CHAPTER 10

Reconstructing the ancestral eukaryote: lessons from the past

Mary J. O'Connell and James O. McInerney

10.1 History

The time will come I believe, . . . , when we shall have fairly true genealogical trees of each great kingdom of nature. (Darwin, 1859)

There is increasing interest in reconstructing the tree of life in its complete sense (prokaryotic and eukaryotic). Charles Darwin envisioned understanding the relationships of the eukaryotic kingdoms of animal and plant life in a single large tree. Understanding the evolution of the great diversity of life is a major goal in biology. However, despite decades of extensive studies to resolve the structure of the deepest branches of the Eukaryota tree, a generally accepted phylogenetic reconstruction has not been achieved (Gouy and Li, 1989; Aguinaldo *et al*., 1997; Baldauf *et al*., 2000; Blair *et al*., 2002).

The previous few chapters have addressed methodological concerns in reconstructing ancestral sequences. Few sources of error are potentially greater than the use of an incorrect multiple sequence alignment or phylogenetic tree. Phylogenetic trees from multiple genes are expected to converge on a species-level relationship. Here, some cases where they do not are presented together with the potential implications for ancestral sequence reconstruction.

In this chapter we shall address two major issues relating to the resolution of relationships in the eukaryotic division. The first is the relationships among the three-crown Eukaryota (fungi, plants, and animals), the second is the interphylum relationships at the root of the animal kingdom itself. The analysis of nucleotide or amino acid substitution patterns has provided the majority of the evidential support for phylogenetic hypotheses to date and is held as the most informative (Woese, 1987). However, the use of sequence data to resolve the branching order of the eukaryotic root has lead to the discovery of three conflicting topologies.

The existence of a separate and distinct fungal kingdom has long been established; however, its exact phylogenetic position in relation to the other eukaryotic kingdoms continues to pose a problem. Traditionally fungi were considered to be more closely related to plants than to animals, but this view was not supported by solid evidence until relatively recently. There are three possible topologies describing the relationship between the plants, animals, and fungi: the traditional grouping is fungi with plants, the second topology consists of animals with fungi (Cavalier-Smith, 1987a), and the third topology consists of animals with plants.

There is also a controversy regarding the deepest branches within the animal kingdom—so much so that Telford (2004a) has dubbed it " . . . an awkward subject for phylogeny". Despite extensive phylogenetic studies, the most likely reconstruction of the earliest animal ancestor remains unresolved. The traditional metazoan topology based on comparative anatomy of tissue organization includes a clade of animals with a true body cavity (coelomates, such as arthropods and chordates), whereas animals that have a pseudocoelome, such as the nematode, occupy a more basal position in the tree and animals with no body cavity such as platyhelminths occupy the most basal position (Mader, 1993). According to this comparative developmental approach to animal

evolution, nematodes are basal and vertebrates and arthropods are more closely related. This hypothesis of relationships, known as the Coelomata hypothesis, dominated animal systematics for more than 50 years. More recently however, the use of molecular sequence data to reconstruct the root of the animal phylogeny has led to the proposition of an alternative topology which describes a clade of molting animals termed the Ecdysozoa (Aguinaldo *et al*., 1997). The Ecdysozoa hypothesis postulates that all phyla composed of animals that grow by shedding a cuticular exoskeleton originate from a common ancestor, thus forming a distinct molting clade. This clade consists of arthropods and nematodes to the exclusion of vertebrates and contradicts the classical Coelomata grouping, which places the arthropods and vertebrates together to the exclusion of the nematodes (Hyman, 1940).

Reconstructing ancient events and hypothesizing about the nature of extinct genomes is reliant on sampling and the ability of current methods to interpret the data. There are no methods of analysis that have been developed that can be shown to be free from any kind of potentially misleading artefact. In addition, at the time of writing, we have a painfully inadequate sampling of genes and genomes across the eukaryotes. The latter has resulted in analyses that have either focused on completed genomes from small numbers of taxa or larger taxon sampling for small numbers of genes. This situation will improve over time with the sequencing of completed genomes or the availability of libraries of expressed sequence tags from a greater phylogenetic distribution.

10.2 Methodological developments

Molecular sequence data have been applied in various different forms to the problem of reconstructing the ancestral eukaryote. These include analyses of raw sequence data, gene content, insertion/deletion events, structures and domains, rRNA secondary structures, and intron analyses.

