BIOINFORMATICS

Vol. 18 no. 1 2002
Pages 92–99

Automated ortholog inference from phylogenetic trees and calculation of orthology reliability

Christian E. V. Storm and Erik L. L. Sonnhammer

Center for Genomics and Bioinformatics, Karolinska Institutet, S-171 77 Stockholm, Sweden

Received on February 2, 2001; revised on July 17, 2001; accepted on August 15, 2001

ABSTRACT

Motivation: Orthologous proteins in different species are likely to have similar biochemical function and biological role. When annotating a newly sequenced genome by sequence homology, the most precise and reliable functional information can thus be derived from orthologs in other species. A standard method of finding orthologs is to compare the sequence tree with the species tree. However, since the topology of phylogenetic tree is not always reliable one might get incorrect assignments.

Results: Here we present a novel method that resolves this problem by analyzing a set of bootstrap trees instead of the optimal tree. The frequency of orthology assignments in the bootstrap trees can be interpreted as a support value for the possible orthology of the sequences. Our method is efficient enough to analyze data in the scale of whole genomes. It is implemented in Java and calculates orthology support levels for all pairwise combinations of homologous sequences of two species. The method was tested on simulated datasets and on real data of homologous proteins.

Availability: Downloadable free of charge from ftp://ftp.cgb.ki.se/pub/prog/orthostrapper/ or on request from the authors.

Contact: christian.storm@cgb.ki.se

INTRODUCTION

Orthologs are proteins in different species that go back to a single protein in the last common ancestor of these species. This definition was given by Fitch (1970) and it implies that because of their phylogenetically close relationships, orthologous proteins are likely to have identical or very similar functions, although function is not part of the definition. With the vast amount of sequence data produced by the genome projects, automated methods for assigning orthology are needed.

Orthology is not necessarily a one-to-one relationship. In case of one or multiple duplication events subsequent to the speciation event, it can be a one-to-many or a many-to-many relationship (see Figure 1a). When assigning orthology to proteins of more than two species, the relationships can become complex. This is due to the fact that orthology is non-transitive (Fitch, 2000), meaning that two proteins in different species that are both orthologous to a protein in a third species are not necessarily orthologous to each other (see Figure 1b). Orthologous relationships may be further complicated because of horizontal gene transfer or gene conversion (Gogarten and Olendzenski, 1999).

Standard methods for finding orthologs are based on he analysis of phylogenetic trees. Reconciled trees (Goodman *et al.*, 1979; Page, 1994) are an effective method to find orthologous relationships (Yuan *et al.*, 1998): a new tree is constructed that reconciles the sequence tree with the species tree by postulating missing sequences, e.g. due to gene loss. Orthologous and paralogous relationships can directly be read from the reconciled tree.

A general problem is that the tree analyzed may not reflect all phylogenetic relationships correctly, since all tree calculation is based on somewhat arbitrary parameters. To get a confidence estimate for a given tree topology, the bootstrap method was applied to phylogenetic trees (Felsenstein, 1985). This technique is frequently used to estimate the confidence level for a given phylogenetic hypothesis, but has been criticized for a systematically bias towards lower values (Zharkikh and Li, 1995; Newton, 1996). However, Efron *et al.* (1996) showed that the bootstrap method can be seen as a first order approximation of the accuracy of the tree's topology. Different methods have been proposed to get a more precise estimate, for instance the complete-and partial-bootstrap (Zharkikh and Li, 1995) or the second level bootstrap (Efron *et al.*, 1996).

We developed and implemented an approach to calculate orthology support levels. This method browses a sequence tree for orthologous relationships between two species. By applying it to a large number of bootstrap trees it is possible to assign a support value to all orthologous pairings in the tree. The bootstrap trees are calculated from pseudosamples that are created by sampling with replacement from the original alignment.

