QAUST: Protein Function Prediction Using Structure Similarity Search, Protein Interaction and Functional Sequence Motifs

Fatima Zohra Smaili[1,a,#], Shuye Tian[2,b,#], Ambrish Roy[3,c], Meshari Alazmi[1,4,d], Stefan T. Arold[5,e], Srayanta Mukherjee[3,f], P. Scott Hefty[6,g], Wei Chen[2,h,*], Xin Gao[1,i,*]

[1]*King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal 23955, Saudi Arabia*

[2]*Department of Biology, Southern University of Science and Technology of China (SUSTC), Shenzhen 518055, China*

[3]*Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA*

[4]*College of Computer Science and Engineering, University of Hail, Hail, Saudi Arabia*

[5]*King Abdullah University of Science and Technology (KAUST), Biological and Environmental Sciences and Engineering (BESE) Division, Thuwal 23955, Saudi Arabia.*

[6]*Department of Molecular Bioscience, University of Kansas, 2030 Becker Dr, Lawrence, KS 66047, USA*

[#]Equal contribution.

*Corresponding authors.

Email: chenw@sustech.edu.cn (Chen W), xin.gao@kaust.edu.sa (Gao X).

Running title: *Smaili FZ et al / Protein Function Prediction*

[a] ORCID: 0000-0001-6439-0659.

[b] ORCID: 0000-0002-1832-5765.

[c] ORCID: 0000-0003-1738-0591.

[d] ORCID: 0000-0001-9074-1029.

[e] ORCID: 0000-0001-5278-0668.

[f] ORCID: 0000-0002-2750-507.

[g] ORCID: 0000-0002-2303-2465.

[h] ORCID: 0000-0003-3263-1627.

[i] ORCID: 0000-0002-7108-3574.

This manuscript contains a total of 8739 words, 6 figures, 3 tables, 4 supplementary figures and 4 supplementary tables.

Abstract

The number of available protein sequences in public databases is increasing exponentially. However, a significant percentage of these sequences lack functional annotation, which is essential for the understanding of how biological systems operate.

We propose a novel method, QAUST, to infer protein functions, specifically Gene Ontology (GO) terms and Enzyme Commission (EC) numbers. Our method uses three sources of information: structure information encoded by global and local structure similarity search, biological network information inferred by protein-protein interaction data, and sequence information extracted from functionally discriminative sequence motifs. The three pieces of information are combined by consensus averaging to make the final prediction. Our approach has been tested on 500 protein targets from the CAFA benchmark set. The results show that our method provides accurate functional annotation and outperforms other prediction methods based on sequence similarity search or threading. We further demonstrate that a previously unknown function of TRIM22 protein predicted by QAUST can be experimentally validated. Availability: http://www.cbrc.kaust.edu.sa/qaust/submit/.

KEYWORDS: Protein function prediction; Gene Ontology (GO) terms; Enzyme Commission (EC) numbers; Protein structure similarity; Protein-protein interaction; Functionally discriminative motifs

Introduction

As of today, over 88 million protein sequences are available in the UniProtKB/TrEMB [1] database. However, this increase in the number of known protein sequences does not reflect a parallel increase in our biological knowledge, as less than 1% of these sequences have a manually annotated function [2]. On the other hand, the functional annotation of these sequences is not only an essential step for the understanding of physiological processes and biological systems in living entities, but also one of the highly challenging tasks in biology, which is why there is an increasing need to provide reliable, automated protein function annotation.

Significant efforts have been made to identify evolutionarily related proteins and automatically transfer functional annotations between homologous protein pairs [3−6]. To make such sequence similarity based functional transfer possible, powerful sequence-alignment methodologies have been developed. In particular, algorithms like BLAST/PSI-BLAST [3] and hidden Markov model (HMM) based techniques [4−6] have been frequently used to transfer functional annotations between homologous proteins. The underlying assumption of these sequence-based methods is that evolutionarily related proteins may inherit the function of a shared common ancestor. However, there are numerous cases in which proteins with high sequence similarity have distinct functions [7,8]. To partially address the problem, several methods have been developed to predict function using annotated conserved sequence motifs that are responsible for the functional aspect of the protein. These methods typically construct the sequence motifs from multiple sequence alignment of proteins belonging to the same protein family with known function [9−11]. They, however, have two major limitations. First, high-quality sequence alignment is typically required for motif construction, which is not trivial to obtain especially when the sequence homology is low. Second, the accuracy is limited by the quality of functional annotation of motifs. To overcome these limitations, we propose in this work to use a protein-specific “functionally discriminative motif” constructed from sequence fragments excised from the template sequence.

From another perspective, the 3D structure of a protein sequence is believed to be more involved in its biological function [12,13] since structures are more conserved than the sequences. The 3D structure of a protein can therefore provide additional information

for function transfer, especially when the sequence similarity between related proteins is too low for sequence homolog detection [14,15]. However, the relationship between the protein function and its structure is not straightforward, as in some cases, similar structures perform the same function while in many cases similar folds perform different functions [16,17]. Therefore, many prediction methods have been relying on local structure similarity search methods rather than global similarity search to identify functionally homologous proteins [18−20]. Most of these approaches scan the query protein against a library of known conserved spatial motifs or known active sites (*e.g.*, binding sites) with known function [21]. Local similarity search methods have been proven to be quite accurate in detecting functional similarity between proteins of different folds, but they also have a high probability of producing false positive matches [22]. One possible solution is to combine global and local structure alignment to overcome the promiscuity of global structure comparison and low specificity of local structure matching [23,24], which we implement in this project.

A number of function prediction methods are based on the information extracted from protein-protein interaction (PPI) networks [25,26]. The assumption in this case is that proteins that physically interact with each other frequently appear at the same sub-cellular location and are part of the same biological process [27]. However, it is not always the case that proteins which interact with each other share the same molecular function (*e.g.*, PD1 and PD-L1), which is why PPI information is not always sufficient to predict very specific functions [28].

Finally, recently there is an emergence of methods which combine multiple sources of information (PPI, domains, sequence alignments, *etc*.) using advanced machine learning algorithms to perform function prediction. These methods have shown to improve the prediction performance over methods that use only one type of information [29−37].

In this work, we propose a new protein function prediction method, Quantitative Annotation of Unknown STructure (QAUST), which combines the global and local structure similarity search with protein-protein interaction networks and functional sequence motif detection. Our approach follows a sequence-to-structure-to-function workflow. Starting from the protein amino acid sequence, we first generate structure

predictions by the Iterative Threading ASSembly Refinement method (I-TASSER) [38]. The predicted structure is then used to identify the proteins with similar functions based on a combination of global and local structure similarity search method that follows the same pipeline used in COFACTOR [24,39]. Protein-protein interaction information is meanwhile extracted from the STRING database [40]. And finally, we extract functionally discriminative sequence motifs as our third main prediction feature. The confidence scores obtained from these three features are combined in a consensus function to obtain our final confidence score.

