BIOINFORMATICS

Vol. 22 no. 14 2006, pages e9–e15
doi:10.1093/bioinformatics/btl213

Automatic clustering of orthologs and inparalogs shared by multiple proteomes

Andrey Alexeyenko[1,2], Ivica Tamas[1,3], Gang Liu[1] and Erik L.L. Sonnhammer[1,2,*]

[1]Center for Genomics and Bioinformatics, Karolinska Institutet, S-17177 Stockholm, [2]Stockholm Bioinformatics Center, Albanova, Stockholm University, SE-10691 Stockholm, Sweden and [3]Present address: Department of Molecular Biology & Functional Genomics, Stockholm University, SE-10691, Stockholm, Sweden

ABSTRACT

Motivation: The complete sequencing of many genomes has made it possible to identify orthologous genes descending from a common ancestor. However, reconstruction of evolutionary history over long time periods faces many challenges due to gene duplications and losses. Identification of orthologous groups shared by multiple proteomes therefore becomes a clustering problem in which an optimal compromise between conflicting evidences needs to be found.

Results: Here we present a new proteome-scale analysis program called MultiParanoid that can automatically find orthology relationships between proteins in multiple proteomes. The software is an extension of the InParanoid program that identifies orthologs and inparalogs in pairwise proteome comparisons. MultiParanoid applies a clustering algorithm to merge multiple pairwise ortholog groups from InParanoid into multi-species ortholog groups. To avoid outparalogs in the same cluster, MultiParanoid only combines species that share the same last ancestor.

To validate the clustering technique, we compared the results to a reference set obtained by manual phylogenetic analysis. We further compared the results to ortholog groups in KOGs and OrthoMCL, which revealed that MultiParanoid produces substantially fewer outparalogs than these resources.

Availability: MultiParanoid is a freely available standalone program that enables efficient orthology analysis much needed in the post-genomic era. A web-based service providing access to the original datasets, the resulting groups of orthologs, and the source code of the program can be found at http://multiparanoid.cgb.ki.se.

Contact: Erik.Sonnhammer@sbc.su.se

Supplementary information: http://multiparanoid.cgb.ki.se/ISMB2006/

1 INTRODUCTION

The increasing availability of complete proteomes provides the opportunity to reconstruct their evolutionary history based on sequence data. This is particularly welcomed by functional and comparative genomics, which is heavily dependent on orthology analysis. Orthologous genes exist in many guises, ranging from proteins with identical functions in identical pathways to proteins that share a common evolutionary origin but have diverged in function. Establishing orthology between genes is today one of the most reliable methods to obtain functional annotation.

In this paper we consider orthologs as defined by Fitch (1970): genes descending from a single gene in the last common ancestor of the species. Such genes are most likely to be functional counterparts. On the other hand, genes arising from duplications are defined as paralogs. Genomes of invertebrates and higher organisms are notorious for high numbers of gene duplications and/or gene losses. Such genomic variation has been explained as an adaptation to different environments (Chervitz *et al*., 1998; Troemel *et al*., 1995; Enmark and Gustafsson, 2001; Maglich *et al*., 2001).

Paralogs may arise from a duplication that occurred either before or after the speciation event that gave rise to the species of interest. If the duplication occurred first, the genes resulting from the duplication cannot be orthologs. Such genes are called outparalogs (Sonnhammer and Koonin, 2002). However, if the duplication happened after the speciation, the resulting genes can be considered co-orthologs. Such genes are called inparalogs. Given that the goal is to identify the complete set of orthologs and avoid non-orthologs, one wants to find all inparalogs while avoiding all outparalogs. A simplification of the problem would be to consider only the most similar inparalogs as true orthologs. However, there is often no clear functional distinction between inparalogs in the same group (Kondrashov *et al*., 2002).

The best orthology analysis is obtained from careful manual inspection of phylogenetic trees, for instance as was done by Wheelan *et al*., (1999) to identify human-mouse-rat-worm orthologs. However, this is very labor-intensive, and to save time many groups have resorted to using high-scoring global BLAST (Altschul *et al*., 1997) matches to approximate orthologs (e.g. Rubin *et al*., 2000). The BLAST approach can be substantially improved by only accepting reciprocally best matching protein pairs as orthologs (Mushegian *et al*., 1998). This approach works reasonably well for the proteomes of bacteria. However, its application to diversified eukaryotic species faces additional problems due to a complex evolutionary past (Xie and Ding, 2000).

The COG method (Tatusov *et al*., 1997) extends the reciprocal best matching method to allow incorporation of multiple species into each ortholog group. It has the ability to include inparalogs, but because it groups sequences of widely different evolutionary distances in a single cluster, out-paralogs are also commonplace. COGs

*To whom correspondence should be addressed.

initially contained only prokaryotic proteomes, but a version of seven eukaryotic species—KOGs—has been released (Tatusov *et al*., 2003). Lee *et al*. (2002) applied the COG method to cDNA sequences of 28 eukaryotes, resulting in the EGO (formerly TOGA) database.

