Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

BIOTECNOLOGIE, FARMACOLOGIA E TOSSICOLOGIA: PROGETTO N. 1 “BIOTECNOLOGIE CELLULARI E MOLECOLARI”

Ciclo XXV

Settore Concorsuale di afferenza: 05/E1

Settore Scientifico disciplinare: BIO/10

A CLUSTERING METHOD FOR ROBUST AND RELIABLE LARGE SCALE FUNCTIONAL AND STRUCTURAL PROTEIN SEQUENCE ANNOTATION

Presentata da

Damiano Piovesan

Coordinatore Dottorato	Relatore
Prof. Santi Mario Spampinato	**Prof.ssa Rita Casadio**

Esame finale anno 2013

ABSTRACT

Bioinformatics, in the last few decades, has played a fundamental role to give sense to the huge amount of data produced. Obtained the complete sequence of a genome, the major problem of knowing as much as possible of its coding regions, is crucial. Protein sequence annotation is challenging and, due to the size of the problem, only computational approaches can provide a feasible solution. As it has been recently pointed out by the Critical Assessment of Function Annotations (CAFA), most accurate methods are those based on the transfer-by-homology approach and the most incisive contribution is given by cross-genome comparisons. In the present thesis it is described a non-hierarchical sequence clustering method for protein automatic large-scale annotation, called “The Bologna Annotation Resource Plus” (BAR+). The method is based on an all-against-all alignment of more than 13 millions protein sequences characterized by a very stringent metric. BAR+ can safely transfer functional features (Gene Ontology and Pfam terms) inside clusters by means of a statistical validation, even in the case of multi-domain proteins. Within BAR+ clusters it is also possible to transfer the three dimensional structure (when a template is available). This is possible by the way of cluster-specific HMM profiles that can be used to calculate reliable template-to-target alignments even in the case of distantly related proteins (sequence identity < 30%).

Other BAR+ based applications have been developed during my doctorate including the prediction of Magnesium binding sites in human proteins, the ABC transporters superfamily classification and the functional prediction (GO terms) of the CAFA targets. Remarkably, in the CAFA assessment, BAR+ placed among the ten most accurate methods. At present, as a web server for the functional and structural protein sequence annotation, BAR+ is freely available at http://bar.biocomp.unibo.it/bar2.0.

1. INTRODUCTION

Life science and biology are now living a flourishing period. We have now a great opportunity to deeply understand the living machinery thanks to the recent progresses in the genomic field and the integration of modern sequencing techniques as the Next Generation Sequencing (NGS) [1].

Bioinformatics, in the last few decades, has played a fundamental role to give sense to the huge amount of data produced, first of all for the need to annotate the DNA, intended as the process of localizing coding sequences along genomes and secondly to understand the role of translated genes in living cells. Proven that genomic data, and in particular sequences of coding genes, are reliable by means of the accuracy of modern sequencing machines, life science has to deal now with tens of millions of sequences coming from thousands of different organisms.

When the complete sequence of an entire genome becomes available the problem of localizing genes along the sequence poses. Depending on the organism, and so on the experimental data available this process can be automatic, based on computational methods, or manually curated by specialists.

All primary DNA data, the sequences, are publicly available in few well organised databases. The first to be published and the most important is GenBank [2], this database maintains and merges data from three different organisations that exchange their content: the DNA DataBank of Japan (DDBJ) [3], the European Molecular Biology Laboratory (EMBL) [4], and GenBank at NCBI.

In the early stage, before the advent of ultra rapid sequencing machines, GenBank included only prokaryote genomes data, now eukaryotic genomes are available as well, and all the data are organized in two principal divisions including sequences from complete and incomplete genomes, the GenBank and the Whole Genome Shotgun division (WGS) respectively [2]. According to the statistics in the website, at the moment of writing (February 2013), the total number of sequences available is about 150 millions for the canonical GenBank division and about 100 millions for the WGS. In figure 1, it is possible to appreciate the exponential growth of the number of sequences in the two divisions since the GenBank birth date.

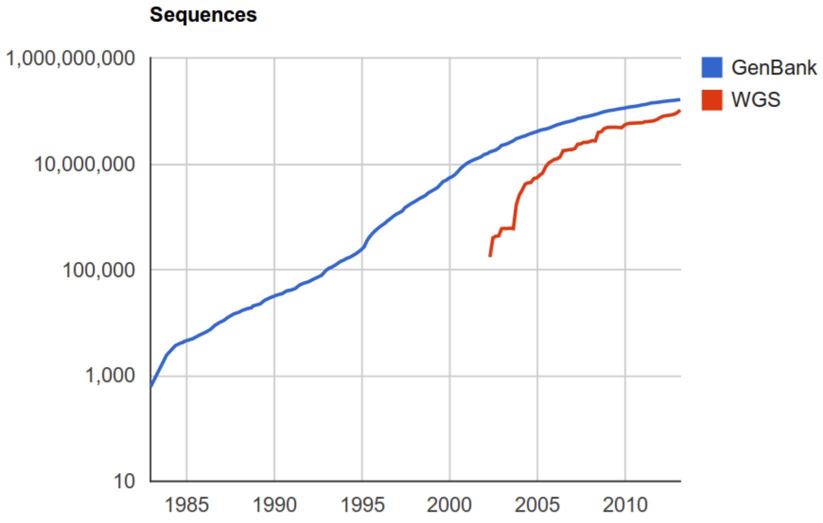


Figure 1 GenBank number of sequences growth.

GenBank collects both complete and incomplete genomes. The WGS indicates the Whole Genome Shotgun sequencing project data coming from incomplete genomes and are collected separately. (Figure obtained from the GenBank website http://www.ncbi.nlm.nih.gov/genbank).

Another database for the collection of genome data is Ensembl [5]. The website was born in the 2000 from a project with the aim to provide annotation for newly sequenced genomes based on automatic pipelines. The database than was expanded including comparative genomics, variation and regulatory data. The Ensembl project is a European collaboration between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI). The database includes 61 eukaryote organisms (February

2013) some of which, like the human and mouse genomes, are manually annotated meaning that transcribed regions are defined after the evaluation of each single experimental evidence by a human curator.

The automatic annotation pipeline in Ensembl, namely Genbuild [6][7], consists in four principal steps. The first one is the localization of all already known specie-specific proteins. The second one is a similarity search of proteins coming from related organisms in the remaining un-annotated regions of the genome, the third stage is the mapping of cDNA and EST data available for that organism to support the predictions made at the preceding stages and to identify the UTRs. The last step is the collection and identification of transcripts that map to the same gene, to remove redundancy.

This genome annotation, intended as the identification of the coding regions, is a nonstop process since new experimental evidences can modify substantially the predicted positions of gene boundaries. Each genome in fact is continuously revised and new annotated versions, also for well-studied genomes, are periodically released. The Genbuild results for the Human genome at the moment of writing were last updated in January 2013.

The availability of protein sequences has made a great difference in the numerous scientific studies of important biological molecules, noticeable are the great advances in the discovery of new associations between DNA mutations in coding regions and diseases [8][9]. However, a comprehensive vision of the complete repertoire of functions performed by all genome regions is still missing and a reliable outline of all biological protein roles is needed to take advantage of the huge amount of sequence data available [10][11].

The amount of protein structures determined by time consuming and expensive experimental methods is significantly smaller when compared to the data produced by large-scale DNA sequencing methods [12]. For example, the number of proteins with trusted and safe manual curated annotations in the UniProtKB [13] database is about 540,000 (release 2013_02) that is 1:55 of the total number of included proteins, considering also automatically annotated proteins.

The experimental data associated with manual curated proteins is the source for computational techniques aiming to fill the gap between the experimental manual protein characterization and the large-scale automatic sequence/structure/function annotation [14][15]. The challenging problem of a reliable large-scale functional annotation has become so important in the last few years that attracted the attention of a large part of the scientific community involved in the development of function predictors [16][17]. To help science to keep pace with this flow of knowledge, bioinformatics continuously develops tools for the management and the integration of many different resources [18][19].

The work described in this thesis is a method for the automatic transfer of structural and functional features from well-annotated proteins to newly un-reviewed targets. The method, called "The Bologna Annotation Resource Plus" (BAR+), relies on a non-hierarchical sequence clustering for protein automatic large-scale annotation. The method is based on an all-against-all alignment of more than 13 millions protein sequences characterized by a very stringent metric that allows a safe transferring of functional and structural features (Gene Ontology functional terms, Pfam domains and PDB structures) by means of a statistical validation and the development of cluster-specific sequence profiles.

During my doctorate BAR+ and many useful related applications have been published and described in some scientific articles (see list of publications section and appendix).

There is also an article describing the results of an evaluation of the state-of-the-art in the field of automatic functional prediction [16]. BAR+ participated in this competition and was judged to be among the best ten prediction methods. This evaluation, called the Critical Assessment of Function Annotations (CAFA) [16], involved more than fifty research groups from all over the world.

1.1 Protein function annotation

Given the entire sequence of a genome, after the identification of coding regions, it is fundamental to understand the specific role of the translated sequences in the living cell. Annotating proteins means to map specific biological functions to sequences. This task unfortunately is one of the most difficult for several reasons, first because is very complicated for experimentalist to test biological functions in living organisms and second because finding the proper definition of a specific biological function for a protein is not trivial [20].

Even if not all proteins are enzymes, the best-known and studied role of proteins is considering them as enzymes. By this, many efforts has been done to classify them by the type of reactions they can catalyze. An example of an available resource for such classification is the Enzyme Commission number (EC number) [21] that organize all reactions in a hierarchical way by an identification code of four digits corresponding to four different levels of increasing specificity.

Of course, to fully understand the biological role of a protein inside the cell considering only the reaction it catalyzes is not enough satisfactory. For example, looking at membrane receptors, knowing that they can phosphorilate a substrate do not tell us so much about their role in signals transduction.

Moreover, it is well known that a large amount of proteins responsible for all biological processes in the cell can perform more than one function per protein and that a single protein can perform different functions depending on the sub-cellular localization or on the type of tissue where it is expressed. This type of proteins are called “moonlight proteins” [22].

As the knowledge about all possible functions performed by proteins is increasing and it is already really extended, a standard and organised ontology has been created, the Gene Ontology [23].

1.2 Gene Ontology

The Gene Ontology (GO) [23] is a controlled vocabulary of functional terms subdivided into three main divisions, namely: i) molecular function (MF); ii) cellular component (CC) and biological process (BP). All GO terms are organized in a directed acyclic graph (DAG), where nodes (terms) represent functional definitions and links represent relationships among terms. It is a directed graph because there is a hierarchy, this means that some terms are more general than other terms and that there is a root term. It is acyclic because it is not possible to have paths that starting from a node point back to the same node. In figure 2 there is an example of a little portion of the gene ontology graph including all the ancestors of two GO terms: “fibroblast growth factor receptor signalling pathway” (GO:0008543) and “transcription corepressor activity” (GO:0003714) with 26 and 4 ancestors respectively. These two terms (and some others not listed here) are associated to the same protein “14-3-3 protein beta/alpha” from Homo Sapiens (P31946 in UniProtKB), an adapter protein involved in many signalling pathways.

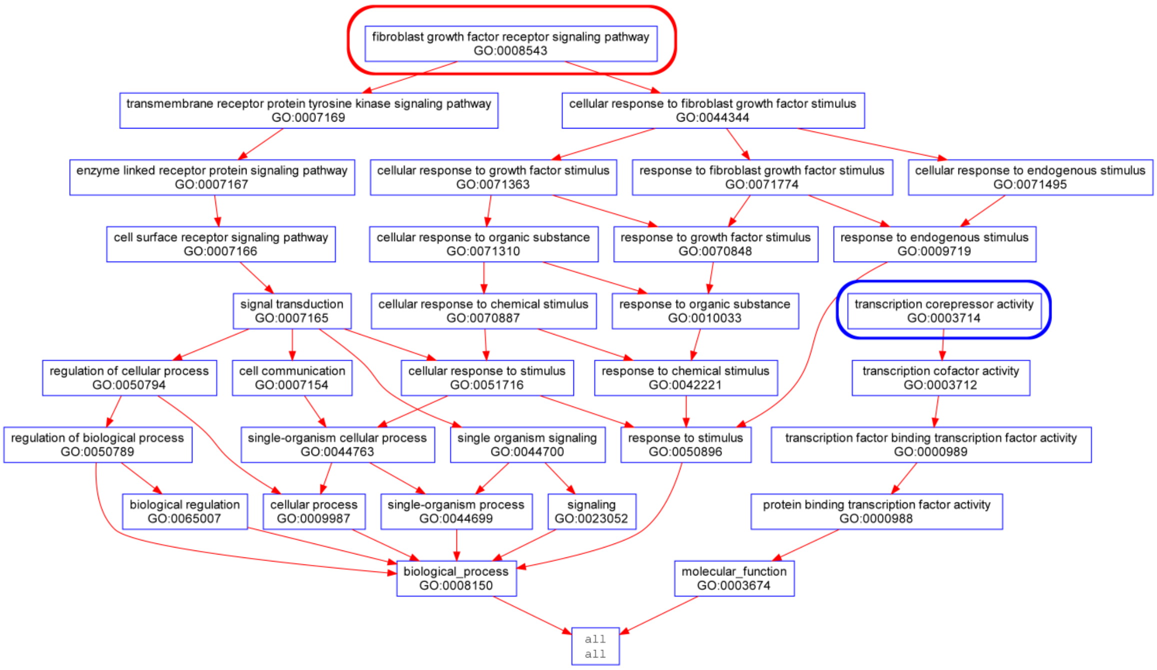

fibroblast growth factor receptor signaling pathway
GO:0008543
transmembrane receptor protein tyrosine kinase signaling pathway
GO:0007169
cellular response to fibroblast growth factor stimulus
GO:0044344
enzyme linked receptor protein signaling pathway
GO:0007167
cellular response to growth factor stimulus
GO:0071363
response to fibroblast growth factor stimulus
GO:0071774
cellular response to endogenous stimulus
GO:0071495
cell surface receptor signaling pathway
GO:0007166
cellular response to organic substance
GO:0071310
response to growth factor stimulus
GO:0070848
response to endogenous stimulus
GO:0009719
signal transduction
GO:0007165
cellular response to chemical stimulus
GO:0070887
response to organic substance
GO:0010033
transcription corepressor activity
GO:0003714
regulation of cellular process
GO:0050794
cell communication
GO:0007154
cellular response to stimulus
GO:0051716
response to chemical stimulus
GO:0042221
transcription cofactor activity
GO:0003712
regulation of biological process
GO:0050789
single-organism cellular process
GO:0044763
single organism signaling
GO:0044700
response to stimulus
GO:0050896
transcription factor binding transcription factor activity
GO:0000989
biological regulation
GO:0065007
cellular process
GO:0009987
single-organism process
GO:0044699
signaling
GO:0023052
protein binding transcription factor activity
GO:0000988
biological_process
GO:0008150
molecular_function
GO:0003674
all
all

Figure 2. A portion of the Gene Ontology graph.

In this example the "fibroblast growth factor receptor signalling pathway" (GO:0008543 red circle) and the "transcription corepressor activity" (GO:0003714 blue circle) are shown. These two terms for example can be associated to the same protein "14-3-3 protein beta/alpha", an adapter protein involved in many signalling pathways.

The three ontologies are quite different considering both the size (the number of terms) and the level of specificity reached in the three main branches. The specificity of a single term is measured as the length of the shortest path that separates that term from the ontology root, and in particular the length (distance) is calculated as the number of nodes traversed along that path. In table 1 there is a statistic on the number of terms and the maximum specificity reached in the Gene Ontology vocabulary calculated for the three main branches separately. When a term is not connected with any children, it is a leaf. Leaf terms are the deepest terms that can be found for each branch of the graph and when a protein is already associated with a leaf term describing a specific function is not possible to obtain further information about that function. Ideally, each protein should be associated only by leaves terms but experimental limitations, very often, allow only a more general annotation. The Biological Process (BP) sub-ontology is the more characterized by considering the number of terms included (24,697 terms) and also the deepest, reaching a maximum depth of 13 nodes. This is a consequence of the much more difficult problem of defining biological processes properly. In fact, the number of biological processes is much larger than the sum of all biochemical reactions carried out by enzyme proteins.

	BP*	CC*	MF*
Terms	24,697	3,146	9,547
Leaves	13,095	2,439	7,646
Obsolete terms°	686	148	894
Max terms depth^	13	9	11
Average terms depth^	6.24	4.21	5.34
Average leaves depth^	6.48	4.28	5.42

Table 1. Statistic of the terms included in the Gene Ontology in the three main sub-ontologies.

** the three main ontologies: BP = Biological Process; CC = Cellular Component; MF = Molecular Function. ^ the term depth is calculated as the minimum number of nodes that separates that term from the root of the ontology. ° obsolete terms are those that were eliminated or substituted in new releases of the Gene Ontology vocabulary. All data refer to the Gene Ontology release of February 2013.*

1.3 Sequence, structure and function relationships in proteins

Encoded in the sequence of a genome there is all information needed to develop a living organism, but the realization of that extremely complex machinery pass through proteins. Proteins are the effective magic tools that carry out almost all biochemical functions in living organisms and participate in all processes inside the cell.

Proteins can be classified in families when sharing similar features at the sequence or structural level and when an evolutionary relationship is established. The basic concept is that if two sequences coming from two different organisms share a certain level of similarity at the sequence or structure level they could be evolutionary related and they could perform an identical or similar function. In such cases, the two proteins are called homologous [24].

The protein function is the key feature that at the end is really subjected to the evolutionary pressure [25] and so as new mutations are collected during evolution the divergence between two proteins is much more extended at the sequence level than for their structures since structures are strictly related to the biological function [25]. Consequently, similarities in protein structures can be more reliable than sequence similarities but when two proteins share similar structures having very different sequences can be considered distantly related homologous proteins [26][27].

However, distinguishing between paralogous and homologous proteins is very important for the correct attribution of the biological function [28]. This specification requires the construction of phylogenetic trees that depends on the challenging ability to identify evolutionary relationships starting from sequence data [29] and on the ability to discriminate between speciation and duplication events. A fundamental resource capturing this distinction for similar proteins is COG (Clusters of Orthologous Groups)

[30], an evolutionary classification that identifies orthologous proteins by means of a large-scale comparative analysis of genomic sequence data.

It is well established that biological functional units can be smaller than an entire protein and that all the functions performed by the protein universe is the sum of different combinations of small units called domains [31]. Many resources actually define families and functions on a domain-based point of view [31].

Sequence similarity search allows clustering procedures to define sets of similar sequences, that can be achieved using similarity-detection tools such as BLAST [32] or profiling tools based on multiple sequence alignments, for example, PSI-BLAST [33]. However, not negligible problems related to this approach are the definition of a similarity threshold for separating families from each other and that is very difficult to safely detect very distantly related proteins based solely on their sequence identity [26].

1.4 Structure-based classification of proteins

The reference resource for protein structures is the Protein Data Bank (PDB) now including some 88,512 structures (February 26, 2013). Even if some researches demonstrated that the number of possible folds is not so extended, the representation of the entire protein structural space is well away to be satisfactory [34].

Different metrics and methods were used to classify proteins and domains on a structural basis. By this point of view the most important projects are the Structural Classification Of Proteins (SCOP) [35] and the CATH Protein Structure Classification (CATH) [36]. Both methods use a hierarchical classification based on structural common properties. The hierarchy levels for the first method, SCOP, are: Class, Fold, Superfamily and Family; for the second, CATH, are: Class, Architecture, Topology and Homologous superfamily.

The two projects differ principally by the method of classification, whereas SCOP is manually curated, CATH involves both automatic and manual classification procedures. Even if they are based on structural features, they differ principally in the identification of domain boundaries inside the same protein.

The analysis of subfamily in SCOP or CATH allows inspecting the functional divergence with respect to structure. Many function prediction methods as SUPERFAMILY [37] and Gene3D [38] take advantage of these data to associate functions to sequence profiles corresponding to the structural classification.

1.5 Classification of proteins by means of sequence clustering

Automatic annotation methods take advantage of clustering techniques to deal with large amount of data. Clustering procedures can be classified as hierarchical and non-hierarchical ones. Hierarchical clusters are based on the construction of a tree representation of similarity between sequences. This approach allows to explore different families at different levels of similarity from closely related homologous to remote distant relationships. CluSTr [39], SYSTERS [40] and ProtoNet [41] are examples of this approach that build their similarity trees starting from a matrix of similarity built upon an all-against-all sequence alignment.

ProtoNet and CluSTr use different linkage criterion to define clusters and different approaches to detect similarity. CluSTr uses the single-linkage clustering where clusters, considering that a cluster is formed by more than one element, are defined on the minimum distance between their members. ProtoNet instead uses the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) where is calculated the average distance between all members of two clusters.

Eventually it can be said that hierarchical clustering does not provide a single partitioning of the data set, but instead it provides an extensive hierarchy of clusters that merge with each other at certain distances, moreover methods based on this approach lead to different results based on the way distances are computed and on the linkage criterion chosen.

Non-hierarchical methods instead provide clusters without any hierarchical relationship between them; these approaches classify proteins by means of an unambiguous partition of the data set generating non-overlapping groups.

ProtoMap [42] takes advantage in generating sequence similarity graphs. TribeMCL [43] applies the Markov clustering approach (MCL). This method operates on a graph that contains similarity information obtained by pair wise alignments of sequences and is rather independent of the presence of multi domain proteins.

1.6 Domain-based classification of proteins

As it has already been commented earlier function is often associated with domains, so the problem of the identification of functional domains from sequence alone poses and solutions provided by current methods are not completely satisfactory [44].

Knowledge of function at the domain level is very useful to increase the accuracy for function prediction methods [31]. Novel functions can arise from different combinations of single domains and an exhaustive library of all possible functional domains is a fundamental resource to inspect the role of unknown proteins.

A great value is given when domains are defined based on the structural classification as Gene3d [38] and SUPERFAMILY [45] do starting respectively from CATH [36] and SCOP [35] domain definitions.

Family domains can be identified by means of multiple sequence alignments (MSA), resources based on MSAs are PROSITE [46] and Pfam [47]. The first, PROSITE, transforms information coming from the MSA in patterns that can be seen as statistical signature profiles family-specific. Pfam similarly uses manually curated MSA, but they are processed to build Hidden Markov Models (HMM), a sophisticated way to create a profile representation that is able to detect specific sequence patterns even in distantly related proteins [48].

Pfam was last updated in November 2011 and contains more than 13,000 families. The library of HMM profiles covers about the 71% of UniProtKB [13] sequences. InterPro [49][50] instead is a consensus method (including also Pfam profiles) that increases the UniProtKB coverage up to 77% thanks to the combination of 11 different resources containing different domain definitions.

2. BAR+

Improvements in bioinformatics gave rise to a noticeable growth in the field of automatic protein annotation. That happened thanks to the computational power reached nowadays and the development of new effective tools for large-scale comparisons of available complete genomes and proteomes sequences.

Based on the notion that similar sequences share similar functions and structures the largest part of current automatic methods just perform a more or less sophisticated homology-based transfer of annotation [16].

A common accepted identity threshold for a safe functional transfer based solely on sequence similarity is about 40-50% [26]. Nevertheless, considering that proteins can contain multiple domains and that different combinations of shared domains can lead to different functions [51], the overlap extent between two aligned sequences (coverage) should be considered to avoid erroneous predictions.

BAR+, is an updated version of the previously developed method BAR (The Bologna Annotation Resource [52]) that is described in [53]. The method relies on a non-hierarchical clustering of the protein sequence universe based on a all-against-all large-scale similarity search. Clustering in BAR+ is characterized by a very stringent metric that ensures a reliable detection of evolutionary relationships among pair of sequences even in the case of multi domain proteins.

Many improvements have been achieved after the first published version of the method [52], firstly by considering the increased size of the sequence space explored and secondly by looking at the quality of annotation transferred. Following there is a brief description of BAR+, further details and the presentation of the last release of the web server are included in this paper [53] (also printed in the appendix).

2.1 Data set

BAR+ includes 988 complete genomes from both prokaryotic and eukaryotic organisms. Bacterial genomes were downloaded from the NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) while eukaryotic ones from both NCBI (ftp://ftp.ncbi.nih.gov/refseq/release/) and Ensembl [5] (ftp://ftp.ensembl.org/pub/). Summing up were obtained complete proteomes sequences from 925 prokaryotes and 63 eukaryotes including 4,096,673 sequences.

Another 9,399,063 sequences were retrieved from the UniProtKB (release 05_2010) excluding fragments. UniProtKB is one of the databases of the Universal Protein Resource (UniProt) [54]. It collects all available proteins with experimental annotations (Swiss-Prot [55] division) and all predicted proteins from genomic data (TrEMBL division). In TrEMBL almost all annotations are assigned with computational procedures.

