Amino Acids (2009) 37:741–749
DOI 10.1007/s00726-008-0212-9

ORIGINAL ARTICLE

# Application of protein grey incidence degree measure to predict protein quaternary structural types

**Xuan Xiao · Wei-Zhong Lin**

Received: 23 August 2008 / Accepted: 10 November 2008 / Published online: 27 November 2008


**Abstract** Many proteins are composed of two or more subunits, each associated with different polypeptide chains. The number and arrangement of subunits forming a protein are referred to as quaternary structure. It has been known for long that the functions of proteins are closely related to their quaternary structure. In this paper the grey incidence degree is introduced that can calculate the numerical relation between various components, expressed the similar or different degree between these components. We have demonstrated that introduction of the grey incidence degree can remarkably enhance the success rates in predicting the protein quaternary structural class. It is anticipated that the grey incidence degree can be also used to predict many other protein attributes, such as subcellular localization, membrane protein type, enzyme functional class, GPCR type, protease type, among many others.

**Keywords** Protein sequence distance measure · Grey system · Grey incidence degree · Protein quaternary structural type · Nearest neighbor algorithm

## Introduction

One key element in understanding the molecular machinery of the cell is to understand the structure and function of each protein encoded in the genome. Advancement of in vitro techniques enables availability of primary structure information of thousands of proteins. The three-dimensional conformational state (tertiary/quaternary structure) of a protein is dependent on the primary structure to a large extent.

Actually, several proteins are a combination of two or more individual polypeptide chains. The arrangement according to which such subunits assemble is called the protein quaternary structure. In the protein universe there are many different classes of subunit construction, such as monomer, dimmer, trimer, tetramer, and so forth. The oligomers may be homo-oligomers or hetero-oligomers; the former consist of identical polypeptide chains, whereas the latter are nonidentical. Biological processes are often influenced by the quaternary structure of proteins involved therein. For example, some critical ligands only bind to dimmers but not to monomers; some marvelous allosteric transitions only occur in tetramers but not other oligomers; and some ion channels are formed by tetramers, whereas others are formed by pentamers; the sodium channel is formed by a monomer (Chen et al. 2002) while the potassium channel by a homo-tetramer (Doyle et al. 1998); the M2 proton channel is formed by a homo-tetramer (Schnell and Chou 2008) while hemoglobin by a hetero-tetramer (Perutz 1964); the phospholamban is formed by homo-pentamer (Oxenoid and Chou 2005; Oxenoid et al. 2007) while the gamma-aminobutyric acid type A ($GABA_A$) receptor (Chou 2004b; Tretter et al. 1997) and α7 nicotinic acetylcholine receptor (Chou 2004a) by a hetero-pentamer. Moreover, the quaternary structural type is also very useful in screening the candidates of proteins for their three-dimensional structure determination by the X-ray cryptography technique.

Many lines of evidences have indicated that mathematical/computational approaches, such as structural bioinformatics (Chou 2004a, b, c, d, 2005a), molecular docking (Chou et al. 2003; Gao et al. 2007; Li et al. 2007;

X. Xiao (✉) · W.-Z. Lin
Computer Department, Jing-De-Zhen Ceramic Institute, 333001 Jing-De-Zhen, China
e-mail: xiaoxuan0326@yahoo.com.cn

Wang et al. 2008a; Zhang et al. 2006a, b; Zheng et al. 2007), molecular packing (Chou et al. 1984, 1988), pharmacophore modelling (Chou et al. 2006; Sirois et al. 2004), Mote Carlo simulated annealing approach (Chou 1992), diffusion-controlled reaction simulation (Chou and Zhou 1982), graph/diagram approach (Andraos 2008; Chou 1981, 1989, 1990; Chou and Forsen 1980; Chou et al. 1979; Chou and Liu 1981; Cornish-Bowden 1979; King and Altman 1956; Kuzmic et al. 1992; Myers and Palmer 1985; Zhou and Deng 1984), bio-macromolecular internal collective motion simulation (Chou 1988), QSAR (Dea-Ayuela et al. 2008; Du et al. 2005, 2008a, b; Gonzalez-Díaz et al. 2006, 2008; Prado-Prado et al. 2008), protein subcellular location prediction (Chou and Shen 2006a, c, 2007a, d, 2008a; Xiao et al. 2005), protein structural class prediction (Chou 1995, 2000; Chou and Cai 2004; Xiao et al. 2006, 2008a, c), identification of membrane proteins and their types (Chou and Shen 2007c), identification of enzymes and their functional classes (Shen and Chou 2007a), identification of GPCR and their types (Chou 2005b; Chou and Elrod 2002; Gao and Wang 2006; Xiao et al. 2008b), identification of proteases and their types (Chou and Shen 2008b), protein cleavage site prediction (Chou 1993, 1996; Shen and Chou 2008a), and signal peptide prediction (Chou and Shen 2007e; Shen and Chou 2007e) can timely provide very useful information and insights for both basic research and drug design and hence are widely welcome by science community. The present study is attempted to develop a computational approach for predicting the quaternary structural type of proteins based on their sequence information alone in hope to provide a useful tool for further stimulating the development of this area.

