37] and demonstrated by Eqs.28–30 therein. However, to reduce the computational time, in this study we adopted the 10-fold cross validation test, as done by many investigators with SVM as the operation engine.

### 2.5. Web-server for iRNA(m6A)-PseDNC

The last, but not the least important, step of the 5-step rules is how to establish a user-friendly web-server for the predictor that is accessible to the public. As pointed out in Ref. [156] and demonstrated in a series of recent publications (see, e.g. [20,24,25,92,94,95,136] [138–142] [144–146] [148–152] [157–162]), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have significantly increased the impacts of bioinformatics on medical science [39], driving medicinal chemistry into an unprecedented revolution [84]. For the convenience of the majority of the experimental scientists, the web-server for the iRNA(m6A)-PseDNC predictor has also been established at http://lin-group.cn/server/iRNA(m6A)-PseDNC.php.

## 3. Results and discussion

The success rates achieved by the proposed predictor by 10-fold cross validation test on the benchmark datasets as described in Section 2.1 are given in Table 2, where for facilitating comparison, the corresponding results by the existing state-of-the-art method RAM-NPPS

**Table 2**
Comparison with the existing state-of-the-art method in identifying m6A sites for *S. cerevisiae* genome.

| Methods | Sn(%)[a] | Sp(%)[a] | Acc(%)[a] | MCC [a] |
|---|---|---|---|---|
| RAM-NPPS[b] | 74.29 | 69.93 | 72.11 | 0.44 |
| iRNA(m6A)-PseDNC[c] | 86.84 | 95.64 | 91.24 | 0.83 |

[a] See Eq. (10) for the metrics definition.
[b] See [16] for RAM-NPPS, which is the existing state-of-the-art method in identifying m6A sites for *S. cerevisiae* genome.
[c] The predictor proposed in this paper with $C = 2^{15}$ and $\gamma = 2^{-7}$.

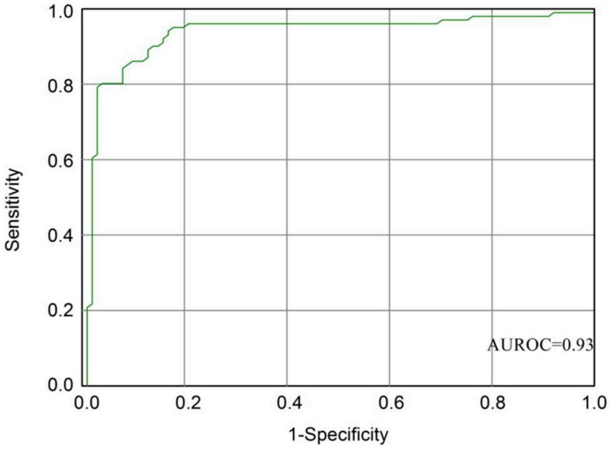

**Fig. 1.** A graphical illustration to show the performances of iRNA(m6A)-PseDNC in identifying m6A sites. See the text in Section 3 for further explanation.

[16] are also listed. As we can see from the table, for the existing best method in identifying m6A sites for the S. *cerevisiae* genome: its Sn (sensitivity) is over 12% lower than that of iRNA(m6A)-PseDNA, its Sp (specificity) is over 25% lower, its Acc (accuracy) is about 20% lower, and its MCC (stability) is 39% lower. Theses compelling results indicate that the iRNA(m6A)-PseDNA predictor is indeed very powerful.

Since graphic analysis is a very useful vehicle for studying complicated biological systems, as demonstrated by a series of previous studies (see, e.g., [163–171]. Shown in Fig. 1 is the graph of Receiver Operating Characteristic (ROC) widely used to reflect the quality of a predictor [172]. The area under the ROC curve, also called AUROC, is used to check the performance quality of the classifier: the value 0.5 of AUROC is equivalent to the outcome obtained by the random guess while 1 of AUROC represents the perfect prediction with 100% correctness. For the case of the current predictor, it is 0.93 clearly indicating its performance is really very good.

