BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 11 2014, pages 1522–1529
doi:10.1093/bioinformatics/btu083

Sequence analysis

Advance Access publication February 6, 2014

iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition

Shou-Hui Guo[1], En-Ze Deng[1], Li-Qin Xu[1], Hui Ding[1], Hao Lin[1,2,*], Wei Chen[2,3,*] and Kuo-Chen Chou[2,4,*]

[1]Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China, [2]Gordon Life Science Institute, Belmont, Massachusetts, USA, [3]Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China and [4]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

Associate Editor: John Hancock

ABSTRACT

Motivation: Nucleosome positioning participates in many cellular activities and plays significant roles in regulating cellular processes. With the avalanche of genome sequences generated in the post-genomic age, it is highly desired to develop automated methods for rapidly and effectively identifying nucleosome positioning. Although some computational methods were proposed, most of them were species specific and neglected the intrinsic local structural properties that might play important roles in determining the nucleosome positioning on a DNA sequence.

Results: Here a predictor called 'iNuc-PseKNC' was developed for predicting nucleosome positioning in *Homo sapiens*, *Caenorhabditis elegans* and *Drosophila melanogaster* genomes, respectively. In the new predictor, the samples of DNA sequences were formulated by a novel feature-vector called 'pseudo k-tuple nucleotide composition', into which six DNA local structural properties were incorporated. It was observed by the rigorous cross-validation tests on the three stringent benchmark datasets that the overall success rates achieved by iNuc-PseKNC in predicting the nucleosome positioning of the aforementioned three genomes were 86.27%, 86.90% and 79.97%, respectively. Meanwhile, the results obtained by iNuc-PseKNC on various benchmark datasets used by the previous investigators for different genomes also indicated that the current predictor remarkably outperformed its counterparts.

Availability: A user-friendly web-server, iNuc-PseKNC is freely accessible at http://lin.uestc.edu.cn/server/iNuc-PseKNC.

Contact: hlin@uestc.edu.cn, wchen@gordonlifescience.org, kcchou@gordonlifescience.org

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

Received on December 20, 2013; revised on January 22, 2014; accepted on February 3, 2014

1 INTRODUCTION

The basic unit of eukaryotic chromatin is nucleosome. Each nucleosome contains a 147-bp core DNA (Richmond and Davey, 2003) that is tightly wrapped in 1.67 left-handed super-helical turns around a histone octamer (Segal *et al.*, 2006) as shown in Figure 1. The histone octamer is formed by eight histones, of which two are of H2A, two of H2B, two of H3 and two of H4, and these histones bear highly conservative property in organism evolution (Kornberg, 1977). Under the effect of histone H1, the nucleosome core particle forms a stable structure by further packaging into an advanced structure (Luger *et al.*, 1997). Adjacent nucleosomes are linked via a short DNA sequence, called the linker DNA, which ranges from 10 to 100 bp (Athey *et al.*, 1990; Mavrich *et al.*, 2008a, b).

By modulating the accessibility of genomic regions to regulatory proteins (Albert *et al.*, 2007; Yuan and Liu, 2008), it was observed that the packaging of DNA around the histone octamer played important roles in many biological processes such as transcriptional control, DNA replication, DNA repair and RNA splicing (Berbenetz *et al.*, 2010; Schwartz *et al.*, 2009; Yasuda *et al.*, 2005). Therefore, it is fundamentally important for in-depth understanding the subsequent steps of gene expression to

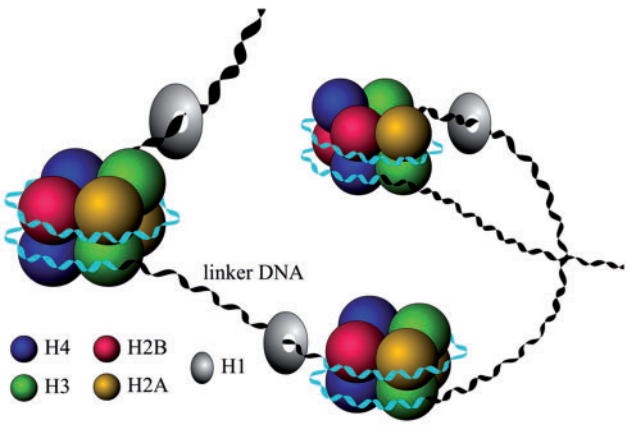

Fig. 1. A schematic illustration to show the basic structure of nucleosome. Each nucleosome consists of ~147bp of DNA wrapped 1.67 turns around a histone octamer. Instead of light blue for nucleosomes, the linker DNAs are colored black. See the text for further explanation

*To whom correspondence should be addressed.

reveal the mechanism involved in controlling nucleosome positioning.

High-throughput techniques, such as chromatin immunoprecipitation (ChIP) coupled with microarrays (ChIP-chip) and ChIP coupled with sequencing techniques (ChIP-Seq), have been developed. Also, high-resolution nucleosome-positioning maps are now available for several model organisms including *Homo sapiens* (Ozsolak *et al*., 2007; Schones *et al*., 2008), *Caenorhabditis elegans* (Valouev *et al*., 2008), *Drosophila melanogaster* (Mavrich *et al*., 2008a, b) and *Sacchoromryces cerevisiae* (Lee *et al*., 2007; Weiner *et al*., 2010). These high-resolution data provided unprecedented opportunities, or made it feasible to develop computational methods for accurately predicting nucleosome positioning by feature extraction approaches.

