Molecular Therapy
Nucleic Acids
Original Article

iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC

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

[1]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; [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; [3]Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China; [4]Gordon Life Science Institute, Boston, MA 02478, USA

There are many different types of RNA modifications, which are essential for numerous biological processes. Knowledge about the occurrence sites of RNA modifications in its sequence is a key for in-depth understanding of their biological functions and mechanism. Unfortunately, it is both time-consuming and laborious to determine these sites purely by experiments alone. Although some computational methods were developed in this regard, each one could only be used to deal with some type of modification individually. To our knowledge, no method has thus far been developed that can identify the occurrence sites for several different types of RNA modifications with one seamless package or platform. To address such a challenge, a novel platform called "iRNA-PseColl" has been developed. It was formed by incorporating both the individual and collective features of the sequence elements into the general pseudo K-tuple nucleotide composition (PseKNC) of RNA via the chemicophysical properties and density distribution of its constituent nucleotides. Rigorous cross-validations have indicated that the anticipated success rates achieved by the proposed platform are quite high. To maximize the convenience for most experimental biologists, the platform's web-server has been provided at http://lin.uestc.edu.cn/server/iRNA-PseColl along with a step-by-step user guide that will allow users to easily achieve their desired results without the need to go through the mathematical details involved in this paper.

INTRODUCTION

Since the first modified RNA ribonucleic acid was found ~60 years ago,[1] ~150 known RNA modifications have been reported.[2] Emerging evidences suggest that RNA modifications are critical components of the gene regulatory landscape and are involved in a variety of biological processes in the post-transcriptional level, such as protein translation and localization,[3] mRNA splicing,[4] affecting ribosome biogenesis,[5] mediating antibiotic resistance,[6] and stem cell pluripotency.[7] However, many aspects of RNA modifications remain unknown.[8] Therefore, detecting the positions of RNA modifications plays an essential role for understanding their molecular mechanisms and functions.

The advent of next-generation sequencing technologies has allowed investigation of RNA modifications on a genome-wide scale.[9–15] For example, the N^1-methyladenosine (m^1A),[9, 10] N^6-methyladenosine (m^6A),[13] and 5-methylcytosine (m^5C)[15] maps are available for the human transcriptome. Although these experimental methods played active roles in promoting the research progress on understanding the biological functions and the identification of RNA modifications, they are still labor-intensive. As excellent complements to experimental techniques, some computational methods (based on the high-resolution experimental data) have been developed to identify RNA modifications.[7, 16–21]

Reminiscent of the regulation of gene expression by histone modifications, it is also possible to mediate biological functions in a collective way by combining different kinds of RNA modifications.[8] Unfortunately, to the best of our knowledge, no computational tool is available for dealing with a system that simultaneously contains several different kinds of RNA modifications. Actually, this kind of multi-modification systems may contain much more interesting things worthy of exploration.

In view of this, the present study was initiated in an attempt to fill such a void by establishing a seamless package or platform that can be used to analyze a biological system that simultaneously contains the three well known types of RNA modifications: m^1A, m^6A, and m^5C (Figure 1).

Received 12 February 2017; accepted 17 March 2017;
http://dx.doi.org/10.1016/j.omtn.2017.03.006.

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
Correspondence: Kuo-Chen Chou, Gordon Life Science Institute, Boston, MA 02478, USA.
E-mail: kcchou@gordonlifescience.org

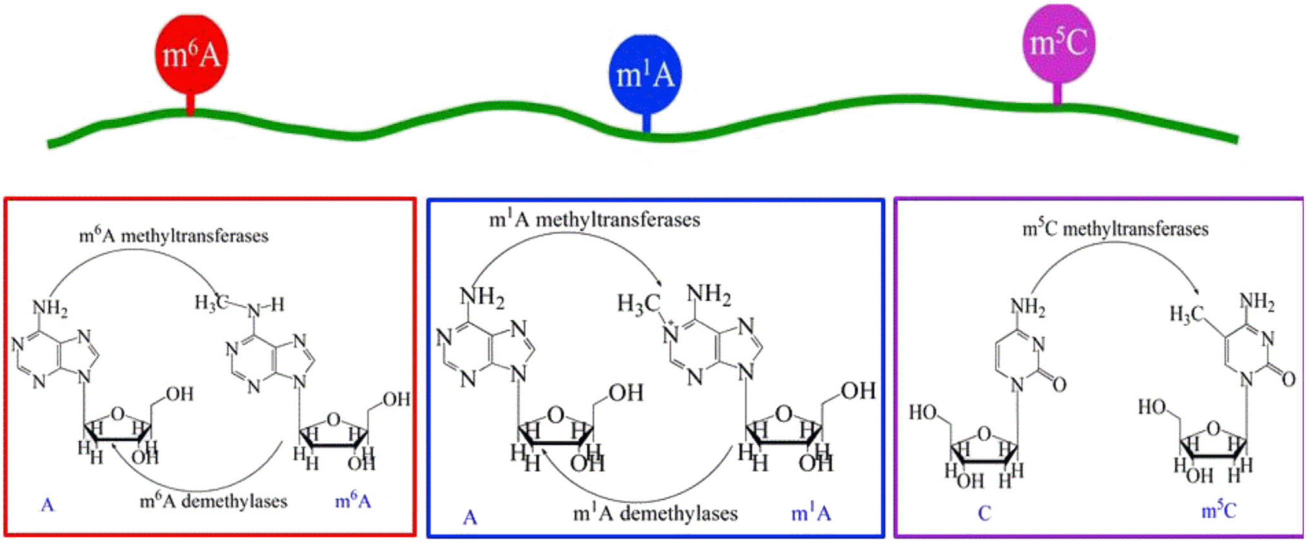

Figure 1. A Schematic Drawing to Show the Three Types of Modifications that May Simultaneously Occur in an RNA Sequence
Three types of modifications (m^1A, m^6A, and m^5C) are shown.

