Ecdysozoan Clade Rejected by Genome-Wide Analysis of Rare Amino Acid Replacements

Igor B. Rogozin, Yuri I. Wolf, Liran Carmel, and Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

As the number of sequenced genomes from diverse walks of life rapidly increases, phylogenetic analysis is entering a new era: reconstruction of the evolutionary history of organisms on the basis of full-scale comparison of their genomes. In addition to brute force, genome-wide analysis of alignments, rare genomic changes (RGCs) that are thought to comprise derived shared characters of individual clades are increasingly used in genome-wide phylogenetic studies. We propose a new type of RGCs designated RGC_CAMs (after Conserved Amino acids-Multiple substitutions), which are inferred using a genome-scale analysis of protein and underlying nucleotide sequence alignments. The RGC_CAM approach utilizes amino acid residues conserved in major eukaryotic lineages, with the exception of a few species comprising a putative clade, and selects for phylogenetic inference only those amino acid replacements that require 2 or 3 nucleotide substitutions, in order to reduce homoplasy. The RGC_CAM analysis was combined with a procedure for rigorous statistical testing of competing phylogenetic hypotheses. The RGC_CAM method is shown to be robust to branch length differences and taxon sampling. When applied to animal phylogeny, the RGC_CAM approach strongly supports the coelomate clade that unites chordates with arthropods as opposed to the ecdysozoan (molting animals) clade. This conclusion runs against the view of animal evolution that is currently prevailing in the evo-devo community. The final solution to the coelomate–ecdysozoa controversy will require a much larger set of complete genome sequences representing diverse animal taxa. It is expected that RGC_CAM and other RGC-based methods will be crucial for these future, definitive phylogenetic studies.

Introduction

The genomic era brought about the opportunity to expand phylogenetic analysis to the whole-genome scale, substantially increasing its resolution power. Most often, this involves construction of phylogenetic trees from concatenated alignments of numerous genes but other types of genomic markers, such as gene composition, gene order, and protein domain combinations, have been employed as well (Wolf et al. 2002; Snel et al. 2005). Analysis of rare genomic changes (RGCs) that have occurred in genomes of specific clades is often considered a particularly promising avenue of phylogenetic study (Rokas and Holland 2000; Nei and Kumar 2001; Delsuc et al. 2005; Boore 2006). The RGCs are, essentially, genomic equivalents of shared derived characters ("Hennigian" markers) that form the basis of classical cladistics (Hennig 1950; Rokas and Holland 2000; Boore 2006). Examples of RGCs include retroposon integrations, insertions and deletions (indels) of introns and large protein segments, evolutionary conserved motifs in proteins, protein domain fusions, changes in gene order, and genetic code variants (Venkatesh et al. 1999; Rokas and Holland 2000; Nei and Kumar 2001; Shedlock et al. 2004). Most RGCs represent changes caused by single (or a few) rare mutational events. In a variety of studies, RGCs have been mapped onto existing phylogenies to gain insight into their mode of evolution or, conversely, were employed to infer phylogenetic trees, typically, by using maximum parsimony (MP) (Nikaido et al. 1999; Rokas and Holland 2000; Nei and Kumar 2001; Boore 2006). The emerging consensus seems to be that RGCs often are phylogenetically informative. In cases where sequence data generate conflicting or equivocal results, RGCs offer an independent way of evaluating alternative phylogenies. Notably, RGC analysis has been recently used to propose substantial revisions of the deep branchings of evolutionary trees for both eukaryotes (Stechmann and Cavalier-Smith 2002, 2003) and prokaryotes (Iyer et al. 2004). However, systematic identification of RGCs is a major challenge.

We propose a new type of RGC (designated RGC_CAMs after Conserved Amino acid-Multiple substitutions) that are inferred by genome-scale analysis of protein sequence alignments and used them to address the coelomate–ecdysozoa controversy, a notorious open problem in animal phylogeny. The traditional, "textbook" tree topology, originally based on the data of comparative anatomy, includes a clade of animals with a true body cavity (coelomates, such as arthropods and chordates), whereas animals that have a pseudocoelom, such as nematodes, and those without a coelome, such as flatworms, occupy more basal positions in the tree (e.g., Brusca RC and Brusca GJ 1990; Raff 1996). The coelomate topology reverberates with the straightforward notions of the hierarchy of morphological and physiological complexity among the considered organisms, which is the main reason why this phylogeny had been accepted since the time of Ernst Haeckel (1866). Early molecular phylogenetic analyses of 18S rRNA supported the monophyly of the coelomates (Field et al. 1988; Turbeville et al. 1991). However, a seminal work of Lake and coworkers reported phylogenetic analysis of 18S rRNAs from a much larger set of animal species and arrived at a new tree topology that clustered arthropods and nematodes in a clade of molting animals termed the Ecdysozoa (Aguinaldo et al. 1997). The ecdysozoan topology was recovered only when certain species of nematodes, which apparently have evolved slowly, were included in the analyzed sample. On the basis of these observations, the classical coelomate topology has been reinterpreted as a case of long-branch attraction (LBA) (Aguinaldo et al. 1997; Telford and Copley 2005), one of the most pervasive artifacts of phylogenetic analysis (Felsenstein 1978; Reyes et al. 2000; Philippe, Zhou et al. 2005). The ecdysozoan scenario was supported by independent phylogenetic analysis of 18S RNA (Giribet et al.

Key words: phylogenetic analysis, cladistics, rare genomic changes, coelomata, ecdysozoa, microsporidia.

E-mail: koonin@ncbi.nlm.nih.gov.

Mol. Biol. Evol. 24(4):1080–1090. 2007
doi:10.1093/molbev/msm029
Advance Access publication February 13, 2007

2000; Peterson and Eernisse 2001), by combined analysis of 18S and 28S rRNA sequences (Mallatt and Winchell 2002), and some protein phylogenies, such as those for Hox (de Rosa et al. 1999). In addition, an argument in support of Ecdysozoa has been raised on the basis of an apparent derived shared character of this clade, a distinct, multimeric form of β-thymosin (Manuel et al. 2000).

The ecdysozoan topology gained rapid recognition and nearly unanimous acceptance in the evo-devo community thanks primarily, to the interpretation of molting as a fundamental developmental feature (Adoutte et al. 2000; Valentine and Collins 2000; Collins and Valentine 2001; Telford and Budd 2003). However, phylogenetic analyses of multiple sets of orthologous proteins seemed to turn the tables again by lending stronger support to the coelomate topology. In particular, Mushegian et al. (1998) reported phylogenetic analysis of 42 sets of probable orthologs, whereas Blair et al. (2002) analyzed ~100 orthologous nuclear proteins using several phylogenetic methods. Both studies found that a significant majority of trees supported the coelomate topology. Further phylogenetic analysis of ~500 eukaryotic orthologous groups (KOGs) of proteins (Tatusov et al. 2003; Koonin et al. 2004) in 6 eukaryotic species using a panel of phylogenetic methods showed the strongest and consistent support for the coelomate topology (Wolf et al. 2004). Blair et al. (2002) further assessed the effect of the evolutionary rate of the analyzed genes on the tree topology and found that the Coelomata hypothesis was supported even with the slowest evolving proteins, suggesting that this topology is not due to LBA. Wolf et al. (2004) also examined the potential effects of branch length effect on the tree topology and concluded that such effects could not explain the observed support of the Coelomata hypothesis. This result is compatible with the topologies of trees produced using nonsequence-based criteria, such as gene content and multidomain protein composition, suggesting a general concordance between tempo and mode in animal evolution (Wolf et al. 2004). The Coelomata hypothesis was further supported by several independent phylogenetic studies (Stuart and Berry 2004; Philip et al. 2005; Zdobnov et al. 2005; Ciccarelli et al. 2006); in addition, the status of multimeric β-thymosin as a derived shared character of Ecdysozoa has been questioned by analysis of the sequenced genomes (Telford 2004b).