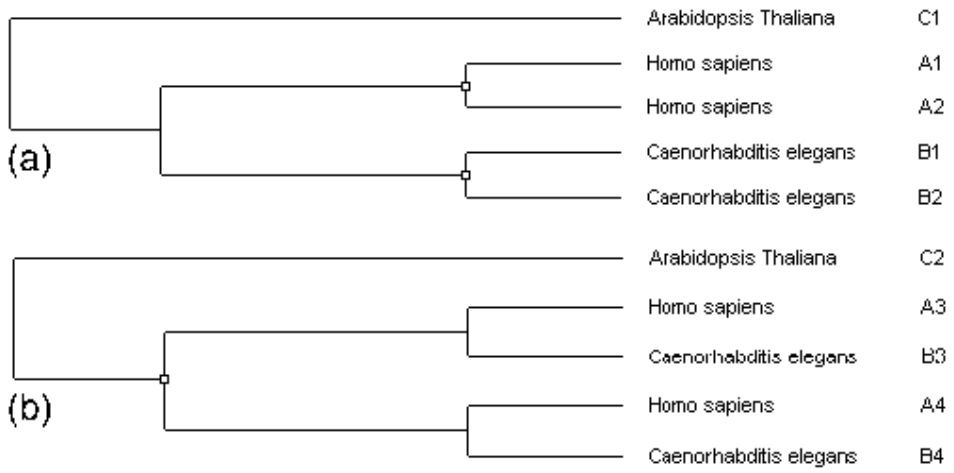


Fig. 1. Hypothetical examples for illustrating different levels of orthology assignments that involve paralogs. Duplication events are marked with a square at the node. (a) Many-to-many orthologous relationships. The group A1/A2 in *H.sapiens* is orthologous to the group B1/B2 in *C.elegans*. They go back to a single sequence in the last common ancestor. The paralogs A1 and A2 were separated by a duplication, and so were B1/B2. The two duplication events occurred independently after the speciation. (b) Orthology between more than two species. In this tree, the duplication event that led to A3/A4 and B3/B4 occurred before the speciation. Therefore A3 is orthologous only to B3 but not to B4, and A4 only to B4 but not to B3. However, the four sequences A3, A4, B3 and B4 are orthologous to C2. But although A3 and B4 are orthologs of C2, they are not orthologous to each other! That means orthology is not transitive. Note that if A3 and B4 had been lost or were not sequenced yet, methods based on relative sequence distance would assign orthology incorrectly to B3/A4. The trees were constructed with GeneTree.

ALGORITHM

Assigning orthology from a phylogenetic tree

The algorithm that detects orthologous relationships in a tree was designed to assign orthology between two (groups of) species. It is possible to combine clades, e.g. all mammals, into one group, and analyze them for orthologs in the group of all yeast species. Applying the algorithm to a series of species pairs and combining the results can detect orthologous relationships between multiple species.

All sequences of the tree must be classified into one of four different groups. The sequences of the two (groups of) species that one wants to find orthologous relationships between must be assigned to the two primary groups. Remaining sequences that come from species distant to these groups should be assigned to the outgroup. Other remaining sequences should either be added to one of the primary groups if they are closely related to only one of them, or to the 'blank' group of sequences that are ignored in the analysis. For instance, when analyzing worm and human orthologs, it is a good idea to ignore all fly sequences, since it is not clear which species they are closer to Mushegian *et al.* (1998). Sequences that are not assigned to any group are by default added to the blank group. Using this information, orthologous assignments are made as follows:

(1) for each sequence m of species-group 1 do;

(2) start at the leaf that is the current sequence;

(3) go up one node and analyze the new branch;

(4) if the leaves of the new branch are all proteins of species-group 1 repeat step 3;

(5) if the leaves of the new branch are all proteins of species-group 2 report orthology between m and all sequences on this branch and go to step 1;

(6) if the new branch contains at least one sequence of the outgroup OR at least one sequence of species group 1 AND at least one sequence of species-group 2 go to step 1.

We estimate the runtime behavior of this algorithm to be:

Average CPU usage: $O(m^*\log n)$.

Worst-case CPU usage: $O(m^*n)$.

m: number of sequences in the group of species.

n: total number of sequences in the tree.

Since the execution time of this algorithm is low compared to most tree building algorithms the number of trees that can be analyzed is really limited by the tree building method used. Therefore analyzing a large number of trees ($\sim$1000) of reasonable size ($\sim$100 sequences) is quick

on a normal desktop computer using common neighbor-joining programs. This makes this approach suitable for sampling with a bootstrapping method.

