Journal of Systematics and Evolution 46 (3): 274–286 (2008) doi: 10.3724/SP.J.1002.2008.08038
(formerly Acta Phytotaxonomica Sinica) http://www.plantsystematics.com

Animal phylogeny and large-scale sequencing: progress and pitfalls

Henner BRINKMANN* Hervé PHILIPPE**

(*Centre Robert Cedergren, Département de Biochimie, Université de Montréal,* 2900 Boulevard Edouard-Montpetit, Montréal, Québec, H3T1J4, Canada)

Abstract Phylogenomics, the inference of phylogenetic trees using genome-scale data, is becoming the rule for resolving difficult parts of the tree of life. Its promise resides in the large amount of information available, which should eliminate stochastic error. However, systematic error, which is due to limitations of reconstruction methods, is becoming more apparent. We will illustrate, using animal phylogeny as a case study, the three most efficient approaches to avoid the pitfalls of phylogenomics: (1) using a dense taxon sampling, (2) using probabilistic methods with complex models of sequence evolution that more accurately detect multiple substitutions, and (3) removing the fastest evolving part of the data (e.g., species and positions). The analysis of a dataset of 55 animal species and 102 proteins (25712 amino acid positions) shows that standard site-homogeneous model inference is sensitive to long-branch attraction artifact, whereas the site-heterogeneous CAT model is less so. The latter model correctly locates three very fast evolving species, the appendicularian tunicate *Oikopleura*, the acoel *Convoluta* and the myxozoan *Buddenbrockia*. Overall, the resulting tree is in excellent agreement with the new animal phylogeny, confirming that "simple" organisms like platyhelminths and nematodes are not necessarily of basal emergence. This further emphasizes the importance of secondary simplification in animals, and for organismal evolution in general.

Key words long-branch attraction (LBA) artifact, new animal phylogeny, phylogenomics, random error, systematic error.

The classical view of animal relationships was essentially based on morphological considerations, and was strongly influenced by the assumption of an evolution towards an ever increasing complexity (Brusca & Brusca, 1990). It generally focused on the evolution of internal body cavities (coeloms) and resulted in a phylogeny grouping mollusks, annelids, arthropods and deuterostomes to the exclusion of nematodes and platyhelminths (the Coelomata hypothesis). Nevertheless, after some decades, it was becoming apparent that morphological studies alone could not reliably resolve relationships among the major groups of animals. Great hope was therefore placed in the newly arising field of molecular phylogeny, and its dominant marker 18S-like rDNA. In contrast to the classical view, the new animal phylogeny, initially based on 18S rDNA (Aguinaldo et al., 1997; Halanych et al., 1995), proposed two major bilaterian groups, deuterostomes and protostomes, and the subsequent division of protostomes into Ecdysozoa and Lophotrochozoa (Adoutte et al., 2000). The new taxonomy led to an intricate mixture of complex and rather simple organisms, rendering the Coelomata concept meaningless and implying that several organisms must have become secondarily simplified over evolutionary time (e.g., loss of coelom).

Small SubUnit (SSU) rDNA, although an excellent phylogenetic marker, contains only ~1,000 alignable positions, many of which are constant. Its resolving power is too limited to solidly infer all animal relationships, because of the preponderance of stochastic (or sampling) errors. An appealing solution lay in using alternative, possibly longer, markers (essentially proteins). The field was first dominated by rather easily amplifiable mitochondrial genes like the 12S and 16S rDNA and protein encoding genes like *cox*1, *cyt*b, or *nad*2. Later, genes encoding highly conserved nuclear proteins, like the translation elongation factors 1 alpha (EF-1α) (Kobayashi et al., 1996; McHugh, 1997) and 2 (EF-2) (Regier & Shultz, 2001), the Na/K ATPase (Anderson et al., 2004), the largest subunit of RNA Polymerase II (Sidow & Thomas, 1994), RAG1 (Groth & Barrowclough, 1999) and RAG2 (Sullivan et al., 2000) were partially sequenced. Although a similar lack of resolution was observed, results were generally congruent with the rRNA tree (Halanych, 2004). Some conflicts exist among these markers, but these can generally be explained as stochastic errors. However, incongruences can sometimes be well supported because they are due to other reasons: (i) hidden paralogy (primarily

Received: 26 March 2008 Accepted: May 2008
* E-mail: <Henner.Brinkmann@umontreal.ca>.
** Author for correspondence. E-mail: Herve.Philippe@umontreal.ca; Tel.: 1-514-343-6720.

in protein markers, but also in rDNA), (ii) horizontal gene transfer (HGT), (iii) incomplete lineage sorting, and (iv) systematic errors (tree reconstruction artifacts).

A possible way to simultaneously improve resolving power and solve the incongruence issue, is to use multiple genes, especially in form of a concatenation (Kluge, 1989). Concatenation of two or more genes/gene fragments soon became the standard approach (Shultz & Regier, 2000), although this only slightly increases the resolving power. With progress in sequencing technology, it became possible to consider complete genome information (especially mitochondrial genomes) or a large part of it (for nuclear genomes). This approach was called phylogenomics and we will now discuss in detail its strengths and weaknesses.

1 Stochastic and systematic errors in phylogenomics

The use of a large number of genes (more than 100) provides more primary information, thereby greatly reducing sampling errors. As a result, statistical support is greatly increased, compared with studies based on single or few genes, often leading to trees with a hundred percent bootstrap support (Blair et al., 2002; Wolf et al., 2004). However, this unprecedented level of resolving power (Delsuc et al., 2005) is unfortunately accompanied by a greatly enhanced sensitivity to systematic errors (Philippe et al., 2005a).

Systematic errors correspond to cases in which tree reconstruction methods will recover an erroneous tree more and more frequently, the more data that are analyzed. Methods are then said to be inconsistent. The most famous example is the inconsistency of maximum parsimony when taxa evolve at a heterogeneous rate, the long-branch attraction artifact (Felsenstein, 1978). Any systematic feature of the sequences, such as compositional bias, will steadily accumulate with the addition of more data and can eventually become dominant. The consequence will be strongly supported tree reconstruction artifacts, e.g., grouping of species with similar G+C content (Phillips et al., 2004). The strong support for artifactual nodes induced by systematic errors in phylogenomics contrasts with random errors in single gene phylogenies, which are usually not strongly supported, since they are not accumulative.

Probabilistic methods (either maximum likelihood (ML) or Bayesian inference (BI)) are consistent, meaning that they recover the correct solution, provided that the sequences have evolved according to the process assumed by the model. Therefore, in a probabilistic framework, tree reconstruction artifacts can only occur if there is a violation of the underlying model of sequence evolution by the data. It should nevertheless be noted that probabilistic methods are rather robust to model violations (Sullivan & Swofford, 2001).