OrthoMCL represents a different approach to finding multi-species ortholog groups. It uses a Markov clustering algorithm based on graph flow theory, and can find clusters of desired tightness depending on the ''inflation parameter'' (Li *et al*., 2003). With the parameters they used, OrthoMCL was much stricter than EGO and KOGs with regard to the inclusion of outparalogs. The OrthoMCL web resource initially included *E. coli* and nine eukaryotic proteomes; the latest release contains 55 proteomes (Chen *et al*., 2006). A drawback with the above methods is that they do not provide confidence values for the predicted orthologs. They also do not necessarily have a unique last common ancestor in each group, which can lead to inclusion of outparalogs in the same cluster.

The InParanoid method was specifically designed to find inparalogs by a special extension of the reciprocal best matching method in pairwise proteome comparisons (Remm *et al*., 2001). It provides confidence scores for both the seed orthologs and the inparalogs. The method was evaluated against a manually curated set of worm-human orthologous transmembrane proteins. The latest release of InParanoid contained 25 eukaryotic proteomes plus *E. coli* (O'Brien *et al*., 2005).

In this paper we employ a new clustering technique to keep the advantages of InParanoid while extending the method to include multiple species. The new method called MultiParanoid reads the output from InParanoid and builds multi-species clusters from these. To benchmark the method on three-species ortholog groups, we extended the manually curated reference dataset by also including fly orthologs.

We then used this curated dataset as a reference in order to estimate the quality and features of MultiParanoid. We also compared the results to KOGs and OrthoMCL, and carried out a detailed analysis of the differences. Each discrepancy was categorized to gain insights into the particular characteristics of each method. We also review the HomoloGene database (Wheeler *et al*., 2006) that was not directly comparable to MultiParanoid clusters.

2 METHODS

2.1 Algorithm

MultiParanoid takes pairwise ortholog clusters (from e.g. InParanoid) and merges them into multi-species clusters. While there is no formal limit on the number of proteomes that can be processed, the following description is given for the case of three species. The input to MultiParanoid for N species consists of $N * (N - 1) / 2$ tables of InParanoid output—one for each pair of species.

Given a list of species A, B, and C, and pairwise ortholog cluster tables A-B, B-C, and A-C, the procedure starts by reading the list of clusters from the A-B table. These are kept as seed clusters that may be extended to include sequences in the other proteomes. The program next looks for the presence of the seed orthologs from the A-B cluster in the A-C and B-C tables. If present, all the members (inparalogs) in corresponding A-C or B-C clusters are added to the seed cluster. This procedure is repeated until all pairwise ortholog groups are processed.

This clustering corresponds to a single-linkage approach. We also implemented additional cluster trimming features in order to exclude outliers. For instance every member was required to have the confidence value—an average of its InParanoid scores—above a cutoff. However, since InParanoid clusters are already strict, trimming the multi-species clusters did not improve the overall quality.

On rare occasions, a gene may be assigned to multiple MultiParanoid clusters. To address this problem, we applied an additional procedure to assure non-redundant presence of the analyzed genes in the clusters. If a gene is not a seed ortholog in any of the clusters, it is assigned to the cluster where it has a higher InParanoid score and removed from the other. If it is assigned as the seed ortholog of a cluster, it is retained in this cluster in order to avoid disrupting the processed cluster and deleted from the other.

2.2 Construction of the reference set

Clustering of worm proteins containing at least two transmembrane segments was originally done as described elsewhere (Remm and Sonnhammer, 2000). To retrieve homologous fly and human sequences, SWISS-PROT, TREMBL and VTS databases were searched using specifically designed HMMs. After a manual curation, the original dataset contained 221 group of proteins based on sequence similarities. The largest observed family consists mainly of G-protein coupled receptors.

Putative worm—fly—human orthologs were extracted via complete phylogenetic analysis as follows:

(i) Multiple sequence alignments were done with the HMMALIGN algorithm from the HMMER package (http://hmmer.wustl.edu). Sequences having gaps (>50%) were removed from alignments.

(ii) Phylogenetic trees were constructed implementing ClustalW with observed distance and Kimura correction (Thompson *et al*., 1994). Bootstrap values were used to estimate reliability of a given branching order. A total of 100 bootstrap tests were run on trees. Only bootstrap values >60% were considered to be significant.

2.3 Comparison of ortholog clusters

Ortholog clusters generated by OrthoMCL, KOG, or manually made, were compared to the output of MultiParanoid in both directions. The comparison program took each cluster (''query'') from the first set and searched for its genes in the second group of clusters. If the genes were found in more than one cluster, the query cluster was labeled as ''Split''. Query clusters with no counterparts were labeled as ''Not found''. Otherwise (exactly one cluster found), its congruity to the query was tested. A result for the query cluster was classified into a number of categories (Supplementary Table 1). For each gene clustered by only first of the compared methods, a series of possible reasons were checked (Figure 3).

