HPO2GO: Prediction of Human Phenotype Ontology Term Associations Using Cross Ontology Annotation Co-occurrences

Tunca Doğan[1,2,3,*]

[1] Cancer Systems Biology Laboratory (CanSyL), Graduate School of Informatics, METU, Ankara, 06800, Turkey

[2] Department of Health Informatics, Graduate School of Informatics, METU, Ankara, 06800, Turkey

[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, CB10 1SD, UK

* corresponding author email address: tdogan@metu.edu.tr

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26663v2 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

ABSTRACT

Analysing the relationships between biomolecules and the genetic diseases is a highly active area of research, where the aim is to identify the genes and their products that cause a particular disease due to functional changes originated from mutations. Biological ontologies are frequently employed in these studies, which provided researchers with extensive opportunities for knowledge discovery through computational data analysis.

In this study, a novel approach is proposed for the identification of relationships between biomedical entities by automatically mapping phenotypic abnormality defining HPO terms with biomolecular function defining GO terms, where each association indicates the occurrence of the abnormality due to the loss of the biomolecular function expressed by the corresponding GO term. The proposed HPO2GO mappings were extracted by calculating the frequency of the co-annotations of the terms on the same genes/proteins, using already existing curated HPO and GO annotation sets. This was followed by the filtering of the unreliable mappings that could be observed due to chance, by statistical resampling of the co-occurrence similarity distributions. Furthermore, the biological relevance of the finalized mappings were discussed over selected cases, using the literature.

The resulting HPO2GO mappings can be employed in different settings to predict and to analyse novel gene/protein - ontology term - disease relations. As an application of the proposed approach, HPO term – protein associations (i.e., HPO2protein) are predicted. In order to test the predictive performance of the method on a quantitative basis, and to compare it with the state-of-the-art, CAFA2 challenge HPO prediction target protein set was employed. The results of the benchmark indicated the potential of the proposed approach, as HPO2GO performance was among the best (*Fmax* = 0.35). The automated cross ontology mapping approach developed in this work can easily be extended to other ontologies as well, to identify unexplored relation patterns at the systemic level. The datasets, results and the source code of HPO2GO are available for download at: https://github.com/cansyl/HPO2GO.

1. INTRODUCTION AND BACKGROUND

Systematic definition of biomedical entities (e.g., diseases, abnormalities, symptoms, traits, gene and protein attributes, activities, functions and etc.) is crucial for computational studies in biomedicine. Ontological systems, composed of standardized controlled vocabularies, are employed for this purpose. Human Phenotype Ontology (HPO) system annotates disease records (i.e., terms and definitions about diseases, recorded in relevant databases) with a standardized phenotypic vocabulary (Robinson *et al.*, 2008; Köhler *et al.*, 2017). The source of the disease information for HPO are Orphanet (Rath *et al.*, 2012), DECIPHER (Firth *et al.*, 2009), and OMIM (Amberger *et al.*, 2014) databases. Each of the phenotype terms define a specific type of abnormality encountered in human diseases (e.g., HP:0001631 - atrial septal defect). The generation of HPO terms and their associations with diseases are carried out with both manual curation efforts and automated procedures (e.g., text mining). The curation job is usually done by experts by reviewing the relevant literature publications along with the disease centric information at various biomedical data resources. The growing library of HPO currently contains nearly 12,000 phenotype terms, providing more than 123,000 annotations to 7,000 different rare (mostly Mendelian) diseases and the newly added 132,000 annotations to 3,145 common diseases (Groza *et al.*, 2015). A long-term goal of the HPO project is that the system to be adopted for clinical diagnostics, which will both provide a standardized approach to medical diagnostics and present structured machine readable biomedical data for the development of novel computational methods using data mining techniques. Apart from phenotype-disease associations, which is the main aim of the HPO project, HPO also provides phenotype-gene associations by using the known rare disease - gene relations (i.e., the information which is in the form of: "certain mutation(s) in *Gene X* causes the hereditary *Disease Y*"), using the abovementioned disease centric resources. The associations between HPO terms and biomolecules, together with the downstream analysis of these associations, help in disease gene identification and prioritization (Köhler *et al.*, 2009). With the mapping of phenotypes to human genes, HPO currently (January 2018) provides 122,166 annotations between 3,698 human genes and 6,729 HPO terms.

The Gene Ontology (GO) is an ontological system to define gene/protein attributes with an extensive controlled vocabulary (GO Consortium, 2014). Each GO term defines a unique aspect of biomolecular attributes. Similar to other ontological systems GO has a directed acyclic graph (DAG) structure, where terms are related to each other mostly with “is_a” or “part_of” relationships. GO is composed of three categories (i.e., aspects) in terms of the type of the defined gene product / protein attribute such as: *(i)* molecular function – MF (i.e., the basic function of the protein at the molecular level; e.g., GO:0016887 - ATPase activity), *(ii)* biological process - BP (i.e., the high level process, in which the protein plays a role; e.g., GO:0005975 - carbohydrate metabolic process), and *(iii)* cellular component – CC (i.e., subcellular location, where the protein carries out its intended activity; e.g., GO:0016020 - membrane). Similar to the other ontological systems, the basic way of annotating a gene or protein with a GO term is the manual curation by reviewing the relevant literature. GO also employs the concept of “evidence codes”, where all annotations are labelled with descriptions indicating the quality of the source information used for the annotation (e.g., ECO:0000006 - experimental evidence, ECO:0000501 – IEA: evidence used in automatic assertion). UniProt-GOA (Gene Ontology Annotation) database (Huntley *et al.*, 2015) houses an extensive collection of GO annotations for UniProt protein sequence and annotation knowledgebase records. In the UniProtKB/Swiss-Prot database (i.e., housing manually reviewed protein entries with highly reliable annotation) version 2018_02, there are a total of 2,850,015 GO term annotations for 529,941 protein records; whereas in UniProtKB/TrEMBL database (i.e., housing mostly electronically translated uncharacterized protein entries) version 2018_02, there are a total of 189,560,296 GO term annotations for 67,760,658 protein records. Most of the annotations for the UniProtKB/TrEMBL database entries are produced by automated predictions (UniProt Consortium, 2017).

Due to the high volume of experimental research that *(i)* discover new associations between biomolecules and ontological terms, and *(ii)* produce completely new and uncharacterized gene/protein sequences; curation efforts are having hard time in keeping up with annotation process. To aid manual curation efforts, automated computational methods come into play. These computational methods exploit the approaches and techniques widely used in the fields of data mining, machine learning and statistics, to produce probabilistic associations between biomedical entities. Critical Assessment of Functional Annotation (CAFA) challenge (Radivojac *et al.*, 2013; Jiang *et al.*, 2016) aims to evaluate the automated methods that produce GO and HPO term association predictions for protein entries, on a standard time-held benchmarking dataset. Now after its third instalment, CAFA organization have already brought together a research community, dedicated to elevate the capabilities of automated function prediction approaches closer to the level of expert review.

Protein function prediction using GO terms is a highly active area of research, where various types of approaches utilizing: amino acid sequence similarities (Hawkins *et al.*, 2009), 3D structure analysis (Roy, Yang & Zhang, 2012), semantic similarities between the ontological terms (Falda *et al.*, 2012), gene expression profiles (Lan *et al.*, 2013), protein-protein interactions - PPIs (Wass, Barton & Sternberg, 2012), shared functional domains and their arrangements (Fang & Gough, 2012; Finn et al, 2016; Doğan *et al.*, 2016) and ensemble approaches that exploit multiple feature types (Wass, Barton & Sternberg, 2012; Cozzetto *et al.*, 2013; Lan *et al.*, 2013; Rifaioglu *et al.*, 2017); are employed to model the proteins and to transfer the functional annotations from characterized proteins (i.e., the ones that have reliable annotation), to the uncharacterized ones with highly similar features. Known GO associations of genes and proteins are also used in different contexts in the literature. For example, the method "MedSim" uses the semantic similarities between GO terms for the prioritization of disease genes (Schlicker, Lengauer & Albrecht, 2010). The method "spgk" uses a shortest-path graph kernel to compute functional similarities between gene products using their GO annotations and the term relations on the GO DAG (Alvarez, Qi & Yan, 2011).

Apart from the machine-produced functional predictions for genes/proteins, automated prediction of the associations between human genes/proteins and phenotype/disease defining ontological terms is a non-trivial task, which can be utilized to identify large-scale novel disease-gene-pathway/system relations. The identification of direct disease-gene relations is a widely studied topic (Moreau & Tranchevent, 2012). A considerable amount of the existing literature about disease-gene associations involve the calculation of semantic similarities between gene products, based on the already existing ontological term annotations (Washington *et al.*, 2009; Smedley *et al.*, 2013; Deng *et al.*, 2015; Rodríguez-García *et al.*, 2017). For example, the method "PhenomeNET" was employed to generate mappings between the highly related terms across similar ontological systems (Rodríguez-García *et al.*, 2017) such as the HPO, Mammalian Phenotype Ontology – MP (Smith, Goldsmith & Eppig, 2005), Human Disease Ontology – DO (Kibbe *et al.*, 2014) and Orphanet Rare Disease Ontology – ORDO (Vasant *et al.*, 2014); for discovering novel gene-disease associations.