## 4. Conclusion

iRNA(m6A)-PseKNC is a powerful predictor for identifying the $m^6A$ sites in *S. cerevisiae* genome. It was developed by using the Euclidean distance to balance out the size of the negative training subset with that of the positive one. In order to consider both the local and global effects, each of the statistical samples in this predictor is formulated with a 22-D PseKNC vector. The learning machine implemented in the new predictor is SVM. It has been observed by the 10-fold cross-validation test that iRNA(m6A)-PseKNC is superior to RAM-NPPS, the state-of-the-art method in this regard. A public-accessible web-server for the new predictor has been established. We anticipate that it will become a very useful high throughput tool for genome analysis.

## Acknowledgements

The authors wish to thank the three anonymous reviewers, whose constructive comments are very helpful for further strengthening the presentation of this paper. This work was supported by the National Nature Scientific Foundation of China (31771471, 61772119), 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), the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015Z006, ZYGX2016J125, ZYGX2016J118).

## Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ab.2018.09.002.

## References

[1] W.A. Cantara, P.F. Crain, J. Rozenski, J.A. McCloskey, K.A. Harris, X. Zhang, F.A. Vendeix, D. Fabris, P.F. Agris, The RNA modification database, RNAMDB: 2011 update, Nucleic Acids Res. 39 (2011) D195–D201.

[2] R. Desrosiers, K. Friderici, F. Rottman, Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells, Proc. Natl. Acad. Sci. U. S. A. 71 (1974) 3971–3975.

[3] G. Cao, H.B. Li, Z. Yin, R.A. Flavell, Recent advances in dynamic m6A RNA modification, Open Biol 6 (2016) 160003.

[4] K. Kariko, M. Buckstein, H. Ni, D. Weissman, Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA, Immunity 23 (2005) 165–175.

[5] K.D. Meyer, S.R. Jaffrey, Rethinking m(6)A readers, writers, and erasers, Annu. Rev. Cell Dev. Biol. 33 (2017) 319–342.

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

[7] T.W. Nilsen, Molecular biology. Internal mRNA methylation finally finds functions, Science 343 (2014) 1207–1208.

[8] K.D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C.E. Mason, S.R. Jaffrey, Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons, Cell 149 (2012) 1635–1646.

[9] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, R. Sorek, G. Rechavi, Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq, Nature 485 (2012) 201–206.

[10] B. Linder, A.V. Grozhik, A.O. Olarerin-George, C. Meydan, C.E. Mason, S.R. Jaffrey, Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome, Nat. Methods 12 (2015) 767–772.

[11] S. Schwartz, D.A. Bernstein, M.R. Mumbach, M. Jovanovic, R.H. Herbst, B.X. Leon-Ricardo, J.M. Engreitz, M. Guttman, R. Satija, E.S. Lander, G. Fink, A. Regev, Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA, Cell 159 (2014) 148–162.

[12] W. Chen, P. Feng, H. Ding, H. Lin, iRNA-Methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33.

[13] W. Chen, H. Tran, Z. Liang, H. Lin, L. Zhang, Identification and analysis of the N (6)-methyladenosine in the Saccharomyces cerevisiae transcriptome, Sci. Rep. 5 (2015) 13859.

[14] Y. Zhou, P. Zeng, Y.H. Li, Z. Zhang, Q. Cui, SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features, Nucleic Acids Res. 44 (2016) e91.

[15] W. Chen, P. Xing, Q. Zou, Detecting N(6)-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines, Sci. Rep. 7 (2017) 40242.

[16] P. Xing, R. Su, F. Guo, L. Wei, Identifying N(6)-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine, Sci. Rep. 7 (2017) 46757.

[17] W. Chen, H. Lin, Recent advances in identification of RNA modifications, Noncoding RNA 3 (2016).

[18] K.C. Chou, H.B. Shen, Recent progresses in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.

[19] J. Jia, Z. Liu, X. Xiao, 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 (2015) 47–56.

[20] B. Liu, L. Fang, F. Liu, X. Wang, J. Chen, Identification of real microRNA precursors with a pseudo structure status composition approach, PLoS One 10 (2015) e0121501.