The renaissance of the ecdysozoan scenario did not take long in the making. Large-scale maximum-likelihood analyses of alignments of multiple genes from an extended range of animal species (Brinkmann et al. 2005; Dopazo H and Dopazo J 2005; Philippe, Lartillot, et al. 2005), putative derived molecular characters in the form of shared orthologs and domain combinations (Copley et al. 2004), and gain and loss of introns (Roy and Gilbert 2005) concordantly provided support for the ecdysozoan topology. The coelomate topology, once again, has been proclaimed an artifact, caused primarily by LBA and related to inadequate taxon sampling (Brinkmann et al. 2005; Philippe, Lartillot, et al. 2005).

Given the multiple lines of support for each of the alternative tree topologies, the coelomate–ecdysozoa conundrum is often considered to stay unresolved and the metazoan tree is accordingly presented as a multifurcation (Hedges 2002; Telford 2004a; Jones and Blaxter 2005). Here, we show that the RGC_CAM approach unequivocally supports the coelomate clade and that this result is robust to branch length effects and taxon sampling.

Materials and Methods

Sequence Alignments

Each of the 716 protein alignments (488,157 sites altogether) constructed from selected KOGs (Tatusov et al. 2003; Koonin et al. 2004) analyzed here included orthologous genes from 8 eukaryotic species with completely sequenced genomes: *Homo sapiens* (Hs), *Caenorhabditis elegans* (Ce), *Drosophila melanogaster* (Dm), *Saccharomyces cerevisiae* (Sc), *Schizosaccharomyces pombe* (Sp), *Arabidopsis thaliana* (At), *Anopheles gambiae* (Ag), and *Plasmodium falciparum* (Pf) (Rogozin, Wolf, et al. 2003). To these KOGs, probable orthologs from 12 other eukaryotic genomes, namely, those of *Mus musculus* (Mm), *Caenorhabditis briggsae* (Cb), *Canis familiaris* (Cf), *Bos taurus* (Bt), *Encephalitozoon cuniculi* (Ec), *Oryza sativa* (Os), *Theileria parva* (Tp), *Dictyostelium discoideum* (Dd), *Cryptococcus neoformans* (Cn), *Neurospora crassa* (Nc), *Apis mellifera* (Am), and *Strongylocentrotus purpuratus* (St), were added using the COGNITOR method (Tatusov et al. 1997). Briefly, all the protein sequences from the new genomes are compared with the protein sequences previously included in the KOGs; a protein is assigned to a KOG when 2 genome-specific best hits to members of the given KOG are detected. Amino acid sequence alignments are available at the authors' Web site at ftp://ftp.ncbi.nlm.nih.gov/pub/koonin/RGC_CAM/. For most of the analyses described here, a data set of 10 species (Hs, Pf, At, Sc, Sp, Dm, Ag, Ce, Cb, and Mm) was employed; in several special cases, additional species were included as indicated in the respective tables. For the analysis of the phylogenetic position of microsporidia, the original set of 8 species (Hs, Pf, At, Sc, Sp, Dm, Ag, and Ce) was used, in order to maximize the number of genes available for analysis (considering the massive gene loss in microsporidia).

To minimize misalignment problems, only conserved, unambiguously aligned regions of the alignments were subject to further analysis. Specifically, all positions containing a deletion or insertion in at least one sequence were removed from the protein sequence alignment together with 5 adjacent positions. Starting methionines were also excluded.

A New Type of RGCs and Its Use for Statistical Testing of Phylogenetic Hypotheses

We propose a new type of RGCs that are inferred from the genome-wide analysis of protein alignments described above. The method utilizes amino acid residues that are conserved in most of the included eukaryotes, with the exception of a few (1–4) species. This is done under the assumption that any character shared by the included major eukaryotic lineages, namely, plants, animals, fungi, and Apicomplexa, is the ancestral state, whereas the deviating species possess a derived state (fig. 1). In order to reduce the

```
Hs MSLICSISNEVPEHPCVSPVS …
Mm MSLICSISNEVPEHPCVSPVS …
Dm MALVCALTNEVPETPVVSPHS …
Ag MSLVCSISNEVPEHPCISPKS …
Ce MSFVCGISGELTEDPVVSQVS …
Cb MSFVCGISGEPTEDPVVSPVS …
At M--NCAISGEVPEEPVVSKKS …
Sc M--LCAISGKVPRRPVLSPKS …
Sp M--FCSISGETPKEPVISRVS …
Pf MSIICTISGQTPEEPVIS-KT …
                  / \
Hs              AAC
Mm              AAT
Dm              AAC
Ag              AAC
Ce              GGT
Cb              GGG
At              GGC
Sc              GGG
Sp              GGA
Pf              GGC
```

FIG. 1.—An example of an RGC_CAM supporting the coelomate clade. A section of the alignments of KOG0289 (a splicing factor) is shown. The RGC_CAM position is shown in green (the amino acid conserved in most eukaryotes) and red (the replacement shared by chordates and arthropods). A fragment of the underlying nucleotide sequence alignment is shown to illustrate the 2 nucleotide substitutions that are required for the glycine to asparagine replacement in this position. Species abbreviations: *Homo sapiens* (Hs), *Caenorhabditis elegans* (Ce), *Drosophila melanogaster* (Dm), *Saccharomyces cerevisiae* (Sc), *Schizosaccharomyces pombe* (Sp), *Arabidopsis thaliana* (At), *Anopheles gambiae* (Ag), *Plasmodium falciparum* (Pf), *Mus musculus* (Mm), *Caenorhabditis briggsae* (Cb).

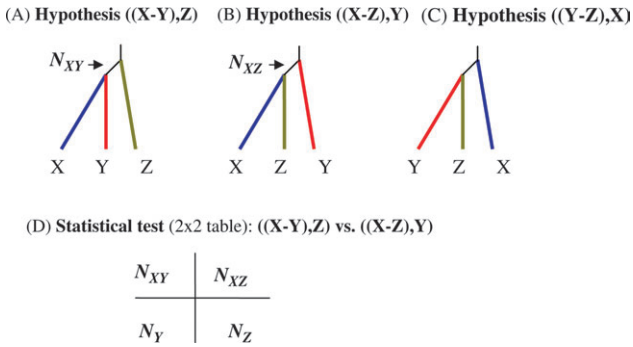

FIG. 2.—Statistical testing of phylogenetic hypotheses using the RGC_CAM analysis results. The upper part of the figure (*A–C*) shows the 3 alternative topologies of a 3-lineage rooted tree and illustrates the 3 hypotheses to be analyzed, irrespective of the specific test. N_{XY} and N_{XZ} are numbers of RGC_CAMs shared by 2 lineages (X–Y and X–Z, respectively). The lower part (*D*) specifically illustrates the Fisher-based (FB) test. N_Y and N_Z are branch lengths of lineages Y and Z (measured as the number of RGC_CAMs).

level of homoplasy (the same amino acid replacements in different lineages that do not reflect common ancestry but rather represent parallel, reverse, or convergent changes [Telford and Budd 2003]), we used only those amino acid replacements that require 2 or 3 nucleotide substitutions. Multiple substitutions are rare, so the chance to encounter homoplasy is much lower compared with amino acid changes that require single nucleotide substitutions (Averof et al. 2000; Matsuda et al. 2001; Silva and Kondrashov 2002; Kondrashov 2003). Thus, these replacements are plausible rare genomic changes (RGC_CAMs). To simplify further presentation, we use the following notation: S1 ≠ S2 = S3 means that, for a conserved amino acid position in an alignment, species S2 and S3 share the same amino acid that is different from the amino acid in the species S1. Under this notation, for example, a human RGC_CAM is denoted by Hs ≠ Mm = Pf = At = Sc = Sp = Dm = Ag = Ce = Cb, whereas an RGC_CAM shared by the 2 mammalian species is denoted by Hs = Mm ≠ Pf = At = Sc = Sp = Dm = Ag = Ce = Cb.

