Associating Gene Ontology Terms with Pfam Protein Domains

Seyed Ziaeddin Alborzi, Marie-Dominique Devignes, David Ritchie

Associating Gene Ontology Terms with Pfam Protein Domains

Seyed Ziaeddin Alborzi, Marie-Dominique Devignes, and David W. Ritchie

Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, 54506, France
CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, 54506, France
Inria Nancy Grand-Est, Villers-lès-Nancy, 54600, France
seyed-ziaeddin.alborzi@inria.fr,
marie-domonique.devignes@loria.fr,
dave.ritchie@inria.fr

Abstract. With the growing number of three-dimensional protein structures in the protein data bank (PDB), there is a need to annotate these structures at the domain level in order to relate protein structure to protein function. Thanks to the SIFTS database, many PDB chains are now cross-referenced with Pfam domains and Gene ontology (GO) terms. However, these annotations do not include any explicit relationship between individual Pfam domains and GO terms. Therefore, creating a direct mapping between GO terms and Pfam domains will provide a new and more detailed level of protein structure annotation. This article presents a novel content-based filtering method called GODM that can automatically infer associations between GO terms and Pfam domains directly from existing GO-chain/Pfam-chain associations from the SIFTS database and GO-sequence/Pfam-sequence associations from the UniProt databases. Overall, GODM finds a total of 20,318 non-redundant GO-Pfam associations with a F-measure of 0.98 with respect to the InterPro database, which is treated here as a "Gold Standard". These associations could be used to annotate thousands of PDB chains or protein sequences for which their domain composition is known but which currently lack any GO annotation. The GODM database is publicly available at http://godm.loria.fr/.

Keywords: Protein Structure, Protein Function, Gene Ontology, Content-Based Filtering

1 Introduction

Proteins carry out many important biological functions. At the molecular level, these functions are often performed by highly conserved regions called "domains". Currently, the Pfam database is one of the most widely used sequence-based classifications of protein domains and domain families [1]. Protein domains may also be considered as building blocks which are combined in different ways

in order to endow different proteins with different functions. A given Pfam domain might exist in several different proteins. It is widely accepted that protein domains often correspond to distinct and stable three-dimensional (3D) structures, and that there is often a close relationship between protein structure and protein function [2]. The Protein Data Bank (PDB) [3, 4] contains more than 107,000 3D structures, that have been determined by X-ray crystallography or NMR spectroscopy. As well as sequence-based and structure-based classifications, proteins may also be classified according to their function. For example, the Gene Ontology (GO) [5] organizes a controlled vocabulary describing the biological process (BP), molecular function (MF), and cellular component (CC) aspects of gene annotation. It provides an ontology of defined terms to unify the representation of the gene and protein roles in cells. The GO vocabulary is structured as a rooted Directed Acyclic Graph (rDAG) in which GO terms are nodes connected by different hierarchical relations. Each GO term within the gene ontology has a term name, a distinct alphanumeric identifier, and a namespace indicating to which ontology it belongs.

Although the GO is very useful, it does not generally provide a direct relationship between biological function and a (sequence-based) Pfam domain. Figure 1 illustrates the different kinds of relationships that can occur when considering GO-protein annotations at the domain level. Except for simple single-domain proteins where the mapping is obvious, it is generally not possible to compare and classify structure-function relationships at the domain level. An interesting exception is the dcGO database which provides multiple ontological annotations (Gene Ontology: GO, EC, pathways, phenotype, anatomy and disease ontologies) for protein domains [6]. In dcGO, an association between an ontology term and a domain is inferred from the principle that if a term tends to be attached to proteins in UniProtKB that contain a certain domain, then the term should be associated with that domain. For each Pfam domain, dcGO compares the number of Uniprot sequences containing that domain and annotated with a certain GO term to what could be obtained if association was random. The statistical significance of the association is then assessed using a hypergeometric distribution, followed by multiple hypotheses testing in terms of false discovery rate. Only significant associations are retained in the dcGO database.

