Using Grey Model to Predict Protein Remote Homologous Family

Xuan Xiao\textsuperscript{a}, Weihua Cao\textsuperscript{b} and Weizhong Lin\textsuperscript{*}

Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403 China

*Corresponding author e-mail: linweizhong@jci.edu.cn; \textsuperscript{a}jdzxiaoxuan@163.com; \textsuperscript{b}2364743501@qq.com

Abstract. Protein remote homology detection is the most basic and core problems of protein structure and function research. The purpose of protein remote homology detection is to detect the remote evolutionary relationship between proteins by computation methods. At present, there are many methods for remote homology detection of proteins, but some methods are not very effective. In bioinformatics, it is urgent to further improve the performance of protein remote homology detection. In this study, we propose a new model called PHom-GRA to detect protein remote homology, which is the integration of various ranking methods via using grey relational analysis. Our experiment constructs benchmark dataset with lower homology, in which any two proteins have not more than 40\% identity or homology. We achieve an ROC1 score of 0.7372 and an ROC50 score of 0.7968 in jack-knife test.

1. Introduction
Detecting remote homology relationship among proteins plays one of the fundamental and central roles in computational proteomics. With the development of sequencing techniques, the protein sequence data rapid raise. To find those proteins structure and function is more and more urgent. Although we can apply the experimental methods such as NMR spectroscopy and X-ray crystallography to determine proteins structures and function, not all protein can be applied those experimental methods [1]. Thus it is significant to develop effectively automatic predictor.

Usually, we can predict a protein’s structure or function by seeking proteins those have similarity in sequence. Unfortunately, these methods have no effective on remote homology protein pairs. The reason is that while having similar structures and functions, these protein pairs does not have similar sequences. Now, some computational models have been proposed to address this problem. Liu et.al developed server classifiers to predict remote homology proteins [2-7]. In their works, they applied alignment methods, discriminative methods, and ranking methods. Among these methods, ranking methods were reported to get the state-of-the-art performance.

Although Liu et.al had achieved great performances predictor, there were some defects. Firstly, the benchmark datasets adopted in their work had high similarity. In the benchmark dataset [3, 5] which containing 7329 proteins from 1070 different super families, the identity of any two sequences was lower than 95\%. So some proteins pairs had higher similarity. Another defect was the ranking algorithm used in those studies, LambdaMART, spends a lot of time to training the learning model.
For example, if the training dataset has N proteins, the LambdaMART need to deal with N^2 proteins pair samples.

In this work, we cut down the identity of any two sequences in the benchmark dataset to 40% and used grey relational analysis (GRA) as the ranking method. Moreover, we developed a new feature representing model of protein and built a novel predictor PHom-GRA to seek remote homologous proteins for a query protein.

2. Materials and Methods

2.1. Benchmark Dataset
The benchmark dataset used in this study was taken from Liu [5]. This dataset included 7329 proteins from 1070 different super families and 1824 families which were constructed based on SCOP database. To reduce the redundancy and homology bias, the program CD-HIT [8] 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. Feature Representation
It is the key work for protein homology detecting to formulate the protein samples with a feature matrices. For a protein:

\[ P = R_1R_2R_3L R_l \quad (1) \]

Its feature vector can be represented as:

\[ P = [\psi_1 \ \psi_2 \ \psi_3 \ \psi_4 \ \psi_5] \quad (2) \]

Here two feature representation models were adopted to generate the formulation (2).

2.2.1. Grey-PSSM. This model was proposed by Lin [9, 10] firstly. It extracted the sequential evolution information by using Position Specific Scoring Matrix (PSSM). The protein P with L amino acid residues (see eq. 1) was expressed by a 20×L matrix which were generated by using PSI-BLAST[11] to search the UniProtKB/Swiss-Prot database.

\[ PSSM = \begin{bmatrix}
m_{1,1} & m_{1,2} & L & m_{1,20} \\
m_{2,1} & m_{2,2} & L & m_{2,20} \\
M & M & M & M \\
m_{L,1} & m_{L,2} & L & m_{L,20}
\end{bmatrix} \quad (3) \]

Where \( m_{i,j} \) (1 ≤ i ≤ L, 1 ≤ j ≤ 20) represents the score of amino acid residue in i-th sequential position of the protein that is being changed to amino acid type j during the evolution process. Here, the numerical codes 1,2,…,20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes.

Finally, A 60 dimension feature vector was derived to represent the protein (eq. 1). This Grey-PSSM model can be seen in [9, 10].

2.2.2. PCA-GLCM. Xiao[12, 13] described a protein as the cellular automation image and constructed its grey level co-occurrence matrix (GLCM) to express the protein. In this study, the four GLCM-
derived features: angular second moment, contrast, inverse different moment, and entropy, were used to formulate the protein as Eq.2.

2.3. Method
We used ranking method to detect proteins remote homology. Given a query protein, the predictor searched it against the benchmark database and return the top ranked proteins. PSI-BLAST is a popular tool to tackle this problem. PSI-BLAST searches a query protein by using multiple sequence alignment and return proteins identity. Similarly, grey relational analysis[14, 15] is also adopted to rank the relationship of proteins. This algorithm was described in [16].