3 RESULTS

3.1 The MultiParanoid algorithm

The MultiParanoid algorithm is in its default form a simple chaining together of overlapping pairwise ortholog groups. It thus depends heavily on the quality of these groups—errors here will be propagated to the multi-species clusters. We therefore used InParanoid with default parameters, which are relatively strict, to generate the pairwise groups. As MultiParanoid provides confidence scores for the cluster members, calculated as mean InParanoid scores from the pairwise clusters, we explored ways to tighten the multi-species clusters by excluding orthologs of lower confidence. However, we found that this mainly increased the false negative rate (data not shown).

It is important to keep in mind that MultiParanoid was designed to only handle multiple proteomes that all diverged at roughly the same time point. If species of unequal relatedness are clustered, e.g. yeast, human, and chimpanzee, an implicit problem is created.

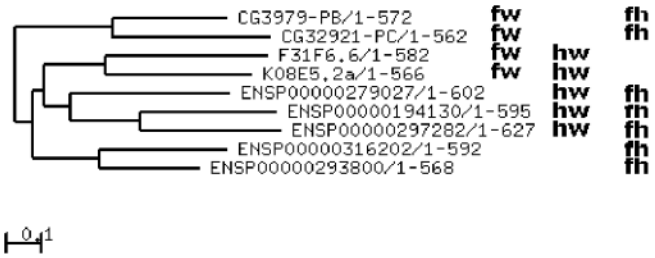

Fig. 1. Illustration of a ''tree conflict'' that may occur when merging multiple InParanoid clusters into one MultiParanoid cluster. All the sequences of the tree belong to a single MultiParanoid cluster 3575 (version 1.00), including five human proteins (ENSP*), two fly proteins (CG*), and two worm proteins. At the InParanoid 2-species level however, only ENSP00000279027, ENSP00000194130, ENSP00000297282 were recognized as human orthologs of the worm genes, yet all five were orthologous to the fly genes (InParanoid cluster members are indicated by the labels fh: fly-human; fw: fly-worm; hw: human-worm).

There is no ancestral node in the species tree of these organisms that represents the last common ancestor for all the species pairs. The resulting clusters will therefore often contain human-chimpanzee outparalogs, which were included because they are bona fide inparalogs relative to yeast.

This constitutes a major principle difference between MultiParanoid and KOGs/OrthoMCL. Both these databases combine species at very different distances, which makes the clusters less strict ortholog groups. Another difference is that only MultiParanoid gives the user confidence values. OrthoMCL is in many ways similar to InParanoid in its treatment of seed and inparalogs, but the algorithm based on Markov clustering is very different. It uses normalized E-values rather than bit scores, and the clustering is done in one step for all proteomes. A drawback is that the ''inflation parameter'' that governs the tightness of the clusters needs to be set in an ad hoc fashion.

3.2 Manual construction of the reference set

When the InParanoid algorithm was originally developed, a manually curated dataset of human-worm ortholog groups was used as a trusted standard to evaluate the accuracy of the predicted groups (Remm *et al*., 2001). Here we have extended the original dataset of human-worm orthologs by including fly orthologs to create a suitable 3-species reference set to test the accuracy of the MultiParanoid algorithm. The original dataset contained 221 groups, and most of these (202) could be extended with fly orthologs. However, in 19 cases, no fly ortholog was found, and in some cases the original group had to be redefined in the light of the fly ortholog. In total, the new reference set contains 221 groups (141 human-worm-fly, 19 human-worm, 28 human-fly, and 33 worm-fly). It is built from 697 human, 307 fly, and 361 worm proteins. This manually curated dataset is available at http://multiparanoid.cgb.ki.se/stats.html and can be used as a reference set by other developers of algorithms for detecting ortholog groups.

3.3 Benchmarking MultiParanoid

We executed MultiParanoid on the same versions of the human, fly, and worm proteomes that were used to create the reference set. To characterize MultiParanoid's ability to reconstruct the manual clusters, we extracted the intersecting and non-overlapping sets between the two clusterings, as shown in supplementary Table 1A. Both clusterings had roughly the same number of clusters: 221 in the reference set and 214 by MultiParanoid. Of these, 132 were identical. Another 45 clusters were almost identical in the sense that one was a subset of the other. This leaves about 40 clusters that clearly differed. Inspection of these cases revealed that the prevalent reason for the disagreement is the different sequence distances obtained by pairwise alignments used in InParanoid and those obtained by multiple alignments used for the manual phylogenetic analyses. Moreover, a manual curator's perception of what constitutes a ''too short'' or ''too weak'' match may differ from the strict InParanoid cutoffs.

