Published online 27 April 2017 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* ***W285–W290***
doi: 10.1093/nar/gkx330

The Bologna Annotation Resource (BAR 3.0): improving protein functional annotation

Giuseppe Profiti, Pier Luigi Martelli* and Rita Casadio

Biocomputing Group, BiGeA/CIG, 'Luigi Galvani' Interdepartmental Center for Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Bologna 40126, Italy

Received January 31, 2017; Revised April 10, 2017; Editorial Decision April 12, 2017; Accepted April 18, 2017

ABSTRACT

BAR 3.0 updates our server BAR (Bologna Annotation Resource) for predicting protein structural and functional features from sequence. We increase data volume, query capabilities and information conveyed 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 are cluster-specific and annotate new sequences entering the cluster after satisfying similarity constraints. 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 modeling. Some other 3 399 026 sequences are singletons. BAR 3.0 offers an improved search interface, allowing queries by UniProtKB-accession, Fasta sequence, GO-term, PFAM-domain, organism, PDB and ligand/s. When evaluated on the CAFA2 targets, BAR 3.0 largely outperforms our previous version and scores among state-of-the-art methods. BAR 3.0 is publicly available and accessible at http://bar.biocomp.unibo.it/bar3.

INTRODUCTION

Sequencing technologies are producing a deluge of biosequences, including protein sequences stored into public databases. Identifying the functional and structural features of proteomes in an experimental way is a time consuming and slow process, as compared to the pace of data production. While waiting for experimental confirmation of protein existence and characterization, bioinformatics tools are widely used to infer structural and functional features. Routinely, different inference algorithms extract information from the annotation of proteins already characterized at the structural and/or functional levels and transfer them to new sequences based on different similarity criteria. As a result, some 50 000 000 protein sequences are still labeled 'predicted' in the last release (2017_01) of UniprotKB (http://www.uniprot.org/statistics/TrEMBL). We developed a system (the Bologna Annotation Resource (BAR, 1–3)) that allows transfer of statistically validated annotation among sequences that enter a cluster after constraining the alignment with stringent similarity criteria. BAR (including different versions of the same system, e.g. BAR+ (2); BAR++) is based on extensive pairwise comparison of the protein sequences included in UniProtKB (4). Sequences sharing at least 40% sequence identity, over at least 90% of the alignment length are grouped together in the same cluster. The over-representation of gene ontology (GO) (5) and PFAM (6) terms, annotating sequences in the same cluster, is statistically validated with a Bonferroni-corrected Fisher test. Validated annotations are then propagated to all the sequences of the cluster. Similarly, PDB structures associated to proteins in a cluster allow propagating structural information to all the sequences in the cluster. BAR+ predictions achieved a good performance in the Critical Assessment of Protein Annotation (CAFA) experiment (7), which compared over than 50 state-of-the-art methods for protein function prediction. Subsequent assessments highlighted the need for an update of the underlying data (8). Here, we present a new version (BAR 3.0) that implements the updates and adds new features to the web server. Improvements with respect to the previous versions include quality of the protein annotation predictions, information returned and user experience. The quality of the predictions was tested on the CAFA2 dataset (8), allowing a comparison with the state-of-the-art methods in protein function prediction. Interestingly, the BAR present version consistently well performs in all the branches of GO. The new web server includes not only GO term, PFAM and PDB annotations for the clusters, but also KEGG pathways (9) and

*To whom correspondence should be addressed. Tel: +39 051 209 4005; Fax: +39 051 209 4005; Email: gigi@biocomp.unibo.it

cross-cluster links, based on IntAct (10) protein–protein interactions and physical interactions due to protein complexes. By this, end users can explore single clusters and network of clusters connected by interacting protein sequences. Furthermore, users can also submit queries using search terms, like PDB, ligands, GO terms and organisms. These queries return a list of all the BAR 3.0 clusters associated to the user's input. Result page exploits modern technologies such as HTML5 and responsive web design for better displaying on different devices, while jQuery and JSON allow filtering of annotations and download of all or some of the data associated to a cluster, in both human and machine-readable formats.

