A Novel Pseudo Amino Acid Composition for Predicting Subcellular Location of Proteins

Wangren Qiu and Xuan Xiao
Department of Computer, Jingdezhen Ceramic Institute, Jingdezhen 333001, P.R. China
Email: {qiuone@163.com, xiaoxuan0326@yahoo.com.cn}

Lidong Wang
Department of Mathematics, Dalian Maritime University, Dalian 116026, P.R. China
Email: wld1979@yahoo.com.cn

Dianxuan Gong
College of Sciences, Hebei Polytechnic University, Tangshan, 063009, P. R. China
Email: dxgong@heut.edu.cn

***Abstract*—Information on subcellular localization of proteins plays a vitally important role in molecular cell biology, proteomics and drug discovery. In this field, finding the most suitable representation for protein sample is one of the most crucial procedures. Inspired by the modes of pseudo amino acid composition (PAA), cellular automaton image (CAI) for protein and the chaos game representation (CGR) for DNA sequence, a 20-dimension CGR-walk mode for representation of protein sample is proposed. In the proposed model, the sequence order effect is discussed and manifested with a point of the 20-dimension space. And then, the track of protein sample is projected to all of the twenty amino acids, in another word, a protein sample is expressed by a 20-dimension vector. Followed with the preparation work, the proposed mode is applied into four protein datasets. The comparison results indicate that the present method may at least serve as an alternative to the existing predictors in this field.**

***Index Terms*—Chaos game representation model, protein sequence, pseudo amino acid composition, fuzzy K-nearest neighbor**

I. Introduction

For the reason that the attributes of protein sequence are closely correlated with its structures, functions and roles in biological processes, many scientists analyzed subcellular localization of protein sequence in a variety of ways. As we know, there are twenty different amino acids in protein sequence, and amino acid sequence is closely related to the biological function of protein. Its change often leads to the change of biological function of protein. The closer the genetic relationship is, the smaller the difference in amino acid composition between them will be [1]. In a sense, the dynamical folding process and stable structure, or native conformation, of a protein is determined by its primary structure, namely its amino acid sequence. Therefore, it is a great challenging and interesting issue to obtain the information on protein from its arrangement order of amino acid. To deal with the issue, the crucial procedure is to formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted [2].

Over the past thirty years, a large number of researchers have been studying the feature of protein sequence and proposed their representations. Two kinds of models are usually used to represent protein sample [2]. One is the sequential model, and the other is the discrete model. The most straightforward sequential model for a protein sample is its entire amino acid sequence. However, its application must be aimed at the sequence-similarity-search-based tools, such as BLAST [3], and this approach failed to work when the query protein did not have significant sequence similarity to any attribute-known proteins. As regards to the discrete model, many methods have been proposed to represent a protein sample. The simplest one is the amino acid (AA) composition or AAC presented by Nakashima et al. in 1986. To avoid completely losing the sequence-order information in using the AAC-discrete model, Chou proposed some different discrete models, or the so-called “pseudo amino acid composition’’ (PseAAC) [4-5]. Xiao presented the complexity measure factor mode and cellular automaton image (CAI) mode [6-8]. Meanwhile, to represent, investigate and visually reveal the patterns of DNA sequences, Jeffrey proposed the chaos game representation (CGR) for DNA sequences. The correlation properties of coding and noncoding DNA sequences were studied by Peng et al in their fractal landscape or DNA walk model [9, 25]. Because the DNA walk model is proposed to study the effects of correlation of DNA sequences on long-range correlations, some improved models were also proposed for the representation of protein [1].

Since the prediction is influenced together by the representation of the protein, the given benchmark dataset, the prediction algorithm and evaluation criterion, we will continue the research enlightened by modes of pseudo

doi:10.4304/jcp.8.3.764-771

amino acid composition (PseAAC) and chaos game representation (CGR) for DNA sequences. In this paper, we construct a similar CGR-walk model based on the similar HP model for protein sample, and a 20-dimension CGR-walk model for representation of protein sample is introduced to predict subcelllualr localization. The prediction results have been compared with the three modes, i.e. amino acid composition, Chou's PseAAC and Xiao's CAI in follows.

II. Materials and methods

In this section, for **e**xperimenting and testifying the proposed method on predicting subcellular localization, four datasets are listed and followed with three sequential representations for the entire amino acid sequence of protein sample.

Table I
NUMBER OF PROTEINS IN EACH OF THE SUBCELLULAR LOCATIONS FOR FOUR DATASETS

Orga nism	Subset	Number of Proteins	Number of location sites covered and entries in each site
Viru s	A	204	7 (26, 11, 12, 6, 83, 10, 56)
	B	180	7 (7, 3, 4, 1, 134, 3, 28)
plant	A	265	11 (2, 124, 41, 2, 10, 37, 28, 2, 8, 6, 5)
	B	406	11 (10, 80, 60, 16, 36, 59, 57, 14, 32, 23, 19)
Gpos	A	232	5 (3, 117, 47, 1, 64)
	B	220	5 (11, 79, 61, 4, 65)
Gneg	A	653	8 (152, 76, 12 , 6, 186, 6, 103, 112)
	B	643	8 (210, 20, 4, 1, 345 , 1, 13 , 49)

