PLOS | ONE

RESEARCH ARTICLE

Exploring Mouse Protein Function via Multiple Approaches

Guohua Huang[1☯], Chen Chu[3☯], Tao Huang[4], Xiangyin Kong[4], Yunhua Zhang[5], Ning Zhang[6]*, Yu-Dong Cai[2]*

1 Department of Mathematics, Shaoyang University, Shaoyang, Hunan, 422000, China, **2** School of Life Sciences, Shanghai University, 99 Shangda Road, Shanghai, 200444, China, **3** Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China, **4** Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China, **5** College of Life Science, Anhui Agricultural University, Hefei, Anhui, 230036, China, **6** Department of Biomedical Engineering, Tianjin Key Lab of Biomedical Engineering Measurement, Tianjin University, Tianjin, China

☯ These authors contributed equally to this work.
* zhni@tju.edu.cn (NZ); cai_yud@126.com (Y-DC)

Abstract

Although the number of available protein sequences is growing exponentially, functional protein annotations lag far behind. Therefore, accurate identification of protein functions remains one of the major challenges in molecular biology. In this study, we presented a novel approach to predict mouse protein functions. The approach was a sequential combination of a similarity-based approach, an interaction-based approach and a pseudo amino acid composition-based approach. The method achieved an accuracy of about 0.8450 for the 1st-order predictions in the leave-one-out and ten-fold cross-validations. For the results yielded by the leave-one-out cross-validation, although the similarity-based approach alone achieved an accuracy of 0.8756, it was unable to predict the functions of proteins with no homologues. Comparatively, the pseudo amino acid composition-based approach alone reached an accuracy of 0.6786. Although the accuracy was lower than that of the previous approach, it could predict the functions of almost all proteins, even proteins with no homologues. Therefore, the combined method balanced the advantages and disadvantages of both approaches to achieve efficient performance. Furthermore, the results yielded by the ten-fold cross-validation indicate that the combined method is still effective and stable when there are no close homologs are available. However, the accuracy of the predicted functions can only be determined according to known protein functions based on current knowledge. Many protein functions remain unknown. By exploring the functions of proteins for which the 1st-order predicted functions are wrong but the 2nd-order predicted functions are correct, the 1st-order wrongly predicted functions were shown to be closely associated with the genes encoding the proteins. The so-called wrongly predicted functions could also potentially be correct upon future experimental verification. Therefore, the accuracy of the presented method may be much higher in reality.

OPEN ACCESS

Citation: Huang G, Chu C, Huang T, Kong X, Zhang Y, Zhang N, et al. (2016) Exploring Mouse Protein Function via Multiple Approaches. PLoS ONE 11 (11): e0166580. doi:10.1371/journal.pone.0166580

Editor: Alexey Porollo, Cincinnati Children's Hospital Medical Center, UNITED STATES

Received: August 14, 2016

Accepted: October 31, 2016

Published: November 15, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Funded by National Natural Science Foundation of China (61401302, 31371335, 81171342, 81201148), the Tianjin Research Program of the Application Foundation and Advanced Technology (14JCQNJC09500), the Innovation Program of the Shanghai Municipal Education Commission (12ZZ087), the National Research Foundation for the Doctoral Program of Higher Education of China (20130032120070, 20120032120073), the Scientific Research Fund of

Hunan Provincial Education Department (15B216), the Science and Technology Program of Hunan (2015JC3099), and the Seed Foundation of Tianjin University (60302064, 60302069).

Competing Interests: The authors have declared that no competing interests exist.

1 Introduction

Recent advances in sequencing technology have identified a large number of proteins that perform a wide variety of functions in cellular activities. Knowledge of protein function is crucial to understanding the mechanisms behind cellular processes and preventing and treating disease. However, most of the proteins identified to date have unknown functions. Approximately 1% of the more than 13 million protein sequences available have been experimentally annotated with essential functions; the remaining proteins have been marked with putative, uncharacterized, hypothetical, unknown or inferred functions [1]. Although physical experimental approaches, including high-throughput screening, are capable of determining the biological functions of proteins, they are expensive and time-consuming. Additionally, these methods are aimed at certain functions, which produce one-sided descriptions of protein function [2].

Computational approaches can make up for the deficiencies of experiments. Following the success of the computational approach in sequence alignment and comparison, many computational techniques have been presented to determine protein functions during the last decade [3]. The most commonly applied approach is to transfer functional annotation from the most similar protein with known functional information. Both sequence and structural similarities are heavily utilized in this type of homology-based annotation transfer. To infer protein function, the servers OntoBlast [4] and GoFigure [5] use the sequence alignment tool BLAST [6]. Confunc [7], the protein function prediction (PFP) algorithm [8] and the extended similarity group method (ESG) [9] employ the sequence alignment tool PSI_BLAST [10]. The Blast2GO suite is the homology transfer-based functional annotation of the gene ontology vocabulary [11]. Similar to the sequence similarity-based method, the structure similarity-based approach generally uses structure alignments via programs such as DaliLite [12–14], STRUCTAL [15], MultiProt [16], Bioinfo3D [17], and 3DCoffee [18] to measure homology among proteins. PHUNCTIONER [19] utilizes structural alignment to identify crucial positions in a protein that might hold clues to specific functions. Pegg *et al.* [20] constructed a structure-function link database and used it to correct the errors in the annotation of enzymes. Some researchers have attempted to combine the sequence and structure similarity approaches to explore protein function. For example, the FRalanyzer [21] uses sequence-structure alignments to elucidate protein function.

Recently, a large body of protein-protein interaction networks has become available to explore the functional relationships between interacting proteins. There are many computational models for predicting protein-protein interactions [22–24]. The commonly accepted hypothesis (called guilt-by-association (GBA) [25]) is that proteins are more likely to share identical or similar functions with interacting proteins than with non-interacting proteins. Since Schwikowski *et al.* [26] pioneered the utility of interaction networks for annotating protein functions in yeast, numerous interaction-based methods have been proposed to infer the functions of proteins. Hishigaki *et al.* [27] presented an improved predictive method called the Chi-square method to elucidate protein function. Chi *et al.* [28] used an iterative strategy to transfer neighboring protein functions. Chua *et al.* [29] extended the neighborhood to indirect neighbors, called 2-neighbors. These types of local network predictions mainly transferred functional annotation from the directly interacting neighborhood. Additionally, some global optimization techniques have been adopted to elucidate protein function. For example, Deng *et al.* [30], Letovsky *et al.* [31] and Kourmpetis *et al.* [32] used the Bayesian Markov random field method to infer protein functions from protein-protein interaction data and functional annotation of the protein interaction partners. The protein-protein interaction network is viewed as a graph, where the nodes represent proteins and the edges represent the interactions between proteins. Some graph-based methods have been presented for function predictions.

Nabieva *et al.* [33] modeled the functional annotation from the interaction network as a minimum multiway cut problem and introduced a network-flow algorithm that simulated the functional flow between proteins. The clustering-based and network alignment-based techniques have been employed to predict protein functions. Altaf-Ul-Amin *et al.* [34] and Arnau *et al.* [35] used different clustering techniques to classify protein functions, whereas Singh *et al.* [36] presented a global alignment of multiple protein interaction networks to infer protein functions. These approaches outperformed sequence similarity and local alignment of networks. Some researchers have presented a routine to predict protein function by combining multiple methods and data sources. For instance, Cozzetto *et al.* [2] integrated PSI-BLAST, text-mining, machine learning, and profile-profile comparisons to predict protein functions. As these authors noted, although considerable progress has been made, the functional annotation of integrative methods can be improved. Most of the above-mentioned networks are binary (*i.e.*, 1 indicates interaction and 0 indicates no interaction). Additionally, the interaction between proteins can be strong or weak. The STRING database [37] is a protein interaction repository that characterizes each interaction into a weight value based on eight different lines of evidence. Hu *et al.* [38] used a weighted interaction to predict protein function and achieved a promising performance.

Great progress has been made in the computational protein function prediction field, where state-of-the-art prediction algorithms substantially outperform first-generation methods and contribute to subsequent experimental studies. However, there still remains considerable need for the improvement of the current tools [39]. To this end, we presented an integrated method to explore mouse protein functions by fusing sequence similarities, weighted interactions from the STRING database and the pseudo amino acid composition of proteins. Unannotated proteins were aligned against a database consisting of proteins with known functions. If the query protein was homologous to well-annotated proteins, the alignment scores were used to infer function. If there were no known homologous proteins, we extracted weighted interactions from the STRING database and used them to predict the query protein function. For proteins whose functions the previous two approaches could not predict, we used the pseudo amino acid composition (PseAAC)-based nearest-neighbor approach to elucidate their function.

