Phylogenomics of Eukaryotes: Impact of Missing Data on Large Alignments

Hervé Philippe, Elizabeth A. Snell,* Eric Bapteste, Philippe Lopez,* Peter W. H. Holland,‡ and Didier Casane

*School of Animal and Microbial Sciences, The University of Reading, Reading, U.K.; †Phylogénie, Bioinformatique et Génome, Université Pierre et Marie Curie, Paris, France; ‡Department of Zoology, University of Oxford, Oxford, U.K.

Resolving the relationships between Metazoa and other eukaryotic groups as well as between metazoan phyla is central to the understanding of the origin and evolution of animals. The current view is based on limited data sets, either a single gene with many species (e.g., ribosomal RNA) or many genes but with only a few species. Because a reliable phylogenetic inference simultaneously requires numerous genes and numerous species, we assembled a very large data set containing 129 orthologous proteins (~30,000 aligned amino acid positions) for 36 eukaryotic species. Included in the alignments are data from the choanoflagellate Monosiga ovata, obtained through the sequencing of about 1,000 cDNAs. We provide conclusive support for choanoflagellates as the closest relative of animals and for fungi as the second closest. The monophyly of Plantae and chromalveolates was recovered but without strong statistical support. Within animals, in contrast to the monophyly of Coelomata observed in several recent large-scale analyses, we recovered a paraphyletic Coelamata, with nematodes and platyhelminths nested within. To include a diverse sample of organisms, data from EST projects were used for several species, resulting in a large amount of missing data in our alignment (about 25%). By using different approaches, we verify that the inferred phylogeny is not sensitive to these missing data. Therefore, this large data set provides a reliable phylogenetic framework for studying eukaryotic and animal evolution and will be easily extendable when large amounts of sequence information become available from a broader taxonomic range.

Introduction

Our understanding of the phylogenetic position of animals within eukaryotes, as well as the relationships within animals, is mainly based on ribosomal RNA analysis (Wainright et al. 1993; Aguinaldo et al. 1997; Mallatt and Winchell 2002). Over a century ago, James-Clark suggested that the choanoflagellates could be the closest living relatives of animals (James-Clark 1866). These protists have a collar of feeding tentacles surrounding a flagellum, an organization reminiscent of the feeding cells of sponges. This clade (choanoflagellates + animals) plus fungi could constitute a monophyletic group called Opisthokonta (Cavalier-Smith and Chao 1996). Within animals, it has long been accepted that sponges, cnidarians, and ctenophores derive from early branching lineages. There is no consensus on the relationships within the bilaterian (or triploblast) animals (Giribet 2002). Aguinaldo et al. (1997) recently proposed that bilaterians are divided into three major groups: deuterostomes (e.g., vertebrates and echinoderms), lophotrochozoans (e.g., molluscs and platyhelminths), and ecdysozoans (e.g., arthropods and nematodes). The latter two groups are usually, but not always, proposed to form a monophyletic group, the lophotrochozoans (Adoutte et al. 2000). Yet, phylogenies based on a single gene are generally not well resolved because of the weakness of phylogenetic signal and, more problematically, are highly sensitive to lateral gene transfers (Doolittle 1999), hidden paralogy (Page 2000), and tree reconstruction artifacts (Philippe and Laurent 1998).

Even if some single protein phylogenies seem to confirm that animals are closely related to fungi (Baldauf and Palmer 1993; Nikoh et al. 1994; Baldauf et al. 2000; Moreira et al. 2000), some others do not (Germot and Philippe 1999; Bouzat et al. 2000; Chihade et al. 2000; Loyttyoja and Milinkovitch 2001). Choanozoa, comprising choanoflagellates, Ichthyosporea, Corallochytrea, and Cristidiscoidea (Cavalier-Smith and Chao 2003), are generally considered to be the sister-group of animals to the exclusion of fungi ([King and Carroll 2001; Snell et al. 2001] but see [Ragan et al. 2003]), but statistical support is generally unsatisfactory. More generally, the knowledge of the relationships among eukaryotic groups is limited and many questions remain (Philippe et al. 2000; Simpson and Roger 2002). It is thus timely to reevaluate the origin and evolution of animals by employing more comprehensive data. Recently, choanoflagellates and animals have been clustered in a phylogeny based on eleven mitochondrial proteins (Lang et al. 2002). Some phylogenies based on multiple nuclear genes (Baldauf et al. 2000; Moreira et al. 2000) and on genome content (Korbel et al. 2002) confirmed the proximity of animals and fungi, but others do not (Veuthey and Bittar 1998; Daubin et al. 2002). Interestingly, the monophyly of ecdysozoans has never been recovered by large-scale inference (Mushegian et al. 1998; Baldauf et al. 2000; Haasdford 2000; Blair et al. 2002; Korbel et al. 2002; Hugues and Friedman 2004; Wolf et al. 2004). These incongruencies could result from the reduced amount of positions used, thereby leading to large stochastic effects; to the limited number of taxa; or to the large distance of the outgroup, leading to tree reconstruction artifacts. Although it is not clear whether increasing taxon sampling or gene sampling is the most efficient approach (Lecointre et al. 1993; 1994; Hillis 1996; Graybeal 1998; Rannala et al. 1998; Poe and Swofford 1999; Mitchell et al. 2000;
Rosenberg and Kumar 2001; Braun and Kimball 2002; Pollock et al. 2002; Zwickl and Hillis 2002), it is expected that increasing both simultaneously should improve the phylogenetic inference (Bapteste et al. 2002; Rokas et al. 2003).