In fact, a number of computational methods have been developed to predict protein quaternary structures. Garian developed a method which used decision tree models and a feature extraction approach (simple binning function) to successfully predict homodimer and non-homodimer (Garian 2001). Chou and Cai also investigated this topic with a pseudo-amino acid (PseAA) composition, or PseAAC, feature extraction method to predict monomer, homodimer, homotrimer, homotetramer, homopentamer, homohexamer and homooctamer (Chou and Cai 2003). Zhang et al. (2003, 2006a, b) successfully predict homodimers and non-homodomers, homodimer, homotrimer, homotetramer, homohexamer with weighted auto-correlation functions feature extraction approach. Carugo (2007) provided a method that allowed one to predict if a chain participated in hetero-oligomeric assemblies based on amino acid composition.

Protein quaternary structural type prediction can be mapped to a standard pattern classification problem. Structural categories of proteins are considered as classes, whereas, structural and functional units of proteins are treated as patterns. The following two modes are often used to express a protein: (1) the sequential mode, and (2) the discrete mode. However, because protein sequences are extremely complicated with much variation in both sequence order and length, it is hardly to establish a feasible predictor by using the sequential mode to represent protein samples, as elaborated by Chou (Chou and Cai 2002). The simplest discrete mode is use the amino acid composition of a protein to represent it (Xiao and Chou 2007).

Accuracy of different types of classifiers depends on classification principle as well as characteristics of patterns. For example, if the classes are linearly separable, then use of minimum distance classifier may be a wise decision, whereas, it is not useful if the classes are linearly non-separable. In that case K-nearest neighbor classifier can produce better results (Ghosh and Parai 2008). In K-nearest neighbor classifier, the distances must be calculated between the new protein and the other protein in data set. Different distance measures can be used for this, e.g., Euclidian distance, city block distance, Mahalanobis distance, etc.

In 1982, Deng proposed a grey system theory to study the uncertainty of a system (Deng 1982). According to this theory, if the information of a system investigated is fully known, it is called a "white system"; if completely unknown, a "black system"; if partially known, a "grey system". It was a new theory and method applicable to the study of problems with unascertained and very few data and/or poor information. Grey incidence degree is one of the major components of the grey systems theory (Liu et al. 2005). The protein prediction is a grey system. The goal of the present study was to explore the properties of grey incidence degree in classifying protein quaternary structural type, using k-nearest sequence classification schemes. From this study, it is found that the method based on grey incidence degree performs better compared to the minimum distance classifier.

## Method

### Protein sample representation by pseudo-amino acid composition (PseAAC)

A protein sequence is generally constituted by 20 native amino acids whose single character codes are: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, M, and Y. Consider a protein chain of $L$ amino acid residues: $R_1$, $R_2$, $R_3$, $R_4$,…$R_L$, where $R_1$ represents the residue at the sequence position 1, $R_2$ at position 2, and so forth. We can expressed it as a vector in a 20-D vector (Chou 1995); that is,

$$\mathrm{P} = [p_1, p_2, \ldots, p_{20}] \quad (1)$$

where $p_1$ is the occurrence frequency of amino acid A in the protein, $p_2$ that of amino acid C, and so forth. However, using the amino acid composition to represent a protein sample as in Eq. 1 would lose all of its sequence-order information. To avoid losing the sequence-order information, a logic approach is to use the entire sequence to represent the protein sample and apply the sequence search-based tools such as BLAST (Altschul 1997; Wootton and Federhen 1993) to perform prediction. However, this kind of approach failed to work when the query protein did not have significant homology to proteins of known characteristics (Chou and Shen 2007d). In order to avoid complete losing the sequence-order information and also enable the prediction more effective for those proteins that do not have significant homology to characterized proteins, a feasible approach is to use the pseudo-amino acid (PseAA) composition to represent the protein sample. The PseAA composition (Chou 2001) was originally proposed for predicting protein subcellular localization and membrane protein type (Chou 2001); while the amphiphilic PseAA composition (Chou 2005c) proposed for predicting the enzyme functional classification. The essence of PseAA composition is to use a discrete model to represent a protein sample yet without complete losing its sequence-order information. According to its definition, the PseAA composition for a given protein sample is expressed by a set of $20 + \lambda$ discrete numbers, where the first 20 represent the 20 components of the classical amino acid composition while the additional $\lambda$ numbers incorporate some of its sequence-order information via various different kinds of coupling modes. Ever since the concept of PseAA composition was introduced, various PseAA composition approaches have been stimulated to deal with different problems in proteins and protein-related systems (see, e.g., Ding and Zhang 2008; Jiang et al. 2008; Li and Li 2008; Lin 2008; Lin et al. 2008; Zhang and Fang 2008; Zhang et al. 2008; Zhou et al. 2007). Owing to its wide usage, recently a very flexible PseAA composition generator, called "PseAAC" (Shen and Chou 2008b), was established at the website http://chou.med.harvard.edu/bioinf/PseAA/, by which users can generate 63 different kinds of PseAA composition. In the current study, we also use the PseAA composition to represent a protein sample. According to (Chou 2001), the PseAA composition of a protein P is defined by a $(20 + \lambda)$-D vector, as given by

