

Article

Numerical Characterization of Protein Sequences Based on the Generalized Chou's Pseudo Amino Acid Composition

Chun Li ^{1,2,*}, Xueqin Li ¹ and Yan-Xia Lin ²¹ Department of Mathematics, Bohai University, Jinzhou 121013, China; 18841655169@163.com² NIASRA—National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong 2522, Australia; yanxia@uow.edu.au

* Correspondence: lichwun@163.com; Tel.: +86-416-3400-145

Academic Editors: Yang Kuang and Wanbiao Ma

Received: 18 September 2016; Accepted: 29 November 2016; Published: 6 December 2016

Abstract: The technique of comparison and analysis of biological sequences is playing an increasingly important role in the field of Computational Biology and Bioinformatics. One of the key steps in developing the technique is to identify an appropriate manner to represent a biological sequence. In this paper, on the basis of three physical–chemical properties of amino acids, a protein primary sequence is reduced into a six-letter sequence, and then a set of elements which reflect the global and local sequence-order information is extracted. Combining these elements with the frequencies of 20 native amino acids, a $(21 + \lambda)$ dimensional vector is constructed to characterize the protein sequence. The utility of the proposed approach is illustrated by phylogenetic analysis and identification of DNA-binding proteins.

Keywords: generalized pseudo amino acid composition; numerical characterization; phylogenetic analysis; identification of DNA-binding proteins

1. Introduction

In the task of comparison and analysis of biological sequences, choosing a type of DNA/protein representation is an important step. The usual representation of the primary structure of DNA is a string of four letters: A (adenine); G (guanine); C (cytosine); and T (thymine). This expression is called a letter sequence representation (LSR) or a DNA primary sequence. Similarly, a protein primary sequence is usually expressed in terms of a series of 20 letters, which denote 20 different amino acids. The sequence encodes information of the corresponding structure and function in a living organism. However, it is difficult to obtain the information from the representation of a primary sequence directly. Therefore, various sequence representation techniques have been developed for encoding bio-sequences and extracting the hidden information.

Graphical representation of DNA is a useful tool for visualizing and analyzing DNA sequences. By using the tool, one can obtain a route to condense the information coded by DNA primary sequences into a set of invariants [1,2]. Early attempts towards graphical representations of DNA were made by Hamori and Ruskin in 1983 [3], Hamori in 1985 [4], and Gates in 1985 [5]. Afterwards, more graphical representations of DNA sequences were well developed by researchers [1,2,6–15]. In comparison with DNA, graphical representations of proteins emerged only very recently [2,16–27]. As a matter of fact, most of the graphical representations of DNA involve some degree of arbitrariness, such as the selection of directions to be assigned to individual bases. For a string like DNA sequence over an alphabet with size 4, there are $4! = 24$ possible ways of assigning 4 directions to 4 nucleic acid bases. If these methods are directly extended to protein sequences, the corresponding figure is

$20! \approx 2.433 \times 10^{18}$. It is impracticable to represent one protein sequence by such an enormous number of graphs. This is probably the most important reason why protein graphical representations have not been advanced [19,23]. It is found that reducing the alphabet or fixing the directions assigned to amino acid residues plays an important role in addressing this problem. For details, we refer to some recent publications [2,16,21,23,24,28].

Matrix representation of a biological sequence is another powerful tool for characterization and comparison of sequences. These matrices include: The frequency matrix; Euclidean-distance matrix (*ED*); graph theoretical distance matrix (*GD*); line distance matrix (*LD*); quotient matrix (*D/D*, *M/M*, *L/L*); and their “higher order” matrices [1,2,12,13,20,21,27,29,30]. Among them, *ED*, *GD*, *L/L*, etc., are derived from a graphical representation. For example, *L/L* is a symmetric matrix whose diagonal entries are zero, while other entries are defined as the quotient of the Euclidean distance between two points of the graph and the sum of geometrical lengths of edges between the two points. Once the matrix is given, some of matrix invariants can be used as descriptors of the sequence. Eigenvalues of a matrix are one of the best-known matrix invariants [31]. In fact, two graphs are isomorphic if and only if their adjacency matrices are similar. It is of interest to note that similar matrices have the same eigenvalues. Among all the eigenvalues, the leading eigenvalue often plays a special role and has been widely used in the field of biological science and chemistry. However, a problem we must face is that the calculation of the eigenvalue will become more and more difficult with the order of the matrix large. ALE-index is an alternative invariant we proposed in 2005 [32]. The ALE-index can be viewed as an Approximation of the Leading Eigenvalue (ALE) of the corresponding matrix (it is just in this sense that it is called ‘ALE’-index), while it is much simpler for calculation than the latter. Therefore, it may be more economical to adopt the ALE-index when one is interested only in the leading eigenvalue.

The third method for formulating a protein sequence is the pseudo amino acid composition (PseAAC), with the advantage of avoiding loss of the sequence-order information. Ever since the concept of PseAAC [33,34] or Chou’s PseAAC [35,36] was proposed, it has rapidly penetrated into nearly all fields of computational proteomics (see a long list papers cited in [36,37]). Stimulated by the great successes of PseAAC in dealing with protein/peptide sequences, the concept of PseAAC has been extended [38–42] to cover DNA/RNA sequences as well via the form of PseKNC (pseudo K-tuple nucleotide composition) [43,44], which has been proven very useful in studying many important genome analysis problems, as summarized in a recent review paper [45]. Also, because the concept of PseAAC has been increasingly and widely used in both computational proteomics and genomics, a very powerful web-server called “Pse-in-One” [46] was established that can be used to generate the pseudo components for both protein/peptide and DNA/RNA sequences.

In this paper, we modify the method of Chou’s PseAAC and propose a novel approach for numerically characterizing a protein sequence. We characterize a protein sequence by a $(21 + \lambda)$ dimensional vector, whose first 20 components are the occurrence frequencies of 20 native amino acids, while the last $\lambda + 1$ components are based on a six-letter sequence derived from the protein primary sequence. The former is used to reflect the effect of the amino acid composition, and the latter is used to reflect the effect of sequence order and property of the residues. It is well known that a sequence naturally contains two pieces of information: the elements of the sequence; and the orders of the elements. Any methodologies based on the amino acid composition alone are worthy of further investigation. However, as pointed out by Chou [33,34], it is not feasible to completely include all sequence order patterns. It was stirring to see that Chou creatively developed an approach as mentioned above to extract the important feature beyond amino acid composition. Our scheme is similar to, but different from, that of Chou. Experiments about phylogenetic analysis on two datasets and identification of DNA-binding proteins illustrate the utility of the proposed method.

