An automated pipeline for multi-species protein function prediction from the UniProt Knowledgebase

Matteo Re, Marco Mesiti, Giorgio Valentini *,
AnacletoLab – Computational Biology and Bioinformatics,
Department of Computer Science, University of Milan, Via Comelico 39, 20135 Milan, Italy
*To whom correspondence should be addressed: valentini@di.unimi.it

1. INTRODUCTION

A recent international challenge, CAFA (Critical Assessment of Functional Annotation of proteins), provided the unique opportunity of critically evaluate and compare state-of-the-art methods for automated protein function prediction (AFP) (1). As pointed out by the CAFA results, one of the key issues that characterizes the AFP problem is the integration of both heterogeneous experimental data and different prediction methods. These items pose serious computational problems because of the ever increasing rate at which bio-molecular data are made available in public databases and for the complexity of the the prediction tasks.
In this contribution we describe an automatic pipeline that addresses both these issues by exploiting data directly available for the UniProt KB. We applied the pipeline to the second and still ongoing CAFA 2 challenge.

2. METHODS

2.1 Extraction of data from the UniProt Knowledgebase.

Instead of collecting large amount of data directly from many databases, we extracted all the available annotations coming from EggNog, InterPro, Pfam, PRINTS, PROSITE, SMART, and SUPFAM included in the plain text file generated concurrently with each UniProt/Knowledgebase release. Besides to collect only data describing the targets at molecular level, we also collected functional keywords assigned to the targets by Swissprot curators. We finally extracted from the aforementioned plain text annotations file all the available Gene Ontology (GO) functional annotations based on the experimental evidence of each target.

2.2. Network construction and integration.

In order to predict the functional labels for the more than 100,000 CAFA 2 targets, we first constructed for each possible target/data-source pair a binary profile vector indicating the presence/absence of any feature considered in the external database (i.e. proteins domains available in Pfam). These profiles have been then used to compute a pairwise similarity score for each pair of targets according to similarity measures specific for each type of data. These values were used to construct connections between targets according to specific filtering policies to avoid noisy links between proteins. In this step of our pipeline we constructed multi-species networks to transfer functional annotations from well annotated species (i.e. *S.cerevisiae, A.thaliana, M.musculus, H.sapiens* and *E.coli*) to poorly annotated species with the aim to make the most effective usage of all the available annotations. Finally the 11 different networks constructed with different types of genomic and proteomic data have been integrated for each of the multi-species networks, giving rise to 8 different multi-species and multi-view data networks.

2.3 GO term prediction.

The AFP task is performed by applying to the resulting multi-species multi-view networks an improved version of a semi-supervised graph kernel-based ranking method, recently applied to AFP and drug repositioning problems (2,3). More precisely, we applied random walk graph kernels with different number of steps to explore the topology of the graph at different depths from each node/protein, and we merged the predictions with a test-and-select ensemble approach.

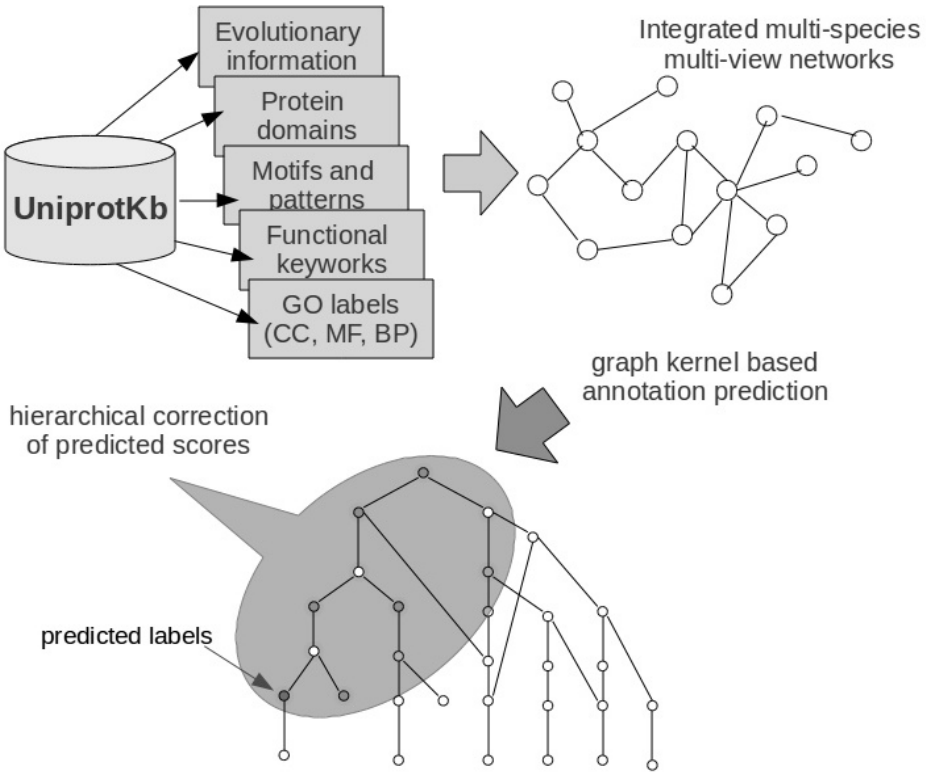

Figure 1: a simplified scheme of the UniprotKb pipeline.

In order to ensure the coherence of the obtained predictions with the true path rule governing the GO ontology, we performed a hierarchical correction of the obtained ranking scores using the method proposed in (4). The resulting predictions have been evaluated at per-protein, per-species and per-GO term level. An overview of our pipeline is presented in Figure 1.

3. CONCLUSIONS

On the basis of the results of the first CAFA edition we constructed a pipeline able to integrate diverse and complementary source of information available from the UniProt Knowledgebase. Our pipeline is fully automated and can be easily reused for novel releases of the UniProt Knowledge base and of the GO. The adopted learning strategy has been specifically tailored to promote an effective inter-species functional annotation transfer, and to allow an automated massive integration of different sources of data, adopting ensemble learning methods to combine multiple prediction algorithms. The pipeline has been realized using public and in house written Perl, R and C software libraries.

REFERENCES

1. Radivojac, P. et al. 2013. A large-scale evaluation of computational protein function prediction. Nature Methods, 10(3):221–227.
2. Mesiti, M., Re, M. and Valentini, G. 2012. A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks, IEEE ACM Transactions on Computational Biology and Bioinformatics, 9(6) pp. 1812-1818
3. Re, M. and Valentini, G. 2013. Network-based drug ranking and repositioning with respect to DrugBank therapeutic categories. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(6):1359–1371.
4. Cozzetto, D., Buchan, D., Bryson, K. and Jones, D. 2013. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC Bioinformatics, 14(Suppl 3:S1).