Phylogenomics

Hervé Philippe, Frédéric Delsuc, Henner Brinkmann, Nicolas Lartillot

PHYLOGENOMICS

Hervé Philippe, Frédéric Delsuc, Henner Brinkmann and Nicolas Lartillot*

Canadian Institute for Advanced Research. Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada

* Laboratoire d'Informatique, de Robotique et de Mathématiques de Montpellier. CNRS – Université de Montpellier 2. 161, rue Ada, 34392 Montpellier Cedex 5, France

Corresponding author:

Hervé Philippe

Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada.

Email: herve.philippe@umontreal.ca

Keywords: inconsistency, molecular phylogeny, systematic bias, taxon sampling, tree of life

CONTENTS

ABSTRACT

The continuous flow of genomic data is creating unprecedented opportunities for the reconstruction of molecular phylogenies. Access to whole-genome data means that phylogenetic analysis can now be performed at different genomic levels, such as primary sequences and gene order, allowing for reciprocal corroboration of the results. We critically review the different kinds of phylogenomic methods currently available, paying particular attention to method reliability. Our emphasis is on methods for the analysis of primary sequences because these are the most advanced. We discuss the important issue of statistical inconsistency and show how failing to fully capture the process of sequence evolution in the underlying models leads to tree reconstruction artifacts. We suggest strategies for detecting and potentially overcoming these problems. These strategies involve the development of better models, the use of an improved taxon sampling and the exclusion of phylogenetically misleading data.

INTRODUCTION

The newly arising discipline of phylogenomics owes its existence to the revolutionizing progress in DNA sequencing technology. The number of complete genome sequences is already high and increases at an ever-accelerating pace. The newly coined term phylogenomics (Eisen 1998, O'Brien & Stanyon 1999) comprises several areas of research at the interplay between molecular biology and evolution. The main issues are: (1) using molecular data to infer species relationships, and (2) using information on species evolutionary history to gain insights into the mechanisms of molecular evolution. The majority of publications on phylogenomics deal with the second aspect (see Sjolander 2004 for review). However, our concern here is the use of data at the genomic scale to reconstruct the phylogeny of organisms.

A novel and interesting aspect of phylogenomics lies in the possibility of using molecular information above the primary sequence level. In particular, trees can be inferred from whole genome features such as gene content (Fitz-Gibbon & House 1999, Snel et al 1999, Tekaia et al 1999), gene-order (Korbel et al 2002, Sankoff et al 1992), intron positions (Roy & Gilbert 2005), or protein domain structure (Lin & Gerstein 2000, Yang et al 2005). A potential advantage of these methods is that the complexity of some of these characters (e.g. gene order) renders the character-state space very large, reducing the risk of homoplasy by convergence and reversal, thus rendering the inferred phylogenies more reliable. However, these integrated approaches imply the use of a reduced number of characters relative to primary sequence based approaches. There are

about 300 times fewer genes than amino acid positions (assuming a mean protein length of about 300 amino acids), thus increasing the risk of stochastic error.

In fact, stochastic or sampling error constitutes one of the major limitations of standard phylogenetics based on single genes. Because the number of positions of a single gene is small, random noise influences the inference of numerous nodes, leading generally to poorly resolved phylogenetic trees. The idea of using large amounts of genomic data as a way to address this problem is not new. For example, to resolve the original tritomy between chimpanzee, human, and gorilla, about ten kilobases were sequenced as early as the late 1980's (Miyamoto et al 1988). However, technical and financial limitations have often confined molecular systematics to the use of a few markers (e.g. ribosomal RNA (rRNA)), for which a large diversity of organisms have been sequenced. Numerous important phylogenetic questions remained unsolved and great hope was placed into the wealth of genomic data soon to be available.

Sampling (or stochastic) error should vanish as the number of genes added to the analysis gets large enough. In practice, this means that statistical support (e.g. bootstrap support) will eventually rise to 100% as more genes are considered. The use of tens of thousands, or millions, of aligned positions that provide a great deal of phylogenetic information should ultimately lead to fully resolved trees. Indeed, several empirical studies confirmed this premise (Bapteste et al 2002, Madsen et al 2001, Murphy et al 2001, Qiu et al 1999, Rokas et al 2003, Soltis et al 1999). This increased resolution leads to the

optimistic view that phylogenomics would “end the incongruencies” observed in single gene phylogenies (Gee 2003). However, whether the resulting, highly supported, phylogenetic trees are the true ones is not certain.

Systematic error and consistency

The most important challenge of phylogenomics is to verify that tree reconstruction methods are consistent, i.e. converge towards the correct answer as more and more characters are considered (Felsenstein 1978, Felsenstein 1988). In principle, at least in a probabilistic framework, a lack of consistency can always be traced back to some violation of model assumptions by the data analyzed. Note that methods that are not explicitly model based, such as Maximum Parsimony (MP), are equivalent to a statistical analysis under an implicit model (Steel & Penny 2000). The best understood causes of method inconsistency stem from models that do not properly account for: (1) variable evolutionary rates, leading to the long branch attraction (LBA) artifact (Felsenstein 1978), (2) heterogeneous nucleotide/amino acid compositions, resulting in the artificial grouping of species that share the same bias (Lockhart et al 1994), and (3) heterotachy, i.e. shift of position-specific evolutionary rates (Kolaczkowski & Thornton 2004, Lockhart et al 1996, Philippe & Germot 2000). These systematic biases could be interpreted respectively as rate signal, compositional signal and heterotachous signal, which we will collectively refer to as non-phylogenetic signals (Ho & Jermiin 2004). In other words, non-phylogenetic signals are due to substitutions that occurred along the true

phylogeny, but that are misinterpreted by tree reconstruction methods as supporting an alternative topology.

Compared to single gene studies, inconsistency is more pronounced in phylogenomic analyses. For example, a reanalysis of the large dataset of Rokas *et al.* (Rokas et al 2003) demonstrates that, depending on the method used, mutually incongruent, yet 100% supported, trees could be obtained (Phillips et al 2004). It is well-accepted that the analysis of phylogenomic datasets will necessarily increase the resolution of the trees through the “increase of the signal-to-noise ratio” (Rokas et al 2003). Indeed, the signal-to-random-noise ratio increases but the phylogenetic-to-non-phylogenetic signal ratio remains constant whatever the number of genes considered (assuming that the gene sampling is not biased). In this review, we will focus on best practices for enhancing the phylogenetic signal in genomic data, while reducing the impact of erroneous signals, in order to obtain accurate and robust trees.

ASSEMBLY OF PHYLOGENOMIC DATASETS

The reliability of a phylogenetic tree depends on the quality of the data and the accuracy of the reconstruction method. In 1988, Felsenstein noted that “molecular evolutionists who use methods for inferring phylogenies do not engage in much discussion of the properties of the methods they use since they focus on the difficult task of collecting the data” p. 523 (Felsenstein 1988). Almost 20 years later, molecular systematists still spend much of their time assembling larger and larger datasets, and the crucial discussion about inference methods

remains neglected. In phylogenomics, the reliability of the inference is often simply justified by the large number of characters used. Nevertheless, the problem of data acquisition deserves further discussion, as it can heavily compromise the subsequent analysis.

Importance of a rich taxon sampling

A long-standing debate in phylogenetics concerns the relative importance of improving taxon versus gene sampling (Graybeal 1998, Hillis et al 2003, Rosenberg & Kumar 2003). In the genomic age, gene sampling would seem not to be an issue. However the limited resources devoted to systematics often prevent sequencing the genomes of all relevant species. Two strategies can be used, depending on the importance accorded to taxon sampling: (1) gathering complete genome sequences from a few key organisms, or (2) gathering incomplete, yet large, genome sequences from a great diversity of organisms.