A. Dataset

The four protein datasets are viral [10], plant [11], Gram-positive bacterial [12] and Gram-negative bacterial [13], respectively, and each one of the datasets consist of two subsets shown in Table I. The first dataset is classified into 7 subcellular locations with respect to their host and virus-infected cells according to the experimental annotations. It consists of 384 viral proteins, of which (1) 33 are in cytoplasm, (2) 14 in endoplasmic reticulum, (3) 16 in extracell, (4) 7 in inner capsid, (5) 217 in nucleus, (6) 13 in outer capsid, and (7) 84 in plasma membrane. According to the experimental annotations, the second dataset is classified into 11 subcellular locations. And it consists of 671 protein sequences, of which 12 belong to cell wall, 204 to chloroplast, 101 to cytoplasm, 18 to endoplasmic reticulum, 46 to extracell, 96 to mitochondrion, 85 to nucleus, 16 to peroxisome, 40 to plasma membrane, 29 to plastid, and 24 to vacuole. The third dataset consists of 452 Gram-positive bacterial proteins and is classified into 5 subcellular locations according to the experimental annotations. Of the dataset, 14 belong to cell wall, 196 to cytoplasm, 108 to extracell, 5 to periplasm and 129 to plasma membrane. According to the experimental annotations, the last dataset is classified into 8 subcellular locations. It consists of 1296, of which 362 in cytoplasm, 96 in extracell, 16 in fimbrium, 7 in flagellum, 531 in inner membrane, 7 in nucleoid, 116 in outer membrane and 161 in periplasm. All of above dataset were constructed with rigorous cutoff thresholds by Shen and Chou (Refer to http: //www.csbio.sjtu.edu.cn/bioinf/).

B. Representation of Protein Samples

By extracting different features from protein sequences, various discrete modes were proposed, for example, amino acid composition (AAC) discrete model, Chou's pseudo amino acid composition (PseAAC) and Xiao's complexity measure factor based on cellular automata image. The three concepts of PseAAC have been widely used to study various problems in proteins and protein-related systems, such as predicting enzymes and their family/sub-family classification. In this paper, the three discrete representations are chosen for the comparison with the proposed method, and they are briefly formulated as follows.

(1) AA composition discrete model

The protein sequence is composed of 20 different kinds of native amino acids, namely Alanine (A), Arginine (R), Asparagine (N), Aspartic acid (D), Cysteine (C), Glutamic acid (E), Glutamine (Q), Glycine (G), Histidine (H), Isoleucine (I), Leucine (L), Lysine (K), Methionine (M), Phenylalanine (F), Proline (P), Serine (S), Threonine (T), Tryptophan (W), Tyrosine (Y) and Valine (V). The simplest discrete representation is based on the amino acid (AA) composition. The AA composition discrete model can be formulated as follows.

Given a protein sequence P with *L* amino acid residues,

$$P = R_1 R_2 R_3 \cdots R_{i-1} R_i R_{i+1} \cdots R_L \tag{1}$$

Where R_1 represents the first residue, R_2 represents the second residue, and so forth. According to the AA model, the protein P of (1) can be expressed by

$$P = (f_1, f_2, \cdots, f_{20})^T \tag{2}$$

where f_u (u = 1, 2, . . . , 20) are the normalized occurrence frequencies of the 20 native amino acids in protein P and T is the transposing operator. The AA composition discrete model has been widely used for predicting the structural class of proteins and their other attributes. However, from (2), it is clearly that all of the sequence order effects are lost by using the AA composition discrete model. This is the main shortcoming of the AA composition discrete model.

(2) PseAA composition

To avoid losing the sequence order information completely, the concept of pseudo amino acid composition (PseAA composition) was proposed first by Chou [8], and many efforts have been made in improving it [4, 15-17]. According to the typical PseAA composition discrete model, the protein P of (1) can be formulated as

$$P = (p_1, p_2, \cdots, p_{20}, p_{20+1}, p_{20+2}, \cdots, p_{20+\lambda})^T, (\lambda < L) \tag{3}$$

where

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\lambda} \tau_k}, & (1 \le u \le 20) \\ \dfrac{\omega \tau_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{k=1}^{\lambda} \tau_k}, & (21 \le u \le 20 + \lambda) \end{cases} \tag{4}$$

ω is the weight factor and τ_k is the kth tier correlation factor, which reflects the sequence order correlation among all of the *k*th most contiguous residues as formulated by

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,i+k} \quad (k < L) \tag{5}$$

with

$$J_{i,i+k} = \tfrac{1}{3}((H_1(R_{i+k}) - H_1(R_i))^2 + (H_2(R_{i+k}) - H_2(R_i))^2 + (H_3(R_{i+k}) - H_3(R_i))^2) \tag{6}$$

where $H_1(R_i)$, $H_2(R_i)$, and $H_3(R_i)$ are the hydrophobicity value, hydrophilicity value and side chain mass for the amino acid R_i respectively. Note that before substituting the values of hydrophobicity, hydrophilicity, and side chain mass into (6), they all are subjected to a standard conversion, as described by the following equation:

$$\begin{cases} H_1(R_i) = \dfrac{H_1^0(R_i) - < H_1^0 >}{SD(H_1^0)} \\ H_2(R_i) = \dfrac{H_2^0(R_i) - < H_2^0 >}{SD(H_2^0)} \\ H_3(R_i) = \dfrac{H_3^0(R_i) - < H_3^0 >}{SD(H_3^0)} \end{cases} \tag{7}$$

where the symbols $H_1^0(R_i)$ and $H_2^0(R_i)$ are the original hydrophobicity and hydrophilicity values for R_i, and $H_3^0(R_i)$ is the side chain mass for R_i which are shown in Appendix A as well as other seven chemical and physical properties of amino acid. In (7) the symbol "< >" means taking the average of the quantity therein over 20 native amino acids, and SD means the corresponding standard deviation. The converted values obtained by (7) will have a zero mean value over the 20 native amino acids, and will remain unchanged if they go through the same conversion procedure again. As we can see from (3-7), the first 20 components in (3) are associated with the conventional amino acid composition of P, whereas the remaining components (p_{20+1}, p_{20+2}, $\cdots$, $p_{20+\lambda}$) are the λ correlation factors that reflect the first tier, second tier, and so forth up to the λ th tier sequence order correlation patterns. It is these additional λ factors that approximately incorporate the sequence order effects. Note that λ is a parameter of integer and that choosing a different integer for λ will lead to a dimension different PseACC. λ =5 in this study.