Nonetheless, we found that there are several GO-Pfam associations from manually curated data sources (e.g. InterPro) which are not present in dcGO. Moreover, based on our previous ECDomainMiner approach [7, 8] to discover associations between EC numbers and protein domains, we found that there are many reliable EC-Pfam associations which are not covered by dcGO. Furthermore, there are thousands of protein structures in the PDB which lack GO annotations. If there is a direct association between protein domains and GO terms, these structures can be annotated through their associated domains. Based on our analysis, we estimated that dcGO associations can only annotate 43% of the unannotated PDB structures. Therefore, we were motivated to develop a more systematic approach, which we call "GODM" ("GO Domain Miner"), with the aim of discovering a much larger set of GO-domain associations than dcGO.

GODM uses a "recommender-based" approach for finding direct associations between GO terms and Pfam domains. We recently developed a similar recommender-based approach called "ECDomainMiner" for assigning enzyme classification (EC) numbers to Pfam domains (manuscript accepted) [8]. Thus, the GODM approach described here represents a natural extension of our previously developed ECDomainMiner approach. Recommender systems are a subclass of information filtering system [9, 10] which seek to predict a list of items that might be of interest to an on-line customer, and are divided into two main types. Collaborative filtering approaches make associations by calculating the similarity between activities of users [11, 12]. In contrast, content-based filters predict associations between user profiles and description of items by identifying common attributes [10, 13]. Here, we use content-based filtering to associate GO terms with Pfam domains from existing GO-chain and Pfam-chain associations from SIFTS [14], and GO-sequence and Pfam-sequence associations from SwissProt and TrEMBL. As well has handling simple one-to-one associations as in dcGO (Figure 1 part A), GODM can also resolve cases where multiple GO terms are associated with multi-domain chains (Figure 1 parts B, C, and D).

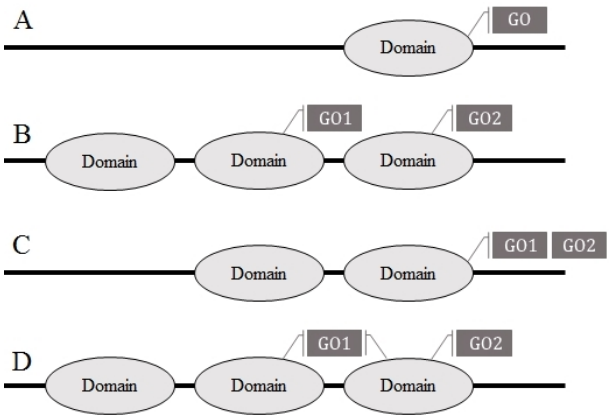

Fig. 1. A graphical representation of different situations of GO-Domain association in a protein sequence or structure.

While SwissProt and TrEMBL were originally developed separately, both databases have since been incorporated in the UniProt resource. SwissProt now represents a non-redundant, high quality, manually curated part of UniProt Knowledge Base (UniProtKB). In contrast, TrEMBL is an automatically annotated and unreviewed part of UniProtKB, and contains around 40 times more entries than SwissProt. In order to parameterise and evaluate our method, we use the InterPro database [15] which contains a large number of manually curated GO-Pfam associations. We assess the performance of our approach against a "Gold Standard" dataset derived from InterPro, and we compare our results

with the GO-Pfam associations available from the dcGO database. We also show how our database of more than 20,000 GO-Pfam associations for molecular function ontology can be exploited for automatic annotation purposes.

2 Methods

2.1 Data Preparation

Flat data files of SIFTS (July 2015), Uniprot (July 2015), and InterPro (version 53.0) were downloaded and parsed using in-house Python scripts. From the SIFTS data, associations between PDB chains and GO terms, and associations between PDB chains and Pfam domains were extracted in which each GO term is a leaf in the hierarchy of the Molecular Function ontology (GO-MF) and each Pfam refers either to a Pfam domain or a Pfam family (i.e. Pfam motifs and repeats were excluded). Associations between Uniprot sequence accession numbers (ANs) and GO terms from molecular function ontology, and AN-Pfam associations were then extracted from the SwissProt and TrEMBL sections of Uniprot to give two datasets of Swissprot associations and TrEMBL associations, respectively. Then, based on the evidence code of the GO term, associations in SwissProt and TrEMBL datasets were divided into two groups namely, associations for which GO terms were assigned in UniProtKB by manual curation, and Inferred from Electronic Annotation (IEA). These four datasets are subsequently called Swissprot, Swissprot-IEA, TrEMBL, and TrEMBL-IEA. Note that there are no evidence codes in the SIFTS.