Model violations are always present, because even a complex model cannot capture all the subtleties of the real world. A good example is the compositional heterogeneity of the sequences contained in an alignment, such as species with a high (or low) GC content. For example, a strong GC-pressure in a given organism leads to a preferential mutation of A and T towards G and C. This biased mutation process will lead to an increased fixation of A or T to G or C substitutions, especially at third codon positions. If another unrelated organism is under the same influence, the probability is high that these two organisms independently acquired the same character state at many homologous positions (homoplasy). However, if a model assumes that the evolutionary process is stationary, i.e., is the same in all parts of the tree, it will predict many fewer nucleotide positions with the same character state in these two unrelated species than observed and the probabilistic methods will erroneously infer that the two GC-rich taxa are closely related. This illustrates how model violations translate into tree reconstruction artifacts, when the model incorrectly interprets multiple substitutions (i.e., the predictions of the model differ significantly from reality, see (Lartillot et al., 2007)). In the case of two GC rich species, disregarding the fast-evolving third codon positions is often sufficient to overcome compositional artifact (Jeffroy et al., 2006).

This also explains why mutational saturation (i.e., an excess of multiple substitutions) is so deleterious to phylogenetic inference. In a perfect world, saturation would simply erase the ancient phylogenetic signal and lead to a lack of resolution in the deepest nodes of the tree. However, especially in phylogenomics, a subtle model violation such as minor variations of amino acid composition may be sufficient to introduce a bias in the interpretation of multiple substitutions, hence creating a strong spurious signal, favoring an incorrect topology (Rodriguez-Ezpeleta et al., 2007).

2 Detecting systematic errors

One of the main problems of phylogenomics is,

therefore, estimating whether systematic errors have had an effect on the usually well-supported tree. First, one looks at the properties of the alignment. For instance, the level of saturation can be evaluated by comparing the number of inferred substitutions with the number of observed differences (Philippe et al., 1994b), the heterogeneity of nucleotide/amino acid composition (Rodriguez-Ezpeleta et al., 2007), and the level of heterotachy (rate heterogeneity within a position) (Baele et al., 2006; Lopez et al., 2002). These analyses are only the starting point for further experiments, such as the removal of the species with the highest compositional bias. Indeed, sensitivity of phylogenetic inference to taxon sampling is currently the most efficient way to detect systematic errors. For instance, using two different species to represent red algae lead to two incongruent, yet strongly supported, eukaryotic phylogenies (Rodriguez-Ezpeleta et al., 2007). Or, in animals, platyhelminths clustered with nematodes within Ecdysozoa, but are grouped with Lophotrochozoa when nematodes were not included in the analysis (Philippe et al., 2005b). Having a rich taxon sampling therefore provides an effective way to evaluate whether systematic errors are playing a role in a phylogenomic analysis.

The first phylogenomic studies of animals (Blair et al., 2002; Philip et al., 2005; Wolf et al., 2004) were based on a very limited number of species (i.e., *Homo*, *Drosophila* and *Caenorhabditis*) and focused on the Coelomata (nematodes basal) versus Ecdysozoa (nematodes together with insects) question. Since only a single distant outgroup was available, the probability of a long-branch attraction (LBA) artifact due to a systematic error was high. Not surprisingly, these early studies all placed the nematodes as basal with solid support and came invariably to the conclusion that the Coelomata hypothesis is correct (Blair et al., 2002; Dopazo et al., 2004; Wolf et al., 2004). We will use this example to illustrate the efficiency of three methods (improved model, improved taxon sampling, and removal of fast-evolving data) for overcoming systematic errors.

Although LBA artifact was initially characterized as a problem specific to parsimony (due to the statistical inconsistency of maximum parsimony (MP) as a reconstruction method), very similar phenomena are often observed in probabilistic reconstructions as well. In the latter, however, the causes are violations of one or more of the hypotheses of the underlying models of evolution, thus preventing a correct interpretation of multiple substitutions that are concentrated along branches of fast-evolving taxa. In the following, we will use, for convenience, the term LBA artifact also in the context of model violation problems.

3 Avoiding systematic errors by improving the model of sequence evolution

Detecting systematic errors is important, but the ultimate goal is to avoid them in the first place. One way to do this is to improve the models of sequence evolution to make them more realistic, thus reducing model violations. The most useful improvements for phylogenetic accuracy are those that allow a better detection of multiple substitutions, not obligatorily the ones that most improve the fit of the model to the data (Yang, 1997). Numerous improvements since the original, completely homogeneous, Jukes and Cantor model (Jukes & Cantor, 1969) have been made to include evolutionary heterogeneities.

The first model that assumed the positions of an alignment to be heterogeneous (Yang, 1994) constituted a major improvement. The evolutionary rate of all amino acid positions is estimated, and then divided into discrete categories (often 4—here described with +Γ4 to 16). The resulting distribution is assumed to follow a gamma law and its shape parameter, alpha, is estimated from the data and used to obtain a greatly improved approximation of the evolutionary process. The assumption of gamma distribution is especially important because interpretation of the fastest evolving positions, which are the most prone to generate misleading signals, are improved. Application of the discrete gamma correction largely reduced the likelihood of LBA-artifacts (Yang, 1996). More recently, models incorporating qualitative heterogeneities in addition to quantitative ones among positions constitute another important advancement (Lartillot & Philippe, 2004; Pagel & Meade, 2004).

In a Markovian model, the substitution process is described by a so-called instantaneous rate matrix, specifying the rate of replacement between each pair of amino acids (or nucleotides). Standard models assume that the same rate matrix can describe every amino acid position of a protein. The exchange rates of this matrix are obtained by empirical means, or are directly estimated from the data. In general they are biochemically meaningful, in that probabilities of substitutions between biochemically similar amino acids are much higher than between non-related pairs of residues. For instance, models based on the WAG matrix (Whelan & Goldman, 2001), which was established empirically (based on alignments of globular

proteins) allow one to predict more convergences and reversions among amino acids with similar biochemical properties. However, the process reaches equilibrium after ~5 substitutions and is only dominated by its equilibrium frequency vector over 20 amino acids: which implies that, at saturated positions, each amino-acid is supposed, by a site-homogeneous model, to appear with a probability essentially reflecting its relative abundance in the overall alignment (Lartillot et al., 2007).