Since the terminology of a "protein function" might be ambiguous, we would like to clarify that the definitions of function followed in this work is Enzyme Commission (EC) numbers [41] and Gene Ontology (GO) terms [42]. EC numbers are used to categorize enzymes into hierarchical families using a numerical classification. Specifically, the EC number (which is composed of four numbers separated by periods *i.e.*, A.B.C.D) refers to the reaction catalyzed by a specific enzyme. On the other hand, the GO terms are a set of controlled vocabulary to formally describe proteins and RNAs based upon their functions. Three aspects of ontologies, Biological Process (BP), Cellular Component (CC) and Molecular Function (MF), are defined in this database. Each one of these three GO aspects is represented by a structured directed acyclic graph (DAG), where nodes represent GO terms which describe gene product functions, while the edges represent the relationships ("is_a" or "part_of") between the GO terms. In GO's functional hierarchy, the more general functions are on the top of the graph while more specific terms are usually present further down the graph.

Our prediction results are compared to the following programs: COFACTOR [39,43], a global and local structure similarity-based method, LOMETS [44], a meta-threading algorithm, HHsearch [5], an HMM-based method that is widely used to detect protein homologs, BLAST [3], which transfers annotations based on sequence similarity, naïve baseline which predicts GO terms based on their annotation frequency, as well as two highly-ranked methods from the CAFA assessment [45], GoFDR [32], and INGA [46].

Methods

Dataset

To evaluate QAUST for EC prediction, we used the benchmark data set of COFACTOR [39,43] as our testing data set. This data set consists of 318 enzymes with unique EC numbers (first three digits) covering all 6 enzyme classes. Similarly, all sequences in our template libraries with a sequence identity > 30% with the query enzymes are excluded from the template libraries.

We evaluate QAUST for GO prediction on a data set of 500 randomly chosen non-redundant proteins from the CAFA 2 targets (https://biofunctionprediction.org/cafa/) annotated with at least one GO term. To eliminate any structure or function homologs to the query, templates having a sequence identity > 30% with the query proteins are excluded from the template libraries both in the I-TASSER threading library and our function prediction template libraries.

Enzyme Commission (EC) number prediction

Global and local similarity search

The first step of our protein function prediction is the generation of the predicted 3D model of the query protein using I-TASSER [38] as outlined in Section S1. The predicted model of the query protein obtained from I-TASSER is then scanned against a non-redundant (pairwise sequence identity no more than 90%) structure template library of 2,385 enzymes with at least the first three digits of EC number annotated by the Catalytic Site Atlas (CSA) database [47]. This library scanning detects homologous structure templates to the query proteins using two types of structure similarity search programs: *global similarity* search and *local similarity* search.

Global similarity search

Templates with a similar global structure to the predicted structure of the query protein are detected from the template library using TM-align [48]. Another important consideration when searching for templates with similar global folds to the query protein is the quality of the structural models. Appraising the accuracy of the structure modeling in the scoring scheme helps to reduce the number of false positive predictions. In this

particular case, the quality of the predicted I-TASSER model generated in the previous step is evaluated using *Cscore* [38].

Local similarity search

The local structural search approach consists of three steps (**Figure 1**). The first of which is the structural match of the specific catalytic/active residue pairs. For a given pair of query and template proteins, we first scan the known catalytic/active residues of the template through the query sequence. The query's residues whose amino acid types are the same as the amino acid types of the template's catalytic/active residues are marked as potential active sites in the query. The structures of all combined sets of marked residues in the query are extracted from the predicted model and used as candidate active sites. The structure of the candidate site is superimposed on the known catalytic/active residues in the template. To make the structure superimposition more reliable, for each residue *i*, the coordinates of C_α atoms and side-chain centers of mass of the two neighboring residues, *i.e.*, the *i*-1 and *i*+1th residues are also included in the superimposition.

The second step is to identify the key local environment residues around the active sites in the query and the template. For this purpose, we superimpose the complete structure of the query and template proteins based on the rotation matrix obtained from the superimposition of the candidate catalytic/active residue structures obtained in previous step. A sphere of radius *r* is then defined around the geometric center of the template's local 3D fragments, where *r* is the maximum distance of the template residues in the local 3D fragment from the geometric center. The sphere represents a local environment or probable active site region, under which the query and template's chemical and structural similarity are compared. Because a sphere comprising of a very small number of catalytic/active residues can easily generate false positive hits, when the template's active site region is small, we set the number of residues inside the sphere to be a minimum of 20 residues. This value is obtained using minimum grid search parameter optimization by evaluating different sphere sizes in the range of [10,50] residues to select the most accurate value.

In the third step, the best alignment of the local active site residues in the spheres between the query and the template is identified using a scoring function similar to TM-

align. Starting from the initial superposition of the query and template protein structures, we perform a Needleman-Wunsch dynamic programming to generate the best alignment for the residues in the selected sphere of the template and the query, where the alignment score matrix S_{ij} for aligning the *i*th residue in the query and the *j*th residue in the template is defined as:

$$S_{ij} = \left[\frac{1}{1 + \frac{d_{ij}}{d_0}} + M_{ij} \right], \qquad (1)$$

where d_{ij} is the Cα distance between residues i and j, d_0 is the distance cutoff given by $d_0 = 1.24\sqrt[3]{L-15} - 1.8$ obtained from TM-align, M_{ij} is the substitution score between the *i*th and *j*th residues taken from the BLOSUM62 mutation matrix with the value normalized by the diagonal element in the mutation matrix. The gap penalty is set as -1. For a given scoring matrix S_{ij}, a new alignment is generated by dynamic programming. A new superposition and scoring matrix are then constructed based on the new alignment to obtain a newer alignment from dynamic programming. This procedure is iteratively repeated until the final alignment is converged. For each alignment, the active site match (AcM) is evaluated using an alignment score defined as:

$$\mathrm{AcM} = \frac{1}{N_t} \sum_{i=1}^{N_{ali}} \frac{1}{1 + \left(\frac{d_{ii}}{d_0}\right)^2} + \frac{1}{N_t} \sum_{i=1}^{N_{ali}} M_{ii}, \qquad (2)$$

where N_t represents the number of residues in the active site sphere of the template, N_{ali} is the number of aligned residue pairs. The maximum AcM score obtained during the heuristic iterations is recorded for each candidate active site. Finally, the set of residues in the candidate active site which has the highest AcM score is selected to evaluate the similarity between the query and the template's active site. The weights and form the AcM score have been derived based on the predicted structures of 100 randomly chosen training proteins from the template library, which are non-homologous (sequence similarity < 30%) to the test proteins in order to maximize the sensitivity and specificity of the predictions.