There have been recent efforts to increase the size of the alignments used to infer the eukaryotic phylogeny. However, studies of eukaryotic/animal phylogenies have been limited either by the number of positions or by the number of species: 1 gene, 2,500 species and 1,500 nucleotide positions (Van de Peer et al. 2000); 4 genes, 61 species and 1,500 amino acid positions (Baldauf et al. 2000); 11 genes, 20 species and 3,000 amino acid positions (Lang et al. 2002); 100 genes, 4 species and 44,000 amino acid positions (Blair et al. 2002); and 500 genes, 6 species and >30,000 positions (Wolf et al. 2004). The inverse relationship between taxa number and sequence length results mainly from a general preference to avoid the inclusion of incomplete taxa (i.e., sequences with many missing data). In fact, the problem of missing data is often considered to be a significant obstacle in phylogenetic reconstruction (Donoghue et al. 1989; Anderson 2001; Kearney 2002; Sanderson et al. 2003). Empirical studies have shown that the use of taxa with many missing data often leads to poorly resolved trees (Wiens and Reeder 2001; Wilkinson and Benton 1995; Gao and Norell 1998). Computer simulations confirm that phylogenetic accuracy is decreased by the inclusion of highly incomplete taxa (Huelsenbeck 1991). Therefore, it is common to exclude taxa or characters that contain too many missing data.

However, we recently analyzed a very large data set of 123 proteins for a sample of 23 eukaryotic species (~25,000 unambiguously aligned positions), and the phylogenetic position of *Mastigamoeba balamuthi*, a species containing 69% of missing data, as sister-group of *Entamoeba histolytica* was robustly inferred (Bapteste et al. 2002), in agreement with other studies (Milyutina et al. 2001; Arisue et al. 2002a; Fahmi et al. 2003). The very good resolution for placing a highly incomplete taxon possibly results from numerous positions that are nevertheless available (~8,000 amino acid residues are used for *Mastigamoeba*), providing a large amount of phylogenetic signal. Indeed a recent computer simulation (Wiens 2003) suggests that it is possible to accurately place highly incomplete taxa as long as they have a sufficient number of complete characters.

Of prime importance is the validation of the hypothesis that missing data do not constitute a serious obstacle for the phylogenetic analysis in the case of very large data sets. In fact, the sequencing of a few thousand ESTs is an inexpensive method for obtaining a large amount of data for an interesting species (several thousands of positions, as in the case of *Mastigamoeba* [Bapteste et al. 2002]) at the cost that missing cells will also be prevalent in the resulting data matrix. The application of this method to tens of eukaryotic species makes it possible to assemble a very comprehensive alignment of 100 species and tens of thousands of positions, which should improve the resolution in the inference of eukaryotic phylogeny. In this study, we sequenced ~1,000 ESTs from *Monosiga ovata*, a representative of the choanoflagellates, which have a key position in the study of animal evolution. We assembled a data set of 129 orthologous proteins for 36 eukaryotic species. We verified, using our alignment and computer simulation, that missing data did not constitute a central problem in our analysis. As expected, the phylogeny inferred from our large alignment shows strong statistical support for most of the nodes (28 of 34). We discuss the possible reasons (mutational saturation, impact of protein function or of taxon sampling, among others) for the limited resolution of the six remaining nodes.

### Materials and Methods

#### Monosiga ESTs

A living culture of *M. ovata* Kent, strain M-1 (ATCC 50635), was purchased from the American Type Culture Collection (http://www.atcc.org/) and cultured in Sonneborn’s Paramecium medium following the supplier’s instructions. An oligoT-primed cDNA library was constructed from growing cells in the vector lambda ZAPII, and 1,152 expressed sequence tags (ESTs) were sequenced starting from the 5’ end. A detailed analysis of the EST sequences will be published elsewhere. The sequences have been deposited in the GenBank EST database (GenBank accession numbers CO434616-CO435569).

#### Construction of the Alignment

During our analysis of the ESTs of *Mastigamoeba* (Bapteste et al. 2002) and *Monosiga*, we performed Blast searches against SwissProt in order to find genes that are sufficiently conserved in eukaryotes to yield a meaningful alignment. We retrieved all of the homologous sequences from GenBank with the program AliBaba (P.L., unpublished data) when the Blast score between *Mastigamoeba* or *Monosiga* EST sequence and other eukaryotic sequences was lower than $10^{-10}$. This nonstringent criterion (which can sometimes retrieve paralogous genes) was used only for the screening of the data bank; for the selected genes, the cutoff was much more stringent (generally Blast e-value $<10^{-50}$) and defined in order to avoid the mixing of paralogous genes. The homologous proteins were then aligned with ClustaW (Thompson et al. 1994), leading to a data set of ~500 different protein encoding genes. A preliminary phylogenetic analysis led us to select 174 genes that could be useful for inferring the eukaryotic tree according to three criteria: (1) showing a reasonable taxonomic distribution, (2) being sufficiently conserved across all eukaryotes, and (3) being very likely orthologous.