Calculating ortholog bootstrap values

The bootstrap method is typically applied to assign the accuracy of a statistical estimation. When applying it to phylogenetic trees, the multiple alignment is used as a data sample. In bootstrapping, the phylogenetic tree inferred from the original multiple alignment is the null hypothesis to test against. Here the null hypothesis becomes multiple subhypotheses: the ensemble of all possible pairwise orthologous relationships in the tree. The orthologous assignments are inferred with the approach described above. To calculate the support levels for these subhypotheses, a series of pseudo samples is generated from the data sample. These pseudo samples are trees generated with the same method as in classical tree bootstrapping.

The bootstrap trees are generated from the original multiple alignment by sampling with replacement. Columns are picked randomly, and a column can be picked more than once. The same number of columns as in the original alignment is picked. From each bootstrap alignment a tree is constructed that is immediately analyzed for orthologs. If an orthology assignment is found, the corresponding subhypothesis is given a score of one. The scores for each subhypothesis are added up for all trees calculated from the bootstrap alignments. This way one gets a total bootstrap support value for each subhypothesis. To distinguish these from 'classical' bootstrap values of phylogenetic trees, we propose to denote them 'ortholog bootstrap', or short 'orthostrap' values. (Phylogenetic bootstrap values could also be denoted 'phylostrap' values to distinguish them from all other applications of bootstrapping.)

An advantage of this method is that all possible pairwise orthology assignments are given a score, thus allowing orthologous relationships not represented in the original tree to be assessesd.

IMPLEMENTATION

The ortholog bootstrapping is implemented as a Java 1.2 program, named Orthostrapper. For sampling the pseudo alignments and calculation of the trees it utilizes Belvu (Sonnhammer, unpublished). Belvu calculates neighbor-joining trees (Saitou and Nei, 1987; Studier and Keppler, 1988) based on uncorrected distances and assigns a root to a tree by using the 'center of tree' approach (Thompson *et al.*, 1994). Orthostrapper analyzes the bootstrap trees and prints out a matrix of the calculated ortholog bootstrap values for all possible combinations between the sequences of the two species groups.

METHODS AND DATA

Simulated data

To test the performance of any orthology finding method one would need a testset of true orthologs. Ideally this would consist of proteins experimentally determined to have the same biochemical function and biological role in different species. Unfortunately no such dataset exists. Therefore we decided to test Orthostrapper on simulated data to get a first impression of its behavior. The simulations also allow assessing the error introduced by the 'center of tree' rooting method.

The simulations were calculated with the program Rose (Stoye *et al.*, 1998), using the rates given in the PAM-matrix for amino acid substitution and the default parameters for the insertion–deletion function. Rose simulates evolution by evolving a sequence following a guidance tree. The guidance trees were randomly created, each with 5–50 sequences of length 100–300 from three different 'species'.

Two different models were used for calculating the trees:

(a) Molecular clock/balanced tree:
the evolution guidance tree is calculated so that all leaves have the same distance from the root. This distance varies slightly between trees, but is on average 40 PAM. All sequences in a tree created with this model are mutated at the same rate. With this model, 28 057 sequences grouped in 1000 families were created. The guidance trees contained 3093 orthologous sequence pairs in total.

(b) Independent evolutionary rates/unbalanced trees:
the distances between the nodes in the evolution guidance tree is chosen randomly from an interval so that the expected maximum distance of two sequences is 80 PAM. This means that the evolutionary rate of each sequence is independent. With this model, 27 263 sequences grouped in 1000 families were created. The guidance trees contained 2881 orthologous sequence pairs in this dataset.

Real data

To test Orthostrapper on real data, a set of 114 families containing worm–mammalian orthologs (Remm and Sonnhammer, 2000) was used. Orthostrapper was run on 1000 bootstrap trees per group. It turned out to be difficult to compare the Orthostrapper results directly to the orthologs assigned manually for these families, because the approaches are rather different. If a branching pattern is unclear, the manual orthology assignment typically includes all sequences of the branch, while Orthostrapper resolves the fine branching pattern by sampling. A comparison of two such different approaches did not seem meaningful.

