https://www.scirp.org/journal/ns Natural Science, 2020, Vol. 12, (No. 3), pp: 181-198

Use Chou's 5-Steps Rule to Predict Remote Homology Proteins by Merging Grey Incidence Analysis and Domain Similarity Analysis

Weizhong Lin[1], Xuan Xiao[1,2], Wangren Qiu[1], Kuo-Chen Chou[2]

[1]Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China; [2]Gordon Life Science Institute, Boston, MA 02478, USA

Correspondence to: Xuan Xiao, xxiao@gordonlifescience.org

Keywords: Remote Homology Proteins, Grey Model, Domain Similarity, Chou's 5-Steps Rules

Received: February 29, 2020 **Accepted:** March 22, 2020 **Published:** March 25, 2020

ABSTRACT

Detecting remote homology proteins is a challenging problem for both basic research and drug development. Although there are a couple of methods to deal with this problem, the benchmark datasets based on which the existing methods were trained and tested contain many high homologous samples as reflected by the fact that the cutoff threshold was set at 95%. In this study, we reconstructed the benchmark dataset by setting the threshold at 40%, meaning none of the proteins included in the benchmark dataset has more than 40% pairwise sequence identity with any other in the same subset. Using the new benchmark dataset, we proposed a new predictor called "dRHP-GreyFun" based on the grey modeling and functional domain approach. Rigorous cross-validations have indicated that the new predictor is superior to its counterparts in both enhancing success rates and reducing computational cost. The predictor can be downloaded from https://github.com/jcilwz/dRHP-GreyFun.

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 [1, 2]. With the avalanche of protein sequences generated in the post-genomic age, it is highly desired to timely detect the remote homology proteins. 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. Membrane proteins are difficult to crystallize and most of them will not dissolve in normal solvents. Therefore, so far very few membrane protein structures have been determined. Although NMR is indeed a very powerful tool in determining the 3D structures of membrane proteins (see, e.g., [3-7]), it is also time-consuming and costly. To acquire the structural information in a

timely manner, a series of 3D protein structures have been developed by means of structural bioinformatics tools (see, e.g., [8-20]). Meanwhile, facing the explosive growth of biological sequences discovered in the post-genomic age, to timely use them for drug development, a lot of important sequence-based information, such as PTM (posttranslational modification) sites in proteins [21, 22], protein-drug interaction in cellular networking [23], DNA-methylation sites [24], recombination spots [25], and sigma-54 promoters [26], have been deducted by various sequential bioinformatics tools such as PseAAC approach [27] and PseKNC approach [28]. Actually, the rapid development in sequential bioinformatics and structural bioinformatics have driven the medicinal chemistry undergoing an unprecedented revolution [29], in which the computational biology has played increasingly important roles in stimulating the development of finding novel drugs. In view of this, the computational methods were also utilized in this study for detecting remote homology.

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., [30]). Unfortunately, such principle cannot cover the cases of remote homology proteins. In view of this, considerable efforts [31-35] have been made to detect remote homology proteins.

Although these methods each had their own merits and did play a 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 [33, 34] 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 benchmark dataset. Secondly, the ranking algorithm used in those studies would spend a lot of time to train or learn the model. For example, if the training dataset had N proteins, the Lambda-MART would need to deal with N^2 proteins pair samples.

As demonstrated by a series of recent publications [23, 25, 26, 36-71], to develop a really useful predictor for a biological system, one needs to follow Chou's 5-step rule to go through the following five steps: 1) select or construct a valid benchmark dataset to train and test the predictor; 2) represent the samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; 3) introduce or develop a powerful algorithm to conduct the prediction; 4) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; 5) establish a user-friendly web-server for the predictor that is accessible to the public. Papers presented for developing a new sequence-analyzing method or statistical predictor by observing the guidelines of Chou's 5-step rules have the following notable merits: 1) crystal clear in logic development, 2) completely transparent in operation, 3) easily to repeat the reported results by other investigators, 4) with high potential in stimulating other sequence-analyzing methods, and 5) very convenient to be used by the majority of experimental scientists. Below, let us elaborate on how to deal with these five steps one by one.

2. MATERIALS AND METHOD

2.1. Benchmark Dataset

According to Chou's 5-step rules [72], 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.* [33]. It contains 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 [73] was adopted to remove those proteins that had ≥40% pairwise sequence identity to any other in the same subset. Meanwhile, removed were also those families that only had one protein sequence. Finally, we obtained 3128 proteins from 540 super-families and 777 families.

2.2. Sample Formulation

Most biological systems have two remarkable features: one is of evolution and the other is of complexity. All biological species have developed beginning from a very limited number of ancestral species. It

is true for protein sequence as well [30]. Their evolution involves changes of single residues, insertions and deletions of several residues, gene doubling, and gene fusion [9, 74]. With these changes accumulated for a long period of time, many similarities between initial and resultant amino acid sequences are gradually eliminated, but the corresponding proteins may still share many common attributes, such as having basically the same biological function, subcellular location and similar binding site. To take into account the evolution information, many investigators used the PSSM (Position-Specific Scoring Matrix) approach [75], as done in a series of previous publications (see, e.g., [76-81]). On the other hand, biological systems are extremely complicated with a lot of uncertainties. According to the grey system theory [82], if the information of an investigated system is fully known, it is called a "white system;" if completely unknown, a "black system;" if partially known, a "grey system." Actually, most biological systems belong to the grey systems, and hence it is particularly effective to treat them with the grey model approach [83-86].

2.2.1. Grey Incidence Analysis of Proteins Formulated by Grey-PSSM

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

$$\boldsymbol{P} = R_1 R_2 R_3 \cdots R_i \cdots R_L \tag{1}$$

where $R_i \left(i = 1, 2, \cdots, L \right)$ is the i-th residue in the protein. Because all the existing machine-learning algorithms (such as "Optimization" algorithm [87], "Covariance Discriminant" or "CD" algorithm [88, 89], "Nearest Neighbor" or "NN" algorithm [90], and "Support Vector Machine" or "SVM" algorithm [90]) can only handle vectors as elaborated in a comprehensive review [29]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition [27] or PseAAC [91] was proposed. Ever since then, it has been widely used in nearly all the areas of computational proteomics (see, e.g., [92-95] as well as a long list of references cited in [96]). Because it has been widely and increasingly used, four powerful open access soft-wares, called "PseAAC" [97], "PseAAC-Builder" [98], "propy" [99], and "PseAAC-General" [100], were established: the former three are for generating various modes of Chou's special PseAAC [101]; while the 4th one for those of Chou's general PseAAC [72], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode (see Eqs.9-10 of [72]), "Gene Ontology" mode (see Eqs.11-12 of [72]), and "Sequential Evolution" or "PSSM" mode (see Eqs.13-14 of [72]). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) [28] was developed for generating various feature vectors for DNA/RNA sequences [102, 103] that have proved very useful as well. Particularly, recently a very powerful web-server called "Pse-in-One" [104] and its updated version "Pse-in-One2.0" [105] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

According to the general PseAAC [72], the protein of Equation (1) can be formulated as

$$\boldsymbol{P} = \left[\Psi_1 \;\; \Psi_2 \;\; \cdots \;\; \Psi_u \;\; \cdots \;\; \Psi_\Omega \right]^{\mathrm{T}} \tag{2}$$

where T is the transposing operator, the subscript Ω is an integer, and its value and the components $\Psi_u \left(u = 1, 2, \cdots \right)$ 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 *et al.* [85, 86] is adopted. It has extracted the sequential evolution information by the Position Specific Scoring Matrix (PSSM). After the Grey-PSSM treatment, we have finally got a 60-D PseKNC vector for Equation (2); *i.e.*, its subscript parameter Ω = 60and each of the 60 components therein has been uniquely defined below. Suppose the set of protein samples is

$$\mathbb{S} = P_1 P_2 P_3 \cdots R_i \cdots R_N \tag{3}$$

where $P_i\left(1 \le i \le N\right)$ is the i-th protein. According to Eqs.6-11 in Lin *et al.* [106], the distance $\Gamma\left(P_i, P_j\right)$ is defined as the grey incidence degree between P_i and P_j. The larger the value of $\Gamma\left(P_i, P_j\right)$, the more similar between P_i and P_j will be.