The first approach is supported by some computer simulation studies (Rosenberg & Kumar 2003) and is the most frequently used in phylogenomic analyzes (Blair et al 2002, Goremykin et al 2004, Misawa & Janke 2003, Philip et al 2005, Rokas et al 2003, Wolf et al 2004). However, the design of computer simulations and the interpretation of their results, make it difficult to draw firm conclusions from this approach (Hillis et al 2003, Rosenberg & Kumar 2003). Empirical evidence seems nevertheless to argue against the taxon-poor approach, as illustrated by the phylogeny of metazoans. Two new clades (Ecdysozoa, the moulting animals, including among others arthropods and

nematodes, and Lophotrochozoa including among others annelids, molluscs and platyhelminthes) were proposed from rRNA analyses (Aguinaldo et al 1997). However, several phylogenomic studies strongly supported the paraphyly of Ecdysozoa, when considering a few model organisms and using a distant outgroup (Blair et al 2002, Dopazo et al 2004, Philip et al 2005, Wolf et al 2004). In contrast, when 49 species and 71 genes (20,705 positions) are used (Philippe et al 2005), the monophyly of Ecdysozoa and Lophotrochozoa is recovered with strong support (Figure 1a). However, the removal of close outgroups leads to drastic changes (Figure 1b): the fast evolving lineages emerge paraphyletically at the base of the tree. Such an asymmetrical tree-shape is expected to result from LBA artifact, when the outgroup is distantly related (Philippe & Laurent 1998), as fungi are. Contrary to the situation with a few species discussed above, the statistical support for these incorrect placements is weak (bootstrap values between 32 and 63), demonstrating that an increased taxon sampling (from 4-10 to 45) has reduced, but not eliminated, the impact of LBA. In fact, the low bootstrap support (Figure 1b) demonstrates that non-phylogenetic signal becomes equivalent to phylogenetic signal when species sampling is impoverished.

A rich-taxon sampling is not the panacea however. First, computation time increases rapidly with the number of species, rendering exhaustive searches impossible with more than 20 species and most heuristic searches with probabilistic methods intractable with more than ~200 species. Second, adding taxa can sometimes degrade the phylogenetic inference (Kim 1996). Third, the

number of extant species can be naturally sparse, forever preventing the assembly of a rich and balanced taxon sample. For example, *Amborella* is proposed to constitute the first, or one of the first, emerging angiosperm lineages (Qiu et al 1999), but is the only extant representative of an ancient group. Even if the heated debate about its placement (Goremykin et al 2004, Soltis et al 2004) could be solved by improved taxon sampling, the assumed basal position of *Amborella* might prove difficult to attest in the absence of closely related extant taxa (Stefanovic et al 2004). In conclusion, an adequate taxon sampling, as balanced as possible, is important to increase the accuracy of phylogenomic trees.

Missing data

Although genome sequencing has become ever easier, it seems unlikely that complete genomes will be soon available for a rich diversity of organisms. In addition, a bias in favor of sequencing small genomes leads to potential problems. Since small genomes are generally derived from larger genomes, whole-genome features, such as gene content or gene order, will evolve much faster, rendering tree reconstruction susceptible to artifacts such as LBA and compositional bias (Copley et al 2004, House & Fitz-Gibbon 2002, Korbel et al 2002, Lake & Rivera 2004, Wolf et al 2001). Moreover, genome reduction is often associated with an accelerated rate of protein evolution (Brinkmann et al accepted, Dufresne et al 2005), or extremely biased nucleotide compositions (Herbeck et al 2005). The sampling of a fraction of the genome from species with

huge genomes is therefore a necessity to represent some key taxa and/or to include less biased representatives.

Two low cost approaches can be used: (1) the selection of a limited set of genes potentially useful for the phylogenetic question of interest, followed by their targeted PCR amplification and sequencing (Murphy et al 2001); (2) the sequencing of thousands of Expressed Sequence Tags (ESTs), which generally provides hundreds of relevant genes (Bapteste et al 2002). The first method is more adapted to phylogeny at small evolutionary scales and has the advantage that genes can be *a priori* selected to obtain an optimal phylogenetic signal. The second might be preferable at larger evolutionary scales (e.g. among protists) and allows the discovery of many other genes, which can shed light on the evolution of important features such as metabolic pathways.

Phylogenomic reconstruction methods based on gene content/order cannot be applied to incomplete genomic sampling, but those based on DNA strings and on primary sequences can. In the latter case, missing data will occur even when complete genomes are used, especially at a large evolutionary scale, since most, if not all, genes can be lost, duplicated or horizontally transferred in some organisms. To our knowledge, no theoretical reasons suggest that sequence-based approaches can not be used on incomplete alignments, i.e. containing cells coded as missing data. Nevertheless, “the problem of missing data is widely considered to be the most significant obstacle (…) in combining datasets (…) that do not include identical taxa”, as suggested by empirical studies and computer simulations (Wiens 2003). Two problems need to be

distinguished: (1) the potential lack of resolution induced by the presence of taxa with too many missing cells, (2) the possible interaction between missing entries and artifact-inducing model violations.

Recent computer simulations using large number of characters (Philippe et al 2004, Wiens 2003) suggest that the inaccurate placement of incomplete taxa is not due to missing data but rather to the insufficient number of informative characters. As an extreme example, the tree reconstruction method remains accurate when positions have an average of four known and 32 unknown character states, because each species is nevertheless represented by about 3,000 amino acids (Philippe et al 2004). However, the presence of missing cells unevenly distributed across the data matrix potentially affects estimates of model parameters. It is not yet clear how the induced model misspecifications in turn influence phylogenetic inference. Interestingly, it seems that the advantage of adding an incomplete taxon that breaks a long branch is greater than the disadvantage of the induced model misspecification (Wiens in press).

Few attempts at assessing the effect of missing data have been made with empirical data (Bapteste et al 2002, House & Fitz-Gibbon 2002, Philippe et al 2004). For instance, a bipartition of the supermatrix (25% of missing data) into the most complete genes and the less complete genes appears to be indistinguishable from random bipartitions of the same size (Bapteste et al 2002, Philippe et al 2004). However, when the level of missing data is extreme (92%), the quality of the inference appears to be affected (e.g. strong support for the paraphyly of Glires and of Ecdysozoa) (Driskell et al 2004), despite the large size

of the data set (70 taxa and 1131 genes). In summary, even if the problem generated by missing data has been overrated, additional work is needed to characterize its impact more precisely.

INFERENCE OF PHYLOGENOMIC TREES

The methods used in phylogenomic inference are of two kinds: (1) primary sequence based methods, which are very similar to the classical tree reconstruction, and for which several excellent reviews and textbooks are available (Felsenstein 2004, Holder & Lewis 2003, Swofford et al 1996), and (2) methods above the sequence level (Wolf et al 2002). The two approaches are fundamentally similar, the main difference being the characters used. In both cases, we believe that probabilistic methods are more powerful and more reliable. First, they have a more robust theoretical justification in that they rely on an explicit account of their assumptions by using stochastic models describing the pattern of molecular evolution (Felsenstein 2004). Second, not only do they allow estimating the phylogeny and a confidence-level, but they also provide general methods to evaluate the fit of the model used (Goldman 1993).