The predictor PHom-GRA was an integrated method to detect proteins remote homology. This model is seen in Figure 1. A query protein can be expressed by Grey-PSSM and PCA-GLCM respectively. Subsequently, relational scores between the query protein and every protein of the benchmark database are computed by using GRA. Finally, PHom-GRA combines the two scores and the alignment score got by PSI-BLAST to search the top ranked homologous proteins of the query protein.

![Figure 1. Flowchart of PHom-GRA](image-url)

3. Result and discussion
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. Because the LambdaMART ranking algorithm used in preview studies [3, 5] consumed more training time and computer memory, the 5-fold cross-validation test was adopted in these studies. In this work, we employed GRA to compute the relationship score between the query protein and benchmark dataset proteins, so the computing time and memory jumped down. Therefore the jackknife test was adopted in our experiment.

The predictive results were listed in Table 1, from which we can extract that PHom-GRA achieved the best performance.
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 |
| PHom-GRA (PSI-BLAST+Grey-PSSM)       | 0.7110| 0.7737 |
| PHom-GRA (PSI-BLAST+PCA-GLCM)        | 0.7138| 0.7652 |
| PHom-GRA (PSI-BLAST+Grey-PSSM+PCA-GLCM) | 0.7371| 0.7968 |

As mentioned in this table, PHom-GRA, which combined the alignment score of PSI-BLAST, the grey incidence degree of Grey-PSSM feature, and the grey incidence degree of PCA-GLCM, achieved the best performance, the score of ROC1 and ROC50 were higher 0.03 than other methods listed in the table.

4. Conclusion
Protein remote homology detection is vital for studying protein structures and functions. It is a challenging task for proteins with low sequence similarities to search remote homologous proteins. In methods proposed, it is reported that learning to rank (LTR) had achieved the state-of-the-art performance. Since this approach need consume too much time and memory in training, a novel predictor PHom-GRA was proposed. It used GRA to compute the similarity between the query protein and proteins in benchmark dataset through the protein feature expression model. Meanwhile it integrate GRA result and PSI-BLAST into one framework. Compared with the PSI-BLAST searching benchmark dataset, PHom-GRA promoted prediction performance. Experimental results disclosed that scores of ROC1 and ROC50 got by PHom-GRA were superior to those got by PSI-BLAST. An additional advantage was that PHom-GRA wasted much less the training time and computer memory.

Acknowledgments
This work was support by the grants from the National Natural Science Foundation of China (No.61462047, 31560316), 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.

References
[1] Zhou, G.P., R.B. Huang, and F.A. Troy, 2nd, 3D structural conformation and functional domains of polysialyltransferase ST8Sia IV required for polysialylation of neural cell adhesion molecules. Protein Pept Lett, 2015. 22(2): p. 137-48.
[2] Chen, J., et al., A comprehensive review and comparison of different computational methods for protein remote homology detection. Brief Bioinform, 2016.
[3] Chen, J., et al., dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. Scientific Reports, 2016. 6: p. 32333.
[4] Liu, B., J. Chen, and S. Wang, Protein Remote Homology Detection by Combining Pseudo Dimer Composition with an Ensemble Learning Method. Current Proteomics, 2016. 13(2): p. 86-91.
[5] Liu, B., J. Chen, and X. Wang, Application of learning to rank to protein remote homology detection. Bioinformatics, 2015. 31(21): p. 3492-3498.
[6] Liu, B., J. Chen, and X. Wang, Protein remote homology detection by combining Chou’s distance-pair pseudo amino acid composition and principal component analysis. Molecular Genetics and Genomics, 2015. 290(5): p. 1919-1931.
[7] Liu, B., et al., Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics, 2014. 30(4):
[8] Huang, Y., et al., CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics, 2010. 26(5): p. 680-2.

[9] Lin, W.Z., et al., iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol Biosyst, 2013. 9(4): p. 634-644.

[10] Lin, W.Z., et al., Predicting secretory proteins of malaria parasite by incorporating sequence evolution information into pseudo amino acid composition via grey system model. PLoS One, 2012. 7(11): p. e49040.

[11] Schaffer, A.A., et al., Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res, 2001. 29(14): p. 2994-3005.

[12] Xiao, X., P. Wang, and K.C. Chou, GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J Comput Chem, 2009. 30(9): p. 1414-23.

[13] Xiao, X., P. Wang, and K.C. Chou, Cellular automata and its applications in protein bioinformatics. Curr Protein Pept Sci, 2011. 12(6): p. 508-19.

[14] Deng, J.L., Introduction to Grey System Theory. The Journal of Grey System, 1989(1): p. 1-24.

[15] Liu, S., Z. Fang, and Y. Lin, A new definition for the degree of grey incidence. Scientific Inquiry, 2006. 7(2): p. 111-124.

[16] Lin, W.Z., X. Xiao, and K.C. Chou, GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis. Protein Eng Des Sel, 2009. 22(11): p. 699-705.