However, semantic similarity based approach sometimes suffers from the low coverage of especially the HPO annotations on the protein space. The authors of two recent studies have investigated this issue (Kulmanov & Hoehndorf, 2017; Peng *et al.*, 2017). In this context, increasing the coverage of HPO annotations by predicting gene/protein-HPO term associations may help semantic similarity based association studies.

There are only a few examples of HPO term-protein association prediction methods in the literature. In the "dcGO" method, the authors mapped ontological terms (including HPO) to protein domains, which are the functional units, and transferred the ontology mapping to proteins according to known domain annotations (Fang & Gough, 2012). The objective in the "PHENOstruct" method is the prediction of gene-HPO term associations using heterogeneous biological data consist of PPIs, GO annotations, literature relations, variants and known HPO annotations, together with a structured SVM classifier (Kahanda *et al.*, 2015). One of the text mining based CAFA2 challenge participating methods "EVEX", was employed for protein-HPO term association prediction. Originally, EVEX utilizes text mining approaches for large-scale integration of heterogeneous biological data and event extraction to generate a structured resource of relations, to be used in pathway curation (Van Landeghem *et al.*, 2013). In the context of HPO term prediction, EVEX scans the literature to detect proteins and phenotypic terms that co-occur on the same text corpus, and associates them with each other based on certain criteria, similar to other text mining based approaches. A network based HPO prediction method participated in CAFA2 was the "RANKS", in which the authors developed a flexible algorithmic scheme for heterogeneous biological network analysis, and used previously generated functional Interaction and functional human gene networks for gene-HPO term association prediction (Valentini *et al.*, 2016). According to the CAFA2 challenge results (Jiang *et al.*, 2016), the participating methods EVEX, RANKS, PHENOstruct and dcGO were among the top performers. In a recent study, the authors proposed two hierarchical ensemble methods: *(i)* the Hierarchical Top-Down, and *(ii)* the True Path Rule, for gene-HPO term associations; in which the hierarchical graph structure of HPO has been utilized together with the RANKS algorithm and the SVM classifier (Notaro, *et al.*, 2017).

The text mining approach is highly effective for predicting gene-disease relations in disease gene prioritization studies (Krallinger, Valencia & Hirschman, 2008). However, this approach suffers from low coverage in some cases, due to knowledge limitation in the literature. In other words, there is a bias towards detecting highly studied and already known relations. If a certain abnormality and a gene/protein has not been studied together in the same concept yet, it is often not possible to identify the relation. Network based methods are proposed on top of either text-mining results, protein-protein interactions and/or pathway data (Bromberg, 2013; Guney & Oliva, 2014; Guala & Sonnhammer, 2017) to detect indirect relations, which greatly increase the coverage; nevertheless, they still moderately rely on the previously reported relations. It is also important to note that, any predictive approach is limited by the quality and the coverage of its source information. However, the predictive output of different approaches often complement each other, contributing to fill different portions of the missing information in the knowledge space. Due to this reason, developing novel approaches to complement text mining based methods is crucial for automated ontological association prediction. The observed low performance of even the best methods in the HPO term prediction track of the CAFA2 challenge displayed the necessity of novel approaches for the biomedical entity relation prediction.

In this study, a new approach is proposed to produce phenotypic abnormality HPO term associations to both GO terms and human genes/proteins with the analysis of co-annotation fractions between the HPO and GO term combinations. For this, HPO and GO terms that are continually co-occurring on different proteins as annotations, are linked to each other (i.e., the system training step), entitled as the HPO2GO mappings. After that, proteins with a linked GO term annotation receives the corresponding HPO term as the phenotypic term prediction (i.e., the application step), entitled as the HPO2protein predictions. The idea here is to associate a HPO term Y with a GO term X in the sense that: "if a protein loses its function defined by the GO term X (or at least a reduction in the defined functionality) as a result of a genetic mutation, the loss of function may cause the disease, which is defined by the phenotype term Y". This idea is based on the nature of annotating genes/proteins with HPO terms; as for example, only the functionally problematic versions of these genes/proteins (e.g., disease causing variants) are associated with the relevant genetic diseases and their defining phenotypic abnormality terms. Mutations often lead to diseases by causing either a loss of existing functionality or a gain of new functionality in the gene products. As a result, if the HPO term Y and the GO term X are observed to be frequently co-occurring on different proteins, then the lost function, which gave way to the corresponding disease may be the one defined by the GO term *X*. This function usually corresponds to a large-scale biological process. This approach exploits the significantly higher coverage of GO term annotations for genes/proteins, compared to the HPO term annotations; to produce novel gene/protein - HPO term associations.

In order to test the biological relevance of this approach, selected HPO2GO mappings were manually examined. Additionally, the proposed methodology was employed to predict HPO terms for the human protein target dataset provided in the CAFA2 challenge. Using the benchmark set, the prediction performance was calculated and compared with the state-of-the-art HPO prediction methods. Another set of HPO2GO mappings were generated for this test, using the time-held training data provided in CAFA2. Finally, the up-to-date HPO2GO mappings were employed to generate HPO term predictions to human protein entries in the UniProtKB/Swiss-Prot database (i.e., HPO2protein predictions). The training and test datasets, along with the source code of the proposed methodology and the analyses are available for download at https://github.com/cansyl/HPO2GO.

2. METHODS

Dataset Construction

In order to generate the training sets, which were employed to generate the HPO2GO mappings, first, gene to HPO term mappings file was downloaded from the HPO web-site (January 2017 version of the file named: "ALL_SOURCES_ALL_FREQUENCIES_ genes_ to_phenotype.txt"). This file contained 153,575 annotations between 3,526 human genes and 6,018 HPO terms. This file is shared in the HPO2GO repository with the filename: "HPO_gene_to_phenotype_annotation_01_2017_ALL_SOURCES_ALL_FREQUENCIES .txt". In HPO, "genes_to_phenotype" file only contains the asserted (i.e., specific) annotations to genes; whereas "phenotype_to_genes" file contains all annotations propagated through the root of the HPO DAG, according to the true path rule. As a result, parents of the asserted terms are included as well. In this study, the asserted annotations are used in the analysis (in terms of both GO and HPO), in order to make sure the training set includes only the most reliable annotations.

Subsequently, all GO term annotations to the human proteins (with the experimental evidence codes: EXP, IDA, IPI, IMP, IGI and IEP) in UniProtKB were downloaded from the UniProt-GOA database 2017_01 version, using QuickGO browser (filename: "GOA_UniProt_human_protein_annotation.tsv"). After eliminating the repeating (i.e., redundant) annotations, the finalized file contained 179,651 GO annotations between 18,577 unique human genes and 14,632 GO terms (filename of the finalized GO annotation file: "GO_annot_human_proteins_UniProtGOA_01_2017.txt"). An additional column containing the corresponding HGNC symbols (i.e., gene symbols) of the coding genes was also included in the downloaded GO annotation file. This column was later used to combine the GO annotations with the HPO annotations, since the HPO annotation file includes the gene symbols.

Applied Methodology

The proposed methodology is divided into 2 steps: *(i)* training of the system (i.e., the generation of the HPO2GO mappings), and *(ii)* the application step (i.e., the prediction of

HPO term-protein associations – HPO2protein, using the previously generated HPO2GO mappings).

Figure 1 represents the whole HPO2GO mapping (i.e., training) procedure. For the training of the system, first, the HPO and GO annotation datasets were prepared (Figure 1.1 and Figure 1.2) and the initial HPO-GO mappings were generated (Figure 1.3) by identifying the genes/proteins shared between individual HPO and GO terms (i.e., the cases where HPO and GO terms are co-annotated to the same genes/proteins). This mapping generated 1,433,208 unique pairs between 6,005 HPO terms and 9,685 GO terms. At this point, it was observed that some of GO and HPO terms were annotated to high number of proteins, and it was highly probable to for them to co-occur on the same protein once or twice just by chance. In order to eliminate the randomly occurred mapping cases, a filtering procedure was required to be applied. For each HPO-GO term pair, a co-occurrence similarity measure, inspired from semantic similarity based approaches, has been calculated. The co-occurrence similarity formulation is given in Equation 1.

$$S_{HPOi,GOj} = \frac{2 * N_{G\ HPOi\&GOj}}{N_{G\ HPOi} + N_{G\ GOj}} \tag{1}$$

Here, $S_{HPOi,GOj}$ is the co-occurrence similarity between the HPO term "*HPOi*" and the GO term "*GOj*", $N_{G\ HPOi\&GOi}$ is the number of genes/proteins where these terms are annotated together, $N_{G\ HPOi}$ is the total number of genes with the annotation "*HPOi*", and $N_{G\ GOi}$ is the total number of genes with the annotation "*GOi*".

The mapping process and the co-occurrence similarity calculation are shown in Figure 2 with a toy example. Following the calculation of the co-occurrence similarities between all HPO-GO pairs, a thresholding operation was applied in order to distinguish between relevant mappings and the random ones. Two parameters were used for the thresholding operation: *(i)* the co-occurrence similarities (*S*), and *(ii)* the number of genes with co-occurring annotations (*n*). The aim behind employing a second parameter (i.e., *n*) was to eliminate the potential random pairing cases, where the co-occurrence similarity is still high. These cases are rare; however, it is still possible to observe a few of them especially when *n* is very small, due to extremely high number of term combinations. In Figure 2, this situation is represented on the toy example, here $S_{HPOD,GO4}$ is equal to $S_{HPOB,GO3}$; however the HPO_D-GO_4 mapping is probably less reliable compared to HPO_B-GO_3 since $n_{HPOD,GO4}$ is equal to 1.