To increase the number of eukaryotes represented, especially animals, we included nucleotide sequences obtained from ongoing EST and genome projects for 26 species. Most of the sequences were retrieved from GenBank through NCBI (http://www.ncbi.nlm.nih.gov/), except for *Candida albicans* (Stanford Genome Technology Center at http://www-sequence.stanford.edu/group/candida), *Cryptococcus neoformans* (C. neoformans cDNA Sequencing Project at http://wwwgenome.ou.edu/cneo.html); and *C. neoformans* Genome Project, Stanford Genome Technology Center and The Institute for

To maximize the number of genes by taxonomic group, chimerical sequences for a few genes were constructed by using closely related species: Schistosoma mansoni with Echinococcus multilocularis and Schistosoma japonicum, Trichinella spiralis with Trichuris muris, Brugia malayi with Onchocerca volvulus, Amblyomma americanum with Sarcoptes scabiei, Tetrahymena thermophila with Paramecium tetraurelia, Toxoplasma gondii with Neospora caninum and Sarcocystis neurona, and Theileria annulata with Babesia bovis. For two important groups, rhodophytes and stramenopiles, we combined more distantly related taxa, since the monophyly of these groups has not been questioned. By combining Porphyra yezoensis with Gracilaria gracilis and Cystodium caldarium on one hand and Phytophthora infestans with Laminaria digitata, Blastocystis hominis, and Thalassiosira weissflogii, we were able to incorporate the complete sequences of some long and potentially informative genes (e.g., chaperonin, EF2, and VATA). These combinations of distantly related taxa concern only 5 of 129 genes.

It should be noted that, despite our efforts to improve taxon sampling, we are still far from a dense taxon sampling of eukaryotes. In particular, except for choanoflagellates, we added species according to their availability (e.g., three insects and three kinetoplastids), whereas the best would have been to add species that break long branch (e.g., crustaceans or euglenoids). However, our sampling is much better (36 species) than those of other studies (e.g., crustaceans or euglenoids). However, our sampling is much better (36 species) than those of other studies using a similar number of amino acid positions (<6 species).

For adding to the alignments the orthologs of the previously mentioned 174 genes from these 26 species, we followed the approach described by Bapteste et al. (2002). The alignment was then manually refined with the ED program (Philippe 1993). In order to detect obvious contaminant sequences (e.g., host genes for parasites), which can occur in large-scale sequencing projects, we constructed a maximum parsimony (MP) tree for each protein with PAUP 4b8 (Swofford 2000). We detected exclusively mammalian contaminants for apicomplexan species and yeast contaminants for Dictyostelium, which were removed. Moreover, we verified that our alignments did not contain paralogous or xenologous genes. In fact, most of the genes we used have distant homologs (e.g., mitochondrial homologs for ribosomal proteins or duplications specific to eukaryotes for chaperonins and proteasome proteins), but it is straightforward to isolate orthologs, which are always much more similar among themselves. An additional complication came from the fact that several genes underwent more recent duplications (plants and vertebrates are well known for their complete (or almost) genome duplication). We discarded genes when the paralogous copies of a given species were not closely related (when all the gene copies of e.g., vertebrates or angiosperms, did not form a monophyletic group). For example, some protein coding genes widely used in phylogenetic inference (e.g., actin, HSP70, and tubulins) were not used here. We, therefore, believe that hidden paralogy problem should have been considerably reduced in our data set.

We selected 36 species for which a large amount of sequences were available (sequences available for at least 50 genes). From our data set of 174 genes, we retained only 129 genes according to the two following criteria: (1) the orthology between all the species was unambiguous, (2) the taxonomic sample was large (at least 26 different species of our 36 species). A summary of the species available for each gene is given in Supplementary Table S1.

Regions of unambiguous alignment were manually selected by using the program MUST (Philippe 1993). We removed almost all of the gap-containing positions, except when the insertion/deletion involved only a limited number of taxa. The limits of the unambiguously aligned blocks were fixed to the first encountered constant amino acid position (or to a very-conserved position displaying amino acids of the same functional category) preceding or following gap-containing parts. All alignments are available at http://mbe.oupjournals.org/.

Phylogenetic Analysis

We used essentially the same tree reconstruction method as in Bapteste et al. (2002). Phylogenetic trees were based on the analysis of amino acid sequences with maximum likelihood (ML), MP, and Neighbor-Joining (NJ) methods with the programs PHYML version 1.0 (Guindon and Gascuel 2003), PROTML version 2.3 (Adachi and Hasegawa 1996), and Tree-Puzzle version 5.0 (Strimmer and von Haeseler 1996), PAUP version 4d8 (Swofford 2000) and MUST version 3.0 (Philippe 1993), respectively. We first concatenated all the genes into a single alignment, which was analyzed with ML (with a JTT+F+Γ model to take into account between-site rate variation through PHYML), MP, and NJ (distances being calculated by a ML method with a JTT+F+Γ model with Tree-Puzzle). The major groups (fungi, ascomycetes, animals, nematodes, arthropods, chordates, alveolates, Apicomplexa, kinetoplastids, and green plants) were each recovered with high support; monophyly of each group was constrained in further analyses.

In subsequent analyses, we treated each protein separately (Yang 1996), instead of concatenating the sequences (see Results). Given the large size of our data set, exhaustive searches for optimal trees are not possible. We, therefore, constrained the relationships among the four animal groups (platyhelminths, nematodes, arthropods, and chordates) as a multifurcation, and computed, with PROTML, the approximate likelihood values for the 135,135 possible topologies connecting fungi, choanoflagellates, animals, amoebozoans, stramenopiles, alveolates, kinetoplastids, red algae, and green plants. Then, for
the 5,000 best topologies, we computed the exact likelihood with a JTT+F model and retained the 1,000 best topologies for more precise calculations, as follows. The likelihood value for each of the 129 genes, for each of these 1,000 topologies, was computed with a JTT+F+Γ model by using the Tree-Puzzle program. The best topology from the 1,000 was chosen according to the sum of the log likelihood for all the genes (Yang 1996). To determine the relationships between animals, we then selected the 20 best topologies and evaluated the 15 possible branching patterns between platyhelminths, nematodes, arthropods, and chordates, which led to the analysis of 300 (20*15) topologies.