2.2.2. Domain Similarity Analysis

In addition to the PseAAC [27, 91] approach, the functional domain [107-112] can also be used to characterize protein sample, $P_i \in \mathbb{S}$, according to the following steps.

Step 1. Searching UniProt release 2018_08 Swiss-Prot FASTA format flatfile by HMMER [113-115] for the homology set of protein P_i, we have obtained $\mathbb{S}_i^{\text{homo}}$. If the outcome has more than 10 protein sequences, only the top 10-ranking ones are used.

Step 2. For the protein in $\mathbb{S}_i^{\text{homo}}$, $P_k^i \in \mathbb{S}_i^{\text{homo}}\left(1 \le k \le 10\right)$, annotate its functional domains by running hmmscan program against Pfam-A database (Pfam release 32.0). The Pfam-A contains 17,929 functional domains and 688 clans, as defined by

$$\begin{cases} \mathbb{F} = \{ f_1, f_2, f_3, \cdots, f_{17929} \} \\ \mathbb{C} = \{ c_1, c_2, c_3, \cdots, c_{688} \} \end{cases} \tag{4}$$

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

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

meaning that all functional domains of P_k^i contains the set D_k^i.

Step 3. The protein P_i can be expressed by the following domains set

$$D^i = \bigcup_{k=1}^{10} D_k^i \tag{6}$$

where $\cup$ denotes union in the set theory.

As we can see from Equations ((5), (6)) the distance (Dis) between P_i and P_j is within the range $0 \le \text{Dis}\left(P_i, P_j\right) \le 1$.

2.3. Operation Engine or Algorithm

In this study, the Grey Relational Analysis [82, 116] and the Domain Similarity Index was utilized to rank the relationship of proteins. Given a query protein, the system will search the benchmark dataset for it and return the top-ranking similar proteins. The predictor thus formed is called "dRHP-GreyFun". 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.

3. RESULTS AND DISCUSSION

Among the independent dataset test, sub-sampling (e.g., 5 or 10-fold cross-validation) test, and jackknife test, which are often used for examining the accuracy of a statistical prediction method [117], the jackknife test was deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [118, 119], as clearly elucidated in a comprehensive review paper [72] and demonstrated by Eqs.28-32 therein. Therefore, the jackknife test has been increasingly recognized and widely adopted by investigators to test the power of various prediction methods (see, e.g., [120-123]). However, to reduce the computational time, we adopted the 5-fold and 10-fold cross-validation in this study as done by many investigators with SVM as the prediction engine. This is also because the LambdaMART ranking algorithm used in preview studies [33, 34] would consume a lot of training time and computer memory. As a compromise, the 5-fold cross-validation test was adopted there. But, now we employed the operation engine

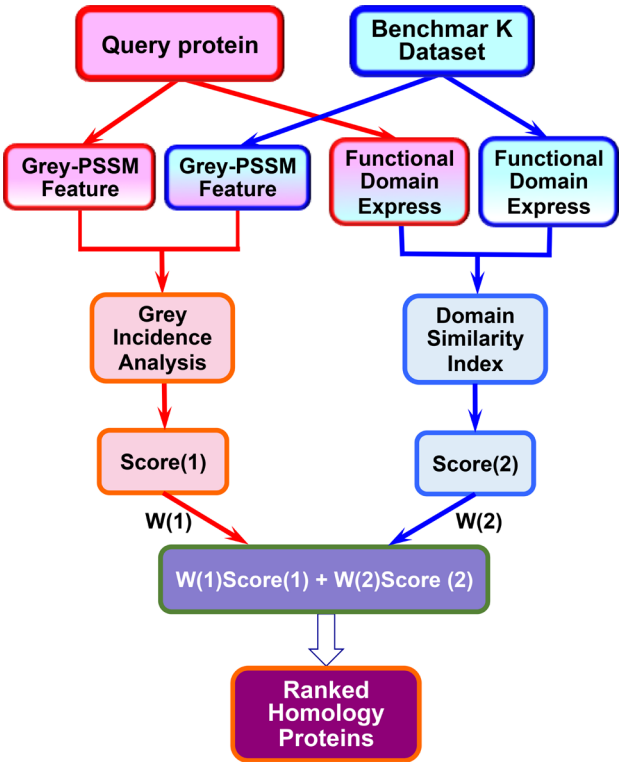

Figure 1. A flowchart to show how the proposed predictor "dRHP-GreyFun" is working by following the guidelines of Chou's 5-steps rule.

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
dRHP-GreyFun	0.9620	0.8861

based on the grey modeling and functional domains to detect the remote homology 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 outcomes thus obtained are given in Table 1, where we can see that dRHP-GreyFun achieved the best performance in both the score of ROC1 and the score of ROC50.

4. CONCLUSIONS

Protein remote homology detection is vitally important for studying protein structures and functions. It is anticipated that the proposed method may become a useful high throughput tool for both basic research and drug design.

As pointed out in [124] and demonstrated in a series of recent publications (see, e.g., [40, 125-144]), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have significantly increased the impacts of bioinformatics on medical science [29], driving medicinal chemistry into an unprecedented revolution [96]. Accordingly, we have also provided a web-server for the prediction method presented in this paper, by which users can easily get their desired results without the need to go through the complicated math equation involved. Also, all the programs can be downloaded from https://github.com/jcilwz/dRHP-GreyFun.

It is illuminating that using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein, as indicated by many previous studies on a series of important biological topics, (see, e.g., [145-158]), particularly what happened is for the topics of enzyme kinetics, protein folding rates [153, 159-161], and low-frequency internal motion [162, 163].

For the remarkable and awesome roles of the "5-steps rule" in driving proteome, genome analyses and drug development, see a series of recent papers [139, 164-188], where the rule and its wide applications have been very impressively presented from various aspects or at different angles.

ACKNOWLEDGEMENTS

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.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Chou, K.C., Watenpaugh, K.D. and Heinrikson, R.L. (1999) A Model of the Complex between Cyclin-Dependent Kinase 5 (Cdk5) and the Activation Domain of Neuronal Cdk5 Activator. *Biochemical & Biophysical Research Communications* (*BBRC*), **259**, 420-428. https://doi.org/10.1006/bbrc.1999.0792
2. Zhang, J., Luan, C.H., Chou, K.C. and Johnson, G.V.W. (2002) Identification of the N-Terminal Functional Domains of Cdk5 by Molecular Truncation and Computer Modeling. *Proteins*: *Structure*, *Function and Genetics*, **48**, 447-453. https://doi.org/10.1002/prot.10173
3. Schnell, J.R. and Chou, J.J. (2008) Structure and Mechanism of the M2 Proton Channel of Influenza a Virus. *Nature*, **451**, 591-595. https://doi.org/10.1038/nature06531
4. Berardi, M.J., Shih, W.M., Harrison, S.C. and Chou, J.J. (2011) Mitochondrial Uncoupling Protein 2 Structure Determined by NMR Molecular Fragment Searching. *Nature*, **476**, 109-113. https://doi.org/10.1038/nature10257
5. Ouyang, B., Xie, S., Berardi, M.J., Zhao, X.M., Dev, J., Yu, W., Sun, B. and Chou, J.J. (2013) Unusual Architecture of the p7 Channel from Hepatitis C Virus. *Nature*, **498**, 521-525. https://doi.org/10.1038/nature12283
6. Oxenoid, K., Dong, Y.S., Cao, C., Cui, T., Sancak, Y., Markhard, A.L., Grabarek, Z., Kong, L., Liu, Z., Ouyang, B., Cong, Y., Mootha, V.K. and Chou, J.J. (2016) Architecture of the Mitochondrial Calcium Uniporter. *Nature*, **533**, 269-273. https://doi.org/10.1038/nature17656
7. Dev, J., Park, D., Fu, Q., Chen, J., Ha, H.J., Ghantous, F., Herrmann, T., Chang, W., Liu, Z., Frey, G., Seaman, M.S., Chen, B. and Chou, J.J. (2016) Structural Basis for Membrane Anchoring of HIV-1 Envelope Spike. *Science*, **353**, 172-175. https://doi.org/10.1126/science.aaf7066