2. Methods

A protein sequence can be viewed as a string of 20 amino acids. Without loss of generality, by the numerical indices 1, 2, ..., 20, we represent the 20 native amino acids according to the alphabetical

order of their single-letter codes: A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W and Y. Then the frequencies of appearance of the 20 amino acids in a protein sequence are often used to construct a vector

$$[f_1, f_2, \dots, f_{20}]$$

This is the conventional amino acid composition. The advantage of such a vector representation is that it is easy in statistical treatment, but it cannot reflect the effect regarding sequence order and property. In what follows, we will take this effect into account through a set of elements in addition to the 20 components.

Hydrophobicity, isoelectric point (pI), and relative distance (RD) are three important physicochemical properties of the 20 native amino acids. Here RD can be viewed as an integration of the information on three side chain properties: composition; polarity; and molecular volume—where composition is defined as the atomic weight ratio of hetero (noncarbon) elements in end groups or rings to carbons in the side chain (for details, see [47]). Listed in Table 1 are the original numerical values for hydrophobicity, pI and RD. As can be seen from Table 1, the values of P_1^0 (Hydrophobicity) is in the range $[-2.53 \sim 1.38]$, and the values of P_2^0 (isoelectric point) are in the range of $2.97 \sim 10.76$, while P_3^0 (relative distance) varies between 1469 and 3355. Therefore, the normalization of these values is needed. Here we normalize them by the formulary below:

$$p'_n(AA_i) = P_n^0(AA_i) - \min_{j=1, \dots, 20} \{P_n^0(AA_j)\},$$

$$P_n^*(AA_i) = \frac{p'_n(AA_i)}{\max_{j=1, \dots, 20} \{p'_n(AA_j)\}} \cdot i = 1, 2, \dots, 20, n = 1, 2, 3. \quad (1)$$

Table 1. The original numerical values for the properties of the 20 native amino acids.

Amino Acid (AA)	Hydrophobicity ^a (P_1^0)	pI ^b (P_2^0)	RD ^b (P_3^0)
A	0.62	6.02	1889
C	0.29	5.02	3355
D	−0.90	2.97	2209
E	−0.74	3.22	1812
F	1.19	5.48	1916
G	0.48	5.97	2078
H	−0.40	7.59	1507
I	1.38	6.02	1765
K	−1.50	9.74	1797
L	1.06	5.98	1822
M	0.64	5.75	1689
N	−0.78	5.42	1943
P	0.12	6.30	1720
Q	−0.85	5.65	1538
R	−2.53	10.76	1697
S	−0.18	5.68	2000
T	−0.05	6.53	1469
V	1.08	5.97	1680
W	0.81	5.89	2317
Y	0.26	5.66	1787

^a Taken from [41]; ^b Taken from [47–49].

Clearly, the normalized values for properties of the 20 native amino acids are in the interval $[0,1]$. The corresponding values are listed in Table 2. The last row in this table gives the average values.

Table 2. The normalized values for the properties of the 20 native amino acids.

AA	P_1^*	P_2^*	P_3^*
A	0.8056	0.3915	0.2227
C	0.7212	0.2632	1.0000
D	0.4169	0	0.3924
E	0.4578	0.0321	0.1819
F	0.9514	0.3222	0.2370
G	0.7698	0.3851	0.3229
H	0.5448	0.5931	0.0201
I	1.0000	0.3915	0.1569
K	0.2634	0.8691	0.1739
L	0.9182	0.3864	0.1872
M	0.8107	0.3569	0.1166
N	0.4476	0.3145	0.2513
P	0.6777	0.4275	0.1331
Q	0.4297	0.3440	0.0366
R	0	1.0000	0.1209
S	0.6010	0.3479	0.2815
T	0.6343	0.4570	0
V	0.9233	0.3851	0.1119
W	0.8542	0.3748	0.4496
Y	0.7136	0.3453	0.1686
\overline{P}_n	0.6471	0.3994	0.2283

For each amino acid (AA), we associate it with a triple $(t(1), t(2), t(3))$, where

$$t(n) = \begin{cases} +1 & \text{if } P_n^*(AA) \geq \overline{P}_n \\ -1 & \text{otherwise} \end{cases} \quad (n = 1, 2, 3) \quad (2)$$

All the amino acids with a same triple form a group. In this way, the 20 native amino acids can be classified into 6 groups:

$G_I = \{A, Y, V, M, L, I\}$,

$G_{II} = \{C, W, G, F\}$,

$G_{III} = \{D, S, N\}$,

$G_{IV} = \{E, Q\}$,

$G_V = \{H, T, R, K\}$,

$G_{VI} = \{P\}$.

For each group, the first amino acid is selected to be the representative. That is, A, C, D, E, H and P are used to stand for the six groups, respectively. The value of the property of a group is defined as the average value of the property of amino acids belonging to the group. Listed in Table 3 are the corresponding values of the six groups.

Table 3. The values for properties of the six groups.

Group	Representative	P_1	P_2	P_3
G_I	A	0.8619	0.3761	0.1607
G_{II}	C	0.8242	0.3363	0.5024
G_{III}	D	0.4885	0.2208	0.3084
G_{IV}	E	0.4437	0.1881	0.1092
G_V	H	0.3606	0.7298	0.0787
G_{VI}	P	0.6777	0.4275	0.1331

At the same time, a protein primary sequence can be reduced into a six-letter sequence by replacing each element in the protein sequence with its representative letter. Suppose $S = S_1 S_2 \dots S_L$ is a given six-letter sequence, we inspect it by stepping one element at a time. For the step k ($k = 1, 2, \dots, L$), a 3-D space point $q_k = (x_k, y_k, z_k)$ can be constructed as follows:

$$(x_k, y_k, z_k) = (x_{k-1}, y_{k-1}, z_{k-1}) + (P_1(S_k), P_2(S_k), P_3(S_k)), \quad (3)$$

where $(x_0, y_0, z_0) = (0, 0, 0)$. When k runs from 1 to L , we get L points q_1, q_2, \dots, q_L . Connecting these points one by one sequentially with straight lines, a three-dimensional curve can be drawn. One can further associate the graph with some structural matrices. Here we adopt the L/L matrix and denote it by M , whose (i, j) -entry is defined as follows:

$$m_{ij} = \begin{cases} \frac{d(i, j)}{d(i, i+1) + d(i+1, i+2) + \dots + d(j-1, j)} & \text{if } i < j \\ 0 & \text{if } i = j, \\ m_{ji} & \text{if } i > j \end{cases} \quad (4)$$

where $d(i, j)$ is the Euclidean distance between points q_i and q_j . It is not difficult to see that $\lim_{t \rightarrow +\infty} {}^t M$ is a $(0, 1)$ matrix; here ${}^t M$ stands for the product of Hadamard multiplication of the matrix M by itself t -times. In this paper, we call the limit matrix as a generalized adjacency matrix (GAM) generated by points q_1, q_2, \dots, q_L , and denote it by M_G . Obviously, $[M_G]_{ij} = 1$ if and only if q_i and q_j lie on a straight line in the graph.