10.2.1 Data concatenation

Assuming that ortholog identification is flawless there are serious concerns over the approach of concatenating data (Phillips *et al*., 2004). Data concatenation is carried out to overcome topological differences that are due to stochastic effects or in order to add up small signals. It is expected that signals are additive and will point in the same direction whereas noise is dispersive and random. Data concatenation is usually carried out using genes that do not conflict with one another—although there is a school of thought that says that you should always concatenate. The major philosophical argument in favor of data concatenation is to remove stochastic error. As individual genes from a group of taxa may have histories that differ from those of direct descent (e.g. duplication and loss), conflicting phylogenies are often produced. A way of overcoming the conflict between individual genes in a data-set is to concatenate these genes into a single alignment. As alignments tend towards infinity we expect values to go from 90 to 100%. The result of applying this approach to real data from yeast species in the past has been the production of a single phylogeny with 100% bootstrap support on all branches (Rokas *et al*., 2003). However, testing this same data-set for systematic biases Phillips *et al* (2004) found conflicting phylogenies using slightly different phylogenetic approaches (in one instance maximum likelihood and maximum parsimony, and in the second minimum evolution), but interestingly both of the resultant conflicting phylogenies had 100% bootstrap support (Phillips *et al*., 2004). By recoding the nucleotides accordingly as either purines or pyrimidines, the original phylogeny was recovered, illustrating that compositional bias existed in the yeast data-set (Phillips *et al*., 2004). It is evident from this analysis that whereas data concatenation has the desired effect of reducing sampling effects, it is unlikely that it will allow the analyst to determine whether the tree is correct or not. Concatenating data together results in higher bootstrap values, but these are only meaningful if there is consistency in models used to analyse and generate the data (Phillips *et al*., 2004). Bootstrapping is only a method of assessing sampling effects and will only report on whether we would expect this signal given a longer alignment analysed in the same way; that is not to say that the signal is not homoplastic in the first place.

10.2.2 Combined protein data-sets conserved across all domains of life

An approach that has been designed to reconstruct the tree of life rather than the eukaryotic clade specifically is that of conserved protein analysis (Brown *et al*., 2001). The idea is to use the most ancient proteins to reconstruct the most ancient branches. On comparing open reading frames from complete (or almost complete) Bacteria, Archea and Eukaryota, 23 orthologous proteins were found to be present across all domains using a sample size of 45 species. To verify that these were single gene orthologs, homology searches and individual gene phylogenies were constructed. In some cases there were both cytoplasmic and mitochondrial copies present in eukaryote species: only the cytoplasmic copies were retained for analysis as they represent the more closely related copy. Poorly conserved regions of alignment were removed, leaving 6591 positions, which were concatenated. Using a number of phylogenetic reconstruction methods including maximum-likelihood quartet puzzling, maximum parsimony, minimum evolution, and neighbor joining, support was found for the Coelomata clade rather than the Ecdysozoa (Brown *et al*., 2001), with support values of 100, 100, and 96% for maximum parsimony, neighbor joining, and quartet puzzling respectively.

10.2.3 Introns

Besides the use of protein-coding sequences to reconstruct the phylogeny of the animals, other characters such as the pattern of spliceosomal intron conservation have been employed. Introns have a very slow rate of insertion and loss, with intron-turnover estimates ranging from around 10^{-9}/year for flies and worms to 10^{-11}/year for mammals (Lynch and Richardson, 2002; Roy *et al*., 2003). A high proportion of introns should therefore persist for very long periods, giving them a desirable slow rate of evolution. It seemed improbable that an intron, once lost, would be regained in exactly the same position; this gave the added benefit of irreversibility to this approach. Possibly the most significant advantage to using introns is that they were believed to be immune to rate variation between branches (Roy and Gilbert, 2005).

A method for analyzing the pattern of shared intron positions for an unresolved tree consisting of molecular data (from complete genomes) for arthropods, nematodes, and deuterostomes, and a plant outgroup, was developed (Roy and Gilbert, 2005). Using 684 identified eukaryotic orthologs and measuring the pattern of intron conservation across all species, support was found for the Ecdysozoa hypothesis with a significance score of $P < 10^{-6}$. The Coelomata grouping received no more support in this analysis than the universally rejected grouping of nematodes with vertebrates. The method described in Roy and Gilbert (2005) takes into account variation in rates of intron loss in a specific lineage but does not incorporate possible differences in rates of loss between introns within a single lineage. The Dollo parsimony approach used in combination with intron data should place species with similarly high or low rates of character loss together to resolve a highly supported and uncontroversial phylogeny. This is not the case with intron data (Wolf *et al*., 2004), suggesting that this is an unsuitable data source. It is conceivable that intron-position data are suitable but that the Dollo parsimony treatment of this new data form (though suitable for nucleotide or protein sequences) is not appropriate. It was necessary to test these data independently; such analysis has shown that the same intron has been lost independently on multiple branches of the *Caenorhabditis* clade (Coghlan and Wolfe, 2004). In the genomes of *Caenorhabditis elegans* and *Caenorhabditis briggsae* there are more than 6000 introns that are lineage-specific, with a very high rate of intron turnover (at least 0.005 intron gains or losses on average over a gene per million years; Coghlan and Wolfe, 2004). Also, intron–exon structure is not retained even within closely related nematode species. Introns therefore, are unlikely to be valid phylogenetic characters (Cho *et al*., 2004).