Statistical resampling was used to determine the optimal parameter values (to be used as thresholds), that separate meaningful mappings from random ones. A permutation (i.e., randomization) test was constructed for this purpose. A randomized HPO-GO term mapping table was generated (Figure 1.4) by first, shuffling the indices of the original "HPO vs. gene" and "GO vs. gene" annotation tables; and second, calculating both the randomized co-occurrence similarities (i.e., S_R) and the number of genes with co-occurring annotations (i.e., n_R) for each random HPO-GO mapping. For each arbitrarily selected S (i.e., $S > 0$, $S \geq 0.1$, $S \geq 0.2$, …, $S \geq 0.6$) and n (i.e., $n \geq 1$, $n \geq 2$, …, $n \geq 5$) threshold value combination, the original GO-HPO mappings with lower than the threshold S and n values were deleted and a co-occurrence similarity distribution histogram was plotted using the remaining mappings (i.e., histograms plots in Figure 1 and in Figure 3). The same procedure was applied for the randomized mapping set as well. Finally, Kolmogorov–Smirnov test -KS test- (Lilliefors, 1967; Hollander, Wolfe & Chicken, 2013) is employed to calculate a test statistic for estimating whether the samples from the random and the original sets (at each S and n selection) are from the same distribution or not. KS is a nonparametric test of 1-dimensional probability distributions that can be used to compare two samples, considering the quantized distance between the samples. The null hypothesis states that the two samples are drawn from the same distribution. Here, the distribution (i.e., histogram) of the S values for the original and the randomized mapping sets represent the two samples. The reason behind using histograms instead of the actual S values was that, both high and low S values were presented in both distributions; as a result, the significance test by checking sample distances approach would not work. However, the frequencies of these high and low S values are different from each other in the original and the random distributions. If the null hypothesis is accepted at a selected threshold value pair (S and n), which means that the distributions are not statistically different from each other, then it is concluded that the selected thresholds failed to eliminate the random pairings in the original mapping (i.e., a higher threshold is required). The lowest threshold values, where the samples from the two distributions became significantly different from each other, were selected as the official thresholds. Excessive threshold values were not considered in order not to eliminate too many GO-HPO mappings. After the determination of the parameter values (i.e., *S* and *n* thresholds), the HPO2GO mappings were finalized, which ended the training process.

HPO2protein prediction step was a simple procedure, where query proteins were annotated with the HPO terms, by taking their already existing GO annotations into account. HPO2GO mappings were employed for this purpose. There were a total of 3 application runs in this study using: *(i)* CAFA2 targets as the query set (for the performance tests and for the comparison with the state-of-the-art), *(ii)* CAFA3 targets as the query set (to officially participate to the CAFA3 challenge, the results of which are yet to be announced), and *(iii)* all human protein entries in the UniProtKB/Swiss-Prot database (to generate the HPO2protein predictions).

Performance Evaluation Metrics

In this study, it was not possible to use a standard fold based cross-validation to measure the performance and to determine the parameter values in the training procedure, since in most cases, the number of genes/proteins that have a co-occurring HPO-GO term annotations were so low. As a result, it was impossible to separate the samples into training and validation sets. Instead, the optimal parameter values were determined by using statistical resampling. However, a performance test was still required in order to assess the success of the proposed approach. For this, CAFA2 challenge benchmark set was employed. Due to the fact that CAFA2 challenge was long before the analysis done in this study, HPO2GO mappings were re-generated using the training data provided in CAFA2. This was followed by the production of the HPO-protein association predictions on the CAFA2 target gene set. This analysis both served as a performance test with time-held data (one of the hardest and most informative tests for predictive models) and a performance comparison with the state-of-the-art (i.e., other HPO prediction methods participated in CAFA2). The most basic definitions of the evaluation metrics used in this test; *recall*, *precision* and *Fmax*, are shown in Equation 2, 3 and 4.

$$Rc_{\tau i} = \frac{TP_{\tau i}}{TP_{\tau i} + FN_{\tau i}} \quad (2)$$

$$Pr_{\tau i} = \frac{TP_{\tau i}}{TP_{\tau i} + FP_{\tau i}} \tag{3}$$

$$F_{max} = \max_{i=1\ldots N} \left\{ \frac{2 * Pr_{\tau i} * Rc_{\tau i}}{Pr_{\tau i} + Rc_{\tau i}} \right\} \tag{4}$$

In equations 2,3 and 4; $TP_{\tau i}$, $FN_{\tau i}$, $FP_{\tau i}$, $Rc_{\tau i}$ and $Pr_{\tau i}$ represent the number of true positives, the number of false negatives, the number of false positives, *recall* and *precision* values, respectively; at the *i*th probabilistic score threshold. *Fmax* correspond to the maximum of the *F-score* values (i.e., harmonic mean of *precision* and *recall*, shown inside the curly brackets in Equation 4) calculated for each arbitrarily selected probabilistic score threshold. Finally, *i*=1…*N* represents there are *N* different arbitrarily selected probabilistic score thresholds.

In the proposed method, probabilistic scores for each HPO-protein association prediction is calculated using the term co-occurrence similarity scores in Equation 1. If the mapping between the terms *HPOi* and *GOj* received the co-occurrence similarity score $S_{HPOi,GOj}$, then all proteins that receive the *HPOi* prediction due to the presence of *GOj* annotation obtains the probabilistic prediction score: $S_{HPOi,GOj}$. The calculation of the score in Equation 1 is set to range between 0 and 1; as a result, it can directly be used as a probabilistic score. Apart from that, probabilistic score thresholds represent values, under which the predictions are discarded. This way, a different set of predictions are given for each arbitrarily selected probabilistic score thresholds, leading to different precision and recall values. It is important to note that, probabilistic score thresholds are different from the thresholds we used to filter out unreliable HPO2GO mappings during the training process. The probabilistic score thresholds are used here (i.e., after the production of HPO2protein predictions) to produce binary predictions from continuous prediction scores, to be able to calculate performances. More details regarding the CAFA2 evaluation metrics are given in Jiang *et al.*, 2016.

3. RESULTS

Statistical Analysis of the Mappings

The initial HPO to GO mappings were generated according to the procedure explained in the Methods section (Figure 2). The initial mapping of the original set resulted in 1,433,208 mappings between 6,005 HPO terms and 9,685 GO terms. The same procedure for the randomized set produced 1,543,917 mappings between 5,995 HPO terms and 9,685 GO terms. The initial HPO-GO mappings for both the original and the randomized sets are available for download in the repository of the study (respective filenames: "HPO_GO_Raw_Original_Mapping.txt" and "HPO_GO_Random_Mapping.txt"). It was expected that the mappings generated from the random set would have lower co-occurrence similarity values on average compared to the original set mappings; in other words, they would contain less number of mappings for a particular co-occurrence similarity value. Table 1 displays the comparison of the number of mappings for different co-occurrence similarity values, between the original and the randomized sets. As observed from Table 1, when $S > 0$ there is no difference between the mappings; however as S is increased, the difference between the mappings becomes clear. Also, when S is increased, the number of mapped HPO and GO terms were decreased since many terms did not have any mappings that satisfied the stringent S values. The parameter n was not taken into account while calculating the statistics in Table 1 (i.e., $n \geq 1$ for all values in the table).

The histograms in Figure 3 display the co-occurrence similarity distributions (i.e., S) for arbitrarily selected n values. As observed from the histograms, when the mappings with low n values are eliminated, the distributions shift to the right (i.e., the mean of S increases), which can be interpreted as the mappings became more reliable. However, excessive values of n thresholds leave only a few mappings to work with, especially at n=25 and n=75 (please see the number of mappings at the vertical axis of Figure 3.C and D). Histograms in Figure 3 also show that thresholding the mappings using only n (not using S at all) would not be sufficient because there are mappings with very low S values even at very high n thresholds (i.e., 25 and 75). This observation verified the decision to use both of the parameters for the filtering operation. At this point, the statistical resampling (i.e., KS test) was applied since it was not possible to determine the optimal *n* threshold by just manually checking the histograms.

In order to find the minimum *S* and *n* values that significantly separate the original mapping from the randomized mapping, 35 different distributions, all combinations of the selected *n* (i.e., $n \geq 1, 2, \ldots, 5$) and *S* (i.e., $S > 0$, $S \geq 0.1, \ldots, 0.6$) values, were prepared and tested individually against the co-occurrence distribution of the random mapping, generated with the same *S* and *n* thresholds. This test resulted in 35 different *p*-value calculations and the minimum parameter values that satisfied the statistical significance (i.e., rejection of the null hypothesis, which states that the two samples are from the same distribution) were selected. Table 2 displays the significance results of all KS tests. The cells with "NaN" indicate the cases, where the test could not be completed due insufficient number of samples to calculate the statistic. However, incomplete tests were not a problem since the aim here was observing the minimum threshold values, where the distributions significantly diverge from each other (NaNs are located far away from this point). In Table 2, the cell with the *p*-value written in bold font (i.e., 0.0057) signifies the point, where the corresponding thresholds $n \geq 2$ and $S \geq 0.1$ yielded the required significance (*p*-value < 0.01); and thus, these values were selected as the finalized thresholds. This means that, all of the mappings with $n < 2$ and $S < 0.1$ were considered unreliable and eliminated from the initial HPO-GO mappings.