The fact that site-homogeneous models are dominated by equilibrium frequency vector is biologically unsound; in all alignments many positions seem to be saturated and yet still display very few different amino acids, such as positions with only D or E (or with only K or R), when the functional constraint is to have a negative (positive) charge at that location. Standard homogeneous models will underestimate multiple substitutions at these positions, because they expect to see many different amino acids after a few substitutions have occurred. In other word, when confronted with highly saturated positions that experienced multiple substitutions, this leads to an inadequate behavior of the standard model (Lartillot et al., 2007). In contrast, a mixture model such as the CAT model (Lartillot & Philippe, 2004) allows dealing with the amino acid specificity of positions by sorting them into groups with various profiles of amino acid stationary frequencies; since the size of the alphabet is close to the actual value (e.g., two for a D/E position), the CAT model predicts that about two amino acids will be present at these positions, even after numerous substitutions. Due to that property, the CAT model is able to more correctly evaluate the possibility of multiple conservative substitutions, thereby successfully avoiding LBA artifacts where other models fail, as detailed below.

Starting from a phylogenomic dataset of animals that strongly rejects the Coelomata hypothesis (Philippe et al., 2005b), Lartillot et al. (Lartillot et al., 2007) reduced the species sampling to include only fungi, arthropods, nematodes and deuterostomes. Because a single distant outgroup (fungi) is used, a spurious attraction of the fast-evolving nematodes towards the base of the tree would be expected and was indeed observed when using a site-homogeneous model (e.g., WAG+Γ4). In contrast, under the same difficult conditions, the CAT+Γ4 model resolves the fast-evolving nematodes as a sister-group to arthropods. Posterior predictive analysis suggests that the CAT model is less sensitive to LBA artifacts than standard site-homogeneous models, because of its more realistic handling of highly saturated positions, due to a more accurate estimate of the true alphabet size per position.

Another interesting case concerning the phylogenetic position of the very fast-evolving acoels represented by *Convoluta* was recently published. The classical view of a platyhelminth affiliation is supported by an analysis of a phylogenomic dataset using the WAG+Γ4 model (probably because of the rapid evolutionary rate of platyhelminths). Interestingly, the CAT+Γ4 model strongly rejects this grouping and instead favors the acoels as an early part of the deuterostome radiation (Philippe et al., 2007), rather than a sister-group of all other Bilateria, as suggested by rRNA analysis (Ruiz-Trillo et al., 1999).

There are other heterogeneities of the evolutionary process that are now incorporated in models. For instance, heterotachy can be handled by a covarion-like model (Galtier, 2001; Huelsenbeck, 2002) or a mixture of branch length model (Kolaczkowski & Thornton, 2004). The first appears to have a better fit (Zhou et al., 2007), whereas the second was shown to avoid LBA artifacts (Kolaczkowski & Thornton, 2008). Similarly, various non-stationary models (Blanquart & Lartillot, 2006; Foster, 2004; Galtier & Gouy, 1995; Yang & Roberts, 1995) deal with compositional bias. To date, various model violations have generally been addressed independently, mainly because of mathematical difficulties and of computation time limitations. However, difficult cases, such as the position of honeybees based on mitochondrial genomes (Blanquart & Lartillot, 2008), require addressing several model violations simultaneously. The use of more complex models could become prohibitive, however, due to their excessive need of memory, computation power and time.

4 Avoiding systematic errors by improving the taxon sampling

Solely improving models is often insufficient to adequately interpret multiple substitutions, because complex models need much more data to estimate the additional parameters. These parameters, such as positional rate heterogeneity, require many species in order to be accurately estimated. Moreover, increasing the number of taxa, especially by breaking long branches, is in itself an excellent way to more efficiently detect multiple substitutions (Hendy & Penny, 1989; Hillis, 1996; Philippe & Douzery, 1994; Zwickl & Hillis, 2002).

A good example illustrating the strong influence of an improved taxon sampling in the context of animal evolution is the taxon-dependent switch between two significantly supported and mutually exclusive topologies (Delsuc et al., 2005). Starting with the three classical ingroup species (i.e., *Homo, Drosophila* and *Caenorhabditis*) and budding yeast as a distant outgroup, a significant support for the Coelomata topology (*Homo* with *Drosophila*) was found. However, the addition of three rather slowly evolving fungi and three closely related outgroups, i.e., two choanoflagellates and a cnidarian, which efficiently break the long branch leading to the budding yeast, leads to a significant support of the Ecdysozoa topology (*Drosophila* with *Caenorhabditis*).

Another example of major improvement in phylogenetic performance that is associated with better species sampling can be seen in the progression of three papers that were all based on the same original dataset, and using the same phylogenetic methods (Baurain et al., 2007; Philippe et al., 2005b; Philippe et al., 2004), with the major difference being the successive addition of more animal and outgroup sequences as they became available. The support for protostomes was only marginal (Bootstrap Value (BV) of 57%) and that for Ecdysozoa non-existant in the original dataset, which contained only seven non-fast-evolving bilaterian sequences and one choanoflagellate (close outgroup) (Philippe et al., 2004). The second dataset had a 100% BV for protostomes (18 non-fast bilaterian sequences and four close outgroups including two diploblasts) but only a 7% BV for Ecdysozoa (Philippe et al., 2005b). Finally, the third dataset, which contained 31 non-fast bilaterian sequences and nine close outgroups, including two sponges and four cnidarians, showed in addition to a 100% BV for protostomes, a 79 % BV for Ecdysozoa (Baurain et al., 2007).

Improvement in taxon sampling to reduce the impact of systematic error can be obtained by increasing the number of species, but also by the exclusion of rogue taxa (Sanderson & Shaffer, 2002). Rogue taxa are generally fast-evolving but also evolve differently from the other taxa (e.g., compositionally biased or a different set of variable positions). They therefore accumulate problems: numerous multiple substitutions and numerous model violations. Systematic errors can often be avoided by changing the representative species of a group from a rogue taxon to a "normal" species. For instance, with a limited taxon sampling (23 species), the simple replacement of the rogue, fast-evolving nematode, *Caenorhabditis,* by a rather slowly evolving one, *Xiphinema,* was sufficient to avoid the LBA artifact between nematodes and platyhelminths and to recover the monophyly of Ecdysozoa (Baurain et al., 2007). Therefore, sequencing a large number of taxa has two advantages: breaking (long) branches and allowing the selection of the slowest evolving representative species.

One should also keep in mind that the outgroup species, which are needed to root molecular phylogenies (at least with reversible models), are in essence rogue taxa, because they diverged earlier and often evolved under different evolutionary constraints. Therefore, an easy way to reduce the systematic error and one that should generally be applied is to perform an additional analysis without the outgroup species (Brinkmann et al., 2005).

There is a tradeoff between these two solutions for a better phylogenetic inference: improved species sampling versus the use of better models of sequence evolution. Since with fewer species one needs a better model to infer multiple changes that could be more readily deduced from a denser species sampling, an improved species sampling would in principle permit the use of more simple models. Taxon sampling in phylogenomics is nevertheless still too sparse at present to evaluate this tradeoff in more detail.