First, we estimated the branch length for each analyzed taxon in RGC_CAM units. For each species, we calculated the number of amino acid residues that are different from all other species (excluding relatively close species, e.g., mouse was excluded when we calculated the branch length for human: Hs ≠ Pf = At = Sc = Sp = Dm = Ag = Ce = Cb). To calculate an internal branch length (fig. 2), a pair of relatively close species was used (e.g., Dm = Ag ≠ Pf = At = Sc = Sp = Hs = Mm = Ce = Cb for insects).

The next step of the RGC_CAM analysis is statistical testing of phylogenetic hypotheses. We developed 2 tests designed to resolve ambiguous phylogenetic relationships by analyzing all possible evolutionary scenarios for 3 lineages (fig. 2*A–C*). In the first test (hereinafter FB [Fisher-based] test), the number of RGC_CAMs shared by 2 lineages (e.g., Hs = Mm = Dm = Ag ≠ Pf = At = Sc = Sp = Ce = Cb for mammals and insects—these shared RGC_CAMs are consistent with the coelomate hypothesis) was used as a variable. The values of this variable for 2 compared alternative topologies, along with the respective branch lengths (excluding the branch that is common to both alternatives), were put in a 2 × 2 contingency table (fig. 2*D*). The test is based on a null model under which, in a comparison of 2 alternative hypotheses, for example, ((X-Y),Z) versus ((X-Z),Y) in figure 2*A* and *B*, the number of RGC_CAMs that are shared by 2 lineages due to chance (N_{XY} and N_{XZ}) is proportional to the length of the branch the position of which differs between the 2 hypotheses, that is, Y and Z, respectively, in the above example. Explicitly, we employed pairwise comparisons, that is, hypothesis ((X-Y),Z) versus hypothesis ((X-Z),Y); ((X-Y),Z) versus ((Y-Z),X), and ((X-Z),Y) versus ((Y-Z),X) (fig. 2*A–C*), using the right tail Fisher exact test (fig. 2*D*). It should be emphasized that all numbers in the contingency tables are independent, that is, each RGC_CAM is counted only once. It is required that the results of the 3 tests were consistent (hereinafter consistency criterion), that is, in order to accept the hypothesis ((X-Y),Z), *P* values associated with this hypothesis should be ≤0.05 for both pairwise comparisons ((X-Y),Z) versus ((X-Z),Y) and ((X-Y),Z) versus ((Y-Z),X), whereas the *P* value associated with the ((X-Z),Y) versus ((Y-Z),X) comparison should be insignificant (>0.05).

The second test (hereinafter BB [binomial-based] test) relies on a simple probabilistic model. It is assumed that we observe a binary irreversible character with an ancestral state "0". Let P_t be the (binomial) probability of the character transitioning to state "1" in a particular site along branch *t*. Denoting by *N* the total number of sites containing a potentially irreversible character, we interpret the number of transitions observed along a branch *t*, N_t, as the number of successes in a binomial process, out of a total of *N* experiments. Let the pattern of the character at a particular site be denoted by the species where the character is in state "1", for example, XY means that a character is in state 1 in X and

Y but is in state "0" in all other species. For this test, the data must contain an out-group to the subtree XYZ such that, for certain patterns, it is possible to ascertain that the last common ancestor of X, Y, and Z was in state 0. Explicitly, the patterns X, Y, Z, XY, XZ, and YZ were counted, and their counts are denoted N_X, N_Y, N_Z, N_{XY}, N_{XZ}, and N_{YZ}, respectively. The binomial probabilities along terminal branches were approximated by $P_X=N_X/N$, $P_Y=N_Y/N$, and $P_Z=N_Z/N$. The hypothesis testing procedure is based on the obvious notion that, if the tree has a certain topology, then the existence of shared characters between nonsiblings is explained by incidental parallel transition (homoplasy). Suppose that we observe N_{XY} patterns XY out of N "experiments." Expanding all subsequent expressions only to the highest order term, the underlying binomial probability of this observation, given the topologies in figure 2*B* and *C*, is $P_X \cdot P_Y$ (second order term), whereas the probability of getting N_{XY} under the topology in figure 2*A* is P_{XY} (first order term). We then perform an exact one-sided binomial test, comparing the null hypothesis $P_{binom}=P_X \cdot P_Y$ to the alternative $P_{binom}>P_X \cdot P_Y$, and obtaining a P value P_{XY}. Rejection of the null hypothesis ($P_{XY}<0.05$) is interpreted as support for the topology in figure 2*A*. Analogous tests can be performed for N_{XZ} and N_{YZ}, obtaining the P values P_{XZ} and P_{YZ}, respectively. The topology in figure 2*A* is considered to be supported only if the binomial exact test is rejected for N_{XY} but is not rejected for both N_{XZ} and N_{YZ}.

A fundamental difficulty with the above procedure is that the number of sites that harbor irreversible characters, N, is unknown. We can only bound it from below by $N_X+N_Y+N_Z+N_{XY}+N_{XZ}+N_{YZ}$. Moreover, for a very large number of sites, $N \to \infty$, all tests necessarily reject the null hypothesis (i.e., $P_{XY}, P_{XZ}, P_{YZ} \approx 0$) as even a small number of shared characters cannot be explained by incidental parallel transition. To alleviate this problem, we compute the 3 P values as a function of N, starting from the lower bound and increasing N until all 3 P values are small enough.

For all analyses with the FB and BB tests, the same data sets were employed.

Phylogenetic Analysis of RGC_CAM Sites

Extractions from multiple alignments consisting entirely of RGC_CAM columns were additionally analyzed using traditional MP, maximum likelihood (ML), and Bayesian methods. First, identical sequences (resulting from the RGC_CAM requirement) were collapsed into a single instance. The MP topology was found using the exhaustive search routine of the PAUP* program; 1,000 bootstrap replications were analyzed using the heuristic search (tree-bisection-reconnection) routine of PAUP* (Swofford 2006). The Adachi–Hasegawa test, as implemented in the ProtML program of the MolPhy package (Adachi and Hasegawa 1992) was run with the frequency-corrected Jones-Taylor-Thornton (JTT) amino acid substitution model on the set of competing topologies. The Kishino–Hasegawa test (Kishino and Hasegawa 1989), implemented in the CODEML program of the PAML package (Yang 1997), was run with either Dayhoff or JTT amino acid substitution model with either uniform or gamma distribution of rates across sites. Bayesian topology estimates were performed using the MrBayes program (Ronquist and Huelsenbeck 2003) by running 1,000,000 Monte Carlo Markov Chain post burn-in generations with mixed amino acid substitution model and uniform distribution of rates across sites. The approximately unbiased test was performed using the Consel program with the default parameters (Shimodaira and Hasegawa 2001).

Results

The RGC_CAM Approach

We aimed at combining the abundance of information contained in numerous alignments of orthologous proteins with the main advantage of RGCs, namely, the low level of homoplasy. To this end, a 2-tier approach was employed. At the first step, positions in multiple alignments were identified that contained one amino acid in a small subset (1–4) of the analyzed species and another conserved amino acid in the rest of the species (fig. 1). Obviously, in such positions, the amino acid that is found in the smaller subset of species is a candidate derived shared character and could support the hypothesis that the species sharing this amino acid comprise a clade. However, because the contribution of homoplasy to the set of positions selected in the first step was likely to be substantial, an additional filtering step was required to identify the likely RGCs. Thus, from the initially selected positions, we chose only those that required 2 or 3 nucleotide substitutions, under the rationale that such multiple substitutions were unlikely to occur independently in different lines of descent. We designated this new class of phylogenetic characters RGC_CAM (after Conserved Amino acid-Multiple substitutions). As detailed under Materials and Methods, RGC_CAMs can be conveniently used to test alternative phylogenetic hypotheses in a statistically rigorous manner (fig. 2). The RGC_CAM approach produced reasonable results for insect (Ag and Dm) monophyly and the relationship of major fungal lineages (see Supplementary Materials online). We then applied the RGC_CAM approach to 3 well-known cases of controversial relationships among metazoan taxa: 1) the phylogeny of mammalian orders, 2) the evolutionary position of microsporidia, and 3) the Coelomata–Ecdysozoa dilemma.

