BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences

Damiano Piovesan¹, Pier Luigi Martelli¹, Piero Fariselli², Andrea Zauli³, Ivan Rossi³ and Rita Casadio¹,*

¹Department of Biology, Bologna Biocomputing Group, Bologna Computational Biology Network, ²Department of Computer Science, University of Bologna, Bologna and ³BioDec srl, Bologna, Italy

Received February 4, 2011; Revised April 4, 2011; Accepted April 13, 2011

ABSTRACT

We introduce BAR-PLUS (BAR⁺), a web server for functional and structural annotation of protein sequences. BAR⁺ is based on a large-scale genome cross comparison and a non-hierarchical clustering procedure characterized by a metric that ensures a reliable transfer of features within clusters. In this version, the method takes advantage of a large-scale pairwise sequence comparison of 13 495 736 protein chains also including 988 complete proteomes. Available sequence annotation is derived from UniProtKB, GO, Pfam and PDB. When PDB templates are present within a cluster (with or without their SCOP classification), profile Hidden Markov Models (HMMs) are computed on the basis of sequence to structure alignment and are cluster-associated (Cluster-HMM). Therefrom, a library of 10 858 HMMs is made available for aligning even distantly related sequences for structural modelling. The server also provides pairwise query sequence–structural target alignments computed from the correspondent Cluster-HMM. BAR⁺ in its present version allows three main categories of annotation: PDB [with or without SCOP (*)] and GO and/or Pfam; PDB (*) without GO and/or Pfam; GO and/or Pfam without PDB (*) and no annotation. Each category can further comprise clusters where GO and Pfam functional annotations are or are not statistically significant. BAR⁺ is available at http://bar.biocomp.unibo.it/bar2.0.

INTRODUCTION

In the post-genomic era, with the advent of rapid sequencing techniques, reliable and efficient functional annotation methods are needed. Routinely, a translated protein sequence is aligned towards a data base of already annotated sequences and by this it is endowed with different features depending on the level of sequence identity (SI). This similarity search is the basis for transfer of annotation by homology. The UniProt Knowledgebase (UniProtKB; http://www.UniProtKB.org/) is presently our major resource of information of protein sequences and of corresponding functions and structures, when available. It provides links also to other resources/data bases, allowing a comprehensive knowledge of experimental and computational characteristics of known/putative proteins and genes. However, only 4.4% of the all protein universe that presently (UniProtKB release 2011_03; 8 March 2011) includes some 14 million of sequences has evidence at the protein and at the transcript level. With this scenario, inference of function and structure among related sequences requires the definition of rules to increase the reliability of annotation. This is routinely obtained with clustering methods by which sequences are included into sets of similarity. Clustering can be hierarchical and non-hierarchical. Hierarchical clustering categorizes sequences into a tree-structure. Examples of hierarchical clustering include SYSTERS (1), Picasso (2) and iProClass (3). CluStr (4,5) and ProtoNet (6,7) are the only web servers that comprise the large number of sequences made available by fully sequenced genomes and the entire UniProtKB. Both CluStr and ProtoNet cluster sequences according to different levels of SI, as set by different E-value thresholds, and with different hierarchical algorithms. Alternatively, non-hierarchical methods consider explicitly proteins containing multiple domains or proteins that sharing common domains do not necessarily have the same function. Proteins with different combinations of shared domains can have different molecular and biological functions, as recently re-discussed (10). In order to address these problems, we...
developed BAR (11), an annotation procedure that relies on a non-hierarchical clustering method and a large-scale genome comparison where pairs of sequences are selected with very strict criteria of similarity and overlapping of the alignment as described in the next section. We provided statistical validation that BAR allows reliable functional and structural annotation in addition to that given by commonly used databases (11). Here, we introduce BAR+, an updated and extended version of BAR that includes: (i) a 5-fold increase in sequences; (ii) GO terms from the three main roots (molecular function, biological process and cellular localization; http://www.geneontology.org/); (iii) Pfam domains (http://pfam.sanger.ac.uk/); (iv) known ligands and (v) for clusters containing PDB structure/s, a Cluster HMM model and the corresponding alignment of the target sequence to the optimal template in the cluster for computing its 3D structure.