5 Avoiding systematic errors by removing fast-evolving data

Improving taxon sampling is not always possible, mainly because of extinctions (e.g., no close relatives are known for *Amborella* and *Latimeria*), and we cannot be sure that models correctly interpret multiple substitutions. An alternative method to avoid systematic error is to remove the data that evolve the fastest or that violate most a given model. For instance, to address the issue of compositional bias, one can easily remove the third codon positions, which accumulate the bias faster, or recode the sequences into a reduced alphabet that is more homogeneous, e.g., RY coding for purine and pyrimydine (Woese et al., 1991).

The most general approach is the selective removal of the fastest evolving positions, since this reduces the level of saturation and concomitantly eliminates the positions that violate most a given model. A simple way to achieve this positional sorting is to use the affiliation to discrete gamma categories (Ruiz-Trillo et al., 1999): the higher the number, the faster the corresponding positions. Unfortunately, this

approach is topology dependent, with different topologies having an effect on the affiliation of positions and a strong impact on further analyses (Rodriguez-Ezpeleta et al., 2007). Alternatively, one can use a compatibility approach to improve the quality of the input data without any *a priori* knowledge of the phylogenetic relationships (Pisani, 2004; Qiu & Estabrook, 2008). One identifies alignment positions that are the least compatible with the other positions and excludes them from the analysis, based on the principle that their random behavior is due to multiple substitutions. In other words, this approach is based on the assumption that the positions that are the most compatible in an alignment matrix share a similar historical signal and evolve slowly. The SF (slow-fast) method is also largely topology independent, since the evolutionary rate is inferred within predefined monophyletic groups for which only the intergroup-relationships are under study (Brinkmann & Philippe, 1999). It was recently used to de-saturate a phylogenomic animal data set, where the distant outgroup (fungi) was suspected to erroneously attract the fast-evolving nematodes to the base of Bilateria, hence supporting the Coelomata hypothesis (Delsuc et al., 2005). The progressive removal of fast-evolving positions leads to a continuous decrease in the support for Coelomata; lack of resolution due to the reduced amount of data was not an explanation, since the support for Ecdysozoa monophyly (arthropods + nematodes) continuously increases.

One can also eliminate genes instead of positions, based on the idea that more information is available to determine if they are fast evolving and that one can selectively pinpoint problematic species (Brinkmann et al., 2005). A good example is provided by a phylogenomic analysis that provides strong support in favor of the strange grouping of platyhelminths+nematodes, possibly due to an ingroup LBA artifact (Philippe et al., 2005b). The progressive elimination of the proteins that evolved the fastest in platyhelminths and nematodes regularly decreases the support for platyhelminths+nematodes, and increases the one for the grouping of platyhelminths with Lophotrochozoa.

Although powerful, the removal of fast-evolving data (positions, genes and species) is potentially dangerous, because one needs to make at least some assumptions about the phylogeny and the evolutionary model to estimate the rate. These assumptions may bias the output of the data removal (Rodriguez-Ezpeleta et al., 2007). This approach should therefore mainly be reserved to exploratory analyses or when long branches cannot be broken.

6 Phylogenomics and lack of resolution

Because of systematic errors, a high statistical support in phylogenomics can be misleading, especially when few species are considered (Hedtke et al., 2006; Jeffroy et al., 2006). Interestingly, a weak support can also be created by systematic errors if the strength of the phylogenetic and of the artifactual signal are about the same (Rodriguez-Ezpeleta et al., 2007). Coming back to the animal phylogeny, it was proposed that the major lines of bilaterian evolution can not be resolved by a phylogenetic approach (Rokas et al., 2005), since they result from a rapid and massive adaptive radiation, the so-called "Cambrian explosion". It is indeed well known that resolving closely spaced speciation events could require a very large amount of data (Philippe et al., 1994a; Saitou & Nei, 1986) but also that noise (i.e., mutational saturation) can seriously reduce the resolving power of molecular phylogenetics. Therefore the lack of resolution observed by Rokas et al. (2005) could alternatively be due to mutational saturation. This explanation is suggested by (i) the use of several rogue taxa as well as a rather poor species sampling and (ii) the use of too simplistic phylogenetic inference methods. We recently applied two of the approaches detailed above simultaneously to reduce systematic error: (i) the use of an improved taxon sampling (from 13 to 46 animals, with elimination of some rogue taxa, e.g. *Caenorhabditis* or *Drosophila*), and (ii) the use of a better method (CAT+Γ4 model). Interestingly, this resulted in a greatly improved statistical support, most of the nodes within Bilateria being supported by a high bootstrap value, e.g., Ecdysozoa and Lophotrochozoa (Baurain et al., 2007). This improved resolution is best explained by the more efficient extraction of the phylogenetic signal that results in a better signal to noise ratio. This study illustrates that, even when considering a very large dataset, interpreting lack of resolution as evidence of radiation is hazardous. It is of prime importance to first exclude the possibility that a systematic error that erased part of the phylogenetic signal is not the explanation.

7 An updated phylogenomic tree of animals

The ongoing major sequencing efforts at the level of both complete genomes and expressed sequence tags (ESTs) allows the construction of more complete

datasets with less missing positions and also with a largely improved species sampling (Philippe & Telford, 2006). Here we updated alignments used in three recent studies (Delsuc et al., 2006; Jimenez-Guri et al., 2007; Philippe et al., 2007). A phylogenomic dataset assembled using SCAFOS (Roure et al., 2007) (55 species, 102 genes and 25712 amino acid positions) contains the major metazoan lineages and two independent and closely related outgroups, choanoflagellates and ichthyosporeans. For the first time, three very fast-evolving species, the appendicularian tunicate *Oikopleura dioica*, the acoel *Convoluta pulchra* and the myxozoan *Buddenbrockia plumatellae,* are simultaneously included. These are all rogue species, which could induce LBA artifacts. We therefore used a rich taxon sampling and a complex model of sequence evolution (CAT+Γ4).

The Bayesian tree shown in Fig. 1, inferred by Phylobayes using a CAT+Γ4 model, is in good agreement with current knowledge. First, in contrast to what would be expected if LBA plays a dominant role, the three very fast evolving species are not clustered together. Second, the tree is in agreement with the new animal phylogeny, especially the monophyly of Ecdysozoa and Lophotrochozoa is recovered. Poriferans are the most basal monophyletic group of metazoans and cnidarians emerge shortly thereafter. Cnidarians are deeply divided into two major and highly discrete monophyletic groups, the anthozoans represented here by sea anemones (*Nematostella*) and stony corals (*Acropora*) and two additional cnidarian groups hydrozoans (*Hydra* and *Hydractinia*) and scyphozoans (*Cyanea*) as well as the divergent sequence of the myxozoan *Buddenbrockia plumatellae*, in agreement with a recent work (Jimenez-Guri et al., 2007). However, even with the site-heterogeneous CAT model, which is less sensitive to LBA (Lartillot et al., 2007), the area surrounding the position of a very fast evolving species is usually associated with an enhanced uncertainty that is reflected here in the low posterior probability (PP) value of 0.88. Alternatively this phenomenon may be seen as being related to the much greater error margin associated with the estimate of its very long branch.