8. Chou, K.C., Tomasselli, A.G. and Heinrikson, R.L. (2000) Prediction of the Tertiary Structure of a Caspase-9/Inhibitor Complex. *FEBS Letters*, **470**, 249-256. https://doi.org/10.1016/S0014-5793(00)01333-8

9. Chou, K.C., Jones, D. and Heinrikson, R.L. (1997) Prediction of the Tertiary Structure and Substrate Binding Site of Caspase-8. *FEBS Letters*, **419**, 49-54. https://doi.org/10.1016/S0014-5793(97)01246-5

10. Chou, K.C. (2004) Insights from Modelling the 3D Structure of the Extracellular Domain of alpha7 Nicotinic Acetylcholine Receptor. *Biochemical and Biophysical Research Communication* (*BBRC*), **319**, 433-438. https://doi.org/10.1016/j.bbrc.2004.05.016

11. Chou, K.C. (2005) Coupling Interaction between Thromboxane A2 Receptor and Alpha-13 Subunit of Guanine Nucleotide-Binding Protein. *Journal of Proteome Research*, **4**, 1681-1686. https://doi.org/10.1021/pr050145a

12. Chou, K.C. and Howe, W.J. (2002) Prediction of the Tertiary Structure of the Beta-Secretase Zymogen. *Biochemical and Biophysical Research Communications* (*BBRC*), **292**, 702-708. https://doi.org/10.1006/bbrc.2002.6686

13. Chou, K.C. (2004) Insights from Modelling the Tertiary Structure of BACE2. *Journal of Proteome Research*, **3**, 1069-1072. https://doi.org/10.1021/pr049905s

14. Chou, K.C. (2004) Insights from Modelling Three-Dimensional Structures of the Human Potassium and Sodium Channels. *Journal of Proteome Research*, **3**, 856-861. https://doi.org/10.1021/pr049931q

15. Chou, K.C. (2005) Modeling the Tertiary Structure of Human Cathepsin-E. *Biochemical and Biophysical Research Communications* (*BBRC*), **331**, 56-60. https://doi.org/10.1016/j.bbrc.2005.03.123

16. Chou, K.C. (2005) Insights from Modeling the 3D Structure of DNA-CBF3b Complex. *Journal of Proteome Research*, **4**, 1657-1660. https://doi.org/10.1021/pr050135+

17. Wang, S.Q., Du, Q.S. and Chou, K.C. (2007) Study of Drug Resistance of Chicken Influenza A Virus (H5N1) from Homology-Modeled 3D Structures of Neuraminidases. *Biochemical and Biophysical Research Communications* (*BBRC*), **354**, 634-640. https://doi.org/10.1016/j.bbrc.2006.12.235

18. Wang, S.Q., Du, Q.S., Huang, R.B., Zhang, D.W. and Chou, K.C. (2009) Insights from Investigating the Interaction of Oseltamivir (Tamiflu) with Neuraminidase of the 2009 H1N1 Swine Flu Virus. *Biochemical and Biophysical Research Communications* (*BBRC*), **386**, 432-436. https://doi.org/10.1016/j.bbrc.2009.06.016

19. Li, X.B., Wang, S.Q., Xu, W.R., Wang, R.L. and Chou, K.C. (2011) Novel Inhibitor Design for Hemagglutinin against H1N1 Influenza Virus by Core Hopping Method. *PLoS ONE*, **6**, e28111. https://doi.org/10.1371/journal.pone.0028111

20. Ma, Y., Wang, S.Q., Xu, W.R., Wang, R.L. and Chou, K.C. (2012) Design Novel Dual Agonists for Treating Type-2 Diabetes by Targeting Peroxisome Proliferator-Activated Receptors with Core Hopping Approach. *PLoS ONE*, **7**, e38546. https://doi.org/10.1371/journal.pone.0038546

21. Xu, Y., Ding, J., Wu, L.Y. and Chou, K.C. (2013) iSNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition. *PLoS ONE*, **8**, e55844. https://doi.org/10.1371/journal.pone.0055844

22. Chou, K.C. (2019) Progresses in Predicting Post-Translational Modification. *International Journal of Peptide Research and Therapeutics* (*IJPRT*). https://link.springer.com/article/10.1007%2Fs10989-019-09893-5 https://doi.org/10.1007/s10989-019-09893-5

23. Xiao, X., Min, J.L., Lin, W.Z., Liu, Z., Cheng, X. and Chou, K.C. (2015) iDrug-Target: Predicting the Interactions between Drug Compounds and Target Proteins in Cellular Networking via the Benchmark Dataset Optimization Approach. *Journal of Biomolecular Structure and Dynamics* (*JBSD*), **33**, 2221-2233. https://doi.org/10.1080/07391102.2014.998710

24. Liu, Z., Xiao, X., Qiu, W.R. and Chou, K.C. (2015) iDNA-Methyl: Identifying DNA Methylation Sites via Pseu-

do Trinucleotide Composition. *Analytical Biochemistry*, **474**, 69-77. https://doi.org/10.1016/j.ab.2014.12.009

25. Chen, W., Feng, P.M., Lin, H. and Chou, K.C. (2013) iRSpot-PseDNC: Identify Recombination Spots with Pseudo Dinucleotide Composition. *Nucleic Acids Research*, **41**, e68. https://doi.org/10.1093/nar/gks1450
26. Lin, H., Deng, E.Z., Ding, H., Chen, W. and Chou, K.C. (2014) iPro54-PseKNC: A Sequence-Based Predictor for Identifying Sigma-54 Promoters in Prokaryote with Pseudo k-Tuple Nucleotide Composition. *Nucleic Acids Research*, **42**, 12961-12972. https://doi.org/10.1093/nar/gku1019
27. Chou, K.C. (2001) Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition. *PROTEINS*: *Structure*, *Function*, *and Genetics*, **43**, 246-255. https://doi.org/10.1002/prot.1035
28. Chen, W., Lei, T.Y., Jin, D.C., Lin, H. and Chou, K.C. (2014) PseKNC: A Flexible Web-Server for Generating Pseudo K-Tuple Nucleotide Composition. *Analytical Biochemistry*, **456**, 53-60. https://doi.org/10.1016/j.ab.2014.04.001
29. Chou, K.C. (2015) Impacts of Bioinformatics to Medicinal Chemistry. *Medicinal Chemistry*, **11**, 218-234. https://doi.org/10.2174/1573406411666141229162834
30. Chou, K.C. (2004) Review: Structural Bioinformatics and Its Impact to Biomedical Science. *Current Medicinal Chemistry*, **11**, 2105-2134. https://doi.org/10.2174/0929867043364667
31. Liu, B., Wang, X., Lin, L., Dong, Q. and Wang, X. (2008) A Discriminative Method for Protein Remote Homology Detection and Fold Recognition Combining Top-n-Grams and Latent Semantic Analysis. *BMC Bioinformatics*, **9**, Article No. 510. https://doi.org/10.1186/1471-2105-9-510
32. Liu, B., Wang, X., Zou, Q., Dong, Q. and Chen, Q. (2013) Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation. *Molecular Informatics*, **32**, 775-782. https://doi.org/10.1002/minf.201300084
33. 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. *Molecular Genetics and Genomics*: *MGG*, **290**, 1919-1931. https://doi.org/10.1007/s00438-015-1044-4
34. Chen, J., Long, R., Wang, X.L., Liu, B. and Chou, K.C. (2016) dRHP-PseRA: Detecting Remote Homology Proteins Using Profile-Based Pseudo Protein Sequence and Rank Aggregation. *Scientific Reports*, **6**, Article No. 32333. https://doi.org/10.1038/srep32333
35. Chen, J., Guo, M., Wang, X. and Liu, B. (2018) A Comprehensive Review and Comparison of Different Computational Methods for Protein Remote Homology Detection. *Brief Bioinform*, **19**, 231-244. https://doi.org/10.1093/bib/bbw108
36. Feng, P.M., Chen, W., Lin, H. and Chou, K.C. (2013) iHSP-PseRAAAC: Identifying the Heat Shock Protein Families Using Pseudo Reduced Amino Acid Alphabet Composition. *Analytical Biochemistry*, **442**, 118-125. https://doi.org/10.1016/j.ab.2013.05.024
37. Chen, W., Feng, P.M., Deng, E.Z., Lin, H. and Chou, K.C. (2014) iTIS-PseTNC: A Sequence-Based Predictor for Identifying Translation Initiation Site in Human Genes Using Pseudo Trinucleotide Composition. *Analytical Biochemistry*, **462**, 76-83. https://doi.org/10.1016/j.ab.2014.06.022
38. Ding, H., Deng, E.Z., Yuan, L.F., Liu, L., Lin, H., Chen, W. and Chou, K.C. (2014) iCTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels. *BioMed Research International* (*BMRI*), **2014**, Article ID: 286419. https://doi.org/10.1155/2014/286419
39. Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) iSuc-PseOpt: Identifying Lysine Succinylation Sites in Proteins by Incorporating Sequence-Coupling Effects into Pseudo Components and Optimizing Imbalanced Training Dataset. *Analytical Biochemistry*, **497**, 48-56. https://doi.org/10.1016/j.ab.2015.12.009
40. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H. and Chou, K.C. (2017) iRNA-AI: Identifying the Adenosine to