RGC_CAM Analysis of Mammalian Phylogeny

The branching order of the mammalian orders is a notoriously hard problem, conceivably due to the burst-like radiation at the outset of the evolution of placental mammals (Novacek 1992, 2001). In the past, most molecular studies have supported a primate–ferungulata (artiodactyls and carnivores) clade, to the exclusion of rodents (Li et al. 1990; Arnason et al. 2000; Cao et al. 2000; Reyes et al. 2000). However, the recent analysis of RGCs, namely, retroposon insertions (Thomas et al. 2003), along with a phylogeny based on concatenation of 19 nuclear and 3 mitochondrial genes (Murphy et al. 2001), suggested a primate–rodent clade. We tested the human–mouse–cow and human–mouse–dog trifurcations using the RGC_CAM approach on concatenated alignments of 683 and 685

FIG. 3.—Amino acid variability associated with a RGC_CAM. A section of the alignment of KOG1651 (phospholipid hydroperoxide glutathione peroxidase) is shown. The RGC_CAM is in position 89; the lysine residue that is conserved in most eukaryotes is shown in green, and the variable residues in mammals are shown by different colors. The universally conserved position 90 is shown in bold. Species abbreviations: *H. sapiens* (Hs), *C. elegans* (Ce), *D. melanogaster* (Dm), *S. cerevisiae* (Sc), *S. pombe* (Sp), *A. thaliana* (At), *A. gambiae* (Ag) *and P. falciparum* (Pf), *M. musculus* (Mm), *Caenorhabditis briggsae* (Cb), Ss (*Sus scrofa*), Rn (*Rattus norvegicus*), Ca (primate *Cebus apella*), and Bt (*Bos taurus*).

genes, respectively (ftp://ftp.ncbi.nlm.nih.gov/pub/koonin/RGC_CAM/). Analysis of the human–mouse–cow trifurcation revealed only 2 RGC_CAMs, both of which supported the human–mouse clade. This support of the human–mouse clade is particularly notable given that the cow branch was extremely long (i.e., contained many apparent RGC_CAMs) compared with the human and mouse branches (the lengths of the branches were 224, 12, and 7 RGC_CAMs, respectively). This was, probably, due to the sequencing errors in the cow genome given that this is, generally, not a fast-evolving species (Murphy et al. 2001). Analysis of the human, mouse, and dog sequences revealed comparable branch lengths (13, 7, and 11 RGC_CAMs, respectively). A single RGC_CAM was shared by human and mouse, and no shared RGC_CAMs were detected for the other 2 pairs of branches. Interestingly, the shared RGC_CAM position is highly variable among mammalian species (fig. 3). Apparently, in this case, the replacement of a highly conserved amino acid was accompanied by a substantial relaxation of evolutionary constraints on this position. Such relaxations of selective constraints are a likely source of homoplasy (Telford 2002) suggesting that RGC_CAMs are not homoplasy free. Thus, a more explicit assessment of the level of homoplasy and statistical hypothesis testing are crucial for this method (see below).

RGC_CAM Analysis of the Phylogenetic Position of Microsporidia

We applied the RGC_CAM approach to a well-known case of problematic phylogeny, namely, the evolutionary position of microsporidia. Microsporidia are amitochondrial unicellular eukaryotes that have been traditionally considered an early branching lineage that diverged from the common ancestor with the rest of eukaryotes prior to the mitochondrial endosymbiosis (Vossbrinck et al. 1987; Leipe et al. 1993). However, several more recent molecular phylogenetic studies have suggested that microsporidia are evolutionarily related to fungi (Peyretaillade et al. 1998; Hirt et al. 1999; Katinka et al. 2001; Vivares et al. 2002; Williams et al. 2002; Thomarat et al. 2004; Fischer and Palmer 2005; Gill and Fast 2006). In our analysis, the raw number of shared RGC_CAMs was the largest for the microsporidia–fungi clade (supplementary table S1, Supplementary Material online). However, the microsporidian branch was extremely long which led to inconsistent results (supplementary table S1, Supplementary Material online). Specifically, the FB test yielded significant *P* values both for the basal position of microsporidia and for the microsporidia–fungi clade (supplementary table S1, Supplementary Material online). Thus, the trifurcation animals–fungi–microsporidia at present cannot be resolved using RGC_CAMs which is not surprising taking into account the exceptionally long branch leading to microsporidia (see Discussion for additional details on this problem).

RGC_CAM Analysis of the Coelomate–Ecdysozoan Conundrum

The case of mammalian phylogeny as well as the examples of RGC_CAM application (see Supplementary Material online) suggested that there are additional sources of uncertainty in the RGC_CAM analysis including sequencing errors and, potentially, population polymorphism. These problems can be alleviated by using pairs of closely related species instead of a single species. We applied this approach to the analysis of the coelomate–ecdysozoa conundrum. The set of 10 analyzed species includes *S. cerevisiae*, *S. pombe*, *A. thaliana*, *P. falciparum*, and 3 pairs of relatively close animal species (human–mouse, mosquito–Drosophila, and 2 nematodes) (694 genes). In agreement with previous findings (Aguinaldo et al. 1997), analysis of the branch lengths suggested that nematodes are a taxon with an extremely long branch which is likely to cause substantial problems for conventional phylogenetic methods (table 1; Reyes et al. 2000; Delsuc et al. 2005; Philippe, Zhou, et al. 2005). The long nematode branch notwithstanding, we observed an excess of shared RGC_CAMs in mammals and insects, in support of the coelomate topology (table 1). Statistical testing (see Materials and Methods for details) of the 3 alternative hypotheses, coelomate (C), ecdysozoa (E), and "bizarre" (B) (grouping of mammals with nematodes to the exclusion of insects) showed strong and consistent support for the coelomate hypothesis (table 1). This result was not affected by the use of more stringent conditions in terms of the sequence conservation in the alignment regions flanking the shared RGC_CAM position (table 1). Notably, the ecdysozoan and bizarre hypotheses were statistically indistinguishable.

Additional Statistical Tests for the Phylogenies Obtained with the RGC_CAM Approach

In addition to the FB-test, we applied the newly developed BB test (see Materials and Methods) and several probabilistic tests commonly used in phylogenetic studies to further assess the validity of the resolution of problematic tree topologies by the RGC_CAM approach. Each of these tests was applied to the Coelomata–Ecdysozoa problem and to the problem of the phylogenetic position of microsporidia (in order to further evaluate the resolution power of the

Table 1
The RGC_CAM Analysis of the Coelomata–Ecdysozoa Conundrum

	Mammals	Insects	Nematodes
Alignment stringency: 0 (68 KOGs with shared RGC_CAMs, 6 KOGs with conflicting RGC_CAMs)			
Branch length, number of RGC_CAMs	86	86	467
Hypothesis	C	E	B
Number of shared RGC_CAMs	34	26	16
Hypothesis testing	**C** versus E	**C** versus B	E versus B
P_{Fisher}	10^{-11}	8×10^{-15}	0.11
Alignment stringency: 1 (57 KOGs with shared RGC_CAMs, 4 KOGs with conflicting RGC_CAMs)			
Branch length, number of RGC_CAMs	73	75	379
Hypothesis	C	E	B
Number of shared RGC_CAMs	29	20	13
Hypothesis testing	**C** versus E	**C** versus B	E versus B
P_{Fisher}	3×10^{-10}	2×10^{-12}	0.2
Alignment stringency: 2 (33 KOGs with shared RGC_CAMs, 1 KOGs with conflicting RGC_CAMs)			
Branch length, number of RGC_CAMs	40	37	203
Hypothesis	C	E	B
Number of shared RGC_CAMs	18	12	5
Hypothesis testing	**C** versus E	**C** versus B	E versus B
P_{Fisher}	9×10^{-7}	7×10^{-10}	0.08

NOTE.—Alignment stringency: 0, no restrictions on the alignment regions flanking the shared RGC_CAMs; 1, at least one conserved amino acid within 5 positions upstream or downstream of each RGC_CAM position; 2, at least one conserved amino acid within 5 positions both upstream and downstream of each RGC_CAM position. Hypothesis testing: The following 3 phylogenetic hypotheses were tested as described under Materials and Methods and shown in figure 2: C, coelomate, that is, (mammals, insects) nematodes; E, ecdysozoa, that is, (insects, nematodes) mammals; B, bizarre, that is, (mammals, nematodes) insects. The hypothesis that received significant statistical support in each pairwise test is shown in bold and underlined.