Scoring function for global and local similarity search

The final score for predicting EC numbers, used to sort the hits from the enzyme library is a combination of the global similarity search score and the AcM score (obtained from the local similarity search) and is defined as:

$$QAUSTEC = C_{norm} \cdot \left[TM + \frac{Cov}{1 + RMSD_{ali}}\right] + 2 \cdot ID_{ali} \cdot Cov + \frac{AcM}{2}. \quad (3)$$

where *Cov* represents the coverage of the structural alignment, $RMSD_{ali}$ is the RMSD (root-mean-square deviation) between the model and the template structure in the structurally aligned region, and ID_{ali} is the sequence identity between query and template based on the alignment generated by TM-align. The hyperbolic-tangent-like normalization is further used to normalize the raw EC score to be between 0 and 1:

$$QAUSTEC_{norm} = \frac{2}{1 + \exp(-QAUSTEC)} - 1. \quad (4)$$

Gene Ontology (GO) term prediction

For GO term prediction, we combine three different predictors. Each one of these predictors generates a confidence score. The three confidence scores obtained are then combined in a consensus function to generate the final prediction score. We use three predictors: first, the global structure similarity, which uses I-TASSER to predict the 3D structure of the query, and then scans a library of templates to identify those which have a similar global structure to the predicted model. The second predictor is based on PPI information, and the third one is based on extracted functional sequence motifs.

Global protein structure similarity

Similar to EC prediction, I-TASSER is also used here to construct the corresponding 3D model to the query sequence. The model obtained is then scanned against a library of templates to identify those which share a similar global structure to the query model (https://zhanglab.ccmb.med.umich.edu/BioLiP/library.html). For the time being, the functionally important residues for most of the proteins in the GO template library are unknown. Therefore, only the global similarity search is taken into consideration when sorting the hits from the GO library. Global similarity search for GO prediction is done in

a similar way to global similarity search for EC prediction described in the previous section. The only difference is that to select the best hits for GO prediction, we rank a template using the Fh-score defined as:

$$Fh_{score} = C_{norm} \times \left(TM_{score} + \frac{1}{1 + RMSD_{ali}} \times Cov\right) + 3 \times ID_{ali} \times Cov. \quad (5)$$

Since each single protein can be annotated with multiple GO terms and the global search may result in many close template structures, a query protein can have multiple GO term predictions with high Fh-scores. Therefore, the confidence score of each GO term is calculated as follow:

$$P_{structure}(\lambda) = \frac{1}{N} \sum_{i=1}^{N_\lambda} Fh\,(i), \quad (6)$$

where λ represents a given GO term, N_λ is the number of templates annotated with the GO term λ, and N is the total number of templates selected for generating the consensus. When multiple close templates are available, we only consider the templates with an Fh-score >1. For those query proteins with less than 10 templates of Fh-score >1, the top 10 templates are selected for generating the consensus prediction regardless of the Fh-score. Also, given the hierarchical nature of the GO DAG, we consider that when a protein is annotated with a given GO term, all its ancestor GO terms (through "is_a" relation) are automatically implied. Therefore, once a GO term λ is scored, we score all its ancestor terms as well. The score of any ancestor GO term μ of term λ is calculated as:

$$P_{structure}\,(\mu) = P_{structure}\,(\lambda) \times \left(1 + \frac{N_\mu}{N_0}\right), \quad (7)$$

where N_μ and N_0 are the number of leaf nodes under node μ and the root node, respectively. Since COFACTOR [39,43] uses a similar structure scoring function, we have highlighted the main differences between QAUST and COFACTOR in Section S2, and compared our method with COFACTOR in the experiments.

Protein-protein interaction network

We exploit the information provided by the STRING [40] database, which is a library of PPI networks, to extend our prediction set. The query protein sequence is mapped to its corresponding STRING entry by BLAST, with minimum sequence identity cutoff of 90%. Extracting the PPI partners of the query, we calculate the confidence score of STRING for a GO term λ ($P_{STRING}\ (\lambda)$) as the frequency of the GO term λ among the experimentally annotated interaction partners of the query protein:

$$P_{STRING}(\lambda) = \frac{n_\lambda}{N}, \qquad (8)$$

where n_λ is the number of interaction partners annotated with the GO term λ and N is the number of partners associated with term λ, according to the corresponding UniProt-GOA (http://www.ebi.ac.uk/GOA) entry of this PPI partner. This score could take any value from 0 to 1.

Functionally discriminative sequence motifs

In addition to the structure similarity search and protein-protein interaction features discussed above, we also include sequentially extracted features to predict GO terms since a sequence is a highly valuable source of information that can especially be useful when dealing with proteins for which we cannot construct a good quality 3D structure model or those with no known protein-protein interaction information.

Our functionally discriminative motif detection algorithm follows three steps: detection of sequence templates, identification of functionally discriminative motifs given a GO term, and scoring the query protein.

Detection of sequence templates for query protein: The sequence homologs of the query sequence are detected by PSI-BLAST [3] from the Uniref90 database [49]. We filter all obtained homologs with sequence identity > 30% to the query.

Identification of functionally discriminative motifs given a GO term: We map all the selected sequence homologs of the query to their corresponding GO annotations in the UniProt-GOA database (http://www.ebi.ac.uk/GOA). GO terms assigned with “Inferred from Electronic Annotation” (IEA) or “No biological Data available” (ND) evidence codes are not considered. We also filter out annotations with evidence code IPI (Inferred from Physical Interactions) since we use protein-protein interaction (PPI) information in our features. After filtering these annotations, we are left with the annotations based on

evidence codes: EXP, IDA, IMP, IGI, IEP, TAS, and IC. For each GO term λ, we build two sets of sequences from the set of homolog sequences detected in the previous step. These two sets are: the “annotated set”, which is the set of sequence homologs annotated with this specific GO term, and the “not-annotated set”, which is the set of sequence homologs not annotated with this given GO term. For each one of these two sets, we extract the ten most frequent motifs by extracting all unique amino acid motifs of length [4,7] from the sequence set using sliding windows. These motifs are ranked in descending order by their occurrences. The top 10 most frequent motifs are the initial "frequent list", while the remaining motifs are in the "waiting list". If, within the "frequent list", a short motif is a substring of another longer motif, the shorter motif is discarded, and the most frequent motif from the "waiting list" is transferred to "frequent list" to ensure that the latter always has 10 motifs. This process is iterated until, in the "frequent list", any motif is not a substring of another motif. The motifs in the "frequent list" are used for matching the query in the next step.