2.2 Sequence alignments

The first step to compare sequences is an all-against-all pair wise comparison of the entire dataset (13,495,736 sequences) with the BLAST (Basic Local Alignment Search Tool) program [32]. Pitfalls of BLAST are that it performs only local alignments and that it is based on a heuristic algorithm so the optimal solution (the optimal alignment) is not guaranteed. On the contrary, it is much faster than any other alignment algorithm and it is a key feature considering that an all-against-all comparison given this dataset means about 10^{14} alignments. Even if BLAST is very fast in a single desktop computer to perform all these alignments it would have taken about 7 years long, so the program was run in parallel by means of a GRID computing environment [52] where, exploiting distributed resources (500 CPU), the process time was reduced to few months.

Each single alignment was run with some fixed parameters to obtain statistically comparable results. The database size parameter was settled to 100,000 and the E-value threshold to 10^{-10}. Some other BLAST parameters were left unchanged with their default values (gap opening penalty = -11, gap extension penalty = -1, substitution matrix = BLOSUM62).

2.3 BAR+ clustering

After the alignments the similarity relationships between sequences was represented by an undirected graph where nodes are protein sequences and links are similarity relationships between proteins. The similarity between proteins is evaluated considering two specific parameters of the pair wise alignment: sequence identity and coverage.

The identity was calculated on the aligned sequences as the percentage of identical residues in the same position. This value was taken as it is in the BLAST output. The coverage instead was calculated as the ratio between the overlap of the two sequences over the total length of the alignment. By this, two proteins were connected in the graph if and only if the following constraints were respected simultaneously:

Sequence identity ≥ 40%

Coverage ≥ 90%

With these stringent criteria the detection of evolutionary relationships is more reliable and it is guaranteed also for multi-domain proteins thanks to the high coverage threshold.

Clusters were defined as the connected components of the graph. This means that all members of a cluster are connected through at least a path and that all clusters are disjoined. This type of clustering ensures a unique partitioning of the sequence space independently from any arbitrary decision on the detection level of evolutionary relationships. As a result, with this approach it is possible to have members of the same cluster that are not directly connected. Moreover, in large clusters coexist pairs of proteins with low sequence identity (< 30%) implying that remote homologous proteins can be grouped together. All sequences without any link with other proteins are called singletons.

2.4 Statistical validation of cluster specific annotations

Given that clusters members are mostly derived from UniProtKB [13] entries, it is possible to collect all annotations associated to these proteins. However, when very different sequences are included in the same BAR+ cluster, it is difficult to assess which pool of annotations can really fit all sequences inside the cluster, and it is true in particular for big clusters. In other words, it is fundamental to define a set of annotations that can be safely transferred to all un-annotated sequences inside the same cluster. The selection of cluster specific annotations is then performed by a statistical validation already described in [52]. This procedure consists of a calculation for each annotation of a P-value representing the probability that a specific annotation can be found in a specific cluster by chance. This procedure was applied for both Gene Ontology and Pfam terms.

For each term inside a cluster the P-value was calculated as:

$$\boldsymbol{P\ value\ (term)} = \sum_{\boldsymbol{i=N}}^{\boldsymbol{min(M,Z)}} \left(\frac{\binom{\boldsymbol{M}}{\boldsymbol{i}}\binom{\boldsymbol{D-M}}{\boldsymbol{Z-i}}}{\binom{\boldsymbol{D}}{\boldsymbol{Z}}} \right)$$

Where N is the number of sequences in a given cluster endowed with the same specific term, D is the number of sequences with at least an annotation, Z is number of sequences with at least an annotation inside the cluster, M is the number of sequences with the same annotation in the entire database. To each calculated value was then applied the Bonferroni correction considering that a cluster can contain more than one term.

In order to assess if a term is statistically significant a bootstrapping procedure was applied to find a P-value threshold. The bootstrapping was performed by reshuffling randomly go terms among clusters but maintaining the specific number of annotations for each cluster. Repeating this procedure for 100 times, it was possible to compare the distribution of the random generated P-values with the observed ones. A P-value threshold of 0.01 was found to maximize the difference between the two distributions and it guarantees that terms under that value are cluster-specific.

2.5 Structural modelling through HMM cluster profiles

Structural modelling is a quite simple task when it is possible to find a template that share at least 30% of sequence identity with the target protein. When it is not the case, but some information relating the target fold are known, homology modelling is still possible. In the worst case, when any suitable template cannot be found, the only un-trusted “ab-initio” prediction methods are feasible. In any case, a good structural model depends directly on the quality and reliability of the template/target sequence alignment [56]. BAR+ offers a powerful solution for modelling targets that fall into clusters containing structural templates. This is particularly interesting when distantly related target/template pairs coexist in the same cluster [53]. Reliable target/template alignments in a BAR+ cluster endowed with a template are possible by means of a HMM model, that is calculated from the multiple sequence alignment (MSA) of the template (or multiple templates if it is the case) and the corresponding neighbours. The neighbours are sequences directly linked with the template and so sharing at least 40% of sequence identity and a coverage ≥ 90%. To increase accuracy of the cluster-specific HMM profiles, when multiple templates coexist in the same cluster, the MSA of the neighbours is refined based on the structural alignment of the templates. The total number of clusters including templates and so endowed with an HMM in BAR+ is 10,858 [53].

3. BAR+ APPLICATIONS

3.1 Protein classification in BAR+

The following three paragraphs refer to a paper accepted by the editor but not printed yet and so not included in the appendix.

The definition of the term “family” is very complex when speaking about proteins; it depends on the metric we consider to group them. If the basic concept is that proteins belonging to the same family share a common ancestor, it is very difficult to determine boundaries between similar families and detect the complete set of proteins belonging to a particular family. Two proteins with very similar structure but with very low sequence identity (for example lower than 20%) are probably remote homologues [25] but it is very difficult to discriminate between ortologous and paralogous proteins or exclude an event of convergent evolution [30][57]. In recent years a number of different classification systems have been developed to organize proteins, both at the sequence and structural level [58].

Among all classification schemes, the most noticeable are those based on: (1) hierarchical families of proteins, such as super-families/families; (2) families of protein domains, such as those in Pfam [47]; (3) sequence motifs or conserved regions, like in PROSITE [46]; (4) structural classes, such as in SCOP [35] and CATH [36]. As a case study of the discriminative power of identifying protein families here an analysis of the biggest cluster in BAR+ is reported. The cluster includes the ATP-binding domain of the ABC (ATP Binding Cassette) transporters.

In the following analysis the Transporter Classification Data Base (TCDB) [59] was considered as reference because its classification scheme is completely manually curated and it classifies proteins based on their main function and source organism.

3.2 ATP-binding cassette cluster in BAR+

The ABC-transporters are members of a protein superfamily that contains both uptake and efflux transport systems. These transporters couple the P-P-bond-hydrolysis of the ATP to drive transport of various substrates across membranes and also to participate in processes of DNA and RNA repair [60][61]. These fundamental roles explain also why these proteins can be found and are extremely conserved over all organisms in every kingdom [62][63].

Members of the family can be found in living cells as complexed subunits where a dimeric Nucleotide Binding Domain (NBD) is coupled with a dimeric Transmembrane Domain (TMD) formed by alfa-helices, and as dimers where the NBD plus the TMD are fused together, the dimeric organization is the one adopted in particular by exporters. Some complexes are also more complicated. There are examples where an accessory subunit is necessary for the recruitment of a specific substrate acting as a receptor, and there are also particular transporters spanning from the interior of the mitochondrion to the cell cytoplasm allowing substrate translocation across two membranes.

In BAR+ the biggest cluster (the cluster number 1) includes the NDB domain of the ABC-transporters. The cluster contains 87,893 proteins mostly from prokaryotes and with an average length of 281 residues. This cluster well represents a typical situation in which structural features and functional annotation can be safely transferred from well-annotated SwissProt entries. The cluster indeed contains 22 PDB from Prokaryotes that remain, after the superimposition, in a RMSD range of 1.89 ± 0.39 Å and 77 validated GO terms that are safely transferred to about 37% of the cluster sequences never annotated before (figure 3).

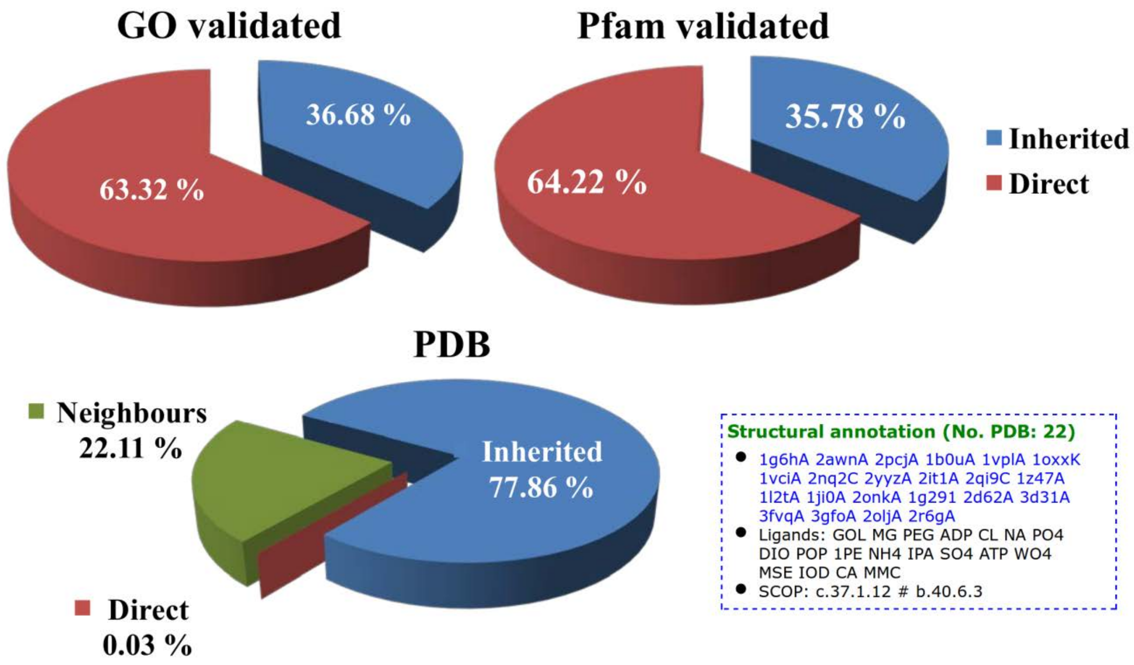


Figure 3. Annotation transfer of the ATP binding domain of the ABC transporters.

In the cluster number 1 it is possible to transfer annotations at the structural and functional level. 22 PDB are available as templates and 73 GO terms and 6 Pfam domains can be safely transferred to un-annotated sequences. Percentage refers to UniProtKB already annotated entries compared to un-annotated sequences. Interestingly more than 53,000 sequences can be modelled by means of the cluster HMM profile with low homology with all templates (sequence identity < 30%).

3.3 Extending the TCDB with BAR+

Similarly to the Enzyme Classification (EC) [21], proteins in the TCDB are associated with a 5 digits code. The first 3 digits of every ABC-transporter is “3.A.1” where “3” is for “Primary Active Transporters”, “A” stays for “P-P-bond-hydrolysis-driven transporters” and “1” indicates the “ATP-binding Cassette (ABC) Superfamily”. The other 2 digits specify the substrate and the organism, for example “3.A.1.1.4” is the lactose porter (“1”) of the Agrobacterium Radiobacter (“4”). At the end, 422 different ABC-transporters can transport 88 different substrates.

For each transporter for each organism the corresponding subunits (multiple chains or just one for fused domains) are mapped in UniProtKB and are labelled based on their role in the complex and their localization in the cell. Because of this, labels correspond to: cytoplasmic proteins (C), transmembrane proteins (TM), receptor proteins (R), proteins where the membrane and the cytoplasmic portion are fused together (MC); proteins with joined membrane and receptor subunits (MR), altogether there is a total of 1,073 chains mapped into UniProtKB sequences.

In figure 4 it is represented the percentage of sequences that inherit TCDB annotation/labelling in BAR+, the most populated cell compartments are the cytoplasm (C), membrane (TM) and outside cell with receptor proteins (R).

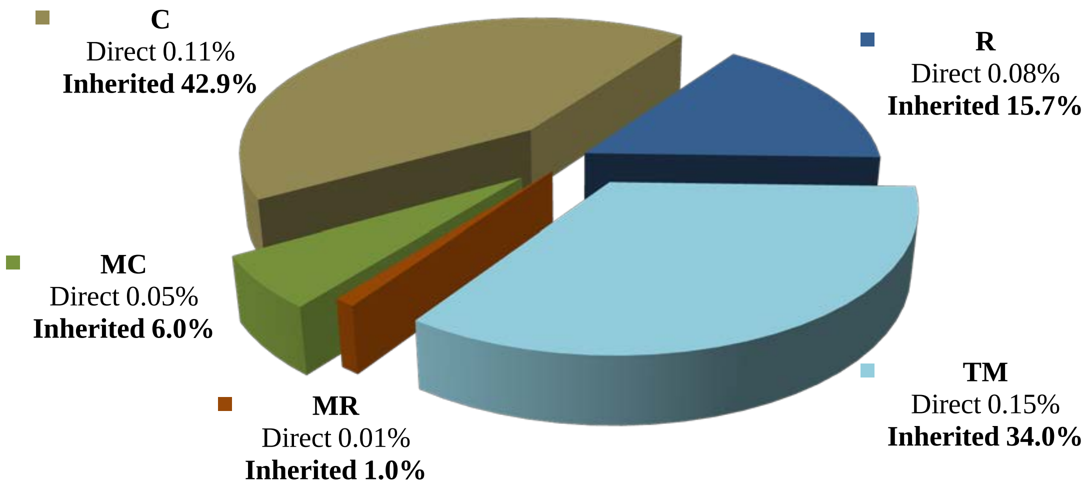


Figure 4. Annotated sequences in BAR+ annotated accordingly to the TCDB classification.

Cytoplasmic proteins (C), transmembrane proteins (TM), receptor proteins (R), proteins with fused membrane and cytoplasmic chains (MC); proteins comprising a membrane and an extra cytoplasmic portion (MR).

In BAR+, these 1,073 sequences map to 396 clusters, containing 256,866 other sequences. This procedure allows to confirm the sub-cellular localization specificity of BAR+ clusters. In fact, TCDB subunits belonging to different organisms that fall in the same cluster are always annotated in agreement with BAR+ validated terms. By exploiting the BAR+ power of transferring annotations, we can extend the size of the TCDB of about 256 times. In particular, considering all clusters containing ABC transporters subunits are transferred: 124 Molecular Function, 201 Biological process and 41 Cellular Component terms to 243,364, 237,657 and 214,558 sequences respectively. 70 Pfam domains are inherited by 256,349 sequences (figure 4).

3.4 Ligands and binding sites in BAR+: the “human magnesome”

When a cluster in BAR+ is endowed with a PDB template, it also includes an HMM profile that allows a reliable target/template alignment. The ability of modelling structures in BAR+ also implies the transferring of all other structural features associated to PDB templates including substrates binding sites.

Following there is a description of the BAR+ ability to transfer magnesium binding sites in proteins. The choice of studying this cation was based on the fact that magnesium binding sites are less conserved through evolution compared to others divalent cations and their detection is very difficult [64]. The work described here was published in [65] and it is also included in the appendix.

Magnesium covers a large amount of different roles in living organisms both at the structural and functional level. It is fundamental as cofactor for more than 300 reactions in cells and it is involved in the stabilization of membranes and nucleic acids thanks to its high positive charge [64].

Magnesium is a divalent cation with a small radius, a great charge density and it is coordinated with an octahedral geometry. In proteins it usually binds no more than three residues (it binds carbonyl oxygen in the backbone or charged side chains atoms) and water molecules to satisfy the total of six bonds of its coordination [66]. Magnesium concentration in living cells is very high (0.5-1mM, [67]), it is the most abundant alkaline cation and its concentration seems to drive the association with proteins.

The only bioinformatics resources available for the magnesium binding sites analysis are the PROCOGNATE [68] web server that maps ligands from PDB to cognate enzymes in SCOP [35] and CATH [36] and PROSITE [46] that defines just few patterns useful to retrieve only very specific domains. Moreover, it has been recently implemented a method to retrieve magnesium binding sites in structures by implementing a structural alphabet [69], but it is relevant for only already structurally solved proteins.

The first step in BAR+ to transfer binding sites among sequences was to map PDB residues that bind Mg into the corresponding template sequence. To identify atoms interactions on known three-dimensional structures it is sufficient to set a cut-off distance based on the type of interaction that has to be detected. To avoid arbitrary choices the Mg interactions with protein atoms have been identified parsing both the “LINK” and “SITE” fields on the PDB files. When multiple PDB refer to the same sequence and different magnesium is bound by different amino acids, all the sequence positions corresponding to the magnesium binding residues were collected.

Binding positions were transferred from the template(s) to the target after a pair-wise alignment calculated by means of the cluster specific HMM with the Hmmalign [48] tool.

The set of human sequences that fall into clusters containing magnesium binding templates was defined as the “human magnesome” [65]. The total number of humans sequences that bind magnesium in BAR+ is 3,751. The number of clusters containing the 1,341 PDB involved in the magnesium binding sites transfer is 251. These clusters were also manually checked for the presence of structures without any published evidence supporting any observed functional or structural role of Mg^{2+} in the cell so far. This was the case for only 119 structures falling into 21 clusters, for these templates a functional role of the magnesium cation is not supported yet by literature.

The number of human sequences that inherited annotation from human templates is 2,688. Other 1,063 sequences inherited magnesium binding sites from structures of other organisms.

4. CONCLUSIONS

Automatic protein annotation is a major challenge for bioinformatics. Many tools and resources have been developed so far, but the computational approaches for predicting protein functions given the sequence are not satisfactory yet. BAR+ is among the most accurate methods for functional automatic annotation, as demonstrated by a recent international benchmark [16]. However, room for improvements still remains. Given the stringent criteria adopted to identify evolutionary relationships, most of the clusters in BAR+ contain only strictly related proteins. Sometimes this detection based on sequence similarity is not sensitive enough and two related proteins could remain in separated clusters. On the other side, there are also clusters so large that proteins with different functions are grouped together. That happens in clusters containing proteins with very strongly conserved domains, like the ATP binding cassette. This domain is so important for life that it is present in all living organisms and many proteins are combinations of this domain and other short sub-domains that are responsible for different secondary specific functions. In such cases, an additional sub-clustering could be necessary.

Summing-up BAR+ clustering procedure can be improved adopting different clustering approaches for different type of families by developing a metric to explore clusters and detect these situations.

Another problem is related to the quality of the source data associated to proteins that are sequences and sequence annotations. The impressive growth rate of available sequences in UniProtKB has already been discussed previously in the introduction. In BAR+, this problem conveys in the all-against-all sequence comparison necessary for the clustering that is the only computational bottleneck of the entire pipeline. The UniProtKB database now contains about twice the number of sequences already included in BAR+, and so the number of new alignments needed is enormous considering that it is proportional to the square of the number of sequences.

On the other hand, annotations associated to sequences are frequently updated and consistently modified even in the case of manually curated proteins. For example, the Gene Ontology functional term “protein binding” has been recently removed from all entries in SwissProt, the UniProtKB division that is supposed to contain only trusted experimental annotations. So even BAR+, that is based on a statistical validation for the annotation transferring, is sensible to systematic errors included in source annotation databases, as all other homology-based methods are. This entails that the BAR+ pool of annotations and the corresponding P-values need to be frequently updated and recalculated.

Presently BAR+ ranges among the most accurate methods for sequence to function and/or sequence to structure and function prediction [16]. It is freely available on the web (http://bar.biocomp.unibo.it/bar2.0) and it represents a unique source of information for protein families.

5. ACKNOWLEDGEMENTS

Foremost I would like to thank my supervisor Professor Rita Casadio, I have been really honoured to be one of her student. I have appreciated so much her dedication and her generosity of time and ideas to make my doctorate experience really productive and stimulating. I am also thankful for the excellent example she has shown me as a successful scientist and professor.

The members of the Biocomputing Group have contributed immensely to my personal and professional time at the University of Bologna. The group has been a source of friendships as well as good advices and collaborations. I am especially grateful for the scientific support and supervision that both Professor Pier Luigi Martelli and Professor Piero Fariselli have given me. I would like to thank also all the other members of the Biocomputing Group for their kindness in supporting me during difficult situations and for their uncountable programming suggestions, in particular: Castrense Savojardo, Gianluca Tasco, Giuseppe Profiti and Pietro Di Lena.

I am also very grateful to Professor David Jones that took me in his laboratory at UCL for three months and his collaborators Domenico Cozzetto and Federico Minneci that gave me a concrete idea on how a profitable scientific collaboration works.

Special thanks to my family and friends for their support and love.

6. BIBLIOGRAPHY

1. Mardis ER: **Next-Generation DNA Sequencing Methods**. *Annual Review of Genomics and Human Genetics* 2008, **9**:387–402.

2. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Research* 2012, **41**:D36–D42.

3. Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y: **DDBJ launches a new archive database with analytical tools for next-generation sequence data**. *Nucleic Acids Res* 2010, **38**:D33–D38.

4. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Hoad G, Kanz C, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Pastor MPG, Plaister S, Sobhany S, Stoehr P, Vaughan R, Wu D, Zhu W, Apweiler R: **EMBL Nucleotide Sequence Database in 2006**. *Nucleic Acids Res* 2007, **35**:D16–D20.

5. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJP, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SMJ: **Ensembl 2012**. *Nucleic Acids Research* 2011, **40**:D84–D90.

6. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: **The Ensembl Automatic Gene Annotation System**. *Genome Res.* 2004, **14**:942–950.

7. Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SMJ, Stabenau A, Storey R, Clamp M: **The Ensembl Analysis Pipeline**. *Genome Res.* 2004, **14**:934–941.

8. Hirschhorn JN, Daly MJ: **Genome-wide association studies for common diseases and complex traits**. *Nat Rev Genet* 2005, **6**:95–108.

9. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: **A comprehensive review of genetic association studies**. *Genet Med* 2002, **4**:45–61.

10. Xie L, Bourne PE: **Functional Coverage of the Human Genome by Existing Structures, Structural Genomics Targets, and Homology Models**. *PLoS Comput Biol* 2005, **1**:e31.

11. Raes J, Harrington ED, Singh AH, Bork P: **Protein function space: viewing the limits or limited by our view?** *Current Opinion in Structural Biology* 2007, **17**:362–369.

12. Yura K, Yamaguchi A, Go M: **Coverage of whole proteome by structural genomics observed through protein homology modeling database**. *J. Struct. Funct. Genomics* 2006, **7**:65–76.

13. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data**. *Database* 2011, **2011**:bar009–bar009.

14. Lee D, Redfern O, Orengo C: **Predicting protein function from sequence and structure**. *Nat Rev Mol Cell Biol* 2007, **8**:995–1005.

15. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment**. *Proteins: Structure, Function, and Bioinformatics* 1991, **9**:56–68.

16. Radivojac P, Clark WT, Oron TR, et al.: **A large-scale evaluation of computational protein function prediction**. *Nat Meth* 2013, **10**:221–227.

17. Friedberg I: **Automated protein function prediction—the genomic challenge**. *Brief Bioinform* 2006, **7**:225–242.

18. Wu CH, Huang H, Yeh L-SL, Barker WC: **Protein family classification and functional annotation**. *Comput Biol Chem* 2003, **27**:37–47.

19. Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, De Castro E, Lachaize C, Baratin D, Phan I, Bougueleret L, Bairoch A: **HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot**. *Nucleic Acids Res.* 2009, **37**:D471–478.

20. Jones CE, Brown AL, Baumann U: **Estimating the annotation error rate of curated GO database sequence annotations**. *BMC Bioinformatics* 2007, **8**:170.

21. Bairoch A: **The ENZYME database in 2000**. *Nucl. Acids Res.* 2000, **28**:304–305.

22. Jeffery CJ: **Moonlighting proteins--an update**. *Mol Biosyst* 2009, **5**:345–350.

23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat. Genet.* 2000, **25**:25–29.

24. Reeck GR, De Haën C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, Zuckerkandl E: **“Homology” in proteins and nucleic acids: A terminology muddle and a way out of it**. *Cell* 1987, **50**:667.

25. Chothia C, Lesk AM: **The Evolution of Protein Structures**. *Cold Spring Harb Symp Quant Biol* 1987, **52**:399–405.

26. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *The EMBO journal* 1986, **5**:823.

27. Wilson CA, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores**. *J. Mol. Biol.* 2000, **297**:233–249.

28. Nehrt NL, Clark WT, Radivojac P, Hahn MW: **Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals**. *PLoS Comput Biol* 2011, **7**:e1002073.

29. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E: **EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates**. *Genome Res.* 2009, **19**:327–335.

30. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution**. *Nucleic Acids Res* 2000, **28**:33–36.

31. Reid AJ, Yeats C, Orengo CA: **Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone**. *Bioinformatics* 2007, **23**:2353–2360.

32. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421.

33. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res.* 1997, **25**:3389–3402.

