Int. J. Mol. Sci. **2014**, *15*, 1746-1766; doi:10.3390/ijms15021746

OPEN ACCESS
International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

iRSpot-TNCPseAAC: Identify Recombination Spots with Trinucleotide Composition and Pseudo Amino Acid Components

Wang-Ren Qiu [1], Xuan Xiao [1,2,4,*] and Kuo-Chen Chou [3,4]

[1] Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen 333046, China; E-Mail: qiuone@163.com
[2] Information School, ZheJiang Textile & Fashion College, Ningbo 315211, China
[3] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia; E-Mail: kcchou@gordonlifescience.org
[4] Gordon Life Science Institute, Belmont, MA 02478, USA

***** Author to whom correspondence should be addressed; E-Mail: xxiao@gordonlifescience.org or jdzxiaoxuan@163.com; Tel.: +86-138-7980-9729; Fax: +86-798-873-2324.

Received: 2 January 2014; in revised form: 14 January 2014 / Accepted: 16 January 2014 / Published: 24 January 2014

Abstract: Meiosis and recombination are the two opposite aspects that coexist in a DNA system. As a driving force for evolution by generating natural genetic variations, meiotic recombination plays a very important role in the formation of eggs and sperm. Interestingly, the recombination does not occur randomly across a genome, but with higher probability in some genomic regions called "hotspots", while with lower probability in so-called "coldspots". With the ever-increasing amount of genome sequence data in the postgenomic era, computational methods for effectively identifying the hotspots and coldspots have become urgent as they can timely provide us with useful insights into the mechanism of meiotic recombination and the process of genome evolution as well. To meet the need, we developed a new predictor called "iRSpot-TNCPseAAC", in which a DNA sample was formulated by combining its trinucleotide composition (TNC) and the pseudo amino acid components (PseAAC) of the protein translated from the DNA sample according to its genetic codes. The former was used to incorporate its local or short-rage sequence order information; while the latter, its global and long-range one. Compared with the best existing predictor in this area, iRSpot-TNCPseAAC achieved higher rates in accuracy, Mathew's correlation coefficient, and sensitivity, indicating that the new predictor may become a useful tool for identifying the recombination hotspots and coldspots, or, at least, become a complementary tool to the existing methods. It has not

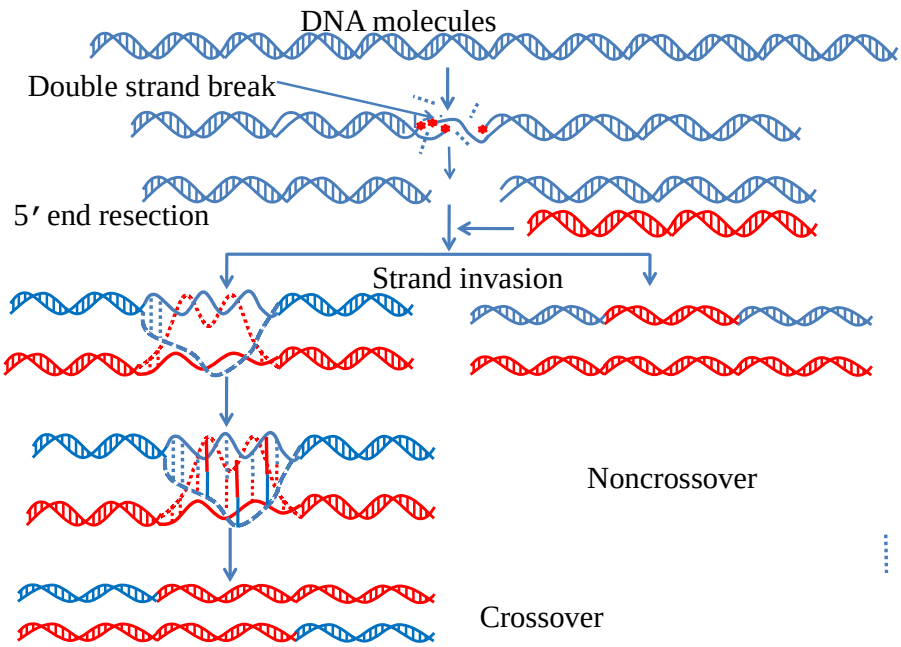
escaped our notice that the aforementioned novel approach to incorporate the DNA sequence order information into a discrete model may also be used for many other genome analysis problems. The web-server for iRSpot-TNCPseAAC is available at http://www.jci-bioinfo.cn/iRSpot-TNCPseAAC. Furthermore, for the convenience of the vast majority of experimental scientists, a step-by-step guide is provided on how to use the current web server to obtain their desired result without the need to follow the complicated mathematical equations.

Keywords: genome; DNA; recombination spots; hotspots; coldspots; trinucleotide composition; pseudo amino acid composition; web-server; iRSpot-TNCPseAAC

1. Introduction

Meiosis and recombination are two indispensible aspects for cell reproduction and growth (Figure 1). The former is a special type of cell division by which the genome is divided in half to generate daughter cells for participating in sexual reproduction, while the latter is to produce single-strand ends that can invade the homologous chromosome [1].

Figure 1. An illustration to show the process of meiosis and recombination in a DNA system. Adapted from [2].



Recombination is initiated by double-strand breaks (or broken DNA ends); defecting in meiosis may lead to male infertility [3–5]. Meiotic recombination ensures accurate chromosome segregation during the first meiotic division and provides a mechanism to increase genetic heterogeneity among the meiotic products. Accordingly, identification of recombination spots may provide very useful information for in-depth understanding the reproduction and growth of cells.

In the past decades, a lot of global mapping studies have been performed to map double-strand break sites on chromosomes [6–13]. The following findings were observed through these studies for the meiotic recombination events. (i) They generally concentrate in 1:2.5 kilobase regions; (ii) They do not occur randomly across the entire genome but with a higher rate in some regions and lower in others; the former is a so-called "hotspot" while the latter, "coldspot"; (iii) They do not share a consensus sequence pattern.

With the rapid increasing number of genome sequences, it is important to address the following problem. Given a genome sequence, how can we predict which part of it is the hotspot for recombination, and which part is not?

Based on the nucleotide sequence contents, Liu *et al*. [14] proposed a computational method to deal with this problem. However, in their method no sequence-order effect whatsoever was taken into account, and, hence, its prediction power might be limited.

Actually, one of the most important, but also most difficult, problems in computational biology is how to formulate a biological sequence with a discrete model or a vector, yet still keep considerable sequence order information. This is as all the existing operation engines, such as covariance discriminant (CD) [15–20], neural network [21–23], support vector machine (SVM) [24–26], random forest [27,28], conditional random field [29], nearest neighbor (NN) [30,31], K-nearest neighbor (KNN) [32–34], OET-KNN (optimized evidence-theoretic k-nearest neighbors) [35–38], and Fuzzy K-nearest neighbor [39–43], can only handle vector, but not sequence, samples. However, a vector defined in a discrete model may completely lose all the sequence-order information.