Approaches based on whole-genome features

Since genomes are the results of evolution, virtually any features comparable between organisms can be used to infer phylogenies, as evidenced by the plethora of new approaches recently published (Fitz-Gibbon & House

1999, Henz et al 2005, House & Fitz-Gibbon 2002, House et al 2003, Korbel et al 2002, Lin & Gerstein 2000, Pride et al 2003, Qi et al 2004, Snel et al 1999, Tekaia et al 1999, Wolf et al 2001, Yang et al 2005). The justifications of whole genome tree approaches are generally that the phylogeny of organisms can not be equated to the phylogeny of single genes (such as rRNA) and that this classical phylogeny is sensitive to hidden paralogy, horizontal gene transfer (HGT) or tree reconstruction artifacts (Fitz-Gibbon & House 1999, Lin & Gerstein 2000, Snel et al 1999, Tekaia et al 1999). Since tree reconstruction artifacts can affect any approach and are difficult to detect, these new approaches based on various character types are of paramount importance to corroborate phylogenetic inference (Miyamoto & Fitch 1995, Swofford 1991, Wolf et al 2002).

Distribution of sequence strings

The frequencies of small oligonucleotides (up to 8) or oligopeptides (up to 6) observed in the genome or the proteome can be transformed into distances then used to construct phylogenies (Blaisdell 1986). Numerous well-accepted clades, including deep ones, were recovered using this approach, confirming that a phylogenetic signal is present in these characters (Edwards et al 2002, Pride et al 2003, Qi et al 2004). However, several undisputed clades were significantly rejected, and the comparison of the branch lengths of the genome tree obtained using tetranucleotide frequencies with those of the tree obtained from the standard analysis of rRNA sequences (Pride et al 2003) suggests that the phylogenetic signal contained in sequence strings saturates rapidly. However,

the methods proposed so far are extremely crude. Oligonucleotide or oligopeptide frequencies are transformed into distances without any underlying model of evolution. It is nevertheless remarkable that something considered as a bias in standard sequence-based methods (Lockhart et al 1994) contains a phylogenetic signal, but it is not yet clear whether accurate methods can be developed to extract it.

Homology and orthology assessment

All other approaches require the definition of homologous, or more often, orthologous genes. By definition, the phylogenetic history displayed by orthologous genes is the organismal phylogeny (Fitch 1970), whereas paralogous or xenologous genes display a combination of organismal and gene-specific history. In practice, the identification of orthologous genes involves a certain amount of circularity because it requires an *a priori* knowledge of the organismal phylogeny. Indeed, finding orthologous genes is difficult since the organismal phylogeny is generally unknown (or at best partially known), and its reconstruction represents the goal. This problem is much akin to the alignment/phylogeny problem, in which alignment and phylogeny should be estimated simultaneously (Sankoff et al 1973, Wheeler 2003), but the practical difficulties are such that the two steps are generally separated. Moreover, a careful phylogenetic reconstruction of all gene families is a Herculean labor most researchers want to avoid (but see Storm & Sonnhammer 2002). Therefore, an operational, yet approximate, definition of orthology is used. Schematically, all

genomes are compared to each other at the amino acid level; only the pairs of sequences that are best reciprocal hits are further considered. The clusters of orthologous groups are then constructed by a single linkage analysis of all orthologous pairs (Tatusov et al 1997) or by Markov cluster algorithms (Remm et al 2001). These approaches, albeit reasonably effective in practice, do not guarantee the identification of orthologous genes, since, if a gene has been transferred from a distant organism and replaced the original copy, it will fulfill all the requirements while being xenologous. Information from synteny will probably improve the accuracy of orthology assignment (Zheng et al 2004), but much more work is needed before obtaining perfect assignment of orthologous genes.

Gene order methods

Since the character space of gene order data is huge, it probably constitutes the most promising genome feature based method. However, it is also technically and computationally the most difficult (see Moret et al 2005 for review). Briefly, distances can be computed by minimising the number of inversions, transpositions, insertions and deletions necessary to transform one unichromosomal genome into another (Sankoff et al 1992), a method which is most often further simplified by considering only inversions (Bourque & Pevzner 2002). Alternatively, the break-point distance between two genomes, defined as the minimum number of pairs of genes next to each other in one of the two genomes, but not in the other (Nadeau & Taylor 1984), can be used to infer phylogeny (Blanchette et al 1997). However, the evolution of gene order involves

additional rearrangement mechanisms, not easily accountable for with these methods, such as translocation (i.e., transposition between chromosomes), and fusion or fission events of chromosomes.

More recently, methods based on maximum-likelihood distances (Wang & Warnow 2005), or on Bayesian inference (Larget et al 2005, York et al 2002) have been proposed that take multiple changes in gene order into account. They revealed a non-negligible level of saturation (York et al 2002) and a limited tree-resolving power (Larget et al 2005). The small size of mitochondrial genomes might explain the predominance of stochastic noise in the latter case. Prokaryotic genomes contain much more information, but are too large for analysis by current software. The drastic simplifying assumption that gene order can be reduced to the presence/absence of gene pairs allows the inference of prokaryotic phylogenies that are similar to the ones based on gene content (Korbel et al 2002, Wolf et al 2001). Further methodological and computational developments are needed to realize the full potential of gene-order methods.

Gene content methods

Phylogenomic analyses based on gene content generally use orthologs, but a few variants exist which use homologous instead of orthologous genes (Fitz-Gibbon & House 1999), protein domain content (Yang et al 2005), or fold occurrence (Lin & Gerstein 2000, Yang et al 2005). Interestingly, an orthologous gene present in all organisms (the Holy Grail of the sequence-based approach) has the same character state and is not informative for gene content methods.

This is an important source of corroboration, since patterns informative for sequence-based approaches are not informative for gene content approaches. Otherwise, the distribution of orthologs will be informative in the cases of (1) gene loss, which have a non-negligible probability of being convergent, (2) gene genesis, which is potentially the most informative, and (3) horizontal gene transfers (i.e. the acquisition of a new gene from a distantly related organism at the base of a clade will create a synapomorphy for this clade and a homoplasy for locating this clade). Gene content, albeit more integrated than primary sequences, is thus far from providing an unambiguous phylogenetic signal.

The binary matrices of presence/absence of homologous or orthologous genes can be analyzed by distance methods (Lin & Gerstein 2000, Snel et al 1999), parsimony (Fitz-Gibbon & House 1999) or Dollo parsimony (Wolf et al 2001). However, big/small genome attraction (Lake & Rivera 2004) appears to affect all these methods, as demonstrated by the artificial grouping of unrelated species with small genomes (e.g. *Mycoplasma*, *Buchnera*, *Chlamydia* or *Rickettsia*). For instance, Copley et al (2004) demonstrated that phylogenies based on gene content and protein domain combinations support the paraphyly of Ecdysozoa, but are biased by a systematic high rate of character loss in nematodes. When this bias is accounted for by computing the number of losses expected randomly, a slight support for the monophyly of Ecdysozoa is recovered. Interestingly, this artifact yields the same inconsistency phenotype as the one displayed by sequence-based analyses (Philippe et al 2005): in both cases, arthropods are sister of vertebrates to the exclusion of nematodes. Such a

convergence between the two methods could be explained by the fact that the fast evolving species are also those that have undergone the most extreme genome reduction. This observation weakens the strength of the corroboration between gene content and primary sequence approaches.

