molecules

MDPI

Article

Identification of D Modification Sites by Integrating Heterogeneous Features in *Saccharomyces cerevisiae*

Pengmian Feng [1,5,*,†], Zhaochun Xu [2,3,†], Hui Yang [2], Hao Lv [2], Hui Ding [2] and Li Liu [4,*]

1 Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China
2 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; jdzxuzhaochun@163.com (Z.X.); huiyang0325@163.com (H.Y.); 13208188368@163.com (H.L.); hding@uestc.edu.cn (H.D.)
3 Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333403, China
4 Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China
5 School of Public Health, North China University of Science and Technology, Tangshan 063000, China
* Correspondence: fengpengmian@gmail.com (P.F.); liliu2010imu@163.com (L.L.); Tel.: +86-315-3725715 (P.F. & L.L.)
† These authors contributed equally to this work.

Academic Editors: Xiangxiang Zeng, Alfonso Rodríguez-Patón and Quan Zou
Received: 2 December 2018; Accepted: 17 December 2018; Published: 22 January 2019

Abstract: As an abundant post-transcriptional modification, dihydrouridine (D) has been found in transfer RNA (tRNA) from bacteria, eukaryotes, and archaea. Nonetheless, knowledge of the exact biochemical roles of dihydrouridine in mediating tRNA function is still limited. Accurate identification of the position of D sites is essential for understanding their functions. Therefore, it is desirable to develop novel methods to identify D sites. In this study, an ensemble classifier was proposed for the detection of D modification sites in the *Saccharomyces cerevisiae* transcriptome by using heterogeneous features. The jackknife test results demonstrate that the proposed predictor is promising for the identification of D modification sites. It is anticipated that the proposed method can be widely used for identifying D modification sites in tRNA.

Keywords: dihydrouridine; nucleotide physicochemical property; pseudo dinucleotide composition; RNA secondary structure; ensemble classifier

1. Introduction

To date, more than 100 kinds of post-transcriptional modifications have been identified in transfer RNAs (tRNAs). It has been demonstrated that these modifications are involved in all core aspects of tRNA function [1]. Among them, dihydrouridine (D) is a prevalent tRNA modification, which has been found in the three domains of life [2].

The D modification is formed by a dihydrouridine synthase [3]. Unlike uridine (U), the ring of D is not aromatic, which precludes its interactions with other bases in tRNA by stacking interactions [4,5]. By destabilizing the tRNA structure, D can enhance the conformational flexibility of tRNA [6]. Therefore, it is concluded that the flexibility and even the folding of tRNA could be affected by D modification [4,7].

Recent studies have also shown that tRNA lacking D degrades significantly faster, suggesting that D modification can protect tRNAs from degradation [1,8]. Despite the abundant occurrence of D modification, our knowledge about its roles in mediating tRNA biological functions is still limited.

Therefore, it is urgent to develop novel methods to describe the distribution of D modification sites. Since it is cost ineffective and labor intensive to detect D modification sites by using experimental techniques, it is necessary to develop theoretical methods for the detection of D modification.

Therefore, in the present study, an ensemble classifier was proposed for the detection of D modification sites in the *Saccharomyces cerevisiae* transcriptome, in which the nucleotide physicochemical property, pseudo dinucleotide composition, and secondary structure component were employed to train the basic predictors, respectively. In the jackknife test, the ensemble classifier obtained an accuracy of 83.09% for identifying D modification sites. This result demonstrated the superiority of the proposed method for identifying D modification sites in the *S. cerevisiae* transcriptome.

2. Results

2.1. Performances of Different Features

In order to demonstrate the effectiveness of the different kinds of features for identifying D sites, we first built support vector machine (SVM) predictors based on each kind of sequence encoding schemes (i.e., nucleotide physicochemical property, pseudo dinucleotide composition, or secondary structure component). Their jackknife test results for identifying D sites in the *S. cerevisiae* transcriptome are reported in Table 1. Although the nucleotide-physicochemical-property-based predictor (NPCP-SVM) obtained the highest accuracy (Acc) for identifying D sites, its sensitivity (Sn) was only 67.65%, indicating that it still could not accurately identify the real D sites. For the predictors based on pseudo dinucleotide composition and secondary structure component (namely PseDNC-SVM and SSC-SVM), their accuracies (Acc) were only 75.74% and 72.79% with the atthews correlation coefficients (MCC) of 0.5 and 0.45, respectively. Taken together, these results indicate that the performances of the aforementioned three predictors were not fully satisfactory. Therefore, there is still scope to improve the performance for identifying D sites.

