Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

A procedure to recruit members to enlarge protein family databases - the building of UECOG (UniRef-Enriched COG Database) as a model

G.R. Fernandes[1]*, D.V.C. Barbosa[1]*, F. Prosdocimi[1], I.A. Pena[1], L. Santana-Santos[1], O. Coelho Junior[1], A. Barbosa-Silva[1], H.M. Velloso[1], M.A. Mudado[2], D.A. Natale[3], A.C. Faria-Campos[4], S.V. A. Campos[4] and J.M. Ortega[1]

[1]Departamento de Bioquímica e Imunologia, Laboratório de Biodados, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[2]Fundação Ezequiel Dias, Belo Horizonte, MG, Brasil
[3]Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA
[4]Departamento de Ciência da Computação, ICEX, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
*These authors contributed equally to this study.

Corresponding author: J.M. Ortega
E-mail: miguel@icb.ufmg.br

Genet. Mol. Res. 7 (3): 910-924 (2008)
Received June 2, 2008
Accepted August 11, 2008
Published September 30, 2008

ABSTRACT. A procedure to recruit members to enlarge protein family databases is described here. The procedure makes use of UniRef50 clusters produced by UniProt. Current family entries are used to recruit additional members based on the UniRef50 clusters to which they belong. Only those additional UniRef50 members that are not fragments and whose length is within a restricted range relative to the original entry are recruited. The enriched dataset is then limited to contain only genomes from selected clades. We used the COG database - used for genome annotation and for studies of phylogenetics and gene evolution - as a model. To validate the method, a

UniRef-Enriched COG0151 (UECOG) was tested with distinct procedures to compare recruited members with the recruiters: PSI-BLAST, secondary structure overlap (SOV), Seed Linkage, COGnitor, shared domain content, and neighbor-joining single-linkage, and observed that the former four agree in their validations. Presently, the UniRef50-based recruitment procedure enriches the COG database for Archaea, Bacteria and its subgroups Actinobacteria, Firmicutes, Proteobacteria, and other bacteria by 2.2-, 8.0-, 7.0-, 8.8-, 8.7-, and 4.2-fold, respectively, in terms of sequences, and also considerably increased the number of species.

Key words: COG; Secondary database; UniRef; UniProt; UECOG

INTRODUCTION

Phylogenetics, genome annotation and studies of gene evolution all benefit from the comparative analysis of protein families. Several databases are dedicated to the clustering of related proteins. Indeed the COG database (Tatusov et al., 1997, 2003) - a collection of Clusters of Orthologous Groups (COGs) of proteins - has not only supported diverse analysis of the protein families, but has stimulated the development of databases derived from the use of distinct procedures, such as OrthoMCL-DB (Chen et al., 2006) and Inparanoid (O'Brien et al., 2005). In some sense, attribution of Gene Ontology terms to amino acid sequences by the GOA project (Camon et al., 2004) also forms collections of genes with the same activity in distinct organisms, therefore being putative orthologs. Other recent databases or procedures to cluster sequences also address the same issue, such as FlowerPower (Krishnamurthy et al., 2007), Ortholuge (Fulton et al., 2006), OrthologID (Chiu et al., 2006), and Seed Linkage (Barbosa-Silva et al., 2008). In addition to these, the Universal Protein Resource Consortium (UniProt) produces the UniProt Knowledgebase (UniProt Consortium, 2007) and the UniProt Reference Clusters (UniRef) containing over five million sequences. The latter set contains equivalents of protein families that are generated triweekly using CD-HIT (Li and Godzik, 2006), resulting in three distinct types of clusters: i) UniRef100, where the best representative of 100% identical entries is selected to stand for the sequences in that cluster; ii) UniRef90 and iii) UniRef50, where members of each cluster show either 90 or 50% identity, respectively, to the seed sequence (Suzek et al., 2007).

An attractive possibility would be to use UniRef50 clusters to enrich a database built with complete genomes such as COG. The rationale for doing so is that the UniRef50 clusters are generated triweekly, and thus provide a basis for rapid enrichment of less-frequently updated databases. Here, we describe the procedure to build UniRef-Enriched COGs (UECOGs) and present a case study using multiple validation procedures for recruitment. These latter procedures include: a) PSI-BLAST (position-specific iterated - basic local alignment search tool), where all recruited UniRef50 members that are hit under a PSI-BLAST (Altschul et al., 1997) search started by the recruiter (a COG member from the same UniRef50 cluster) are labeled as valid recruited entries; b) secondary structure overlap (SOV), where recruited members whose indices of secondary structure overlap as determined by SSPro4 (Geourjon et al., 2001) and SOV (Rost et al., 1994) are labeled as valid if over a given threshold; c) neighbor-joining tree neighboring, where all recruited entries that are in single linkage (continuously consecutive) to the recruiter are labeled as valid; d) domain structure, where recruited members that share the same content of domains as determined by RPS-BLAST using SMART (Schultz et al., 1998), Pfam (Finn et al., 2006) and COG domain databases,

and a threshold of 75% coverage and 1e-3 E-value cutoff are applied; e) COGnitor at the NCBI website, using the old COG version as database, determining if the recruited entries belong to the expected COG; f) Seed Linkage, a software developed by our group to enlarge clusters using their members as seed for recruitment of cognate proteins (Barbosa-Silva et al., 2008).

