Causes, consequences and solutions of phylogenetic incongruence

Anup Som

Submitted: 24th January 2014; Received (in revised form): 5th April 2014

Abstract

Phylogenetic analysis is used to recover the evolutionary history of species, genes or proteins. Understanding phylogenetic relationships between organisms is a prerequisite of almost any evolutionary study, as contemporary species all share a common history through their ancestry. Moreover, it is important because of its wide applications that include understanding genome organization, epidemiological investigations, predicting protein functions, and deciding the genes to be analyzed in comparative studies. Despite immense progress in recent years, phylogenetic reconstruction involves many challenges that create uncertainty with respect to the true evolutionary relationships of the species or genes analyzed. One of the most notable difficulties is the widespread occurrence of incongruence among methods and also among individual genes or different genomic regions. Presence of widespread incongruence inhibits successful revealing of evolutionary relationships and applications of phylogenetic analysis. In this article, I concisely review the effect of various factors that cause incongruence in molecular phylogenies, the advances in the field that resolved some factors, and explore unresolved factors that cause incongruence along with possible ways for tackling them.

Keywords: molecular phylogeny; phylogenetic incongruence; phylogenomics; systematic error; heterotachy; phylogenetic network

INTRODUCTION

Recovering true historical relationships between any groups of species, genes or proteins is one of the principal goals of evolutionary research. Knowledge of phylogeny tells us about the pattern of evolutionary relationships, revealing the historical pattern of speciation and divergence and enabling us to classify life according to evolutionary scheme. Moreover, it is important because of its wide applications throughout biology such as in studies of human diseases: for epidemiological investigations [1–3], for identifying and characterizing newly discovered pathogens [4, 5] and for identifying and tracking natural reservoirs of zoonotic diseases [6]. In addition, phylogenetic analysis has been widely used to test the a priori hypothesis of epidemiological clustering in suspected transmission chains of HIV [7]. Phylogenetic analysis of molecular sequences is also one of the principal interpretive tools for understanding the organization and evolution of genes and genomes [8]. However, the success of revealing evolutionary relationships and applications of phylogenetic analysis greatly depends on the quality of the inferred phylogeny.

Despite immense progress in recent years, phylogenetic reconstruction involves many challenges that create uncertainty with respect to the true historical relationships of the organisms or genes analyzed [9]. One of the most notable difficulties is the widespread occurrence of incongruence (see glossary) among methods [10, 11] and also among individual genes or different genomic regions [12–15], and the existence of pervasive incongruences impedes achievement of the primary goals of evolutionary research. Evolutionary biologists expected that the advent of large volumes of data with sophisticated computational tools would resolve the problem. Surprisingly, the advent of molecular and genomic...
data has increased the variety of classifications rather than reducing the problem [10, 16]. Thus, despite the rise of phylogenomics [17], many important nodes remain unresolved [18]. In reality, large amounts of data with available efficient algorithms have made immense progress, but still many important nodes remain unresolved [15, 19]. This issue raises a critical question: why important nodes remain unresolved, and how do we overcome this problem. Molecular data are continuously accumulating at a rapid pace, analytical methods are refined and new ones are developed but time has not stopped. Incongruence is still a problem that scientists are trying to tackle.

In this review, I emphasize the various factors that cause incongruence in molecular phylogenies, the advances in the field that resolved some factors, and explore unresolved factors that cause incongruence along with possible ways of tackling them. Reviewing the causes, consequences and solutions of phylogenetic incongruence is a difficult task, and may be too big to be addressed in a single article. However, some of these causes, consequences and solutions have already been reviewed separately in the different literatures [11, 19–21], but no such effort to combine all the possible causes and their consequences has been concisely discussed in a single article.

