Published online 8 January 2013

Nucleic Acids Research, 2013, Vol. 41, No. 6 ***e68***
doi:10.1093/nar/gks1450

iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition

Wei Chen[1,2,*], Peng-Mian Feng[3], Hao Lin[4,*] and Kuo-Chen Chou[2,*]

[1]Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan, China, [2]Bioinformatics and Computer-Aided Drug Discovery, Gordon Life Science Institute, San Diego, CA, USA, [3]School of Public Health, Hebei United University, Tangshan 063000, China and [4]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, China

Received October 9, 2012; Revised November 27, 2012; Accepted December 12, 2012

ABSTRACT

Meiotic recombination is an important biological process. As a main driving force of evolution, recombination provides natural new combinations of genetic variations. Rather than randomly occurring across a genome, meiotic recombination takes place in some genomic regions (the so-called 'hotspots') with higher frequencies, and in the other regions (the so-called 'coldspots') with lower frequencies. Therefore, the information of the hotspots and coldspots would provide useful insights for in-depth studying of the mechanism of recombination and the genome evolution process as well. So far, the recombination regions have been mainly determined by experiments, which are both expensive and time-consuming. 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 the recombination regions. In this study, a predictor, called 'iRSpot-PseDNC', was developed for identifying the recombination hotspots and coldspots. In the new predictor, the samples of DNA sequences are formulated by a novel feature vector, the so-called 'pseudo dinucleotide composition' (PseDNC), into which six local DNA structural properties, i.e. three angular parameters (twist, tilt and roll) and three translational parameters (shift, slide and rise), are incorporated. It was observed by the rigorous jackknife test that the overall success rate achieved by iRSpot-PseDNC was >82% in identifying recombination spots in *Saccharomyces cerevisiae*, indicating the new predictor is promising or at least may become a complementary tool to the existing methods in this area. Although the benchmark data set used to train and test the current method was from *S. cerevisiae*, the basic approaches can also be extended to deal with all the other genomes. Particularly, it has not escaped our notice that the PseDNC approach can be also used to study many other DNA-related problems. As a user-friendly web-server, iRSpot-PseDNC is freely accessible at http://lin.uestc.edu.cn/server/iRSpot-PseDNC.

INTRODUCTION

Genetic recombination describes the generation of new combinations of alleles that occurs at each generation in diploid organisms. It is an important biological process and results from a physical exchange of chromosomal material (1). As a main driving force of evolution, recombination provides new combinations of genetic variations and accelerates the evolution of sexual reproductive organisms. A schematic illustration to show the meiotic recombination pathways is given in Figure 1.

As recombination is crucial to genome evolution, identification and characterization of recombination spots are substantially important. In the past decades, several global mapping studies have been performed to map double-strand breaks sites on chromosomes in yeast to determine the distribution pattern of recombination regions across genome (3–5). They found that meiotic recombination events generally concentrate in $1 \sim 2.5$ kilobase regions and does not occur randomly across the genome. Regions that exhibit elevated rates of recombination

*To whom correspondence should be addressed. Tel: +86 315 3725715; Fax: +86 315 3725715; Email: wchen@gordonlifescience.org; chenwei_imu@yahoo.com.cn
Correspondence may also be addressed to Hao Lin. Tel: +86 28 8320 8232; Fax: +86 28 8320 8238; Email: hlin@uestc.edu.cn
Correspondence may also be addressed to Kuo-Chen Chou. Tel: +1 858 380 4623; Fax: +1 858 380 4623; Email: kcchou@gordonlifescience.org

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

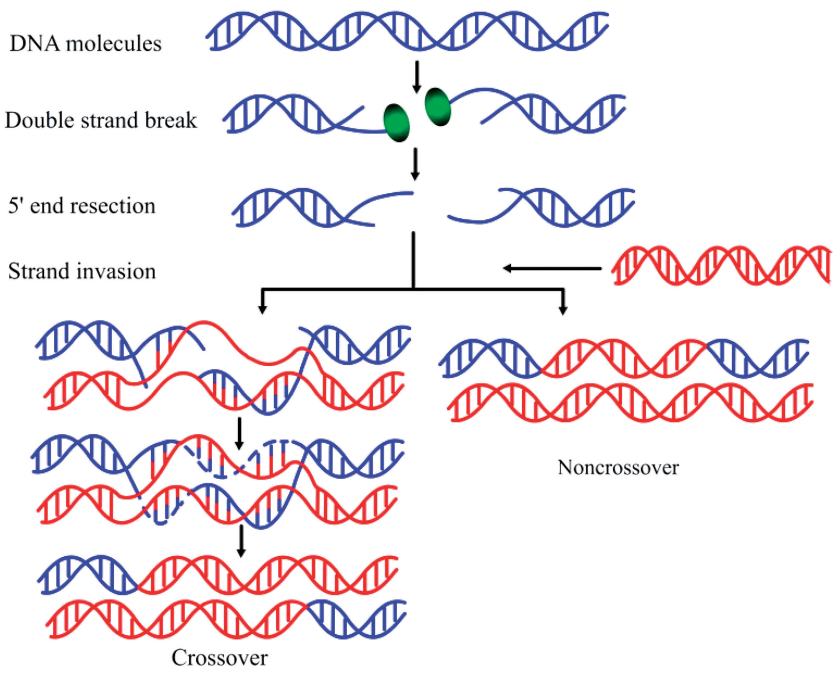

Figure 1. A schematic drawing to show the meiotic recombination pathways in a DNA system. Recombination is initiated by a double-strand break (DSB) catalysed by the Spo11 protein (green ball), a relative of archaeal topoisomerase VI. After DSBs are formed, Spo11 is removed from the DNA molecule (blue helix) and the single-stranded 3′ ends are formed. These tails undergo strand invasion of intact homologous duplexes (red helix), ultimately yielding mature recombinant products. The repair of meiotic DSB can result in either reciprocal exchange of the chromosome arms flanking the break (a crossover) as shown in the left lower panel, or no exchange of flanking arms (a non-crossover or parental configuration) as shown in the right lower panel. Adapted from (2).