(3) Complexity measure factor based on cellular automata image (CAI)

It is very difficult to find protein's characteristic vector particularly when the sequence is very long. To cope with this situation and contain the lost information of order effects, complexity measure factor based on cellular automata images was proposed by Xiao [8]. The images are derived from the amino acid sequence through the space-time evolution of cellular automata. At the first step, the 20 amino acids are coded in a binary mode as given in Table II, which can reflect the chemical and physical properties of an amino acid better, as well as its structure and degeneracy. Through the above encoding procedure, a protein sequence is transformed to a serial of digital signals. For example, the sequence "MASAA..." is transformed to "100111100101001 1100111001...".

We adopt the circulating boundary condition, with the iterative formula given below:

$$\begin{cases} D(i,j) = F(D(i-1,j-1), D(i-1,j), D(i-1,j+1)) \; (2 \le i \le n, 2 \le j \le 5L-1) \\ D(i,1) = F(D(i-1,5L), D(i-1,1), D(i-1,2)) \quad (2 \le i \le n) \\ D(i,5L) = F(D(i-1,5L-1), D(i-1,5L), D(i-1,1)) \quad (2 \le i \le n) \end{cases} \tag{8}$$

where, D (1: n, 1: 5L) is a two-dimensional (2D) array to present the amino acid sequence image, the first row of array D deposit the protein 01 sequence after digital coding, F is the iterative rule, *n* is the iterative time. Data derived by the process with the evolving rule is saved in the rows starting from the second, and data in each row is derived from those in its previous row.

TABLE II.
THREE DIFFERENT TYPES FOR CODING AMINO ACIDS

Type	Coding				
Character	P	L	Q	H	R
Decimal	1	3	4	5	6
Binary	00001	00011	00100	00101	00110
Character	S	F	Y	W	C
Decimal	9	11	12	14	15
Binary	01001	01011	01100	01110	01111
Character	T	I	M	K	N
Decimal	16	18	19	20	21
Binary	10000	10010	10011	10100	10101
Character	A	V	D	E	G
Decimal	25	26	28	29	30
Binary	11001	11010	11100	11101	11110

The evolution rule for image formation must be able to obviously distinguish whether the proteins concerned are similar to each other or not. With plentiful experiments, the 84th rule is found be the best one in serving such a purpose among all the 256 kinds of evolving rules. The time that the rule evolves determines the width of the images. It was found that the image structure is basically steady when the time is 100. When the 2D array (matrix) was transformed into an image with visualization techniques, the predicating subcelluar localization is transformed into image recognition. Since the protein images are saved in 2D arrays, every row of gene images is a 01 sequence. The Ziv-Lempel complexity of these 01 sequences then is regarded as pseudo amino acid composition. Moreover, the Ziv-Lempel complexity of a sequence can be measured by the minimal number of steps required for its synthesis in a certain process.

There are 100 complexities if the image has 100 rows, and these complexities all can be regarded as pseudo amino acid components. However, the best predict accuracy can be gained under the first 5 complexities used in plentiful experiments. Thus, by following exactly the same procedure as [4], a protein can be expressed by a vector or a point in a 25 dimensional space; i.e.

$$X = (x_1, x_2, x_3, \cdots, x_{25})^T \tag{9}$$

where x_i (i=1, 2, ..., 20) are the occurrence frequencies of the 20 amino acids in the protein, arranged alphabetically according to their single letter codes, x_j (j=21, 22, ..., 25) are the complexity measure factors for the protein sequence, T represents the transpose operator.

III. CGR-walk Model for the Representation of Protein Sample

Since each coding sequence in the complete genome of an organism can be translated into a protein sequence by using the genetic code and the secondary and the space structures of a protein are determined by its amino acid sequence. In order to discover the correlation among in amino acids, we link all protein sequences according to the order of the coding sequence, and present a novel representation of protein based on CGR-walk model in 20-dimension.

The Chaos Game is an algorithm which allows one to produce pictures of fractal structures. However, as a 20-D image, the "picture" of a protein cannot be viewed. We then represent it in 20-D vector. The Chaos Game model for obtaining representation of protein sequence still can, in simplest form, be proceed as follows:

(1). Locate 20 dots in 20-dimensional space. The dots must be orthogonally distributing in the space to character the points effectively. We call these dots vertices. For example, let $(\overline{c}_1, \overline{c}_2, ..., \overline{c}_{20})$ represents the set of 20 amino acid (P, L, H, R, F, Y, C, S, Q, W, T, I, M, E, N, A, K, G, D, V) which can better reflect the chemical and physical properties of an amino acid, the 20 points then are denoted as:

$$\overline{c}_i = (c_1^i, c_2^i, ..., c_{20}^i) \tag{10}$$

where $c_j^i = 1$ when $i = j$ and $c_j^i = 0$ when $i \neq j$ $i, j = 1, 2, ..., 20$.

(2). Pick a point as the origin, and mark it as CGR_0 , i.e. the coordinate of the point is (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0). The moving point then is rolling according to the representation of amino acids in the protein although its initial position is the initial point.