**PRESENCE OF INCONGRUENCE: A SURVEY**

The widespread occurrences of phylogenetic incongruence among methods and also among genes are well-known [10, 11]. Although there is no doubt about the existence of statistically significant incongruence among different tree reconstruction methods, but statistically significant incongruence among genes has been questioned [10]. However, it has been accepted for a long time that gene trees could be incongruent with each other because of the genetic drift in natural populations, gene duplication, horizontal gene transfer or lineage sorting [10, 22]. In an investigation, Rokas et al. [13] has shown that the single-gene phylogenies reveal extensive incongruence when a particular tree reconstruction method is used. They analyzed 106 genes of eight yeast species and showed that there are several strongly supported (defined here as a bootstrap value >70%) alternative phylogenies. Analyses of 106 genes resulted in >20 alternative maximum likelihood (ML) or maximum parsimony (MP) trees. In a similar study, Kopp and True [12] also reported that single-gene datasets of *Drosophila melanogaster* species group produce several strongly supported conflicting clades, which were also consistently seen in combined analyses. Lack of resolution in combined data analyses stems from conflicting phylogenetic signal between individual genes either because of real differences in their evolutionary history or because of different statistical biases [25]. Many studies have also reported that there are several strongly supported trees among different genes [23–25]. On the other hand, Jeffroy et al. [10] have analyzed the same 106 genes assembled by Rokas et al. [13] from 14 yeast species and shown that although individual genes recover different trees, they are not statistically significant (i.e. no incongruence among the 106 individual genes), and stated ‘all single gene trees are different because of the predominant effect of stochastic error’. This result contradicted other results and also stands in sharp contrast because gene trees are different not only due to statistical biases, but also due to real differences in their evolutionary history [26, 27].

Despite evidence of incongruence in recent studies, a further investigation has been conducted using nucleotide sequence data published in the scientific literatures (Table 1). These multigene datasets were used for testing different evolutionary hypotheses, and therefore quality of data had been verified. Reanalysis of the multigene datasets reveal that the phylogenetic incongruence is present in both cases (i.e. among methods and also among genes) even when robust reconstruction methods (i.e. ML and Bayesian method) are used. It is admissible that the incongruence among methods is very common and statistically robust (Figure 1), which agreed with other observations. Figure 1 shows the subtrees of six-taxon phylogenies that are based on the concatenation of four genes of Pryer et al. [28] data. Trees were inferred using neighbor-joining (NJ) and ML methods under the generalized time reversible (GTR) model with rate heterogeneity ($\alpha = 0.2652$). Pryer et al. [28] data contained an alignment of 5072 nucleotide characters based on the concatenation of four genes from 35 taxa. In Figure 1, instead of 35 taxon trees, two six-taxon sub-trees have been shown where incongruence is significant. The ML tree shows that Pinus and Gnetum are monophyletic with 78% bootstrap value, whereas the NJ tree shows Pinus and Gnetum...
are paraphyletic. This result shows a distinct example of statistically significant incongruence among methods. On the other hand, topological differences among genes are also widespread, but most of them are lacking statistical support. In Figure 2, phylogenies obtained from a single gene dataset of Springer et al. [29] data. Gene trees were obtained from ATP7A (690 bp, 42 taxa and rate of heterogeneity $\alpha = 0.751$) and mt-RNA (1647 bp, 42 taxa and $\alpha = 0.751$) genes by using neighbor-joining–maximum composite likelihood (NJ-MCL) method of phylogenetic inference [36–38]. Gene ATP7A shows Elephant and Hyrax are monophyletic with 65% bootstrap support whereas mt-RNA tree shows that Elephant and Sirenian are monophyletic with 81% bootstrap support.