To avoid completely losing the sequence-order information for proteins, the pseudo amino acid composition [44,45] or Chou's pseudo amino acid components (PseAAC) [46] was proposed. Ever since the concept of PseAAC was proposed in 2001 [44], it has penetrated into almost all the areas of computational proteomics, such as identifying cysteine S-nitrosylation sites in proteins [29], predicting bacterial virulent proteins [47], predicting antibacterial peptides [48], identifying bacterial secreted proteins [49], predicting supersecondary structure [50], predicting protein subcellular location [51–59], predicting membrane protein types [60,61], discriminating outer membrane proteins [62], identifying antibacterial peptides [48], identifying allergenic proteins [63], predicting metalloproteinase family [64], predicting protein structural class [65], identifying GPCRs (G protein-coupled receptors) and their types [66,67], identifying protein quaternary structural attributes [68,69], predicting protein submitochondria locations [70–73], identifying risk type of human papillomaviruses [74], identifying cyclin proteins [75], predicting GABA(A) receptor proteins [76], classifying amino acids [77], predicting the cofactors of oxidoreductases [78], predicting enzyme subfamily classes [79], detecting remote homologous proteins [80], analyzing genetic sequences [81], predicting anticancer peptides [82], among many others (see a long list of papers cited in the References section of [83]). Recently, the concept of PseAAC was further extended to represent the feature vectors of nucleotides [15], as well as other biological samples [84–86]. As it has been widely and increasingly used, recently two powerful soft-wares, called "PseAAC-Builder" [87] and "propy" [88], were established for generating various special Chou's pseudo-amino acid compositions, in addition to the web-server "PseAAC" [89], built in 2008.

Encouraged by the success of introducing PseAAC for proteins, recently, Chen *et al*. [25] proposed the pseudo dinucleotide composition or PseDNC to represent DNA sequences for identifying the

recombination spots by counting some sequence effects, remarkably improving the prediction results in comparison with those by Liu *et al*. [14], without including any sequence information. However, in PseDNC, only the correlations of dinucleotides along a DNA sequence were considered, and, hence, some important sequence order effects might be missed.

The present study was initiated in an attempt to incorporate the long-range or global correlations of trinucleotides along a DNA sequences in hope to further improve the prediction quality in indentifying the recombination spots.

As demonstrated in a series of recent publications [24,42,90–92] and summarized in a comprehensive review [83], to establish a really useful statistical predictor for a biological system, one needs 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; 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.

2. Results and Discussion

2.1. Benchmark Dataset

The benchmark dataset S used in this study was taken from Liu *et al*. [14], which contains 490 recombination hotspots and 591 recombination coldspots, as can be formulated by:

$$S = S^{+} \bigcup S^{-} \tag{1}$$

where subset S^{+} and S^{-} are respectively for the hot and cold spots, while $\bigcup$ represents the symbol for "union" in the set theory. For reader's convenience, the 490 DNA sequences in S^{+} and 591 sequences in S^{-} are given in the Supplementary Information S1.

2.2. Formulate DNA Samples by Combining Trinucleotide Composition and Pseudo Amino Acid Components

Suppose a DNA sequence **D** with L nucleotides; *i.e*.,

$$\mathbf{D} = N_1 N_2 N_3 N_4 N_5 N_6 N_7 \cdots N_L \tag{2}$$

where

$$N_i \in \{\text{A (adenine)}, \quad \text{C (cytosine)} \quad \text{G (guanine)} \quad \text{T (thymine)}\} \tag{3}$$

denotes the i-th (i = 1, 2, …, L) nucleotide in the DNA sequence. If the feature vector of the DNA sequence is formulated by its mononucleotide composition (MNC), we have:

$$\begin{aligned} \mathbf{D} &= \begin{bmatrix} f(\text{A}) & f(\text{C}) & f(\text{G}) & f(\text{T}) \end{bmatrix}^{\mathbf{T}} \\ &= \begin{bmatrix} f_1^{(1)} & f_2^{(1)} & f_3^{(1)} & f_4^{(1)} \end{bmatrix}^{\mathbf{T}} \end{aligned} \tag{4}$$

where $f_1^{(1)} = f(\mathrm{A})$, $f_2^{(1)} = f(\mathrm{C})$, $f_3^{(1)} = f(\mathrm{G})$, and $f_4^{(1)} = f(\mathrm{T})$ are the normalized occurrence frequencies of adenine (A), cytosine (C), guanine (G), and thymine (T), respectively, in the DNA sequence; and the symbol **T** is the transpose operator. As we can see from Equation (4), all the sequence order information is missed if using MNC to represent a DNA sequence. If using the dinucleotide composition (DNC) to represent the DNA sequence, instead of the four components as shown in Equation (4), the corresponding feature vector will contain 4 × 4 = 16 components, as given below:

$$\begin{aligned} \mathbf{D} &= \left[\begin{array}{cccccc} f(\mathrm{AA}) & f(\mathrm{AC}) & f(\mathrm{AG}) & f(\mathrm{AT}) & \cdots & f(\mathrm{TT}) \end{array} \right]^{\mathbf{T}} \\ &= \left[\begin{array}{cccccc} f_1^{(2)} & f_2^{(2)} & f_3^{(2)} & f_4^{(2)} & \cdots & f_{16}^{(2)} \end{array} \right]^{\mathbf{T}} \end{aligned} \tag{5}$$

where $f_1^{(2)} = f(\mathrm{AA})$ is the normalized occurrence frequency of AA in the DNA sequence; $f_2^{(2)} = f(\mathrm{AC})$, that of AC; $f_3^{(2)} = f(\mathrm{AG})$, that of AG; and so forth. If represented by the trinucleotide composition (TNC), the corresponding feature vector will contain $4 \times 4 \times 4 = 4^3 = 64$ components, as given below:

$$\begin{aligned} \mathbf{D} &= \left[\begin{array}{cccccc} f(\mathrm{AAA}) & f(\mathrm{AAC}) & f(\mathrm{AAG}) & f(\mathrm{AAT}) & \cdots & f(\mathrm{TTT}) \end{array} \right]^{\mathbf{T}} \\ &= \left[\begin{array}{cccccc} f_1^{(3)} & f_2^{(3)} & f_3^{(3)} & f_4^{(3)} & \cdots & f_{64}^{(3)} \end{array} \right]^{\mathbf{T}} \end{aligned} \tag{6}$$

where $f_1^{(3)} = f(\mathrm{AAA})$ is the normalized occurrence frequency of AAA in the DNA sequence; $f_2^{(3)} = f(\mathrm{AAC})$, that of AAC; and so forth. Generally speaking, if a DNA sequence is represented by the *K*-tuple nucleotide composition, the corresponding vector D for the DNA sequence will contain 4^K components; *i.e.*,

$$\mathbf{D} = \left[\begin{array}{cccccc} f_1^{(K)} & f_2^{(K)} & f_3^{(K)} & f_4^{(K)} & \cdots & f_{4^K}^{(K)} \end{array} \right]^{\mathbf{T}} \tag{7}$$

As we can see from Equations (5–7), with increasing the tuple number, although the base sequence-order information within a local or very short range could be gradually 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 exactly 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 by incorporating their local sequence order information [93–97], their global or long-range sequence order information still could not be reflected. As mentioned above, to deal with this kind of problems in proteomics, the concept of PseAAC [44,45] was introduced.