RESULTS AND DISCUSSION

By incorporating collective effects of nucleotides into PseKNC,[22, 23] a seamless platform called "iRNA-PseColl" has been developed for identifying the occurrence sites of different RNA modifications.

It has been observed by the most rigorous cross-validation, the jackknife test,[24] that the success rates achieved by the new predictor are quite high for the three different types of RNA modification sites, respectively (Table 1).

Because it is the first platform predictor ever developed for simultaneously identifying three different types of RNA modification sites based on its sequence information alone, it is not possible to demonstrate its power by a comparison with its counterparts because there is no such a counterpart yet for exactly the same purpose. Nevertheless, as we can see from Table 1, all the scores are quite high, particularly for the overall accuracy (Acc) and Mathew's correlation coefficient (MCC).

Let us use graphic analysis to further demonstrate the proposed platform's quality. As it is, the graphical approach is a useful vehicle for studying complicated biological systems because it can provide intuitive insights, as demonstrated by a series of previous studies.[25–34] Therefore, it would be instructive and illuminative to give an intuitive illustration for the current study as well. To realize this, the graph of receiver operating characteristic (ROC)[35, 36] was adopted as shown in Figure 2, where the ROC curves for the current method in identifying m^1A, m^6A, and m^5C modifications were given, respectively. The best possible prediction method would yield a point with the coordinate (0, 1) representing 100% sensitivity and 0 false-positive rate or 100% specificity. Therefore, the (0, 1) point is also called a perfect classification. A completely random guess would give a point along a diagonal from the point (0, 0) to (1, 1). The area under the ROC curve, also called AUROC, is used to indicate the performance quality of the classifier: the value 0.5 of AUROC is equivalent to random prediction while 1 of AUROC represents a perfect one. The AUROC for the case of m^1A, m^6A, or m^5C is 0.998, 0.849, or 0.911, respectively, indicating that the proposed platform is quite promising, holding very high potential to become a useful high throughput tool for genome analyses.

Inspired by a series of recent publications,[20, 21, 37–53] papers published with a publicly accessible web server will significantly enhance their impacts; this is particularly true for those papers aimed at developing novel prediction methods.[54] Accordingly, the web server for the current platform has been established. Moreover, for the convenience of the scientific community, a user guide is given in the Supplemental Materials and Methods.

MATERIALS AND METHODS

According to the Chou's[55] five-step guidelines that have been followed by many investigators in a series of recent publications,[21, 39, 41, 43–52, 56–63] to develop a new prediction method that not only can be easily used by most experimental scientists but also can inspire theoretical scientists to develop many other relevant prediction methods, we should make the following five procedures very clear: (1) how to construct or select a valid benchmark dataset to train and test the prediction model, (2) how to represent a biological sequence sample with a mathematical formulation or vector that is really correlated with the target concerned, (3) how to introduce or develop a powerful engine (or algorithm) to run the prediction model, (4) how to properly perform the cross-validation tests to objectively evaluate the anticipated accuracy, and (5) how to design a user-friendly web server to make it easy for people to get their desired

Table 1. The Success Rates Obtained by the Proposed Model in Identifying Three Different Types of RNA Modification Sites

	Metrics[a]			
Modification Type	Sn (%)	Sp (%)	Acc (%)	MCC
(1) m^1A	98.38	99.89	99.13	0.98
(2) m^6A	81.86	99.11	90.38	0.82
(3) m^5C	75.83	79.17	77.50	0.55

The results were obtained by the jackknife tests on the three benchmark datasets given in Supplemental Materials and Methods, respectively. Acc, overall accuracy; MCC, Mathew's correlation coefficient; Sn, sensitivity; Sp, specificity.

[a] See Equation 13 and the relevant text for the definition of metrics.

results. Below, we elaborate the five procedures in establishing the new predictor.

Benchmark Dataset

Owing to the fast development of high-throughput experimental techniques, the experimentally confirmed m^1A, m^6A, and m^5C modification data is available for the human genome.[9, 10, 13, 15] By mapping the experimental data to the human genome, the sequence samples with statistical significance were obtained for the three kinds of RNA modification sites as well. For facilitating the formulation, let us use the following scheme to represent a potential RNA modification-site-containing sample

$$R_{\xi}(\circledast) = N_{-\xi}N_{-(\xi-1)}\cdots N_{-2}N_{-1}\circledast N_{+1}N_{+2}\cdots N_{+(\xi-1)}N_{+\xi}, \quad \text{(Equation 1)}$$

where the symbol $\circledast$ denotes the single nucleic acid code A (adenine) or C (cytosine), the subscript ξ is an integer, $N_{-\xi}$ represents the ξ -th upstream nucleotide from the center, the $N_{+\xi}$ represents the ξ -th downstream nucleotide, and so forth. The $(2\xi+1)$ -tuple RNA sample, $R_{\xi}(\circledast)$, can be further classified into the following two categories:

$$\mathbf{R}_{\xi}(\circledast)\in\begin{cases}\mathbf{R}_{\xi}^{+}(\circledast), & \text{if its center can be of } 2'\text{-O-methylation}\\ \mathbf{R}_{\xi}^{-}(\circledast), & \text{otherwise}\end{cases}, \quad \text{(Equation 2)}$$

where $R_{\xi}^{+}(\circledast)$ denotes a true modification segment with A or C at its center, $R_{\xi}^{-}(\circledast)$ denotes a false modification segment with A or C at its center, and the symbol $\in$ means "a member of" in the set theory.

