Bioinformatics, 32(3), 2016, 362–369
doi: 10.1093/bioinformatics/btv604
Advance Access Publication Date: 17 October 2015
Original Paper

OXFORD

Sequence analysis

iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo *k*-tuple nucleotide composition

Bin Liu[1,2,3,*], Longyun Fang[1], Ren Long[1], Xun Lan[4,*] and Kuo-Chen Chou[3,5,*]

[1]School of Computer Science and Technology, [2]Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China, [3]Computational Biology, Gordon Life Science Institute, Belmont, MA 02478, USA, [4]Department of Genetics, Stanford University, Stanford, CA 94305, USA and [5]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on September 25, 2015; revised on October 9, 2015; accepted on October 12, 2015

Abstract

Motivation: Enhancers are of short regulatory DNA elements. They can be bound with proteins (activators) to activate transcription of a gene, and hence play a critical role in promoting gene transcription in eukaryotes. With the avalanche of DNA sequences generated in the post-genomic age, it is a challenging task to develop computational methods for timely identifying enhancers from extremely complicated DNA sequences. Although some efforts have been made in this regard, they were limited at only identifying whether a query DNA element being of an enhancer or not. According to the distinct levels of biological activities and regulatory effects on target genes, however, enhancers should be further classified into strong and weak ones in strength.

Results: In view of this, a two-layer predictor called '**iEnhancer-2L**' was proposed by formulating DNA elements with the 'pseudo *k*-tuple nucleotide composition', into which the six DNA local parameters were incorporated. To the best of our knowledge, it is the first computational predictor ever established for identifying not only enhancers, but also their strength. Rigorous cross-validation tests have indicated that **iEnhancer-2L** holds very high potential to become a useful tool for genome analysis.

Availability and implementation: For the convenience of most experimental scientists, a web server for the two-layer predictor was established at http://bioinformatics.hitsz.edu.cn/iEnhancer-2L/, by which users can easily get their desired results without the need to go through the mathematical details.

Contact: bliu@gordonlifescience.org, bliu@insun.hit.edu.cn, xlan@stanford.edu, kcchou@gordonlifescience.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Enhancers are distinct genomic regions that can upregulate transcription of target genes through interaction with transcription factors (TFs) (Shlyueva *et al.*, 2014). In contrast to gene proximal elements, such as promoters, enhancers are distal elements that can be 20 kb or further away from a gene, or even located on a different chromosome. This feature makes the prediction of enhancer a challenging task.

Experimental methods, such as chromatin immunoprecipitation followed by deep sequencing (Heintzman and Ren, 2009), can successfully detect many regulatory enhancers by targeting enhancer binding TF, such as P300 (Heintzman *et al.*, 2007; Visel *et al.*, 2009). But TF only binds to enhancers in a cell-specific manner. Not all of the functional enhancers are occupied by a certain TF. Besides, active enhancers are nucleosome-depleted open regions, which display deoxyribonuclease or DNase I hypersensitivity (Boyle *et al.*, 2011). But, many other DNA elements, such as promoters, are also accessible by DNase I. The histones in the flanking nucleosomes of enhancer often carry characteristic post-translational modifications. Accordingly, genome-wide mapping of histone modifications (Ernst *et al.*, 2011; Erwin *et al.*, 2014; Fernandez and Miranda-Saavedra, 2012; Firpi *et al.*, 2010; Kleftogiannis *et al.*, 2015; Rajagopal *et al.*, 2013) allows accurate determination and classification of enhancers and promoters. For instance, the open chromatin regions flanked by histone modifications, such as H3K4me3 and H3K27ac, have been classified as strong enhancers, while the regions flanked by H3K4me1 classified as weak enhancers (Fig. 1). These kinds of techniques have now been widely used and the results thus obtained are quite consistent with the enhancer activity assays.

Unfortunately, the experimental methods by the aforementioned authors are expensive and time consuming. Therefore, several computational methods were developed in an attempt to timely predict enhancers in genomes. These methods differ in operation algorithms and input features. For instance, Firpi *et al.* (2010) developed a predictor called CSI-ANN by using an Artificial Neural Network (ANN) approach combined with efficient data transformation and feature extraction. Fernandez and Miranda-Saavedra (2012) proposed a support vector machine (SVM) classifier called ChromeGenSVM by employing a genetic algorithm to optimize specific histone epigenetic marks in feature selection process. Using the random forest algorithm, Rajagopal *et al.* (2013) proposed a predictor called RFECS to identify enhancers. By combining a multiple kernel learning method and evolutionary conservation, Erwin *et al.* (2014) developed a different enhancer predictor called EnhancerFinder. Considering the inherent limitation of the conventional *k*-mers approach due to the high-dimension and over-fitting problems, Ghandi *et al.* (2014) utilized the gapped *k*-mers to develop a classifier named GKM–SVM for identifying enhancers. Very recently, using an ensemble prediction framework, Kleftogiannis *et al.* (2015) developed a novel predictor called DEEP and they claimed that it performed better than the four state-of-the-art methods in identifying enhancers.