10.2.4 Analysis of protein domains

Rather than comparing aligned sequences it is possible to adopt a molecular cladistic approach by examining the presence or absence of specific domains, any shared/derived higher-order character

would therefore be indicative of a close relationship. Using this tactic it is possible to hypothesize unique or shared patterns/combinations of protein domains. Applying this approach to the Metazoa and the unicellular choanoflagellates, the pattern of domains that are present clusters these groups of organisms into a single clade (King and Carroll, 2001). A derivative of this approach has also been applied to bacterial phylogeny reconstruction, where the presence of specific gene orthologs and families are used as the heritable characters (House and Fitz-Gibbon, 2002). This approach held promise for the Coelomata/Ecdysozoa topology. To date the largest data-set assembled in this fashion consisted of 1712 orthologous genes and 2906 protein domains from completed metazoan genomes, the result of which was greater support for the Coelomata (Copley *et al*., 2004). However, it is known that the nematode lineage has a higher rate of character loss than the arthropods and vertebrates. An increased rate of secondary loss in one taxon produces the same effect on phylogeny reconstruction as rapidly evolving species, and therefore is subject to long-branch attraction (LBA; see section 10.3). An attempt to correct these data for LBA was proposed (Copley *et al*., 2004) that calculates a coefficient of secondary loss for the nematodes, arthropods, and vertebrates. This is done by examining characters that are present in the yeast outgroup and in at least one of the animal ingroups. For each ingroup the tendency to lose characters is calculated as a ratio of the number of characters lost in that lineage compared to the total number of characters that existed in the metazoan common ancestor. On correcting these data for LBA in this way, the opposing phylogeny (Ecdysozoa) is supported, therefore highlighting the importance of LBA and also showing that protein-domain combination data and orthologous genes, regardless of the large data-sets available, are not exempt from this issue.

10.2.5 rRNA secondary structure

The secondary structure for RNAs transcribed from rRNA is complex, driven by the base pairing between regions of the rRNA molecule, and forming the well-known structure of loops and stems. The selective pressure to retain the secondary structure of this molecule results in different evolutionary rates in the stem and loop regions, with stem mutations having a different probability of fixation than their equivalent mutations in loops. Not all mutations between base pairs are equally likely. A novel method has been developed that uses a 16-state model rather than a simpler four-state single-nucleotide model. This model takes the single substitution rate, double substitution rate, double transversion rate, and the substitutions between paired and mismatched states into account for calculating the phylogeny (Telford *et al*., 2005). Using this more sophisticated approach, small-subunit rRNA data from bilateria were tested. The resultant topology supported the ecdysozoan hypothesis (Telford *et al*., 2005).

10.2.6 Insertion/deletion/fusion events

Possibly one of the most important observations supporting the Ecdysozoa hypothesis is the presence of a mulitmeric form of the β-thymosin gene in the genomes of *Drosophila melanogaster* and *C. elegans*, whereas other metazoans used in the analysis showed the presence of a monomeric form (Manuel *et al*., 2000). This gene was therefore taken to be a molecular synapomorphy that consists of a change from the primitive monomeric character to the derived multimeric form. This represents a rare event and was provided as a convincing line of evidence linking the arthropods and nematodes (Manuel *et al*., 2000). However, the multimeric form was also found to be present in a deuterosome, *Ciona intestinalis*, and a lophotrochozoan, *Hemissenda crassicornis*, and also exists outside the Metazoa in a fungus, thereby demoting the multimeric form from being an ecdysozoan-specific state (Telford, 2004b). The β-thymosin gene is therefore not a valid character to show support for the ecdysozoan clade, and it is advisable to take a more comprehensive taxon sample size when considering molecular synapomorphies.

The identification of a rare genomic event, such as the insertion of a 12-amino acid sequence in a primarily highly conserved region of the EF-1α protein, seemed to add serious weight to the

argument for the grouping of the animal and fungal clades (Baldauf and Palmer, 1993). Later studies showed that this insertion, while conserved in position and strikingly conserved among fungi, varies extensively in both length and sequence (Baldauf, 1999) and is not present at all in some platyhelminths (Littlewood *et al*., 2001). The EF-1α protein is also present in multiple copies in most genomes, and in the case of some flatworms phylogenetic reconstruction using this protein does not produce uncontroversial monophyletic groups (Littlewood *et al*., 2001).