In literature, the benchmark dataset usually consists of a training dataset and a testing dataset: the former is for the use of training a model, while the latter for testing the model. However, as elucidated in a comprehensive review,[64] there is no need to artificially separate a benchmark dataset into the aforementioned two parts if the prediction model is examined by the jackknife test or subsampling (K-fold) cross-validation, because the outcome thus obtained is actually from a combination of many different independent dataset tests.

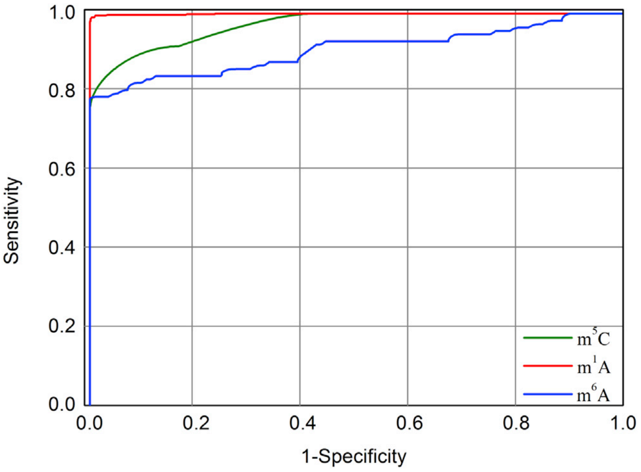

Figure 2. A Graphical Illustration to Show the Performances of iRNA-PseColl in Identifying m^1A, m^6A, and m^5C Modification Sites, Respectively

The performances are illustrated by means of the ROC curves.[35, 36] The area under the ROC curve is called AUROC. The greater the AUROC value is, the better the performance will be. See the text for further explanation.

Thus, the benchmark datasets for the current study can be further formulated as

$$\begin{cases}\mathbb{S}_{\xi}(m^1A) = \mathbb{S}_{\xi}^{+}(m^1A) \cup \mathbb{S}_{\xi}^{-}(m^1A), & \text{when } \circledast = A\\ \mathbb{S}_{\xi}(m^6A) = \mathbb{S}_{\xi}^{+}(m^6A) \cup \mathbb{S}_{\xi}^{-}(m^6A), & \text{when } \circledast = A\\ \mathbb{S}_{\xi}(m^5C) = \mathbb{S}_{\xi}^{+}(m^5C) \cup \mathbb{S}_{\xi}^{-}(m^5C), & \text{when } \circledast = C\end{cases}, \quad \text{(Equation 3)}$$

where the positive subset $\mathbb{S}_{\xi}^{+}(m^1A)$ only contains those RNA samples that can have m^1A modification, and the negative subset $\mathbb{S}_{\xi}^{-}(m^1A)$ only contains those RNA samples that cannot have m^1A modification, while U denotes the symbol of "union" in the set theory,[64] and so forth.

The benchmark datasets were derived from the RNA sequences in human genome that have the experimentally confirmed m^1A, m^6A, and m^5C modification sites.[9, 10, 13, 15] The detailed procedures to construct the benchmark dataset are as follows. First, as done in Chou,[65] by sliding the $(2\xi+1)$ -tuple nucleotide window (Figure 3) along each of the aforementioned RNA sequences, only those RNA segments with $\circledast =$ A or C at the center were collected. Second, if the upstream or downstream in a RNA sequence was less than ξ or greater than $L-\xi$ where L is the length of the RNA sequence concerned, the lacking code was filled with the same code of its nearest neighbor. Third, the RNA segment samples thus obtained were put into the positive subset $\mathbb{S}_{\xi}^{+}(m^1A)$, $\mathbb{S}_{\xi}^{+}(m^6A)$, or $\mathbb{S}_{\xi}^{+}(m^5C)$ if their centers were experimentally annotated as the m^1A, m^6A, or m^5C sites; otherwise, into the corresponding negative subset $\mathbb{S}_{\xi}^{-}(m^1A)$, $\mathbb{S}_{\xi}^{-}(m^6A)$, or $\mathbb{S}_{\xi}^{-}(m^5C)$. Fourth, to reduce redundancy and bias, none of the included RNA segments had pairwise sequence identity with any other in a same subset. By strictly following the above procedures, we obtained an array of benchmark datasets with

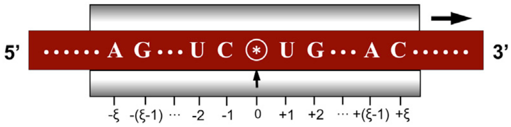

Figure 3. An Illustration to Show the Process of Collecting the RNA Samples by Sliding the (2ξ + 1) -nt Scaled Window along an RNA Sequence
Adapted from Chou[65] with permission. See the text for further explanation.