34. Leonov H, Mitchell JSB, Arkin IT: **Monte Carlo estimation of the number of possible protein folds: Effects of sampling bias and folds distributions**. *Proteins: Structure, Function, and Bioinformatics* 2003, **51**:352–359.

35. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG: **Data growth and its impact on the SCOP database: new developments**. *Nucl. Acids Res.* 2007.

36. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA: **New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures**. *Nucleic Acids Res.* 2013, **41**:D490–498.

37. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J: **SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny**. *Nucl. Acids Res.* 2009, **37**:D380–D386.

38. Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S, Orengo CA: **Gene3D: modelling protein structure, function and evolution**. *Nucl. Acids Res.* 2006, **34**:D281–D284.

39. Kriventseva EV, Fleischmann W, Zdobnov EM, Apweiler R: **CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins**. *Nucleic Acids Res.* 2001, **29**:33–36.

40. Krause A, Stoye J, Vingron M: **Large scale hierarchical clustering of protein sequences**. *BMC Bioinformatics* 2005, **6**:15.

41. Kaplan N, Sasson O, Inbar U, Friedlich M, Fromer M, Fleischer H, Portugaly E, Linial N, Linial M: **ProtoNet 4.0: a hierarchical classification of one million protein sequences**. *Nucleic Acids Res.* 2005, **33**:D216–218.

42. Yona G, Linial N, Linial M: **ProtoMap: automatic classification of protein sequences and hierarchy of protein families**. *Nucleic Acids Res* 2000, **28**:49–55.

43. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Res.* 2002, **30**:1575–1584.

44. Tress M, Cheng J, Baldi P, Joo K, Lee J, Seo J-H, Lee J, Baker D, Chivian D, Kim D, Ezkurdia I: **Assessment of predictions submitted for the CASP7 domain prediction category**. *Proteins* 2007, **69 Suppl 8**:137–151.

45. De Lima Morais DA, Fang H, Rackham OJL, Wilson D, Pethica R, Chothia C, Gough J: **SUPERFAMILY 1.75 including a domain-centric gene ontology method**. *Nucleic Acids Res.* 2011, **39**:D427–434.

46. Sigrist CJA, Cerutti L, De Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **PROSITE, a protein domain database for functional characterization and annotation**. *Nucleic acids research* 2010, **38**:D161.

47. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A: **The Pfam protein families database**. *Nucl. Acids Res.* 2010, **38**:D211–D222.

48. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucl. Acids Res.* 2011, **39**:W29–W37.

49. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, De Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJA, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong S-Y: **InterPro in 2011: new developments in the family and domain prediction database**. *Nucleic Acids Res* 2012, **40**:D306–D312.

50. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier**. *Nucleic acids research* 2005, **33**:W116.

51. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes**. *J. Mol. Biol.* 2001, **310**:311–325.

52. Bartoli L, Montanucci L, Fronza R, Martelli PL, Fariselli P, Carota L, Donvito G, Maggi GP, Casadio R: **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.* 2009, **8**:4362–4371.

53. Piovesan D, Luigi Martelli P, Fariselli P, Zauli A, Rossi I, Casadio R: **BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences**. *Nucleic Acids Research* 2011.

54. The UniProt Consortium: **Reorganizing the protein space at the Universal Protein Resource (UniProt)**. *Nucleic Acids Research* 2011, **40**:D71–D75.

55. Bairoch A, Boeckmann B, Ferro S, Gasteiger E: **Swiss-Prot: juggling between evolution and stability**. *Brief. Bioinformatics* 2004, **5**:39–55.

56. Nayeem A, Sitkoff D, Krystek S: **A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models**. *Protein Science* 2006, **15**:808–824.

57. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE: **Protein molecular function prediction by Bayesian phylogenomics**. *PLoS Comput. Biol.* 2005, **1**:e45.

58. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A: **Protein function annotation by homology-based inference**. *Genome Biology* 2009, **10**:207.

59. Saier MH Jr, Yen MR, Noto K, Tamang DG, Elkan C: **The Transporter Classification Database: recent advances**. *Nucleic Acids Res.* 2009, **37**:D274–278.

60. Davidson AL, Dassa E, Orelle C, Chen J: **Structure, function, and evolution of bacterial ATP-binding cassette systems**. *Microbiol. Mol. Biol. Rev.* 2008, **72**:317–364, table of contents.

61. Wagner K, Moolenaar GF, Goosen N: **Role of the two ATPase domains of Escherichia coli UvrA in binding non-bulky DNA lesions and interaction with UvrB**. *DNA Repair (Amst.)* 2010, **9**:1176–1186.

62. Anjard C, Loomis WF: **Evolutionary Analyses of ABC Transporters of Dictyostelium discoideum**. *Eukaryot Cell* 2002, **1**:643–652.

63. Berntsson RP-A, Smits SHJ, Schmitt L, Slotboom D-J, Poolman B: **A structural classification of substrate-binding proteins**. *FEBS Lett.* 2010, **584**:2606–2617.

64. Cowan JA: **Metal Activation of Enzymes in Nucleic Acid Biochemistry**. *Chem. Rev.* 1998, **98**:1067–1088.

65. Piovesan D, Profiti G, Martelli PL, Casadio R: **The human “magnesome”: detecting magnesium binding sites on human proteins**. *BMC Bioinformatics* 2012, **13**:S10.

66. Dudev T, Cowan JA, Lim C: **Competitive Binding in Magnesium Coordination Chemistry: Water versus Ligands of Biological Interest**. *Journal of the American Chemical Society* 1999, **121**:7665–7673.

67. Cowan J: **Structural and catalytic chemistry of magnesium-dependent enzymes**. *Biometals* 2002, **15**:225–235.

68. Bashton M, Nobeli I, Thornton JM: **PROCOGNATE: a cognate ligand domain mapping for enzymes**. *Nucleic Acids Research* 2007, **36**:D618–D622.

69. Dudev M, Lim C: **Discovering structural motifs using a structural alphabet: application to magnesium-binding sites**. *BMC bioinformatics* 2007, **8**:106.

7. LIST OF PUBLICATIONS

7.1 Printed papers

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

Piovesan D, Martelli PL, Fariselli P, Zauli A, Rossi I and Casadio R.

Nucleic Acids Research **(2011) doi: 10.1093/nar/gkr292**

The human "magnesome": detecting magnesium binding sites on human proteins

Piovesan D, Profiti G, Martelli PL, Casadio R.

BMC Bioinformatics **(2012) doi: 10.1186/1471-2105-13-S14-S10**

How to inherit statistically validated annotation within BAR+ protein clusters

Piovesan D, Martelli PL, Fariselli P, Profiti G, Zauli A, Rossi I, Casadio R.

BMC Bioinformatics **(2013) doi:10.1186/1471-2105-14-S3-S4**

A large-scale evaluation of computational protein function prediction

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML,

Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Törönen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DW, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kaßner R, Seemayer S, Vicedo E, Schaefer C, Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Hönigschmid P, Hopf TA, Kaufmann S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Björne J, Salakoski T, Wong A, Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJ, Skunca N, Supek F, Bošnjak M, Panov P, Džeroski S, Smuc T, Kourmpetis YA, van Dijk AD, Braak CJ, Zhou Y, Gong Q, Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric N, Guo Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney SD, Friedberg I.

***Nature Methods* (2013) doi: 10.1038/nmeth.23**

7.2 Accepted papers

Extended and robust protein sequence annotation over conservative non hierarchical clusters: the case study of the ABC transporters

Piovesan D, Profiti G, Martelli PL, Fariselli P, Casadio R.

ACM Journal on Emerging Technologies in Computing System (In print)

7.3 Poster presentations

The human “magnesome”: how to detect human proteins that can bind Mg ions

Piovesan D, Casadio R

Presentation at EUROMAG Bologna 2011 , Bologna, Italy, June 8-10, 2011

From protein sequence to function and structure with BAR+

Piovesan D, Martelli PL, Fariselli P, Rossi I, Guerzoni D, Donvito G, Maggi GP, Casadio R.

Presentation at CAFA Sig. (ISMB), Vienna, Austria, July 15-16, 2011

7.4 Oral presentations

Protein structure prediction in the genomic era: annotation-facilitated remote homology detection

Piovesan D

National Congress of Chemical Division of Biological Systems 2010, San Vito di Cadore, Italy, September 9-11, 2010

Extended and robust protein sequence annotation over conservative non hierarchical clusters. The Bologna Annotation Resource v 2.0

Piovesan D, Martelli PL, Fariselli P, Rossi I, Guerzoni D, Donvito G, Maggi GP, Casadio R.

14th International Biotechnology Symposium and Exhibition 2010, Rimini, Italy, September 14-18, 2010

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

Piovesan D, Martelli PL, Fariselli P, Rossi I, Guerzoni D, Donvito G, Maggi GP, Casadio R.

Presentation at BITS 2011, Pisa, Italy, July 20-22, 2011

APPENDIX

In the following pages are included the printed versions of all these papers:

- **BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences**
- **The human "magnesome": detecting magnesium binding sites on human proteins**
- **How to inherit statistically validated annotation within BAR+ protein clusters**
- **A large-scale evaluation of computational protein function prediction**

Nucleic Acids Research Advance Access published May 26, 2011

Nucleic Acids Research, 2011, **1–6**
doi:10.1093/nar/gkr292

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

Damiano Piovesan[1], Pier Luigi Martelli[1], Piero Fariselli[2], Andrea Zauli[3], Ivan Rossi[3] and Rita Casadio[1,*]

[1]Department of Biology, Bologna Biocomputing Group, Bologna Computational Biology Network, [2]Department of Computer Science, University of Bologna, Bologna and [3]BioDec 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 clustering partitions a sequence data set into disjoint clusters (8,9). However, neither hierarchical nor 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

*To whom correspondence should be addressed. Tel: +39 0512094005; Fax: +39 0512094005; Email: casadio@biocomp.unibo.it 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

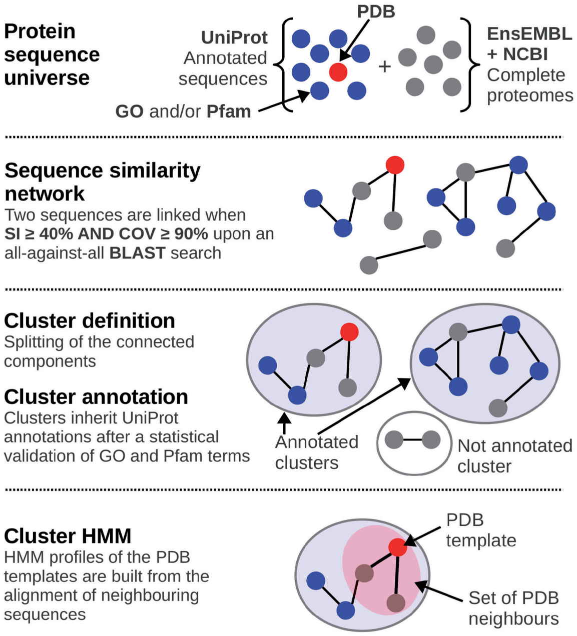

Protein sequence universe
UniProt
Annotated sequences
PDB
GO and/or Pfam
+
EnsEMBL + NCBI
Complete proteomes
Sequence similarity network
Two sequences are linked when
SI ≥ 40% AND COV ≥ 90% upon an all-against-all BLAST search
Cluster definition
Splitting of the connected components
Cluster annotation
Clusters inherit UniProt annotations after a statistical validation of GO and Pfam terms
Annotated clusters
Not annotated cluster
Cluster HMM
HMM profiles of the PDB templates are built from the alignment of neighbouring sequences
PDB template
Set of PDB neighbours

Figure 1. BAR$^+$ implementation. Our method collects sequences from the protein universe (UniProtKB) including also some 988 genomes. By this, all the features [PDB (± SCOP classification) (red circles), GO terms (including Molecular Function, Biological Process and Cellular Localization) and Pfam models (blue circles) are also included. An extensive BLAST alignment is performed of all the 13 495 736 sequences in a GRID environment. The sequence similarity network is built by connecting two sequences only if their SI is $\geq$40% with an overlapping COV $\geq$ 90%. About 913 762 clusters are obtained by splitting of the connected components. By this, any cluster may contain from 2 up to 87 893 sequences (one cluster containing ABC transporters from Prokaryotes, Eukaryotes and Archaea). Stand alone sequences are called Singletons (30.4% of the total protein universe). Sequences inherit the annotations within a cluster. When clusters are endowed with PDB template/s, a Cluster-HMM is generated by considering all the sequences that have an identity $\geq$ 40% and a COV $\geq$ 90% with the structure/s (pink subset). The Cluster-HMM can be used to align all the other sequences in the cluster to template/s.

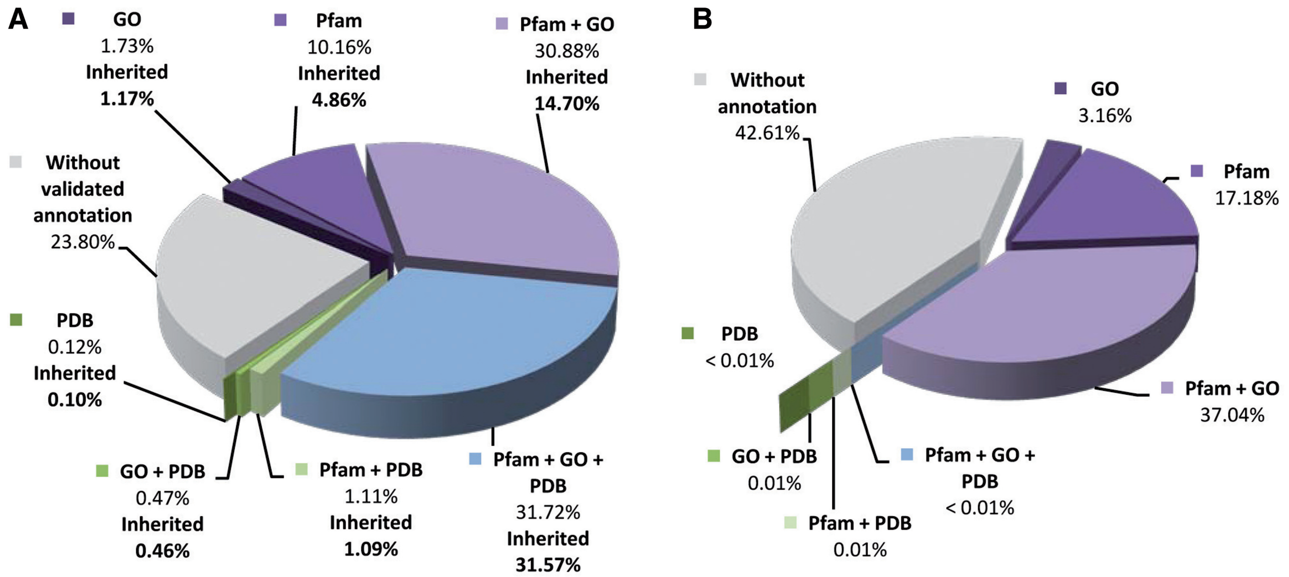


Figure 2. Different types of annotations are possible with BAR$^+$. After clustering and depending on the features (structure, domains and function) annotated in the cluster, sequences within a cluster can inherit different types of annotation. The percentage of sequences endowed with a given annotation type and inheriting validated annotation ($P < 0.01$) is indicated. (**A**) Sequences within clusters. Percentage is computed with respect to 9 401 223 comprised in 913 762 clusters. Inherited: sequences that inherit annotations by falling into a cluster. Without validated annotation: the slice comprises sequences with no annotation and not validated annotations. (**B**) Singletons (stand alone sequences). Percentage is computed with respect to 4 091 908 singleton sequences.

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 $\geq$ 40% and COV $\geq$ 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 correspondent 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 $P \leq 0.01$ (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.

Table 1. The fine grain types of annotation with BAR^+

	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	2 982 449 (22.10)	1 408 542 (10.44)	1 028 565 (7.62)	2 903 431 (21.51)
Inherited	**2 967 743 (21.99)**	**1 404 011 (10.40)**	**1 026 154 (7.60)**	**1 382 310 (10.24)**
Pfam				
Clusters	8334 (0.91)	3647 (0.40)	1463 (0.16)	85 886 (9.40)
Sequences	2 984 057 (22.11)	1 409 647 (10.45)	1 028 569 (7.62)	2 922 876 (21.66)
Inherited	**2 969 285 (22.00)**	**1 405 095 (10.41)**	**1 026 156 (7.60)**	**1 398 603 (10.36)**
Without Pfam				
Clusters	320 (0.04)	123 (0.01)	25[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	3 042 649 (22.55)	1 450 437 (10.75)	1 029 707 (7.63)	3 311 421 (24.54)
Inherited	**3 026 916 (22.43)**	**1 445 521 (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	3 045 465 (22.57)	1 451 928 (10.76)	1 029 755 (7.63)	3 739 076 (27.71)
Inherited	**3 029 337 (22.45)**	**1 446 890 (10.72)**	**1 027 247 (7.61)**	**1 852 223 (13.72)**
Singletons	2608 (0.02)	10[a]	5[a]	1 515 720 (11.23)
Without Pfam				
Clusters	452 (0.05)	176 (0.02)	30[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[a]	2[a]	2[a]	129 212 (0.96)
Without GO				
Pfam validated				
Clusters	679 (0.07)	345 (0.04)	15[a]	54 314 (5.94)
Sequences	44 172 (0.33)	27 775 (0.21)	654[a]	547 459 (4.06)
Inherited	**43 416 (0.32)**	**27 410 (0.20)**	**633**[a]	**221 585 (1.64)**
Pfam				
Clusters	779 (0.09)	377 (0.04)	16[a]	122 236 (13.38)
Sequences	44 582 (0.33)	27 983 (0.21)	656[a]	695 684 (5.15)
Inherited	**43 735 (0.32)**	**27 592 (0.20)**	**634**[a]	**301 792 (2.24)**
Singletons	205[a]	**1**[a]	**0**[a]	702 834 (5.21)
Without Pfam				
Clusters	270 (0.03)	83 (0.01)	5[a]	412 192 (45.11)
Sequences	5308 (0.04)	1771 (0.01)	154[a]	1 494 443 (11.07)
Inherited	**5023 (0.04)**	**1689 (0.01)**	**149**[a]	
Singletons	129[a]	1[a]	0[a]	1 743 526 (12.92)

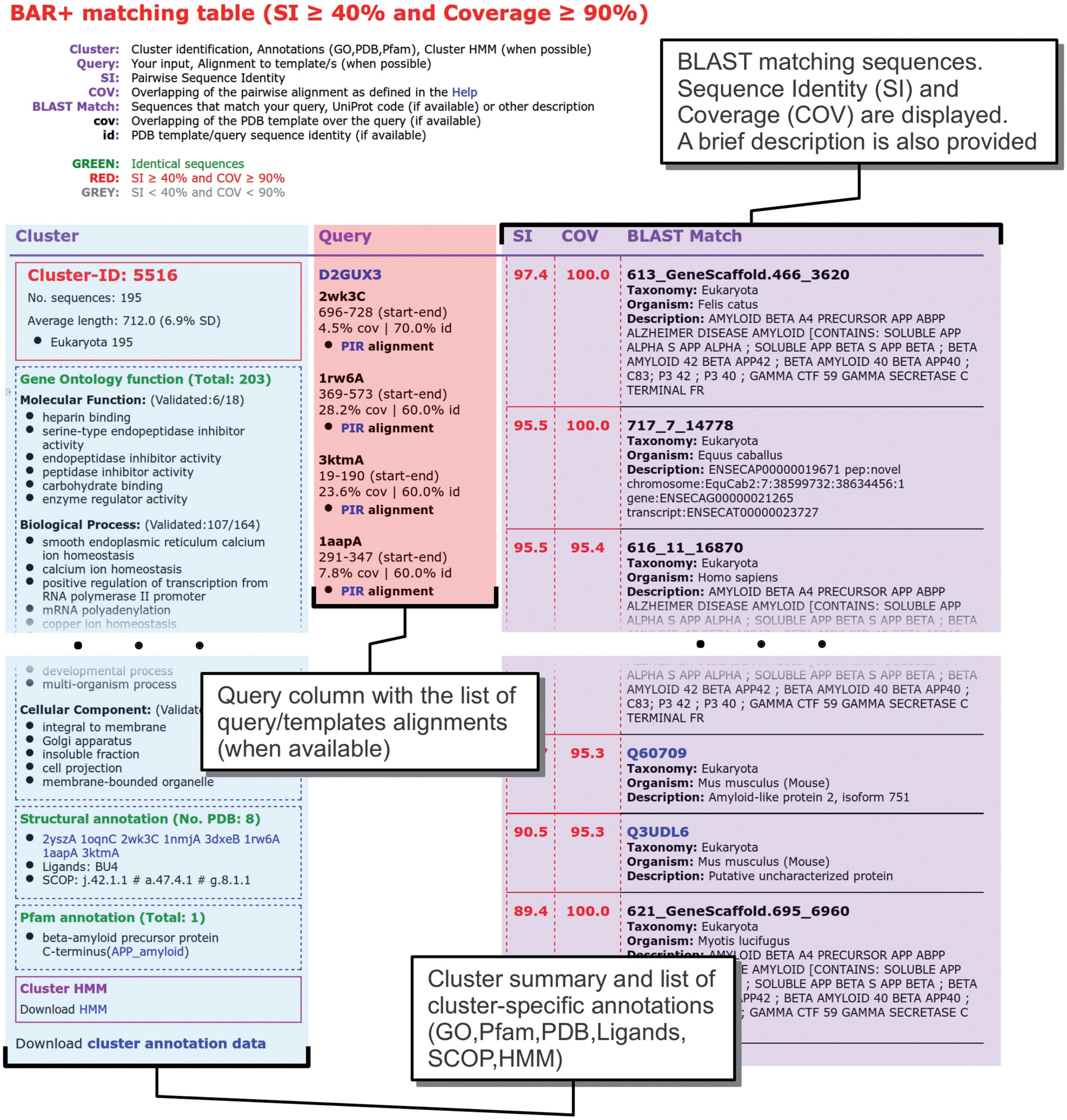
Percentage is evaluated with respect to the total number of sequences in the data base (13 495 736 sequences). Bold character: sequences that inherit the annotation type

[a]Values are negligible. Validated: $P \leq 0.01$ (See text for details, 11). Within BAR^+ 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.