### Table 1: Information about the datasets used in this study, including source paper, number of taxa, sequence length, genes, organism and evolutionary model

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Reference</th>
<th>Taxa</th>
<th>Length</th>
<th>Gene(s)</th>
<th>Organism</th>
<th>Evolution model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pryer</td>
<td>[28]</td>
<td>35</td>
<td>5072</td>
<td>atpB + rbcL + rps4 + SSU-rRNA</td>
<td>Seed plant</td>
<td>GTR + G</td>
</tr>
<tr>
<td>Springer</td>
<td>[29]</td>
<td>42</td>
<td>19397</td>
<td>19 nuclear &amp; three mtRNA genes*</td>
<td>Placental mammal</td>
<td>FP4 + G</td>
</tr>
<tr>
<td>Murphy</td>
<td>[30]</td>
<td>66</td>
<td>9779</td>
<td>15 nuclear &amp; three mtDNA genes*</td>
<td>Placental mammal</td>
<td>K2P, HKY + G</td>
</tr>
<tr>
<td>Teeling</td>
<td>[31]</td>
<td>29</td>
<td>7187</td>
<td>A2AB + BRCAl + RAG1 + RAG2</td>
<td>Rhinolophid microbat</td>
<td>GTR + G + I</td>
</tr>
<tr>
<td>Yoder</td>
<td>[32]</td>
<td>31</td>
<td>4454</td>
<td>Cytochrome b, ND2, IRBP</td>
<td>Malagasy carnivore</td>
<td>GTR + G + I</td>
</tr>
<tr>
<td>Madsen</td>
<td>[33]</td>
<td>30</td>
<td>6518</td>
<td>A2AB + IRBP + vWF + rRNA</td>
<td>Placental mammal</td>
<td>HKY85</td>
</tr>
<tr>
<td>James</td>
<td>[34]</td>
<td>214</td>
<td>6436</td>
<td>18S + 28S + 5.8S + EFI$\alpha$ + RPBI + RPB2</td>
<td>Fungus</td>
<td>GTR + I + G</td>
</tr>
</tbody>
</table>

**Note:** "Genes name can be found in the corresponding reference article.

**Figure 1:** Incongruence among phylogenetic methods (Pinus and Gnetum case). Phylogenies based on the combined dataset (concatenation of four genes) containing alignments of 5072 nucleotide characters from 35 taxa [28]. Trees were inferred using (A) NJ and (B) ML methods under GTR model with rate heterogeneity ($\alpha = 0.2652$). The bootstrap supports of each tree are indicated to the left of the corresponding node; 1000 bootstrap supports for NJ method using MEGA4 [38] and 500 bootstrap supports for ML method using PAUP [35] were performed. The ML tree shows that Pinus and Gnetum are monophyletic with 78% bootstrap support, whereas the NJ tree shows that Pinus and Gnetum are paraphyletic. This result shows the statistically significant incongruence among methods. [Note: In this figure, instead of 35 taxa trees, two six-taxon subtrees are shown where incongruence was significant]

### FACTORS THAT CAUSE PHYLOGENETIC INCONGRUENCE

#### Analytical factors

The causes of negative impact of various factors are analytical and biological. Analytical factors affecting phylogenetic reconstruction include: (i) the choice of optimality criterion [39], (ii) taxon sampling [40, 41] and (iii) specific assumptions in the modeling of sequence evolution [42]. The existing probability-based optimality criteria (i.e. ML and Bayesian method) and substitution models are statistically robust, but they have limitations to represent reality [43–45]. It might be possible that small variation from reality makes the problem. Models of sequence evolution are designed to capture the most essential aspects of processes shaping sequence evolution. However, these models rely on assumptions that render them sensitive to a number of influences. For example, extreme variability in rates of evolution among lineages has been shown to lead to long-branch attraction [46]. Complex models of sequence evolution may better approximate the evolution of the sequences, and therefore might be expected to
give more accurate results. On the other hand, more complex models require the estimation of more parameters, each of which is subjected to some errors. Several studies reported that the simple models sometimes perform better than the complex models [47–49], and this finding highlighted the complexity of phylogeny reconstruction and demands for more theoretical work on statistical models [50, 51].