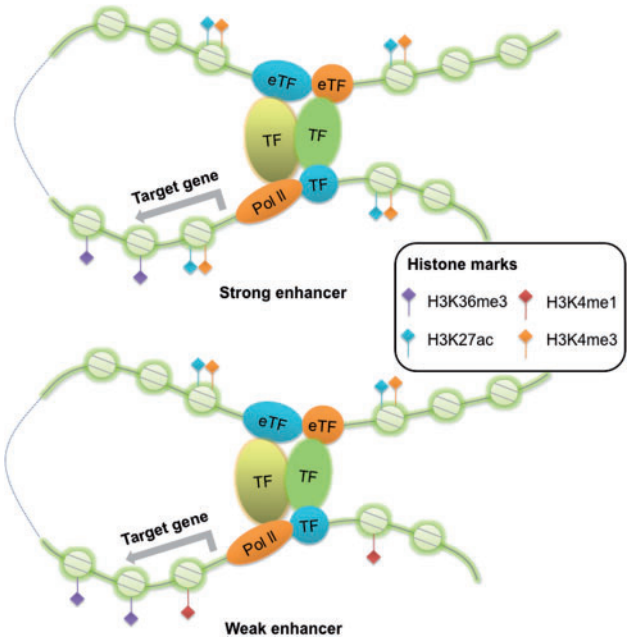

Fig. 1. A schematic drawing to show the strong and weak enhancers. The open chromatin regions flanked by histone modifications (such as H3K4me3 and H3K27ac) are of strong enhancer (top), while those flanked by H3K4me1 are of weak enhancer (bottom)

Although the aforementioned computational methods can yield quite encouraging results and each of them has its own advantage, further work is needed due to the following reasons. Most of those studies were focused on discriminating enhancers from other regulatory elements. But it was shown that enhancers are a large group of functional elements with many different subgroups, such as, strong enhancers, weak enhancers and poised or inactive enhancers (Shlyueva *et al.*, 2014), implying that enhancers of distinct subgroups will have different biological activities and regulatory effects on target genes. Therefore, in order to really understand the mechanism underlying the gene regulation through enhancers, it is indispensable to classify them according to their attributes to these subgroups. This study was initiated in an attempt to address this problem.

According to the Chou's 5-step rule (Chou, 2011) and performed in a series of recent publications (Chen *et al.*, 2014; Ding *et al.*, 2014; Jia *et al.*, 2015a; Lin *et al.*, 2014; Liu *et al.*, 2014; Xu *et al.*, 2014a, 2015), to establish a really useful sequence-based statistical predictor for a biological system, we should consider the following five procedures: (1) construct or select a valid benchmark dataset to train and test the predictor; (2) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor and (5) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by-one.

2 Materials and methods

2.1 Benchmark dataset

The benchmark dataset was constructed based on the chromatin state information of nine cell lines, including H1ES, K562, GM12878, HepG2, HUVEC, HSMM, NHLF, NHEK and HMEC. The chromatin state information was annotated by ChromHMM (Ernst and Kellis, 2012; Ernst *et al.*, 2011) with the whole genome profile of multiple histone marks, such as H3K4me1, H3K4me3, H3K27ac, etc.

As pointed out by a comprehensive review (Chou and Shen, 2007a), there is no need to separate a benchmark dataset into a training dataset and a testing dataset for validating a prediction method if it 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. Thus, the benchmark dataset set $\mathbb{S}$ for this study can be formulated as

$$\begin{cases} \mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \\ \mathbb{S}^+ = \mathbb{S}^+_{\text{strong}} \cup \mathbb{S}^+_{\text{weak}} \end{cases}, \quad (1)$$

where $\mathbb{S}^-$ is the set that contains non-enhancers only, $\cup$ is the symbol for union in the set theory, $\mathbb{S}^+$ is the set containing enhancers only, $\mathbb{S}^+_{\text{strong}}$ is the subset containing the strong enhancers only and $\mathbb{S}^+_{\text{weak}}$ is the subset containing the weak enhancers only. The strong

enhancers in $\mathbb{S}^{+}_{strong}$ and the non-enhancers in $\mathbb{S}^{-}$ are clearly annotated in all the nine tissues mentioned above, but the weak enhancers in $\mathbb{S}^{+}_{weak}$ vary significantly with different tissues. To cope with this kind of situation, the weak enhancers were constructed based on the human embryonic stem cell.

All the samples obtained via the above procedure were divided into 200 bp fragments in order to be consistent with the length of nucleosome and linker DNA. All those samples with length <200 bp were removed. Thus, we obtained 742 strong enhancers, 370 517 weak enhancers, and 5 257 994 non-enhancers. To get rid of redundancy and avoid bias, the CD-HIT software (Li and Godzik, 2006) was used by setting the cutoff threshold at 80% to remove those DNA fragments with high sequence similarity (note that the most stringent cutoff threshold for DNA sequences allowed by CD-HIT was 75%). The numbers of weak enhancers and non-enhancers are much larger than that of strong enhancers. To get a benchmark dataset with balanced subsets in size, we randomly picked 742 samples from the weak enhancers and 1484 samples from the non-enhancers. The final benchmark dataset $\mathbb{S}$ thus obtained contains 2968 samples, of which 742 are strong enhancers belonging to $\mathbb{S}_{strong}$, 742 weak enhancers belonging to $\mathbb{S}^{+}_{weak}$, and 1484 non-enhancers belonging to $\mathbb{S}^{-}$ (cf. Equation 1). The sequences of the 2968 samples classified into three subsets are given in Supplementary Information S1.