different ξ values and hence different lengths of RNA samples $(2\xi+1)$ as well (see Equation 1), as illustrated below

$$\mathbb{S}_{\xi}(\circledast) \subset \begin{cases} 23 \text{ nucleotides, when } \xi = 11 \\ 25 \text{ nucleotides, when } \xi = 12 \\ 27 \text{ nucleotides, when } \xi = 13 \\ \qquad \vdots \\ 39 \text{ nucleotides, when } \xi = 19 \\ 41 \text{ nucleotides, when } \xi = 20 \\ 43 \text{ nucleotides, when } \xi = 21 \end{cases}, \qquad \text{(Equation 4)}$$

where the symbol ⊂ means "formed by." It was observed via preliminary tests as well as many reports[19, 43, 66] that when $\xi = 20$ (i.e., the RNA samples formed by 41 nucleotides [nt]), the corresponding results were most promising. Accordingly, hereafter we only consider the 41-nt RNA sequences. By doing so, we obtained 6,366, 1,130 and 120 sequence samples for the positive subsets $\mathbb{S}_{\xi}^{+}(\mathrm{m}^1\mathrm{A})$, $\mathbb{S}_{\xi}^{+}(\mathrm{m}^6\mathrm{A})$, and $\mathbb{S}_{\xi}^{+}(\mathrm{m}^5\mathrm{C})$, respectively. The numbers of samples thus obtained in the corresponding negative subsets are much greater, and hence the benchmark datasets would be very imbalanced. Using such highly skewed benchmark dataset to train predictors would lead to the outcome that many positive cases might be mispredicted as negative ones.[42, 44, 56, 67] To balance out the size between the positive subset and the negative subset, we randomly picked out 6,366, 1,130, and 120 from the corresponding negative samples to for the negative subsets $\mathbb{S}_{\xi}^{-}(\mathrm{m}^1\mathrm{A})$, $\mathbb{S}_{\xi}^{-}(\mathrm{m}^6\mathrm{A})$, and $\mathbb{S}_{\xi}^{-}(\mathrm{m}^5\mathrm{C})$, respectively, as done in Chen et al.[16] and Feng et al.[19]

Finally, the detailed RNA sequence samples thus obtained for the benchmark dataset $\mathbb{S}_{\xi=20}(\mathrm{m}^1\mathrm{A})$, $\mathbb{S}_{\xi=20}(\mathrm{m}^6\mathrm{A})$, and $\mathbb{S}_{\xi=20}(\mathrm{m}^5\mathrm{C})$ are given in the Supplemental Materials and Methods, which can also be directly downloaded from http://lin.uestc.edu.cn/server/iRNA-PseColl/dataset.htm.

Formulating RNA Sequence Samples

One of the most challenging problems in computational biology today is how to formulate a biological sequence with a vector that can reflect its key pattern important for the function or mechanism concerned. The importance of such a challenge is due to the fact that nearly all the existing machine-learning algorithms were developed to handle vector rather than sequence samples, as elucidated in a review article.[54] Unfortunately, a vector defined in a discrete model may lose many important sequence pattern features. To deal with such a problem for protein/peptide sequences, the pseudo amino acid composition (PseAAC)[68–72] was developed. Ever since it was introduced, the concept of PseAAC has penetrated into nearly all the areas of computational proteomics (see a long list of references cited in two review papers[55, 73]). Inspired by the concept of PseAAC and encouraged by its great successes, the pseudo nucleotide composition (PseKNC)[22, 74–76] was proposed and has been increasingly used in various fields of genome analysis.[20, 21, 23, 37, 39, 40, 42, 43, 51–53, 58–60, 77–85] With both PseAAC and PseKNC being increasingly and widely used, it is highly desired to design a seamless package that can generate various modes of PseAAC and PseKNC according to users' needs for protein/peptide and DNA/RNA sequences, respectively. This was exactly the driving force of establishing the web server called Pse-in-One[86] and what it is about.

The general form of PseKNC for an RNA sequence sample is given by[23]

$$\mathrm{R} = [\phi_1 \quad \phi_2 \quad \cdots \quad \phi_u \quad \cdots \quad \phi_\Gamma]^{\mathbf{T}}, \qquad \text{(Equation 5)}$$

where T is a transpose operator, while the subscript Γ an integer and its value as well as the components ϕ_u $(u = 1, 2, \cdots, \Gamma)$ will depend on how to extract the desired features from the RNA sequence sample. In order to make Equation 4 able to reflect both the local feature of its individual constituent nucleotides and that of their collective effect, let us define the components in Equation 4 from the following two different approaches.

Local Features of Individual Nucleotides

RNA consists of four types of nucleotides: A (adenosine), C (cytidine), G (guanosine), and U (uridine). They can be classified into three different categories (Table 1): (1) from the angle of ring number, A and G have two rings, whereas C and U only one; (2) from the chemical functionality, A and C belong to amino group, while G and U to keto group; and (3) from the angle of hydrogen bonding, C and G can be bonded to each other with three hydrogen bonds, but A and U with only two (Figure 4). All these properties would have different impacts to RNA's low-frequency internal motion[87, 88] and its biological function[89–91].

To reflect the aforementioned features, let us denote the i-th nucleotide of Equation 1 by[92, 93]

$$\mathrm{N}_i = (x_i, \ y_i, \ z_i), \qquad \text{(Equation 6)}$$

where x_i, y_i, and z_i refer to the attributes of (1) ring structure, (2) functional group, and (3) hydrogen bonding in Table 2, respectively. Accordingly, the nucleotide A can be formulated as (1, 1, 1), C as (0, 1, 0), G as (1, 0, 0), and U as (0, 0, 1); or generally we have

$$\begin{gathered} x_i = \begin{cases} 1, \text{ if } \mathrm{N}_i \in \{\mathrm{A}, \ \mathrm{G}\} \\ 0, \text{ if } \mathrm{N}_i \in \{\mathrm{C}, \ \mathrm{U}\} \end{cases}; \quad y_i = \begin{cases} 1, \text{ if } \mathrm{N}_i \in \{\mathrm{A}, \ \mathrm{C}\} \\ 0, \text{ if } \mathrm{N}_i \in \{\mathrm{G}, \ \mathrm{U}\} \end{cases}; \\ z_i = \begin{cases} 1, \text{ if } \mathrm{N}_i \in \{\mathrm{A}, \ \mathrm{U}\} \\ 0, \text{ if } \mathrm{N}_i \in \{\mathrm{C}, \ \mathrm{G}\} \end{cases}. \end{gathered} \qquad \text{(Equation 7)}$$