Presently, UECOG enriches the COG database for Archaea, Bacteria and its subgroups Actinobacteria, Firmicutes, Proteobacteria, and other bacteria by 2.2-, 8.0-, 7.0-, 8.8-, 8.7-, and 4.2-fold, respectively, in terms of sequence, and also increased the number of species. Users can download UECOG for the distinct clades from our server at http://biodados.icb.ufmg.br/uecog. UECOG is updated monthly using the latest available iProClass table and UniProtKB file.

MATERIAL AND METHODS

Enrichment of COG database with members from UniRef50 clusters

Two tables of tabulated data were used for this purpose: i) the COG file whog, downloaded from NCBI [ftp://ftp.ncbi.nlm.nih.gov/pub/COG/COG] and ii) iProClass (Huang et al., 2003) downloaded September 14, 2007 from the PIR FTP site [ftp://ftp.pir.georgetown.edu/databases/iproclass/]. The table obtained from COG was updated to contain the txid NCBI taxonomy information and the taxon group IDs or clade Archaea or Bacteria (which was further subdivided into Actinobacteria, Firmicutes, Proteobacteria, and other bacteria). From the COG table, the genbank identifier (*gi*) ID was extracted and the iProClass table was consulted to determine which COG members would act as recruiters. The UniRef50 cluster ID for recruiters was obtained. All members from the UniRef50 recruited clusters were then recruited.

Filtering of UniParc entries and fragments

Certain entries included in UniRef50 are not from the richly annotated set included in UniProtKB. Entries from the UniProt Archive (UniParc) - not useful for further analysis - were filtered out of the database based on the "UPI" prefix of the entry identifier. After the FASTA file for the other entries was obtained, a parser was used to inspect the annotation in FASTA file in order to filter out all those containing the string "(Fragment)", which denotes entries without a functional start or stop codon.

Filtering by taxonomic clade

One possibly undesirable effect of this enrichment procedure is the addition of eukaryotic sequences to the (largely prokaryotic) COG clusters. Thus, we decided to filter the recruitment using the taxon group (clade) ID. Clade subtrees can be obtained at the NCBI taxomomy site using, for example, the query "txid2 [Subtree]" and formatting the results as txid list. The final UECOG thus contains only Archaea and Bacteria subtrees. The eukaryotic organisms *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Encephalitozoon cuniculi* were not included in UECOG.

Filtering by size selection

The size of each potential recruit was compared to that of its recruiter. If the ratio

ranged from 0.9 to 1.1 - where the recruited protein was not more than 10% shorter or longer than the recruiter - the recruitment was allowed.

Validation by PSI-BLAST

The PSI-BLAST validation uses COG members as queries (recruiters) and cognate UniRef50 members (recruited) plus recruiters as a formatted database. The blastpgp program was performed with parameter -h 1 x 10^{-5} (an E-value cutoff that limits the inclusion of sequences in the alignment that is used to construct the position-specific scoring matrices, PSSM) and run to convergence. Recruited sequences hitting at least one recruiter were considered to be PSI-BLAST validated. The search was run either using the entire UECOG as database or concentrated on a specific clade (e.g., Proteobacteria).

Validation by domain conservation

Conserved domains were mapped to recruiters using RPS-BLAST with the CDD database (which contains SMART, Pfam and COG) (Marchler-Bauer et al., 2002). A stringent procedure was executed to map the domains, requiring an E-value lower than 1E-3 and over 75% coverage of the domain size in the CDD database. The occurrence of domains was examined in recruiters to determine the list of domains shared among all recruiters that are from the same UniRef50 cluster within a given COG. All recruited sequences from this same UniRef50 were considered domain validated if they showed the same domains shared by recruiters in the cognate UniRef50 cluster.

Validation by branch distance in neighbor-joining trees

Recruitment of UniRef50 members was evaluated by construction of neighbor-joining trees. Sequences from a given clade were aligned with Clustal W using BLOSUM62 and default parameters, submitted to SEQBOOT to generate five adjusted multiple alignments, and the distance between them was estimated with PHYLIP PROTDIST. Moreover, five trees were generated with PHYLIP NEIGHBOR, and the distance and number of branches between all proteins were calculated and stored. To determine the neighboring in a tree, a single-linkage procedure was started with the original COG members that acted as recruiters and the number of sequences clustered by an iterative one branch distance was determined. When the search reached a COG member, the count was not incremented but the search was continued. Sequences that were one branch apart in single-linkage iterations were considered neighbor-joining validated. Merging of UniRef50 clusters in neighbor-joining trees was the prominent cause of search interruption.

Validation by secondary structure overlap

The secondary structure of recruiters and recruited proteins was determined using a local implementation of the SSPro4 software. Percentage of structural overlap was determined with the SOV parameter described by Rost et al. (1994) for tuples of sequences from the same UniRef50 to determine the minimum SOV value amongst them. Recruiters were aligned with recruited candidates and the maximum SOV values were stored for each recruited sequence. In the case that a given UniRef50 member was the sole recruiter, SOV values of alignments of recruited sequences

to this recruiter were used. The results were processed by classes of SOV. Sequences that passed the 80% SOV against recruiters were considered SOV validated.

Validation by Seed Linkage

Recruiters from the Edited COG0151 were used as seed in the Seed Linkage program (Barbosa-Silva et al., 2008) using as a database the entire UECOG0151. No seed remained as a singlet and a single cluster was formed.

Availability

UECOG is available from our server at http://biodados.icb.ufmg.br/uecog. UECOG will be updated on a monthly basis. Future versions will incorporate web services.

