BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 6 2006, pages 699–707
doi:10.1093/bioinformatics/btk040

Phylogenetics

OrthologID: automation of genome-scale ortholog identification within a parsimony framework

Joanna C. Chiu[1,†], Ernest K. Lee[2], Mary G. Egan[2], Indra Neil Sarkar[2,3], Gloria M. Coruzzi[1,*] and Rob DeSalle[2]

[1]Department of Biology, New York University, New York, NY 10003, USA, [2]Division of Invertebrate Zoology and [3]Division of Library Services, American Museum of Natural History, New York, NY 10024, USA

Received on July 10, 2005; revised and accepted on January 4, 2006
Advance Access publication January 12, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Motivation: The determination of gene orthology is a prerequisite for mining and utilizing the rapidly increasing amount of sequence data for genome-scale phylogenetics and comparative genomic studies. Until now, most researchers use pairwise distance comparisons algorithms, such as BLAST, COG, RBH, RSD and INPARANOID, to determine gene orthology. In contrast, orthology determination within a character-based phylogenetic framework has not been utilized on a genomic scale owing to the lack of efficiency and automation.

Results: We have developed OrthologID, a Web application that automates the labor-intensive procedures of gene orthology determination within a character-based phylogenetic framework, thus making character-based orthology determination on a genomic scale possible. In addition to generating gene family trees and determining orthologous gene sets for complete genomes, OrthologID can also identify diagnostic characters that define each orthologous gene set, as well as diagnostic characters that are responsible for classifying query sequences from other genomes into specific orthology groups. The OrthologID database currently includes several complete plant genomes, including *Arabidopsis thaliana, Oryza sativa*, *Populus trichocarpa*, as well as a unicellular outgroup, *Chlamydomonas reinhardtii*. To improve the general utility of OrthologID beyond plant species, we plan to expand our sequence database to include the fully sequenced genomes of prokaryotes and other non-plant eukaryotes.

Availability: http://nypg.bio.nyu.edu/orthologid/

Contact: gloria.coruzzi@nyu.edu

INTRODUCTION

The vast and rapidly increasing amounts of genomic and expressed sequence tags (ESTs) data provide researchers with great opportunities for advancement in many areas of biology. Very often, the identification of orthologous genes across species is a prerequisite for comparative genomic studies. By definition, orthologs are genes that have evolved from a common ancestry through speciation, and therefore are often predicted to retain similar biochemical functions. Identification of orthologous gene sets in multiple species has allowed researchers to apply the molecular and biochemical data gained from studying model organisms to study other organisms that are not as easily manipulated genetically. Gene expression analyses, gene content studies and developmental biology are but a few areas that have already benefited from this comparative genomic approach.

Another area of biology that requires the identification of orthologous genes is phylogenetics. In order to generate meaningful phylogenetic hypotheses for species evolution through character-based or distance-based analysis, it is essential that only orthologous gene sets are aligned and analyzed. Traditionally, phylogenetic trees have been generated from a single or a small number of orthologous genes because of the lack of available orthologous gene sets. It has become increasingly clear that small-scale analyses, based on only a few (or single) gene regions may not reflect the evolutionary history of the species but rather, may reflect the evolutionary history of the molecules themselves. One approach to this 'gene tree, species tree' problem is to perform simultaneous analyses that combine larger numbers of data partitions (e.g. orthologous gene sets) (Miyamoto, 1985; Kluge, 1989, 1997; Chippindale and Wiens, 1994; Olmstead and Sweere, 1994; Nixon and Carpenter, 1996; Gatesy *et al*., 1999, 2002, 2003; Gatesy and Arctander, 2000; Rokas *et al*., 2003; Matthee *et al*., 2004; Bardeleben *et al*., 2005; Bruvo-Madaric *et al*., 2005; Wahlberg *et al*., 2005). Numerous studies have demonstrated that simultaneous analyses of multiple data partitions can result in an increase in overall character support, despite conflict among the characters, due to emergent properties not evident in the separate analyses of individual data partitions (Gatesy *et al*., 2002, 2003). Rokas *et al*. (2003) obtained a fully resolved species tree with high bootstrap support for seven species of *Saccharomyces* by analyzing a supermatrix (a data matrix comprised of multiple data partitions) that included 106 genes. In a subsequent study, they examined the relative benefit of increasing the numbers of genes or taxa included to assess phylogenetic consistency, and concluded that increasing gene number had a significantly positive effect (Rokas and Carroll, 2005).

*To whom correspondence should be addressed.
†Present address: Department of Molecular Biology and Biochemistry, Rutgers University, Center for Advanced Biotechnology and Medicine, Piscataway, NJ 08854, USA
The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

The increasing amount of genomic and EST sequences currently available provide researchers the raw materials with which to generate phylogenies on a genomic scale, instead of using only a few annotated genes. However, gene duplication events and subsequent diversification of duplicated sequences present difficulties for the assembly of comparative datasets due to the presence of families of genes. One major difficulty in the use of sequence data for both phylogenetic studies and comparative genomics is the determination of gene orthology, which is especially time-consuming and problematic in cases of large gene families. A further limitation for researchers working within a character-based parsimony framework is the lack of character-based automated tools for phylogenetic analysis and orthology determination. Here, we present OrthologID, an automated Web-based tool for the identification of orthologous genes within a character-based parsimony framework.

OrthologID was developed as a collaborative project by the New York Plant Genomics Consortium (NYPG; http://nypg.bio.nyu.edu), to facilitate the identification of gymnosperm EST sequences that are orthologous to the sequences in the completed genomes of *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa* and *Chlamydomonas reinhardtii*. In order to address several unresolved questions regarding the evolution of seed plants using molecular data, it was necessary to greatly increase the amount of data brought to bear on these questions, especially since complete gymnosperm genomes have not been sequenced. A large amount of EST data from primitive gymnosperm species was generated by NYPG to provide the raw materials. The development of OrthologID provides the necessary automated tools for the identification of orthologous gene regions across plant taxa. This represents the first step towards making numerous additional gene regions available for genome-scale phylogenetic analysis.