(3). For the given protein sequence shown in (1) with length L, every amino acid of the protein sequence corresponds to a vertex. In the 20-dimension space, let CGR_k denotes the site of the moving point after k times rolling and $CGR_k = CGR_{k-1} + 0.4 * w_i * (CGR_{k-1} - \overline{c}_i)$ when R_k is $\overline{c}_i$. P_k^i denotes the projected length of the ith amino acids after k times rolling. The relation between these variables is summarized in following.

$$\mathrm{P}_k^i = \begin{cases} \mathrm{P}_{k-1}^i + 0.4 * w_i * \left\| CGR_{k-1} - \overline{c}_i \right\| & R_k \quad is \quad \overline{c}_i \\ \mathrm{P}_{k-1}^i & R_k \; isnot \, \overline{c}_i \end{cases} \tag{11}$$

where $P_0^i = 0$, $w_i = \dfrac{0.9*(H_l{}^i\text{-min}H_l)}{\text{max}H_l\text{-min}H_l} + \dfrac{\text{mean}(|H_l|)}{\text{max}(|H_l|)}$, $i = 1, 2, ..., 20$, $k = 1, 2, ..., L$. $l = 1, 2, ..., 10$. maxH_l and minH_l are the maxim and minimum of vector respectively. H_1 represents hydrophobicity; H_2 represents hydrophilicity; H_3 represents side-chain mass; H_4 represents pK1 (alpha-COOH); H_5 represents pK2 (NH3); H_6 represents PI; H_7 represents average volume of buried residue; H_8 represents molecular weight; H_9 represents side chain volume; H_{10} represents mean polarity.

(4). After the moving point has been rolled L times according to the given protein sequence, the protein can be expressed by a vector or a point in a 20 dimensional space, i.e.

$$P_L = (p_L^1, p_L^2, p_L^3, \cdots, p_L^{20})^T \tag{12}$$

Since the site CGR_k on which the moving point stood at time k is related with all of the previous k-1 amino acids which would affect those of all amino acids followed with the kth amino acid, the vector $P_L = (p_L^1, p_L^2, p_L^3, \cdots, p_L^{20})^T$ is the summation weighted projected length of each amino acid. Therefore, P_L reflects the order effects of the whole protein sequence as well as the information of occurrence frequency. This effect also will be testified and presented with the experimental results in the following section.

IV. Experimental Method and Steps

A. Prediction Algorithm, Measure and Test Method

There are many different prediction algorithms introduced to address for predicting subcellular localization, such as discriminant algorithm [19], neural network algorithm [20], genetic algorithm [21], support vector machine (SVM) [22], and K-nearest Neighbor algorithm [10, 12]. For the reason that the K-nearest neighbor (KNN) classifier has good performance and simple-to-use feature. Fuzzy K-nearest neighbor classifier gains more accuracy than the conventional KNN besides having the above merits. Thus, we shall focus on the fuzzy K-nearest neighbor algorithm in this paper.

As regard to the measure for the "nearness" of the fuzzy KNN classifier, there are many different definitions, such as Euclidean distance, and Mahalanobis distance [19]. Although the Mahalanobis distance may give rise to high accuracy, it has higher levity than Euclidean distance. In this study, we only concern the effects of the pseudo amino acid compositions on predicting subcellualr localization, not those of distance or algorithms. For the simple computation, the Euclidean distance is the preferable measure for the experiment with fuzzy K-nearest neighbor classifier in this paper.

For the reason that, in jackknife test, all the proteins in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining protein samples, and each protein sample will be in turn moved between the training and testing dataset. The outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset. So, the jackknife test can exclude the "memory" effect and avoid the arbitrariness problem met in the independent dataset test and subsampling test. Furthermore, the jackknife test has been increasingly and widely used by investigators to examine the quality of various predictors. Jackknife test then is the ideal one for this study.

B. Experimental Steps

Four datasets are parted into two subsets A_i and B_i $(i = 1, 2, 3, 4)$. There are 204, 265, 232, and 653 elements in A_i $(i = 1, 2, 3, 4)$, respectively. Correspondingly, there are 180, 406, 220, and 643 elements in B_i $(i = 1, 2, 3, 4)$, respectively.

All A_i and B_i $(i = 1, 2, 3, 4)$ are the testing datasets without distinguishing them as learning dataset or testing datasets mentioned in literatures [10-12].

Every protein sequence has been expressed as ten vectors according to the ten chemical and physical properties of amino acid, i.e. H_l ($l = 1, 2, ..., 10$), by CGR-walk model as well as AA composition discrete model (AA), Chou's PseAA composition and Xiao's complexity measure factor based on cellular automata image. Then the thirteen representations of a protein sequence are tested by fuzzy *K*-nearest neighbor classifier with *K*=1, *K*=2, ..., *K*=10 with the test method of jackknife test. The classify accuracies are listed and discussed in the next section.

TABLE III.
COMPARISON RESULTS OF THE JACKKNIFE TEST ON SUBSET A OF VIRAL PROTEIN

k	1	2	3	4	5	6	7	8	9	10
$CGRH_1$	81	78	80	**78**	**78**	75	74	73	74	73
$CGRH_2$	83	78	**81**	**78**	76	75	**76**	74	74	74
$CGRH_3$	**84**	78	78	74	74	74	74	73	72	72
$CGRH_4$	83	79	79	**78**	77	76	74	74	74	74
$CGRH_5$	83	78	78	**78**	76	75	75	**75**	74	**75**
$CGRH_6$	83	**81**	80	77	77	74	**76**	74	**75**	74
$CGRH_7$	**84**	79	79	75	75	75	74	73	72	72
$CGRH_8$	83	78	78	75	73	76	73	73	74	74
$CGRH_9$	**84**	78	79	76	74	**77**	75	73	73	72
CH_{10}	82	79	**81**	77	77	**77**	75	**75**	73	72
AAC	83	79	79	**78**	74	74	73	72	73	73
PAA	83	79	79	**78**	74	74	74	72	73	73
CAI	83	79	79	**78**	74	74	73	73	73	73
Best.	1	1	1	1	1	1	1	1	1	1

V. RESULTS AND DISCUSSION

Given the prediction method and test method described in the above sections, the datasets are experimented by using fuzzy *K*-nearest neighbor classifier with different *K*. We will outline the results in detail and followed with the discussion in this section.

