Published online July 11, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 11 ***3309–3316***
doi:10.1093/nar/gkl433

Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits

Christophe Dessimoz*, Brigitte Boeckmann[1], Alexander C. J. Roth and Gaston H. Gonnet

ETH Zurich, Institute of Computational Science, CH-8092 Zürich and [1]Swiss Institute of Bioinformatics, CMU, Michel-Servet 1, CH-1211 Genève

Received March 14, 2006; Revised May 23, 2006; Accepted June 1, 2006

ABSTRACT

Correct orthology assignment is a critical prerequisite of numerous comparative genomics procedures, such as function prediction, construction of phylogenetic species trees and genome rearrangement analysis. We present an algorithm for the detection of non-orthologs that arise by mistake in current orthology classification methods based on genome-specific best hits, such as the COGs database. The algorithm works with pairwise distance estimates, rather than computationally expensive and error-prone tree-building methods. The accuracy of the algorithm is evaluated through verification of the distribution of predicted cases, case-by-case phylogenetic analysis and comparisons with predictions from other projects using independent methods. Our results show that a very significant fraction of the COG groups include non-orthologs: using conservative parameters, the algorithm detects non-orthology in a third of all COG groups. Consequently, sequence analysis sensitive to correct orthology assignments will greatly benefit from these findings.

INTRODUCTION

The identification of orthologous genes is a central problem in bioinformatics. Orthologs are genes that evolve from a common ancestor through speciation events, as opposed to paralogs, that result from gene duplication (1). Discriminating orthologs from paralogs is an important, but non-trivial task. It is important, because function conservation is considerably higher among orthologs (2), and also because only orthologs reflect the history of their species (1), meaning that phylogeny inferences must be based on orthologs. It is non-trivial because this distinction requires precise estimates of evolutionary distances from data that are often noisy. Other complications include gene deletion, variations in evolutionary rates, lateral gene transfer (LGT), or simply the fact that orthology and paralogy are non-transitive relations, meaning that the relation of every pair of genes must be analyzed separately.

So far, several projects have addressed this problem systematically. Of those, the COGs database (3,4) is by far the best established, probably due to its early inception, its wide scope, its reasonable performance and its presence on the NCBI website. The significance of COG in the community is reflected by hundreds of references in scientific articles. Even more importantly, most current initiatives for the identification of orthologs use ideas derived from the methodology of COG, in particular the idea of genome-specific best hit (5–7). Of all those projects depending either on the methods or results from COG, few question the accuracy of them.

In its last accessible release (2003), the COGs database groups 138 458 proteins from 66 prokaryotes into 4873 groups that consist of orthologs and in-paralogs. The term in-paralog was coined by Remm and coworkers (6) and describes in this context paralogs inside the same species ('trivial paralogs'), as opposed to out-paralogs that result from a duplication event prior to the last speciation event. [Strictly speaking, in/out-paralogy is a relation defined over two sequences and a speciation event of reference. When that event is omitted, it is here the last speciation event that is implied.] The inclusion of in-paralogs is usually justified by the fact that such sequences are orthologous to every other sequence within their group. Consequently, the relation of every pair of sequences inside the same COG is unambiguous: pairs of sequences from the same species are paralogs, otherwise, they are expected to be orthologous. The construction of COG groups is based on the fact that orthologous genes almost always have a higher level of sequence conservation than paralogs. Hence, genome-specific best hits ('BeTs') are likely to be formed between orthologs. Yet, if the corresponding ortholog is missing, a BeT might link paralogous sequences. That problem is partly taken care of by COG's approach: BeTs are only grouped when they

*To whom correspondence should be addressed. Tel: +41 44 6327472; Fax: +41 44 6321172; Email: cdessimoz@inf.ethz.ch

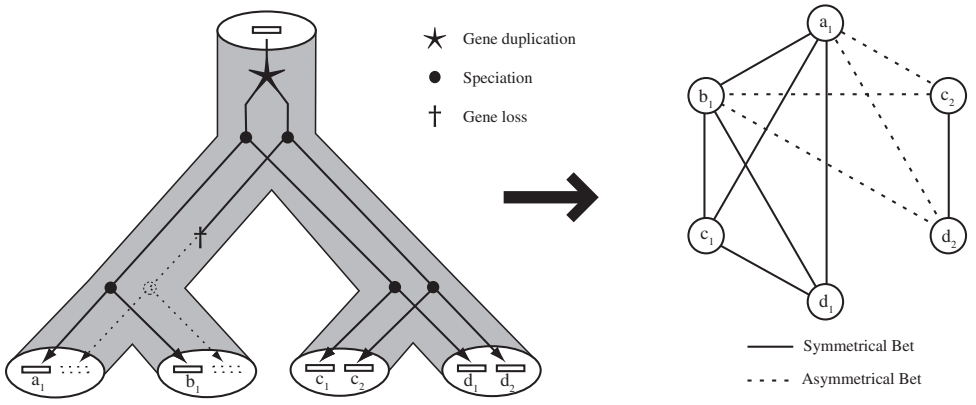

Figure 1. A simple evolutionary scenario under which the COG algorithm groups paralogous sequences.

form triangles, and triangles are merged only when they have a common side. However, if more than one species have lost the corresponding ortholog, the construction over triangles will not suffice to prevent paralogs from being clustered together. This scenario is far from being unlikely, because losses occurring before speciation events get replicated, and therefore the problem becomes very significant as more species and strains are included for analysis. In fact, simple situations, such as the one illustrated on Figure 1 are sufficient to have paralogs clustered together. It is then up to the human curation step at the end of the COG building process (3) to resolve all such cases.

The difficulty caused by a single missing ortholog can be easily avoided by requiring that all BeTs be symmetrical, which is what most other projects do. However, if the corresponding ortholog is missing in both genomes, even a symmetrical BeT will link paralogs. Therefore, BeTs, even symmetrical, are not necessarily linking orthologs.