To estimate the robustness of the phylogenetic inference, we used the bootstrap method (Felsenstein 1985). First, for the concatenated alignment, bootstrap values (BV) were calculated with an exact procedure: 100 replicates were generated using SEQBOOT (Felsenstein 2001), and trees were inferred by using the ML method without Γ distribution (for computing-time reasons) using PHYML. Second, we computed BVs by drawing genes with replacement (instead of positions as in the standard bootstrap) (Bapteste et al. 2002). In practice, gene BVs were computed through a modification of the RELL method (Kishino et al. 1990; Bapteste et al. 2002), which is known to overestimate BVs when the number of sampled topologies is limited (Hirt et al. 1999). However, BVs computed using gene resampling appeared to be conservative with respect to the classical site resampling (unpublished results). The gene BVs corresponding to the relationships between eukaryotic groups were computed with 10,000 replicas by using the 1,000 best topologies with a multifurcation within animals. The sample of topologies was likely sufficient, since the same BVs were obtained using the 100 best ranking topologies (data not shown). Indeed, very similar results were obtained by using only the best 20 topologies. This suggests that our BVs were not overestimated. Within animals, the bootstrap values were obtained from the 300 topologies representing all the possible relationships between the four animal clades.

Impact of Missing Data

The inclusion of a substantial amount of missing data allowed us to use more genes and more species. In principle, ML correctly handles missing data. However, missing data are often viewed as being problematic (Donoghue et al. 1989; Anderson 2001; Kearney 2002; Sanderson et al. 2003). We thus checked that missing sequences did not affect our results according to the method described in Bapteste et al. (2002). First, we partitioned the complete data set between genes displaying fewer than five missing species (73 genes) and the remaining ones (56 genes). For each partition, we computed the likelihoods of the best 1,000 topologies, and correlation coefficient was used to estimate the effect of partitioning. The significance of this correlation was then estimated by computing the same coefficient correlation on 10,000 random partitions.

Second, we introduced, in the real data set, an additional amount of 25% missing cells, yielding an alignment with 50% of question marks. Such random replacement was repeated 100 times, and phylogenetic inference was performed for each replica. The results were summarized by a majority rule consensus tree of the 100 topologies (computed with CONSENSE (Felsenstein 2001)).

Third, we performed computer simulations to estimate the effect of missing data on topological accuracy. The best tree obtained in the separate analysis was used as the model topology. Branch lengths were estimated on the concatenated sequences with a JTT+F+Γ model. Using Pseq-gen (Rambaut and Grassly 1997), we simulated 100 alignments of 30,399 positions with a JTT+Γ model. The amino acid equilibrium frequencies and the α parameter were fixed to their estimates obtained on the real data set. Because missing data were not homogeneously distributed in our alignment (e.g., Amblyomma shows ~75% and Saccharomyces 1% of missing data), we first replaced amino acids by question marks in the simulated data sets at exactly the place where missing data are present in the real alignment. Second, we randomly introduced question marks in the simulated data sets at various proportions (50%, 75%, and 90%). Phylogenetic trees were computed from all these alignments with PHYML (without Γ distribution, for computing-time reasons), and the majority-rule consensus tree was then inferred to summarize the results.

Results and Discussion

Concatenate Analysis

We assembled a large data set of 129 orthologous proteins from 36 eukaryotic species including 15 animals, hoping to gain a more robust insight into eukaryotic phylogeny. Sequences for most taxa were obtained from DNA sequence databases, thanks to several genomic and EST sequencing projects. These included representatives of several major eukaryotic lineages, with the exception of choanoflagellates, the most likely sister-group of animals. To fill this important gap, we constructed a cDNA library from the choanoflagellate M. ovata and randomly sequenced 1,152 ESTs. This approach yielded sequences for 85 of the 129 genes in our data set of orthologous proteins. The 129 proteins were aligned giving a total alignment length of approximately 90,000 residues. After removal of regions of uncertain homology, we used ~30,000 positions for the phylogenetic analyses.

The ML phylogeny inferred from the concatenation of the 129 genes is shown in figure 1. It was rooted between opisthokonts and all other eukaryotes, according to a gene-fusion event (Philippe et al. 2000; Stechmann and Cavalier-Smith 2002). The alpha parameter is 0.7, indicating that rate across site heterogeneity is important. All of the undisputed monophyletic groups were recovered with a BV of 100%, except for arthropods (95%), nematodes (98%), and alveolates (99%). Surprisingly, within insects, we obtained the topology (Diptera, (Lepidoptera, Hymenoptera)), whereas ((Diptera, Lepidoptera), Hymenoptera) is expected based on morphological data (Kristensen 1991), rRNA (Pashley et al. 1993), mitochondrial DNA (Delsuc et al. 2003) and intron insertion (Rokas et al. 1999). This result is
out of the scope of our study but deserves further investigation using an improved species sample (at least one species for each holometabolous insect order).