Figure 4 displays the total number of unique mappings (vertical axis) with co-occurrence similarity values greater than the corresponding threshold value (horizontal axis), for the original and the randomized distributions on the blue and red coloured curves, respectively. Figure 4.A shows the plot for the combination with greater than or equal to one co-annotated gene (i.e., $n \geq 1$), Figure 4.B displays the same value for $n \geq 2$, Figure 4.C and D for $n \geq 3$ and 4; respectively. The differences between Figure 3 and Figure 4 is that, *(i)* in Figure 4 cumulative number of mappings are given (i.e., all mappings left after thresholding with $S \geq 0.1, 0.2, \ldots$), whereas in Figure 3, the number of mappings that fall into each *S* bin is given; and *(ii)* in Figure 4, plots are given for $n \geq 1$, 2, 3 and 4 since the aim was to display the curves around the selected threshold *n* value; whereas in Figure 3, there are plots for $n \geq 1$, 5, 25 and 75 to visually indicate the distribution shifts especially at high *n* values (i.e., *n*=25 and *n*=75). Figure 4 was drawn as a visual representation of the likeness between the original and the randomized distributions at different parameter selections. As observed from Figure

4, the distributions diverged from each other at $n \geq 2$, which also is consistent with the KS test results. Considering the co-occurrence similarity parameter, $S \geq 0.1$ produced a clear separation between the original and the randomized distributions as long as n is greater than 1. Following the HPO-GO mapping elimination according to the selected thresholds, finalized HPO2GO mappings contained 45,805 associations between 3,693 HPO terms and 2,801 GO terms. HPO2GO mappings are available for download in the repository of the study (filename: "HPO2GO_Finalized_Mapping.txt").

It was only possible to use a small portion of the input GO annotations for the generation of the HPO2GO mappings because the number of HPO annotated genes were only 3,526; whereas, the number of GO annotated human genes were 18,577. Since mappings can be done over the genes/proteins with co-occurring GO and HPO annotations, only 3,526 genes/proteins were used in the process. The remaining 15,051 human genes with GO annotations were only used in the application step (i.e., HPO2protein), to predict HPO term associations.

The Biological Relevance of the Selected HPO2GO Mappings – A Case Study

Two different examples were selected and examined to discuss the biological relevance of HPO2GO mappings. The first case is the mapping between the phenotypic abnormality HPO term "absence of bactericidal oxidative respiratory burst in phagocytes" (HP:0002723) and the GO term "respiratory burst after phagocytosis" (GO:0045730), which is in the BP category. The exact definition of this GO term in the UniProt-GOA database is: "*A phase of elevated metabolic activity, during which oxygen consumption increases; this leads to the production, by an NADH dependent system, of hydrogen peroxide (H2O2), superoxide anions and hydroxyl radicals*" (URL: https://www.ebi.ac.uk/QuickGO/term/GO:0045730). These two terms are mapped to each other in HPO2GO with high confidence (i.e., $S = 0.89$ and $n = 4$). The symbols of the co-annotated genes were *CYBA*, *CYBB*, *NCF2* and *NCF1*. As observed from the names of both terms and from the description of the GO term, the HPO term defines an abnormal condition that corresponds to the absence of the biological process portrayed by the mapped GO term. This is in accordance with the logic behind mapping HPO terms with GO terms, which stated the occurrence of an abnormality (i.e., the HPO term) due to the loss of the biomolecular function defined by the mapped GO term. There also is a GO term named "respiratory burst after phagocytosis" (GO:0045728), which is related to the mapped term (GO:0045730) on the GO DAG. This term (GO:0045728) defines a more specific function that is the exact opposite of the mapped HPO term (HP:0002723), semantically. There also is an evidence for the relation between HP:0002723 and GO:0045728 in the OBO formatted term definitions of HPO (URL: http://purl.obolibrary.org/obo/hp.obo). However, in HPO2GO, GO:0045728 could not be mapped to HP:0002723 due to low coverage in the source GO annotation set. GO:0045728 was only annotated to one gene (symbol: *HCK*), which was not annotated to HP:0002723, as a result, the mapping could not be generated. Nevertheless, the mapped GO term (GO:0045730) still defined a sufficiently related function.

The second selected case was the mapping between the HPO term "cerebellar hemisphere hypoplasia" (HP:0100307) and the MF category GO term "tRNA-intron endonuclease activity" (GO:0000213). The exact definition of this specific GO term in the UniProt-GOA database is: "*Catalysis of the endonucleolytic cleavage of pre-tRNA, producing 5'-hydroxyl and 2',3'-cyclic phosphate termini, and specifically removing the intron*" (URL: https://www.ebi.ac.uk/QuickGO/term/GO:0000213). These two terms were mapped to each other in HPO2GO with high confidence (i.e., S = 0.86 and n = 3). The symbols of the co-annotated genes were *TSEN2*, *TSEN34* and *TSEN54*. The HPO term HP:0100307 is associated with the disease entry "Pontocerebellar Hypoplasia, Type 2C (PCH2C)" (OMIM:612390) in the OMIM database. According to the disease definition, pontocerebellar hypoplasia is a heterogeneous group of neurodegenerative disorders associated with abnormally small cerebellum and brainstem, and the type 2C is characterized by a progressive microcephaly from child birth (Barth, 1993). The occurrence of the disease is associated with missense mutations in either *TSEN2*, *TSEN34* or *TSEN54* genes, which are parts of the tRNA splicing endonuclease complex (Budde *et al.*, 2008). It was reported that, due to the abovementioned mutations, there was a partial loss in the function of cleaving the pre-tRNAs by the endonuclease complex (Budde *et al.*, 2008). This is another clear example for a HPO term defining an abnormal condition, that is caused by the perturbation in the function defined by the mapped GO term.

Performance Comparison with the State-of-the-art

The test for the comparison with the state-of-the-art had two objectives: *(i)* measuring the performance of the method on a time-held dataset to observe the relevance of the proposed approach, and *(ii)* investigating how the proposed method competes with the best performing methods in the literature. For this, we have re-generated the HPO2GO mappings using the CAFA2 training set, which contained 133,175 annotations between 5,586 HPO terms and 4,418 proteins, from October 2015. Whereas, CAFA2 evaluation set (i.e., benchmarking set) contained 37,090 annotations between 2,838 HPO terms and 440 proteins. The reason behind the presence of low number of annotations (and proteins) in the evaluation set was that, only the HPO annotations produced between the time of the challenge participation deadline and the end of the annotation collection period (a total duration of nearly 8 months) were used to generate the time-held evaluation set. All of the datasets, the source code and the supplementary files used in the CAFA2 challenge, and thus in this benchmarking experiment, is available through the CAFA project repositories (URLs: https://github.com/yuxjiang/CAFA2 and https://ndownloader.figshare.com/files/3658395).

HPO2GO mappings generated using the CAFA2 training set contained 27,424 mappings between 2,640 HPO terms and 2,488 GO terms. Considering the whole CAFA2 human target protein set, this mapping produced 1,922,333 HPO predictions for 16,256 proteins and 2,640 HPO terms. The calculated performance of this prediction set was low (*Fmax* = 0.30), mainly due to high number of false positive (FP) hits. However, it is also probable that many of these false positives were actually non-documented HPO annotations of the corresponding protein, as the benchmark annotation set is incomplete. Increasing the thresholds with the aim of reducing the number of false positives resulted in a matching increase in the number of false negatives (FN), with a similar *Fmax* value. With the aim of enriching the mappings (to be able to reduce FPs without a significant increase in FNs), HPO annotations of genes from January 2014 (i.e., the CAFA2 training set) were propagated to the root of HPO DAG according to the true path rule. The propagated training set contained 379,513 annotations between 4,418 human proteins and 6,576 HPO terms; as opposed to 133,175 annotations between 4,418 human proteins and 5,586 HPO terms in the asserted CAFA2 set. As observed from the dataset statistics, propagating the annotations have only added about one thousand new terms to the set; however, the number of annotations were significantly increased.

Repeating the CAFA2 benchmark analysis using propagated HPO annotations and the same GO annotations set resulted in the same performance (*Fmax* = 0.30). Next, automated GO annotations (i.e., evidence code: IEA) have been included in the source GO annotation set, which increased the number of unique GO annotations from 128,947 to 214,235 (a 66% increase). Using the propagated HPO annotations together with enlarged GO annotation set, the new HPO-GO mappings, namely "HPOprop2GOall", were generated. The finalized HPOprop2GOall contained 198,928 mappings between 4,780 HPO terms and 5,196 GO terms; as opposed to 27,424 mappings between 2,640 HPO terms and 2,488 GO terms in the original mappings. The drastic difference between the numbers have indicated the enrichment provided by annotation propagation and GO set enlargement. Subsequently, HPOprop2GOall mappings were used to predict HPO associations for all CAFA2 targets, producing 13,022,574 predictions (as opposed to 1,922,333 predictions with the asserted set). Considering only the CAFA2 benchmark proteins, the predictions generated by using the optimized parameters (i.e., $n \geq 170$ and $S \geq 0.11$) resulted in 34,486 HPO predictions for 221 benchmark proteins and 235 HPO terms, with a performance of *Fmax* = 0.35 (no-knowledge benchmark sequences in the full evaluation mode), which is among the top performances considering all of the models from 38 participating groups in the CAFA2 HPO prediction track. The *Fmax* performance of the top model in the challenge was 0.36 (Jiang *et al.*, 2016), and the performance of the naïve baseline classifier was also the same. In Figure 5, each bar displays the overall performance (*Fmax*) of the CAFA2 participators, baseline classifiers and HPO2GO. At this point in the study, additional HPO2GO mapping sets were generated using different *n* and *S* threshold selections, and tested on the CAFA2 benchmark; however, these mappings produced performances slightly inferior to the one generated using the optimal thresholds (data not shown). HPO2GO CAFA2 benchmark test prediction results are available in the repository of the study (filename: "HPO_CAFA2_benchmark_predictions.txt").