As mentioned above, once a symmetric matrix is given, one can calculate its ALE-index by the following formula:

$$\chi = \frac{1}{2} \left(\frac{1}{L} \|\cdot\|_{m1} + \sqrt{\frac{L-1}{L}} \|\cdot\|_F \right), \quad (5)$$

where L is the order of the matrix, $\|\cdot\|_{m1}$ and $\|\cdot\|_F$ are the $m1$ - and F -norms of a matrix, respectively. In order to reduce variations caused by comparison of matrices with different sizes, we consider instead of $\chi(M_G)$ a normalized ALE-index $\chi'(M_G) = \frac{\chi(M_G)}{\sqrt{6L}}$.

In addition, following the similar procedures in capturing the sequence-order information of a protein [33,34], for the six-letter sequence $S = S_1 S_2 \dots S_L$, we extract a set of new order-correlated factors as defined below:

$$\begin{aligned} \theta_1 &= \frac{1}{L-1} \times \frac{1}{3} \times \sum_{n=1}^3 g_n(S, 1), \\ \theta_2 &= \frac{1}{L-2} \times \frac{1}{3} \times \sum_{n=1}^3 g_n(S, 2), \\ &\dots \\ \theta_\lambda &= \frac{1}{L-\lambda} \times \frac{1}{3} \times \sum_{n=1}^3 g_n(S, \lambda). \end{aligned} \quad (\lambda < L) \quad (6)$$

where θ_k ($k = 1, 2, \dots, \lambda$) is called the k -th tier correlation factor, $g_n(S, k)$ represents the coupling mode function as given by

$$g_n(S, k) = \sqrt{\sum_{i=1}^{L-k} (P_n(S_i) - P_n(S_{i+k}))^2} \quad (7)$$

Factor θ_1 reflects the coupling mode between the most contiguous elements along a six-letter sequence (Figure 1a); θ_2 reflects the coupling mode between the second-most contiguous (Figure 1b); θ_3 reflects the coupling mode between the third-most contiguous (Figure 1c), and so on. λ is the highest rank of the coupling mode.

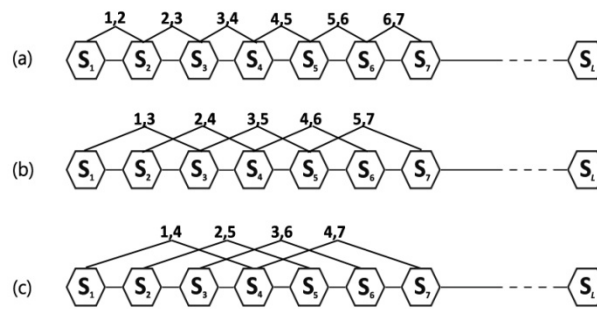


Figure 1. A schematic diagram to show: (a) the first-tier; (b) the second tier; and (c) the third-tier sequence order correlation mode along a sequence. Where the regular hexagon is used to show that each element of the sequence corresponds to one of the six amino acid groups.

Consequently, a protein sequence can be characterized by a $(21 + \lambda)$ dimensional vector V :

$$V = (v_1, v_2, \dots, v_{20}, v_{20+1}, \dots, v_{20+\lambda}, v_{20+\lambda+1}), \quad (8)$$

where

$$v_i = \begin{cases} f_i & 1 \leq i \leq 20 \\ w_1 \theta_{i-20} & 20 + 1 \leq i \leq 20 + \lambda \\ w_2 \chi' & i = 20 + \lambda + 1 \end{cases} \quad (9)$$

Here w_1 and w_2 are weight factors. It is easy to see that the first 20 components reflect the effect of the amino acid composition, whereas the last $\lambda + 1$ components reflect the effect of sequence order and property of the residues. For convenience, a set of such $21 + \lambda$ components as formulated by Equations (8) and (9) is called the generalized pseudoamino acid composition of a protein sequence, and denoted by G-PseAAC.

3. Results

In this section, we will illustrate the use of the new quantitative characterization of protein sequences with two experiments. As we can see from Equations (8) and (9), there are three adjustable parameters for the G-PseAAC: λ , w_1 , and w_2 . It is not known beforehand which λ , w_1 , and w_2 are best for a given problem. Three datasets are considered in this paper. The first one is used for determining these parameters and others for testing purpose.

3.1. Experiment I: Phylogenetic Analysis on Two Datasets

The first dataset used in this paper is composed of β -globin protein of 17 species (see Table 4). According to the method proposed, we associate each of the 17 protein sequences with a $\tau = 21 + \lambda$ dimensional vector. These vectors are then used to define a pair-wise evolutionary distance between any two protein sequences i and j :

$$D(i, j) = d(V_i, V_j) = \sqrt{\sum_{k=1}^{\tau} (v_{ik} - v_{jk})^2} \quad (10)$$

where $V_i = (v_{i1}, v_{i2}, \dots, v_{i,\tau})$ and $V_j = (v_{j1}, v_{j2}, \dots, v_{j,\tau})$ are the corresponding vectors for sequences i and j , respectively. Thus, a 17×17 real symmetric matrix D_{17} is obtained. On the basis of the achieved distance matrix D_{17} , a phylogenetic tree can be constructed using a UPGMA (Unweighted Pair Group Method with Arithmetic Mean) program included in the MEGA4 package. It is found that, when $\lambda = 7$ and $w_1 = w_2 = 1.6$, the non-mammals, including Guttata, Gallus and Muscovy duck, appear to cluster together and stay outside of the mammals, while Opossum is distinguished from the remaining mammals. In addition, Primate group {Human, Chimpanzee, Gorilla}, Cetartiodactyla

group {Cattle, Banteng, Sheep, Goat}, Lagomorpha group {Rabbit, European hare}, and Rodentia group {House mouse, Western wild mouse, Spiny mouse, Norway rat} form separate branches, respectively (cf. Figure 2). This result is in accordance with the accepted taxonomy and the literature [1,12,30].

Table 4. The β -globin protein of 17 species.

No.	Species	Accession Number	Length (aa)
1	Human	ALU64020	147
2	Gorilla	P02024	147
3	Chimpanzee	P68873	147
4	Cattle	CAA25111	145
5	Banteng	BAJ05126	145
6	Goat	AAA30913	145
7	Sheep	ABC86525	145
8	European hare	CAA68429	147
9	Rabbit	CAA24251	147
10	House mouse	ADD52660	147
11	Western wild mouse	ACY03394	147
12	Spiny mouse	ACY03377	147
13	Norway rat	CAA29887	147
14	Opossum	AAA30976	147
15	Guttata	ACH46399	147
16	Gallus	CAA23700	147
17	Muscovy duck	CAA33756	147

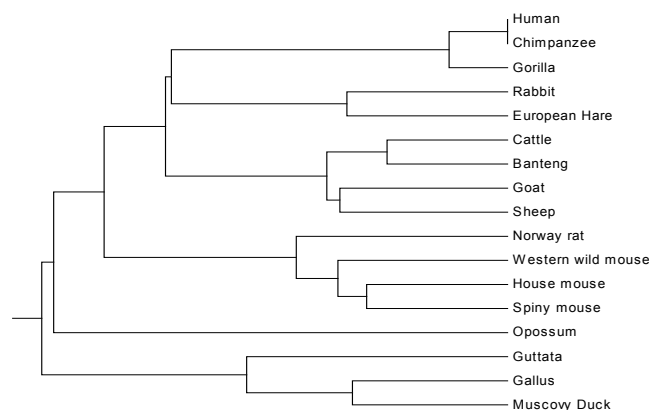


Figure 2. The relationship tree of 17 species.