RESULTS

The use of COGs for phylogenetic studies or as a source of information for the annotation of novel genomes would benefit from a reliable update. One example is given in Figure 1. Two strains of *Helicobacter pylori* are present in COGs, and often proteins from both of these strains are grouped as brothers in neighbor-joining trees (for example, see the branch labeled "A" in Figure 1, left panel), but sometimes an expansion is exclusive to a given genome (see branch "B", also in the left panel). The enrichment with additional sequences obtained from an updated dataset such as the UniRef50 database corroborates the observation that gene B is not shared equally between *H. pylori* genomes for strains 26695 and J99 (Figure 1, right panel). Thus, we set out to enrich COG with UniRef50 entries.

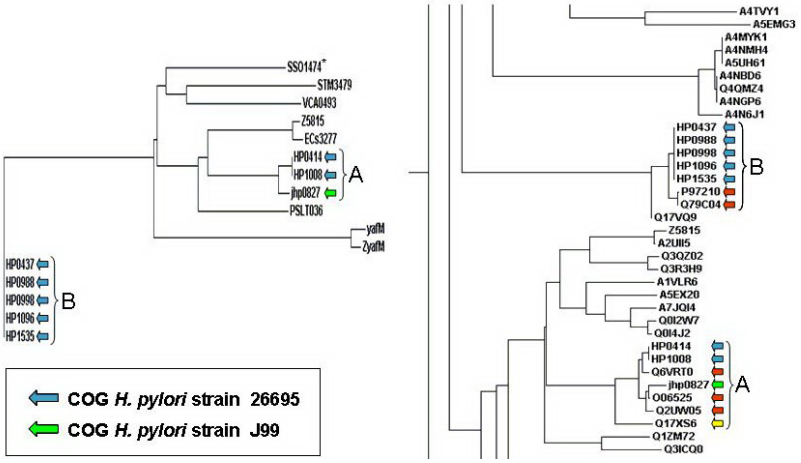

Figure 1. Neighbor-joining trees using COG or UECOG databases. On the left, a region of the phylogram showing two putatively distant transposases (COG1943), one (gene A) present in both *Helicobacter pylori* strains 26695 (blue arrows) and J99 (green arrows) and the other potentially exclusive to one of these strains (gene B). On the right, a region of the phylogram obtained with UECOG confirming this exclusivity after the addition of 154 Proteobacteria sequences that included sequences from other *H. pylori* strains (red arrows) and from *H. acinonychis* (yellow arrow). Sequence SSO1474* from *Sulfolobus solfataricus* (Archaea) was used to represent the root of the unrooted tree on the left.

Production of an updated version of COG

COG is the most used ortholog database for genome and gene annotation analyses. It was initially built in 1997 and further updated in 2001. UniProt produces a triweekly updated database containing well-annotated protein sequences from a plethora of organisms, and includes a clustered set called UniRef50. Our aim was the production of a UniRef-Enriched COG database to allow updated gene annotation based on ortholog groups and phylogenetic studies with a more complete source of information. In order to produce UECOG, we first produced an updated version of COG (Edited COG), since this database now contains some original sequences that are no longer valid. In this initial analysis, we identified 4506 NCBI *gi* in COG, which were discontinued. Of these invalid *gi*'s, 2091 could be mapped to new valid *gi*'s using Batch Entrez (http://www.ncbi.nlm.nih.gov/sites/batchentrez) that linked old *gi*'s to new ones based on accession number data. Moreover, to be maintained in Edited COG, all entries must have a corresponding accession in UniProtKB. To find the UniProtKB equivalent to those proteins that failed to return a UniProtKB accession by consulting iProClass with either its *gi* or the updated *gi*, we performed a BLAST search against the UniProtKB database using the amino acid sequence in COG as query. We analyzed the BLAST results to select proteins identical to the COG sequence; this UniProtKB entry was put into the Edited COG. After this procedure, we applied a last filter to delete the entries that were labeled as "Fragment" in the UniProtKB FASTA descriptions. The whole procedure is illustrated in Figure 2.

