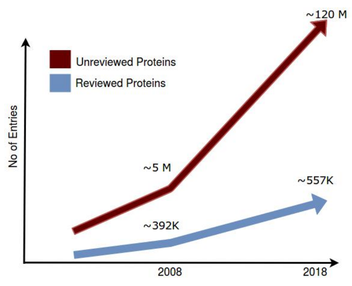
GrAPFI: Graph Based Inference for Automatic Protein Function Annotation

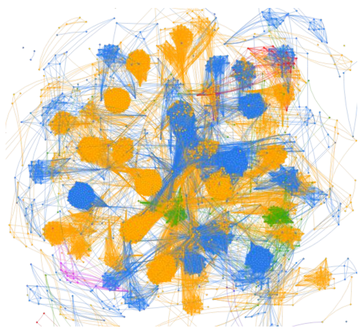
Bishnu Sarker, David W Ritchie and Sabeur Aridhi
University of Lorraine, CNRS, INRIA, LORIA, Vandoeuvre-les-Nancy, France
bishnu.sarker@inria.fr

Scan me

Introduction



1(a)



1(b)

The growing number of protein sequences in UniProtKB (Fig-1(a)) makes it increasingly expensive to annotate them manually. Here, we present GrAPFI (Graph based Automatic Protein Function Inference), a tool for the automatic functional annotation of proteins with Enzyme Commission (EC) numbers. The EC System uses a four digit numbering with a hierarchical structure. GrAPFI utilizes the domain composition of the proteins. Our general observation is that protein domains shares functional properties. Fig-1(b). illustrates this for proteins from viruses dataset. The 6 different colors in this figure correspond to the 6 different top level EC classes.

Graph Construction

GrAPFI constructs the protein graph based on domain composition of each protein. Each node of the graph represents a protein, while a link between two nodes means that the proteins exhibit a given minimum level of domain similarity. Each node u is identified by a set of labels $L(u)$ (one or more annotations to propagate), has a set of neighbours $N(u)$, and for every neighbour v it has an associated weight $W_{u,v}$. If u and v have domain composition D1=(d1,d2,d3,d4) and D2=(d1,d3,d5), then,

$$W_{u,v} = \frac{|(d1,d2,d3,d4) \cap (d1,d3,d5)|}{|(d1,d2,d3,d4) \cup (d1,d3,d5)|}$$
$$= \frac{|(d1,d3)|}{|(d1,d2,d3,d4,d5)|}$$
$$= \frac{2}{5} = 0.4.$$

