Large scale prediction of protein domain functions using shared annotations

Erva Ulusoy[1,2], Tunca Doğan[1,2,*]

[1]Biological Data Science Lab, Department of Computer Engineering, Hacettepe University, Ankara, Turkey
[2]Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, Ankara, Turkey
*To whom correspondence should be addressed: tuncadogan@gmail.com

Discovering the unknown functions of proteins is a major step toward understanding how biological processes work. The expensive and time-consuming nature of wet-lab experimental approaches prompted researchers to develop computational strategies for biomolecular function identification. The main idea behind these approaches, let it be network analysis- or machine learning-based, is that annotations can be transferred among proteins sharing similar characteristics (e.g., sequence, structure, protein-protein interactions, phylogenetic profiles, etc.). Considering that the biological functions of genes and proteins are multifaceted, and there is a vast amount of scientific knowledge on this subject, it is important to define biomolecular functions in a systematic and machine-readable way for function annotation. Biological ontologies (e.g., gene ontology - GO) are frequently used to meet this need by providing standardized vocabularies of functional information. Another term that is relevant in this context is protein domains. Domain composition of a protein can reveal important properties, as domains are structural and functional units that dictate how the protein should act at the molecular level.

In this study, we proposed a new method called Domain2GO with the aim of identifying unknown functions of proteins by associating their domains with Gene Ontology terms, thus redefining the problem as domain function prediction (Figure 1 displays the overall methodology). Domain2GO mappings are generated using information about the domain content of proteins together with their documented GO annotations, obtained from the InterPro and UniProt - Gene Ontology Annotation (GOA) databases, respectively. In order to obtain highly reliable associations, we employed statistical resampling and analyzed the co-occurrence patterns of domains and GO terms on the same proteins. Furthermore, three different probabilistic association measures were calculated via the expectation-maximization (EM) algorithm, in order to assess the significance of Domain2GO mappings and calculate the predictive performance of the proposed method in an ablation setting. Finalized domain-GO mappings were generated via thresholding the association scores.

For protein function prediction performance evaluation and comparison against other methods, we employed Critical Assessment of Function Annotation 3 (CAFA3) challenge datasets. The results demonstrated the potential of Domain2GO, especially when predicting molecular function and biological process terms, as it performed better than baseline predictors, curated GO associations, and various challenge participating methods (with Fmax = 0.48 and 0.36 for MFO and BPO, respectively). Furthermore, we developed a hybrid/ensemble function prediction approach by combining the domain-based Domain2GO and sequence-based BLAST (to benefit from using a larger fraction of the biomolecular knowledge space), which performed especially well in terms of predicting cellular component annotations, indicating the complementarity between these approaches. We conducted use-case studies and observed that Domain2GO predicts more specific/informative function terms, compared to the manually curated GO annotations of the same proteins. Finally, Domain2GO was applied to predict currently unknown functions of the proteins in the UniProtKB/Swiss-Prot database by propagating domain-associated GO terms to full proteins that contain those domains.

Apart from high performance, another advantage of using Domain2GO is its speed, as it is multiple orders of magnitude faster compared to machine/deep learning methods that have compatible or slightly higher prediction performance. Furthermore, its results are explainable, as opposed to black box models, since Domain2GO's function predictions are localized to specific regions/structural units in proteins. The methodology of Domain2GO can easily be adapted to predict different types of biomolecular relationships, such as the disease, phenotype, ligand/drug associations of genes and proteins. The source code, datasets, and results of the study are fully available at https://github.com/HUBioDataLab/Domain2GO.