Collective Features of the Constituent Nucleotides

There are some methods to reflect the coupling of a biological sequence or the collective effect of its constituent elements, such as

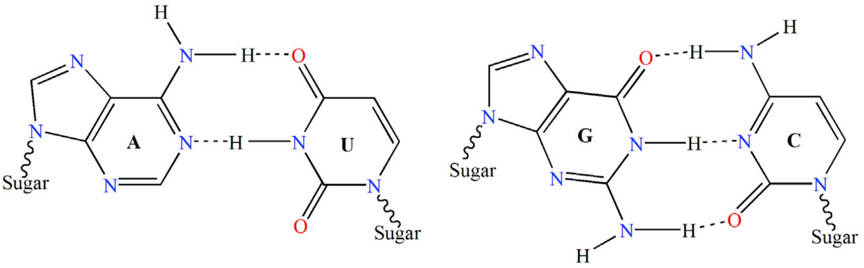

Figure 4. Illustration to Show the Structure of Paired Nucleic Acid Residues

Left: A-U pair bonded to each other with two hydrogen bonds. Right: G-C pair with three hydrogen bonds. Adapted from Chou[87] with permission.

the conditional probability approach,[94] degenerate Kmer strategy,[40] and g-gap dipeptide mode.[41] In this study, we would like to use a different approach; i.e., consider the occurrence frequency of a nucleotide not only for its local site but also for its distribution along the sequence of an RNA sample, as defined by the following equation

$$D_i = \frac{1}{\|L_i\|}\sum_{j=1}^{\ell} f(\mathrm{N}_j), \qquad \text{(Equation 8)}$$

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

$$f(N_j) = \begin{cases} 1, & \text{if } N_j = \text{the nucleotide concerned} \\ 0, & \text{otherwise} \end{cases}. \qquad \text{(Equation 9)}$$

For instance, suppose a RNA sequence is “CACGUC.” The density of “A” at the sequence position 1, 2, 3, 4, 5, or 6 is $0 = 0/1$, $0.5 = 1/2$, $0.33 \approx 1/3$, $0.25 = 1/4$, $0.20 = 1/5$, or $0.16 = 1/6$, respectively; that of “C” is $1 = 1/1$, $0 = 0/2$, $0.66 \approx 2/3$, $0.5 = 2/4$, $0.4 = 2/5$ or $0.5 = 3/6$, respectively; and so forth.

By combing Equations 6 and 9, the i-th nucleotide of Equation 1 can be uniquely defined by a set of four variables; i.e.,

$$\mathrm{N}_i = (x_i,\ y_i,\ z_i,\ D_i). \qquad \text{(Equation 10)}$$

For example, the RNA sequence “CACGUC” can be expressed by the following five sets of digital numbers: (0, 1, 0, 1), (1, 1, 1, 0.5), (0, 1, 0, 0.66), (1, 0, 0, 0.25), (0, 0, 1, 0.2), and (0, 1, 0, 0.5). Submitting these numbers into Equation 5, we have

$$\begin{aligned} &\mathrm{R(CACGUC)} \\ &\quad = [0\ 1\ 0\ 1\ 1\ 1\ 1\ 0.5\ 0\ 1\ 0\ 0.66\ 1\ 0\ 0\ 0.25\ 0\ 0\ 1\ 0.2\ 0\ 1\ 0\ 0.5]^{\mathrm{T}}, \end{aligned} \qquad \text{(Equation 11)}$$

meaning that the 6-nt nucleotide example can be defined by a $6 \times 4 = 24$ -D (dimensional) PseKNC vector.

Accordingly, all the samples in the current benchmark datasets (Supplemental Materials and Methods) can be formulated with a $41 \times 4 = 164$ -D vector.

Operation Engine

The prediction was operated by SVM (support vector machine), which has been widely used in various areas of bioinformatics and computational biology.[20, 40, 42, 59, 67, 77–81, 95–103] Its basic idea has been elaborated in the aforementioned the papers, and there is no need to repeat it here.

In the current study, the LibSVM package 3.18 was used to implement SVM, which can be downloaded for free from http://www.csie.ntu.edu.tw/∼cjlin/libsvm/. The SVM algorithm contains two uncertain quantities: one is the regularization parameter C and the other is the kernel width parameter γ. They were optimized via an optimization procedure using the grid search approach as described by

$$\begin{cases} 2^{-5} \le C \le 2^{15} & \text{with step } \Delta C = 2 \\ 2^{-15} \le \gamma \le 2^{-5} & \text{with step } \Delta\gamma = 2^{-1} \end{cases}, \qquad \text{(Equation 12)}$$

where ΔC and $\Delta\gamma$ represent the step gaps for C and γ, respectively.

For those readers who are interested in knowing more about SVM, see Chou and Cai[104] and Cai et al.[105] or a monograph[106] where a brief introduction or detailed description were given, respectively.