Stimulated by the PseAAC approach [44,45] in computational proteomics, below let us propose a novel feature vector to represent the DNA sequence (cf. Equation (2)) by combining its TNC (see Equation (2)) and the pseudo amino acid components of its translated protein chain.

As is well known, three nucleotides encode an amino acid (see Figure 2). Thus, according the conversion table from DNA codons to amino acids (Table 1), the DNA sequence in Equation (2) can be translated into a protein sequence expressed by:

$$\mathbf{P} = A_1 A_2 A_3 \cdots A_{L^*} \tag{8}$$

with

$$\begin{cases} A_i \in \{20 \text{ native amino acids}\} \\ L^* = \text{Int}\{L/3\} \end{cases} \tag{9}$$

where the symbol "Int" is an integer truncation operator meaning to take the integer part for the number in the brackets immediately after it.

Figure 2. A graph to show how a DNA codon of three nucleotides is converted to an amino acid. The characters in the first three rings from the center represent four bases in DNA, while those in the fourth ring represent the single-letter codes of the 20 native amino acids in protein. The symbol ***** means the "Stop" sign.

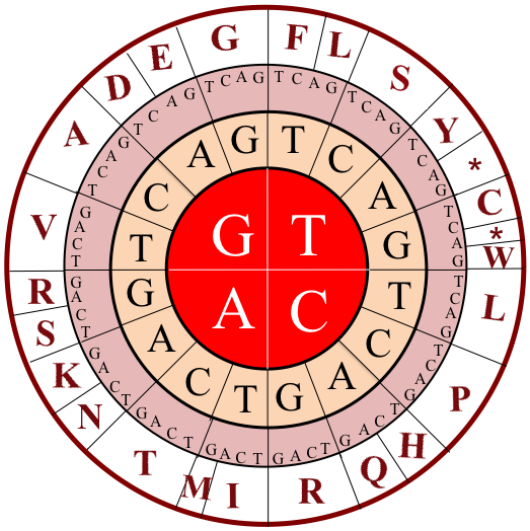


Table 1. The conversion code of the 64 trinucleotides in DNA to the 20 amino acids in protein.

Trinucleotide	Amino acid	Trinucleotide	Amino acid
AAA	Lys (K)	GAA	Glu (E)
AAC	Asn (N)	GAC	Asp (D)
AAG	Lys (K)	GAG	Glu (E)
AAT	Asn (N)	GAT	Asp (D)
ACA	Thr (T)	GCA	Ala (A)
ACC		GCC	
ACG		GCG	
ACT		GCT	
AGA	Arg (R)	GGA	Gly (G)
AGC	Ser (S)	GGC	
AGG	Arg (R)	GGG	
AGT	Ser (S)	GGT	

Table 1. *Cont.*

Trinucleotide	Amino acid	Trinucleotide	Amino acid
ATA	Ile (I)	GTA	Val (V)
ATC		GTC	
ATG	Met (M)	GTG	
ATT	Ile (I)	GTT	
CAA	Gln (Q)	TAA	Stop!
CAC	His (H)	TAC	Tyr (Y)
CAG	Gln (Q)	TAG	Stop!
CAT	His (H)	TAT	Tyr (Y)
CCA	Pro (P)	TCA	Ser (S)
CCC		TCC	
CCG		TCG	
CCT		TCT	
CGA	Arg (R)	TGA	Stop!
CGC		TGC	Cys (C)
CGG		TGG	Trp (W)
CGT		TGT	Cys (C)
CTA	Leu (L)	TTA	Leu (L)
CTC		TTC	Phe (F)
CTG		TTG	Leu (L)
CTT		TTT	Phe (F)

Now, according to the formulation of Chou's PseAAC approach [44,45], for the protein chain of Equation (8), we have:

$$\begin{cases} \theta_1 = \dfrac{1}{L^*-1}\sum_{i=1}^{L^*-1}\Theta(A_i, A_{i+1}) \\ \theta_2 = \dfrac{1}{L^*-2}\sum_{i=1}^{L^*-2}\Theta(A_i, A_{i+2}) \\ \theta_3 = \dfrac{1}{L^*-3}\sum_{i=1}^{L^*-3}\Theta(A_i, A_{i+3}) \\ \vdots \\ \theta_\lambda = \dfrac{1}{L^*-\lambda}\sum_{i=1}^{L^*-1}\Theta(A_i, A_{i+\lambda}) \end{cases} \quad (\lambda < L^*) \tag{10}$$

where θ_k $(k = 1, 2, 3, \cdots, \lambda)$ is called the *k*-th tier correlation factor that reflects the sequence order correlation between all the *k*-th most contiguous residues along a protein chain. In this study, the correlation function in Equation 10 is given by:

$$\Theta(A_i, A_j) = \frac{1}{6}\sum_{n=1}^{6}\left[H_n(A_j) - H_n(A_i)\right]^2 \tag{11}$$

where $H_n(A_j)$ $(n = 1, 2, \cdots, 6)$ is the six physicochemical properties of amino acid A_j; they are, respectively, hydrophobicity, hydrophilicity, side-chain mass, pK1 (α-COOH), pK2 (NH3), and PI. Note that before substituting these physicochemical values into Equation (11), they were all subjected to a standard conversion as described by the following equation:

$$H_n(A_i) = \frac{H_n^0(A_i) - \left\langle H_n^0 \right\rangle}{\mathrm{SD}(H_n^0)} \tag{12}$$

where $H_n(A_i)$ $(n = 1, 2, \cdots, 6)$ is the n-th original physicochemical property value for the amino acid A_i as given in Table 2, the symbol < and > means taking the average of the quantity therein over 20 native amino acids, and SD means the corresponding standard deviation. Listed in Table 3 are the converted values obtained by Equation (12) that will have a zero mean value over the 20 native amino acids, and will remain unchanged if going through the same conversion procedure again.

Table 2. List of the original values of the six physical-chemical properties for each of the 20 native amino acids.

