OXFORD UNIVERSITY PRESS | Bioinformatics

# Predict Remote Homology Proteins by Merging Grey Incidence Analysis and Domain Similarity Analysis

| | |
|---|---|
| Journal: | *Bioinformatics* |
| Manuscript ID | Draft |
| Category: | Original Paper |
| Date Submitted by the Author: | n/a |
| Complete List of Authors: | Lin, Weizhong; Jingdezhen Ceramic Institute, School of Information Engineering<br>Xiao, Xuan; Jingdezhen Ceramic Institute<br>Qiu, Wang-Ren; Jingdezhen Ceramic Institute<br>Chou, Kuo-Chen; University of Electronic Science and Technology of China |
| Keywords: | Protein sequence analysis, Information extraction, Bioinformatics, Remote homology proteins prediction |
| | |

*Bioinformatics*, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

*Subject Section*

# Predict Remote Homology Proteins by Merging Grey Incidence Analysis and Domain Similarity Analysis

Wei-Zhong Lin[1], Xuan Xiao[1,2,*], Wangren Qiu[1],Kuo-Chen Chou[2,3]

[1] Information Engineering School, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333046 China; 2 Gordon Life Science Institute, Boston, MA 02478, USA; 3 Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX



## Abstract

**Motivation:** Protein remote homology detection is a challenging problem for drug development. Although there are a couple of methods to deal with this problem, the benchmark datasets based on which the existing models were trained and tested contained many high homologous samples due to the fact that the cutoff threshold was set at 95%. Thus, in this study, we reconstructed the benchmark dataset by setting the threshold at 40%, meaning none of the proteins included has more than 40% pairwise sequence identity with any other. Using the new benchmark dataset, we proposed a new method called PHom-GRADSI to detect the remote homologous proteins by integrating various ranking approaches via grey incidence analysis and function domain similarity index.

**Results:** Rigorous jack-knife test have indicated that the new predictor is superior to its counterparts in both enhancing successes rates and reducing computational cost. Phom-GRADSI get an ROC1 score of 0.9620 and an ROC50 score of 0.8861. These result show the domain similarity index is an effective means to measure the distance of two proteins.

**Availability:** We deliver programs of this noval method in https://github.com/jcilwz/RemoteHomology/tree/master/program. Everyone can download programs from this website and the usage is described in ReadMe.txt.

**Contact:** linweizhongjci@sina.com

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Detecting remote homology relationship among proteins plays one of the fundamental and central roles in computational proteomics. It is particularly useful for drug development (see, e.g.,(Chou, et al., 1999; Zhou, et al., 2015)). With the development of sequencing techniques, the protein sequence data rapidly raise. To find those proteins structure and function is more and more urgent. Although X-ray crystallography is a powerful tool in determining protein 3D structures, it is time-consuming and expensive. Particularly, not all proteins can be successfully crystallized, particularly for membrane proteins. The NMR technique is indeed a very powerful tool in determining the 3D structures for membrane proteins as indicated by a series of recent publications (see, e.g., (Berardi, et al., 2011; Dev, et al., 2016; OuYang, et al., 2013; Oxenoid, et al., 2016; Schnell and Chou, 2008)), it is time-consuming and costly. To acquire the structural information in a timely manner, one has to resort to various structural bioinformatics tools based on the sequence similarity principle (see, e.g., (Chou, 2004)). Unfortunately, such principle cannot cover the cases of remote homology proteins. In view of this, considerable efforts (Chen, et al., 2016; Chen, et al., 2016; Liu, et al., 2016; Liu, et al., 2015; Liu, et al., 2015; Liu, et al., 2014) have been made to detect remote homology proteins.

Although these methods each had their own merits and did play stimulating role in this area, further work is needed. Firstly, the benchmark datasets used in their studies had high similarity. For instance, the benchmark dataset in (Chen, et al., 2016; Liu, et al., 2015) contains 7329 proteins from 1070 different super families, with pairwise sequence identity cutoff set at 95%. In other words, it would allow those proteins with higher than 80% similarity in the data set. Secondly, the ranking algorithm used in those studies would spend a lot of time to training the learning model. For example, if the training dataset has N proteins, the LambdaMART need to deal with N2 proteins pair samples.