The Application of the Method to Generate Finalized HPO2protein Predictions

Up-to-date HPO2GO mappings were employed to predict HPO terms for the human protein entries in the UniProtKB/Swiss-Prot database (i.e., 20,258 protein records), and the resulting prediction set was marked as the finalized HPO2protein predictions. This set contained 3,468,582 HPO predictions for 18,101 proteins and 3,693 HPO terms. HPO2protein predictions are available in the repository of the study (filename: "HPO2protein_Predictions.txt").

Finally, up-to-date HPO2GO model was run on the CAFA3 human protein targets, which produced 3,453,130 predictions on 16,609 human proteins with 3,719 HPO terms. A more stringent subset of this prediction set (i.e., predictions produced from mappings with $S \geq 0.2$) has been officially submitted to the CAFA3 challenge. HPO2GO CAFA3 target predictions are available in the repository of the study (filename: "HPO_CAFA3_target_predictions.txt"). There was a small difference between the number of query proteins in HPO2protein and the CAFA3 target sets (20,258 as opposed to 20,197, respectively). At the time of writing this manuscript, the CAFA3 challenge results have not been announced yet.

4. DISCUSSION

As a part of the main HPO project, a sub-set of the HPO terms had already been mapped to the relevant terms from different ontology systems (e.g., anatomy, Gene Ontology process or cell type) to yield semantic interoperability with these systems. However, this mapping has been done by manually comparing the term definitions, only for a sub-set of GO terms; as a result, the coverage of this mapping was quite limited. In our approach, we linked all GO-HPO term combinations that satisfy the co-occurrence similarity tests. This way, the non-documented relations are also identified. In this sense, it is expected that the HPO2GO mappings will be valuable for the research community. It would also be interesting to compare the HPO2GO mappings with the abovementioned manually curated associations; however, it is not possible to access this data in the HPO repository anymore.

In this study, individual terms from both ontologies are mapped to each other considering the co-annotated genes/proteins. However, the initial design of the experiment considered the mapping of an HPO term to a trio of GO terms, one from each GO category (i.e., biological process – BP, molecular function – MF and cellular component – CC). This way, the corresponding phenotypic abnormality would be associated with a problem in a specific molecular event (defined by the MF term), as a part of a defined large-scale process (BP term), occurring at a particular sub-cellular location (CC term). This approach would have been more biologically relevant compared to the current design; however, the initial design failed due to the scarcity of both HPO annotations and GO annotations containing MF, BP and CC term trios (data not shown). After that, a second option was considered, where HPO terms were mapped to MF and BP term pairs to enrich the set of proteins with the required GO annotations (i.e., MF and BP at the same time); nevertheless, the same problem was encountered again. Reliable annotation sets with higher coverage, which may become available in the future with more curation efforts, may solve this problem and make the abovementioned mapping approach practical. However at present, even for the currently applied one to one term mapping approach, the main challenge is the low coverage of the predicted associations due to the small size of the source annotation sets. There can be a few alternative solutions to this problem. First of all, the training sets with enriched GO annotation (for the genes/proteins with HPO annotations) may be obtained by including the annotations with evidence codes of reduced reliability (e.g., IEA – electronically generated).

Another option for enlarging the GO annotation set can be incorporating the genes (and their respective annotations) from other organisms, that are orthologous to human genes. Scaling up the coverage of HPO set can be provided by propagating the annotations to the parent terms according to the hierarchical structure of HPO. Another option here would be taking a more elaborate approach in the mapping procedure by taking the hierarchical term relationships into account while generating the HPO2GO mappings (i.e., the parent and child terms of the target HPO-GO term pair, that are co-annotated to different genes/proteins, will also contribute to the calculation of the co-occurrence similarity of the target HPO-GO pair).

The official CAFA2 challenge results have indicated that, the methods based on sequence similarities (e.g., the baseline classifier BLAST and a few models from the participating groups) can achieve a good predictive performance considering the GO terms in the molecular function (MF) category. This was expected since it is possible to detect most of the signatures related to the molecular functions by analysing the amino acid sequence. However, most of the sequence-similarity based methods failed in predicting the cellular component (CC) GO term and HPO term associations. This can be explained for CC terms as either by the cleavage of the signals from the sequence post-translationally or the difficulties in detecting weak signals used for directing proteins to different compartments. Considering the HPO prediction, the case may completely be different. As opposed to GO terms, which define the attributes the proteins contain, HPO terms define phenotypic abnormalities caused by the protein when it loses one (or more) of its functions, usually due to certain mutations in the gene that codes the protein. Due to this reason, transferring a HPO annotation from one protein to another based on sequence similarity does not have a biological relevance, which explains the poor performance of the BLAST classifier.

An important observation regarding the CAFA tests done in this study is that, there was a large difference between the number of HPO predictions for CAFA2 and CAFA3 targets, using HPO2GO with default parameters (i.e., 1,922,333 in CAFA2 as opposed to 3,453,130 in CAFA3). There was also an increase in the number of predicted HPO terms (i.e., 2,640 in CAFA2 as opposed to 3,719 in CAFA3), and there were no significant increase in the number of targets. The increase in the number of predictions and the predicted HPO terms can be attributed to the training set getting larger and more informative in time. The training set used for CAFA2 contained 133,175 annotations; whereas, it was 153,575 for CAFA3. The comparison of the predictive performances of HPO2GO trained by the CAFA2 and the CAFA3 training sets may reveal more about the situation.

5. CONCLUSION

In this study, a simple and effective strategy, HPO2GO, is proposed to semantically map phenotypic abnormality defining HPO terms with biomolecular function defining GO terms, considering the cross-ontology annotation co-occurrences on different genes/proteins. This approach can easily be translated into novel HPO term predictions for genes/proteins, as well as into new HPO-disease or gene-disease associations. A literature based case study was carried to discuss the biological relevance of the selected HPO2GO mappings. This work also presents an application of the cross-ontology term mapping approach by generating HPO-protein associations. HPO2GO was benchmarked on CAFA2 challenge protein targets and it was revealed that the method was among the best performers of the HPO term prediction track participators (i.e., the state-of-the-art methods). Also, the up-to-date trained system was employed to predict HPO associations for all human proteins in the UniProtKB/Swiss-Prot database (i.e., HPO2protein predictions). The methodology proposed here was only meant to support the already established approaches (e.g., text mining), since different techniques with different data sources and perspectives produce results that complement distinct missing pieces of the knowledge space. It would also be interesting to analyse the complementarity between the results of the proposed method and the results of the conventional approaches participated in CAFA2 challenge; however, this was not possible since the actual predictions of the participant groups are not publicly available.

As for the future work, it is first planned to map the HPO terms to GO term trios (i.e., MF, BP and CC terms at the same time) using enriched annotation datasets, as explained at the Discussion section. Another future task is the integration of HPO2GO mappings to our freely available GO based automated protein function prediction tool/server UniGOPred (Rifaioglu *et al.*, 2018); so that, query proteins that receive a GO term prediction will be automatically associated with the HPO term(s) that are mapped to the corresponding GO term. It is expected that this approach would produce large-scale HPO predictions for uncharacterized proteins without any curated annotation, where the only available information is the amino acid sequence. The knowledge extraction methodology proposed here can easily be combined with various types of protein features employed in other predictive methods (e.g., variant information, PPIs, gene expression profiles, etc.) to generate an ensemble HPO term prediction tool that produces novel HPO-gene/protein-disease associations.

6. REFERENCES

Alvarez MA, Qi X, Yan C. 2011. A shortest-path graph kernel for estimating gene product semantic similarity. *Journal of biomedical semantics*, 2(1), 3.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2014. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1), D789-D798.

Barth PG. 2014. Pontocerebellar hypoplasias: an overview of a group of inherited neurodegenerative disorders with fetal onset. *Brain Dev.*, 15: 411-422.

Bromberg Y. 2013. Disease gene prioritization. *PLoS computational biology*, 9(4), e1002902.

Budde BS, Namavar Y, Barth PG, Nürnberg G, Becker C, van Ruissen F, ... van der Knaap MS. 2008. tRNA splicing endonuclease mutations cause pontocerebellar hypoplasia. *Nature genetics*, 40(9), 1113.

Cozzetto D, Buchan DW, Bryson K, Jones DT. 2013. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC bioinformatics*, 14(3), p.S1.

Deng Y, Gao L, Wang B, Guo X. 2015. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PloS one*, 10(2), e0115692.

Doğan T, MacDougall A, Saidi R, Poggioli D, Bateman A, O'Donovan C, Martin MJ. 2016. UniProt-DAAC: domain architecture alignment and classification, a new method for automatic functional annotation in UniProtKB. *Bioinformatics*, 32(15), 2264-2271.