Scoring the query protein: For each of the two sets (annotated and not-annotated sets) we check the number of frequent motifs extracted in the previous step that are also present in the query sequence. Then, we calculate the confidence score of the GO term given the query sequence as follows:

$$P_{MOTIF}(\lambda) = \frac{n_q(\lambda)}{N(\lambda)}\left[1 - \frac{n_q(\lambda^c)}{N(\lambda^c)}\right], \qquad (9)$$

where λ is the given GO term, $N(\lambda)$ and $N(\lambda^c)$ are the number of frequent patterns from the “annotated set” and “not-annotated set”, both of which equal to 10. $n(\lambda)$ and $n(\lambda^c)$ are the corresponding number of matched patterns at the query sequence. This score can take any value from 0 to 1. An ideal value of this score would be equal to 1, which happens when all the sequences in the annotated set contain these frequent motifs and none of the sequences in the not-annotated set contains these same motifs. This scoring function has been designed to penalize the prediction in case the query sequence matches a high number of frequent motifs from the not-annotated set. This way, the scoring function accounts for two essential pieces of information: which set has the maximum number of frequent motifs matched in the query, and how significant is the difference between the number of matched motifs from the annotated set to that from the not-

annotated set. **Figure 2** shows a flowchart detailing the three steps of extracting functional sequence motifs.

Consensus

To predict GO terms, the three main scores obtained from the three different predictors (the structure search, the PPI network, and the functional motifs) are combined by consensus averaging to calculate the final confidence score $P_{consensus}(\lambda)$ for a GO term λ:

$$P_{consensus}(\lambda) = 1 - \prod_{m \in \{structure, STRING, MOTIF\}} \left(1 - P_m(\lambda)\right) \qquad (10)$$

This equation used to calculate the consensus has been previously used by other methods for protein function prediction [46]. If one or more predictors are not available for a given term (*e.g.*, no interaction partners are known for the given query), only the available predictors are used to obtain the confidence score. Also, since GO uses the true-path rule (*i.e.*, if a protein is associated by a term, it is also implicitly annotated by its ancestors), for every predicted GO term, all its ancestors are considered to be predicted as well since they are more general terms.

Results

Prediction of EC numbers

We compared the EC prediction performance of our method to five methods: HHsearch [5], LOMETS [44] , BLAST [3], COFACTOR [39,43] and DEEPre webserver [50]. We compared the performance of these methods based on precision (positive predictive value) and recall (sensitivity) rates. **Figure 3** shows the precision-recall graph corresponding to four baseline methods as well as QAUST. Since the DEEPre webserver does not report the confidence score with the annotation, we could not draw the precision-recall curves but compared QAUST to DEEPre based on accuracy. An EC number prediction is considered to be “true” if the first three digits of the EC number from the hit are identical to those of the query protein; otherwise the hit is considered to be “false”. As shown in Figure 3, the rate of true positive predictions using the EC-score is much higher than that of HHsearch, LOMETS, BLAST and COFACTOR at most recall rates. QAUST has also

an area under precision-recall curve (AUPRC) of 0.712 which is higher than that of COFACTOR (0.643), LOMETS (0.510), and HHsearch (0.489). **Table 1** reports the accuracy of QAUST compared to five other methods including DEEPre and COFACTOR. The results show that DEEPre has a slightly higher performance than QAUST in terms of accuracy, which is probably due to the fact that DEEPre is a machine learning method trained on a large number of enzymes with known functions that overlap or contain close homologs to our test data.

Prediction of GO terms

To assess the contribution of individual predictors to the GO prediction performance by QAUST, we visualize the precision-recall curve of the structure similarity search alone ($P_{structure}$), the precision-recall curve of structure similarity search combined with PPI information ($P_{structure}$ and P_{STRING}), and that of the final QAUST prediction ($P_{structure}$, P_{STRING} and P_{MOTIF}). Additionally, we compared the prediction performance on our dataset (please see subsection Dataset under section Methods) to COFACTOR [39,43], BLAST [3], LOMETS [44], HHsearch [5], INGA [46] webserver, a method that combines BLAST, PPI information and Pfam in one predictor, and GoFDR [32], one of the top function prediction methods at the CAFA assessment [45] which uses a machine learning model as classifier and discriminative residues as the main feature.

The performance has been primarily evaluated using precision-recall curves computed at each prediction score threshold. We also used the F_{max} measure as a quantitative measure to evaluate the overall performance of the precision-recall curves. Precision, recall, and F_{max} are defined in the same way as the CAFA evaluation [51]. The F_{max} measure has been computed as the maximum value of the $F_{measure}$ which is computed at each threshold as $\frac{2\times precision\times recall}{precision+recall}$.

Precision at threshold $\boldsymbol{t}$ is defined as $\frac{|\mathbf{P_x}(\mathbf{t})\cap\mathbf{C_x})|}{|\mathbf{P_x}(\mathbf{t})|}$, while recall is defined as $\frac{|\mathbf{P_x}(\mathbf{t})\cap\mathbf{C_x})|}{|\mathbf{C_x}|}$ where $\mathbf{x}$ is a query protein, $\mathbf{P_x(t)}$ is the set of predicted terms for $\mathbf{x}$ at threshold $\boldsymbol{t}$ and $\mathbf{C_x}$is the set of correct terms that $\mathbf{x}$ is experimentally annotated with.

Similar to the CAFA evaluation [34], we also reported the minimum semantic distance (S_{min}) as an additional evaluation metric for GO prediction. S_{min} is defined as

$min_t\{\sqrt{ru(\boldsymbol{t})^2 + mi(\boldsymbol{t})^2}\}$, where *ru*(***t***) is the remaining uncertainty at threshold ***t*** defined as $\frac{1}{n_e}\sum_{i=1}^{n_e}\sum_f ic(f) \mid f \notin P_i(t) \wedge f \in C_i$ and *mi*(***t***) is the misinformation at threshold ***t*** defined as $\frac{1}{n_e}\sum_{i=1}^{n_e}\sum_f ic(f) \mid f \in P_i(t) \wedge f \notin C_i$, where n_e is the number of proteins in our dataset, $\boldsymbol{P_i(t)}$ is the set of predicted GO term for protein ***i*** at threshold ***t***, $\boldsymbol{C_i}$ is the set of terms that protein ***i*** is actually annotated with and ***ic(f)*** is the information content of the GO term ***f***. The very general and unspecific GO terms such as "Molecular Function", "Biological Process", "Cellular Component", "Binding" and "Protein Binding" are excluded from the evaluation.