Amino acid	**Hydro-phobicity** [a] H_1^0	**Hydro-philicity** [b] H_2^0	**Side-chain mass** [c] H_3^0	**pK1** [d] H_4^0	**pK2** [e] H_5^0	**PI** [f] H_6^0
A	0.62	−0.5	15	2.35	9.87	6.11
C	0.29	−1.00	47	1.71	10.78	5.02
D	−0.90	3.00	59	1.88	9.60	2.98
E	−0.74	3.00	73	2.19	9.67	3.08
F	1.19	−2.50	91	2.58	9.24	5.91
G	0.48	0.00	1	2.34	9.60	6.06
H	−0.40	−0.50	82	1.78	8.97	7.64
I	1.38	−1.80	57	2.32	9.76	6.04
K	−1.50	3.00	73	2.20	8.90	9.47
L	1.06	−1.80	57	2.36	9.60	6.04
M	0.64	−1.30	75	2.28	9.21	5.74
N	−0.78	0.20	58	2.18	9.09	10.76
P	0.12	0.00	42	1.99	10.60	6.30
Q	−0.85	0.20	72	2.17	9.13	5.65
R	−2.53	3.00	101	2.18	9.09	10.76
S	−0.18	0.30	31	2.21	9.15	5.68
T	−0.05	−0.40	45	2.15	9.12	5.60
V	1.08	−1.50	43	2.29	9.74	6.02
W	0.81	−3.40	130	2.38	9.39	5.88
Y	0.26	−2.30	107	2.20	9.11	5.63

[a] Taken from [98]; [b] Taken from [99]; [c] Taken from any biochemistry text book; [d] Taken from [100] for C^{α}-COOH; [e] Taken from [100] for NH_3; [f] Taken from [101].

Table 3. The corresponding values obtained by the standard conversion of Equation12 on the original values in Table 2.

Amino acid	H_1	H_2	H_3	H_4	H_5	H_6
A	0.62	−0.15	−1.55	0.78	0.77	−0.10
C	0.29	−0.41	−0.52	−2.27	2.57	−0.64
D	−0.90	1.67	−0.13	−1.46	0.24	−1.65
E	−0.74	1.67	0.33	0.01	0.37	−1.61
F	1.19	−1.19	0.91	1.87	−0.48	−0.20
G	0.48	0.11	−2.00	0.73	0.24	−0.13
H	−0.40	−0.15	0.62	−1.94	−1.01	0.65
I	1.38	−0.82	−0.19	0.63	0.55	−0.14
K	−1.50	1.67	0.33	0.06	−1.15	1.56
L	1.06	−0.82	−0.19	0.82	0.24	−0.14
M	0.64	−0.56	0.39	0.44	−0.54	−0.29
N	−0.78	0.22	−0.16	−0.03	−0.77	2.20
P	0.12	0.11	−0.68	−0.94	2.21	−0.01
Q	−0.85	0.22	0.29	−0.08	−0.69	−0.33
R	−2.53	1.67	1.23	−0.03	−0.77	2.20
S	−0.18	0.27	−1.03	0.11	−0.65	−0.32
T	−0.05	−0.10	−0.58	−0.18	−0.71	−0.36
V	1.08	−0.67	−0.65	0.49	0.51	−0.15
W	0.81	−1.65	2.17	0.92	−0.18	−0.22
Y	0.26	−1.08	1.43	0.06	−0.73	−0.34

By combining the λ correlation factors with the 64 components in TNC (see Equation (6)), the DNA sequence is formulated by:

$$\mathbf{D} = \begin{bmatrix} d_1 & d_2 & \cdots & d_{64} & d_{64+1} & \cdots & d_{64+\lambda} \end{bmatrix}^{\mathbf{T}} \tag{13}$$

where:

$$d_u = \begin{cases} \dfrac{f_u^{(3)}}{\sum_{i=1}^{64} f_i^{(3)} + w\sum_{k=1}^{\lambda} \theta_k}, & (1 \le u \le 64) \\ \dfrac{w\theta_{u-64}}{\sum_{i=1}^{64} f_i^{(3)} + w\sum_{k=1}^{\lambda} \theta_k}, & (64+1 \le u \le 64+\lambda) \end{cases} \tag{14}$$

where w is the weight factor which is determined by optimizing the outcome as will be mentioned later. The rationale of using Equation (13) to represent the DNA sequence is that the local or short-range sequence order effect can be directly reflected via the occurrence frequencies of its 64 trinucleotides, while the global or long-range sequence order effect can be indirectly reflected via the λ pseudo amino acid components of its translated protein chain. As three nucleotides encode an amino acid, the above approach is both quite rational and natural.

2.3. Use Support Vector Machine as an Operation Engine

Support vector machine (SVM) has been widely to make classification prediction (see, e.g., [24,102–105]. The basic idea of SVM is to transform the input 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 [103,106]. Here, the DNA samples as formulated by Equation (13) were used as inputs for the SVM. Its software was downloaded from the LIBSVM package [107,108], which provided a simple interface. Due to this advantages, the users can easily perform classification prediction by properly selecting the built-in parameters C and γ. In order to maximize the performance of the SVM algorithm, the two parameters in the RBF kernel were preliminarily optimized through a grid search strategy in this study. To obtain the optimized parameters, the search function "SVMcgForClass" was downloaded from http://www.matlabsky.com.

The predictor obtained via the aforementioned procedures is called iRSpot-TNCPseAAC, where "i" means "identify", "RSpot" means "Recombination Spots", while TNCPseAAC means a combination of "Tri-Nucleotide Composition" and "Pseudo Amino Acid Components."

To objectively evaluate the quality of a new predictor, one should use proper metrics [109] and rigorous cross-validation [83] to test it. Below, let us address these problems.

2.4. Four Different Metrics for Measuring the Prediction Quality

In literature, the following metrics are often used for examining the performance quality of a predictor:

$$\begin{cases} Sn = \dfrac{TP}{TP+FN} \\ Sp = \dfrac{TN}{TN+FP} \\ Acc = \dfrac{TP+TN}{TP+TN+FP+FN} \\ MCC = \dfrac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \tag{15}$$

where *TP* represents the number of the true positive; *TN*, the number of the true negative; *FP*, the number of the false positive; *FN*, the number of the false negative; *Sn*, the sensitivity; *Sp*, the specificity; Acc, the accuracy; MCC, the Mathew's correlation coefficient. To most biologists, however, the four metrics as formulated in Equation (15) are not quite intuitive and easier-to-understand, particularly for the Mathew's correlation coefficient. Here let us adopt the formulation proposed recently [25,29] based on the Chou's symbol and definition [110]; *i.e.*,

$$\begin{cases} Sn = 1-\dfrac{N_-^+}{N^+} \\ Sp = 1-\dfrac{N_+^-}{N^-} \\ Acc = 1-\dfrac{N_-^+ + N_+^-}{N^+ + N^-} \\ Mcc = \dfrac{1-\left(\dfrac{N_-^+ + N_+^-}{N^+ + N^-}\right)}{\sqrt{\left(1+\dfrac{N_+^- - N_-^+}{N^+}\right)\left(1+\dfrac{N_-^+ - N_+^-}{N^-}\right)}} \end{cases} \tag{16}$$

where N^+ is the total number of the hotspot samples investigated while N_-^+ the number of the hotspot samples incorrectly predicted as coldspots; N^- the total number of the coldspot samples investigated while N_+^- the number of the coldspot samples incorrectly predicted as the hotspots [111].