$$\mathrm{P} = [p_1, p_2, \ldots, p_{20}, p_{20+1}, \ldots p_{20+\lambda}] \quad (2)$$

where the first 20 elements are the same as in Eq. 1, and $p_{20+j}$ $(j = 1,\ldots,\lambda)$ are the pseudo-amino acid components that represent the $j$th rank sequence-order correlation factors (see Fig. 1 of Ref. Chou 2001). Given the sequence of a protein, the $(20 + \lambda)$ elements in Eq. 2 can be easily derived via Eqs. 4–7 of Chou (2001); they can also be generated by using the web-server called "PseAAC" at http://chou.med.harvard.edu/bioinf/PseAA/.

## Nearest neighbor (NN) classifier

Although neural network based methods give higher accuracy, they suffer form some draw-backs. Black-box nature of neural networks makes it difficult to view how the structures are actually predicted. Neural based methods along with hidden Markov models perform well when many homologies of query protein are available (Karplus et al. 1998). This goes against generalization of prediction. Nearest neighbor classifiers have been successfully for predicting protein subcellular location and other attributes (see, e.g., Chou and Shen 2006a, b, 2007a, b, c; Shen and Chou 2007a, b, c, d, f).

Assume a system include $N$ proteins which are classified into $M$ subsets (GPCRs main families),

$$S = \bigcup_{m=1}^{M} \mathrm{S}_m = \{\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_N\} \quad (3)$$

where each subset $\mathrm{S}_m$ $(m = 1, 2,\ldots,M)$ is composed of proteins with the same type and its size (the number of proteins therein) is $N_m$. Obviously, $N = \sum_{m=1}^{M} N_m$. According to Eq. 2, the $i$th $(i = 1, 2,\ldots,N)$ protein in the set $S$ (see Eq. 3) is formulated by

$$\mathrm{P_i} = [p_1^i, p_2^i, \ldots, p_{20+\lambda}^i], \quad (4)$$

Similarly, a target protein (query protein) should be represented by

$$\mathrm{P_?} = [p_1^?, p_2^?, \ldots, p_{20+\lambda}^?]. \quad (5)$$

Now, for the target protein $\mathrm{P_?}$, how we identify which family it belonged to? In our study we used the K-Nearest neighbor rule to handle this problem. According to the NN rule, the target protein should be assigned to the subset containing the majority of its nearest neighbor. Owing to its good performance and simple-to-use feature, the NN rule, also named as "voting NN rule", is quite popular in pattern recognition community. There are many different definitions to measure the "nearness" for the NN classifier, such as Euclidean distance, Hamming distance, and Mahalanobis distance. The grey incidence degree between the protein sequence's can calculate the numerical relation between various components, expressed the similar or different degree between these components. Therefore, we used the degree of grey incidence to measure the nearness between the target protein $\mathrm{P_?}$ and the comparable protein $\mathrm{P}_i$.

Grey incidence degree (GID)

Assume $P = \{P_1, P_2, \ldots, P_N\}$ are the set of compared series, namely samples of protein sequence, and $P_?$ is the target sequence. The grey relational coefficient is defined as

$$\gamma(p_k^?, p_k^i) = \frac{\Delta_{\text{Min}} + \xi\Delta_{\text{Max}}}{\Delta_k^{?,i} + \xi\Delta_{\text{Max}}} \tag{6}$$

where

$$\Delta_k^{?,i} = |p_k^? - p_k^i| \tag{7}$$

$$\Delta_{\max} = \max_{\forall j} \max_{\forall k} |p_k^? - p_k^j|, \quad (j = 1, 2, \ldots, N; k = 1, 2, \ldots, 20 + \lambda) \tag{8}$$

$$\Delta_{\min} = \min_{\forall j} \min_{\forall k} |p_k^? - p_k^j|, \quad (j = 1, 2, \ldots, N; k = 1, 2, \ldots, 20 + \lambda) \tag{9}$$

$$\xi = \text{distinguishing coefficient}, \in [0, 1] \tag{10}$$

where $\in$ is symbol in the set theory meaning "member of" and the symbol $\forall$ is a logical statement denote "for every".