Using the above-determined values for λ , w_1 , and w_2 , we infer the relationship of 72 coronavirus spike (S) proteins. The coronavirus, whose name is derived from its crown-like shape, is a positive-sense, single-stranded RNA virus in the family *Coronaviridae*. It was first identified in the 1960s from the nasal cavities of patients with the common cold. Most coronaviruses are not dangerous, but some strains could cause severe, sometimes fatal, diseases in humans and other animals. The MERS coronavirus (commonly shortened to MERS-CoV) is the virus that causes the Middle East respiratory syndrome (MERS). MERS was first reported in 2012 in Saudi Arabia and then in other countries in the Middle East, Africa, Asia, Europe and America. As of July 2016, 1769 laboratory-confirmed cases of MERS-CoV infection, including at least 630 related deaths (the case fatality rate is >30%), have been reported in over 27 countries (<http://www.who.int/emergencies/mers-cov/en/>). People also died from a severe acute respiratory syndrome (SARS), which first emerged in 2002 in Guangdong Province, China, and then spread globally. SARS resulted in more than 8000 infections with a case-fatality rate of ~10%. The virus that causes SARS is officially called SARS coronavirus (SARS-CoV). Both MERS-CoV and SARS-CoV are identified as members of the beta group of coronavirus, *Betacoronavirus*, while

they are distinct from each other. The name, accession number, and abbreviation of the 72 sequences are listed in Table 5. According to the existing taxonomic groups, sequences 1–5 belong to group alpha (formerly known as Coronavirus group 1 (CoV-1)), sequences 6–8 are members of group gamma (formerly CoV-3), and the remaining belongs to group beta (formerly CoV-2). Refer to Table 5 for details.

Table 5. The accession number, name and abbreviation for 72 coronavirus spike proteins.

NO.	Accession Number	Virus Name/Strain	Abbreviation
1	CAB91145	Transmissible gastroenteritis virus, genomic RNA	TGEVG
2	NP_058424	Transmissible gastroenteritis virus	TGEV
3	AAK38656	Porcine epidemic diarrhea virus strain CV777	PEDVC
4	NP_598310	Porcine epidemic diarrhea virus	PEDV
5	BAL45637	Human coronavirus 229E	HCoV-229E
6	AAP92675	Avian infectious bronchitis virus isolate BJ	IBVBJ
7	AAS00080	Avian infectious bronchitis virus strain Ca199	IBVC
8	NP_040831	Avian infectious bronchitis virus	IBV
9	NP_937950	Human coronavirus OC43	HCoV-OC43
10	AAK83356	Bovine coronavirus isolate BCoV-ENT	BCoVE
11	AAL57308	Bovine coronavirus isolate BCoV-LUN	BCoVL
12	AAA66399	Bovine coronavirus strain Mebus	BCoVM
13	AAL40400	Bovine coronavirus strain Quebec	BCoVQ
14	NP_150077	Bovine coronavirus	BCoV
15	AAB86819	Mouse hepatitis virus strain MHV-A59C12 mutant	MHVA
16	YP_209233	Murine hepatitis virus strain JHM	MHVJHM
17	AAF69334	Mouse hepatitis virus strain Penn 97-1	MHVP
18	AAF69344	Mouse hepatitis virus strain ML-10	MHVM
19	NP_045300	Mouse hepatitis virus	MHV
20	AAU04646	SARS coronavirus civet007	civet007
21	AAU04649	SARS coronavirus civet010	civet010
22	AAU04664	SARS coronavirus civet020	civet020
23	AAV91631	SARS coronavirus A022	A022
24	AAV49730	SARS coronavirus B039	B039
25	AAP51227	SARS coronavirus GD01	GD01
26	AAS00003	SARS coronavirus GZ02	GZ02
27	AAP30030	SARS coronavirus BJ01	BJ01
28	AAP13567	SARS coronavirus CUHK-W1	CUHK-W1
29	AAP37017	SARS coronavirus TW1	TW1
30	AAR87523	SARS coronavirus TW2	TW2
31	BAC81348	SARS coronavirus TWH genomic RNA	TWH
32	BAC81362	SARS coronavirus TWJ genomic RNA	TWJ
33	AAQ01597	SARS coronavirus Taiwan TC1	TaiwanTC1
34	AAQ01609	SARS coronavirus Taiwan TC2	TaiwanTC2
35	AAP97882	SARS coronavirus Taiwan TC3	TaiwanTC3
36	AAP13441	SARS coronavirus Urbani	Urbani
37	AAP72986	SARS coronavirus HSR 1	HSR1
38	AAQ94060	SARS coronavirus AS	AS
39	AAP94737	SARS coronavirus CUHK-AG01	CUHK-AG01
40	AAP94748	SARS coronavirus CUHK-AG02	CUHK-AG02
41	AAP94759	SARS coronavirus CUHK-AG03	CUHK-AG03
42	AAP30713	SARS coronavirus CUHK-Su10	CUHK-Su10
43	AAP33697	SARS coronavirus Frankfurt 1	Frankfurt1
44	AAR14803	SARS coronavirus PUMC01	PUMC01
45	AAR14807	SARS coronavirus PUMC02	PUMC02
46	AAR14811	SARS coronavirus PUMC03	PUMC03
47	AAP41037	SARS coronavirus TOR2	TOR2
48	AAP50485	SARS coronavirus FRA	FRA
49	AAR23250	SARS coronavirus Sin01-11	Sino1-11
50	AHX00731	MERS coronavirus	KFU-HKU1
51	AHX00711	MERS coronavirus	KFU-HKU13
52	AHX00721	MERS coronavirus	KFU-HKU19Dam
53	AIY60578	MERS coronavirus	Abu-Dhabi_UAE_9
54	AIY60568	MERS coronavirus	Abu-Dhabi_UAE_33
55	AIZ74417	MERS coronavirus	Hu-France(UAE)-FRA1

Table 5. Cont.