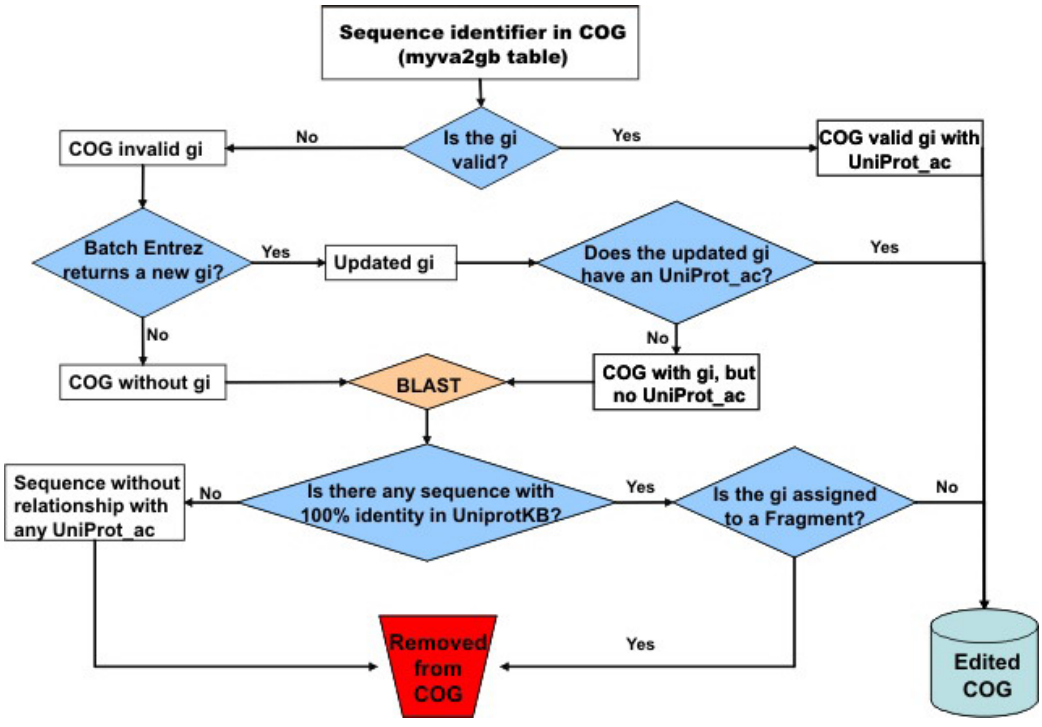

Figure 2. Methodology of COG edition. Discontinued *gi* in COG were updated with Batch Entrez or with BLAST to UniProt, and those entries without a reference in UniProt were removed from Edited COG.

Database recruitment procedure

All the Edited COG entries have a corresponding UniProtKB entry, and have thus been assigned to a UniRef50 cluster as well. The COG entries that are members of a UniRef50 cluster were called recruiters. In the next step, we selected all non-fragment UniProtKB entries that share a UniRef50 cluster with one recruiter protein. Each of these recruited proteins joined its recruiter COG cluster. All the steps are shown in Figure 3.

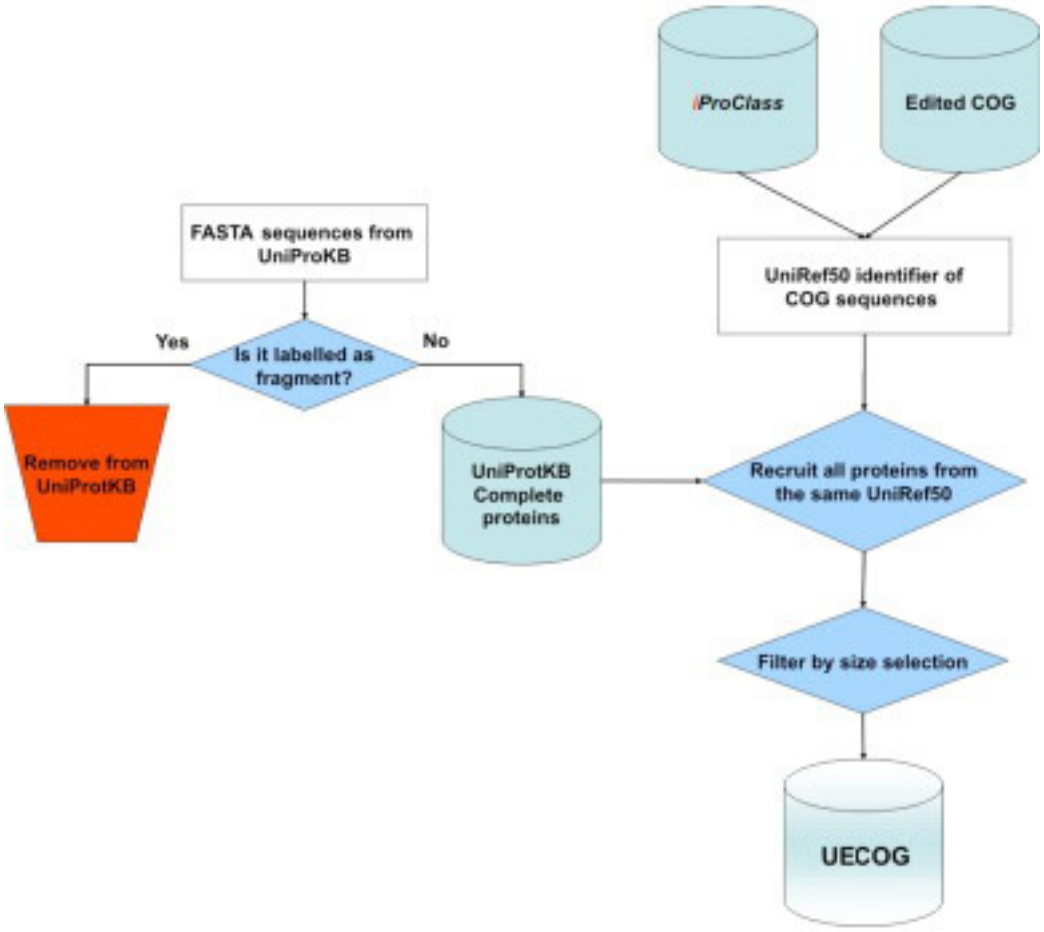

Figure 3. Methodology to build UECOG. Members of Edited COG acted as recruiters of members of their UniRef50 clusters, but only complete proteins were recruited. Size selection was further added at the recruitment step (see below).

Enrichment of COG database with members from UniRef50 clusters

UECOG enrichment was performed by recruiting non-fragment members of UniRef50 clusters that share similar length with one or more COG members (see Material and Methods).

Only prokaryotic organisms comprise the UECOG database. The three eukaryotic organisms were removed from Edited COG, because there is a specific database for the eukaryotic organisms (KOG), and the filtering by clade would not be easily determined. The final database, consisting of prokaryotic sequences, was enriched with 961,725 proteins (7.01-fold compared to the original COG database), as shown in Table 1. Sequences from Proteobacteria proved to be the greatest source of UECOG enrichment (8.74-fold), while Actinobacteria showed the greatest source of enrichment of species growing from 4 species in COG to 6736 in UECOG. Archaea yielded the smallest enrichment of sequence and genomes as compared to other groups, probably because it is very distantly related to other bacterium groups, and there is no great number of Archaea already sequenced.