**BAR+ IMPLEMENTATION**

BAR+ is constructed by performing an all-against-all pairwise alignment of all protein sequences (collected from the entire UniProtKB 05_2010, with the exclusion of fragments (9,399,063 sequences), and from the proteome of complete sequenced genomes available on the same date at the National Center for Biotechnology Information (NCBI) [www.ncbi.nlm.nih.gov/genomes/lproks.cgi (Prokaryotes); www.ncbi.nlm.nih.gov/genomes/leuks.cgi (Eukaryotes)] and at Ensembl (http://www.ensembl.org/info/data/ftp/index.html) for a total of 988 complete proteomes (the list of the species is available at BAR+ web site). For the sake of comparison, we also used the entire SwissProt 03_2011 (8 March). Similarly to BAR (11), BAR+ is also a non-hierarchical clustering method relying on a comparative large-scale genome analysis. The method relies on a non-hierarchical clustering procedure characterized by a stringent metric that ensures a reliable transfer of features within clusters. In this new version, the method takes advantage of a larger scale pairwise sequence comparison than BAR, including 13,495,736 protein sequences. Alignment is performed with BLAST (12) in a GRID environment (11). From this we compute for each pair both the SI and the Coverage (COV) defined as the ratio of the length of the intersection of the aligned regions on the two sequences and the overall length of the alignment (namely the sum of the lengths of the two sequences minus the intersection length). Each protein is then taken as a node and a graph is built allowing links among nodes only when the following similarity constraints are found among two proteins: their SI is \( \geq 40\% \) and COV is \( \geq 90\% \). By this, clusters are simply the connected components of the graph (11). A workflow of the method is shown in Figure 1. Seventy percent of the whole data set (9,401,223 sequences) falls into 913,962 clusters. Noticeably, 55% of the clusters include 84% of the cluster-included sequences. The number of sequence in the clusters ranges from two up to 87,893 in the most populated (Molecular Function: ABC transporter). Given our stringent criteria, 87% of the clusters contain sequences whose standard deviation (SD) of the protein length is \( \leq 5 \) residues. The remaining sequences (30% of the total) originate singletons (containing just one sequence). Well annotated sequences are characterized by functional and structural annotations derived from UniProtKB entries (Figure 1). These include GO, Pfam, PDB and SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) (when available). To assess whether GO and Pfam terms are significant in a cluster, we compute \( P \)-values and given the multiplicity of the terms, we applied the Bonferroni correction (11). We evaluated the cumulative distribution of Bonferroni corrected \( P \)-values by adopting a bootstrapping procedure. From this we set the threshold \( P \)-value at 0.01 in order to discriminate among random and significant (cluster associated) features (11). Validated features (significant for the cluster) are those endowed with \( P \leq 0.01 \). According to our procedure when hypothetical and or putative proteins fall into an annotated and validated cluster, they can safely inherit GO terms and Pfam domain/s even in the case of very low SI with the most annotated proteins. These sequences can
therefore be labelled as distantly related homologues and inherit function and structure (when available) in a validated manner. We previously discussed that this procedure can increase the level of annotation of UniProtKB (11). Here we increase the level of structural and functional annotations of cluster-included sequences by 54% (Figure 2A). When sequences are standing alone (according to our criteria) they are singletons. They can anyway carry along information (Figure 2B), provided that each singleton is endowed with PDB and/or Pfam and/or GO annotation.

**CLUSTER-HMMs**

In BAR+, when PDB templates are present within a cluster (with or without their SCOP classification), profile HMMs are computed on the basis of sequence to structure alignment and are cluster associated (Cluster-HMM) (Figure 1). When different templates are present in a cluster the structural alignment among them is computed with MUSTANG (13). Multiple alignments comprising all the overlapping templates and the sequences similar to them (with SI ≥ 40% and COV ≥ 90%) are computed with MUSCLE (14) and fed to HMMER 2.3 (15) in order to train the profile-HMM. By this, a library of 10 858 HMMs is made available for aligning even distantly related sequences to a given PDB template/s. The server also provides the pairwise query sequence–structural target alignment computed with the Viterbi decoding implemented in HMMER from the corresponding Cluster-HMM and useful for further processing and/or computing the corresponding 3D structure.