Falda M, Toppo S, Pescarolo A, Lavezzo E, Di Camillo B, Facchinetti A, ... Fontana P. 2012. Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. *BMC bioinformatics*, 13(4), S14.

Fang H, Gough J. 2012. DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic acids research*, 41(D1), D536-D544.

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, ... Gough J. 2016.

InterPro in 2017—beyond protein family and domain annotations. *Nucleic acids research*, 45(D1), D190-D199.

Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, ... Carter NP. 2009. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, 84(4), 524-533.

Gene Ontology Consortium. 2014. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1), D1049-D1056.

Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, ... Vasant D. 2015. The human phenotype ontology: semantic unification of common and rare disease. *The American Journal of Human Genetics*, 97(1), 111-124.

Guala D, Sonnhammer EL. 2017. A large-scale benchmark of gene prioritization methods. *Scientific reports*, 7, 46598.

Guney E, Oliva B. 2014. Analysis of the robustness of network-based disease-gene prioritization methods reveals redundancy in the human interactome and functional diversity of disease-genes. *PLoS one*, 9(4), e94686.

Hawkins T, Chitale M, Luban S, Kihara D. 2009. PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins: Structure, Function, and Bioinformatics*, 74(3), 566-582.

Hollander M, Wolfe DA, Chicken E. 2013. *Nonparametric statistical methods*. John Wiley & Sons.

Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, *et al.* 2015. The GOA database: gene ontology annotation updates for 2015. *Nucleic acids research*, 43(D1):1057–63.

Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, ... Penfold-Brown D. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1), 184.

Kahanda I, Funk C, Verspoor K, Ben-Hur A. 2015. PHENOstruct: Prediction of human phenotype ontology terms using heterogeneous data sources. *F1000Research*, 4.

Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, Parkinson H, Schriml LM . 2014. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research*, 43:1071–8.

Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, ... Robinson PN. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4), 457-464.

Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, ... Brudno M. 2016. The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1), D865-D876.

Krallinger M, Valencia A, Hirschman L. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome biology*, 9(2), S8.

Kulmanov M, Hoehndorf R. 2017. Evaluating the effect of annotation size on measures of semantic similarity. *Journal of biomedical semantics*, 8(1), 7.

Lan L, Djuric N, Guo Y, Vucetic S. 2013. MS-k NN: protein function prediction by integrating multiple data sources. *BMC bioinformatics*, 14(3), p.S8.

Lilliefors HW. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, 62(318), 399-402.

Moreau Y, Tranchevent LC. 2012. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature reviews genetics*, 13(8), 523.

Notaro M, Schubach M, Robinson PN, Valentini G. 2017. Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods. *BMC bioinformatics*, 18(1), 449.

Peng J, Li Q, Shang X. 2017. Investigations on factors influencing HPO-based semantic similarity calculation. *Journal of biomedical semantics*, 8(1), 34.

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, ... Pandey G. 2013. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3), 221.

Rath A, Olry A, Dhombres F, Brandt MMC, Urbero B, Ayme S. 2012. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.*, 33:803–808.

Rifaioglu AS, Doğan T, Saraç ÖS, Ersahin T, Saidi R, Atalay MV, Martin MJ, Cetin-Atalay R. 2018. Large-scale automated function prediction of protein sequences and an experimental case study validation on PTEN transcript variants. *Proteins: Structure, Function, and Bioinformatics*, 86(2), 135-151.

Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, *83*(5), 610-615.

Rodríguez-García MÁ, Gkoutos GV, Schofield PN, Hoehndorf R. 2017. Integrating phenotype ontologies with PhenomeNET. *Journal of biomedical semantics*, 8(1), 58.

Roy A, Yang J, Zhang Y. 2012. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research*, 40(W1), W471-W477.

Schlicker A, Lengauer T, Albrecht M. 2010. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*, 26(18), i561-i567.

Smedley D, Oellrich A, Köhler S, Ruef B, Sanger Mouse Genetics Project, Westerfield M, ... Mungall C. 2013. PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database*, bat025.

Smith CL, Goldsmith CAW, Eppig JT. 2005. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1), R7.

UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic acids research*, 45(D1), D158-D169.

Valentini G, Armano G, Frasca M, Lin J, Mesiti M, Re M. 2016. RANKS: a flexible tool for node label ranking and classification in biological networks. *Bioinformatics*, 32(18), 2872-2874.

Van Landeghem S, Björne J, Wei CH, Hakala K, Pyysalo S, Ananiadou S, ... Ginter F. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PloS one*, 8(4), e55814.

Vasant D, Chanas L, Malone J, Hanauer M, Olry A, Jupp S, ... Rath A. 2014. Ordo: An ontology connecting rare disease, epidemiology and genetic data. *In Proceedings of ISMB*.

Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. 2009. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology*, 7(11), e1000247.

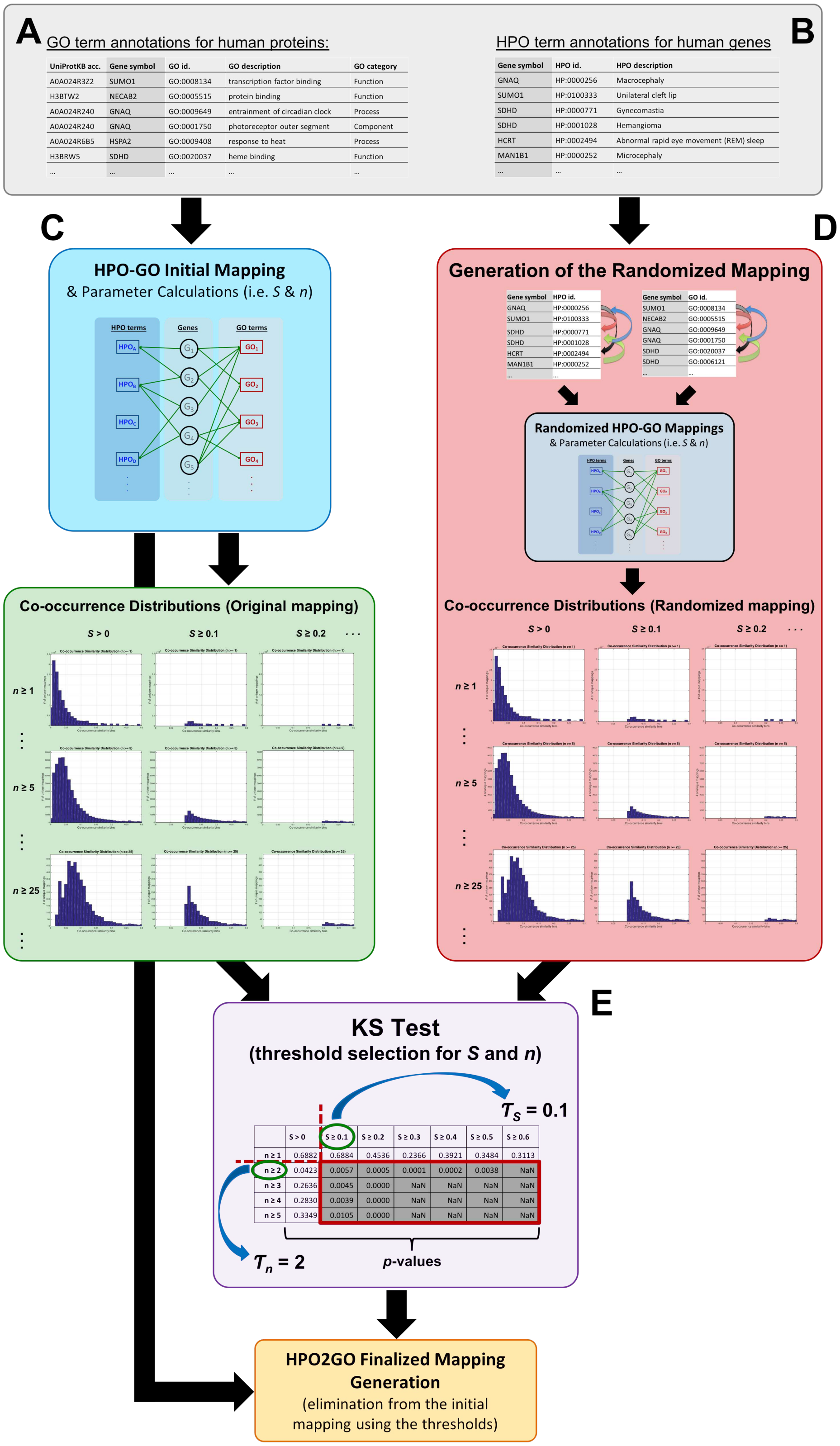
Wass MN, Barton G, Sternberg MJ. 2012. CombFunc: predicting protein function using heterogeneous data sources. *Nucleic acids research*, 40(W1), W466-W470.

Table 1. Statistics of the initial (i.e., raw) original and randomized HPO-GO mappings ($n \geq 1$).