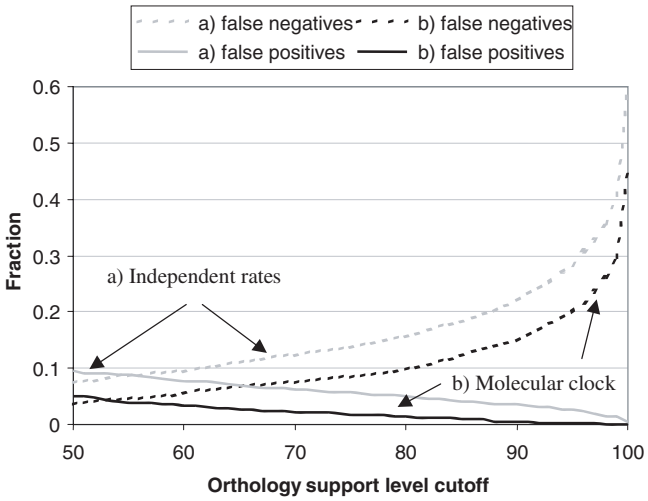


Fig. 2. The fraction of false orthologous assignments (false positives) and undetected orthologs (false negatives), dependent on the orthology support level cutoff chosen for assigning orthology. Shown are the results for the two different models used in the simulations: (a) independent evolutionary rates; (b) molecular clock. The true/false assignments were made by comparing the Orthostrapper orthologs to the true orthologs in simulated data.

Instead, we took the 114 families and generated a new set of orthologs by building neighbor-joining trees and feeding them to GeneTree (Page, 1998). GeneTree reconciles a sequence tree with a species tree, and one can easily read a list of orthologs from the reconciled tree. In our dataset, GeneTree assigned 352 worm proteins to at least one orthologous counterpart in mammalia, and 1083 mammalian proteins to have at least one orthologous counterpart in worm. These assignments gave a total of 2105 pairwise orthologous relationships.

RESULTS

Simulated data

The simulated trees, for which the true tree topology is known, can be used to assess the performance of the ortholog bootstrapping method. Although the trees are artificial, they allow us to assess the accuracy of ortholog detection by comparing the orthologs calculated from reconstructed trees to the true orthologs in the dataset. We also took advantage of the possibility to generate trees with equal or unequal rates of mutation, in order to study the robustness of the ortholog bootstrapping method to such factors. Figure 2 shows a plot of false negative and false positive assignments for all ortholog bootstrap values.

Equal rates of evolution (or 'molecular clock') produces a well-balanced tree, while unequal rates for different branches will make tree reconstruction more difficult due to problems with placing the root correctly and 'long branch attraction'. Correct rooting is particularly important for correct ortholog assignment. As shown in Figure 2, the probability for a false negative assignment is increased by ~4–10% in unequal rate trees, and for false positive assignment by ~4.5–6%, depending on the cutoff chosen for assigning orthology. Because the unequal rates were chosen to be rather extreme, we do not expect trees from real data to suffer more than a few percent inaccuracy due to incorrect rooting or long branch attraction.

How much confidence should one have in a given ortholog bootstrap value? Simulated data can be of some help here. For instance when doing a first scan of large datasets one might apply a cutoff of 0.5. Then a sequence pair would be considered orthologous if orthology is supported by more than 50% of the bootstrap trees. Figure 2 shows that by applying a cutoff of 0.5 in the simulation one would find over 95% of the true orthologous relationships, but at the cost of up to 10% false assignments. For other applications, for instance functional inferences, one might apply a cutoff of 0.95 for assigning orthology. In the simulated data the number of false positives drops close to zero at this value. But on the other hand one might miss up to 30% of the true orthologs.

Analysis of the worm–mammalian homologs

To make sure the tree-parsing algorithm was implemented correctly in Orthostrapper, we used it to analyze the optimal trees of the 114 families for orthologous relationships. This gave exactly the same orthology assignments as with GeneTree.

We then calculated ortholog bootstrap levels for all 9968 worm–mammalian sequence pairs in the complete set of 114 families. 2105 of these pairs were predicted as orthologs by tree reconciliation in the original tree. The distribution of ortholog bootstrap values for orthologous pairs found with the reconciliation in the optimal tree is shown in Figure 3 as black bars. The distribution of the 7963 ortholog bootstrap values of pairs not found by tree reconciliation are shown in Figure 3 as white bars. A table of all results is available at ftp://ftp.cgb.ki.se/pub/prog/Orthostrapper/results/orthostrapping.xls. We note that of the pairs predicted as orthologs by tree reconciliation in the original tree, 25% (523) have an ortholog bootstrap support less than 0.5, suggesting that they may well be false ortholog assignments.