10.3 Methodological biases

10.3.1 Gene sampling and taxon sampling

The choice to use large data-sets or indeed completed genomes in phylogenetic reconstructions is generally made at the expense of taxon sampling. The number of characters and taxa required for accurate phylogenetic reconstruction is debatable (Rokas and Carroll, 2005). Clearly the sample of taxa used had a profound effect on the analyses that lead to the reconstruction of the moulting clade. More slowly evolving nematode species produce the ecdysozoa grouping while faster-evolving species produce the Coelomata relationship. A number of studies suggested that gene or character sampling may have an even greater effect on phylogeny reconstruction than taxon sampling (Mitchell *et al*., 2000; Rosenberg and Kumar, 2001). This finding spurred on a plethora of phylogenetic studies using large amounts of sequence data supporting the coelomata hypothesis (Mushegian *et al*., 1998; Blair *et al*., 2002). Complete eukaryotic genomes and clusters of orthologous groups of proteins permitted large-scale analysis of over 500 eukaryotic orthologous genes (known as KOGs) using a variety of phylogenetic methods, and support was found for the Coelomata hypothesis (Koonin *et al*., 2004). Using complete genomes of 11 eukaryotic species, homologous sequences derived from 18 human chromosomes (25 000 amino acid sequences). Following adjustment for unequal evolutionary rates among lineages, the Coelomata grouping was favoured using distance, maximum parimony, and Bayesian phylogeny-reconstruction methods (Dopazo *et al*., 2004). This study highlighted the large number of exons/characters required to reliably reconstruct the animal phylogeny, stating that those analyses supporting the Ecdysozoa hypothesis did not reach the sample-size requirement.

Over the following years, the analyses became larger, with data-sets growing from 100 (Blair *et al*., 2002) to 500 genes (Wolf *et al*., 2004) and with the most recent boasting more than 800 genes (Philip *et al*., 2005). These large-scale gene analyses support, without exception, the grouping of coelomate arthropods and vertebrates to the exclusion of the pseudocoelomate nematodes. So it is clear that smaller numbers of genes from a wide taxon sampling support the Ecdysozoa and large scale or genome wide analyses from a small number of taxa support the Coelomata.

We might be tempted at this stage to retire the debate, giving victory over to the Coelomate followers, and dismissing the Ecdysozoa claim as a result of poor gene sample size. It seems that this might be a little premature, however, as recent studies suggest that the phylogenies from large-scale analyses might be the result of LBA (Felsenstein, 1978).

One of the first molecular studies into the structure of the three kingdoms of euakryotic life involved sequence data from both large- and small-subunit rRNA, 10 isoaccepetor tRNA families, and six highly conserved proteins from all three kingdoms (Gouy and Li, 1989). Applying a transformed distance method and a maximum-parsimony method, using these three distinct data-sets, a single phylogenetic tree was obtained that placed the fungi at the base, with plant and animal kingdoms as closest neighbors. This was the first sequence analysis to provide statistically robust reconstructions of the base of the eukaryotic tree.

This traditional phylogeny has been challenged with data from ultrastructural characters such as the presence of a uniflagellate reproductive stage grouping the fungi with the animals (Cavalier-Smith, 1987b). These ultrastructural characters are not unique or consistent synapomorphies for animals and fungi; the uniflagellate condition and flattened mitochondrial cristae have been detected outside of the opisthokonts (Steenkamp

et al., 2006), calling for the need for molecular synapomorphies. Sequence data used include studies of four protein-encoding genes (α-tubulin, β-tubulin, EF-1α, and enolase; Baldauf and Palmer, 1993; Keeling and Doolittle, 1996). Support for a topology that grouped animals and fungi together to the exclusion of plants was found, although the support varied considerably depending on the protein. On closer analysis of the tubulin data, it was found that these proteins are members of highly paralogous families, with α-tubulin for example consisting of 23 members. In addition the enolase protein has a history of duplication, loss, and horizontal gene transfer, therefore making identification of true orthologs for these proteins difficult (Harper and Keeling, 2004). The use of these characters for phylogeny reconstruction is therefore dubious.

As the analysis of paralogs is difficult due to the lack of sequence data or indeed complete genomes a different approach is to completely disgard multigene families and focus instead on only single gene families from completed genomes. This way we are sure that we are comparing like with like (Blair *et al*., 2002; Philip *et al*., 2005). This method of single-gene ortholog identification and analysis ensures that hidden paralogy is minimized. The conservative approach for single-gene ortholog identification involves a two-tier process (see Figure 10.1). Supertree construction, consensus-tree construction, and total evidence methods have all been applied to the data (Creevey and McInerney, 2005; Philip *et al*., 2005).