Table 1. Performances of different methods for identifying dihydrouridine (D) sites.

Methods	Sn (%)	Sp (%)	Acc (%)	MCC
NPCP-SVM	67.65	100	83.82	0.59
PseDNC-SVM	73.53	77.94	75.74	0.50
SSC-SVM	70.59	75.00	72.79	0.45
Ensemble SVM	76.47	89.71	83.09	0.62

2.2. Improving Predictive Performance Using Ensemble Learning

Several recent works have demonstrated that the ensemble learning scheme can improve the performance of predictors [9–13]. In order to improve the performance of identifying D sites, we constructed an ensemble predictor based on SVM by using different kinds of features. Therefore, three basic SVM-based predictors were built by using nucleotide physicochemical property, pseudo dinucleotide composition, and secondary structure component, respectively. Figure 1 shows the prediction process with the ensemble classifier. The three predictors were integrated as an ensemble predictor via a voting strategy (see Materials and Methods). By combining the results of the three predictors together, a sequence in the benchmark dataset was predicted as a D-site-containing sequence if its prediction probabilities yielded by more than two predictors were all greater than 0.5.

The jackknife test results of the ensemble predictor for identifying D sites in *S. cerevisiae* transcriptome are also listed in Table 1. It was found that the sensitivity of the ensemble predictor was improved to 76.47%. Although its specificity and accuracy was a little lower than NPCP-SVM, the MCC of the ensemble predictor was 0.62, which was higher than that of any single SVM-based predictor, indicating the ensemble predictor was much more stable than NPCP-SVM, PseDNC-SVM, and SSC-SVM for the detection of D modification sites.

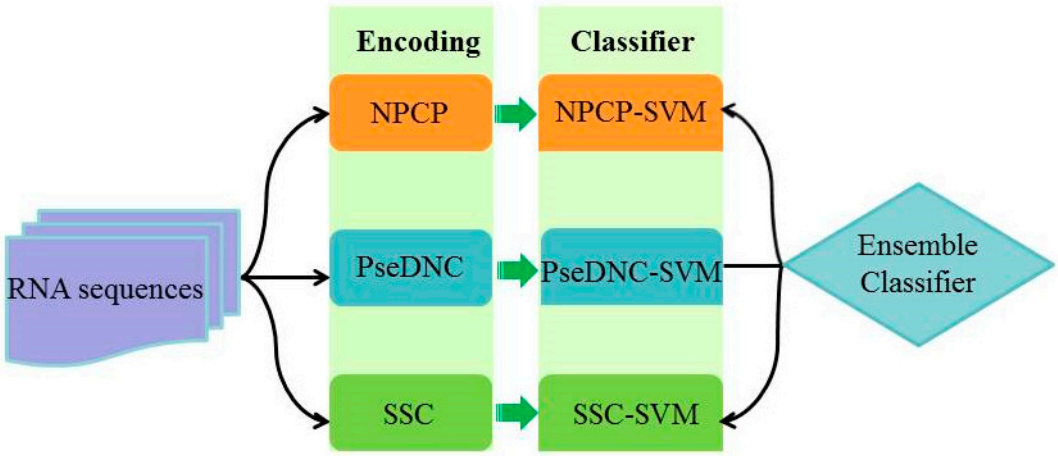

Figure 1. The flaw chart of the ensemble classifiers. NPCP-SVM stands for nucleotide-physicochemical-property-based predictor; PseDNC-SVM stands for pseudo-dinucleotide-composition-based predictor; SSC-SVM stands for secondary-structure-based predictor.