Table 1. Enrichment in UECOG.

Clade	COG		Edited COG		UECOG		Fold
	Genomes	Proteins	Genomes	Proteins	Genomes	Proteins	
COG	66	144320	n/a	n/a	n/a	n/a	n/a
Prokaryotes	63	137122	63	124369	3477	961725	7.01
Archaea	13	22374	13	21310	248	49836	2.23
Bacteria	53	114748	50	103059	3229	911889	7.95
Actinobacteria	4	9391	4	6736	391	65871	7.01
Firmicutes	12	20921	12	19961	747	184403	8.81
Proteobacteria	24	67737	24	60741	1594	592000	8.74
Other bacteria	14	16699	10	15621	497	69615	4.17

n/a = not available.

Moreover, the overall enrichment was measured for each UECOG cluster. The majority of the COG clusters were enriched with new proteins. The data in Figure 4 present the amount of enrichment for each COG. Each dot represents a single COG. The original number of sequences for each COG was plotted on the X-axis while the enriched UECOG number is shown on the Y-axis. The dots above the dashed line represent the enriched clusters. The data show that some UECOG clusters are enriched up to about 30-fold.

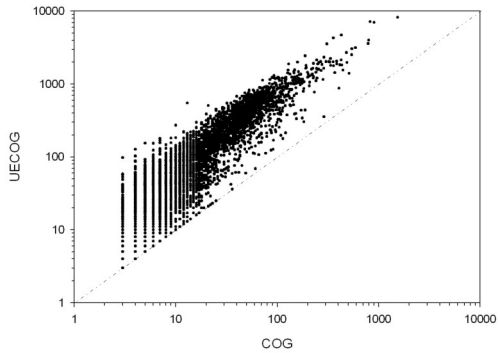

Figure 4. Enrichment in UECOG. The number of entries in UECOG is shown as a function of the entries in Edited COG. A dashed line indicates the cases where no enrichment was obtained.

Recruitment validation procedures

To validate the recruitment of UniRef50 proteins to each COG, we performed PSI-BLAST and RPS-BLAST searches, comparisons of SOV, and single linkage inspection of neighbor-joining trees, and used the Seed Linkage program (Barbosa-Silva et al., 2008) using recruiters as seed. The validation experiments described below were conducted on UECOGs constructed without size selection.

We chose for a case analysis the cluster COG0151 (phosphoribosylamine-glycine ligase) containing 53 proteins in COG; of these, the 51 prokaryote sequences (excluding *S. cerevisiae*, *S. pombe*, and *E. cuniculi*) were used as queries to illustrate the validation. The corresponding UECOG was selected because it was greatly enriched (with 501 proteins), where its protein sequences have broadly conserved domains and it possesses proteins from all clades of the 66 organisms of COG. Of the 501 candidates, 42 would be discarded by size filtering; however, they were retained in these experiments to illustrate the appropriateness of the filtering by size. With PSI-BLAST, the validation was executed in two flavors: validation of the whole UECOG as database or using six sub-sets (Archaea, Bacteria, Actinobacteria, Firmicutes, Proteobacteria, and other bacteria) as databases. PSI-BLAST is a BLAST module that, with iterative searches creates the PSSM related to query and the protein sequences that are found in each round. Therefore, PSI-BLAST searches can be more sensitive, finding more similar proteins and obtaining more orthologs than BLASTp. The searches were carried out using each recruiter protein (from COG) as query against its corresponding UECOG until the convergence of sequences was found. The recruited protein was considered "validated" when it was found as a hit by at least one recruiter.

In the search using the whole set, the 51 queries against the 552 proteins of UECOG0151 (recruited + recruiters) were used and 94% of recruited candidates were validated. Conversely, the six sub-sets were generated in accordance to the taxonomy clades and an analysis was performed to verify the differences in the performance of validation caused by the restriction of recruiters and recruited groups to a clade, thus resulting in a search focused on that clade. In this way, we could validate an equal or greater number of recruited candidates. The results are shown in Table 2 and indicate that validation by both means is rather high.

Table 2. PSI-BLAST, Seed Linkage, and secondary structure overlap (SOV) validation of UECOG0151.

Clade	Recruiter	Recruited	PSI-BLAST		Seed linkage	SOV
			UECOG database validated	Clade database validated	Validated	Validated
Prokaryotes	51	501	470 (94%)	470 (94%)	463 (92%)	470 (94%)
Archaea	12	30	30 (100%)	30 (100%)	29 (97%)	29 (97%)
Bacteria	41	471	438 (93%)	446 (95%)	434 (92%)	441 (94%)
Actinobacteria	4	41	38 (93%)	41 (100%)	40 (98%)	41 (100%)
Firmicutes	8	93	92 (99%)	92 (99%)	92 (99%)	92 (99%)
Proteobacteria	21	313	279 (89%)	286 (91%)	282 (90%)	288 (92%)
Other bacteria	6	24	20 (83%)	20 (83%)	20 (83%)	20 (83%)

Another validation procedure was conducted by the Seed Linkage software produced by our group (Barbosa-Silva et al., 2008). This software produces a clustering of cognate proteins from multiple organisms beginning with a single sequence through connectivity saturation with that seed sequence. Thus, recruiters were used as seed and a file containing both recruiters and recruited proteins was used as database. Seed Linkage returned a single cluster as expected, but excluded 38 sequences, distributed into diverse clades (Table 2).