GrAPFI Annotation Workflow

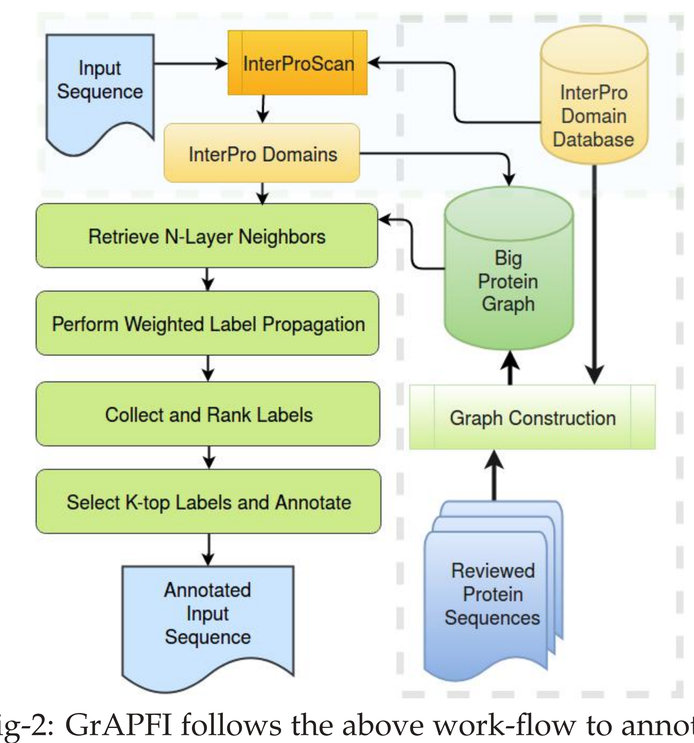
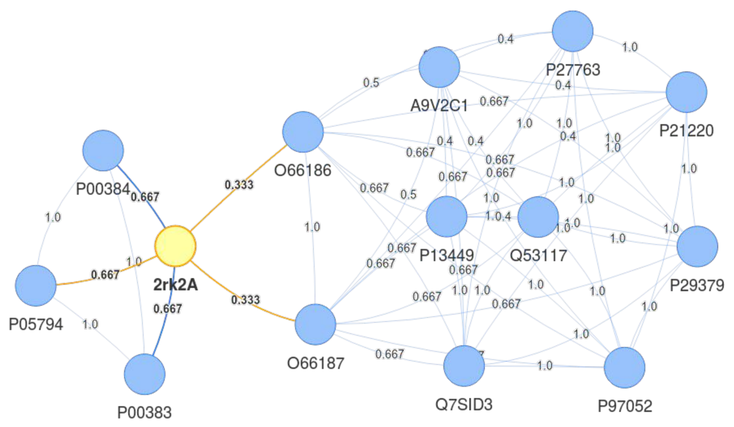
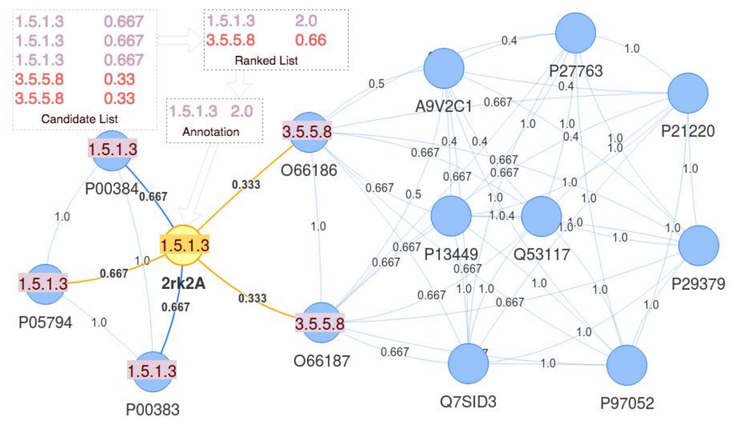


Fig-2: GrAPFI follows the above work-flow to annotate an un-reviewed protein.

Example Function Annotation of 2rk2A Enzyme



3(a)