[21] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J. Biomol. Struct. Dyn. 33 (2015)

2221–2233.
[22] J. Jia, Z. Liu, X. Xiao, B. Liu, iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset, Anal. Biochem. 497 (2016) 48–56.
[23] J. Jia, Z. Liu, X. Xiao, B. Liu, pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, J. Theor. Biol. 394 (2016) 223–230.
[24] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, Oncotarget 8 (2017) 4208–4217.
[25] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Mol. Ther. Nucleic Acids 7 (2017) 155–163.
[26] B. Liu, F. Yang, D.S. Huang, iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, Bioinformatics 34 (2018) 33–40.
[27] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites, Mol. Ther. Nucleic Acids 11 (2018) 468–474.
[28] B. Liu, F. Weng, D.S. Huang, iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty312/4978052.
[29] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, Genomics (2018), https://doi.org/10.1016/j.ygeno.2018.01.005.
[30] B. Liu, K. Li, D.S. Huang, iEnhancer-EL, Identifying enhancers and their strength with ensemble learning approach, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty458.
[31] Z.D. Su, Y. Huang, Z.Y. Zhang, Y.W. Zhao, D. Wang, W. Chen, H. Lin, iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty508.
[32] Y.D. Khan, N. Rasool, W. Hussain, S.A. Khan, iPhosT-PseAAC: identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC, Anal. Biochem. 550 (2018) 109–116.
[33] H. Yang, W.R. Qiu, G. Liu, F.B. Guo, W. Chen, H. Lin, iRSpot-Pse6NC, Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general, PseKNC International Journal of Biological Sciences 14 (2018) 883–891.
[34] X. Cheng, W.Z. Lin, X. Xiao, pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty628.
[35] L. Cai, T. Huang, J. Su, X. Zhang, W. Chen, F. Zhang, L. He, Implications of newly identified brain eQTL genes and their interactors in Schizophrenia, Mol. Ther. Nucleic Acids 12 (2018) 433–442.
[36] Y. Zhang, R. Xie, J. Wang, A. Leier, T.T. Marquez-Lago, T. Akutsu, G.I. Webb, J. Song, Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework, Briefings Bioinf. (2018), https://doi.org/10.1093/bib/bby079.
[37] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review), J. Theor. Biol. 273 (2011) 236–247.
[38] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[39] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.
[40] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins: Structure, Function, and Genetics (Erratum: ibid. 44 (2001) 246–255 60) 43 (2001).
[41] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[42] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo amino acid composition and support vector machine for prediction of enzyme subfamily classes, J. Theor. Biol. 248 (2007) 546–551.
[43] L. Nanni, A. Lumini, Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization, Amino Acids 34 (2008) 653–660.
[44] D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres, Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition, J. Theor. Biol. 257 (2009) 17–26.
[45] M. Esmaeili, H. Mohabatkar, S. Mohsenzadeh, Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses, J. Theor. Biol. 263 (2010) 203–209.
[46] H. Mohabatkar, Prediction of cyclin proteins using Chou's pseudo amino acid composition, Protein Pept. Lett. 17 (2010) 1207–1214.
[47] S.S. Sahu, G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction, Comput. Biol. Chem. 34 (2010) 320–327.
[48] H. Mohabatkar, M. Mohammad Beigi, A. Esmaeili, Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo amino acid composition and support vector machine, J. Theor. Biol. 281 (2011) 18–23.
[49] B.M. Mohammad, M. Behjati, H. Mohabatkar, Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach, J. Struct. Funct. Genom. 12 (2011) 191–197.
[50] M. Hayat, A. Khan, Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of chou's PseAAC, Protein Pept. Lett. 19 (2012) 411–421.
[51] S. Mei, Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning, J. Theor. Biol. 310 (2012) 80–87.
[52] L. Nanni, S. Brahnam, A. Lumini, Wavelet images and Chou's pseudo amino acid composition for protein classification, Amino Acids 43 (2012) 657–665.
[53] M.K. Gupta, R. Niyogi, M. Misra, An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition, SAR QSAR Environ. Res. 24 (2013) 597–609.