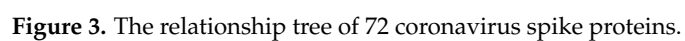
NO.	Accession Number	Virus Name/Strain	Abbreviation
56	AIZ74433	MERS coronavirus	Hu-France-FRA2
57	ALJ54502	MERS coronavirus	Hu/Qunfidhah-KSA-Rs1338
58	AKN24821	MERS coronavirus	KFMC-1
59	AKN24830	MERS coronavirus	KFMC-7
60	ALJ76282	MERS coronavirus	Hu/Taif, KSA-2083
61	ALJ76281	MERS coronavirus	Hu/Taif, KSA-5920
62	ALJ54493	MERS coronavirus	Hu/Makkah-KSA-728
63	ALB08267	MERS coronavirus	KOREA/Seoul/014-1
64	ALB08278	MERS coronavirus	KOREA/Seoul/014-2
65	ALR69641	MERS coronavirus	D2731.3
66	AKQ21055	MERS coronavirus	ADFC-A-HKU1
67	AKQ21064	MERS coronavirus	ADFC-A-HKU2
68	AKQ21073	MERS coronavirus	ADFC-A-HKU3
69	ALA50001	MERS coronavirus	camel/Taif/T68
70	ALA50012	MERS coronavirus	camel/Taif/T89
71	ALT66813	MERS coronavirus	Jordan_1
72	ALT66802	MERS coronavirus	Jordan_10

The corresponding phylogenetic tree constructed by our method is shown in Figure 3. Observing Figure 3, we find that TGEVG, TGEV, PEDVC, PEDV and HCoV-229E, which belong to group alpha, are clearly clustered together, and so do the three gamma coronaviruses IBV, IBVBJ, IBVC. In the subtree of the group beta, MERS-CoVs appear to cluster together, and SARS-CoVs are situated at an independent branch, while BCoV, BCoV-M, BCoV-Q, BCoV-E, BCoV-L, HCoV-OC43, MHV, MHV-A, MHV-M, MHV-P and MHV-JHM form a separate branch. The resulting cluster agrees well with the established taxonomic groups.

3.2. Experiment II: Identification of DNA-Binding Proteins

Numerous biological mechanisms depend on nucleic acid-protein interactions. The first step for understanding these mechanisms is to identify the interacting molecules. There are different strategies for determining DNA sequences that bind specifically to a known protein. However, it is difficult to accurately identify DNA-binding proteins [50]. Existing experimental techniques have low practical value due to time consumption and expensive costs [51]. Therefore, developing an efficient computational approach for identifying DNA-binding proteins is becoming increasingly important. In this section, we explore the application of the G-PseAAC to the identification of DNA-binding proteins. The parameters λ , w_1 , and w_2 used here are the same as those determined in Section 3.1.

The dataset used here is taken from [51]. Its original version was created in 2009 by Kumar et al. [52], in which the DNA-binding proteins are extracted from the Pfam database [53] with keywords of “DNA-binding domain” and pairwise sequence identity cutoff of 25%, while the non DNA-binding domains are randomly selected from Pfam protein families that are unrelated to the DNA-binding protein family. Xu et al. [51] removed some sequences from the original dataset, and its current version is composed of 1585 protein sequences. This benchmark dataset contains 770 DNA-binding proteins and 815 non DNA-binding proteins, which form the positive sample set and negative sample set, respectively. We randomly divide the 770 DNA-binding proteins into two parts, one has 410 sequences and the other 360 sequences. Also, we randomly select 410 and 405 sequences from the 815 non DNA-binding proteins, respectively. We conduct two sets of data. Set I contains 410 DNA-binding proteins and 410 non DNA-binding proteins. This set serves as a training set. The remaining protein sequences (360 DNA-binding proteins and 405 non DNA-binding proteins) form Set II, which serves as a test set.



Support vector machine (SVM) is employed as the classifier, and its implementation is based on the package LIBSVM (a Library for Support Vector Machines) v3.17 [54], which is open sourced and can be freely downloaded from <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. There are four types of kernel functions in LIBSVM: linear kernel; polynomial kernel; radial basis function (RBF) kernel; and sigmoid kernel. Among them, the RBF kernel is deemed a reasonable first choice [55]. The main reason is that, taking the form $K(V_i, V_j) = e^{-\gamma \|V_i - V_j\|^2}$, the RBF kernel can non-linearly map samples into a higher dimensional space so it can handle the non-linearly separable data. Accordingly, the RBF kernel is also adopted in this paper. The model selection of this kernel involves two parameters to be decided: the penalty parameter C and the kernel parameter γ . We first convert each of the 1585 protein sequences into a 28-D vector, and then the vectors belonging to Set I are scaled and fed to the SVM. With an optimization procedure using a grid search strategy in LIBSVM, the parameter pair (C, γ) is determined as (8, 0.5) (It should be pointed out that the optimal values for one round of cross-validation may not be the same for another.). In literature, a set of metrics are often used to measure the prediction quality. To make it intuitive and easy to understand for readers, here we adopt the definition and notations used in [40,41,56–60] to describe the corresponding evaluation metrics:

$$\begin{aligned}
 S_n &= 1 - \frac{N_{-}^{+}}{N^{+}}, \\
 S_p &= 1 - \frac{N_{+}^{-}}{N^{-}}, \\
 Acc &= 1 - \frac{N_{-}^{+} + N_{+}^{-}}{N^{+} + N^{-}}, \\
 MCC &= \frac{1 - \left(\frac{N_{-}^{+}}{N^{+}} + \frac{N_{+}^{-}}{N^{-}} \right)}{\sqrt{\left(1 + \frac{N_{-}^{+} - N_{+}^{-}}{N^{+}} \right) \left(1 + \frac{N_{+}^{-} - N_{-}^{+}}{N^{-}} \right)}}, \\
 F1_M &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},
 \end{aligned}$$

where N^{+} is the total number of DNA-binding proteins investigated, while N_{-}^{+} the number of DNA-binding proteins incorrectly predicted to be of non DNA-binding proteins; N^{-} the total number of non DNA-binding proteins investigated, while N_{+}^{-} the number of non DNA-binding proteins incorrectly predicted as DNA-binding proteins. Precision = $\frac{N^{+} - N_{-}^{+}}{N^{+} - N_{-}^{+} + N_{+}^{-}}$, Recall = $1 - \frac{N_{+}^{-}}{N^{-}}$. It should be pointed out that the set of metrics above is valid only for the single-label system (such as the case at hand). For the multi-label systems whose existence has become more frequent in system biology [61–64] and system medicine [65], a completely different set of metrics as defined in [66] is needed.

With the best pair (C, γ) obtained in the training stage, Set II is fed to the SVM. We find that $N_{-}^{+} = 17$ and $N_{+}^{-} = 22$. We thus have

$$S_n = 95.28\%, S_p = 94.57\%, Acc = 94.90\%, MCC = 0.8978, F1_M = 94.62\%.$$

Repeating the above random division procedure three times, we perform three cross-validation tests and list the results in Table 6. As can be seen, the accuracy (Acc), Matthew's correlation coefficient (MCC), and F1-measure ($F1_M$) in each cross-validation test are greater than 94.90%, 0.8977, and 94.59%, respectively. This result indicates that our method is promising in identifying DNA-binding proteins.