As shown in **Figure 4**, our method combining structure, PPI and functional motif information achieves higher precision than most other methods at most recall points, in particular for MF and BP. For our dataset, structure and motif information has been used for all proteins. However, the PPI information from STRING was missing for 74 proteins. In this case, only the structure and motif information is used. The F_{max} measure values are reported in **Table 2** while the minimum semantic distance (S_{min}) values are reported in **Table 3**. Surprisingly, in CC prediction, naïve baseline which predicts GO terms based on their annotation frequency, achieves higher performance than all other methods including QAUST. In fact, in the CAFA assessment [45], naïve baseline also outperformed most of the other methods in predicting CC terms. One possible explanation for why naïve baseline has higher F_{max} for CC terms prediction is because the most frequently used CC terms in protein annotation are usually part of a small set of very general terms such as "cytoplasm" or "intracellular part". Since the naïve baseline is solely based on frequency, it increases the chance of predicting a true positive [45]. We also reported the p-values obtained from the Mann-Whitney U test to assess the significance of the difference in performance of QAUST compared to all other methods in Section S3 and Table S1.

As a further analysis, to investigate if the performance of our method is solely due to the power of the I-TASSER structure prediction we used, we replaced the I-TASSER structure prediction component of our method by HHsearch and LOMETS structure prediction, respectively. Our results show that no matter which structure prediction method is used, our scoring function, $P_{structure,}$ can significantly improve the performance

on predicting the GO terms. Meanwhile, among the three structure prediction methods, I-TASSER with $P_{Structure}$ consistently performs the best over all three GO hierarchy branches of MF, BP and CC, whereas LOMETS with $P_{Structure}$ has the second best performance on MF and CC, and HHsearch with $P_{Structure}$ is the second best on predicting BP terms (Section S4 and Figure S1). Additionally, we have evaluated the performance of our method when only PPI and motif information are used without including any structure-based information. The results show that the function prediction performance drops when structure features are not used (Section S5 and Figure S2).

How do protein-protein interaction information and functional sequence motifs improve the prediction?

Protein-protein interaction information extracted from STRING is an important feature used in our prediction. In Figure 4, we show how protein-protein interaction information alone improves the performance achieved by the structure similarity search (orange dash lines versus magenta dash lines). The precision-recall curves in Figure 4 show that the contribution of protein-protein information from STRING is very significant for CC and BP terms, especially for large recall rates. Moreover, the precision-recall curves confirm our initial hypothesis on the utility of PPI information for function annotation. As shown in the figure, while there is some improvement in predicting MF terms, this improvement is not substantial. The reason PPI is not particularly helpful in MF term prediction is most probably because proteins that interact with each other do not necessarily share the same specific molecular function, even when they are part of the same biological process.

In addition to the structure similarity search and the protein-protein interaction features, the results show that the functional motifs extracted improve the performance of the prediction significantly. As a further analysis, we have evaluated the performance of our functional motif detection method when both predicted and experimentally annotated GO terms are taken into consideration instead of considering experimental annotations only. The results of this experiment are reported in Section S6 and Table S2. In addition to comparing the performance of our method to BLAST, LOMETS, HHsearch, and COFACTOR, we also compared it to INGA and GoFDR, two top methods from CAFA

[45] in particular for MF and BP terms prediction, and to naïve baseline which is one of the performance references used in CAFA.

Case study

To better illustrate the performance of QAUST and the contribution of each component to the prediction, we used as an example Bacteriophage T4 gene 59 helicase assembly protein (P13342) (the cyan structure in **Figure 5**A), which is a DNA binding protein required mainly for DNA replication in the late stage of T4 infection [52]. Figure 5B shows the set of BP terms associated with this protein. In this particular example, both BLAST and INGA did not predict any correct term for this protein (the naïve root term is not counted here). When solely using global structure similarity ($P_{structure}$), we could only predict one single correct BP term. This makes sense because all the queries in our test set are difficult targets, which do not have close homologs in the template database. For instance, the closest template for this query P13342 is the methionine-tRNA ligase (the magenta structure in Figure 5A), which corresponds to the PDB ID 2CT8A. The sequence identity between P13342 and 2CT8A is only 6.84% and the TM-score between the two structures is only 0.24. Therefore, structure similarity or homology-based methods are not expected to predict the function of the query well. Structure information ($P_{Structure}$) combined with PPI predicted three correct terms out of six. On the other hand, QAUST predicted four correct terms out of six. In addition, the prediction of QAUST is at least one level deeper in the GO hierarchy than the other methods. Meanwhile, the predicted MF and CC terms for this protein by QAUST are at least as accurate as other methods. Two other case studies illustrating examples for predicting MF and CC terms can be found in Sections S7 and S8, respectively.

Experimental validation of TRIM22 dimerization

To provide an experimental assessment of the performance of QAUST, we chose the human tripartite motif-containing 22 (TRIM22) protein as an example. TRIM22 is known as an interferon-inducible protein which shows antiviral activity, such as HIV, HBV and HCV [53−55]. Recent studies also showed that TRIM22 mediates autophagy in human

macrophages [56]. However, the function of TRIM22 is still not comprehensively understood as the protein only exists in primates.

We applied QAUST to predict the function for TRIM22. Among the predicted GO terms with high consensus scores (Section S9 and Table S3), some of the CC and BP terms agree well with the previously known functions of TRIM22, such as the CC term "nucleus" and the BP term "response to virus". However, the only two predicted MF terms have quite high consensus scores, "protein binding" and "protein homodimerization activity", suggesting that TRIM22 binds to itself to form a dimer.

We thus set out to test if human TRIM22 can form homodimer using coimmunoprecipitation (**Figure 6**A). We first expressed Flag or GFP-tagged human TRIM22 protein by co-transfecting the two plasmids into HEK293T cells. After 48-hour incubation, cells were harvested and lysed (20 mM Tris-HCl pH7.5, 150 mM NaCl, 1 mM EDTA, 1% NP-40 with proteinase inhibitor). Both Flag and GFP-tagged TRIM22 were detected in cell lysate by western blot (Figure 6B). We then pulled down Flag-tagged TRIM22 from the cell lysate. For each sample, 25 μl protein A/G beads were incubated with 1 μg Flag antibody at 4 degree. Mouse IgG was used as a negative control. After 2 hours, beads were washed with lysis buffer and then incubated with 500 μg cell lysate at 4 degree for another 2 hours. Western blot showed that when Flag-tagged TRIM22 was pulled down, GFP-tagged TRIM22 can be detected by GFP antibody (Figure 6C), which showed that GFP-tagged and Flag-tagged TRIM22 bind together in HEK293T. To further confirm this binding, we did co-IP in the opposite way. Flag-tagged TRIM22 was also detected in immunoprecipitation of GFP-tagged TRIM22 (Figure 6D). These results reveal that TRIM22 can bind to itself, which is most likely to form the homodimer.