2 Data and Methods

2.1 Data

A total of 14,732 mouse protein sequences with their functional annotations were downloaded from the Mouse Functional Genome Database (MfunGD, http://mips.gsf.de/genre/proj/mfungd/) [40], which is an important repository of protein sequences that provides high-quality protein function annotations with respect to cellular function exclusively for mice. To extensively examine the model for independency of homology, we used the sequence cluster program CD-HIT [41] to remove or reduce similarities between sequences. We obtained 12,478 proteins with a similarity threshold of 0.7. The mouse proteins in the MfunGD are annotated using the Functional Catalogue (FunCat) annotation scheme, which is widely used for the analysis of protein networks [42]. Compared with the GO categories, the FunCat category structure is simpler and more hierarchical.

As shown in **Table 1**, there are a total of 24 functional categories. The balance between the specificity of the categories, human usability and requirements for subsequent bioinformatic applications is a general consideration in the design of an annotation scheme [42]. In line with this notion, the 24-category-scheme for protein function classification is not performed at the most specific level, but it keeps our system descriptive and compact, which complies with the

Table 1. The number of mouse proteins in each category in our dataset.

Functional Number	Functional categories	Number of proteins
1	METABOLISM	2,401
2	ENERGY	522
3	CELL CYCLE AND DNA PROCESSING	971
4	TRANSCRIPTION	1,921
5	PROTEIN SYNTHESIS	399
6	PROTEIN FATE (folding modification destination)	2,187
7	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	7,330
8	REGULATION OF METABOLISM AND PROTEIN FUNCTION	972
9	CELLULAR TRANSPORT TRANSPORT FACILITIES AND TRANSPORT ROUTES	2,078
10	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM	3,143
11	CELL RESCUE DEFENSE AND VIRULENCE	656
12	INTERACTION WITH THE ENVIRONMENT	1,212
13	SYSTEMIC INTERACTION WITH THE ENVIRONMENT	1,454
14	TRANSPOSABLE ELEMENTS VIRAL AND PLASMID PROTEINS	9
15	CELL FATE	1,180
16	DEVELOPMENT (Systemic)	939
17	BIOGENESIS OF CELLULAR COMPONENTS	769
18	CELL TYPE DIFFERENTIATION	317
19	TISSUE DIFFERENTIATION	313
20	ORGAN DIFFERENTIATION	491
21	SUBCELLULAR LOCALIZATION	8,467
22	CELL TYPE LOCALIZATION	232
23	TISSUE LOCALIZATION	261
24	ORGAN LOCALIZATION	542
Total	—	38,766

doi:10.1371/journal.pone.0166580.t001

main goal of our study. The fact that the functions outnumber the proteins indicates that some proteins perform multiple functions. For details, see **S1 Table**.

Protein-protein interaction pairs in mice were retrieved from STRING (Version 9.1, http://string-db.org/) [37], which is a protein-protein interaction database that collects known or predicted, direct (physical) or indirect (functional) associations. STRING quantifies each pair of protein interactions into a combined score. Currently, STRING contains 5,214,234 proteins from 1,133 organisms.

Because the manner in which the entries in the MfunGD are numbered differs from the method in STRING, comparison requires the mapping of associations between them. The mapping was performed using the BioMart database [43]. A total of 10,539 of the 12,478 proteins in MfunGD were mapped to the proteins in STRING.

2.2 Methods

The aim of this study is to predict the function of a given protein P based on n known-function proteins $P_1, P_2, \ldots, P_n$, assuming that the function categories are $f_1, f_2, \ldots, f_{24}$. One protein may belong to several function categories (*e.g.*, the protein mc10000007 belongs to categories f_8 'REGULATION OF METABOLISM AND PROTEIN FUNCTION' and f_{10} 'CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM'). Thus, we used a 24-dimensional vector $F_i = (d_{1,i}, d_{2,i}, \ldots, d_{24,i})$ to indicate the function categories of a protein

P_i, where $d_{j,i}$ is

$$d_{j,i} = \begin{cases} 1 & P_i \text{ has function } f_j \\ 0 & P_i \text{ does not have function } f_j \end{cases} \tag{1}$$

Three methods were used in this study to achieve this goal.

2.2.1 Sequence similarity-based approach

Proteins with similar sequences likely share the same or similar functions. Therefore, it is possible to predict protein functions based on sequence similarities. Herein, we used the PSI--BLAST program (E-value 0.001, iteration 3) to align the given unknown-function protein (P) against the known-function proteins ($P_1, P_2, \ldots, P_n$) in our dataset. The alignment score between P and P_i represents their similarity. This score is denoted as s_i. The predicted protein function scores of protein P are given by a 24-dimensional vector W and are calculated by

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{24} \end{bmatrix} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{24,1} & d_{24,2} & \cdots & d_{24,n} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} \tag{2}$$

where w_j denotes the score of a protein having function f_j. Elements in vector W are sorted from highest to lowest to obtain the predicted functions of protein P. A function receiving a high score is more likely to be an actual function of a given protein according to GBA [44] because there are several known proteins similar to the given protein that have this function. Thus, a function sequence can be constructed according to W. We provide an example to elaborate this point. If we obtain $w_{23} \geq w_2 \geq \ldots \geq w_5$, protein P is most likely to have function f_{23}, followed by function f_2 and so forth. The least likely function is f_5. For convenience, we call function f_{23} the 1st-order prediction, function f_2 the 2nd-order prediction and function f_5 the 24th-order prediction. This scheme to define the predicted results for multi-label classification problems has been used in previous studies [38, 45, 46].

2.2.2 Weighted interaction-based approach. Proteins in a cell interact with each other to perform particular functions. Following the GBA rule [25], interacting proteins may possess similar functions. We used the combined scores in the STRING database as weighted values between proteins. These values represent a fusion of eight types of evidence, including co-expression, gene fusion and experimental evidence. We assume that the combined score between P and P_i ($i = 1,2,\cdots,n$) is t_i. The predictive functional value is given by Y, a 24-dimensional vector computed by

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{24} \end{bmatrix} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{24,1} & d_{24,2} & \cdots & d_{24,n} \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} \tag{3}$$

where y_j denotes the score of a given protein with function f_j. Similar to the sequence similarity-based approach, each element y_j in vector Y is sorted from highest to lowest to obtain the function sequence of protein P. For example, if we obtain $y_{12} \geq y_{21} \geq \ldots \geq y_2$, protein P is most likely to have function f_{12}, followed by f_{21}, and the least likely function is f_2. In this study, we call function f_{12} the 1st-order prediction, function f_{21} the 2nd-order prediction, and function f_2 the 24th-order prediction.

2.2.3 PseAAC-based approach. Protein sequences can be characterized by pseudo amino acid composition, which was proposed by Chou to predict protein subcellular localization [47] and has become popular in the prediction of post-translational modification sites [48, 49] and membrane protein types [50–52]. PseAAC maps a protein sequence into a numerical vector. If a protein sequence is $X_1X_2\cdots X_N$, where X_i is an amino acid residue, then $L(X)$ is the property value of amino acid X in the physicochemical and biochemical respects. The normalized property value is computed by

$$F(X) = \frac{L(X) - \frac{1}{20}\sum_{Y\in\Phi} L(Y)}{\sqrt{\sum_{X\in\Phi}\left(L(X) - \frac{1}{20}\sum_{Y\in\Phi} L(Y)\right)^2 \Big/ 20}} \tag{4}$$

where Φ is the set of 20 types of amino acids. The correlation factor between residues in the protein sequence is computed by

$$C_i = \frac{1}{N-i}\sum_{k=1}^{N-i}((F(X_k) - F(X_{k+i}))^2,\ i = 1, 2, \cdots, \lambda,\ \lambda < N \tag{5}$$