**DIFFERENT ANNOTATIONS with BAR+**

BAR+ allows 35 possible fine grain types of annotations (plus no annotation) (Table 1). The most complete type of annotation is the one with PDB (with and without SCOP annotation) and GO terms and Pfam domains with \( P < 0.01 \) (validated) (first row in Table 1). Interestingly, enough 0.11% of the total sequences in our database are sufficient to annotate in a validated manner and with the most complete annotation another 21.99% sharing common clusters (8251; 0.90% of the total), with an annotation gain factor higher than 200. Summing up (along the first row of Table 1), we can conclude that validated functional annotation is possible within 10% of the clusters. Eleven percent of the sequences remains without annotation and are included in 45% of the clusters. About 57% of singletons (corresponding to 17% of the total set) are annotated with different features (Figure 2B and Table 1).

**SUBMITTING A PROTEIN SEQUENCE TO BAR+**

When a query sequence is submitted, there are three possible outcomes (Figure 3). The sequence can match a sequence already present in the cluster (or in a singleton). By this, non-annotated proteins can inherit functional and structural annotation from other proteins within the same cluster. Validated annotations are inherited when clusters are endowed with validated GO and Pfam \( (P < 0.01) \). Alternatively a BLAST alignment starts. The query sequence may then align with any other sequence in BAR+ with the stringent criteria of our procedure and, therefore, find a cluster from where it can safely inherit all the corresponding structural and functional features.
Alternatively, when the criteria are not met, all the BLAST matches are returned. This allows anyway locating the sequence within a cluster. However, in this case, annotation through inheritance should be manually curated. Singletons may be or not source of information depending on their annotation.

BAR\textsuperscript{+} UPDATE

BAR\textsuperscript{+} collects sequences and their features from UniProtKB and genome repositories. Our re-clustering is programmed on a yearly base. BAR\textsuperscript{+} cluster annotation will be updated every 6 months. This is based on the notion that indeed the BAR\textsuperscript{+} annotation system increases its capacity only when we add information. This is achieved when proteins with evidence at the transcript and protein level (e.g.: PDB new files and/or proteins with GO/Pfam terms) are included in the system. For example, by comparing UniprotKB 05_2010 with SwissProt 03_2011, we collected some 2445 sequences carrying information according to our criteria (evidence at protein/transcript level). By aligning this set towards BAR\textsuperscript{+} clusters, we find that 62% of the sequences fall into already validated clusters. About 8% aligns with singletons and only 0.03% of the total number of BAR\textsuperscript{+}