10.3.2 Running out of steam

When considering the analysis of molecular data it is necessary for any given data-set to ask whether or not it is capable of reconstructing the phylogeny, a key facet of this being the number of

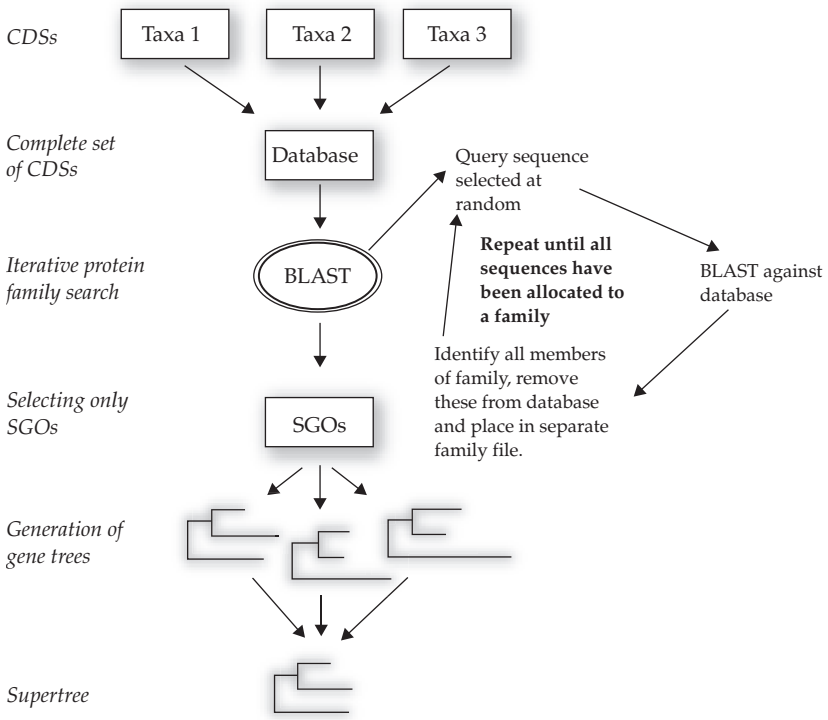

Figure 10.1 Schematic showing the steps involved in identifying single-gene orthologous protein families using coding DNA sequences (CDSs) from completed (or other) genomes. This method is believed to remove many of the biases that can be encountered in generating data-sets for the reconstruction of ancestral relationships. SGO, single-gene ortholog.

characters required to reliably reconstruct a phylogeny. Analysis has shown that molecular data run quickly out of steam when the process of evolution is rapid at the edges of the tree. It is estimated that in these circumstances a polynomial number of samples is required for the phylogeny to be resolved (Mossel and Steel, 2004).

10.3.3 Markov chain Monte Carlo (MCMC) methods of phylogenetic reconstruction and data concatenation

MCMC algorithms play a critical role in Bayesian inference of phylogeny. The rate of convergence of many of the widely used Markov chains has been tested. The practical application of this is that Bayesian MCMC methods can be misleading when the data are generated from a mixture of trees. Thus, in cases of data containing conflicting/potentially conflicting phylogenetic signals, phylogenetic reconstruction should be performed separately on each signal (Mossel and Vigoda, 2005).

10.3.4 LBA

Site-stripping has been applied to analyse sites with different mutation rates in the data-set to reduce the effect of LBA (Philip *et al*., 2005). Site-stripping is a method of dividing the data-set into different categories of site depending on mutation rate. The purpose of treating the data in this way is to reduce the effect of LBA. In the case of Philip *et al*. (2005), eight different rate categories were defined (see Figure 10.2). The progress of the phylogeny was then followed as the faster-evolving sites are methodologically removed and using different combinations of these rate categories the emerging phylogeny with absence of (or at least reduced) LBA is generated (see Figure 10.2a and b; this method has also been applied to