MATERIALS AND METHODS

Databases

BAR3.0 was obtained by clustering 32 268 689 sequences out of the 2016_05 release of the UniProtKB repository (4). Fragments and sequences shorter than 40 residues are not considered. GO terms (5), PFAM (6), KEGG (9) and interaction annotations from IntAct (10) derive from the same UniProtKB release. PDB (11) chains are associated to the corresponding UniProtKB entries using SIFTS (12).

UniRef90 (13) clusters (release 2016_05, collecting UniProt sequences that share more than 90% identity) were exploited to populate the BAR 3.0 clusters, allowing to save alignment time. This procedure was possible thanks to the introduction (in August 2013) of length constraints inside UniRef90 clusters.

By construction, different UniRef100 clusters are aggregated into the same UniRef90 cluster when their representative sequences share more than 90% identity and their lengths are at least 80% of the longest one. Since the same length constraint does not hold in the UniRef100 clusters, we discarded from the UniRef90 clusters all the sequences violating the length constraint.

Graph building, clustering and statistical validation of annotations

BAR 3.0 building involves the full UniProtKB database and the following steps. (i) All the SwissProt sequences up to release 2016_05 and the TrEMBL sequences up to release 2013_01 are compared with BLAST to search for pairs of proteins sharing a sequence identity (SI) $\geq$ 40%, on an alignment coverage (COV) $\geq$ 90%. COV is defined as the ratio between the number of overlapping positions and the alignment length. AlignBucket algorithm (14) is used to speed-up the alignment procedure exploiting the constraint on COV. (ii) A graph is built by connecting sequence pairs that fulfill both identity and coverage constraints. (iii) Clusters are obtained by isolating the connected components of the graph. (iv) UniRef90 clusters are mapped to BAR 3.0 clusters, allowing to include the remaining TrEMBL sequences (up top the release 2016_05). (v) Each cluster is annotated by collecting GO annotations, PFAM domains and PDB structures of its members. To assess whether GO and Pfam terms are significant in a cluster, we computed the over-representation *P*-values with the Fisher's exact test and, given the multiplicity of the terms, we applied the Bonferroni correction (15). The significance level on the corrected *P*-values was set to 0.01. Non-protein ligands present in the PDB files associated to the clusters were also collected.

The system will be updated, at least yearly by (i) adding new sequences to BAR 3.0 and reshaping clusters accordingly, and (ii) downloading new annotations from UniProtKB and performing statistical validation. The former operation requires a variable time, depending on the number of new sequences to be aligned. However, AlignBucket (14) and the use of UniRef90 clusters allow to speed up the computation. The update of annotations and the validation procedure requires just few hours.

Cluster-HMMs

When structural information from the PDB is present within a cluster, a profile hidden Markov model (HMM) is computed (Cluster-HMM) to facilitate the sequence alignment of the proteins in the cluster with their structural templates. The HMM building involves the following steps. (i) When different templates are present in a cluster, their structural alignment is computed with MUSTANG (16); (ii) For each template structure in the cluster, a multiple alignment of similar sequences (with SI $\geq$ 40% and COV $\geq$ 90%) is computed with Clustal Omega (17) and in case of multiple overlapping templates, a comprehensive multiple sequence alignment is built guided by the structural alignment; (iii) A HMM is trained on the multiple sequence alignment with HMMER3 (18). Cluster-HMMs allow aligning all the sequences in a cluster (even if distantly related) to the corresponding template/s. The Viterbi decoding implemented in HMMER3 retrieves the target-to-template alignments in PIR format to be fed to external modeling programs.

Web server implementation

The front-end for the Web server follows the Model-View-Controller paradigm, and it is optimized to work with all common Web browsers. The web server is implemented as a REST web service and exploits technologies like Ajax (jQuery—https://jquery.com—and Bootstrap—https://getbootstrap.com/—libraries), JSON and a queuing service (using Sun Grid Engine). For aligning new sequences against the BAR 3.0 dataset, the server runs BLAST after exploiting the speed-up techniques implemented in AlignBucket (14). The alignment runs asynchronously: after submitting the query, the server displays a bookmarkable page reporting the status of the job. Such status could be 'queued', 'running' or 'completed' and, at the end of the alignment procedure, a link to the results is provided to the user. The web server is open to all and it does not require registration.