Now, it can be clearly seen from Equation (16) that when $N_-^+ = 0$ meaning none of the hotspots was incorrectly predicted to be a coldspot, we have the sensitivity $Sn = 1$. When $N_-^+ = N^+$ meaning that all the hotspots were incorrectly predicted to be the coldspots, we have the sensitivity $Sn = 0$. Likewise, when $N_+^- = 0$ meaning none of the coldspots was incorrectly predicted to be the hotspot, we have the specificity $Sp = 1$; whereas $N_+^- = N^-$ meaning all the coldspots were incorrectly predicted as the hotspots, we have the specificity $Sp = 0$. When $N_-^+ = N_+^- = 0$ meaning that none of hotspots in the positive dataset and none of the coldspots in the negative dataset was incorrectly predicted, we have the overall accuracy $Acc = 1$ and $MCC = 1$; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the hotspots in the positive dataset and all the coldspots in the negative dataset were incorrectly predicted, we have the overall accuracy $Acc = 0$ and $MCC = -1$; whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have $Acc = 0.5$ and $MCC = 0$ meaning no better than random guess. As we can see from the above discussion based on Equation (16), the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient have become much more intuitive and easier-to-understand.

It should be pointed out that the metrics as given in Equation (15) and Equation (16) are valid only for the single-label systems as in the current case. For the multi-label systems in which emergence has become increasingly frequent in cell's molecular systems [112–118] and biomedical systems [43,119], a completely different set of metrics as defined in [109] is needed.

2.5. Evaluate the Anticipated Success Rates by Jackknife Tests

The following three cross-validation methods are often used in statistical prediction to evaluate the anticipated accuracy of a predictor: independent dataset test, subsampling (*K*-fold cross-validation) test, and jackknife test [120]. However, as elucidated by a review article [83], among the three methods, the jackknife test is deemed the least arbitrary and most objective as it can always yield a unique outcome for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictor [48,60,63,65,69,76,121,122]. Accordingly, in this study we also used the results obtained by jackknife tests to optimizing the uncertain parameters and to compare with the other predictors in this area.

3. Experimental Section

The results obtained with iRSpot-TNCPseAAC on the benchmark dataset S of Supplementary Information S1 by the jackknife test are given in Table 4, where for facilitating comparison the corresponding results by the iRSpot-PseDNC [25] on the same benchmark dataset are also given.

Table 4. A comparison of iRSpot-TNCPseAAC with the best existing method.

Predictor	Test method	*Sn* (%)	*Sp* (%)	*Acc* (%)	MCC
iRSpot-PseDNC [a]	Jackknife	73.06	89.49	82.04	0.638
iRSpot-KNCPseAAC [b]	Jackknife	87.14	79.59	83.72	0.671

[a] From [25]; [b] This paper with $\lambda = 5$, $w = 1.1$, $C = 32$ and $\gamma = 0.5$ for the LIBSVM operation engine [107,108].

As we can clearly see from the table, the iRSpot-TNCPseAAC predictor is superior to iRSpot-PseDNC [25] in three of the four metrics as defined by Equation (16); *i.e.*, it can yield higher accuracy *Acc*, higher Mathew's correlation coefficient MCC, and higher sensitivity *Sn*. Therefore, it is anticipated that the new predictor will become a useful tool for identifying the recombination spots in DNA, or at the very least become a complementary tool to iRSpot-PseDNC, the best existing prediction method in this area.

4. Conclusions

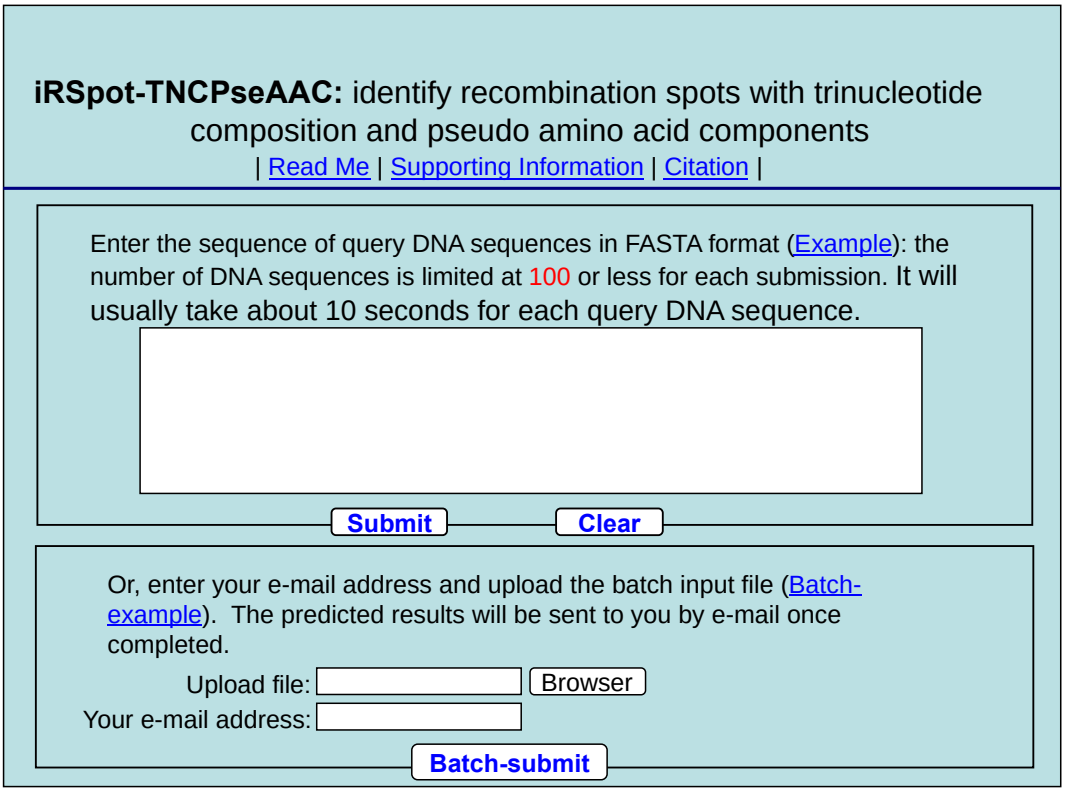
The above fact has also proved that it is indeed a feasible and promising approach to extend the concept of pseudo amino acid composition [44,45,123] developed in computational proteomics to the area of computational genomics. As shown by Equation (13) and the related equations in defining its $64+\lambda$ components, each of the DNA samples investigated in this study was formulated by a combination of its trinucleotide composition (TNC) with the pseudo amino acid components (PseAAC) that were derived from the protein translated from the DNA sample according to its genetic codes. The former can better incorporate its local or short-rage sequence order information in comparison with the dinucleotide composition (DNC) used in iRSpot-PseDNC [25]; while the latter can incorporate its global or long-range sequence order effects in a more natural or logical manner. Accordingly, it is anticipated that the idea or approach by extending the Chou's pseudo amino acid composition [44,45,123] for protein sequences to the pseudo oligonucleotide composition for DNA or RNA sequences may also be used to deal with many other genome analysis problems.