Altogether, gene-content phylogenies are not in excellent agreement with previous knowledge, even if we ignore the problem due to small genome attraction. The monophyly of the three domains (Archaea, Bacteria and Eukaryota) is always recovered, but the phylogeny within domains is much more problematic. For instance, *Halobacterium* never clusters with *Methanosarcina*, probably because of many HGTs from Bacteria which attract it towards the base of Archaea (Korbel et al 2002); except with threshold parsimony (House et al 2003), the monophyly of Proteobacteria is never recovered (Dutilh et al 2004, Gu & Zhang 2004, Henz et al 2005, Wolf et al 2001).

Several technical improvements have recently been proposed. Since big/small genome attraction is akin to the problem of compositional bias in sequence-based approaches, this non-phylogenetic signal can be reduced by the use of the logdet/paralinear transformation (Lake & Rivera 2004). A simple model of gene genesis and gene loss allows ML estimates of evolutionary distances (Gu & Zhang 2004, Huson & Steel 2004), but simulations showed that their performance appears to be slightly poorer than the performance of Dollo parsimony (Huson & Steel 2004).

In summary, the methods for inferring trees based on whole genome features are at an early stage of their development, which might be comparable

to that of sequence-based methods in the early 1970's. In particular, they generally lack a global probabilistic modeling. Numerous works are ongoing and it will be important to extensively evaluate the accuracy of present and future methods.

Approaches based on primary sequences

Supermatrix versus supertree

The question of how to analyze multiple datasets has been the subject of intense debate (for review see Bull et al 1993, de Queiroz et al 1995). In brief, three approaches are mainly used: (1) total evidence (Kluge 1989), in which all datasets are combined together, called hereafter the supermatrix approach, (2) separate analysis (Miyamoto & Fitch 1995), in which the datasets are analyzed individually and resulting topologies are combined using consensus or supertree methods, called hereafter the supertree approach, (3) conditional combination (Bull et al 1993, Lecointre & Deleporte 2005), in which only the datasets considered as congruent are combined, called hereafter the conditional supermatrix approach. Generally, their respective advantages are: (1) minimizing stochastic error, (2) increasing the significance of corroboration, and (3) minimizing conflicting signal. In the practice of phylogenomics, the supermatrix is by far preferred (Murphy et al 2001, Philippe et al 2005, Qiu et al 1999, Rokas et al 2003), followed by a few cases of the supertree method (Daubin et al 2002, Philip et al 2005).

Although Bull et al (1993) state that “no rational systematist would suggest combining genes with different histories to produce a single reconstruction”, the reliance in the power of the supermatrix approach is quite strong, and this methodology is generally applied. Against Bull et al., one may argue that discordant genes will each display a different discordant history, which will be averaged away through a combined analysis (but see Matte-Tailliez et al 2002). In most of the large multigene studies, the possibility of incongruencies is voluntarily minimised by selecting genes having *a priori* the same evolutionary history (e.g. single-copy genes (Lerat et al 2003, Murphy et al 2001, Philip et al 2005), organellar genes (Qiu et al 1999, Soltis et al 1999), or orthology assessment based on synteny (Rokas et al 2003)). But generally, the homogeneity of the datasets was not tested, indicating a strict application of the total evidence principle.

The problem of homogeneity is important since several recurrent processes (gene duplication, HGT, or lineage sorting) can lead to incongruent gene trees. In particular the very notion of a Tree of Life has been questioned because of rampant HGTs (Doolittle 1999). Several studies have nevertheless argued that a strong phylogenetic signal is present in the prokaryotic genomes (for review see Brown 2003, Philippe & Douady 2003) and therefore that HGTs do not wipe out the notion of organismal phylogeny. It is clear however, that few, and probably none, of the genes have followed exactly the organismal phylogeny during the entire history of life on Earth. Thus, HGTs constitute a source of nuisance that should be addressed to improve the accuracy of phylogenetic

reconstructions.

The infrequent use of the conditional supermatrix approach is likely not due to a philosophical rejection of its principle, but rather to the difficulty of detecting homogeneous datasets in practice. The Incongruence Length Difference (ILD) test (Farris et al 1995) was initially designed for parsimony and has been recently adapted to distance methods (Zelwer & Daubin 2004). However, the interpretation of this test is complicated by the fact that stochastic noise can generate by itself significant results (Dolphin et al 2000). Its efficiency in detecting incongruence and determining data combinability has been repeatedly questioned (Darlu & Lecointre 2002, Dowton & Austin 2002, Yoder et al 2001). Parametric incongruence tests have also been proposed for ML methods (Huelsenbeck & Bull 1996), or in a Bayesian framework (Nylander et al 2004), although they are not yet in widespread use. A promising model has been proposed, in which each gene can, with a certain prior probability, choose between either conforming to the common global topology or relying on its own topology (Suchard et al 2003).

Nevertheless, the conditional supermatrix approach is used, despite the fact that the criteria used to discard gene/sequence are not well validated (Brochier et al 2002, Brown et al 2001, Lecointre & Deleporte 2005, Matte-Tailliez et al 2002). However, the nature of the test used to decide whether a gene significantly supports a different topology yields divergent interpretations of the same data with regards to the importance of HGTs (Bapteste et al 2004, Lerat et al 2003, Zhaxybayeva et al 2004). An improvement of these

incongruence tests (Goldman et al 2000) therefore constitutes an important avenue of future research.

The supertree approach is most often used to combine trees from the literature that were obtained from diverse sources of data (Sanderson et al 1998), the most popular method being Matrix Representation with Parsimony (MRP, Baum 1992, Ragan 1992). The comparative efficiency of supermatrix and supertree approaches is poorly studied, especially in a phylogenomic context (Gatesy et al 2004, Philip et al 2005). We have therefore analyzed the dataset of 71 genes (Figure 1a) using a supertree MRP approach. Interestingly, almost all nodes inferred from the supermatrix or using the supertree are identical, except the position of urochordates within deuterostomes (not shown) and the relationships among protostomes, indicating an excellent congruence of the supermatrix and supertree (only 3 differences for 46 bipartitions). When the supertree is reconstructed with MP (Figure 1c), platyhelminthes are grouped with tardigrads+nematods instead of with other lophotrochozoans (annelids and molluscs), disrupting the monophyly of both Ecdysozoa and Lophotrochozoa. When the supertree is reconstructed using Bayesian inference (Figure 1d), the results were slightly more consistent, since the monophyly of Ecdysozoa, but not Lophotrochozoa, was recovered. The MRP supertree approach appears to have difficulty in placing the fast-evolving lineages (e.g. Platyhelminthes). The synergy among all positions in the supermatrix might explain the ability of the method to better deal with LBA artifacts. However, refined studies are urgently required to evaluate the relative accuracy and efficiency of the supermatrix and supertree

methods.

Supermatrix: scaling up current methods

In contrast to the genome-feature approaches mentioned above, the main advantage of sequence-based methods is that their properties have been intensively explored, tested and validated, so that many of their strengths and weaknesses are known. Indeed, in most sequence-based phylogenomic analyses published to date, almost the same protocols as for single gene studies have been applied. The congruence among results obtained by different methods is high, with some notable exceptions (Canback et al 2004, Goremykin et al 2004, Soltis et al 2002, Stefanovic et al 2004). These incongruencies confirm the presence of non-phylogenetic signal, and suggest that the increase in resolution obtained by analyzing larger datasets is not in itself a guarantee of accuracy. Conversely, the agreement between the methods does not mean that the obtained tree is correct (see Brinkmann et al accepted).