Appropriate and extensive taxon sampling is one of the most important determinants of accurate phylogenetic estimation. In phylogenetic analyses, the inclusion or exclusion of certain taxa may influence the accuracy of phylogenetic estimations. Considerable bodies of work have addressed the effect of taxon sampling on phylogenetic reconstruction, especially in relation to density of sampling, and have stated that an adequate increase of taxon sampling strongly improves the phylogenetic estimation [40, 52–55]. Heath et al. [56] performed a comprehensive study on taxon sampling that included an exploration of the effect and strategies of taxon sampling and concluded that there are many benefits of dense taxon sampling and advised evolutionary biologists to appreciate taxon sampling when interpreting phylogenetic analysis results to explain any unexpected patterns. For further details about the issue of taxon sampling and phylogenetic analysis readers are advised to go through the comprehensive review article by Nabhan and Sarkar [21].

**Biological factors**

Biological factors cause incongruence between two phylogenies because of (i) violations of the orthology due to three major mechanisms, namely, lineage sorting, hidden paralogy and horizontal gene transfer [57, 58]; (ii) stochastic error or character sampling bias related to the length of the genes; and (iii) systematic error due to the presence of a nonphylogenetic signal in the data [10].

A growing corpus of studies has revealed conflicting evolutionary histories among genes [11, 59]. The conflicting evolutionary histories among gene trees could mainly be due to coalescent stochasticity [11, 22, 60] when incomplete sorting of the ancestral polymorphism occurs during successive speciation events, leading to gene genealogies that differ from the species phylogeny (Figure 3). This phenomenon, called ‘incomplete lineage sorting’, has been detected in several different taxa [24, 60], including hominids [61]. A recent genome-wide analysis of hominid primates indicated that, as a consequence of incomplete lineage sorting, roughly 30% of our genome support the [chimpanzee, (human, gorilla)] or [human, (chimpanzee, gorilla)] branching order, i.e. topologies different from the true [gorilla, (human, chimpanzee)] species tree [60, 62]. A very different reason why gene trees can be truly incongruent is gene duplications and extinctions. If a dataset includes paralogous copies, then the true phylogeny will partly reflect the duplication history of the gene that is independent of species divergence history. The third
mechanism is horizontal gene transfer. If genetic exchanges occur between species, then the phylogeny of individual genes will be influenced by the number and nature of transfers they have undergone. It is well-known that the horizontal gene transfers are an important source of conflicts between gene trees in bacteria [63, 64]. Quantifying the amount of phylogenetic conflict caused by horizontal gene transfer versus other factors would appear worthwhile for a correct interpretation of bacterial phylogenomic data: how frequently, and under which conditions, should we invoke horizontal gene transfer? Horizontal gene transfer is expected to be very rare or absent in animals. Thus, phylogenetic agreement in animals is much more common than disagreement, indicating that horizontal gene transfer is not widespread enough to erase the vertical signal [11, 15]. Beside the three major mechanisms, hybridization, recombination and natural selection, are cited as potential causes of gene tree incongruence [11]. Figure 4 summarizes the negative factors that cause incongruence in species and genes trees.

INCONGRUENCE DUE TO SEQUENCE ALIGNMENTS

The reconstruction of phylogenetic history is predicated on being able to accurately establish hypotheses of character homology, and mistaken hypotheses of homology are a primary source of error in evolutionary studies. Phylogenies are often thought to be more dependent upon the specifics of the sequence alignment rather than on the method of reconstruction [65–67]. One of the earlier studies to see the effects of nucleotide sequences alignment on phylogeny, Morrison and Ellis [68] did a case study of 18s rDNAs of Apicomplexa and concluded that many of the literature disagreements concerning the phylogeny of the Apicomplexa are probably based on differences in sequence alignment strategies rather than differences in data or tree-building methods [68].