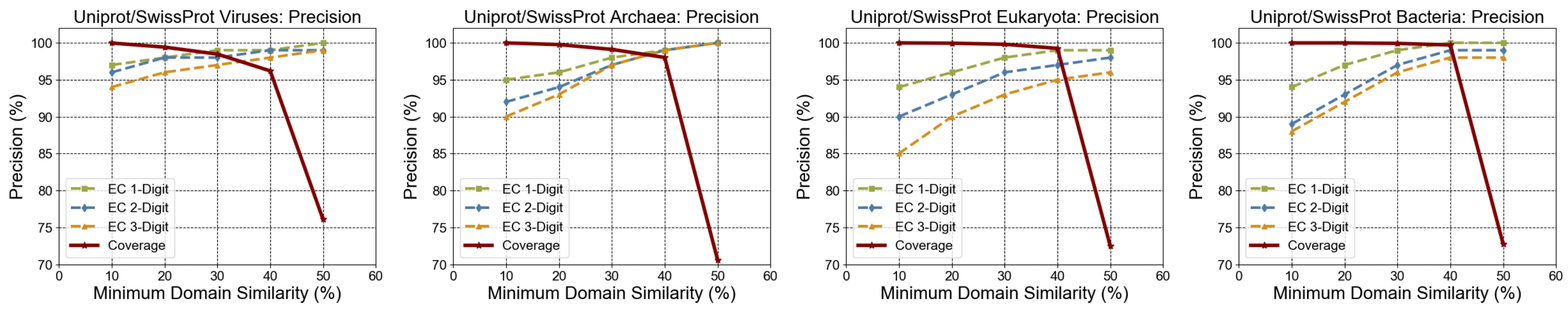


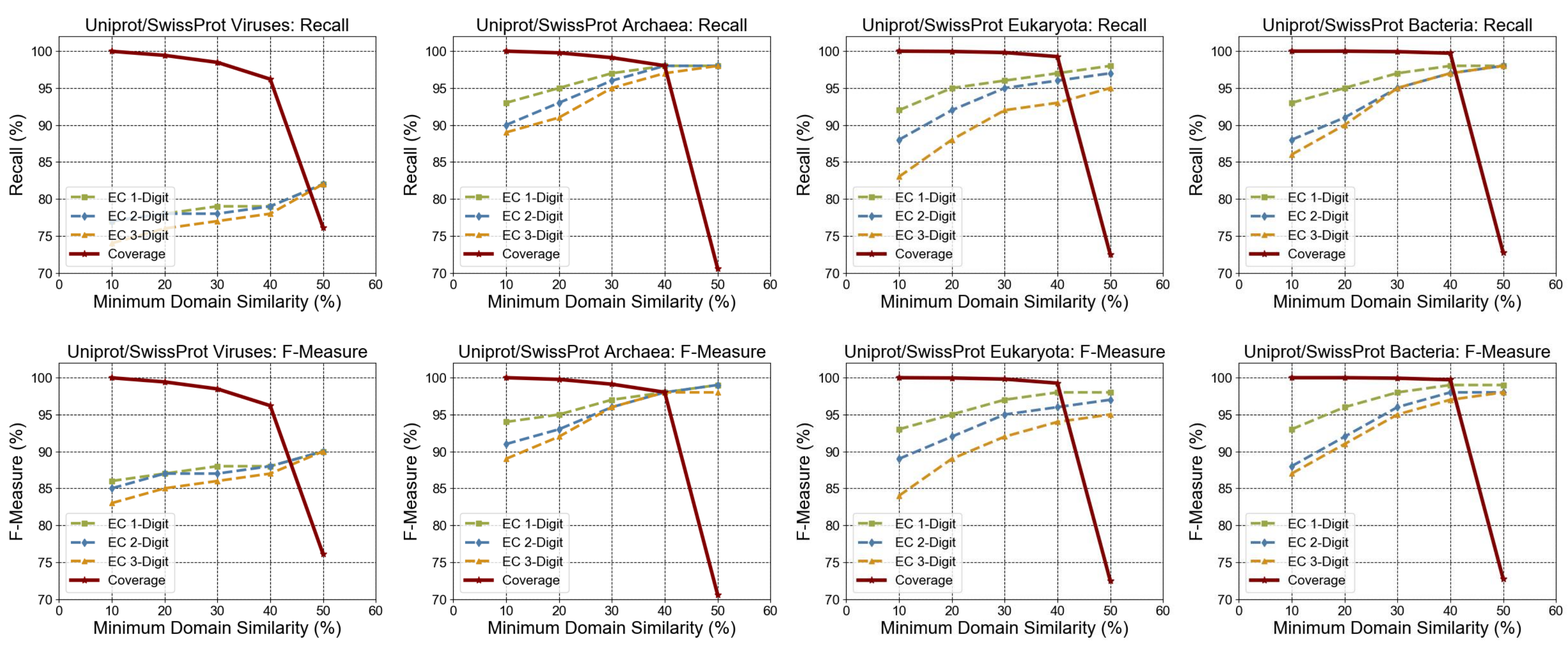
3(b)