Table 1. The fine grain types of annotation with BAR\textsuperscript{+}

|                      | PDB (%) | SCOP Mono | SCOP Multi | Without PDB |
|----------------------|---------|-----------|------------|-------------|
| **GO validated**     |         |           |            |             |
| Pfam validated       |         |           |            |             |
| Clusters             | 8251 (0.90) | 3613 (0.40) | 1461 (0.16) | 83 266 (9.11) |
| Sequences            | 2982 449 (22.10) | 1408 542 (10.44) | 1028 565 (7.62) | 2 903 431 (21.51) |
| Inherited            | 2907 743 (21.99) | 1404 011 (10.40) | 1026 154 (7.60) | 1 382 310 (10.24) |
| Pfam                 |         |           |            |             |
| Clusters             | 8334 (0.91) | 3647 (0.40) | 1463 (0.16) | 85 886 (9.40) |
| Sequences            | 2984 057 (22.11) | 1409 647 (10.45) | 1028 569 (7.62) | 2 922 876 (21.66) |
| Inherited            | 2909 285 (22.00) | 1405 095 (10.41) | 1026 156 (7.60) | 1 398 603 (10.36) |
| Without Pfam         |         |           |            |             |
| Clusters             | 320 (0.04) | 123 (0.01) | 25\textsuperscript{a} | 6251 (0.68) |
| Sequences            | 42 202 (0.31) | 15 415 (0.11) | 7363 (0.05) | 143 533 (1.06) |
| Inherited            | 41 825 (0.31) | 15 303 (0.11) | 7331 (0.05) | 93 568 (0.69) |
| **GO**               |         |           |            |             |
| Pfam validated       |         |           |            |             |
| Clusters             | 8938 (0.98) | 3887 (0.43) | 1504 (0.16) | 133 895 (14.65) |
| Sequences            | 30 426 649 (22.55) | 14 503 437 (10.75) | 10 297 07 (7.63) | 3 311 421 (24.54) |
| Inherited            | 30 261 916 (22.43) | 1 445 52 (10.71) | 1 027 219 (7.61) | 1 617 763 (11.99) |
| Pfam                 |         |           |            |             |
| Clusters             | 9357 (1.02) | 4033 (0.44) | 1526 (0.17) | 322 937 (35.34) |
| Sequences            | 30 454 65 (22.57) | 14 512 928 (10.76) | 10 297 55 (7.63) | 3 739 076 (27.71) |
| Inherited            | 30 293 337 (22.45) | 1 446 890 (10.72) | 1 027 247 (7.61) | 1 852 223 (13.72) |
| Singletons           | 2608 (0.02) | 10\textsuperscript{a} | 5\textsuperscript{a} | 1 515 720 (11.23) |
| Without Pfam         |         |           |            |             |
| Clusters             | 452 (0.05) | 176 (0.02) | 30\textsuperscript{a} | 45 539 (4.98) |
| Sequences            | 46 311 (0.34) | 17 020 (0.13) | 7400 (0.05) | 330 354 (2.45) |
| Inherited            | 45 803 (0.34) | 16 862 (0.12) | 7362 (0.05) | 226 500 (1.68) |
| Singletons           | 279\textsuperscript{a} | 2\textsuperscript{a} | 2\textsuperscript{a} | 129 212 (0.96) |
| Without GO           |         |           |            |             |
| Pfam validated       |         |           |            |             |
| Clusters             | 679 (0.07) | 345 (0.04) | 15\textsuperscript{a} | 54 314 (5.94) |
| Sequences            | 44 172 (0.33) | 27 775 (0.21) | 654\textsuperscript{a} | 547 459 (4.06) |
| Inherited            | 43 416 (0.32) | 27 410 (0.20) | 633\textsuperscript{a} | 221 585 (1.64) |
| Pfam                 |         |           |            |             |
| Clusters             | 779 (0.09) | 377 (0.04) | 16\textsuperscript{a} | 122 236 (13.38) |
| Sequences            | 44 582 (0.33) | 27 983 (0.21) | 656\textsuperscript{a} | 695 684 (5.15) |
| Inherited            | 43 735 (0.32) | 27 592 (0.20) | 634\textsuperscript{a} | 301 792 (2.24) |
| Singletons           | 205\textsuperscript{a} | 1\textsuperscript{a} | 0\textsuperscript{a} | 702 834 (5.21) |
| Without Pfam         |         |           |            |             |
| Clusters             | 270 (0.03) | 83 (0.01) | 5\textsuperscript{a} | 412 192 (45.11) |
| Sequences            | 5308 (0.04) | 1771 (0.01) | 154\textsuperscript{a} | 1 494 443 (11.07) |
| Inherited            | 5023 (0.04) | 1689 (0.01) | 149\textsuperscript{a} | 1 743 526 (12.92) |
| Singletons           | 129\textsuperscript{a} | 1\textsuperscript{a} | 0\textsuperscript{a} | 702 834 (5.21) |

Percentage is evaluated with respect to the total number of sequences in the data base (13 495 736 sequences). Bold character: sequences that inherit the annotation type