3. Materials and Methods

3.1. Benchmark Dataset

The original 208 positive samples (D-site-containing sequences) were fetched from the RMBase database [14]. All of these sequences in RMBase were 41 nt long with the D site in the center. Preliminary tests indicated that the best prediction results were achieved when the sequence was 41 nt long. In order to avoid redundancy, sequences with more than 80% sequence similarity were removed using the CD-HIT program [15]. Accordingly, we obtained 68 D-site-containing sequences from the *S. cerevisiae* transcriptome.

Negative samples were obtained by selecting 41-nt-long sequences that satisfied the following rules: (1) uridine is the center of the sequence, and (2) no dihydrouridine modification of the centered uridine has been identified experimentally. Accordingly, we could obtain a huge number of negative samples, from which we randomly picked 68 samples to form the negative subset for the purpose of using a balance benchmark dataset to train the model. In summary, our benchmark dataset comprised 68 D-site-containing sequences and 68 false D-site-containing sequences from the *S. cerevisiae* transcriptome, which is available at https://github.com/chenweiimu/D-Pred.

3.2. Sequence Encoding Scheme

3.2.1. Nucleotide Physicochemical Property (NPCP)

Adenosine (A), cytosine (C), guanine (G), and uridine (U) have different chemical properties [16,17]. In terms of ring structures, A and G are purines containing two rings, whereas C and U are pyrimidines containing one ring. When forming secondary structures, C and G form strong hydrogen bonds, whereas A and U form weak hydrogen bonds. In terms of amino/keto bases, A and C belong to the amino group, while G and U belong to the keto group [16,17].

In order to encode RNA sequences using these properties, the (x, y, z) coordinates were used to describe the chemical properties of the four nucleotides, and a value of 0 or 1 was assigned to (x, y, z), respectively. If x, y, and z coordinates stand for the ring structure, the hydrogen bond, and the amino/keto bases, A, C, G, and U can be represented by (1, 1, 1), (0, 0, 1), (1, 0, 0), and (0, 1, 0), respectively.

Accordingly, by using nucleotide chemical properties, each sequence could be encoded by a 123 (3×41)-dimensional vector, as given bellow:

$$\mathbf{R}_1 = \left[\begin{array}{ccccccc} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \cdots & \varepsilon_i & \cdots & \varepsilon_{123} \end{array} \right]^{\mathbf{T}} \tag{1}$$

where ε_i indicates the abovementioned nucleotide chemical properties, and its value is 0 or 1.

3.2.2. Pseudo Dinucleotide Composition

The pseudo *k*-tuple nucleotide composition (PseKNC), proposed by Chen et al. [18,19], has been successfully and widely applied in computational genomics [20–22]. PseKNC not only includes local sequence order information but also the global sequence pattern [23]. In the current study, the pseudo dinucleotide composition (PseDNC) was used to encode the RNA sequences and is defined as follows [18,19]:

$$\mathbf{R} = \left[\begin{array}{ccccccc} d_1 & d_2 & \cdots & d_{16} & d_{16+1} & \cdots & d_{16+\lambda} \end{array} \right]^{\mathbf{T}} \tag{2}$$

where

$$d_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (1 \le u \le 16) \\ \frac{w\theta_{u-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (16 < u \le 16 + \lambda) \end{cases}. \tag{3}$$

In Equation (3), f_u $(u = 1, 2, \cdots, 16)$ is the normalized occurrence frequency of the u-th nonoverlapping dinucleotide in the RNA sequence, and

$$\theta_j = \frac{1}{\mathrm{L}-j-1} \sum_{i=1}^{\mathrm{L}-j-1} \mathrm{C}_{i,\, i+j} \quad (j = 1, 2, \cdots, \lambda;\ \lambda < \mathrm{L}) \tag{4}$$

where θ_j is the j-tier correlation factor that reflects the sequence order correlation between all the j-th most contiguous dinucleotides. The coupling factor $\mathrm{C}_{i,\, i+j}$ is defined as

$$\mathrm{C}_{i,\, i+j} = \frac{1}{\mu} \sum_{g=1}^{\mu} \left[P_g(\mathrm{D}_i) - P_g\left(\mathrm{D}_{i+j}\right) \right]^2 \tag{5}$$

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

Inspired by a recent study [24], the three RNA physicochemical properties, namely, enthalpy [25], entropy [25], and free energy [26], were used to define PseDNC. Thus, in Equation (4), μ is equal to 3. The normalized numerical values of the three physicochemical properties of the 16 different RNA dinucleotides were obtained from our previous work [24].