The dramatic change of scale of the data matrices implies the need for a corresponding increase in computational power, in particular for probabilistic methods. Two factors have to be considered. First, there is a simple scaling-up of both memory requirement and computational load. Second, the reliability of the heuristic search procedures underlying ML programs, or the Monte Carlo devices of the Bayesian samplers, is anything but guaranteed. A particular concern is that conflicting signals result in the presence of many secondary maxima in the space of tree topologies, separated by high potential barriers. Standard procedures are

likely to get trapped in these local optima (Salter 2001), and thus do not yield reliable phylogenetic estimates.

A number of algorithmic innovations have led to better heuristic searches in the space of topologies: genetic algorithms (Brauer et al 2002, Lemmon & Milinkovitch 2002), disk-covering methods (Huson et al 1999), or parallelized computing (Keane et al 2005). In the case of the Bayesian methods, an interesting approach has been proposed consisting of using coupled 'heated' Monte Carlo Markov Chains (MCMC), which can be easily parallelized (Altekar et al 2004, Feng et al 2003). Thanks to these advances, maximum likelihood and Bayesian phylogenetic reconstructions will soon be able to handle phylogenomic datasets. However, their overall reliability has been evaluated mainly on simulated data, which are probably much more 'funnel-shaped' towards the true phylogeny than are real sequences (Brinkmann et al accepted, Stamatakis et al 2005).

Another possible stance towards efficient tree space searches is to restrict the analysis, by constraining nodes that have been found with high support when each gene of the concatenation was analyzed separately (Philippe et al 2005). The number of trees compatible with these constraints is still large, but accessible to an exhaustive analysis. Such approaches may not be considered as a definitive method, but could provide a good proxy.

Toward more complex models

If the availability of large data matrices poses new computational

challenges to probabilistic methods, it allows the development of more realistic, parameter-rich, models. As long as the number of parameters increases more slowly than the number of sites, a model does not fall into the infinitely many parameter trap (Felsenstein 2004), and thus, has good consistency properties. Of course, no probabilistic model will ever capture evolutionary patterns in their full complexity, but their most important aspects, at least the ones that cause inconsistency of the current methods, can be accounted for (Steel 2005). The main idea would not so much be an increased realism, but an improved flexibility accounting for the diverse kinds of heterogeneities and disparities of the substitution processes.

One possible research direction is to account for disparities in the evolutionary process across the genes that make up the concatenation. A simple solution is to constrain the model to have a global topology, but gene-specific branch lengths (Yang 1996). More generally, any parameter other than the topology can be considered as gene-specific in such "separate" (or partitioned) models. Another possible avenue of research is to account for site-specific patterns of substitution, using mixture models (Kolaczkowski & Thornton 2004, Lartillot & Philippe 2004, Pagel & Meade 2004). A mixture model combines several different classes to describe the substitution process, each of which is characterized by its own set of parameters (e.g. equilibrium frequencies or exchangeability probabilities).

Thus far, few studies have tried to address the relative performances of alternative probabilistic models on phylogenomic datasets. A recent study

(Brinkmann et al accepted) has confirmed that accounting for site-specific rates, or having a good empirical substitution rate matrix are important factors yielding a higher phylogenetic accuracy. On the other hand, separate models, in spite of their overall better statistical fit, do not seem to fundamentally improve phylogenetic inference. Since separate models handle a substantial part of heterotachy, this suggests that heterotachy, despite recent interest, may not constitute a major source of systematic bias. In any case, a much wider analysis of the impact of model choice on the prevalence of artifacts in phylogenomic inference has to be carried out.

Reducing systematic errors through data exclusion

Phylogenomic datasets contain a large amount of genuine phylogenetic signal but they also contain non-phylogenetic signals that current methods of tree reconstruction are not able to perfectly handle. To avoid the perils of inconsistency, one can take advantage of the fact that the quantity of phylogenetic signal is no longer a serious limiting factor. More precisely, the part of the datasets that contains mainly non-phylogenetic signals can be excluded, allowing the concentration of phylogenetic signal in the remaining dataset. This increase of the phylogenetic to non-phylogenetic signal ratio reduces the probability of inconsistency, even without the use of improved tree reconstruction methods.

The rationale of most methods of data exclusion is straightforward: tree reconstruction artifacts are due to multiple substitutions that are not correctly

identified as convergences or reversions by inference methods (Olsen 1987). The simplest possibility consists in removing the fast evolving species, which by definition accumulate multiple substitutions. This approach efficiently reduces the misleading effect of the rate signal (Philippe et al 2005, Stefanovic et al 2004). In many cases, the exclusion of odd taxa (Sanderson & Shaffer 2002) is implicit, since investigators never envision using them in their analyses (e.g. microsporidia as a fungal representative).

When all of the available species representing a clade of interest are fast-evolving, the specific removal of the fastest evolving sequences of this clade from the supermatrix appears to be efficient. For example, when a complete supermatrix of 133 genes is used, all tree reconstruction methods strongly, albeit artifactually, locate microsporidia at the base of eukaryotes, but probabilistic methods with a complex model (WAG+F+Γ) avoid this LBA artifact when >70% of the microsporidial sequences are coded as missing data (Brinkmann et al accepted). It is interesting to note that a highly incomplete taxon (70% missing data) is more accurately located than a complete one. A similar approach, in which genes in their totality are discarded, was successful at avoiding the attraction between the fast evolving nematodes and platyhelminthes (Philippe et al 2005) or between nematodes and outgroup (Dopazo & Dopazo 2005). Interestingly, in the first case, the statistical support for the monophyly of Ecdysozoa and Lophotrochozoa increases when more and more genes are discarded (as long as more than 40 genes are considered).

These approaches are rather crude, since sequences from genes and/or

species are discarded in their totality. Nevertheless, even if these sequences contain much non-phylogenetic signal, they likely also contain some phylogenetic signal. Refined methods have been proposed to selectively eliminate fast evolving characters in part (Lopez et al 1999) or completely (Brinkmann & Philippe 1999, Burleigh & Mathews 2004, Dutilh et al 2004, Pisani 2004, Ruiz-Trillo et al 1999). In several cases, taxa that emerged at the base of the tree when all the characters are used are relocated later in the tree when fast evolving positions are removed, strongly suggesting that their observed basal position is due to an LBA artifact (Brochier & Philippe 2002, Philippe et al 2000, Pisani 2004). However, the number of remaining slowly evolving positions is often too small to yield high statistical support for most of the clades in single-gene analyses, but not in phylogenomic analyses (Burleigh & Mathews 2004, Delsuc et al 2005).

The RY coding strategy (Woese et al 1991) discards all fast-evolving transitions and improves inference without drastically compromising the resolution (Phillips et al 2004). Importantly, this coding not only addresses the problem of rate signal but also of compositional signal. Indeed, the GC content can be extremely variable among homologous sequences from various organisms whereas the frequency of purines is remarkably homogeneous (Woese et al 1991). This constitutes a method of choice to avoid inconsistency due to compositional bias.

Finally, it is possible to remove characters whose evolutionary history violates most the assumptions of the underlying model of sequence evolution

instead of the fastest evolving characters/species. In fact, the RY coding eliminates transitions that are mainly responsible for the non-stationarity of the nucleotide composition (see Hrdy et al 2004 for a similar approach in the case of proteins). Similarly, the constant sites violate the assumptions about the distribution of site rates (Lockhart et al 1996) and their elimination constitutes an efficient way of improving inference (Hirt et al 1999, Phillips et al 2004). Heterotachous sites violate the assumption that the evolutionary rate of a position is constant through time, made by all current models except the covarion model (Fitch & Markowitz 1970). As expected, the elimination of these sites had reduced LBA artifacts in the case of the eukaryotic phylogeny (Inagaki et al 2004, Philippe & Germot 2000).