The correlation factors reflect information about the position and category of amino acids in the protein sequence. The PseAAC of a protein sequence is computed by

$$v_i = \begin{cases} \dfrac{f_i}{\sum_{j=1}^{20} f_j + \varpi\sum_{k=1}^{\lambda} C_k} & 1 \leq i \leq 20 \\ \dfrac{\varpi C_{i-20}}{\sum_{j=1}^{20} f_j + \varpi\sum_{k=1}^{\lambda} C_k} & 21 \leq i \leq \lambda + 20 \end{cases} \tag{6}$$

where ϖ is the sequence order effects, and f_i is the occurrence frequency of amino acids. In this article, we set λ and ϖ to 50 and 0.15, respectively. Five physicochemical and biochemical properties of amino acids, i.e., codon diversity, electrostatic charge, molecular volume, polarity and secondary structure, are used to compute the PseAAC of protein sequences. These properties are retrieved from references [53–55], as listed in **Table 2**. For each category of property, we used the last 50 digits in the formula (6). In addition to the frequencies of 20 amino acids, we used a 270 (20+5*50)-dimensional vector to represent a protein sequence. The cosine distance between the query protein P and the known-function proteins P_i is given by

$$\Delta(P_i, P) = \frac{V_p \bullet V_{P_i}}{\|V_p\|\|V_{P_i}\|} \tag{7}$$

where the operators • and || || indicate the inner product and module of vectors, respectively, and V_P and V_{P_i} are the 270-dimensional PseAACs of proteins P and P_i, respectively. The

Table 2. The physicochemical and biochemical properties of the 20 amino acids.

Amino acid	Polarity	Second structure	Molecular volume	Codon diversity	Electrostatic charge
A	-0.591	-1.302	-0.733	1.57	-0.146
C	-1.343	0.465	-0.862	-1.02	-0.255
D	1.05	0.302	-3.656	-0.259	-3.242
E	1.357	-1.453	1.477	0.113	-0.837
F	-1.006	-0.59	1.891	-0.397	0.412
G	-0.384	1.652	1.33	1.045	2.064
H	0.336	-0.417	-1.673	-1.474	-0.078
I	-1.239	-0.547	2.131	0.393	0.816
K	1.831	-0.561	0.533	-0.277	1.648
L	-1.019	-0.987	-1.505	1.266	-0.912
M	-0.663	-1.524	2.219	-1.005	1.212
N	0.945	0.828	1.299	-0.169	0.933
P	0.189	2.081	-1.628	0.421	-1.392
Q	0.931	-0.179	-3.005	-0.503	-1.853
R	1.538	-0.055	1.502	0.44	2.897
S	-0.228	1.399	-4.76	0.67	-2.647
T	-0.032	0.326	2.213	0.908	1.313
V	-1.337	-0.279	-0.544	1.242	-1.262
W	-0.595	0.009	0.672	-2.128	-0.184
Y	0.26	0.83	3.097	-0.838	1.512

doi:10.1371/journal.pone.0166580.t002

predicted function value of the query protein was computed by

$$R = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{24} \end{bmatrix} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{24,1} & d_{24,2} & \cdots & d_{24,n} \end{bmatrix} \begin{bmatrix} \Delta(P_1, P) \\ \Delta(P_2, P) \\ \vdots \\ \Delta(P_n, P) \end{bmatrix} \tag{8}$$

Similar to the two above approaches, the elements in the vector R are sorted from high to low, such as $r_3 > r_{12} > \cdots > r_1$, where protein P is most likely to have function f_3, second most likely to have function f_{12}, and least likely to have function f_1.

2.3 Cross-Validation and Assessment

We used two cross-validation methods: leave-one-out cross-validation and ten-fold cross-validation to examine the performance of the presented methods. In the ten-fold cross-validation method, the original dataset are randomly and equally divided into ten parts. Samples in each part are singled out as testing samples, while samples in other nine parts are used as training samples. For the leave-one-out cross-validation approach, each sample in the original dataset is taken as a testing sample in turn and the remaining samples are used as training samples. To assess the experimental results, the prediction accuracy for the j^{th}-order prediction is given by

$$ACC_j = \frac{1}{n}\sum_{i=1}^{n} U_{j,i},\ j = 1, 2, \cdots, 24 \tag{9}$$

where $U_{j,i} = 1$ if the function category of the j^{th}-order prediction is actually the function category of protein P according to current knowledge. Otherwise, $U_{j,i} = 0$.

Table 3. Prediction accuracies of three methods and the combined method in the first three order predictions.

Method	Number of proteins of testing dataset	1st-order	2nd-order	3rd-order
Similarity-based	10,252	0.8756	0.7132	0.5158
Interaction-based	10,539	0.7535	0.6296	0.5299
PseAAC-based	12,478	0.6786	0.5874	0.2519
Combined	12,478	0.8464	0.6814	0.4996

doi:10.1371/journal.pone.0166580.t003

3 Results and Discussion

3.1 Performance of the Simple Approach

The performance of the three approaches for a dataset consisting of 12,478 proteins evaluated by the leave-one-out method is listed in Table 3. The similarity-based approach yielded the best accuracy of 0.8756 in the 1st-order prediction but could not predict functions of 2,226 proteins because they have no homologues with annotated proteins in the dataset. The interaction-based approach produced a lower prediction accuracy of 0.7535 than the similarity-based approach and could not predict the functions of 1,939 proteins that have no interactions with annotated proteins. The PseAAC-based nearest-neighbor approach performed worst in terms of the prediction accuracy, but it was able to predict the functions of all the test proteins. The results indicated that each approach has its strengths and limitations. Table 3 also shows that the 1st-order prediction performed best, followed by the 2nd-order prediction and the 3rd-order prediction, indicating the predicted function sequence for each test protein is quite reasonable.

The three approaches were compared on different testing datasets in the above paragraph. For a fair comparison, we generated a common dataset where each protein could be tested by the leave-one-out method. The common dataset consisted of 8,481 proteins. The accuracies of the three approaches versus order are plotted in Fig 1. The similarity-based approach performed best, followed by the interaction-based approach and the PseAAC-based nearest-neighbor approach. The similarity-based approach was much more accurate (by 0.11) than the interaction-based approach in the 1st-order prediction and more accurate (by 0.07) in the 2nd-order prediction, while the latter was much more accurate (by 0.09) than the PseAAC-based approach in the 1st-order prediction and more accurate (by 0.02) in the 2nd-order prediction. The results confirmed the advantage of the similarity-based approach over the other two approaches in terms of the prediction accuracy. As mentioned previously, the similarity-based approach cannot address non-homologous proteins, and the PseAAC-based approach can predict the functions off all proteins despite the lower prediction accuracy. Therefore, it is wise to jointly utilize the three methods to predict the protein functions.

3.2 Prediction by the Combined Approach

We combined the three approaches to predict the functions of proteins to make use of their respective advantages and disadvantages. For a given protein, we first employed the similarity-based approach. If the protein had no homologues, we applied the interaction-based approach. If the protein could not be predicted by the interaction-based approach, we used the PseAAC-based nearest-neighbor approach. The performance of the combined approach based on leave-one-out validation on the 12,478 proteins is shown in the fifth row of Table 3. The accuracy of the combined method was much higher than the interaction-based and PseAAC-based approaches and slightly lower than the similarity-based approach. However, the combined approach could predict all proteins, whereas the similarity-based approach could not.