The two parameters w and λ were optimized in the following ranges [0, 1] and [1, 10] with steps of 0.1 and 1, respectively. In the current work, the optimal values for w and λ were 0.5 and 4, respectively. Hence, the RNA sequence can be formulated by a (16 + 4) = 20-dimensional vector as given below:

$$\mathbf{R}_2 = \left[\begin{array}{ccccccc} d_1 & d_2 & \cdots & d_{16} & d_{17} & \cdots & d_{20} \end{array} \right]^{\mathbf{T}} \tag{6}$$

3.2.3. Secondary Structure Component (SSC)

Considering the fact that RNA modification is affected by its structures [27], the RNA sequences were also encoded using the RNA secondary structures. By using the RNAfold tool (version 2.1.9) in ViennaRNA package with default parameters [28], we obtained the secondary structure status at each position, which was represented by brackets ("(" or ")") indicating paired nucleotides and by dots (".") indicating unpaired nucleotides. In the current study, we did not distinguish "(" and ")" and

used "(" for both situations. For a given trinucleotide, there were eight (2^3) possible structure statuses (i.e., "(((", "((.", "(..", "(.(", ".((", ".(.", "..(", and " ... "). If the first nucleotide in the trinucleotide was further considered, there would be 32 (4×8) possible sequence-structure modes, which were denoted as "A-(((", "A-((.", "A-(..", ... , and "U- ... ". Therefore, a given sequence could be represented by using the following sequence-structure:

$$\mathbf{R}_3 = \left[f^A_{(((}, f^A_{((.}, f^A_{(..}, \ldots, f^A_{...}, f^C_{(((}, \ldots, f^U_{...} \right]^{\mathbf{T}}. \quad (7)$$

The elements in the vector of R_3 indicate the frequency of the 32 sequence-structure modes.

3.3. Support Vector Machine

SVM is a well-known machine learning method for pattern recognition and has been widely used in bioinformatics [29–35]. In the current study, the LibSVM package 3.18 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used to perform SVM. Due to its effectiveness and speed in training process, the radial basis kernel function (RBF) of SVM was often used to find the classification hyperplane. The regularization parameter C and kernel parameter γ of the SVM operation engine was optimized in the ranges of $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{-5}]$ with steps of 2 and 2^{-1}, respectively. The prediction was made according to the probability score yielded from SVM. If its probability score was greater than 0.5, a uridine would be predicted as a D site, otherwise, a non-D-site.

3.4. Ensemble Classifiers

By using the NPCP, PseKNC, and SSC features, three basic classifiers were built, which voted for the final result according to the following rule [9]:

$$\mathrm{V}_i = \sum_{k=1}^{3} f(pre(C_k), Class_i) \quad (i = 1, 2) \quad (8)$$

where V_i is the voting score for the sequence belonging to the Class_i. $f(pre(C_k),\mathrm{Class}_i)$ is defined as

$$f(pre(C_k), Class_i) = \begin{cases} 1 & if\ pre(C_k) \in Class_i \\ 0 & if\ pre(C_k) \notin Class_i \end{cases} (i = 1, 2; k = 1, 2, 3). \quad (9)$$

The final prediction is determined by

$$Sgn(i) = \mathrm{argmax}_i\{V_i\} \quad (i = 1, 2)\ . \quad (10)$$

$\mathrm{Sgn}(i)$ is the argument that maximizes the voting score V_i.

3.5. Performance Evaluation

The performance of the method were evaluated by using sensitivity (Sn), specificity (Sp), accuracy (Acc), and the Matthews correlation coefficient (MCC), as given below [36–40]:

$$\begin{cases} \mathrm{Sn} = 1 - \frac{N^+_-}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \frac{N^-_+}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = 1 - \frac{N^+_- + N^-_+}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \frac{1 - \left(\frac{N^+_-}{N^+} + \frac{N^-_+}{N^-} \right)}{\sqrt{\left(1 + \frac{N^-_+ - N^+_-}{N^+} \right) \left(1 + \frac{N^+_- - N^-_+}{N^-} \right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \quad (11)$$

where N^+ represents the total number of D-site-containing sequences, while N_-^+ is the number of D-site-containing sequences incorrectly predicted to be of false D-site-containing sequences. N^- is the total number of false D-site-containing sequences, while N_+^- the number of the false D-site-containing sequences incorrectly predicted to be of D-site-containing sequences.