Evaluation procedure

The second Critical Assessment of Automated protein Function Annotation (CAFA2) (8) is an international experiment run in 2013 to evaluate the performance of large-scale algorithms adopted for protein function prediction. The CAFA2 benchmark dataset consists of 3649

Table 1. Distribution of sequences in clusters and singletons with their annotations

	In clusters	In singletons
# Sequences	28 869 663	3 399 026
From SwissProt	519 015	17 478
From TrEMBL	28 350 648	3 381 548
# Sequences with experimental GO annotations	82 672	6 092
From SwissProt	57 391	3 684
From TrEMBL	25 281	2 408
# Sequences with GO annotations	20 556 103	1 506 125
From SwissProt	494 047	14 277
From TrEMBL	20 062 056	1 491 848
# Sequences with PFAM annotation	23 263 014	1 509 339
From SwissProt	487 946	12 111
From TrEMBL	22 775 068	1 497 228
# Sequences with PDB	35 660	1 185

As defined by gene ontology Consortium, experimental GO terms are those associated to evidence codes EXP, IDA, IPI, IMP, IGI, IEP (http://geneontology.org/page/guide-go-evidence-codes)

Table 2. Statistics of inherited annotations in BAR 3.0

	# Clusters	# Sequences included	# Sequences inheriting new annotations
Total number of clusters	1 361 773		
With any validated annotation	674 463	25 448 877	16 430 135
With validated GO terms	302 159	22 241 661	15 938 828
With validated PFAM	645 502	24 555 055	16 105 082
With at least one PDB	19 015	11 653 046	11 626 119

sequences that lacked any GO annotation (in any sub-ontology) in UniProtKB release 2013_12 and acquired experimental GO annotations (at least in one sub-ontology) in release 2014_10. The considered evidence codes are 'Inferred from experiment' (EXP), 'Inferred from direct assay' (IDA), 'Inferred from mutant phenotype' (IMP), 'Inferred from genetic interaction' (IGI), 'Inferred from expression pattern' (IEP), 'Traceable author statement' (TAS) and 'Inferred by curator' (IC). These evidence codes are used for CAFA evaluation only, given that they are slightly different from the ones considered as experimental by GO (http://geneontology.org/page/guide-go-evidence-codes). For fairly evaluating the performances of the BAR 3.0 system on the CAFA2 targets, we implemented a specific version of BAR 3.0 (BAR 3.0_{CAFA2}) containing sequences and annotations only from UniProtKB release 2013_01. The benchmark sequences were aligned to the clusters of BAR 3.0_{CAFA2}. The corresponding validated annotations were collected and compared to the experimental GO annotations present in version 2014_10 of UniProfKB. For each protein, we computed the precision and recall by dividing the number of correctly predicted terms by the number of predicted terms and the number of terms to be predicted, respectively. All the ancestors of a given GO term are considered in the computation. Overall precision and recall are computed as the averages of the per-protein values over the benchmark dataset. F1 value is computed as the harmonic average of overall precision and recall.

RESULTS

BAR 3.0 statistics

BAR 3.0 contains 28 869 663 sequences grouped in 1 361 773 clusters, along with 3 399 026 isolated sequences, called singleton. About 3% of the SwissProt sequences in BAR 3.0 are singletons, while the remaining 97% belongs to clusters, conveying experimentally validated annotations to be used in the statistical validation of GO terms and PFAM domains. Details about the distribution of sequences and annotations among clusters and singletons are presented in Table 1.