The present study was initiated to address the two problems with the aim to develop a more powerful method in this regard.

## 2 Methods

### 2.1 Benchmark Dataset

According to Chou's 5-step rules (Chou, 2011) that have been widely and increasingly used by many investigators (see, e.g., (Chen, et al., 2016; Chen, et al., 2016; Cheng, et al., 2018; Cheng, et al., 2018; Cheng, et al., 2018; Cheng, et al., 2017; Ehsan, et al., 2018; Feng, et al., 2017; Feng, et al., 2018; Jia, et al., 2016; Jia, et al., 2016; Jia, et al., 2016; Lin, et al., 2012; Liu, et al., 2017; Liu, et al., 2017; Liu, et al., 2018; Song, et al., 2018)), the first prerequisite in establishing a new predictor is to construct or select an effective benchmark dataset.

In this study, the benchmark dataset was taken from Liu et al. (Liu, et al., 2015). It included 7329 proteins from 1070 different super families and 1824 families derived from SCOP database. To reduce the redundancy and homology bias, the program CD-HIT(Huang, et al., 2010) was adopted to cut down those proteins that had ≥40% pairwise sequence identity to any other in the dataset. Furthermore we removed those families that just had one protein sequence. Finally, we obtained 3128 proteins from 540 super-families and 777 families.

### 2.2 Grey Incidence Analysis of proteins formulated by Grey-PSSM

Given a protein with L amino acid residues, it is usually expressed by

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\cdots\mathrm{R}_i\cdots\mathrm{R}_L \tag{1}$$

where $\mathrm{R}_i$ $(i = 1, 2, \cdots, L)$ is the i-th residue in the protein. Since all the existing machine-learning algorithms can only handle vector but not sequence samples (Chou, 2015), one has to convert Eq.1 into a vector model. But a biological sequence expressed as a vector in the discrete framework may lose all the sequence-order or pattern information.

To avoid completely losing this kind of information for proteins, the pseudo amino acid composition (PseAAC)(Chou, 2005; Chou, 2001) was proposed. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics (see, e.g., (Behbahani, et al., 2016; Dehzangi, et al., 2015; Meher, et al., 2017; Rahimi, et al., 2017; Tahir, et al., 2017; Tripathi and Pandey, 2017; Yu, et al., 2017; Zhang and Duan, 2018) as well as a long list of references cited in(Chou, 2009; Chou, 2017)). According to the general PseAAC(Chou, 2011), the protein of Eq.1 can be formulated as

$$\mathbf{P} = [\Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_u \quad \cdots \quad \Psi_\Omega]^{\mathbf{T}} \tag{2}$$

where T is the transposing operator, the subscript $\Omega$ is an integer, and its value and the components $\Psi_u$ $(u = 1, 2, \cdots)$ will depend on how to extract the desired features and properties from the protein sequence.

In this study, the model, Grey-PSSM proposed by Lin(Lin, et al., 2012; Lin, et al., 2013), is adopted. It extracted the sequential evolution information by the Position Specific Scoring Matrix (PSSM). For the concrete procedures, refer to the original papers (Lin, et al., 2012; Lin, et al., 2013).

After the Grey-PSSM treatment, we have finally got a 60-D PseKNC vector for Eq.2; i.e., its subscript parameter $\Omega = 60$ and each of the 60 components therein has been uniquely defined.

Assume

$$\mathbb{S} = \{P_1, P_2, \cdots, P_N\} \tag{3}$$

are the set of protein samples, and $P_i (1 \le i \le N)$ is the $i^{th}$ protein. According to Equals [6]~[11] in Ref. (Lin, et al., 2009), the distance $\Gamma(P_i, P_j)$ is defined as the grey incidence degree between $P_i$ and $P_j$. The larger the value of $\Gamma(P_i, P_j)$, the more similar they are.

