Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations

Seyed Ziaeddin Alborzi, Marie-Dominique Devignes, Sabeur Aridhi, Rabie Saidi, Alexandre Renaux, Maria J. Martin, David W. Ritchie

Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations

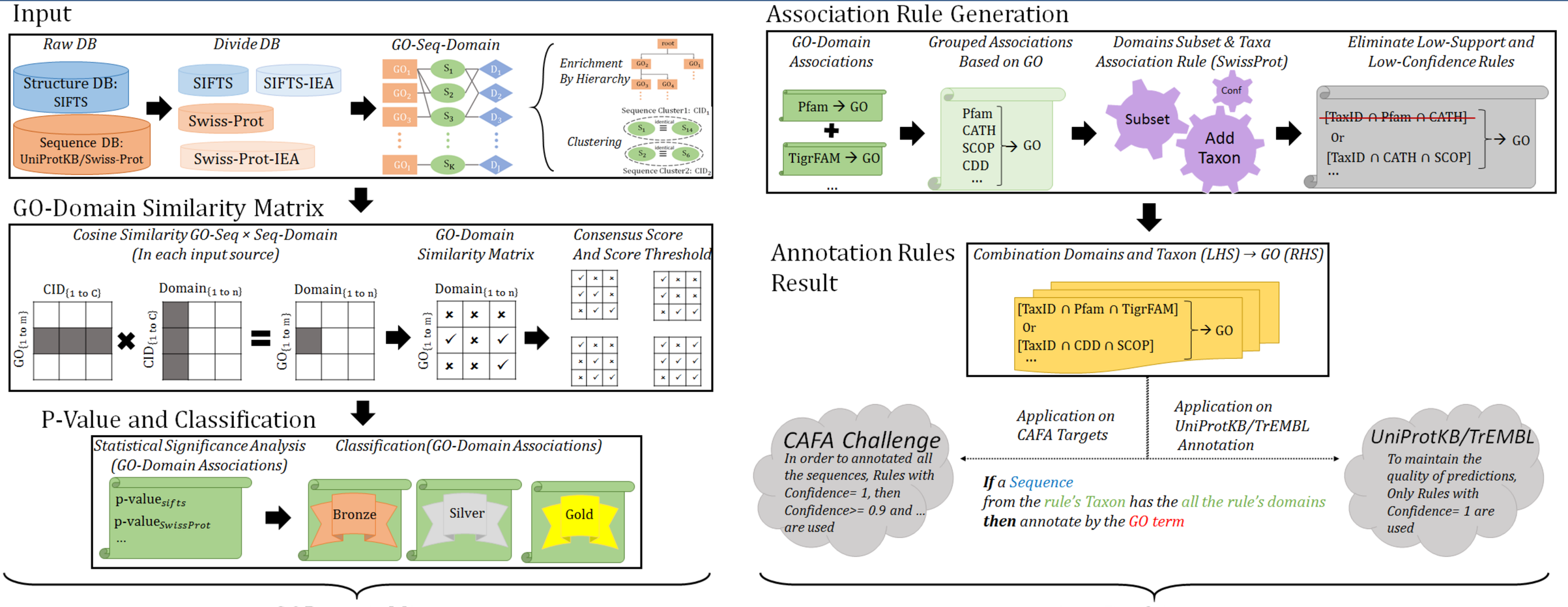
Seyed Ziaeddin Alborzi[1,3], Sabeur Aridhi[3], Marie-Dominique Devignes[2,3] Rabie Saidi[4], Alexandre Renaux[4], Maria J. Martin[4], David W. Ritchie[1]

[1]INRIA Nancy Grand-Est, [2]CNRS Nancy, [3]Université de Lorraine, [4]European Bioinformatics Institute

Introduction

The GO ontology is widely used for functional annotation of genes and proteins. It describes biological processes (BP), molecular function (MF), and cellular components (CC) in three distinct hierarchical controlled vocabularies. At the molecular level, functions are often performed by highly conserved parts of proteins, identified by sequence or structure alignments and classified into domains or families (SCOP, CATH, PFAM, TIGRFAMs, etc.). The InterPro database provides a valuable integrated classification of protein sequences and domains which is linked to nearly all existing other classifications. Interestingly, several InterPro families have been manually annotated with GO terms using expert knowledge and the literature. However, the list of such annotations is incomplete (only 20% of Pfam domains and families possess MF GO functional annotation). We therefore developed the GODomainMiner approach to expand the available functional annotations of protein domains and families (1). Based on our ECDomainMiner approach (2), we use the respective associations of protein sequences with GO terms and protein domains to infer direct associations between GO terms and protein domains. Finally, we used our calculated GO-Domain associations to devise a systematic way, called AutoProf-Annotator, to generate high confidence rules for protein sequence (or structure) annotation.