Statistical validation of annotations in clusters leads to 674 463 clusters having at least one validated annotation, comprising a total of 25 448 877 (88% of clustered sequences, 79% of all BAR 3.0 sequences). Among clusters, 15 195 (containing 38.9% of clustered sequences) are endowed with at least one statistically validated GO term, at least one statistically validated PFAM and at least one PDB structure, leading to 11 206 902 sequences inheriting a full annotation they did not have before. Other 258 814 clusters are endowed with statistically validated GO terms and PFAM, resulting in 4 425 042 sequences inheriting a new annotation. Details about clusters with statistically validated annotations and sequences inheriting new annotations are listed in Table 2 and shown in Figure 1A. The distribution of annotation among singletons is quite different: about 43% of them do not have any type of annotation, as shown in Figure 1B. In any case when a new sequence matches a singleton a new cluster is formed and by this it inherits what is available or it brings in what it carries along.

Input

The server can process different types of inputs. As in the previous versions of BAR, the user can provide a protein sequence in the form of a UniprotKB accession or FASTA. If the sequence is not included in the BAR 3.0 clusters, the sequence is aligned with Blast against the BAR 3.0 sequences of similar length as determined with AlignBucket. The sequence is associated with a cluster or a singleton if the alignment fulfills the SI (≥40%) and COV (≥90%) constraints.