Satchwell *et al*. (1986) for the first time found that a 10-bp interval repetition of AA/TT/TA occurred in the 147-bp core region of nucleosomes. Widlund *et al*. (1999) demonstrated that CA dinucleotide played an important role in nucleosome positioning, and the sequences containing the fragment TATAAACGCC had high binding affinity to histone. Segal *et al*. (2006) found that ~50% of nucleosome placements were prefigured by genome sequence. It was also observed that nucleosome deficiency always appeared in poly (dA:dT) fragments (Segal and Widom, 2009). Subsequently, Liu *et al*. (Liu *et al*., 2011a, b) found that the 10–11 bp periodicity signals for some particular dinucleotides, such as AA, TT, TA and GC, were more pronounced in the DNA nucleosomal sequences than in the linker DNA sequences. The above findings have demonstrated that nucleosome positioning is sequence-dependent to some extent.

Based on the characteristics of nucleosome positioning sequence (or nucleosomal sequences), various computational methods (Chen *et al*., 2010, 2012b; Gupta *et al*., 2008; Peckham *et al*., 2007; Xing *et al*., 2011, 2013; Zhang *et al*. 2012a,b; Zhao *et al*., 2010) were proposed for predicting nucleosome positioning in different genomes. All these methods could yield quite encouraging results, and each of them did play a role in stimulating the development of this area. However, further work is needed due to the following reasons. (i) The datasets constructed in those methods were too small to reflect the statistical profile of nucleosomes. (ii) No cutoff threshold (Chou and Shen, 2007) was imposed to rigorously exclude the redundant samples or those with high sequence similarity with others in a same dataset. (iii) No web-server was provided to most of these methods, and hence their usage is quite limited, especially for the majority of experimental scientists. (iv) All the local DNA structural properties (Miele *et al*., 2008; Nozaki *et al*., 2011) and their impacts to the global sequence effects were ignored; however, it was demonstrated that this kind of properties might play important roles in determining the rotational positioning of DNA around the histone octamer (Chen *et al*., 2012b).

The present article was initiated in an attempt to improve the prediction of nucleosomes from the above four aspects.

According to a comprehensive review (Chou, 2011) and demonstrated by a series of recent publications (Chen *et al*., 2013; Xiao *et al*., 2013), to establish a really useful statistical predictor for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these procedures one by one.

2 MATERIALS AND METHODS

2.1 Benchmark datasets for the nucleosomal and linker sequences

In this article, we considered the following three species: (i) *H.sapiens*; (ii) *C.elegans*; and (iii) *D.melanogaster*. The experimental data for nucleosome positions in the first species (Schones *et al*., 2008) were downloaded from http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.aspx; those for the second species from http://hgdownload.cse.ucsc.edu; and those for the third species from (Mavrich *et al*., 2008a, b) and http://atlas.bx.psu.edu/. The entire genome sequences for the three species were downloaded from the UCSC genome database at http://hgdownload.cse.ucsc.edu/, where the hg18 version, WS170/ce4 version and BDGP Release 5 version were used for (i) *H.sapiens*, (ii) *C.elegans* and (iii) *D.melanogaster* genomes, respectively.

Since the *H.sapiens* genome and its nucleosome map contain a huge amount of data, according to Liu's strategy (Liu *et al*., 2011a) the nucleosome-forming sequence samples (positive data) and the linkers or nucleosome-inhibiting sequence samples (negative data) were extracted from chromosome 20. As for the other two species, namely *C.elegans* and *D.melanogaster*, the positive and negative data were extracted from their entire genomes. In the datasets thus formed from the three organisms, each of the DNA fragments was assigned with a nucleosome formation score to reflect its propensity to form nucleosome: the higher the score was, the more likely the fragment would be in forming a nucleosome. The DNA fragments with the highest nucleosome formation scores were selected as the nucleosomal sequences, while those with the lowest scores as the linker sequences.

As elaborated in (Chou, 2011), a dataset containing many redundant samples with high similarity would be lack of statistical representativeness. A predictor, if trained and tested by such a biased benchmark dataset, might yield misleading results with overestimated accuracy (Chou and Shen, 2006; Ding, 2013). To get rid of redundancy and avoid bias, the CD-HIT software (Fu *et al*., 2012) was used with the cutoff threshold set at 80% to remove those DNA fragments with high sequence similarity (note that the most stringent cutoff threshold for DNA sequences by CD-HIT was 75%).

Finally, we obtained three benchmark datasets as formulated by

$$\mathbb{S}_k = \mathbb{S}_k^+ \cup \mathbb{S}_k^-, \quad k = \begin{cases} 1 & \text{for } H.sapiens \\ 2 & \text{for } C.elegans \\ 3 & \text{for } D.melanogaster \end{cases} \tag{1}$$

here the positive dataset $\mathbb{S}_1^+$ contains 2273 nucleosome-forming sequences while the negative dataset $\mathbb{S}_1^-$ contains 2300 nucleosome-inhibiting sequences; $\mathbb{S}_2^+$ contains 2567 nucleosome-forming sequences while $\mathbb{S}_2^-$ contains 2608 nucleosome-inhibiting sequences; $\mathbb{S}_3^+$ contains 2900 nucleosome-forming sequences while $\mathbb{S}_3^-$ contains 2850 nucleosome-inhibiting sequences; and the symbol $\cup$ means the union in the set theory. All the sequence samples are 147-bp long; none of them has >80% pairwise sequence identity with any other. The detailed sequences in the three benchmark datasets $\mathbb{S}_1, \mathbb{S}_2$ and $\mathbb{S}_3$ are given in Supplementary Materials S1, S2 and S3, respectively.