Inosine Editing Sites in RNA Sequences. *Oncotarget*, **8**, 4208-4217. https://doi.org/10.18632/oncotarget.13758

41. Chen, W., Ding, H., Zhou, X., Lin, H. and Chou, K.C. (2018) iRNA(m6A)-PseDNC: Identifying N6-Methyladenosine Sites Using Pseudo Dinucleotide Composition. *Analytical Biochemistry*, **561**-**562**, 59-65. https://doi.org/10.1016/j.ab.2018.09.002
42. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H. and Chou, K.C. (2018) iRNA-3typeA: Identifying 3-Types of Modification at RNA's Adenosine Sites. *Molecular Therapy*: *Nucleic Acid*, **11**, 468-474. https://doi.org/10.1016/j.omtn.2018.03.012
43. Butt, A.H. and Khan, Y.D. (2018) Prediction of S-Sulfenylation Sites Using Statistical Moments Based Features via Chou's 5-Step Rule. *International Journal of Peptide Research and Therapeutics* (*IJPRT*). https://doi.org/10.1007/s10989-019-09931-2
44. Awais, M., Hussain, W., Khan, Y.D., Rasool, N., Khan, S.A. and Chou, K.C. (2019) iPhosH-PseAAC: Identify Phosphohistidine Sites in Proteins by Blending Statistical Moments and Position Relative Features According to the Chou's 5-Step Rule and General Pseudo Amino Acid Composition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://www.ncbi.nlm.nih.gov/pubmed/31144645 https://doi.org/10.1109/TCBB.2019.2919025
45. Barukab, O., Khan, Y.D., Khan, S.A. and Chou, K.C. (2019) iSulfoTyr-PseAAC: Identify Tyrosine Sulfation Sites by Incorporating Statistical Moments via Chou's 5-Steps Rule and Pseudo Components. *Current Genomics*, **20**, 306-320. http://www.eurekaselect.com/174277/article https://doi.org/10.2174/1389202920666190819091609
46. Butt, A.H. and Khan, Y.D. (2019) Prediction of S-Sulfenylation Sites Using Statistical Moments Based Features via Chou's 5-Step Rule. *International Journal of Peptide Research and Therapeutics* (*IJPRT*). https://doi.org/10.1007/s10989-019-09931-2
47. Chen, Y. and Fan, X. (2019) Use Chou's 5-Steps Rule to Reveal Active Compound and Mechanism of Shuangsheng Pingfei San on Idiopathic Pulmonary Fibrosis. *Current Molecular Medicine*, **20**, 220-230. https://doi.org/10.2174/1566524019666191011160543
48. Du, X., Diao, Y., Liu, H. and Li, S. (2019) MsDBP: Exploring DNA-Binding Proteins by Integrating Multi-Scale Sequence Information via Chou's 5-Steps Rule. *Journal of Proteome Research*, **18**, 3119-3132. https://doi.org/10.1021/acs.jproteome.9b00226
49. Dutta, A., Dalmia, A., Singh, K.K. and Anand, A. (2019) Using the Chou's 5-Steps Rule to Predict Splice Junctions with Interpretable Bidirectional Long Short-Term Memory Networks. *Computers in Biology and Medicine*, **116**, Article ID: 103558. https://doi.org/10.1016/j.compbiomed.2019.103558
50. Ehsan, A., Mahmood, M.K., Khan, Y.D., Barukab, O.M., Khan, S.A. and Chou, K.C. (2019) iHyd-PseAAC (EPSV): Identify Hydroxylation Sites in Proteins by Extracting Enhanced Position and Sequence Variant Feature via Chou's 5-Step Rule and General Pseudo Amino Acid Composition. *Current Genomics*, **20**, 124-133. https://doi.org/10.2174/1389202920666190325162307
51. Hussain, W., Khan, S.D., Rasool, N., Khan, S.A. and Chou, K.C. (2019) SPalmitoylC-PseAAC: A Sequence-Based Model Developed via Chou's 5-Steps Rule and General PseAAC for Identifying S-Palmitoylation Sites in Proteins. *Analytical Biochemistry*, **568**, 14-23. https://doi.org/10.1016/j.ab.2018.12.019
52. Hussain, W., Khan, Y.D., Rasool, N., Khan, S.A. and Chou, K.C. (2019) SPrenylC-PseAAC: A Sequence-Based Model Developed via Chou's 5-Steps Rule and General PseAAC for Identifying S-Prenylation Sites in Proteins. *Journal of Theoretical Biology*, **468**, 1-11. https://doi.org/10.1016/j.jtbi.2019.02.007
53. Ju, Z. and Wang, S.Y. (2020) Prediction of Lysine Formylation Sites Using the Composition of k-Spaced Amino Acid Pairs via Chou's 5-Steps Rule and General Pseudo Components. *Genomics*, **112**, 859-866. https://doi.org/10.1016/j.ygeno.2019.05.027

54. Kabir, M., Ahmad, S., Iqbal, M. and Hayat, M. (2020) iNR-2L: A Two-Level Sequence-Based Predictor Developed via Chou's 5-Steps Rule and General PseAAC for Identifying Nuclear Receptors and Their Families. *Genomics*, **112**, 276-285. https://doi.org/10.1016/j.ygeno.2019.02.006

55. Khan, Z.U., Ali, F., Khan, I.A., Hussain, Y. and Pi, D. (2019) iRSpot-SPI: Deep Learning-Based Recombination Spots Prediction by Incorporating Secondary Sequence Information Coupled with Physio-Chemical Properties via Chou's 5-Step Rule and Pseudo Components. *Chemometrics and Intelligent Laboratory Systems* (*CHEMOLAB*), **189**, 169-180. https://doi.org/10.1016/j.chemolab.2019.05.003

56. Lan, J., Liu, J., Liao, C., Merkler, D.J., Han, Q. and Li, J. (2019) A Study for Therapeutic Treatment against Parkinson's Disease via Chou's 5-Steps Rule. *Current Topics in Medicinal Chemistry*, **19**, 2318-2333. http://www.eurekaselect.com/175887/article https://doi.org/10.2174/1568026619666191019111528

57. Le, N.Q.K. (2019) iN6-Methylat (5-Step): Identifying DNA N(6)-Methyladenine Sites in Rice Genome Using Continuous Bag of Nucleobases via Chou's 5-Step Rule. *Molecular Genetics and Genomics*: *MGG*, **294**, 1173-1182. https://doi.org/10.1007/s00438-019-01570-y

58. Le, N.Q.K., Yapp, E.K.Y., Ho, Q.T., Nagasundaram, N., Ou, Y.Y. and Yeh, H.Y. (2019) iEnhancer-5Step: Identifying Enhancers Using Hidden Information of DNA Sequences via Chou's 5-Step Rule and Word Embedding. *Analytical Biochemistry*, **571**, 53-61. https://doi.org/10.1016/j.ab.2019.02.017

59. Le, N.Q.K., Yapp, E.K.Y., Ou, Y.Y. and Yeh, H.Y. (2019) iMotor-CNN: Identifying Molecular Functions of Cytoskeleton Motor Proteins Using 2D Convolutional Neural Network via Chou's 5-Step Rule. *Analytical Biochemistry*, **575**, 17-26. https://doi.org/10.1016/j.ab.2019.03.017