This problem could be solved through phylogenetic analysis of the relevant gene families, in particular tree reconciliation (8), but this procedure is not yet practical in large-scale, automated contexts (2). In the following, we present an algorithm that detects non-orthology without the need of gene tree construction, then report its application on the last version of the COGs database. The algorithm was developed in the context of our own orthology classification project OMA (9), in which it is used to verify every predicted orthologous relation.

MATERIALS AND METHODS

The algorithm presented here is designed to detect non-trivial paralogous relations within groups of orthologs such as COG groups. Knowing that a paralogous relation within a group is likely to be caused by the loss of the corresponding ortholog in both species, the algorithm looks for a third-party species, which we call the 'witness of non-orthology', in which both corresponding orthologs are present (Figure 2). Under the assumptions of good and complete data, and similar evolutionary rates among orthologs, such a situation is characterized by the following three requirements on the evolutionary distances: (i) In Z, z_3 is the closest protein to x_1 and z_4 is the closest protein to y_2. (ii) The pair (x_1, z_3) must

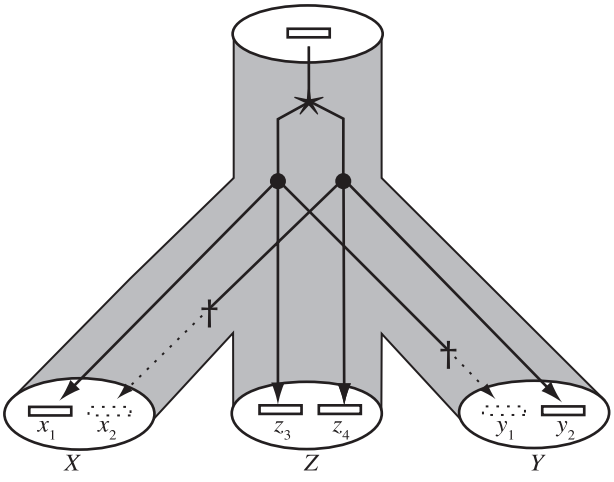

Figure 2. Suitable case of a witness. A duplication occurred before all speciations and Z is a witness of the non-orthology between the sequences x_1 and y_2.

be significantly closer than (x_1, z_4), and conversely, (y_2, z_4), must be significantly closer than (y_2, z_3), That excludes cases where z_3 and z_4 are in-paralogs (Figure 3, left), because for in-paralogs to fulfill those conditions, convergent evolution at the sequence level would be required, a phenomenon that is so unlikely that we ignore it (10). (iii) The distance between (x_1, z_4), must be similar to (y_2, z_3). That excludes cases where X (respectively Y) speciated before the duplication event, in which case x_1 (respectively y_2) is orthologous to all three other genes (Figure 3, right).

We finish this overview of the algorithm by considering the impact of LGT and gene fusion/fission. Clearly, the algorithm presented here was not designed to detect LGT events between x_1 and y_2, an interesting problem in itself that remains largely unsolved. More importantly here, an LGT in a third-party species Z can lead to a situation where Z wrongly appears to be witness of non-orthology: consider three orthologous proteins x_1, y_2 and z_3 in three species X, Y and Z. At some point, Z acquires through LGT a member of that orthologous family, which we now refer to as z_4. Z keeps both copies z_3 and z_4. Furthermore, Z happens to be closer to X than Y, while the donor of z_4 is closer to Y than X. This situation leads to a misclassification by our

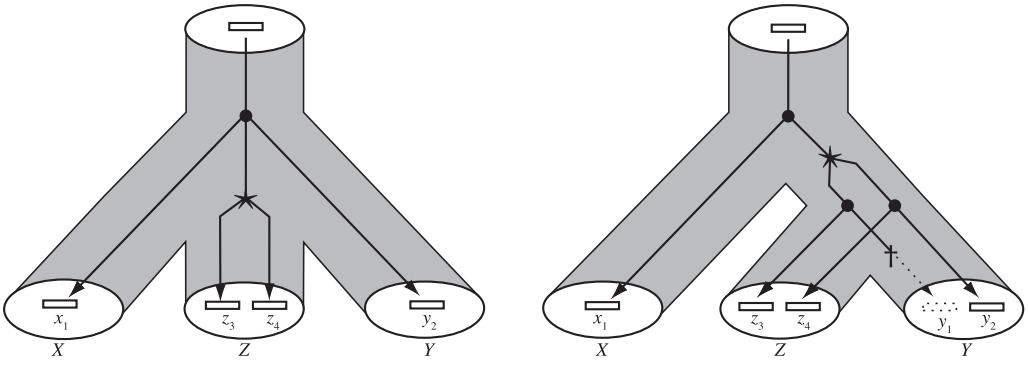

Figure 3. Unsuitable cases of witnesses. To the left, duplication occurred only in *Z*, and therefore z_3 and z_4 are in-paralogs with respect to (*X*, *Y*) and cannot act as witness of non-orthology. To the right, *X* speciated before the duplication event. Hence, x_1 is orthologous to all three other proteins and cannot act as witness of non-orthology.

algorithm. Although such cases cannot be ruled out, we did not encounter any among the numerous case-by-case analysis performed on the results. It could be that orthologous gene displacement of z_3 by z_4 through homologous recombination is a much more likely scenario, and besides, the frequency of LGT appears to be higher among closely related species (11). As for gene fusion or gene fission, the units for amino acid sequence analysis are no longer proteins but domains. Even though the analysis of homologous domains from distinct proteins is scientifically meaningful, our analysis remains at the level of entire proteins to simplify matters.

Note that the complications caused by LGT events and, probably to a lesser extent, by gene fusion/fission are not specific to our method and pose challenges to other approaches as well, in particular tree reconciliation.