OVERVIEW OF ORTHOLOGY DETERMINATION METHODS

Currently, many researchers use pairwise sequence comparison schemes, such as BLAST (Altschul *et al*., 1990), COG (Clusters of Orthologous Groups; Tatusov *et al*., 1997, 2000, 2001, 2003), INPARANOID (Remm *et al*., 2001; O'Brien *et al*., 2005), RBH (Reciprocal Blast Hits; Hirsch and Fraser, 2001; Jordan *et al*., 2002) and RSD (Reciprocal Smallest Distance Algorithm; Wall *et al*., 2003) to determine gene orthology on a genomic scale.

Although highly efficient, it has been shown that BLAST may not be a reliable approach for determining orthology relative to evolutionary history (Koski and Golding, 2001). RBH has limitations in cases where the forward BLAST best hit identifies an ortholog, and the reverse best hit identifies a paralog, resulting in these gene pairs being excluded. This scenario may be more common in multi-gene families with numerous gene duplication events, resulting in closely related paralogs. As a result, the number of ortholog pairs identified from two genomes using RBH may be an underestimate. To improve the sensitivity of reciprocal genome queries, Wall *et al*. (2003) developed the RSD algorithm. Instead of relying solely on BLAST scores to determine reciprocal best hits, RSD also uses character-based maximum likelihood estimation of evolutionary distances to identify orthologs between genomes. The number of ortholog gene sets identified from two test genomes using RSD increased significantly when compared with that identified using RBH (Wall *et al*., 2003).

The algorithms mentioned above identify one-to-one orthology relationships between two genomes. However, orthologs between two species may not be expressed as simple one-to-one relationships; one-to-many or many-to-many relationships will naturally occur in cases where duplication events following speciation produce a pair of true orthologs, termed in-paralogs by Remm *et al*. (2001). Whereas two in-paralogs from one species can both be regarded as true orthologs to a single gene in a second species; out-paralogs, which are paralogs resulting from a duplication preceding speciation, cannot be classified as orthologs. In terms of comparative functional studies, since in-paralogs in one species are most closely related to their ortholog in another species, the function of the ortholog may be useful for predicting the function of the in-paralogs. However, it is important to consider that in-paralogs may have subsequently diverged functionally.

Two widely used algorithms that are built on all-against-all pairwise comparisons, INPARANOID (Remm *et al*., 2001; O'Brien *et al*., 2005) and COG (Tatusov *et al*., 1997, 2000, 2001, 2003), identify both one-to-many and many-to-many orthologous relationships. COG identifies groups of sequences from at least three species that most likely represent orthologs (Tatusov *et al*., 2000). Although COG determines orthology based on pairwise sequence comparisons as in RBH, RSD and INPARANOID, it also takes advantage of structural data when available (Tatusov *et al*., 2000). Of the two methods, COG and INPARANOID, only INPARANOID is capable of differentiating between in-paralogs (true orthologs by definition) and out-paralogs (Remm *et al*., 2001). However, owing to the nature of its algorithm, INPARANOID can only identify orthologs from two genomes at a time.

PHYLOGENETIC ORTHOLOGY DETERMINATION USING OrthologID

Tree building methods are an alternate approach to the identification of orthologous gene regions. In this approach, gene family trees are constructed to explore the evolutionary history of gene family members. Complete genomes from multiple species can be included in the analysis. The resulting trees are screened to identify orthologous groups (clades, clusters, nodes). Within a character-based parsimony framework, shared derived characters provide the information for group membership. Stated another way, nodes are defined by shared derived characters. Tree building approaches are computationally labor intensive. Tree building involves a number of manual steps: (1) searching and downloading sequences for putative members of a gene family from fully sequenced genomes of the species of interest; (2) performing alignments of retrieved sequences using multiple sets of alignment parameters; (3) combining multiple alignments into a single robust alignment either by elision (Wheeler *et al*., 1995) or culling (Gatesy *et al*., 1993); (4) transforming the resulting alignment into the appropriate format for downstream analyses, e.g. compatible with PAUP* (Swofford, 2003) and (5) performing searches within a parsimony framework using PAUP*, and saving resultant trees and computing the strict consensus in the case of multiple trees.

To use this tree building approach to identify orthologous gene sets from complete genomes would require that all of the manual steps described be repeated thousands of times. Without automation, high-throughput parsimony-based orthology determination would be practically impossible owing to the manually intensive

and time-consuming procedures for generating gene trees. We have therefore developed OrthologID, a Web-based tool that fully automates the phylogenetic approach for the identification of orthologs within a character-based parsimony framework. OrthologID automatically searches the local database of completely sequenced genomes and clusters gene sequences into putative gene families, performs sequence alignments using multiple sets of alignment parameters and culls alignment-ambiguous regions. It performs tree searches within a parsimony framework, saves resultant trees and computes the strict consensus when multiple equally parsimonious trees are obtained from the analysis.

OrthologID uses sequences from completely sequenced genomes to generate gene family trees and identify orthologous gene sets. This avoids the potential for error that missing in-paralogs could pose if partially sequenced genomes were used to construct gene trees. OrthologID uses the most ancestral taxon to root each gene family tree. The choice of an outgroup taxon is essential in order to adequately define the ingroup in terms of evolutionary history. The choice of the outgroup taxon need only be the most ancestral to the taxa that are considered to be in the ingroup; it need not necessarily be the sister taxon (Nixon and Carpenter, 1993). From a phylogenetic perspective, the outgroup need only be a distant relative to a set of ingroup taxa such that the ingroup taxa are more closely related to each other than to the outgroup taxon (Smith, 1994). In this context, *C.reinhardtii* was chosen as an outgroup taxon deemed ancestral to the other taxa considered for each gene family.