60. Liang, R., Xie, J., Zhang, C., Zhang, M., Huang, H., Huo, H., Cao, X. and Niu, B. (2019) Identifying Cancer Targets Based on Machine Learning Methods via Chou's 5-Steps Rule and General Pseudo Components. *Current Topics in Medical Chemistry*, **19**, 2301-2317. https://doi.org/10.2174/1568026619666191016155543

61. Liang, Y. and Zhang, S. (2019) Identifying DNase I Hypersensitive Sites Using Multi-Features Fusion and F-Score Features Selection via Chou's 5-Steps Rule. *Biophysical Chemistry*, **253**, Article ID: 106227. https://doi.org/10.1016/j.bpc.2019.106227

62. Liu, Z., Dong, W., Jiang, W. and He, Z. (2019) csDMA: An Improved Bioinformatics Tool for Identifying DNA 6 ma Modifications via Chou's 5-Step Rule. *Scientific Reports*, **9**, Article No. 13109. https://doi.org/10.1038/s41598-019-49430-4

63. Malebary, S.J., Rehman, M.S.U. and Khan, Y.D. (2019) iCrotoK-PseAAC: Identify Lysine Crotonylation Sites by Blending Position Relative Statistical Features According to the Chou's 5-Step Rule. *PLoS ONE*, **14**, e0223993. https://doi.org/10.1371/journal.pone.0223993

64. Nazari, I., Tahir, M., Tayari, H. and Chong, K.T. (2019) iN6-Methyl (5-Step): Identifying RNA N6-Methyladenosine Sites Using Deep Learning Mode via Chou's 5-Step Rules and Chou's General PseKNC. *Chemometrics and Intelligent Laboratory Systems* (*CHEMOLAB*), **193**, Article ID: 103811. https://doi.org/10.1016/j.chemolab.2019.103811

65. Ning, Q., Ma, Z. and Zhao, X. (2019) dForml(KNN)-PseAAC: Detecting Formylation Sites from Protein Sequences Using K-Nearest Neighbor Algorithm via Chou's 5-Step Rule and Pseudo Components. *Journal of Theoretical Biology*, **470**, 43-49. https://doi.org/10.1016/j.jtbi.2019.03.011

66. Tahir, M., Tayara, H. and Chong, K.T. (2019) iDNA6mA (5-Step Rule): Identification of DNA N6-Methyladenine Sites in the Rice Genome by Intelligent Computational Model via Chou's 5-Step Rule. *CHEMOLAB*, **189**, 96-101. https://doi.org/10.1016/j.chemolab.2019.04.007

67. Vishnoi, S., Garg, P. and Arora, P. (2020) Physicochemical n-Grams Tool: A Tool for Protein Physicochemical Descriptor Generation via Chou's 5-Step Rule. *Chemical Biology & Drug Design*, **95**, 79-86.

https://doi.org/10.1111/cbdd.13617

68. Wiktorowicz, A., Wit, A., Dziewierz, A., Rzeszutko, L., Dudek, D. and Kleczynski, P. (2019) Calcium Pattern Assessment in Patients with Severe Aortic Stenosis via the Chou's 5-Steps Rule. *Current Pharmaceutical Design*, **25**, 3769-3775. https://doi.org/10.2174/1381612825666190930101258

69. Yang, L., Lv, Y., Wang, S., Zhang, Q., Pan, Y., Su, D., Lu, Q. and Zuo, Y. (2019) Identifying FL11 Subtype by Characterizing Tumor Immune Microenvironment in Prostate Adenocarcinoma via Chou's 5-Steps Rule. *Genomics*, **112**, 1500-1515. https://doi.org/10.1016/j.ygeno.2019.08.021

70. Vundavilli, H., Datta, A., Sima, C., Hua, J., Lopes, R. and Bittner, M. (2020) Using Chou's 5-Steps Rule to Model Feedback in Lung Cancer. *IEEE Journal of Biomedical and Health Informatics*. (In Press) https://doi.org/10.1109/JBHI.2019.2958042

71. Khan, Y.D., Amin, N., Hussain, W., Rasool, N., Khan, S.A. and Chou, K.C. (2020) iProtease-PseAAC(2L): A Two-Layer Predictor for Identifying Proteases and Their Types Using Chou's 5-Step-Rule and General PseAAC. *Analytical Biochemistry*, **588**, Article ID: 113477. https://doi.org/10.1016/j.ab.2019.113477

72. Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition (50th Anniversary Year Review, 5-Steps Rule). *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024

73. Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W. (2010) CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences. *Bioinformatics*, **26**, 680-682. https://doi.org/10.1093/bioinformatics/btq003

74. Chou, K.C. (1995) The Convergence-Divergence Duality in Lectin Domains of the Selectin Family and Its Implications. *FEBS Letters*, **363**, 123-126. https://doi.org/10.1016/0014-5793(95)00240-A

75. Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) Improving the Accuracy of PSI-BLAST Protein Database Searches with Composition-Based Statistics and Other Refinements. *Nucleic Acids Research*, **29**, 2994-3005. https://doi.org/10.1093/nar/29.14.2994

76. Chou, K.C. and Shen, H.B. (2007) MemType-2L: A Web Server for Predicting Membrane Proteins and Their Types by Incorporating Evolution Information through Pse-PSSM. *Biochemical and Biophysical Research Communications* (*BBRC*), **360**, 339-345. https://doi.org/10.1016/j.bbrc.2007.06.027

77. Shen, H.B. and Chou, K.C. (2007) EzyPred: A Top-Down Approach for Predicting Enzyme Functional Classes and Subclasses. *Biochemical and Biophysical Research Communications* (*BBRC*), **364**, 53-59. https://doi.org/10.1016/j.bbrc.2007.09.098

78. Shen, H.B. and Chou, K.C. (2009) QuatIdent: A Web Server for Identifying Protein Quaternary Structural Attribute by Fusing Functional Domain and Sequential Evolution Information. *Journal of Proteome Research*, **8**, 1577-1584. https://doi.org/10.1021/pr800957q

79. Chou, K.C. and Shen, H.B. (2010) A New Method for Predicting the Subcellular Localization of Eukaryotic Proteins with Both Single and Multiple Sites: Euk-mPLoc 2.0. *PLoS ONE*, **5**, e9931. https://doi.org/10.1371/journal.pone.0009931

80. Wu, Z.C., Xiao, X. and Chou, K.C. (2011) iLoc-Plant: A Multi-Label Classifier for Predicting the Subcellular Localization of Plant Proteins with Both Single and Multiple Sites. *Molecular BioSystems*, **7**, 3287-3297. https://doi.org/10.1039/c1mb05232b

81. Chou, K.C., Wu, Z.C. and Xiao, X. (2012) iLoc-Hum: Using Accumulation-Label Scale to Predict Subcellular Locations of Human Proteins with Both Single and Multiple Sites. *Molecular BioSystems*, **8**, 629-641. https://doi.org/10.1039/C1MB05420A

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

83. Xiao, X., Wang, P. and Chou, K.C. (2009) GPCR-CA: A Cellular Automaton Image Approach for Predicting

G-Protein-Coupled Receptor Functional Classes. *Journal of Computational Chemistry*, **30**, 1414-1423. https://doi.org/10.1002/jcc.21163