The grey incidence degree is actually a weighting sum of grey relational coefficient and can be derived from

$$\Gamma(P_?, P_i) = \sum_{k=1}^{20+\lambda} w_k \gamma(p_k^?, p_k^i) \tag{11}$$

The weighting factor, $w_k$, must satisfy $\sum_{k=1}^{20+\lambda} w_k = 1$. Normally, if all the process parameters are of equal weighting, the distinguishing coefficient $\xi$ is 0.5 and the grey incidence degree is performed in equal weighting fashion, $w_k = 1/(20 + \lambda), (k = 1, 2, \ldots 20 + \lambda)$(Deng 1982; Tsai et al. 2005), which is the value selected in the present study.

The grey incidence degree, $\Gamma(P_?, P_i)$, stands for the level of correlation between the target series $P_?$ and the comparable series $P_i$. Thus, if one of the comparable series influences more on the target series than the others, the corresponding value of grey incidence degree is larger than those values of the other grades. It also indicates the degree of influence on the target series exerted by the comparable series. According to Eqs. 6–10, when $P_? \equiv P_i$, we have $\Gamma(P_?, P_i) = 1$, indicating that these two proteins have perfect or 100% similarity.

Let's assume sequences: $P_0$ = [0.0173, 0.0006, 0.0037, 0.0099, 0.0068], $P_1$ = [0.0395, 0.0093, 0.0031, 0.0099, 0.0142], $P_2$ = [0.0068, 0.0185, 0.0191, 0.0074, 0.0056], $P_3$ = [0.0333, 0.0204, 0.0111, 0.0012, 0.0062], $P_4$ = [0.0150, 0.0137, 0.0136, 0.0156, 0.0138]. First, we compute the $\Delta_i = |x_i - x_0|$, $\Delta_1$ = [0.0222, 0.0087, 0.0006, 0.0, 0.0074], $\Delta_2$ = [0.0105, 0.0179, 0.0154, 0.0025, 0.0012], $\Delta_3$ = [0.0160, 0.0198, 0.0074, 0.0087, 0.0006], $\Delta_4$ = [0.0023, 0.0131, 0.0099, 0.0057, 0.0070]. So, $\Delta_{\max} = 0.0222$, $\Delta_{\min} = 0$. Second, we calculate the $\gamma_i$, $\gamma_1$ = [0.8333, 0.9273, 0.9946, 1.0000, 0.9375], $\gamma_2$ = [0.9136, 0.8611, 0.8782, 0.9780, 0.9893], $\gamma_3$ = [0.8740, 0.8486, 0.9375, 0.9273, 0.9946], $\gamma_4$ = [0.9797, 0.8944, 0.9181, 0.9512, 0.9407]. Lastly, we get the grey incidence degree: $\Gamma(x_0, x_1) = 0.9386$, $\Gamma(x_0, x_2) = 0.9240$, $\Gamma(x_0, x_3) = 0.9164$, $\Gamma(x_0, x_4) = 0.9368$.

The method that uses the grey incidence degree to analysis a system is also called grey incidence analysis.

## Results and discussion

The training dataset and independent dataset taken from Chou and Cai (2003) are used to test the current method. The training dataset consists of 3,174 protein sequences, of which 382 are with annotation of monomer, 817 of dimer, 593 of trimer, 884 of tetramer, 54 of pentamer, 287 of hexamer, and 157 of octamer. The independent dataset consists of 332 protein sequences, of which 50 are with annotation of monomer, 102 of dimer, 56 of trimer, 80 of tetramer, 6 of pentamer, of 28 hexamer, and 10 of octamer.

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test (Chou and Zhang 1995). However, as elucidated in (Chou and Shen 2008a) and demonstrated by Eq. 50 of Chou and Shen (2007d), among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors (see, e.g., Chou and Shen 2008b; Ding et al. 2007; Jiang et al. 2008; Jin et al. 2008; Kannan et al. 2008; Li and Li 2008; Lin 2008; Lin et al. 2008; Niu et al. 2008; Shi et al. 2008; Tian et al. 2008; Wang et al. 2008b; Wu and Yan 2008; Xiao and Chou 2007; Zhang and Fang 2008; Zhang et al. 2008; Zhou 1998; Zhou and Assa-Munt 2001; Zhou and Doctor 2003; Zhou et al. 2007).

During the jackknife test, each protein sample in the dataset is singled out in turn as a "test sample" and all the rule-parameters are determined from the remaining $N - 1$ samples. The success rates by jackknife test for the aforementioned 3,174 proteins classified into sever quaternary structural classes are given in Table 1, where for facilitating comparison the corresponding rates obtained by the CD (Covariant Discriminant) and Support Vector Machine are also listed. It can be seen from Table 1 that the overall success rate by the current approach is 87.3%, which is remarkably higher than those by the other approaches. It can be seen by comparing CD and GID that

