OXFORD ACADEMIC

iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework

FREE

Bin Liu ✉, Ren Long, Kuo-Chen Chou

Bioinformatics (2016) 32 (16): 2411-2418. **DOI:** https://doi.org/10.1093/bioinformatics/btw186
Published: 08 April 2016 **Article history** ▾

Abstract

Motivation: Regulatory DNA elements are associated with DNase I hypersensitive sites (DHSs). Accordingly, identification of DHSs will provide useful insights for in-depth investigation into the function of noncoding genomic regions.

Results: In this study, using the strategy of ensemble learning framework, we proposed a new predictor called *iDHS-EL* for identifying the location of DHS in human genome. It was formed by fusing three individual Random Forest (RF) classifiers into an ensemble predictor. The three RF operators were respectively based on the three special modes of the general pseudo nucleotide composition (PseKNC): (i) kmer, (ii) reverse complement kmer and (iii) pseudo dinucleotide composition. It has been demonstrated that the new predictor remarkably outperforms the relevant state-of-the-art methods in both accuracy and stability.

Availability and Implementation: For the convenience of most experimental scientists, a web server for *iDHS-EL* is established at http://bioinformatics.hitsz.edu.cn/iDHS-EL, which is the first web-server predictor ever established for identifying DHSs, and by which users can easily get their desired results without the need to go through the mathematical details. We anticipate that **iDHS-EL** will become a very useful high throughput tool for genome analysis.

Contact:bliu@gordonlifescience.org or bliu@insun.hit.edu.cn

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

Issue Section: SEQUENCE ANALYSIS

1 Introduction

In genetics, DNase I hypersensitive sites (DHSs) are the regions of chromatin that are sensitive to cleavage by the DNase I enzyme. In these specific regions of the genome, chromatin has lost its condensed structure. As a consequence, the corresponding DNA region will become more exposing and easier to be accessible by enzymes, such as DNase I, and hence enhance its degradation. These accessible chromatin zones are functionally related to transcriptional activity because of the necessity to bind with proteins, such as transcription factors. Since its discovery about 30 years ago (Wu *et al.*, 1979), DHSs have been used as the markers for detecting the regulatory DNA regions.

In general, these specific regions are usually nucleosome-free and associated with a wide variety of genomic regulatory elements, such as promoters, enhancers, insulators, silencers and suppressors (Chen *et al.*, 2016b; Felsenfeld, 1992; Felsenfeld and Groudine, 2003; Gross and Garrard, 1988; Liu *et al.*, 2016b). Accordingly, one effective approach for discovering functional DNA elements from the noncoding sequences is to identify DHSs.

The gold-standard approach for identifying DHS is the Southern blot technique, but it is a tricky, time-consuming and inaccurate to acquire the DHS information by using the Southern blot approach (Crawford *et al.*, 2006). In 2010, by combining the DNase I digestion and high throughput sequencing technology, the DNase-seq technique was proposed (Song and Crawford, 2010), leading to a remarkable enhance in resolution. Unfortunately, there is no effective methodology for analysing the DNase-seq data (Madrigal and Krajewski, 2012). Therefore, one has to resort to the computational approaches for identifying DHSs. In fact some efforts have been made in this regard. For example, Noble *et al.* (2005) proposed a predictor based on the Support Vector Machine (SVM) in which the nucleotide composition was used to formulate the feature vector for predicting DHSs in K562 cell line. Recently, by using the pseudo nucleotide composition (PseKNC) (Chen *et al.*, 2014c, 2015; Liu *et al.*, 2015d), which was developed based on the idea of pseudo amino acid composition for

proteins (Chou, 2001a), Feng *et al.* (2014) proposed a more powerful predictor to identify DHSs by incorporating both the local and global sequence-order effects of DNA.

Although the aforementioned computational approaches yielded quite encouraging results, and they did stimulate the development of this area, some further work is needed due to the following reasons. (i) These existing methods were based on different features extracted from a DNA sequence, and hence lacking the elegance and efficiency of uniform treatment; in other words, a new framework is needed to combine them into one framework. (ii) None of these methods has ever provided a web-server or stand-alone tool, and hence their practical usage value is quite limited, particularly for the majority of experimental biologists (Chou, 2015).

This study was initiated in an attempt to address these shortcomings by developing a more powerful and also more uniform predictor for identifying DHSs. As manifested in a series of recent publications (Chen *et al.*, 2016a; Jia *et al.*, 2016a,b,c; Liu *et al.*, 2015b,f, 2016a,b,c; Xiao *et al.*, 2015), to develop a really useful sequence-based statistical predictor for a biological system and also to make the developing process logically clearer and easier to follow, according to Chou's five-step guidelines (Chou, 2011) we should make the following five procedures crystal clear: (i) benchmark dataset; (ii) sample representation; (iii) operation engine; (iv) cross validation; (v) web-server. Below, let us describe how to deal with these steps one-by one.