RGC_CAM approach). As shown in figure 4*A*, the BB-test supported the Coelomata hypothesis for any N in the interval [715, 2201] (see Material and Methods). At $N=2201$, the Ecdysozoa hypothesis could no longer be rejected, with $P_{\text{Ecdysozoa}}$ reaching the value of 0.05. However, by then, $P_{\text{Coelomata}}$ is indistinguishable from zero. Like the FB-test, the BB-test failed to provide support for the fungal–microsporidian clade and yielded lower P values for the fungal–metazoan clade although the former topology could not be rejected for a wide range of N values; by contrast, the animal–microsporidian clade was rejected for most of the range of N (fig. 4*B*).

To further assess the validity of the resolution of problematic tree topologies by the RGC_CAM approach, we applied several standard probabilistic tests. The RGC_CAM columns were extracted from multiple alignments and analyzed using MP, ML, and Bayesian inference methods. Each of the tests provided unequivocal support for the coelomate topology over the ecdysozoa topology (table 3). By contrast, the results on the phylogenetic position of the microsporidia were ambiguous, with the MP and BI supporting the (presumably, correct) fungi–microsporidia clade, but the ML tests preferring the fungi–metazoa clade (table 3). However, in this case, with the exception of one of the ML tests, none of the methods could reject any of the topologies at a statistically significant level (table 3).

Assessment of the Robustness of the RGC_CAM Approach

The statistical tests employed here are based on the assumption that RGC_CAMs within a gene evolve independently of each other. This could be questioned under the premise of possible epistatic interactions between RGC_CAM positions. We examined the distributions of RGC_CAMs across the analyzed genes for nematodes, insects, and mammals and found no obvious indication that conserved positions in some genes are much more prone to changes compared with other genes (fig. 5). We used Monte Carlo simulations to test the hypothesis that some genes were enriched in RGC_CAMs. The RGC_CAMs were randomly shuffled across protein sequences taking into account the length of each alignment. The mean RGC_CAM density in the top 10% quantile of the distribution was used as the weight function. The weight of the observed distributions was not significantly greater than weights of the simulated distributions for all 6 animal species ($P > 0.05$; 10,000 replicates). Thus, independence of RGC_CAMs seems to be a reasonable approximation.

The analysis of the mouse–human–dog trifurcations revealed a potential source of homoplasy, that is, a replacement in a highly conserved amino acid resulting in a substantial relaxation of evolutionary constraints on the respective position (fig. 3). To address this concern, we analyzed the distribution of identical and different amino acids in pairs of relatively close species under the condition that all other species have different amino acid in this position (e.g., Hs = Mm $\neq$ Pf = At = Sc = Sp = Dm = Ag = Ce = Cb vs. Hs $\neq$ Mm $\neq$ Pf = At = Sc = Sp = Dm = Ag = Ce = Cb) (table 2). A relatively high fraction of differences was observed between Drosophila and Anopheles, which is compatible with the high rate of evolution in flies (Savard et al. 2006). By contrast, in the other 2 pairs of relatively close species, the vast majority of amino acids were conserved (table 2). These observations suggest that, although some of the shared RGC_CAMs are, probably, due to homoplasy, this is unlikely to be the major factor behind these shared characters.

The extent of homoplasy among the RGC_CAMs was further assessed by analysis of conflicting RGC_CAMs

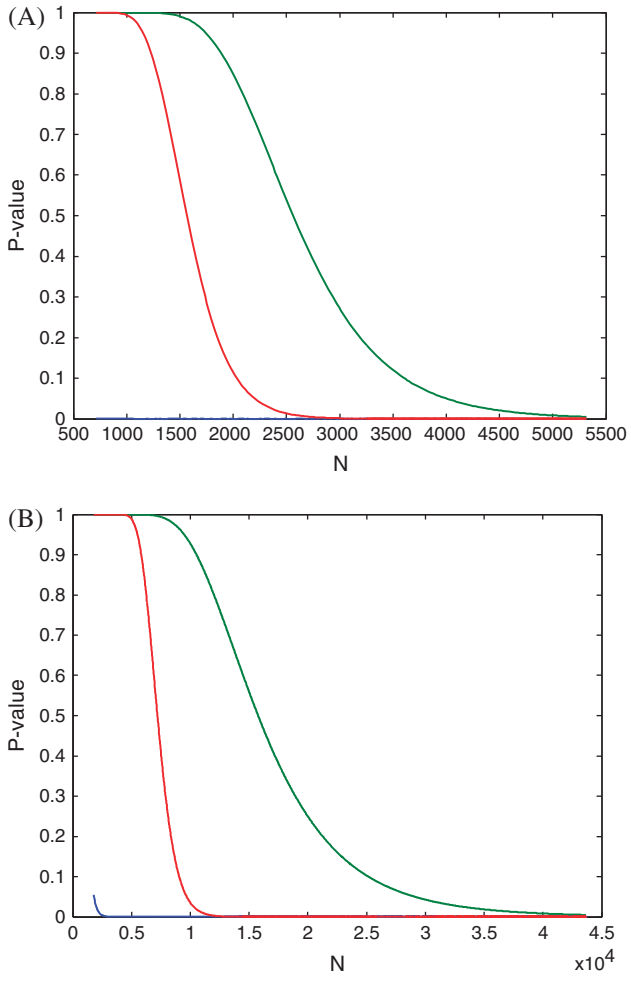

FIG. 4.—Testing phylogenetic hypotheses using the binomial-based (BB) test. (*A*) Coelomata versus Ecdysozoa. $P_{\text{Ceolomata}}$ (blue), $P_{\text{Ecdysozoa}}$ (green), and P_{Bizarre} (red) as a function of *N*. (*B*) Phylogenetic position of microsporidia. $P_{\text{A–F}}$ (blue), $P_{\text{A–Ec}}$ (green), and $P_{\text{F–Ec}}$ (red) as a function of *N*; A–F, animal–fungi, A–Ec, animals–*Encephalitozoon cuniculi* (microsporidia), F–Ec, fungi–*E. cuniculi*.

that supported alternative hypotheses in the same alignment (table 1). The genes containing such incompatible RGC_CAMs comprised ~5–10% of all genes with shared RGC_CAMs (table 1). These results suggest that, although RGC_CAMs are not homoplasy free, the level of homoplasy is not exceedingly high, and there is a strong phylogenetic signal in the whole-genome analysis of RGC_CAMs.

Table 2
Distribution of Identical and Different Amino Acids in Pairs of Relatively Close Species in Positions Where All Other Species Have a Different, Conserved Amino Acid

Pair of species	Dm–Ag	Hs–Mm	Ce–Cb
Different amino acids (YZ-X)	22	1	20
The same amino acid (YY-X)	86	86	467

NOTE.—The notation is as follows: X stands for an amino acid that is conserved in all compared species other than the given pair; YZ denotes different amino acids in the given pair; and YY denotes identical amino acids. Species abbreviations: *Homo sapiens* (Hs), *Caenorhabditis elegans* (Ce), *Drosophila melanogaster* (Dm), *Anopheles gambiae* (Ag), *Mus musculus* (Mm), and *Caenorhabditis briggsae* (Cb).