### 2.3 Domain Similarity Analysis

There are also other models used to formulating proteins except for the PseAAC. Here, we propose a novel mothed to describe the proteins. For a protein $P_i \in \mathbb{S}$, we describe its functional domains set by the following steps.

Step 1, $\mathbb{S}_i^{homo}$ is the homology set of protein $P_i$ and it is extracted by searching against UniProt release 2018_08 Swiss-Prot FASTA format flatfile by HMMER(Finn, et al., 2015; Finn, et al., 2011; Potter, et al., 2018). We just use the top 10 sequences if the search results have more than 10 sequences. Therefore there are at most 10 proteins in $\mathbb{S}_i^{homo}$.

Step 2, for a protein in $\mathbb{S}_i^{homo}$, $h_k^i \in \mathbb{S}_i^{homo}$ $(1 \le k \le 10)$,we annotate its functional domains by running hmmscan program against Pfam-A database (Pfam release 32.0). The Pfam-A includes 17,929 functional domains and 688 clans. We define the sets F and C as following.

$$\mathbb{F} = \{f_1, f_2, \cdots, f_{17929}\}$$
$$\mathbb{C} = \{c_1, c_2, \cdots, c_{688}\} \tag{4}$$

where $f_i$ $(1 \le \mathrm{i} \le 17929)$ denote the i-th functional domain in $\mathbb{F}$ and $c_i$ $(1 \le i \le 688)$ the i-th clan in $\mathbb{C}$. Some functional domains have same clan. For example, the domains of "PF15884" and "PF17050" have the same clan "CL0683". Therefore, the functional domains set of protein $h_k^i$, the k-th homology protein of protein $P_i$, is expressed as a set

$$D_k^i = \{f_i | f_i \in \mathbb{F}\} \tag{5}$$

It is mean that all functional domains of $h_k^i$ constitute the set $D_k^i$.

Step 3, the protein $P_i$ is expressed as a domains set

$$D_i = \cup_{k=1}^{10} D_k^i \tag{6}$$

$D_i$ is a set which is unioned together the functional domain set of each homology protein of the protein $P_i$. We define $D_i$ as the functional domain set formulation of $P_i$.

As above steps, a protein is expressed a set including some functional domains from Pfam-A. For the proteins in same family or clan have similar functional domains, new distance between two proteins, named as Domain Similarity Index (DSI), can be defined based on the functional domains.

The algorithm of distance between $P_i$, $P_j$ is described as follows.

1) If $D_i \cap D_j \ne \emptyset$, $\mathrm{DSI}(P_i, P_j) = \frac{|D_i \cap D_j|}{|D_i \cup D_j|}$

2) Else If $D_i \cap D_j = \emptyset$

We define $Clan_i$ to denote the clans set of $P_i$. $Clan_i$ includes the clans of each element in $D_i$

2.1) If $Clan_i \cap Clan_j \neq \emptyset$, $\mathrm{DSI}(P_i, P_j) = F$. F is a constant and in this study F is equal to 0.2.

2.2) Else, $\mathrm{DSI}(P_i, P_j) = 0$

Where, |•| means the count of set, $\cap$ is the intersection operator of two sets, and $\cup$ is the union operator of two sets. From above description, we have $0 \leq \mathrm{DSI}(P_i, P_j) \leq 1$ and the larger the $\mathrm{DSI}(P_i, P_j)$, the more similarity they are.

## 2.4 Operation Engine or Algorithm

In this study, the Grey Relational Analysis (Deng, 1989; Liu, et al., 2006) and the Domain Similarity Index was utilized to rank the relationship of proteins. Given a query protein, the system will search it against the benchmark dataset and return the top ranked proteins. The predictor thus formed is called "PHom-GRADSI". Illustrated in Figure 1 is a flowchart to show how the proposed predictor is working. In this paper, w(1) and w(2) are equal to 0.5.