The bilaterian part of the metazoan tree (Fig. 1) has a pronounced basal branch that separates it from the other metazoans, this may indicate either a long evolutionary period and/or a shorter phase of accelerated evolution potentially associated with the origin of bilaterian animals. This tree supports a division of bilaterians into the two major groups deuterostomes and protostomes. The location of the very fast evolving acoel *Convoluta* and of the chaetognath *Spadella* at the base of protostomes is interesting since their phylogenetic position is a longstanding and ongoing dispute. The chaetognaths were originally considered to be part of the deuterostomes, essentially based on morphological characters and on their deuterostome-like development. Our analysis confirms the protostome affinity of chaetognaths, as the sister-group to Ecdysozoa + Lophotrochozoa (Marletaz et al., 2006), and not to Lophotrochozoa (Matus et al., 2006). If chaetognaths are the most basal protostomes, one could consider their unusual features as an indication for the potential ancestry of the deuterostome developmental program and of deuterostome-like organisms within bilaterians (Marletaz et al., 2006).

The evolutionary position of the acoel flatworms is much more disputed. Classically they were considered to be a part of the platyhelminths, i.e. flatworms. However, several molecular phylogenetic analyses (rRNA, myosin, and mitochondrial genome) found strong support for a much more basal position within the metazoan tree as the sister-group to all extant bilaterian animals (Baguna & Riutort, 2004; Ruiz-Trillo et al., 2002; Ruiz-Trillo et al., 2004; Ruiz-Trillo et al., 1999). A major problem of these studies is the very fast evolutionary rate of the acoels,which could easily lead to an LBA artifact, with acoels being attracted by the distantly related non-bilaterian outgroup. An EST based phylogenomic analysis using the CAT+Γ4 model significantly rejects a sister-group relationship to the platyhelminths, but did not recover acoels as the most basal bilaterian, suggesting instead that they are rather close to, or included within, deuterostomes (Philippe et al., 2007). The current analysis (Fig. 1) confirms the rejection of the classical platyhelminths link, but suggests a slightly different position, as a sister-group of the protostomes. The moderate support (PP 0.88) argues in favor of an improved taxon sampling within acoels before making any firm conclusion.

Although, the monophyly of both deuterostomes and chordates is recovered in our analysis (Fig. 1), the basal and the subsequent branches of deuterostomes are very short suggesting that the diversification of major lineages, such as chordates or cephalochordates, could have happened rapidly after the origin of the deuterostomes themselves (at least more rapidly than within protostomes, where the branches at the base of Lophotrochozoa and Ecdysozoa are longer). Accordingly, a very similar analysis (same inference method and similar set of genes and of species) did not recover the monophyly of deuterostomes, and the

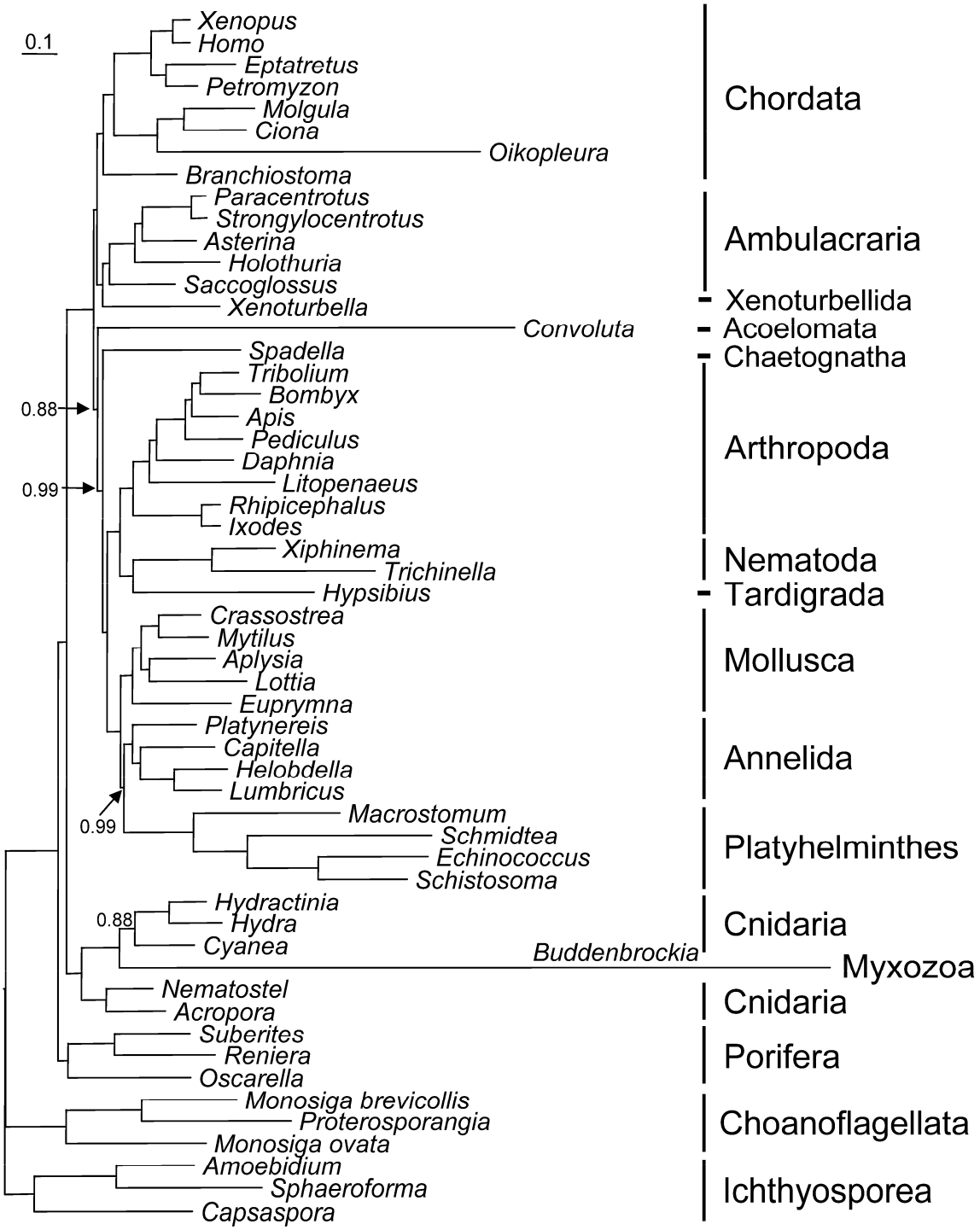

Fig. 1. The tree was inferred based on a phylogenomic dataset (102 genes and 25712 amino acid positions) using the program phylobayes (Bayesian inference) and the CAT model with four discrete gamma categories (CAT+Γ4) after elimination of the constant positions. The methods are the same as in (Lartillot & Philippe, 2008). Posterior probability values are only indicated if they are lower than 1.