[54] M. Khosravian, F.K. Faramarzi, M.M. Beigi, M. Behbahani, H. Mohabatkar, Predicting antibacterial peptides by the concept of chou's pseudo amino acid composition and machine learning methods, Protein Pept. Lett. 20 (2013) 180–186.
[55] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, J. Theor. Biol. 341 (2014) 34–40.
[56] Y. Xu, X. Wen, L.S. Wen, L.Y. Wu, N.Y. Deng, iNitro-Tyr, Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PLoS One 9 (2014) e105018.
[57] M. Hayat, N. Iqbal, Discriminating protein structure classes by incorporating pseudo average chemical shift to chou's general PseAAC and support vector machine, Comput. Meth. Progr. Biomed. 116 (2014) 184–192.
[58] S. Mondal, P.P. Pai, Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction, J. Theor. Biol. 356 (2014) 30–35.
[59] H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, H. Lin, W. Chen, iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels, BioMed Res. Int. 2014 (2014) 286419.
[60] L. Nanni, S. Brahnam, A. Lumini, Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition, J. Theor. Biol. 360 (2014) 109–116.
[61] S. Ahmad, M. Kabir, M. Hayat, Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into chou's general PseAAC, Comput. Meth. Progr. Biomed. 122 (2015) 165–174.
[62] R. Kumar, A. Srivastava, B. Kumari, M. Kumar, Prediction of beta-lactamase and its class by Chou's pseudo amino acid composition and support vector machine, J. Theor. Biol. 365 (2015) 96–103.
[63] K.C. Chou, Y.D. Cai, Predicting protein quaternary structure by pseudo amino acid composition, Protein Struct. Funct. Genet. 53 (2003) 282–289.
[64] M. Behbahani, H. Mohabatkar, M. Nosrati, Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition, J. Theor. Biol. 411 (2016) 1–5.
[65] P.K. Meher, T.K. Sahu, V. Saini, A.R. Rao, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC, Sci. Rep. 7 (2017) 42362.
[66] M. Rahimi, M.R. Bakhtiarizadeh, A. Mohammadi-Sangcheshmeh, OOgenesis_Pred: a sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition, J. Theor. Biol. 414 (2017) 128–136.
[67] M. Tahir, M. Hayat, M. Kabir, Sequence based predictor for discrimination of enhancer and their types by applying general form of Chou's trinucleotide composition, Comput. Meth. Progr. Biomed. 146 (2017) 69–75.
[68] P. Tripathi, P.N. Pandey, 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 (2017) 49–54.
[69] M. Arif, M. Hayat, Z. Jan, iMem-2LSAAC: a two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into Chou's pseudo amino acid composition, J. Theor. Biol. 442 (2018) 11–21.
[70] S. Akbar, M. Hayat, iMethyl-STTNC: identification of N(6)-methyladenosine sites by extending the Idea of SAAC into Chou's PseAAC to formulate RNA sequences, J. Theor. Biol. 455 (2018) 205–211.
[71] E. Contreras-Torres, Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chou's PseAAC, J. Theor. Biol. 454 (2018) 139–145.
[72] Z. Ju, S.Y. Wang, Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition, Gene 664 (2018) 78–83.
[73] Y. Liang, S. Zhang, Identify Gram-negative bacterial secreted protein types by incorporating different modes of PSSM into Chou's general PseAAC via Kullback-Leibler divergence, J. Theor. Biol. 454 (2018) 22–29.
[74] M. Mousavizadegan, H. Mohabatkar, Computational prediction of antifungal peptides via Chou's PseAAC and SVM, J. Bioinf. Comput. Biol. (2018) 1850016.
[75] S.M. Rahman, S. Shatabda, S. Saha, M. Kaykobad, M. Sohel Rahman, DPP-PseAAC: a DNA-binding Protein Prediction model using Chou's general PseAAC, J. Theor. Biol. 452 (2018) 22–34.
[76] E.S. Sankari, D.D. Manimegalai, Predicting membrane protein types by incorporating a novel feature set into Chou's general PseAAC, J. Theor. Biol. 455 (2018) 319–328.
[77] A. Srivastava, R. Kumar, M. Kumar, BlaPred: predicting and classifying beta-lactamase using a 3-tier prediction system via Chou's general PseAAC, J. Theor. Biol. (2018), https://doi.org/10.1016/j.jtbi.2018.08.030.
[78] J. Mei, J. Zhao, Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers, Sci. Rep. 8 (2018) 2359.
[79] K.C. Chou, Y.D. Cai, A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology, Biochem. Biophys. Res. Commun. 311 (2003) 743–747.
[80] J. Mei, J. Zhao, Analysis and prediction of presynaptic and postsynaptic neurotoxins by Chou's general pseudo amino acid composition and motif features, J. Theor. Biol. 427 (2018) 147–153.
[81] M.S. Krishnan, 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 (2018) 62–74.