**Table 1** Success rates of jackknife cross-validation with different approaches on the 3,174 proteins from Chou and Cai (2003)

| Method | Input | Monomer | Dimer | Trimer | Tetramer | Pentamer | Hexamer | Octamer | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Covariant discriminant algorithm (Chou and Cai 2003) | Pseudo amino acid composition[a] | $\frac{309}{382} = 80.9\%$ | $\frac{700}{817} = 85.5\%$ | $\frac{462}{593} = 77.9\%$ | $\frac{755}{884} = 85.4\%$ | $\frac{1}{54} = 1.9\%$ | $\frac{180}{287} = 62.7\%$ | $\frac{85}{157} = 54.1\%$ | $\frac{2492}{3174} = 78.5\%$ |
| Support Vector Machine (Zhang et al. 2007) | Pseudo amino acid composition[b] | $\frac{314}{382} = 82.2\%$ | $\frac{763}{817} = 93.4\%$ | $\frac{471}{593} = 79.4\%$ | $\frac{803}{884} = 90.8\%$ | $\frac{36}{54} = 66.7\%$ | $\frac{183}{287} = 63.8\%$ | $\frac{113}{157} = 72.0\%$ | $\frac{2683}{3174} = 84.5\%$ |
| This paper | Pseudo amino acid composition[a] | $\frac{338}{382} = 88.48\%$ | $\frac{697}{817} = 85.31\%$ | $\frac{513}{593} = 86.51\%$ | $\frac{828}{884} = 93.67\%$ | $\frac{46}{54} = 85.19\%$ | $\frac{225}{287} = 78.40\%$ | $\frac{133}{157} = 84.74\%$ | $\frac{2780}{3174} = 87.6\%$ |

[a] Using the sequence-order correlation factors for pseudo amino acid composition, $\lambda = 45$

[b] Using the multi-scale energy feature vector for pseudo amino acid composition

misallocation errors are remarkably reduced by GID against CD, particularly for pentamer and octamer proteins. The GID can manifest the similar degree of two sequences and investigate a trend among given sequences. When calculate the similar degree of two sequences, GID not only take into account the relations of every corresponding parameters but also $\Delta_{max}$ and $\Delta_{min}$ in all data set. It is why GID can improve the overall prediction success rate under the same data and pseudo-amino acid composition in predicting the protein quaternary structure type.

The performance of the prediction system can be affected by $\lambda$, the number of the sequence-order correlation factors for pseudo-amino acid composition. The results obtained using the jackknife test are shown in Fig. 1. Form Fig. 1, it is clear that compared with the case of $\lambda = 1$, the overall success rates are significantly enhanced along with increasing of $\lambda$, indicating that long-range interaction is very important in determining the protein quaternary structure type. However, the overall success rate does not always monotonously increase with $\lambda$. The highest overall success rate is 87.6 ($\lambda = 45$).

To measure the performance of predictive methods, there exist some standard statistical scoring techniques. The most frequently used measures is Matthew's correlation coefficient (MCC) indexes. The definition of MCC is given by

$$\mathrm{MCC} = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{[\mathrm{TP} + \mathrm{FP}][\mathrm{TP} + \mathrm{FN}][\mathrm{TN} + \mathrm{FP}][\mathrm{TN} + \mathrm{FN}]}} \tag{12}$$

where TP represents the true positive; TN, the true negative; FP, the false positive and FN, the false negative (Xiao

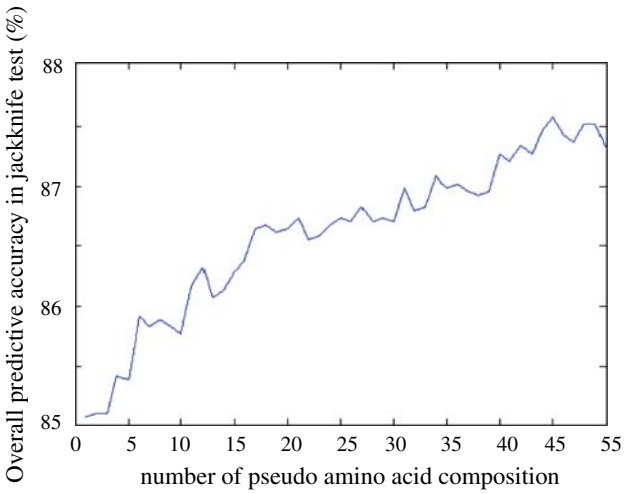


**Fig. 1** The relationship between the number of the sequence-order correlation factors for pseudo amino acid composition and the prediction accuracy in the jackknife test. Note that the highest accuracy is achieved at $\lambda = 45$