One more validation procedure was applied by comparing secondary structure of recruited candidates to every recruiter from the same UniRef50 cluster, in a pair-wise fashion. First, secondary structures of all proteins were predicted by the SSPro4 software (Geourjon et al., 2001). Next, structural overlap between recruited candidates and recruiters was determined as in Rost et al. (1994), and the highest SOV index obtained was saved. Figure 5 shows the distribution of SOV in classes. Most of the recruited candidates show a SOV value over 80%, a value that splits the distribution into two groups. Thus, 80% SOV was used as cutoff for validation. The distribution of validation into distinct clades is shown in Table 2. Validation with SOV was also very high.

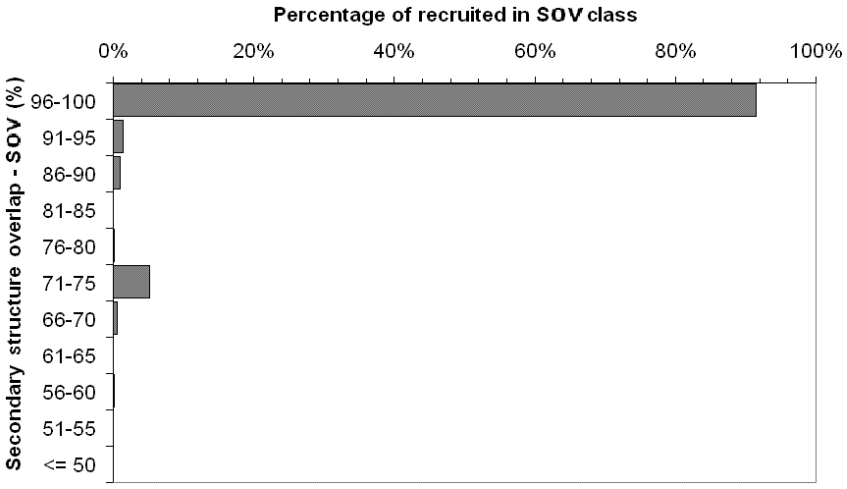

Figure 5. Distribution of secondary structure overlap (SOV) between recruited candidates and recruiters. SOV values were determined for each recruited candidate and the respective recruiters, and the highest value was considered.

An additional verification of the enrichment was conducted by posting the recruited candidates to a search in the COGnitor program available at the NCBI site (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html). This analysis confirmed that 29 of the 30 entries that were not validated by three approaches (PSI-BLAST, SOV and Seed Linkage) actually belong to a different cluster, COG0041. The only sequence that COGnitor mapped to COG0151 actually is a fragment derived from genome annotation. A Venn diagram analyzing the four procedures is shown in Figure 6A. Twenty-nine sequences were not validated by the four procedures and most of the sequences (463 of 501: 92.4%) were four times validated. Figure 6B shows the diagram using only sequences that passed the size selection; with a cost of generating some false negatives, the data indicate that false positives were discarded.

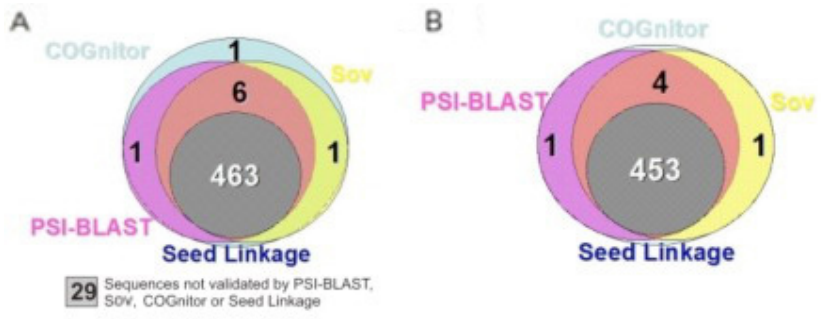

Figure 6. Diagrams combining four procedures of validation of UECOG0151. **A.** Five hundred and one recruited candidates were analyzed and 29 were not validated by any procedure. **B.** Same analysis but after 10% size selection filtering.

If a size filter of 10% is applied (see Material and Methods, and below), none of the 29 non-validated (false positive) entries were recruited. However, this filter also caused the exclusion of 10 of the 4-fold validated sequences (of 463), 6 of the PSI-BLAST and SOV validated sequences, plus the single candidates validated only by PSI-BLAST or only by SOV. To prevent these 18 candidates from being filtered out, the size selection limit would need to be increased. However, doing so might risk incorporating false positives. Figure 7 shows the fraction of all recruited candidates that are not incorporated as a function of the divergence of size selection allowed. Using a 10% cutoff, less than 10% of candidates are not recruited; the curve tends to saturate around 30%. However, using 30% would be acceptable only if coupling filtering with the validation procedures investigated here.