An illustration of the difference between ortholog bootstrapping and phylogenetic bootstrapping is shown in Figure 4. From the reconciled tree in Figure 4b one can read that BAA91192.1/K07H8.2 is the only ortholog pair present in the tree shown in Figure 4a. Notable is the difference of the ortholog bootstrap value from the phylogenetic bootstrap value for this orthologous pairing. This can be explained by the different possible tree

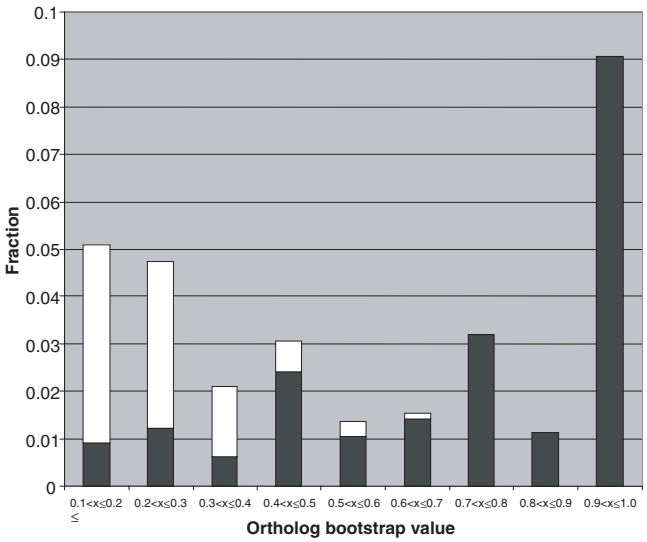


Fig. 3. Distribution of the ortholog bootstrap values for intervals $a < x \leqslant b$. Shown are the results from the analysis of all possible 9968 worm–mammalian pairings of proteins in the dataset that have a value higher 10%. The height of the bars reflect the fraction of sequence pairs in the given ortholog bootstrap value interval. The black parts of the bars show the fraction of pairs in the analyzed dataset that were assigned orthologs by tree reconciliation of the optimal tree. The white part stands for the fraction of worm–mammalian sequence pairs that was not reported orthologous by reconciled trees. The ortholog bootstrap values were calculated with 1000 pseudosamples. 69% of the possible pairs have a value below 10% (not shown). Only four of these were assigned orthologs by tree reconciliation of the optimal trees.

topologies. The node one level up has a bootstrap support of 0.999, meaning that the three sequences group together in 99.9% of all bootstrap trees. The grouping can happen with three possible different topologies:

(1) As observed in the optimal tree.

(2) ZK185.2 together with BAA91192.1 on a branch.

(3) ZK185.2 with K07H8.2 on a branch—then both are reported as orthologous to BAA91192.1 by the tree parser.

The fact that the ortholog bootstrap value for BAA91192.1 and K07H8.2 is ∼10% higher than the phylogenetic bootstrap value indicates that topology 3 occurs in ∼10% of the bootstrap trees. Hence even in this simple example of a one-to-one relationship the ortholog bootstrap value gives a more realistic picture for confidence of orthologous relationships than the phylogenetic bootstrap value.

Figure 5 shows an example where the orthologous relationships are unclear. Only looking at the optimal tree would lead to the result, that the three worm proteins are orthologous to the two human ones. But the phylogenetic bootstrap values shown in the tree already indicate that this branching pattern only has a low support. The ortholog bootstrap values show a high support orthology for the pairing O15431/Y58A7A.1, whereas all other possible pairings in this branch have significantly lower values. This indicates that O15431/Y58A7A.1 are the only orthologs on this branch.

Of the pairings that were not assigned as orthologs in the original tree, 41 have an ortholog bootstrap support higher than 0.5. 14 of these cases come from two trees with orthology-assignments that are close to the root-node, making them vulnerable to slight changes in the position of the root. The other 27 cases indicate orthologous relationships that are not present in the optimal tree. An example is shown in Figure 6, in which four additional potential orthologous relationships are found.

DISCUSSION

We have presented a new algorithm for finding orthologs in combination with the bootstrap method. The main advantage of this approach is that ortholog bootstrap values are assigned to all possible orthologous pairings. This makes it possible to resolve complicated many-to-many orthologous relationships. Orthology assignments in the optimal tree that might be incorrect can be identified by their low ortholog bootstrap value.