relative to a neutral expectation are called recombination hotspots, whereas those with low rates of recombination are recombination coldspots. Additionally, they also found that recombination regions do not share a consensus sequence. With the rapid increasing number of sequenced genomes, it is highly desired to develop reliable automated methods for timely identifying the recombination spots.

Although considerable progress has been made in this regard, the computational predictive accuracy of recombination spots still needs further improvements. The existing computational algorithm for recombination spots prediction was based on the nucleotide sequence contents (6), in which little sequence-order effect was taken into account. To improve the prediction quality, it is necessary to take into account this kind of effect. However, the number of possible patterns for DNA sequences is extremely large, and their lengths vary widely, making it difficult to incorporate the sequence-order information into a statistical predictor. Facing such a difficulty, how can we take into account the sequence-order effect to improve the prediction quality? If it is not feasible to count all the sequence-order information, can we find an approximate way to partially take into account it? Similar problems were also encountered in computational proteomics. To cope with this kind of problems, the concept of pseudo amino acid composition (PseAAC) was proposed by Chou (7). Since then, the concept of PseAAC has penetrated into almost all the fields of computational proteomics, such as predicting protein submitochondrial localization (8), predicting protein structural class (9), predicting DNA-binding proteins (10), identifying bacterial virulent proteins (11), predicting metalloproteinase family (12), predicting protein folding rate (13), predicting GABA(A) receptor proteins (14), predicting protein supersecondary structure (15), predicting cyclin proteins (16), classifying amino acids (17), predicting enzyme family class (18), identifying risk type of human papillomaviruses (19), predicting allergenic proteins (20), identifying G protein-coupled receptors and their types (21) and discriminating outer membrane proteins (22), among many others [see a long list of references cited in a review (23)]. Because of its wide and increasing usage, in 2012, a powerful software called PseAAC-Builder (http://www.pseb.sf.net) (24) was established for generating various special modes of PseAAC, in addition to the earlier web-server PseAAC (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC) (25) built in 2008.

Encouraged by the successes of introducing the PseAAC approach (7,26) into computational proteomics, the present study was initiated in an attempt to propose a novel feature vector, called 'pseudo dinucleotide composition' (PseDNC), to represent DNA sequence samples by incorporating more sequence-order effects so as to improve the quality of predicting the recombination spots.

As summarized in a review (23) and demonstrated by a series of recent publications [see, e.g. (27–29)], 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 data set 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; and (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us elaborate how to deal with these procedures one by one.

MATERIALS AND METHODS

Benchmark data set

The benchmark data set $\mathbb{S}$ for the recombination hotspots and coldspots was taken from Liu *et al.* (6). It contains 490 recombination hotspots and 591 recombination coldspots, as can be formulated by

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

where $\cup$ represents the symbol for 'union' in the set theory; the subset $\mathbb{S}^+$ contains the recombination hotspots only, whereas $\mathbb{S}^-$ recombination coldspots only. For the convenience of readers, the 490 sequences in $\mathbb{S}^+$ and 591 sequences in $\mathbb{S}^-$ are given in the Supplementary Information S1.

PseDNC

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 represents the nucleic acid residue at the sequence position 1, R_2 the nucleic acid residue at position 2 and so forth. If the feature vector of the DNA sequence is formulated by its nucleic acid composition (NAC), we have

$$\mathbf{D} = \begin{bmatrix} f(\mathrm{A}) & f(\mathrm{C}) & f(\mathrm{G}) & f(\mathrm{T}) \end{bmatrix}^{\mathbf{T}} \tag{3}$$

where $f(\mathrm{A})$, $f(\mathrm{C})$, $f(\mathrm{G})$, and $f(\mathrm{T})$ are the normalized occurrence frequencies of adenine (A), cytosine (C), guanine (G) and thymine (T), respectively, in the DNA sequence; the symbol **T** is the transpose operator. As we can see from Equation 3, all the sequence-order information is missed if using NAC to represent a DNA sequence. If using the DNC to represent the DNA sequence, instead of the four components as shown in Equation 3, the corresponding feature vector will contain $4 \times 4 = 16$ components, as given below

$$\begin{aligned} \mathbf{D} &= \begin{bmatrix} f(\mathrm{AA}) & f(\mathrm{AC}) & f(\mathrm{AG}) & f(\mathrm{AT}) & \cdots & f(\mathrm{TT}) \end{bmatrix}^{\mathbf{T}} \\ &= \begin{bmatrix} f_1 & f_2 & f_3 & f_4 & \cdots & f_{16} \end{bmatrix}^{\mathbf{T}} \end{aligned} \tag{4}$$

where $f_1 = f(\mathrm{AA})$ is the normalized occurrence frequency of AA in the DNA sequence; $f_2 = f(\mathrm{AC})$, that of AC; $f_3 = f(\mathrm{AG})$, that of AG and so forth. Although the most contiguous local sequence-order information is included in Equation 4, none of the global sequence-order information is reflected by the formulation. DNC is the most simple pseudo NAC, or PseNAC, according to the terminology similar to that used in (7).