84. Lin, W.Z., Fang, J.A., Xiao, X. and Chou, K.C. (2011) iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *PLoS ONE*, **6**, e24756. https://doi.org/10.1371/journal.pone.0024756
85. Lin, W.Z., Fang, J.A., Xiao, X. and Chou, K.C. (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. https://doi.org/10.1371/journal.pone.0049040
86. Lin, W.Z., Fang, J.A., Xiao, X. and Chou, K.C. (2013) iLoc-Animal: A Multi-Label Learning Classifier for Predicting Subcellular Localization of animal Proteins. *Molecular BioSystems*, **9**, 634-644. https://doi.org/10.1039/c3mb25466f
87. Zhang, C.T. and Chou, K.C. (1992) An Optimization Approach to Predicting Protein Structural Class from Amino Acid Composition. *Protein Science*, **1**, 401-408. https://doi.org/10.1002/pro.5560010312
88. Chou, K.C. and Elrod, D.W. (2002) Bioinformatical Analysis of G-Protein-Coupled Receptors. *Journal of Proteome Research*, **1**, 429-433. https://doi.org/10.1021/pr025527k
89. Chou, K.C. and Cai, Y.D. (2003) Prediction and Classification of Protein Subcellular Location: Sequence-Order Effect and Pseudo Amino Acid Composition. *Journal of Cellular Biochemistry*, **90**, 1250-1260. https://doi.org/10.1002/jcb.10719
90. Hu, L., Huang, T., Shi, X., Lu, W.C., Cai, Y.D. and Chou, K.C. (2011) Predicting Functions of Proteins in Mouse Based on Weighted Protein-Protein Interaction Network and Protein Hybrid Properties. *PLoS ONE*, **6**, e14556. https://doi.org/10.1371/journal.pone.0014556
91. Chou, K.C. (2005) Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics*, **21**, 10-19. https://doi.org/10.1093/bioinformatics/bth466
92. Kabir, M. and Hayat, M. (2016) iRSpot-GAEnsC: Identifying Recombination Spots via Ensemble Classifier and Extending the Concept of Chou's PseAAC to Formulate DNA Samples. *Molecular Genetics and Genomics*, **291**, 285-296. https://doi.org/10.1007/s00438-015-1108-5
93. Meher, P.K., Sahu, T.K., Saini, V. and Rao, A.R. (2017) Predicting Antimicrobial Peptides with Improved Accuracy by Incorporating the Compositional, Physico-Chemical and Structural Features into Chou's General PseAAC. *Scientific Reports*, **7**, Article ID: 42362. https://doi.org/10.1038/srep42362
94. Ju, Z. and He, J.J. (2017) Prediction of Lysine Propionylation Sites Using Biased SVM and Incorporating Four Different Sequence Features into Chou's PseAAC. *Journal of Molecular Graphics and Modelling*, **76**, 356-363. https://doi.org/10.1016/j.jmgm.2017.07.022
95. Yu, B., Li, S., Qiu, W.Y., Chen, C., Chen, R.X., Wang, L., Wang, M.H. and Zhang, Y. (2017) Accurate Prediction of Subcellular Location of Apoptosis Proteins Combining Chou's PseAAC and PsePSSM Based on Wavelet Denoising. *Oncotarget*, **8**, 107640-107665. https://doi.org/10.18632/oncotarget.22585
96. 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. https://doi.org/10.2174/1568026617666170414145508
97. Shen, H.B. and Chou, K.C. (2008) PseAAC: A Flexible Web-Server for Generating Various Kinds of Protein Pseudo Amino Acid Composition. *Analytical Biochemistry*, **373**, 386-388. https://doi.org/10.1016/j.ab.2007.10.012
98. Du, P., Wang, X., Xu, C. and Gao, Y. (2012) PseAAC-Builder: A Cross-Platform Stand-Alone Program for Generating Various Special Chou's Pseudo AMINO Acid Compositions. *Analytical Biochemistry*, **425**, 117-119. https://doi.org/10.1016/j.ab.2012.03.015

99. Cao, D.S., Xu, Q.S. and Liang, Y.Z. (2013) Propy: A Tool to Generate Various Modes of Chou's PseAAC. *Bioinformatics*, **29**, 960-962. https://doi.org/10.1093/bioinformatics/btt072

100. Du, P., Gu, S. and Jiao, Y. (2014) PseAAC-General: Fast Building Various Modes of General Form of Chou's Pseudo Amino Acid Composition for Large-Scale Protein Datasets. *International Journal of Molecular Sciences*, **15**, 3495-3506. https://doi.org/10.3390/ijms15033495

101. Chou, K.C. (2009) Pseudo Amino Acid Composition and Its Applications in Bioinformatics, Proteomics and System Biology. *Current Proteomics*, **6**, 262-274. https://doi.org/10.2174/157016409789973707

102. Chen, W., Lin, H. and Chou, K.C. (2015) Pseudo Nucleotide Composition or PseKNC: An Effective Formulation for Analyzing Genomic Sequences. *Molecular BioSystems*, **11**, 2620-2634. https://doi.org/10.1039/C5MB00155B

103. Liu, B., Yang, F., Huang, D.S. and Chou, K.C. (2018) iPromoter-2L: A Two-Layer Predictor for Identifying Promoters and Their Types by Multi-Window-Based PseKNC. *Bioinformatics*, **34**, 33-40. https://doi.org/10.1093/bioinformatics/btx579

104. Liu, B., Liu, F., Wang, X., Chen, J., Fang, L. and Chou, K.C. (2015) Pse-in-One: A Web Server for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Nucleic Acids Research*, **43**, W65-W71. https://doi.org/10.1093/nar/gkv458

105. Liu, B., Wu, H. and Chou, K.C. (2017) Pse-in-One 2.0: An Improved Package of Web Servers for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Natural Science*, **9**, 67-91. https://doi.org/10.4236/ns.2017.94007

106. 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* (*PEDS*), **22**, 699-705. https://doi.org/10.1093/protein/gzp057

107. Chou, K.C., Liu, W., Maggiora, G.M. and Zhang, C.T. (1998) Prediction and Classification of Domain Structural Classes. *Proteins*: *Structure*, *Function and Genetics*, **31**, 97-103. https://doi.org/10.1002/(SICI)1097-0134(19980401)31:1<97::AID-PROT8>3.0.CO;2-E

108. Chou, K.C. and Maggiora, G.M. (1998) Domain Structural Class Prediction. *Protein Engineering*, **11**, 523-538. https://doi.org/10.1093/protein/11.7.523

109. Chou, K.C. and Cai, Y.D. (2002) Using Functional Domain Composition and Support Vector Machines for Prediction of Protein Subcellular Location. *The Journal of Biological Chemistry*, **277**, 45765-45769. https://doi.org/10.1074/jbc.M204161200

110. Chou, K.C. and Cai, Y.D. (2004) Predicting Protein Structural Class by Functional Domain Composition. *Biochemical and Biophysical Research Communications* (*BBRC*), **321**, 1007-1009. https://doi.org/10.1016/j.bbrc.2004.07.059

111. Chou, K.C. and Cai, Y.D. (2004) Predicting Subcellular Localization of Proteins by Hybridizing Functional Domain Composition and Pseudo Amino Acid Composition. *Journal of Cellular Biochemistry*, **91**, 1197-1203. https://doi.org/10.1002/jcb.10790

112. Cai, Y.D. and Chou, K.C. (2005) Using Functional Domain Composition to Predict Enzyme Family Classes. *Journal of Proteome Research*, **4**, 109-111. https://doi.org/10.1021/pr049835p

113. Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Research*, **39**, W29-W37. https://doi.org/10.1093/nar/gkr367

114. Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A. and Eddy, S.R. (2015) HMMER Web Server: 2015 Update. *Nucleic Acids Research*, **43**, W30-W38. https://doi.org/10.1093/nar/gkv397

115. Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R. and Finn, R.D. (2018) HMMER Web Server: 2018 Up-

date. *Nucleic Acids Research*, **46**, W200-W204. https://doi.org/10.1093/nar/gky448