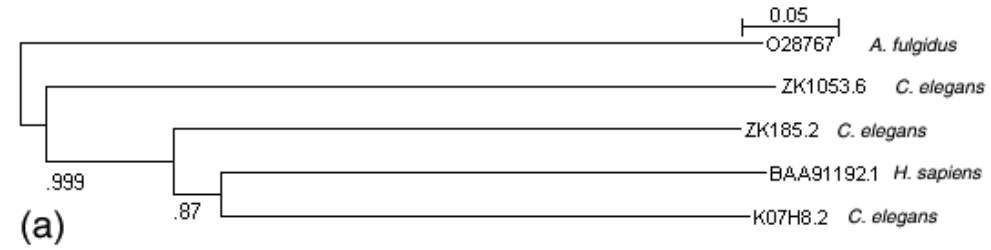
Our implementation is efficient enough to allow ortholog bootstrapping on large amounts of data. The complete analysis of the 114 groups/2624 sequences with 1000 pseudosamples took less than 10 h on a 500 MHz UNIX workstation.

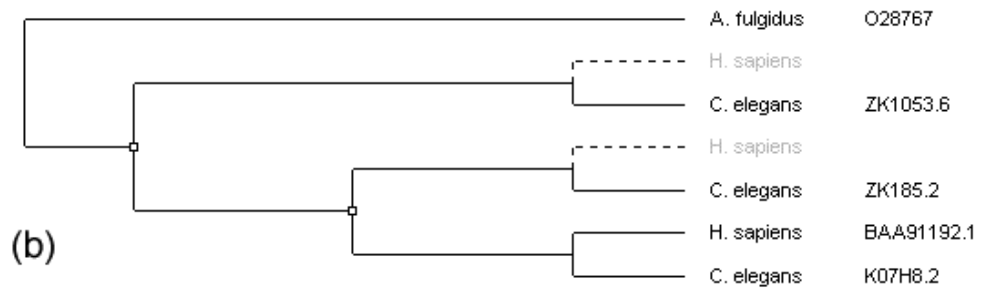
The results of the analysis show that the tree-parsing algorithm, when used to parse only the optimal trees, assigns the same orthologous relationships as one finds by reconciling trees. This indicates that the algorithm works correctly when looking for orthologs.

In case of one-to-one orthologous relationships the calculated ortholog bootstrap value gives a more realistic view of the possible orthology of two sequences than one would get using the phylogenetic bootstrap value as is done frequently. This is because the ortholog bootstrap value reflects all possible branching that support orthology for the given pair.

For potential one-to-many or many-to-many orthologous relationships Orthostrapper makes it possible to assign support values to pairs within these multiple relationships. Additionally, as demonstrated in Figure 5, it is possible to make a statement about orthologous relationships that can not be resolved by phylogenetic bootstrap values in combination with tree reconciliation.

The results of the simulation show that the calculated ortholog bootstrap levels have a non-linear relation to confidence intervals. For instance a 95% confidence





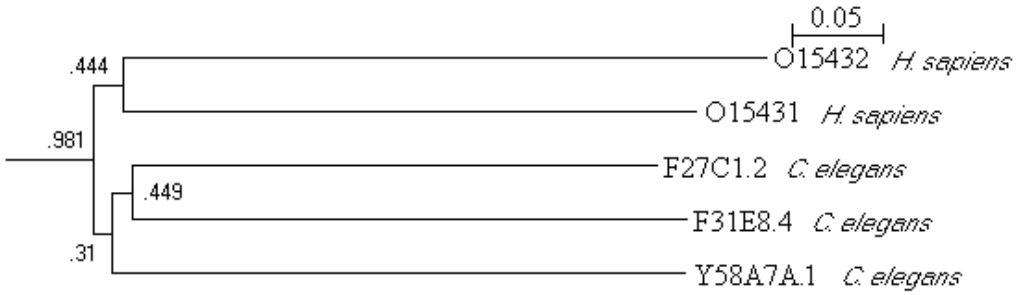
(c) Ortholog bootstrap confidence values