\textsuperscript{a}Values are negligible. Validated: $P < 0.01$ (See text for details, 11). Within BAR\textsuperscript{+} clusters, 35 different types of annotations are possible: (i) +GO+Pfam+PDB [with or without SCOP (Monodomain, Multidomain)*]; GO and Pfam are or not validated (no. of levels = 12). (ii) +Pfam+PDB (with or without SCOP)* (no. of levels = 6). (iii) +GO+PDB (with or without SCOP)* (number of levels = 6). (iv) +Pfam+GO (no. of levels = 4). (v) +PDB (with or without SCOP)* (number of levels = 3). (vi) +GO (no. of levels = 2). (vii) +Pfam (no. of levels = 2). Seventy percent of the initial set fall into clusters (913 962) and 53% in validated clusters. Some 6% of the sequences are annotated without validation and the remaining 11% are not annotated (rightmost bottom cell). About 17 and 13% of the sequences are singletons with and without annotations, respectively.
singleton become new clusters (with two protein sequences). Another 7% fall into non-validated clusters without affecting the statistical significance of the cluster-specific annotation. The remaining 23% originate new singletons. We are currently planning to include other annotation resources in order to extend our annotation process with more protein domains and their interactions.

**ACKNOWLEDGEMENTS**

The authors would like to thank INFN (Istituto Nazionale di Fisica Nucleare) and CNAF (Centro Nazionale per la Ricerca e Sviluppo delle Tecnologie Informatiche e Telematiche) for support in GRID computing.
FUNDING

D.P. is the recipient of a MIUR (Ministero Istruzione Università Ricerca) fellowship supporting his Ph.D. program; MIUR-FIRB (Fondo per gli Investimenti della Ricerca di Base) 2003/LIBI-International Laboratory for Bioinformatics delivered (to R.C., in part). Funding for open access charge: Fondo Ordinario per le Università (FFO) 2010 delivered (to R.C. and P.L.M.).

Conflict of interest statement. None declared.

REFERENCES

1. Krause, A., Stoye, J. and Vingron, M. (2002) The SYSTERS protein sequence cluster set. *Nucleic Acids Res.*, 28, 270–272.
2. Heger, A. and Holm, L. (2001) Picasso: generating a covering set of protein family profiles. *Bioinformatics*, 17, 272–279.
3. Wu, C.H., Huang, H., Nikolskaya, A., Hu, Z. and Barker, W.C. (2001) The iProClass integrated data base for protein functional analysis. *Nucleic Acids Res.*, 29, 52–54.
4. Kriventseva, E.V., Fleischmann, W., Zdobnov, E.M. and Apweiler, R. (2001) CluSTr: a data base of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res.*, 29, 33–36.
5. Petrysak, R., Kretschmann, E., Wieser, D. and Apweiler, R. (2005) The predictive power of the CluSTr data base. *Bioinformatics*, 21, 3604–3609.
6. Kaplan, N., Sasson, O., Inbar, U., Friedlich, M., Fromer, M., Fleischer, H., Portugaly, E., Linial, N. and Linial, M. (2005) ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res.*, 33, D216–D218.
7. Loewenstein, Y., Portugaly, E., Fromer, M. and Linial, M. (2008) Efficient algorithms for accurate hierarchical clustering of huge data sets: tackling the entire protein space. *Bioinformatics*, 24, i41–i49.
8. Sperisen, P. and Pagni, M. (2005) JACOP: a simple and robust method for the automated classification of protein sequences with modular architecture. *BMC Bioinformatics*, 6, 216–227.
9. Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30, 1575–1584.
10. Cuff, A.L., Sillitoe, I., Lewis, T., Clegg, A.B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J. and Orengo, C.A. (2011)Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, 39, D420–D426.
11. Bartoli, L., Montanucci, L., Fronza, R., Martelli, P.L., Fariselli, P., Carota, L., Donvito, G., Maggi, G. 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.
12. McGinnis, S. and Madden, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, 32 (Web Server issue), W20–W25.
13. Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J. and Lesk, A.L. (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, 64, 559–574.
14. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32, 1792–1797.
15. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, 14, 755–763.