Can we go beyond sequence similarity to predict protein function ?

Paolo Fontana[1] , Tiziana Sanavia[2], Andrea Facchinetti[2], Enrico Lavezzo[3], Marco Falda[3], Duccio Cavalieri[1], Barbara Di Camillo[2], Stefano Toppo[3]*

[1] Istituto Agrario San Michele all'Adige Research and Innovation Centre, Foundation Edmund Mach, via E. Mach 1, I-38010, San Michele all'Adige (Trento), Italy
[2] Department of Information Engineering, University of Padova, via Gradenigo 6, I-35131, Padova, Italy
[3] Department of Molecular Medicine, University of Padova, v.le G. Colombo 3, I-35131, Padova, Italy

*To whom correspondence should be addressed: stefano.toppo@unipd.it

1. INTRODUCTION

Recent results of CAFA experiment have given us the unique opportunity to rethink the strategy used so far to make function prediction. Our approach, Argot2 (1,2), is mainly based on evaluating sequence similarities provided by BLAST and HMMER searches vs Uniprot (3) and PFAM (4) databases respectively. After extracting from GOA (5) database the Gene Ontology (GO) annotations of sequence hits and empirically weighting their e-value scores, we let Argot2 algorithm to make some considerations about how these annotations distribute and cluster in the graph by means of semantic similarity. Looking at the obtained results we have realized that, to some extent, PFAM searches have been used improperly. What is most surprising is that, rather than extending GO annotations, the PFAM results have simply reinforced those hits already found by BLAST. It is as if we have some GO terms weighted twice and consequently their final scores have been overestimated, while the improperly added terms were simply trimmed because belonging to poorly weighted paths. The take home message is that we are still bound to sequence similarity paradigm and we are still entrapped in this idea. CAFA has confirmed that top performing methods still rely on this. After all, what else can we do if the only thing we have is the amino acid sequence of the target to predict ? Can we go beyond sequence similarity to predict protein function?

Exploring weak signals of similarity is dangerous as false positive hits are the majority. On the other hand, we know that function, as well as the protein fold, can be conserved despite a great divergence in amino acid sequence. For these reasons we are figuring out how this peculiarity can be exploited. We are planning to look the other way round i.e. how function is distributed in GOA using the semantic similarity, in order to investigate if there is a correlation with sequence similarity or exceptions to take care of.

We are exploring the distribution of GO terms in GOA database and the possible correlation of their homogeneity with their sequence similarity (see Fig. 1). Eukaryotic proteins of same length have been extracted from Uniprot and clustered using CD-HIT (6) at 100%, 80%, 60%, 40%, and 20% sequence identity (Uniprot_100, Uniprot_80, Uniprot_60, Uniprot_40, Uniprot_20). Real median of sequence identity of the obtained groups has been recalculated and reported in the x-axis of Fig 1. In order to understand how domains and protein architectures are built looking at their associated functions, the same analysis has been performed for PFAM and TIGRFAMs (7) (the “equivalogs” groups) databases. The sequence identity has been calculated and reported in the x-axis of Fig.1 for TIGRFAMs, PFAM-A seed models (PFAM seed), and Protein architectures extracted from PFAM-A models (i.e. proteins having the same domains in the same order and number - PFAM Arch.). For each group of proteins the level of functional homogeneity has been assessed using the semantic similarity based on Lin's formula (8) and the percentage of these "homogeneous" groups has been reported in the y-axis of Fig. 1.

The present scenario of how GO terms are spread in the protein databanks seems to demonstrate that function is conserved up to 40%-50% sequence identity but dramatically drops when moving to 30% or lower values. Indeed, this result may be biased given that automatic annotations of IEA terms in GOA are mainly based on sequence similarity. So, if on the one hand it is not surprising to see that function conservation drops at low levels of identities, on the other hand it is interesting to observe that the majority of groups of proteins are still semantically homogeneous.

The final outcome, though preliminary, will help us to design a better solution in the future Argot3 algorithm hoping to have a more comprehensive view to automate functional inference even for those difficult cases that do not share high sequence similarities with known proteins.