2.2 Pseudo k-tuple nucleotide composition

Suppose a DNA sequence **D** with L nucleic acid residues; i.e.

$$\mathbf{D} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7 \ldots \mathrm{R}_L \tag{2}$$

where R_1 denotes the nucleic acid residue at the sequence position 1, R_2 the nucleic acid residue at position 2 and so forth. If the DNA sequence is represented by the k-tuple nucleotide composition (Ioshikhes *et al.*, 1996), the corresponding feature vector will contain 4^k components, as given by

$$\mathbf{D} = \begin{bmatrix} f_1 & f_2 & f_3 & f_4 & \ldots & f_{4^k} \end{bmatrix}^{\mathbf{T}} \tag{3}$$

As we can see from the above equation, with the gradual increase of k, although most of the base sequence-order information within a local or short range could be included, none of the global or long-range sequence-order information would be reflected by the formulation.

Actually, in computational proteomics, we have also faced the same situation; i.e. although the dipeptide composition, tripeptide composition, and k-tuple peptide composition were used by many investigators to represent protein sequences, their global or long-range sequence-order information could still not be reflected. To deal with this problem, the concept of pseudo amino acid composition (Chou, 2001a) or Chou's PseAAC (Lin and Lapointe, 2013) was proposed. Since then, the PseAAC approach has rapidly penetrated into many areas of computational proteomics (see, e.g. Chen and Li, 2013; Esmaeili *et al.*, 2010; Hajisharifi *et al.*, 2014; Liu *et al.*, 2013; Mohabatkar *et al.*, 2011, 2013; Mohammad Beigi *et al.*, 2011; Nanni and Lumini, 2008; Nanni *et al.*, 2012; Sahu and Panda, 2010) and a long list of references cited in a review (Chou, 2011). Owing to its wide usage, recently two powerful softwares, called 'PseAAC-Builder' (Du *et al.*, 2012) and 'propy' (Cao *et al.*, 2013), were established for generating various special pseudo-amino acid compositions.

Stimulated by the PseAAC approach (Chou, 2001a, 2005) in computational proteomics, below we are to propose a novel feature vector, called 'pseudo k-tuple nucleotide composition' (PseKNC), to represent DNA-sequence samples by incorporating the global or long-range sequence-order effects so as to improve the prediction quality in identifying nucleosomes.

Similar to Equation (5) of Chou (2001a) or Equation (3) of Chou (2009), the PseKNC can be formulated as

$$\mathbf{D} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{4^k} & d_{4^k+1} & \cdots & d_{4^k+\lambda} \end{bmatrix}^{\mathbf{T}} \tag{4}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j} & \left(1 \le u \le 4^k\right) \\ \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_i + w\sum_{j=1}^{\lambda}\theta_j} & \left(4^k \le u \le 4^k+\lambda\right) \end{cases}. \tag{5}$$

In the above equation, λ is the number of the total counted ranks (or tiers) of the correlations along a DNA sequence; f_u ($u = 1, 2, \ldots, 4^k$) are the same as Equation (3) that are now normalized to $\sum_{i=1}^{4^k} f_i = 1$; while w is the weight factor. The concrete values of λ and w as well as k will be further discussed later, while θ_j is given by

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+j}\mathrm{R}_{i+j+1}) \quad (j = 1, 2, \ldots, \lambda; \quad \lambda < L) \tag{6}$$

which represents the j-tier structural correlation factor between all the j-th most contiguous dinucleotide. For example, θ_1 is 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 reflects the second-tier correlation factor between all the second-most contiguous dinucleotide (Fig. 2b); θ_3 reflects the third-tier correlation factor between all the third-most contiguous dinucleotide (Fig. 2c); and so forth.

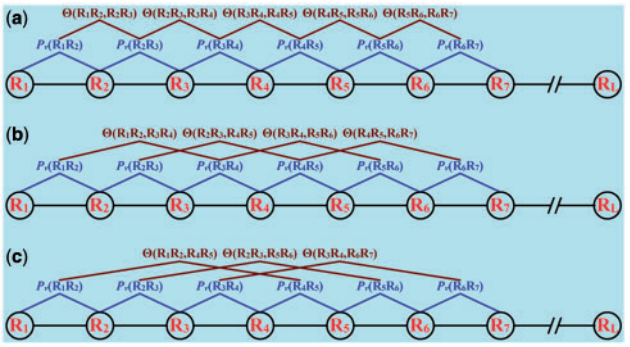

Fig. 2. A schematic drawing to show the correlations of dinucleotides along a DNA sequence for (A) the first-tier coupling that reflects the correlation mode between all the most contiguous dinucleotide, (B) the second-tier coupling between all the second-most contiguous dinucleotide and (C) the third-tier coupling between all the third-most contiguous dinucleotide

Accordingly, the parameter λ actually represents the highest counted rank (or tier) of the correlation along a DNA sequence, and hence must be an integer. The correlation function $\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+j}\mathrm{R}_{i+j+1})$ in Equation (6) is defined by

$$\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+j}\mathrm{R}_{i+j+1}) = \frac{1}{\mu}\sum_{v=1}^{\mu}[P_v(\mathrm{R}_i\mathrm{R}_{i+1}) - P_v(\mathrm{R}_{i+j}\mathrm{R}_{i+j+1})]^2 \tag{7}$$

where μ is the number of local DNA structural properties considered that is equal to 6 in the current article as will be explained below; $P_v(\mathrm{R}_i\,\mathrm{R}_{i+1})$, the numerical value of the v-th ($v = 1, 2, \ldots, \mu$) DNA local structural property for the dinucleotide $\mathrm{R}_i\,\mathrm{R}_{i+1}$ at position i and $P_v(\mathrm{R}_{i+j}\,\mathrm{R}_{i+j+1})$ the corresponding value for the dinucleotide $\mathrm{R}_{i+j}\,\mathrm{R}_{i+j+1}$ at position $i+j$.