3.6. Jackknife Cross-Validation

Among the three methods (i.e., independent dataset test, K-fold cross-validation test, and jackknife cross-validation), the jackknife cross-validation is deemed to be the least arbitrary, as demonstrated by in a recent review paper [41]. In the jackknife cross-validation, each sample in the training dataset is in turn singled out as an independent test sample and all the rule parameters are calculated without including the one being identified [42–46]. Accordingly, jackknife cross-validation was also used to examine the performance of the method proposed in the current study.

4. Conclusions

In this study, by integrating heterogeneous sequence-based features, a SVM-based ensemble classifier was proposed to identify D modification sites in the *S. cerevisiae* transcriptome. In this predictor, not only was the local and global sequence information included by encoding RNA sequences using PseDNC, but the nucleotide chemical properties and structures were also considered by representing RNA sequences using nucleotide physicochemical properties and predicted RNA secondary structures. The jackknife test results demonstrate that the proposed predictor is promising for the identification of D modification sites. It is anticipated that the proposed method will become an essential computational tool for identifying D modification sites in tRNA.

However, the proposed method has two flaws. The limited number of experimentally verified D modification data hindered us from extracting effective features to describe the D modification sites containing sequences. The other shortcoming is that the present method directly uses the entirety of the features, which may reduce the generalization capacity of the model and increase the computational time. Therefore, in future work, we shall make efforts to collect more D modification data and also employ the feature selection method to winnow out the optimal features.

Author Contributions: P.F. and L.L. conceived and designed the experiments; P.F., Z.X., H.Y., H.L., and H.D. performed the experiments; P.F. and L.L. wrote the paper.

Funding: This work was supported by the National Nature Scientific Foundation of China (31771471, 61772119) and the Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dyubankova, N.; Sochacka, E.; Kraszewska, K.; Nawrot, B.; Herdewijn, P.; Lescrinier, E. Contribution of dihydrouridine in folding of the D-arm in tRNA. *Organ. Biomol. Chem.* **2015**, *13*, 4960–4966. [CrossRef] [PubMed]
2. Sprinzl, M.; Horn, C.; Brown, M.; Ioudovitch, A.; Steinberg, S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **1998**, *26*, 148–153. [CrossRef] [PubMed]
3. Yu, F.; Tanaka, Y.; Yamashita, K.; Suzuki, T.; Nakamura, A.; Hirano, N.; Suzuki, T.; Yao, M.; Tanaka, I. Molecular basis of dihydrouridine formation on tRNA. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 19593–19598. [CrossRef]
4. Jones, C.I.; Spencer, A.C.; Hsu, J.L.; Spremulli, L.L.; Martinis, S.A.; DeRider, M.; Agris, P.F. A counterintuitive Mg^{2+}-dependent and modification-assisted functional folding of mitochondrial tRNAs. *J. Mol. Biol.* **2006**, *362*, 771–786. [CrossRef] [PubMed]
5. Dalluge, J.J.; Hashizume, T.; Sopchik, A.E.; McCloskey, J.A.; Davis, D.R. Conformational flexibility in RNA: The role of dihydrouridine. *Nucleic Acids Res.* **1996**, *24*, 1073–1079. [CrossRef]
6. Kasprzak, J.M.; Czerwoniec, A.; Bujnicki, J.M. Molecular evolution of dihydrouridine synthases. *BMC Bioinform.* **2012**, *13*, 153. [CrossRef]