2.2 Pseudo k-tuple nucleotide composition (PseKNC)

Suppose a DNA sample $\mathbf{D}$ with L nucleotides; i.e.

$$\mathbf{D} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3\cdots\mathrm{N}_i\cdots\mathrm{N}_L \tag{2}$$

where N_1 represents the first nucleotide at the sample sequence position 1, N_2 the second nucleotide at the position 2, and so forth. They can be any of the four nucleotides; i.e.

$$\mathrm{N}_i \in \{\mathrm{A\,(adenine)}\quad \mathrm{C\,(cytosine)}\quad \mathrm{G\,(guanine)}\quad \mathrm{T\,(thymine)}\} \tag{3}$$

If the DNA sequence sample is represented by the k-tuple nucleotide (or k-mers) composition (Ioshikhes *et al.*, 1996), the corresponding feature vector will contain 4^k components, as given by

$$\mathbf{D} = [f_1 \quad f_2 \quad f_3 \quad \cdots \quad f_i \quad \cdots \quad f_{4^k}]^{\mathrm{T}}, \tag{4}$$

where the symbol $\mathbf{T}$ is the transpose operator, f_i represents the normalized occurrence frequency of the ith k-mer. As we can see from Equation (4), with the incensement of k values, although longer-range information can be incorporated, the vector's dimension will increase rapidly. As we can see from Equation (4), when $k > 4$, the number of the vector components will rapidly increase, causing the so-called 'high-dimension disaster' (Wang *et al.*, 2008) or overfitting problem that will significantly reduce the deviation tolerance or cluster-tolerant capacity (Chou, 1999) so as to lower down the prediction success rate or stability. Therefore, the k-mers approach is useful only when the value of k is very small. In other words, it can only be used to incorporate the local or short-range sequence-order information, but certainly not the global or long-range sequence-order information. To approximately cover the long-range sequence-order effects, one popular and well-known method is to use the pseudo components that were originally introduced in dealing with protein/peptide sequences (Chou, 2001, 2005) and recently extended to deal with DNA/RNA sequences (Chen *et al.*, 2014a, b, c, 2015a, b; Liu *et al.*, 2015a, b, c).

According to the concept of pseudo components, the DNA sample of Equation (2) can be formulated by the vector called pseudo k-tuple nucleotide composition of PseKNC (Chen *et al.*, 2014a); i.e.

$$\mathbf{D} = [d_1 \quad d_2 \quad \cdots \quad d_{4^k} \quad d_{4^k+1} \quad \cdots \quad d_{4^k+\lambda}]^{\mathrm{T}} \tag{5}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 4^k) \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (4^k < u \le 4^k + \lambda) \end{cases} \tag{6}$$

In Equation (6), the definition of f_u ($u = 1, 2, \ldots, 4^k$) is the same as the components in Equation (4), meaning the normalized occurrence frequency of the uth non-overlapping k-mer in the DNA sequence, and

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

where θ_1 is called the first-tier correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotide along a DNA sequence (Fig. 2a), θ_2, the second-tier correlation factor between all the second most contiguous dinucleotide (Fig. 2b); θ_3, the third-tier correlation factor between all the third most contiguous dinucleotide (Fig. 2c); and so forth. $\mathrm{C}_{i,i+j}$ denotes the coupling factor of the ith dinucleotide with the $(i+j)$th one along a DNA sequence, and its detailed definition is given by

$$C_{i,i+j} = \frac{1}{\mu}\sum_{g=1}^{\mu}[P_g(D_i) - P_g(D_{i+j})]^2, \tag{8}$$

where μ is the number of DNA physicochemical properties considered that is equal to 6 in this study and will be further explained bellow.

Now, it is clear that the first 4^k components in Equation (5) are used to incorporate the short-range or local sequence-order information of the DNA sequence, while the remaining components used for its long-range or global sequence order information. Obviously, λ can also be viewed as the number of the total pseudo components used to reflect the long-range or global sequence effect (Liu *et al.*,