Prediction flowchart



Rules Statistics

AR Confidence > 0.5	Molecular Function	Biological Process	Cellular Component
Combination of Domains	1,723,497	1,841,000	1,543,333
Distinct Taxon	8,337	8,237	8,276
Prediction Rules	4,705	11,676	1,870

Table 1. Numbers of rules and the combination of domains result in the all rules.

AR Confidence = 1	Molecular Function	Biological Process	Cellular Component
Combination of Domains	1,692,547	1,826,347	1,4,96,772
Distinct Taxon	8,332	7966	8,266
Prediction Rules	4,673	11,582	1,853

Association Rule Samples

Rule (Confidence = 1)
- *({{PF02423 ∩ CATH:3.30.1780.10} ∩ Mammalia} → GO:0047127)*
- PF02423: Ornithine cyclodeaminase/mu-crystallin family.
- CATH: 3.30.1780.10: Ornithine cyclodeaminase.
- MF GO:0047127: hiomorpholine-carboxylate dehydrogenase.

UniProtKB/Swiss-Prot:
- Hits: 5 sequences, all are annotated with the GO term.

UniProtKB/TrEMBL Annotation
- Hits: 47 Sequences.
- 1 Sequence is annotated with the GO term.
- 7 Sequence are annotated with ancestors of the GO term (General)
- 39 Sequence are annotated by AutoProf-Annotator

Rule (Confidence = 1)
- *({{CD01399} ∩ Proteobacteria} → GO:0046348)*
- *CD01399* : GlcN6P_deaminase.
- BP GO:0046348 : amino sugar catabolic process.

UniProtKB/Swiss-Prot:
- Hits: 103 sequences, all are annotated with the GO term.

UniProtKB/TrEMBL Annotation
- Hits: 1930 Sequences.
- 1171 Sequence is annotated with the GO term.
- 569 Sequence are annotated with ancestors of the GO term (General)
- 190 Sequence are annotated by AutoProf-Annotator

Summary of CAFA Challenge Predictions

Molecular Function GO | Biological Process GO | Cellular Component GO

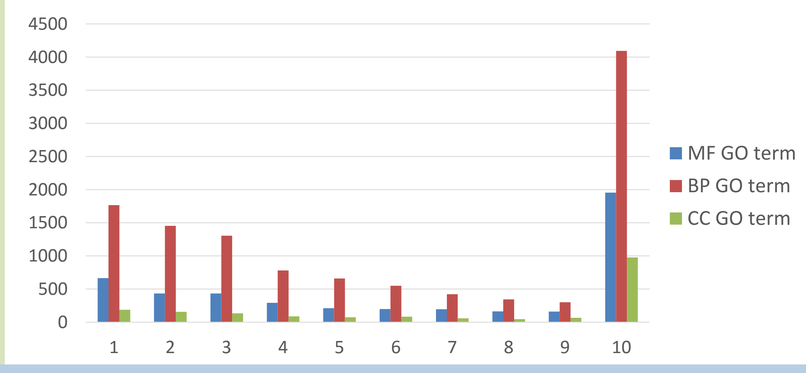
Table 2. GO function prediction for 130,000 CAFA Targets.