chordates were inferred as the first emerging bilaterian group with the remaining deuterostomes forming the sister-group to the protostomes (Lartillot & Philippe, 2008). This high sensitivity to species/gene sampling indicates that the monophyly of deuterostomes is an open question, requiring all the improvements in phylogenomic inference discussed above to be solved (in particular, improved sampling within Xenambulacraria and improved models). An important reason explaining this difficulty is that only a distant outgroup is available, hence the deuterostome monophyly could be due to an attraction of protostomes by the non-bilaterian outgroup (Lartillot & Philippe, 2008).

The topology within "core" protostomes does not reveal any surprises. Within Ecdysozoa the arthropods represent the sister-group of nematodes and tardigrades (water bears); within arthropods there is a basal division between chelicerates and the pan-crustaceans with paraphyletic crustaceans, i.e., with the malacostracan *Litopenaeus* basal and the phyllopod *Daphnia* as the sister-group to insects, here represented by four species of winged insects (Pterygota). Among the three lophotrochozoan groups represented here, the traditional view of a sister-group relationship between the two trochozoans annelids and mollusks (Neotrochozoa) is rejected by the data. Instead a grouping between annelids and platyhelminths is supported (PP 0.99), suggesting that the flatworms are also trochozoans. This unorthodox relationship was also recovered in several recent analyses (Lartillot & Philippe, 2008; Lavrov & Lang, 2005; Passamaneck & Halanych, 2006). However, caution is necessary using this interpretation since many protostome lineages (e.g. priapulids, onychophorans, rotiferans, acanthocephalans, nematomorphans, bryozoans, phoronids, brachiopods, sipunculans, nemertines, pogonophores and echiurids) are not present in our analysis (see (Dunn et al., 2008) for an improvement of taxon sampling).

Finally, we studied both the impact of the model of sequence evolution and of the heuristic search. In addition to phylobayes (CAT+Γ4; Fig. 1), we performed standard phylogenetic analyses with a site-homogeneous model (WAG+Γ4), employing widely used software, including MrBayes (Ronquist & Huelsenbeck, 2003), PhyML (Guindon & Gascuel, 2003), PhyML-SPR (subtree pruning and regrafting) (Hordijk & Gascuel, 2005), Treefinder (Jobb et al., 2004), RaxML (Stamatakis et al., 2005) and IQPNNI (Vinh le & Von Haeseler, 2004). Twelve different topologies (Table 1) were obtained and compared, assuming the WAG+Γ4 model, using the program Tree-Puzzle (Schmidt et al., 2002) to perform likelihood ratio tests, including the Shimodaira-Hasegawa and the Expected Likelihood Weight (ELW) test. The most obvious

Table 1 Estimation of the quality of the model of sequence evolution and of the heuristic search

Software	Observed differences	ΔLnL	S-H	ELW
IQPNNI	(C(E((Ac,P),(M,A))))	4.60	0.94	0.34
MrBayes	(E((C,P),(M,A))) Ac+My basal to Bilateria	83.07	0.36	0.02 #
"	(C(E(((Ac,My)P),(M,A))))	69.38	0.48	0.02 #
"	(E((P,C),(M,A))) Ac basal to Bilateria	59.00	0.52	0.03 #
"	(E(C(((Ac,My)P),(M,A))))	72.10	0.46	0.03 #
"	(C(E(((Ac,My)P),(M,A))))	69.38	0.48	0.02 #
"	((Ac,C),(E(P(M,A))))	55.59	0.55	0.02 #
Treefinder	((E,C),((((Ac,My)T)P),(M,A)))	817.21	0.00 #	0.00 #
PhyML	(C(E((Ac,P),(M,A)))) My+T basal to Meta	787.95	0.00 #	0.00 #
PhyML-SPR	(E(((Ac,P)C),(M,A)))	0.00	1.00	
RAxML-MP	(E(((Ac,P)C),(M,A)))	0.00	1.00	
RAxML-6xRan	(E(((Ac,P)C),(M,A)))	0.00	1.00	
RAxML-1xRan	(Cni(C(E((M,A) P)))) T basal to Xamb; D sister-group to other Meta; Ac basal to Meta	1322.69	0.00 #	0.00 #
RAxML-1xRan	(C(E(Ac,P),(M,A)))) T basal to Bilateria	368.97	0.00 #	0.00 #
RAxML-2xRan	(E(((Ac,P)C),(M,A))) Agnaths not mono	249.09	0.02	0.00 #
Phylobayes	*(Ac(C(E(M(A,P)))))*	126.54	0.18	0.00 #

Abbreviations: A, annelids; Ac, acoels; C, chaetognaths; Cni, Cnidarians; D, deuterostomes; E, Ecdysozoa; M, mollusks; Meta, metazoa; mono, monophyletic; My, myxozoans; P, platyhelminths; T, Tunicate; Xamb, Xenambulacraria.

Six independent chains were run for MrBayes and produced six different topologies. Ten random trees were tested for RAxML. Note that all computations were done with the WAG+Γ4 model, except for phylobayes (CAT+Γ4 model). In that case, topological difference is likely due to the model, and not to tree search, because four independent chains gave the same topology. # indicates that the difference is significant (P<0.05).

result is the major impact of heuristic search; only two software (PHYML-SPR and RAxML) found the same topology. Some software (Treefinder and PHYML) have been trapped in distant local minima (~800 log likelihood values of difference). This confirms that heuristic search strategies are potentially problematic when very long sequences are used, even if the number of taxa (55) is relatively small: the potential barriers between local minima are high in phylogenomics and difficult to overcome. MrBayes clearly illustrates this problem, because six independent MCMCMC chains remain trapped (during ~500,000 generations) in different, albeit close, local minima. It should be noted that our best tree, despite being found twice independently, is not necessarily the ML tree, because an exhaustive search was not performed.