Table 6. The results of three different cross-validation tests.

Test	1	2	3	Average
S_n (%)	95.28	94.72	95.00	95.00
S_p (%)	94.57	95.06	95.06	94.90
Acc (%)	94.90	94.90	95.03	94.94
MCC	0.8978	0.8977	0.9004	0.8986
F1_M (%)	94.62	94.59	94.73	94.65

4. Discussion

4.1. Selection of Properties for Amino Acids

In addition to the three physical–chemical properties mentioned above, both hydrophilicity and molecular weight of amino acids can play important roles for characterization of proteins. Therefore, one can consider r -combinations of the five properties to describe a protein sequence. The purpose of this paper is to find an appropriate way for converting a protein sequence of 20 kinds of amino acids into a string over a “small” alphabet. If we take r to be 3, by the scheme described in Section 2, the triple $(t(1), t(2), t(3))$ has at most $2^3 = 8$ different forms. This means that the 20 native amino acids can thus be classified into no more than eight groups, whereas if the 5-combination or 4-combination is selected, by the similar scheme, $(t(1), t(2), \dots, t(r))$ will have $2^5 = 32$ or $2^4 = 16$ possible forms. Compared with “20,” the figure is not “small.” Therefore, r is taken to be 3 in this paper. By means of each of the 3-combinations of the five properties, the same experiments are performed. As a result, we find that hydrophobicity, isoelectric point, and relative distance form the best 3-combination.

4.2. Feature Analysis

As we see from Equations (8) and (9), the 28-D feature vector consists of three parts: 20 amino acid compositions; 7 correlation factors; and 1 ALE-index. One may be interested in knowing whether or not the last two parts are significant. First and foremost, let us see what would happen if only the first part was used? Without loss of generality, suppose S is a protein sequence and the counts of 20 native amino acids are n_1, n_2, \dots, n_{20} , respectively. Then we have a multi-set $M(S) = \{n_1 \cdot A, n_2 \cdot C, \dots, n_{20} \cdot Y\}$. Based on the knowledge of combinatorics, it is not difficult to see that there are a total of $\frac{|S|!}{n_1! \cdot n_2! \cdot \dots \cdot n_{20}!} = \frac{(n_1 + n_2 + \dots + n_{20})!}{n_1! \cdot n_2! \cdot \dots \cdot n_{20}!}$ different sequence/strings possessing the same amino acid composition. This suggests that the amino acid composition alone is not sufficient to represent and compare protein sequences. What would happen if only the first two parts were used (i.e., without using the ALE-index)? By using the vector with the first 27 components, experiments I and II are performed. For the first dataset, there is no significant difference between the tree constructed with the 27-D vector and that with the 28-D vector. For the second dataset, the corresponding relationship tree of coronavirus spike proteins is shown in Figure 4. From Figure 4, it is easy to see that MERS-CoVs belonging to *Betacoronavirus* appear to cluster together with the three *Gammacoronaviruses*, instead of the other *Betacoronaviruses*. This phenomenon is disappointing. For the third dataset, we repeat the three cross-validation tests with the 27-D vector and list the corresponding results in Table 7. By comparing Table 7 with Table 6, we can find that the prediction quality diminished slightly. These results indicate that the ALE-index can make a very positive contribution to the performance of experiments.

Table 7. Results of the three cross-validation tests with the 27-D vector.

Test	1	2	3	Average
S_n (%)	95.00	93.61	94.44	94.35
S_p (%)	94.32	94.32	95.06	94.57
Acc (%)	94.64	93.99	94.78	94.47

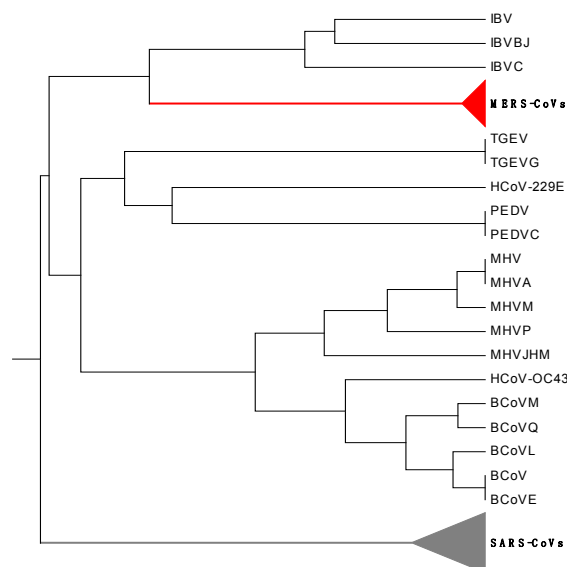


Figure 4. The relationship tree of the coronavirus spike proteins with the 27-D vector.

5. Conclusions

By means of three important physicochemical properties of amino acids, we first classify the 20 native amino acids into six groups, and assign to each group a representative symbol. Then, by substituting each letter with its representative letter, we convert a protein primary sequence into a six-letter sequence, which can be regarded as a coarse-grained description of the protein primary sequence. In comparison with the string composed of 20 kinds of amino acids, the reduced sequence not only makes the generalization from representations of DNA sequences to those of proteins easier, but also enables us to focus more on the information of our interest. On the basis of the six-letter sequence, we obtain a generalized adjacency matrix (GAM) and then its normalized ALE-index. Also, we extract λ order-correlated factors via the reduced sequence. Combining these elements with the frequencies of occurrence of 20 native amino acids, we construct a $(21 + \lambda)$ dimensional vector to characterize a protein sequence. Our method is tested by phylogenetic analysis and identification of DNA-binding proteins. The feature analysis implies that the $\lambda + 1$ components beyond the amino acid composition play very important roles in the performance of the experiment. As shown in a series of recent publications (see, e.g., [58,67–72]) in demonstrating new methods or approaches, user-friendly and publicly accessible web-servers will significantly enhance their impacts [73]. We will make efforts in our future work to further improve our method and provide a web-server for the new method presented.

Acknowledgments: The authors wish to thank the four anonymous referees for their valuable suggestions and support. The authors also thank Ren Zhang, University of Wollongong, Australia, for helpful discussions. This work was partially supported by the National Natural Science Foundation of China (No. 11171042), the Program for Liaoning Innovative Research Team in University (LT2014024), the Natural Science Foundation of Liaoning Province (201602005), and the Liaoning Bai Qian Wan Talents Program (2012921060).