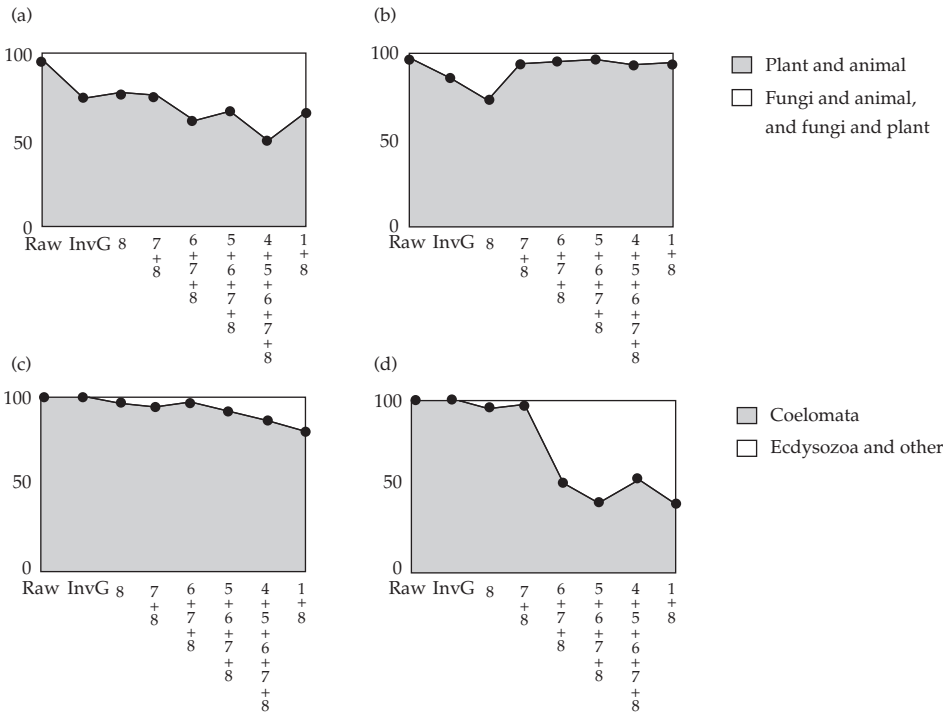

Figure 10.2 (a, b) Data from completed plant, animal, and fungal genomes. (c, d) Data from completed animal genomes. (b, d) Data controlled for sequence length and ability to recover uncontroversial parts of the tree. The *x* axis indicates rate categories removed from data, going from the most slowly to the most rapidly evolving. Raw, represents the data in its original form and InvG indicates data analysed using an invariable gamma model of rate variation. Modified from Philip *et al*. (2005).

the Coelomata/Ecdysozoa data-set and the results are shown in Figure 10.2c and d). A number of statistical and randomization tests were carried out on the data which consisted of 1452 aligned positions and in no case using this taxon selection was support found for the Opisthokont grouping; instead the animals were grouped with the plants; see Figure 10.2a and b (Philip *et al*., 2005). The issue still remains, however, that we do not have a large-enough taxa sample size. Indeed, the need for complete genome sequences of lower plants, animals, and protists is evident from all the literature to date. The inclusion of data from sister protista has the effect of producing the Opisthokonta grouping (Steenkamp *et al*., 2006).

It is most interesting that in a recent analysis of 500 genes, the least support for the Coelomata group comes from the maximum-likelihood analyses (Wolf *et al*., 2004). Maximum-likelihood methods are expected to be the most robust when dealing with rate heterogeneity, and therefore we would expect that, if the phylogenetic signal is strong enough, the use of maximum-likelihood methods would retrieve the most probable tree. The largest data-set applied includes 780 single-gene orthologous protein families (representing some 436 450 amino acid positions) and the site-stripping method described above (see Figure 10.2) and in the vast majority of analyses (24 out of 26) the Coelomata topology was favored over the Ecdysozoa (Philip *et al*., 2005). This analysis, while sensitively and thoroughly examining available eukaryotic complete genomes, has a limitation in that the available completed genomes are biased towards the higher eukaryotes. A major concern therefore is that the phylogeny supported by this and other large-scale analyses is not due to true phylogenetic signal but rather the rapid evolution of the nematode, causing the nematode to be dragged to the root of the phylogeny. It has been shown that multiple gene analyses

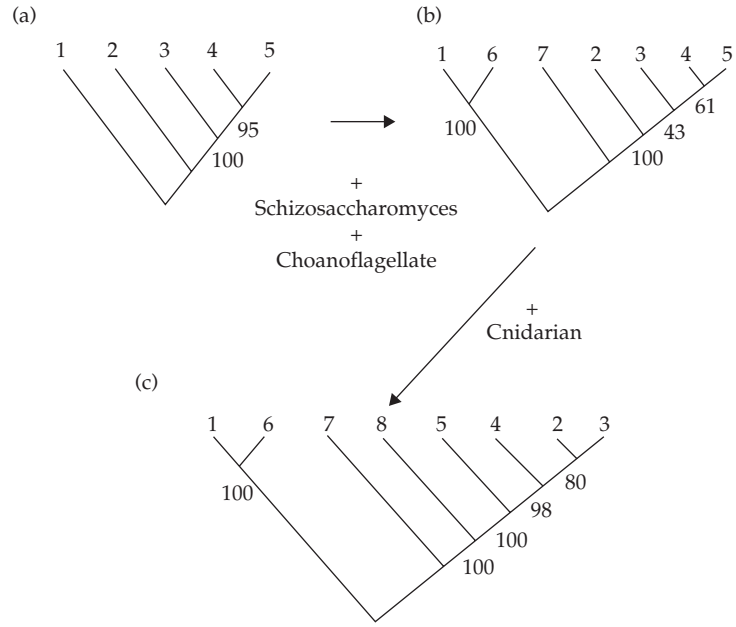

Figure 10.3 This figure shows the importance of taxa sampling on the resulting animal phylogeny. The numbers (1–8) correspond to the following species respectively: *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Schistosoma mansoni*, *Drosophila melanogaster*, *Homo sapiens*, *Schizosaccharomyces cerevisiae*, *Monosiga brevicolis*, *Hydra* sp. Using distant outgroups such as yeast in the analysis draws the nematode towards the base of the tree, Coelomata grouping (a). The addition of a second yeast species (*Schizosaccharomyces* sp.) and the choanoflagellate (*Monosiga* sp.) results in a decrease in the support for the grouping of *Drosophila* sp., *H. sapiens*, and *C. elegans* (b). The addition of the Cnidarian (*Hydra* sp.) causes a rearrangement of the species, placing the nematodes as a sister taxa to *Schistosoma* sp. with very high confidence/support (c). Modified from Philippe *et al*. (2005).