To reduce bias due to the various numbers of identical sequences and sequences of chains in the five source datasets, all PDB chains and Uniprot sequences were grouped into clusters having identical sequences using the Uniref non-redundant cluster annotations [16]. Each cluster was assigned a unique identifier (CID), and the source GO-chain and GO-AN associations were then mapped to the corresponding cluster in order to make five sets of GO-CID associations. A similar mapping was applied to the source Pfam-chain and Pfam-AN associations to make five sets of Pfam-CID associations.

For the InterPro reference data, we extracted a total of 1,561 GO-Pfam associations in which each GO term is a leaf node of the molecular function ontology and each Pfam refers to either a Pfam domain or a Pfam family. These associations were considered to be "true" associations. However, for training and filtering purposes, we also needed some examples of "false" associations. We therefore selected a set of the lowest-scoring GO-Pfam associations with the same size as InterPro dataset from the other datasets. These associations have to belong to at least two out of five datasets with no intersection with InterPro dataset. Because these associations have very little support in the data, we consider them to be "false" associations. Then, we randomly divided the InterPro dataset and our calculated "false" associations into two "Training" and "Test" subsets of the same size (each having half of the "true" and "false" associations). These two subsets were used for training and evaluation purposes respectively.

In the rest of this article, we will refer to the InterPro dataset as our "Gold Standard" dataset.

2.2 Finding GO-Pfam Associations by Content-Based Filtering

For each of the five datasets, all GO-CID relations are encoded in a binary (GO × CID) matrix, where a 1 represents the presence of a GO annotation and a 0 represents no annotation. This matrix is then row-normalised such that each row has unit magnitude when considered as a vector. Similarly, all CID-Pfam relations are encoded in a second binary (CID × Pfam) matrix which is column-normalised. Consequently, calculating the product of the two normalised matrices corresponds to calculating a matrix of cosine similarity scores between the rows of the first matrix and the columns of the second matrix. Thus, the product matrix represents an array of raw GO-Pfam association scores. Because we wish to draw upon the relations from all five input datasets, we combine the five scores to give a single normalized confidence score (CS):

$$CS_{go,d} = \frac{\sum_i w_i S_i(go, d)}{\sum_i w_i} \tag{1}$$

where $i \in \{$*SIFTS, Swissprot, Swissprot-IEA, TrEMBL, TrEMBL-IEA*$\}$ enumerates the five datasets, w_i are weight factors, to be determined, and where an individual association score, $S_i(go, d)$ is set to zero whenever there is no data for a given go and d. In order to calculate the weight factors, we calculated Receiver-Operator-Characteristic (ROC) curves [17] using the true associations from the Interpro Training set and all other associations as background associations. The weights were varied from 0.0 to 1.0 in steps of 0.1, and for each combination, associations were scored and ranked, and area under the curve (AUC) was calculated. Finally, we selected the combination of weights that gave the best area under the curve (AUC) of the ROC curve.

2.3 Defining a Confidence Score Threshold

Having determined the best weight for each data source, we next wished to determine a threshold for the confidence score. We scored and ranked the members of the training set of InterPro, and we divided the ranked list into two subsets according to a threshold value that was varied from 0.0 to 1.0 in steps of 0.01. For each threshold value, we counted the number of true associations above the threshold, here called true positives (TPs), false associations above the threshold, false positives (FPs), false associations below the threshold, true negatives (TNs), and true associations below the threshold, false negatives (FNs). We then calculated the "F-measure" which is a harmonic mean of recall and precision using:

$$F = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2}$$

The score threshold that gave the best F-measure was confirmed by verifying that the F-measure calculated on the Test dataset is also very high. This threshold was thus selected as the best threshold to use for accepting predicted associations.

2.4 Hypergeometric Statistical Analysis

While the above procedure provides a systematic way to infer GO-Pfam associations, we wished to estimate the statistical significance, and thus the degree of confidence, that might be attached to those predictions. More specifically, we wished to calculate the probability, or "p-value", that a GO term and a Pfam domain could be found to be associated simply by chance. For example, it is natural to suppose such associations can be predicted at random if *go* or *d* are highly represented in the structure/sequence CIDs. In principle, in order to estimate the probability of getting our GO-Pfam associations by chance, one could generate random datasets by shuffling the relations between GO terms and CIDs on the one hand, and between Pfam domains and CIDs on the other hand. However, this is quite impractical given the very large numbers of CIDs, GO terms, and Pfam domains, and the complexity of the filtering procedure that would have to be repeated for each shuffled version of the dataset. Therefore, following [6], we assume that within each dataset (SIFTS, Swissprot, Swissprot-IEA, TrEMBL, or TrEMBL-IEA), the random hypothesis for the (*go*, *d*) association is represented by the hypergeometric distribution of the expected number of CIDs associated with both *go* and *d*.