The relationship between the eukaryotic groups (fig. 1) was in good agreement with previous studies (Baldauf et al. 2000; Moreira et al. 2000; Bapteste et al. 2002; Lang et al. 2002). Red algae were closely related to green plants (BV 74%) and animals to fungi (BV 100%). The kinetoplastids, a group of uncertain affinity (Philippe et al. 2000; Simpson et al. 2002), were close to alveolates, but their position within this group is unstable (BV 64%), probably owing to their long unbroken branch (fig. 1). This analysis of multiple nuclear genes strongly supports the sharing of a common ancestor by choanoflagellates with animals to the exclusion of fungi and all other sampled eukaryotes (BV 100%); this is in agreement with single-gene studies (Wainright et al. 1993; King and Carroll 2001; Snell et al. 2001) and with a multi-gene mitochondrial phylogeny (Lang et al. 2002). Furthermore, our EST survey also found eighteen genes shared only by choanoflagellates and animals, including several transcription factors, actin-binding proteins, proteins implicated in vesicle transports, and seven proteins of unknown function. In contrast, only one gene (ascorbate peroxidase) was present in Monostiga and several eukaryotic groups but was absent in animals; we interpret this as a gene loss in the metazoan ancestor. From our study, we cannot exclude that choanoflagellates are included within animals, since no representative of sponges, cnidarians, and ctenophores were included. It is, however, unlikely that choanoflagellates are secondarily degenerate animals. This is suggested by the secondary compaction of the mitochondrial genome in sponges (Dennis Lavrov, personal communication), cnidarians, and bilaterians, but not choanoflagellates and diverse protists (e.g., loss of numerous genes between cox2 and atp8, and between atp6 and cox3; (King et al. 2003)).

## Separate Analysis

The analysis of each protein separately allowed branch lengths and evolutionary rates to vary between proteins (Yang 1996); therefore, one important cause of phylogenetic reconstruction artifact is reduced. Consequently, we performed a separate analysis of our 129 genes by using the method described in Bapteste et al. (2002). This separate analysis amounts to an ML reconstruction under a heterogeneous model allowing branch length and evolutionary rate to vary between proteins. To perform this analysis, the undisputed monophyletic groups were constrained and an exhaustive ML search with a JTT model allowed the selection of the 1,000 topologies of highest likelihood. For each topology, log likelihood was maximized separately for each gene with a JTT+F+Γ model and then summed to identify the topology that maximized the sum of the log likelihoods (fig. 2).

The homogeneous model, corresponding to the analysis of the concatenated sequences, is nested within the heterogeneous one, since its constraint is that branch
The lengths and the α parameter are the same for all genes. We thus compared the fit of the two models with a log likelihood ratio test by using the topology of figure 2. The separate model had a better likelihood (lnL = -2,764,511) than did the concatenate model (lnL = -2,778,840), which was as expected, since it had 79*128 = 11,392 additional free parameters. This number corresponds to 69 branches plus the alpha parameter plus the 19 equilibrium amino acid frequencies (79) multiplied by the number of genes minus one (128). Nevertheless, the separate model gave a significantly better fit to the data than did the simpler one (i.e., the concatenate approach): 2ΔlnL = 28,658 (for P = 0.01, the X² limit is 17,088), indicating that the evolutionary rates on the branches of the phylogeny were significantly different between the genes studied. However, because the hypothesis about rates is derived from the sequence data and tested using the same data, the X² test may not be reliable (Ota et al. 2000; Yoder and Yang 2000; Pupko et al. 2002), and it is preferable to use the Akaike Information Criterion (AIC) (Akaike 1973) for this model comparison. The AIC for the concatenate model (1,557,858 = -2*lnL + 2*α) is much higher than for the separate model (1,551,984 = -2*lnL + 2*α) (i.e., the concatenate approach). The separate analysis showed a better fit to the data than the concatenate analysis did, the corresponding phylogeny (fig. 2) should be preferred over the other one (fig. 1), although it should be noted that the best model does not always produce the best phylogeny (Yang 1997; Guindon and Gascuel 2002).

Nevertheless, the two topologies (figs. 1 and 2) are very similar. Apart from minor variation of the BVs (e.g., from 57% to 69% for the monophyly of protostomes), the only difference is the phylogenetic position of the stramenopiles. It has been proposed that Alveolates and stramenopiles form the chromalveolate clade (Cavalier-Smith 1998). In the concatenate analysis, stramenopiles are weakly placed as the sister group of Plantae (BV 39%). Interestingly, the monophyly of chromalveolates is recovered by the separate analysis, albeit with a limited support (BV 61%), in agreement with several recent works (Baldauf et al. 2000; Fast et al. 2001; Arisue et al. 2002b; Bapteste et al. 2002; Dacks et al. 2002).

**Choanoflagellates as a Useful Outgroup for Rooting Animal Phylogeny**

The phylogenetic tree obtained by analysis of the 129 proteins separately (fig. 2) strongly indicates that the choanoflagellates are the closest relatives of the Metazoa, among the protist groups studied here (100% bootstrap). This phylogenetic position is also recovered by the concatenated analysis (fig. 1). We conclude that the choanoflagellates is the sister group to the Metazoa. This finding is consistent with several previous analyses based on smaller data sets (Wainright et al. 1993; King and Carroll 2001; Snell et al. 2001; Lang et al. 2002). Furthermore, our analyses strongly indicate that fungi are the sister group to choanoflagellates plus animals (100% bootstrap); a clade referred to as the Opisthokonta.