QUERY CLASSIFICATION USING OrthologID: AUTOMATION OF THE GUIDE TREE/CAOS APPROACH

Prior to the development of OrthologID, the placement of query sequences (e.g. ESTs) into orthology groups using a character-based approach required manual rebuilding of gene family trees for each new query to be classified. This laborious process made high-throughput classification of query sequences difficult, and hindered the use of sequence data for genome-scale combined phylogenetic analysis as well as for comparative functional genomics. OrthologID overcomes this limitation by classifying query sequences using the CAOS algorithm (Sarkar *et al*., 2002) and the 'guide tree' approach. CAOS is a rapid algorithm for determining gene orthology based on derived traits shared between orthologous genes. By incorporating the CAOS algorithm, OrthologID classifies new query sequences (full-length cDNA or EST) from genomes that are not completely sequenced, based on the phylogenetic and orthology relationships that are already determined through the analysis of complete genomes. This is similar to the manner in which COGNITOR is used to classify query sequences into COGs using the existing COG database (Tatusov *et al*., 2000). In the guide tree/CAOS approach implemented in OrthologID, a complete parsimony gene family tree that is used to identify orthologous groups from complete genomes is used as a guide tree for classifying query sequences from other species. This guide tree is fed to the CAOS algorithm for the identification of characters that are diagnostic of each node and each orthologous gene set. In order to place query sequences into orthology groups assembled from complete genomes, CAOS screens the query sequence for the presence of characters that are diagnostic of nodes on the guide tree. The CAOS algorithm and the use of guide trees are an improvement over traditional tree building approaches since the guide tree/CAOS approach eliminates the need to manually rebuild a gene family tree for each new query to be classified.

OrthologID SYSTEM ARCHITECTURE

Backend of OrthologID

The backend of OrthologID is composed of a database of sequences and phylogenetic trees and four interconnected modules [a Gene Family Creator (GFC), an Alignment Constructor, a Tree Builder, and a Diagnostics Generator]. Each module can be easily upgraded as better algorithms become available. The flowchart in Figure 1 shows the components of OrthologID and their interactions.

The GFC clusters genes from complete genomes into gene families. GFC searches each ingroup gene against both ingroup and outgroup genomes using NCBI BLAST (Altschul *et al*., 1997). For clustering purposes, an expectation value cutoff of 1e−20 is used. For a pair of genes g_1 and g_2, g_1 is defined as clusterable with g_2 if the *E*-value in the BLAST of g_1 against g_2 is within the aforementioned cutoff, and the alignable regions of the two genes are at least 80% of the longer sequence. The latter criterion is used to avoid the clustering of genes that only share one structural domain with high sequence similarity. A gene g is considered a member of the gene family $\mathtt{F}$ if at least one other gene in $\mathtt{F}$ is clusterable with g. After performing all-against-all BLAST searches, GFC randomly picks a gene g from one of the ingroup genomes and looks for clusterable genes in the BLAST result of g. Each clusterable gene is added to the current family, and this gene's BLAST result is again searched for new members. This process is repeated until no more genes can be clustered to the current family. GFC then starts a new gene family, and the above steps are repeated. Algorithmically, GFC treats each gene g_i as a vertex in a directed graph G. An edge exists from g_i to g_j if g_i is clusterable with g_j. The clustering algorithm starts with a vertex that has not been visited, and traverses the graph G in a depth-first manner. Each gene encountered during the traversal is added to the current family $\mathtt{F}$. If a gene that belongs to a previously constructed gene family $\mathtt{F}'$ is encountered, $\mathtt{F}'$ is merged into $\mathtt{F}$. This process is iterated until all vertices in G have been visited. Gene family membership in the OrthologID database is based on the above criteria.

The Alignment Constructor creates robust alignments for each gene family. The results of tree building analyses and subsequent character-based orthology determination depend heavily on alignment. The program MAFFT version 5 is considered one of the more efficient and reliable multiple alignment programs based on benchmark tests comparing MUSCLE (Edgar, 2004), TCoffee (Notredame *et al*., 2000) and ClustalW (Thompson *et al*., 1994), relative to the BAliBASE benchmark dataset (Katoh *et al*., 2005). The Alignment Constructor makes use of the MAFFT L-INS-i algorithm, which is an iterative refinement method with local pairwise alignment information. The Alignment Constructor uses different sets of alignment parameters to create three different alignments for each gene family. The three pairs of gap open penalty and offset values are (1.53, 0.123), (2.4, 0.1) and (1.0, 0.2). Alignments are compared and alignment-ambiguous regions culled (Gatesy *et al*., 1993). The resulting, culled alignment is then passed on to the Tree Builder.

The Tree Builder module generates gene family trees within a parsimony framework. Where possible (for small gene families with