Letting N denote the total number of CIDs, N_d the number of CIDs related to the Pfam domain d, and N_{go} the number of CIDs related to the GO term *go*, the hypergeometric probability distribution is given by

$$p(X_{go,d} \geqslant K_{go,d}) = \frac{\sum_{i=K_{go,d}}^{\min(N_d,N_{go})} \binom{N_{go}}{i}\binom{N-N_{go}}{N_d-i}}{\binom{N}{N_d}}, \tag{3}$$

where $p(X_{go,d} \geqslant K_{go,d})$ represents, in each dataset, the probability of having a number $X_{go,d}$ equal to or greater than the observed number $K_{go,d}$ of CIDs associated with both d and *go*. Traditionally, a p-value of less than 0.05 is taken to be statistically significant. However, because this test is applied to a large number of GO-Pfam associations, we apply a Bonferoni correction which takes into account the so-called family-wise error rate (FWER) [18]. We therefore consider any p-value less than $0.05/T$ as denoting a statistically significant inferred GO-Pfam association in a dataset, with T the total number of tested GO-Pfam associations for that dataset.

2.5 Gold, Silver, And Bronze Associations

In order to differentiate associations based on their quality and reliability, our method categorizes associations into three classes of "Gold", "Silver", and "Bronze" using their calculated similarity scores and p-values. An association belongs to the Gold class if all its available p-values are statistically significant. The Silver class consists of associations for which the number of statistically significant p-values among the five datasets is greater than or equal to the number of statistically insignificant p-values (e.g. GO-Pfam is a Silver associations if its p-values

are significant in SIFTS, SwissProt, and TrEMBL-IEA). The remaining associations are assigned to the Bronze class. An illustration of the whole procedure is shown in Figure 2.

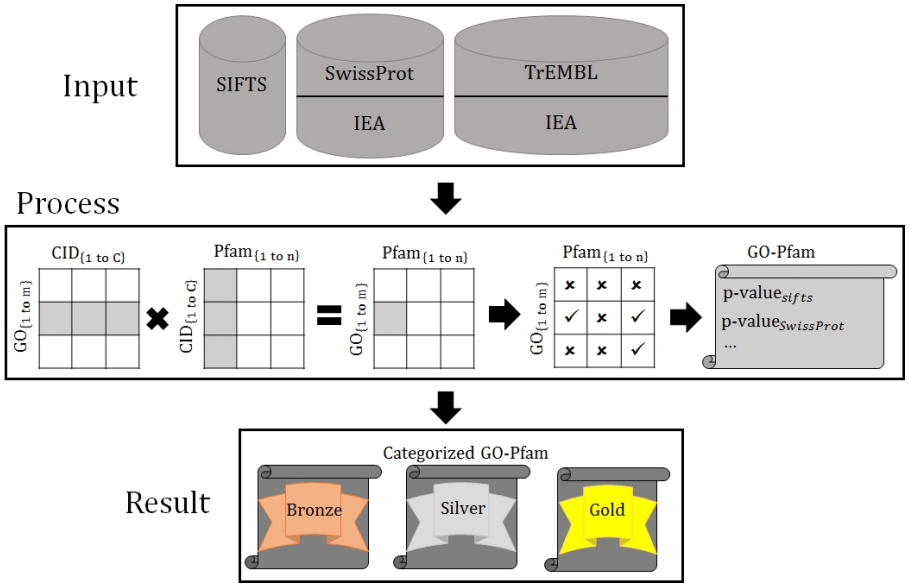

Fig. 2. A schematic overview of the GODM procedure.