The platform predictor obtained via the aforementioned procedures is called “iRNA-PseColl,” where “i” stands for “identify,” “Pse” for “pseudo component approach,” and “Coll” for “collective effects of nucleotides.”

Quality Control or Examination

Quality control is a very important process in industries; it is even more important for a predictor. To deal with this problem, we need to address the following two issues: (1) what standard or metrics should we adopt to measure the predictor’s quality, and (2) what test process or method we should take to calculate the metrics. Below, we address the two problems.

A Set of Four Intuitive Metrics

The current prediction is belonging to the category called “binary classification” widely existing in genome analyses. To measure the prediction quality of this kind, a set of four metrics are usually used in literature[107]: (1) sensitivity or Sn, (2) specificity or Sp, (3) overall accuracy or Acc, and (4) Mathew’s correlation coefficient or MCC. Unfortunately, their formulations were directly taken from mathematical literature and difficult to be understood by most biological scientists. Fortunately, using the symbols introduced by Chou[108] in

Table 2. Classification of Nucleotides

Angle of View	Attribute	Nucleotides
(1) Ring structure	purine	A, G
	pyrimidine	C, U
(2) Functional group	amino	A, C
	keto	G, U
(3) Hydrogen bonding	stronger	C, G
	weaker	A, U

See Local Features of Individual Nucleotides for further explanation.

studying signal peptides, Xu et al.[109] and Chen et al.[110] have derived a new set of metrics that is equivalent to the old one but much more intuitive and easier to be understood by most biologists, as given below

To address this, we need to consider two issues: one is what metrics should be used to reflect the predictor's success rates; the other is what test method should be adopted to derive the metrics rates.

To quantitatively evaluate the quality of a binary classification predictor, four metrics are generally needed.[107] They are: (1) Acc for the predictor's overall accuracy; (2) MCC for its stability; (3) Sn for its sensitivity; and (4) Sp for its specificity. Unfortunately, the conventional formulations for the four metrics are not quite intuitive, and most biologists have difficulty understanding them, particularly the stability of MCC. Fortunately, as elaborated in Yu et al.[109] and Chen et al.,[110] by using the Chou's[111] symbols and derivation in studying signal peptides, the conventional metrics can be converted into a set of four intuitive equations, as formulated below:

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases}, \tag{Equation 13}$$

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

With the metrics of Equation 13, the meanings of Sn, Sp, Acc, and MCC have become crystal clear as discussed and used in a series of follow-up studies for many different areas.[20, 21, 38, 40, 42, 44–49, 56, 57, 61, 67, 80, 82, 84, 97, 99, 112–115] It is instructive to point out that more multi-label sequence samples have been emerging in system biology and medicine.[49, 116–119] To deal with this kind of multi-label system, a much more sophisticated set of metrics is needed as elaborated in Chou.[120]

Jackknife Validation

Three different cross-validation methods are often adopted in literature. These methods include[24]: (1) an independent dataset test, (2) a subsampling (or K-fold cross-validation) test, and (3) the jackknife test. However, as elucidated in Chou[55] in the above three choices, the jackknife test has been demonstrated to be the least arbitrary that can always yield a unique outcome for a given benchmark dataset. Therefore, the jackknife test has been widely recognized and increasingly adopted by researchers to analyze the quality of various predictors.[83, 103, 121–131] In view of this, we also used the jackknife test to examine the quality of the current prediction method. The jackknife test can exclude the "memory" effect because both the training dataset and testing dataset in a jackknife system are actually open, and each sample will be, in turn, moved between the two. The arbitrariness problem intrinsic to the independent dataset and subsampling tests[55] no longer exists, because the outcome derived via the jackknife test for a predictor is always the same on a given benchmark dataset.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Materials and Methods and can be found with this article online at http://dx.doi.org/10.1016/j.omtn.2017.03.006.

AUTHOR CONTRIBUTIONS

W.C., H.L., and K.-C.C. conceived and designed the study. P.F. and H.D. conducted the experiments. P.F., H.D., and W.C. implemented the algorithms. H.Y. established the web server. W.C., H.L., and K.-C.C. performed the analysis and wrote the paper. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors wish to thank the two anonymous reviewers for their constructive comments, which were very helpful for strengthening the presentation of this paper. This work was supported by the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (BJ2014028), the Outstanding Youth Foundation of North China University of Science and Technology (JP201502), the China Postdoctoral Science Foundation (2015M582533), and the Fundamental Research Funds for the Central Universities, China (ZYGX2015J144 and ZYGX2015Z006).

REFERENCES

1. Davis, F.F., and Allen, F.W. (1957). Ribonucleic acids from yeast which contain a fifth nucleotide. J. Biol. Chem. *227*, 907–915.

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. Meyer, K.D., and Jaffrey, S.R. (2014). The dynamic epitranscriptome: N6-methyladenosine and gene expression control. Nat. Rev. Mol. Cell Biol. *15*, 313–326.

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

5. Peifer, C., Sharma, S., Watzinger, P., Lamberth, S., Kötter, P., and Entian, K.D. (2013). Yeast Rrp8p, a novel methyltransferase responsible for m1A 645 base modification of 25S rRNA. Nucleic Acids Res. *41*, 1151–1163.

6. Ballesta, J.P., and Cundliffe, E. (1991). Site-specific methylation of 16S rRNA caused by pct, a pactamycin resistance determinant from the producing organism, Streptomyces pactum. J. Bacteriol. *173*, 7213–7218.

7. Chen, T., Hao, Y.J., Zhang, Y., Li, M.M., Wang, M., Han, W., Wu, Y., Lv, Y., Hao, J., Wang, L., et al. (2015). m(6)A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. Cell Stem Cell *16*, 289–301.