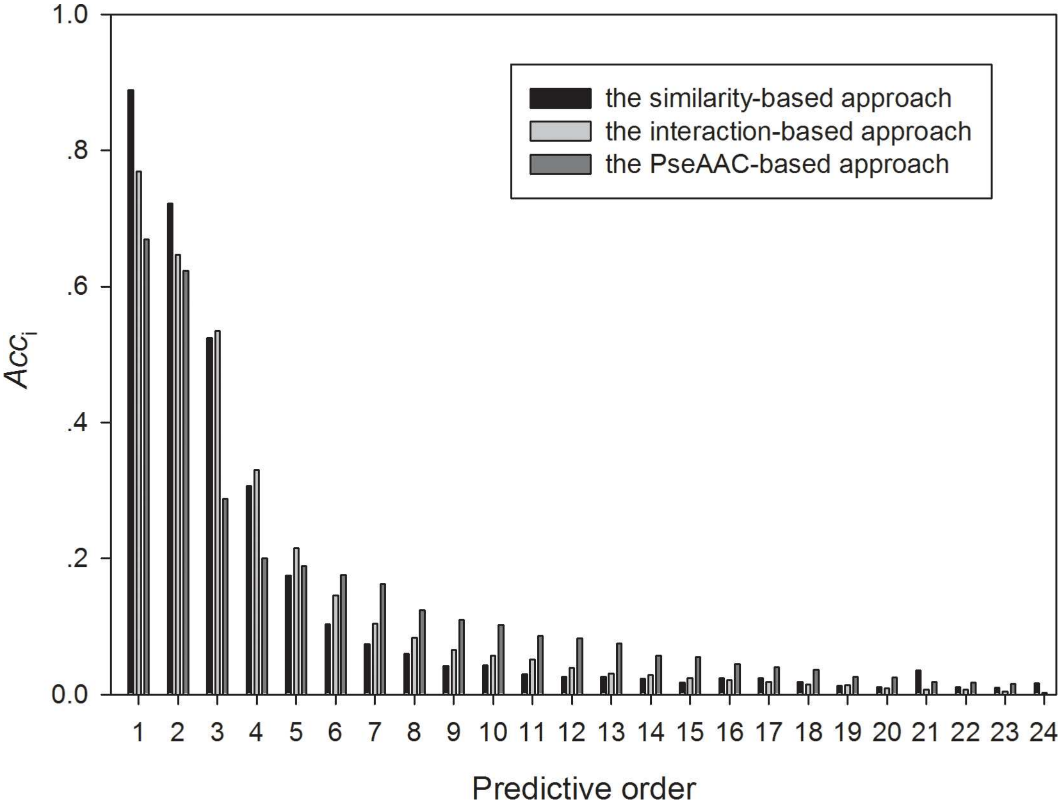

Fig 1. The prediction accuracies of 24 order predictions for these three methods on the common dataset.

doi:10.1371/journal.pone.0166580.g001

Therefore, the combined method has wide application at the cost of reduced prediction accuracy. For proteins with no homologues or interactions with annotated proteins, the best alternative is to use the combined approach. The contributions of the three approaches to the final predictive performance are shown in Table 4. The similarity-based approach contributed most, predicting more than 80% of all proteins and yielding an Acc_1 of 0.8756, followed by the interaction-based approach and the PseAAC-based approach.

To fully indicate the effectiveness of the combined method, we also used ten-fold cross-validation to examine this method. Because the predicted results yielded by this cross-validation method may influenced by the division of the dataset, the combined method was executed five times with different divisions. The prediction accuracies for the 1st-order, 2nd-order and 3rd-order predictions in each time are listed in Table 5. Compared to the prediction accuracies yielded by the leave-one-out cross-validation that are listed in Table 3, the performances of these two cross-validation methods are almost at the same level, which indicates that the combined method is still quite effective when there are no close homologs are available. Furthermore, it can be observed from Table 5 that the standard deviations for the 1st-order, 2nd-order and 3rd-order predictions are quite low, indicating the stability of the combined method.

3.3 Possible Protein Functions

In this study, the assessment of the predicted results was based on currently annotated proteins. Therefore, "right" and "wrong" predictions were relatively defined. For example, if the studied protein had function F_A and the predicted function was F_B, the prediction was not correct. It is conceivable that with the development of our knowledge, the protein could be found

Table 4. Contributions of the three approaches to the predicted results.

Method	Number of proteins	Proportion	Acc_1
Similarity-based approach	10,252	82.16%	0.8756
Interaction-based approach	1,876	15.03%	0.7154
PseAAC-based approach	350	2.81%	0.6943

doi:10.1371/journal.pone.0166580.t004

to possess F_B; thus, the prediction could be correct in the future. The currently annotated functions of the proteins are a subset of their actual functions. In this respect, some "wrong" predictions by our method in the current dataset may be correct. Next, we explore these wrong predictions.

It is worth performing further analysis on the wrongly predicted proteins. Because the 1st-order prediction is the most important, we investigated proteins with "wrong" 1st-order prediction but with "right" 2nd-order prediction. Because these proteins might possess the predicted 1st-order functions, we called them "false-wrong" 1st-order predicted proteins. As mentioned above, the combined method was evaluated by both the leave-one-out and ten-fold cross-validations. Because the predicted results yielded by the ten-fold cross-validations are not unique, we selected the predicted results yielded by the leave-one-out cross-validation to further analyze wrongly predicted proteins. In the leave-one-out test on the 12,478 proteins, we identified 966 such proteins: 658 proteins from the similarity-based approach, 258 proteins from the interaction-based approach and 50 proteins from the PseAAC-based approach. All these proteins are listed in S2 Table.

The goal of this process was to further validate our method. If we found evidence indicating that any of these proteins possessed the "wrong-predicted" functions, the actual prediction accuracy of our method would be much higher than presented above. This would allow the method to be applied to new protein function discoveries, but further experimental validations may be required for these proteins.

3.4 Possible Function Analysis of Significant "False-Wrong" 1st-Order Predicted Proteins

We explored the functions of proteins whose predicted 1st-order functions were wrong and whose predicted 2nd-order functions were correct. There were 966 such proteins. Forty protein genes were closely related to "false-wrong" 1st-order predicted functions, of which sixteen were predicted by the similarity-based approach, twenty-two were predicted by the interaction-based approach, and two were predicted by the PseAAC-based approach, as listed in Table 6, Table 7 and Table 8, respectively.

As shown in Table 6, sixteen significant proteins were predicted by the similarity-based approach. The proteins MYO1G, NEO1 and SDK1 were predicted to have the 1st-order function 'subcellular localization', suggesting that these gene products have specific cellular

Table 5. Performances of the combined method evaluated by ten-fold cross validation.

Order	1	2	3	4	5	Mean ± std [a]
1st	0.8429	0.8416	0.8420	0.8440	0.8419	0.8425 ±0.0010
2nd	0.6768	0.6792	0.6781	0.6787	0.6802	0.6786 ±0.0013
3rd	0.5023	0.4972	0.5007	0.4977	0.4998	0.4995 ±0.0021

a: std is the abbreviation of standard deviation.

doi:10.1371/journal.pone.0166580.t005

Table 6. The sixteen significant proteins with "wrong" 1st-order predictions but "right" 2nd-order predictions based on the sequence similarity-based approach.

Protein ID	Name	"wrong" predicted function in 1st-order prediction
mc11000118	MYO1G	SUBCELLULAR LOCALIZATION
mc9001073	NEO1	SUBCELLULAR LOCALIZATION
mc5002204	SDK1	SUBCELLULAR LOCALIZATION
mc17000153	PLG	PROTEIN FATE (folding, modification, destination)
mc2000415	GM711	PROTEIN FATE (folding, modification, destination)
mc15000840	MAPK15	PROTEIN FATE (folding, modification, destination)
mc7000273	PRKD2	PROTEIN FATE (folding, modification, destination)
mc11002342	STRADA	PROTEIN FATE (folding, modification, destination)
mc7001424	NTRK3	PROTEIN FATE (folding, modification, destination)
mc14000439	BMPR1A	PROTEIN FATE (folding, modification, destination)
mc11001586	KSR1	PROTEIN FATE (folding, modification, destination)
mc6000496	EPHB6	PROTEIN FATE (folding, modification, destination)
mc7000874	KLK9	PROTEIN FATE (folding, modification, destination)
mc15001663	KRT2	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
mc17001082	PTK7	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
mc1000962	SPEG	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)

doi:10.1371/journal.pone.0166580.t006

localizations. MYO1G has been reported as a hematopoietic-specific myosin that localizes to the plasma membrane [56]. Moreover, neogenin 1 (NEO1) and sidekick cell adhesion

Table 7. The twenty-two significant proteins with "wrong" 1st-order predictions but "right" 2nd-order predictions based on the weighted interaction-based approach.