S	**# of mappings**		**# of mapped HPO terms**		**# of mapped GO terms**	
	Original mapping	**Random mapping**	**Original mapping**	**Random mapping**	**Original mapping**	**Random mapping**
= 1	2 433	1 898	844	877	1 108	1 265
≥ 0.9	2 440	1 898	848	877	1 109	1 265
≥ 0.8	2 658	1 899	962	878	1 179	1 266
≥ 0.7	2 805	1 899	1 028	878	1 212	1 266
≥ 0.6	7 355	5 249	1 941	1 653	2 577	2 844
≥ 0.5	8 075	5 252	2 188	1 655	2 712	2 847
≥ 0.4	15 462	9 724	3 014	2 243	4 053	4 207
≥ 0.3	32 393	21 615	4 082	3 017	6 011	6 081
≥ 0.2	63 439	43 593	5 032	3 662	7 569	7 490
≥ 0.1	181 048	134 038	5 920	5 199	8 884	9 005
> 0.0	1 433 208	1 543 917	6 005	5 995	9 685	9 685

Table 2. KS test significance values for the comparison of original vs. randomized distributions at different co-occurrence similarity (*S*) and the number of co-annotated genes (*n*) thresholds.

KS test statistic		**Co-occurrence similarity threshold**						
		$S > 0$	**$S \geq 0.1$**	**$S \geq 0.2$**	**$S \geq 0.3$**	**$S \geq 0.4$**	**$S \geq 0.5$**	**$S \geq 0.6$**
# of co-annotated genes threshold	**$n \geq 1$**	0.6882	0.6884	0.4536	0.2366	0.3921	0.3484	0.3113
	$n \geq 2$	0.0423	**0.0057**	0.0005	0.0001	0.0002	0.0038	NaN
	$n \geq 3$	0.2636	0.0045	0.0000	NaN	NaN	NaN	NaN
	$n \geq 4$	0.2830	0.0039	0.0000	NaN	NaN	NaN	NaN
	$n \geq 5$	0.3349	0.0105	0.0000	NaN	NaN	NaN	NaN


A
B
C
D
E
HPO-GO Initial Mapping
& Parameter Calculations (i.e. S & n)
HPO terms
Genes
GO terms
HPO_A
HPO_B
HPO_C
HPO_D
G_1
G_2
G_3
G_4
G_5
GO_1
GO_2
GO_3
GO_4
Generation of the Randomized Mapping
Randomized HPO-GO Mappings
& Parameter Calculations (i.e. S & n)
Co-occurrence Distributions (Original mapping)
Co-occurrence Distributions (Randomized mapping)
S > 0
S ≥ 0.1
S ≥ 0.2
. . .
n ≥ 1
n ≥ 5
n ≥ 25
KS Test
(threshold selection for S and n)
T_S = 0.1
T_n = 2
p-values
HPO2GO Finalized Mapping Generation
(elimination from the initial mapping using the thresholds)

GO term annotations for human proteins:

UniProtKB acc.	Gene symbol	GO id.	GO description	GO category
A0A024R3Z2	SUMO1	GO:0008134	transcription factor binding	Function
H3BTW2	NECAB2	GO:0005515	protein binding	Function
A0A024R240	GNAQ	GO:0009649	entrainment of circadian clock	Process
A0A024R240	GNAQ	GO:0001750	photoreceptor outer segment	Component
A0A024R6B5	HSPA2	GO:0009408	response to heat	Process
H3BRW5	SDHD	GO:0020037	heme binding	Function
...	...	...	...	...

HPO term annotations for human genes

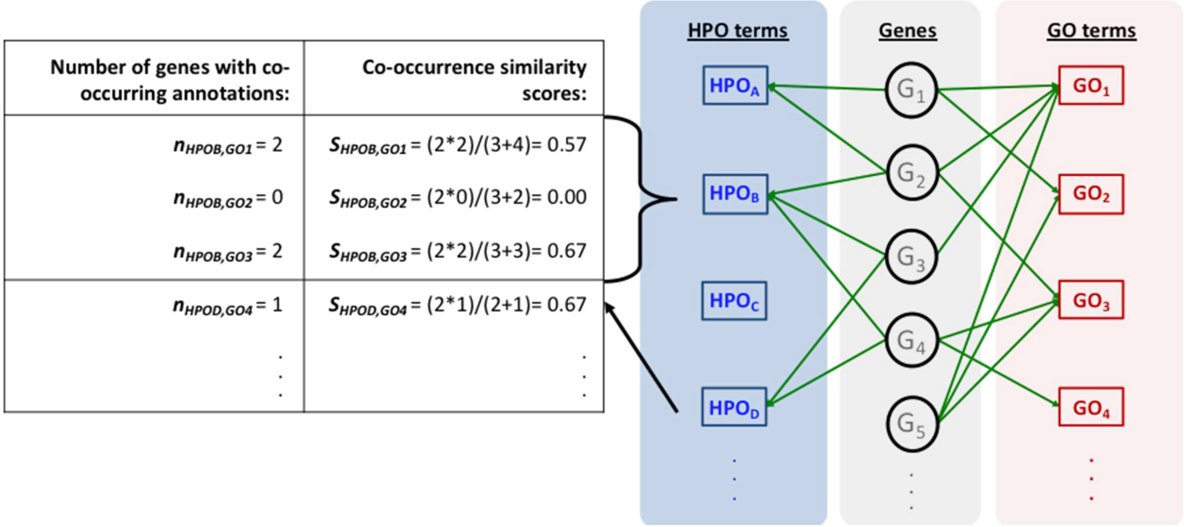
Gene symbol	HPO id.	HPO description
GNAQ	HP:0000256	Macrocephaly
SUMO1	HP:0100333	Unilateral cleft lip
SDHD	HP:0000771	Gynecomastia
SDHD	HP:0001028	Hemangioma
HCRT	HP:0002494	Abnormal rapid eye movement (REM) sleep
MAN1B1	HP:0000252	Microcephaly
...	...	...

Gene symbol	HPO id.
GNAQ	HP:0000256
SUMO1	HP:0100333
SDHD	HP:0000771
SDHD	HP:0001028
HCRT	HP:0002494
MAN1B1	HP:0000252
...	...

Gene symbol	GO id.
SUMO1	GO:0008134
NECAB2	GO:0005515
GNAQ	GO:0009649
GNAQ	GO:0001750
SDHD	GO:0020037
SDHD	GO:0006121
...	...

	S > 0	S ≥ 0.1	S ≥ 0.2	S ≥ 0.3	S ≥ 0.4	S ≥ 0.5	S ≥ 0.6
n ≥ 1	0.6882	0.6884	0.4536	0.2366	0.3921	0.3484	0.3113
n ≥ 2	0.0423	0.0057	0.0005	0.0001	0.0002	0.0038	NaN
n ≥ 3	0.2636	0.0045	0.0000	NaN	NaN	NaN	NaN
n ≥ 4	0.2830	0.0039	0.0000	NaN	NaN	NaN	NaN
n ≥ 5	0.3349	0.0105	0.0000	NaN	NaN	NaN	NaN

Figure 1. Schematic representation of the whole HPO2GO mapping (i.e., training) procedure.



Number of genes with co-occurring annotations:	Co-occurrence similarity scores:
$n_{HPOB,GO1}$ = 2	$S_{HPOB,GO1}$ = (2*2)/(3+4)= 0.57
$n_{HPOB,GO2}$ = 0	$S_{HPOB,GO2}$ = (2*0)/(3+2)= 0.00
$n_{HPOB,GO3}$ = 2	$S_{HPOB,GO3}$ = (2*2)/(3+3)= 0.67
$n_{HPOD,GO4}$ = 1	$S_{HPOD,GO4}$ = (2*1)/(2+1)= 0.67
⋮	⋮

Figure 2. Representation of the initial HPO-GO mapping process together with the calculation of co-occurrence similarities (*S*) and the number of genes with co-occurring annotations (*n*), on a toy example.