To incorporate the global sequence-order information into the feature vector for the DNA sequence, let us consider the following approach. As shown in Equation 2, the first dinucleotide in the DNA sequence is $\mathrm{R}_1\mathrm{R}_2$, the second dinucleotide is $\mathrm{R}_2\mathrm{R}_3$ and so forth; the last one is $R_{L\text{-}1}\mathrm{R}_L$. Thus, by following the similar procedures as described in (7) to reflect the global sequence-order information of a protein with a set of sequence-order-correlated factors, for the DNA sequence of Equation 2, we also have the corresponding factors as defined below:

$$\begin{cases} \theta_1 = \frac{1}{L-2}\sum_{i=1}^{L-2} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+1}\mathrm{R}_{i+2}) \\ \theta_2 = \frac{1}{L-3}\sum_{i=1}^{L-3} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+2}\mathrm{R}_{i+3}) \\ \theta_3 = \frac{1}{L-4}\sum_{i=1}^{L-4} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+3}\mathrm{R}_{i+4}) \\ \quad \cdots\cdots \\ \theta_\lambda = \frac{1}{L-1-\lambda}\sum_{i=1}^{L-1-\lambda} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+\lambda}\mathrm{R}_{i+\lambda+1}) \end{cases} \quad (\lambda < L) \tag{5}$$

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 (Figure 2a); θ_2, the second-tier correlation factor between all the second most contiguous dinucleotide (Figure 2b); θ_3, the third-tier correlation factor between all the third most contiguous dinucleotide (Figure 2c) and so forth.

In Equation 5, the parameter λ is an integer, representing the highest counted rank (or tier) of the correlation along a DNA sequence, and the correlation function is given by

$$\Theta\left(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_j\mathrm{R}_{j+1}\right) = \frac{1}{\mu}\sum_{u=1}^{\mu}\left[P_u(\mathrm{R}_i\mathrm{R}_{i+1}) - P_u\left(\mathrm{R}_j\mathrm{R}_{j+1}\right)\right]^2 \tag{6}$$

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

DNA local structural properties

Multiple lines of evidences have indicated that some local DNA structural properties, i.e. angular parameters (twist, tilt and roll) and translational parameters (shift, slide and rise), have important roles in biological processes, such as protein–DNA interactions, formation of chromosomes and higher-order organization of the genetic material (30–32). Accordingly, these six structural properties

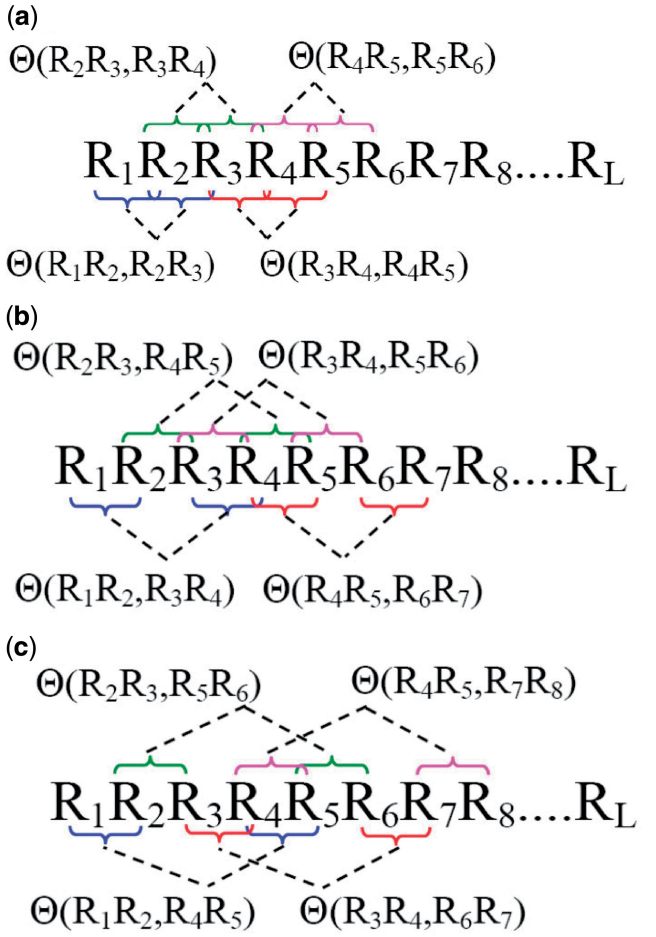

Figure 2. A schematic illustration to show the correlations of dinucleotides along a DNA sequence. (**a**) The first-tier correlation reflects the sequence-order mode between all the most contiguous dinucleotide. (**b**) The second-tier correlation reflects the sequence-order mode between all the second-most contiguous dinucleotide. (**c**) The third-tier correlation reflects the sequence-order mode between all the third-most contiguous dinucleotide.

might have impact on DNA binding of regulatory proteins, either directly by hampering or favoring complex formation or indirectly through the modulation of the chromatin structures and hence the DNA accessibility (33). Listed in Table 1 are their original numerical values derived from (32) for twist $P_1(R_iR_{i+1})$, tilt $P_2(R_iR_{i+1})$, roll $P_3(R_iR_{i+1})$, shift $P_4(R_iR_{i+1})$, slide $P_5(R_iR_{i+1})$, and rise $P_6(R_iR_{i+1})$, respectively, where R_iR_{i+1} represents the 16 possible dinucleotides AA, AC, AG, AT,..., TT. It was these six DNA local physical structural properties that were to be used as correlation functions to derive the PseDNC for the current study. Meanwhile, it is also self-evident why $\mu = 6$ in Equation 6 for the current case.