Input data

The algorithm uses two inputs: the COGs database and pairwise sequence alignments between all proteins involved in the analysis. As introduced above, the orthology of two sequences is verified through an exhaustive search of the corresponding sequences in complete, third-party genome. Therefore, a large number of genomes is desirable. However, since the relation between every pair of sequence is needed, such searches require the computation of a very large number of pairwise alignments. For practical reasons, all results presented here use results from the Smith–Waterman (12) all-against-all protein alignments precomputed in the scope of the OMA project (9).

For each alignment, a PAM distance estimate and the corresponding variance is computed using maximum likelihood and numeric integration (13,14).

Comparison of evolutionary distances

The algorithm uses evolutionary distances to detect paralogs. However, the distances estimates are subject to perturbation, which must be taken into account when comparing them. Therefore, assuming that errors are normally distributed, the difference $\Delta(d_1, d_2)$ of two distances d_1, d_2 has expected value:

$$E[\Delta(d_1, d_2)] = E(d_1) - E(d_2)$$

with variance

$$\sigma^2[\Delta(d_1, d_2)] = \sigma^2(d_1) + \sigma^2(d_2) - 2\mathrm{Cov}(d_1, d_2)$$

If the two distances are independent, the covariance term disappears and the variance of the difference can be obtained directly from the individual variances. But more often than not, d_1 and d_2 involve a common protein and are therefore not independent, meaning that not taking the covariance into account overestimates the error. We have developed a method to approximate the covariance of two evolutionary distances, which will be the subject of a separate article.

Algorithm

The algorithm goes through each COG group, and verifies inside each of them that every two genes x_1, y_2 coming from different species have a significant alignment, and are indeed orthologs. Alignments are considered significant if the score is above 130 (47 bits, which typically corresponds to an *E*-value around 2e−6) and the length of the alignment not <50% of the smallest sequence. The verification of orthology is performed through the search, in each third-party genome *Z*, of two genes z_3 and z_4 that fulfill the three conditions (i–iii) presented at the beginning of this section:

$$\begin{aligned} &\forall z_i \neq z_3 : \Delta(x_1 z_3, x_1 z_i) < k \cdot \sigma[\Delta(x_1 z_3, x_1 z_i)] \\ &\forall z_j \neq z_4 : \Delta(y_2 z_4, y_2 z_j) < k \cdot \sigma[\Delta(y_2 z_4, y_2 z_j)] \end{aligned} \quad \mathbf{1}$$

$$\begin{aligned} &\Delta(x_1 z_4, x_1 z_3) > k \cdot \sigma[\Delta(x_1 z_4, x_1 z_3)] \\ &\Delta(y_2 z_3, y_2 z_4) > k \cdot \sigma[\Delta(y_2 z_3, y_2 z_4)] \end{aligned} \quad \mathbf{2}$$

$$| \Delta(x_1 z_4, y_2 z_3) | < k \cdot \sqrt{[\sigma^2(x_1 z_4) + \sigma^2(y_2 z_3)]}, \quad \mathbf{3}$$

where *k* is the confidence level, which we set to 1.96. If the quartet (x_1, y_2, z_3, z_4), fulfills all three conditions, there is enough evidence to consider x_1, y_2 paralogs. The algorithm was implemented in the programming environment Darwin (15).

A note about parameter choice. As mentioned previously, the classification of protein pairs in orthologs and non-orthologs can be very difficult or even impossible, especially when a speciation event immediately follows a duplication event, or in the situation of frequent gene gain and gene

loss, as it is observed in certain groups of proteins, such as metabolic enzymes. Here, the choice of $k = 1.96$ standard deviations was established empirically such that the false-positive rate (orthologs misclassified as non-orthologs) is much smaller than the false-negatives rate (missed non-orthologs). In other words, we expect that our algorithm reports only clear-cut cases of paralogy.

Phylogenetic analysis

To verify individual cases reported by the algorithm, phylogenetic trees were constructed using independent, common software packages, as follows: sequences were aligned using Muscle (16) and ClustalW (17). Whenever they differed, the one that seemed more likely was selected. Short sequences, suspicious regions and most gap-containing columns removed. Distance matrices (JTT, gamma) generated with protdist (18) were used to construct phylogenetic trees using neighbor (18). Clusters of interest were selected for detailed analysis. Alignments of the selected data were performed using Tcoffee (19) and the result subsequently modified as described above, and considering the Tcoffee CORE (consistency of overall residue evaluation) values for the alignment. Information on the stability of the tree topology was assessed building an extended majority rule consensus tree using consense (18) from BIONJ (20) searches performed on 1000 bootstrap replicates, which were constructed with seqboot (18). Protein trees of the data subset were constructed using the Bayesian tree-building method MrBayes (21) (JTT; invgamma-4; 1 000 000 generations). The trees were rooted using an outgroup whenever a suitable ancient paralog could be found. Note that since the analysis attempts at clustering homologs into clans, and not at predicting their hierarchical order, placement of the root is not critical here.

Validation

The performances of the algorithm were evaluated using the HAMAP database (22), a collection of orthologous microbial protein families generated manually by expert curators in the Swiss–Prot group. The database was retrieved on November 23, 2005. Proteins from the 99 most represented species also present in our OMA project were used in the analysis: of all 29 245 proteins, there were 21 831 proteins (75.6%), grouped in 1189 orthologous families. That yielded 309 829 pairwise relations to be verified by our procedure.