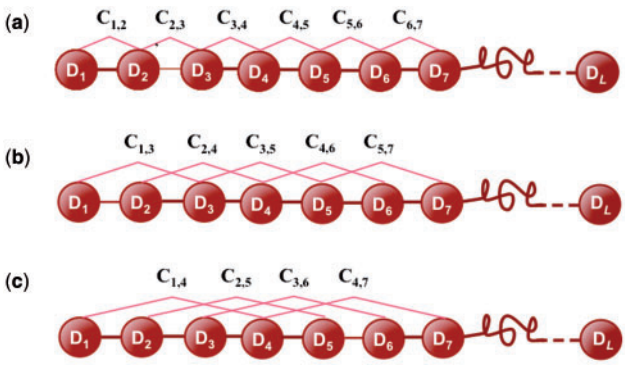

Fig. 2. A schematic drawing to illustrate the correlations of dinucleotides (or 2-mers) along a DNA sequence. The symbol $\mathrm{C}_{i,j}$ denotes the coupling of the ith dinucleotide with the jth one, and their value will be defined later. (**a**) The first-tier coupling reflects the sequence-order mode between all the most contiguous non-overlapping dinucleotide. (**b**) The second-tier coupling reflects the sequence-order mode between all the second-most contiguous non-overlapping dinucleotide. (**c**) The third-tier correlation reflects the sequence-order mode between all the third-most contiguous non-overlapping dinucleotide

2015d, e), and w of Equation (6) is the weight factor (Chou, 2001, 2005). The concrete values for λ and w will be further discussed later.

2.3 DNA local structural property parameters

Many reports indicate that DNA physicochemical properties play important roles in gene expression regulation (Brukner *et al.*, 1995; Fukue *et al.*, 2005; Gowers and Halford, 2003), and that they are also useful for conducting genome analysis (Miele *et al.*, 2008). Actually, being evolutionarily more constrained than the underlying actual sequence, they are also closely correlated with the functional non-coding elements such as enhancers (Parker *et al.*, 2009). Accordingly, it is reasonable to use the physicochemical properties of nucleotides to define the coupling factors of Equation (8).

According to Dickerson (1989), the spatial arrangements of two neighboring base pairs are usually characterized by six parameters, of which three are local translational parameters (shift, slide and rise) and other three are the local angular parameters (twist, tilt and roll).

The detailed values for the six DNA local parameters are given in Table S2-1 of Supplementary Information S2, which were used to calculate the coupling factors of Equation (8) to reflect global or long-range sequence pattern information for the DNA sequences concerned. Note that before substituting these parameters into Equation (8), all the original values $P_g(D_i)$ $(i = 1, 2, 3, \ldots, 6)$ were subjected to a standard conversion, as described by

$$P_g(\mathrm{D}_i) \Leftarrow \frac{P_g(\mathrm{D}_i) - P_g(\mathrm{D}_i)}{\mathrm{SD}\{\langle P_g(\mathrm{D}_i)\rangle\}}, \qquad (9)$$

where the symbol $\langle\ \rangle$ means taking the average of the quantity therein over the $4^2 = 16$ different dinucleotides, and SD means the corresponding standard deviation. For the detailed mathematical formulation of SD, see Equation (4) of Chou (2001) or Equation (4) of Chou (2005). The advantage to do so is that the converted values obtained by Equation (9) will have a zero mean value over the 16 different dinucleotides, and will remain unchanged if going through the same conversion procedure again (Chou and Shen, 2007b). Listed in Table S2-2 of Supplementary Information S2 are the corresponding values obtained via the standard conversion of Equation (9) from the data in Table S2-1 of Supplementary Information S2.

Thus, any DNA sequence can be uniquely defined by a feature vector of Equation (5), where the short-range or local sequence pattern can be reflected by the 4^k k-tuple nucleotides, while the long-range or global sequence correlation can be reflected by the λ pseudo components as given by Equations (6–8.).

2.4 Support vector machine

Being widely used in the realm of bioinformatics (see, e.g. Chen *et al.*, 2015c; Liu *et al.*, 2015e, f; Qiu *et al.*, 2015; Xiao *et al.*, 2015; Xu *et al.*, 2015), SVM is a machine-learning algorithm based on the statistical learning theory (Cortes and Vapnik, 1995). An SVM training algorithm builds a non-probabilistic binary linear classifier to assign new samples into one of the two categories. In an SVM model, the samples are mapped to points in space, where the separate categories are divided by a clear wide gap. New samples are then mapped into that same space and their categories are predicted according to which side of the gap they fall on. In addition to linear classification, SVMs can efficiently perform a non-linear classification by using the so-called kernel trick, implicitly mapping the inputs into high-dimensional feature spaces.

In this study, we used the functions provided in LIBSVM (Chang and Lin, 2009), a library for SVM classification and regression. The RBF kernel function contains two parameters C and γ, which will be optimized for the benchmark dataset via the grid search tool provided in LIBSVM (Chang and Lin, 2011), as will be further discussed later.

For a brief formulation of SVM and how it works, see Cai *et al.* (2003), Chou and Cai (2002); for more details about SVM, see a monograph (Cristianini and Shawe-Taylor, 2000).

2.5 Two-layer classification framework

To make the prediction method not only able to identify whether a DNA sample is an enhancer but also make it able to identify the strength type, here we are to develop a two-layer predictor similar to the treatment in identifying membrane proteins and their types (Chou and Shen, 2007a,b).

The first-layer predictor was trained and tested by the benchmark dataset in the first equation of Equation (1); while the second-layer predictor was trained and tested by its second equation.