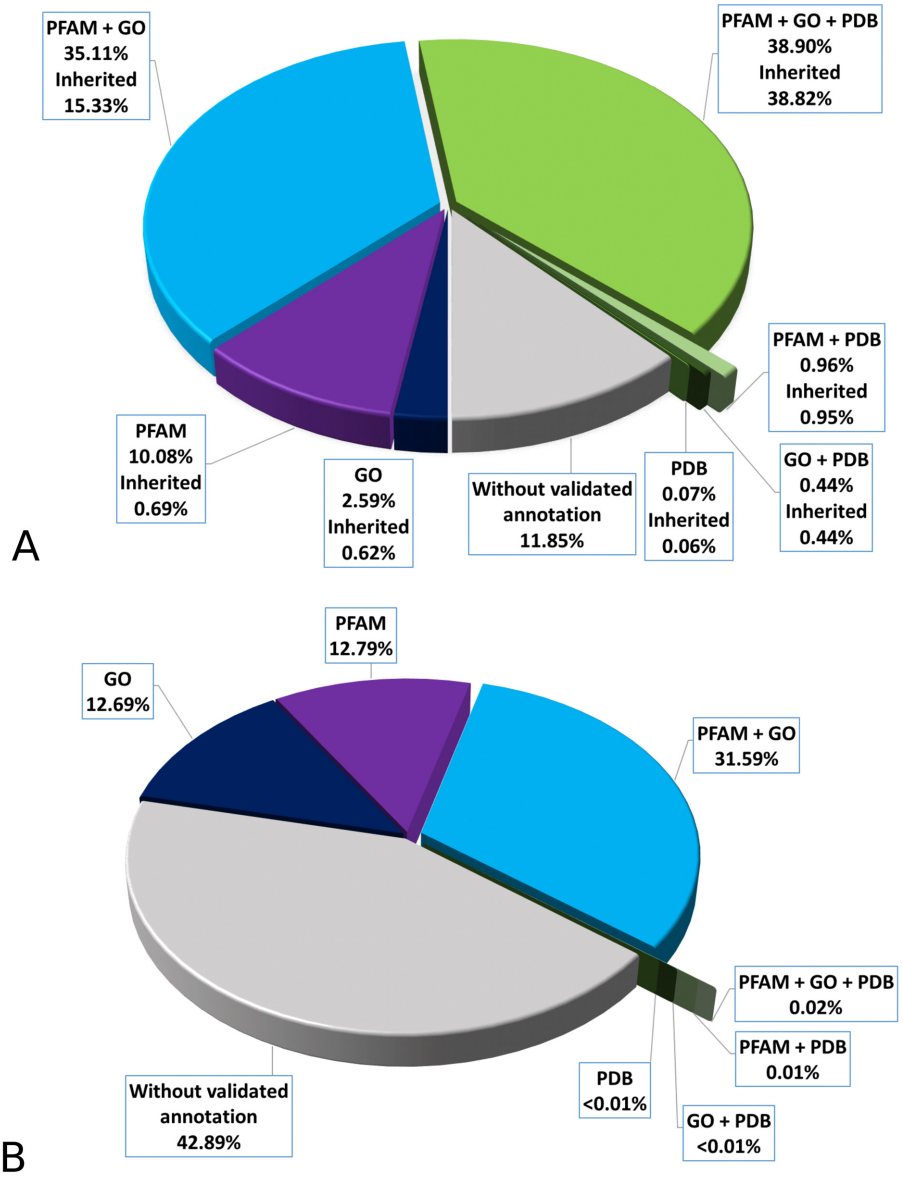

Figure 1. Distribution of statistically validated annotation among sequences in BAR 3.0. On the first chart (**A**), percentage of clustered sequences per statistically validated annotation type. First value refers to percentage of sequences falling in clusters with that annotation. Second value refers to percentage of sequences inheriting at least one annotation they did not have in UniProt. On the second chart (**B**), percentage of singleton sequences by annotation type.

In the BAR 3.0 web server, we also introduced new queries: the user can search for the lists of clusters associated with specific: (i) validated GO term, (ii) validated PFAM annotation, (iii) PDB code, (iv) ligand code or (v) NCBI organism identifier.

Output

General features. For GO term, PFAM, PDB, organism and ligand queries, the result consists of a list of clusters matching the query term. For searches using UniprotKB and FASTA sequences, only the matching cluster is returned (if any).

Each cluster is described in a page containing the following statistics: (i) number of sequences in the cluster, (ii) their average length and (iii) number of organisms for each domain that are represented in the cluster. Details about the distribution by organism of sequences included in the cluster can be downloaded as a text file.

PDB structures associated to sequences in the cluster, if any, along with their specific chain, are also listed. When a 3D structure is present, a link to download the cluster HMM is provided. When the user enters BAR 3.0 with a query sequence, if templates are available, the cluster HMM is used to compile a PIR alignment that can be downloaded to be fed in comparative modeling programs.