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^+ UPDATE

BAR^+ collects sequences and their features from UniProtKB and genome repositories. Our re-clustering is programmed on a yearly base. BAR^+ cluster annotation will be updated every 6 months. This is based on the notion that indeed the BAR^+ 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^+ 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^+



BAR+ matching table (SI ≥ 40% and Coverage ≥ 90%)

Cluster: Cluster identification, Annotations (GO,PDB,Pfam), Cluster HMM (when possible)
Query: Your input, Alignment to template/s (when possible)
SI: Pairwise Sequence Identity
COV: Overlapping of the pairwise alignment as defined in the Help
BLAST Match: Sequences that match your query, UniProt code (if available) or other description
cov: Overlapping of the PDB template over the query (if available)
id: PDB template/query sequence identity (if available)

GREEN: Identical sequences
RED: SI ≥ 40% and COV ≥ 90%
GREY: SI < 40% and COV < 90%

BLAST matching sequences. Sequence Identity (SI) and Coverage (COV) are displayed. A brief description is also provided

Cluster

Cluster-ID: 5516

No. sequences: 195

Average length: 712.0 (6.9% SD)

- Eukaryota 195

Gene Ontology function (Total: 203)

Molecular Function: (Validated:6/18)

- heparin binding
- serine-type endopeptidase inhibitor activity
- endopeptidase inhibitor activity
- peptidase inhibitor activity
- carbohydrate binding
- enzyme regulator activity

Biological Process: (Validated:107/164)

- smooth endoplasmic reticulum calcium ion homeostasis
- calcium ion homeostasis
- positive regulation of transcription from RNA polymerase II promoter
- mRNA polyadenylation
- copper ion homeostasis

● ● ●

- developmental process
- multi-organism process

Cellular Component: (Validat

- integral to membrane
- Golgi apparatus
- insoluble fraction
- cell projection
- membrane-bounded organelle

Structural annotation (No. PDB: 8)

- 2yszA 1oqnC 2wk3C 1nmjA 3dxeB 1rw6A 1aapA 3ktmA
- Ligands: BU4
- SCOP: j.42.1.1 # a.47.4.1 # g.8.1.1

Pfam annotation (Total: 1)

- beta-amyloid precursor protein C-terminus(APP_amyloid)

Cluster HMM

Download HMM

Download **cluster annotation data**

Query

D2GUX3

2wk3C
696-728 (start-end)
4.5% cov | 60.0% id
- **PIR alignment**

1rw6A
369-573 (start-end)
28.2% cov | 60.0% id
- **PIR alignment**

3ktmA
19-190 (start-end)
23.6% cov | 60.0% id
- **PIR alignment**

1aapA
291-347 (start-end)
7.8% cov | 60.0% id
- **PIR alignment**

Query column with the list of query/templates alignments (when available)

SI	COV	BLAST Match
97.4	100.0	**613_GeneScaffold.466_3620** **Taxonomy:** Eukaryota **Organism:** Felis catus **Description:** AMYLOID BETA A4 PRECURSOR APP ABPP ALZHEIMER DISEASE AMYLOID [CONTAINS: SOLUBLE APP ALPHA S APP ALPHA ; SOLUBLE APP BETA S APP BETA ; BETA AMYLOID 42 BETA APP42 ; BETA AMYLOID 40 BETA APP40 ; C83; P3 42 ; P3 40 ; GAMMA CTF 59 GAMMA SECRETASE C TERMINAL FR
95.5	100.0	**717_7_14778** **Taxonomy:** Eukaryota **Organism:** Equus caballus **Description:** ENSECAP00000019671 pep:novel chromosome:EquCab2:7:38599732:38634456:1 gene:ENSECAG00000021265 transcript:ENSECAT00000023727
95.5	95.4	**616_11_16870** **Taxonomy:** Eukaryota **Organism:** Homo sapiens **Description:** AMYLOID BETA A4 PRECURSOR APP ABPP ALZHEIMER DISEASE AMYLOID [CONTAINS: SOLUBLE APP ALPHA S APP ALPHA ; SOLUBLE APP BETA S APP BETA ; BETA
● ● ●		
		ALPHA S APP ALPHA ; SOLUBLE APP BETA S APP BETA ; BETA AMYLOID 42 BETA APP42 ; BETA AMYLOID 40 BETA APP40 ; C83; P3 42 ; P3 40 ; GAMMA CTF 59 GAMMA SECRETASE C TERMINAL FR
	95.3	**Q60709** **Taxonomy:** Eukaryota **Organism:** Mus musculus (Mouse) **Description:** Amyloid-like protein 2, isoform 751
90.5	95.3	**Q3UDL6** **Taxonomy:** Eukaryota **Organism:** Mus musculus (Mouse) **Description:** Putative uncharacterized protein
89.4	100.0	**621_GeneScaffold.695_6960** **Taxonomy:** Eukaryota **Organism:** Myotis lucifugus **Description:** AMYLOID BETA A4 PRECURSOR APP ABPP E AMYLOID [CONTAINS: SOLUBLE APP ; SOLUBLE APP BETA S APP BETA ; BETA PP42 ; BETA AMYLOID 40 BETA APP40 ; ; GAMMA CTF 59 GAMMA SECRETASE C

Cluster summary and list of cluster-specific annotations (GO,Pfam,PDB,Ligands, SCOP,HMM)

Figure 3. BAR$^+$ at work. A query sequence has been submitted. Provided that the sequence after running BLAST has a level of SI ≥ 40% with a COV ≥ 90% to any sequence of BAR$^+$, it is included into a cluster. In the above example, the cluster is well annotated and the sequence inherits all the possible annotations from the cluster including GO terms (203), PDB/s, ligands, SCOP and Pfam annotations and the Cluster-HMM. Furthermore in PIR format alignment/alignments of the query sequence to the cluster template/s with Cluster HMM is/are also provided. All the sequences that align with the query are returned. (●●●) Only the top and bottom portions of the page are shown.

singletons 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. Petryszak,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.

BMC Bioinformatics

RESEARCH Open Access

The human "magnesome": detecting magnesium binding sites on human proteins

Damiano Piovesan[1], Giuseppe Profiti[1,2], Pier Luigi Martelli[1], Rita Casadio[1,2*]

From NETTAB 2011 Workshop on Clinical Bioinformatics
Pavia, Italy. 12-14 October 2011

Abstract

Background: Magnesium research is increasing in molecular medicine due to the relevance of this ion in several important biological processes and associated molecular pathogeneses. It is still difficult to predict from the protein covalent structure whether a human chain is or not involved in magnesium binding. This is mainly due to little information on the structural characteristics of magnesium binding sites in proteins and protein complexes. Magnesium binding features, differently from those of other divalent cations such as calcium and zinc, are elusive. Here we address a question that is relevant in protein annotation: how many human proteins can bind Mg^{2+}? Our analysis is performed taking advantage of the recently implemented Bologna Annotation Resource (BAR-PLUS), a non hierarchical clustering method that relies on the pair wise sequence comparison of about 14 millions proteins from over 300.000 species and their grouping into clusters where annotation can safely be inherited after statistical validation.

Results: After cluster assignment of the latest version of the human proteome, the total number of human proteins for which we can assign putative Mg binding sites is 3,751. Among these proteins, 2,688 inherit annotation directly from human templates and 1,063 inherit annotation from templates of other organisms. Protein structures are highly conserved inside a given cluster. Transfer of structural properties is possible after alignment of a given sequence with the protein structures that characterise a given cluster as obtained with a Hidden Markov Model (HMM) based procedure. Interestingly a set of 370 human sequences inherit Mg^{2+} binding sites from templates sharing less than 30% sequence identity with the template.

Conclusion: We describe and deliver the "human magnesome", a set of proteins of the human proteome that inherit putative binding of magnesium ions. With our BAR-hMG, 251 clusters including 1,341 magnesium binding protein structures corresponding to 387 sequences are sufficient to annotate some 13,689 residues in 3,751 human sequences as "magnesium binding". Protein structures act therefore as three dimensional seeds for structural and functional annotation of human sequences. The data base collects specifically all the human proteins that can be annotated according to our procedure as "magnesium binding", the corresponding structures and BAR+ clusters from where they derive the annotation (http://bar.biocomp.unibo.it/mg).

Background

Magnesium is the most abundant divalent alkaline ion in living cells and it is an indispensable element for many biological processes. Magnesium deficiency in humans is responsible for many diseases including osteoporosis [1] or metabolic syndrome (MetS), a combination of different metabolic disorders that increase the risk of developing cardiovascular diseases and diabetes [2]. Magnesium is characterised by specific chemico-physical properties: it is redox inert, it has a small ionic radius and is consequently endowed with a high charge density [3,4]. In cells magnesium ions have both structural and functional roles. Magnesium plays a key role in stabilising protein structures, phosphate groups of membrane lipids and negatively charged phosphates of nucleic acids. Concomitantly, it is

* Correspondence: casadio@biocomp.unibo.it
[1]Biocomputing Group, Department of Biology, University of Bologna, Bologna, 40126, Italy
Full list of author information is available at the end of the article

also involved in catalytic roles, such as the activation/inhibition of many enzymes [3,4].

Observations on the structural geometry of Mg^{2+} binding sites in proteins known with atomic resolution may be derived from PROCOGNATE, a cognate ligand domain mapping for enzymes [5] and from the Protein Data Bank [PDB, http://www.rcsb.org]. Typical magnesium binding sites on proteins show three or fewer direct binding contacts with carbonyl oxygen atoms of the backbone and/or protein side chains, with a tendency to bind water molecules given the octahedral coordination geometry of the divalent cation [3,6]. It is known that Mg^{2+} binding sites are less specific than those of other divalent cations such as Zn^{2+} and Ca^{2+}, and that in particular conditions, Zn^{2+} can dislocate Mg^{2+} from its pocket [3,7]. Apparently metal binding sites on proteins seem to satisfy constraints related to the physiological availability of the ions [4]. Magnesium binds weakly to proteins and enzymes ($Ka \leq 10^5\ M^{-1}$) [8] and its binding affinity appears to be dependent on its high cellular concentration. Free Mg^{2+} concentration is higher than that of any other ion (0.5-1mM, [4]). As a consequence magnesium binding sites are less conserved through evolution than those of others divalent cations [4] and their detection is therefore difficult. Mg^{2+} binding sequence motifs have been described to be conserved in similar RNA and DNA polymerases [9,10]. Three dimensional Mg^{2+} binding pockets derived from 70 Mg^{2+} binding proteins solved at atomic resolution were recognised in protein structures by implementing a structural alphabet [11].

In this work we describe how to assign putative Mg^{2+} binding sites to human proteins that lack structural information and also to proteins that share less than 30% sequence identity with any available Mg^{2+} binding protein template. This is possible within our BAR-PLUS annotation resource (BAR+), a non hierarchical clustering method that has been recently described and relies on the pair wise sequence comparison of about 14 millions proteins, including 998 complete proteomes of different species and *Homo sapiens* [12,13]. This paper to our knowledge describes the first large scale investigation of magnesium binding sites at the human proteome level. The results highlight that residues involved in magnesium binding in protein structures (derived from the PDB) falling into the same BAR+ cluster are conserved and can be transferred to all the human sequences sharing the same cluster on the basis of structure to sequence alignment with a cluster specific hidden Markov model (HMM). Magnesium binding sites within a given cluster are also conserved when pair-wise sequence identity among the target and the template/s is less than 30%. A data base (BAR-hMG) is made available from where for a given human input sequence the predicted magnesium binding site/s can be retrieved with the corresponding structural template/s and the annotating BAR+ cluster.

Methods

The dataset of Mg^{2+} binding protein structures

A list of 4,710 magnesium binding protein structures was retrieved from the Ligand-Expo database [14] by searching “MG” as Mg^{2+} ligand identifier. The Expo database is a data warehouse that integrates databases, services and tools related to small molecules bound to macromolecules and based on PDB. It allows users to extract ligand information directly from the PDB, to perform chemical substructure searches of PDB ligands using a graphical interface and also to browse other relevant small molecule resources on the Web. It is updated daily and therefore provides the most current information on small molecules present in the PDB. Its reliability is based on the reliability of the structures from where information is derived and ultimately on the resolution of the electron density map of the molecule. Our set includes PDBs with an average Resolution (R) factor of 0.23 nm. The list of magnesium binding residues and corresponding positions in the sequence for each PDB was obtained parsing both the “LINK” and “SITE” fields on the coordinate files [15]. In order to guarantee that magnesium is part of a biologically significant PDB structure, we filtered out fragments and chimeric structures by constraining the coverage of the template PDB structure to its UniProtKB corresponding sequence (without signal peptide, when present) to be ≥70%. This bound guarantees a satisfactory overlapping of the sequence to its structure and this is essential in building by homology procedures. Applying this criterion, we ended up with 1,341 PDB templates. For each PDB structure the reference sequence and the corresponding UniProtKB [16] accession are obtained from the Sifts web server [17]. In case of multiple PDBs containing different magnesium binding sites and referring to the same sequence, all the sites are mapped into the protein sequence. Human sequences are collected from UniProtKB (release 2011_02), including also splicing isoforms, for a total of 110,464 sequences. Most of these sequences are annotated in UniProtKB in an automatic way and lack any experimental evidence. When fragments are filtered out, the total number of human sequences adopted for our analysis is 84,520.

The BAR-PLUS annotation resource

BAR+ is an annotation resource based on the notion that sequences with high identity value to a counterpart can inherit from this the same function/s and structure, if available (http://bar.biocomp.unibo.it/bar2.0/). The method has been recently described [13]. Briefly, an extensive BLAST alignment [18] was performed for some 13,495,736 sequences in a GRID environment [13]. The sequence similarity network was built by connecting two proteins only if their sequence identity is ≥40% with an overlap (Coverage, COV) ≥90%. 913,762 clusters were obtained by splitting of the connected components of the similarity network. Mapping of PDB, Pfam functional domains (http://pfam.sanger.ac.uk/) and GO terms (Gene Ontology terms, http://www.geneontology.org/) as listed in the UniProtKB protein files allows different annotation types within each cluster. Enrichment of Pfam domains [http://www.sanger.ac.uk/resources/databases/pfam.html] and GO terms [http://www.geneontology.org/] for each cluster was statistically validated (by computing a Bonferroni corrected P-value and by selecting its significance threshold with a bootstrapping procedure) [13]. Only when P<0.01, terms are transferred from one protein to another one in the same cluster and annotation is inherited by all the sequences in the cluster. When a sequence falls into a validated cluster it can inherit in a validated manner functional and structural annotation (PDB +/SCOP +/Pfam +/GOterms +/). Stand alone sequences are called Singletons (30.4% of the total protein universe). Clusters can contain distantly related proteins that by this procedure can be annotated with high confidence. We verified that the magnesium containing 1,341 PDB structures were in BAR+ clusters and when not present, we included them in the corresponding cluster. In any case we verified that backbone structure was conserved in the same cluster (average Root Mean Square Deviation (RMSD) was about 2.0±0.2 Å) (for the definition of RSMD see: http://cnx.org/content/m11608/latest/). The human sequences were then aligned against BAR+ clusters and only those satisfying the BAR+ constraints (ID≥40% and COV≥90%) were retained. Out of the 84,520 human sequences aligned towards BAR+ with the required criteria, some 61,106 fell into 22,858 clusters and some 2,791 aligned with singletons. The remaining portion of the human proteome (aligned with sequences contained in BAR+ clusters with lower sequence identity and coverage than those required for a validated transfer of annotation) is not considered in the present analysis. In BAR+, each cluster endowed with structure/s is characterised by a computed cluster Hidden Markov Model (HMM) that is derived from a structure-to-sequence alignment within the cluster and can be adopted to model the cluster sequences on the structure template/s of the cluster [12]. We took advantage of the cluster HMM both for structural alignments of the newly introduced PDB structures and for sequence-to-structure alignment.

Selection of the "human magnesome"

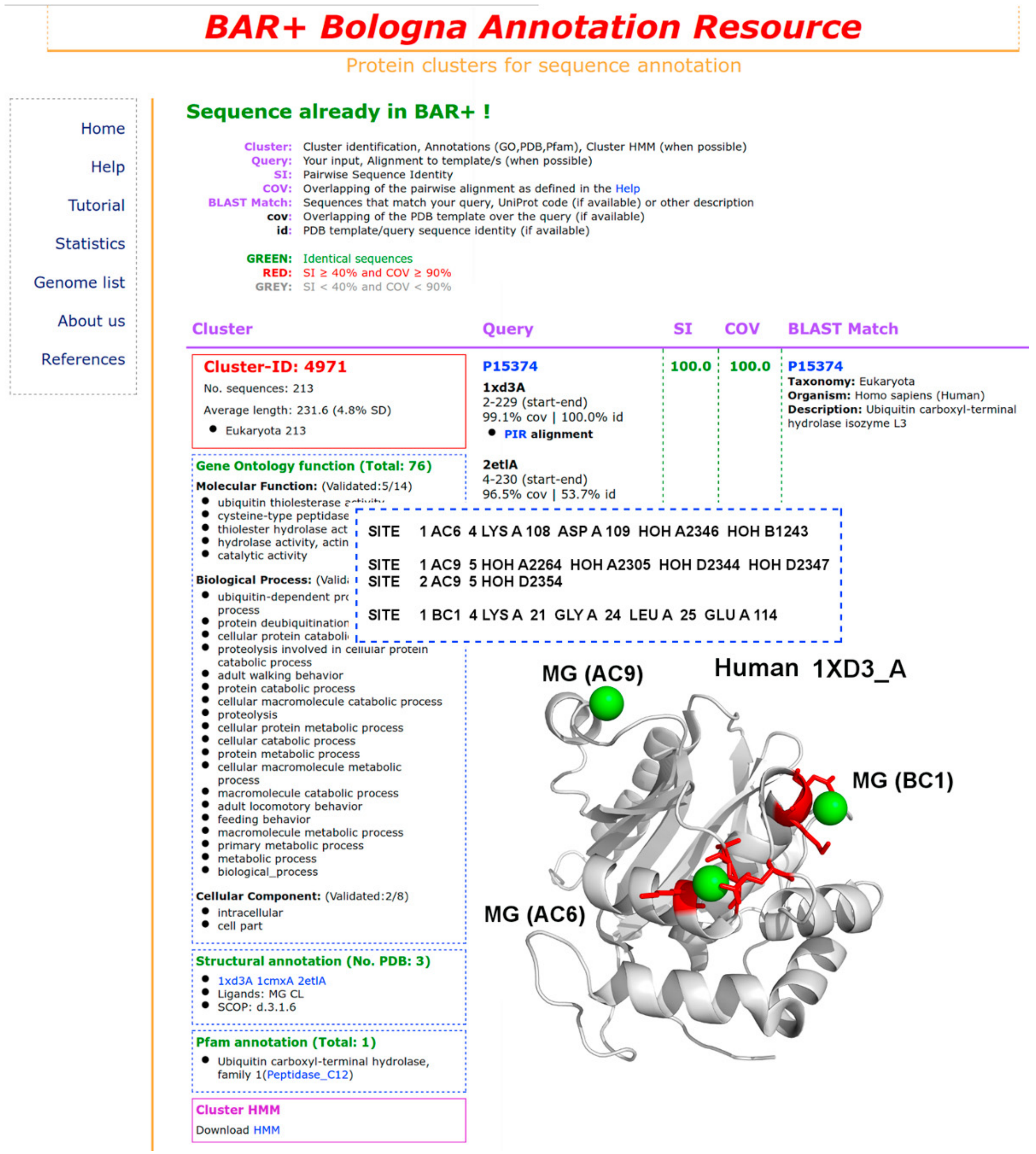
Out of the above selected 61,106 human sequences, we focused on the subset that comprises all the chains included in 251 clusters endowed with magnesium containing PDB structures. In our clusters, we deal with 1,341 PDBs. We therefore checked all the PDB files, the corresponding UniProtKB files and the related literature. From this effort we were able to verify that for only 119 structures (9% of the total) in 21 clusters there is no published observation supporting so far any functional or structural role of MG. Within the clusters, sequences could also safely inherit validated Pfam functional domains and GO functional terms (Molecular Function, Biological Process and Cellular Component, http://www.geneontology.org/).

Binding positions were transferred from the template/s to the target after pair-wise alignment/s based on the cluster HMM. 251 clusters contain Mg binding templates and there from an equivalent number of HMM models were used to transfer Mg binding position/s to the human sequences in the clusters. 141 clusters contain 827 magnesium binding protein structures derived from non human species (25 different Eukaryota, 42 different bacteria, 9 different Archaea and 1 virus). 110 clusters contain 514 human templates.

Results and discussion

Finding Magnesium binding sites with BAR+

When a human sequence has a counterpart in BAR+ with sequence identity ≥ 40% over at least 90% of the alignment length, it falls into the same cluster of the similar chain. In the example of Figure 1, when human sequence P09936 is aligned towards the BAR+ data base, the result web page identifies cluster #4791 that comprises 213 sequences from Eukaryotes with an average length of 232 residues (Standard Deviation (SD)=4.8%) and 3 PDB structures with magnesium and chloride ions as ligands (1CMX_A from *Saccharomyces cerivisiae*; 2ETL_A and 1XD3_A from *Homo sapiens*). The three templates are however highly similar (the average root mean square deviation is 1.62+/-0.35Å). Here we focus only on magnesium binding sites and for clarity we show only the structure of the human Ubiquitin hydrolase UCH-L3 (1XD3_A). As shown, the structure contains 3 Mg ions. The Site field of the corresponding PDB file indicates that of the three Magnesium ions one is coordinated only by water molecules and it is not considered in our analysis. The remaining two are coordinated by four and two residues, respectively (the remaining coordination sites are probably occupied by water). With the cluster HMM based alignment only the coordination sites including residues of the template/s are transferred to the human sequences falling into the cluster. From the cluster, the human sequence inherited all the validated features that are reported in the corresponding web page: validated GO terms, the SCOP classification, and the Pfam domain PF01088 (Ubiquitin carboxyl-terminal hydrolase, family 1). BAR+ gives the HMM based target/



BAR+ Bologna Annotation Resource

Protein clusters for sequence annotation

Home

Help

Tutorial

Statistics

Genome list

About us

References

Sequence already in BAR+ !

Cluster: Cluster identification, Annotations (GO,PDB,Pfam), Cluster HMM (when possible)
Query: Your input, Alignment to template/s (when possible)
SI: Pairwise Sequence Identity
COV: Overlapping of the pairwise alignment as defined in the Help
BLAST Match: Sequences that match your query, UniProt code (if available) or other description
cov: Overlapping of the PDB template over the query (if available)
id: PDB template/query sequence identity (if available)

GREEN: Identical sequences
RED: SI ≥ 40% and COV ≥ 90%
GREY: SI < 40% and COV < 90%

Cluster	Query	SI	COV	BLAST Match
Cluster-ID: 4971 No. sequences: 213 Average length: 231.6 (4.8% SD) ● Eukaryota 213	P15374 1xd3A 2-229 (start-end) 99.1% cov | 100.0% id ● PIR alignment 2etlA 4-230 (start-end) 96.5% cov | 53.7% id	100.0	100.0	P15374 Taxonomy: Eukaryota Organism: Homo sapiens (Human) Description: Ubiquitin carboxyl-terminal hydrolase isozyme L3

Gene Ontology function (Total: 76)

Molecular Function: (Validated:5/14)

- ubiquitin thiolesterase activity
- cysteine-type peptidase
- thiolester hydrolase act
- hydrolase activity, actin
- catalytic activity

Biological Process: (Valid

- ubiquitin-dependent pr process
- protein deubiquitination
- cellular protein cataboli
- proteolysis involved in cellular protein catabolic process
- adult walking behavior
- protein catabolic process
- cellular macromolecule catabolic process
- proteolysis
- cellular protein metabolic process
- cellular catabolic process
- protein metabolic process
- cellular macromolecule metabolic process
- macromolecule catabolic process
- adult locomotory behavior
- feeding behavior
- macromolecule metabolic process
- primary metabolic process
- metabolic process
- biological_process

Cellular Component: (Validated:2/8)

- intracellular
- cell part

Structural annotation (No. PDB: 3)

- 1xd3A 1cmxA 2etlA
- Ligands: MG CL
- SCOP: d.3.1.6

Pfam annotation (Total: 1)

- Ubiquitin carboxyl-terminal hydrolase, family 1(Peptidase_C12)

Cluster HMM

Download HMM

```
SITE     1 AC6  4 LYS A 108  ASP A 109  HOH A2346  HOH B1243

SITE     1 AC9  5 HOH A2264  HOH A2305  HOH D2344  HOH D2347
SITE     2 AC9  5 HOH D2354

SITE     1 BC1  4 LYS A  21  GLY A  24  LEU A  25  GLU A 114
```

Figure 1 A BAR+ cluster with a magnesium binding template. The BAR+ output. When the query is the UniProtKB accession code P15374, the corresponding annotation cluster comprises 213 sequences from Eukaryotes with an average length of 232 residues and 3 PDB structures. Only one of them (human Ubiquitin hydrolase UCH-L3, PDB:1XD3_A) is shown using PyMol (http://www.pymol.org) with the three Mg ions. Of the three ions (as shown in the inset where the PDB SITE fields are reported) only two are coordinated by lateral side chains (in red in the protein structure representation). The cluster contains 26 validated GO terms and 1 validated Pfam term (PF01088, Ubiquitin carboxyl-terminal hydrolase, family 1) that are also inherited by the human query sequence. See text for details.

template alignment for computational modelling of the 3D structure of all the other sequences in the cluster. Among these, 4 are from *Homo sapiens* and inherit all the cluster specific annotation, including the Mg binding sites.

Bound Mg in this structure is not as yet supported by any experimental observation highlighting a specific functional role. The whole BAR-hMG data base contains 21 out of 251 clusters with templates binding Mg without any experimental (still) determined functional or structural role. This information can be retrieved for each template from the corresponding PDB and UniProtKB files and the quoted literature therein. It should be considered that Mg ions may play a role on protein stability still not fully described or even a role in protein-protein interaction that is at the basis of many relevant biological processes. In many instances the formation of protein complexes has not yet been recognized due to its transient characteristics. Therefore the question is still open and we therefore included also these cases in our data set for a comprehensive analysis of putative Mg binding sites. Clusters containing templates where Mg has a documented structural and functional role are labelled with a yellow star, and a yellow star and the corresponding EC number, respectively. For this reason no label is present in the figure.

Annotation of Mg^{2+} binding sites in human proteins

A structural analysis of the magnesium containing 1,341 PDB templates indicates that the ion can be present in different ways. For this reason we list our annotation results considering that the ion co-crystallises with the protein chain either alone (Mg) or concomitantly with other ions (Mg and Ions) or ligands (Mg and Ligands) or with other ions and ligands (MG, Ions and Ligands). In some instances PDB structures can combine two or more of the binding modes (Mixed). Results are listed by splitting human sequences that inherited annotation from human templates (2,688) from those that inherit annotation from structures of other organisms (1,063). The results are shown in Table 1 and 2, respectively, where the number of sequences with low sequence identity to the cluster templates is also reported. Clusters are split depending on the role of bound Mg ion: functional, structural, not yet determined.

The number of PDB human protein structures with bound magnesium (514) univocally identifies 172 template sequences; within the BAR+ environment this number reaches 2,688 (Annotation inherited from human templates). Some other 1,063 human sequences inherit annotation within BAR+ clusters where the structural templates are from other organisms (Table 2) (Annotation inherited from other organisms).