3.4 Comparison to other methods

MultiParanoid was compared to two alternative methods: KOGs (Tatusov *et al*., 2003) and OrthoMCL (Li *et al*., 2003). To ensure a direct comparison, we ran MultiParanoid on the data used in the KOG and OrthoMCL publications. Both KOGs and OrthoMCL original clusters contained sequences of additional species, but to simplify the comparison, only sequences from human, worm, and fly were considered.

A detailed analysis was performed between MultiParanoid and the two other databases. Corresponding clusters were identified and their content was compared (see Methods). When the clusters differed, we categorized the differences into the following types: split, subset, mismatch (partial overlap), and absence. The number of clusters and genes in these categories are listed in Supplementary Table 1.

The genes that were clustered by only one of the methods were further analyzed to establish a plausible cause of discrepancy. A visual inspection of selected clusters pointed to a number of typical reasons for the observed differences. We decided to use these main categories: tree conflict, too short match, too weak match, outparalog, and other (reason not established). The classification was done in this priority order.

Tree conflict describes the case when a set of inparalogs in proteome A from the comparison A-B disagree with the inparalogs from A-C. Tree conflicts typically occur when combining species at different evolutionary distances (which thus should be avoided), or if one species has lost the original genes. A tree conflict is illustrated in Figure 1: the human-worm InParanoid clustering produced three human inparalogs while human-fly produced five. This can sometimes happen although human/fly/worm descend from roughly the same last common ancestor (of the Bilateria clade); here it was caused by a rather arbitrary clustering of the human/worm genes when the BLAST scores of the alternatives were very close. Tree conflicts are relatively common, and only result in a warning. The total number of clusters generated by MultiParanoid, run on updated human, fly, worm proteomes, that were affected by the tree conflict was 1026 of 6348 (16.1%).

Genes were classified as outparalogs when (1) a paralog (from the same species) exists in the cluster, and it is found in the corresponding cluster of the other method, and (2) a gene from another species is found closer to the second paralog than the paralogs are to each other.

The most striking difference when comparing MultiParanoid to KOGs for human/fly/worm (Supplementary Table 1A) is that although KOGs contain fewer clusters (4543 compared to 5755

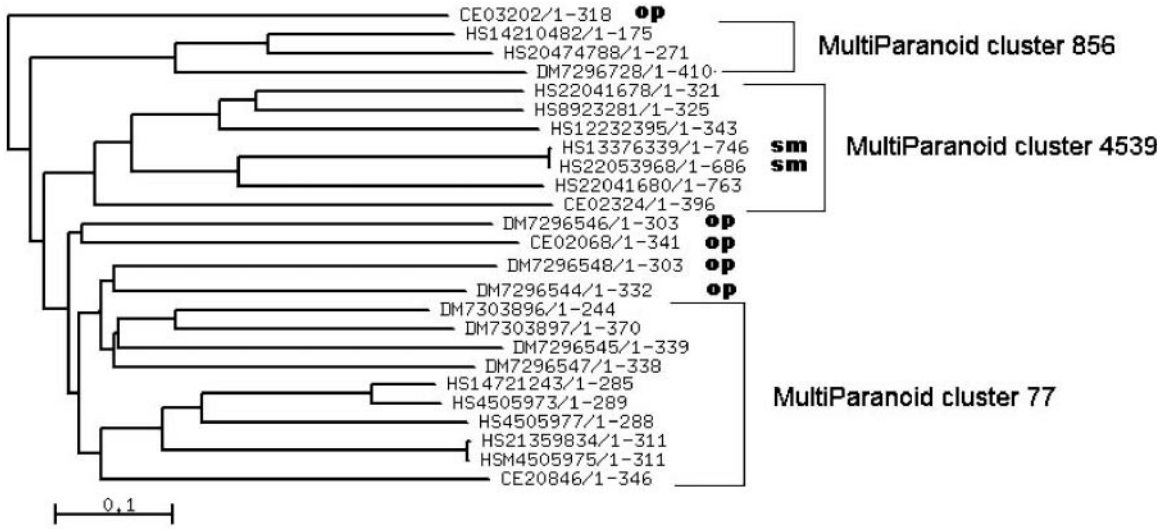

Fig. 2. Example of differences between KOG and MultiParanoid. The sequences in the tree are all the members of KOG cluster 3030. Three subtrees were identified as independent ortholog groups by MultiParanoid. HS*: human proteins, DM*: fly proteins, CE*: worm proteins. Labels: sm: sequence with a ''short match'' to the tree neighbours and therefore not clustered by MultiParanoid; op: outparalog. Note that the op-labeled fly sequences DM7296548 and DM7296544 look like inparalogs in this tree built from a multiple alignment, yet they fell just outside the cluster in InParanoid. This illustrates the clustering differences that may result from different ways of producing the sequence distance matrix.