Both the two predictors were operated with the SVM algorithm as described in the last section. But their parameters k, λ and w were optimized separately, as will be further discussed later.

The two-layer predictor thus obtained is called **Enhancer-2L**. To provide an intuitive view, a flowchart to show the process of how the 2-layer classifier works is given in Figure 3.

2.6 Performance metrics and cross validation

To more completely and objectively evaluate the quality of a predictor, one needs to consider two things. One is the metrics used to quantitatively measure its performance, and the other is the test method adopted during the cross validation.

The predictions performed by both the layers in **Enhancer-2L** are actually in dealing with a binary (two-lass) classification problem. Its first layer is to address whether a query DNA sample belonging to an enhancer or non-enhancer. If it is the former, the prediction will be automatically continued by its second layer to find out where the enhancer is of the strong type or weak one. For this kind of

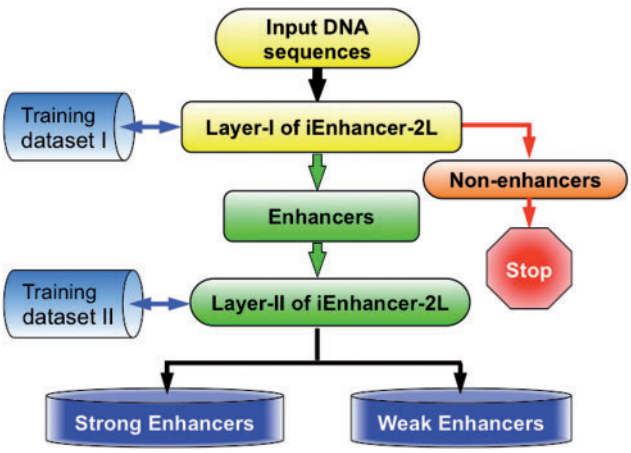

Fig. 3. A flowchart to show how **iEnhancer-2L** works. The input DNA sequences are first identified by its Layer-I sub-predictor as enhancers or non-enhancers. Subsequently, the predicted enhancers are further identified by Layer-II as strong or weak ones. Training datasets I and II mean $\mathbb{S}$ and $\mathbb{S}^+$ of Equation (1), used to train Layer-I and Layer-II sub-predictors, respectively

binary classification problem, the following set of metrics were often used to measure the prediction quality

$$\begin{cases} \mathrm{Sn} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Sp} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{Acc} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{MCC} = \dfrac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})\ (\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \end{cases} \tag{10}$$

where TP represents the true positive; TN, the true negative; FP, the false positive; FN, the false negative; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; MCC, the Mathew's correlation coefficient (Chen *et al.*, 2007). But the metrics formulated in Equation (10) is not easy-to-understand for most experimental scientists, and hence here we would prefer to use the following formulation as done by many investigators in a series of recent publications (see, e.g. Chen *et al.*, 2012, 2013; Guo *et al.*, 2014; Jia *et al.*, 2015b; Lin *et al.*, 2014; Qiu *et al.*, 2014, 2015; Xiao *et al.*, 2015; Xu *et al.*, 2013a, b, 2014a, b):

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_-^+ + N_+^-}{N^+ + N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^+}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \tag{11}$$

where N^+ is the total number of the positive samples or enhancers investigated while N_-^+ the number of enhancer samples incorrectly predicted to be of non-enhancer; N^- the total number of the negative samples or non-enhancers investigated while N_+^- the number of the non-enhancers incorrectly predicted to be of enhancer.

Based on Equation (11), the following are evident. When $N_-^+ = 0$ meaning none of the enhancers was incorrectly predicted belonging to non-enhancers, we have the sensitivity Sn = 1. When $N_-^+ = N^+$ meaning that all the enhancers were incorrectly predicted belonging to non-enhancers, we have the sensitivity Sn = 0. Similarly, when $N_+^- = 0$ meaning none of the non-enhancers was mispredicted, we have the specificity Sp = 1; whereas $N_+^- = N^-$ meaning that all the non-enhancers were incorrectly predicted as enhancers, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that none of enhancers in the positive dataset and none of the non-enhancers in the negative dataset were incorrectly predicted, we have the overall accuracy Acc = 1 and MCC = 1; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the enhancers in the positive dataset and all the non-enhancers in the negative dataset were incorrectly predicted, we have the overall accuracy Acc = 0 and MCC = −1; whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have Acc = 0.5 and MCC = 0 meaning no better than random guess. As we can see from the above discussion, it would make the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient much more intuitive and easier-to-understand by using the metrics formulated in Equation (11) instead of Equation (10), particularly for the meaning of MCC.

Note that, however, the set of metrics in Equation (11) is valid only for the single-label systems. For the multi-label systems whose emergence has become more frequent in system biology (Chou *et al.*, 2012; Lin *et al.*, 2013; Wang *et al.*, 2015; Xiao *et al.*, 2011) and system medicine (Xiao *et al.*, 2013), a completely different set of metrics is needed as elaborated in Chou (2013).