5. Web Server and User Guide

To enhance the value of its practical applications, a web-server for the iRSpot-TNCPseAAC predictor was established. Moreover, for the convenience of the vast majority of experimental scientists, here a step-to-step guide is provided for how to use the web server to get the desired results without the need to follow the mathematic equations that were presented just for the integrity in developing the predictor.

Step 1. Open the web server at http://www.jci-bioinfo.cn/iRSpot-TNCPseAAC and you will see the top page of the predictor on your computer screen, as shown in Figure 3. Click on the Read Me button to see a brief introduction about the **iRSpot-TNCPseAAC** predictor and the caveat when using it.

Figure 3. A semi-screenshot for the top page of the web-server iRSpot-TNCPseAAC at http://www.jci-bioinfo.cn/iRSpot-TNCPseAAC.



Step 2. Either type or copy/paste the query DNA sequences into the input box at the center of Figure 3. The input sequence should be in the FASTA format. For the examples of sequences in FASTA format, click 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 three query DNA sequences in the Example window as the input, after clicking the Submit button, you will see the following message shown on the screen of your computer: the outcome for the 1st query sample is "**recombination hotspot**"; the outcome for the 2nd query sample is "**recombination coldspot**". All these results are fully consistent with the experimental observations as summarized in the Supplementary Information S1. However, no result was given for the 3rd query sample as it contains some invalid characters as warned in the output screen. It takes about a 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.

Step 4. As shown on the lower panel of Figure 3, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format) via the ''**Browse**'' button. To see the sample of batch input file, click on the button Batch-example. After clicking the button Batch-submit, you will see "Your batch job is under computation; once the results are available, you will be notified by e-mail."

Step 5. Click the Supporting Information button to download the benchmark dataset used to train and test the **iRSpot-TNCPseAAC** predictor.

Step 6. Click the Citation button to find the relevant papers that document the detailed development and algorithm of **iRSpot-TNCPseAAC**.

Supplementary Information

Supplementary Information S1. The benchmark dataset S consists of a positive dataset S^+ and a negative dataset S^-. The positive dataset contains 490 recombination hot spots, while the negative dataset contains 591 recombination cold spots.

Acknowledgments

The authors wish to thank the two anonymous reviewers for their constructive suggestions, which were very helpful for strengthening the presentation of this paper. This work was partially supported by the National Nature Science Foundation of China (No. 31260273, 61261027), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No.20120BDH80023), Natural Science Foundation of Jiangxi Province, China (No.20114BAB211013, 20122BAB211033, 20122BAB201044, 20122BAB2010), the Department of Education of JiangXi Province (GJJ12490), the LuoDi plan of the Department of Education of JiangXi Province(KJLD12083), and the JiangXi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008). The funders had no role in the design of this study, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Hansen, L.; Kim, N.K.; Marino-Ramirez, L.; Landsman, D. Analysis of biological features associated with meiotic recombination hot and cold spots in *Saccharomyces cerevisiae*. *PLoS One* **2011**, *6*, e29711.
2. Keeney, S. Spo11 and the formation of DNA double-strand breaks in meiosis. *Genome Dyn. Stab.* **2008**, *2*, 81–123.
3. Ferguson, K.A.; Wong, E.C.; Chow, V.; Nigro, M.; Ma, S. Abnormal meiotic recombination in infertile men and its association with sperm aneuploidy. *Hum. Mol. Genet.* **2007**, *16*, 2870–2879.
4. Griffin, J.; Emery, B.R.; Christensen, G.L.; Carrell, D.T. Analysis of the meiotic recombination gene *REC8* for sequence variations in a population with severe male factor infertility. *Syst. Biol. Reprod. Med.* **2008**, *54*, 163–165.
5. Hann, M.C.; Lau, P.E.; Tempest, H.G. Meiotic recombination and male infertility: From basic science to clinical reality? *Asian J. Androl.* **2011**, *13*, 212–218.
6. Baudat, F.; Nicolas, A. Clustering of meiotic double-strand breaks on yeast chromosome III. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 5213–5218.
7. Klein, S.; Zenvirth, D.; Dror, V.; Barton, A.B.; Kaback, D.B.; Simchen, G. Patterns of meiotic double-strand breakage on native and artificial yeast chromosomes. *Chromosoma* **1996**, *105*, 276–284.
8. Zenvirth, D.; Arbel, T.; Sherman, A.; Goldway, M.; Klein, S.; Simchen, G. Multiple sites for double-strand breaks in whole meiotic chromosomes of Saccharomyces cerevisiae. *EMBO J.* **1992**, *11*, 3441–3447.

9. Petes, T.D. Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.* **2001**, *2*, 360–369.
10. Kohl, K.P.; Sekelsky, J. Meiotic and mitotic recombination in meiosis. *Genetics* **2013**, *194*, 327–334.
11. Lichten, M.; Goldman, A.S. Meiotic recombination hotspots. *Ann. Rev. Genet.* **1995**, *29*, 423–444.
12. Jeffreys, A.J.; Holloway, J.K.; Kauppi, L.; May, C.A.; Neumann, R.; Slingsby, M.T.; Webb, A.J. Meiotic recombination hot spots and human DNA diversity. *Philos. Trans. R. Soc. Lond. Ser. B* **2004**, *359*, 141–152.
13. Wahls, W.P. Meiotic recombination hotspots: Shaping the genome and insights into hypervariable minisatellite DNA change. *Curr. Top. Dev. Biol.* **1998**, *37*, 37–75.
14. Liu, G.; Liu, J.; Cui, X.; Cai, L. Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae. *J. Theor. Biol.* **2012**, *293*, 49–54.
15. Chen, W.; Lin, H.; Feng, P.M.; Ding, C.; Zuo, Y.C.; Chou, K.C. iNuc-PhysChem: A sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One* **2012**, *7*, e47843.
16. Chou, K.C. Prediction of G-protein-coupled receptor classes. *J. Proteome Res.* **2005**, *4*, 1413–1418.
17. Chou, K.C.; Elrod, D.W. Prediction of enzyme family classes. *J. Proteome Res.* **2003**, *2*, 183–190.
18. Wang, M.; Yang, J.; Xu, Z.J.; Chou, K.C. SLLE for predicting membrane protein types. *J. Theor. Biol.* **2005**, *232*, 7–15.
19. Xiao, X.; Wang, P.; Chou, K.C. Predicting protein structural classes with pseudo amino acid composition: An approach using geometric moments of cellular automaton image. *J. Theor. Biol.* **2008**, *254*, 691–696.
20. Chou, K.C. A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. *Proteins: Struct. Funct. Genet* **1995**, *21*, 319–344.
21. Feng, K.Y.; Cai, Y.D.; Chou, K.C. Boosting classifier for predicting protein domain structural class. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 213–217.
22. Cai, Y.D.; Chou, K.C. Artificial neural network for predicting alpha-turn types. *Anal. Biochem.* **1999**, *268*, 407–409.
23. Thompson, T.B.; Chou, K.C.; Zheng, C. Neural network prediction of the HIV-1 protease cleavage sites. *J. Theor. Biol.* **1995**, *177*, 369–379.
24. Feng, P.M.; Chen, W.; Lin, H.; Chou, K.C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* **2013**, *442*, 118–125.
25. Chen, W.; Feng, P.M.; Lin, H.; Chou, K.C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **2013**, *41*, e69.
26. Xiao, X.; Wang, P.; Chou, K.C. iNR-PhysChem: A sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. *PLoS One* **2012**, *7*, e30869.
27. Lin, W.Z.; Fang, J.A.; Xiao, X.; Chou, K.C. iDNA-Prot: Identification of DNA binding proteins using random forest with grey model. *PLoS One* **2011**, *6*, e24756.
28. Kandaswamy, K.K.; Chou, K.C.; Martinetz, T.; Moller, S.; Suganthan, P.N.; Sridharan, S.; Pugalenthi, G. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol.* **2011**, *270*, 56–62.