One implication of these findings is that the choanoflagellates constitute a better outgroup for studying animal phylogeny than do fungi or plants, which are generally used (Blair et al. 2002; Hausdorf 2000; Hugues and Friedman 2004; Wolf et al. 2004). Not only are they more
Closely related to animals, but also we note that their genes have evolved rather slowly (see the branch lengths on fig. 1), which reduces the potential for problems caused by “random outgroup” phenomenon (Wheeler 1990). We, therefore, used the same large data set to re-investigate a few key evolutionary questions within the animals, i.e., the monophyly of protostomes and of Ecdysozoa. Within animals, the monophyly of arthropods, of deuterostomes, and of nematodes was recovered with strong statistical support in our preliminary analyses (data not shown) as well as in our unconstrained ML analysis of the concatenated sequences (fig. 1), and has been constrained in the time-consuming ML analysis shown in figure 2. In agreement with rRNA phylogeny (Aguinaldo et al. 1997), the protostomes (arthropods, nematodes, and platyhelminths) are monophyletic (BV 69%). However, our analyses did not include the acoelomorphs, which have recently been proposed as an outgroup to the rest of the Bilateria (Ruiz-Trillo et al. 2002). We found Ecdysozoa to be paraphyletic: in our analysis nematodes clustered with platyhelminths. However, we also find no support for the alternative clade of Coelomata (Hausdorf 2000; Blair et al. 2002; Wolf et al. 2004), because arthropods and chordates are not united in a clade.

The branch lengths of the tree in figure 2 provide a possible explanation for this result: the evolutionary rates of nematodes and platyhelminths are far larger than those of arthropods and deuterostomes. It seems possible, therefore, that the ecdysozoan clade has been artificially broken by the long-branch attraction (LBA) artifact (Felsenstein 1978). We tried to eschew this problem by including Trichocephalida (e.g., Trichinella), a taxon of nematodes shown to have slowly evolving rRNA sequences and previously found to support a monophyletic Ecdysozoa in rRNA analysis (Aguinaldo et al. 1997). Unfortunately, we found this nematode not to be slowly evolving for our sample of 129 protein coding genes (figure 2) as also occurs for an independent set of 6 proteins (Blair et al. 2002).

The monophyly of protostomes weakly recovered here with moderate support (i.e., the rejection of the Coelomata hypothesis) contrast with several recent large scale analyses (Mushgian et al. 1998; Baldauf et al. 2000; Hausdorf 2000; Blair et al. 2002; Korbel et al. 2002; Hugues and Friedman 2004; Wolf et al. 2004). There are several explanations for this discrepancy. First, we note that some earlier studies have retained areas of the alignment for which homology is uncertain (Blair et al. 2002), which likely introduces some noise in the phylogenetic reconstruction; we opted for a conservative approach, retaining only ~30,000 positions from a complete alignment of ~90,000. Second, in the study of Baldauf et al. (2000), hidden paralogy can be problematic, since many paralogs exist within animals for three of the four genes used (actin, and a- and b-tubulins) and finding the correct orthologs is difficult. Third, and more importantly, the species sampling used in all these studies is very poor (generally, three animals, Caenorhabditis, Drosophila, and Homo, plus Schistosoma (Hausdorf 2000), plus five additional species (Baldauf et al. 2000)). When so few taxa are used, LBA becomes a major concern (Philippe and Laurent 1998). In fact, when we reduced the number of species used in our study, we also found that arthropods and deuterostomes are grouped to the exclusion of nematodes (data not shown). We suggest that the paraphyly of protostomes observed in these studies results from the fast evolutionary rate of nematodes. Increasing species sampling above the 15 animal species used here, especially inclusion of annelids, molluscs, cnidarians, and sponges, should confirm or reject the monophyly of Ecdysozoa or Coelomata.

Ribosomal Versus Other Proteins

Because highly expressed proteins are more often sequenced in EST projects, ribosomal proteins were over-represented in our data set. There were 65 ribosomal proteins (yielding 9,159 positions, with 8.3% of missing data) and 64 nonribosomal proteins (yielding 21,240 positions, with 33.7% of missing data). It is, therefore, important to evaluate whether our phylogeny was biased by this protein sampling. As explained in Materials and Methods, we partitioned the data into two sets (ribosomal proteins and nonribosomal proteins) and computed the correlation coefficient of the lnL over the 1,000 best topologies for each partition. This was compared with the distribution obtained from 10,000 random partitions, which had the same size as the original one (9,000 vs. 21,000 positions). Although the correlation coefficient between ribosomal and nonribosomal proteins was low (r = 0.35), this partition was not significantly different from random partitions (P = 0.28). This suggests that unequal representation of different functional protein families has little influence on the phylogenetic inference.

The low value of the correlation prompted us to explore this issue in greater detail. We inferred phylogeny from these two data sets separately. For both partitions, the groups whose monophyly was constrained in the separate analysis were recovered in ML inference with a concatenate JTT+F+G model (data not shown). Except for these 25 nodes, the “ribosomal” and the “nonribosomal” phylogenies obtained with a separate JTT+F+G model were different (e.g., stramenopiles are sister groups of red algae and of alveolates, respectively), but they do not differ in any strongly supported clade, according to BVs (fig. 2). This suggests that the discrepancies are mostly stochastic and that the two protein categories are not mutually in an essential way.