[82] L. Zhang, L. Kong, iRSpot-ADPM: identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components, J. Theor. Biol. 441 (2018) 1–8.

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

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

[85] P. Du, X. Wang, C. Xu, Y. Gao, PseAAC-Builder, A cross-platform stand-alone program for generating various special Chou's pseudo amino acid compositions, Anal. Biochem. 425 (2012) 117–119.

[86] D.S. Cao, Q.S. Xu, Y.Z. Liang, propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (2013) 960–962.

[87] P. Du, S. Gu, Y. Jiao, PseAAC-General: fast building various modes of general form of Chou's pseudo amino acid composition for large-scale protein datasets, Int. J. Mol. Sci. 15 (2014) 3495–3506.

[88] K.C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, Curr. Proteonomics 6 (2009) 262–274.

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

[90] W. Chen, H. Lin, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. Biosyst. 11 (2015) 2620–2634.

[91] W. Chen, H. Tang, J. Ye, H. Lin, iRNA-PseU: identifying RNA pseudouridine sites Molecular Therapy -, Nucleic Acids 5 (2016) e332.

[92] B. Liu, L. Fang, R. Long, X. Lan, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics 32 (2016) 362–369.

[93] B. Liu, R. Long, iDHS-EL, Identifying DNase I hypersensi-tivesites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework, Bioinformatics 32 (2016) 2411–2418.

[94] B. Liu, S. Wang, R. Long, iRSpot-EL: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (2017) 35–41.

[95] B. Liu, F. Yang, 2L-piRNA, A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function, Mol. Ther. Nucleic Acids 7 (2017) 267–277.

[96] T. Wang, J. Yang, H.B. Shen, Predicting membrane protein types by the LLDA algorithm, Protein Pept. Lett. 15 (2008) 915–921.

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

[98] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Q. Dong, Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, Bioinformatics 30 (2014) 472–479.

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

[100] J. Chen, R. Long, X.L. Wang, B. Liu, dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation, Sci. Rep. 6 (2016) 32333.

[101] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, iDNA6mA-PseKNC: Identifying DNA N (6)-methyladenosine Sites by Incorporating Nucleotide Physicochemical Properties into PseKNC, Genomics (2018).

[102] H. Lin, Z.Y. Liang, H. Tang, W. Chen, Identifying sigma70 promoters with novel pseudo nucleotide composition, IEEE ACM Trans. Comput. Biol. Bioinf (2017).

[103] J. Zhang, P. Feng, H. Lin, W. Chen, Identifying RNA N(6)-methyladenosine sites in Escherichia coli genome, Front. Microbiol. 9 (2018) 955.

[104] J. Chen, H. Liu, J. Yang, Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, Amino Acids 33 (2007) 423–428.

[105] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.

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

[107] W. Chen, P.M. Feng, H. Lin, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.

[108] Y. Xu, X.J. Shao, L.Y. Wu, N.Y. Deng, iSNO-AAPair, Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, PeerJ 1 (2013) e171.

[109] H. Lin, E.Z. Deng, H. Ding, W. Chen, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972.

[110] P.M. Feng, W. Chen, H. Lin, iHSP-PseRAAAC, Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, Anal. Biochem. 442 (2013) 118–125.

[111] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, PLoS One 9 (2014) e106691.