Over the decade, continuous efforts have been made to improve the quality of multiple sequence alignments, and as a result, several improved algorithms of sequence alignment have been developed such as SATCHMO [69], RASCAL [70], MUSCLE [71], MAVID [72] and CLUSTAL OMEGA [73]. Different alignment algorithms claimed high alignment accuracy, but a recent study finds the pitfalls of the existing algorithms and raised questions about the quality of multiple alignments and their effects on subsequent analyses [74, 75]. Multiple alignments are typically produced by feeding sequences into the different alignment programs from left to right. Landan and Graur [74] used existing alignment programs and instructed them to read the sequences both from left to right and also from right to left, and found that the two resulting alignments can be quite different and consequently produce different trees even though both alignments start with identical information and were processed by identical algorithms. Therefore, one of the potential sources of incongruence in phylogeny is due to artifacts in multiple sequence alignments. So what are the possible solutions for resolving incongruence cause by alignment artifacts? The problems are hard and so far no perfect remedy exists. However, a good approach is choosing highly conserved, recently diverged and long sequences where the effect of alignments in both directions seems to be relatively minor. Furthermore, for alignments that are not too short, removal of problematic regions (i.e. eliminate regions that cannot be aligned with confidence) leads to better trees [76, 77].

INCONGRUENCE CAUSED BY HETROTACHY

Heterotachy refers to the variation in the evolutionary rate of a given site (i.e. nucleotide/ amino acid) of a gene or protein through time. It has been convincingly demonstrated that the evolutionary rate of a
given site is not always constant throughout time. Functional constraints on sites in a gene sequence often change through time, causing shifts in site-specific evolutionary rates, a phenomenon called heterotachy [78]. This phenomenon has been recently confirmed as an important process of sequence evolution [79], and can lead to phylogenetic reconstruction artifacts [80]. Unlike other types of biases, heterotachy does not leave any evident footprints in sequences, and therefore leads to artifacts that are particularly difficult to detect [81, 82].

Recently, several simulation studies have been conducted to compare the relative performance of the phylogenetic inference methods under the heterotachous conditions, and studies have shown that ML outperforms MP, but that in the majority of cases, both methods failed to recover the correct topology under the heterotachous conditions [51, 80, 83–86]. Recent controversies over the relative performances of the ML and MP methods can not address the important issue of how to reconstruct a correct phylogeny when the concatenations suffer from heterotachy. Particularly, adopting a genome-scale approach increases the chance of heterotachy because genes and genomes are unlikely to evolve in a similar way as we concatenate sequence data from species that are extremely distantly related or have large differences in their environmental habitat, biological composition and life history traits.

So, what are the implications of these findings? Are all previously accepted multigene trees in need of reinspection? To find an acceptable answer, a simulation study was conducted to measure the performance of the phylogenetic inference methods when different level of heterotachy is present in the data. The simulation strategy of Kolaczkowski and Thornton [80] was used to introduce heterotachy in the data. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

**Figure 4:** A summarization of the negative factors that cause incongruence in species and genes trees.

**Figure 5:** A histogram plot shows the performance of the phylogenetic methods with the variation of heterotachous signal (see text). The simulation strategy of Kolaczkowski and Thornton [80] was used to introduce heterotachy in the data. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.
Box I: Glossary

**Molecular phylogenetics:** the study of evolutionary relationships among organisms or genes by a combination of molecular biology and statistical techniques.

**Phylogenetic incongruence:** two (or more) phylogenetic trees are said to be incongruent when they exhibit conflicting branching orders (i.e. topologies) and can not be superimposed. This implies that at least one node (also known as a bipartition) present in one tree is not found in the other(s), where it is replaced by alternative groupings of taxa.

**Phylogenetic signals:** tendency for related taxa to resemble each other more than they resemble species drawn at random from the tree. The strength of the phylogenetic signal is proportional to the number of substitutions occurring along the branch.