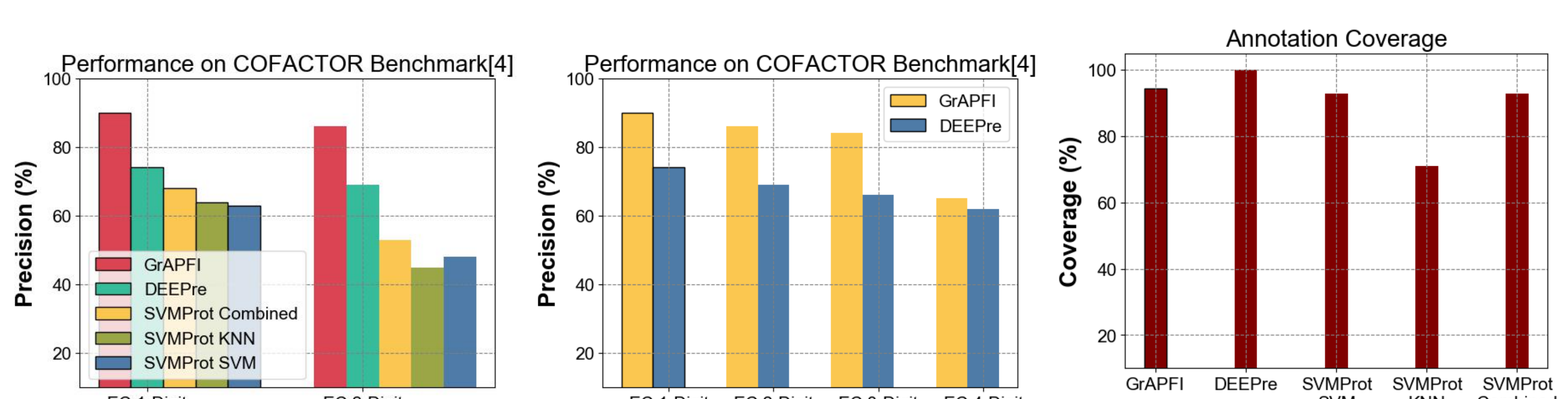
Fig-3(a) shows a query protein (yellow node) connected with its five neighbors. Fig-3(b) shows the result of label propagation applied on the graph in Fig-3(a).

Results

Dataset	# Nodes	# Edges	Average Degree	# Domains	Total EC
SwissProt Viruses	3208	478447	298.28	1031	150
SwissProt Archaea	10619	1168710	220.12	2499	727
SwissProt Eukaryota	55042	30753219	1117.45	6744	2832
SwissProt Bacteria	193429	409837148	4237.6	6480	2902







A comparative study is presented based on a benchmark of 318 proteins taken from COFACTOR[4]. After removing proteins that do not have ground truth EC annotation, we had 297 Enzymes to annotate.

Conclusion

GrAPFI is a novel graph based approach for automatic protein functional annotation. It utilizes a domain based graph representation of the UniProtKB/SwissProt protein database.

To evaluate the performance, leave one out cross validation is used on the graph built on four species from UniProt/SwissProt namely Viruses, Archaea, Eukaryota and Bacteria. Average precision, $Pr_{avg} = \frac{1}{M}\sum_{\forall p \in P} Pr_p$, recall, $Re_{avg} = \frac{1}{M}\sum_{\forall p \in P} Re_p$ and F-Measure, $F_{Measure} = \frac{2 \times Pr_{avg} \times Re_{avg}}{Pr_{avg} + Re_{avg}}$ are used as performance metrics. Our performance comparison shows *GrAPFI* outperforms other methods e.g. DEEPre[3], SVMProt[2]. The future work aims at using *GrAPFI* for protein Annotation with GO terms.

References

1. T. U. Consortium. Uniprot: a hub for protein information. Nucleic Acids Research, 2014.
2. Li, Ying Hong, et al. SVM-prot 2016: A web-server for machine learning prediction of protein functional families from sequence irrespective of similarity, 2016.
3. Li, Yu, et al. DEEPre: sequence-based enzyme EC number prediction by deep learning, 2017.
4.Chengxin Zhang, Peter L. Freddolino, and Yang Zhang. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. Nucleic Acids Research,2017.

Acknowledgement Bishnu Sarker is funded by an INRIA Cordi-S Doctoral Grant.