We believe that these data removal approaches are complementary to the improvement of tree reconstruction methods through the implementation of more realistic models of sequence evolution. So far, they are also more readily accessible in practice, since they are less demanding in terms of the complexity of bioinformatic methods and computational time.

CONCLUDING REMARKS

In this review, we have emphasized that inconsistency of tree reconstruction methods constitutes the major limitation of phylogenomics. We have explored ways to reduce its impact, whether at the level of data assembly or at the level of tree building *per se*. Adequate taxon sampling, probabilistic methods and the exclusion of phylogenetically misleading data constitute the

three most important criteria to obtain reliable phylogenomic trees. However, this might not be sufficient to ensure that the inferred tree is the correct one. We believe that an important test of the reliability of tree reconstruction methods is their robustness with respect to species sampling. In particular, a reliable method should be able to recover exactly the same topology with a taxon-poor and a taxon-rich sampling, which is far from being the case at present (Figure 1). This is particularly important to be confident in the phylogenetic location of poorly diversified clades (e.g. *Amborella*, or monotremes). The stability of phylogenies in face of variation of species sampling will constitute one of the best guarantees that the non-phylogenetic signal has been correctly handled.

The correctness of inferences should also be verified *via* corroboration from independent sources (Miyamoto & Fitch 1995, Swofford 1991). When complete genomes are used, this may appear hopeless (even if internal verifications of homogeneity are possible). However, genomes can be conveniently subdivided into various character types that can be considered as more or less independent: oligonucleotide composition, sequences of orthologous gene, gene content, and gene order. If inferences based on these "independent" sets of characters converge to the same results, an increased confidence can be placed in the corresponding phylogeny, although the same bias can theoretically affect several approaches. We therefore encourage the developments of sophisticated probabilistic methods for all types of data, and not only for primary sequences.

In the long term, one might envision that the Tree of Life, or at least its

global scaffold, will be established within the next 10 years. Then, to numerous molecular systematists the important question would be: what next?

Acknowledgements

We wish to thank David Bryant, Emmanuel Douzery, David Moreira, Davide Pisani, Nicolas Rodrigue, Naiara Rodríguez-Ezpeleta, Béatrice Roure and Brad Shaffer for critical readings of the manuscript. This work was supported by operating funds from Genome Québec. H.P. is a member of the Program in Evolutionary Biology of the Canadian Institute for Advanced Research (CIAR), which is acknowledged for salary and interaction support. H.P. is also grateful to the Canada Research Chairs Program and the Canadian Foundation for Innovation for salary and equipment support.

Figure legends

Figure 1 Super-matrix, taxon sampling and super-tree in animal phylogenomics. A dataset constituted of 71 slowly evolving nuclear proteins corresponding to 20,705 amino acid positions (Philippe et al 2005) was used for phylogenetic analyses based on Maximum Likelihood (ML) with a separate WAG+F+Γ model. Panel (a) presents the ML tree obtained with 49 species (redrawn from Philippe et al 2005). This tree strongly supports the new animal phylogeny (Aguinaldo et

al 1997) dividing Bilateria into Deuterostomia and Protostomia which comprises Lophotrochozoa and Ecdysozoa. Note that the major division between Deuterostomia and Protostomia is supported by a 100% bootstrap value. In panel (b), the exclusion of only four close outgroup sequences (Choanoflagellata, Cnidaria and Ctenophora) creates an LBA artifact via the distant fungal outgroup leading to the successive early emergence of the fast evolving Platyhelmintes and then Nematodes plus Tardigrada. Note that the overall bootstrap support substantially decreases, requiring caution when the existence of a radiation is extrapolated from low statistical supports, since limited resolution can also be due to poor taxon sampling or to unreliable tree reconstruction methods (not shown). Branch lengths are drawn proportionally to evolutionary rates and the height of triangles represents the taxonomic diversity of the different groups. Panel (c) presents the supertree obtained by maximum parsimony analysis conducted with PAUP* (Swofford 2000) of the matrix representation of 71 source trees obtained from Bayesian analyses of individual genes with MrBayes (Ronquist & Huelsenbeck 2003) using a WAG+F+Γ model. This supertree recovers both the monophyly of Deuterostomia and Protostomia with strong bootstrap support, but fails to find Ecdysozoa and Lophotrochozoa. Panel (d) shows the supertree obtained on the same matrix from a Bayesian analysis using MrBayes and a simple two-state model. This supertree strongly supports the respective monophyly of Deuterostomia, Protostomia and Ecdysozoa, but not of Lophotrochozoa. Numbers on branches are bootstrap values (a-c) or posterior probabilities (d).

Literature cited

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, et al. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387: 489-93

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20: 407-15

Bapteste E, Boucher Y, Leigh J, Doolittle WF. 2004. Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol.* 12: 406-11

Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc. Natl. Acad. Sci. USA* 99: 1414-9

Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41: 3–10

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

Blaisdell BE. 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci U S A* 83: 5155-9

Blanchette M, Bourque G, Sankoff D. 1997. *Breakpoint phylogenies*. Presented at Eighth Genome Informatics Conference (GIW 1997)

Bourque G, Pevzner PA. 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.* 12: 26-36

Brauer MJ, Holder MT, Dries LA, Zwickl DJ, Lewis PO, Hillis DM. 2002. Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. *Mol. Biol. Evol.* 19: 1717-26

Brinkmann H, Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* 16: 817-25

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. accepted. An empirical assessment of long branch attraction artifacts in phylogenomics. *Syst. Biol.*

Brochier C, Bapteste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.* 18: 1-5

Brochier C, Philippe H. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature* 417: 244

Brown JR. 2003. Ancient horizontal gene transfer. *Nat. Rev. Genet.* 4: 121-32

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. *Nat. Genet.* 28: 281-5

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Wadell PJ. 1993.

Partitioning and combining characters in phylogenetic analysis. *Syst. Biol.* 42: 384-97

Burleigh JG, Mathews S. 2004. Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. *Am.J. Bot.* 91: 1599-613

Canback B, Tamas I, Andersson SG. 2004. A phylogenomic study of endosymbiotic bacteria. *Mol. Biol. Evol.* 21: 1110-22

Copley RR, Aloy P, Russell RB, Telford MJ. 2004. Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. *Evol. Dev.* 6: 164-9

Darlu P, Lecointre G. 2002. When does the incongruence length difference test fail? *Mol. Biol. Evol.* 19: 432-7

Daubin V, Gouy M, Perriere G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* 12: 1080-90

de Queiroz A, Donoghue MJ, Kim J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annu. Rev. Ecol. Syst.* 26: 657-81

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6: 361-75

Dolphin K, Belshaw R, Orme CD, Quicke DL. 2000. Noise and incongruence: interpreting results of the incongruence length difference test. *Mol. Phylogenet. Evol.* 17: 401-6

Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284: 2124-9

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

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-i21

Dowton M, Austin AD. 2002. Increased congruence does not necessarily indicate increased phylogenetic accuracy--the behavior of the incongruence length difference test in mixed-model analyses. *Syst. Biol.* 51: 19-31

Driskell AC, Ane C, Burleigh JG, McMahon MM, O'Meara B C, Sanderson MJ. 2004. Prospects for building the tree of life from large sequence databases. *Science* 306: 1172-4

Dufresne A, Garczarek L, Partensky F. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* 6: R14

Dutilh BE, Huynen MA, Bruno WJ, Snel B. 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J. Mol. Evol.* 58: 527-39

Edwards SV, Fertil B, Giron A, Deschavanne PJ. 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst. Biol.* 51: 599-613

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8: 163-7

Farris JS, Källerjo M, Kluge AG, Bult C. 1995. Testing significance of incongruence. *Cladistics* 10: 315-9

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27: 401-10

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet* 22: 521-65

Felsenstein J. 2004. *Inferring phylogenies*. Sunderland, MA, USA: Sinauer Associates, Inc. 645 pp.