2.3 DNA local structural property parameters

It has been reported that DNA structural properties play important roles in many biological processes, such as prokaryotic transcription initiation, protein–DNA interactions, formation of chromosomes and meiotic recombination (Abeel *et al.*, 2008; Chen *et al.*, 2013; Goni *et al.*, 2007, 2008). Recently, Miele *et al.* (2008) developed a model to predict nucleosome occupancy by using basic physical properties. Their model captures a substantial part of chromatin's structural complexity, thus leading to a much better prediction of nucleosome occupancy than the methods based only on periodic curved-DNA motifs (Miele *et al.*, 2008).

Illuminated by Miele's work (Miele *et al.*, 2008), in this article, the DNA local structural properties were considered to define PseKNC. Generally speaking, the spatial arrangements of two neighboring base pairs are characterized by six parameters (Dickerson, 1989), of which three are local translational parameters and other three the local angular parameters, as summarized in Equation (8)

$$\text{Translational} = \begin{cases} \text{Rise} \\ \text{Slide} \\ \text{Shift} \end{cases} \qquad \text{Angular} = \begin{cases} \text{Twist} \\ \text{Roll} \\ \text{Tilt} \end{cases} \tag{8}$$

and illustrated in Figure 3. The detailed values for the six DNA local structural property parameters are given in Table S1 of Supplementary Material S4, which will be used to calculate the global or long-rangesequence-order effects for the nucleosome and linker sequences via Equation (7) as well as Equations (4–6).

Note that before substituting them into Equation (7), all the original values in Table S1 of Supplementary Material S4 for $P_v(\mathrm{R}_i\ \mathrm{R}_{i+1})$

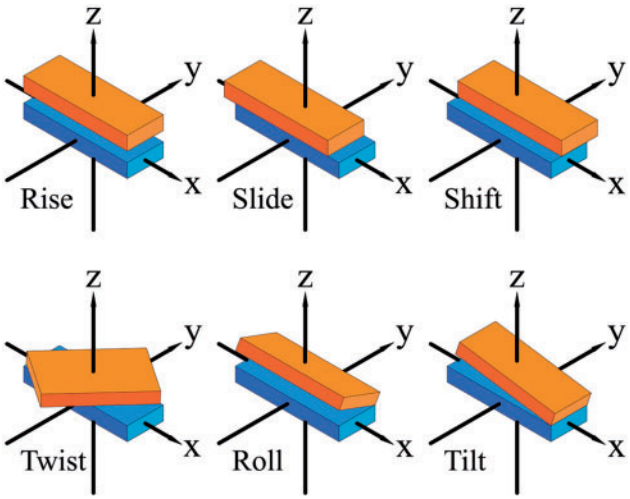

Fig. 3. A schematic illustration to show the six spatial arrangements between two neighboring base pairs in DNA, where one is colored orange and the other blue. See the text and Equation (8) for further explanation

($v = 1, 2, \ldots, 6$) were subjected to a standard conversion (Chou, 2005) as described by the following equation:

$$P_v(\mathrm{R}_i\mathrm{R}_{i+1}) \Leftarrow \frac{P_v(\mathrm{R}_i\mathrm{R}_{i+1}) - <P_v>}{\mathrm{SD}(P_v)} \tag{9}$$

where the symbol $<>$ means taking the average of the quantity therein over 16 different dinucleotides, and SD means the corresponding standard deviation. 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. Listed in Table S2 of Supplementary Material S4 are the values of $P_v(\mathrm{R}_i\ \mathrm{R}_{i+1})$ ($v = 1, 2, \ldots, 6$) obtained via the standard conversion of Equation (9) from those of Table S1 Supplementary Material S4.

2.4 Support vector machine

Support vector machine (SVM) is a powerful and popular method for pattern recognition that has been widely used in the realm of bioinformatics (Bhasin and Raghava, 2004; Mohabatkar *et al.*, 2011; Wan, 2013; Chen *et al.*, 2012c). The basic idea of SVM is to transform the data into a high dimensional feature space, and then determine the optimal separating hyper plane using a kernel function. To handle a multi-class problem, 'one-versus-one (OVO)' and 'one-versus-rest (OVR)' are generally applied to extend the traditional SVM. For a brief formulation of SVM and how it works, see (Chou and Cai, 2002). For more details about SVM, see a monograph Cristianini and Shawe-Taylor (2000).

In the current article, the LIBSVM 2.86 package (Fan *et al.*, 2005) was used as an implementation of SVM, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. The radial basis function was selected as the kernel function due to its effectiveness and speed in training process. The optimal penalty constant C and width parameter γ were determined via an optimization procedure using a grid search approach.

The predictor obtained via the above procedures is called iNuc-PseKNC, where 'i' stands for 'identify', 'Nuc' for 'nucleosome', 'Pse' for 'pseudo', 'K' for '*k*-tuple', 'N' for 'nucleotide' and 'C' for 'composition'. Moreover, its web-server has also been established as will be further described later.