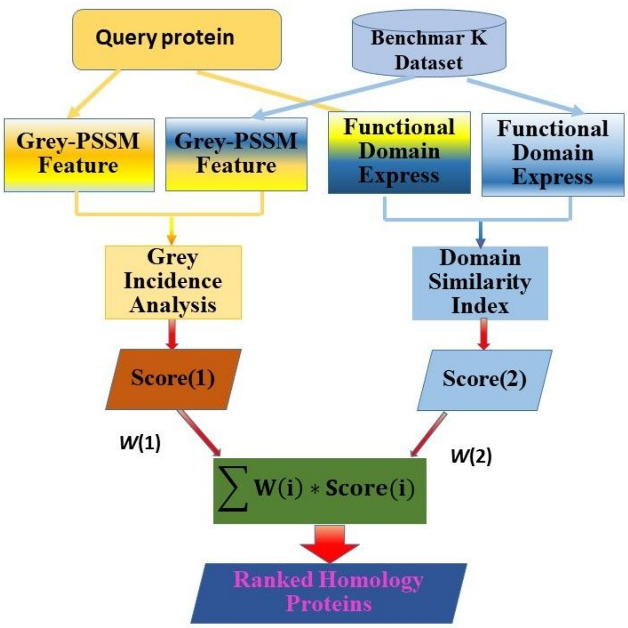


**Fig.1.** This flowchart illustrate how the proposed predictor is working.

# 3 Results

The jackknife test is deemed the least arbitrary and most objective among three cross-validation methods: independent dataset test, K-fold cross-validation test and jackknife test (Chou and Zhang, 1995). Because the LambdaMART ranking algorithm used in preview studies (Chen, et al., 2016; Liu, et al., 2015) consumed more training time and computer memory, as a compromise the 5-fold cross-validation test was adopted there. Now, we employed GRA and DSI to compute the relationship score between the query protein and benchmark dataset proteins, significantly reducing the computing time and memory. Therefore it would be feasible to use the most rigorous jackknife test to examine the prediction quality. The outcome thus obtained are given in Table 1, where we can see that PHom-GRADSI achieved the best performance in both the score of ROC1 and the score of ROC50.

In the same time, we used the Jaccard Index (JI) to calculate the similarity of the proteins. The JI is used to measure the similarity of two sets A and B. It is defined as following:

$$\mathrm{JI(A,B)} = \frac{|A \cap B|}{|A \cup B|}$$

The proteins were formulated as the functional domain set as Eq.(5). In the method of Jaccard Index, the score of ROC and ROC50 were 0.8196 and 0.8070 respectively. In contrast to this, the score of ROC and ROC50 achieved by DSI were 0.9053 and 0.8454 respectively. It can be concluded that DSI preforms better than Jaccard Index. The main reason is that some different domains have same clan. If there are not same elements in two proteins' functional domain set, the distance of Jaccard Index is zero, but the distance of DSI is greater zero. It is evident that the non zero distance of these proteins is more reasonable because their some functional domain have same clan.

Because not every protein can be found its homology proteins, its functional domain set formulation is empty set (see Eq.(4)). For example, there are 23 proteins who cannot be formulated as functional domain set in the Benchmark Dataset. The distances between these proteins and other proteins are zero according to the definition of JI and DSI . So we cannot distinguish the similarity of these proteins who have no homology proteins. This situation is the worst failure DSI. In order to overcome the failure we merged GIA and Jaccard Index and merging GIA and DSI in predicting, respectively. From the Table 1, it is showed that the values of ROC1 and ROC50 are improved at least 0.04 after merging GIA into Jaccard Index or DSI. We can draw the grey incidence analysis can compensate for the worst failure of DSI.

**Table 1**. A comparison of the jackknife test results for protein remote homology detection on the benchmark dataset

| Methods | ROC1 | ROC50 |
|---|---|---|
| PSI-BLAST | 0.7113 | 0.7647 |
| GRA (Grey-PSSM) | 0.8937 | 0.7149 |
| Jaccard Index | 0.8196 | 0.8070 |
| Domain Similarity Index (DSI) | 0.9053 | 0.8454 |
| GRA and Jaccard Index | 0.9301 | 0.8533 |
| **PHom-GRADSI** | **0.9620** | **0.8861** |

# 4 Conclusion

Protein remote homology detection is vital for studying protein structures and functions. It is anticipated that the proposed method may become a useful high throughput toll for both basic research and drug design. In this study, a novel method DSI is proposed. It describes a protein as a functional domain set and measure the distance of two proteins by comparing two proteins' functional domain set similarity. The work testifies the DSI method is effective. This method formulating proteins and calculating distance between proteins may be used in other fields of predicting protein function or structure. We deliver programs of this novel method in https://github.com/jcilwz/RemoteHomology/tree/master/program. Everyone can download programs from this website and the usage is described in ReadMe.txt.