[112] Y. Xu, X. Wen, X.J. Shao, N.Y. Deng, iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition, Int. J. Mol. Sci. 15 (2014) 7594–7610.

[113] W.R. Qiu, X. Xiao, W.Z. Lin, Imethyl-pseaac: identification of protein methylation sites via a pseudo amino acid composition approach, BioMed Res. Int. 2014 (2014) 947416.

[114] Y.N. Fan, X. Xiao, J.L. Min, iNR-Drug, Predicting the interaction of drugs with nuclear receptors in cellular networking, Intenational Journal of Molecular Sciences (IJMS) 15 (2014) 4915–4937.

[115] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, W. Chen, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.

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

[117] W. Chen, P.M. Feng, H. Lin, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, BioMed Res. Int. 2014 (2014) 623149.

[118] W. Chen, P.M. Feng, E.Z. Deng, H. Lin, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83.

[119] W.R. Qiu, X. Xiao, W.Z. Lin, iUbiq-Lys, Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model, J. Biomol. Struct. Dyn. 33 (2015) 1731–1742.

[120] Z. Liu, X. Xiao, W.R. Qiu, iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77.

[121] B. Liu, L. Fang, S. Wang, X. Wang, H. Li, Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, J. Theor. Biol. 385 (2015) 153–159.

[122] R. Xu, J. Zhou, B. Liu, Y.A. He, Q. Zou, X. Wang, Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach, J. Biomol. Struct. Dyn. 33 (2015) 1720–1730.

[123] W. Chen, P. Feng, H. Ding, H. Lin, Using deformation energy to analyze nucleosome positioning in genomes, Genomics 107 (2016) 69–75.

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

[125] Z. Liu, X. Xiao, D.J. Yu, J. Jia, W.R. Qiu, pRNAm-PC: predicting N-methyladenosine sites in RNA sequences via physical-chemical properties, Anal. Biochem. 497 (2016) 60–67.

[126] X. Xiao, H.X. Ye, Z. Liu, J.H. Jia, iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition, Oncotarget 7 (2016) 34180–34189.

[127] J. Jia, Z. Liu, X. Xiao, B. Liu, Ippbs-opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets, Molecules 21 (2016) E95.

[128] W.R. Qiu, X. Xiao, Z.C. Xu, iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier, Oncotarget 7 (2016) 51270–51283.

[129] C.J. Zhang, H. Tang, W.C. Li, H. Lin, W. Chen, iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition, Oncotarget 7 (2016) 69783–69793.

[130] L. Cai, W. Yuan, Z. Zhang, L. He, In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data, Sci. Rep. 6 (2016) 36540.

[131] B. Liu, L. Fang, F. Liu, X. Wang, iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach, J. Biomol. Struct. Dyn. 34 (2016) 223–235.

[132] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC, Oncotarget 7 (2016) 44310–44321.

[133] J. Jia, Z. Liu, X. Xiao, B. Liu, Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC), J. Biomol. Struct. Dyn. 34 (2016) 1946–1961.

[134] J. Jia, Z. Liu, X. Xiao, B. Liu, iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC, Oncotarget 7 (2016) 34558–34570.

[135] W. Chen, H. Ding, P. Feng, H. Lin, iACP: a sequence-based tool for identifying anticancer peptides, Oncotarget 7 (2016) 16895–16909.

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

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

[138] Y. Xu, C. Li, iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC, Med. Chem. 13 (2017) 544–551.

[139] W.R. Qiu, B.Q. Sun, X. Xiao, D. Xu, iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory, Molecular Informatics 36 (2017) UNSP 1600010.

[140] L.M. Liu, Y. Xu, 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 (2017) 552–559.

[141] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, J.H. Jia, iKcr-PseEns, Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics 110 (2018) 239–246.

[142] J. Song, Y. Wang, F. Li, T. Akutsu, N.D. Rawlings, G.I. Webb, iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Briefings Bioinf. (2018), https://doi.org/10.1093/bib/bby028.