3 Results

Our method takes as input five large datasets of GO-chain associations from SIFTS, and GO-sequence associations from SwissProt, SwissProt-IEA, TrEMBL and TrEMBL-IEA as well as five large datasets of Pfam-Chain and Pfam-sequence associations. These source datasets were merged to give a global dataset of $1,161,372$ non-redundant GO-Pfam associations. The Training dataset consisting of 1,560 "true" and "false" GO-Pfam associations. The best ROC-plot AUC value of 0.99 was obtained with the weights $w_{SIFTS} = 10$, $w_{SwissProt} = 1$, $w_{SwissProt-IEA} = 10$, $w_{TrEMBL} = 1$, and $w_{TrEMBL-IEA} = 8$. These weights clearly give a greater importance to the GO-Pfam associations from SIFTS and the IEA (Inferred from Electronic Annotation) section of SwissProt and TrEMBL compared to those derived from TrEMBL and the manually curated section of SwissProt.

In order to reduce the number of false associations predicted by our approach (and not just to simply optimise the overall AUC performance), various threshold values of the confidence score (using the above weights) were tested on the

Training dataset using the F-measure (Section 2.3) with respect to the number of true and false associations having scores above or below the threshold. This gave an optimal threshold score of 0.01 for a maximum F-Measure of 0.99. Applying this threshold to the Test dataset yielded a recall value of 0.965 and a precision value of 1.0 to give a F-measure of 0.98. This threshold was then used to filter GO-Pfam associations from the merged dataset according to their confidence score. It is worth noting that if the ranked list of Test associations is evaluated with respect to the median rank (since the dataset contains equal numbers of true and false instances), our scoring function gives recall and precision values of 0.965, and thus a F-measure of only 0.965. This shows that using the chosen score threshold provides an objective way to achieve high recall with good precision (i.e. a low rate of false positive associations).

3.1 Analysis of Calculated GO-Pfam Associations

The summary of our calculated GO-Pfam associations are shown in Table 1. This table shows the numbers of GO-Pfam associations along with the numbers of distinct GO terms and Pfam entries involved in those associations for the five source datasets, our merged global dataset before and after filtering (the latter corresponding to our "GODM" GO-Pfam associations), and for the InterPro dataset of true associations. The overlap between these two last datasets is shown in the last line of the table.

Dataset	GO-Pfam associations	GO terms	Pfam entries
SIFTS	10,064	2,763	3,370
SwissProt	22,435	4,220	4,669
SwissProt-IEA	28,982	3,228	4,469
TrEMBL	22,031	2,766	3,613
TrEMBL-IEA	1,136,711	4,254	9,342
Merged	1,161,372	5,510	9,929
InterPro	1,561	591	1,390
Our Associations (GODM)	**20,318**	**5,047**	**6,154**
Common with InterPro	1,519	586	1,362

Table 1. Statistics on the given and our result GO-Pfam associations.

Overall, Table 1 shows that our approach yielded a total of 20,318 GO-Pfam associations that include 1,519 associations already present in InterPro. While this shows that our method finds 97.3% of the "correct" GO-Pfam associations in InterPro, it also shows that only 2.7% of the correct InterPro associations have confidence scores below our optimal score threshold of 0.01. This relatively high proportion of common associations reflects the fact that our method is designed to give relatively strong support (Confidence Score) to the correct associations in InterPro based on the five input sources.

3.2 Comparison Between our GODM and InterPro GO-Pfam Associations

Figure 3 (A) shows the average number of GO-Pfam associations per GO term and Pfam entry both for InterPro (shown in grey) and our calculated GODM dataset (in black). The ratio for our method is higher for GO terms (4.03 versus 2.64) and Pfam entries (3.3 versus 1.12), which reflects: i) a significant enrichment in the annotation of Pfam domains; and ii) participation of Pfam domains in different functions as either a single domain or a part of a complex.

Figure 3 (B) shows the distribution of GO terms (in grey) and Pfam entries (in black) according to the number of associations they are involved in. More than 1,800 GO terms and 2,500 Pfam entries are involved in single associations, i.e. associated with a single Pfam domain and a single GO term respectively. Intersection of these single association sets yields a list of 135 one-to-one GO-Pfam associations. Nevertheless, the distribution also shows that our collection of associations rather favours multiple associations, thereby reflecting the complex many-to-many relationships that exist within the original datasets.