Before substituting into Equation 6, the original values as listed in Table 1 for $P_u(R_iR_{i+1})$ $(u = 1, 2, \cdots, 6)$, they were all subjected to a standard conversion (26), as described by the following equation

$$P_u(R_iR_{i+1}) = \frac{P_u(R_iR_{i+1}) - < P_u >}{SD(P_u)} \quad (7)$$

where the symbol $< >$ means taking the average of the quantity therein for 16 different dinucleotides (cf. Equation 4), and SD means the corresponding standard deviation. The converted values obtained by Equation 7 will have a zero mean value for the 16 different dinucleotides and will remain unchanged if going through the same conversion procedure again. Listed in Table 2 are the values of $P_u(R_iR_{i+1})$ $(u = 1,2,\cdots,6)$ obtained via the standard conversion of Equation 7 from those of Table 1.

Now we can see from Figure 2 that the sequence-order effect of a DNA sequence can be, to some extent, reflected through a set of sequence-correlation factors θ_1 θ_2, θ_3, $\cdots$, θ_λ, as clearly defined by Equations 5 and 6. Similar to the procedure as described in (7) for converting the amino acid composition to the PseACC, let us augment the DNC of Equation 4 to the PseDNC as given later in the text

$$\mathbf{D} = \left[d_1 \quad d_2 \quad \cdots \quad d_{16} \quad d_{16+1} \quad \cdots \quad d_{16+\lambda} \right]^{\mathbf{T}} \quad (8)$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (17 \le k \le 16+\lambda) \end{cases} \quad (9)$$

where f_k $(k = 1,2,\cdots,16)$ are the same as those in Equation 4, θ_j $(j = 1,2,\cdots,\lambda)$ are given by Equation 5, λ is the number of the total counted ranks (or tiers) of the correlations along a DNA sequence and w is the weight factor. The concrete values for λ and w will be discussed further. Thus, instead of a 16-D (dimensional) vector (cf. Equation 4), the DNA sequence is now formulated by a $(16+\lambda)$−D vector as shown in Equation 8. It is through the additional λ correlation factors (Figure 2) that not only considerable global sequence-order effects can be incorporated but the DNA sequences with extreme difference in length can also be converted into a set of feature vectors with a same dimension. The latter is an important pre-requisite for formulating the statistical samples because many powerful classification engines, such as Covariant Discriminant (34,35), Support Vector Machine (SVM) (36) and K-Nearest Neighbor (37–39) algorithms, require the input to be a set of digital vectors with a fixed number of components.

SVM

SVM is an effective method for supervised pattern recognition and has been widely used in the realm of bioinformatics [see, e.g. (14,40–45)]. The basic idea of SVM is to transform the data into a high dimensional feature space and then determine the optimal separating hyperplane. A brief introduction about the formulation of SVM was given in (46). In this study, the SVM implementation was based on the freely available package LIBSVM 2.84 written by Chang and Lin (47). Because of its effectiveness and speed in training process, the radial basis kernel function was used to obtain the best classification hyperplane. The

Table 1. The original numerical values for the six DNA dinucleotide physical structures

Dinucleotide	Physical structures[a]					
	$P_1(R_iR_{i+1})$	$P_2(R_iR_{i+1})$	$P_3(R_iR_{i+1})$	$P_4(R_iR_{i+1})$	$P_5(R_iR_{i+1})$	$P_6(R_iR_{i+1})$
AA	0.026	0.038	0.020	1.69	2.26	7.65
AC	0.036	0.038	0.023	1.32	3.03	8.93
AG	0.031	0.037	0.019	1.46	2.03	7.08
AT	0.033	0.036	0.022	1.03	3.83	9.07
CA	0.016	0.025	0.017	1.07	1.78	6.38
CC	0.026	0.042	0.019	1.43	1.65	8.04
CG	0.014	0.026	0.016	1.08	2.00	6.23
CT	0.031	0.037	0.019	1.46	2.03	7.08
GA	0.025	0.038	0.020	1.32	1.93	8.56
GC	0.025	0.036	0.026	1.20	2.61	9.53
GG	0.026	0.042	0.019	1.43	1.65	8.04
GT	0.036	0.038	0.023	1.32	3.03	8.93
TA	0.017	0.018	0.016	0.72	1.20	6.23
TC	0.025	0.038	0.020	1.32	1.93	8.56
TG	0.016	0.025	0.017	1.07	1.78	6.38
TT	0.026	0.038	0.020	1.69	2.26	7.65

[a]In this table, the following symbols were used to represent the six physical structures of dinucleotide (32): P_1 for 'twist, P_2 for 'tilt', P_3 for 'roll', P_4 for 'shift', P_5 for 'slide' and P_6 for 'rise'.