Conclusion

In this work, we developed QAUST, a method to predict biological functions of protein molecules using three main features: global and local protein structure similarity, protein-protein interaction and functional sequence motifs. In our method, we constructed the 3D structure from the amino acid sequence using I-TASSER. Functional analogs are then identified by performing global and local structural similarity search through the

functional libraries, with the scoring function involving the confidence score of structural predictions, sequence and structural similarity of the I-TASSER model with the functional templates, and the local active site matches. We have also tried to improve the performance of GO prediction by incorporating protein-protein interaction information, especially in order to improve the prediction of GO terms under BP and CC aspects. We further developed a novel predictor that extracts functional motifs that are related to a specific GO term and used it as our third predictor.

On a set of 500 non-redundant proteins, QAUST is shown to have higher function prediction accuracy than all other competing methods on most prediction tasks. This performance advantage is mainly a result of combining three different predictors which cover major aspects of proteins. Additionally, our three prediction components complement each other in the sense that they contribute differently to the prediction of the three aspects of GO. While PPI information improves significantly the prediction of BP and CC terms, functional motifs detection is mainly useful in improving MF term prediction. However, QAUST has a number of limitations that give room for possible improvement in the future. One main limitation is that QAUST is much more expensive in terms of running time compared to the other methods as reported in Section S10 and Table S4. The second limitation is that our method cannot be directly used to infer functions that are not included in EC or GO systems since it solely infers protein functions from existing protein annotations. Finally, given that the three components we used work differently in predicting different aspects of GO, it may be helpful to weight their scores differently depending on the nature of the GO term evaluated instead of combining the scores in a simple consensus. In particular, advanced machine learning methods such as deep learning [57−61], could help weight and combine the scores in a more efficient way to obtain better prediction results which could be a possible future improvement of this work.

Authors' contributions

FS designed and carried out the experiments and drafted the manuscript. AR drafted the initial manuscript and provided the EC data set and template libraries. MA helped design the webserver. STA revised the manuscript. SM and PSH helped draft the initial

manuscript. ST and WC designed and conducted experimental validation. XG supervised the algorithm design and the experiments, and helped draft and revise the manuscript.

Competing interests

The authors have declared no competing interest.

Acknowledgement

We thank Mr. Chengxin Zhang, Dr. Wei Zhang and Professor Yang Zhang for helpful discussions. The research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/1976-04 and URF/1/1976-06. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) [62], which is supported by National Science Foundation grant number ACI-1053575.

References

[1] U. Consortium, "UniProt: a hub for protein information," *Nucleic acids research*, vol. 43, no. D1, pp. D204-D212, 2014.

[2] E. Boutet *et al.*, "UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View," (in English), *Plant Bioinformatics: Methods and Protocols, 2nd Edition*, vol. 1374, pp. 23-54, 2016.

[3] S. F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," (in English), *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389-3402, Sep 1 1997.

[4] S. R. Eddy, "Profile hidden Markov models," (in English), *Bioinformatics*, vol. 14, no. 9, pp. 755-763, 1998.

[5] J. Soding, "Protein homology detection by HMM-HMM comparison," (in English), *Bioinformatics*, vol. 21, no. 7, pp. 951-960, Apr 1 2005.

[6] K. Karplus, C. Barrett, and R. Hughey, "Hidden Markov models for detecting remote protein homologies," (in English), *Bioinformatics,* vol. 14, no. 10, pp. 846-856, 1998.

[7] W. Tian and J. Skolnick, "How well is enzyme function conserved as a function of pairwise sequence identity?," *Journal of molecular biology,* vol. 333, no. 4, pp. 863-882, 2003.

[8] B. Rost, "Enzyme function less conserved than anticipated," (in English), *Journal of Molecular Biology,* vol. 318, no. 2, pp. 595-608, Apr 26 2002.

[9] R. D. Finn *et al.*, "InterPro in 2017-beyond protein family and domain annotations," (in English), *Nucleic Acids Research,* vol. 45, no. D1, pp. D190-D199, Jan 4 2017.

[10] D. A. D. L. Morais *et al.*, "SUPERFAMILY 1.75 including a domain-centric gene ontology method," (in English), *Nucleic Acids Research,* vol. 39, pp. D427-D434, Jan 2011.

[11] R. Rentzsch and C. A. Orengo, "Protein function prediction using domain families," (in English), *Bmc Bioinformatics,* vol. 14, Feb 28 2013.

[12] G. Lopez, A. Rojas, M. Tress, and A. Valencia, "Assessment of predictions submitted for the CASP7 function prediction category," (in English), *Proteins-Structure Function and Bioinformatics,* vol. 69, pp. 165-174, 2007.

[13] Y. Zhang, "Protein structure prediction: when is it useful?," (in English), *Current Opinion in Structural Biology,* vol. 19, no. 2, pp. 145-155, Apr 2009.

[14] J. Skolnick, J. S. Fetrow, and A. Kolinski, "Structural genomics and its importance for gene function analysis," (in English), *Nature Biotechnology,* vol. 18, no. 3, pp. 283-287, Mar 2000.

[15] P. Aloy, E. Querol, F. X. Aviles, and M. J. E. Sternberg, "Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking," (in English), *Journal of Molecular Biology,* vol. 311, no. 2, pp. 395-408, Aug 10 2001.

[16] A. Roy, N. Srinivasan, and V. S. Gowri, "Molecular and structural basis of drift in the functions of closely-related homologous enzyme domains: implications for function

annotation based on homology searches and structural genomics," *In silico biology,* vol. 9, no. 1, 2, pp. S41-S55, 2009.

[17] P. Bork, C. Sander, and A. Valencia, "Convergent Evolution of Similar Enzymatic Function on Different Protein Folds - the Hexokinase, Ribokinase, and Galactokinase Families of Sugar Kinases," (in English), *Protein Science,* vol. 2, no. 1, pp. 31-40, Jan 1993.

[18] R. V. Spriggs, P. J. Artymiuk, and P. Willett, "Searching for patterns of amino acids in 3D protein structures," (in English), *Journal of Chemical Information and Computer Sciences,* vol. 43, no. 2, pp. 412-421, Mar-Apr 2003.