2 Materials and Methods

2.1 Benchmark datasets

To develop a statistical predictor, the first important thing is to establish a reliable and stringent benchmark dataset for training and testing the predictor. In this study, the benchmark dataset $\mathbb{S}$ constructed by Noble *et al.* (2005) was adopted; it can be formulated as

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \tag{1}$$

Skip to Main Content

where the positive subset $\mathbb{S}^+$ contains 280 DHS sequences collected from the human genome, the negative subset $\mathbb{S}^-$ contains 737 non-DHS sequences, and $\cup$ denotes the 'union' in the set theory. For readers' convenience, the benchmark dataset are given in Supplementary Material.

2.2 Feature description

Three different features are used to construct three kinds of predictors. They are (i) kmer (Lee *et al.*, 2011), (ii) reverse complement kmer (Gupta *et al.*, 2008) and (iii) pseudo dinucleotide composition (PseDNC) (Chen *et al.*, 2013). These features can be used to reflect the characteristics of a DNA sequence from its different angles, as elaborated below.

2.2.1 Kmer

For a DNA sequence

$$\mathbf{D} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_\mathrm{L} \tag{2}$$

where

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

($i = 1, 2, \cdots, L$) and $\in$ is a symbol in the set theory meaning 'member of'. If using kmer (Lee *et al.*, 2011) or k-tuple nucleotide composition to represent the DNA sequence, we have (Chen *et al.*, 2014c; Liu *et al.*, 2015c)

$$\mathbf{D} = \left[\begin{array}{cccccc} f_1^{\mathrm{kmer}} & f_2^{\mathrm{kmer}} & f_3^{\mathrm{kmer}} & f_4^{\mathrm{kmer}} & \cdots & f_{4^k}^{\mathrm{kmer}} \end{array}\right]^{\mathbf{T}} \tag{4}$$

where f_u^{kmer} $(u = 1,\ 2,\ \cdots,\ 4^k)$ is the occurrence frequency of the u-th k-tuple nucleotide in the DNA sequence and T is the transpose operator. For example, when $k = 2$, Equation 4 will be reduced to the case of 2mer; i.e.

$$\mathbf{D} = \left[\begin{array}{cccccc} f(\mathrm{AA}) & f(\mathrm{AC}) & f(\mathrm{AG}) & f(\mathrm{AT}) & \cdots & f(\mathrm{TT}) \end{array}\right]^{\mathbf{T}}$$

Skip to Main Content

$$= \left[\begin{array}{cccccc} f_1^{2\text{mer}} & f_2^{2\text{mer}} & f_3^{2\text{mer}} & f_4^{2\text{mer}} & \cdots & f_{16}^{2\text{mer}} \end{array} \right]^{\mathbf{T}} \tag{5}$$

2.2.2 Reverse complement kmer

The reverse complement kmer (Gupta *et al.*, 2008; Noble *et al.*, 2005) is a variant of the basic kmer, in which the kmers are not expected to be strand-specific, so reverse complements are degenerated into a single feature. For example, if k = 2, there are totally 16 basic dinucleotides ('AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'CT', 'GA', 'GC', 'GG', 'GT', 'TA', 'TC', 'TG', 'TT'), but by removing the reverse complement 2mers, there are only 10 dinucleotides in the reverse compliment kmer approach ('AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'GA', 'GC', 'TA'). For more information of this approach, please refer to (Gupta *et al.*, 2008; Noble *et al.*, 2005). Accordingly, instead of **Equation 5**, we have

$$D = \left[\begin{array}{cccccc} f(\mathrm{AA}) & f(\mathrm{AC}) & f(\mathrm{AG}) & f(\mathrm{AT}) & \cdots & f(\mathrm{TA}) \end{array} \right]^{\mathbf{T}}$$

$$= \left[\begin{array}{cccccc} f_1^{\mathrm{RC2mer}} & f_2^{\mathrm{RC2mer}} & f_3^{\mathrm{RC2mer}} & f_4^{\mathrm{RC2mer}} & \cdots & f_{10}^{\mathrm{RC2mer}} \end{array} \right]^{\mathbf{T}} \tag{6}$$

where f_u^{RC2mer} $(u = 1,\ 2,\ \cdots,\ 10)$ is the occurrence frequency of the u-th reverse complement 2-tuple nucleotide in the DNA sequence.

2.2.3 Pseudo dinucleotide composition

PseDNC is an approach by incorporating the contiguous local sequence-order information (via 2mer) and the global sequence-order pattern (via the concept of pseudo components (Chou, 2001a)) into the feature vector of the DNA sequence (Chen *et al.*, 2013).