**Table 2** Success rates of independent data set test with different approaches on the 332 proteins from (Chou and Cai 2003)

| Method | Input | Monomer | Dimer | Trimer | Tetramer | Pentamer | Hexamer | Octamer | Overall |
|---|---|---|---|---|---|---|---|---|---|
| Covariant discriminant algorithm (Chou and Cai 2003) | Pseudo amino acid composition[a] | $\frac{35}{50} = 70.0\%$ | $\frac{85}{102} = 83.3\%$ | $\frac{44}{56} = 78.6\%$ | $\frac{72}{80} = 90.0\%$ | $\frac{1}{6} = 16.7\%$ | $\frac{21}{28} = 75.0\%$ | $\frac{8}{10} = 80.0\%$ | $\frac{266}{332} = 80.1\%$ |
| This paper | Pseudo amino acid composition[a] | $\frac{44}{50} = 88.0\%$ | $\frac{82}{102} = 80.1\%$ | $\frac{48}{56} = 85.7\%$ | $\frac{76}{80} = 95.0\%$ | $\frac{4}{6} = 66.7\%$ | $\frac{27}{28} = 96.4\%$ | $\frac{10}{10} = 100\%$ | $\frac{291}{332} = 87.7\%$ |

[a] Using the sequence-order correlation factors for pseudo amino acid composition, $\lambda = 45$

et al. 2008a, b, c). The MCC indexes for the sever quaternary structural classes obtained by the jackknife tests with the GID predictor are 0.9223, 0.8633, 0.8610, 0.8359, 0.9117, 0.7234, 0.8040, respectively. It can be seen that the results obtained by the current predictor not only possess higher success rates but also are more stable than those by the CD approach, indicating that the new approach is indeed very powerful and promising.

Moreover, as a demonstration for practical application, predictions were also performed for the 332 independent proteins, based on the rule-parameters derived from the training data set. The predicted results thus obtained are summarized in Table 2, the overall success prediction rate is 87.7%. This is 8% higher than the rate by the CD based on the same pseudo-amino acid composition.

The functional domain composition has been used for predicting protein quaternary structure type (Yu et al. 2006). There is a limitation when prediction based on functional domain composition, how to predict a sequence without functional domain composition information. This paper represents a new attempt to predict these proteins. It is a complementary to functional domain composition method.

## Conclusion

The grey system theory was developed to deal with those systems for which only partial information is available, and hence is particularly useful to deal with biological problems. It is demonstrated in this study that the overall success rate in predicting protein quaternary structural classes can be remarkably improved by using the GID. It has not escaped our notice that the similar approach can be also used to deal with many other complicated problems in biology, such as predicting protein subcellular localization, membrane protein type, enzyme functional class, GPCR type, signal peptides, protease type, among many others.

**Acknowledgments** The author is indebted to Professor Dr. Chou for illuminating discussion. The work in this research was supported by the grants from the National Natural Science Foundation of China (No. 60661003), the Province National Natural Science Foundation of Jiangxi (No. 0611060), and the plan for training youth scientists (stars of Jing-Gang) of province Jiangxi.

## References

Altschul SF (1997) Evaluating the statistical significance of multiple distinct local alignments. In: Suhai S (ed) Theoretical and computational methods in genome research. Plenum, New York, pp 1–14

Andraos J (2008) Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Can J Chem 86:342–357