TABLE IV.
COMPARISON RESULTS OF THE JACKKNIFE TEST ON SUBSET B OF VIRAL PROTEIN

k	1	2	3	4	5	6	7	8	9	10
$CGRH_1$	83	81	83	82	83	82	82	81	81	79
$CGRH_2$	83	83	83	82	82	83	82	81	79	78
$CGRH_3$	83	**84**	**84**	**83**	83	83	82	**82**	82	**81**
$CGRH_4$	82	**84**	83	82	83	82	**83**	**82**	**83**	**81**
$CGRH_5$	82	81	82	82	82	83	**83**	**82**	82	**81**
$CGRH_6$	81	83	**84**	82	83	82	**83**	81	81	**81**
$CGRH_7$	82	83	**84**	**83**	**84**	**84**	**83**	**82**	82	**81**
$CGRH_8$	83	**84**	**84**	82	83	83	**83**	**82**	82	**81**
$CGRH_9$	82	83	**84**	**83**	**84**	**84**	**83**	**82**	82	**81**
CH_{10}	**84**	82	83	**83**	83	83	**83**	81	81	79
AAC	83	81	83	82	82	82	**83**	81	81	**81**
PAA	83	81	83	82	82	82	**83**	81	81	**81**
CAI	83	81	83	82	83	82	**83**	81	81	**81**
Best.	1	1	1	1	1	1	1	1	1	1

Take the subset *A* of viral proteins as an example, the success rates of CGR by using the jackknife cross-validation test are listed in Table III of which, the second row shows the success rates with Hydrophobicity, i.e. H_1, of the amino acid, the third row shows the success rates with Hydrophilicity, i.e. H_2, of the amino acid, and so forth up to the last chemical and physical properties of amino acid. And the success rates of the three counterparts, i.e. AAC, PAA and CAI are listed in following three rows.

In this study, we the comparison of success rates are compared with fuzzy KNN, and the highest success rates of different representations with the same K are shown in bold. In the last row of Table III, the sign that the CGR mode is able to obtain the highest success rates or not be denoted as 1 or 0. With the same work, the success rates of the other seven datasets are listed in Table IV-X.

TABLE V.
COMPARISON RESULTS OF THE JACKKNIFE TEST ON SUBSET *A* OF PLANT PROTEIN

k	1	2	3	4	5	6	7	8	9	10
$CGRH_1$	40	40	43	45	47	48	47	47	48	**49**
$CGRH_2$	43	42	44	**48**	46	47	46	45	45	48
$CGRH_3$	**44**	44	45	47	**49**	48	45	46	48	47
$CGRH_4$	42	42	44	45	48	47	46	46	**49**	48
$CGRH_5$	40	41	45	44	47	46	47	48	48	48
$CGRH_6$	43	44	**48**	**48**	47	48	46	47	47	48
$CGRH_7$	**44**	44	47	47	48	47	46	46	46	46
$CGRH_8$	43	43	**48**	46	48	47	45	45	48	46
$CGRH_9$	43	**45**	**48**	45	47	48	47	46	48	47
CH_{10}	40	40	45	47	46	46	47	**49**	48	48
AAC	40	36	45	45	45	**49**	**48**	47	48	46
PAA	41	35	45	45	46	**49**	**48**	48	48	46
CAI	38	37	43	45	46	46	**48**	46	46	46
Best.	1	1	1	1	1	0	0	1	1	1

TABLE VI.
COMPARISON RESULTS OF THE JACKKNIFE TEST ON SUBSET *B* OF PLANT PROTEIN

k	1	2	3	4	5	6	7	8	9	10
$CGRH_1$	26	29	32	34	33	34	33	32	34	33
$CGRH_2$	27	28	31	33	35	34	35	36	36	36
$CGRH_3$	30	32	28	33	35	36	34	35	36	35
$CGRH_4$	30	31	29	32	35	35	35	34	33	33
$CGRH_5$	30	33	32	34	36	34	34	35	35	32
$CGRH_6$	28	31	29	31	32	34	33	35	34	35
$CGRH_7$	29	30	30	33	35	35	35	35	35	34
$CGRH_8$	30	31	28	33	36	**37**	**36**	35	36	34
$CGRH_9$	26	29	30	32	32	35	34	33	35	34
CH_{10}	27	28	29	32	31	32	31	31	33	33
AAC	**32**	**34**	**34**	**35**	**38**	36	**36**	36	36	**37**
PAA	31	**34**	**34**	**35**	37	36	**36**	**37**	36	36
CAI	29	32	33	34	35	36	**36**	36	**38**	**37**
Best.	0	0	0	0	0	1	1	0	0	0

TABLE VII.
COMPARISON RESULTS OF THE JACKKNIFE TEST ON SUBSET *A* OF GRAM-POSITIVE PROTEIN

k	1	2	3	4	5	6	7	8	9	10
$CGRH_1$	69	74	77	77	78	76	77	76	76	77
$CGRH_2$	74	79	**81**	**81**	80	**81**	**80**	80	**80**	**80**
$CGRH_3$	**76**	**80**	80	**81**	79	**81**	79	79	79	**80**
$CGRH_4$	72	**80**	**81**	77	79	78	78	78	77	77
$CGRH_5$	72	75	**81**	80	79	79	79	77	77	78
$CGRH_6$	73	78	80	**81**	**81**	80	79	**81**	79	78
$CGRH_7$	75	78	79	**81**	79	**81**	79	78	79	**80**
$CGRH_8$	**76**	**80**	80	**81**	79	80	**80**	80	79	**80**
$CGRH_9$	**76**	79	**81**	80	79	80	79	78	78	79
CH_{10}	70	73	78	78	78	78	77	77	76	76
AAC	73	78	**81**	78	80	78	**80**	80	79	78
PAA	72	78	**81**	79	80	78	**80**	79	78	78
CAI	75	78	80	79	**81**	80	**80**	78	76	76
Best.	1	1	1	1	1	1	1	1	1	1

From the tables III-X, we can see that the CGR mode gains most of the highest success rates. There is only one exception which is the experiments on subset *B* of plant protein shown in Table VI. The table also tells us that the best performance is AAC not Chou's PseAA or Xiao's CAI. Since Chou's PseAA and Xiao's CAI have concerned the sequence order lost by AAC, this fact shows that the amount of information of occurrence frequencies is greater than that of sequence order, and has drowned the last one. This also can explain the not well performance in Table IX.