Impact of Missing Data

In order to analyze simultaneously as many species and genes as possible, we chose to have an elevated level of missing data (25%). Their distribution was not homogenous among the data set (see table S1***), given that some species are almost always present (e.g., Saccharomyces, Drosophila, Cryptococcus, Arabidopsis, and Dictyostelium), and a few show >50% of missing data (Ascaris, 56%, Meliodogyne, 60%, Trichinella, 61%, and Amblyomma, 76%). By using the same method as that used for the comparison of ribosomal/nonribosomal proteins, we partitioned the complete data set between genes dis-
playing fewer than five missing species (73 genes) and the remaining ones (56 genes) and measured the effect on the likelihood of the 1,000 selected topologies. This partition had a correlation coefficient of 0.30 and was undistinguishable from random ones ($P = 0.18$), suggesting that missing data did not bias our phylogenetic inference.

To further verify that our results were not biased by the 25% of missing data, randomly chosen amino acid residues were replaced by question marks, leading to an alignment with 50% of missing data. This was analogous to a standard jackknife analysis, but, instead of removing a position for all species simultaneously, we removed a residue for a single species (several residues can be removed at the same position). Interestingly, the consensus tree obtained from 100 different replicas (fig. 3) was almost identical to the phylogeny obtained with the same method on the complete alignment (i.e., PHYML with a concatenate JTT+F model, see figure 1). The only difference was the position of nematodes/platyhelminthes (sister to arthropods/deuterostomes or to arthropods). But this node was not significantly supported. Our alignment was therefore sufficiently large (~30,000 positions and 36 species) so that the phylogenetic inference based on it was not very sensitive to a large amount of missing data (25% or 50%).

To gain more insight into the effect of missing data, we also performed computer simulations. The tree of figure 2 was used as a model for generating sequences that were the same size as our data set under a JTT+F model. The phylogeny was inferred without taking into account among-site rate variation. This choice was mainly guided by computing-time issues, but was also made to be closer to the bona fide conditions (i.e., real sequences did never evolve under the model used to infer the phylogeny). When the simulated data sets were used without introducing missing cells, the recovered phylogeny was the same as the model tree for 100 replicates. This was expected, since the number of positions was very large (30,399). Interestingly, when missing data were introduced exactly as in the real data set (i.e., heterogeneously, see table S1), the model phylogeny was again inferred in 100% of the replicates. This strongly suggests that a high level of missing data (25%), even if unevenly distributed, does not disturb phylogenetic inference. Indeed, this result was not surprising, since the alignments used corresponded to about 22,500 positions without missing data. This constituted a large amount of information to determine the phylogeny, which likely explained why the correct topology was more often recovered here when compared to other simulation studies (Hasegawa and Fujiwara 1993; Kuhner and Felsenstein 1994; Takahashi and Nei 2000; Rosenberg and Kumar 2001; Wiens 2003).

Because the correct phylogeny was always recovered with 25% of missing data, we incorporated much higher levels of missing data. The question marks were randomly inserted into the alignment to a level of 50% and 90%. With a level of 50%, the correct phylogeny was again recovered in all the 100 cases. Because the data set remained quite large (15,000 equivalent positions for 36 species), a large quantity of phylogenetic information remained. More surprisingly, with 90% of missing data, the inferred phylogeny remained similar to the model. The majority-rule consensus tree of the 100 simulations is shown in figure 4. Most of the nodes (26 of 33) were found >80% of the time, and only two (corresponding to the
shortest branches of the model tree) were recovered in ~60% of the cases. This high efficiency in recovering nodes was striking because, on average, each position of the alignment had only 3.6 characters that were not a question mark (14,525 positions were expected to have <3 determined characters and thus to be noninformative). This suggests that, through networking pieces of the phylogenetic signal dispersed over all the 30,000 positions, the ML tree reconstruction method conveyed global and congruent information.

We performed additional simulations with the same quantity of information (i.e., 3,039 positions without any missing data) in order to compare the efficiency of phylogenetic signal recovery from complete or highly patchy alignment. Except for one node that was recovered in only 98 of 100 simulations, all the nodes were always recovered. As expected, for the same amount of sequenced amino acid residues, the phylogenetic inference is slightly more efficient when the alignment does not contain missing data.

These simulation studies demonstrate that missing data do not constitute a serious limitation to phylogenetic inference as long as the quantity of information is globally sufficient (e.g., several thousand of aligned positions). This is in full agreement with the very recent work of Wiens (2003), who concluded that “the reduced accuracy associated with including incomplete taxa is caused by these taxa bearing too few complete characters rather than too many missing data cells.” In our case, all the species were represented by sequences containing >10,000 amino acid residues, a very large amount. Therefore, the phylogeny of eukaryotes shown on figure 2 should not be biased by the 25% of missing data.

Why 30,000 Positions Did Not Yield a Fully Resolved Tree?

Surprisingly, despite the use of 129 genes, the relationships between five principal eukaryotic lineages (stramenopiles, alveolates, Euglenozoa, red algae, and green plants) and between the four animal phyla examined here (arthropods, deuterostomes, nematodes, and platyhelmints) were not well resolved. However, it should be emphasized that most of the nodes (27 of 33) were highly supported (BV ~100%), were recovered even when a third of the alignment was randomly removed (fig. 3), and were congruent with previous phylogenies (based on morphology, ultrastructure, biochemistry, or molecular data). Nevertheless, one has to wonder why such a massive data set gave such poor results in two parts of the tree.