[19] K. Kinoshita and H. Nakamura, "Identification of protein biochemical functions by similarity search using the molecular surface database eF-site," (in English), *Protein Science,* vol. 12, no. 8, pp. 1589-1595, Aug 2003.

[20] D. T. H. Chang, C. Y. Chen, W. C. Chung, Y. J. Oyang, H. F. Juan, and H. C. Huang, "ProteMiner-SSM: a web server for efficient analysis of similar protein tertiary substructures," (in English), *Nucleic Acids Research,* vol. 32, pp. W76-W82, Jul 1 2004.

[21] P. F. Gherardini and M. Helmer-Citterich, "Structure-based function prediction: approaches and applications," *Briefings in functional genomics & proteomics,* vol. 7, no. 4, pp. 291-302, 2008.

[22] R. A. Laskowski, J. D. Watson, and J. M. Thornton, "Protein function prediction using local 3D templates," *Journal of molecular biology,* vol. 351, no. 3, pp. 614-626, 2005.

[23] C. Zhang, W. Zheng, P. L. Freddolino, and Y. Zhang, "MetaGO: Predicting Gene Ontology of Non-homologous Proteins Through Low-Resolution Protein Structure Prediction and Protein–Protein Network Mapping," *Journal of molecular biology,* 2018.

[24] A. Roy, J. Y. Yang, and Y. Zhang, "COFACTOR: an accurate comparative algorithm for structure-based protein function annotation," (in English), *Nucleic Acids Research,* vol. 40, no. W1, pp. W471-W477, Jul 2012.

[25] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq, "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network," (in English), *Genome Biology,* vol. 5, no. 1, 2004.

[26] H. N. Chua, W. K. Sung, and L. Wong, "Using indirect protein interactions for the prediction of Gene Ontology functions," (in English), *Bmc Bioinformatics,* vol. 8, 2007.

[27] A. L. Barabasi, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," (in English), *Nature Reviews Genetics,* vol. 12, no. 1, pp. 56-68, Jan 2011.

[28] R. Magnez, B. Thiroux, S. Taront, Z. Segaoula, B. Quesnel, and X. Thuru, "PD-1/PD-L1 binding studies using microscale thermophoresis," *Scientific reports,* vol. 7, no. 1, p. 17623, 2017.

[29] L. Lan, N. Djuric, Y. Guo, and S. Vucetic, "MS-k NN: protein function prediction by integrating multiple data sources," in *BMC bioinformatics,* 2013, vol. 14, no. 3, p. S8: BioMed Central.

[30] R. You, Z. Zhang, Y. Xiong, F. Sun, H. Mamitsuka, and S. Zhu, "GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank," *Bioinformatics,* vol. 1, p. 9, 2018.

[31] M. Kulmanov, M. A. Khan, and R. Hoehndorf, "DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier," *Bioinformatics,* vol. 34, no. 4, pp. 660-668, 2017.

[32] Q. Gong, W. Ning, and W. Tian, "GoFDR: a sequence alignment based method for predicting protein functions," *Methods,* vol. 93, pp. 3-14, 2016.

[33] Z. Zou, S. Tian, X. Gao, and Y. Li, "mldeepre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning," *Frontiers in genetics,* vol. 9, 2018.

[34] X. Gao, D. Bu, J. Xu, and M. Li, "Improving consensus contact prediction via server correlation reduction," *BMC structural biology,* vol. 9, no. 1, p. 28, 2009.

[35] P. Chen *et al.*, "A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 13, no. 5, pp. 901-912, 2016.

[36] P. Chen, J. Z. Huang, and X. Gao, "LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone," in *BMC bioinformatics,* 2014, vol. 15, no. 15, p. S4: BioMed Central.

[37] P. Chen, J. Li, L. Wong, H. Kuwahara, J. Z. Huang, and X. Gao, "Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences," *Proteins: Structure, Function, and Bioinformatics,* vol. 81, no. 8, pp. 1351-1362, 2013.

[38] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC bioinformatics,* vol. 9, no. 1, p. 1, 2008.

[39] C. Zhang, P. L. Freddolino, and Y. Zhang, "COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information," *Nucleic acids research,* p. gkx366, 2017.

[40] D. Szklarczyk *et al.*, "STRING v10: protein-protein interaction networks, integrated over the tree of life," (in English), *Nucleic Acids Research,* vol. 43, no. D1, pp. D447-D452, Jan 28 2015.

[41] E. C. Webb, *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes* (no. Ed. 6). Academic Press, 1992.

[42] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," (in English), *Nature Genetics,* vol. 25, no. 1, pp. 25-29, May 2000.

[43] A. Roy, J. Yang, and Y. Zhang, "COFACTOR: an accurate comparative algorithm for structure-based protein function annotation," *Nucleic acids research,* vol. 40, no. W1, pp. W471-W477, 2012.

[44] S. T. Wu and Y. Zhang, "LOMETS: A local meta-threading-server for protein structure prediction," (in English), *Nucleic Acids Research,* vol. 35, no. 10, pp. 3375-3382, May 2007.

[45] Y. X. Jiang *et al.*, "An expanded evaluation of protein function prediction methods shows an improvement in accuracy," (in English), *Genome Biology,* vol. 17, Sep 7 2016.

[46] D. Piovesan, M. Giollo, E. Leonardi, C. Ferrari, and S. C. E. Tosatto, "INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity," (in English), *Nucleic Acids Research,* vol. 43, no. W1, pp. W134-W140, Jul 1 2015.

[47] C. T. Porter, G. J. Bartlett, and J. M. Thornton, "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data," (in English), *Nucleic Acids Research,* vol. 32, pp. D129-D133, Jan 1 2004.

[48] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," (in English), *Nucleic Acids Research,* vol. 33, no. 7, pp. 2302-2309, 2005.

[49] U. Consortium, "The universal protein resource (UniProt) in 2010," *Nucleic acids research,* vol. 38, no. suppl_1, pp. D142-D148, 2009.

[50] Y. Li *et al.*, "DEEPre: sequence-based enzyme EC number prediction by deep learning," *Bioinformatics,* vol. 34, no. 5, pp. 760-769, 2017.

[51] P. Radivojac *et al.*, "A large-scale evaluation of computational protein function prediction," *Nature methods,* vol. 10, no. 3, p. 221, 2013.

[52] T. C. Mueser, C. E. Jones, N. G. Nossal, and C. C. Hyde, "Bacteriophage T4 gene 59 helicase assembly protein binds replication fork DNA. The 1.45 Å resolution crystal structure reveals a novel α-helical two-domain fold1," *Journal of molecular biology,* vol. 296, no. 2, pp. 597-612, 2000.