**Phylogenomics:** reconstruction of phylogenies using a large number of genes or genomics regions. Two fundamentally different approaches are used for reconstructing phylogenies from multiple datasets. In one, the phylogenetic reconstruction is done after the gene sequences are concatenated head-to-tail to form a super-gene alignment—called ‘supermatrix’ approach. In the other, phylogenies are inferred separately for each gene and the resulting gene trees are used to generate a majority rule consensus phylogeny—called ‘supertree’ approach. The size of homologous sequence datasets has increased dramatically in recent years, and many of these datasets now involve few hundred to several thousand species or genes [16, 18, 108].

**Model of sequence evolution:** a statistical description of the process of substitution in nucleotide or amino acid sequences. Complex models better approximate the evolutionary process but at the expense of more parameters and computational time. As parameter-rich models require more data to behave properly, they have become really useful with the advent of phylogenomic datasets.

**Gene tree:** a phylogenetic tree that shows the evolutionary relationships between a group of genes.

**Species tree:** a phylogenetic tree that shows the evolutionary relationships between a group of species.

**Paralogy:** two genes are said to be paralogous when they derive from a gene duplication event.

**Horizontal gene transfer:** the transfer of genetic material from one organism to another one that is not its offspring.

**Incomplete lineage sorting:** the failure of two or more lineages in a population to coalesce, leading to the possibility that at least one of the lineages first coalesces with a lineage from a less closely related population.

**Ancestral polymorphism:** the variants that arose by mutation prior to the speciation event that generated the species in which they segregate.

**Coalescent:** the process of joining ancestral lineages when the genealogical relationships of a random sample of sequences from a modern population are traced back.

**Heterotachy:** refers to the fact that the evolutionary rate of a given nucleotide/ amino acid position varies throughout time.

**Long-branch attraction:** the tendency of species at the ends of long branches in a phylogenetic tree to be grouped artificially close to each other. Long branches can result from a fast evolutionary rate and/or a long time span. In particular, the branch of the outgroup is, in essence, long and often artifactualy attracts fast-evolving lineages.

**Stochastic error:** the error in phylogenetic estimation caused by the finite length of the sequences used in the inference. As the size of the sequences (n) increases, the magnitude of the error decreases (stochastic error $\propto \frac{1}{\sqrt{n}}$).

**Systematic error:** the error in phylogenetic estimation that is due to the failure of the reconstruction method to account fully for the properties of the data.

**Natural selection:** the process in nature by which, according to Darwin’s theory of evolution, only the organisms best adapted to their environment tend to survive and produce more offspring while those less adapted tend to be eliminated.

**Hybridization (DNA):** the process of combining two complementary single-stranded DNA or RNA molecules and allowing them to form a single double-stranded molecule through base pairing.

**Recombination:** the exchange of a segment of DNA between two homologous chromosomes during meiosis leading to a novel combination of genetic material in the offspring.

**Monophyletic:** monophyletic taxa include all the species that are derived from a single common ancestor.

**Polyphyletic:** taxa is composed of unrelated organisms descended from more than one ancestor.

**Optimality criteria:** an objective function that is used to evaluate a given tree. The tree that maximizes or minimizes the function is chosen as the best estimate of phylogeny.

**Phylogenetic networks:** a phylogenetic network is any graph used to visualize evolutionary relationships (either abstractly or explicitly) between nucleotide sequences, genes, chromosomes, genomes or species. They are used when reticulate events such as hybridization, horizontal gene transfer, recombination or gene duplication and loss are believed to be involved.
situation is not so dire because when we are reconstructing a large-scale phylogeny by concatenating several genes, there is a fair chance of occurring heterotachy. However, for a set of conserved homologous genes, level of heterotachy is expected to be low or intermediate, and under such situations ML and BI are preferred methods to avoid incongruence caused by heterotachy signal. Recently, two models of sequence evolution have been proposed for dealing with heterotachy. Pagel and Meade [87] proposed a general likelihood-based covarion ‘mixture-model’ for inferring phylogenetic tree from multigene data that suffer from heterotachy. Another model is the site-heterogeneous mixture model, named CAT, was developed for dealing with heterotachy [88]. However, it has been claimed that these models accommodate cases in which different sites in the alignment evolve with different rates, but evidently these models of evolution could not detect or overcome heterotachy [30].