With the performance metrics well defined, now let us consider the test methods. In statistical prediction, the following three cross-validation methods are often used to calculate the metrics values for a predictor: independent dataset test, subsampling (or *K*-fold cross-validation) test and jackknife test (Chou and Zhang, 1995). Of the three methods, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset as elucidated in Chou (2011) and demonstrated by Equations (28)–(32) therein. Therefore, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g. Dehzangi *et al.*, 2015; Hajisharifi *et al.*, 2014; Khan *et al.*, 2015; Kumar *et al.*, 2015; Liu *et al.*, 2015a; Mondal and Pai, 2014; Zhou and Assa-Munt, 2001).

Accordingly, in this study we also use the jackknife test to evaluate the accuracy of the current predictor. During the jackknife test, each of the samples in the benchmark dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the sample being identified. Although the jackknife test may take more computational time, it is worthwhile because it will always yield a unique outcome for a given benchmark dataset.

3 Results and discussion

3.1 Parameter optimization

As we can see from Equations (5–7), the present model depends on the three parameters k, w and λ, where k reflects the local or short-range sequence-order effect, w is the weight factor usually within the range from 0 to 1 and λ is the number of correlation tiers considered to reflect the global sequence pattern effect (Fig. 2). In general, the greater the k or λ is, the more local or global sequence-order information the model will contain; but if k or λ is too large, it may cause the high-dimension problem, reducing the cluster–tolerant capacity (Chou, 1999) so as to lower down the cross-validation accuracy due to over-fitting or 'high dimension disaster' problem (Wang *et al.*, 2008). Therefore, our searching for the optimal values of the three parameters was within the ranges defined below

$$\begin{cases} 1 \le k \le 10 & \text{with step } \Delta = 1 \\ 0 \le w \le 1 & \text{with step } \Delta = 0.1 \\ 1 \le \lambda \le 20 & \text{with step } \Delta = 1 \end{cases} \tag{12}$$

It can be seen from the above equation that to optimize the values for the three parameters, we need to consider $10 \times 11 \times 20 = 2200$ different cases for each of the two layers in **iEnhancer-2L**. To reduce the computational time, we used the 5-fold cross-validation to determine the optimal parameters, and the results thus obtained for the two layers of **iEnhancer-2L** are

$$\begin{cases} k = 6 \text{ for both 1st and 2nd layers} \\ \lambda = 9 \text{ for both 1st and 2nd layers} \\ w = 0.1 \text{ for the 1st layer} \\ w = 0.4 \text{ for the 2nd layer} \end{cases} \tag{13}$$

Table 1. The jackknife success rates achieved by **iEnhancer-2L** on the benchmark dataset of Equation (1) (see also Supplementary Information S1)

Layer	Benchmark dataset	Sn (%)	Sp (%)	Acc (%)	Mcc	AUC
I[a]	$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^-$	78.09	75.88	76.89	0.54	0.85
II[b]	$\mathbb{S}^+ = \mathbb{S}^+_{\text{strong}} \cup \mathbb{S}^+_{\text{weak}}$	62.21	61.82	61.93	0.24	0.66

[a]See Equation (13) for the parameters used for the first layer of **iEnhancer-2L**.
[b]See Equation (13) for the parameters used for the second layer of **iEnhancer-2L**.

3.2 The performance in identifying enhancers and classifying their strength types

Subsequently, the parameter values in Equation (13) are used to conduct the rigorous jackknife tests on the **iEnhancer-2L** predictor for calculating the Sn, Sp, Acc and MCC as defined in Equation (11). The results thus obtained are given in Table 1, from which we can see that the overall accuracy (Acc) by the first layer of **iEnhancer-2L** is ~77% with MCC being 0.54. Although these rates are only slightly better than the corresponding jackknife rates obtained by gkm-SVM (Ghandi *et al.*, 2014) and kmer (Ioshikhes *et al.*, 1996) on the same benchmark dataset used in this study, the success rates (see the third row of Table 1) achieved by the second layer of **iEnhancer-2L** in identifying the types of enhancers is beyond the reach of gkm-SVM (Ghandi *et al.*, 2014) and kmer (Ioshikhes *et al.*, 1996). To the best of our knowledge, **iEnhancer-2L** is so far the first computational predictor ever established for being able to identifying not only the enhances but also their types in strength.

As pointed out in Chou (2015), the availability of web-server for a prediction method will make it much more convenient to users. Unfortunately, neither kmer (Ioshikhes *et al.*, 1996) nor gkm-SVM (Ghandi *et al.*, 2014) had a web-server, and hence their practical usage was limited. In contrast, the web-server for the new **iEnhancer-2L** is available, as detailed below.

3.3 Web-server and its user guide

For the convenience of users, particularly for the vast majority of experimental scientists, a publicly accessible web-server for **iEnhancer-2L** has been established. Moreover, to maximize users' convenience, given below is a step-by-step guide on how to use it to get the desired results without the need to go through the above mathematical details.

Step 1: Open the web-server by clicking the link at http://bioinformatics.hitsz.edu.cn/iEnhancer-2L/ and you will see the top page of **iEnhancer-2L** as shown in Figure 4. Click on the *Readme* button to see a brief introduction about the server.