3 RESULTS AND DISCUSSION

3.1 Criteria for performance evaluation

One of the important procedures in developing a useful statistical predictor (Chou, 2011) is to objectively evaluate its performance or anticipated success rate. To realize this, we first need a set of metrics to quantitatively measure the performance of a predictor. Here, let us use the criterion proposed in (Chou, 2001b, c) to develop a set of more intuitive and easier-to-understand metrics. According to that criterion, the rates of correct predictions for the nucleosome-forming sequences and the nucleosome-inhibiting sequences are, respectively, defined by

$$\begin{cases} \Lambda^+ = \frac{N^+ - N^+_-}{N^+}, & \text{for the nucleosome-forming sequences} \\ \Lambda^- = \frac{N^- - N^-_+}{N^-}, & \text{for the nucleosome-inhibiting sequences} \end{cases} \tag{10}$$

where N^+ is the total number of the nucleosome-forming sequences investigated while N^+_- the number of the nucleosome-forming sequences incorrectly predicted as the nucleosome-inhibiting sequences; N^- the total number of the nucleosome-inhibiting sequences investigated while N^-_+ the number of the nucleosome-inhibiting sequences incorrectly predicted as nucleosome-forming sequences. Based on the symbols in Equation (10), the following set of metrics can be obtained (Xu *et al.*, 2013)

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

where Sn stands for the sensitivity, Sp for the specificity, Acc for the accuracy and MCC for the Mathew's correlation coefficient. Such four metrics are generally used in statistical prediction for quantitatively measuring the performance of a predictor from four different angles. In some statistical analysis, Sn is also called the 'true positive rate' and (1 − Sp) the 'false positive rate', as will be further discussed later.

From Equation (11), we can easily see the following. When $N^+_- = 0$ meaning none of the nucleosome-forming sequences was incorrectly predicted to be a nucleosome-inhibiting sequence, we have the sensitivity Sn = 1. When $N^+_- = N^+$ meaning that all the nucleosome-forming sequences were incorrectly predicted to be the nucleosome-inhibiting sequences, we have the sensitivity Sn = 0. Likewise, when $N^-_+ = 0$ meaning none of the nucleosome-inhibiting sequences was incorrectly predicted to be a nucleosome-forming sequence, we have the specificity Sp = 1; whereas $N^-_+ = N^-$ meaning all the nucleosome-inhibiting sequences were incorrectly predicted to be the nucleosome-forming sequences, we have the specificity Sp = 0. When $N^-_+ = N^+_- = 0$ meaning that none of the nucleosome-forming sequences and none of the nucleosome-inhibiting sequences were incorrectly predicted, we have the overall accuracy Acc = 1 and Mathew's correlation coefficient MCC = 1; when

$N_-^+ = N^+$and $N_+^- = N^-$ meaning that all the nucleosome-forming sequences and all the nucleosome-inhibiting sequences were incorrectly predicted, we have 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 prediction. As we can see from the above discussion based on Equation (11), the meanings of the four metrics have become much more intuitive and easier-to-understand, particularly for the Mathew's correlation coefficient, which is usually used for measuring the quality of binary (two-class) classifications as in the case of current article.

It is instructive to point out that the aforementioned metrics are valid only to the single-label system in which a sample investigated belongs to one, and only one class. In other words, a same nucleotide sequence cannot belong to both nucleosome-forming class and nucleosome-inhibiting class. However, it has been observed recently that some molecular biosystems and biomedical systems are actually the multi-label systems in which some of their constituent molecules may belong to two or more attributes (Chen *et al.*, 2013; Lin *et al.*, 2013; Xiao *et al.*, 2013), and hence need two or more labels to tag them (Chou, 2013).

3.2 Cross validation

Three cross-validation methods, i.e. independent dataset test, sub-sampling (or *K*-fold cross validation) test and jackknife test, are often used to evaluate the anticipated success rate of a predictor (Chou and Zhang, 1995). Among the three methods, however, the jackknife test is deemed the least arbitrary and most objective as elucidated in (Chou and Shen, 2008) and demonstrated by Equations (28–32) of Chou (2011), and hence has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors (see, e.g. Chen *et al.*, 2012a, 2013; Chen and Li, 2013; Chou *et al.*, 2012; Esmaeili *et al.*, 2010; Gupta *et al.*, 2013; Mei, 2012; Mohabatkar *et al.*, 2011, 2013). Accordingly, the jackknife test was also used to examine the performance of the model proposed in the current article. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample and all the rule-parameters are calculated without including the one being identified.

3.3 Parameter optimization

As we can see from Equations (4 and 5), the current prediction model was based on three parameters, namely k, λ and w, where w is the weight factor usually within the range from 0 to 1, k reflects the local or short-range sequence-order effect, and λ the global or long-range sequence-order effect. Generally speaking, the greater the k is, the more local sequence-order information the model contains. Also, the greater the λ is, the more global sequence-order information the model contains. However, if k or λ is too large, it would reduce the cluster-tolerant capacity (Chou, 1999) so as to lower down the cross-validation accuracy due to overfitting or 'high dimension disaster' problem (Wang *et al.*, 2008). Therefore, our searching for the optimal values for the three parameters were carried out according to the following

$$\begin{cases} 2 \le k \le 6, & \text{with step } \Delta = 1 \\ 1 \le \lambda \le 20, & \text{with step } \Delta = 1 \\ 0 \le w \le 1, & \text{with step } \Delta = 0.1 \end{cases} \quad (12)$$

As we can see from above, there are $5 \times 20 \times 11 = 1100$ combinations (or points in the 3D parameter space) that need to be considered for finding the optimal parameter values. To reduce the computational time, let us first use the 5-fold cross-validation approach to deal with the parameter optimization. For example, a histogram is given in Figure 4 to show how different k values would affect the predicted results.

Once the optimal values of the three parameters are determined, the rigorous jackknife test will be performed to finally evaluate the anticipated success rate of the predictor.