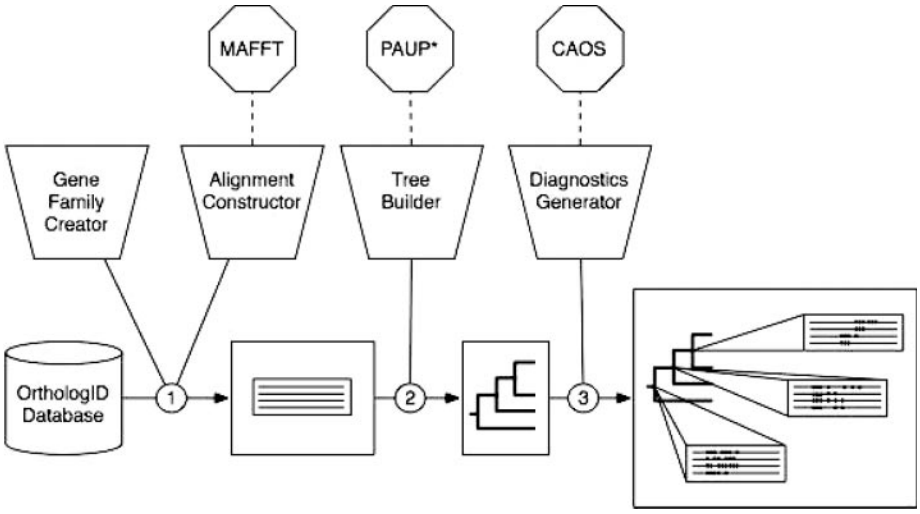

Fig. 1. Overview of OrthologID. Maximum parsimony trees are generated and diagnostic characters are determined through an automated process: (1) sequences are retrieved from OrthologID Database and clustered using the Gene Family Creator and aligned, using the Alignment Constructor (which interfaces with MAFFT); (2) phylogenetic trees are generated using the Tree Builder (which interfaces with PAUP*) and (3) diagnostic characters are ascertained using the Diagnostic Generator (which interfaces with CAOS). Each OrthologID module, shown as trapezoids, are designed to function independently and allow the use of any processing tool (e.g. one could use ClustalW instead of MAFFT for sequence alignment).

fewer than 13 sequences), exhaustive or branch and bound tree searches are performed (as implemented in PAUP*; Swofford, 2003). For large gene families, the enumeration of all possible trees is infeasible because finding the most parsimonious tree is an NP-complete problem (Felsenstein, 1978). Therefore, for alignments with larger numbers of sequences, tree space is rigorously explored using the parsimony ratchet (Nixon, 1999). Each iteration of a ratchet starts with a limited TBR search to generate an initial tree. This tree is used as a 'starting tree' for a search with 10–15% of characters reweighed. The shortest tree is again used as a starting tree to perform another TBR search with all the weights reset. Each ratchet consists of 200 such iterations. The Tree Builder computes 20 ratchets and performs a final TBR swap on the best trees, in order to visit multiple islands of tree space. Where more than one equally parsimonious tree results from the analysis, a strict consensus is computed. This consensus tree is used to identify orthology relationships in complete genomes, and used as a gene family guide tree for Query Orthology Classification. Currently, the Tree Builder module implements the parsimony ratchet using PAUP*. However, the module can easily be changed to accommodate maximum likelihood or Bayesian methods for generating phylogenetic trees. The automation of the calculation of confidence measures, such as Bremer support, is in progress. This module will be added in the near future.

Finally, the Diagnostics Generator is invoked to identify diagnostic characters for orthologous groups using the CAOS algorithm (Sarkar *et al*., 2002).

OrthologID database

At present, the completed genomes of three ingroup species, *A.thaliana*, *O.sativa* and *P.trichocarpa*, and an outgroup species, *C.reinhardtii,* are included in the OrthologID database. The analysis of the above complete genomes resulted in 136 781 gene sequences clustered into 8314 gene families and phylogenetic trees. *A.thaliana* and *O.sativa* sequences were obtained from TIGR, and the *P.trichocarpa* and *C.reinhardtii* gene sets from JGI. As whole genomes from additional plant species become available, new gene family trees and tree node diagnostics will be regenerated from scratch, thus making the database increasingly more comprehensive. In addition to plant genomes, future versions of OrthologID database will include complete genomes from prokaryotic and non-plant eukaryotic species.

Frontend of OrthologID: web interface

The web interface (OrthologID http://nypg.bio.nyu.edu/orthologid/) allows users to (1) search for orthologous gene sets in complete genomes that are available in the OrthologID database (Orthologous Group Search) and (2) classify query sequences into existing orthology groups from complete genomes (Query Orthology Classification).

Orthologous group search A gene locus tag (for *A.thaliana* and *O.sativa* only in current database) is submitted to obtain the gene family tree containing the orthology group the input gene belongs to. The OrthologID Tree and Diagnostics Viewer presents the orthology groups in an interactive phylogenetic tree format. The culled alignment of the gene family and the diagnostic characters defining each of the tree nodes are also presented simultaneously on a split screen. By clicking on the position of the nodes defining each orthologous group reveals the underlying characters that support the grouping. Support measures for gene family guide trees and a searchable text format for orthology groups will be available in the future.

Query orthology classification The user inputs a query for orthology determination by submitting a nucleotide or amino acid sequence in FASTA format. OrthologID performs an initial BLAST search to identify the relevant gene family in the database, and then invokes the CAOS-based classifier to place the query sequence into the corresponding guide tree. In theory, CAOS places