Author Contributions: Chun Li and Yan-Xia Lin conceived the study and wrote the paper. Xueqin Li participated in the design of the study and analysis of the results.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Randic, M.; Vracko, M.; Nandy, A.; Basak, S.C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235–1244. [[CrossRef](#)] [[PubMed](#)]
2. Yao, Y.H.; Yan, S.; Han, J.; Dai, Q.; He, P.-A. A novel descriptor of protein sequences and its application. *J. Theor. Biol.* **2014**, *347*, 109–117. [[CrossRef](#)] [[PubMed](#)]

3. Hamori, E.; Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.* **1983**, *258*, 1318–1327. [[PubMed](#)]
4. Hamori, E. Novel DNA sequence representations. *Nature* **1985**, *314*, 585–586. [[CrossRef](#)] [[PubMed](#)]
5. Gates, M.A. Simpler DNA sequence representations. *Nature* **1985**, *316*, 219. [[CrossRef](#)] [[PubMed](#)]
6. Jeffrey, H.J. Chaos game representation of gene structure. *Nucleic Acids Res.* **1990**, *18*, 2163–2170. [[CrossRef](#)] [[PubMed](#)]
7. Nandy, A. A new graphical representation and analysis of DNA sequence structure: I Methodology and application to globin genes. *Curr. Sci.* **1994**, *66*, 309–314.
8. Nandy, A. Graphical representation of long DNA sequences. *Curr. Sci.* **1994**, *66*, 821.
9. Leong, P.M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci.* **1995**, *11*, 503–507. [[CrossRef](#)] [[PubMed](#)]
10. Zhang, R.; Zhang, C.T. Z curves, an intuitive tool for visualizing and analyzing DNA sequences. *J. Biomol. Str. Dyn.* **1994**, *11*, 767–782. [[CrossRef](#)] [[PubMed](#)]
11. Zhang, R.; Zhang, C.T. A brief review: The Z-curve theory and its application in genome analysis. *Curr. Genomics* **2014**, *15*, 78–94. [[CrossRef](#)] [[PubMed](#)]
12. Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* **2003**, *371*, 202–207. [[CrossRef](#)]
13. Randic, M.; Novic, M.; Plavsic, D. Milestones in graphical bioinformatics. *Int. J. Quantum Chem.* **2013**, *113*, 2413–2446. [[CrossRef](#)]
14. Li, C.; Fei, W.C.; Zhao, Y.; Yu, X.Q. Novel graphical representation and numerical characterization of DNA sequences. *Appl. Sci.* **2016**, *6*, 63. [[CrossRef](#)]
15. Sen, D.; Dasgupta, S.; Pal, I.; Manna, S.; Basak, S.C.; Nandy, A. Intercorrelation of major DNA/RNA sequence descriptors—A preliminary study. *Curr. Comput. Aided Drug Des.* **2016**, *12*, 216–228. [[CrossRef](#)] [[PubMed](#)]
16. Feng, Z.P.; Zhang, C.T. A graphic representation of protein sequence and predicting the subcellular locations of prokaryotic proteins. *Int. J. Biochem. Cell Biol.* **2002**, *34*, 298–307. [[CrossRef](#)]
17. Randic, M. 2-D Graphical representation of proteins based on virtual genetic code. *SAR QSAR Environ. Res.* **2004**, *15*, 147–157. [[CrossRef](#)] [[PubMed](#)]
18. Randic, M.; Zupan, J.; Balaban, A.T. Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.* **2004**, *397*, 247–252. [[CrossRef](#)]
19. Randic, M.; Butina, D.; Zupan, J. Novel 2-D graphical representation of proteins. *Chem. Phys. Lett.* **2006**, *419*, 528–532. [[CrossRef](#)]
20. Randic, M.; Zupan, J.; Balaban, A.T.; Vikić-Topić, D.; Plavsic, D. Graphical representation of proteins. *Chem. Rev.* **2011**, *111*, 790–862. [[CrossRef](#)] [[PubMed](#)]
21. Novic, M.; Randic, M. Representation of proteins as walks in 20-D space. *SAR QSAR Environ. Res.* **2008**, *19*, 317–337.
22. Aguero-Chapin, G.; Gonzalez-Diaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; Gonzalez-Diaz, Y. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. *FEBS Lett.* **2006**, *580*, 723–730. [[CrossRef](#)] [[PubMed](#)]
23. Li, C.; Xing, L.L.; Wang, X. 2-D graphical representation of protein sequences and its application to coronavirus phylogeny. *BMB Rep.* **2008**, *41*, 217–222. [[CrossRef](#)] [[PubMed](#)]
24. Nandy, A.; Ghosh, A.; Nandy, P. Numerical characterization of protein sequences and application to voltage-gated sodium channel α subunit phylogeny. *Silico Biol.* **2009**, *9*, 77–87.
25. Ghosh, A.; Nandy, A. Graphical representation and mathematical characterization of protein sequences and applications to viral proteins. *Adv. Protein Chem. Struct. Biol.* **2011**, *83*, 1–42. [[PubMed](#)]
26. Sun, D.D.; Xu, C.R.; Zhang, Y.S. A novel method of 2D graphical representation for proteins and its application. *MATCH Commun. Math. Comput. Chem.* **2016**, *75*, 431–446.
27. Qi, Z.H.; Jin, M.Z.; Li, S.L.; Feng, J. A protein mapping method based on physicochemical properties and dimension reduction. *Comput. Biol. Med.* **2015**, *57*, 1–7. [[CrossRef](#)] [[PubMed](#)]
28. Randic, M.; Balaban, A.T. On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532–539. [[CrossRef](#)] [[PubMed](#)]
29. Li, C.; Yang, Y.; Jia, M.D.; Zhang, Y.Y.; Yu, X.Q.; Wang, C.Z. Phylogenetic analysis of DNA sequences based on k -word and rough set theory. *Physica A* **2014**, *398*, 162–171. [[CrossRef](#)]