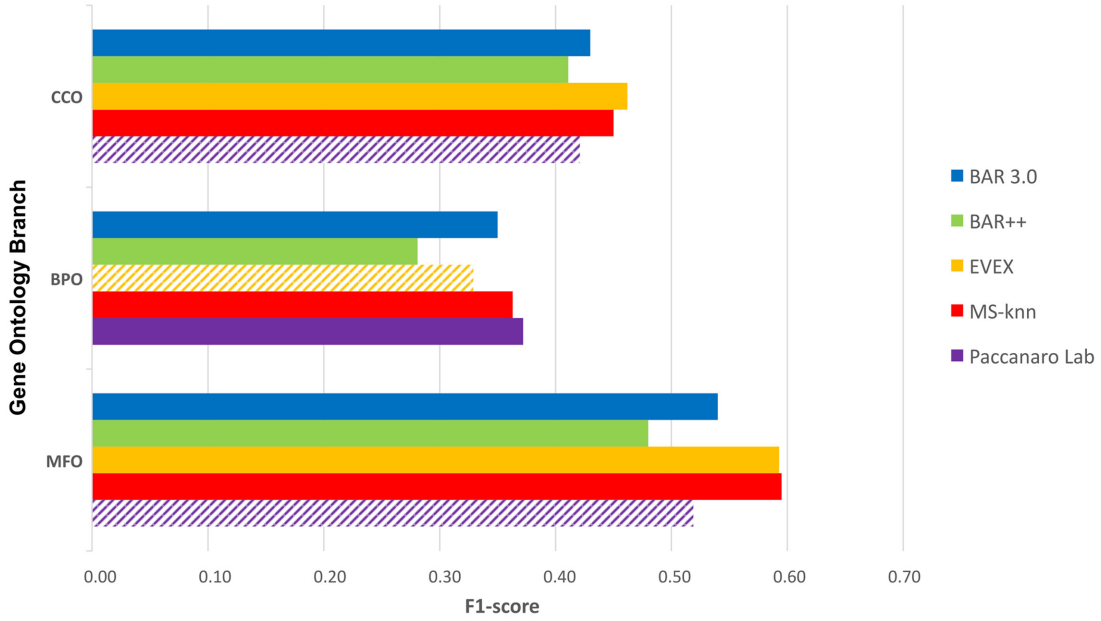

Figure 2. BAR 3.0 benchmarking toward the three best performing methods in CAFA2. Besides BAR 3.0, values are as reported in the assessment (8). F1-score is evaluated as the harmonic mean of precision and recall, where precision is the ratio of correct annotation over all the predicted annotation and recall is the ratio of correct annotation over the real annotation. Other methods shown are MS-knn (21), EVEX (22) and the one from Paccanaro Lab. Dashed bars show the upper limits of the performance when exact values are not available. The CAFA2 paper (8) does not list the exact performance of the Paccanaro Lab in CC and BP sub-ontologies and of EVEX on the MF sub-ontology, since the methods did not classify among the best 10 methods. The performance reported with dashed bars correspond to the 10th classified method in the corresponding sub-ontology.

The lists of GO and PFAM validated annotations are presented in dedicated sections. Validation of a PFAM annotation is extended to the GO terms associated to it in the InterPro2GO mapping (19). For each GO term, we display its depth in the ontology, i.e. the number of ancestors to be visited while looking at the shortest path up to the ontology root (Molecular Function (MF), Biological Process (BP) or Cellular Component (CC)). Non-validated annotations are also available. The tables can be sorted by p-value, depth or validation state, to facilitate the retrieval of more relevant terms.

Moreover in BAR 3.0 we added the 'KEGG Pathways' section, listing all the KEGG (9) pathways associated to sequences in the cluster, along with the supporting UniProt accessions.

All the identifiers point to their respective resource databases URLs (i.e. UniProt for sequences, gene ontology for GO terms and so on).

Intercluster links. A novelty of BAR 3.0 is the information about the links among different clusters: two clusters are linked when they contain sequences that are either connected in the IntAct protein–protein interaction network or part of a PDB complex.

The information on protein–protein interactions are presented in a table listing the linked clusters. For each linked cluster, BAR 3.0 reports: (i) the number of IntAct interactions between the clusters, along with the UniProt accession of the interacting proteins, (ii) the IntAct interactions involving the query sequence, (iii) the presence in the linked cluster of sequences from the same organism as the query sequence, for suggesting new possible interactors.

Similarly, the 'PDB Complexes' table reports link between BAR 3.0 clusters containing different chains of the same PDB complex. The UniprotKB accessions and PDB chains are shown for any interacting pair.

DISCUSSION

Evaluation on the CAFA2 targets

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 (see 'Materials and Methods' section for details). Figure 2 shows the performance of BAR 3.0_{CAFA2} as compared to BAR++ and to the best scoring methods in each sub-ontology, as reported in the CAFA2 assessment (8).

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 MF, BP and CC, respectively. BAR++ predictions submitted to CAFA2 for CC included also an ensemble method exploiting subcellular localization predictors (our BaCelLo (20)) when the annotation was not available from the cluster. These score are at the state-of-the-art: BAR 3.0_{CAFA2} scores are among the first 10 best for both BP and MF.

To better analyze the performance of BAR 3.0_{CAFA2}, we compared the F1-score with those of the three best performing methods in each GO branch, i.e. MS-kNN (21) for MF, EVEX (22) for CC and the Paccanaro Lab method for BP. As shown in Figure 2, even if not reaching the best score,

BAR 3.0_{CAFA2} has a consistent behavior across the different sub-ontologies. The only method performing consistently better than BAR 3.0_{CAFA2} is MS-kNN (21). Besides sequence similarity, it also relies on protein–protein interactions and gene expression data. BAR 3.0, as detailed in the paper, only includes in the present version sequence similarity and shows protein–protein interactions in the output, but it does not use such information for protein function prediction. In future versions, one possibility is to integrate our input with additional information.