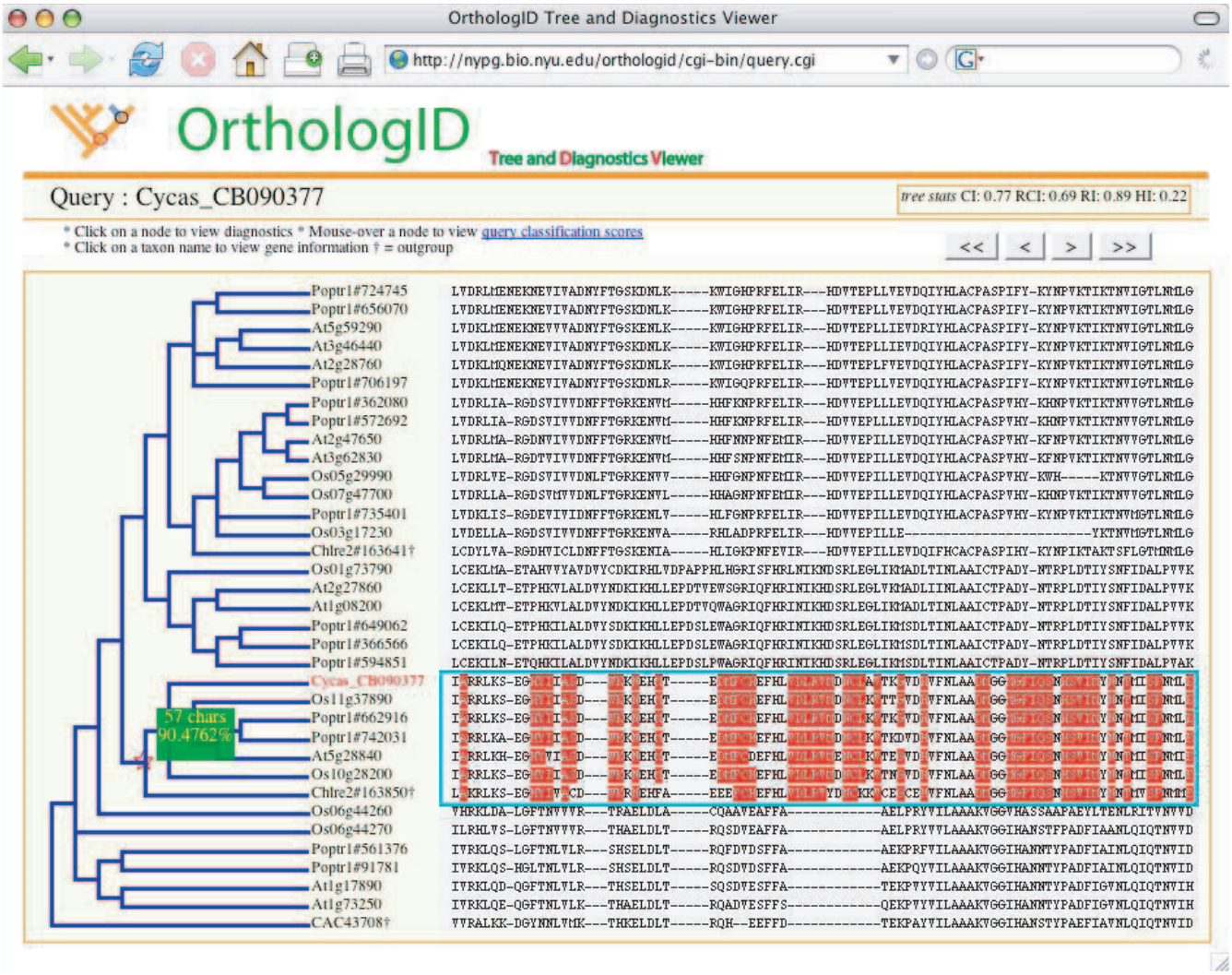

Fig. 2. OrthologID Tree and Diagnostics Viewer showing placement of query by Query Orthology Classification. OrthologID places the query sequence in the gene family guide tree. Tree statistics after query placement are presented in the top right corner of the Viewer (CI, consistency index; RCI, rescaled consistency index; RI, retention index; HI, homoplasy index). The user clicks on a node along the path of classification, and the diagnostic characters used to classify the query at that node are highlighted in the alignment. In addition, a pop-up at each node along the classification path shows (1) the number of diagnostic characters used in the classification and (2) the query placement score.

a query at a tree node into a specific descendant clade even if the classification support index (CSI; the difference in the number of diagnostics in the descendant clades) for the node is very small, as long as it is positive. Owing to possible errors in sequencing and ORF prediction, especially in genome-scale projects, query placement should be treated with caution when the CSI is small. In order to screen for potential problems from these types of errors, a most parsimonious placement filter (MPPF) for the CAOS classifier has been applied. Given a guide tree and a query to be classified, MPPF finds the most parsimonious placement of the query given the guide tree as the constraint, by inserting the query into every possible node of the guide tree, and selecting the shortest tree. The node where the query is placed in the best tree is then used as a cutoff for CAOS classification.

The Tree and Diagnostics Viewer displays the gene family guide tree showing the placement of the query sequence with the alignment in a split view format (Fig. 2). The user can click on any node along the classification path and see the highlighted diagnostic characters used for classifying the query in the alignment. Additional data concerning the query placement at each of the relevant nodes are presented in a pop-up box as users click on each of these nodes. First, the pop-up shows the number of diagnostic characters responsible for placing the query into a particular clade as opposed to other clades at each node. Since the absolute number of diagnostic characters may not be a true indication of how decisively the query is being placed into a clade, an additional measure is generated. The query placement score is defined by the function $S(a) = (k_a/k_a + k_{b1} + k_{b2} + \cdots + k_{bn}) \times 100\%$, where k_a is the number of diagnostic characters shared between the query and the sequence(s) in clade a, and k_{bi} is the number of diagnostic characters shared between the query and the sequence(s) in one of the n sister clades, b_i, of a. Figure 3 shows two examples in which

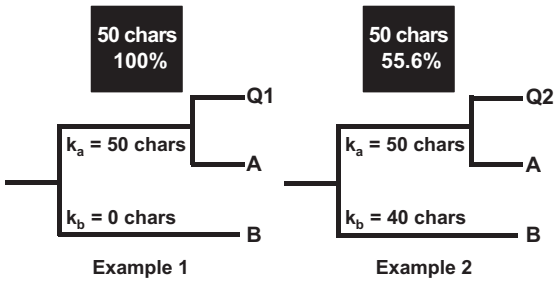

Fig. 3. Two examples illustrating the calculation of query placement score given in the pop-up boxes in the Tree and Diagnostics Viewer. In Example 1, query sequence Q1 shares 50 characters ($k_a = 50$) with the gene(s) in clade A and zero characters ($k_b = 0$) with the gene(s) in clade B. As a result, OrthologID places Q1 into clade A. The query placement score is expressed as $(k_a/k_a + k_b) \times 100\%$. The resulting score (100%) and the number of diagnostic characters responsible for placing Q1 (50) are shown in the pop-up box. In Example 2, $k_a = 50$ and $k_b = 40$; as a result, the strength of the placement of query Q2 is weaker than that of Q1 in Example 1. This is illustrated by the lower query placement score in Example 2 (55.6%).