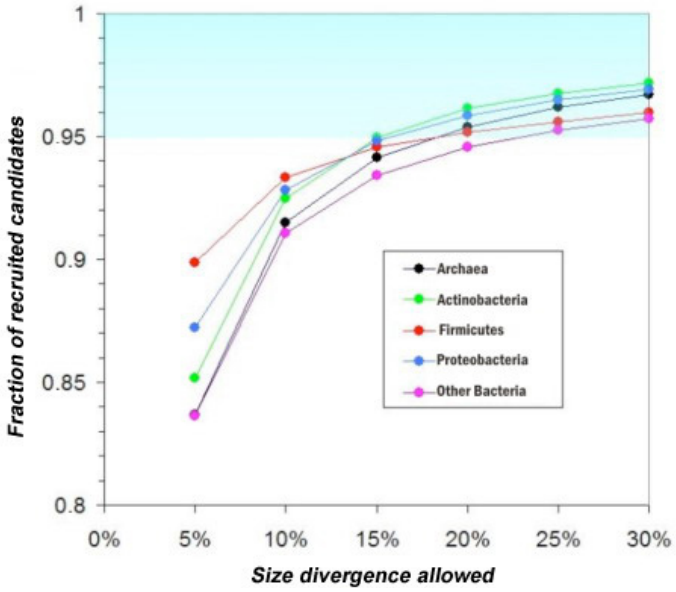

Figure 7. Fraction of recruitment as a function of size selection divergence. The experiment was conducted for the clades indicated.

In addition to the validation methods described above, we mapped domains using RPS-BLAST and the CDD database to determine if the domains shared by the recruiters for a given UniRef50 cluster are present also in the recruited candidates from that same cluster. The results are shown in Table 3. This procedure is stringent but nonetheless returned high values except for Archaea (73%) and Proteobacteria (84%), both of which fell below the statistically expected range; the results for other bacteria (83%) are in accordance with the results above.

Table 3. Validation of UECOG0151 with shared domains.

Clade	Recruiter	With domain	(%)	Recruited	Validated	(%)
Prokaryotes	51	38	75%	501	394	79%
Archaea	12	12	100%	30	22	73%
Bacteria	41	37	90%	471	393	83%
Actinobacteria	4	3	75%	41	39	95%
Firmicutes	8	7	87%	93	92	98%
Proteobacteria	21	20	95%	313	288	84%
Other bacteria	6	3	50%	24	20	83%

Finally, we examined whether recruited candidates are brothers in neighbor-joining trees. This experiment focused on Actinobacteria, Firmicutes and Proteobacteria since the inclusion of all clades could compromise the resolution in the tree. The entire set of sequences of each clade from UECOG0151 was used to construct neighbor-joining trees in five experiments and each recruiter was inspected to verify if the neighbor in the tree was from the same UniRef50 or was an original COG member. In the case that the neighbor was an original member, the algorithm continued the search without incrementing the score; the score was incremented only when the neighbor was a recruited member of the same UniRef50 cluster as the recruiter. The results shown in Table 4 together with manual inspection suggest that UniRef50 clusters are merged in neighbor-joining trees. Thus, the procedure does not seem to be appropriate for validation, but illustrates the important contribution of the enrichment for a better phylogenetic analysis of COGs.

Table 4. Analysis of UECOG0151 recruitment in neighbor-joining trees.

Cluster	Neighbor-joining validation						
	Recruiter	Recruited	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Actinobacteria							
UniRef50_P65894	4	36	11	11	11	10	11
Firmicutes							
UniRef50_Q9HUV8	4	50	52	35	49	45	52
UniRef50_Q9ZF44	1	2	0	0	0	0	0
UniRef50_O66949	1	36	0	0	0	0	0
UniRef50_Q8K8Y4	1	19	19	19	19	19	19
UniRef50_Q5HH10	1	15	15	15	15	15	15
Sub-total	8	122	86	69	83	79	86
Proteobacteria							
UniRef50_O25817	2	2	2	2	2	2	2
UniRef50_Q8KBV8	2	67	8	8	8	4	3
UniRef50_Q9ABD2	1	8	1	1	1	2	1
UniRef50_Q9HUV8	1	9	4	4	5	6	6
UniRef50_Q9PC09	14	207	88	86	86	96	89
UniRef50_Q9PN47	1	18	16	16	16	13	16
Sub-total	21	311	119	117	118	123	117
Total	33	469	216	197	212	212	214

Reassessing a genome with UECOG procedure

The enrichment procedure was expected to add, in an identity range of 50% around each recruiter, a group of proteins from the respective UniRef50 cluster. Therefore, several proteins from organisms not included in COG would not be recruited unless a recruiter exists in a reasonable identity range. To illustrate this problem, we artificially deleted some genomes from COG and asked how many sequences of it would be recovered from other recruiters within COG (Edited COG). The results are presented in Table 5. While for *Escherichia coli* (txid 83333) the recovery reached 82% of all proteins present in UECOG, for some organisms this yield was not satisfactory, probably due to the lack of an organism closely related to this one in COG. Thus, the BBH procedure conducted with complete genomes seems to be very broad in comparison to the coverage attained by UniRef50 clusters, although the enrichment of the COG database of sequences closely related to the ones present in the database is significant.

Table 5. Recovery of a genome deleted from COG by recruiters in Edited COG.

Genome	Proteins present in			Recovered	(%)
	COG	Edited COG	UECOG		
Escherichia coli K12	3762	3242	3580	2938	82%
Bacillus halodurans	3262	3089	3182	1216	38%
Corynebacterium glutamicum	2249	2146	2693	531	20%
Archaeoglobus fulgidus	2034	1917	1920	378	20%

DISCUSSION

Databases of related proteins are a useful source for bioinformatics research, including annotation of novel genomes and phylogenetic studies. However, some approaches to generate such databases are limited by the need for complete genomes. One example is the COG database. Here we report a procedure to update and enrich the COG database to build UniRef-Enriched COGs, which will be maintained and made available at our website (http://biodados.icb.ufmg.br/uecog). The procedure consists of recruiting non-COG members of UniRef50 clusters that share one or more members with COGs. Only candidates that are not fragments are allowed to be incorporated. For this filtering, a parser of the UniProt FASTA files was necessary, but the current release of UniRef now contains a file with this information. We then took the opportunity to remove from Edited COGs the sequences that UniProtKB access labeled as a fragment. Edited COGs were obtained by updating entries that had new *gi* identifiers and by deleting entries that could not be reliably updated. Thus, UECOGs represent Edited COGs plus enrichment. A second important filter was to limit the enrichment to the clades present in the COG database. This procedure ensures enrichment with closely related sequences and avoids recruiting sequences from organisms that are too far apart from the ones possessing complete genomes. Using this approach, we safely obtained more data for analysis.