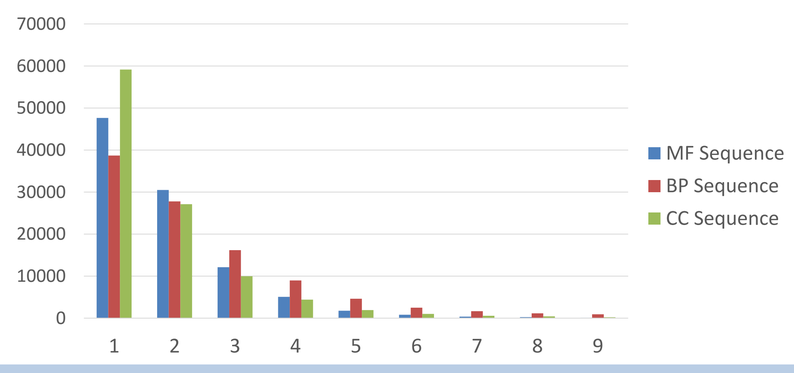
	CAFA Targets	CAFA Targets (Conf. 1)
Prediction	188,549	164,359
Sequence	98,849	81,248
GO term	4,705	4,673
Common to existing GO terms	ISMB/ECCB Function SIG	ISMB/ECCB Function SIG

	CAFA Targets	CAFA Targets (Conf. 1)
Prediction	315,310	229,006
Sequence	106,346	72,543
GO term	11,676	11,582
Common to existing GO terms	ISMB/ECCB Function SIG	ISMB/ECCB Function SIG

	CAFA Targets	CAFA Targets (Conf. 1)
Prediction	191,835	150,411
Sequence	105,274	76,233
GO term	1,870	1,853
Common to existing GO terms	ISMB/ECCB Function SIG	ISMB/ECCB Function SIG

Fig2. CAFA3: Distribution according to the number of GO terms for each sequence (right), and sequences for each GO term (left)





Annotation Examples

PRS56_Human is a target of CAFA3
- Very well annotated protein sequence in UniProtKB/Swiss-Prot
- Annotation Score: 5 - Experimental evidence at protein level
- Existing information in UniProtKB/Swiss-Prot:
  - MF **GO:0004252**
  - BP GO:0043010
  - BP GO:0006508
  - CC GO:0005783
- AutoProf-Annotator predicts following GO terms:
  - MF **GO:0004252** (Exact Match) (Conf. = 1)
  - BP GO:0044699 (Ancestor of GO:0043010) (Conf. = 0.6)
  - BP GO:0019538 (Parent of GO:0006508) (Conf. = 0.7)
  - CC GO:0044464 (Ancestor of GO:0005783) (Conf. = 0.7)

6PGL_SALCH is a target of CAFA3
- Annotated protein sequence in UniProtKB/Swiss-Prot
- Annotation Score: 2 – Protein inferred from homology
- Existing information in UniProtKB/Swiss-Prot :
  - MF **GO:0017057**
  - BP GO:0006006
  - BP GO:0006508
- AutoProf-Annotator predicts following GO terms:
  - MF **GO:0017057** (Exact Match) (Conf. = 1)
  - BP GO:0006006 (Exact Match) (Conf. = 1)
  - BP GO:0006508 (Exact Match) (Conf. = 1)
  - CC GO:0042597 (New Prediction) (Conf. = 0.9)

CONCLUSION

Our GODomainMiner approach provides a substantial enrichment of functional annotations at the protein domain level which has been exploited to develop a novel system here called AutoProf-Annotator for protein functional annotation. We used the AutoProf-Annotator to annotate target sequences in CAFA challenge.

LITERATURE CITED

1. Alborzi, Seyed Ziaeddin, Marie-Dominique Devignes, and David W. Ritchie. "Associating Gene Ontology Terms with Pfam Protein Domains." International Conference on Bioinformatics and Biomedical Engineering. Springer, Cham, 2017.
2. Alborzi, Seyed Ziaeddin, Marie-Dominique Devignes, and David W. Ritchie. "ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains." BMC bioinformatics 18.1 (2017): 107.

ACKNOWLEDGMENTS

This project is funded by the Agence Nationale de la Recherche (grant reference ANR-11-MONU-006-02), Inria and the Lorraine Region. Travel to the meeting was made possible, in part, by a travel award from the NSF.

Contact INRIA Nancy Grand Est, LORIA, Bureau B133, France *Email*: seyed-ziaeddin.alborzi@inria.fr *Website*: https://members.loria.fr/SAlborzi/ *Phone*: +33 (0)7 83 29 89 83