50 characters are used in both cases to place the query into clade A versus B, even though the query placement in Example 1 is much stronger than in Example 2. This is reflected by the higher query placement score in Example 1 (100%), as compared with that in Example 2 (55.6%). For ease of explanation, both examples in Figure 3 show cases in which the query can be placed into either one of two descendant clades. In reality, the query placement score can be calculated even when there are more than two descendant clades. At present, it is not feasible to generate standard support measures, such as Bremer support values, for real-time query placement, as this will greatly increase processing time. However, given the design of the CAOS algorithm used for query placement, the query placement score is an efficient and representative measure of query placement strength.

IMPLEMENTATION

The backend logic of OrthologID is implemented in Perl and C++. Gene families, guide trees and diagnostics are stored in MySQL databases. The tree and diagnostics viewer makes extensive use of JavaScript, and is best viewed with standards compliant browsers, such as Netscape and Firefox.

RESULTS OF EFFICACY TESTS OF OrthologID

Since OrthologID uses the guide tree/CAOS approach as an automated approximation to the full-scale phylogenetic approach to orthology determination of query sequences, these two approaches should give comparable results. In order to test this assumption, comparisons were made by examining the placement of query sequences by OrthologID against the placement generated using full-scale parsimony analysis. A total of 36 plant sequences from a diverse range of species other than the ones whose genomes are included in OrthologID database were randomly selected from NCBI and NYPG databases. These sequences were submitted to OrthologID for orthology determination against the complete genomes of *A.thaliana*, *O.sativa*, *P.trichocarpa* and *C.reinhardtii*. The matrix used for full-scale parsimony analysis for each plant query sequence was retrieved from the OrthologID database, and was the same alignment used by OrthologID to classify that query sequence. Parsimony analysis was performed using the same protocol employed by OrthologID Tree Builder to generate guide trees.

For 77.8% of the 36 plant query sequences, OrthologID and full-scale parsimony analysis resulted in the same orthology classification. In the remaining test cases, although OrthologID and full-scale parsimony analysis placed each of the query sequences in the same clade, their orthology classifications were different. OrthologID placed 13.9% of the 36 queries closer to the terminal nodes when compared with the query placement based on full-scale parsimony analysis (i.e. OrthologID narrowed down the orthologous genes to a fewer number of genes). The more 'precise' placement of the query by OrthologID may be explained by the fact that the phylogenetic tree generated from full-scale parsimony analysis is a strict consensus from equally parsimonious hypotheses. OrthologID addresses polytomies in guide trees by examining character reconstructions for each equally parsimonious resolution of the polytomy. Therefore, the query placement by OrthologID may actually represent one of the most parsimonious hypotheses, and yet be represented as a polytomy in the strict consensus parsimony tree. On the other hand, full-scale parsimony analysis placed 8.3% of the queries closer to the terminal nodes when compared with OrthologID, indicating a lack of diagnostic characters in the sequences at the terminal nodes.

In addition to comparing OrthologID with full-scale parsimony analysis as a benchmark, the effectiveness of OrthologID for identifying orthologous gene sets, that include both single orthologs and in-paralogs, for query sequences, was also examined. Of the 36 plant query sequences tested against the current OrthologID plant database, 66.7% were successfully placed into orthology groups with single orthologs or groups of in-paralogs. This success rate is quite high, given the frequent occurrence of gene duplication in plants.

One desirable feature of non-phylogenetic approaches to orthology determination is their ability to screen through genomes of widely divergent taxa for the presence of distant homologs. A limitation of the phylogenetic method is that it would require the construction of 'tree-of-life-scale' gene trees to accomplish a similar task. To examine the current general utility of the plant-specific OrthologID database for non-plant query sequences, 32 animal query sequences were submitted to OrthologID for orthology placement. The comparison was restricted to the same set of gene families to ensure that any observed differences are due to the fact that query sequences are from closely related versus more distant species, rather than differences in tree topologies. To choose these gene families for testing, a single gene family member from each of the 36 plant gene families previously tested for plant query sequences was randomly selected and submitted to BLAST search against non-plant eukaryotic genomes, ranging from that of *Caenorhabditi elegans* to *Homo sapiens*. Of the 36 plant gene families examined, only 32 were found to have non-plant eukaryotic homologs. When submitted as queries, OrthologID was unable to place any of them into orthology groups that contain single orthologs or in-paralogs. In fact, OrthologID placed 62.5% of the animal queries just inside of the outgroup, and sister to all plant sequences.

This demonstration of the restricted utility (in terms of comparative genomic searches for distant homologs) of the current version of OrthologID is not surprising since the guide trees used for

classification were solely composed of plant genomes. Construction of a gene family guide tree from multiple animal as well as plant genomes might be of greater utility, yet within a character-based phylogenetic framework, it is unlikely that a reliable guide tree could be constructed since it would require available completed genomes of a broad and representative taxonomic sampling as well as appropriate outgroups. Nevertheless, OrthologID may have applicability to comparative genomics. Rather than attempting to construct 'tree-of-life-scale' gene family trees, several databases of curated gene family trees, each from a different and distant taxonomic group, could be probed sequentially by a query sequence using OrthologID. This could be accomplished with minimal modification to OrthologID. It would require that additional genomic databases be added to OrthologID, and gene family tree and diagnostics databases be generated using the automated tools currently available in OrthologID. In this way, OrthologID may provide a character-based alternative to comparative genomic searches for distant homologs.

DISCUSSION

OrthologID represents an alternative character-based automated approach for genome-scale orthology determination. The most significant difference between OrthologID and the existing orthology determination tools is that OrthologID classifies orthology by employing phylogenetic analysis within a character-based parsimony framework. Parsimony phylogenetic analysis explores the evolutionary history of genes and species, and thus is a natural way to detect orthologs (Remm *et al*., 2001). The first step of OrthologID uses the results from all-against-all pairwise comparisons between multiple genomes, not unlike that used in COG, as a criterion to designate gene family membership. This is followed by parsimony analysis that places genes from completely sequenced genomes into orthology groups. With the exception of RSD (Wall *et al*., 2003), which uses maximum likelihood distance estimation as a 'best hit' criteria in addition to pairwise sequence comparisons, all the other current methods do not involve any phylogenetic analysis component. One obvious advantage of OrthologID over the other tools is that since it is the only tool that uses phylogenetic tree analysis to determine orthology, it is also the only tool that provides a bona fide phylogenetic tree in the output. The COG output includes a cluster dendrogram generated using BLAST scores between COG members as the measure of similarity, and cannot be used to replace comprehensive phylogenetic analysis (Tatusov *et al*., 2000).