Table 2. The normalized values for the six DNA dinucleotide physical structures

Dinucleotide	Physical structures[a]					
	$P_1(R_iR_{i+1})$	$P_2(R_iR_{i+1})$	$P_3(R_iR_{i+1})$	$P_4(R_iR_{i+1})$	$P_5(R_iR_{i+1})$	$P_6(R_iR_{i+1})$
AA	0.06	0.5	0.27	1.59	0.11	−0.11
AC	1.50	0.50	0.80	0.13	1.29	1.04
AG	0.78	0.36	0.09	0.68	−0.24	−0.62
AT	1.07	0.22	0.62	−1.02	2.51	1.17
CA	−1.38	−1.36	−0.27	−0.86	−0.62	−1.25
CC	0.06	1.08	0.09	0.56	−0.82	0.24
CG	−1.66	−1.22	−0.44	−0.82	−0.29	−1.39
CT	0.78	0.36	0.09	0.68	−0.24	−0.62
GA	−0.08	0.5	0.27	0.13	−0.39	0.71
GC	−0.08	0.22	1.33	−0.35	0.65	1.59
GG	0.06	1.08	0.09	0.56	−0.82	0.24
GT	1.50	0.50	0.80	0.13	1.29	1.04
TA	−1.23	−2.37	−0.44	−2.24	−1.51	−1.39
TC	−0.08	0.5	0.27	0.13	−0.39	0.71
TG	−1.38	−1.36	−0.27	−0.86	−0.62	−1.25
TT	0.06	0.5	0.27	1.59	0.11	−0.11

[a]See footnote a of Table 1 for further explanation.

regularization parameter C and the kernel width parameter γ were determined via an optimization procedure using a grid search approach, and their actual values thus obtained for the current study were $C = 32$ and $\gamma = 0.5$.

iRSpot-PseDNC and its parameters

The predictor obtained via the aforementioned procedures is called iRSpot-PseDNC. The PseDNC as formulated in Equations 8 and 9 contains two uncertain parameters λ and w. The former represents the total number of correlation ranks counted (cf. Equation 5 and Figure 2), which is an integer and should be smaller than the length of any of the DNA sequences involved in this study, whereas the latter is the weight factor ranged from 0 to 1 (26). Generally speaking, the greater the value of λ, the more sequence-order effects will be incorporated. However, if the value of λ is too large, it might cause the overfitting problem (48) or 'high dimension disaster' (49). Preliminary tests indicated that in using the iRSpot-PseDNC predictor on the benchmark data set $\mathbb{S}$ (Supplementary Information S1), a peak was observed for the overall accuracy Λ (cf. Equation 11) or Acc (cf. Equation 12) when $\lambda = 3$ and $w = 0.05$. Accordingly, the two numerical values were respectively used for the two uncertain parameters in iRSpot-PseDNC.

RESULTS AND DISCUSSIONS

One of the important procedures in developing a useful statistical predictor (23) is to objectively evaluate its

performance or anticipated success rate. Now let us address this problem.

Criteria for performance evaluation

To provide a more intuitive and easier-to-understand method to measure the prediction quality, the criteria proposed in (50) was adopted here. According to that criteria, the rates of correct predictions for the recombination hotspots in data set $\mathbb{S}^+$ and the recombination coldspots in data set $\mathbb{S}^-$ are respectively defined by (cf. Equation 1)

$$\begin{cases} \Lambda^+ = \frac{N^+ - N_-^+}{N^+}, \text{ for the recombination hotspots} \\ \Lambda^- = \frac{N^- - N_+^-}{N^-}, \text{ for the recombination coldspots} \end{cases} \quad (10)$$

where N^+ is the total number of the recombination hotspots investigated, whereas N_-^+ the number of the recombination hotspots incorrectly predicted as the coldspots; N^- the total number of the recombination coldspots investigated, whereas N_+^- the number of the recombination coldspots incorrectly predicted as the hotspots. The overall success prediction rate is given by (51)

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \quad (11)$$

It is obvious from Equations 10 and 11 that, if and only if none of the recombination hotspots and the recombination coldspots are mispredicted, i.e. $N_-^+ = N_+^- = 0$ and $\Lambda^+ = \Lambda^- = 1$, we have the overall success rate $\Lambda = 1$. Otherwise, the overall success rate would be <1.

On the other hand, it is instructive to point out that the following equation set is often used in literatures for examining the performance quality of a predictor

$$\begin{cases} \text{Sn} = \frac{\text{TP}}{\text{TP+FN}} \\ \text{Sp} = \frac{\text{TN}}{\text{TN+FP}} \\ \text{Acc} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \\ \text{MCC} = \frac{(\text{TP}\times\text{TN}) - (\text{FP}\times\text{FN})}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \end{cases} \quad (12)$$

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.

The relations between the symbols in Equation 11 and those in Equation 12 are given by

$$\begin{cases} \text{TP} = N^+ - N_-^+ \\ \text{TN} = N^- - N_+^- \\ \text{FP} = N_+^- \\ \text{FN} = N_-^+ \end{cases} \quad (13)$$

Substituting Equation 13 into Equation 12 and also noting Equation 11, we obtain

$$\begin{cases} \text{Sn} = 1 - \frac{N_-^+}{N^+} \\ \text{Sp} = 1 - \frac{N_+^-}{N^-} \\ \text{Acc} = \Lambda = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \\ \text{MCC} = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N^+}{N^+}\right)\left(1 + \frac{N_-^+ - N^-}{N^-}\right)}} \end{cases} \quad (14)$$