The full version of BAR 3.0 will be benchmarked on the next CAFA experiment, starting February 2017.

CONCLUSION

In this paper, we presented a new version of the BAR for protein function computational annotation. The database was expanded to include new sequences and more information, like ligands and organism, allowing new queries with respect to the previous versions.

BAR 3.0 annotation was tested against the CAFA2 experiment dataset, producing competitive results.

Besides protein function annotation, the BAR 3.0 web server could be now used to investigate cross-cluster connections, like PDB structures with chains falling in different BAR clusters and checking for clusters having a specific validated GO term or PFAM domain.

BAR 3.0 results will be integrated and exploited for new services similarly to what done with previous versions (23,24).

ACKNOWLEDGEMENTS

G.P. thanks ELIXIR-IIB and ELIXIR Europe for supporting his research.

FUNDING

Italian Ministry of Education, University and Research [PRIN 2010–2011 project 20108XYHJS to P.L.M., PON projects PON01_02249, PAN Lab PONa3_00166 to R.C., P.L.M]; European Union RTD Framework Program [COST BMBS Action TD1101, Action BM1405 to R.C.]; University of Bologna [FARB 2012 to R.C.]. Funding for open access charge: University of Bologna [R.F.O. 2016 to P.L.M.].

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. 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–W202.
3. Piovesan,D., Martelli,P. L., Fariselli,P., Profiti,G., Zauli,A., Rossi,I. and Casadio,R. (2013) How to inherit statistically validated annotation within BAR+ protein clusters, *BMC Bioinformatics*, **14**, S4.
4. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
5. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
6. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
7. Radivojac,P., Clark,W.T., Oron,T.R., Schnoes,A.M., Wittkop,T., Sokolov,A., Graim,K., Funk,C., Verspoor,K., Ben-Hur,A. *et al.* (2013). A large-scale evaluation of computational protein function prediction. *Nat Meth* **10**, 221–227.
8. 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 Biol.* **17**, 184.
9. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361.
10. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363.
11. Rose,P.W., Prlić,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J. *et al.* (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.
12. Velankar,S., Dana,J.M., Jacobsen,J., van Ginkel,G., Gane,P., Luo,J., Oldfield,T., O'Donovan,C., Martin,M.J. and Kleywegt,G. (2013). SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.* **41**, D483–D489.
13. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B. and Wu,C.H., (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932.
14. Profiti,G., Fariselli,P. and Casadio,R. (2015) AlignBucket: a tool to speed up 'all-against-all' protein sequence alignments optimizing length constraints. *Bioinformatics*, **31**, 3841–3843.
15. Noble,W.S. (2009) How does multiple testing correction work? *Nat. Biotechnol.*, **27**, 1135–1137.
16. Konagurthu,A.S., Whisstock,J.C., Stuckey,P. J. and Lesk,A.L. (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.
17. Sievers,F., Wilm,A., Dineen,D., Gibson,T. J., Karplus,K., Li,W. and Higgins,D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
18. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
19. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
20. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006). BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* **22**, e408–e416.
21. Lan,L., Djuric,N., Guo,Y. and Vucetic,S. (2013) MS-kNN: protein function prediction by integrating multiple data sources. *BMC Bioinformatics*, **14**, S8.
22. Van Landeghem,S., Hakala,K., Rönnqvist,S., Salakoski,T., Van de Peer,Y. and Ginter,F. (2012) Exploring biomolecular literature with EVEX: connecting genes through events, homology, and indirect associations. *Adv. Bioinformatics.*, **2012**, 582765.
23. Piovesan,D., Profiti,G., Martelli,P. L. and Casadio,R. (2012) The human 'magnesome': detecting magnesium binding sites on human proteins. *BMC Bioinformatics*. **13**, S10.
24. Piovesan,D., Profiti,G., Martelli,P.L., Fariselli,P., Fontanesi,L. and Casadio,R. (2013) SUS-BAR: a database of pig proteins with statistically validated structural and functional annotation. *Database*, **2013**, bat065.