Step 2: You can either type or copy and paste the query sequences into the input box at the center of Figure 4, or directly upload your input data by the *Browse* button. The input sequence should be in the FASTA format. A potential sequence in FASTA format consists of a single initial line beginning with the symbol, >, in the first column, followed by lines of sequence data in which nucleotides are represented using single-letter codes. Except for the mandatory symbol >, all the other characters in the single initial line are optional and only used for the purpose of identification and description. The sequence ends if another line starting with the symbol > appears; this indicates the start of another sequence. Example sequences in FASTA format can be seen by clicking on the *Example* button right above the input box.

Step 3: Click on the *Submit* button to see the predicted results. The predicted results will be shown on the screen. For example, if

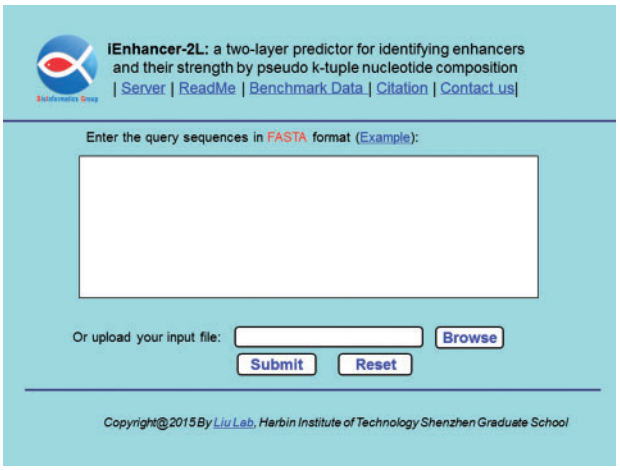

Fig. 4. A semi-screenshot to show the top page of the web-server **iEnhancer-2L**. Its web-site address is http://bioinformatics.hitsz.edu.cn/iEnhancer-2L/

using the query sequences in the *Example* window, you will see the following on the screen: (1) the first query sequence contains nine strong enhancers: sub-sequences 1—200, 2—201, 3—202, 4—203, 5—204, 6—205, 7—206, 8—207 and 9—208; (2) the second query sequence contains one weak enhancer at sub-sequence 1—200; (3) the third query sequence contains one weak enhancer at sub-sequence 1—200; (4) the fourth query sequence contains no enhancer; (5) the fifth query sequence contains no enhancer. All these predicted results are fully consistent with experimental observations.

Step 4: You can also download the predicted results into a file by clicking the *Download* button on the aforementioned screen.

4 Conclusion

It is a big challenge to identify the enhancers from enormous amount of DNA sequences generated in the postgenomic era. One of the keys to develop a sequence-based predictor is how to effectively formulate the sequence samples concerned. Inspired by the successes of introducing the pseudo amino acid composition or PseAAC to formulate protein/peptide samples (Chou 2001, 2005), in the currently proposed **iEnhancer-2L** predictor, we have used the pseudo *k*-tuple nucleotide composition or PseKNC to formulate DNA samples, in which the local sequence patterns of DNA samples are reflected by the *k*-mer composition, while their global sequence patterns reflected by the pseudo components (Liu *et al.*, 2015f).

iEnhancer-2L is a two-layer predictor. Its first layer is to identify whether a query DNA element is of enhancer or not. If the outcome is yes, then the second layer will automatically continue to identify its strength: strong or weak. To the best of our knowledge, it is the first predictor ever developed that can be also used to classify enhancers according to their strength.

A user-friendly web server for **iEnhancer-2L** has been established at http://bioinformatics.hitsz.edu.cn/iEnhancer-2L/, by which users can easily obtain their desired results without the need to go through the complicated mathematics involved, which were presented here just for its integrity. It is also the first web-server ever established for identification/classification of enhancers.

It is anticipated that **iEnhancer-2L** may become a very useful high throughput tool for studying enhancers or, at the very least, play an important complementary role to the existing methods in this area.

Funding

This work was supported by the National Natural Science Foundation of China [No. 61300112, 61573118 and 61272383], the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, the Natural Science Foundation of Guangdong Province [2014A030313695] and National High Technology Research and Development Program of China (863 Program) [2015AA015405].

Conflict of Interest: none declared.

References

Boyle,A.P. *et al.* (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, **21**, 456–464.

Brukner,I. *et al.* (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.*, **14**, 1812–1818.

Cai,Y.D. *et al.* (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.*, **84**, 3257–3263.

Chang,C.C. and Lin, C.J. (2011) LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, **2**, 1–27.

Chen,J. *et al.* (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, **33**, 423–428.

Chen,W. *et al.* (2012) iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One*, **7**, e47843.

Chen,W. *et al.* (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.*, **41**, e68.

Chen,W. *et al.* (2014a) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed. Res. Int. (BMRI)*, **2014**, 623149.

Chen,W. *et al.* (2014b) PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.*, **456**, 53–60.

Chen,W. *et al.* (2014c) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.*, **462**, 76–83.

Chen,W. *et al.* (2015a) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol. BioSyst.*, **11**, 2620–2634.