Feng XZ, Buell DA, Rose JR, Waddell PJ. 2003. Parallel algorithms for Bayesian phylogenetic inference. *J. Parallel Distr. Com.* 63: 707-18

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19: 99-113

Fitch WM, Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet* 4: 579-93

Fitz-Gibbon ST, House CH. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res* 27: 4218-22.

Gatesy J, Baker RH, Hayashi C. 2004. Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. *Syst Biol* 53: 342-55

Gee H. 2003. Evolution: ending incongruence. *Nature* 425: 782

Goldman N. 1993. Statistical tests of models of DNA substitution. *J Mol Evol* 36: 182-98

Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49: 652-70

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2004. The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. *Mol Biol Evol* 21: 1445-54

Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* 47: 9-17

Gu X, Zhang H. 2004. Genome phylogenetic analysis based on extended gene contents. *Mol Biol Evol* 21: 1401-8

Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC. 2005. Whole-genome prokaryotic phylogeny. *Bioinformatics* 21: 2329-35

Herbeck JT, Degnan PH, Wernegreen JJ. 2005. Non-homogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the Enterobacteriales ({gamma}-Proteobacteria). *Mol. Biol. Evol.*

Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. 2003. Is sparse taxon sampling a problem for phylogenetic inference? *Syst Biol* 52: 124-6

Hirt RP, Logsdon JM, Jr., Healy B, Dorey MW, Doolittle WF, Embley TM. 1999. Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Natl Acad Sci U S A* 96: 580-5

Ho SY, Jermiin L. 2004. Tracing the decay of the historical signal in biological sequence data. *Syst Biol* 53: 623-37

Holder M, Lewis PO. 2003. Phylogeny estimation: Traditional and Bayesian approaches. *Nat. Rev. Genet.* 4: 275-84

House CH, Fitz-Gibbon ST. 2002. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J Mol Evol* 54: 539-47.
House CH, Runnegar B, Fitz-Gibbon ST. 2003. Geobiological analysis using whole genome-based tree building applied to the Bacteria, Archaea and Eukarya. *Geobiology* 1: 15-26
Hrdy I, Hirt RP, Dolezal P, Bardonova L, Foster PG, et al. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* 432: 618-22
Huelsenbeck JP, Bull JJ. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. *Syst Biol* 45: 92-8
Huson DH, Nettles SM, Warnow TJ. 1999. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *J. Comput. Biol.* 6: 369-86
Huson DH, Steel M. 2004. Phylogenetic trees based on gene content. *Bioinformatics* 20: 2044-9
Inagaki Y, Susko E, Fast NM, Roger AJ. 2004. Covarion Shifts Cause a Long-Branch Attraction Artifact That Unites Microsporidia and Archaebacteria in EF-1{alpha} Phylogenies. *Mol Biol Evol* 21: 1340-9
Keane TM, Naughton TJ, Travers SA, McInerney JO, McCormack GP. 2005. DPRml: distributed phylogeny reconstruction by maximum likelihood. *Bioinformatics* 21: 969-74
Kim J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Syst. Biol.* 45: 363-74
Kluge AG. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Syst. Zool.* 38: 7-25
Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431: 980-4
Korbel JO, Snel B, Huynen MA, Bork P. 2002. SHOT: a web server for the construction of genome phylogenies. *Trends Genet.* 18: 158-62
Lake JA, Rivera MC. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol Biol Evol* 21: 681-90
Larget B, Simon DL, Kadane JB, Sweet D. 2005. A Bayesian analysis of metazoan mitochondrial genome arrangements. *Mol. Biol. Evol.* 22: 486-95
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21: 1095-109
Lecointre G, Deleporte P. 2005. Total evidence requires exclusion of phylogenetically misleading data. *Zool. Script.* 31: 101-17
Lemmon AR, Milinkovitch MC. 2002. The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. *Proc Natl Acad Sci U S A* 25: 25
Lerat E, Daubin V, Moran NA. 2003. From Gene Trees to Organismal Phylogeny in Prokaryotes:The Case of the gamma-Proteobacteria. *PLoS Biol* 1: E19.
Lin J, Gerstein M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels.

Genome Res 10: 808-18.
Lockhart P, Steel M, Hendy M, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11: 605-12
Lockhart PJ, Larkum AW, Steel M, Waddell PJ, Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. USA* 93: 1930-4
Lopez P, Forterre P, Philippe H. 1999. The root of the tree of life in the light of the covarion model. *J. Mol. Evol.* 49: 496-508
Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, et al. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409: 610-4.
Matte-Tailliez O, Brochier C, Forterre P, Philippe H. 2002. Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol* 19: 631-9.
Misawa K, Janke A. 2003. Revisiting the Glires concept--phylogenetic analysis of nuclear sequences. *Mol Phylogenet Evol* 28: 320-7
Miyamoto MM, Fitch WM. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst.Biol.* 44: 64-76
Miyamoto MM, Koop BF, Slightom JL, Goodman M, Tennant MR. 1988. Molecular systematics of higher primates: genealogical relations and classification. *Proc. Natl. Acad. Sci. USA* 85: 7627-31
Moret BME, Tang J, T. W. 2005. Reconstructing phylogenies from gene-content and gene-order data. In *Mathematics of Evolution and Phylogeny*, ed. O Gascuel, pp. 321-52. Oxford: Oxford University Press
Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* 409: 614-8.
Nadeau JH, Taylor BA. 1984. Lengths of Chromosomal Segments Conserved since Divergence of Man and Mouse. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* 81: 814-8
Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. *Syst Biol* 53: 47-67
O'Brien SJ, Stanyon R. 1999. Phylogenomics. Ancestral primate viewed. *Nature* 402: 365-6
Olsen G. 1987. Earliest phylogenetic branching: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harb. symp. Quant. Biol.* LII: 825-37
Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol* 53: 571-81
Philip GK, Creevey CJ, McInerney JO. 2005. The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. *Mol. Biol. Evol.* 22: 1175-84
Philippe H, Douady CJ. 2003. Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol* 6: 498-505.

Philippe H, Germot A. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol Biol Evol* 17: 830-4

Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* 22: 1246-53

Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? *Curr Opin Genet Dev* 8: 616-23

Philippe H, Lopez P, Brinkmann H, Budin K, Germot A, et al. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc. R. Soc. Lond. BS* 267: 1213-21

Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol* 21: 1740-52

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21: 1455-8

Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda. *Systematic Biology* 53: 978-89

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* 13: 145-58

Qi J, Wang B, Hao BI. 2004. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *J. Mol. Evol.* 58: 1-11

Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, et al. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402: 404-7.

Ragan MA. 1992. Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. *Biosystems* 28: 47-55

Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314: 1041-52

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425: 798-804.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572-4.