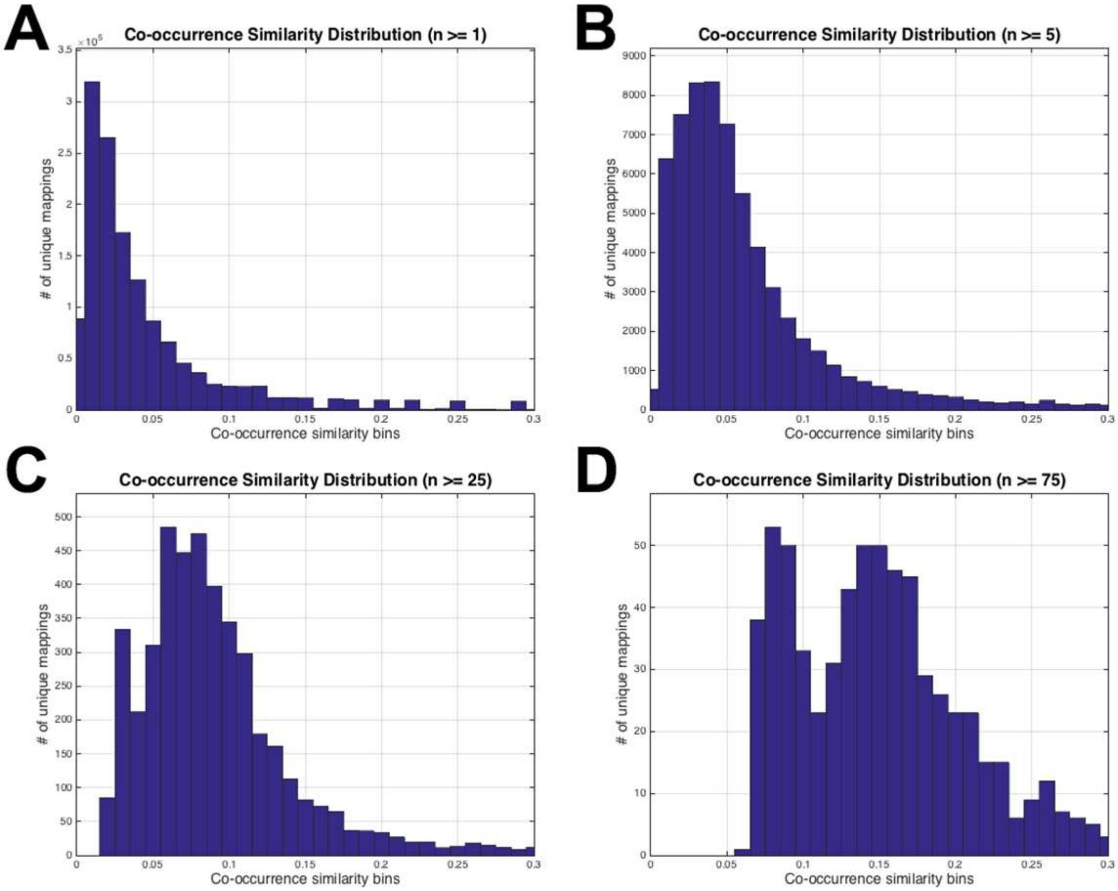


Figure 3. HPO-GO initial mappings co-occurrence similarity distributions. Each plot is drawn for a different value of the number of co-annotated genes (i.e., *n*).

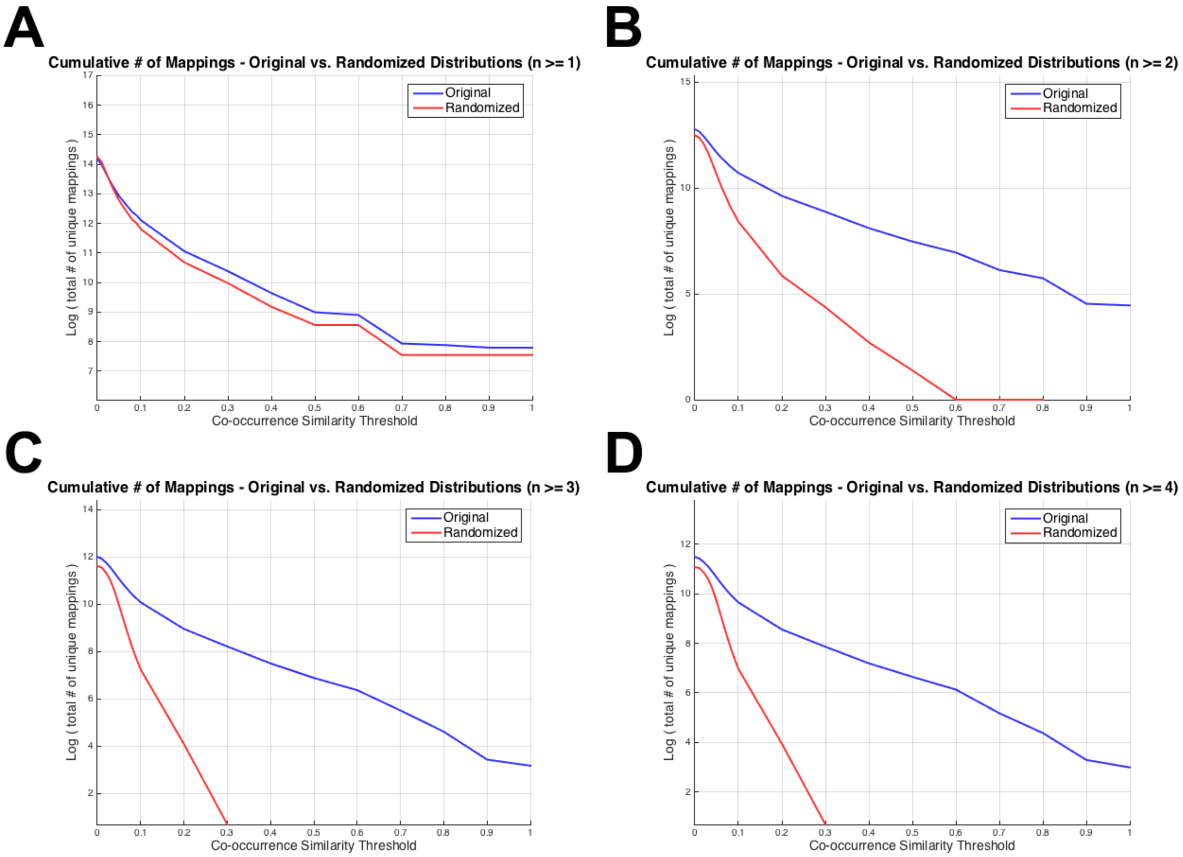


Figure 4. Cumulative plots displaying the number of HPO-GO mappings for the original (blue curve) and the randomized (red curve) distributions. Horizontal axis displays the arbitrarily selected co-occurrence similarity thresholds (i.e., τ_S), and the vertical axis represents the logarithm of the total number of mappings left after the application of the corresponding threshold. Each plot is drawn for a different value of the number of co-annotated genes (i.e., n). As the threshold (i.e., the minimum required co-occurrence similarity value to keep a mapping in the system) increase, more mappings are eliminated; thus, a monotonic decrease was observed for all plots.

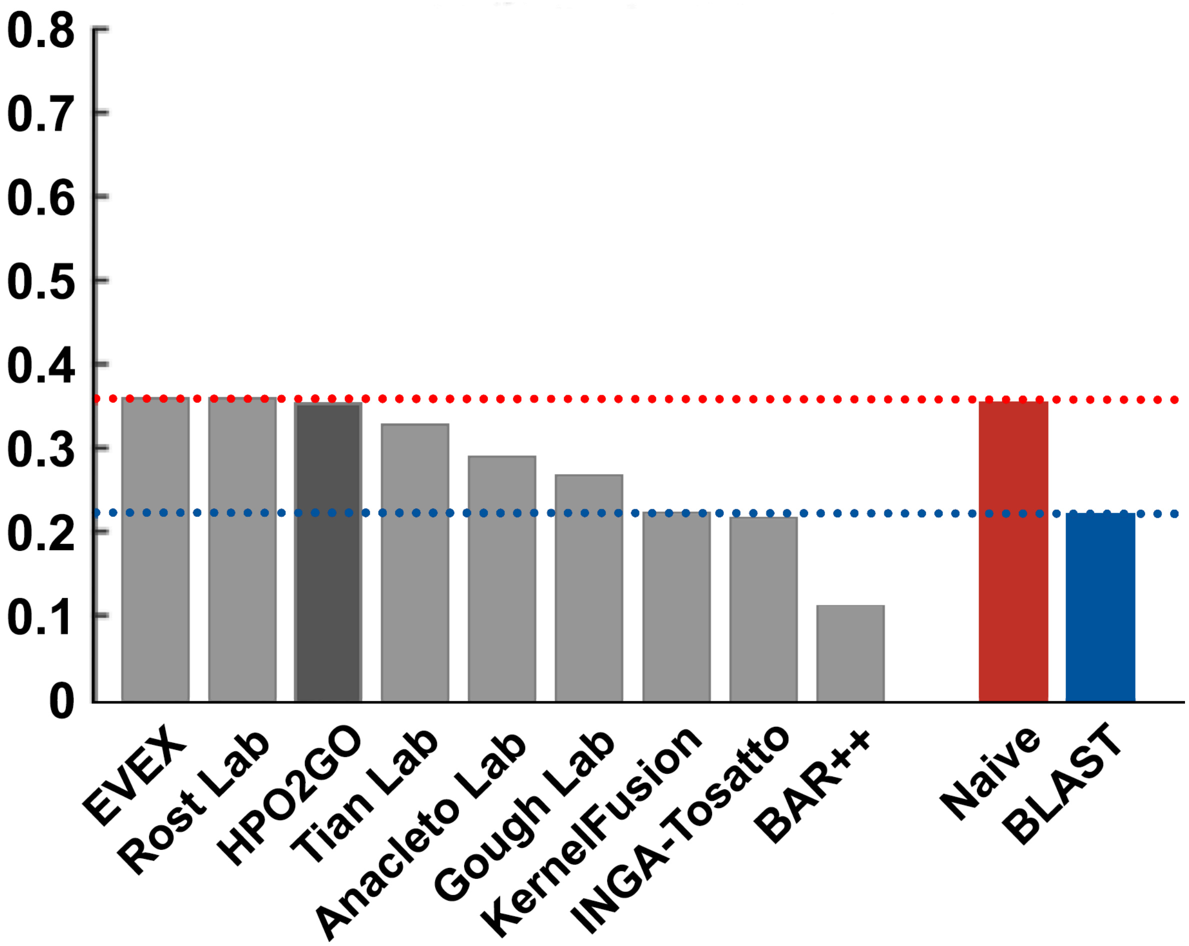


Figure 5. F1-score performance results (*Fmax*) of the top performing groups (grey bars), baseline classifiers (red and blue bars) and HPO2GO (dark grey bar) in CAFA2 HPO prediction benchmark. The lengths of the bars are directly proportional to the performance.