Table 3
Probabilistic Tests of Competing Tree Topologies Applied to the Set of RGC_CAM Positions

	Position of nematodes[a]		Position of microsporidia[b]	
	(N,(C,A))	(C,(N,A))	(A,(M,F))	(M,(A,F))
Maximum Parsimony	0.852	$<10^{-3}$	1.000	$<10^{-3}$
MolPhy (JTT)[c]	1.000	$<10^{-4}$	0.139	0.861
PAML (JTT, uniform)[c,d]	1.000	$<10^{-3}$	0.008	0.990
PAML (Dayhoff, uniform)[c,d]	1.000	$<10^{-3}$	0.213	0.786
Approximately unbiased test	1.000	$<10^{-3}$	0.209	0.791
MrBayes[e]	1.000	$<10^{-3}$	0.909	0.091

[a] N, Nematodes; C, Chordates, A, Arthropoda.
[b] A, Animals; M, Microsporidia, F, Fungi.
[c] RELL bootstrap probabilities.
[d] Under the assumption of a Gamma distribution of evolutionary rates, the estimation of the optimal value of the distribution parameter showed that the distribution was effectively uniform ($a \geq 99$).
[e] Frequency of the topology at equilibrium.

In the above analysis, the coelomate–ecdysozoa problem was addressed by analysis of a 10 species data set. Adding more species might increase the quality of RGC_CAMs by reducing homoplasy but this simultaneously leads to a decrease in the number of RGC_CAM sites and a substantial loss of statistical power (e.g., see supplementary tables S2 and S3, Supplementary Materials online). Nevertheless, taxon sampling is known to be important for the outcome of phylogenetic analysis and cannot be ignored (Soltis et al. 2004; Rokas and Carroll 2005). We performed taxon sampling on an extended set of 15 species (556 genes, in addition to the 10 species used to obtain the results in table 1, probable orthologs from the plant *O. sativa*, the apicomplexan *T. parva*, the social amoeba *D. discoideum*, and fungi *C. neoformans* and *N. crassa* were included). We required that at least 1 plant, 1 fungus, and 1 apicomplexan were present in a sampled set of species. With this restriction, all combinations including from 9 to 15 species (287 samples altogether) were analyzed. Only for one combination of species (the 6 animal species, *T. parva*, *A. thaliana*, and the fungi *C. neoformans* and *N. crassa*), the number of RGC_CAMs compatible with the coelomate topology was smaller than the number of RGC_CAMs compatible with the ecdysozoa topology (20 and 21, respectively), and 4 combinations produced the same number of RGC_CAMs for the coelomate and ecdysozoa topologies. For all other combinations of species (>98%), the number of RGC_CAMs supporting the coelomate hypothesis was greater than the number of RGC_CAMs supporting the ecdysozoa hypothesis. These comparisons do not take into account branch lengths; if these are considered, given the long nematode branch, there was no support for the ecdysozoan hypothesis from any of the 287 samples. Thus, the support of the coelomate hypothesis obtained with RGC_CAMs does not seem to depend on the selection of the analyzed species.

We further analyzed the effect of including additional, deep-branching species of deuterostomes and insects in the analyzed data set. To this end, the 10 species data set on which the results shown in table 1 were obtained was amended with the sequences of the probable orthologs from

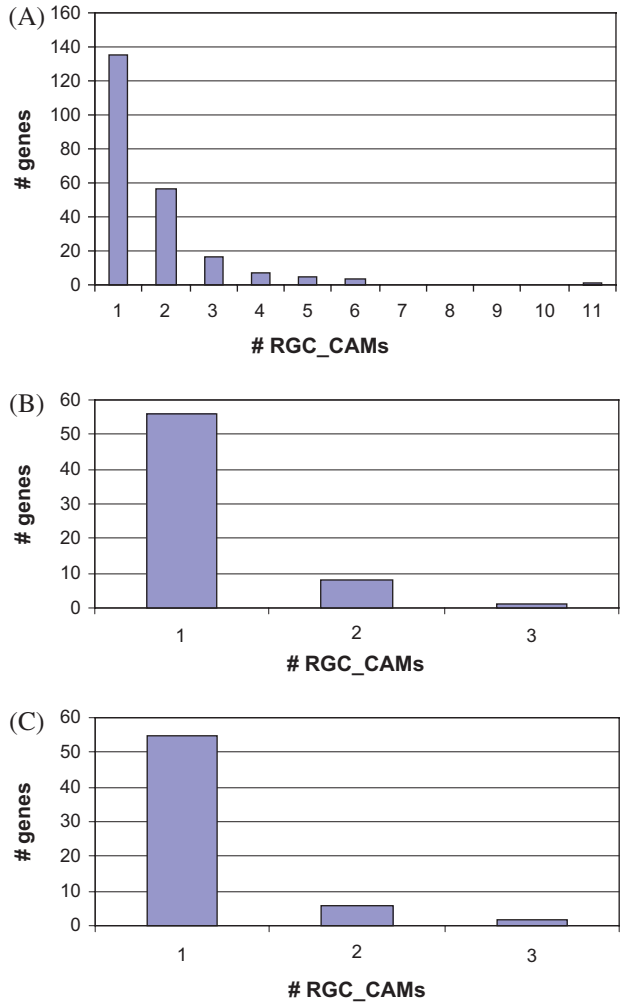

FIG. 5.—Distribution of RGC_CAMs per gene alignment for nematodes (*A*), insects (*B*), and mammals (*C*).

the honeybee *A. mellifera* and the sea urchin *S. purpuratus*. As a result, the insect and the deuterostome branches become extremely short compared with the nematode branch (table 4). Nevertheless, statistical testing of the 3 alternative hypotheses showed highly significant support for the Coelomata hypothesis (table 4). This result indicates that the RGC_CAM approach is robust to the addition of more distant in-group species to pairs of relatively close species used as representative of the analyzed clades.

Discussion

The use of whole-genome data is thought to increase the resolution of phylogenetic analyses (Wolf et al. 2002; Snel et al. 2005). However, analysis of even extremely long alignments of concatenated genes (proteins) does not necessarily eliminate artifacts because of the existing systematic biases, for example, long or short branches (Delsuc et al. 2005). The present study employs a new class of phylogenetic characters (RGC_CAMs) that were inferred using genome-wide identification of amino acid replacements that met 3 criteria: 1) were located in unambiguously aligned regions of orthologous genes, 2) were shared by 2 or more taxa in positions that contain a different, conserved amino acid in a much broader range of taxa, and 3) require 2 or 3 nucleotide substitutions. Examination of several test cases suggests that RGC_CAMs are one of the closest known approximations of irreversible phylogenetic characters (Rokas and Holland 2000) and have the potential to substantially reduce homoplasy, which is one of the major problems plaguing phylogenetic reconstructions. Our tests showed that RGC_CAMs are not free of homoplasies but we attempted to alleviate this problem by using rigorous statistical testing of competing phylogenetic hypotheses.

Table 4
The RGC_CAM Analysis of the Coelomata–Ecdysozoa Conundrum on an Extended Species Set with 2 Deep-Branching in-group Species Added

	Deuterostomes	Insects	Nematodes
Branch length, number of RGC_CAMs	7	37	361
Hypothesis	C	E	B
Number of shared RGC_CAMs	16	7	5
Hypothesis testing	**C** versus E	**C** versus B	E versus B
P_{Fisher}	$\mathbf{5 \times 10^{-18}}$	$\mathbf{6 \times 10^{-12}}$	0.068

NOTE.—The following 3 phylogenetic hypotheses were tested as described under Materials and Methods and shown in figure 2: C, coelomate, that is, (deuterostomes, insects) nematodes; E, ecdysozoa, that is, (insects, nematodes) deuterostomes; B, bizarre, that is, (deuterostomes, nematodes) insects. The hypothesis that received significant statistical support in each pairwise test is shown in bold and underlined. 26 KOGs contained shared RGC_CAMs, 1 KOG contained conflicting RGC_CAMs supporting 2 alternative hypotheses. To the 10 species included in table 1, probable orthologs from *Apis mellifera* and *Strongylocentrotus purpuratus* were added.