Carugo O (2007) Applied Crystallography 40:986–989
Chen Z, Alcayaga C, Suarez-Isla BA, O'Rourke B, Tomaselli G, Marban E (2002) A "minimal" sodium channel construct consisting of ligated S5-P-S6 segments forms a toxin-activatable ionophore. J Biol Chem 277:24653–24658
Chou KC (1981) Two new schematic rules for rate laws of enzyme-catalyzed reactions. J Theor Biol 89:581–592
Chou KC (1988) Review: low-frequency collective motion in biomacromolecules and its biological functions. Biophys Chem 30:3–48
Chou KC (1989) Graphical rules in steady and non-steady enzyme kinetics. J Biol Chem 264:12074–12079
Chou KC (1990) Review: applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys Chem 35:1–24
Chou KC (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. J Mol Biol 223:509–517
Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 268:16938–16948
Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins Struct Funct Genet 21:319–344
Chou KC (1996) Review: prediction of HIV protease cleavage sites in proteins. Anal Biochem 233:1–14
Chou KC (2000) Review: prediction of protein structural classes and subcellular locations. Curr Protein Pept Sci 1:171–208
Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Genet 43:246–255 (Erratum: ibid, 2001, vol 44, 60)
Chou KC (2004a) Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. Biochem Biophys Res Commun 319:433–438
Chou KC (2004b) Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. Biochem Biophys Res Commun 316:636–642
Chou KC (2004c) Molecular therapeutic target for type-2 diabetes. J Proteome Res 3:1284–1288
Chou KC (2004d) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134
Chou KC (2005a) Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. J Proteome Res 4:1681–1686
Chou KC (2005b) Prediction of G-protein-coupled receptor classes. J Proteome Res 4:1413–1418
Chou KC (2005c) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19
Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277:45765–45769
Chou KC, Cai YD (2003) Predicting protein quaternary structure by pseudo amino acid composition. Proteins Struct Funct Genet 53:282–289
Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321:1007–1009 (Corrigendum: ibid., 2005, Vol.329, 1362)
Chou KC, Elrod DW (2002) Bioinformatical analysis of G-protein-coupled receptors. J Proteome Res 1:429–433
Chou KC, Forsen S (1980) Graphical rules for enzyme-catalyzed rate laws. Biochem J 187:829–835
Chou KC, Liu WM (1981) Graphical rules for non-steady state enzyme kinetics. J Theor Biol 91:637–654
Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347:150–157
Chou KC, Shen HB (2006b) Large-scale predictions of Gram-negative bacterial protein subcellular locations. J Proteome Res 5:3420–3428
Chou KC, Shen HB (2006c) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5:1888–1897
Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6:1728–1734
Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. J Cell Biochem 100:665–678
Chou KC, Shen HB (2007c) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360:339–345
Chou KC, Shen HB (2007d) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16
Chou KC, Shen HB (2007e) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357:633–640
Chou KC, Shen HB (2008a) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162
Chou KC, Shen HB (2008b) ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem Biophys Res Commun 376:321–325
Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349
Chou KC, Zhou GP (1982) Role of the protein outside active site on the diffusion-controlled reaction of enzyme. J Am Chem Soc 104:1409–1413
Chou KC, Jiang SP, Liu WM, Fee CH (1979) Graph theory of enzyme kinetics: 1. Steady-state reaction system. Sci Sin 22:341–358
Chou KC, Nemethy G, Scheraga HA (1984) Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices. J Am Chem Soc 106:3161–3170
Chou KC, Maggiora GM, Nemethy G, Scheraga HA (1988) Energetics of the structure of the four-alpha-helix bundle in proteins. Proc Natl Acad Sci USA 85:4295–4299
Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. Biochem Biophys Res Commun 308:148–151 (Erratum: ibid., 2003, Vol.310, 675)
Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ (2006) Review: progress in computational approach to drug development against SARS. Curr Med Chem 13:3263–3270
Cornish-Bowden A (1979) Fundamentals of enzyme kinetics, Chap. 4. Butterworths, London
Dea-Ayuela MA, Perez-Castillo Y, Meneses-Marcel A, Ubeira FM, Bolas-Fernandez F, Chou KC, Gonzalez-Diaz H (2008) HP-Lattice QSAR for dynein proteins: Experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence. Bioorg Med Chem 16:7770–7776
Deng JL (1982) Control problems of grey systems. Syst Control Lett 1(5):288–294
Ding YS, Zhang TL (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recognit Lett 29:1887–1892
Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14:811–815
Doyle DA, Morais CJ, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R (1998) The structure of the potassium