## Acknowledgements

We thank Carla Roberts for editing the final paper.

## Funding

This work was support by the grants from the National Natural Science Foundation of China (No.61462047, 31560316, 31760315). Natural Science Foundation of Jiangxi Province, China (No. 20171ACB20023), the Department of Education of JiangXi Province (GJJ160866), The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

*Conflict of Interest:* none declared.


## References


Behbahani, M., Mohabatkar, H. and Nosrati, M. (2016) Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition, *Journal of Theoretical Biology*, **411**, 1-5.

Berardi, M.J.*, et al.* (2011) Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching, *Nature*, **476**, 109.

Chen, J.*, et al.* (2016) A comprehensive review and comparison of different computational methods for protein remote homology detection, *Brief Bioinform*.

Chen, J.*, et al.* (2016) dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation, *Scientific reports*, **6**, 32333.

Chen, W.*, et al.* (2016) Using deformation energy to analyze nucleosome positioning in genomes, *Genomics*, **107**, 69-75.

Chen, W.*, et al.* (2016) iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, *Oncotarget*, **8**, 4208-4217.

Cheng, X., Xiao, X. and Chou, K.-C. (2018) pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC, *Genomics*, **110**, 231-239.

Cheng, X., Xiao, X. and Chou, K.C. (2018) pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, *Genomics*, **110**, 50-58.

Cheng, X., Xiao, X. and Chou, K.C. (2018) pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information, *Bioinformatics*, **34**, 1448-1456.

Cheng, X.*, et al.* (2017) pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, *Bioinformatics*, **33**, 3524-3531.

Chou, K.-C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, *Bioinformatics*, **21**, 10-19.

Chou, K.-C. (2009) Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology, *Current Proteomics*, 262-274.

Chou, K.-C. (2017) An Unprecedented Revolution in Medicinal Chemistry Driven by the Progress of Biological Science, *Current Topics in Medicinal Chemistry*, **17**, 2337-2358.

Chou, K.-C., Watenpaugh, K.D. and Heinrikson, R.L. (1999) A Model of the Complex between Cyclin-Dependent Kinase 5 and the Activation Domain of Neuronal Cdk5 Activator, *Biochem Bioph Res Co*, **259**, 420-428.

Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo amnio acid composition, *Proteins: Structure, Function, and Bioinformatics*, 246-255.

Chou, K.C. (2004) Structural bioinformatics and its impact to biomedical science, *Curr Med Chem*, **11**, 2105-2134.

Chou, K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition, *J Theor Biol*, **273**, 236-247.

Chou, K.C. (2015) Impacts of bioinformatics to medicinal chemistry, *Med Chem*, **11**, 218-234.

Chou, K.C. and Zhang, C.T. (1995) Prediction of protein structural classes, *Crit Rev Biochem Mol Biol*, **30**, 275-349.

Dehzangi, A.*, et al.* (2015) Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, *Journal of Theoretical Biology*, **364**, 284-294.

Deng, J.L. (1989) Introduction to Grey System Theory, *The Journal of Grey System*, 1-24.

Dev, J.*, et al.* (2016) Structural basis for membrane anchoring of HIV-1 envelope spike, *Science*, aaf7066.

Ehsan, A.*, et al.* (2018) A Novel Modeling in Mathematical Biology for Classification of Signal Peptides, *Scientific reports*, **8**, 1039.

Feng, P.*, et al.* (2017) iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC, *Molecular Therapy - Nucleic Acids*, **7**, 155-163.