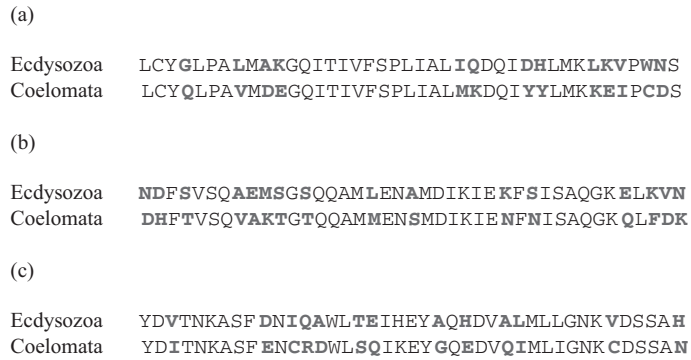

Figure 10.4 Ancestral reconstructions of each of the following proteins from the KOG database were generated using the PAML software package for the Ecdysozoan ancestor (top row of each alignment) and for the Coelomata ancestor (bottom row of each alignment). (a) ATP-dependent DNA helicase (KOG0352), positions 210–250 of the reconstructed amino acid sequences; (b) eIF2-interacting protein ABC50 (KOG0066), positions 330–370 of the reconstructed amino acid sequence; and (c) GTPase Rab26/Rab37, small G-protein superfamily (KOG0083); positions 550–590 of the amino acid sequence.

reinforce LBA if it is present in the data (Phillips *et al*., 2004).

To address the issue of LBA, Wolf *et al*. (2004) assumed that Ecdysozoa was in fact the correct phylogeny (null hypothesis). Then, by differing the degree of branch-length inequality, they determined the frequency of recovering the Coelomata. By simulating a large number of sequences evolving according to the Ecdysozoan phylogeny, giving the nematode ever-increasing branch lengths relative to the fly and vertebrate lineages, it was found that with longer nematode branch lengths the Coelomata tree was recovered. However, for any given nematode branch length, and comparing the simulated and real data-sets, it was found that the frequency of support for Coelomata was significantly lower. Therefore these results show that the real data-set contains phylogenetic signal that is not fully due to LBA (Telford, 2004a).

It is clear that a larger sampling of taxa is needed, as we have too few completed genomes. The analysis of 129 orthologous proteins from 35 animal species (many of which are expressed sequence tags), including the use of an early-branching and slowly evolving animal (*Hydra magnipapillata*) rather than a fungal species such as yeast to root the phylogeny, was an important advance (Philippe *et al*., 2005). By increasing the number of taxa sampled and removing rapidly evolving species and sites the Ecdysozoa phylogeny is supported most strongly. It is highly probable that the use of very-long-branched species such as yeast to root the phylogeny has attracted nematodes to the base of the phylogeny in previous analyses (see Figure 10.3; taken from Philippe *et al*., 2005).

What are the consequences of the two alternative phylogenies with regard to the ancestral sequences that they infer? We analysed three KOG protein families using PAML and produced an ancestral sequence for each based on both the Coelomata phylogeny and the Ecdysozoa phylogeny independently (see Figure 10.4). The resultant ancestral sequences were similar at many positions; this is reflected in the pairwise distances between the Ecdysozoa and Coelomata reconstructed ancestors for genes a, b and c (relating to panels of Figure 10.4) being 0.221, 0.243, and 0.105 respectively. However, there are a number of positions which differ between the two reconstructed ancestral proteins, a sample of which is shown in Figure 10.4.

In examining controversial issues in eukaryote phylogeny reconstruction, it must be borne in mind that the reasons why these issues have remained controversial is that it has been possible to recover contradictory signals. The reasons why these signals exist have been discussed and include simple stochastic error and sampling,

systematic biases, and the selection of genes for analysis that are not orthologs.

References

Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., and Lake, J.A. (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**: 489–493.

Baldauf, S.L. (1999) A search for the origins of animals and fungi: comparing and combining molecular data. *Am. Nat.* **154**: S178–S188.

Baldauf, S.L. and Palmer, J.D. (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. USA* **90**: 11558–11562.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**: 972–927.