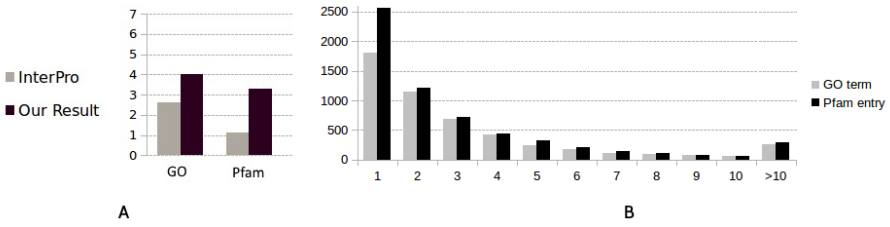

Fig. 3. A: average number of GO-Pfam associations per GO terms and per Pfam entry for the InterPro (grey) and our calculated GODM (black) datasets. B: distribution of GO terms according to their numbers of associations with Pfam entries (grey). distribution of Pfam entries according to their numbers of associations with GO terms (black).

3.3 Comparing GODM and dcGO GO-Pfam Associations

In order to compare our results with dcGO [6], we extracted the Pfam2GO associations from the dcGO website (http://supfam.org/SUPERFAMILY/dcGO) where GO terms are leaves in the MF hierarchy of GO terms. This Pfam2GO dataset includes 3,086 GO-Pfam associations. Figure 4 shows that a total of 2,401 GO-Pfam associations are common to dcGO and our results (overlap B) while only 404 GO-Pfam associations are common between InterPro and dcGO (overlap C). Furthermore, this comparison shows that our GODM dataset contains 17,917 (20,318-2,401) additional GO-Pfam associations that are not available in the dcGO dataset. In a more detailed analysis, the overlap between the GODM

and Pfam2GO datasets was studied with respect to our three quality classes. As summarized in the Table 2, the overlap between two datasets contains 1,621, 600, and 180 Gold, Silver, and Bronze associations, respectively.

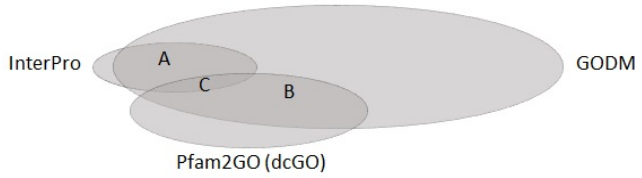

Fig. 4. Venn diagram showing the intersection between Pfam2GO (3,086 associations) from dcGO, our GODM associations (20,318 associations), and manually curated associations (1,561 associations) from InterPro. Region A (1,519 associations) is the overlap between our result and InterPro associations. Region B (2,401 associations) is the common associations between our result and Pfam2GO. Region C (404 associations) is the overlap between Pfam2GO and InterPro associations.

Dataset	GODM	Overlap with	
		Pfam2GO	InterPro
Gold	9,771	1,621	922
Silver	4,280	600	455
Bronze	6,267	180	72
Total	20,318	2401	1,519

Table 2. Overlap between associations in GODM classes, Pfam2GO of dcGO, and InterPro.

3.4 Annotating PDB Chains with GO terms

Our analysis of the July 2015 release of the SIFTS database reveals that some 41% of PDB entries currently lack a leaf GO term annotation. More specifically, we found that a total of 48,409 PDB chains lacking GO annotations in SIFTS include at least one of the 6,154 Pfam domains present in our calculated GODM associations. For those chains, GODM finds 19,371, 7,176 and 12,530 Gold, Silver, and Bronze GO-Pfam associations, respectively, giving a total of 39,077 PDB chains that could benefit from the annotations inferred by GODM. Moreover, 153 PDB chains could benefit from non ambiguous one-to-one GO-Pfam associations.

To give an example, GODM finds a Gold association between PF03018 (Dirigent-like protein) and GO term GO:0042349 ("Guiding stereospecific syn-

thesis activity"). Interestingly, the PF03018 domain is present in the PDB chain 4REV A ("Structure of the dirigent protein DRR206") which is not annotated by any GO term from the molecular function ontology. Consequently the GODM recommendation is to annotate the 4REV PDB entry with GO:0042349 term, which explicitly describes the possible function of this protein. Another example is PDB structure 2YRB, which is described only as "the solution structure of the first C2 domain from human KIAA1005 protein", and for which its previously assigned Pfam domain (PF11618) is annotated as a "protein of unknown function (DUF3250)". In this case, GODM finds a Gold association between PF11618 and GO:0031870 (thromboxane A2 receptor binding) thus indicating that this structure could be annotated with that GO term.