Rosenberg MS, Kumar S. 2003. Taxon sampling, bioinformatics, and phylogenomics. *Syst Biol* 52: 119-24

Roy SW, Gilbert W. 2005. Resolution of a deep animal divergence by the pattern of intron conservation. *Proc Natl Acad Sci U S A* 102: 4403-8

Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguna J. 1999. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. *Science* 283: 1919-23

Salter LA. 2001. Complexity of the likelihood surface for a large DNA dataset. *Syst Biol* 50: 970-8.

Sanderson MJ, Purvis A, Henze C. 1998. Phylogenetic supertrees: assembling

the trees of live. *Tree* 13: 105-9
Sanderson MJ, Shaffer HB. 2002. Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics*: 49-72
Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R. 1992. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc Natl Acad Sci U S A* 89: 6575-9
Sankoff D, Morel D, Cedergren R. 1973. Evolution of 5S RNA and the nonrandomness of base replacement. *Nature New Biol* 245: 232-4
Sjolander K. 2004. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 20: 170-9
Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. *Nat Genet* 21: 108-10
Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, et al. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. *Trends Plant Sci* 9: 477-83
Soltis DE, Soltis PS, Zanis MJ. 2002. Phylogeny of seed plants based on evidence from eight genes. *American Journal of Botany* 89: 1670-81
Soltis PS, Soltis DE, Chase MW. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402: 402-4.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21: 456-63
Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant'? *Trends Genet* 21: 307-9
Steel M, Penny D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* 17: 839-50
Stefanovic S, Rice DW, Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? *BMC Evol Biol* 4: 35
Storm CE, Sonnhammer EL. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 18: 92-9
Suchard MA, Kitchen CM, Sinsheimer JS, Weiss RE. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst Biol* 52: 649-64
Swofford DL. 1991. When are phylogeny estimates from molecular and morphological data incongruent? In *Phylogenetic analysis of DNA sequences*, ed. MM Miyamoto, J Cracraft, pp. 295-333. New York: Oxford University Press
Swofford DL. 2000. PAUP*: Phylogenetic Analysis Using Parsimony and other methods. Sinauer, Sunderland, MA
Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In *Molecular systematics*, ed. DM Hillis, C Moritz, BK Mable, pp. 407-514. Sunderland: Sinauer Associates
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278: 631-7

Tekaia F, Lazcano A, Dujon B. 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res* 9: 550-7

Wang L-S, Warnow T. 2005. Distance-based genome rearrangement phylogeny. In *Mathematics of Evolution and Phylogeny*, ed. O Gascuel, pp. 353-83. Oxford: Oxford University Press

Wheeler WC. 2003. Iterative pass optimization of sequence data. *Cladistics* 19: 254-60

Wiens JJ. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. *Syst Biol* 52: 528-38.

Wiens JJ. in press. Can Incomplete Taxa Rescue Phylogenetic Analyses from Long-Branch Attraction? *Syst Biol*

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. *Syst Appl Microbiol* 14: 364-71.

Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. *Trends Genet* 18: 472-9.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* 1: 8.

Wolf YI, Rogozin IB, Koonin EV. 2004. Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res* 14: 29-36

Yang S, Doolittle RF, Bourne PE. 2005. Phylogeny determined by protein domain content. *Proc. Natl. Acad. Sci. USA* 102: 373-8

Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42: 587-96

Yoder AD, Irwin JA, Payseur BA. 2001. Failure of the ILD to determine data combinability for slow loris phylogeny. *Syst Biol* 50: 408-24.

York TL, Durrett R, Nielsen R. 2002. Bayesian estimation of the number of inversions in the history of two chromosomes. *Journal of Computational Biology* 9: 805-18

Zelwer M, Daubin V. 2004. Detecting phylogenetic incongruence using BIONJ: an improvement of the ILD test. *Mol. Phylogenet. Evol.* 33: 687-93

Zhaxybayeva O, Lapierre P, Gogarten JP. 2004. Genome mosaicism and organismal lineages. *Trends Genet* 20: 254-60

Zheng XH, Lu F, Wang ZY, Zhong F, Hoover J, Mural R. 2004. Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics* 21: 703-10

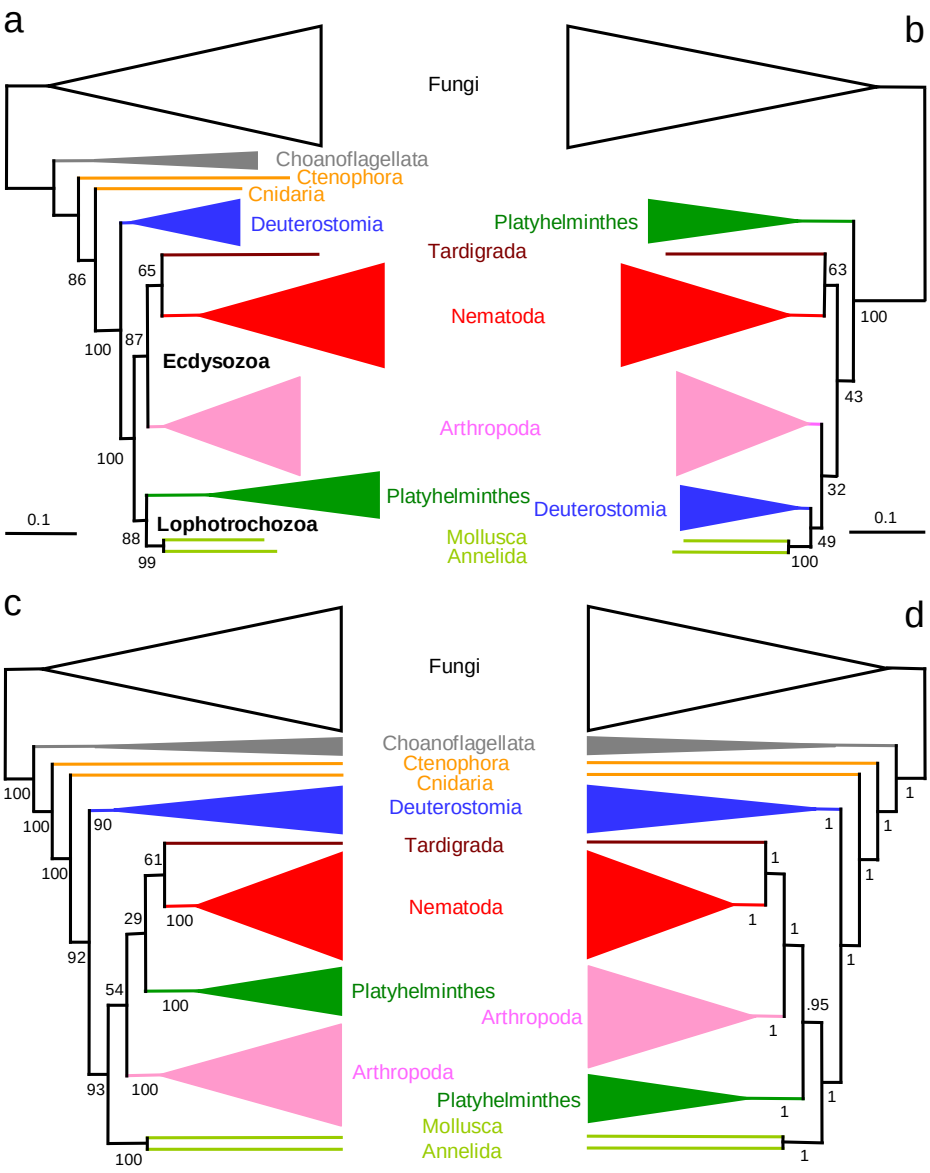

Figure 1: Philippe et al.