[53] S. D. Barr, J. R. Smiley, and F. D. Bushman, "The interferon response inhibits HIV particle production by induction of TRIM22," *PLoS pathogens,* vol. 4, no. 2, p. e1000007, 2008.

[54] A. Di Pietro *et al.*, "TRIM22 inhibits influenza A virus infection by targeting the viral nucleoprotein for degradation," *Journal of virology,* vol. 87, no. 8, pp. 4523-4533, 2013.

[55] C. Yang *et al.*, "Interferon alpha (IFNα)-induced TRIM22 interrupts HCV replication by ubiquitinating NS5A," *Cellular & molecular immunology,* vol. 13, no. 1, p. 94, 2016.

[56] J. Lou, Y. Wang, X. Zheng, and W. Qiu, "TRIM22 regulates macrophage autophagy and enhances Mycobacterium tuberculosis clearance by targeting the nuclear factor–multiplicity κB/beclin 1 pathway," *Journal of cellular biochemistry,* vol. 119, no. 11, pp. 8971-8980, 2018.

[57] Z. Xia *et al.*, "DeeReCT-PolyA: a robust and generic deep learning method for PAS identification," 2018.

[58] R. Umarov, H. Kuwahara, Y. Li, X. Gao, and V. Solovyev, "Promoter analysis and prediction in the human genome using sequence-based deep learning models," *Bioinformatics,* vol. 1, p. 8, 2019.

[59] J.-S. Kim, X. Gao, and A. Rzhetsky, "RIDDLE: Race and ethnicity Imputation from Disease history with Deep LEarning," *PLoS computational biology,* vol. 14, no. 4, p. e1006106, 2018.

[60] Y. Li *et al.*, "Dlbi: deep learning guided bayesian inference for structure reconstruction of super-resolution fluorescence microscopy," *Bioinformatics,* vol. 34, no. 13, pp. i284-i294, 2018.

[61] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, and X. Gao, "Deep learning in bioinformatics: introduction, application, and perspective in big data era," *arXiv preprint arXiv:1903.00342*, 2019.

[62] J. Towns *et al.*, "XSEDE: Accelerating Scientific Discovery," (in English), *Computing in Science & Engineering,* vol. 16, no. 5, pp. 62-74, Sep-Oct 2014.

Figure legends

Figure 1 A schematic diagram of the local similarity search procedure for functional site identification

The residues of the query protein (yellow) in the active site region are shown in cyan, while those of the template protein (grey) are shown in magenta.

Figure 2 Workflow for sequence motif based function prediction in QAUST

The query sequence is searched against UniRef90 database by PSI-BLAST to identify sequence homologs with GO term annotation. For a GO term of interest, λ, the identified homologs are divided into two sets: the "annotated set" (purple) which contains homologs annotated with λ, and "not-annotated set" (green) which consists of homologs not associated with λ. From each of the two sets, frequent motifs, i.e. continuous sequence fragments, are extracted. For illustration purposes, only three five-residue-long motifs from each set are drawn. The GO term λ is predicted with confidence score $n_q(\lambda)/N(\lambda) \cdot \left[1 - n_q(\lambda^c)/N(\lambda^c)\right]$. Here, $N(\lambda)$ and $N(\lambda^c)$ are the total number of extracted frequent motifs for "annotated set" and "not-annotated set", correspondingly; while $n_q(\lambda)$ and $n_q(\lambda^c)$ are the number of frequent motifs from "annotated set" and "not-annotated set" that match the query sequence, respectively. In this example, only the motif "CLPFD" from "annotated set" matches the query, making the confidence score equals to $1/3 \cdot [1 - 0/3] = 1/3$.

Figure 3 Precision-recall curves for EC prediction by QAUST, COFACTOR, LOMETS, HHsearch and BLAST

Figure 4 Precision-recall curves for GO prediction

GO prediction performance of our method based on different sets of features, and seven other methods for each of the three GO branches.

Figure 5 A study case for protein function prediction using QAUST

A. The superimposition between the query (P13342, in cyan) and the closest template in the database (PDB ID 2CT8A, in magenta) based on the structural alignment generated

by TM-align. **B.** Predicted BP terms for protein P13342. The six BP terms (the root term, Biological Process, is a naïve term, which is not counted) shown are the experimentally annotated terms. The colored contours represent the BP terms that are predicted by the corresponding methods.

Figure 6 Experimental validation of homodimerization function of TRIM22 **A.** Illustration of the coimmunoprecipitation method to validate the homodimerization of TRIM22. We expressed Flag- or GFP-tagged human TRIM22 protein by co-transfecting two plasmid into HEK293T cells. If TRIM22 forms a homodimer, when Flag-tagged TRIM22 or GFP-tagged TRIM22 is pulled down, both Flag-tagged and GFP-tagged TRIM22 should be detected by the corresponding antibodies (4 combinations in total). **B.** Both Flag- and GFP-tagged TRIM22 expressed in HEK293T cells, detected by Western Blot. **C.** For Flag-immunoprecipitation, both Flag-tagged and GFP-tagged TRIM22 are detected by the corresponding antibodies, whereas mouse IgG is used as a negative control. **D.** For GFP-immunoprecipitation, both Flag-tagged and GFP-tagged TRIM22 are detected by the corresponding antibodies, whereas rabbit IgG is the negative control.

Tables

Table 1 Accuracy values of EC prediction for QAUST as well as five other methods

Table 2 Fmax values of each branch of GO for QAUST as well as prediction methods

Table 3 Smin (minimum semantic distance) values of each branch of GO for QAUST as well as other prediction methods

Supplementary materials

Figure S1 Precision-recall curves for GO prediction using different structure methods

Precision-recall curves for GO prediction by QAUST's structure similarity based pipeline (I-TASSER + $P_{structure}$), alternative implementations using low resolution homology

models (HHsearch + $P_{structure}$ and LOMETS + $P_{structure}$) and baseline algorithms (HHsearch and LOMETS).

Figure S2 GO prediction performance of QAUST compared to the prediction performance when using PPI and motif features only for each one of the three GO branches

Figure S3 Set of predicted MF GO terms for protein P01574

Figure S4 Set of predicted CC GO terms for protein P32157

Table S1 *P* values from the Mann-Whitney U test to assess improvement/decrease of QAUST performance in GO prediction compared to other methods

Table S2 F_{max} values from GO prediction using only experimental GO annotations for motif detection (QAUST compared to using both experimental and predicted terms)

Table S3 Predicted GO functions for protein TRIM22 with the obtained confidence score for each one of the three GO branches

Table S4 Running time of QAUST and other prediction methods for GO prediction on our dataset (500 sequences)