According to PseDNC (Chen *et al.*, 2014c), the DNA sequence **D** of **Equation 2** can be formulated by a vector given by

Skip to Main Content

$$\mathbf{D} = \left[\begin{array}{cccccc} d_1 & d_2 & \cdots & d_{16} & d_{16+1} & \cdots & d_{16+\lambda} \end{array} \right]^{\mathbf{T}} \tag{7}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_{\mathrm{j}}} \left(1 \le k \le 16\right) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} \left(17 \le k \le 16+\lambda\right) \end{cases} \tag{8}$$

where $fk(k = 1, 2,\ldots,16)$ is the normalized occurrence frequency of dinucleotide in the DNA sequence; the parameter λ is an integer, representing the highest counted rank (or tier) of the correlation along a DNA sequence; w is the weight factor ranged from 0 to 1; θj ($j = 1, 2,\ldots,\lambda$) is called the j-tier correlation factor that reflects the sequence-order correlation between all the most contiguous dinucleotides along a DNA sequence, which is defined by Chen *et al.* (2014c):

$$\begin{cases} \theta_1 = \frac{1}{L-2} \sum_{i=1}^{L-2} \Theta\left(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+1}\mathrm{R}_{i+2}\right) \\ \theta_2 = \frac{1}{L-3} \sum_{i=1}^{L-3} \Theta\left(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+2}\mathrm{R}_{i+3}\right) \\ \theta_3 = \frac{1}{L-4} \sum_{i=1}^{L-4} \Theta\left(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+3}\mathrm{R}_{i+4}\right) \qquad (\lambda < L) \\ \vdots \\ \theta_\lambda = \frac{1}{L-1-\lambda} \sum_{i=1}^{L-1-\lambda} \Theta\left(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+\lambda}\mathrm{R}_{i+\lambda+1}\right) \end{cases} \tag{9}$$

where the correlation function is given by

$$\Theta\left(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_j\mathrm{R}_{j+1}\right) = \frac{1}{\mu} \sum_{\mu=1}^{\mu} \left[P_u\left(\mathrm{R}_i\mathrm{R}_{i+1}\right) - P_u\left(\mathrm{R}_j\mathrm{R}_{j+1}\right)\right]^2 \tag{10}$$

where $\mu = 6$ is the number of physicochemical properties considered in this study (see Table 1); $P_u\left(R_iR_{i+1}\right)$ and $P_u\left(R_jR_{j+1}\right)$ are the normalized numerical values of the u-th physicochemical index for the dinucleotides $\mathrm{R}_i\mathrm{R}_{i+1}$ and $\mathrm{R}_j\mathrm{R}_{j+1}$, respectively.

Table 1.

The normalized values for the six DNA dinucleotide physicochemical properties

Dinucleotide[a]	Physicochemical property					
Sk						
	Roll	Tilt	Twist	Slide	Shift	Rise

AA	0.11	0.27	0.5	0.06	1.59	−0.11
AC	1.29	0.80	0.5	1.5	0.13	1.04
AG	−0.24	0.09	0.36	0.78	0.68	−0.62
AT	2.51	0.62	0.22	1.07	−1.02	1.17
CA	−0.62	−0.27	−1.36	−1.38	−0.86	−1.25
CC	−0.82	−0.09	1.08	0.06	0.56	0.24
CG	−0.29	−0.44	−1.22	−1.66	−0.82	−1.39
CT	−0.24	0.09	0.36	0.78	0.68	−0.62
GA	−0.39	0.27	0.5	−0.08	0.13	0.71
GC	0.65	1.33	0.22	−0.08	−0.35	1.59
GG	−0.82	0.09	1.08	0.06	0.56	0.24
GT	1.29	0.80	0.5	1.5	0.13	1.04
TA	−1.51	−0.44	−2.37	−1.23	−2.24	−1.39
TC	−0.39	0.27	0.5	−0.08	0.13	0.71
TG	−0.62	−0.27	−1.36	−1.38	−0.86	−1.25
TT	0.11	0.27	0.5	0.06	1.59	−0.11

[a] See the main text for further explanation.

The aforementioned three types of feature vectors are actually three special modes of the general PseKNC (Chen *et al.*, 2015c), as can be formulated as

$$\mathbf{D} = [\phi_1 \phi_2 \ \cdots \ \phi_u \ \cdots \ \phi_Z]^{\mathbf{T}} \tag{11}$$

where Z is the dimension of the general PseKNC vector; e.g.

$$Z = \begin{cases} 16 & \text{for 2mer vector} \\ 10 & \text{for RC2mer vector} \\ 16+\lambda & \text{for PseDNC vector} \end{cases} \tag{12}$$

Therefore, they can be easily generated by the web-server called '*Pse-in-One*' (Liu *et al.*, 2015e) established very recently.

2.3 Random Forests algorithm

Widely used in various areas of computational biology [e.g. (Jia *et al.*, 2015a,b, 2016a,b,c; Kandaswamy *et al.*, 2011; Lin *et al.*, 2011; Pugalenthi *et al.*, 2012)], the random forests (RF) algorithm is a powerful algorithm. Its detailed formulation has been clearly described in Breiman (2001), and hence there is no need to repeat here.