[143] A. Ehsan, K. Mahmood, Y.D. Khan, S.A. Khan, A novel modeling in mathematical biology for classification of signal peptides, Sci. Rep. 8 (2018) 1039.

[144] F. Li, C. Li, T.T. Marquez-Lago, A. Leier, T. Akutsu, A.W. Purcell, A.I. Smith, T. Lightow, R.J. Daly, J. Song, Quokka: a comprehensive tool for rapid and

accurate prediction of kinase family-specific phosphorylation sites in the human proteome, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty522.

[145] X. Cheng, X. Xiao, pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC, Mol. Biosyst. 13 (2017) 1722–1727.

[146] X. Cheng, X. Xiao, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, Gene (Erratum: ibid. 644 (2018) 315–321 156-156) 628 (2017).

[147] X. Xuao, X. Cheng, G. Chen, Q. Mao, pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC, Genomics (2018), https://doi.org/10.1016/j.ygeno.2018.05.017.

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

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

[150] X. Cheng, X. Xiao, pLoc-mGneg: predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC, Genomics 110 (2018) 231–239.

[151] X. Cheng, X. Xiao, pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, Genomics 110 (2018) 50–58.

[152] X. Cheng, S.G. Zhao, X. Xiao, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics (Corrigendum, ibid. 33 (2017) 341–346 2610) 33 (2017).

[153] X. Cheng, S.G. Zhao, X. Xiao, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (2017) 58494–58503.

[154] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iPTM-mLys: identifying multiple lysine PTM sites and their different types, Bioinformatics 32 (2016) 3116–3123.

[155] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. Biosyst. 9 (2013) 1092–1100.

[156] K.C. Chou, H.B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2009) 63–92.

[157] J. Wang, B. Yang, J. Revote, A. Leier, T.T. Marquez-Lago, G. Webb, J. Song, T. Lithgow, POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles, Bioinformatics 33 (2017) 2756–2758.

[158] Z. Chen, P.Y. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, J. Song, iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (2018) 2499–2502.

[159] J. Song, F. Li, A. Leier, T.T. Marquez-Lago, T. Akutsu, G. Haffari, G.I. Webb, R.N. Pike, PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy, Bioinformatics 34 (2018) 684–687.

[160] J. Song, F. Li, K. Takemoto, G. Haffari, T. Akutsu, G.I. Webb, PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural and network features in a machine learning framework, J. Theor. Biol. 443 (2018) 125–137.

[161] X. Cheng, X. Xiao, pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information, Bioinformatics 34 (2018) 1448–1456.

[162] J. Wang, B. Yang, A. Leier, T.T. Marquez-Lago, M. Hayashida, A. Rocker, Z. Yanju, T. Akutsu, R.A. Strugnell, J. Song, T. Lithgow, Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors, Bioinformatics 34 (2018) 2546–2555.

[163] K.C. Chou, S. Forsen, Graphical rules for enzyme-catalyzed rate laws, Biochem. J. 187 (1980) 829–835.

[164] G.P. Zhou, M.H. Deng, An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways, Biochem. J. 222 (1984) 169–176.

[165] K.C. Chou, Graphic rules in steady and non-steady enzyme kinetics, J. Biol. Chem. 264 (1989) 12074–12079.

[166] I.W. Althaus, A.J. Gonzales, J.J. Chou, M.R. Diebel, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase, J. Biol. Chem. 268 (1993) 14875–14880.

[167] I.W. Althaus, J.J. Chou, A.J. Gonzales, M.R. Diebel, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E, Biochemistry 32 (1993) 6548–6554.

[168] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen, Using cellular automata to generate Image representation for biological sequences, Amino Acids 28 (2005) 29–35.

[169] Z.C. Wu, X. Xiao, 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids, J. Theor. Biol. 267 (2010) 29–34.

[170] K.C. Chou, W.Z. Lin, X. Xiao, Wenxiang: a web-server for drawing wenxiang diagrams, Nat. Sci. 3 (2011) 862–865.

[171] G.P. Zhou, The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism, J. Theor. Biol. 284 (2011) 142–148.

[172] J.A. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2005) 861–874.