116. Liu, S.F., Fang, Z.G. and Lin, Y. (2006) A New Definition for the Degree of Grey Incidence. *Scientific Inquiry*, **7**, 111-124.
117. Chou, K.C. and Zhang, C.T. (1995) Review: Prediction of Protein Structural Classes. *Critical Reviews in Biochemistry and Molecular Biology*, **30**, 275-349. https://doi.org/10.3109/10409239509083488
118. Chou, K.C. and Shen, H.B. (2008) Cell-PLoc: A Package of Web Servers for Predicting Subcellular Localization of Proteins in Various Organisms. *Nature Protocols*, **3**, 153-162. https://doi.org/10.1038/nprot.2007.494
119. Chou, K.C. and Shen, H.B. (2010) Cell-PLoc 2.0: An Improved Package of Web-Servers for Predicting Subcellular Localization of Proteins in Various Organisms. *Natural Science*, **2**, 1090-1103. https://doi.org/10.4236/ns.2010.210136
120. Mohabatkar, H. (2010) Prediction of Cyclin Proteins Using Chou's Pseudo Amino Acid Composition. *Protein & Peptide Letters*, **17**, 1207-1214. https://doi.org/10.2174/092986610792231564
121. Sahu, S.S. and Panda, G. (2010) A Novel Feature Representation Method Based on Chou's Pseudo Amino Acid Composition for Protein Structural Class Prediction. *Computational Biology and Chemistry*, **34**, 320-327. https://doi.org/10.1016/j.compbiolchem.2010.09.002
122. Zia-ur-Rehman and Khan, A. (2012) Identifying GPCRs and Their Types with Chou's Pseudo Amino Acid Composition: An Approach from Multi-Scale Energy Representation and Position Specific Scoring Matrix. *Protein & Peptide Letters*, **19**, 890-903. https://doi.org/10.2174/092986612801619589
123. Fan, G.L. and Li, Q.Z. (2013) Discriminating Bioluminescent Proteins by Incorporating Average Chemical Shift and Evolutionary Information into the General form of Chou's Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **334**, 45-51. https://doi.org/10.1016/j.jtbi.2013.06.003
124. Chou, K.C. and Shen, H.B. (2009) Recent Advances in Developing Web-Servers for Predicting Protein Attributes. *Natural Science*, **1**, 63-92. https://doi.org/10.4236/ns.2009.12011
125. Cheng, X., Xiao, X. and Chou, K.C. (2017) pLoc-mPlant: Predict Subcellular Localization of Multi-Location Plant Proteins via Incorporating the Optimal GO Information into General PseAAC. *Molecular BioSystems*, **13**, 1722-1727. https://doi.org/10.1039/C7MB00267J
126. Cheng, X., Xiao, X. and Chou, K.C. (2017) pLoc-mVirus: Predict Subcellular Localization of Multi-Location Virus Proteins via Incorporating the Optimal GO Information into General PseAAC. *Gene*, **628**, 315-321. https://doi.org/10.1016/j.gene.2017.07.036
127. 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. https://doi.org/10.1016/j.ygeno.2017.08.005
128. 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. https://doi.org/10.1016/j.ygeno.2017.10.002
129. Cheng, X., Zhao, S.G., Lin, W.Z., Xiao, X. and Chou, K.C. (2017) pLoc-mAnimal: Predict Subcellular Localization of Animal Proteins with Both Single and Multiple Sites. *Bioinformatics*, **33**, 3524-3531. https://doi.org/10.1093/bioinformatics/btx476
130. Xiao, X., Cheng, X., Su, S., Nao, Q. and Chou, K.C. (2017) pLoc-mGpos: Incorporate Key Gene Ontology Information into General PseAAC for Predicting Subcellular Localization of Gram-Positive Bacterial Proteins. *Natural Science*, **9**, 331-349. https://doi.org/10.4236/ns.2017.99032
131. 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. https://doi.org/10.1093/bioinformatics/btx711

132. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., Jia, J.H. and Chou, K.C. (2018) iKcr-PseEns: Identify Lysine Crotonylation Sites in Histone Proteins with Pseudo Components and Ensemble Classifier. *Genomics*, **110**, 239-246. https://doi.org/10.1016/j.ygeno.2017.10.008
133. Cheng, X., Zhao, S.G., Xiao, X. and Chou, K.C. (2017) iATC-mISF: A Multi-Label Classifier for Predicting the Classes of Anatomical Therapeutic Chemicals. *Bioinformatics*, **33**, 341-346. https://doi.org/10.1093/bioinformatics/btx387
134. Feng, P., Ding, H., Yang, H., Chen, W., Lin, H. and Chou, K.C. (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. https://doi.org/10.1016/j.omtn.2017.03.006
135. Liu, B., Wang, S., Long, R. and Chou, K.C. (2017) iRSpot-EL: Identify Recombination Spots with an Ensemble Learning Approach. *Bioinformatics*, **33**, 35-41. https://doi.org/10.1093/bioinformatics/btw539
136. 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. https://doi.org/10.1016/j.omtn.2017.04.008
137. Qiu, W.R., Jiang, S.Y., Xu, Z.C., Xiao, X. and Chou, K.C. (2017) iRNAm5C-PseDNC: Identifying RNA 5-Methylcytosine Sites by Incorporating Physical-Chemical Properties into Pseudo Dinucleotide Composition. *Oncotarget*, **8**, 41178-41188. https://doi.org/10.18632/oncotarget.17104
138. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, D. and Chou, K.C. (2017) iPhos-PseEvo: Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into General PseAAC via Grey System Theory. *Molecular Informatics*, **36**, UNSP 1600010. https://doi.org/10.1002/minf.201600010
139. Chou, K.C., Cheng, X. and Xiao, X. (2019) pLoc_bal-mEuk: Predict Subcellular Localization of Eukaryotic Proteins by General PseAAC and Quasi-Balancing Training Dataset. *Medicinal Chemistry*, **15**, 472-485. https://doi.org/10.2174/1573406415666181218102517
140. Cheng, X., Xiao, X. and Chou, K.C. (2018) pLoc_bal-mGneg: Predict Subcellular Localization of Gram-Negative Bacterial Proteins by Quasi-Balancing Training Dataset and General PseAAC. *Journal of Theoretical Biology*, **458**, 92-102. https://doi.org/10.1016/j.jtbi.2018.09.005
141. Cheng, X., Xiao, X. and Chou, K.C. (2018) pLoc_bal-mPlant: Predict Subcellular Localization of Plant Proteins by General PseAAC and Balancing Training Dataset. *Current Pharmaceutical Design*, **24**, 4013-4022. https://doi.org/10.2174/1381612824666181119145030
142. Chou, K.C., Cheng, X. and Xiao, X. (2019) pLoc_bal-mHum: Predict Subcellular Localization of Human Proteins by PseAAC and Quasi-Balancing Training Dataset. *Genomics*, **111**, 1274-1282. https://doi.org/10.1016/j.ygeno.2018.08.007
143. Xiao, X., Cheng, X., Chen, G., Mao, Q. and Chou, K.C. (2019) pLoc_bal-mGpos: Predict Subcellular Localization of Gram-Positive Bacterial Proteins by Quasi-Balancing Training Dataset and PseAAC. *Genomics*, **111**, 886-892. https://doi.org/10.1016/j.ygeno.2018.05.017
144. Cheng, X., Lin, W.Z., Xiao, X. and Chou, K.C. (2019) pLoc_bal-mAnimal: Predict Subcellular Localization of Animal Proteins by Balancing Training Dataset and PseAAC. *Bioinformatics*, **35**, 398-406. https://doi.org/10.1093/bioinformatics/bty628
145. Chou, K.C., Jiang, S.P., Liu, W.M. and Fee, C.H. (1979) Graph Theory of Enzyme Kinetics: 1. Steady-State Reaction System. *Scientia Sinica*, **22**, 341-358.
146. Chou, K.C. and Forsen, S. (1980) Graphical Rules for Enzyme-Catalyzed Rate Laws. *Biochemical Journal*, **187**, 829-835. https://doi.org/10.1042/bj1870829

147. Chou, K.C., Forsen, S. and Zhou, G.Q. (1980) Three Schematic Rules for Deriving Apparent Rate Constants. *Chemica Scripta*, **16**, 109-113.

148. Chou, K.C., Carter, R.E. and Forsen, S. (1981) A New Graphical Method for Deriving Rate Equations for Complicated Mechanisms. *Chemica Scripta*, **18**, 82-86.

149. Chou, K.C. and Forsen, S. (1981) Graphical Rules of Steady-State Reaction Systems. *Canadian Journal of Chemistry*, **59**, 737-755. https://doi.org/10.1139/v81-107

150. Zhou, G.P. and Deng, M.H. (1984) An Extension of Chou's Graphic Rules for Deriving Enzyme Kinetic Equations to Systems Involving Parallel Reaction Pathways. *Biochemical Journal*, **222**, 169-176. https://doi.org/10.1042/bj2220169

151. Chou, K.C. (1989) Graphic Rules in Steady and Non-Steady Enzyme Kinetics. *The Journal of Biological Chemistry*, **264**, 12074-12079.

152. Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1993) Steady-State Kinetic Studies with the Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitor U-87201E. *The Journal of Biological Chemistry*, **268**, 6119-6124.