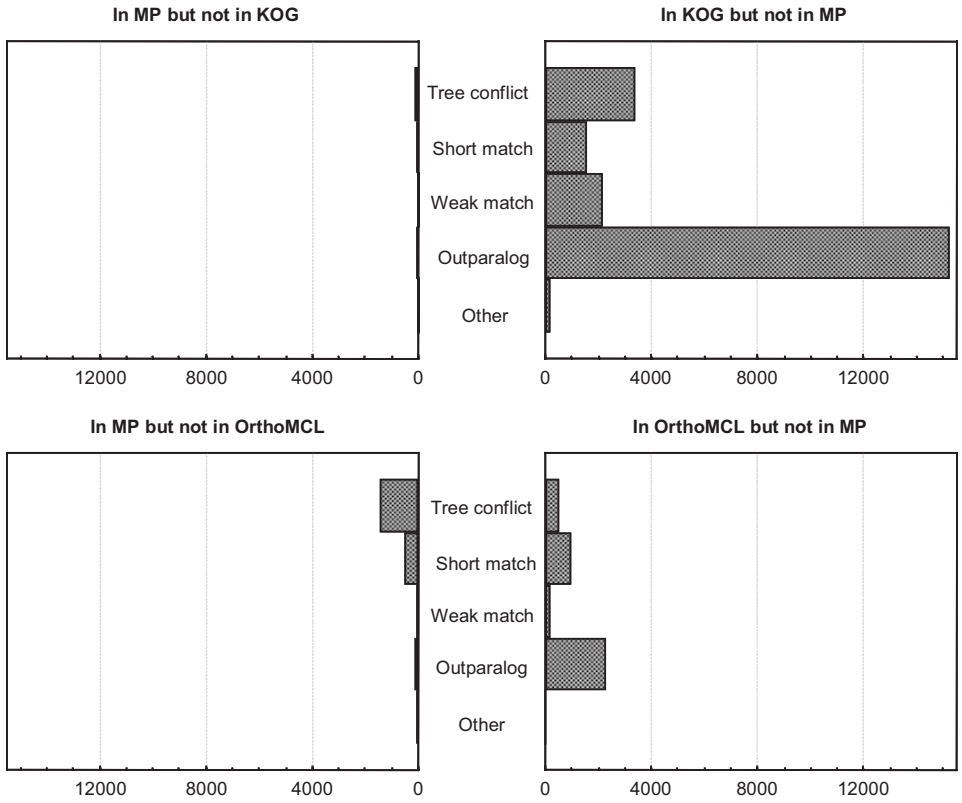

Fig. 3. Comparison of MultiParanoid to KOG and OrthoMCL. A detailed analysis was made of features and possible reasons for observed differences of corresponding ortholog clusters. Sequences clustered in one database but missing from the corresponding cluster in the other database were classified into the following categories* *Tree conflict*: conflict when merging pairwise ortholog groups in MultiParanoid (MP); *Short match:* matches the other cluster proteins with less than 50% of the length; *Weak match:* matches the other cluster proteins below the BLAST cutoff (50 bits); *Outparalog:* the protein is part of another subtree that includes the last common ancestor; *Other:* none of the reasons indicated above. *In case of multiple features per protein, only one is counted in the priority of the list above.

for MultiParanoid), they contain many more sequences (37737 compared to 23122). The average cluster size is thus twice as large in KOGs (8.3 versus 4.0). Only 1451 clusters were identical between KOG and MultiParanoid, while 2094 KOG clusters were supersets of the corresponding MultiParanoid clusters. An example of a typical situation is shown in Figure 2, in which one KOGs cluster contains three separate MultiParanoid clusters. In most of these cases it is clear that the MultiParanoid clusters represent more realistic ortholog groups in which all members derive from a single gene in the last common ancestor (of *Bilateria*). Similar cases have thus been classified as outparalogs in Figure 3. Indeed, of the 22590 KOG genes not found in the corresponding MultiParanoid cluster, 67.6% (15271) were classified outparalogs, which should be seen as an error in KOGs. The second largest reason was tree conflict (15.1%, 3411) followed by weak and short matches (9.6% and 6.9%). The latter two discrepancies may be explained by the fact that InParanoid does not accept matches below 50 bits and 50% of the length. Note that although many of the genes categorized as tree conflict probably also represent outparalogs, we chose to not classify them as such because the tree conflict casts some doubt about the whole cluster. In other words, our figures underestimate the number of outparalogs in KOGs.

The high outparalog rate in KOGs is partly due to the fact that most clusters were built with regard to a higher last common ancestor, e.g. the one of eukaryotes, and contain species beyond the animal clade. Indeed, only 1147 KOG clusters (of 4852) were animal-specific. But even when looking at 50 randomly selected pure human/fly/worm KOG clusters, 32 contained outparalogs by visual inspection of the gene trees. Many of the two-species KOG clusters (TWOGs) with >2 genes also contained outparalogs. Thus, KOGs appears to generally favor inclusion of outparalogs.