Obviously, when $N_-^+ = 0$, meaning none of the recombination hotspots was mispredicted to be a coldspots, we have the sensitivity $\text{Sn} = 1$, whereas $N_-^+ = N^+$, meaning that all the recombination hotspots were mispredicted to be the coldspots, we have the sensitivity $\text{Sn} = 0$. Likewise, when $N_+^- = 0$, meaning none of the recombination coldspots was mispredicted, we have the specificity $\text{Sp} = 1$, whereas $N_+^- = N^-$ meaning all the recombination coldspots were incorrectly predicted as recombination hotspots, we have the specificity $\text{Sp} = 0$. When $N_-^+ = N_+^- = 0$, meaning that none of the recombination hotspots in the data set S^+ and none of the recombination coldspots in S^- was incorrectly predicted, we have the overall accuracy $\text{Acc} = \Lambda = 1$, whereas $N_-^+ = N^+$ and $N_+^- = N^-$, meaning that all the recombination hotspots in the data set $S^+$and all the recombination coldspots in S^- were mispredicted, we have the overall accuracy $\text{Acc} = \Lambda = 0$. The MCC correlation coefficient is usually used for measuring the quality of binary (two-class) classifications. When $N_-^+ = N_+^- = 0$, meaning that none of the recombination hotspots in the data set $S^+$and none of the recombination coldspots in S^- was mispredicted, we have $\text{Mcc} = 1$; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have $\text{Mcc} = 0$, meaning no better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$ we have $\text{MCC} = -1$, meaning total disagreement between prediction and observation. As we can see from the aforementioned discussion, it is much more intuitive and easier to understand when using Equation 14 to examine a predictor for its sensitivity, specificity, overall accuracy and Mathew's correlation coefficient.

Cross-validation

In literatures, the following three cross-validation methods are often used to evaluate the quality of a predictor: independent data set test, subsampling (K-fold cross-validation) test and jackknife test. However, as elaborated by an analysis in (52) and demonstrated by Equations 28–32 of (23), among the three cross-validation methods, the jackknife test is deemed the least arbitrary and most objective because it can always yield a unique result for a given benchmark data set, and hence has been widely recognized and increasingly used by investigators to examine the quality of various predictors [see, e.g. (11,16,21,22,29,53–57)]. Accordingly, the jackknife test was also adopted in this study to examine the anticipated success rates of the current predictor. In the jackknife test, all the samples in the benchmark data set $\mathbb{S}$ will be singled out one by one and tested by the predictor trained by the remaining samples. During the jackknifing process, both

the training data set and testing data set are actually open, and each sample will be in turn moved between the two.

The results obtained with iRSpot-PseDNC on the benchmark data set $\mathbb{S}$ of Supplementary Information S1 by the jackknife test are given in Table 3, where for facilitating comparison, the corresponding results by the IDQD predictor (6) on the same benchmark data set are also given. As indicated in Table 3, the results reported by Liu *et al.* (6) were derived by the 5-fold cross-validation test. As elucidated in (23), this would make their test without a unique result as demonstrated later in the text. For the current case, the benchmark data set $\mathbb{S}$ consists of $\mathbb{S}^+$ and $\mathbb{S}^-$, where $\mathbb{S}^+$ contains 490 recombination hotspots, and $\mathbb{S}^-$ contains 591 recombination coldspots. Substituting these data into Equations 28 and 29 of (23) with $M = 2$ (number of groups for classification) and $\Gamma = 5$ (number of folds for cross-validation), we obtain

$$\Pi = \frac{490!}{[490 - \mathrm{Int}(490/5)]!\mathrm{Int}(490/5)!} \cdot \frac{591!}{[591 - \mathrm{Int}(591/5)]!\mathrm{Int}(591/5)!} = \frac{490!}{(490-98)!98!} \cdot \frac{591!}{(591-118)!118!} > 1.17 \times 10^{232} \quad (15)$$

where the symbol Int is the integer-truncating operator meaning to take the integer part for the number in the bracket right after it. The result of Equation 15 indicates that the number of possible combinations of taking one-fifth samples from each of the two subsets, $\mathbb{S}^+$ and $\mathbb{S}^-$, for conducting the 5-fold cross-validation will be $>10^{232}$, which is an astronomical figure, too large to be practical. Therefore, in their study (6), Liu *et al.* only randomly picked one of $\sim 1.17 \times 10^{232}$ possible combinations (cf. Equation 15) to perform the 5-fold cross-validation. To make the comparison between iRspot-PseNDC and IDQD (6) with the same test method, we also randomly picked one of the possible combinations from the same benchmark data set to perform the 5-fold cross-validation test with iRspot-PseNDC, and the corresponding results thus obtained are given in Table 3 as well.

As we can see from the table, not only the overall accuracy (Acc) achieved by iRSpot-PseDNC using the 5-fold cross validation test is remarkably higher than that by the IDQD (6) but the overall accuracy achieved by iRSpot-PseDNC using the rigorous jackknife test is also higher than that by the IDQD. Besides the overall accuracy, the MCC rates achieved by the iRSpot-PseDNC predictor derived from both 5-fold cross-validation and jackknife tests are also higher than those by the IDQD predictor.

To further demonstrate its performance, we used iRSpot-PseDNC to identify the 452 experimentally annotated recombination hotspots by Pan *et al.* (58) for the *S. cerevisiae* chromosome IV. The results are given in the Supplementary Information S2, from which we can see that 347 outcomes by the predictor were consistent with the experimental observations. The overall success rate was 76.77%, indicating that the method as proposed in this article is promising in identifying recombination hot/cold spots, or can at the very least play a complementary role to the existing method in this area.

Web-server guide

For the convenience of the vast majority of experimental scientists, let us give a step-by-step guide on how to use the iRSpot-PseDNC 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

Open the web server at http://lin.uestc.edu.cn/server/iRSpot-PseDNC and you will see the top page of iRSpot-PseDNC on your computer screen, as shown in Figure 3. Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it.