When more PDB structures fall into the same cluster (Table 1 and 2) their RMSDs are very low (<1 Å) for all the groups. This indicates that the BAR+ clusters preserve the structural specificity. Therefore when a target sequence falls into a cluster characterised by Mg binding, the corresponding site annotation can be safely inherited. This is so also for very distantly related sequences (sequence identity <30%, last column) that are in the same cluster.

In BAR-hMG some 3,751 human sequences are annotated as Mg binding. About 98% of this set is annotated for the first time. For these sequences the corresponding UniProtKB entry neither has any information on Mg binding nor contains any GO term related to Mg binding.

Characteristics of Mg^{2+} binding sites can be detected from a simple counting on the retrieved 1,341 PDB structures contained in the 251 clusters of the BAR-hMG data base. Results (shown in Figure 2) are split

Table 1 Human sequences annotated with human structural templates

	Cluster (#)			PDB (#)	Cluster RMSD (Å)	Template sequence (#)	Annotated sequence (#)	Newly annotated sequence (#)	Annotated sequence (ID<30%)*
	$	^	°						
Mg	8	1	0	9	-	9	55	54	1
Mg and Ions	7	1	0	9	0.30	8	53	52	6
Mg and Ligands	24	4	2	73	0.77	32	159	158	33
Mg , Ions and Ligands	22	5	4	57	0.52	31	1948	1947	19
Mixed	22	6	4	366	0.68	92	473	455	120
Total	83	17	10	514		172	2688	2666	179

Human sequences that inherit annotation from human structural templates are listed as a function of the different typologies of magnesium binding in the PDB files. The table lists the number of clusters, of structural templates, of annotated sequences (sequences that inherit Mg binding positions) according to our procedure, of sequences never annotated before as Mg binding proteins according to UniProtKB and of *sequences annotated when the target/template identity is below the 30%. Three different types of clusters are identified and listed in the first column: $ cluster with structures binding MG with a recognized functional role and whit an EC number, ^ clusters with structures binding MG with a recognized structural role (without an EC number), ° cluster containing structures (119 out of 1,341) binding MG without recognized physiological role.

Table 2 Human sequences annotated with structural templates from other organisms

	Cluster (#)			PDB (#)	Cluster RMSD (Å)	Template sequence (#)	Annotated sequence (#)	Newly annotated sequence (#)	Annotated sequence (ID<30%)*
	$	^	°						
Mg	12	10	0	75	0.73	33	105	105	24
Mg and Ions	5	5	0	160	0.38	10	51	50	22
Mg and Ligands	20	22	3	81	0.86	54	359	352	51
Mg , Ions and Ligands	12	6	2	66	0.52	23	278	276	28
Mixed	21	17	6	445	0.83	95	270	243	66
Total	70	60	11	827		215	1063	1026	191

Table legend is as in Table 1.

into binding sites stabilised by lateral side chains and by backbone carbonyl groups. The highest frequency is observed for Asp and Glu residues. Similar frequency distribution is obtained when counting is done on the newly annotated human sequences (Figure 2). Here binding is referred only to the residue type.

Localising the human Mg^{2+} binding sequences

In Table 3 we list the most populated cellular localizations (Cellular Component of the Gene Ontology) of the human sequences (the "human magnesome") sorted out according to the different magnesium binding modes. For each GO term, the number of human sequences is reported. The selected terms are those that are the most distant from the ontology root in the corresponding BAR+ cluster of each sequence. Similarly GO terms of biological process and molecular function can be obtained for each sequence (data not shown; the data can be retrieved when a sequence falls into a validated cluster).

The "Human Magnesome" database

The "Human Magnesome" is a data base of human sequences generated after annotation with the procedure here described. The main page allows a sequence search either with a UniprotKB accession code or the FASTA format of the sequence. When the sequence is present in the database it is returned with the putative magnesium binding sites, the structural templates from where it inherits magnesium binding and the number of magnesium ions present in the structural templates. Different colors are displayed when the binding residues are identical, similar or different to the template reference/s. Residue substitution is scored with Blosum62 matrix. In Figure 3 a typical output is shown. The data base is available at http://bar.biocomp.unibo.it/mg.

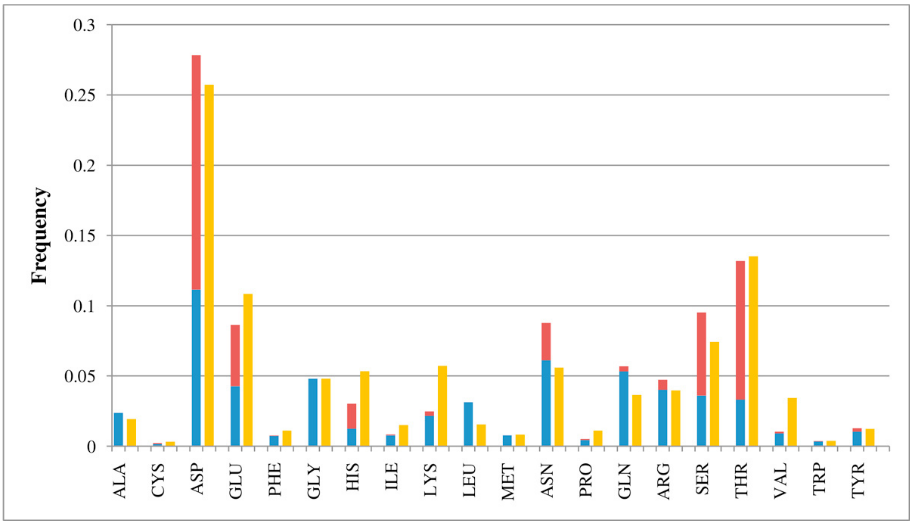


Figure 2 Frequency distribution of Magnesium binding residues in PDB templates and in annotated human sequences. Distribution of the frequency of residues coordinating magnesium ions in the PDB structures (1,341, blue color: Mg is coordinated by the backbone carbonyl oxygen, red color: Mg is coordinated by the lateral side chain) and in the putatively annotated human sequences (3,751, yellow color).

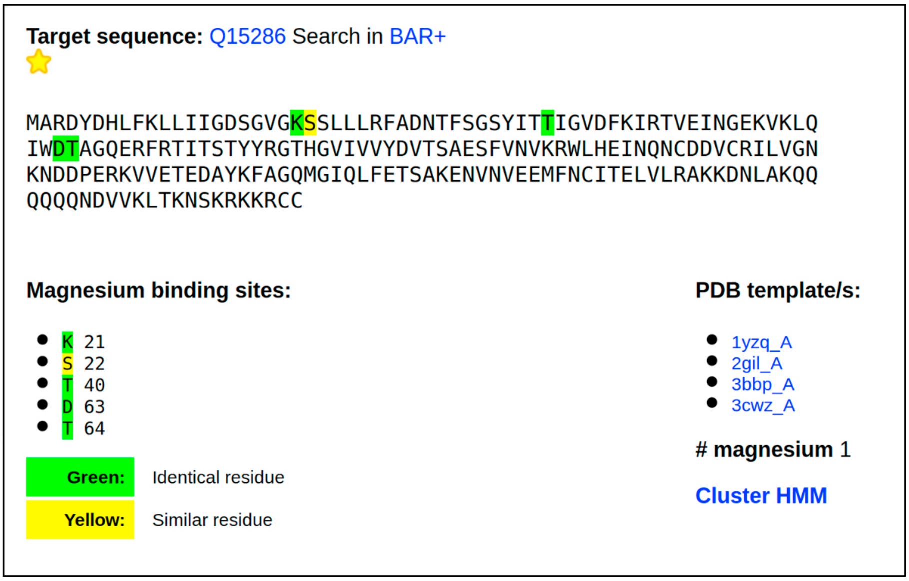
Table 3 Localising the human magnesium binding sequences

Sequence (#)	GO terms (Cellular Component)	Sequence (#)	GO terms (Cellular Component)
	Mg		**Mg + Ions + Ligands**
23	endoplasmic reticulum lumen	1817	cell surface
21	cell body	117	endoplasmic reticulum part
	Mg + Ions	92	dendrite cytoplasm
33	site of polarized growth	56	mitochondrial matrix
13	membrane-bounded organelle	48	cell division site
	Mg + Ligands	47	ruffle
118	azurophil granule	44	cell septum
37	cytoplasmic mRNA processing body	44	membrane raft
19	cytoplasmic membrane-bounded vesicle	37	endoplasmic reticulum
16	intracellular	24	cell leading edge
15	intracellular membrane-bounded organelle	23	plasma membrane enriched fraction
14	mitochondrion	22	internal side of plasma membrane
11	neuron projection	15	cell cortex
11	cell part	15	intracellular membrane-bounded organelle

For explanation see text.

Conclusion

In this work we address the problem of annotating magnesium binding sites in proteins starting from their sequence. We take advantage of an annotation resource recently introduced (BAR+, [13]), where functional and structural features derived from PDB structures are implemented into HMM models that allows sequence to template alignment even when sequence identity is below 30%. This procedure is based on the notion of "cluster", a set of sequences retrieved as connected components of a graph where two proteins are linked together when they share a sequence identity greater or equal than 40% in at least 90% of the pair wise alignment length. By restricting our analysis to clusters containing human sequences and magnesium binding PDB structures, we align with the cluster HMMs some 3,751



Target sequence: Q15286 Search in BAR+

⭐

```
MARDYDHLFKLLIIGDSGVGKSSLLLRFADNTFSGSYITTIGVDFKIRTVEINGEKVKLQ
IWDTAGQERFRTITSTYYRGTHGVIVVYDVTSAESFVNVKRWLHEINQNCDDVCRILVGN
KNDDPERKVVETEDAYKFAGQMGIQLFETSAKENVNVEEMFNCITELVLRAKKDNLAKQQ
QQQQNDVVKLTKNSKRKKRCC
```

Magnesium binding sites:

- K 21
- S 22
- T 40
- D 63
- T 64

Green:	Identical residue
Yellow:	Similar residue

PDB template/s:

- 1yzq_A
- 2gil_A
- 3bbp_A
- 3cwz_A

# magnesium 1

Cluster HMM

Figure 3 The Human Magnesome output. A typical output of the human magnesome site (BAR-hMG). The test sequence inherits, after cluster based HMM alignment to the corresponding templates (listed in the inset), five binding residues (K 21, S 22, T 40, D 63 and T 64). The residues are color coded depending on the BLOSUM 62 alignment scoring matrix. From the result page is also possible to retrieve the matching BAR+ cluster page and the corresponding UniProtKB page of the target entry. The green color in the output indicates residues identical to the original template/s. Similar residues are highlighted in yellow. The yellow star indicates that the protein is located in a cluster where Mg binds to PDB templates (listed) in a documented structural way. Cluster HMM can be downloaded.

human sequences that fall in the same clusters and inherit by this the magnesium binding feature. Some 370 human sequences share an identity to the template less than 30%.

We therefore prove feasible that magnesium binding sites can be inherited from a given template when the sequence falls inside a well annotated cluster from where it derives also validated Pfam functional domains and GO functional terms. Presently we can annotate some 5% of the human genome as inheriting the capability of binding magnesium ions. All the analysed sequences, their binding sites, and the corresponding clusters from where they derive annotation are included in the Human Magnesome data set (BAR-hMG), freely available at http://bar.biocomp.unibo.it/mg.

Acknowledgements
RC thanks the following grants: PRIN 2009 project 009WXT45Y (Italian Ministry for University and Research: MIUR), COST BMBS Action TD1101 (European Union RTD Framework Programme), and PON project PON01_02249 (Italian Ministry for University and Research: MIUR). DP is a recipient of a PHD fellowship from the Ministry of the Italian University and Research. GP is a recipient of a research contract from Health Science and Technologies-ICIR.
This article has been published as part of *BMC Bioinformatics* Volume 13 Supplement 14, 2012: Selected articles from Research from the Eleventh International Workshop on Network Tools and Applications in Biology (NETTAB 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/13/S14

Author details
[1]Biocomputing Group, Department of Biology, University of Bologna, Bologna, 40126, Italy. [2]Health Science and Technologies-ICIR, University of Bologna, Bologna, 40126, Italy.

Authors' contributions
DP carried out all the calculations. GP developed the web site. RC, DP, GP and PM conceived the study, analyzed the data and wrote the manuscript. All the authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 7 September 2012

References

1. Rude RK, Singer FR, Gruber HE: **Skeletal and hormonal effects of magnesium deficiency.** *J Am Coll Nutr* 2009, **28(2)**:131-141[http://www.jacn.org/content/28/2/131.long].
2. Belin RJ, He K: **Magnesium physiology and pathogenic mechanisms that contribute to the development of the metabolic syndrome.** *Magnes Res* 2007, **20(2)**:107-129.
3. Bertini I, Gray HB, Stiefel EI, Valentine EI: *Biological Inorganic Chemistry: Structure and Reactivity* Sausalito (CA): University Science Books; 2007.
4. Cowan JA: **Metal Activation of Enzymes in Nucleic Acid Biochemistry.** *Chem Rev* 1998, **98(3)**:1067-1088.
5. Bashton M, Nobeli I, Thornton JM: **PROCOGNATE: a cognate ligand domain mapping for enzymes.** *Nucleic Acids Res* 2007, **36**:D618-D622.
6. Dudev T, Cowan JA, Lim C: **Competitive Binding in Magnesium Coordination Chemistry: Water versus Ligands of Biological Interest.** *J Am Chem Soc* 1999, **121(33)**:7665-7673.
7. Dudev T, Lim C: **Metal Selectivity in Metalloproteins: Zn^{2+} vs Mg^{2+}.** *J Phys Chem B* 2001, **105(19)**:4446-4452.
8. Cowan J: **Structural and catalytic chemistry of magnesium-dependent enzymes.** *Biometals* 2002, **15(3)**:225-235.
9. Zaychikov E, Martin E, Denissova L, Kozlov M, Markovtsov V, Kashlev M, Heumann H, Nikiforov V, Goldfarb A, Mustaev A: **Mapping of Catalytic Residues in the RNA Polymerase Active Center.** *Science* 1996, **273(5271)**:107-109.
10. Joyce CM, Steitz TA: **Function and Structure Relationships in DNA Polymerases.** *Annu Rev Biochem* 1994, **63**:777-822.
11. Dudev M, Lim C: **Discovering structural motifs using a structural alphabet: Application to magnesium-binding sites.** *BMC bioinformatics* 2007, **8(1)**:106.
12. Bartoli L, Montanucci L, Fronza R, Martelli PL, Fariselli P, Carota L, Donvito G, Maggi GP, Casadio R: **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* 2009, **8**:4362-4371.
13. Piovesan D, Martelli PL, Fariselli P, Zauli A, Rossi I, Casadio R: **BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences.** *Nucleic Acids Res* 2011, **39(Web Server issue)**:W197-W202.
14. Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J: **Ligand Depot: a data warehouse for ligands bound to macromolecules.** *Bioinformatics* 2004, **20(13)**:2153-2155.
15. Berman HM, Henrick K1, Nakamura H: **Announcing the worldwide Protein Data Bank.** *Nat Struct Biol* 2003, **10(12)**:98.
16. The UniProt Consortium: **Ongoing and future developments at the Universal Protein Resource.** *Nucleic Acids Res* 2011, **39**:D214-D219.
17. Velankar S, McNeil P, Mittard-Runte V, Suarez A, Barrell D, Apweiler R, Henrick K: **E-MSD: an integrated data resource for bioinformatics.** *Nucleic Acids Res* 2005, , **33 Database:** D262-D265.
18. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

doi:10.1186/1471-2105-13-S14-S10
Cite this article as: Piovesan *et al.*: The human "magnesome": detecting magnesium binding sites on human proteins. *BMC Bioinformatics* 2012 **13**(Suppl 14):S10.

BMC Bioinformatics

PROCEEDINGS Open Access

How to inherit statistically validated annotation within BAR+ protein clusters

Damiano Piovesan[1,2], Pier Luigi Martelli[1,2], Piero Fariselli[1,3], Giuseppe Profiti[1,4], Andrea Zauli[5], Ivan Rossi[5], Rita Casadio[1,2,4*]

From Automated Function Prediction SIG 2011 featuring the CAFA Challenge: Critical Assessment of Function Annotations
Vienna, Austria. 15-16 July 2011

Abstract

Background: In the genomic era a key issue is protein annotation, namely how to endow protein sequences, upon translation from the corresponding genes, with structural and functional features. Routinely this operation is electronically done by deriving and integrating information from previous knowledge. The reference database for protein sequences is UniProtKB divided into two sections, UniProtKB/TrEMBL which is automatically annotated and not reviewed and UniProtKB/Swiss-Prot which is manually annotated and reviewed. The annotation process is essentially based on sequence similarity search. The question therefore arises as to which extent annotation based on transfer by inheritance is valuable and specifically if it is possible to statistically validate inherited features when little homology exists among the target sequence and its template(s).

Results: In this paper we address the problem of annotating protein sequences in a statistically validated manner considering as a reference annotation resource UniProtKB. The test case is the set of 48,298 proteins recently released by the Critical Assessment of Function Annotations (CAFA) organization. We show that we can transfer after validation, Gene Ontology (GO) terms of the three main categories and Pfam domains to about 68% and 72% of the sequences, respectively. This is possible after alignment of the CAFA sequences towards BAR+, our annotation resource that allows discriminating among statistically validated and not statistically validated annotation. By comparing with a direct UniProtKB annotation, we find that besides validating annotation of some 78% of the CAFA set, we assign new and statistically validated annotation to 14.8% of the sequences and find new structural templates for about 25% of the chains, half of which share less than 30% sequence identity to the corresponding template/s.

Conclusion: Inheritance of annotation by transfer generally requires a careful selection of the identity value among the target and the template in order to transfer structural and/or functional features. Here we prove that even distantly remote homologs can be safely endowed with structural templates and GO and/or Pfam terms provided that annotation is done within clusters collecting cluster-related protein sequences and where a statistical validation of the shared structural and functional features is possible.

Background

When a new protein sequence becomes available the problem of its annotation poses. Most of our expertise in trying to endow the new sequence with structural and functional features is based on similarity search [1-4]. Methods are mainly based on the knowledge that structure is more conserved than sequence through evolution and that structural alignment is conserved as long as sequence identity (SI) is ≥ 30% over the alignment length. This was observed originally by Chothia and Lesk [5] and once in a while revisited at increasing number of proteins solved with atomic resolution and deposited in the Protein Data Bank (PDB) [6]. The observation is at the basis of

* Correspondence: casadio@biocomp.unibo.it
[1]Bologna Biocomputing Group, University of Bologna, Italy
Full list of author information is available at the end of the article

one of the most popular method for computing the three dimensional structure of the target on a template, when found, after a sequence similarity search against the PDB [7]. Recently maps of the protein structure space have revealed fundamental relationship between protein structure and function [8]. When a target sequence well aligns with a template of known structure, its functional properties can be derived on the basis of structural conservation. Proteins sharing some 40-60% of sequence identity are likely to share also similar function [9,10].

However a problem is at hand: how to recognize structural and functional templates when sequence identity is below 30%. In this case proteins are categorized to be distantly related to their homologous counterparts, since they may perform the same function, and possibly be endowed with the same structure although sharing very little sequence homology [11,12]. To this purpose methods have developed trying to grasp local sequence conservation by modeling protein conserved structural and functional domains. The most popular is Pfam ([13], http://pfam.sanger.ac.uk). In this case function can be inferred when a protein is significantly retained by a specific Pfam model that is again based on a local sequence-to-profile alignment and its scoring. SUPERFAMILY (http://supfam.cs.bris.ac.uk/SUPERFAMILY), based on hidden Markov models as Pfam, has been recently modified to address specifically the problem of function assignment by including a domain-based Gene Ontology [14].

When function is to be assigned only on the basis of sequence, the problem still remains unsolved, since very little is known on the relationship among sequence similarity and transfer of function [1,9]. Functions can be described with specific terms following the Gene Ontology vocabulary and comprising three main functional branches: Molecular Function (MFO), Biological Process (BPO), and Cellular Component (CCO) [15]. UniProtKB, the largest resource of protein sequences curates automatically annotated protein records ([16], http://www.uniprot.org/help/biocuration). Here annotation integrates previous knowledge on protein structure and function from various sources, when available, again mainly based on sequence similarity search (UniProtKB/TrEMBL). Eventually the records are manually curated (UniProtKB/SwissProt). However out of the over 18 millions sequence entries presently available (Release 2011_12 of 14-Dec-2011), 75% are proteins inferred by homology or predicted whose features in most instances are far from being attributed even with computational methods.

Several methods have been developed to predict protein function from structures and sequences trying to infer features from selected and well annotated sets of proteins by mean of different computational approaches, including machine learning, and generally aiming at integrating different source of information (see for recent reviews [17,18]).

Here we take advantage of the recently released set of proteins selected by CAFA (http://biofunctionprediction.org/) for function prediction in order to discuss how inheritance of annotation can be statistically validated. Validation is indeed an added value to the annotation process, when possible. For this we developed BAR+ [19,20], a non hierarchical clustering annotation procedure that allows different types of annotation by means of a cluster-mediated transfer of annotation. We also show that our method allows a gain of annotation over a direct Pfam prediction and GOA electronic annotation (http://www.ebi.ac.uk/GOA/).

Databases and methods

Databases

The test set includes 48,298 sequences made available during the 2011 CAFA experiment (CAFA set, http://biofunctionprediction.org). 41,003 sequences of this set (85% of the CAFA set) could be mapped towards UniProtKB Release 2010_05 (CAFA/UniProtKB set); 96% of the CAFA/UniProtKB set were manually curated (UniProtKB/SwissProt) and 2,047 proteins have also a PDB structure; 13,684 of the set are proteins inferred from homology and predicted. We found that 44,495 sequences of the CAFA set (92% of the CAFA set) could be mapped into BAR+ (CAFA/BAR+ set).

BAR+

BAR+, the Bologna Annotation Resource, is our annotation system (BAR+ is available at http://bar.biocomp.unibo.it/bar2.0/). BAR+ allows transfer of validated annotation [19,20]. The method relies on the concept that sequences can inherit the same function/s and structure from their counterparts, provided that they fall into a cluster endowed with validated annotations. BAR+ is based on a clustering procedure with the constraint that sequence identity (SI) is ≥ 40% on at least 90% of the pairwise alignment overlapping (Coverage, Cov). Clusters in BAR+, as previously reported [20], allow 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 (see below). Depending on the categories of annotation in the cluster and provided that they are statistically validated, all new targets that fall into a cluster can inherit statistically validated annotations by transfer.

For generating BAR+ clusters we analyzed a total of over 13 million protein sequences from 988 genomes and UniProtKB release 2010_05. The BAR+ cluster building pipeline starts with an all-against-all sequence comparison with BLAST in a GRID environment [19]. The alignment results are then regarded as an undirected graph where nodes are proteins and links are allowed only among chains that are 40% identical over at least 90% of the alignment length. All the connected nodes fall within the same cluster; when a cluster incorporates a UniProtKB entry, it inherits its annotations (GO and Pfam terms, PDB structures, SCOP classifications). Within a cluster GO and Pfam terms are statistically validated by means of a procedure that includes P-value evaluation with a Bonferroni correction and estimate of the significance threshold value after a bootstrapping procedure [19]; validated terms are those endowed with P-values< 0.01[19]. Clusters can contain distantly related proteins that therefore can be annotated with high confidence and eventually can also inherit a structural template, if present. In BAR+, when PDB templates are present within a cluster profile HMMs (Hidden Markov Models) are computed on the basis of sequence-to-structure alignment and are cluster associated (Cluster-HMM) [20].

Results and discussion

BAR+ contains clusters with statistically validated annotation

70% of the 13,495,736 sequences of BAR+ are collected in 913,762 clusters (the number of sequences in a cluster ranges from 2 to 87,893). Interestingly 87% of the clusters contain sequences whose standard deviation of the protein length is ≤ 5 residues. 1.2% of the clusters, containing 23% of the whole set, contains also PDB structures and is endowed with a cluster specific structural HMM [20]. 30% of the sequences are singletons that eventually can carry along structural and/or functional information.

A cluster collects specific Pfam and GO terms directly from the corresponding UniProtKB protein sequence files. Validation of the terms within a cluster is based on a Bonferroni corrected P-value analysis [19]. We performed a statistical evaluation of the P-values by computing the statistical significance of Pfam and GO terms associated to each cluster and by adopting a bootstrapping procedure. By this procedure we determine the threshold at which significance is different from random and we define a P-value equal to 0.01 as the discriminative value for a single term to be validated or not (see also [19]). In Figure 1 the number of clusters is reported as a function of the corresponding Bonferroni corrected P-value for Pfam and the GO terms of the three main roots. The threshold level discriminates among clusters with statistically validated and not validated annotation. 11% of the clusters have one validated GO term allowing in the present version of BAR+ 45% of the total number of sequences (13,495,736) to be included in clusters endowed with validated terms.