The algorithm classified 279 568 (90.2%) relations as orthologous and 9420 (3.0%) as paralogous. The remaining 20 841 (6.7%) relations had alignments below our significance threshold and could therefore not be processed. The accuracy of the algorithm, in particular its very low false-positive rate was confirmed by following observations:

First, paralogy is often reflected by different Swiss–Prot ID names (e.g. GREA/GREB) (23). From the 9420 predicted paralogs, only 2728 (29.0%) of them have identical ID names. Second, the distribution of the paralogs among HAMAP families was investigated: all 9420 cases of paralogy found by the algorithm are concentrated in only 150 (12.6%) of the 1189 HAMAP families. This is consistent with the fact that the inclusion of just one paralogous protein into an orthologous family is likely to result in several paralogous relations inside that family. And indeed, in all except 8 of these 150 families, more than one paralogous pair was detected. Third, these 8 improbable cases were inspected individually using phylogenetic analysis, which confirmed that they are bona fide paralogs (possibly xenologs). Fourth, the predicted cases of paralogy were compared to the gene trees over HAMAP families built by the group of Laurent Duret (http://pbil.univ-lyon1.fr/help/HAMAP.html), in a similar way as HOBACGEN (24). 7217 predicted cases could be mapped to those trees. In 6418 (88.9%) instances, paralogy was confirmed by the trees, a remarkably high level of consistency considering that the two methods are very different. As for the conflicting 799 cases, which are distributed among 51 families, we believe that most of them are caused by inaccuracies on the gene trees, which are constructed using a variant of Neighbor Joining on observed divergence, a rather crude measure of evolutionary distance.

RESULTS AND DISCUSSION

The algorithm was run on the current release of the COGs database (4) (http://www.biomedcentral.com/1471–2105/4/41). We used the precomputed all-against-all results from 107 complete genomes, of which 52 are represented in COGs, whereas the remaining 55 genomes were only used as potential witnesses of non-orthology. [The complete list is available in the Supplementary Data.] From all 4654 COGs, there is a total of 5 537 713 pairwise relations. Pairs between proteins from the same species (484 043) were not considered further. Additionaly, 2 733 371 relations involve at least one protein from a species outside our set of 107 genomes. Consequently, the following results were obtained through the verification of 2 320 199 relations, 45.9% of all potential orthologous relations.

The results are presented in Table 1. Surprisingly, 44% of the relations had alignment scores below our significance threshold of 130, which corresponds to an *E*-value of about 2e−6, and could therefore not be verified. This implies that an important fraction of relations within COGs cannot be, on the basis of pairwise alignments, reliably considered homologous.

The other result is the significant proportion of non-orthologous relations found by the algorithm, more than a quarter of the pairs that could be verified. They are distributed among about a third of all COGs. The list of such groups, along with all detected non-orthology cases are available in the Supplementary Data.

If we require the presence of at least two witnesses of non-orthology for a pair to be considered non-orthologous, the

Table 1. Results of the algorithm on the COGs database

	#	%
Pairs with score below threshold, not tested	1 021 764	44.0
Pairs with score above threshold	1 298 435	66.0
Non-orthologous pairs	360 856	27.8
Orthologous pairs	937 579	72.2
COG groups with non-orthology	1604	34.5
COG groups without non-orthology	3050	65.5

algorithm still finds 251 391 (19.4%) such pairs within 1146 (24.6%) COGs. When removing the sequence with the most non-orthologous relations from each COG group, the total number of non-orthologous pairs decreases by only 24 868 (1.9%).

The majority (70%) of the groups predominantly non-orthologs are involved in metabolic processes, according to the functional description of the COGs database, although they only constitute a minority of all COGs. In contrast, groups involved in information storage and processing (8%) or cellular processing and signaling (11%) include less frequently non-orthologs. The remainder 11% are poorly characterized proteins. This result is in agreement with previous studies, which state that in prokaryotes, metabolic functions are under high evolutionary pressure from changing environments (25).

Phylogenetic analysis of selected COG groups

The presence of non-orthology in some COG groups is hardly a surprise and was in fact recently acknowledged by Koonin, coauthor of COG, in a review article (2). What is surprising here is rather the extent of non-orthology detected by the algorithm. That prompted us to verify, in addition to the validation work reported in the previous section, a number of our predictions using detailed phylogenetic analysis. In this section, we report the conclusion of such analysis on three COGs, for which we could build Bayesian likelihood trees of high confidence, confirmed by consensus NJ trees with high bootstrap values. Clan assignments were made based on those trees, and considering lineage and function, whenever reliable annotations could be found. We strongly expect that pairs of proteins across clans be non-orthologous, and use these results to evaluate the accuracy of the predictions made by the algorithm.

COG0508 consists of complex-forming acyltransferases that are composed of an N-terminal biotin or lipoic acid attachment domain, a central protein–protein interaction domain, followed by the catalytic 2-oxoacid dehydrogenases acyltransferase domain. The phylogenetic analysis of roughly half of the proteobacterial sequence data from COG0508 suggests the existence of at least four distinct subgroups (see Figure 4): clan 1 is formed by sequences from gamma-proteobacteria, including the dihydrolipoyllysine-residue acetyltransferase component of the pyruvate dehydrogenase complex (EC 2.3.1.12) (AceF) from *Escherichia coli*. Clan 2 consists of proteins highly similar to the *Bacillus subtilis* lipoamide acyltransferase component of the branched-chain alpha-keto acid dehydrogenase complex (EC 2.3.1.168). All sequences in clan 2 are alphaproteobacterial, except for *Pseudomonas aeruginosa* proteins, which are found in both clan 1 and clan 2. As mentioned in section 2, such situation could arise through lateral gene transfer from an alphaproteobacteria to *P.aeruginosa*. If that was the case, there would be strong evidence that clans 1 and 2 should be merged. However, in the present case, it is possible to populate both clans with additional sequences from more distant species (data not shown), legitimating the separation in two clans. Additionally, the long distance between the two clans and the distinct function of at least one family member of each subgroup also supports this conclusion. Clan 3 includes the