4 Conclusion

We have presented a systematic content-based filtering approach for assigning GO terms to protein domains and then categorizing those associations by a statistical method. This was achieved by first collecting existing annotations of protein chains or sequences, namely Pfam domain compositions on one hand and GO-MF leaf term annotations on the other.

We then applied the content-based filtering method to find a list of direct associations between GO-MF leaf terms and Pfam domains. Our approach is able to infer a total of 20,318 direct GO-Pfam associations. Thus, compared to the 1,561 manually curated GO-Pfam associations from InterPro database, our approach discovers over 13 times as many associations in a completely automatic way. We have also proposed some possible ways to further analyze the coverage of the our approach. We believe that the large numbers of GO-Pfam associations calculated using our approach can considerably contribute to enriching the annotations of PDB protein chains, and that this will facilitate a better understanding and exploitation of structure-function relationships at the protein domain level.

Acknowledgments

This project is funded by Agence Nationale de la Recherche (grant number ANR-11-MONU-006-02), the Institut National de Recherche en Informatique et Automatique, and Région Lorraine.

References

1. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., L., S.E.L., Tate, J., Punta, M.: Pfam: the protein families database. Nucleic Acids Research **42**(D1) (2014) D222–D230
2. Berg, J.M., Tymoczko, J.L., Stryer, L.: Protein structure and function. W.H. Freeman (2002)

3. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M.: The protein data bank. European Journal of Biochemistry **80**(2) (1977) 319–324
4. Gutmanas, A., Alhroub, Y., Battle, G.M., Berrisford, J.M., Bochet, E., Conroy, M.J., Dana, J.M., Montecelo, M.A.F., van Ginkel, G., Gore, S.P., Haslam, P., Hatherley, R., Hendrickx, P.M.S., Hirshberg, M., Lagerstedt, I., Mir, S., Mukhopadhyay, A., Oldfield, T.J., Patwardhan, A., Rinaldi, L., Sahni, G., Sanz-García, E., Sen, S., Slowley, R.A., Velankar, S., Wainwright, J., M.E.K.G.: PDBe: protein data bank in europe. Nucleic Acids Research **42**(D1) (2014) D285–D291
5. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature genetics **25**(1) (2000) 25–29
6. Fang, H., Gough, J.: dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic acids research **41**(D1) (2013) D536–D544
7. Alborzi, S.Z., Devignes, M.D., Ritchie, D.W.: EC-PSI: associating enzyme commission numbers with Pfam domains. bioRxiv (2015) 022343
8. Alborzi, S.Z., Devignes, M.D., Ritchie, D.W.: ECDomainMiner: discovering hidden associations between enzyme commission numbers and pfam domains. BMC Bioinformatics (2017)
9. Hanani, U., Shapira, B., Shoval, P.: Information filtering: Overview of issues, research and systems. User Modeling and User-Adapted Interaction **11**(3) (2001) 203–259
10. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. Springer (2011)
11. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc. (1998) 43–52
12. Koren, Y., Bell, R.: Advances in collaborative filtering. In: Recommender Systems Handbook. Springer (2015) 77–118
13. Basu, C., Hirsh, H., Cohen, W., et al.: Recommendation as classification: Using social and content-based information in recommendation. In: AAAI/IAAI. (1998) 714–720
14. Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., ODonovan, C., Martin, M.J., Kleywegt, G.J.: SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Research **41**(D1) (2013) D483–D489
15. Mitchell, A., Chang, H.Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.Y., Bateman, A., Punta, M., Attwood, T.K., Sigrist, C.J.A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D.A., Wu, C.H., Orengo, C., Sillitoe, I., Mi, H., Paul D. Thomas, P.D., D., F.R.: The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Research **43**(D1) (2015) D213–D221
16. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., Wu, C.H.: UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23**(10) (2007) 1282–1288
17. Fawcett, T.: An introduction to ROC analysis. Pattern recognition letters **27**(8) (2006) 861–874
18. Cui, X., Churchill, G.A., et al.: Statistical tests for differential expression in cDNA microarray experiments. Genome Biol **4**(4) (2003) 210