The OrthoMCL clusters were in much better agreement with the MultiParanoid results—both produced roughly 6000 clusters containing about 26000 genes. About 4000 of the clusters were identical, suggesting that these ortholog groups are very trustworthy. In the roughly 2000 clusters with differences, a couple of trends stood out. Outparalog inclusion was about 15 times more common among the OrthoMCL-unique genes (2267) compared to MultiParanoid-unique ones (145). The fact that tree conflicts are three times more common in clusters with MultiParanoid-unique genes than OrthoMCL-unique ones (1453 versus 518) suggests that OrthoMCL builds slightly tighter clusters than MultiParanoid. Genes missing due to short or weak matches were about twice as common in OrthoMCL, indicating that MultiParanoid is stricter in these respects.

The main difference between MultiParanoid and OrthoMCL thus seems to be OrthoMCL's tendency to include outparalogs. This can be explained by the fact that the original OrthoMCL clusters included 10 proteomes, some of which have very different last shared ancestors. For instance, human, mouse, and *E. coli* are included at the same time. Combining proteome pairs with such different relationships inevitably leads to inclusion of outparalogs: the eukaryotic genes underwent multiple common duplications since the divergence from *E. coli*. This problem has been worsened in the latest version of OrthoMCL, which includes 55 proteomes (Chen *et al*., 2006). For example, taking the top 10 (by E-value) OrthoMCL clusters that contained >8 genes from human, *Ciona*, *D. melanogaster*, and *C. elegans*, 8 clusters (111, 1057, 489, 88, 1300, 335, 1428, 123) contained outparalogs in at least 1 species (usually in 2-4). In the previous 10-species OrthoMCL version, only cluster 1057 (rather its prototype, as the numbering was changed) had outparalogs. The corresponding MultiParanoid 4-species clusters had no outparalogs.

Another database of eukaryotic orthologs is HomoloGene (Wheeler *et al*. 2006), which in addition to sequence similarity also uses synteny and DNA substitution rates to build ortholog groups (http://www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html). This database is however very different in nature from MultiParanoid, OrthoMCL and KOGs. HomoloGene is extreme in the opposite way that KOGs is—it splits up ortholog groups into smaller groups, putting inparalogs into different clusters.

For example, only 29 of 3814 HomoloGene clusters that could include both human and yeast genes (labeled ''*Eukaryota*'' or ''*Fungi*/*Metazoa*'') contained more than a single human gene. As a comparison, InParanoid had 2138 human-yeast clusters, and 816 of them contained more than one human orthologs.

Genes that are considered inparalogs by InParanoid are normally not missing from HomoloGene but are found in other clusters, usually with different labeling of the last common ancestor (e.g. human inparalogs could be in clusters labeled ''*Eukaryota*'', ''*Coelomata*'', ''*Amniota*'') or with the same label but another species content.

For instance, the biggest MultiParanoid cluster built from human, *Ciona*, *D. melanogaster*, and *C. elegans* proteins contained more than 400 human genes (zinc finger proteins with Pfam domain zf-C2H2, PF00096), but only a few from other species. The human part thus constituted a vertebrate-specific expansion according to MultiParanoid. Yet, in HomoloGene most of the human genes were split into 6 different clusters labeled higher than vertebrates (''*Coelomata*'' and ''*Fungi*/*Metazoa*''). These clusters contained a set of human genes plus an insect or worm gene (all from the same MultiParanoid cluster), even though the human genes are closer to each other than to any gene outside the vertebrate clade. Some human genes from the MultiParanoid cluster were placed in pure vertebrate clusters (''*Amniota*'', ''*Eutheria*'', ''*Euarchontoglires*'', and human-specific expansions).

HomoloGene thus tends spread inparalogs over isolated small clusters. This property makes the clusters very tight, practically inparalog-free, and misleading in defining complete ortholog sets.

4 DISCUSSION

Functional genomics has driven a demand for fast and efficient orthology analysis tools. The algorithm presented here enables an automated orthology analysis to be performed on multiple proteomes, and is therefore a welcome extension of the previously published InParanoid (Remm *et al*., 2001) algorithm. We found a satisfying high degree of congruence between the results generated by MultiParanoid and the manually curated dataset used as a reference. The ability of the algorithm to correctly identify orthologous sequences was also evaluated by executing MultiParanoid and similar algorithms published by other groups, namely KOGs (Tatusov *et al*., 2003) and OrthoMCL (Li *et al*., 2003), on the same datasets. This showed that the quality of MultiParanoid's clusters is high, and therefore the method should make an important contribution to the bioinformatics tools currently available for orthology analyses.