Feng, P.*, et al.* (2018) iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, *Genomics*.

Huang, Y.*, et al.* (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics*, **26**, 680-682.

Jia, J.*, et al.* (2016) Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition, *Journal of Biomolecular Structure and Dynamics*, **34**, 1946-1961.

Jia, J.*, et al.* (2016) pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, *J Theor Biol*, **394**, 223-230.

Jia, J.*, et al.* (2016) pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC, *Bioinformatics*, **32**, 3133-3141.

Lin, W.-Z.*, et al.* (2012) Predicting Secretory Proteins of Malaria Parasite by Incorporating Sequence Evolution Information into Pseudo Amino Acid Composition via Grey System Model, *PLOS ONE*, **7**, e49040.

Lin, W.-Z., Xiao, X. and Chou, K.-C. (2009) GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis, *Protein Engineering, Design and Selection*, **22**, 699-705.

Lin, W.Z.*, et al.* (2012) Predicting secretory proteins of malaria parasite by incorporating sequence evolution information into pseudo amino acid composition via grey system model, *PLoS One*, **7**, e49040.

Lin, W.Z.*, et al.* (2013) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, *Mol Biosyst*, **9**, 634-644.

Liu, B., Chen, J. and Wang, S. (2016) Protein Remote Homology Detection by Combining Pseudo Dimer Composition with an Ensemble Learning Method, *Current Proteomics*, **13**, 86-91.

Liu, B., Chen, J. and Wang, X. (2015) Application of learning to rank to protein remote homology detection, *Bioinformatics*, **31**, 3492-3498.

Liu, B., Chen, J. and Wang, X. (2015) Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis, *Mol Genet Genomics*, **290**, 1919-1931.

Liu, B.*, et al.* (2017) iRSpot-EL: identify recombination spots with an ensemble learning approach, *Bioinformatics*, **33**, 35-41.

Liu, B., Yang, F. and Chou, K.-C. (2017) 2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function, *Molecular Therapy - Nucleic Acids*, **7**, 267-277.

Liu, B.*, et al.* (2018) iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, *Bioinformatics*, **34**, 33-40.

Liu, B.*, et al.* (2014) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection, *Bioinformatics*, **30**, 472-479.

Liu, S., Fang, Z. and Lin, Y. (2006) A new definition for the degree of grey incidence, *Scientific Inquiry*, **7**, 111-124.

Meher, P.K.*, et al.* (2017) Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC, *Scientific reports*, **7**, 42362.

OuYang, B.*, et al.* (2013) Unusual architecture of the p7 channel from hepatitis C virus, *Nature*, **498**, 521.

Oxenoid, K.*, et al.* (2016) Architecture of the mitochondrial calcium uniporter, *Nature*, **533**, 269.

Rahimi, M., Bakhtiarizadeh, M.R. and Mohammadi-Sangcheshmeh, A. (2017) OOgenesis_Pred: A sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition, *Journal of Theoretical Biology*, **414**, 128-136.

Schnell, J.R. and Chou, J.J. (2008) Structure and mechanism of the M2 proton channel of influenza A virus, *Nature*, **451**, 591.

Song, J.*, et al.* (2018) PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework, *Journal of Theoretical Biology*, **443**, 125-137.

Tahir, M., Hayat, M. and Kabir, M. (2017) Sequence based predictor for discrimination of enhancer and their types by applying general form of Chou's trinucleotide composition, *Comput Meth Prog Bio*, **146**, 69-75.

Tripathi, P. and Pandey, P.N. (2017) A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition, *Journal of Theoretical Biology*, **424**, 49-54.

Yu, B.*, et al.* (2017) Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising, *Oncotarget*, **8**, 107640-107665.

Zhang, S. and Duan, X. (2018) Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC, *Journal of Theoretical Biology*, **437**, 239-250.

Zhou, G.P., Huang, R.B. and Troy, F.A., 2nd (2015) 3D structural conformation and functional domains of polysialyltransferase ST8Sia IV required for polysialylation of neural cell adhesion molecules, *Protein Pept Lett*, **22**, 137-148.