TABLE VIII.
COMPARISON RESULTS OF THE JACKKNIFE TEST ON SUBSET *B* OF GRAM-POSITIVE PROTEIN

k	1	2	3	4	5	6	7	8	9	10
$CGRH_1$	**67**	64	66	68	66	69	68	68	69	68
$CGRH_2$	65	**68**	**72**	72	**73**	**74**	**75**	73	72	73
$CGRH_3$	65	65	68	69	70	71	72	72	**73**	73
$CGRH_4$	**67**	**68**	69	70	69	71	69	73	**73**	71
$CGRH_5$	64	**68**	71	**73**	**73**	71	69	70	72	70
$CGRH_6$	**67**	**68**	69	70	**73**	**74**	70	71	**73**	70
$CGRH_7$	66	67	68	68	71	72	70	72	71	72
$CGRH_8$	66	64	68	70	69	71	70	72	72	**74**
$CGRH_9$	66	**68**	69	69	70	70	71	73	70	72
CH_{10}	66	64	65	66	66	69	66	66	67	68
AAC	**67**	65	71	**73**	71	72	72	**74**	72	71
PAA	66	65	70	72	70	72	72	73	71	71
CAI	65	65	70	67	70	71	73	71	72	71
Best.	1	1	1	1	1	1	1	0	1	1

TABLE IX.
COMPARISON RESULTS OF THE JACKKNIFE TEST ON SUBSET *A* OF GRAM-NEGATIVE PROTEIN

k	1	2	3	4	5	6	7	8	9	10
$CGRH_1$	58	60	61	61	63	64	63	63	64	62
$CGRH_2$	**59**	60	62	64	64	63	64	65	**66**	65
$CGRH_3$	**59**	**62**	63	63	64	64	65	64	64	65
$CGRH_4$	57	59	63	64	64	**67**	65	65	65	65
$CGRH_5$	58	60	63	64	64	64	65	65	64	64
$CGRH_6$	56	59	63	63	65	65	65	65	65	**67**
$CGRH_7$	**59**	60	63	64	64	64	65	**66**	**66**	65
$CGRH_8$	**59**	**62**	62	63	64	64	**66**	65	65	66
$CGRH_9$	58	60	63	64	62	63	64	64	65	65
CH_{10}	57	61	63	61	63	63	62	62	61	63
AAC	58	**62**	63	**65**	**66**	66	**66**	**66**	**66**	65
PAA	58	**62**	**64**	**65**	**66**	66	65	**66**	**66**	66
CAI	58	**62**	63	**65**	65	66	65	**66**	**66**	65
Best.	1	1	0	0	0	1	1	1	1	1

TABLE X.
COMPARISON RESULTS OF THE JACKKNIFE TEST ON SUBSET *B* OF GRAM-NEGATIVE PROTEIN

k	1	2	3	4	5	6	7	8	9	10
$CGRH_1$	71	71	73	**74**	72	73	72	73	73	73
$CGRH_2$	72	71	73	**74**	74	74	74	74	74	**75**
$CGRH_3$	72	72	**74**	**74**	74	75	74	**76**	74	**75**
$CGRH_4$	71	**74**	73	**74**	**75**	**76**	73	75	74	**75**
$CGRH_5$	70	72	71	**74**	73	74	74	74	**75**	**75**
$CGRH_6$	71	73	73	73	73	73	73	73	74	**75**
$CGRH_7$	72	73	73	**74**	74	**76**	**75**	**76**	**75**	**75**
$CGRH_8$	**73**	72	**74**	**74**	74	75	74	75	**75**	**75**
$CGRH_9$	72	73	73	**74**	74	75	74	75	**75**	**75**
CH_{10}	70	71	72	73	72	72	72	72	72	71
AAC	72	72	**74**	73	74	74	**75**	74	**75**	**75**
PAA	**73**	72	**74**	73	74	74	74	74	**75**	**75**
CAI	**73**	72	**74**	73	74	74	**75**	74	**75**	**75**
Best.	1	1	1	1	1	1	1	1	1	1

The highest success rates with different *K* and chemical and physical properties of amino acids are shown in Table XI and XII respectively. As shown in Table XI, the best *K* for fuzzy *K*-nearest neighbor is 4 followed by 10, 1 and 3. From Table XII, the best performance is H_8 followed by H_7, H_3 and H_2, *and* then H_4, H_6. Since the prediction result is affected by the representation of the protein, the given benchmark dataset, these discussions may be only accounted for the fuzzy *K*-nearest neighbor classifier applied on predicting subcellular localization. However, it is still a clue for the proposed mode applied on other datasets or areas.