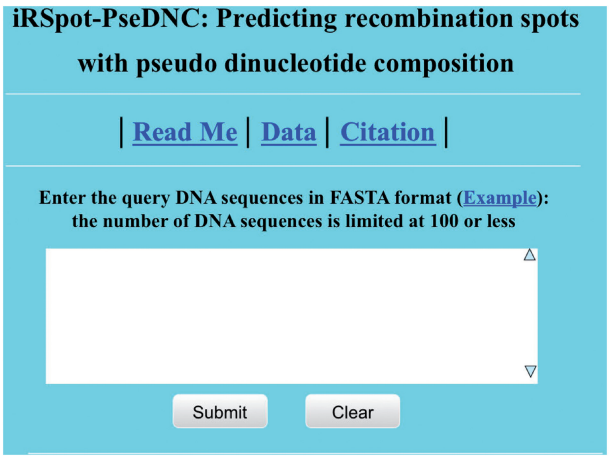

Figure 3. A semi-screenshot to show the top page of the iRSpot-PseDNC web-server. Its website address is at http://lin.uestc.edu.cn/server/iRSpot-PseDNC.

Table 3. A comparison of between iRSpot-PseDNC with the existing method

Predictor	Test method	Sn (%)	Sp (%)	Acc (%)	MCC
iRSpot-PseDNC[a]	Jackknife	73.06	89.49	82.04	0.638
	5-fold cross	81.63	88.14	85.19	0.692
IDQD[b]	5-fold cross	79.40	81.00	80.30	0.603

[a]The parameters used: $\lambda = 3$ and $w = 0.05$ for Equation 9; $C = 32$ and $\gamma = 0.5$ for the LIBSVM operation engine (47).
[b]From Liu *et al.* (6).

Step 2

Either type or copy/paste the query DNA sequence into the input box at the center of Figure 3. 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. All lines should not be longer than 120 characters and usually do not exceed 80 characters. 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

Click on the Submit button to see the predicted result. For example, if you use the query DNA sequences in the Example window as the input, after clicking the Submit button, you will see the following shown on the screen of your computer: the outcome for the first query sample is 'recombination hotspot'; the outcome for the second query sample is 'recombination coldspot'. All these results are fully consistent with the experimental observations as summarized in the Supplementary Information S1. It takes a few seconds for the aforementioned 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.

Step 4

Click on the Citation button to find the relevant papers that document the detailed development and algorithm of iRSpot-PseDNC.

Step 5

Click on the Data button to download the benchmark data sets used to train and test the iRSpot-PseDNC predictor.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Data sets 1 and 2.

ACKNOWLEDGEMENTS

The authors wish to thank the two anonymous reviewers for their constructive comments, which were indeed very helpful for strengthening the presentation of this study.

FUNDING

Funding for open access charge: The National Nature Scientific Foundation of China [61100092, 61202256].

Conflict of interest statement. None declared.

REFERENCES

1. Lewin,B. (2008) *Genes IX*, Chap.18. Jones & Bartlett, Massachusetts, pp. 428–456.
2. Keeney,S. (2008) Spo11 and the Formation of DNA Double-Strand Breaks in Meiosis. *Genome Dyn. Stab.*, **2**, 81–123.
3. Baudat,F. and Nicolas,A. (1997) Clustering of meiotic double-strand breaks on yeast chromosome III. *Proc. Natl Acad. Sci. USA*, **94**, 5213–5218.
4. Klein,S., Zenvirth,D., Dror,V., Barton,A.B., Kaback,D.B. and Simchen,G. (1996) Patterns of meiotic double-strand breakage on native and artificial yeast chromosomes. *Chromosoma*, **105**, 276–284.
5. Zenvirth,D., Arbel,T., Sherman,A., Goldway,M., Klein,S. and Simchen,G. (1992) Multiple sites for double-strand breaks in whole meiotic chromosomes of Saccharomyces cerevisiae. *EMBO J.*, **11**, 3441–3447.
6. Liu,G., Liu,J., Cui,X. and Cai,L. (2012) Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae. *J. Theor. Biol.*, **293**, 49–54.
7. Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins*, **43**, 246–255.
8. 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.
9. 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.
10. Fang,Y., Guo,Y., Feng,Y. and Li,M. (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. *Amino Acids*, **34**, 103–109.
11. Nanni,L., Lumini,A., Gupta,D. and Garg,A. (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 Trans. Comput. Biol. Bioinform.*, **9**, 467–475.
12. Mohammad Beigi,M., Behjati,M. and Mohabatkar,H. (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.
13. Guo,J., Rao,N., Liu,G., Yang,Y. and Wang,G. (2011) Predicting protein folding rates using the concept of Chou's pseudo amino acid composition. *J. Comput. Chem.*, **32**, 1612–1617.
14. Mohabatkar,H., Mohammad Beigi,M. and Esmaeili,A. (2011) Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.*, **281**, 18–23.
15. Zou,D., He,Z., He,J. and Xia,Y. (2011) Supersecondary structure prediction using Chou's pseudo amino acid composition. *J. Comput. Chem.*, **32**, 271–278.
16. Mohabatkar,H. (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett.*, **17**, 1207–1214.
17. Georgiou,D.N., Karakasidis,T.E., Nieto,J.J. and Torres,A. (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.*, **257**, 17–26.
18. Zhou,X.B., Chen,C., Li,Z.C. and Zou,X.Y. (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.*, **248**, 546–551.
19. Esmaeili,M., Mohabatkar,H. and Mohsenzadeh,S. (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.*, **263**, 203–209.
20. Mohabatkar,H., Beigi,M.M., Abdolahi,K. and Mohsenzadeh,S. (2012) Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.*
21. Zia Ur,R. and Khan,A. (2012) Identifying GPCRs and their types with Chou's Pseudo amino acid composition: an approach from