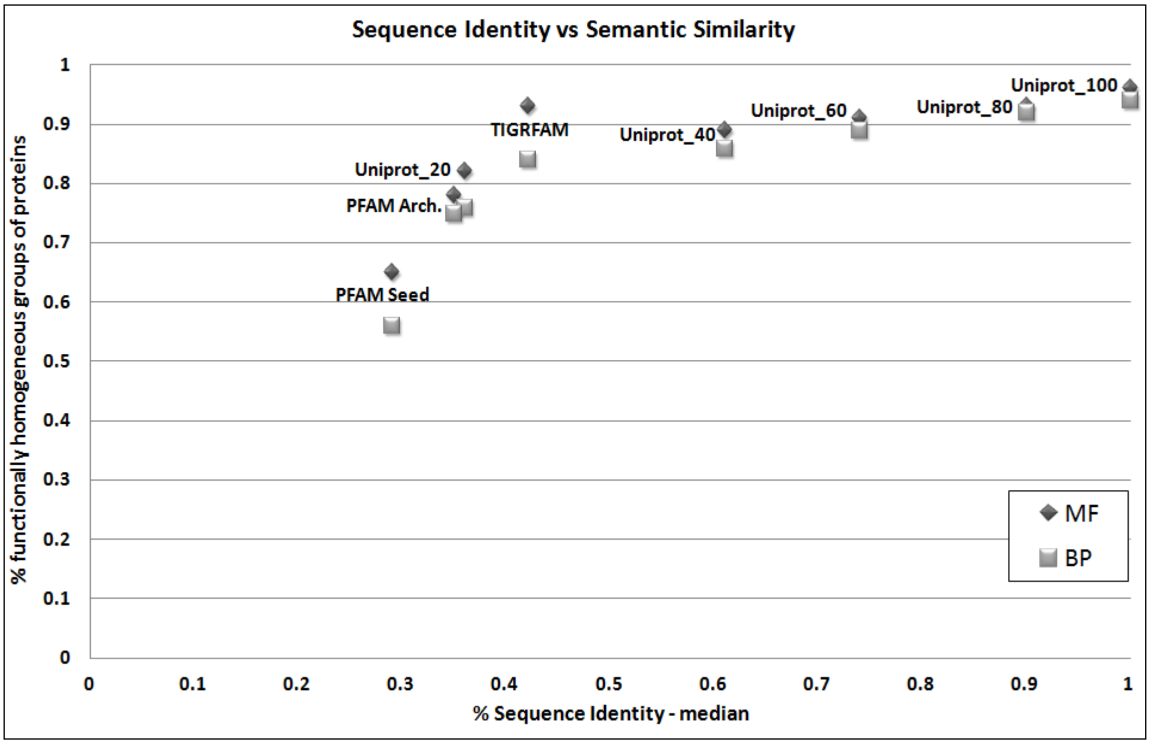

Fig. 1: percentage of sequence identity vs percentage of semantically homogeneous functions calculated for groups of proteins in PFAM, TIGRFAMs, and Uniprot databases. The data are reported for both Molecular Function (MF) and Biological Process (BP). See text for details

2. REFERENCES

1. Falda M., S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo, A. Facchinetti, A. Cilia, R. Velasco, and P. Fontana, *Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms.* Bmc Bioinformatics, 2012. **13**(4).
2. Fontana P., A. Cestaro, R. Velasco, E. Formentin, and S. Toppo, *Rapid Annotation of Anonymous Sequences from Genome Projects Using Semantic Similarities and a Weighting Scheme in Gene Ontology*. Plos One, 2009. **4**(2).
3. Bairoch A., R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33**(Database issue): p. D154-9.
4. Punta M., P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, et al., *The Pfam protein families database.* Nucleic Acids Research, 2012. **40**(D1): p. D290-D301.
5. Dimmer E.C., R.P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M.J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, et al., *The UniProt-GO Annotation database in 2011.* Nucleic Acids Res, 2012. **40**(Database issue): p. D565-70.
6. Huang Y., B. Niu, Y. Gao, L. Fu, and W. Li, *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-2.
7. Haft D.H., J.D. Selengut, and O. White, *The TIGRFAMs database of protein families.* Nucleic Acids Res, 2003. **31**(1): p. 371-3.
8. Lin D., *An Information-Theoretic Definition of Similarity*, in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98)*. 1998, Morgan Kaufmann Publishers Inc. p. 296-304.