TABLE XI
SUM OF THE HIGHEST SUCCESS RATES WITH DIFFERENT KS

Data set		1	2	3	4	5	6	7	8	9	10
Virus	A	3	1	2	4	1	2	2	2	1	1
	B	1	3	5	4	2	2	7	6	1	7
Plant	A	2	1	3	2	1	0	0	1	1	1
	B	0	0	0	0	0	1	1	0	0	0
Gpos	A	3	3	4	5	1	3	2	1	1	4
	B	3	5	1	1	3	2	1	0	3	1
Gneg	A	4	2	0	0	0	1	1	1	2	1
	B	1	1	2	8	1	2	1	2	4	8
Sum		17	16	17	24	9	13	15	13	13	23

TABLE XII.
SUM OF THE HIGHEST SUCCESS RATES WITH DIFFERENT AMINO ACID CHARACTERS

Data set		H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}
Virus	A	2	3	1	1	3	3	1	0	2	3
	B	0	0	5	5	3	3	7	5	7	3
Plant	A	1	1	2	1	0	2	1	1	2	1
	B	0	0	0	0	0	0	0	2	0	0
Gpos	A	0	6	5	2	1	3	3	5	2	0
	B	1	5	1	3	3	5	0	1	1	0
Gneg	A	0	2	2	1	0	1	3	3	0	0
	B	1	2	4	5	3	1	6	5	3	0
Sum		5	19	20	18	13	18	21	22	17	7

VI. CONCLUSION

The chaos game representation of sequences is a method to ordinate the entire domain of possibilities in a continuous two, or higher dimensional space. The CGR transformation makes a certain sequence become an entire new set of statistical analysis tools. Therefore, CGR is a bridge between sequences of discrete units and numeric coordinates in a continuous space. Although some research has been carried out by taking into consideration sequence order and correlations in protein sequences, the methods are yet potentiality for their perfection and continual improvement. In this paper, we convert the CGR coordinates into a new mode for representing protein sample and express the protein sequence with a 20-dimension vector. And the vector is computed from the track of protein sequence which contains the sequence order as well as occurrence frequency. This is the essence why the success rates predicted by the current method are superior to those by many other methods on the same datasets.

The results indicate that the CGR approach is an effective mode for representing protein samples and might at least serve as an alternative improve the prediction quality for other protein attributes [20, 21], such as membrane [19], G-protein-coupled receptor types [6, 24] and enzyme family classes [4, 22].

ACKNOWLEDGMENT

This work was partially supported by the National Nature Science Foundation of China (No. 60961003), the Key Project of Chinese Ministry of Education (no.210116), Natural Science Foundation of Jiangxi Province, China (No. 2010GQS0127), the Doctor Startup Foundation of Liaoning Province (No. 20111030), and the Fundamental Research Funds for the Central Universities (No. 2011QN057).

REFERENCES

[1] J. Gao, L. L. Jiang, Z. Y. Xu, “Chaos game representation walk model for the protein sequences,” *Chin. Phys. Soc.*, vol.18 (10), pp. 4571-4579, April 2009.

[2] K.C. Chou, “Some remarks on protein attribute prediction and pseudo amino acid composition,” *J. of Theor. Biol.*, vol.273 (1), pp. 236-247, Mar. 2011.

[3] S. F. Altschul, “Evaluating the statistical significance of multiple distinct local alignments,” In: Suhai, S. (Ed.), *Theor. Comput. Methods in Genome Research. Plenum, New York*, 1997, pp. 1-14.

[4] K. C. Chou, “Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes,” *Bioinformatics*, vol.21 (1), pp. 10-19, Jan. 2005.

[5] K. C. Chou, D.W. Elrod, “Protein subcellular location prediction, ” *Protein Eng*., vol.12 (2), 107-118, Feb. 1999.

[6] X. Xiao, P. Wang, K. C. Chou, “GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions, ” *Mol. BioSyst.*, 7, pp. 911-919, Mar. 2011.

[7] X. Xiao, S. H. Shao, Z. D. Huang, K. C. Chou, “Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor,” *J.Comput. Chem.*, vol.27 (4), pp.478-482, Mar. 2006.

[8] X. Xiao, S. H. Shao, Y. S. Ding, Z. D. Huang, K. C. Chou, “Using cellular automata images and pseudo amino acid composition to predict protein subcellular location, ” *Amino Acids*, vol.30 (1), pp. 49-54, Feb. 2006.

[9] C. K. Peng, S. V. Buldyrev, A. L. Goldberg, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley, “Long-range correlations in nucleotide sequences,” Nature, vol.356 (6365), pp.168-170, Mar.1992.

[10] H. B. Shen, K. C. Chou, “Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells,” *Biopolymers*, vol.85 (3), pp. 233-240, Feb. 2007.

[11] K. C. Chou, H. B. Shen, “Large-scale plant protein subcellular location prediction,” *J. Cell. Biochem.* Vol.100 (3), pp. 665-678, Feb 2007.

[12] H. B. Shen, K. C. Chou, “Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins,” *Protein Eng. Des. Sel.* , vol.20 (1), pp. 39-46, Jan. 2007.

[13] K. C. Chou, H. B. Shen, “Large-scale predictions of Gram-negative bacterial protein subcellular locations,” *J. Proteome Res*., vol.5 (12), pp. 3420-3428, Dec 2006.

[14] K. C. Chou, H. B. Shen, “Review: recent progresses in protein subcellular location prediction,” *Anal. Biochem.* , *vol.*370, pp. 1-16 , Nov 2007.

[15] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, A. Torres, “Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition, ” *J. Theor. Biol.*, vol.257 (1), pp. 17-26, Mar 2009.

[16] X. Xiao, P. Wang, K. C. Chou, “Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition, ” *J. Appl. Cryst.*, vol.42 (2), pp. 169-173, April 2009.

[17] Y. H. Zeng, Y. Z. Guo, R. Q. Xiao, L.Yang, L. Z. Yu, M. L. Li, “Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach, ” *J.Theor. Biol.*, vol.259 (2), pp. 366-372, Jul. 2009.

[18] Y. L. Chen, Q. Z. Li, “Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition, ” *J.Theor. Biol.*, vol.248 (2), pp. 377-381, Sep. 2007.