As shown above, by using 2mer, RC2mer, and PseDNC, the sample of **Equation 2** can be defined by three different PseKNC vectors, as indicated in Equations 5–7, respectively. Accordingly, we have three different basic RF predictors; i.e.

$$\begin{cases} \mathrm{RF}(1), \text{ when the sample is based on 2mer or Eq.5} \\ \mathrm{RF}(2), \text{ when the sample is based on RC2mer or Eq.6} \\ \mathrm{RF}(3), \text{ when the sample is based on PseDNC or Eq.7} \end{cases} \tag{13}$$

2.4 Ensemble random forests

As demonstrated by a series of previous studies, such as signal peptide prediction (Chou and Shen, 2007c; Shen and Chou, 2007c), membrane protein type classification (Chou and Shen, 2007a; Shen and Chou, 2007d), protein subcellular location prediction (Chou and Shen, 2006; Shen and Chou, 2007b), protein fold pattern recognition (Shen and Chou, 2006), enzyme functional classification (Shen and Chou, 2007a), protein-proteins interaction prediction (Jia *et al.*, 2015a) and protein-protein binding site identification (Jia *et al.*, 2015b), the ensemble predictor formed by fusing an array of individual predictors via a voting system can generate much better prediction quality.

Skip to Main Content

Here, the ensemble predictor is formed by fusing the aforementioned three different individual RF predictor of **Equation 13**; i.e.

$$\mathbb{RF}^{\mathrm{E}} = \mathrm{RF}(1) \forall \mathrm{RF}(2) \forall \mathrm{RF}(3) = \forall_{i=1}^{3} \mathrm{RF}(i) \tag{14}$$

where $\mathbb{RF}^{\mathrm{E}}$ denotes the ensemble predictor, and the symbol $\forall$ denotes the fusing operator (Chou and Shen, 2007b). In this study, the concrete fusion process can be formulated as follows.

For a given DNA sequence sample **D** (see Eq.2), suppose

$$\mathrm{RF}(1) \triangleright \mathbf{D} = P_i \ \ (i = 1,\ 2,\ 3) \tag{15}$$

where the symbol $\triangleright$ is an action operator (Chou and Shen, 2007b) meaning using RF(i) to identify the query sequence **D**, and P_i is the probability thus obtained for the sample query sample **D** belonging to the DHS sequence. Define

$$Y = \tfrac{1}{3} \textstyle\sum_{i=1}^{3} F_i P_i \tag{16}$$

where *Fi* is the fractional factor, and their optimal values were determined via the grid search as given by

$$\begin{cases} F_1 = 0.05 \\ F_2 = 0.90 \\ F_3 = 0.05 \end{cases} \tag{17}$$

Thus, we have

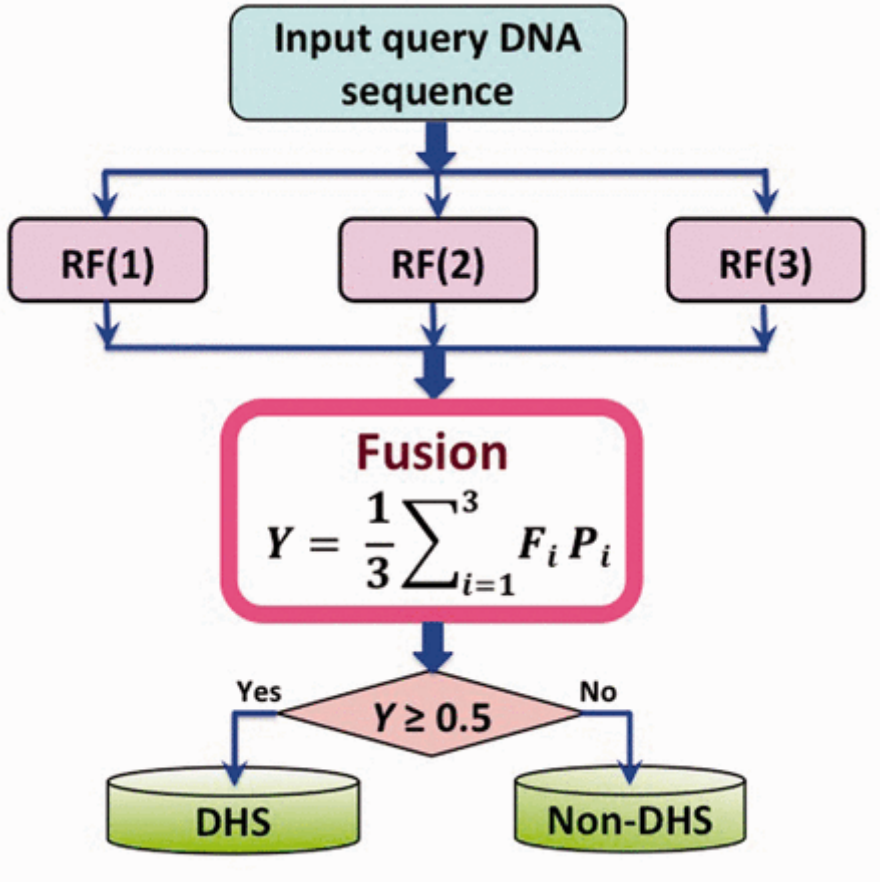
$$\mathbf{D} \in \begin{cases} \mathrm{DHS}, & \text{if } \mathrm{Y} \geq 0.5 \\ \text{non-DHS}, & \text{otherwise} \end{cases} \tag{18}$$