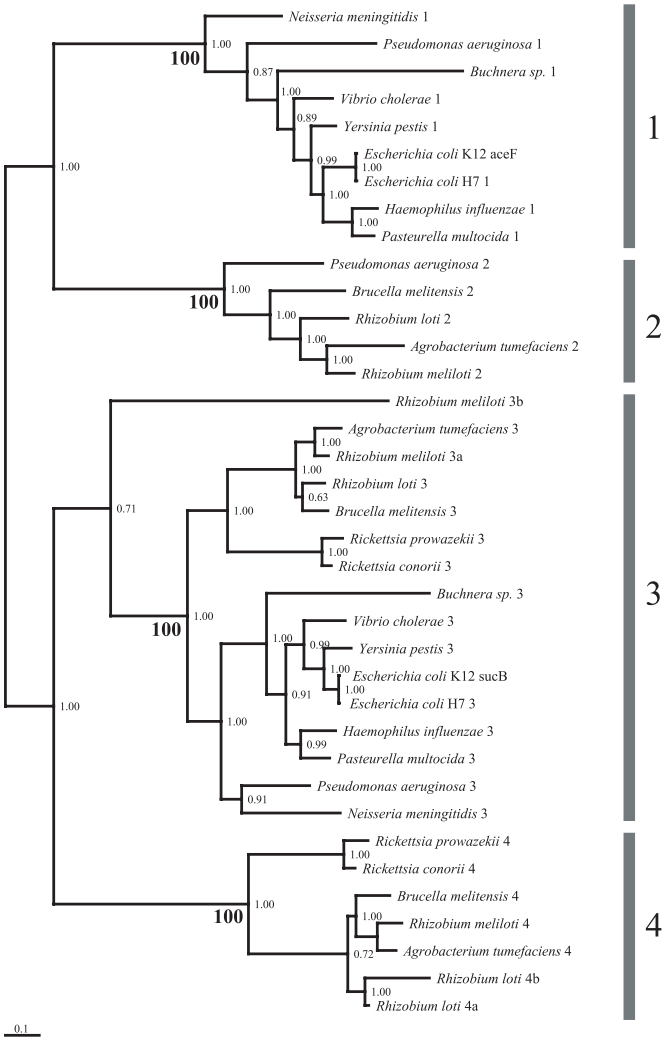

Figure 4. Unrooted phylogenetic consensus tree constructed from a Bayesian analysis of a subgroup from COG0508. Posterior probabilities are indicated to the right of the nodes and clan-supporting bootstrap values are indicated below the probability value. Predicted clans are indicated by the vertical bars on the right side. The leaf labels correspond to the following COG identifiers: *Agrobacterium tumefaciens* (2: AGl2719, 3: AGc4775, 4: AGc2641), *Brucella melitensis* (2: BMEII0746, 3: BMEI0141, 4: BMEI0856), *Buchnera sp.* (1: BU206, 3: BU303), *E.coli* K12 (COG identifier corresponds to the gene name: aceF, sucB), *E.coli* H7 (1: ECs0119, 3: ECs0752), *Haemophilus influenzae* (1: HI1232, 3: HI1661), *Neisseria meningitidis* (1: NMB1342, 3: NMB0956), *Pasteurella multocida* (1: PM0894, 3: PM0278), *Pseudomonas aeruginosa* (1: PA5016, 2: PA2249, 3: PA1586), *Rhizobium loti* (2: mll4471, 3: mll4300, 4a: mlr0385, 4b: mll3627), *Rhizobium meliloti* (2: SMc03203, 3a: SMc02483, 3b: SMb20019, 4: SMc01032), *Rickettsia conorii* (3: RC0226, 4: RC0764), *Rickettsia prowazekii* (3: RP179, 4: RP530), *Vibrio cholerae* (1: VC2413, 3: VC2086), *Y.pestis* (1: YPO3418, 3: YPO1114).

dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex (EC 2.3.1.61) (SucB) of *E.coli*. Note that clan 3 includes two protein sequences of *Rhizobium meliloti*, but those are clearly ancient duplicates, and thus sequence 3b is likely to form yet a separate clan on its own. Finally clan 4 is formed by a presumably further dehydrogenase component from alpha-proteobacteria. The algorithm predicted 382 cases of non-orthologous relations within the sequences considered here. An extract of the result list is given in Table 2 (the full list