As mentioned earlier, orthology relationships between two species cannot always be expressed as simple one-to-one relationships, as in cases where a duplication event occurs after speciation, resulting in multiple in-paralogs being orthologous to a single gene in another species. Whereas OrthologID, INPARANOID and COG are all capable of detecting one-to-many and many-to-many orthology relationships, only OrthologID and INPARANOID can reliably differentiate between in-paralogs and out-paralogs. INPARANOID achieves this by using pairwise sequence comparisons within and between genomes; while OrthologID uses phylogenetic information from outgroup sequences.

Another advantage of OrthologID over RBH, RSD and INPARANOID is that it is not restricted by reciprocal genome queries, and can therefore determine orthology relationships between more than two complete genomes simultaneously. The number of complete genomes OrthologID uses for generating gene family phylogenetic trees determines the number of species from which genes are considered for orthology determination.

In addition to determining orthology for genes from complete genomes, OrthologID employs the CAOS algorithm to identify diagnostic characters that define all the nodes on the phylogenetic trees it generates, including the tree nodes that group orthologous gene sets. OrthologID can therefore screen query sequences (full-length cDNA or EST) from new genomes for diagnostic characters, and place them in orthology groups compiled using completely sequenced genomes when appropriate. While this approach is similar to the manner in which COGNITOR is used to classify query sequences into COGs using the existing COG database (Tatusov *et al*., 2000), the underlying principles and algorithms of the two approaches are different. In contrast to COGNITOR, which does not clearly differentiate between in-paralogs (true orthologs by definition) and out-paralogs (Remm *et al*., 2001), OrthologID is able to differentiate the two based on the placement of the sequences on the output phylogenetic trees. As with other analyses involving genomes that are not completely sequenced, the possibility exists that additional in-paralogs to the query sequence may be present in the unsequenced portion of the genome. However, query classification by OrthologID using the CAOS algorithm and gene family guide trees generated from complete genomes, is perhaps the most reliable high-throughput method available when query sequences are from genomes that are not completely sequenced. The rationale for using only completely sequenced genomes for constructing guide trees with OrthologID is to minimize the possibility of the erroneous placement of query sequences due to missing data. In other words, had gene family trees been constructed using partially sequenced genomes, it is possible that some gene family members would be missing, in which case it could be possible that queries orthologous to these missing gene family members would be incorrectly placed. The fact that OrthologID can efficiently screen through EST sequences from genomes that are not completely sequenced and place them into orthologous gene sets from complete genomes, allows the use of available ESTs for genome-scale phylogenies and comparative functional studies. OrthologID was developed to explore the utility of EST data in parsimony phylogenetic analysis for resolving some key questions in the evolution of land plants. The EST approach is an especially cost effective way to obtain partial sequences of a large number of genes from a given organism with a large genome (Brenner *et al*., 2003; Rudd, 2003; Theodorides *et al*., 2002).

An exclusive feature of OrthologID that we want to emphasize is its ability to identify diagnostic characters of orthologous gene sets using the CAOS algorithm (Sarkar *et al*., 2002). When users input a new query sequence to the OrthologID Web interface for orthology determination, OrthologID displays the diagnostic characters (responsible for classifying the query into a specific orthologous gene set) in the Tree and Diagnostics Viewer. OrthologID thus provides a new tool in the genomics toolbox that may allow researchers to rapidly identify new avenues of research; since the diagnostic characters displayed represent potential functionally important amino acid residues, which might be targeted for future structure-function studies.

The current OrthologID database includes completely sequenced genomes from plants. To increase its scope and general utility to

levels comparable with COG and INPARANOID, the OrthologID database will be expanded to include complete genomes from other phylogenetic lineages, including prokaryotes and non-plant eukaryotes. Currently, the orthologous gene sets identified in OrthologID are presented in phylogenetic tree as well as alignment formats. The existing output format is valuable since it presents the orthologous groups in the context of the evolutionary history of their respective gene families, as well as highlighting the important diagnostic characters. To allow for rapid extraction of orthologous groups, a searchable text output will be added to complement the existing tools on the OrthologID Web interface.

At the moment, since the complete genomes used in OrthologID database do not overlap with those used in INPARANOID or COG, except for the genomes of *A.thaliana* and *O.sativa*, it is premature to compare the performance of the different algorithms. More detailed comparisons on the effectiveness of these algorithms will be performed once the OrthologID database is expanded to include prokaryotic or non-plant eukaryotic complete genomes.

SUMMARY

To date, the lack of readily available automated tools has hindered the development of high-throughput, genome-scale orthology determination within a character-based parsimony framework. Using improved alignment tools and automation of rigorous traditional character-based procedures, we have developed OrthologID, a Web application tool that makes orthology determination within a character-based framework on a genomic scale possible. One unique advantage OrthologID has over existing orthology determination methods is that it identifies diagnostic characters responsible for defining orthologous gene sets, as well as diagnostic characters that are responsible for classifying a query sequence into a specific orthologous gene set. These additional data may be important in structure-function studies and may help in the elucidation of gene function.