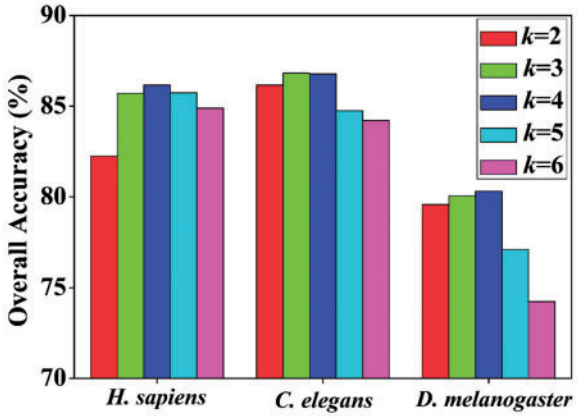

Fig. 4. A histogram to show the overall accuracy by iNuc-PseKNC in identifying nucleosomes with different k values. The accuracy for *H.sapiens* or *D.melanogaster* reaches a peak when $k = 4$, while that for *C.elegans* reaches a peak when $k = 3$

3.4 Prediction quality

Listed in Table 1 are the prediction quality measured by the four metrics defined in Equation (11) for the iNuc-PseKNC predictor in identifying nucleosomes via the rigorous jackknife cross validation. The optimal values for the predictor's three parameters k, λ and w are also given in the table.

Meanwhile, to provide a graphical illustration to show the performance of the current binary classifier iNuc-PseKNC as its discrimination threshold is varied, a 2D plot, called ROC (receiver operating characteristic) curve, is given in Figure 5, where its vertical coordinate Y is for the true positive rate or Sn [cf. Equation (11)] while horizontal coordinate X for the false positive rate or 1−Sp. The best possible prediction method would yield a point with the coordinate (0,1) representing 100% true positive rate (sensitivity Sn) 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 AUROC is often used to indicate the performance quality of a

Table 1. The prediction quality of iNuc-PseKNC measured by four metrics via jackknife tests

Species	Optimal parameters			Metrics			
	k	λ	w	Acc (%)	Sn (%)	Sp (%)	MCC
H.sapiens[a]	4	6	0.5	86.27	87.86	84.70	0.73
C.elegans[b]	3	11	0.5	86.90	90.30	83.55	0.74
D.melanogaster[c]	4	7	0.2	79.97	78.31	81.65	0.60

[a]Using the benchmark dataset given in Supplementary Material S1.
[b]Using the benchmark dataset given in Supplementary Material S2.
[c]Using the benchmark dataset given in Supplementary Material S3.

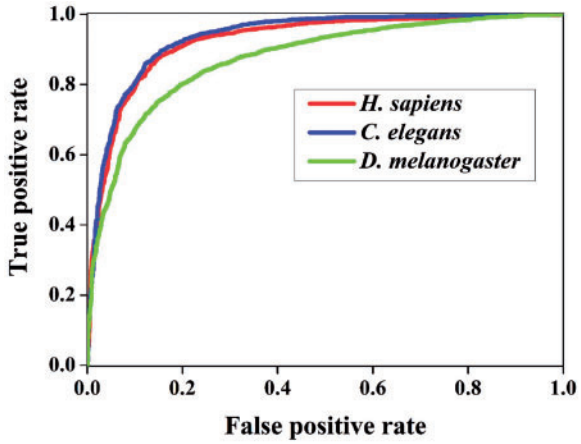

Fig. 5. A graphical illustration to show the performance of the iNuc-PseKNC by means of the ROC curves. The areas under the ROC curves, or AUROC, are 0.925, 0.935 and 0.874 for *H.sapiens*, *C.elegans* and *D.melanogaster*, respectively

binary classifier: the value 0.5 of AUROC is equivalent to random prediction while 1 of AUROC represents a perfect one.

As we can see from Table 1 and Figure 5, even for such large and stringent benchmark datasets, the rates obtained by iNuc-PseKNC were all considerably high, indicating the current predictor is not only accurate, but also quite stable.

It is instructive to point out that the prediction accuracy by the current method for *D.melanogaster* is not as high as those for *H.sapiens* and *C.elegant* accuracies. The reason is probably due to the fact that the features of *D.melanogaster* nucleosomes are not fully extracted. As is well known, the nucleosomal positions are neither fixed at all developmental stages (or tissues) nor uniformly phased with 100% (Gupta *et al.*, 2008; Peckham *et al.*, 2007; Segal *et al.*, 2006). Accordingly, it is rational to calculate the nucleosomal-forming probability. If the probability of a sequence is $\geq$0.5, the sequence is predicted as nucleosome; otherwise, linker.

3.5 Comparison with the existing predictor

As mentioned in the Section 1, the accuracy rates reported by the existing methods in identifying nucleosomes were based on small benchmark datasets without removing high similarity sequences therein, and hence might be over-estimated as shown in Tables S3–S6 of Supplementary Material S4, where the success rates obtained with the current predictor by using the same test methods and same benchmark datasets as used in the existing predictors (Chen *et al.*, 2012b; Zhang *et al.* 2012a, b; Zhao *et al.*, 2010) are given along with the reported rates in those papers. As we can see from these tables, the current predictor iNuc-PseKNC obviously outperformed its counterparts in identifying nucleosomes measured by all the four metrics as defined in Equation (11) as well as by AUROC, indicating that the novel approach by introducing the 'PseKNC' to represent DNA samples is really very useful.

For example, it was reported last year that the overall success rate in identifying nucleosomes achieved by iNuc-PhysChem (Chen *et al.*, 2012b) via the 5-fold cross-validation test was 96%, higher than that by any of its counterparts. Such a high rate, however, was derived from the benchmark dataset collected from *S.cerevisiae* without undergoing a rigorous screening procedure to exclude the high similarity sequences, just like the benchmark datasets used by its then counterparts. Now, let us see what happened if the identification was made by the current predictor iNuc-PseKNC on the same benchmark dataset via the same test method. The results thus obtained are given in Table S6 Supplementary Material S4, from which we can see that the rates for Sn, Sp and Acc by iNuc-PseKNC on the same benchmark dataset as used by iNuc-PhysChem (Chen *et al.*, 2012b) were all 100%!