Protein ID	Name	"wrong" predicted function in 1st-order prediction
mc2003319	ADRM1	SUBCELLULAR LOCALIZATION
mc6000275	ATP6V1F	SUBCELLULAR LOCALIZATION
mc4002507	AURKAIP1	SUBCELLULAR LOCALIZATION
mc17001119	BYSL	SUBCELLULAR LOCALIZATION
mc13001367	DHFR	SUBCELLULAR LOCALIZATION
mc1001293	DTYMK	SUBCELLULAR LOCALIZATION
mc4000473	GNE	SUBCELLULAR LOCALIZATION
mc5001787	HPD	SUBCELLULAR LOCALIZATION
mc3000151	HPS3	SUBCELLULAR LOCALIZATION
mc4001314	MAGOH	SUBCELLULAR LOCALIZATION
mc9000131	MED17	SUBCELLULAR LOCALIZATION
mc4001915	NUDC	SUBCELLULAR LOCALIZATION
mc11000229	PNOL	SUBCELLULAR LOCALIZATION
mcx000234	RGN	SUBCELLULAR LOCALIZATION
mc9000734	RPS25	SUBCELLULAR LOCALIZATION
mc8000054	SHCBP1	SUBCELLULAR LOCALIZATION
mc6000048	SHFM1	SUBCELLULAR LOCALIZATION
mc2002263	NCAPH	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
mc2000861	RIF1	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
mc19000070	CDCA5	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
mc7001471	PRC1	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
mc15001589	NPFF	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM

doi:10.1371/journal.pone.0166580.t007

Table 8. The two significant proteins with "wrong" 1st-order predictions but "right" 2nd-order predictions based on the PseAAC-based approach.

Protein ID	Name	"wrong" predicted function in 1st-order prediction
mc4000691	AKAP2	SUBCELLULAR LOCALIZATION
mc1001669	KISS1	SUBCELLULAR LOCALIZATION

doi:10.1371/journal.pone.0166580.t008

molecule 1 (SDK1) are likely to localize on the plasma membrane based on their biological functions. The proteins PLG, GM711, MAPK15, PRKD2, STRADA, NTRK3, BMPR1A, KSR1, EPHB6 and KLK9 were predicted to have the 1st-order function 'protein fate (folding/modification/destination)'. MAPKs, BMP, KSR1, PRKD2, STRADA, NTRK3 and EPHB6 are responsible for protein phosphorylation and signal transduction [57–63]. KLK9 belongs to the family of kallikrein-related peptidases (KLKs), which possess trypsin-like proteolytic activity [64, 65]. Plasminogen (PLG) is a precursor of the key enzyme of the fibrinolytic system plasmin, which serves as a physiological backup enzyme for ADAMTS13 (a disintegrin and metalloproteinase with a thrombospondin type I motif, member 13) in the degradation of pathological platelet-VWF (Von Willebrand factor) complexes [66]. KRT2, PTK7 and SPEG were predicted to have the 1st-order function 'protein with binding function or cofactor requirement'. Protein tyrosine kinase 7 (PTK7) was reported to interact with the Wnt family proteins [67] and play a pivotal role in planar cell polarity [68]. The intermediate filament keratin proteins, including Keratin 2 (KRT2), bind and interact with signaling molecules, such as CFTR [69], trichoplein [70] and Albatross complexes [71]. SPEG complex locus (SPEG) is a myotubularin (MTM1)-binding protein, and its deficiency has been proven to cause centronuclear myopathy with dilated cardiomyopathy [72].

As shown in **Table 7**, twenty-two significant proteins were predicted by the interaction-based approach. ADRM1, ATP6V1F, AURKAIP1, BYSL, DHFR, DTYMK, GNE, HPD, HPS3, MAGOH, MED17, NUDC, PNO1, RGN, RPS25, SHCBP1 and SHFM1 were predicted to have the 1st-order function 'subcellular localization'. ATPase, H^{+} transporting, lysosomal 14 kDa, V1 subunit F (ATP6V1F) and adhesion regulating molecule 1 (ADRM1) are likely to localize on the plasma membrane based on their biological functions. Several gene products are specifically localized in the nucleus, including AURKAIP1, DHFR, MAGOH, MED17, NUDC, PNO1 and RGN. Among them, NUDC is a nuclear movement protein that interacts with dynein [73]. Mediator complex subunit 17 (MED17) is localized in the nucleus and is involved in transcription regulation [74, 75]. The Bystin-like (BYSL) protein was reported to colocalize with trophinin, tastin and cytokeratins in the cytoplasm, forming a complex in trophectoderm cells that is essential for embryo implantation and ribosomal biogenesis [76]. The ribosomal protein S25 (RPS25) is also located in the cytoplasm and is responsible for protein synthesis [77]. The protein 4-hydroxyphenylpyruvate dioxygenase (HPD) is enriched in the liver cell cytoplasm and encodes an enzyme involved in the catabolic pathway of tyrosine, which catalyzes the conversion of 4-hydroxyphenylpyruvate to homogentisate [78]. SHFM1 (split hand/foot malformation (ectrodactyly) type 1, also known as DSS1) localizes to proteasomes [79]. Additionally, we predicted the specific subcellular localization of Hermansky-Pudlak syndrome 3 (HPS3), which encodes a novel protein with largely unknown function [80], together with aurora kinase A interacting protein 1 (AURKAIP1), deoxythymidylate kinase (DTYMK), glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase (GNE), partner of NOB1 homologue (PNO1), dihydrofolate reductase (DHFR), and SHC SH2-domain binding protein 1 (SHCBP1). Our data provide clues for the future study of these genes. NCAPH, RIF1, CDCA5 and PRC1 were predicted to have the 1st-order function 'protein with binding function or cofactor requirement'. NCAPH (also known as CAP-H) binds to the chromosome and regulates the cell cycle [81]. CDCA5 (also known as SORORIN) binds to sister chromatids and regulates their separation [82]. Protein regulator of cytokinesis 1 (PRC1) was shown to bind to several motor proteins, including KIF4, MKLP1 and CENP-E, and play pivotal roles in the formation of microtubule architecture [83]. Replication timing regulatory factor (RIF1) is responsible for regulating the replication-timing program in mammalian cells [84]. It was shown to bind to aberrant telomeres and to align along the anaphase midzone microtubules [85]. NPFF was predicted to have the 1st-order function 'cellular communication'. NPFF (neuropeptide FF) is an FMRFamide-like peptide with antiopiate properties that is involved in cellular communication as a part of the neurotransmitter system [86, 87].

As shown in **Table 8**, two significant proteins were predicted by the PseAAC-based approach. A-kinase anchor protein 2 (AKAP2) has the known function of 'protein fate (folding, modification, destination)' as it regulates cyclic AMP-dependent protein kinase (PKA) signaling in both a spatial and temporal manner. The specific subcellular localization of AKAP2 is closely related to its function [88]. AKAP2 has both cytosolic and endosomal localizations, and a fraction of endosomal AKAP2 is involved in regulating the expression of several downstream proteins, such as Rab4 and Rab11, and endosomal functions [89]. As another example, kisspeptins (KISS1) have known functions related to 'protein with binding function or cofactor requirement.' The versatile and complex pathways of KISS1 and their receptors play essential roles in the development of the brain and the reproductive system [90] and induce apoptosis in various cancers [91, 92]. Previous publications have shed light on both the cytosolic and nuclear localization of KISS1 receptors, which were linked to distinct functions, such as cytosolic calcium elevation and potential nuclear transactivation activity [93, 94]. These lines of evidence support our prediction of the important 'subcellular localization' function of these proteins.

4. Conclusion

The accurate identification of protein functions remains challenging in the post-genomic era. In this article, we employed protein sequence homology, weighted interactions and pseudo amino acid composition to explore protein functions. The experimental results indicate that homologous proteins are more likely to share functions than interacting proteins, which in turn share more functions than proteins with similar physicochemical and biochemical properties. Weighted interactions can be used to annotate the functions of proteins with no known homologues. The PseAAC-based approach was used for the functional annotation of proteins. These three approaches are complementary and represent an optimal combination for predicting protein functions. Further analyses of wrongly predicted functions will validate the effectiveness of the proposed method.

Supporting Information

S1 Table. The dataset used in this study. The first column of the file is the protein entry ID in MfunGD. The other columns are the functional categories to which the protein belongs. (CSV)

S2 Table. The "false-wrong" 1st-order predicted proteins. Proteins with "wrong" 1st-order function predictions but "right" 2nd-order function predictions in our dataset were called "false-wrong" 1st-order predicted proteins. There were 658 such proteins based on the similarity-based approach, 258 based on the interaction-based approach and 50 based on the PseAAC-based approach. These proteins are listed on three separate sheets. The proteins may possess the function indicated by the 1st-order prediction and are worthwhile subjects for

future analysis.
(XLSX)

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (61401302, 31371335, 81171342, 81201148), the Tianjin Research Program of the Application Foundation and Advanced Technology (14JCQNJC09500), the Innovation Program of the Shanghai Municipal Education Commission (12ZZ087), the National Research Foundation for the Doctoral Program of Higher Education of China (20130032120070, 20120032120073), the Scientific Research Fund of Hunan Provincial Education Department (15B216), the Science and Technology Program of Hunan (2015JC3099), and the Seed Foundation of Tianjin University (60302064, 60302069).