The current OrthologID database catalogs several completely sequenced plant genomes, including unicellular *C.reinhardtii*, and the more derived angiosperms, *A.thaliana*, *O.sativa* and *P.trichocarpa*. Testing of OrthologID utilizing the current plant database shows that it performs well in assigning orthology for EST and full-length cDNA query sequences from other plant species whose genomes have not been completely sequenced, in that its output matches the results obtained using full-scale parsimony analysis. Out of the four complete genomes included in the current OrthologID database, two of them (*P.trichocarpa* and *C.reinhardtii*) are currently not included in existing orthology databases, such as COG and INPARANOID. As a result, OrthologID is able to generate orthology data and relationships that cannot be obtained in other existing databases. To improve the utility and scope of OrthologID beyond the plant kingdom, the database will be expanded to include prokaryotic and non-plant eukaryotic genomes. We anticipate that OrthologID and its database can greatly facilitate orthology determination for genome-scale phylogeny, comparative genomic studies, as well as annotation of EST and sequences from newly sequenced genomes.

ACKNOWLEDGEMENTS

We thank all the members of the New York Plant Genomics Consortium for discussion and invaluable suggestions in the development of OrthologID including D. W. Stevenson and E. D. Brenner (New York Botanical Garden), M. S. Katari and E. de la Torre (New York University), R. A. Martienssen and R. W. McCombie (Cold Spring Harbor Laboratory), and P. J. Planet (American Museum of Natural History). This study was supported by NSF Plant Genome Grant DBI-0421604 and NSF SGER Grant DBI-0326436 to G.M.C. and R.D. of the New York Plant Genomics Consortium, and by the Lewis B. and Dorothy Cullman Program for Molecular Systematics at the American Museum of Natural History.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Bardeleben,C. *et al*. (2005) A molecular phylogeny of the Canidae based on six nuclear loci. *Mol. Phylogenet. Evol*., **37**, 815–831.

Brenner,E.D. *et al*. (2003) Expressed sequence tag analysis in *Cycas*, the most primitive living seed plant. *Genome Biol*., **4**, R78.

Bruvo-Madaric,B. *et al*. (2005) Phylogeny of pholcid spiders (Araneae: Pholcidae): combined analysis using morphology and molecules. *Mol. Phylogenet. Evol*., **37**, 661–673.

Chippindale,P. and Wiens,J. (1994) Weighting, partitioning, and combining characters in phylogenetic analysis. *Syst. Biol*., **43**, 278–287.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.

Felsenstein,J. (1978) The number of evolutionary trees. *Syst. Zool*., **27**, 27–33.

Gatesy,J. and Arctander,P. (2000) Hidden morphological support for the phylogenetic placement of *Pseudoryx nghetinhensis* with bovine bovids: a combined analysis of gross anatomical evidence and DNA sequences from five genes. *Syst. Biol*., **49**, 515–538.

Gatesy,J. *et al*. (1993) Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phylogenet. Evol*., **2**, 152–157.

Gatesy,J. *et al*. (1999) Stability of cladistic relationships between Cetacea and higher-level artiodactyls taxa. *Syst. Biol*., **48**, 6–20.

Gatesy,J. *et al*. (2002) Resolution of a Supertree/Supermatrix paradox. *Syst. Biol*., **51**, 652–664.

Gatesy,J. *et al*. (2003) Combined support for wholesale taxic atavism in gavialine crocodilians. *Syst. Biol*., **52**, 403–422.

Hirsh,A.E. and Fraser,H.B. (2001) Protein dispensability and rate of evolution. *Nature*, **411**, 1046–1049.

Jordan,I.K. *et al*. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*., **12**, 962–968.

Katoh,K. *et al*. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*., **33**, 511–518.

Kluge,A.G. (1989) A concern for the evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Syst. Zool*., **38**, 7–25.

Kluge,A.G. (1997) Testability and the refutation and corroboration of cladistics hypotheses. *Cladistics*, **13**, 81–96.

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol*., **52**, 540–542.

Matthee,C.A. *et al*. (2004) A molecular supermatrix of the rabbits and hares (Leporidae) allows for the identification of five intercontinental exchanges during the Miocene. *Syst. Biol*., **53**, 433–447.

Miyamoto,M. (1985) Consensus cladograms and general classifications. *Cladistics*, **1**, 186–189.

Nixon,K.C. (1999) The Parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics*, **15**, 407–414.

Nixon,K.C. and Carpenter,J.M. (1993) On outgroups. *Cladistics*, **9**, 413–426.

Nixon,K.C. and Carpenter,J.M. (1996) On simultaneous analysis. *Cladistics*, **12**, 221–241.

Notredame,C. *et al*. (2000) T-Coffee: a novel method for multiple sequence alignments. *J. Mol. Biol*., **302**, 205–217.

O'Brien,K.P. *et al*. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*., **33**, D476–D480.

Olmstead,R. and Sweere,J. (1994) Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. *Syst. Biol*., **43**, 467–481.

Remm,M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol*., **314**, 1041–1052.

Rokas,A. and Carroll,S.B. (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy *Mol. Biol. Evol*., **22**, 1337–1344.

Rokas,A. *et al*. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 799–804.

Rudd,S. (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci*., **8**, 321–329.

Sarkar,I.N. *et al*. (2002) An automated phylogenetic key for classifying homeoboxes. *Mol. Phylogenet. Evol*., **24**, 388–399.

Smith,A.B. (1994) Rooting molecular trees: problems and strategies. *Biol. J. Linn. Soc. Lond*., **51**, 279–292.

Swofford,D.L. (2003) *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, MA.

Tatusov,R.L. *et al*. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tatusov,R.L. *et al*. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*., **28**, 33–36.

Tatusov,R.L. *et al*. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*., **29**, 22–28.

Tatusov,R.L. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Theodorides,K. *et al*. (2002) Comparison of EST libraries from seven beetle species: towards a framework for phylogenomics of the Coleoptera. *Insect Mol. Biol*., **11**, 467–475.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Wahlberg,N. *et al*. (2005) Synergistic effects of combining morphological and molecular data in resolving the phylogeny of butterflies and skippers. *Proc. Biol. Sci*., **272**, 1577–1586.

Wall,D.P. *et al*. (2003) Detecting putative orthologs. *Bioinformatics*, **19**, 1710–1711.

Wheeler,W.C. *et al*. (1995) Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Mol. Phylogenet. Evol*., **4**, 1–9.