8. Hoernes, T.P., Hüttenhofer, A., and Erlacher, M.D. (2016). mRNA modifications: dynamic regulators of gene expression? RNA Biol. *13*, 760–765.

9. Dominissini, D., Nachtergaele, S., Moshitch-Moshkovitz, S., Peer, E., Kol, N., Ben-Haim, M.S., Dai, Q., Di Segni, A., Salmon-Divon, M., Clark, W.C., et al. (2016). The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. Nature *530*, 441–446.

10. 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.

11. Cai, L., Yuan, W., Zhang, Z., He, L., and Chou, K.C. (2016). In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci. Rep. *6*, 36540.

12. Khoddami, V., and Cairns, B.R. (2014). Transcriptome-wide target profiling of RNA cytosine methyltransferases using the mechanism-based enrichment procedure Aza-IP. Nat. Protoc. *9*, 337–361.

13. Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature *485*, 201–206.

14. Schwartz, S., Agarwala, S.D., Mumbach, M.R., Jovanovic, M., Mertins, P., Shishkin, A., Tabach, Y., Mikkelsen, T.S., Satija, R., Ruvkun, G., et al. (2013). High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. Cell *155*, 1409–1421.

15. Squires, J.E., Patel, H.R., Nousch, M., Sibbritt, T., Humphreys, D.T., Parker, B.J., Suter, C.M., and Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. *40*, 5023–5033.

16. 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.

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

18. 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.

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

20. Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.C. (2015). iRNA-methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. *490*, 26–33.

21. Liu, Z., Xiao, X., Yu, D.J., Jia, J., Qiu, W.R., and Chou, K.C. (2016). pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. Anal. Biochem. *497*, 60–67.

22. 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.

23. 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.

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

25. Chou, K.C., Jiang, S.P., Liu, W.M., and Fee, C.H. (1979). Graph theory of enzyme kinetics: 1. Steady-state reaction system. Sci. Sin. *22*, 341–358.

26. Chou, K.C., and Forsén, S. (1980). Graphical rules for enzyme-catalysed rate laws. Biochem. J. *187*, 829–835.

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

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

29. Althaus, I.W., Gonzales, A.J., Chou, J.J., Romero, D.L., Deibel, M.R., Chou, K.C., Kezdy, F.J., Resnick, L., Busso, M.E., So, A.G., et al. (1993). The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J. Biol. Chem. *268*, 14875–14880.

30. Althaus, I.W., Chou, J.J., Gonzales, A.J., Deibel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Palmer, J.R., Thomas, R.C., Aristoff, P.A., et al. (1993). Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry *32*, 6548–6554.

31. Wu, Z.C., Xiao, X., and Chou, K.C. (2010). 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*, 29–34.

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

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

34. Zhou, G.P., Chen, D., Liao, S., and Huang, R.B. (2016). Recent progresses in studying helix-helix interactions in proteins by incorporating the Wenxiang diagram into the NMR spectroscopy. Curr. Top. Med. Chem. *16*, 581–590.

35. Fawcett, T. (2005). An introduction to ROC analysis. Pattern Recognit. Lett. *27*, 861–874.

36. Davis, J., and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240.

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

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

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

41. Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K.C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget *7*, 16895–16909.

42. Liu, Z., Xiao, X., Qiu, W.R., and Chou, K.C. (2015). iDNA-methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal. Biochem. *474*, 69–77.

43. 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.

44. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal. Biochem. *497*, 48–56.

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

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

47. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J. Theor. Biol. *394*, 223–230.

48. 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.

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

50. Meher, P.K., Sahu, T.K., Saini, V., and Rao, A.R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci. Rep. *7*, 42362.

51. 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.

52. 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.

53. Liu, B., Wu, H., Zhang, D., Wang, X., and Chou, K.C. (2017). Pse-Analysis: a python package for DNA/RNA and protein/ peptide sequence analysis based on pseudo components and kernel methods. Oncotarget *8*, 4208–4217.

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

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

56. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. Molecules *21*, E95.

57. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget *7*, 34558–34570.

58. Liu, B., Fang, L., Liu, F., Wang, X., and Chou, K.C. (2016). iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J. Biomol. Struct. Dyn. *34*, 223–235.

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

60. Liu, B., Long, R., and Chou, K.C. (2016). iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics *32*, 2411–2418.

61. Qiu, W., Sun, B.Q., Xiao, X., and Chou, K.C. (2016). iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. Mol. Inform. http://dx.doi.org/10.1002/minf.201600010.

62. Cheng, X., Zhao, S.G., Xiao, X., and Chou, K.C. (2016). iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics *33*, 341–346.

63. Khan, M., Hayat, M., Khan, S.A., and Iqbal, N. (2017). Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. J. Theor. Biol. *415*, 13–19.

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

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

66. Chen, W., Tang, H., and Lin, H. (2016). MethyRNA: a web server for identification of N6-methyladenosine sites. J. Biomol. Struct. Dyn. *35*, 683–687.

67. Xiao, X., Min, J.L., Lin, W.Z., Liu, Z., Cheng, X., and Chou, K.C. (2015). iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach. J. Biomol. Struct. Dyn. *33*, 2221–2233.

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

69. Chou, K.C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics *21*, 10–19.

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

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

72. Du, P., Gu, S., and Jiao, Y. (2014). 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*, 3495–3506.

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

74. 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.

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

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

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

78. 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.