	K07H8.2	ZK185.2	ZK1053.6
BAA91192	.953	.127	0

Fig. 4. Ortholog bootstrapping provides information on all possible pairwise orthology assignments. (a) Tree calculated with neighbor-joining, using uncorrected distance and 1000 bootstrap samples, root is set at the center of the tree. From this tree one would assign BAA91192.1 to be orthologous to K07H8.2 and might use the bootstrap value at the connecting node of 0.87 as a confidence value. (b) A reconciled tree constructed with GeneTree from the tree shown in (a). Duplication events are marked with small boxes at the nodes, orthologous relationships can be easily assigned. (c) Results from the Orthostrapper program. An ortholog bootstrap value is calculated for each possible sequence pairing in the tree between the two (groups of) species the analysis is done for. The values were calculated from 1000 pseudosamples.

level is given at an ortholog bootstrap value of 58% in the case of equal mutation rates (see Figure 2). There are improvements to the bootstrap method that allow a better estimate of the confidence, namely the second-level bootstrap. But this would mean calculating at least 200 additional bootstrap trees for each orthologous pairing. The CPU time needed for this would contradict the goal of large-scale analysis. For a more precise look at a small set of sequences, a Monte Carlo Markov Chain (MCMC) approach (Yang and Rannala, 1997) might be better suited, however also here computation times become prohibitively long for trees with over 50 sequences. Furthermore, these approaches are unlikely to lead to significantly improved rooting, which is one of the major problems in correct ortholog assignment. Rooting with an outgroup is often not possible when looking for orthologs: in order to get the correct position of the root the phylogenetic relationships of outgroup sequences to the rest of the tree have to reflect the species tree. They either have to be paralogous to all or orthologous to all other sequences in the tree—but the phylogenetic relationships of the sequences in tree are often unknown. The rather *ad-hoc* method used here by finding the center of the tree assumes similar divergent rates of the sequences. This is not true for some protein families. Thus in case of very unbalanced trees the root will be placed at the wrong position. The simulations indicate that the probability for a wrong assignment is increased by approximately 5–10% in case of a very unbalanced tree. A way to decrease this error is to add distant sequences to the trees. By doing this the phylogenetic relationships one is interested in will be more distant from the center of the tree and therefore less affected by a slightly wrong position of the root.

Some general pitfalls that are not directly connected to the ortholog bootstrapping method are still present. As shown in Figure 1b, incomplete genomes/partial gene loss



Ortholog bootstrap values

	Y58A7A.1	F27C1.2	F31E8.4
O15431	**.766**	.467	.407
O15432	.375	.475	.373

Fig. 5. Because of the low bootstrap support for the branching pattern of the subtrees, it is not possible in this case to confidently assign orthologs based on the original tree. However, the ortholog bootstrap values indicate that only O15431 and Y58A7A.1 are frequently (76.6% of all bootstrap trees) found in an orthologous relationship, whereas the other sequences are found in ortholog assignments with a significantly lower frequency. Shown is a subtree of a larger tree, calculated by neighbor joining, using uncorrected distances and 1000 bootstrap samples, root is set at the center of the tree. The ortholog bootstrap values were calculated from 1000 pseudosamples.

