BAR 3.0: going beyond protein function annotation

Giuseppe Profiti, Pier Luigi Martelli* and Rita Casadio
Biocomputing Group, BiGeA, University of Bologna, Bologna, 40126, Italy
*To whom correspondence should be addressed: gigi@biocomp.unibo.it

1. INTRODUCTION

BAR 3.0 (1) updates our server BAR (Bologna Annotation Resource) (2,3,4) for predicting protein structural and functional features from sequence. This new version is built on a bigger database and features new query capabilities and information presented to the user. The core of BAR 3.0 is a graph-based clustering procedure of UniProtKB sequences, following strict pairwise similarity criteria (sequence identity ≥40% with alignment coverage ≥90%). Each cluster contains the available annotation downloaded from UniProtKB, GO, PFAM and PDB. After statistical validation, GO terms and PFAM domains become cluster-specific and annotate new sequences entering the cluster according to the similarity criteria.
BAR 3.0 includes 28,869,663 sequences in 1,361,773 clusters, of which 22.2% (22,241,661 sequences) and 47.4% (24,555,055 sequences) have at least one validated GO term and one PFAM domain, respectively. 1.4% of the clusters (36% of all sequences) include PDB structures and the cluster is associated to a Hidden Markov Model that allows building template-target alignment suitable for structural modelling. Singleton sequences are a total of 3,399,02. BAR 3.0 offers an improved search interface, allowing queries by UniProtKB-accession, Fasta sequence, GO-term, PFAM-domain, organism, PDB and ligand/s.

2. EVALUATION

When evaluated on the CAFA2 targets, BAR 3.0 largely outperforms our previous version. We benchmarked BAR 3.0, simulating an in-house CAFA2 experiments. The benchmark dataset of CAFA2 was predicted with BAR 3.0_{CAFA2}, containing only UniProtKB sequences and annotations released before January 2014. The predictions were evaluated on the experimental annotations acquired by the benchmark sequences till September 2014. We compared the performance of BAR $3.0_{CAFA2,}$ to the results of BAR++ and the best scoring methods in each sub-ontology as reported in the CAFA2 assessment (5). In the first case, the new version greatly outperform our previous method. When compared to the best scoring methods, BAR 3.0_{CAFA2} consistently performs well.
It appears that BAR 3.0_{CAFA2} outperforms the previous version BAR++ in all the sub-ontologies, reaching F1-scores as high as 0.54, 0.35 and 0.42 for Molecular Function (MF), Biological Process (BP) and Cellular Component (CC), respectively. Predictions from BAR 3.0 have been submitted to CAFA3.
Besides that, BAR 3.0 features also cross-cluster links derived from IntAct protein-protein interactions and PDB protein complexes. That approach may be useful to gain further insights that go beyond the assignment of protein functions.

3. REFERENCES

1. Profiti, G, Martelli, P.L., Casadio, R. 2017. The Bologna Annotation Resource (BAR 3.0): improving protein functional annotation. *NAR Web Server issue* (in press).
2. Bartoli, L., Montanucci, L., Fronza, R., Martelli, P.L., Fariselli, P., Carota, L., Donvito, G., Maggi, G.P., and Casadio, R. 2009. The Bologna annotation resource: a non hierarchical method

for the functional and structural annotation of protein sequences relying on a comparative large-scale genome analysis. *J Proteome Res*., 8:4362-4371.
3. Piovesan, D., Martelli, P.L., Fariselli, P., Zauli, A., Rossi, I., and Casadio, R. 2011. BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences. *Nucleic Acids Res.,* 39:W197-202.
4. Piovesan, D., Martelli, P. L. , Fariselli, P. , Profiti, G., Zauli, A., Rossi, I., Casadio, R. 2013. How to inherit statistically validated annotation within BAR+ protein clusters, *BMC Bioinformatics*, vol. 14, no. Suppl 3, p. S4.
5. Jiang, Y., Oron, R.T., Clark, T.W., Bankapur, R.A., D'Andrea, D., Lepore, R., Funk, S.C., Kahanda, I., Verspoor, M.K., Ben-Hur, A. et al. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology* 17:184.