Author Contributions

Conceptualization: NZ YDC.

Data curation: GH CC TH.

Formal analysis: CC TH XK YZ.

Investigation: GH CC.

Methodology: GH YDC.

Resources: GH CC TH.

Supervision: YDC.

Validation: GH CC NZ.

Writing – original draft: GH CC.

Writing – review & editing: NZ YDC.

References

1. Erdin S, Lisewski AM, Lichtarge O. Protein function prediction: towards integration of similarity metrics. Current opinion in structural biology. 2011; 21(2):180–8. doi: 10.1016/j.sbi.2011.02.001 PMID: 21353529; PubMed Central PMCID: PMC3120633.
2. Cozzetto D, Buchan DW, Bryson K, Jones DT. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC Bioinformatics. 2013; 14 Suppl 3:S1. Epub 2013/03/27. doi: 10.1186/1471-2105-14-s3-s1 PMID: 23514099; PubMed Central PMCID: PMCPmc3584902.
3. Pandey G, Kumar V, Steinbach M. Computational Approaches for Protein Function: A Review. 2006.
4. Zehetner G. OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Research. 2003; 31(13):3799–803. doi: 10.1093/nar/gkg555 PMID: 12824422
5. Khan S, Situ G, Decker K, Schmidt CJ. GoFigure: automated Gene Ontology annotation. Bioinformatics. 2003; 19(18):2484–5. PMID: 14668239.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. Epub 1990/10/05. doi: 10.1016/s0022-2836(05)80360-2 PMID: 2231712.
7. Wass MN, Sternberg MJ. ConFunc—functional annotation in the twilight zone. Bioinformatics. 2008; 24(6):798–806. doi: 10.1093/bioinformatics/btn037 PMID: 18263643.
8. Hawkins T, Chitale M, Luban S, Kihara D. PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins. 2009; 74(3):566–82. doi: 10.1002/prot.22172 PMID: 18655063.

9. Chitale M, Hawkins T, Park C, Kihara D. ESG: extended similarity group method for automated protein function prediction. Bioinformatics. 2009; 25(14):1739–45. doi: 10.1093/bioinformatics/btp309 PMID: 19435743; PubMed Central PMCID: PMC2705228.

10. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–402. Epub 1997/09/01. PMID: 9254694; PubMed Central PMCID: PMCPmc146917.

11. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008; 36(10):3420–35. doi: 10.1093/nar/gkn176 PMID: 18445632; PubMed Central PMCID: PMC2425479.

12. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics. 2000; 16(6):566–7. PMID: 10980157.

13. Holm L, Kaariainen S, Rosenstrom P, Schenkel A. Searching protein structure databases with DaliLite v.3. Bioinformatics. 2008; 24(23):2780–1. doi: 10.1093/bioinformatics/btn507 PMID: 18818215; PubMed Central PMCID: PMC2639270.

14. Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. Nucleic Acids Res. 2010; 38(Web Server issue):W545–9. doi: 10.1093/nar/gkq366 PMID: 20457744; PubMed Central PMCID: PMC2896194.

15. Kolodny R, Linial N. Approximate protein structural alignment in polynomial time. Proc Natl Acad Sci U S A. 2004; 101(33):12201–6. doi: 10.1073/pnas.0404383101 PMID: 15304646; PubMed Central PMCID: PMC514457.

16. Shatsky M, Nussinov R, Wolfson HJ. A method for simultaneous alignment of multiple protein structures. Proteins. 2004; 56(1):143–56. doi: 10.1002/prot.10628 PMID: 15162494.

17. Shatsky M, Dror O, Schneidman-Duhovny D, Nussinov R, Wolfson HJ. BioInfo3D: a suite of tools for structural bioinformatics. Nucleic Acids Res. 2004; 32(Web Server issue):W503–7. doi: 10.1093/nar/gkh413 PMID: 15215437; PubMed Central PMCID: PMC441551.

18. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol. 2004; 340(2):385–95. doi: 10.1016/j.jmb.2004.04.058 PMID: 15201059.

19. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci U S A. 2004; 101(41):14754–9. doi: 10.1073/pnas.0404569101 PMID: 15456910; PubMed Central PMCID: PMC522026.

20. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, et al. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry. 2006; 45(8):2545–55. doi: 10.1021/bi052101l PMID: 16489747.

21. Saini HK, Fischer D. FRalanyzer: a tool for functional analysis of fold-recognition sequence-structure alignments. Nucleic Acids Res. 2007; 35(Web Server issue):W499–502. doi: 10.1093/nar/gkm367 PMID: 17537819; PubMed Central PMCID: PMC1933221.

22. An JY, Meng FR, You ZH, Chen X, Yan GY, Hu JP. Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model. Protein science: a publication of the Protein Society. 2016; 25(10):1825–33. doi: 10.1002/pro.2991 PMID: 27452983; PubMed Central PMCID: PMC5029537.

23. Huang YA, You ZH, Chen X, Chan K, Luo X. Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. BMC Bioinformatics. 2016; 17(1):184. doi: 10.1186/s12859-016-1035-4 PMID: 27112932; PubMed Central PMCID: PMC4845433.

24. Wong L, You Z-H, Ming Z, Li J, Chen X, Huang Y-A. Detection of Interactions between Proteins through Rotation Forest and Local Phase Quantization Descriptors. International journal of molecular sciences. 2016; 17(1):21. doi: 10.3390/ijms17010021 PMID: 26712745

25. Oliver S. Proteomics: Guilt-by-association goes global. Nature. 2000; 403(6770):601–3. doi: 10.1038/35001165 PMID: 10688178

26. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol. 2000; 18(12):1257–61. Epub 2000/12/02. doi: 10.1038/82360 PMID: 11101803.

27. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T. Assessment of prediction accuracy of protein function from protein—protein interaction data. Yeast. 2001; 18(6):523–31. Epub 2001/04/03. doi: 10.1002/yea.706 PMID: 11284008.

28. Chi X, Hou J. An iterative approach of protein function prediction. BMC Bioinformatics. 2011; 12:437. doi: 10.1186/1471-2105-12-437 PMID: 22074332; PubMed Central PMCID: PMC3224793.

29. Chua HN, Sung WK, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics. 2006; 22(13):1623–30. doi: 10.1093/bioinformatics/btl145 PMID: 16632496.

30. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. Journal of computational biology: a journal of computational molecular cell biology. 2003; 10 (6):947–60. Epub 2004/02/26. doi: 10.1089/106652703322756168 PMID: 14980019.

31. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics. 2003; 19 Suppl 1:i197–204. Epub 2003/07/12. PMID: 12855458.

32. Kourmpetis YA, van Dijk AD, Bink MC, van Ham RC, ter Braak CJ. Bayesian Markov Random Field analysis for protein function prediction based on network data. PLoS One. 2010; 5(2):e9293. Epub 2010/03/03. doi: 10.1371/journal.pone.0009293 PMID: 20195360; PubMed Central PMCID: PMC2827541.

33. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 2005; 21 Suppl 1:i302–10. Epub 2005/06/18. PMID: 15961472. doi: 10.1093/bioinformatics/bti1054

34. Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinformatics. 2006; 7:207. doi: 10.1186/1471-2105-7-207 PMID: 16613608; PubMed Central PMCID: PMC1473204.

35. Arnau V, Mars S, Marin I. Iterative cluster analysis of protein interaction data. Bioinformatics. 2005; 21 (3):364–78. doi: 10.1093/bioinformatics/bti021 PMID: 15374873.

36. Singh R, Xu J, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Natl Acad Sci U S A. 2008; 105(35):12763–8. doi: 10.1073/pnas.0806627105 PMID: 18725631; PubMed Central PMCID: PMC2522262.

37. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research. 2013; 41(Database issue):D808–15. Epub 2012/12/04. doi: 10.1093/nar/gks1094 PMID: 23203871; PubMed Central PMCID: PMC3531103.