4 WEB-SERVER GUIDE OR PROTOCOL

For the convenience of the vast majority of experimental scientists, let us give a step-by-step guide on how to use the iNuc-PseKNC web-server to get their desired results without the need to follow the complicated mathematic equations that were presented just for the integrity in developing the predictor. The detailed steps are as follows.

Step 1. On opening the web server at http://lin.uestc.edu.cn/server/iNuc-PseKNC, the top page of iNuc-PseKNC on computer screen will be seen, as shown in Figure 6. Clicking the 'Read Me' button will give a brief introduction about the predictor and the caveat when using it.

Step 2. On clicking the open circle, the organism concerned will be selected. Either typed or copy/pasted the query DNA sequences into the input box at the center of Figure 6. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with a greater than symbol ('>') in the first column, followed by lines of sequence data. The words right after the '>' symbol 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 a '>' 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. To see the predicted result, the 'Submit' button has to be clicked. For example, if the three query DNA sequences

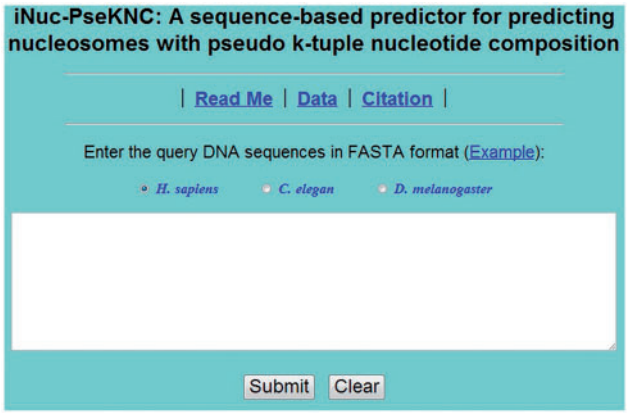

Fig. 6. A semi-screenshot for the top page of the iNuc-PseKNC web-server at http://lin.uestc.edu.cn/server/iNuc-PseKNC

are used from the *H.sapiens* species in the 'Example' window as the input and checking on the '*H.sapiens*' button, after clicking the 'Submit' button, the following shown on the screen of the computer will be seen: the outcome for the first query example (with 147-bp long) is 'nucleosome'; the outcome for the second query sample (with 147-bp long) is 'linker'; the outcome for the third query sample (with 238-bp long) contains $238 - 147 + 1 = 92$ sub-results, in which the outcomes for the segments from #1 to #11 are of 'linker', those for the segments from #72 to #92 are of 'nucleosome' and those from #49 to #71 are of 'linker'. The nucleosome-forming probabilities of these 92 sub-results are also provided. All these results are fully consistent with the experimental observations. It takes about few seconds for the above computation before the predicted result appears on your computer screen; the more number of query sequences and longer of each sequence, the more time it is usually needed. To get the anticipated prediction accuracy, 'the species button consistent with the source of query sequences always be checked': if the query sequences are from *H.sapiens* species, the ''*H.sapiens*'' button is checked; if from *C.elegan*, the '*C.elegan*' button is checked; if from *D.melalogaster*, the '*D.melalogaster*' button is checked.

Step 4. The 'Data' button is clicked to download the benchmark datasets used to train and test the iNuc-PseKNC predictor.

Step 5. The 'Citation' button has to click to find the relevant papers that document the detailed development and algorithm of iNuc-PseKNC.

Caveats. Each of the input query sequences must be 147 bp or longer and only contains valid characters: 'A', 'C', 'G', 'T'.

ACKNOWLEDGEMENTS

The authors would like to thank the three anonymous reviewers for their constructive comments. Many thanks are also due to Dr Zhiqian Zhang for providing the benchmark dataset used in this study.

Funding: National Nature Scientific Foundation of China (grant numbers 61202256, 61301260 and 61100092); Nature Scientific Foundation of Hebei Province (grant number C2013209105); Fundamental Research Funds for the Central Universities (ZYGX2012J113, ZYGX2013J102).

Conflict of Interest: none declared.

REFERENCES

Abeel,T. *et al.* (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18**, 310–323.

Albert,I. *et al.* (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature*, **446**, 572–576.

Athey,B.D. *et al.* (1990) The diameters of frozen-hydrated chromatin fibers increase with DNA linker length: evidence in support of variable diameter models for chromatin. *J. Cell Biol.*, **111**, 795–806.

Berbenetz,N.M. *et al.* (2010) Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS Genet.*, **6**, e1001092.

Bhasin,M. and Raghava,G.P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **32**, W414–W419.

Cao,D.S. *et al.* (2013) propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, **29**, 960–962.

Chen,L. *et al.* (2012a) Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS One*, **7**, e35254.

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

Chen,W. *et al.* (2012c) Prediction of replication origins by calculating DNA structural properties. *FEBS Lett.*, **586**, 934–938.

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

Chen,W. *et al.* (2010) The organization of nucleosomes around splice sites. *Nucleic Acids Res.*, **38**, 2788–2798.

Chen,Y.K. and Li,K.B. (2013) Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.*, **318**, 1–12.

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

Chou,K.C. (2001a) Prediction of protein cellular attributes using pseudo-amino acid composition. *PROTEINS Struct. Funct. Genet.*, **43**, 246–255.

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

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

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

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

Chou,K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. *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. 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. (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. *J. Proteome Res.*, **5**, 1888–1897.

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

Chou,K.C. and Shen,H.B. (2008) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, **3**, 153–162.