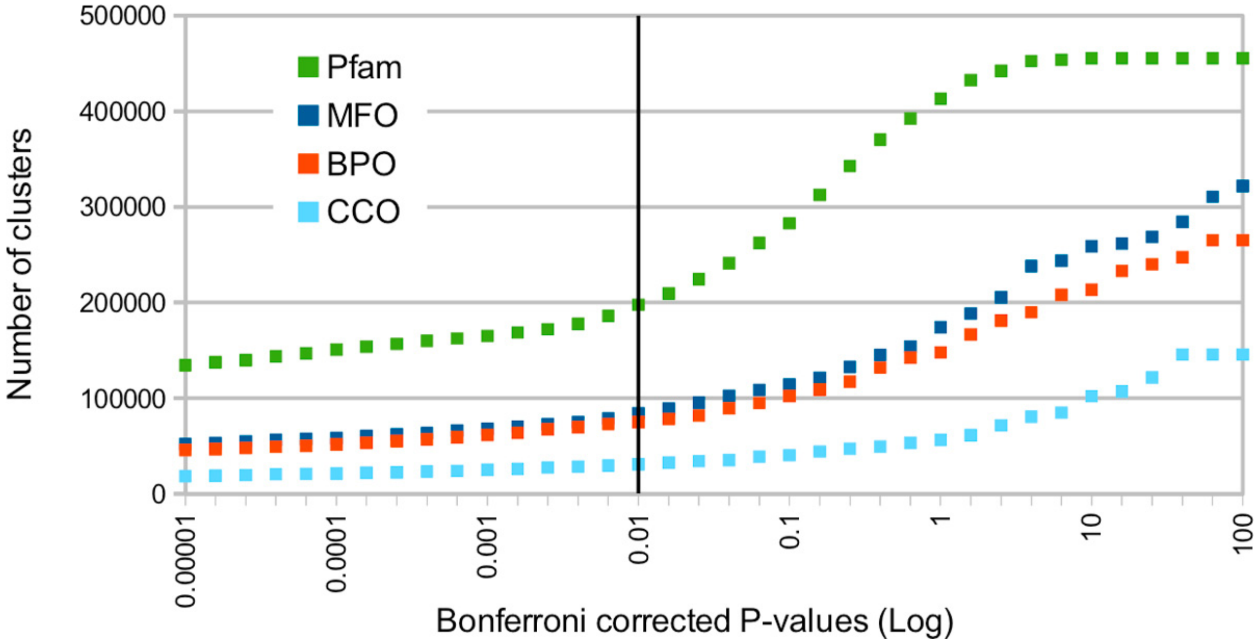


Figure 1 Discriminating among validated and not validated BAR+ clusters. The number of clusters containing GO terms of three main roots and Pfam terms is reported as a function of the Bonferroni-corrected P-value. The black vertical line sets the boundary among validated and not validated terms. It can be proven (data not shown) that that a P-value ≤ 0.01 is a discriminative value good enough to discriminate among the real and the random distribution of each type of GO and Pfam terms (for mathematical details see [15]. Green colour: Pfam terms; Blue colour: Molecular Function (MFO); Red colour: Biological Process (BPO); Pale blue: Cellular Component (CCO). For the different curves the number of validated clusters as compared to the total number of BAR+ clusters is: Pfam 197,826/455,309; MFO 84,506/321,748; BPO 75,147/265,164; CCO 31,042/145,677. The total number of cluster with at least a GO validated term is 100,791.

Within BAR+, inheritance of validated annotation is possible only when a given sequence after alignment towards BAR+ finds a counterpart whose Sequence Identity (SI) is ≥ 40% over at least 90% of the pairwise alignment overlapping (Coverage, Cov).

Inheritance of statistically validated annotation

We aligned all the CAFA target sequences against BAR +clusters. More than 92% of the CAFA set was retained by BAR+ (CAFA/BAR+ set), including singletons (stand alone sequences in BAR+). The statistically validated annotations transferred within BAR+ clusters, including Pfam terms and PDB templates (SI≥ 40% and Cov≥ 90%) of the CAFA/ BAR+ set are detailed in Table 1. The set of CAFA sequences that received a statistically validated annotation (ALL-O OR Pfam in Table 1) includes 37,516 sequences (77.7% of the CAFA set). The list of predicted proteins is grouped by different target sets including sequences from Eukaryotes, Prokaryotes and "Unknown" organisms. In Table 1 annotations are sorted out by the three different types of GO ontologies and Pfam terms. Values relative to sequences endowed with the union of different ontologies is also shown (MFO OR BPO; ALL-O).

For sake of exploring the relevance of the alignment length on the annotation system, we decreased the Cov value to ≥ 70%) while keeping SI≥ 40%. In this case the number of annotated CAFA targets increased by only 3% (Table 1), suggesting that the original 90% Cov value together with SI≥ 40% ensures that most of the CAFA set is already retained within validated clusters.

With our method it is also possible to model distantly related targets that fall into a cluster by aligning them to the template/s in the cluster by means of a cluster HMM, as previously described [20]. By this about 25% of the CAFA set inherits also a PDB structural template/s (11,935 sequences, Table 1) and about 50% of these targets share a sequence identity with the template structure of the cluster lower than 30% (12.5% of the CAFA set). Concomitantly the sequence also inherits validated Pfam domains and GO ontologies and this allows a validation of the functional annotation directly on the protein computed structure.

Statistically validated GO ontologies of the three main roots (MFO, BPO and CCO) are differently distributed among Prokaryotic and Eukaryotic sequences of the CAFA/BAR+ set (Figure 2). Here for sake of simplicity we group all the predicted GO ontologies under the first branches of each principal root. In "Binding" main category "Nucleotide binding" (GO:0000166) and "Protein binding" (GO:0005515) are the most represented in Prokaryotes and Eukaryotes, respectively. In "CatalyticActivity", "Transferase activity" (GO:0016740) and "Hydrolase activity" (GO:0016787) are the most represented in Prokaryotes and Eukaryotes, respectively. The most frequently predicted BPO main category is "Cellular process", with "Cellular biosynthetic process" (GO:0044249) for Prokaryotes and "Cellular macromolecule metabolic process" (GO:0044260) for Eukaryotes. Finally for CCO, the most abundant term both in Prokaryotes and Eukaryotes is "Intracellular" (GO:0005622). The data confirm the variety of statistically validated functional annotations that can be retrieved by adopting BAR+ as an annotation resource and also highlight the main functional features that characterize the proteins of the CAFA set sorted out according to Prokaryotes and Eukaryotes.

In Figure 3 the different validated and inherited Pfam terms are grouped into clans, a collection of Pfam similar entries [12] and shown as a function of the number of sequences from Eukaryotes and Prokaryotes. The most populated clan is "P-loop containing nucleoside triphosphate hydrolase superfamily" (CL0023). Within the clan, the most frequent Pfam domains are Ras family (PF00071)

Table 1 Annotating the CAFA set with BAR+

	Cov	*MFO OR BPO*	*MFO*	*BPO*	*CCO*	*ALL-O*	*Pfam*	*ALL-O OR Pfam*	*PDB°*
Eukaryotes	*90%*	20,532	17,389	17,131	16,430	22,733	24,038	26,378	8,054
[32,143]^	*70%*	1,448							
Prokaryotes	*90%*	9,660	8,915	8,202	4,723	9,843	10,772	11,088	5,924
[12,295]^	*70%*	224							
Unknown	*90%*	36	32	32	10	36	50	50	4
[57]^	*70%*	4							
Total		30,228	26,336	25,365	21,163	32,612	34,860	**37,516**	13,982
[44,495]^									2,047*

Cov: Coverage, 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). For both Cov values Sequence Identity (SI) is ≥ 40%. MFO: Molecular Function Ontology; BPO: Biological Process Ontology; CCO: Cellular Component Ontology. ALL-O: number of sequences with predicted MFO *OR* BPO *OR*CCO. Pfam terms. ALL-O OR Pfam: the union of ALL-O and Pfam. °PDB: sequences that inherit a structural template from a cluster HMM within BAR+ [20]. ^ CAFA/BAR+ set sequences from Eukaryotes, Prokaryotes, and Unknown organisms. *Sequences with a corresponding PDB structure.

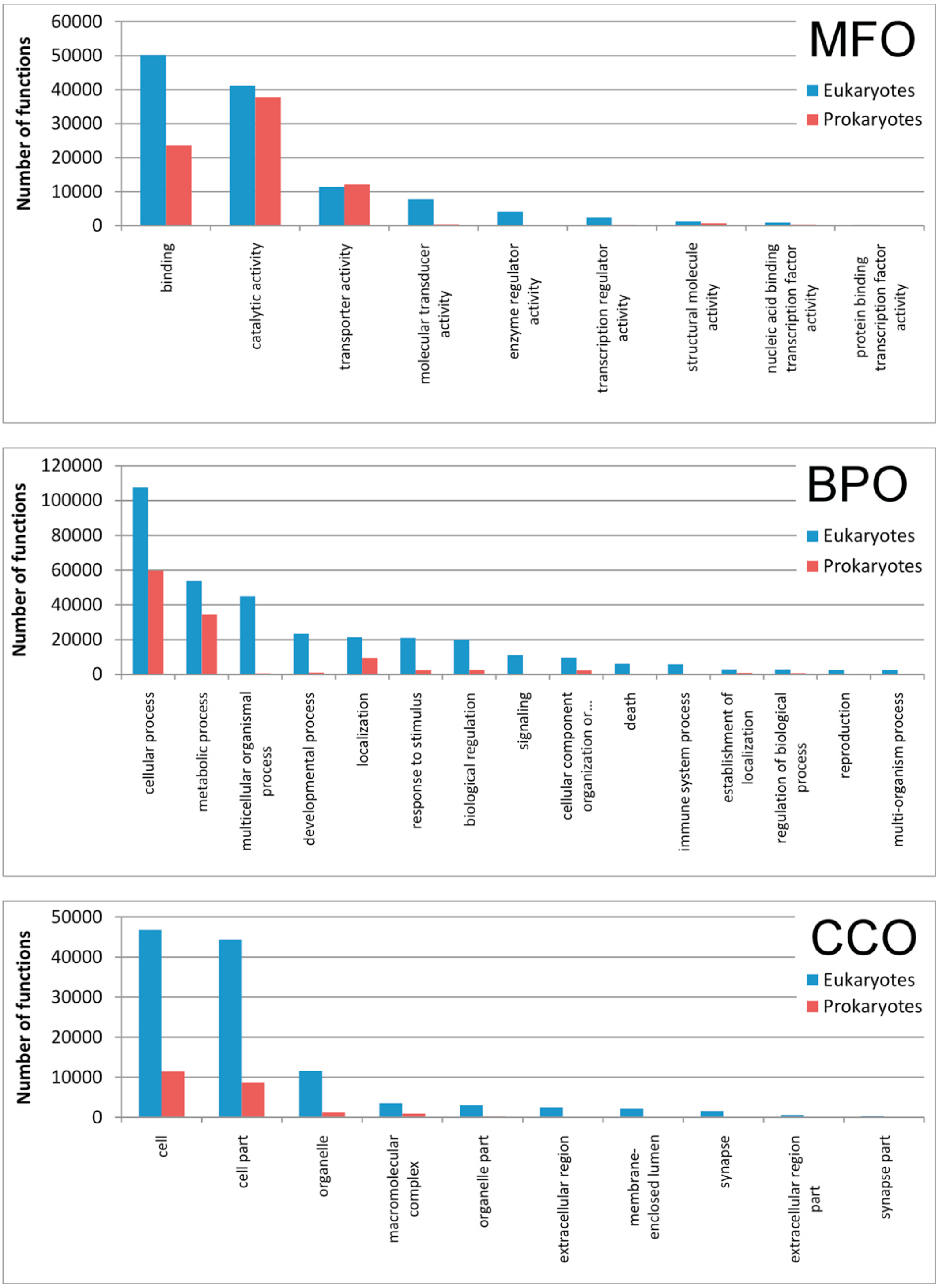


Figure 2 Statistically validated GO ontologies of the CAFA/BAR+ set. Histograms of the main statistically validated GO Molecular Functions (MFO), Biological Processes (BPO), Cellular Component (CCO) ontologies are shown after annotation within validated BAR+ clusters. GO terms are included in main categories and listed with respect to Eukaryotes and Prokaryotes.

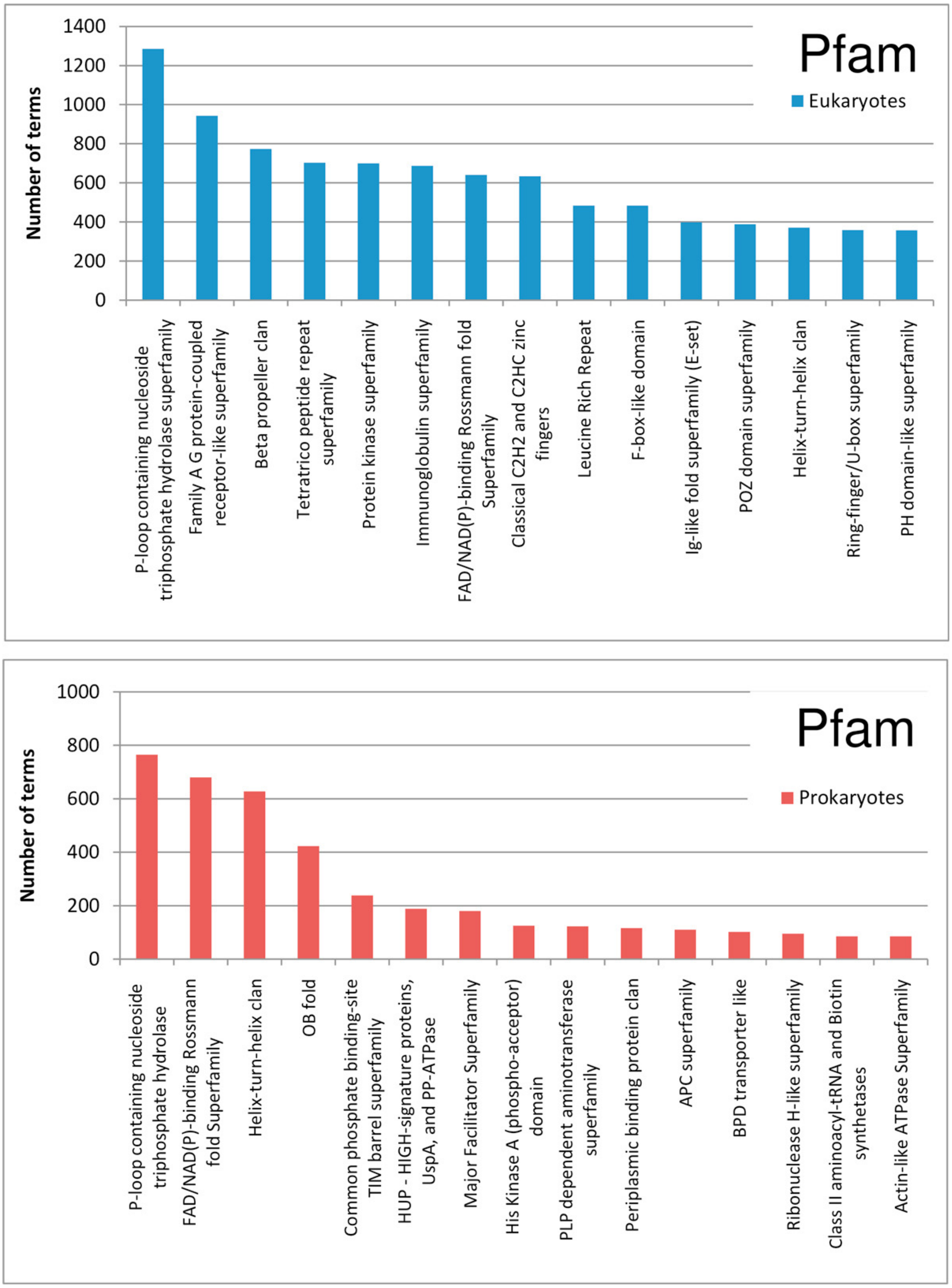


Figure 3 Statistically validated Pfam terms of the CAFA/BAR+ set. Histograms of the most populated clans of Pfam terms are shown after annotation within validated BAR+ clusters. A clan is a collection of Pfam-A entries that are judged likely to be homologous [12]. Clans are sorted out discriminating among Prokaryotes (a) and Eukaryotes (b).

Table 2 Comparing UniProtKB direct annotation with BAR+ annotation

	CAFA/UniProtKB*		BAR+ Validated°		
	Sequences	**Terms**	**Sequences with validated annotation**	**Validated Terms**	**Sequences with new validated annotation**
Total°	34,065	10,628	34,065	13,558	3,659§
Pfam^	30,767	5,293	31,190	5,365	423§
MFO^	20,790	2,048	21,758	2,698	968§
BPO^	19,739	2,719	21,585	4,879	1,846§
CCO^	16,503	568	17,589	616	1,086§
-	-	-	3,451#	5,886#	3,451#
PDB+	2,047+	-	13,084+	-	11,935+

*The CAFA/UniProt KB set (the CAFA sequences that have a UniprotKB file) comprises 41,003 sequences, 3,767 of which do not contain any GO ontology and Pfam terms. °Here the CAFA/UniProtKB subset that can be validated in BAR+ is considered (BAR+validated). The number of sequences and the number of Pfam and GO terms are listed. Sequences that receive new validated terms are also listed according to Pfam, MFO, BPO and CCO. # Sequences of the CAFA set, out of a total of 7,295 that are not present in UniProtKB and are annotated in BAR+. +Number of sequences that have and also receive in BAR+ a PDB template.

and ABC transporter (PF00005) in Eukaryotes and Prokaryotes, respectively.

Comparison with direct UniProtKB annotation

34,065 sequences of CAFA/UniProtKB set found a match in 14,747 BAR+ clusters where their annotation is validated (about 71% of the CAFA set) and for 3,659 sequences the number of validated and annotated terms also increases (Table 2: BAR+ validated). The remaining CAFA/UniProtKB sequences (6,938 sequences of which 54% are not annotated) find a counterpart in BAR+ clusters without a statistically validated annotation and are not considered in Table 2. Furthermore, some 15% of the CAFA set (7,295 sequences) does not have a counterpart in UniproKB and they can be aligned towards BAR+ to receive annotation. Out of these, 3,451 sequences receive a statistically validated annotation (Table 2).

5,215 clusters are also endowed with a cluster HMM, suitable for sequence alignment of the target with the corresponding template/s of 11,935 sequences that by this can inherit also a structure (Table 2). Interestingly 50% of these sequences have a sequence identity to the corresponding template lower than 30%.

BAR+ web site

For the present analysis, BAR+ was updated by distinguishing two sets of clusters: those that are endowed with a statistically validated annotation (labeled with a yellow star), and those that are not statistically validated. A sequence can inherit annotation from a cluster in a statistically validated manner when upon alignment it falls into a statistically validated cluster; however at the web site for a sequence falling into BAR+ clusters we also provide all the cluster-associated and not validated terms. This is so also when the target aligns towards BAR+ singletons. Each cluster endowed with PDB templates is also endowed with a cluster HMM based alignment that for each sequence falling in the cluster allows building of the corresponding three dimensional protein structure. BAR+ is freely available at http://bar.biocomp.unibo.it/bar2.0/.

Conclusion

Functional annotation of protein sequences is one of the most important issues in annotation processes. When annotation is done electronically, mainly based on sequence similarity search, a robust validation process can help in the inheritance of Pfam and GO terms by transfer of annotation. Using our cluster-centric BAR+ annotation system and adopting as a test case the recently released CAFA set of sequences, we can annotate 84.9% of the CAFA set, 77.7% of which in a validated manner.

As compared with UniProtKB that annotates with GO and Pfam terms 77.1% of the CAFA set (Table 2), we validate 10,628 terms for 62.9% of the sequences, we increase the annotation for 7.6% of the set with some additional and validated 2,930 terms and annotate without validation the remaining 6.6% of the set.

Considering also that 7.2% of the CAFA set is newly annotated with validation, the gain in annotation within BAR+ is 14.8% with respect to UniProtKB, suggesting again that cluster specificity for a sequence is a necessary filter to inherit functional and structural features from well known proteins.

Furthermore we can endow with structural models some 25% of the whole CAFA set. At least 50% of the proteins that in BAR+ inherit a structural model share a sequence similarity with the template/s less than 30%, indicating that with our procedure also distantly related homologs can be safely annotated.

Authors' contributions

DP carried out the computational analysis. DP, AZ and IR developed BAR+ under the supervision of PF, PM and RC. GP developed the web site. DP, PF, PM and RC analyzed the data and wrote the manuscript. All the authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements
RC thanks the following grants: PRIN 2009 project 009WXT45Y (Italian Ministry for University and Research: MIUR), COST BMBS Action TD1101 (European Union RTD Framework Programme), and PON project PON01_02249 (Italian Ministry for University and Research: MIUR). DP is a recipient of a PHD fellowship from the Ministry of the Italian University and Research. GP is a recipient of a research contract from CIRI.
This article has been published as part of *BMC Bioinformatics* Volume 14 Supplement 3, 2013: Proceedings of Automated Function Prediction SIG 2011 featuring the CAFA Challenge: Critical Assessment of Function Annotations. The full contents of the supplement are available online at URL. http://www.biomedcentral.com/bmcbioinformatics/supplements/14/S3

Author details
[1]Bologna Biocomputing Group, University of Bologna, Italy. [2]Department of Biology, University of Bologna, Italy. [3]Department of Computer Science, University of Bologna, Italy. [4]Health Science and Technologies-ICIR, University of Bologna, Italy. [5]BioDec srl, Bologna, Italy.

Published: 28 February 2013

References

1. Lesk AM: *Introduction to Bioinformatics.* 3 edition. Oxford: Oxford University Press; 2008.
2. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A: **Protein function annotation by homology-based inference.** *Genome Biology* 2009, **10**:207.
3. Petryszak R, Kretschmann E, Wieser D, Apweiler R: **The predictive power of the CluSTr database.** *Bioinformatics* 2005, **21**:3604-3609.
4. Kaplan N, Sasson O, Inbar U, Friedlich M, Fromer M, Fleischer H, Portugaly E, Linial N, Linial M: **ProtoNet 4.0: a hierarchical classification of one million protein sequences.** *Nucleic Acids Research* 2005, **33**:D216-D218.
5. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5**:823-826.
6. Rost B: **Twilight zone of protein sequence alignments.** *Protein Eng* 1999, **12**:85-94.
7. Sánchez R, Pieper U, Melo F, Eswar N, Martí-Renom MA, Madhusudhan MS, Mirković N, Sali A: **Protein structure modeling for structural genomics.** *Nat Struct Biol* 2000, **7**:986-990.
8. Osadchy M, Kolodny R: **Maps of protein structure space reveal a fundamental relationship between protein structure and function.** *Proc Natl Acad Sci USA* 2011, **108**:12301-6.
9. Rost B: **Enzyme function less conserved than anticipated.** *J Mol Biol* 2002, **318**:595-608.
10. Tian W, Skolnick J: **How well is enzyme function conserved as a function of pairwise sequence identity?** *J Mol Biol* 2003, **333**:863-882.
11. Dietmann S, Fernandez-Fuentes N, Holm L: **Automated detection of remote homology.** *Curr Opin Struct Biol* 2002, 12:362-367.
12. Fariselli P, Rossi I, Capriotti E, Casadio R: **The WWWH of remote homolog detection: the state of the art.** *Brief Bioinform* 2007, **8**:78-87.
13. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunesekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**:D211-222.
14. de Lima Morais DA, Fang H, Rackham OJ, Wilson D, Pethica R, Chothia C, Gough J: **SUPERFAMILY 1.75 including a domain-centric gene ontology method.** *Nucleic Acids Res* 2011, **39**:D427-34.
15. The Gene Ontology Consortium: **Gene ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25-29.
16. The UniProt Consortium: **Ongoing and future developments at the Universal Protein Resource.** *Nucleic Acids Res* 2011, **39**:D214-D219.
17. Clark WT, Radivojac P: **Analysis of protein function and its prediction from amino acid sequence.** *Proteins* 2011, **79**:2086-96.
18. Rentzsch R, Orengo CA: **Protein function prediction–the power of multiplicity.** *Trends Biotechnol* 2009, **27**:210-9.
19. Bartoli L, Montanucci L, Fronza R, Martelli PL, Fariselli P, Carota L, Donvito G, Maggi G, Casadio R: **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* 2009, **8**:4362-4371.
20. Piovesan D, Martelli PL, Fariselli P, Zauli A, Rossi I, Casadio R: **BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences.** *Nucleic Acids Res* 2011, **39**:W197-W202.

doi:10.1186/1471-2105-14-S3-S4
Cite this article as: Piovesan *et al.*: How to inherit statistically validated annotation within BAR+ protein clusters. *BMC Bioinformatics* 2013 **14**(Suppl 3):S4.