29. Xu, Y.; Ding, J.; Wu, L.Y.; Chou, K.C. iSNO-PseAAC: Predict cysteine *S*-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One* **2013**, *8*, e55844.
30. Cai, Y.D.; Chou, K.C. Predicting subcellular localization of proteins in a hybridization space. *Bioinformatics* **2004**, *20*, 1151–1156.
31. Chou, K.C.; Cai, Y.D. Prediction of protease types in a hybridization space. *Biochem. Biophys. Res. Commun.* **2006**, *339*, 1015–1020.
32. Chou, K.C.; Shen, H.B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Proteome Res.* **2006**, *5*, 1888–1897.
33. Chou, K.C.; Shen, H.B. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* **2006**, *347*, 150–157.
34. Chou, K.C.; Shen, H.B. Large-scale predictions of Gram-negative bacterial protein subcellular locations. *J. Proteome Res.* **2006**, *5*, 3420–3428.
35. Chou, K.C.; Shen, H.B. Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **2007**, *6*, 1728–1734.
36. Chou, K.C.; Shen, H.B. Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* **2007**, *357*, 633–640.
37. Shen, H.B.; Chou, K.C. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 288–292.
38. Shen, H.B.; Chou, K.C. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.* **2009**, *394*, 269–274.
39. Xiao, X.; Wang, P.; Chou, K.C. GPCR-2L: Predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. *Mol. Biosyst.* **2011**, *7*, 911–919.
40. Shen, H.B.; Yang, J.; Chou, K.C. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. *J. Theor. Biol.* **2006**, *240*, 9–13.
41. Xiao, X.; Min, J.L.; Wang, P.; Chou, K.C. iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PLoS One* **2013**, *8*, e72234.
42. Xiao, X.; Min, J.L.; Wang, P.; Chou, K.C. iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *J. Theor. Biol.* **2013**, *337C*, 71–79.
43. Xiao, X.; Wang, P.; Lin, W.Z.; Jia, J.H.; Chou, K.C. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **2013**, *436*, 168–177.
44. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Struct. Funct. Genet.* **2001**, *43*, 246–255.
45. Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10–19.
46. Lin, S.X.; Lapointe, J. Theoretical and experimental biology in one—A symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers. *J. Biomed. Sci. Eng.* **2013**, *6*, 435–442.

47. Nanni, L.; Lumini, A.; Gupta, D.; Garg, A. 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.* **2012**, *9*, 467–475.
48. Khosravian, M.; Faramarzi, F.K.; Beigi, M.M.; Behbahani, M.; Mohabatkar, H. Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. *Protein Pept. Lett.* **2013**, *20*, 180–186.
49. Yu, L.; Guo, Y.; Li, Y.; Li, G.; Li, M.; Luo, J.; Xiong, W.; Qin, W. SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. *J. Theor. Biol.* **2010**, *267*, 1–6.
50. Zou, D.; He, Z.; He, J.; Xia, Y. Supersecondary structure prediction using Chou's pseudo amino acid composition. *J. Comput. Chem.* **2011**, *32*, 271–278.
51. Zhang, S.W.; Zhang, Y.L.; Yang, H.F.; Zhao, C.H.; Pan, Q. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: An approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids* **2008**, *34*, 565–572.
52. Kandaswamy, K.K.; Pugalenthi, G.; Moller, S.; Hartmann, E.; Kalies, K.U.; Suganthan, P.N.; Martinetz, T. Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. *Protein Pept. Lett.* **2010**, *17*, 1473–1479.
53. Mei, S. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *J. Theor. Biol.* **2012**, *310*, 80–87.
54. Chang, T.H.; Wu, L.C.; Lee, T.Y.; Chen, S.P.; Huang, H.D.; Horng, J.T. EuLoc: A web-server for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 91–103.
55. Fan, G.L.; Li, Q.Z. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2012**, *304*, 88–95.
56. Huang, C.; Yuan, J. Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. *Biosystems* **2013**, *113*, 50–57.
57. Lin, H.; Wang, H.; Ding, H.; Chen, Y.L.; Li, Q.Z. Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. *Acta Biotheor.* **2009**, *57*, 321–330.
58. Wan, S.; Mak, M.W.; Kung, S.Y. 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.* **2013**, *323*, 40–48.
59. Huang, C.; Yuan, J.Q. Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. *J. Theor. Biol.* **2013**, *335*, 205–212.
60. Chen, Y.K.; Li, K.B. 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.* **2013**, *318*, 1–12.

61. Huang, C.; Yuan, J.Q. A Multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types. *J. Membr. Biol.* **2013**, *246*, 327–334.
62. Hayat, M.; Khan, A. Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC. *Protein Pept. Lett.* **2012**, *19*, 411–421.
63. Mohabatkar, H.; Beigi, M.M.; Abdolahi, K.; Mohsenzadeh, S. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.* **2013**, *9*, 133–137.
64. Mohammad Beigi, M.; Behjati, M.; Mohabatkar, H. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genomics* **2011**, *12*, 191–197.
65. Sahu, S.S.; Panda, G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.* **2010**, *34*, 320–327.
66. Zia Ur, R.; Khan, A. 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.* **2012**, *19*, 890–903.
67. Xie, H.L.; Fu, L.; Nie, X.D. Using ensemble SVM to identify human GPCRs *N*-linked glycosylation sites based on the general form of Chou's PseAAC. *Protein Eng. Des. Sel.* **2013**, *26*, 735–742.
68. Zhang, S.W.; Chen, W.; Yang, F.; Pan, Q. Using Chou's pseudo amino acid composition to predict protein quaternary structure: A sequence-segmented PseAAC approach. *Amino Acids* **2008**, *35*, 591–598.
69. Sun, X.Y.; Shi, S.P.; Qiu, J.D.; Suo, S.B.; Huang, S.Y.; Liang, R.P. Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol. BioSyst.* **2012**, *8*, 3178–3184.
70. Nanni, L.; Lumini, A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids* **2008**, *34*, 653–660.
71. Fan, G.L.; Li, Q.Z. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. *Amino Acids* **2012**, *43*, 545–555.
72. Mei, S. Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. *J. Theor. Biol.* **2012**, *293*, 121–130.
73. Zeng, Y.H.; Guo, Y.Z.; Xiao, R.Q.; Yang, L.; Yu, L.Z.; Li, M.L. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol.* **2009**, *259*, 366–372.
74. Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.* **2010**, *263*, 203–209.
75. Mohabatkar, H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett.* **2010**, *17*, 1207–1214.
76. Mohabatkar, H.; Mohammad Beigi, M.; Esmaeili, A. Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* **2011**, *281*, 18–23.

77. Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J.; Torres, A. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.* **2009**, *257*, 17–26.
78. Zhang, G.Y.; Fang, B.S. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. *J. Theor. Biol.* **2008**, *253*, 310–315.
79. Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.* **2007**, *248*, 546–551.
80. Liu, B.; Wang, X.; Zou, Q.; Dong, Q.; Chen, Q. Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. *Mol. Informa.* **2013**, *32*, 775–782.
81. Georgiou, D.N.; Karakasidis, T.E.; Megaritis, A.C. A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory. *Open Bioinforma. J.* **2013**, *7*, 41–48.
82. Hajisharifi, Z.; Piryaiee, M.; Mohammad Beigi, M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40.
83. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.* **2011**, *273*, 236–247.
84. Li, B.Q.; Huang, T.; Liu, L.; Cai, Y.D.; Chou, K.C. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. *PLoS One* **2012**, *7*, e33393.
85. Huang, T.; Wang, J.; Cai, Y.D.; Yu, H.; Chou, K.C. Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma. *PLoS One* **2012**, *7*, e34460.
86. Jiang, Y.; Huang, T.; Lei, C.; Gao, Y.F.; Cai, Y.D.; Chou, K.C. Signal propagation in protein interaction network during colorectal cancer progression. *BioMed Res. Int.* **2013**, *2013*, 287019.
87. Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* **2012**, *425*, 117–119.
88. Cao, D.S.; Xu, Q.S.; Liang, Y.Z. Propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics* **2013**, *29*, 960–962.
89. Shen, H.B.; Chou, K.C. PseAAC: A flexible web-server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* **2008**, *373*, 386–388.
90. Min, J.L.; Xiao, X.; Chou, K.C. iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. *BioMed Res. Int.* **2013**, *2013*, 701317.
91. Xu, Y.; Shao, X.J.; Wu, L.Y.; Deng, N.Y.; Chou, K.C. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ* **2013**, *1*, e171.
92. Liu, B.; Zhang, D.; Xu, R.; Xu, J.; Wang, X.; Chen, Q.; Dong, Q.; Chou, K.C. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* **2013**, doi:10.1093/bioinformatics/btt709.

93. Lin, H.; Ding, H. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. *J. Theor. Biol.* **2011**, *269*, 64–69.
94. Liu, W.; Chou, K.C. Protein secondary structural content prediction. *Protein Eng.* **1999**, *12*, 1041–1050.
95. Lin, H.; Li, Q.Z. Using pseudo amino acid composition to predict protein structural class: Approached by incorporating 400 dipeptide components. *J. Comput. Chem.* **2007**, *28*, 1463–1466.
96. Chou, K.C. Using pair-coupled amino acid composition to predict protein secondary structure content. *J. Protein Chem.* **1999**, *18*, 473–480.
97. Lin, H.; Ding, C.; Yuan, L.F.; Chen, W.; Ding, H.; Li, Z.Q.; Guo, F.B.; Hung, J.; Rao, N.N. Predicting subchloroplast locations of proteins based on the general form of Chou's pseudo amino acid composition: Approached from optimal tripeptide composition. *Int. J. Biomath.* **2013**, *6*, 1350003, doi:10.1142/S1793524513500034.
98. Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.* **1962**, *84*, 4240–4274.
99. Hopp, T.P.; Woods, K.R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* **1981**, *78*, 3824–3828.
100. Robert, C.W. *CRC Handbook of Chemistry and Physics*, 66th ed.; CRC Press: Boca Raton, FL, USA, 1985.
101. Dawson, R.M.C.; Elliott, D.C.; Elliott, W.H.; Jones, K.M. *Data for Biochemical Research*, 3rd ed.; Clarendon Press: Oxford, UK, 1986.
102. Chen, J.; Liu, H.; Yang, J.; Chou, K.C. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* **2007**, *33*, 423–428.
103. Chou, K.C.; Cai, Y.D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **2002**, *277*, 45765–45769.
104. Lin, W.Z.; Fang, J.A.; Xiao, X.; Chou, K.C. Predicting secretory proteins of malaria parasite by incorporating sequence evolution information into pseudo amino acid composition via grey system model. *PLoS One* **2012**, *7*, e49040.
105. Wang, S.Q.; Yang, J.; Chou, K.C. Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. *J. Theor. Biol.* **2006**, *242*, 941–946.
106. Cai, Y.D.; Zhou, G.P.; Chou, K.C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.* **2003**, *84*, 3257–3263.
107. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27.
108. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000; p. 189.
109. Chou, K.C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* **2013**, *9*, 1092–1100.
110. Chou, K.C. Using subsite coupling to predict signal peptides. *Protein Eng.* **2001**, *14*, 75–79.
111. Chou, K.C. Prediction of protein signal sequences and their cleavage sites. *Proteins: Struct. Funct. Genet.* **2001,** 42, 136–139.
112. Chou, K.C.; Wu, Z.C.; Xiao, X. iLoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One* **2011**, *6*, e18258.

113. Wu, Z.C.; Xiao, X.; Chou, K.C. iLoc-Plant: A multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. BioSyst.* **2011**, *7*, 3287–3297.
114. Wu, Z.C.; Xiao, X.; Chou, K.C. iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. *Protein Pept. Lett.* **2012**, *19*, 4–14.
115. Xiao, X.; Wu, Z.C.; Chou, K.C. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.* **2011**, *284*, 42–51.
116. Xiao, X.; Wu, Z.C.; Chou, K.C. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. *PLoS One* **2011**, *6*, e20592.
117. Chou, K.C.; Wu, Z.C.; Xiao, X. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* **2012**, *8*, 629–641.
118. Lin, W.Z.; Fang, J.A.; Xiao, X.; Chou, K.C. iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. Biosyst.* **2013**, *9*, 634–644.
119. Chen, L.; Zeng, W.M.; Cai, Y.D.; Feng, K.Y.; Chou, K.C. Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS One* **2012**, *7*, e35254.
120. Chou, K.C.; Zhang, C.T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349.
121. Fan, G.L.; Li, Q.Z. Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2013**, *334*, 45–51.
122. Qiu, J.D.; Huang, J.H.; Liang, R.P.; Lu, X.Q. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. *Anal. Biochem.* **2009**, *390*, 68–73.
123. Chou, K.C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics* **2009**, *6*, 262–274.