Chen,W. *et al.* (2015b) iRNA-Methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*, **490**, 26–33.

Chen,W. *et al.* (2015c) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics*, **31**, 119–120.

Chou,K.C. (1999) A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun. (BBRC)*, **264**, 216–224.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins (Erratum: ibid., 2001, Vol. 44, 60)*, **43**, 246–255.

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

Chou,K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.*, **273**, 236–247.

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

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

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.

Chou,K.C., and Shen, H.B. (2007a) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun. (BBRC)*, **360**, 339–345.

Chou,K.C., and Shen, H.B. (2007b) Review: recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **370**, 1–16.

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

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

Cortes,C., and Vapnik, V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Cristianini,N., and Shawe-Taylor, J. (2000) *An Introduction of Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK.

Dehzangi,A. *et al.* (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.

Dickerson,R.E. (1989) Definitions and nomenclature of nucleic acid structure components. *Nucleic Acids Res.*, **17**, 1797–1803.

Ding,H. *et al.* (2014) iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed. Res. Int. (BMRI)*, **2014**, 286419.

Ernst,J., and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Erwin,G.D. *et al.* (2014) Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput. Biol.*, **10**, e1003677.

Fernandez,M., and Miranda-Saavedra,D. (2012) Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Res.*, **40**, e77.

Firpi,H.A. *et al.* (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, **26**, 1579–1586.

Fukue,Y. *et al.* (2005) A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance. *Nucleic Acids Res.*, **33**, 3821–3827.

Ghandi,M. *et al.* (2014) Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLoS Comput. Biol.*, **10**, e1003711.

Gowers,D.M., and Halford,S.E. (2003) Protein motion from non-specific to specific DNA by three-dimensional routes aided by supercoiling. *EMBO J.*, **22**, 1410–1418.

Guo,S.H. *et al.* (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, **30**, 1522–1529.

Hajisharifi,Z. *et al.* (2014) Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.*, **341**, 34–40.

Heintzman,N.D., and Ren, B. (2009) Finding distal regulatory elements in the human genome. *Curr. Opin. Genet. Dev.*, **19**, 541–549.

Heintzman,N.D. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.

Ioshikhes,I. *et al.* (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.*, **262**, 129–139.

Jia,J. *et al.* (2015a) 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.

Jia,J. *et al.* (2015b) Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. *J. Biomol. Struct. Dyn. (JBSD)*, **16**, 1–38, doi:10.1080/07391102.2015.1095116.

Khan,Z.U. *et al.* (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.

Kleftogiannis,D. *et al.* (2015) DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.*, **43**, e6.

Kumar,R. *et al.* (2015) Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.*, **365**, 96–103.

Li,W., and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Lin,W.Z. *et al.* (2013) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. BioSyst.*, **9**, 634–644.

Lin,H. *et al.* (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.

Liu,B. *et al.* (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.

Liu,B. *et al.* (2015a) iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. *J. Biomol. Struct. Dyn. (JBSD)*, **3**, 1–13, doi:10.1080/07391102.2015.1014422.

Liu,B. *et al.* (2015b) Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *J. Theor. Biol.*, **385**, 153–159.

Liu,B. *et al.* (2015c) 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.

Liu,B. *et al.* (2015d) repRNA: a web server for generating various feature vectors of RNA sequences. *Mol. Genet. Genom.* doi:10.1007/s00438-015-1078-7.

Liu,B. *et al.* (2015e) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.*, **43**, W65–W71.

Liu,Z. *et al.* (2015f) iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem. (also Data Brief., 2015, 4, 87–89)*, **474**, 69–77.

Miele,V. *et al.* (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.*, **36**, 3746–3756.

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.

Parker,S.C.J. *et al.* (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science*, **324**, 389–392.

Qiu,W.R. *et al.* (2014) iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. *Biomed. Res. Int. (BMRI)*, **2014**, 947416.

Qiu,W.R. *et al.* (2015) iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model. *J. Biomol. Struct. Dyn. (JBSD)* **33**, 1731–1742.

Rajagopal,N. *et al.* (2013) RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.*, **9**, e1002968.

Shlyueva,D. *et al.* (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.

Visel,A. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.

Wang,T. *et al.* (2008) Predicting membrane protein types by the LLDA algorithm. *Protein Pept. Lett.*, **15**, 915–921.

Wang,X. *et al.* (2015) MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. *Bioinformatics*, **31**, 2639–2645.

Xiao,X. *et al.* (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.

Xiao,X. *et al.* (2013) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.*, **436**, 168–177.

Xiao,X. *et al.* (2015) iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. *J. Biomol. Struct. Dyn. (JBSD)*, **33**, 2221–2233.

Xu,Y. *et al.* (2013a) 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.

Xu,Y. *et al.* (2013b) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *Peer J.*, **1**, e171.

Xu,Y. *et al.* (2014a) 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.

Xu,Y. *et al.* (2014b) iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS One*, **9**, e105018.

Xu,R. *et al.* (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. (JBSD)*, **33**, 1720–1730.

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