This best ML tree under the WAG+Γ4 model, however, contains nodes that are very likely the result of an LBA artifact, e.g. the node that unites the acoel *Convoluta,* platyhelminths, and the chaetognath *Spadella*. These three taxa correspond to long, unbroken, branches that are not grouped with the CAT+Γ4 model (Fig. 1). Moreover, to our knowledge, there are no morphological characters to support the grouping of chaetognaths with acoels or platyhelminths. The CAT topology is, under the WAG model, far from being the best tree, since it has a log likelihood value more than 125 units worse and is even rejected by the ELW test. Since several studies have demonstrated a much better fit for the CAT than for the WAG model (Lartillot et al., 2007; Lartillot & Philippe, 2004; Lartillot & Philippe, 2008), the CAT topology should be preferred, even if we cannot be sure that this topology is the best under that model (although it has been found in four chains independently).

8 Biological implications

The new animal phylogeny that assumes a basal division in deutero- and protostomes (with Lophotrochozoa and Ecdysozoa) proposed by rRNA analysis is now confirmed, in some detail, by phylogenomics and can be used with confidence. Comparison with the older view of metazoan evolution essentially represented in the Coelomata hypothesis illustrates the strong influence of the "great chain of being" of Aristotle on scientists (Adoutte et al., 1999). Almost all of the major differences between the two classification schemes involve simple organisms, of which the best known groups are platyhelminths (acoelomates) and nematodes (pseudocoelomates) which were considered to be the two oldest lineages within Bilateria under the Coelomata hypothesis. These groups are not primitively simple, but are the product of secondary simplification, which transforms formerly rather complex organisms into more or less simple ones that look at least superficially primitive. Interestingly, many "simple" animals (e.g., nematodes, platyhelminths, acoels, myzostomids, myxozoans) are often fast-evolving from a molecular point of view, suggesting that their secondary simplification, which implies gene losses, would have relaxed functional constraints. The same phenomenon is observed in eukaryotes, especially when mitochondria are lost (Philippe et al., 2000). The absence of simple and early-branching bilaterian animals strongly suggests that the urbilaterian ancestor was rather complex. This does not imply that it appears suddenly from a simpler ancestor, but instead that, among the intermediate, simpler, ancestral species, only one lineage survives.

The strong influence of the "great chain of being" is harmful. The complex thinking of Aristotle, elaborated long before the Darwinian theory, was unfortunately too often reduced to the implicit assumption that evolution invariantly leads towards more complexity. Molecular phylogenetics refutes this hypothesis, by demonstrating that "simple" organisms almost invariantly emerge from within "complex" organisms, not at their base, hence being secondarily simplified. Unfortunately, the erroneous "great chain of being" ideology remains influential, especially thanks to the incorrect use of the adjectives "higher" and "lower" to characterize taxa, a practice that should be prohibited (Mogie, 2007). Instead of uselessly searching for simple and primitive organisms to understand evolution, one should rather look for simple evolutionary mechanisms that could explain the characters of complex and simple extant organisms.

9 Perspectives

The guidelines we suggest to obtain reliable phylogenomic trees (Philippe et al., 2005a) are, not surprisingly, very similar to those recommended to obtain a reliable single gene phylogeny (Sanderson & Shaffer, 2002): (1) increase the taxon sampling, with at least two species per group of interest (although the evolutionary process renders it often inherently scarce by the extinction of related species), (2) improve the models of sequence evolution (especially the properties that enhance detection of multiple substitutions), and (3) remove the fastest-evolving data, especially species (although the last approach is potentially dangerous, and intellectually not satisfying, since the

first two points should be sufficient).

We have only explored phylogenomic inference based on primary sequences (in fact, only on super-matrices, but super-trees generally give similar results), because this approach is the most advanced. This offers the unique opportunity to explore the strengths and weaknesses of genome-based inference. Several fundamentally different approaches have been proposed (for review see Delsuc et al., 2005), such as similarity in oligonucleotide frequencies (not dependent on homology), gene presence/absence, or gene order. Although they are promising (especially gene order because character space is huge, greatly reducing the possibility of homoplasy), the inference methods need to be improved. The use of intron positions is illustrative, by supporting either Coelomata (Zheng et al., 2007) or Ecdysozoa (Roy & Gilbert, 2005), depending only on the methods used. In the long run, these alternative phylogenomic approaches are expected to furnish an important corroboration, since they are derived from completely independent characters that evolve according to their proper rules.

Acknowledgements This work was supported by operating funds from the Natural Sciences and Engineering Research Council and the Canada Research Chairs Program. I wish to thank Nicolas Lartillot for discussions and helpful comments.

References

Adoutte A, Balavoine G, Lartillot N, de Rosa R. 1999. Animal evolution. The end of the intermediate taxa? Trends in Genetics 15: 104–108.

Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R. 2000. The new animal phylogeny: reliability and implications. Proceedings of the National Academy of Sciences 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.

Anderson FE, Cordoba AJ, Thollesson M. 2004. Bilaterian phylogeny based on analyses of a region of the sodium-potassium ATPase beta-subunit gene. Journal of Molecular Evolution 58: 252–268.

Baele G, Raes J, Van de Peer Y, Vansteelandt S. 2006. An improved statistical method for detecting heterotachy in nucleotide sequences. Molecular Biology and Evolution 23: 1397–1405.

Baguna J, Riutort M. 2004. The dawn of bilaterian animals: the case of acoelomorph flatworms. Bioessays 26: 1046–1057.

Baurain D, Brinkmann H, Philippe H. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Molecular Biology and Evolution 24: 6–9.

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

Blanquart S, Lartillot N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Molecular Biology and Evolution 23: 2058–2071.

Blanquart S, Lartillot N. 2008. A Site- and Time-Heterogeneous Model of Amino-Acid Replacement. Molecular Biology and Evolution 25: 842–858.

Brinkmann H, Giezen M, Zhou Y, Raucourt GP, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Systematic Biology 54: 743–757.

Brinkmann H, Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Molecular Biology and Evolution 16: 817–825.

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

Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439: 965–968.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6: 361–375.

Dopazo H, Santoyo J, Dopazo J. 2004. Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. Bioinformatics 20: i116–i121.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SH, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452: 745–749.

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

Foster PG. 2004. Modeling compositional heterogeneity. Systematic Biology 53: 485–495.

Galtier N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Molecular Biology and Evolution 18: 866–873.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proceedings of the National Academy of Sciences USA 92: 11317–11321.

Groth JG, Barrowclough GF. 1999. Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. Molecular Phylogenetics and Evolution 12: 115–123.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704.

Halanych KM. 2004. The new view of animal phylogeny. Annual Review of Ecology, Evolution, and Systematics 35: 229–256.

Halanych KM, Bacheller JD, Aguinaldo AM, Liva SM, Hillis DM, Lake JA. 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science 267: 1641–1643.

Hedtke SM, Townsend TM, Hillis DM. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Systematic Biology 55: 522–529.

Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. Systematic Zoology 38: 297–309.