Chou,K.C. and Zhang,C.T. (1995) Prediction of protein structural classes. *Crit. Rev. Bioch. 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.
Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction of Support Vector Machines and other Kernel-based Learning Methodds.* Cambridge University Press, Cambridge, UK.
Dickerson,R.E. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Biomol. Struct. Dynam.*, **6**, 627–634.
Ding,H. *et al.* (2013) Prediction of Golgi-resident protein types by using feature selection technique. *Chemometr. Intell. Lab. Syst.*, **124**, 9–13.
Du,P. *et al.* (2012) PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.*, **425**, 117–119.
Esmaeili,M. *et al.* (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.*, **263**, 203–209.
Fan,R.E. *et al.* (2005) Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.*, **6**, 1889–1918.
Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
Goni,J.R. *et al.* (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.*, **8**, R263.
Goni,J.R. *et al.* (2008) DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics*, **24**, 1731–1732.
Gupta,S. *et al.* (2008) Predicting human nucleosome occupancy from primary sequence. *PLoS Comput. Biol.*, **4**, e1000134.
Gupta,M.K. *et al.* (2013) An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition. *SAR QSAR Environ. Res*, **24**, 597–609.
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.
Ioshikhes,I. *et al.* (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.*, **262**, 129–139.
Kornberg,R.D. (1977) Structure of chromatin. *Ann. Rev. Biochem.*, **46**, 931–954.
Lee,W. *et al.* (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.
Lin,S.X. and Lapointe,J. (2013) Theoretical and experimental biology in one. *J. Biomed. Sci. Engineer.*, **6**, 435–442.
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.
Liu,B. *et al.* (2013) Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. *Mol. Inform.*, **32**, 775–782.
Liu,H. *et al.* (2011a) Analysis of nucleosome positioning determined by DNA helix curvature in the human genome. *BMC Genom.*, **12**, 72.
Liu,H. *et al.* (2011b) Role of 10-11bp periodicities of eukaryotic DNA sequence in nucleosome positioning. *BioSystems*, **105**, 295–299.
Luger,K. *et al.* (1997) Crystal structure of the nucleosome core particle at 2.8A resolution. *Nature*, **389**, 251–260.
Mavrich,T.N. *et al.* (2008a) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–1083.
Mavrich,T.N. *et al.* (2008b) Nucleosome organization in the Drosophila genome. *Nature*, **453**, 358–362.
Mei,S. (2012) Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *J. Theor. Biol.*, **310**, 80–87.
Miele,V. *et al.* (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.*, **36**, 3746–3756.
Mohabatkar,H. *et al.* (2013) Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.*, **9**, 133–137.
Mohabatkar,H. *et al.* (2011) Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.*, **281**, 18–23.
Mohammad Beigi,M. *et al.* (2011) Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genom.*, **12**, 191–197.
Nanni,L. and Lumini,A. (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids*, **34**, 653–660.
Nanni,L. *et al.* (2012) Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of chou's pseudo amino acid composition and on evolutionary information. *IEEE/ACM Transact. Comput. Biol. Bioinform. IEEE, ACM*, **9**, 467–475.
Nozaki,T. *et al.* (2011) Computational analysis suggests a highly bendable, fragile structure for nucleosomal DNA. *Gene*, **476**, 10–14.
Ozsolak,F. *et al.* (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.*, **25**, 244–248.
Peckham,H.E. *et al.* (2007) Nucleosome positioning signals in genomic DNA. *Genome Res.*, **17**, 1170–1177.
Richmond,T.J. and Davey,C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150.
Sahu,S.S. and Panda,G. (2010) A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.*, **34**, 320–327.
Satchwell,S.C. *et al.* (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.
Schones,D.E. *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.
Schwartz,S. *et al.* (2009) Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.*, **16**, 990–995.
Segal,E. and Widom,J. (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, **19**, 65–71.
Segal,E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
Valouev,A. *et al.* (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.*, **18**, 1051–1063.
Wan,S. *et al.* (2013) GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. *J. Theor. Biol.*, **323**, 40–48.
Wang,T. *et al.* (2008) Predicting membrane protein types by the LLDA algorithm. *Protein Peptide Lett.*, **15**, 915–921.
Weiner,A. *et al.* (2010) High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res.*, **20**, 90–100.
Widlund,H.R. *et al.* (1999) Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *J. Biol. Chem.*, **274**, 31847–31852.
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.
Xing,Y. *et al.* (2011) Prediction of nucleosome occupancy in Saccharomyces cerevisiae using position-correlation scoring function. *Genomics*, **98**, 359–366.
Xing,Y.Q. *et al.* (2013) An analysis and prediction of nucleosome positioning based on information content. *Chromos. Res.*, **21**, 63–74.
Xu,Y. *et al.* (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.
Yasuda,T. *et al.* (2005) Nucleosomal structure of undamaged DNA regions suppresses the non-specific DNA binding of the XPC complex. *DNA Repair*, **4**, 389–395.
Yuan,G.C. and Liu,J.S. (2008) Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.*, **4**, e13.
Zhang,Z. *et al.* (2012a) Predicting nucleosome positions in yeast: using the absolute frequency. *J. Biomol. Struct. Dnam.*, **29**, 1081–1088.
Zhang,Z. *et al.* (2012b) Prediction of nucleosome positioning using the dinucleotide absolute frequency of DNA fragment. *MATCH Commun. Math. Comput. Chem.*, **63**, 639–650.
Zhao,X. *et al.* (2010) Prediction of nucleosome DNA formation potential and nucleosome positioning using increment of diversity combined with quadratic discriminant analysis. *Chromos. Res.*, **18**, 777–785.