The RGC_CAM implementation described here is only one of a family of possible methods based on the analysis of potential RGCs derived from multiple sequence alignments. In particular, the RGC_CAMs can be defined not as sites that contain an invariant amino acid in all sequences other than those in a putative clade, but by reconstruction of ancestral states, for example, using MP. This would result in a greater number of RGC_CAMs available for analysis but also in an increased level of homoplasy. Obviously, the RGC_CAM approach remains to be optimized with regard to this inevitable trade-off.

The RGC_CAM analysis strongly supports the Coelomata topology of the animal tree over the Ecdysozoa topology; a broad variety of the applied tests, either those developed specifically for the use with this approach or standard ones and based on several different principles, were unanimous and unequivocal in preferring the Coelomata topology. Previously, the coelomate topology has received support from phylogenetic analysis of multiple families of conserved proteins (Blair et al. 2002; Wolf et al. 2004; Philip et al. 2005) and from complementary approaches such as trees based on the distribution of domain combinations (Wolf et al. 2004) and on total evidence for several highly conserved genes (Philip et al. 2005). However, it has been argued that all the evidence in support of the coelomate topology stems from one or another form of the LBA artifact caused, in part, by inadequate choice of the analyzed taxa, in particular, inclusion of only fast-evolving nematodes of the genus *Caenorhabditis*

(Aguinaldo et al. 1997; Philippe, Lartillot, et al. 2005; Telford and Copley 2005; Baurain et al. 2006). The analysis of a larger, more representative set of species appeared to support the ecdysozoan topology and to survive several tests for LBA (Philippe, Lartillot, et al. 2005). The support for the Ecdysozoa clade critically depended on the elimination from the analysis of an increasing fraction of fast-evolving genes and/or sites (Brinkmann et al. 2005; Delsuc et al. 2005; Baurain et al. 2006). This constitutes a potential problem because this procedure, by design, will favor the Ecdysozoa topology inasmuch as the Coelomata hypothesis predicts a longer nematode branch than the Ecdysozoa hypothesis. Furthermore, at least 2 recent simulations studies suggested that increasing the number of analyzed genes improves phylogenetic resolution to a much greater extent than increasing the number of species (Rosenberg and Kumar 2001; Rokas and Carroll 2005); the latter might even have an adverse effect (Rokas and Carroll 2005).

Additional support for the ecdysozoan topology has been harnessed by analysis of putative derived characters represented by shared genes and protein domain combinations (Copley et al. 2004), and shared intron positions (Roy and Gilbert 2005). The problem with these analyses, however, is that neither orthologous genes nor introns are good RGCs as massive parallel losses or gains might occur independently in different lineage, resulting in a high level of homoplasy. In particular, both nematodes and arthropods are prone to extensive loss of genes and introns (Rogozin, Babenko, et al. 2003; Rogozin, Wolf, et al. 2003; Koonin et al. 2004), an effect that might invalidate the support for the ecdysozoan topology obtained with these approaches.

The present analysis of RGC_CAMs confirmed that nematodes comprise an extremely long branch. Nevertheless, with the branch lengths explicitly taken into account, the statistical support for the coelomate topology was overwhelming. Although, because of the missing data problem, we did not have the opportunity to analyze a large number of species, taxon sampling on a 15 species data set as well as inclusion of deeper branching species of insects and deuterostomes demonstrated remarkable robustness of the support for Coelomata. Nevertheless, the level of homoplasy in the coelomate–ecdysozoa tests was considerable, with all 3 alternative hypotheses supported by at least a few shared RGC_CAMs (table 1). In part, this might be caused by the long nematode branch but it cannot be ruled out that at least some of the apparent homoplasies reflect evolutionary reality. Specifically, the different topologies could result from a duplication of multiple genes (perhaps, whole-genome duplication) predating the divergence of mammals, insects, and nematodes (Wolf et al. 2004). Under this scenario, the most strongly supported hypothesis still reflects the actual order of lineage divergence, but alternative topologies result from lineage-specific, differential loss of paralogs.

The inability of the RGC_CAM approach to reliably recover the fungal–microsporidian clade reveals limitations of this approach in resolving phylogenies that include extremely fast-evolving lineages. Nevertheless, we expect that, with further increase in the representation of sequenced genomes, it will become possible to overcome such limitations, at least, partially.

Conclusions

The method of phylogenetic analysis developed here is based on a special class of rare genomic changes, the RGC_CAMs, which are clade-specific replacements of otherwise conserved amino acids requiring 2 or 3 nucleotide substitutions. This approach to RGC selection results in a substantial reduction of homoplasy, the inevitable trade-off being that the number of RGC_CAMs is relatively small. Nevertheless, the RGC_CAM analysis showed considerable potential for solving hard phylogenetic problems provided that a large number of alignments of orthologous proteins is available for the analyzed taxa. Hence, the utility of this approach is expected to grow with the further progress of genome sequencing. The RGC_CAM method retained its resolution power even in the presence of long branches and was notably robust with respect to taxon sampling. When applied to the phylogeny of animals, the RGC_CAM approach unequivocally supported the coelomate topology over the ecdysozoan topology. Since the first report on the new topology of the animal tree that included the ecdysozoan clade (Aguinaldo et al. 1997), multiple lines of evidence have been presented in support of each of the conflicting topologies. In particular, it has been suggested that the coelomate clade is an LBA artifact caused by inadequate selection of taxa for phylogenetic analysis (Philippe, Lartillot, et al. 2005; Telford and Copley 2005). The RGC_CAM analysis seems to render this explanation unlikely. Conceivably, the final solution to the problem will be reached only when a much larger set of animal species becomes amenable to genome-wide phylogenetic analysis. It can be expected that RGC_CAMs and other types of known and new RGCs will be of major importance for these future, definitive phylogenetic studies.

Supplementary Material

Supplementary tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Masatoshi Nei, Aleksey Kondrashov, Galina Glazko, and Teresa Przytycka for useful discussions. This work was supported in part by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/Department of Health and Human Services.

Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health Intramural Research Program.

Literature Cited

Adachi J, Hasegawa M. 1992. MOLPHY: programs for molecular phylogenetics. Tokyo (Japan): Institute of Statistical Mathematics.

Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R. 2000. The new animal phylogeny: reliability and implications. Proc Natl Acad Sci USA. 97:4453–4456.

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature. 387:489–493.

Arnason U, Gullberg A, Burguete AS, Janke A. 2000. Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. Hereditas. 133:217–228.

Averof M, Rokas A, Wolfe KH, Sharp PM. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science. 287:1283–1286.

Baurain D, Brinkmann H, Philippe H. 2006. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol. 24:6–9.

Blair JE, Ikeo K, Gojobori T, Hedges SB. 2002. The evolutionary position of nematodes. BMC Evol Biol. 2:7.

Boore JL. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends Ecol Evol. 21:439–446.

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol. 54:743–757.

Brusca RC, Brusca GJ. 1990. Invertebrates. Sunderland (MA): Sinauer Associates.

Cao Y, Fujiwara M, Nikaido M, Okada N, Hasegawa M. 2000. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. Gene. 259: 149–158.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311:1283–1287.

Collins AG, Valentine JW. 2001. Defining phyla: evolutionary pathways to metazoan body plans. Evol Dev. 3:432–442.

Copley RR, Aloy P, Russell RB, Telford MJ. 2004. Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of Caenorhabditis elegans. Evol Dev. 6:164–169.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.

de Rosa R, Grenier JK, Andreeva T, Cook CE, Adoutte A, Akam M, Carroll SB, Balavoine G. 1999. Hox genes in brachiopods and priapulids and protostome evolution. Nature. 399:772–776.

Dopazo H, Dopazo J. 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biol. 6:R41.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27:401–410.

Field KG, Olsen GJ, Lane DJ, Giovannoni SJ, Ghiselin MT, Raff EC, Pace NR, Raff RA. 1988. Molecular phylogeny of the animal kingdom. Science. 239:748–753.