The predictor thus established is called *iDHS-EL*, where ‘i’ stands for ‘identify’, ‘DHS’ for ‘DNase I hypersensitive site’ and ‘EL’ for ‘ensemble learning’. To provide an intuitive picture, a flowchart is provided in Figure 1 to illustrate the prediction process of *iDHS-EL*.

Article Navigation

Skip to Main Content

Fig. 1.



View large Download slide

The flowchart to show how the iDHS-EL predictor works. The three operation engines RF^1, RF^2 and RF^3 are based on the 2mer, RC2mer and PseDNC feature vectors, respectively. See **Equations 5–7** and the relevant text for further explanation

3 Results and discussion

As pointed out in Section 1, among the five guidelines in developing a useful predictor, one of them is how to objectively evaluate its anticipated success rates (Chou, 2011). To fulfil this, the following two things need to consider: one is what metrics should be used to measure the predictor's quality; the other is what kind of test method should be taken to derive the metrics rates. Below, let us to address such two problems.

3.1 Metrics used to reflect the success rates

Skip to Main Content

A set of four metrics are usually used in literature to measure the quality of a predictor: (i) overall accuracy or Acc; (ii) Mathew's correlation coefficient or MCC;

(iii) sensitivity or Sn and (iv) specificity or Sp (Chen *et al.*, 2007). But the four metrics have the following two problems.

First, they are seriously affected by the imbalance degree of a benchmark dataset $\mathbb{S}$ as defined by

$$\mathfrak{P}(\mathbb{S}) = \frac{\text{Number of samples in } \mathbb{S}^+}{\text{Number of samples in } \mathbb{S}^-} = \frac{N(\mathbb{S}^-)}{N(\mathbb{S}^+)} \tag{19}$$

When $\mathfrak{P}(\mathbb{S}) = 1$, the benchmark dataset $\mathbb{S}$ is completely balanced; when $\mathfrak{P}(\mathbb{S}) > 1$, it is negatively imbalanced; when $\mathfrak{P}(\mathbb{S}) < 1$, it is positively imbalanced. The larger the $\mathfrak{P}(\mathbb{S})$, the more skewed the benchmark dataset will be. For the case of this study, $N(\mathbb{S}^-) = 737$ and $N(\mathbb{S}^+) = 280$ (see **Eq. 1** and Supplementary Material), we have $\mathfrak{P}(\mathbb{S}) \approx 2.63$, meaning that the dataset is very skewed in favour to the negative case. To make the performance measurement more objectively reflect a prediction method for a system with high imbalance degree, two additional metrics have been incorporated. One is the product of sensitivity Sn and specificity Sp (Jin and Dunbrack, 2005), denoted here as Pt; the other is the property excess (Yang *et al.*, 2005) denoted here as Py (Jin and Dunbrack, 2005).

Second, the conventional formulations for the four metrics are not intuitive, and most experimental scientists feel hard to understand them, particularly for the MCC. To overcome this problem, let us adopt the formulations proposed in Chen *et al.* (2013) and Xu *et al.* (2013) based on the symbols used by Chou (2001b) in studying the signal peptides.

Thus, we finally have a set of six new metrics as given below

$$\begin{cases} \mathrm{Sn} = 1 - \frac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \frac{N_-^+}{N^+} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \Lambda = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \\ \mathrm{Pt} = \left(1 - \frac{N_-^+}{N^+}\right)\left(1 - \frac{N_+^-}{N^-}\right) & 0 \le \mathrm{Pt} \le 1 \\ \mathrm{Py} = 1 - \frac{N_-^+}{N^+} - \frac{N_+^-}{N^-} & -1 \le \mathrm{Py} \le 1 \end{cases} \quad (20)$$

where N^+ represents the total number of true-phosphorylation samples investigated, whereas N_-^+ the number of phosphorylation samples incorrectly predicted to be of false- phosphorylation sample; N^- the total number of false-phosphorylation samples, whereas N_+^- the number of false-phosphorylation samples incorrectly predicted to be of true-phosphorylation sample.