Table 2. Predicted non-orthologous relations for the data shown in Figure 4

	Predicted non-orthologs	Pair of witnesses
A.tumefaciens 2	*Buchnera sp. 1*	*P.aeruginosa 2 + 1*
A.tumefaciens 2	*E.coli H7 1*	*P.aeruginosa 2 + 1*
A.tumefaciens 2	*E.coli K12 acef*	*P.aeruginosa 2 + 1*
A.tumefaciens 2	*H.influencae 1*	*P.aeruginosa 2 + 1*
A.tumefaciens 2	*Neisseria menininginitidis 1*	*P.aeruginosa 2 + 1*
A.tumefaciens 2	*P.multocida 1*	*P.aeruginosa 2 + 1*
A.tumefaciens 2	*R.loti 4a*	*B.melitensis 2 + 4*
A.tumefaciens 2	*R.meliloti 4*	*B.melitensis 2 + 4*
A.tumefaciens 2	*V.cholerae 1*	*P.aeruginosa 2 + 1*
A.tumefaciens 2	*Y.pestis 1*	*P.aeruginosa 2 + 1*
A.tumefaciens 3	*B.melitensis 2*	*Buchnera sp. 3 + 1*
A.tumefaciens 3	*Buchnera sp. 1*	*P.aeruginosa 3 + 1*
A.tumefaciens 3	*P.aeruginosa 2*	*B. melitensis 3 + 2*
A.tumefaciens 3	*P.aeruginosa 1*	*Buchnera sp. 3 + 1*
A.tumefaciens 3	*P.multocida 1*	*Buchnera sp. 3 + 1*
A.tumefaciens 3	*R.conorii 4*	*B.melitensis 3 + 4*
A.tumefaciens 3	*R.loti 2*	*B.melitensis 3 + 2*
A.tumefaciens 3	*R.loti 4a*	*B.melitensis 3 + 4*
A.tumefaciens 3	*R.meliloti 2*	*B.melitensis 3 + 2*
A.tumefaciens 3	*R.meliloti 4*	*B.melitensis 3 + 4*
A.tumefaciens 3	*R.prowazekii 4*	*B.melitensis 3 + 4*
A.tumefaciens 4	*B.melitensis 2*	*R.loti 4a + 2*
A.tumefaciens 4	*Buchnera sp. 1*	*B.melitensis 4 + 2*
A.tumefaciens 4	*E.coli H7 3*	*B.melitensis 4 + 3*
A.tumefaciens 4	*E.coli K12 sucB*	*B.melitensis 4 + 3*
A.tumefaciens 4	*H.influencae 1*	*R.loti 4a + 2*
A.tumefaciens 4	*H.influencae 3*	*B.melitensis 4 + 3*
A.tumefaciens 4	*N.menininginitidis 1*	*B.melitensis 4 + 2*
A.tumefaciens 4	*N.menininginitidis 3*	*B.melitensis 4 + 3*
A.tumefaciens 4	*P.aeruginosa 2*	*B.melitensis 4 + 2*
A.tumefaciens 4	*P.aeruginosa 3*	*B.melitensis 4 + 3*
A.tumefaciens 4	*P.multocida 1*	*B.melitensis 4 + 2*
A.tumefaciens 4	*P.multocida 3*	*B.melitensis 4 + 3*
A.tumefaciens 4	*R.conorii 3*	*B.melitensis 4 + 3*
A.tumefaciens 4	*R. oti 2*	*B.melitensis 4 + 2*
A.tumefaciens 4	*R.loti 3*	*B.melitensis 4 + 3*
A.tumefaciens 4	*R.meliloti 2*	*B.melitensis 4 + 2*
A.tumefaciens 4	*R.meliloti 3a*	*B.melitensis 4 + 3*
A.tumefaciens 4	*R.prowazekii 3*	*B.melitensis 4 + 3*
A.tumefaciens 4	*V.cholerae 3*	*B.melitensis 4 + 3*
A.tumefaciens 4	*Y.pestis 3*	*B.melitensis 4 + 3*
B.melitensis 2	*E.coli H7 1*	*P.aeruginosa 2 + 1*
B.melitensis 2	*E.coli H7 3*	*A.tumefaciens 2 + 3*
B.melitensis 2	*E.coli K12 acef*	*P.aeruginosa 2 + 1*
B.melitensis 2	*E.coli K12 sucB*	*A.tumefaciens 2 + 3*
B.melitensis 2	*H.influencae 1*	*P.aeruginosa 2 + 1*
B.melitensis 2	*H.influencae 3*	*A.tumefaciens 2 + 3*
B.melitensis 2	*N.menininginitidis 1*	*P.aeruginosa 2 + 1*
B.melitensis 2	*N.menininginitidis 3*	*A.tumefaciens 2 + 3*
B.melitensis 2	*P.aeruginosa 3*	*A.tumefaciens 2 + 3*
B.melitensis 2	*P.multocida 1*	*P.aeruginosa 2 + 1*
B.melitensis 2	*P.multocida 3*	*A.tumefaciens 2 + 3*
B.melitensis 2	*R.conorii 3*	*A.tumefaciens 2 + 3*
B.melitensis 2	*R.conorii 4*	*A.tumefaciens 2 + 4*
B.melitensis 2	*R.loti 3*	*A.tumefaciens 2 + 3*

The sequences in the first two columns are predicted to be non-orthologous by the pair of witnesses in the third column.

of paralogy is available in the Supplementary Data). A total of 379 predictions are consistent with the clan assignment, while the remaining three predictions support the exclusion of *R.meliloti* 3b from clan 3. Furthermore, comparison with the clan assignment reveals that the algorithm missed 24 non-orthologous relations, which implies a false-negative rate of 6.0%.

COG0513 includes various DEAD-box containing RNA helicases. The phylogenetic analysis of the proteobacterial