Although probabilistic methods (i.e. BI and ML) are very efficient up to certain levels of heterotachous signal. The existence of heterotachy compels us to improve the phylogenetic reconstruction methods. So far, there is no such way to accommodate heterotachous signal in the methods of phylogenetics inferences. Moreover, it is well-known that existence of heterotachy dramatically affects phylogenetic accuracy [80]. Therefore, phylogenetic incongruence might be because of heterotachy, and phylogenetic inference methods should be improved so that methods can successfully accommodate within-site rate variation (i.e. heterotachous signals), and successful modeling of within-site rate variation will help to resolve many unresolved nodes.

RESOLVING INCONGRUENCE USING PHYLOGENOMICS APPROACH

To resolve the effect of analytical and biological factors, evolutionary biologists had tried to reconstruct a true (or at least a fair) evolutionary history by analysis of concatenated datasets [13, 16, 29, 89–93]. Concatenated datasets offer two major advantages. First, our confidence in the inferred relationships is greatly improved when they are supported by several independent datasets. Second, each individual dataset may not contain sufficient information to resolve all relationships; therefore, combining the data may increase the amount of phylogenetic signal and provide the necessary resolution. But a problem occurs when different genes support conflicting phylogenies. Perfect agreement among the gene-specific phylogenies is rare, and therefore phylogenetic analyses of different sets of concatenated genes often do not converge on the same tree [26, 27]. There is growing evidence that trees built from concatenated sequences have to be treated with caution [14, 27, 93, 94]. Some studies have yielded results at odds with widely accepted phylogenies [95].

Over the past two decades, tremendous progress in this field has resulted in resolving various analytical and biological factors that cause incongruence in phylogenies. Improvement of phylogenetic inference methods, probabilistic models of sequence evolution and genome-scale approach (i.e. phylogenetic reconstruction is done after the gene sequences are concatenated head-to-tail to form a super gene alignment) have resolved a significant amount of incongruence, but still there are large number of incongruences among methods and also among genes [10–15]. Adopting phylogenomics approach resolved two biological factors. Those are violation of orthology assumption and stochastic error. The nonorthologous comparisons are gene-specific and are buffered in a multigene analysis; and stochastic error is related to gene lengths, that is, naturally vanishes when more and more genes are concatenated. On the other hand, systematic errors (i.e. nonhistorical signals where reality differs from simplified models) are not expected to disappear with the addition of data. Moreover, longer sequences exacerbate the potential for biases (systematic error) to be positively misleading. Systematic errors result from nonphylogenetic signals being present in the data; those are compositional signal due to heterogeneity of nucleotide compositions among genes or species, rate-heterogeneity signal (i.e. rate variation across lineages) and also heterotachous signal due to within-site rate variation [19]. The bias causing systematic errors creates an erroneous signal that could dominate the true phylogenetic signal causing the tree reconstruction method to be inconsistent and lead to an incorrect, but highly supported tree [96].