channel: molecular basis of $K^+$ conduction and selectivity. Science 280:69–77
Du QS, Mezey PG, Chou KC (2005) Heuristic molecular lipophilicity potential (HMLP): a 2D-QSAR study to LADH of molecular family pyrazole and derivatives. J Comput Chem 26:461–470
Du QS, Huang RB, Chou KC (2008a) Review: recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Curr Protein Pept Sci 9:248–259
Du QS, Huang RB, Wei YT, Du LQ, Chou KC (2008b) Multiple field three dimensional quantitative structure-activity relationship (MF-3D-QSAR). J Comput Chem 29:211–219
Gao QB, Wang ZZ (2006) Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 19(11):511–516
Gao WN, Wei DQ, Li Y, Gao H, Xu WR, Li AX, Chou KC (2007) Agaritine and its derivatives are potential inhibitors against HIV proteases. Med Chem 3:221–226
Garian R (2001) Prediction of quaternary structure from primary structure. Bioinformatics 17:551–556
Ghosh A, Parai B (2008) Protein secondary structure prediction using distance based classifiers. Int J Approx Reason 47:37–44
Gonzalez-Diaz H, Sanchez-Gonzalez A, Gonzalez-Diaz Y (2006) 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. J Inorg Biochem 100:1290–1297
Gonzalez-Díaz H, Gonzalez-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks, and connectivity indices. Proteomics 8:750–778
Jiang X, Wei R, Zhang TL, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept Lett 15:392–396
Jin Y, Niu B, Feng KY, Lu WC, Cai YD, Li GZ (2008) Predicting subcellular localization with AdaBoost learner. Protein Pept Lett 15:286–289
Kannan S, Hauth AM, Burger G (2008) Function prediction of hypothetical proteins without sequence similarity to proteins of known function. Protein Pept Lett 15:1107–1116
Karplus K, Barrett C, Hughey R (1998) Hidden markov models for detecting remote preotein homologies. Bioinformatics 14:846–856
King EL, Altman C (1956) A schematic method of deriving the rate laws for enzyme-catalyzed reactions. J Phys Chem 60:1375–1378
Kuzmic P, Ng KY, Heath TD (1992) Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. Anal Biochem 200:68–73
Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept Lett 15:612–616
Li Y, Wei DQ, Gao WN, Gao H, Liu BN, Huang CJ, Xu WR, Liu DK, Chen HF, Chou KC (2007) Computational approach to drug design for oxazolidinones as antibacterial agents. Med Chem 3:576–582
Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252:350–356
Lin H, Ding H, Feng-Biao Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 15:739–744
Liu SF, Fang ZG, Lin Y (2005) A new definition for the degree of grey incidence. Sci Inq 7(2):111–124
Myers D, Palmer G (1985) Microcomputer tools for steady-state enzyme kinetics. Bioinformatics (original: Comput Appl Biosci) 1:105–110
Niu B, Jin YH, Feng KY, Liu L, Lu WC, Cai YD, Li GZ (2008) Predicting membrane protein types with bagging learner. Protein Pept Lett 15:590–594
Oxenoid K, Chou JJ (2005) The structure of phospholamban pentamer reveals a channel-like architecture in membranes. Proc Natl Acad Sci USA 102:10870–10875
Oxenoid K, Rice AJ, Chou JJ (2007) Comparing the structure and dynamics of phospholamban pentamer in its unphosphorylated and pseudo-phosphorylated states. Protein Sci 16:1977–1983
Perutz MF (1964) The hemoglobin molecule. Sci Am 211:65–76
Prado-Prado FJ, Gonzalez-Diaz H, de la Vega OM, Ubeira FM, Chou KC (2008) Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorg Med Chem 16:5871–5880
Schnell JR, Chou JJ (2008) Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451:591–595
Shen HB, Chou KC (2007a) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59
Shen HB, Chou KC (2007b) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20:39–46
Shen HB, Chou KC (2007c) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355:1006–1011
Shen HB, Chou KC (2007d) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel 20:561–567
Shen HB, Chou KC (2007e) Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Commun 363:297–303
Shen HB, Chou KC (2007f) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85:233–240
Shen HB, Chou KC (2008a) HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. Anal Biochem 375:388–390
Shen HB, Chou KC (2008b) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388
Shi MG, Huang DS, Li XL (2008) A protein interaction network analysis for yeast integral membrane protein. Protein Pept Lett 15:692–699
Sirois S, Wei DQ, Du QS, Chou KC (2004) Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. J Chem Inf Comput Sci 44:1111–1122
Tian F, Lv F, Zhou P, Yang Q, Jalbout AF (2008) Toward prediction of binding affinities between the MHC protein and its peptide ligands using quantitative structure-activity relationship approach. Protein Pept Lett 15:1033–1043
Tretter V, Ehya N, Fuchs K, Sieghart W (1997) Stoichiometry and assembly of a recombinant GABAA receptor subtype. J Neurosci 17:2728–2737
Tsai L, Liou HY, Jiang GF (2005) Application of grey relational analysis to the influential factors on natural frequencies of helical springs. J Grey Syst 8(2):141–156
Wang JF, Wei DQ, Chen C, Li Y, Chou KC (2008a) Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. Protein Pept Lett 15:27–32
Wang T, Yang J, Shen HB, Chou KC (2008b) Predicting membrane protein types by the LLDA algorithm. Protein Pept Lett 15:915–921
Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem 17:149–163
Wu G, Yan S (2008) Prediction of mutations in H3N2 hemagglutinins of influenza a virus from North America based on different datasets. Protein Pept Lett 15:144–152

Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Pept Lett 14:871–875

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28:57–61

Xiao X, Shao SH, Huang ZD, Chou KC (2006) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27:478–482

Xiao X, Lin WZ, Chou KC (2008a) Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J Comput Chem 29:2018–2024

Xiao X, Wang P, Chou KC (2008b) GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J Comput Chem (in press)

Xiao X, Wang P, Chou KC (2008c) Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J Theor Biol 254:691–696

Yu X, Wang C, Li Y (2006) Classification of protein quaternary structure by function domain composition. BMC Bioinformatics 7:187

Zhang GY, Fang BS (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J Theor Biol 253:310–315

Zhang SW, Pan Q, Zhang HC, Zhang YL, Wang HY (2003) Classification of protein quaternary structure with support vector machine. Bioinformatics 19:2390–2396

Zhang R, Wei DQ, Du QS, Chou KC (2006a) Molecular modeling studies of peptide drug candidates against SARS. Med Chem 2:309–314

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006b) Prediction protein homooligomer types by pesudo amino acid composition: approached with an improved feature extraction and Naive Bayes feature fusion. Amino Acids 30:461–468

Zhang SW, Chen W, Zhao CH, Cheng YM, Pan Q (2007) Predicting protein quaternary structure with multi-scale energy of amino acid factor solution scores and their combination. Lecture Notes in Computer Science, pp 65–72

Zhang GY, Li HC, Fang BS (2008) Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein Pept Lett 15:1132–1137

Zheng H, Wei DQ, Zhang R, Wang C, Wei H, Chou KC (2007) Screening for new agonists against Alzheimer's disease. Med Chem 3:488–493

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins Struct Funct Genet 44:57–59

Zhou GP, Deng MH (1984) An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochem J 222:169–176

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins Struct Funct Genet 50:44–48

Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248:546–551