Fischer WM, Palmer JD. 2005. Evidence from small-subunit ribosomal RNA sequences for a fungal origin of Microsporidia. Mol Phylogenet Evol. 36:606–622.

Gill EE, Fast NM. 2006. Assessing the microsporidia-fungi relationship: combined phylogenetic analysis of eight genes. Gene. 375:103–109.

Giribet G, Distel DL, Polz M, Sterrer W, Wheeler WC. 2000. Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. Syst Biol. 49:539–562.

Haeckel E. 1866. Generelle morphologie der organismen. Berlin (Germany): G. Reimer.

Hedges SB. 2002. The origin and evolution of model organisms. Nat Rev Genet. 3:838–849.

Hennig W. 1950. Grundzuge einer Theorie der Phylogenetischen Systematik. Berlin (Germany): Deutscher Zentralverlag.

Hirt RP, Logsdon JM Jr, Healy B, Dorey MW,Doolittle WF, Embley TM. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci USA. 96:580–585.

Iyer LM, Koonin EV, Aravind L. 2004. Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. Gene. 335:73–88.

Jones M, Blaxter M. 2005. Evolutionary biology: animal roots and shoots. Nature. 434:1076–1077.

Katinka MD,Duprat S, Cornillot E, et al. (17 co-authors). 2001. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature. 414:450–453.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol. 29:170–179.

Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat. 21:12–27.

Koonin EV, Fedorova ND, Jackson JD, et al. (18 co-authors). 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. 5:R7.

Leipe DD, Gunderson JH, Nerad TA, Sogin ML. 1993. Small subunit ribosomal RNA+ of Hexamita inflata and the quest for the first branch in the eukaryotic tree. Mol Biochem Parasitol. 59:41–48.

Li WH, Gouy M, Sharp PM, O'hUigin C, Yang YW. 1990. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. Proc Natl Acad Sci USA. 87:6703–6707.

Mallatt J, Winchell CJ. 2002. Testing the new animal phylogeny: first use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes. Mol Biol Evol. 19:289–301.

Manuel M, Kruse M, Muller WE, Le Parco Y. 2000. The comparison of β-thymosin homologues among metazoa supports an arthropod-nematode clade. J Mol Evol. 51:378–381.

Matsuda T, Bebenek K, Masutani C, Rogozin IB, Hanaoka F, Kunkel TA. 2001. Error rate and specificity of human and murine DNA polymerase η. J Mol Biol. 312:335–346.

Murphy WJ, Eizirik E, O'Brien SJ, et al. (11 co-authors). 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science. 294:2348–2351.

Mushegian AR, Garey JR, Martin J, Liu LX. 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.

Nei M, Kumar S. 2001. Molecular evolution and phylogenetics. Oxford: Oxford University.

Nikaido M, Rooney AP, Okada N. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. Proc Natl Acad Sci USA. 96: 10261–10266.

Novacek MJ. 1992. Mammalian phylogeny: shaking the tree. Nature. 356:121–125.

Novacek MJ. 2001. Mammalian phylogeny: genes and supertrees. Curr Biol. 11:R573–R575.

Peterson KJ, Eernisse DJ. 2001. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol Dev. 3:170–205.

Peyretaillade E, Biderre C, Peyret P, Duffieux F, Metenier G, Gouy M, Michot B, Vivares CP. 1998. Microsporidian Encephalitozoon cuniculi, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. Nucleic Acids Res. 26: 3513–3520.

Philip GK, Creevey CJ, McInerney JO. 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. 22:1175–1184.
Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol. 22: 1246–1253.
Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol. 5:50.
Raff RA. 1996. The shape of life: genes, development, and the evolution of animal form. Chicago (IL): University of Chicago Press.
Reyes A, Pesole G, Saccone C. 2000. Long-branch attraction pheonomenon and the impact of among-site rate variation on rodent phylogeny. Gene. 259:177–187.
Rogozin IB, Babenko VN, Fedorova ND, et al. (20 co-authors). 2003. Evolution of eukaryotic gene repertoire and gene structure: discovering the unexpected dynamics of genome evolution. Cold Spring Harb Symp Quant Biol. 68:293–301.
Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol. 13:1512–1517.
Rokas A, Carroll SB. 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, Holland PW. 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol. 15:454–459.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: 1572–1574.
Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci USA. 98:10751–10756.
Roy SW, Gilbert W. 2005. Resolution of a deep animal divergence by the pattern of intron conservation. Proc Natl Acad Sci USA. 102:4403–4408.
Savard J, Tautz D, Lercher MJ. 2006. Genome-wide acceleration of protein evolution in flies (Diptera). BMC Evol Biol. 6:7.
Shedlock AM, Takahashi K, Okada N. 2004. SINEs of speciation: tracking lineages with retroposons. Trends Ecol Evol. 19: 545–553.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17:1246–1247.
Silva JC, Kondrashov AS. 2002. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. Trends Genet. 18:544–547.
Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. Annu Rev Microbiol. 59:191–209.
Soltis DE, Albert VA, Savolainen V, et al. (11 co-authors). 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci. 9:477–483.
Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science. 297:89–91.
Stechmann A, Cavalier-Smith T. 2003. The root of the eukaryote tree pinpointed. Curr Biol. 13:R665–R666.
Stuart GW, Berry MW. 2004. An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage. BMC Bioinformatics. 5:204.
Swofford D. 2006. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates, Inc.
Tatusov RL, Fedorova ND, Jackson JD, et al. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:41.
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science. 278:631–637.
Telford MJ. 2002. Cladistic analyses of molecular characters: the good, the bad and the ugly. Contrib Zool. 71:93–100.
Telford MJ. 2004a. Animal phylogeny: back to the coelomata? Curr Biol. 14:R274–R276.
Telford MJ. 2004b. The multimeric beta-thymosin found in nematodes and arthropods is not a synapomorphy of the Ecdysozoa. Evol Dev. 6:90–94.
Telford MJ, Budd GE. 2003. The place of phylogeny and cladistics in Evo-Devo research. Int J Dev Biol. 47:479–490.
Telford MJ, Copley RR. 2005. Animal phylogeny: fatal attraction. Curr Biol. 15:R296–R299.
Thomarat F, Vivares CP, Gouy M. 2004. Phylogenetic analysis of the complete genome sequence of Encephalitozoon cuniculi supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. J Mol Evol. 59:780–791.
Thomas JW, Touchman JW, Blakesley RW, et al. (71 co-authors). 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature. 424:788–793.
Turbeville JM, Pfeifer DM, Field KG, Raff RA. 1991. The phylogenetic status of arthropods, as inferred from 18S rRNA sequences. Mol Biol Evol. 8:669–686.
Valentine JW, Collins AG. 2000. The significance of moulting in Ecdysozoan evolution. Evol Dev. 2:152–156.
Venkatesh B, Ning Y, Brenner S. 1999. Late changes in spliceosomal introns define clades in vertebrate evolution. Proc Natl Acad Sci USA. 96:10267–10271.
Vivares CP, Gouy M, Thomarat F, Metenier G. 2002. Functional and evolutionary analysis of a eukaryotic parasitic genome. Curr Opin Microbiol. 5:499–505.
Vossbrinck CR, Maddox JV, Friedman S, Debrunner-Vossbrinck BA, Woese CR. 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature. 326:411–414.
Williams BA, Hirt RP, Lucocq JM, Embley TM. 2002. A mitochondrial remnant in the microsporidian Trachipleistophora hominis. Nature. 418:865–869.
Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. Trends Genet. 18:472–479.
Wolf YI, Rogozin IB, Koonin EV. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. 14:29–36.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Zdobnov EM, von Mering C, Letunic I, Bork P. 2005. Consistency of genome-based methods in measuring Metazoan evolution. FEBS Lett. 579:3355–3361.

Jianzhi Zhang, Associate Editor

Accepted February 7, 2007