A large-scale evaluation of computational protein function prediction

Predrag Radivojac[1], Wyatt T Clark[1], Tal Ronnen Oron[2], Alexandra M Schnoes[3], Tobias Wittkop[2], Artem Sokolov[4,5], Kiley Graim[4], Christopher Funk[6], Karin Verspoor[6,7], Asa Ben-Hur[4], Gaurav Pandey[8,9], Jeffrey M Yunes[10], Ameet S Talwalkar[11], Susanna Repo[8,12], Michael L Souza[13], Damiano Piovesan[14], Rita Casadio[14], Zheng Wang[15], Jianlin Cheng[15], Hai Fang[16], Julian Gough[16], Patrik Koskinen[17], Petri Törönen[17], Jussi Nokso-Koivisto[17], Liisa Holm[17], Domenico Cozzetto[18], Daniel W A Buchan[18], Kevin Bryson[18], David T Jones[18], Bhakti Limaye[19], Harshal Inamdar[19], Avik Datta[19], Sunitha K Manjari[19], Rajendra Joshi[19], Meghana Chitale[20], Daisuke Kihara[20,21], Andreas M Lisewski[22], Serkan Erdin[22], Eric Venner[22], Olivier Lichtarge[22], Robert Rentzsch[23], Haixuan Yang[24], Alfonso E Romero[24], Prajwal Bhat[24], Alberto Paccanaro[24], Tobias Hamp[25], Rebecca Kaßner[25], Stefan Seemayer[25], Esmeralda Vicedo[25], Christian Schaefer[25], Dominik Achten[25], Florian Auer[25], Ariane Boehm[25], Tatjana Braun[25], Maximilian Hecht[25], Mark Heron[25], Peter Hönigschmid[25], Thomas A Hopf[25], Stefanie Kaufmann[25], Michael Kiening[25], Denis Krompass[25], Cedric Landerer[25], Yannick Mahlich[25], Manfred Roos[25], Jari Björne[26], Tapio Salakoski[26], Andrew Wong[27], Hagit Shatkay[27,28], Fanny Gatzmann[29], Ingolf Sommer[29], Mark N Wass[30,31], Michael J E Sternberg[30], Nives Škunca[32], Fran Supek[32], Matko Bošnjak[32], Pance Panov[33], Sašo Džeroski[33], Tomislav Šmuc[32], Yiannis A I Kourmpetis[34,35], Aalt D J van Dijk[34,36], Cajo J F ter Braak[34], Yuanpeng Zhou[37], Qingtian Gong[37], Xinran Dong[37], Weidong Tian[37], Marco Falda[38], Paolo Fontana[39], Enrico Lavezzo[38], Barbara Di Camillo[40], Stefano Toppo[38], Liang Lan[41], Nemanja Djuric[41], Yuhong Guo[41], Slobodan Vucetic[41], Amos Bairoch[42,43], Michal Linial[44], Patricia C Babbitt[3], Steven E Brenner[8], Christine Orengo[23], Burkhard Rost[25], Sean D Mooney[2] & Iddo Friedberg[45,46]

Automated annotation of protein function is challenging. As the number of sequenced genomes rapidly grows, the overwhelming majority of protein products can only be annotated computationally. If computational predictions are to be relied upon, it is crucial that the accuracy of these methods be high. Here we report the results from the first large-scale community-based critical assessment of protein function annotation (CAFA) experiment. Fifty-four methods representing the state of the art for protein function prediction were evaluated on a target set of 866 proteins from 11 organisms. Two findings stand out: (i) today's best protein function prediction algorithms substantially outperform widely used first-generation methods, with large gains on all types of targets; and (ii) although the top methods perform well enough to guide experiments, there is considerable need for improvement of currently available tools.

The accurate annotation of protein function is key to understanding life at the molecular level and has great biomedical and pharmaceutical implications. However, with its inherent difficulty and expense, experimental characterization of function cannot scale up to accommodate the vast amount of sequence data already available[1]. The computational annotation of protein function has therefore emerged as a problem at the forefront of computational and molecular biology.

Many solutions have been proposed in the last four decades[2–10], yet the task of computational functional inference in a laboratory often relies on traditional approaches such as identifying domains or finding Basic Local Alignment Search Tool (BLAST)[11] hits among proteins with experimentally determined function. Recently, the availability of genomic-level sequence information for thousands of species, coupled with massive high-throughput experimental data, has created new opportunities for function prediction. A large number of methods have been proposed to exploit these data, including function prediction from amino acid sequence[12–16], inferred evolutionary relationships and genomic context[17–21], protein-protein interaction networks[22–25], protein structure data[26–28], microarrays[29] or a combination of data types[30–34]. An unbiased evaluation of these different methods can provide insight into their ability to characterize proteins functionally and can guide biological experiments. So far, however, a comprehensive assessment incorporating a large and diverse set of target sequences has not been conducted because of practical difficulties in providing an accurately annotated target set.

A full list of author affiliations appears at the end of the paper.

RECEIVED 2 APRIL 2012; ACCEPTED 10 DECEMBER 2012; PUBLISHED ONLINE 27 JANUARY 2013; DOI:10.1038/NMETH.2340

In this report, we present the results of the first CAFA experiment, a worldwide effort aimed at analyzing and evaluating protein function prediction methods. Although protein function can be described in multiple ways, we focus on classification schemes provided by the Gene Ontology (GO) Consortium[35]. Over the course of 15 months, 30 teams associated with 23 research groups participated in the effort, testing 54 function annotation algorithms. Short descriptions of published methods and detailed descriptions of unpublished methods can be found in the **Supplementary Note**. These methods were evaluated on a target set of 866 protein sequences from 11 species.

RESULTS

Protein function is a concept that can have different interpretations in different biological contexts. Generally, it describes biochemical, cellular and phenotypic aspects of the molecular events that involve the protein, including how the protein interacts with the environment (such as with small compounds or pathogens). From the various classification schemes developed to standardize descriptions of protein function, we chose the "Molecular Function" and "Biological Process" categories from GO. Each category in GO is a hierarchical set of terms and relationships among them that capture functional information; such a system facilitates computation, and its outputs can be interpreted by humans. GO's consistency across species and its widespread adoption make it suitable for large-scale computational studies. In CAFA, given a new protein sequence, the task of a protein function prediction method is to provide a set of terms in GO along with the confidence scores associated with each term.

The experiment was organized as follows. A set of 48,298 proteins lacking experimentally validated functional annotation was provided to the community 4 months before the submission deadline for predictions (**Fig. 1**). Proteins were annotated by the predicting groups, and these annotations were submitted to the assessors. After the submission deadline, GO experimental annotations for those sequences were allowed to accumulate over a period of 11 months. Methods were then evaluated on 866 targets from 11 species that had accumulated functional annotations during the waiting period (**Supplementary Table 1**). The Swiss-Prot database[36] was selected as the gold standard because of its relatively high reliability[37].

The selection of proteins was ineluctably biased owing to experimenter and annotator choice during the evaluation time frame. Thus, the set of targets was first analyzed to establish that it was representative of those sequences experimentally annotated before the submission deadline. In terms of organismal representation, the eukaryotic targets provided reasonable coverage of taxa (**Fig. 1**). In contrast, the set of prokaryotic targets was heavily biased toward *Escherichia coli* K-12. The distribution of terms over the target sequences was representative of the annotations in Swiss-Prot (data not shown); however, we note that in the Molecular Function category a large fraction of target sequences (38%) were associated with "protein binding" as their most specific term. The distribution of term depths over all targets is shown in **Supplementary Figure 1** for both ontologies.

Overall predictor performance

The quality of protein function prediction can be measured in different ways that reflect differing motivations for understanding function. In some cases, imprecise experimental characterization means that it is not entirely clear whether a prediction is correct. For CAFA, we principally report a simple metric, the maximum *F*-measure ($F_{\max}$; Online Methods), which considers predictions across the full spectrum from high to low sensitivity. This approach, however, has limitations, such as penalization of specific predictions (see Discussion). We note that the choice of evaluation metric differentially affects different prediction methods, depending on their application objectives.

Top predictor performance, based on maximum *F*-measure and calculated over all targets, is shown in **Figure 2** (precision-recall curves are shown in **Supplementary Fig. 2**; the performance evaluation for the Molecular Function ontology when proteins annotated with only the "protein binding" term were included is shown in **Supplementary Fig. 3**). All methods were compared with two baseline tools: (i) BLAST, in which all GO terms of an experimentally annotated sequence (template) from Swiss-Prot were transferred to the target sequence such that

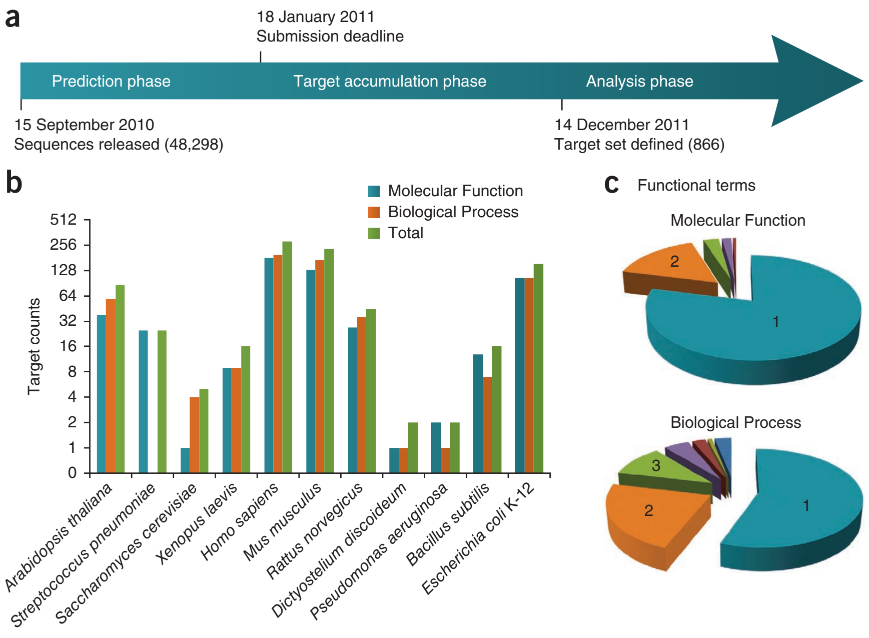


Figure 1 | Experiment timeline and target analysis. (**a**) Timeline for the CAFA experiment. (**b**) Number of target sequences per organism. The graph shows the number of target sequences for each of the ontologies (Molecular Function and Biological Process) as well as the total number of targets, obtained as a union between sequences in the two ontologies. Of 866 proteins, 531 had Molecular Function annotations and 587 had Biological Process annotations. (**c**) Distribution of target sequences in each ontology according to the number of leaf terms available for each protein sequence. For example, in the Molecular Function category, 79% of proteins had one leaf term, 16% had two leaf terms, and so on. A term is considered a leaf term for a particular target if no other GO term associated with that sequence is its descendant.

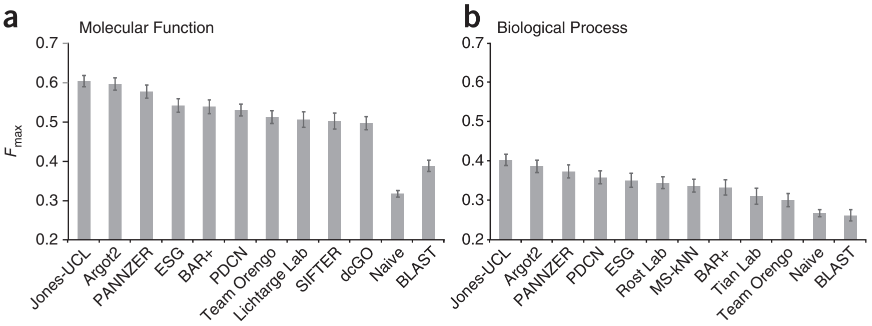


Figure 2 | Overall performance evaluation. (**a**,**b**) The maximum *F*-measure for the top-performing methods for Molecular Function ontology (**a**) and Biological Process ontology (**b**). All panels show the top ten participating methods in each category as well as the BLAST and Naive baseline methods. Note that 33 models outperformed BLAST in the Molecular Function category, whereas 26 models outperformed BLAST in the Biological Process category (cutoff scores below which methods were excluded from the panels were 0.468 and 0.300 for the Molecular Function and Biological Process categories, respectively). In the Molecular Function category, proteins with "protein binding" as their only leaf term were excluded from the analysis because the protein binding term was not considered informative (results that include those proteins are presented in **Supplementary Fig. 3**). A perfect predictor would be characterized with $F_{max} = 1$. Confidence intervals (95%) were determined using bootstrapping with $n = 10,000$ iterations on the set of target sequences. For cases in which a principal investigator participated in multiple teams, only the results of the best-scoring method are presented.

the scores equaled pairwise sequence identity between the template and the target (terms with multiple hits retained the highest score), and (ii) a naive method (Naive), in which each GO term for each target was scored with the relative frequency of this term in Swiss-Prot over all annotated proteins (Online Methods). We also evaluated the quality of position-specific iterated (PSI)-BLAST predictions, but we found that it did not provide any advantage over BLAST: specifically, F_{max}(PSI-BLAST) = F_{max}(BLAST) = 0.38 for Molecular Function; F_{max}(PSI-BLAST) = 0.24 and F_{max}(BLAST) = 0.26 for Biological Process. We believe that the improved ability of PSI-BLAST to identify remote homologs has been canceled out by its reranking of close hits.

We observed a substantial performance difference in the ability to predict the two GO categories (Molecular Function versus Biological Process). This can be partly explained by the topological differences between the ontologies (respectively: number of terms, 8,728 and 18,982; branching factor, 5.9 and 6.4; maximum depth, 11 and 10; number of leaf terms, 7,003 and 8,125). However, more fundamentally, terms in the Biological Process ontology were associated with a more abstract level of function. Such terms were less likely to be predictable solely from amino acid sequence, which was the data source used by most methods in this experiment and may critically depend on the cellular and organismal context.

Predictor performance on categories of targets

We divided the target sequences into a variety of different categories to compare predictor performance across each category. The first division was between easy and difficult targets. A target was considered easy if it had a 60% or higher sequence identity with any experimentally annotated protein. We manually chose the threshold of 60% after plotting the distribution of sequence identities between targets and annotated proteins (**Supplementary Fig. 4**). This resulted in 188 easy and 343 difficult targets in the Molecular Function category and 247 easy and 340 difficult targets in the Biological Process category. **Supplementary Figure 5** shows the precision-recall curves for both categories. Perhaps unsurprisingly, whereas BLAST outperformed Naive in the easy target category, their performance was similar for the difficult targets. However, because of the similar performance among top-ranked predictors over easy and difficult targets, the sequence identity–based classification of targets does not seem to accurately reflect the uncertainty associated with a protein's true function (except for with BLAST). This may be because the methods can compensate for the differences in sequence similarity of the best hit by using multiple sequence hits as well as other data sources.

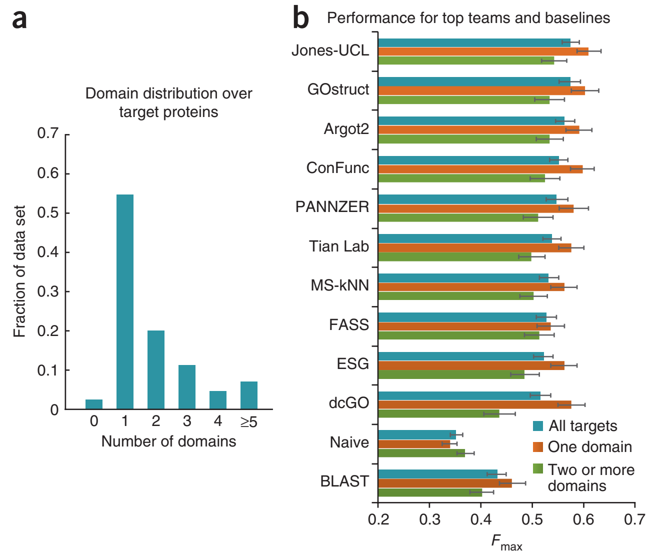
Next we compared prediction performance on eukaryotic versus prokaryotic targets (**Supplementary Fig. 6**). Performance was generally similar in the Molecular Function category, but in the Biological Process category we observed high prediction accuracy for prokaryotic targets. We believe this is because most prokaryotic targets came from *E. coli*, for which reliable experimental data are available, whereas the data for eukaryotic targets came from sources with highly variable coverage and quality. It is important to note that the particular calculation of precision and recall (Online Methods) adversely affected methods that predicted on only eukaryotic targets (BMRF, ConFunc, GOstruct and Tian Lab) and resulted in lower overall performance for these methods. Detailed results for eukaryotic and prokaryotic targets, as well as several individual organisms, are shown in **Supplementary Figures 6** and **7**.

Finally we separated targets into sequences containing a single domain versus sequences containing multiple protein domains, with domains defined according to Pfam-A classification[38] (targets without any Pfam-A hits were grouped together with single-domain proteins). Multidomain proteins were generally longer; however, they were not associated with more functional terms than single-domain proteins. By analyzing the performance of the top ten methods in each category, we found that although the overall accuracy was higher on single-domain proteins, results were significant in only the Molecular Function category and for eukaryotic targets ($P = 1.4 \times 10^{-5}$, $n = 10$, paired *t*-test; **Fig. 3**). Though generally expected, the higher performance on single-domain proteins further emphasizes the need for developing methods that can optimally combine sequence information from multiple domains along with other information to produce a relatively small set of predicted terms.

Predictor performance on functional terms

We assessed the ability of methods to predict individual GO terms by calculating the area under the receiver operating characteristic (ROC) curve (AUC; Online Methods). To more confidently assess the performance in predicting individual terms, we considered only terms for which at least 15 targets were annotated.

Figure 3 | Domain analysis and performance evaluation for single-domain versus multidomain eukaryotic targets. (**a**) Distribution of target proteins with respect to the number of Pfam domains they contain. (**b**) Performance evaluation in the Molecular Function category. Each of the ten top-performing methods showed higher accuracy (higher F_{max}) on single-domain proteins. Confidence intervals (95%) were determined using bootstrapping with n = 10,000 iterations on the set of target sequences.



Average AUC values were then calculated from the five top-performing models in each ontology, excluding those models that provide only single-score predictions.

Using the above criteria, we were able to calculate average AUC values for 28 Molecular Function and 223 Biological Process terms (**Supplementary Table 2**). We found a clear distinction between the average AUC of Molecular Function terms generally associated with catalytic and transporter activity and those associated with binding. In general, the prediction of terms associated with binding showed lower AUC values, even though proteins were biased toward being annotated with binding terms. Among the Biological Process terms, we found, as expected, low AUC values associated with less specific terms such as "locomotion", "cellular process" and "response to stress." We also found that prediction of terms associated with "cell adhesion", "metabolic process", "transcription" and "regulation of gene expression" showed high performance. We tested whether a high predictor AUC value on individual terms was due to high levels of sequence similarity among sequences experimentally annotated with those terms, and we found a moderate level of correlation (data not shown).

Case study

Here we illustrate some challenges associated with computational protein function prediction. We provide a detailed analysis of the human mitochondrial polynucleotide phosphorylase 1 (hPNPase, encoded by *PNPT1*), a large (783-amino-acid) protein with seven Pfam domains (**Fig. 4a**). Human PNPase is characterized by several experimentally determined functions, which makes it an attractive target with which to evaluate the performance of prediction methods. hPNPase belongs to a family of exoribonucleases, which hydrolyze single-stranded RNA in the 3′-to-5′ direction. In complex with other components of the mitochondrial degradasome, hPNPase mediates the translocation of small RNAs into the mitochondrial matrix[39]. It is also proposed to be involved in several biological processes including cell-cycle arrest[40], cellular senescence and response to oxidative stress[41].

Owing to its involvement in several molecular functions and biological processes, the comprehensive and accurate listing of functions of hPNPase is a challenging task. Furthermore, though PNPase is prevalent in bacteria and eukarya, it has accumulated several lineage-specific functions. Specifically, whereas bacterial and chloroplast PNPase have demonstrated exoRNase and polyadenylation activities, hPNPase functions predominantly as an RNA importer[39], showing exoRNase activity only *in vitro*[42]. Finally, hPNPase is a mitochondrial protein found in the intermembrane matrix. Taken together with its involvement in the rRNA import process, this suggests the need to predict the cellular compartment as part of a comprehensive understanding of function.

Figure 4b shows the experimental GO-term annotation of hPNPase as well as the terms predicted by a representative set of the ten top-performing methods. Within the Molecular Function terms, none of the methods predicted poly(U) or poly(G) RNA binding[43] or microRNA binding. However, most methods that did predict function correctly predicted 3′-to-5′ exoRNase activity and polyribonucleotide nucleotidyltransferase activity. It should

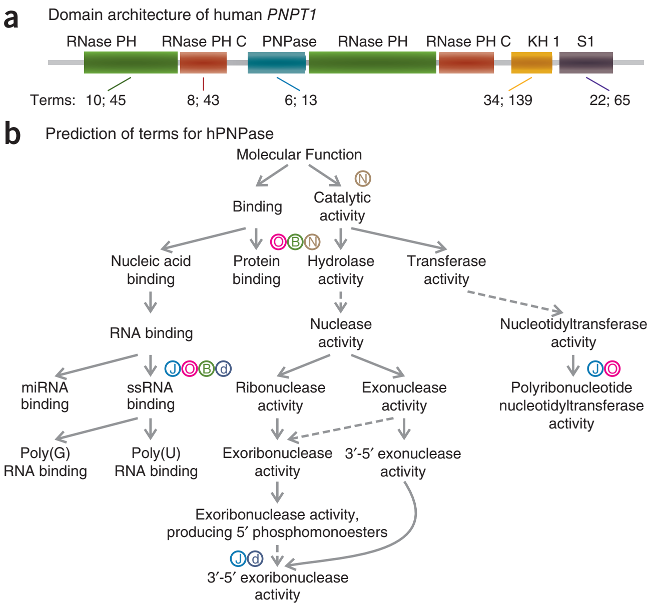


Figure 4 | Case study on the human *PNPT1* gene. (**a**) Domain architecture of human *PNPT1* gene according to the Pfam classification. For each domain, the numbers of different leaf terms (for the Molecular Function and Biological Process categories) associated with any protein in Swiss-Prot database containing this domain are shown. (**b**) Molecular Function terms (six of which are leaves) associated with the human *PNPT1* gene in Swiss-Prot as of December 2011. Colored circles represent the predicted terms for three representative methods as well as two baseline methods. The prediction threshold for each method was selected to correspond to the point in the precision-recall space that provides the maximum *F*-measure. J (blue), Jones-UCL; O (magenta), Team Orengo; d (navy blue), dcGO; B (green), BLAST; N (brown), Naive. Dashed lines indicate the presence of other terms between the source and destination nodes.

be noted that poly(U) and poly(G) binding and microRNA binding are uncommon throughout the PNPase lineage. This may be the reason why none of the programs predicted these terms.

In the Biological Process category, the most prominent function of hPNPase in the literature is the import of nuclear 5S rRNA into the mitochondrion[39]; indeed, it is hypothesized that this is the reason for hPNPase's location in the intermembrane matrix. However, this function, along with other important terms, such as cellular senescence, was not predicted by any of the top-performing methods at the optimal threshold levels. Generally, the Biological Process predictions were highly nonspecific for most models. In sum, the multidomain architecture of hPNPase, its pleiotropy and the different functions it assumes in different taxa all contribute to the challenge of correctly predicting hPNPase function.

DISCUSSION

Protein function is difficult to predict for several reasons. First, function is studied from various aspects and at multiple levels: for example, it describes the biochemical events involving the protein and also how each protein affects pathways, cells, tissues and the entire organism. Second, protein function and its experimental characterization are context dependent: a particular experiment is unlikely to determine a protein's entire functional repertoire under all conditions (such as temperature, pH or the presence of interacting partners). Third, proteins are often multifunctional[44] and promiscuous[45]; in fact, of the experimentally annotated proteins in Swiss-Prot, 30% have more than one leaf term in the Molecular Function ontology, as do 60% in the Biological Process ontology[16]. Fourth, in addition to being incomplete, available functional annotations are error prone because of experiment interpretation or curation issues[37,46]. Finally, current efforts largely map protein function to gene names, thus confounding the functions of potentially diverse isoforms. Despite these challenges, the CAFA experiment revealed progress in automated function annotation over the past decade.

Top algorithms are useful and outperform BLAST considerably. The first generation of function prediction methods performed a simple function transfer via pairwise sequence similarity: that is, the most similar annotated hit was used as the basis of function prediction[47]. Several studies have been aimed at characterizing performance of these methods[3,16,48]. The CAFA experiment provides evidence that the best algorithms universally outperform simple functional transfer. The experiment also showed that BLAST is largely ineffective at predicting functional terms related to the Biological Process ontology. This is possibly due to homologs assuming different biological roles in different tissues and organisms[49].