We also developed a series of approaches to validate the recruitment and illustrated their usage with the analysis of a chosen cluster UECOG0151. The approaches validating the highest percentage of sequences were PSI-BLAST and SOV validation, followed by Seed Link-

age software. COGnitor, available as a service in NCBI for single sequences, confirmed that 29 of the 30 sequences not validated by the three mentioned approaches were from a different cluster, COG0041, and suggested the use of a size filter to prevent false positives. Inspection of UECOG database indicates that it is feasible to use the validation procedures to restrict the inclusion of recruited candidates; inclusion of such validation steps is being considered for a second release of UECOG. The version under development will allow users to download only recruited candidates verified by the validation approaches described here. However, size selection was efficient for elimination of all recruited candidates that were not capable of validation. Further tests for the size selection are warranted, but there is indication that the use of 30% (rather than the 10% used in this construction), coupled with the validation procedures, could yield additional recruited candidates without diminishing the accuracy of recruitment. However, the routine use of such validations would compromise the speed of the update, and therefore was not considered an integral part of a generally applicable updated methodology.

One limitation of the approach described - possibly unique to COGs - is that it will fail to find remote orthologs; that is, it is limited to finding new members whose sequence identity is greater than or equal to 50% of any current member. COGs are noted for being independent of any similarity-based cutoff, and thus, the prototypical COG recruitment procedure (COGnitor) is able to recruit quite distant proteins. To estimate the contribution of low-identity orthologs, we examined the status of proteins in the current COG set. Specifically, we determined the number of proteins whose BLAST best hit was <50% identical. Of the 144,320 prokaryotic proteins, 52,608 (36.5%) would be missed by the UniRef50 recruitment procedure, and 95% of COGs would lack at least one member. The impact of the large number of false negatives largely depends on the purpose for which COGs can be used. On the one hand, COGs can be used for propagating annotation from one protein to another. For this purpose, the only real need is to have the two proteins in the same family. The impact of failing to cluster protein Y with protein X (when they should be together) is that annotation would not be propagated to protein Y (assuming protein X has some known function), and thus, protein Y would keep its current annotation. If protein Y is annotated and all the members of the COG that contain protein X are not, then these proteins would not get the annotation of Y, and again they would keep the current annotation. Accordingly, the failure to recruit all possible members does not really create a scientific problem above and beyond what already exists. Another purpose for which COGs can be used is to examine the metabolic suite of a given organism. For this purpose, it is imperative that all proteins from the organism are properly classified, lest one mistakenly concludes that a given set of proteins is missing from the organism. For this reason, UECOGs should not be used as is for such studies. However, it is expected that the grouping of more genomes in novel COG versions, or in any equivalent database built with the use of BBH relationships and complete genomes, can add important recruiters to positions in the evolutionary tree that could then recruit the missing proteins.

We envision that the UniRef50 enrichment procedure itself is applicable to any protein classification database, and may indeed perform quite well in certain systems. The advantage of the method is that it is not computationally intensive. This offers flexibility: use as is to enrich a database, perhaps for interim releases or for a quick update before annotation of new sequences, or couple the UniRef50 enrichment procedure with any flavored clustering methodology to reduce computation time and resources. No matter which database one wishes to update, the UniRef-based enrichment offers speed and accuracy.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Barbosa-Silva A, Satagopam VP, Schneider R and Ortega JM (2008). Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence. *BMC Bioinformatics* 9: 141.

Camon E, Magrane M, Barrell D, Lee V, et al. (2004). The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res.* 32: D262-D266.

Chen F, Mackey AJ, Stoeckert CJ Jr and Roos DS (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34: D363-D368.

Chiu JC, Lee EK, Egan MG, Sarkar IN, et al. (2006). OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* 22: 699-707.

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, et al. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res.* 34: D247-D251.

Fulton DL, Li YY, Laird MR, Horsman BG, et al. (2006). Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics* 7: 270.

Geourjon C, Combet C, Blanchet C and Deléage G (2001). Identification of related proteins with weak sequence identity using secondary structure information. *Protein Sci.* 10: 788-797.

Huang H, Barker WC, Chen Y and Wu CH (2003). iProClass: an integrated database of protein family, function and structure information. *Nucleic Acids Res.* 31: 390-392.

Krishnamurthy N, Brown D and Sjölander K (2007). FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evol. Biol.* 7 (Suppl 1): S12.

Li W and Godzik A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658-1659.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, et al. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* 30: 281-283.

O'Brien KP, Remm M and Sonnhammer EL (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 33: D476-D480.

Rost B, Sander C and Schneider R (1994). Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235: 13-26.

Schultz J, Milpetz F, Bork P and Ponting CP (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.* 95: 5857-5864.

Suzek BE, Huang H, McGarvey P, Mazumder R, et al. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23: 1282-1288.

Tatusov RL, Koonin EV and Lipman DJ (1997). A genomic perspective on protein families. *Science* 278: 631-637.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.

UniProt Consortium (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 35: D193-D197.