38. Hu LL, Huang T, Shi X, Lu WC, Cai YD, Chou KC. Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. PLoS ONE. 2011; 6(1): e14556. doi: 10.1371/journal.pone.0014556 PMID: 21283518

39. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nature methods. 2013; 10(3):221–7. Epub 2013/01/29. doi: 10.1038/nmeth.2340 PMID: 23353650; PubMed Central PMCID: PMC3584181.

40. Ruepp A, Doudieu ON, van den Oever J, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. The Mouse Functional Genome Database (MfunGD): functional annotation of proteins in the light of their cellular context. Nucleic Acids Res. 2006; 34(Database issue):D568–71. doi: 10.1093/nar/gkj074 PMID: 16381934; PubMed Central PMCID: PMC1347437.

41. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–9. Epub 2006/05/30. doi: 10.1093/bioinformatics/btl158 PMID: 16731699.

42. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 2004; 32 (18):5539–45. doi: 10.1093/nar/gkh894 PMID: 15486203; PubMed Central PMCID: PMC524302.

43. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart—biological queries made easy. BMC Genomics. 2009; 10:22. doi: 10.1186/1471-2164-10-22 PMID: 19144180; PubMed Central PMCID: PMC2649164.

44. Oliver S. Guilt-by-association goes global. Nature. 2000; 403(6770):601–3. doi: 10.1038/35001165 PMID: ISI:000085288200029.

45. Chen L, Zeng WM, Cai YD, Feng KY, Chou KC. Predicting Anatomical Therapeutic Chemical (ATC) Classification of Drugs by Integrating Chemical-Chemical Interactions and Similarities. PLoS ONE. 2012; 7(4):e35254. doi: 10.1371/journal.pone.0035254 PMID: 22514724

46. Chen L, Lu J, Zhang N, Huang T, Cai Y-D. A hybrid method for prediction and repositioning of drug Anatomical Therapeutic Chemical classes. Molecular BioSystems. 2014; 10(4):868–77. doi: 10.1039/C3MB70490D PMID: 24492783

47. Chou K. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001; 43(3):246–55.

48. Xu Y, Ding J, Wu LY, Chou KC. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One. 2013; 8(2):e55844. doi: 10.1371/journal.pone.0055844 PMID: 23409062; PubMed Central PMCID: PMC3567014.

49. Qiu WR, Xiao X, Lin WZ, Chou KC. iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. Biomed Res Int. 2014; 2014:947416. doi: 10.1155/2014/947416 PMID: 24977164; PubMed Central PMCID: PMC4054830.

50. Jia C, Lin X, Wang Z. Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition. International journal of molecular sciences. 2014; 15(6):10410–23. doi: 10.3390/ijms150610410 PMID: 24918295; PubMed Central PMCID: PMC4100159.

51. Hayat M, Khan A. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J Theor Biol. 2010; 271(1):10–7. doi: 10.1016/j.jtbi.2010.11.017 PMID: 21110985.

52. Chen YK, Li KB. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. J Theor Biol. 2013; 318:1–12. doi: 10.1016/j.jtbi.2012.10.033 PMID: 23137835.

53. Atchley WR, Zhao J, Fernandes AD, Druke T. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A. 2005; 102(18):6395–400. Epub 2005/04/27. doi: 10.1073/pnas.0408677102 PMID: 15851683; PubMed Central PMCID: PMC1088356.

54. Rubinstein ND, Mayrose I, Pupko T. A machine-learning approach for predicting B-cell epitopes. Molecular immunology. 2009; 46(5):840–7. doi: 10.1016/j.molimm.2008.09.009 PMID: 18947876

55. Huang T, Shi X, Wang P, He Z, Feng K, Hu L, et al. Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS ONE. 2010; 5(6):e10972. doi: 10.1371/journal.pone.0010972 PMID: 20532046

56. Olety B, Wälte M, Honnert U, Schillers H, Bähler M. Myosin 1G (Myo1G) is a haematopoietic specific myosin that localises to the plasma membrane and regulates cell elasticity. FEBS Letters. 2010; 584 (3):493–9. doi: 10.1016/j.febslet.2009.11.096 PMID: 19968988

57. Therrien M, Chang HC, Solomon NM, Karim FD, Wassarman DA, Rubin GM. KSR, a novel protein kinase required for RAS signal transduction. Cell. 1995; 83(6):879–88. PMID: 8521512.

58. Morrison KB, Tognon CE, Garnett MJ, Deal C, Sorensen PH. ETV6-NTRK3 transformation requires insulin-like growth factor 1 receptor signaling and is associated with constitutive IRS-1 tyrosine phosphorylation. Oncogene. 2002; 21(37):5684–95. doi: 10.1038/sj.onc.1205669 PMID: 12173038.

59. Taylor EB, Ellingson WJ, Lamb JD, Chesser DG, Winder WW. Long-chain acyl-CoA esters inhibit phosphorylation of AMP-activated protein kinase at threonine-172 by LKB1/STRAD/MO25. American journal of physiology Endocrinology and metabolism. 2005; 288(6):E1055–61. doi: 10.1152/ajpendo.00516.2004 PMID: 15644453.

60. Cordenonsi M, Montagner M, Adorno M, Zacchigna L, Martello G, Mamidi A, et al. Integration of TGF-beta and Ras/MAPK signaling through p53 phosphorylation. Science. 2007; 315(5813):840–3. doi: 10.1126/science.1135961 PMID: 17234915.

61. Yuan J, Rozengurt E. PKD, PKD2, and p38 MAPK mediate Hsp27 serine-82 phosphorylation induced by neurotensin in pancreatic cancer PANC-1 cells. J Cell Biochem. 2008; 103(2):648–62. doi: 10.1002/jcb.21439 PMID: 17570131.

62. Yu J, Bulk E, Ji P, Hascher A, Koschmieder S, Berdel WE, et al. The kinase defective EPHB6 receptor tyrosine kinase activates MAP kinase signaling in lung adenocarcinoma. International journal of oncology. 2009; 35(1):175–9. PMID: 19513565.

63. Liu JA, Wu MH, Yan CH, Chau BK, So H, Ng A, et al. Phosphorylation of Sox9 is required for neural crest delamination and is regulated downstream of BMP and canonical Wnt signaling. Proc Natl Acad Sci U S A. 2013; 110(8):2882–7. doi: 10.1073/pnas.1211747110 PMID: 23382206; PubMed Central PMCID: PMC3581920.

64. Stefansson K, Brattsand M, Ny A, Glas B, Egelrud T. Kallikrein-related peptidase 14 may be a major contributor to trypsin-like proteolytic activity in human stratum corneum. Biol Chem. 2006; 387(6):761–8. Epub 2006/06/28. doi: 10.1515/BC.2006.095 PMID: 16800737.

65. Lizama AJ, Andrade Y, Colivoro P, Sarmiento J, Matus CE, Gonzalez CB, et al. Expression and bioregulation of the kallikrein-related peptidases family in the human neutrophil. Innate immunity. 2015; 21 (6):575–86. doi: 10.1177/1753425914566083 PMID: 25563717.

66. Tersteeg C, de Maat S, De Meyer SF, Smeets MW, Barendrecht AD, Roest M, et al. Plasmin cleavage of von Willebrand factor as an emergency bypass for ADAMTS13 deficiency in thrombotic microangiopathy. Circulation. 2014; 129(12):1320–31. doi: 10.1161/CIRCULATIONAHA.113.006727 PMID: 24449821.

67. Peradziryi H, Kaplan NA, Podleschny M, Liu X, Wehner P, Borchers A, et al. PTK7/Otk interacts with Wnts and inhibits canonical Wnt signalling. The EMBO journal. 2011; 30(18):3729–40. doi: 10.1038/emboj.2011.236 PMID: 21772251; PubMed Central PMCID: PMC3173783.

68. Lu X, Borchers AG, Jolicoeur C, Rayburn H, Baker JC, Tessier-Lavigne M. PTK7/CCK-4 is a novel regulator of planar cell polarity in vertebrates. Nature. 2004; 430(6995):93–8. doi: 10.1038/nature02677 PMID: 15229603.

69. Duan Y, Sun Y, Zhang F, Zhang WK, Wang D, Wang Y, et al. Keratin K18 increases cystic fibrosis transmembrane conductance regulator (CFTR) surface expression by binding to its C-terminal hydrophobic patch. J Biol Chem. 2012; 287(48):40547–59. doi: 10.1074/jbc.M112.403584 PMID: 23045527; PubMed Central PMCID: PMC3504769.