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

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., and Stanhope, M.J. (2001) Universal trees based on large combined protein sequence data sets. *Nat. Genet.* **28**: 281–285.

Cavalier-Smith, T. (1987a) The simultaneous symbiotic origin of mitochondria, chloroplasts, and microbodies. *Ann. NY Acad. Sci.* **503**: 55–71.

Cavalier-Smith, T. (1987b) *The Origin of Fungi and Pseudofungi*. Cambridge University Press, Cambridge.

Cho, S., Jin, S., Cohen, A., and Ellis, R.E. (2004) A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Res.* **14**: 1207–1220.

Coghlan, A. and Wolfe, K.H. (2004) Origins of recently gained introns in Caenorhabditis. *Proc. Natl. Acad. Sci. USA* **101**: 11362–11367.

Copley, R.R., Aloy, P., Russell, R.B. and Telford, M.J. (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.

Creevey, C.J. and McInerney, J.O. (2005) Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* **21**: 390–392.

Darwin, C. (1859) *On the Origin of the Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

Dopazo, H., Santoyo, J., and Dopazo, J. (2004) Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. *Bioinformatics* **20** (suppl. 1): I116–I121.

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

Gouy, M. and Li, W.H. (1989) Molecular phylogeny of the kingdoms Animalia, Plantae, and Fungi. *Mol. Biol. Evol.* **6**: 109–122.

Harper, J.T. and Keeling, P.J. (2004) Lateral gene transfer and the complex distribution of insertions in eukaryotic enolase. *Gene* **340**: 227–235.

House, C.H. and Fitz-Gibbon, S.T. (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J. Mol. Evol.* **54**: 539–547.

Hyman, L.H. (1940) *The Invertebrates*. McGraw-Hill, New York.

Keeling, P.J. and Doolittle, W.F. (1996) Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Mol. Biol. Evol.* **13**: 1297–1305.

King, N. and Carroll, S.B. (2001) A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. *Proc. Natl. Acad. Sci. USA* **98**: 15032–15037.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S. *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**: R7.

Littlewood, D.T., Olson, P.D., Telford, M.J., Herniou, E. A., and Riutort, M. (2001) Elongation factor 1-alpha sequences alone do not assist in resolving the position of the acoela within the metazoa. *Mol. Biol. Evol.* 437–442.

Lynch, M. and Richardson, A.O. (2002) The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.* **12**: 701–710.

Mader, S.S. (1993) *Biology*. Wm. C. Brown, Dubuque.

Manuel, M., Kruse, M., Muller, W.E., and Le Parco, Y. (2000) The comparison of beta-thymosin homologues among metazoa supports an arthropod-nematode clade. *J. Mol. Evol.* **51**: 378–381.

Mitchell, A., Mitter, C., and Regier, J.C. (2000) More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera). *Syst. Biol.* **49**: 202–224.

Mossel, E. and Steel, M. (2004) A phase transition for a random cluster model on phylogenetic trees. *Math. Biosci.* **187**: 189–203.

Mossel, E. and Vigoda, E. (2005) Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* **309**: 2207–2209.

Mushegian, A.R., Garey, J.R., Martin, J., and Liu, L.X. (1998) Large-scale taxonomic profiling of eukaryotic

model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res.* **8**: 590–598.

Philip, G.K., Creevey, C.J., and McInerney, J.O. (2005) The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. *Mol. Biol. Evol.* **22**: 1175–1184.

Philippe, H., Lartillot, N., and Brinkmann, H. (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.* **22**: 1246–1253.

Phillips, M.J., Delsuc, F., and Penny, D. (2004) Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **21**: 1455–1458.

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

Rokas, A., Williams, B.L., King, N., and Carroll, S.B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**: 798–804.

Rosenberg, M.S. and Kumar, S. (2001) Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well. *Mol. Biol. Evol.* **18**: 1823–1827.

Roy, S.W. and Gilbert, W. (2005) Resolution of a deep animal divergence by the pattern of intron conservation. *Proc. Natl. Acad. Sci. USA* **102**: 4403–4408.

Roy, S.W., Fedorov, A., and Gilbert, W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. USA* **100**: 7158–7162.

Steenkamp, E.T., Wright, J., and Baldauf, S.L. (2006) The protistan origins of animals and fungi. *Mol. Biol. Evol.* **23**: 93–106.

Telford, M.J. (2004a) Animal phylogeny: back to the coelomata? *Curr. Biol.* **14**: R274–R276.

Telford, M.J. (2004b) The multimeric beta-thymosin found in nematodes and arthropods is not a synapomorphy of the Ecdysozoa. *Evol. Dev.* **6**: 90–94.

Telford, M.J., Wise, M.J., and Gowri-Shankar, V. (2005) Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the bilateria. *Mol. Biol. Evol.* **22**: 1129–1136.

Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev.* **51**: 221–271.

Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res.* **14**: 29–36.