Unlike KOGs where the minimal cluster consists of three genes, one per species (''triangles'', Tatusov *et al*., 1997), MultiParanoid is based on pairwise groups of orthologs. For genomes A, B, C, protein pairs {A1, B1} and {B1, C1} can be reciprocally best hits, whereas {A1, C1} may not be. Hence, clusters exist where a triangle is not secured. This leads to what we call a ''tree conflict'' when merging pairwise orthologs from three species. If the species have roughly the same last ancestor we believe that the best action in such cases is to combine all genes from the pairwise clusters. Still, only a minor fraction (∼15%) of MultiParanoid clusters had tree conflicts when clustering human, fly, and worm. In very few cases (100-200 genes) did the conflict lead to ambiguous cluster membership.

We here consider human, fly, and worm to descend from roughly the same last ancestor. Yet, two different subgroupings have been proposed: the ''*Ecdysozoa*'' (worm-fly) and ''*Coelomata*'' (fly-human) hypotheses (Blair *et al*., 2002; Dopazo and Dopazo 2005; Philip *et al*., 2005). Neither of these gets full support from molecular data. Looking at gene trees, the *Coelomata* grouping is found in about 60% of the trees, *Ecdysozoa* in 25%, and worm-human in 15%. The question is therefore probably unresolvable and we consider the three species to be roughly equally related. It is thus wiser to use molecular data to group species than to use the classical taxonomy, especially since the latter can be ambiguous or vague.

The requirement of only clustering species with shared last ancestor can be a drawback for MultiParanoid, as it only allows few eukaryotic species to be included in multi-species groups. However, a possibility is to consider several species in a clade as a 'pseudo-species', e.g. mammals or arthropods. If one treats all mammalian genes as 'pseudo-inparalogs' when compared to arthropods and nematodes, it is possible to avoid outparalogs. This is done by labeling the included outparalogs as pseudo-inparalogs, and not transferring functional information between them. We are developing a new version of MultiParanoid with multiple species in the same clade with precise labeling of what are orthologs and what are not within each cluster. Using this framework, which is similar to the HOPS database (Storm and Sonnhammer, 2003), we can build clusters that include all completely sequenced eukaryotic species. Even incomplete proteomes can be included, as long as one complete proteome is part of the clade.

5 DATA

The manually curated data set of transmembrane proteins was based on the older proteome versions:

Human: 35118 sequences from SwissProt and TREMBL.

Fly: 14100 predicted proteins sequences from FlyPep Release 1. (http://www.fruitfly.org/sequence/download.html).

Worm: 19099 predicted proteins from WormPep 20 (ftp://ftp.wormbase.org/pub/wormbase/).

These original protein sets and the manually curated clusters are available at http://multiparanoid.cgb.ki.se/download.

The KOG clusters were published as a supplementary material by Tatusov *et al*. (2003) and were downloaded at http://www.ncbi.nlm.nih.gov/COG/new/. For the purpose of this work, only human, fly and worm genes were extracted from the seven species in total. In addition, we included the 2-species clusters (TWOGS). The version numbers of the proteomes are not available, but the original complete sets of proteins of all KOG proteomes in FASTA format can be downloaded from the same location.

OrthoMCL clusters were gathered via queries to the Web service http://www.cbil.upenn.edu/gene-family/. The following datasets were downloaded: {human, fly, worm}, {human, fly}, {human, worm}, {fly, worm}. The respective versions of complete protein sets were obtained from the original sites listed in their article (Li *et al*., 2003).

Alternative splice forms of the same gene may sometimes end up in different clusters. We therefore only used the longest spliced form of each gene.

MultiParanoid scripts, FASTA sequence and data files are available from the web site http://multiparanoid.cgb.ki.se/. The final clusters generated by MultiParanoid can be downloaded as a single text file. The web-based version 1.00 of MultiParanoid with search by gene/protein ID and cross-links to protein/domain databases currently includes four genomes—*C. elegans*, *D. melanogaster*, *C. intestinalis*, and *H. sapiens*—and will be expanded.

ACKNOWLEDGEMENTS

This work was supported by a grant from Pfizer Corporation.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389-3402.

Blair,J.E., Ikeo,K., Gojobori,T. and Hedges,S.B. (2002) The evolutionary position of nematodes. *BMC Evol. Biol.* **2,** 7.

Bono,H., Goto,S., Fujibuchi,W., Ogata,H. and Kanehisa,M. (1998) Systematic prediction of orthologous units of genes in the complete genomes. *Genome Inform Ser Workshop Genome Inform.* **9**, 32–40.

Chen,F., Mackey,A.J., Stoeckert,C.J.Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* Jan 1; **34,** (Database issue):D363–8.

Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T., Weng,S., Cherry,J.M. and Botstein,D. (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, Dec 11;**282**(5396), 2022–8.