[19] H. Lin, “The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition, ” *J. Theor. Biol.*, vol. 252 (2), pp. 350-356, May 2008.

[20] Y. D. Cai, X. J. Liu, K. C. Chou, “Artificial neural network model for predicting membrane protein types, ” *J. Biomol. Struct. Dynam*, vol.18 (4), pp. 607-610, Feb. 2001.

[21] Y. S. Ding, T. L. Zhang, “Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier, ” *Pattern Recog. Lett.*, vol.29 (13), pp. 1887-1892, Oct. 2008.

[22] X. B. Zhou, C. Chen, Z. C. Li, X. Y. Zou, “Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, ” *J. Theor. Biol.*, vol.248 (3), pp. 546-551, Oct. 2007.

[23] H. B. Shen, K. C. Chou, “Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition, ” *Biochem. Biophys. Res. Commun.*, vol.337 (3), pp. 752-756, NoV. 2005.

[24] J. D. Qiu, J. H. Huang, R. P. Liang, X. Q. Lu, “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.* , vol.390 (1), pp. 68-73, Jan. 2009.

[25] R. Zhu, Y. Qin and J. Wang, “Energy-aware distributed intelligent date gathering algorithm in wireless sensor networks,” International Journal of Distributed Sensor Networks, vol. 2011, Article ID 235724, pp. 1-13, 2011.

Wangren Qiu received his BS and MS degrees in Mathematics from Nanchang University, Nanchang, China, in 2000 and Dalian Maritime University, Dalian, China, in 2006, respectively.

He is currently a candidate for doctor's degree of the Research Center of Information and Control, Dalian University of Technology, Dalian, China, from 2009, and an associate professor in the Department of Computer, Jingdezhen Ceramic Institute, Jingdezhen, China. His research interests include fuzzy sets and its applications, bioinformatics, and data mining.

Xuan Xiao received his BS degrees in in Television from Nanchang University, Nanchang, China, in 1991 and college of machine of Tianjing University, Tianjing, China, in 2002, respectively. And his PhD degree in control theory and control engineering from the Institute of information of Donghua university, Shanghai, China, in 2006.

He is currently a Professor in the Department of Machine and Electron, Jingdezhen Ceramic Institute, Jingdezhen, China. His research interests include pattern recognition, bioinformatics, sensory evaluation.

Lidong Wang received his BS and MS degrees in Mathematics from Inner Mongolia University, Hohhot, China, in 2003 and Dalian University of Technology, Dalian, China, in 2006, respectively, and his PhD degree in control theory and control engineering from the Research Center of Information and Control, Dalian University of Technology, Dalian, China, in 2009.

He is currently a Lecturer in the Department of Mathematics, Dalian Maritime University, Dalian, China. His research interests include fuzzy sets and its applications, granular computing, and data mining.

Dianxuan Gong received his BS degrees in Mathematics from Qufu Normal University, Qufu, China, in 2003 and his MS and PhD degrees in Mathematics from Dalian University of Technology, Dalian, China, in 2006 and 2009, respectively, and his PhD degree in Control Theory and Control Engineering from the Research Center of Information and Control, Dalian University of Technology, Dalian, China, in 2009.

He is currently a Lecturer in the College of Sciences, Hebei Polytechnic University, Tangshan, China. His research interests include numerical computation and data mining.

Appendixe A Chemical and physical properties of amino acid

H R	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}
A	0.62	-0.5	15	2.35	9.87	6.11	91.5	89.09	27.5	-0.1
C	0.29	-1	47	1.71	10.8	5.02	117.7	121.15	44.6	1.36
D	-0.9	3	59	1.88	9.6	2.98	124.5	133.1	40	-0.8
E	-0.7	3	73	2.19	9.67	3.08	155.1	147.13	62	-0.8
F	1.19	-2.5	91	2.58	9.24	5.91	203.4	165.19	115.5	1.27
G	0.48	0	1	2.34	9.6	6.06	66.4	75.07	0	-0.4
H	-0.4	-0.5	82	1.78	8.97	7.64	167.3	155.16	79	0.49
I	1.38	-1.8	57	2.32	9.76	6.04	168.8	131.17	93.5	1.31
K	-1.5	3	73	2.2	8.9	9.47	171.3	146.19	100	-1.2
L	1.06	-1.8	57	2.36	9.6	6.04	167.9	131.17	93.5	1.21
M	0.64	-1.3	75	2.28	9.21	5.74	170.8	149.21	94.1	1.27
N	-0.8	0.2	58	2.18	9.09	10.76	135.2	132.12	58.7	-0.5
P	0.12	0	42	1.99	10.6	6.3	129.3	115.13	41.9	0
Q	-0.9	0.2	72	2.17	9.13	5.65	161.1	146.15	80.7	-0.7
R	-2.5	3	101	2.18	9.09	10.76	202	174.2	105	-0.8
S	-0.2	0.3	31	2.21	9.15	5.68	99.1	105.09	29.3	-0.5
T	-0.1	-0.4	45	2.15	9.12	5.6	122.1	119.12	51.3	-0.3
V	1.08	-1.5	43	2.29	9.74	6.02	141.7	117.15	71.5	1.09
W	0.81	-3.4	130	2.38	9.39	5.88	237.6	204.24	145.5	0.88
Y	0.26	-2.3	107	2.2	9.11	5.63	203.6	181.19	117.3	0.33

Where H_1: Hydrophobicity; H_2: Hydrophilicity; H_3: side-chain mass; H_4: pK1 (alpha-COOH); H_5: pK2 (NH3); H_6: PI; H_7: Average volume of buried residue; H_8: Molecular weight; H_9: Side chain volume; H_{10}: Mean polarity.