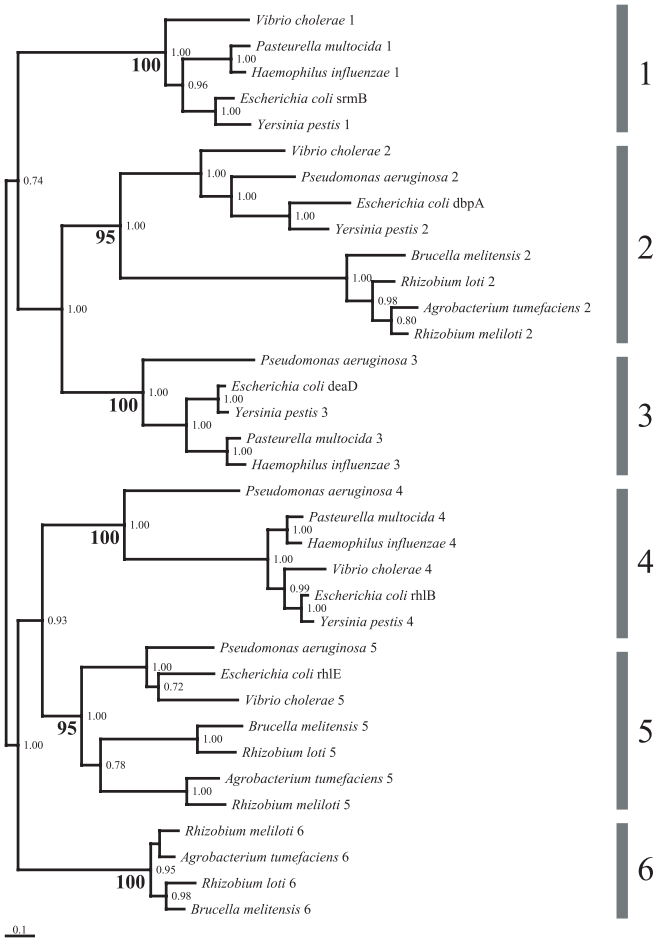

Figure 5. Unrooted phylogenetic consensus tree for COG0513, constructed from a Bayesian analysis. Posterior probabilities are drawn to the right of the nodes and clan-supporting bootstrap values are below the relevant nodes. The vertical bars bars to the right indicate the prediced clans. The leaf labels correspond to the COG identifiers: *A.tumefaciens* (2: AGl1362, 5: AGc4238, 6: AGc3366), *B.melitensis* (2: BMEI1824, 5: BMEI0934, 6: BMEI1035), *E.coli* K12 (COG identifier corresponds to the gene name: dbpA, deaD, rhlB, rhlE, srmB), *H.influenzae* (1: HI0422, 3: HI0231, 4: HI0892), *P.multocida* (1: PM1840, 3: PM1112, 4: PM1921), *P.aeruginosa* (2: PA0455, 3: PA2840, 4: PA3861, 5: PA0428), *R.loti* (2: mlr4393, 5: mlr0349, 6: mll0224), *R.meliloti* (2: SMc01090, 5: SMb20880, 6: SMc00522), *V.cholerae* (1: VC0660, 2: VC2564, 4: VC0305, 5: VCA0204), *Y.pestis* (1: YPO2708, 2: YPO1776, 3: YPO3488, 4: YPO3869).

data from this group suggests the existence of six clans (see Figure 5), of which five are formed around the following proteins from *E.coli*: (i) the ATP-dependent RNA helicase SrmB, which is involved in an early assembly step of 50S ribosomal subunits (26); (ii) the cold-shock DEAD-box protein A (DeaD), required for cell division and normal cell growth at low temperature (27); (iii) the DEAD-box RNA helicase B (RhlB), a component of the RNA degradosome, which seems to have little activity unless being activated by the endoribonuclease RNase E (28); (iv) the putative RNA helicase RhlE, which has been shown to be non-essential for normal cell growth (29); (v) the ATP-independent RNA 3′→5′ helicase DbpA (30) and (vi) the subgroup includes RNA helicases that are conserved in some alphaproteobacteria. The algorithm predicted 408 cases of non-orthology, 88.9% of the 459 non-orthologous

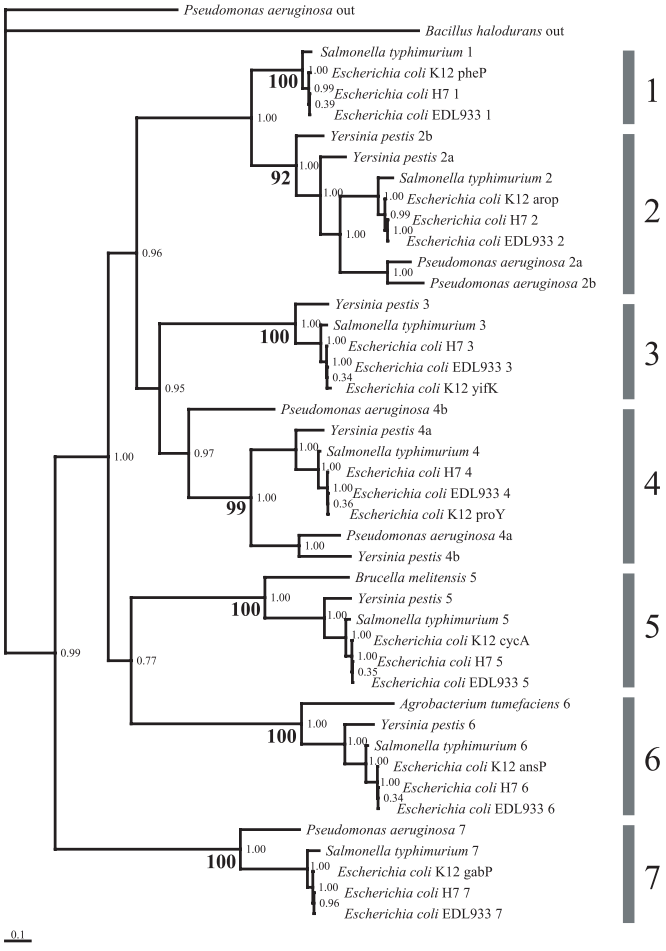

Figure 6. Phylogenetic consensus tree rooted by outgroups for COG1113, constructed from a Bayesian analysis of a data subgroup from COG1113. Posterior probabilities of the Bayesian analysis are drawn to the right of the nodes and clan-supporting bootstrap values below relevant nodes. Predicted clans are indicated by vertical bars to the right. The leaf labels correspond to the COG identifiers: *A.tumefaciens* C58 (6: AGl2082), *Bacillus halodurans* (out: BH2171), *B.melitensis* (5: BMEII0038), *E.coli* K12 (COG identifier corresponds to the gene name: ansP, aroP, cycA, gabP, pheP, proY, yifK), *E.coli* H7 EDL933 (1: ZpheP, 2: ZaroP, 3: ZyifK, 4: ZproY, 5: ZcycA, 6: ZansP, 7: ZgabP), *E.coli* H7 (1: ECs0614, 2: ECs0116, 3: ECs4729, 4: ECs0452, 5: ECs5186, 6: ECs2057, 7: ECs3524), *P.aeruginosa* (2a: PA3000, 2b: PA0866, 4a: PA5097, 4b: PA0789, 7: PA0129, out: PA2079), *Salmonella typhimurium* LT2 (1: STM0568, 2: STM0150, 3: STM3930, 4: STM0400, 5: STM4398, 6: STM1584, 7: STM2793), *Y.pestis* (2a: YPO3421, 2b: YPO1743, 3: YPO3854, 4a: YPO3201, 4b: YPO4015, 5: YPO1859, 6: YPO1937).

relations that can be deduced from the clan assignment. In this case, there was no false-positive prediction.