multi-scale energy representation and position specific scoring matrix. *Protein Pept. Lett.*, **19**, 890–903.
22. Hayat,M. and Khan,A. (2012) Discriminating outer membrane proteins with fuzzy K-Nearest neighbor algorithms based on the general form of Chou's PseAAC. *Protein Pept. Lett.*, **19**, 411–421.
23. 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.
24. Du,P., Wang,X., Xu,C. and Gao,Y. (2012) PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.*, **425**, 117–119.
25. Shen,H.B. and Chou,K.C. (2008) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.*, **373**, 386–388.
26. Chou,K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21**, 10–19.
27. Lin,W.Z., Fang,J.A., Xiao,X. and Chou,K.C. (2011) iDNA-Prot: identification of DNA binding proteins using random forest with grey model. *PLoS One*, **6**, e24756.
28. Xiao,X., Wu,Z.C. and Chou,K.C. (2011) iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.*, **284**, 42–51.
29. Chou,K.C., Wu,Z.C. and Xiao,X. (2012) iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Molecular Biosystems*, **8**, 629–641.
30. Abeel,T., Saeys,Y., Bonnet,E., Rouze,P. and Van de Peer,Y. (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome. Res.*, **18**, 310–323.
31. Goni,J.R., Fenollosa,C., Perez,A., Torrents,D. and Orozco,M. (2008) DNAlive: a tool for the physical analysis of DNA at the genomic scale. *Bioinformatics*, **24**, 1731–1732.
32. Goni,J.R., Perez,A., Torrents,D. and Orozco,M. (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.*, **8**, R263.
33. Miele,V., Vaillant,C., d'Aubenton-Carafa,Y., Thermes,C. and Grange,T. (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.*, **36**, 3746–3756.
34. Chou,K.C., Liu,W., Maggiora,G.M. and Zhang,C.T. (1998) Prediction and classification of domain structural classes. *Proteins*, **31**, 97–103.
35. Chou,K.C. and Cai,Y.D. (2005) Prediction of membrane protein types by incorporating amphipathic effects. *J. Chem. Inf. Model.*, **45**, 407–413.
36. Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer Verlag, New York.
37. Denoeux,T. (1995) A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.*, **25**, 804–813.
38. Shen,H.B. and Chou,K.C. (2007) Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. *Biopolymers*, **85**, 233–240.
39. Chou,K.C. and Shen,H.B. (2007) Review: recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **370**, 1–16.
40. Chen,W. and Lin,H. (2010) Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information. *Biochem. Biophys. Res. Commun.*, **401**, 382–384.
41. Cai,Y.D., Zhou,G.P. and Chou,K.C. (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.*, **84**, 3257–3263.
42. Lin,H. and Ding,H. (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. *J. Theor. Biol.*, **269**, 64–69.
43. Zhu,L., Yang,J. and Shen,H.B. (2009) Multi label learning for prediction of human protein subcellular localizations. *Protein J.*, **28**, 384–390.
44. Chen,W., Feng,P. and Lin,H. (2012) Prediction of replication origins by calculating DNA structural properties. *FEBS Lett.*, **586**, 934–938.
45. Chen,C., Shen,Z.B. and Zou,X.Y. (2012) Dual-Layer wavelet SVM for predicting protein structural class via the general form of Chou's pseudo amino acid composition. *Protein Pept. Lett.*, **19**, 422–429.
46. 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.
47. Chang,C.C. and Lin,C.J. (2001) LIBSVM: a library for support vector machines, Software. http://www.csie.ntu.edu.tw/_cjlin/libsvm.
48. Chou,K.C. (1999) A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun.*, **264**, 216–224.
49. Wang,T., Yang,J., Shen,H.B. and Chou,K.C. (2008) Predicting membrane protein types by the LLDA algorithm. *Protein Pept. Lett.*, **15**, 915–921.
50. Chou,K.C. (2001) Using subsite coupling to predict signal peptides. *Protein Eng.*, **14**, 75–79.
51. Chou,K.C. (2001) Prediction of signal peptides using scaled window. *Peptides*, **22**, 1973–1979.
52. Chou,K.C. and Shen,H.B. (2010) Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Natural Science*, **2**, 1090–1103.
53. Hayat,M. and Khan,A. (2012) MemHyb: Predicting membrane protein types by hybridizing SAAC and PSSM. *J. Theor. Biol.*, **292**, 93–102.
54. Sun,X.Y., Shi,S.P., Qiu,J.D., Suo,S.B., Huang,S.Y. and Liang,R.P. (2012) Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol. Biosyst.*, **8**, 3178–3184.
55. Hayat,M. and Khan,A. (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. *J. Theor. Biol.*, **271**, 10–17.
56. 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.
57. Chou,K.C., Wu,Z.C. and Xiao,X. (2011) iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One*, **6**, e18258.
58. Pan,J., Sasaki,M., Kniewel,R., Murakami,H., Blitzblau,H.G., Tischfield,S.E., Zhu,X., Neale,M.J., Jasin,M., Socci,N.D. *et al.* (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, **144**, 719–731.