According to **Equation 20**, the following are crystal clear. (i) When $N_-^+ = 0$ meaning none of the DHS samples is incorrectly predicted to be of non-DHS sample, we have the sensitivity $\mathrm{Sn} = 1$; whereas $N_-^+ = N^+$ meaning that all the DHS samples are incorrectly predicted to be of non-DHS sample, we have the sensitivity $\mathrm{Sn} = 0$. (ii) When $N_+^- = 0$ meaning none of the non-DHS samples is incorrectly predicted to be of DHS sample, we have the specificity $\mathrm{Sp} = 1$; whereas $N_+^- = N^-$ meaning that all the non-DHS samples are incorrectly predicted to be of DHS sample, we have the specificity $\mathrm{Sp} = 0$. (iii) When $N_-^+ = N_+^- = 0$ meaning that none of the DHS samples in the positive dataset and none of the non-DHS samples in the negative dataset is incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 1$ and $\mathrm{MCC} = 1$; whereas $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the DHS samples in the positive dataset and all the DHS samples in the negative dataset are incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 0$ and $\mathrm{MCC} = -1$. (iv) When $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have $\mathrm{Acc} = 0.5$ and $\mathrm{MCC} = 0$ meaning no better than random guessing.

Skip to Main Content

As we can see from the above discussion, the set of metrics formulated in **Equation 20** has made the meanings of sensitivity, specificity, overall accuracy, and MCC much

more intuitive and easier-to-understand, particularly for the meaning of MCC, as concurred and adopted by many authors in a series of recent publications [e.g. (Chen *et al.*, 2014a,b; Ding *et al.*, 2014; Jia *et al.*, 2015a; Lin *et al.*, 2014, 2015a; Xiao *et al.*, 2015)].

Note that, of the six metrics in **Equation 20**, the most important are the Pt and Py in dealing with a system in which the number of negative samples is overwhelmingly greater than that of positive samples, as elaborated in Jin and Dunbrack (2005) and Yang *et al.* (2005). The metrics Acc and MCC reflect the overall accuracy of a predictor and its stability. The metrics Sn and Sp are used to measure a predictor from two opposite angles. When, and only when, both Sn and Sp of the predictor A are higher than those of the predictor B, we can say A is better than B.

Also, it is instructive to point out that the set of equations given in **Equation 20** is valid for the single-label systems only. As for the multi-label systems existing in the system biology (Chou *et al.*, 2012; Lin *et al.*, 2013; Xiao *et al.*, 2011) and system medicine (Xiao *et al.*, 2013), a completely different set of metrics is needed as elucidated in Chou (2013).

3.2 Cross-validation

With a set of intuitive evaluation metrics clearly defined, the next step is what kind of validation method should be adopted to derive the metrics values.

The following three cross-validation methods are often used in literature: (i) independent dataset test, (ii) subsampling (or K-fold cross-validation) test and (iii) jackknife test (Chou and Zhang, 1995). Of these three, 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). Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors [e.g. (Ahmad *et al.*, 2015; Chou and Cai, 2005; Dehzangi *et al.*, 2015; Khan *et al.*, 2015; Kumar *et al.*, 2015; Liu *et al.*, 2015e; Nanni *et al.*, 2014; Shen and Chou, 2007e)].

In this study, however, to reduce the computational time, we adopted the 5-fold cross-validation method, as done by many investigators with RF as the prediction engine. To do this, we first randomly divided the benchmark dataset $\mathbb{S}$ of **Equation 1**

Skip to Main Content

into five groups that they were approximately equal to each other in the size of their subsets, as formulated below

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 = \cup_{i=1}^{5} \mathbb{S}_i \tag{21}$$

where

$$\mathbb{S}_i = \mathbb{S}_i^+ \cup \mathbb{S}_i^- \ \left(i = 1,\ 2,\ \cdots,\ 5\right) \tag{22}$$

with

$$\begin{cases} \left|\mathbb{S}_1^+\right| \approx \left|\mathbb{S}_2^+\right| \approx \left|\mathbb{S}_3^+\right| \approx \left|\mathbb{S}_4^+\right| \approx \left|\mathbb{S}_5^+\right| \\ \left|\mathbb{S}_1^-\right| \approx \left|\mathbb{S}_2^-\right| \approx \left|\mathbb{S}_3^-\right| \approx \left|\mathbb{S}_4^-\right| \approx \left|\mathbb{S}_5^-\right| \end{cases} \tag{23}$$

where $\left|\mathbb{S}_1^+\right|$ denotes the number of samples (or cardinalities) in $\mathbb{S}_1^+$, and so forth. Actually, **Equations 21–23** can also be formulated as

$$\mathbb{S}_1 \triangleq \mathbb{S}_2 \triangleq \mathbb{S}_3 \triangleq \mathbb{S}_4 \triangleq \mathbb{S}_5 \tag{24}$$

where the symbol $\triangleq$ means that the divided five benchmark datasets in **Equation 21** are about the same in size, and so are their subsets (Jia *et al.*, 2016a). Thus, each of the five sub-benchmark datasets was singled out one-by-one and tested by the model trained with the remaining four sub-benchmark datasets. The cross-validation process was repeated for five times, with their average as the final outcome. In other words, during the process of 5-fold cross-validation, both the training dataset and testing dataset were actually open, and each sub-benchmark datasets was in turn moved between the two. The 5-fold cross-validation test can exclude the 'memory' effect, just like conducting 5 different independent dataset tests.