7. Whelan, F.; Jenkins, H.T.; Griffiths, S.C.; Byrne, R.T.; Dodson, E.J.; Antson, A.A. From bacterial to human dihydrouridine synthase: Automated structure determination. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2015**, *71*, 1564–1571. [CrossRef]
8. Alexandrov, A.; Chernyakov, I.; Gu, W.; Hiley, S.L.; Hughes, T.R.; Grayhack, E.J.; Phizicky, E.M. Rapid tRNA decay can result from lack of nonessential modifications. *Mol. Cell* **2006**, *21*, 87–96. [CrossRef] [PubMed]
9. Chen, W.; Xing, P.; Zou, Q. Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. *Sci. Rep.* **2017**, *7*, 40242. [CrossRef] [PubMed]
10. Jia, C.; Zuo, Y.; Zou, Q. O-GlcNAcPRED-II: An integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics* **2018**, *34*, 2029–2036. [CrossRef]
11. Zou, Q.; Guo, J.; Ju, Y.; Wu, M.; Zeng, X.; Hong, Z. Improving tRNAscan-SE Annotation Results via Ensemble Classifiers. *Mol. Inform.* **2015**, *34*, 761–770. [CrossRef] [PubMed]
12. Chen, W.; Feng, P.M.; Lin, H.; Chou, K.C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **2013**, *41*, e68. [CrossRef] [PubMed]
13. Wan, S.; Duan, Y.; Zou, Q. HPSLPred: An Ensemble Multi-label Classifier for Human Protein Subcellular Location Prediction with Imbalanced Source. *Proteomics* **2017**, *17*, 1700262. [CrossRef] [PubMed]
14. Xuan, J.J.; Sun, W.J.; Lin, P.H.; Zhou, K.R.; Liu, S.; Zheng, L.L.; Qu, L.H.; Yang, J.H. RMBase v2.0: Deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Res.* **2018**, *46*, D327–D334. [CrossRef] [PubMed]
15. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
16. Zhang, J.; Feng, P.; Lin, H.; Chen, W. Identifying RNA N(6)-Methyladenosine Sites in *Escherichia coli* Genome. *Front. Microbiol.* **2018**, *9*, 955. [CrossRef] [PubMed]
17. Feng, P.; Ding, H.; Yang, H.; Chen, W.; Lin, H.; Chou, K.-C. iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC. *Mol. Ther.-Nucleic Acids* **2017**, *7*, 155–163. [CrossRef]
18. Chen, W.; Lei, T.-Y.; Jin, D.-C.; Lin, H.; Chou, K.-C. PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* **2014**, *456*, 53–60. [CrossRef]
19. Chen, W.; Zhang, X.; Brooker, J.; Lin, H.; Zhang, L.; Chou, K.-C. PseKNC-General: A cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* **2014**, *31*, 119–120. [CrossRef]
20. Chen, W.; Feng, P.-M.; Deng, E.-Z.; Lin, H.; Chou, K.-C. iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.* **2014**, *462*, 76–83. [CrossRef]
21. Chen, W.; Feng, P.-M.; Lin, H.; Chou, K.-C. iSS-PseDNC: Identifying Splicing Sites Using Pseudo Dinucleotide Composition. *BioMed Res. Int.* **2014**. [CrossRef] [PubMed]
22. Lin, H.; Liang, Z.Y.; Tang, H.; Chen, W. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**. [CrossRef]
23. Chen, W.; Lin, H.; Chou, K.-C. Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences. *Mol. BioSyst.* **2015**, *11*, 2620–2634. [CrossRef]
24. Chen, W.; Feng, P.; Ding, H.; Lin, H.; Chou, K.-C. iRNA-Methyl: Identifying N-6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* **2015**, *490*, 26–33. [CrossRef]
25. Freier, S.M.; Kierzek, R.; Jaeger, J.A.; Sugimoto, N.; Caruthers, M.H.; Neilson, T.; Turner, D.H. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 9373–9377. [CrossRef] [PubMed]
26. Xia, T.; SantaLucia, J., Jr.; Burkard, M.E.; Kierzek, R.; Schroeder, S.J.; Jiao, X.; Cox, C.; Turner, D.H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **1998**, *37*, 14719–14735. [CrossRef] [PubMed]
27. Lu, X.J.; Olson, W.K.; Bussemaker, H.J. The RNA backbone plays a crucial role in mediating the intrinsic stability of the GpU dinucleotide platform and the GpUpA/GpA miniduplex. *Nucleic Acids Res.* **2010**, *38*, 4868–4876. [CrossRef]
28. Lorenz, R.; Bernhart, S.H.; Honer Zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB* **2011**, *6*, 26. [CrossRef]