30. Randic, M.; Guo, X.F.; Basak, S.C. On the characterization of DNA primary sequences by triplet of nucleic acid bases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619–626. [[CrossRef](#)] [[PubMed](#)]
31. Randic, M.; Vracko, M. On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599–606. [[CrossRef](#)] [[PubMed](#)]
32. Li, C.; Wang, J. New invariant of DNA sequences. *J. Chem. Inf. Model.* **2005**, *36*, 115–120. [[CrossRef](#)] [[PubMed](#)]
33. Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Bioinform.* **2001**, *43*, 246–255. [[CrossRef](#)] [[PubMed](#)]
34. Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10–19. [[CrossRef](#)] [[PubMed](#)]
35. 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. [[CrossRef](#)] [[PubMed](#)]
36. Du, P.; Gu, S.; Jiao, Y. PseAAC-General: Fast building various modes of general form of Chou's pseudo amino acid composition for large-scale protein datasets. *Int. J. Mol. Sci.* **2014**, *15*, 3495–3506. [[CrossRef](#)] [[PubMed](#)]
37. Chou, K.C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteom.* **2009**, *6*, 262–274. [[CrossRef](#)]
38. Kabir, M.; Hayat, M. iRSpot-GAEnsC: Identifying recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. *Mol. Genet. Genom.* **2016**, *291*, 285–296. [[CrossRef](#)] [[PubMed](#)]
39. Tahir, M.; Hayat, M. iNuc-STNC: A sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC. *Mol. Biosyst.* **2016**, *12*, 2587–2593. [[CrossRef](#)] [[PubMed](#)]
40. Chen, W.; Feng, P.M.; Lin, H.; Chou, K.C. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **2013**, *41*, e68. [[CrossRef](#)] [[PubMed](#)]
41. Qiu, W.R.; Xiao, X.; Chou, K.C. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* **2014**, *15*, 1746–1766. [[CrossRef](#)] [[PubMed](#)]
42. Li, L.Q.; Yu, S.J.; Xiao, W.D.; Li, Y.S.; Huang, L.; Zheng, X.Q.; Zhou, S.W.; Yang, H. Sequence-based identification of recombination spots using pseudo nucleic acid representation and recursive feature extraction by linear kernel SVM. *BMC Bioinform.* **2014**, *15*, 340. [[CrossRef](#)] [[PubMed](#)]
43. Chen, W.; Lei, T.Y.; Jin, D.C.; Chou, K.C. PseKNC: A flexible web-server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* **2014**, *456*, 53–60. [[CrossRef](#)] [[PubMed](#)]
44. Chen, W.; Zhang, X.; Brooker, J.; Lin, H.; Zhang, L.Q.; Chou, K.C. PseKNC-General: A cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* **2015**, *31*, 119–120. [[CrossRef](#)] [[PubMed](#)]
45. Chen, W.; Lin, H.; Chou, K.C. Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences. *Mol. Biosyst.* **2015**, *11*, 2620–2634. [[CrossRef](#)] [[PubMed](#)]
46. Liu, B.; Liu, F.; Wang, X.L.; Chen, J.; Fang, L.; Chou, K.C. Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [[CrossRef](#)] [[PubMed](#)]
47. Grantham, R. Amino acid difference formula to help explain protein. *Science* **1974**, *185*, 862–864. [[CrossRef](#)] [[PubMed](#)]
48. Ma, F.; Wu, Y.T.; Xu, X.F. Correlation analysis of some physical chemistry properties among genetic codons and amino acids. *J. Anhui Agric. Univ.* **2003**, *30*, 439–445.
49. Li, C.; Wang, J.; Zhang, Y.; Wang, J. Similarity analysis of protein sequences based on the normalized relative entropy. *Comb. Chem. High Throughput Scr.* **2008**, *11*, 477–481. [[CrossRef](#)]
50. Hagarat, N.; Francois, J.C.; Praseuth, D. Modern tools for identification of nucleic acid-binding proteins. *Biochimie* **2008**, *90*, 1265–1272. [[CrossRef](#)] [[PubMed](#)]
51. Xu, R.F.; Zhou, J.Y.; Liu, B.; Yao, L.; He, Y.L.; Zou, Q.; Wang, X.L. enDNA-Prot: Identification of DNA-binding proteins by applying ensemble learning. *Biomed. Res. Int.* **2014**, *2014*, 294279. [[CrossRef](#)] [[PubMed](#)]
52. Kumar, K.K.; Pugalenth, G.; Suganthan, P.N. DNA-Prot: Identification of DNA binding proteins from protein sequence information using random forest. *J. Biomol. Struct. Dyn.* **2009**, *26*, 679–686. [[CrossRef](#)] [[PubMed](#)]

53. Sonnhammer, E.L.; Eddy, S.R.; Durbin, R. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* **1997**, *28*, 405–420. [[CrossRef](#)]
54. Chang, C.C.; Lin, C.J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [[CrossRef](#)]
55. Hsu, C.W.; Chang, C.C.; Lin, C.J. A Practical Guide to Support Vector Classification. Available online: <https://www.csie.ntu.edu.tw/~cjlin/libsvm> (accessed on 17 August 2014).
56. Lin, H.; Deng, E.Z.; Ding, H.; Chen, W.; Chou, K.C. iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* **2014**, *42*, 12961–12972. [[CrossRef](#)] [[PubMed](#)]
57. Liu, B.; Fang, L.; Long, R.; Lan, X.; Chou, K.C. iEnhancer-2L: A two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* **2016**, *32*, 362–369. [[CrossRef](#)] [[PubMed](#)]
58. Jia, J.; Zhang, L.; Liu, Z.; Xiao, X.; Chou, K.C. pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. *Bioinformatics* **2016**, *32*, 3133–3141. [[CrossRef](#)] [[PubMed](#)]
59. Chen, W.; Feng, P.; Ding, H.; Lin, H.; Chou, K.C. Using deformation energy to analyze nucleosome positioning in genomes. *Genomics* **2016**, *107*, 69–75. [[CrossRef](#)] [[PubMed](#)]
60. Chen, W.; Tang, H.; Ye, J.; Lin, H.; Chou, K.C. iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* **2016**, *5*, e332.
61. 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. [[CrossRef](#)] [[PubMed](#)]
62. 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. [[CrossRef](#)] [[PubMed](#)]
63. 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. [[CrossRef](#)] [[PubMed](#)]
64. 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. [[CrossRef](#)] [[PubMed](#)]
65. 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. [[CrossRef](#)] [[PubMed](#)]
66. Chou, K.C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* **2013**, *9*, 1092–1100. [[CrossRef](#)] [[PubMed](#)]
67. Qiu, W.R.; Sun, B.Q.; Xiao, X.; Xu, Z.C.; Chou, K.C. iPTM-mLys: Identifying multiple lysine PTM sites and their different types. *Bioinformatics* **2016**, *32*, 3116–3123. [[CrossRef](#)] [[PubMed](#)]
68. Qiu, W.R.; Sun, B.Q.; Xiao, X. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget* **2016**, *7*, 44310–44321. [[CrossRef](#)] [[PubMed](#)]
69. Qiu, W.R.; Xiao, X.; Xu, Z.H.; Chou, K.C. iPhos-PseEn: Identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. *Oncotarget* **2016**, *7*, 51270–51283. [[CrossRef](#)] [[PubMed](#)]
70. Chen, W.; Ding, H.; Feng, P.; Lin, H.; Chou, K.C. iACP: A sequence-based tool for identifying anticancer peptides. *Oncotarget* **2016**, *7*, 16895–16909. [[CrossRef](#)] [[PubMed](#)]
71. Jia, J.; Liu, Z.; Xiao, X.; Liu, B.X.; Chou, K.C. iCar-PseCp: Identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. *Oncotarget* **2016**, *7*, 34558–34570. [[CrossRef](#)] [[PubMed](#)]
72. Xiao, X.; Ye, H.X.; Liu, Z.; Jia, J.H.; Chou, K.C. iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. *Oncotarget* **2016**, *7*, 34180–34189. [[CrossRef](#)] [[PubMed](#)]
73. Chou, K.C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **2015**, *11*, 218–234. [[CrossRef](#)] [[PubMed](#)]