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

80. 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.

81. 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.

82. Liu, B., Fang, L., Liu, F., Wang, X., Chen, J., and Chou, K.C. (2015). Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE *10*, e0121501.

83. 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.

84. Chen, W., Feng, P., Ding, H., Lin, H., and Chou, K.C. (2016). Using deformation energy to analyze nucleosome positioning in genomes. Genomics *107*, 69–75.

85. Tahir, M., and Hayat, M. (2016). iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC. Mol. Biosyst. *12*, 2587–2593.

86. Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. *43* (W1), W65–W71.

87. Chou, K.C. (1984). Low-frequency vibrations of DNA molecules. Biochem. J. *221*, 27–31.

88. Chou, K.C., Maggiora, G.M., and Mao, B. (1989). Quasi-continuum models of twist-like and accordion-like low-frequency motions in DNA. Biophys. J. *56*, 295–305.

89. Chou, K.C., Chen, N.Y., and Forsen, S. (1981). The biological functions of low-frequency phonons: 2. Cooperative effects. Chem. Scr. *18*, 126–132.

90. Chou, K.C., and Mao, B. (1988). Collective motion in DNA and its role in drug intercalation. Biopolymers *27*, 1795–1815.

91. Chou, K.C. (1988). Low-frequency collective motion in biomacromolecules and its biological functions. Biophys. Chem. *30*, 3–48.

92. Chou, K.C., and Zhang, C.T. (1992). Diagrammatization of codon usage in 339 human immunodeficiency virus proteins and its biological implication. AIDS Res. Hum. Retroviruses *8*, 1967–1976.

93. Zhang, C.T., and Chou, K.C. (1994). A graphic approach to analyzing codon usage in 1562 Escherichia coli protein coding sequences. J. Mol. Biol. *238*, 1–8.

94. Chou, K.C. (1995). A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. Protein Sci. *4*, 1365–1383.

95. 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.

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

97. Ding, H., Deng, E.Z., Yuan, L.F., Liu, L., Lin, H., Chen, W., and Chou, K.C. (2014). iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Res. Int. *2014*, 286419.

98. Fan, Y.N., Xiao, X., Min, J.L., and Chou, K.C. (2014). iNR-Drug: predicting the interaction of drugs with nuclear receptors in cellular networking. Int. J. Mol. Sci. *15*, 4915–4937.

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

100. Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., and Chou, K.C. (2014). 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*, e106691.

101. Qiu, W.R., Xiao, X., Lin, W.Z., and Chou, K.C. (2015). iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. J. Biomol. Struct. Dyn. *33*, 1731–1742.

102. Xu, R., Zhou, J., Liu, B., He, Y., Zou, Q., Wang, X., and Chou, K.C. (2015). 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*, 1720–1730.

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

104. Chou, K.C., and Cai, Y.D. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. *277*, 45765–45769.

105. Cai, Y.D., Zhou, G.P., and Chou, K.C. (2003). Support vector machines for predicting membrane protein types by using functional domain composition. Biophys. J. *84*, 3257–3263.

106. Cristianini, N., and Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, *Chapter 3* (Cambridge University Press).

107. 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.

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

109. 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.

110. 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.

111. Chou, K.C. (2001). Prediction of protein signal sequences and their cleavage sites. Proteins *42*, 136–139.

112. 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.

113. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. J. Biomol. Struct. Dyn. *34*, 1946–1961.

114. Xu, Y., and Chou, K.C. (2016). Recent progress in predicting posttranslational modification sites in proteins. Curr. Top. Med. Chem. *16*, 591–603.

115. Xu, Y., Wen, X., Wen, L.S., Wu, L.Y., Deng, N.Y., and Chou, K.C. (2014). iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE *9*, e105018.

116. Xiao, X., Wu, Z.C., and Chou, K.C. (2011). iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J. Theor. Biol. *284*, 42–51.

117. Chou, K.C., Wu, Z.C., and Xiao, X. (2012). iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. *8*, 629–641.

118. Xiao, X., Wang, P., Lin, W.Z., Jia, J.H., and Chou, K.C. (2013). iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal. Biochem. *436*, 168–177.

119. Lin, W.Z., Fang, J.A., Xiao, X., and Chou, K.C. (2013). iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol. Biosyst. *9*, 634–644.

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

121. Zhou, G.P., and Assa-Munt, N. (2001). Some insights into protein structural class prediction. Proteins *44*, 57–59.

122. Zhou, G.P., and Doctor, K. (2003). Subcellular location prediction of apoptosis proteins. Proteins *50*, 44–48.

123. Chou, K.C., and Cai, Y.D. (2005). Prediction of membrane protein types by incorporating amphipathic effects. J. Chem. Inf. Model. *45*, 407–413.

124. 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.

125. Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., and Sattar, A. (2015). Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J. Theor. Biol. *364*, 284–294.

126. Khan, Z.U., Hayat, M., and Khan, M.A. (2015). Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J. Theor. Biol. *365*, 197–203.

127. Kumar, R., Srivastava, A., Kumari, B., and Kumar, M. (2015). Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. J. Theor. Biol. *365*, 96–103.

128. Ali, F., and Hayat, M. (2015). Classification of membrane protein types using voting feature interval in combination with Chou's pseudo amino acid composition. J. Theor. Biol. *384*, 78–83.

129. 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.

130. Ju, Z., Cao, J.Z., and Gu, H. (2016). Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. J. Theor. Biol. *397*, 145–150.

131. Behbahani, M., Mohabatkar, H., and Nosrati, M. (2016). 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*, 1–5.