Principles underlying best methods. The methods evaluated in CAFA used a variety of biological and computational concepts. Most methods used sequence alignments with an underlying hypothesis that sequence similarity is correlated with functional similarity. Recent studies have shown that this correlation is weak when applied to pairs of proteins[16] and that domain assignments alone are not sufficient to resolve function[50]. Therefore, the main challenge for the alignment-based methods was to devise ways of combining multiple hits or identified domains into a single prediction score. More than half the methods used data beyond sequence similarity, such as types of evolutionary relationships, protein structure, protein-protein interactions or gene expression data. The challenge for these methods was finding ways to integrate disparate data sources and properly handle incomplete and noisy data. For example, the protein-protein interaction network for yeast is nearly complete (although noisy), whereas the sets of available interactions for *Arabidopsis thaliana* and *Xenopus laevis* are rather sparse (but less noisy, given a smaller fraction of high-throughput data). Finally, some methods used literature mining, which could also be related to the task of retrieving the correct function rather than predicting it from the set of textual descriptions about a protein. As information retrieval is still a challenging research problem, it was useful to evaluate performance accuracy of the methods that exploited literature searching.

On the computational side, most methods used machine learning principles: that is, they typically found combinations of sequence-based or other features that correlated with a specific function in a training set of experimentally annotated proteins. Although these methods automate the task of learning and inference, they also require experience in selecting classification models (for example, a support vector machine), learning parameters, features or the training data that would result in good performance. In addition, the sets of rules according to which these methods score new proteins may be difficult to interpret. Despite the added layer of complexity, machine learning generally played a positive role in increasing prediction accuracy. Thus, it may be expected that top-performing methods in the future will be based on well-founded principles of statistical learning and inference.

With few exceptions, the same methods that performed well for the Molecular Function category also performed well in the Biological Process category; however, their overall performance in the latter category was inferior. We believe that this is because homologs may perform their biochemical roles in different pathways, and prediction methods are less able to discern those differences at this time. Because sequence similarity is less predictive of the biological roles of proteins, a key to improving the prediction of a protein's biological function will be our ability to generate better-quality systems data and to develop computational tools that exploit them.

Evaluation metrics. The choice of evaluation metrics was another interesting aspect of the experiment. We decided to use simple and easily interpretable metrics (Online Methods), although simple measures based on precision and recall have limitations in this domain. First, such metrics are sensitive to problems related to the nonuniform distribution of proteins over GO terms due to the equal weight given to all terms. Second, proteins are weighted equally regardless of the depth of their experimental annotation: that is, a correct prediction on a protein annotated with a shallow term (and its ancestors) is considered as good as a correct prediction on a protein annotated with a deep term. Third, a method that reports only high-confidence deep annotations for a small number of proteins will be penalized (in terms of recall) compared to a method that annotates all proteins with frequently occurring general terms. Finally, in some cases, it is not clear whether to consider a prediction correct or erroneous; with our current approach, we consider only the experimental annotation and more general predictions to be correct. As such, correct and highly specific predictions will be penalized if the protein has been experimentally annotated only in a more generic way. For those reasons, we encourage the development of a diverse set of metrics to understand better the strengths and weaknesses of function prediction in different application contexts.

Summary. The CAFA experiment was designed to enable the community to periodically reassess the performance of computational methods as experimental evidence accumulates. In addition, the large set of targets released to the community provided us with prediction scores for most proteins across multiple methods. If the experiment is repeated, we expect to be able to evaluate future methods against those that deposited predictions in the first CAFA experiment and therefore monitor progress in the field over time.

Though the CAFA experiment has seen positive outcomes, it is also clear that there is significant room for the improvement of protein function prediction. In the Molecular Function category, performance may be considered accurate. However, in the Biological Process category, the overall performance of the top-scoring methods was below our expectations. This was true for any subset of targets. Another area in need of improvement is the availability of tools that can easily be used by experimental scientists and that can be maintained and upgraded on a regular basis. As the community moves beyond the initial algorithm development stage, there is a need to provide stand-alone tools (similar to the BLAST package) capable of predicting protein function at several different levels.

Given its significance, its intellectual challenge and the growing need for accurate functional annotations, protein function prediction is likely to remain an active and expanding research field. As the quality of data improves and the number of experimentally annotated proteins grows, we expect that computational prediction will become more accurate. On the basis of the CAFA experiment, it seems that the most powerful methods will be those that will devise principled ways to integrate a variety of experimental evidence and weigh different data appropriately and separately for each functional term. Novel ideas and approaches are necessary as well.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We gratefully acknowledge I. Landsberg-Halperin for coining the term "CAFA," T. Theriault for the initial graphical design of **Figure 1**, G. Schuster for illuminating discussions on hPNPase and A. Facchinetti, R. Velasco, E. Cilia, D.A. Lee, P. Vats, R. Banerjee and A. Bayaskar for their participation in various individual projects. The Automated Function Prediction Special Interest Group meeting at the ISMB 2011 conference was supported by the US National Institutes of Health (NIH) grant R13 HG006079-01A1 (P.R.) and Office of Science (Biological and Environmental Research), US Department of Energy (DOE BER), grant DE-SC0006807TDD (I.F.). Individual projects were partially supported by the following awards: US National Science Foundation (NSF) DBI-0644017 (P.R.), ABI-0965768 (A.B.-H.), DMS0800568 (D. Kihara), CCF-0905536 and DBI-1062455 (O.L.), DBI-0965768 (K.V.) and ABI-1146960 (I.F.); Marie Curie International Outgoing Fellowship PIOF-QA-2009-237751 (S.R.); PRIN 2009 project 009WXT45Y Italian Ministry for University and Research MIUR (R.C.); NIH GM093123 (J.C.), GM075004 and GM097528 (D. Kihara), GM079656 and GM066099 (O.L.), LM00945102 (C.F.), R01 GM071749 (S.E.B.) and LM009722 and HG004028 (S.D.M.); FP7 "Infrastructures" project TransPLANT Award 283496 (A.D.J.v.D.); UK Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/G022771/1 (J.G.), BB/K004131/1 (A.P.) and BB/F020481/1 (M.N.W. and M.J.E.S.); BBSRC (D.T.J.); Marie Curie Intra European Fellowship Award PIEF-GA-2009-237292 (D.T.J.); Department of Information Technology, Government of India (R.J.); EU, BBSRC and NIH Awards (C.O.); Natural Sciences and Engineering Research Council of Canada Discovery Award #298292-2009, Discovery Accelerator Award #380478-2009, Canada Foundation for Innovation New Opportunities Award 10437 and Ontario's Early Researcher Award #ER07-04-085 (H.S.); Netherlands Genomics Initiative (Y.A.I.K. and C.J.F.t.B.); National Information and Communication Technology Australia (K.V.); National Natural Science Foundation of China grants 31071113 and 30971643 (W.T.); DOE BER KP110201 (S.E.B.); and Alexander von Humboldt Foundation (B.R.). P.R. acknowledges the Indiana University high-performance computing resources (NSF grant CNS-0723054). I.F. acknowledges the assistance of the high-performance computing group at Miami University.

AUTHOR CONTRIBUTIONS

P.R. and I.F. conceived of the CAFA experiment, supervised the project and wrote most of the manuscript. S.D.M. participated in the design of and supervised the method assessment. W.T.C. performed the analysis of feasibility of the experiment and most of the target and performance analysis and contributed to writing. P.R. and W.T.C. designed and produced figures. T.R.O. developed the web interface, including the portal for submission and the storage of predictions. T.R.O. and T.W. verified the assessment code and participated in analysis. A.M.S. designed and performed the analysis of targets. A. Bairoch, M.L., P.C.B., S.E.B., C.O. and B.R. steered the CAFA experiment, provided critical guidance and participated in writing. The remaining authors participated in the experiment, provided writing and data for their methods and contributed comments on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2340.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Liolios, K. *et al.* The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **38**, D346–D354 (2010).
2. Bork, P. *et al.* Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707–725 (1998).
3. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O. & Ofran, Y. Automatic prediction of protein function. *Cell Mol. Life Sci.* **60**, 2637–2650 (2003).
4. Watson, J.D., Laskowski, R.A. & Thornton, J.M. Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* **15**, 275–284 (2005).
5. Friedberg, I. Automated protein function prediction—the genomic challenge. *Brief. Bioinform.* **7**, 225–242 (2006).
6. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007).
7. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* **8**, 995–1005 (2007).
8. Punta, M. & Ofran, Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput. Biol.* **4**, e1000160 (2008).
9. Rentzsch, R. & Orengo, C.A. Protein function prediction—the power of multiplicity. *Trends Biotechnol.* **27**, 210–219 (2009).
10. Xin, F. & Radivojac, P. Computational methods for identification of functional residues in protein structures. *Curr. Protein Pept. Sci.* **12**, 456–469 (2011).
11. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
12. Jensen, L.J. *et al.* Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.* **319**, 1257–1265 (2002).
13. Wass, M.N. & Sternberg, M.J. ConFunc—functional annotation in the twilight zone. *Bioinformatics* **24**, 798–806 (2008).
14. Martin, D.M., Berriman, M. & Barton, G.J. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics* **5**, 178 (2004).

15. Hawkins, T., Luban, S. & Kihara, D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci.* **15**, 1550–1556 (2006).
16. Clark, W.T. & Radivojac, P. Analysis of protein function and its prediction from amino acid sequence. *Proteins* **79**, 2086–2096 (2011).
17. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
18. Marcotte, E.M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
19. Enault, F., Suhre, K. & Claverie, J.M. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. *BMC Bioinformatics* **6**, 247 (2005).
20. Engelhardt, B.E., Jordan, M.I., Muratore, K.E. & Brenner, S.E. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput. Biol.* **1**, e45 (2005).
21. Gaudet, P., Livstone, M.S., Lewis, S.E. & Thomas, P.D. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinform.* **12**, 449–462 (2011).
22. Deng, M., Zhang, K., Mehta, S., Chen, T. & Sun, F. Prediction of protein function using protein-protein interaction data. *J. Comput. Biol.* **10**, 947–960 (2003).
23. Letovsky, S. & Kasif, S. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **19** (suppl. 1), i197–i204 (2003).
24. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* **21**, 697–700 (2003).
25. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. & Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21** (suppl. 1), i302–i310 (2005).
26. Pazos, F. & Sternberg, M.J. Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl. Acad. Sci. USA* **101**, 14754–14759 (2004).
27. Pal, D. & Eisenberg, D. Inference of protein function from protein structure. *Structure* **13**, 121–130 (2005).
28. Laskowski, R.A., Watson, J.D. & Thornton, J.M. Protein function prediction using local 3D templates. *J. Mol. Biol.* **351**, 614–626 (2005).
29. Huttenhower, C., Hibbs, M., Myers, C. & Troyanskaya, O.G. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics* **22**, 2890–2897 (2006).
30. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. USA* **100**, 8348–8353 (2003).
31. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
32. Costello, J.C. *et al.* Gene networks in *Drosophila melanogaster*: integrating experimental data to predict gene function. *Genome Biol.* **10**, R97 (2009).
33. Kourmpetis, Y.A., van Dijk, A.D., Bink, M.C., van Ham, R.C. & ter Braak, C.J. Bayesian Markov Random Field analysis for protein function prediction based on network data. *PLoS ONE* **5**, e9293 (2010).
34. Sokolov, A. & Ben-Hur, A. Hierarchical classification of gene ontology terms using the GOstruct method. *J. Bioinform. Comput. Biol.* **8**, 357–376 (2010).
35. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
36. Bairoch, A. *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005).
37. Schnoes, A.M., Brown, S.D., Dodevski, I. & Babbitt, P.C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, e1000605 (2009).
38. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
39. Wang, G. *et al.* PNPASE regulates RNA import into mitochondria. *Cell* **142**, 456–467 (2010).
40. Sarkar, D. *et al.* Down-regulation of Myc as a potential target for growth arrest induced by human polynucleotide phosphorylase (hPNPaseold-35) in human melanoma cells. *J. Biol. Chem.* **278**, 24542–24551 (2003).
41. Wu, J. & Li, Z. Human polynucleotide phosphorylase reduces oxidative RNA damage and protects HeLa cell against oxidative stress. *Biochem. Biophys. Res. Commun.* **372**, 288–292 (2008).
42. Wang, D.D., Shu, Z., Lieser, S.A., Chen, P.L. & Lee, W.H. Human mitochondrial SUV3 and polynucleotide phosphorylase form a 330-kDa heteropentamer to cooperatively degrade double-stranded RNA with a 3′-to-5′ directionality. *J. Biol. Chem.* **284**, 20812–20821 (2009).
43. Portnoy, V., Palnizky, G., Yehudai-Resheff, S., Glaser, F. & Schuster, G. Analysis of the human polynucleotide phosphorylase (PNPase) reveals differences in RNA binding and response to phosphate compared to its bacterial and chloroplast counterparts. *RNA* **14**, 297–309 (2008).
44. Jeffery, C.J. Moonlighting proteins. *Trends Biochem. Sci.* **24**, 8–11 (1999).
45. Khersonsky, O. & Tawfik, D.S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).
46. Brenner, S.E. Errors in genome annotation. *Trends Genet.* **15**, 132–133 (1999).
47. Doolittle, R.F. *Of URFS and ORFS: A Primer on How to Analyze Derived Amino Acid Sequences* (University Science Books, 1986).
48. Addou, S., Rentzsch, R., Lee, D. & Orengo, C.A. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. *J. Mol. Biol.* **387**, 416–430 (2009).
49. Nehrt, N.L., Clark, W.T., Radivojac, P. & Hahn, M.W. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.* **7**, e1002073 (2011).
50. Brown, S.D., Gerlt, J.A., Seffernick, J.L. & Babbitt, P.C. A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol.* **7**, R8 (2006).

[1]School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA. [2]Buck Institute for Research on Aging, Novato, California, USA. [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California, USA. [4]Department of Computer Science, Colorado State University, Fort Collins, Colorado, USA. [5]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, California, USA. [6]Computational Bioscience Program, University of Colorado School of Medicine, Aurora, Colorado, USA. [7]National ICT Australia, Victoria Research Laboratory, Melbourne, Australia. [8]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California, USA. [9]Mount Sinai School of Medicine, New York, New York, USA. [10]Joint Graduate Group in Bioengineering, University of California, Berkeley, Berkeley, California, USA. [11]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, California, USA. [12]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [13]Biophysics Graduate Program, University of California, Berkeley, Berkeley, California, USA. [14]Department of Biology, University of Bologna, Bologna, Italy. [15]Department of Computer Science, University of Missouri, Columbia, Missouri, USA. [16]Department of Computer Science, University of Bristol, Bristol, UK. [17]Department of Biological and Environmental Sciences & Institute of Biotechnology, Viikki Biocentre, University of Helsinki, Helsinki, Finland. [18]Department of Computer Science, University College London, London, UK. [19]Bioinformatics Group, Centre for Development of Advanced Computing, Pune University Campus, Pune, India. [20]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA. [21]Department of Biological Sciences, Purdue University, West Lafayette, Indiana, USA. [22]Department of Molecular and Human Genetics, Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, Texas, USA. [23]University College London, Institute for Structural and Molecular Biology, London, UK. [24]Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway, University of London, Egham, UK. [25]Technische Universität München, Bioinformatik-I12, Informatik, Garching, Germany. [26]Department of Information Technology, University of Turku, Turku Centre for Computer Science, Turku, Finland. [27]School of Computing, Queen's University, Kingston, Ontario, Canada. [28]Department of Computer and Information Sciences, University of Delaware, Newark, Delaware, USA. [29]Max Planck Institute for Informatics, Saarbrücken, Germany. [30]Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College, London, UK. [31]Structural Computational Biology Group, Spanish National Cancer Research Centre, Madrid, Spain. [32]Division of Electronics, Rudjer Boskovic Institute, Zagreb, Croatia. [33]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. [34]Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands. [35]Bioinformatics Systems, Nestlé Institute of Health Sciences, Lausanne, Switzerland. [36]Applied Bioinformatics, Plant Research International, Wageningen, The Netherlands. [37]Institute of Biostatistics, School of Life Sciences, Fudan University, Shanghai, China. [38]Department of Molecular Medicine, University of Padova, Padova, Italy. [39]Istituto Agrario San Michele all'Adige Research and Innovation Centre, Trento, Italy. [40]Department of Information Engineering, University of Padova, Padova, Italy. [41]Department of Computer and Information Sciences, Temple University, Philadelphia, Pennsylvania, USA. [42]Swiss Institute of Bioinformatics, Geneva, Switzerland. [43]Department of Human Protein Sciences, University of Geneva, Geneva, Switzerland. [44]Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. [45]Department of Microbiology, Miami University, Oxford, Ohio, USA. [46]Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio, USA. Correspondence should be addressed to P.R. (predrag@indiana.edu) or I.F. (i.friedberg@miamioh.edu).

ONLINE METHODS

Experiment design. The CAFA experiment was conceived in the fall of 2009. The Organizing, Steering and Assessment Committees were designated by March 2010. During the same period a feasibility study was conducted to determine the rate at which experimental annotations accumulated in Swiss-Prot between 2007 and 2010. We concluded that a period of 6 months or more would result in annotations of at least 300–500 proteins, which would be sufficient for statistically reliable comparisons between algorithms. The experiment was announced in July 2010 and subsequently heavily advertised. The set of targets was announced on 15 September 2010 with a prediction submission deadline of 18 January 2011 (**Fig. 1**).

Predictors were asked to submit predictions for each target along with scores ranging between 0 and 1 that would indicate the strength of the prediction (ideally, posterior probabilities). To reduce the amount of data submitted, we allowed no more than 1,000 term annotations for each target. Prediction algorithms were also associated with keywords from a predetermined set, which were used to provide insight into the types of approaches that performed well. A list of all participating teams, principal investigators and methods is provided in **Supplementary Table 3**.

Initial comparative evaluation of models was conducted in July 2011 during the Automated Function Prediction (AFP) Special Interest Group (SIG) meeting associated with the ISMB 2011 conference. This study provides the analysis on a set of targets from the Swiss-Prot database from 14 December 2011.

Target proteins. A set of 48,298 target amino acid sequences was announced in September 2010. Because our feasibility study showed that only a handful of species were steadily accumulating experimental annotations, target proteins were selected from predominantly those species. The targets contained all the sequences in Swiss-Prot from 7 eukaryotic and 11 prokaryotic species that were not associated with any experimental GO terms. A protein was considered experimentally annotated if it was associated with GO terms having EXP, IDA, IMP, IGI, IEP, TAS or IC evidence codes. An additional set of targets was announced consisting of 1,301 enzymes from multiple species and metagenomic studies that were the focus of the Enzyme Function Initiative project[51].

18 January 2011 was set as the deadline for the submission of function predictions. To exclude targets that had accumulated annotations before the submission deadline, we obtained annotated proteins from the January version of Swiss-Prot, GO[35] and UniProt-GOA[52] databases. We refer to those sets of proteins as Swiss-Prot(t_0), GO(t_0) and GOA(t_0), respectively.

We later determined the evaluation set of target proteins by downloading a newer version of the Swiss-Prot database, denoted as Swiss-Prot(t). The set of target proteins for the CAFA experiment was then selected using the following scheme

$$\text{Targets}(t) = \text{Swiss-Prot}(t) - \text{Swiss-Prot}(t_0) - \text{GO}(t_0) - \text{GOA}(t_0)$$

Note that this experiment was designed to allow for reassessment of algorithm performance at some later point in time.

Evaluation metrics. Algorithms were evaluated in two scenarios: (i) protein centric and (ii) term centric. These two types of evaluations were chosen to address the following related questions: (i) what is the function of a particular protein? and (ii) what are the proteins associated with a particular functional term?

1. Protein-centric metrics. The main evaluation metric in CAFA was the precision-recall curve. For a given target protein i and some decision threshold $t \in [0, 1]$, the precision and recall were calculated as

$$\text{pr}_i(t) = \frac{\sum_f I\left(f \in P_i(t) \wedge f \in T_i\right)}{\sum_f I\left(f \in P_i(t)\right)}$$

and

$$\text{rc}_i(t) = \frac{\sum_f I\left(f \in P_i(t) \wedge f \in T_i\right)}{\sum_f I\left(f \in T_i\right)}$$

where f is a functional term in the ontology, T_i is a set of experimentally determined (true) nodes for protein i, and $P_i(t)$ is a set of predicted terms for protein i with score greater than or equal to t. Note that f ranges over the entire ontology (separately for Molecular Function and Biological Process), excluding the root. Function $I(\cdot)$ is the standard indicator function. For a fixed threshold t, a point in the precision-recall space is then created by averaging precision and recall across targets. Precision at threshold t is calculated as

$$\text{pr}(t) = \frac{1}{m(t)} \cdot \sum_{i=1}^{m(t)} \text{pr}_i(t)$$

where $m(t)$ is the number of proteins on which at least one prediction was made above threshold t. On the other hand, recall is calculated over all n proteins in a target set, i.e.,

$$\text{rc}(t) = \frac{1}{n} \cdot \sum_{i=1}^{n} \text{rc}_i(t)$$

regardless of the prediction threshold. The maximum ratio between $m(t)$ and n (over all thresholds t) is referred to as the prediction coverage. If a particular algorithm outputs only a fixed score (for example, 1), its performance will be described by a single point in the precision-recall space instead of by a curve.

For submissions with unpropagated functional annotations, the organizers recursively propagated all scores toward the root of the ontology such that each parent term received the highest score among its children. The annotations were propagated regardless of the type of relationship between terms. We note that it may be useful to associate different weights with different ontological terms and therefore reward algorithms that are better at predicting more difficult or less frequent terms. However, for simplicity, in our main evaluation, each term was associated with an equal weight of 1 (weighted precision-recall curves are shown in **Supplementary Fig. 8**).

The main appeal of the precision-recall evaluation stems from its interpretability: if, for a particular threshold, a method has a precision of 0.7 at a recall of 0.5, this indicates that on average 70% of the predicted terms will be correct and that about 50% of the true annotations will be revealed for a previously unseen protein.

doi:10.1038/nmeth.2340

On the other hand, a limitation of this evaluation method is that the terms are not independent because of ontological relationships, and the unequal level of specificity of functional terms at the same depth in the ontology was not taken into account.

To provide a single number for comparisons between methods, we calculated the *F*-measure (a harmonic mean between precision and recall) for each threshold and calculated its maximum value over all thresholds. More specifically, we used

$$F_{\max} = \max_t \left\{ \frac{2 \cdot \mathrm{pr}(t) \cdot \mathrm{rc}(t)}{\mathrm{pr}(t) + \mathrm{rc}(t)} \right\}$$

2. Term-centric metrics. For each functional term *f*, we calculated the area under the ROC curve (AUC) using a sliding threshold approach. The ROC curve is a plot of sensitivity (or recall) for a given false positive rate (or 1 − specificity). The sensitivity and specificity for a particular functional term *f* and threshold *t* were calculated as

$$\mathrm{sn}_f(t) = \frac{\sum_i I\left(f \in P_i(t) \wedge f \in T_i\right)}{\sum_i I\left(f \in T_i\right)}$$

and

$$\mathrm{sp}_f(t) = \frac{\sum_i I\left(f \notin P_i(t) \wedge f \notin T_i\right)}{\sum_i I\left(f \notin T_i\right)}$$

where $P_i(t)$ is the set of predicted terms for protein *i* with a score greater than or equal to threshold *t*, and T_i is the set of true terms for protein *i*. Once the sensitivity and specificity for a particular functional term were determined over all proteins for different values of the prediction threshold, the AUC was calculated using the trapezoid rule. The AUC has a useful probabilistic interpretation: given a randomly selected protein associated with functional term *f* and a randomly selected protein not associated with *f*, the AUC is the probability that the former protein will receive a higher score than the latter protein[53].

Baseline methods. In addition to the methods implemented by the community, we used two additional methods as baselines. The first such method is based on BLAST[11] hits to the database of proteins with experimentally annotated functions (roughly 37,000 proteins). The score for a particular term was calculated as the maximum sequence identity between the target protein and any protein experimentally annotated with that term. More specifically, if a particular protein was hit with the local sequence identity 75%, all its functional terms were transferred to the target sequence with the score of 0.75. If a term was hit with multiple sequence identity scores, the highest one was retained. BLAST was selected as a baseline method because of its ubiquitous use. We note that the same method was tested using the BLAST bit scores, which resulted in slightly better performance. In addition to BLAST, we also tested PSI-BLAST[11], in which the profiles were created using the most recent "nr" database and –j 3 –h 0.0001 parameters. These profiles were then searched against a database of experimentally annotated proteins with *E*-values used to rank the hits. The second baseline method, referred to as Naive, used the prior probability of each term in the database of experimentally annotated proteins as the prediction score for that term. If a term "protein binding" occurs with relative frequency 0.25, each target protein was associated with score 0.25 for that term. Thus, the Naive method assigned the same predictions to all targets.

51. Gerlt, J.A. *et al.* The Enzyme Function Initiative. *Biochemistry* **50**, 9950–9962 (2011).
52. Barrell, D. *et al.* The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.* **37**, D396–D403 (2009).
53. Hanley, J.A. & McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).

doi:10.1038/nmeth.2340