29. Feng, C.Q.; Zhang, Z.Y.; Zhu, X.J.; Lin, Y.; Chen, W.; Tang, H.; Lin, H. iTerm-PseKNC: A sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* **2018**. [CrossRef]
30. Su, Z.D.; Huang, Y.; Zhang, Z.Y.; Zhao, Y.W.; Wang, D.; Chen, W.; Chou, K.C.; Lin, H. iLoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* **2018**. [CrossRef]
31. Chen, W.; Yang, H.; Feng, P.; Ding, H.; Lin, H. iDNA4mC: Identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* **2017**, *33*, 3518–3523. [CrossRef] [PubMed]
32. Li, D.; Ju, Y.; Zou, Q. Protein Folds Prediction with Hierarchical Structured SVM. *Curr. Proteomics* **2016**, *13*, 79–85. [CrossRef]
33. Wang, S.P.; Zhang, Q.; Lu, J.; Cai, Y.D. Analysis and Prediction of Nitrated Tyrosine Sites with the mRMR Method and Support Vector Machine Algorithm. *Curr. Bioinform.* **2018**, *13*, 3–13. [CrossRef]
34. Yang, H.; Lv, H.; Ding, H.; Chen, W.; Lin, H. iRNA-2OM: A Sequence-Based Predictor for Identifying 2′-O-Methylation Sites in Homo sapiens. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **2018**, *25*, 1266–1277. [CrossRef] [PubMed]
35. Dao, F.Y.; Lv, H.; Wang, F.; Feng, C.Q.; Ding, H.; Chen, W.; Lin, H. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics* **2018**. [CrossRef] [PubMed]
36. Feng, P.-M.; Chen, W.; Lin, H.; Chou, K.-C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* **2013**, *442*, 118–125. [CrossRef] [PubMed]
37. Song, J.; Wang, Y.; Li, F.; Akutsu, T.; Rawlings, N.D.; Webb, G.I.; Chou, K.C. iProt-Sub: A comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Briefings Bioinform.* **2018**. [CrossRef]
38. Zhu, X.J.; Feng, C.Q.; Lai, H.Y.; Chen, W.; Lin, H. Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl.-Based Syst.* **2018**. [CrossRef]
39. Yang, H.; Qiu, W.R.; Liu, G.Q.; Guo, F.B.; Chen, W.; Chou, K.C.; Lin, H. iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci.* **2018**, *14*, 883–891. [CrossRef]
40. Tang, H.; Zhao, Y.W.; Zou, P.; Zhang, C.M.; Chen, R.; Huang, P.; Lin, H. HBPred: A tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* **2018**, *14*, 957–964. [CrossRef]
41. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef] [PubMed]
42. Feng, P.M.; Lin, H.; Chen, W. Identification of antioxidants from sequence information using naive Bayes. *Comput. Math. Methods Med.* **2013**, *2013*, 567529. [CrossRef] [PubMed]
43. Feng, P.M.; Ding, H.; Chen, W.; Lin, H. Naive Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.* **2013**, *2013*, 530696. [CrossRef] [PubMed]
44. Lai, H.Y.; Chen, X.X.; Chen, W.; Tang, H.; Lin, H. Sequence-based predictive modeling to identify cancerlectins. *Oncotarget* **2017**, *8*, 28169–28175. [CrossRef] [PubMed]
45. Yang, H.; Tang, H.; Chen, X.X.; Zhang, C.J.; Zhu, P.P.; Ding, H.; Chen, W.; Lin, H. Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition. *BioMed Res. Int.* **2016**, *2016*, 5413903. [CrossRef] [PubMed]
46. Chen, X.X.; Tang, H.; Li, W.C.; Wu, H.; Chen, W.; Ding, H.; Lin, H. Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. *BioMed Res. Int.* **2016**, *2016*, 1654623. [CrossRef] [PubMed]

Sample Availability: Samples of the compounds are not available from the authors.