153. Chou, K.C. (1990) Review: Applications of Graph Theory to Enzyme Kinetics and Protein Folding Kinetics. Steady and Non-Steady State Systems. *Biophysical Chemistry*, **35**, 1-24. https://doi.org/10.1016/0301-4622(90)80056-D

154. Althaus, I.W., Gonzales, A.J., Chou, J.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1993) The Quinoline U-78036 Is a Potent Inhibitor of HIV-1 Reverse Transcriptase. *The Journal of Biological Chemistry*, **268**, 14875-14880.

155. Chou, K.C. (2010) Graphic Rule for Drug Metabolism Systems. *Current Drug Metabolism*, **11**, 369-378. https://doi.org/10.2174/138920010791514261

156. Zhou, G.P. (2011) The Disposition of the LZCC Protein Residues in Wenxiang Diagram Provides New Insights into the Protein-Protein Interaction Mechanism. *Journal of Theoretical Biology*, **284**, 142-148. https://doi.org/10.1016/j.jtbi.2011.06.006

157. Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G. and Reusser, F. (1993) Kinetic Studies with the Nonnucleoside HIV-1 Reverse Transcriptase Inhibitor U-88204E. *Biochemistry*, **32**, 6548-6554. https://doi.org/10.1021/bi00077a008

158. Chou, K.C., Lin, W.Z. and Xiao, X. (2011) Wenxiang: A Web-Server for Drawing Wenxiang Diagrams. *Natural Science*, **3**, 862-865. https://doi.org/10.4236/ns.2011.310111

159. Chou, K.C. and Forsen, S. (1980) Diffusion-Controlled Effects in Reversible Enzymatic Fast Reaction System: Critical Spherical Shell and Proximity Rate Constants. *Biophysical Chemistry*, **12**, 255-263. https://doi.org/10.1016/0301-4622(80)80002-0

160. Chou, K.C., Li, T.T. and Forsen, S. (1980) The Critical Spherical Shell in Enzymatic Fast Reaction Systems. *Biophysical Chemistry*, **12**, 265-269. https://doi.org/10.1016/0301-4622(80)80003-2

161. Shen, H.B., Song, J.N. and Chou, K.C. (2009) Prediction of Protein Folding Rates from Primary Sequence by Fusing Multiple Sequential Features. *Journal of Biomedical Science and Engineering*, **2**, 136-143. https://doi.org/10.4236/jbise.2009.23024

162. Chou, K.C., Chen, N.Y. and Forsen, S. (1981) The Biological Functions of Low-Frequency Phonons: 2. Cooperative Effects. *Chemica Scripta*, **18**, 126-132.

163. Chou, K.C. (1988) Review: Low-Frequency Collective Motion in Biomacromolecules and Its Biological Functions. *Biophysical Chemistry*, **30**, 3-48. https://doi.org/10.1016/0301-4622(88)85002-6

164. Chou, K.C. (2019) The Cradle of Gordon Life Science Institute and Its Development and Driving Force. *Int J Biol Genetics*, **1**, 1-28.

165. Chou, K.C. (2019) Showcase to Illustrate How the Web-Server iDNA6mA-PseKNC Is Working. *Journal of Pathology Research Reviews & Reports*, **1**, 1-15.

166. Chou, K.C. (2019) The pLoc_bal-mPlant Is a Powerful Artificial Intelligence Tool for Predicting the Subcellular Localization of Plant Proteins Purely Based on Their Sequence Information. *International Journal of Nutrition Sciences*, **4**, 1-4.

167. Chou, K.C. (2019) Showcase to Illustrate How the Web-Server iNitro-Tyr Is Working. *Glo J of Com Sci and Infor Tec.*, **2**, 1-16.

168. Chou, K.C. (2019) Gordon Life Science Institute: Its Philosophy, Achievements, and Perspective. *Annals of Cancer Therapy and Pharmacology*, **2**, 1-26. https://onomyscience.com/onomy/cancer_archive_volume2_issue2.html

169. Chou, K.C. (2019) The pLoc_bal-mAnimal Is a Powerful Artificial Intelligence Tool for Predicting the Subcellular Localization of Animal Proteins Based on Their Sequence Information Alone. *Scientific Journal of Biometrics & Biostatistics*, **2**, 1-13.

170. Chou, K.C. (2020) Showcase to Illustrate How the Webserver pLoc_bal-mEuk Is Working. *Biomedical Journal of Scientific & Technical Research*. https://doi.org/10.18483/ijSci.2247

171. Chou, K.C. (2020) The pLoc_bal-mGneg Predictor Is a Powerful Web-Server for Identifying the Subcellular Localization of Gram-Negative Bacterial Proteins Based on Their Sequences Information Alone. *ijSci*, **9**, 27-34. https://doi.org/10.18483/ijSci.2248

172. Chou, K.C. (2020) How the Artificial Intelligence Tool iRNA-2methyl Is Working for RNA 2'-Omethylation Sites. *Journal of Medical Care Research and Review*, **3**, 348-366.

173. Chou, K.-C. (2020) Showcase to Illustrate How the Web-Server iKcr-PseEns Is Working. *Journal of Medical Care Research and Review*, **3**, 331-347. https://doi.org/10.18483/ijSci.2247

174. Chou, K.C. (2020) The pLoc_bal-mVirus Is a Powerful Artificial Intelligence Tool for Predicting the Subcellular Localization of Virus Proteins According to Their Sequence Information Alone. *Journal of Genetics and Genomics*, **4**.

175. Chou, K.C. (2019) How the Artificial Intelligence Tool iSNO-PseAAC Is Working in Predicting the Cysteine s-Nitrosylation Sites in Proteins. *Journal of Stem Cell Research and Medicine*, **4**, 1-9.

176. Chou, K.C. (2020) Showcase to Illustrate How the Web-Server iRNA-Methyl Is Working. *Journal of Molecular Genetics*, **3**, 1-7.

177. Chou, K.C. (2020) How the Artificial Intelligence Tool iRNA-PseU Is Working in Predicting the RNA Pseudouridine Sites. *Biomedical Journal of Scientific & Technical Research*, **24**.

178. Chou, K.C. (2020) Showcase to Illustrate How the Web-Server iSNO-AAPair Is Working. *Journal of Genetics and Genomics*, **4**. https://doi.org/10.18483/ijSci.2247

179. Chou, K.C. (2020) The pLoc_bal-mHum Is a Powerful Web-Serve for Predicting the Subcellular Localization of Human Proteins Purely Based on Their Sequence Information. *Advances in Bioengineering and Biomedical Science Research*, **3**, 1-5.

180. Chou, K.C. (2020) Showcase to Illustrate How the Web-Server iPTM-mLys Is Working. *Infotext Journal of Infectious Diseases and Therapy*, **1**, 1-16.

181. Chou, K.C. (2020) The pLoc_bal-mGpos Is a Powerful Artificial Intelligence Tool for Predicting the Subcellular Localization of Gram-Positive Bacterial Proteins According to Their Sequence Information Alone. *Glo J of Com Sci and Infor Tec*, **2**, 1-13.

182. Chou, K.C. (2020) Showcase to Illustrate How the Web-Server iPreny-PseAAC Is Working. *Glo J of Com Sci and Infor Tec.*, **2**, 1-15.

183. Chou, K.C. (2020) Some Illuminating Remarks on Molecular Genetics and Genomics as Well as Drug Development. *Molecular Genetics and Genomics*, **295**, 261-274. https://doi.org/10.1007/s00438-019-01634-z

184. Chou, K.C. (2020) The Problem of Elsevier Series Journals Online Submission by Using Artificial Intelligence. *Natural Science*, **12**, 37-38. https://doi.org/10.4236/ns.2020.122006

185. Chou, K.C. (2020) The Most Important Ethical Concerns in Science. *Natural Science*, **12**, 35-36. https://doi.org/10.4236/ns.2020.122005

186. Chou, K.C. (2020) Other Mountain Stones Can Attack Jade: The 5-Steps Rule. *Natural Scienc*e, **12**, 59-64. https://doi.org/10.4236/ns.2020.123011

187. Chou, K.C. (2020) Using Similarity Software to Evaluate Scientific Paper Quality Is a Big Mistake, *Natural Science*, **12**, 42-58. https://doi.org/10.4236/ns.2020.123008

188. Chou, K.C. (2020) Gordon Life Science Institute and Its Impacts on Computational Biology and Drug Development. *Natural Science*, **12**, 125-161. https://doi.org/10.4236/ns.2020.123013