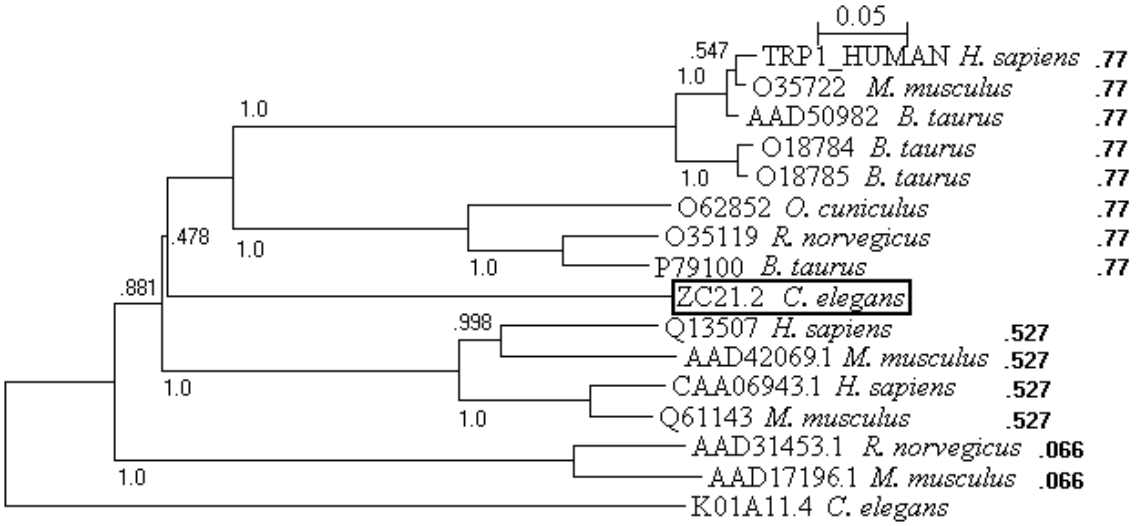


Fig. 6. Potential orthologous relationships found with ortholog bootstrapping that are not present in the optimal tree. Ortholog bootstrap values between the *C.elegans* gene ZC21.2 and mammalian homologs are shown on the right. ZC21.2 is assigned a high value of 0.77 to the mammalian genes it clusters with in the optimal tree. However, the values indicate that the optimal tree does not reflect all potential orthologous relationships. In 52.7% of the analyzed bootstrap trees orthology is reported between ZC21.2 and Q61143/CAA06943/AAD42069/Q13507, suggesting that ZK21.2 is orthologous to these sequences as well. All ortholog bootstrap values between K01A11.4 and the mammalian sequences were 0.0. The tree was calculated with neighbor joining, using uncorrected distance and 1000 bootstrap samples; the root was set at the center of the tree. Calculation of ortholog bootstrap values was done with 1000 samples. All mammalian sequences in the tree were put in one group and analyzed for orthologs in *C.elegans*. Orthologous relationships within the mammalian group were not analyzed.

will increase the risk of assigning orthology incorrectly. This can be prevented by including orthologous sequences from an outgroup species, but often none are available. Another way is to enlarge the species groups. For instance when looking for orthologs between *Homo sapiens* and *Caenorhabditis elegans*: instead of only comparing

sequences from those two species one should include sequences from other vertebrates and nematodes. Assume a gene that was lost in *H.sapiens* is present in another vertebrate. Orthology would then be assigned correctly between the *C.elegans* sequence and this gene. With only human sequences in the analysis, an ancient paralog of this gene in human might be incorrectly assigned as an ortholog to the *C.elegans* gene.

It is also a good idea to use several different methods for tree construction and compare the results when assigning orthology, since different methods often produce trees of different topology. Orthostrapper can read trees from any program that produces output in the Newick format (Felsenstein *et al.*, 2000).

ACKNOWLEDGEMENT

We thank Maido Remm for providing the testset and helpful discussions.

REFERENCES

Efron,B., Halloran,E. and Holmes,S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proc. Natl Acad. Sci. USA*, **93**, 13 429–13 434.

Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Felsenstein,J., Archie,J., Day,W.H.E., Maddinson,W., Meacham,C., Rohlf,F.J. and Swofford,D. (2000) The Newick tree format. htpp://www.evolution.genetics.washington.edu/phylip/newicktree.htm1.

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.

Gogarten,J.P. and Olendzenski,L. (1999) Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.*, **9**, 630–636.

Goodman,M., Czelusniak,J., Moore,G.W., Romero-Herrera,A.E. and Matsuda,G. (1979) Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, **28**, 132–168.

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 ecoded by the human, fly, nematode, and yeast genomes. *Genome Res.*, **8**, 590–598.

Newton,M.A. (1996) Bootstrapping phylogenies: large deviations and dispersion effects. *Biometrika*, **82**, 315–328.

Page,R.D. (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, **14**, 819–820.

Page,R.D.M. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.*, **43**, 58–77.

Remm,M. and Sonnhammer,E. (2000) Classification of transmembrane protein families in the *Caenorhabditis elegans* genome and identification of human orthologs. *Genome Res.*, **10**, 1679–1689.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Stoye,J., Evers,D. and Meyer,F. (1998) Rose: generating sequence families. *Bioinformatics*, **14**, 157–163.

Studier,J.A. and Keppler,K.J. (1988) A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.*, **5**, 729–731.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.*, **10**, 19–29.

Yang,Z. and Rannala,B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.*, **14**, 717–724.

Yuan,Y.P., Eulenstein,O., Vingron,M. and Bork,P. (1998) Towards detection of orthologues in sequence database. *Bioinformatics*, **14**, 285–289.

Zharkikh,A. and Li,W.H. (1995) Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. *Mol. Phylogenet. Evol.*, **4**, 44–63.