70. Nishizawa M, Izawa I, Inoko A, Hayashi Y, Nagata K, Yokoyama T, et al. Identification of trichoplein, a novel keratin filament-binding protein. Journal of cell science. 2005; 118(Pt 5):1081–90. doi: 10.1242/jcs.01667 PMID: 15731013.

71. Sugimoto M, Inoko A, Shiromizu T, Nakayama M, Zou P, Yonemura S, et al. The keratin-binding protein Albatross regulates polarization of epithelial cells. The Journal of cell biology. 2008; 183(1):19–28. doi: 10.1083/jcb.200803133 PMID: 18838552; PubMed Central PMCID: PMC2557036.

72. Agrawal PB, Pierson CR, Joshi M, Liu X, Ravenscroft G, Moghadaszadeh B, et al. SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy. American journal of human genetics. 2014; 95(2):218–26. doi: 10.1016/j.ajhg.2014.07.004 PMID: 25087613; PubMed Central PMCID: PMC4129406.

73. Aumais JP, Williams SN, Luo W, Nishino M, Caldwell KA, Caldwell GA, et al. Role for NudC, a dynein-associated nuclear movement protein, in mitosis and cytokinesis. Journal of cell science. 2003; 116 (10):1991–2003. doi: 10.1242/jcs.00412 PMID: 12679384

74. Liu Z, Myers LC. Med5(Nut1) and Med17(Srb4) are direct targets of mediator histone H4 tail interactions. PLoS One. 2012; 7(6):e38416. Epub 2012/06/14. doi: 10.1371/journal.pone.0038416 PMID: 22693636; PubMed Central PMCID: PMC3367926.

75. Kikuchi Y, Umemura H, Nishitani S, Iida S, Fukasawa R, Hayashi H, et al. Human mediator MED17 subunit plays essential roles in gene regulation by associating with the transcription and DNA repair machineries. Genes Cells. 2015; 20(3):191–202. doi: 10.1111/gtc.12210 PMID: 25482373.

76. Fukuda MN, Miyoshi M, Nadano D. The role of bystin in embryo implantation and in ribosomal biogenesis. Cell Mol Life Sci. 2008; 65(1):92–9. Epub 2007/10/06. doi: 10.1007/s00018-007-7302-9 PMID: 17917702; PubMed Central PMCID: PMC2771125.

77. Landry DM, Hertz MI, Thompson SR. RPS25 is essential for translation initiation by the Dicistroviridae and hepatitis C viral IRESs. Genes & development. 2009; 23(23):2753–64. Epub 2009/12/03. doi: 10.1101/gad.1832209 PMID: 19952110; PubMed Central PMCID: PMC2788332.

78. Awata H, Endo F, Matsuda I. Structure of the human 4-hydroxyphenylpyruvic acid dioxygenase gene (HPD). Genomics. 1994; 23(3):534–9. doi: 10.1006/geno.1994.1540 PMID: 7851880.

79. Gudmundsdottir K, Lord CJ, Ashworth A. The proteasome is involved in determining differential utilization of double-strand break repair pathways. Oncogene. 2007; 26(54):7601–6. Epub 2007/06/15. doi: 10.1038/sj.onc.1210579 PMID: 17563742.

80. Di Pietro SM, Falcon-Perez JM, Dell'Angelica EC. Characterization of BLOC-2, a complex containing the Hermansky-Pudlak syndrome proteins HPS3, HPS5 and HPS6. Traffic. 2004; 5(4):276–83. Epub 2004/03/20. doi: 10.1111/j.1600-0854.2004.0171.x PMID: 15030569.

81. Lai SK, Wong CH, Lee YP, Li HY. Caspase-3-mediated degradation of condensin Cap-H regulates mitotic cell death. Cell death and differentiation. 2011; 18(6):996–1004. Epub 2010/12/15. doi: 10.1038/cdd.2010.165 PMID: 21151026; PubMed Central PMCID: PMC3131938.

82. Diaz-Martinez LA, Gimenez-Abian JF, Clarke DJ. Regulation of centromeric cohesion by sororin independently of the APC/C. Cell cycle. 2007; 6(6):714–24. Epub 2007/03/16. 3935 [pii]. PMID: 17361102. doi: 10.4161/cc.6.6.3935

83. Kurasawa Y, Earnshaw WC, Mochizuki Y, Dohmae N, Todokoro K. Essential roles of KIF4 and its binding partner PRC1 in organized central spindle midzone formation. The EMBO journal. 2004; 23 (16):3237–48. Epub 2004/08/07. doi: 10.1038/sj.emboj.7600347 PMID: 15297875; PubMed Central PMCID: PMC514520.

84. Cornacchia D, Dileep V, Quivy JP, Foti R, Tili F, Santarella-Mellwig R, et al. Mouse Rif1 is a key regulator of the replication-timing programme in mammalian cells. The EMBO journal. 2012; 31(18):3678–90. Epub 2012/08/02. doi: 10.1038/emboj.2012.214 PMID: 22850673; PubMed Central PMCID: PMC3442270.

85. Xu L, Blackburn EH. Human Rif1 protein binds aberrant telomeres and aligns along anaphase midzone microtubules. The Journal of cell biology. 2004; 167(5):819–30. doi: 10.1083/jcb.200408181 PMID: 15583028; PubMed Central PMCID: PMC2172464.

86. Demichel P, Rodriguez JC, Roquebert J, Simonnet G. NPFF, a FMRF-NH2-like peptide, blocks opiate effects on ileum contractions. Peptides. 1993; 14(5):1005–9. Epub 1993/09/01. PMID: 8284250.

87. Mollereau C, Gouarderes C, Dumont Y, Kotani M, Detheux M, Doods H, et al. Agonist and antagonist activities on human NPFF(2) receptors of the NPY ligands GR231118 and BIBP3226. Br J Pharmacol. 2001; 133(1):1–4. doi: 10.1038/sj.bjp.0704049 PMID: 11325787; PubMed Central PMCID: PMC1572765.

88. Sarma GN, Kinderman FS, Kim C, von Daake S, Chen L, Wang BC, et al. Structure of D-AKAP2:PKA RI complex: insights into AKAP specificity and selectivity. Structure. 2010; 18(2):155–66. doi: 10.1016/j.str.2009.12.012 PMID: 20159461; PubMed Central PMCID: PMC3090270.

89. Eggers CT, Schafer JC, Goldenring JR, Taylor SS. D-AKAP2 interacts with Rab4 and Rab11 through its RGS domains and regulates transferrin receptor recycling. J Biol Chem. 2009; 284(47):32869–80. Epub 2009/10/03. doi: 10.1074/jbc.M109.022582 PMID: 19797056; PubMed Central PMCID: PMC2781703.

90. Li D, Yu W, Liu M. Regulation of KiSS1 gene expression. Peptides. 2009; 30(1):130–8. Epub 2008/11/11. doi: 10.1016/j.peptides.2008.09.025 PMID: 18996159.

91. Kostakis ID, Agrogiannis G, Vaiopoulos AG, Mylona E, Patsouris E, Kouraklis G, et al. KISS1 expression in colorectal cancer. APMIS: acta pathologica, microbiologica, et immunologica Scandinavica. 2013; 121(10):1004–10. doi: 10.1111/apm.12161 PMID: 24033850.

92. Wang H, Jones J, Turner T, He QP, Hardy S, Grizzle WE, et al. Clinical and biological significance of KISS1 expression in prostate cancer. Am J Pathol. 2012; 180(3):1170–8. doi: 10.1016/j.ajpath.2011.11.020 PMID: 22226740; PubMed Central PMCID: PMC3349884.

93. Kroll H, Bolsover S, Hsu J, Kim SH, Bouloux PM. Kisspeptin-evoked calcium signals in isolated primary rat gonadotropin- releasing hormone neurones. Neuroendocrinology. 2011; 93(2):114–20. doi: 10.1159/000321678 PMID: 21051881.

94. Onuma TA, Duan C. Duplicated Kiss1 receptor genes in zebrafish: distinct gene expression patterns, different ligand selectivity, and a novel nuclear isoform with transactivating activity. FASEB J. 2012; 26(7):2941–50. doi: 10.1096/fj.11-201095 PMID: 22499582.