Dopazo,H and Dopazo,J. (2005) Genome-scale evidence of the nematode-arthropod clade. *Genome Biol.*; **6**(5), R41. Epub 2005 Apr 28.

Enmark,E., Gustafsson,J.A. (2001) Comparing nuclear receptors in worms, flies and humans. *Trends Pharmacol Sci*. **22**(12), 611–5. Review.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool*. **19**, 99–113.

Fitch,W.M. (2000) Homology: A personal view on some of the problems. *Trends Genet.* **16**, 227–231.

Kondrashov,F.A., Rogozin,I.B., Wolf,Y.I., Koonin,E.V. (2002) Selection in the evolution of gene duplications. *Genome Biol.*, **3**(2), RESEARCH0008. Epub 2002 Jan 14.

Lee,Y., Sultana,R., Pertea,G., Cho,J., Karamycheva,S., Tsai,J., Parvizi,B., Cheung,F., Antonescu,V., White,J., Holt,I., Liang,F. and Quackenbush,J. (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* **12**(3), 493–502.

Lespinet,O., Wolf,Y.I., Koonin,E.V. and Aravind,L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* **12**(7), 1048-59.

Li,L., Stoeckert,C.J., Roos,D.S. (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes *Genome Res.* **13**(9), 2178–89.

Maglich,J.M., Sluder,A., Guan,X., Shi,Y., McKee,D.D., Carrick,K., Kamdar,K., Willson,T.M. and Moore,I.T. (2001) Comparison of complete nuclear receptor sets from the human, *Caenorhabditis elegans* and *Drosophila* genomes. *Genome Biol*. **2**(8), RESEARCH0029. Epub 2001 Jul 24.

Mushegian,A.R., Garey,J.R., Martin,J. and Liu,L.X. (1998) Large-scale taxonomic profiling of eukaryotic model organisms: A comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res.* **8**, 590-598.

Philip,G.K., Creevey,C.J. and McInerney,J.O. (2005) The *Opisthokonta* and the *Ecdysozoa* may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the *Coelomata* than *Ecdysozoa*. *Mol. Biol. Evol.* **22**(5), 1175-84. Epub 2005 Feb 9.

Remm,M., Storm,C.E. and Sonnhammer,E.L.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052.

Rubin,G.M., Yandell,M.D., Wortman,J.R., Gabor Miklos,G.L., Nelson,C.R., Hariharan,I.K., Fortini,M.E., Li,P.W., Apweiler,R., Fleischmann,W., Cherry,J.M., Henikoff,S., Skupski,M.P., Misra,S., Ashburner,M., Birney,E., Boguski,M.S., Brody,T., Brokstein,P., Celniker,S.E., Chervitz,S.A., Coates,D., Cravchik,A., Gabrielian,A., Galle,R.F., Gelbart,W.M., George,R.A., Goldstein,L.S., Gong,F., Guan,P., Harris,N.L., Hay,B.A., Hoskins,R.A., Li,J., Li,Z., Hynes,R.O., Jones,S.J., Kuehl,P.M., Lemaitre,B., Littleton,J.T., Morrison,D.K., Mungall,C., O'Farrell,P.H., Pickeral,O.K., Shue,C., Vosshall,L.B., Zhang,J., Zhao,Q. and Zheng,X.H. (2000) Comparative genomics of the eukaryotes. *Science*, Mar. 24 **287**(5461), 2204–15.

Sonnhammer,E.L.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* **18**, 619-620.

Storm,C.E. and Sonnhammer,E.L.L. (2003) Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res.* **13**(10), 2353-62.

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

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., Rao,B.S., Smirnov,S., Sverdlov,A.V., Vasudevan,S., Wolf,Y.I., Yin,J.J. and Natale,D.A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. Sep 11; **4**(1), 41.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* Nov 11; **22**(22), 4673–80.

Troemel,E.R., Chou,J.H., Dwyer,N.D., Colbert,H.A. and Bargmann,C.I. (1995) Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans*. *Cell*, Oct 20; **83**(2), 207–18.

Wheelan,S.J., Boguski,M.S., Duret,L., Makalowski,W. (1999) Human and nematode orthologs – lessons from the analysis of 1800 human genes and the proteome of *Caenorhabditis elegans*. *Gene*, Sep 30; **238**(1), 163–70.

Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Geer,L.Y., Helmberg,W., Kapustin,Y., Kenton,D.L., Khovayko,O., Lipman,D.J., Madden,T.L., Maglott,D.R., Ostell,J., Pruitt,K.D., Schuler,G.D., Schriml,L.M., Sequeira,E., Sherry,S.T., Sirotkin,K., Souvorov,A., Starchenko,G., Suzek,T.O., Tatusov,R., Tatusova,T.A., Wagner,L. and Yaschenko,E. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, Jan 1; **34** (Database issue):D173–80.

Xie T. and Ding D. (2000) Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale. *Gene,* **261**, 305–310.