Hillis DM. 1996. Inferring complex phylogenies. Nature 383: 130–131.

Hordijk W, Gascuel O. 2005. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics 21: 4338–4347.

Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. Molecular Biology and Evolution 19: 698–707.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends in Genetics 22: 225–231.

Jimenez-Guri E, Philippe H, Okamura B, Holland PW. 2007. *Buddenbrockia* is a cnidarian worm. Science 317: 116–118.

Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evolutionary Biology 4: 18.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN ed. Mammalian protein metabolism. New York: Academic Press.

Kluge AG. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Systematic Zoology 38: 7–25.

Kobayashi M, Wada H, Satoh N. 1996. Early evolution of the Metazoa and phylogenetic status of diploblasts as inferred from amino acid sequence of elongation factor-1 alpha. Molecular Phylogenetics and Evolution 5: 414–422.

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431: 980–984.

Kolaczkowski B, Thornton JW. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. Molecular Biology and Evolution. (in press)

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evolutionary Biology 7 Suppl 1: S4.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution 21: 1095–1109.

Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philosophical Transactions of the Royal Society B: Biological Sciences 363: 1463–1472.

Lavrov DV, Lang BF. 2005. Poriferan mtDNA and animal phylogeny based on mitochondrial gene arrangements. Systematic Biology 54: 651–659.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Molecular Biology and Evolution 19: 1–7.

Marletaz F, Martin E, Perez Y, Papillon D, Caubit X, Lowe CJ, Freeman B, Fasano L, Dossat C, Wincker P, Weissenbach J, Le Parco Y. 2006. Chaetognath phylogenomics: a protostome with deuterostome-like development. Current Biology 16: R577–578.

Matus DQ, Copley RR, Dunn CW, Hejnol A, Eccleston H, Halanych KM, Martindale MQ, Telford MJ. 2006. Broad taxon and gene sampling indicate that chaetognaths are protostomes. Current Biology 16: R575–576.

McHugh D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. Proceedings of the National Academy of Sciences USA 94: 8006–8009.

Mogie M. 2007. Drop “higher” and “lower” to raise descriptive standards. Nature 448: 865.

Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology 53: 571–581.

Passamaneck Y, Halanych KM. 2006. Lophotrochozoan phylogeny assessed with LSU and SSU data: evidence of lophophorate polyphyly. Molecular Phylogenetics and Evolution 40: 20–28.

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. Molecular Biology and Evolution 22: 1175–1184.

Philippe H, Brinkmann H, Martinez P, Riutort M, Baguna J. 2007. Acoel flatworms are not Platyhelminthes: evidence from phylogenomics. PLOS One 2: e717.

Philippe H, Chenuil A, Adoutte A. 1994a. Can the Cambrian explosion be inferred through molecular phylogeny? Development 120: S15–S25.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005a. Phylogenomics. Annual Review of Ecology, Evolution, and Systematics 36: 541–562.

Philippe H, Douzery E. 1994. The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationships. Journal of Mammalian Evolution 2: 133–152.

Philippe H, Germot A, Moreira D. 2000. The new phylogeny of eukaryotes. Current Opinion in Genetics and Development 10: 596–601.

Philippe H, Lartillot N, Brinkmann H. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Molecular Biology and Evolution 22: 1246–1253.

Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Molecular Biology and Evolution 21: 1740–1752.

Philippe H, Sörhannus U, Baroin A, Perasso R, Gasse F, Adoutte A. 1994b. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. Journal of Evolutionary Biology 7: 247–265.

Philippe H, Telford MJ. 2006. Large-scale sequencing and the new animal phylogeny. Trends in Ecology & Evolution 21: 614–620.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Molecular Biology and Evolution 21: 1455–1458.

Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda. Systematic Biology 53: 978–989.

Qiu Y-L, Estabrook GF. 2008. Inference of phylogenetic relationships among key angiosperm lineages using a

compatibility method on a molecular data set. Journal of Systematics and Evolution 46: 130–141.
Regier JC, Shultz JW. 2001. Elongation factor-2: a useful gene for arthropod phylogenetics. Molecular Phylogenetics and Evolution 20: 136–148.
Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Systematic Biology 56: 389–399.
Rokas A, Kruger D, Carroll SB. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science 310: 1933–1938.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. BMC Evolutionary Biology 7 Suppl 1: S2.
Roy SW, Gilbert W. 2005. Resolution of a deep animal divergence by the pattern of intron conservation. Proceedings of the National Academy of Sciences USA 102: 4403–4408.
Ruiz-Trillo I, Paps J, Loukota M, Ribera C, Jondelius U, Baguna J, Riutort M. 2002. A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proceedings of the National Academy of Sciences USA 99: 11246–11251.
Ruiz-Trillo I, Riutort M, Fourcade HM, Baguna J, Boore JL. 2004. Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. Molecular Phylogenetics and Evolution 33: 321–332.
Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguna J. 1999. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science 283: 1919–1923.
Saitou N, Nei M. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. Journal of Molecular Evolution 24: 189–204.
Sanderson MJ, Shaffer HB. 2002. Troubleshooting molecular phylogenetic analyses. Annual Review of Ecology and Systematics 33: 49–72.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
Shultz JW, Regier JC. 2000. Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. Proceedings of the Royal Society: Biological Sciences 267: 1011–1019.
Sidow A, Thomas WK. 1994. A molecular evolutionary framework for eukaryotic model organisms. Current Biology 4: 596–603.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–463.
Sullivan J, Swofford DL. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Systematic Biology 50: 723–729.
Sullivan JP, Lavoue S, Hopkins CD. 2000. Molecular systematics of the African electric fishes (Mormyroidea: Teleostei) and a model for the evolution of their electric organs. Journal of Experimental Biology 203: 665–683.
Vinh le S, Von Haeseler A. 2004. IQPNNI: moving fast through tree space and stopping in time. Molecular Biology and Evolution 21: 1565–1571.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution 18: 691–699.
Woese CR, Achenbach L, Rouviere P, Mandelco L. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Systematic and Applied Microbiology 14: 364–371.
Wolf YI, Rogozin IB, Koonin EV. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Research 14: 29–36.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39: 306–314.
Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution 11: 367–370.
Yang Z. 1997. How often do wrong models produce better phylogenies? Molecular Biology and Evolution 144: 105–108.
Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Molecular Biology and Evolution 12: 451–458.
Zheng J, Rogozin IB, Koonin EV, Przytycka TM. 2007. Support for the Coelomata clade of animals from a rigorous analysis of the pattern of intron conservation. Molecular Biology and Evolution 24: 2583–2592.
Zhou Y, Rodrigue N, Lartillot N, Philippe H. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evolutionary Biology 7: 206.
Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology 51: 588–598.