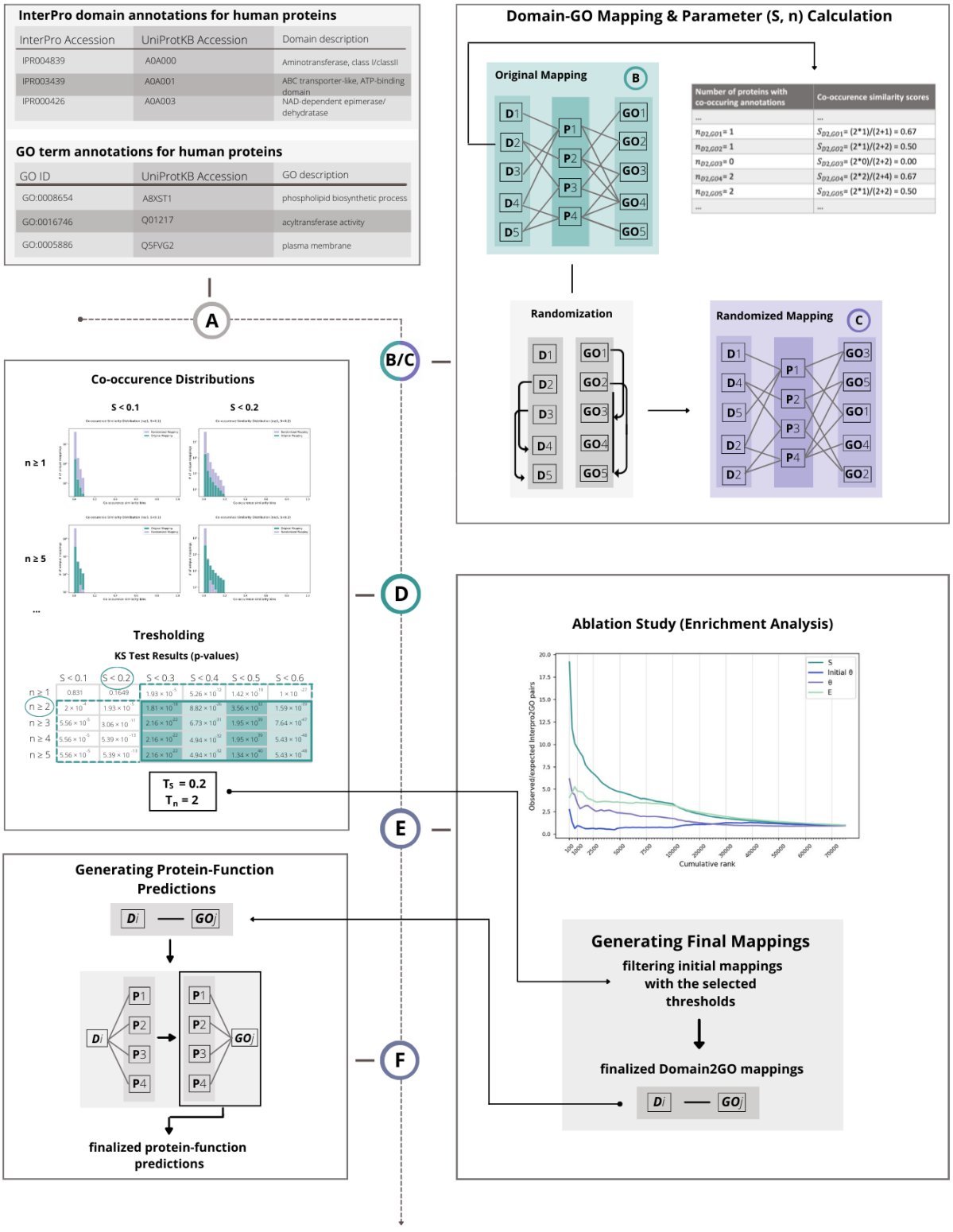

Figure 1. Schematic representation of the Domain2GO methodology; (A) obtaining the source domain (InterPro) and GO annotation (UniProt-GOA) datasets; **(B)** initial mapping of the InterPro domains and GO terms, and the parameter calculation; **(C)** generation of the randomized annotation and mapping sets to compare with the original ones; **(D)** plotting the co-occurence similarity distributions, statistical resampling of the distributions and threshold selection; **(E)** ablation study (using the EM algorithm) and the enrichment analysis of top predictions ranked by different statistical measures, generation of the finalized Domain2GO mappings by filtering initial mappings with the selected threshold; and **(F)** generation of the final protein-function association predictions.