3.3 Comparison with the existing methods

Listed in Table 2 are the 5-fold cross-validation results by *iDHS-EL* on the benchmark dataset of **Equation 1** (see Supplementary Material). For facilitating comparison, listed in that table are also the corresponding results obtained by the *SVM-RevcKmer* predictor (Noble *et al.*, 2005) and *SVM-PseDNC* predictor (Feng *et al.*, 2014),

respectively. From the table, we can see the following. (i) Among the three predictors the newly proposed one achieved the highest success rates in both Pt and Py, the two most important metrics used to measure the quality of a predictor as elucidated in the follow-up text to Equation 20. (**ii**) Although the Sp rate by the proposed predictor was slightly (2.04%) lower than that by *SVM-RevcKmer*, its Sp rate was 5.71% higher than that by *SVM-RevcKmer*. As mentioned in Section 3.1, the two metrics are used to measure a predictor from two opposite angles, and they are constrained with each other (Chou 1993). Therefore, it is meaningless to use only one of the two for comparing the quality of two predictors. In other words, a meaningful comparison in this regard should count the rates of both Sn and Sp, or even better, the rate of their combination that is none but Pt and Py. As shown in Table 2, the Pt and Py rates achieved by *iDHS-EL* are remarkably higher than those by the existing predictors.

Table 2.

A comparison of the proposed predictor with the existing ones via the 5-fold cross-validation on a same benchmark dataset of Supplementary Material S1

Predictor	Sn[a] (%)	Sp[a](%)	Acc[a](%)	MCC[a](%)	Pt[a](%)	Py[a](%)
SVM-RevcKmer[b]	65.36	92.81	85.25	0.616	60.66	58.17
SVM-PseDNC[c]	61.07	92.26	83.68	0.571	56.34	53.33
iDHS-EL[d]	64.64	94.30	86.14	0.636	64.51	61.84

[a] See Equation 20 for the definition of the metrics.

[b] The prediction method developed by (Noble et al. 2005).

[c] The prediction method developed by Feng et al. (2014).

[d] The prediction method proposed in this paper.

3.4 Feature analysis

RF is a combination of decision trees, which have the ability to select important ones from many features and ignore others. Furthermore, because decision trees generate explicit models describing the relationship between features and the predictions,

Skip to Main Content

which facilities the interpretation of the models. A measure of how each feature contributes to the prediction can be calculated in the training process. A random noise value is used to replace a feature. If the performance is obviously degraded, it means that this feature contribute to the prediction. On the other hand, if the performance is stable, it means that this feature is irrelevant. Thus, we can calculate each relative importance of features according to the following procedure (Jiang *et al.*, 2007). For each tree, the average prediction accuracy of the OOB (Out Of Bag) portion of the data is calculated. For each feature, replace its value with random noise, and then the average accuracy is recalculated. Finally, the difference between the two accuracies is then averaged over all trees, and normalized by the standard error. As a result, the mean decrease accuracy represents the relative importance of each feature. As is shown in Table 3, it lists the top 10 most important features of the three individual RF classifiers (see **Eq. 13**). From the table, we can see the following. **(i)** Among the three RF classifiers, there are some common important features, such as CG, GC, CA, which are fully consistent with previous studies (Wang *et al.*, 2012; Zhang *et al.*, 2012). **(ii)** Some features are only important for one RF classifier but not for the others, such as CT, and $\lambda = 1, 2, 3$. These features describe the characteristics of DHSs in different aspects, and therefore, the predictive performance can be improved by combining these complementary features via the proposed ensemble learning framework.

Table 3.

Ranking the top 10 most important features for the three different individual RF classifiers (see Eq. 13)

RF(1)			RF(2)			RF(3)		
Rank	**Feature**	**MDA[a] (%)**	**Rank**	**Feature**	**MDA[a] (%)**	**Rank**	**Feature**	**MDA[a] (%)**
1	CG	35.46	1	CG	11.12	1	CG	29.70
2	GC	14.13	2	CA	9.27	2	GC	14.82
3	CA	12.46	3	GC	7.29	3	CA	10.86
4	TG	11.19	4	AA	3.04	4	TG	10.38

Skip to Main Content

RF(1)			RF(2)			RF(3)		
Rank	**Feature**	**MDA (%)**	**Rank**	**Feature**	**MDA (%)**	**Rank**	**Feature**	**MDA (%)**
5	AT	8.63	5	CC	2.82	5	AT	8.58
6	TT	7.20	6	AG	2.58	6	$\lambda = 2$	6.14
7	TA	6.98	7	AT	2.13	7	TT	6.09
8	AA	6.87	8	GA	1.87	8	TA	5.99
9	CC	6.56	9	AC	1.86	9	$\lambda = 3$	5.67
10	CT	6.05	10	TA	1.76	10	$\lambda = 1$	5.44

[a] The abbreviation of mean decrease accuracy.