COG1113 consists of members of the amino acid-polyamine-organo-cation (APC) superfamily from bacteria, specifically those integral membrane proteins that are involved in the transport of amino acids in prokaryotes. The phylogenetic analysis of this group suggests the existence of various clans (see Figure 6), including those formed around the seven proteins found in *E.coli*: (i) phenylalanine-specific permease (PheP), (ii) aromatic amino acid transport protein (AroP), (iii) probable transport protein YifK, (iv) proline-specific permease (ProY), (v) D-serine/D-alanine/glycine transporter (CycA), (vi) L-asparagine permease (AnsP), (vii) GABA (4-aminobutyrate) permease (GabP). The seven clans were predicted with high probability and their clusterings confirmed by significant bootstrap values (99–100%) except for one (92%). The analyzed dataset includes members of quite related organisms, but most clans can already be populated with further members from other species of COG1113. The algorithm predicted 257 pairs of non-orthologs, of which 254 are consistent with the phylogenetic analysis. That represents 97.7% of the 260 non-orthologous relations that can be deduced from the clan assignment. The conflicting three predictions suggest that *P.aeruginosa* 4a is non-orthologous to *E.coli K12* ProY and to *E.coli H7 EDL933* 4, and that *P.aeruginosa* 4b is non-orthologous to *Yersinia pestis* 4b. But here too, the extension of the phylogenetic analysis using additional sequences from the UniProtKB database supports the division of clan 4 into further subgroups (data not shown).

CONCLUSION

We present here a new algorithm for the detection of non-orthologous relations caused by the limitations of genome-specific best hit methods, such as the COGs database. The algorithm, rather than building gene trees, a process both computationally expensive and error-prone, works with pairwise distance estimates. The accuracy of the algorithm was evaluated through verification of the distribution of predicted cases, case-by-case phylogenetic analysis and comparisons with prediction from other projects using independent methods. Using conservative parameters, the algorithm detected non-orthology in a third of the COG groups. Methods sensitive to correct orthology assignments, such as function prediction, phylogenetic trees or genome rearrangement analysis, will profit from both the algorithm and the results presented here.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank G. Cannarozzi, D. Margadant, A. Schneider and two anonymous reviewers for their comments and suggestions on the manuscript.

Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool*., **19**, 99–113.
2. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet*., **39**, 309–338.
3. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
4. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al*. (2003) The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
5. Fujibuchi,W., Ogata,H., Matsuda,H. and Kanehisa,M. (2000) Automatic detection of conserved gene clusters in multiple genomes by

graph comparison and p-quasi grouping. *Nucleic Acids Res*., **28**, 4029–4036.
6. Remm,M., Storm,C. and Sonnhammer,E. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol*., **314**, 1041–1052.
7. Lee,Y., Sultana,R., Pertea,G., Cho,J., Karamycheva,S., Tsai,J., Parvizi,B., Cheung,F., Antonescu,V., White,J. *et al.* (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res*., **12**, 493–502.
8. Goodman,M., Czelusniak,J., Moore,G.W. and Romero-Herrara,A.E. (1979) Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool*., **28**, 132–168.
9. Dessimoz,C., Cannarozzi,G., Gil,M., Margadant,D., Roth,A., Schneider,A. and Gonnet,G.H. (2005) OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In McLysath,A. and Huson,D.H. (eds), *Lecture Notes in Computer Science*, Vol. 3678, Springer-Verlag, pp. 61–72.
10. Doolittle,R.F. (1994) Convergent evolution: the need to be explicit. *Trends Biochem. Sci*., **19**, 15–18.
11. Lawrence,J.G. and Hendrickson,H. (2003) Lateral gene transfer: when will adolescence end? *Mol. Microbiol*., **50**, 739–749.
12. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.
13. Gonnet,G.H. (1994) A *Tutorial Introduction to Computational Biochemistry Using Darwin. Technical Report Informatik*. ETH Zurich, Switzerland.
14. Muller,T. and Vingron,M. (2000) Modeling amino acid replacement. *J. Comput. Biol*., **7**, 761–776.
15. Gonnet,G.H., Hallett,M.T., Korostensky,C. and Bernardin,L. (2000) Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics*, **16**, 101–103.
16. Edgar,R.C. (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
17. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*., **31**, 3497–3500.
18. Felsenstein,J. (1993) Phylip (phylogeny inference package) version 3.5c. distributed by the author.
19. Poirot,O., O'Toole,E. and Notredame,C. (2003) Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res*., **31**, 3503–3506.
20. Gascuel,O. (1997) Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Mol. Biol. Evol*., **14**, 685–695.
21. Ronquist,F. and Huelsenbeck,J.P. (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
22. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J.A., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in swiss-prot. *Comput. Biol. Chem*., **27**, 49–58.
23. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*., **31**, 365–370.
24. Perriere,G., Duret,L. and Gouy,M. (2000) Hobacgen: database system for comparative genomics in bacteria. *Genome Res*., **10**, 379–385.
25. Pal,C., Papp,B. and Lercher,M.J. (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genet*., **37**, 1372–1375.
26. Charollais,J., Pflieger,D., Vinh,J., Dreyfus,M. and Iost,I. (2003) The DEAD-box RNA helicase srmb is involved in the assembly of 50s ribosomal subunits in *Escherichia coli*. *Mol. Microbiol*., **48**, 1253–1265.
27. Jones,P.G., Mitta,M., Kim,Y., Jiang,W. and Inouye,M. (1996) Cold shock induces a major ribosomal-associated protein that unwinds double-stranded RNA in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **93**, 76–80.
28. Carpousis,A.J. (2002) The *Escherichia coli* RNA degradosome: structure, function and relationship in other ribonucleolytic multienzyme complexes. *Biochem. Soc. Trans*., **30**, 150–155.
29. Ohmori,H. (1994) Structural analysis of the rhle gene of *Escherichia coli*. *Jpn. J. Genet*., **69**, 1–12.
30. Diges,C.M. and Uhlenbeck,O.C. (2005) *Escherichia coli* dbpa is a $3' \rightarrow 5'$ RNA helicase. *Biochemistry*, **44**, 7903–7911.