There are several explanations for this limited resolution. First, multiple substitutions at the same position are expected to be frequent because the speciations at the base of eukaryotes occurred several hundreds of million years ago. The mutational saturation was estimated as previously described (Philippe and Forterre 1999) and appeared to be important (fig. 5). We found saturation to be of the same order of magnitude as for other markers within eukaryotes such as rRNA, tubulins, elongation factors, and actin (Philippe and Adoutte 1998; Roger et al. 1999), but less important than when the three domains of life are compared (Philippe and Forterre 1999). This level of saturation will certainly reduce the resolving power, but it should affect node recovery as a function of its depth. This cannot explain the lack of resolution at both the base of animals and of eukaryotes, since the base of opisthokonts (lying between these two nodes) was well resolved.

Another explanation for limited resolution could be that the various proteins have strong but incongruent signals. These incongruencies would be the result of tree reconstruction artifacts (such as LBA, biased amino acid composition, or covarion-structure), or the incorrect orthology assignment (hidden paralogy or horizontal gene transfer). To investigate this, we inferred phylogeny for each gene individually using PHYML and JTT+F model. As expected, because of the limited number of positions available (a mean of 235), these phylogenies were poorly resolved. They did not strongly support nodes that were not present in the concatenate analysis (data not shown).
suggesting that we indeed only retained orthologous sequences. There is a single exception, the pyruvate kinase, which strongly supports the grouping of kinetoplastids with opisthokonts (more precisely with fungi). This could result from hidden paralogy or horizontal gene transfer, but did not affect our inference (fig. 2), since the very same results were obtained when this gene is discarded. In summary, we suggest that a weak but consistent phylogenetic signal is present in these various genes (Budin and Philippe 1998; Dacks et al. 2002).

Finally, a straightforward explanation of this limited resolution is that speciation events leading to the sampled taxa in these areas were much more closely spaced than elsewhere (Philippe et al. 1994). In fact, it has been proposed that both animal (Conway Morris 2000) and eukaryotic (Knoll 1992; Philippe and Adoutte 1998) diversification, occurred relatively rapidly. However, these hypotheses need to be more precisely defined, especially with a quantification of the time intervals.

Conclusion

Our analysis of a large data set (129 proteins) strongly supported that animals are closely related to choanoflagellates and slightly more distantly related to fungi. These results highlight the pivotal importance of choanoflagellate studies to the understanding of the origin of animal multicellularity (Brooke and Holland 2003). Furthermore, assuming that the eukaryotic phylogeny can be rooted with opisthokonts (Philippe et al. 2000; Stechmann and Cavalier-Smith 2002), support, albeit non conclusive, is provided for the monophyly of two super-ensembles of eukaryotic phyla, Plantae and Chromalveolata. Nonetheless, we find that the resolving power of molecular phylogeny is still limited for some clades, despite the use of ~30,000 homologous amino acid positions. In addition, long-branch attraction is, as expected, a problem (e.g., nematodes or kinetoplastids), although we argue that this can be overcome with a larger species sampling. Until now, it has not been possible to combine dense species sampling with the assembly and analysis of a large, multi-protein data set. Because DNA sequencing costs are dropping, this limitation is less acute, and, as shown here, computational methods are available to extract phylogenetic information from such a large amount of data. It is important to note that our computer simulations have shown that an important level of missing data (e.g., 25%) constitutes a minor problem for the analysis of a large data set. This finding is especially valuable, since creating a large alignment without including the missing data is quite difficult: most, if not all, of the genes can be absent for a few species (because the data were lost or because the gene is not yet sequenced) or cannot be used for some species (because of paralogy or xenology). We suggest that application of high-throughput EST sequencing could be readily applied to a wide range of animals and eukaryotes, yielding data that could be easily combined with the data set assembled for the present study. This strategy should provide a cost-effective route to refining further our view of the eukaryotic evolution.

Acknowledgment

This paper is dedicated to the memory of André Adoutte.

We thank Stephane Guindon for his help with the PHYML software and Henner Brinkmann, Franz Lang, Nicolas Lartillot, David Moreira, Naiara Rodriguez-Ezepeleta, and two anonymous referees for critical comments on the manuscript. PWHH and EAS acknowledge support from the BBSRC. We acknowledge the contributions of genome and cDNA projects that have generated some sequences used in these analyses. Candida albicans sequences were generated by the Stanford Genome Technology Center with the support of the NIDR and the Burroughs Wellcome Fund; Cryptococcus neoformans data from the C. neoformans Sequencing Project, NIH-NIAID grant number AI147079, and Bruce A. Roe, Doris Kupfer, Heather Bell, Sun So, Yuong Tang, Jennifer Lewis, Sola Yu, Kent Buchanan, Dave Dyer, and Juneann Murphy supported by NIH-NIAID grant number AI147079; C. neoformans genome data courtesy of the Stanford Genome Technology Center, funded by the NIAID/NIH under cooperative agreement AI47087, and The Institute for Genomic Research, funded by the NIAID/NIH under cooperative agreement U01 AI48594; Cryptosporidium parvum data courtesy of the C. parvum Genome Sequencing Project, Virginia Commonwealth University/Tufts University School of Veterinary Medicine, directed by Gregory A. Buck and Giovanni Widmer, funded by NIAID (AI46418); Dictyostelium discoideum data courtesy of the D. discoideum Genome Project at the Institute of Biochemistry I, Cologne and the Department of Genome Analysis, IMB, Jena, supported by the Deutsche Forschungsgemeinschaft (No 113/10-1 and 10-2); Neurospora crassa data courtesy
of the \textit{N. crassa} Genome Project, Whitehead Institute Center for Genome Research.

\section*{Literature Cited}


Felsenstein, J. 2001. PHYLLIP (Phylogenetic Inference Package). Distributed by the author, Department of Genetics, University of Washington, Seattle.