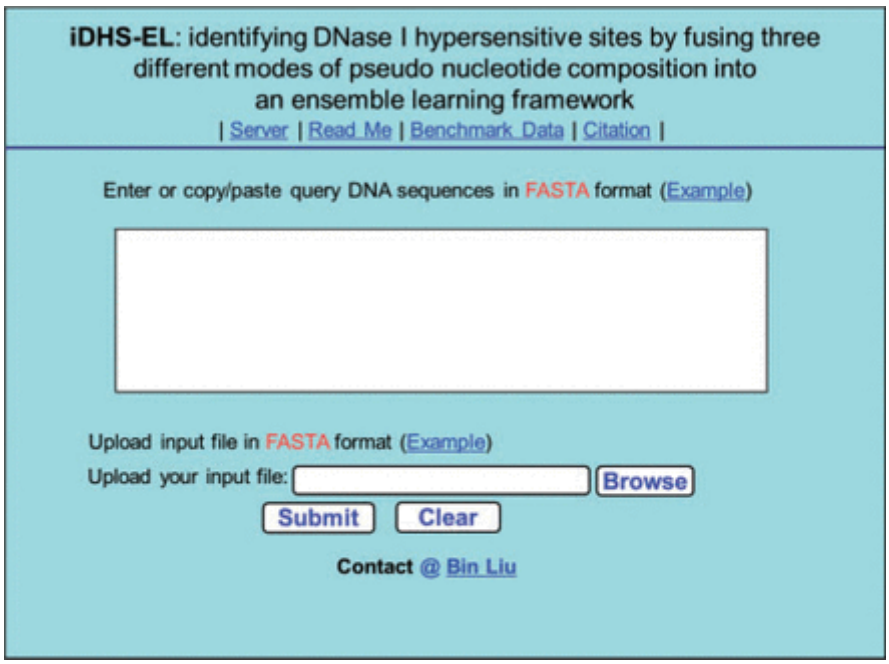
3.5 Web server and user guide

For the convenience of the vast majority of experimental scientists, a web server for the *iDHS-EL* predictor has been established. To our best knowledge, it is the first web-server ever established for predicting the DHSs in human genome. Moreover, to maximize users' convenience, a step-by-step guide on how to use it to get the desired results is given below.

Step 1.Open the web server at http://bioinformatics.hitsz.edu.cn/iDHS-EL, and you will see its top page as shown in Figure 2. Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it.

Fig. 2.

Skip to Main Content



View large Download slide

A semi-screenshot to show the top page of the iDHS-EL web-server. Its web-site address is http://bioinformatics.hitsz.edu.cn/iDHS-EL/

Step 2. You can either type or copy/paste the query DNA sequence into the input box at the center of Figure 2, or directly upload your input data by the 'Browse' button. The input sequence should be in the FASTA format. For the examples of DNA 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. For example, if you use the four query DNA sequences in the Example window as the input, you will see the following shown on the screen of your computer: (**i**) the first query DNA sequence (misc_ppid_8090) is of DHS; (**ii**) the second query sequence (misc_ppid_7576) is of DHS; (**iii**) the third query sequence (misc_ppid_7953) is of non-DHS; (**iv**) the fourth query sequence (misc_ppid_6460) is of non-DHS. All these predicted results are fully consistent with the experimental observations.

Step 4. Click on the 'Benchmark Data' button to download the datasets used to train and test the model.

Step 5. Click on the 'Citation' button to find the relevant papers that document the detailed development and algorithm of *iDHS-EL*.

4 Conclusion

A novel predictor called *iDHS-EL* was proposed for identifying the location of DHS in human genome by fusing the kmer approach, reverse complement kmer approach, dinucleotide-based auto cross covariance approach, and pseudo dinucleotide composition approach into an ensemble classifier.

It was demonstrated by cross-validations on a same benchmark dataset that the new predictor outperformed the state-of-the-art methods (Feng *et al.*, 2014; Noble *et al.*, 2005) in this area. Furthermore, a user-friendly a web server for *iDHS-EL* was provided at http://bioinformatics.hitsz.edu.cn/iDHS-EL/, 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. Also, it is the first web-serve predictor ever established for identifying the location of DHS in human genome.

Acknowledgements

The authors would like to thank Shumin Li for her helpful discussions. The authors also wish to thank the three anonymous reviewers for their constructive comments, which were very helpful in strengthening the presentation of this study. 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), Shenzhen Foundational Research Funding (Grant No. JCYJ20150626110425228), and Development Program of China (863 Program) [2015AA015405].

Conflict of Interest: none declared.

References

Skip to Main Content

Ahmad S . et al. . (2015) Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into Chou's general PseAAC. *Comput. Methods Programs*