APPROACHES TO OVERCOME SYSTEMATIC ERRORS

Phylogenetic inference from nucleotides or amino acids can be misled by both sampling (stochastic) error and systematic error. Systematic errors are due
to heterogeneity of nucleotide compositions, rate variation across lineages and also within-site rate variation. Heterogeneity of nucleotide compositions causes incorrect phylogenetic inference because unrelated clades with similar nucleotide frequencies (due to convergence rather than shared ancestral frequencies) will have greater similarity and may group together in a phylogenetic analysis, sometimes with strong statistical support. Several methods have been developed to account for the problem of compositional heterogeneity among sequences, including distance approaches [97–99], parsimony approaches [100] and maximum-likelihood approaches [101]. Recently, the RY-coding strategy (coding the nucleotides as purines and pyrimidines) was recognized a way to deal with compositional signal [96]. Indeed, the G + C content can be extremely variable among homologous sequences from various organisms, whereas the frequency of purines is remarkably homogeneous [102]. This constitutes a method of choice to avoid incongruence resulting from compositional bias. A recent research has reported that GC-rich genes induced a higher amount of conflict among gene trees and performed worse than AT-rich genes in retrieving well-supported, consensual nodes on the placental tree [103]. This GC effect is mainly as a consequence of genome-wide variations in recombination rate. Indeed, recombination is known to drive GC-content evolution through GC-biased gene conversion and might be problematic for phylogenetic reconstruction, for instance, in an incomplete lineage sorting context. Beside compositional heterogeneity, rate variation across lineages has been well-studied and successfully modeled by the use of the gamma distribution [104, 105]. The mentioned implementations improved the reconstruction of the phylogenies, especially when very divergent species or genes were included and long-branch attraction artifacts were widespread.

As part of the continuous effort to resolve systematic errors, recently, several complementary approaches have been applied, namely, (i) variation in taxon sampling, (ii) removal of fast-evolving species, genes or sites [106–108] and (iii) allowing a more efficient detection of multiple substitutions. In addition to the mentioned complementary approaches, amino acid coding into functional categories has potential to overcome some types of systematic errors in genome-scale datasets [30].

Although recent studies indicated that reconstructing phylogeny from alignments of concatenated genes greatly reduces the stochastic error, the potential for systematic error still remains, heightening the need for reliable methods to analyze multigene datasets. Recently, a new methodology, called ‘phylogenetic networks’, has been adopted to deal with incongruences of gene trees [109, 110]. Phylogenetic networks provide an alternative to phylogenetic trees when analyzing datasets whose evolution involves significant amounts of reticulate events such as horizontal gene transfer, hybridization or recombination [111, 112]. However, it has been claimed that the phylogenetic networks approach has potential to visualize contradictory evidence for species phylogenies and incorporate additional information [113], but evidently this approach needs more theoretical work to establish an efficient network algorithm and simultaneously need comprehensive tests using large-scale data to assess the accuracy of this approach. These further analyses will find out the potential applications and pitfalls of phylogenetic networks approach.

**CONCLUSIONS, IMPLICATIONS AND FUTURE RESEARCH**

Although recovering a highly resolved tree of life is yet to be achieved, wealth of genomic data with probabilistic tree reconstruction methods and complex models of sequence evolution have taken phylogenetic analysis to a new height. Furthermore, cautionary application of several contemporary approaches such as variation of taxon sampling, removal of fast evolving species (genes or sites), removal of ambiguously aligned regions or use of amino acid sequences instead of nucleotides resolved several uncertainties. For further improvement, an accurate probabilistic modeling of heterotachy (i.e. within-site rate variation) is needed. In the future, it should be examined whether the phylogenomics approach itself can be fixed in case of heterotachy. To achieve this goal, it would be necessary to define a model of sequence evolution that considers heterotachy and to implement it in phylogenetic reconstruction methods. Secondly, accurate establishment of character homology (i.e. multiple sequence alignment) is an essential requirement. Evolutionary biologists should pay great attention to improve the quality of multiple sequence alignments because recent findings have demonstrated that the existing alignment algorithms fail to accurately establish character homology and mistaken hypotheses of
homology are a primary source of error in phylogenetic analysis.

Despite several known factors, there are possibly other hidden factors that may cause incongruence, which prevent the achievement of Darwin’s dream of having ‘...true genealogical trees for each great kingdom of Nature...’. Evolutionary biologists should explore the possibility of hidden negative factors and find out their best possible solutions.

**Key Points**

- Resolving phylogenetic incongruence is not an easy problem; particularly, the problem becomes more complicated when attempts of resolving one negative factor may introduce a new negative factor that causes incongruence.
- Adopting phylogenomics approach resolved few negative factors that caused phylogenetic incongruence, but simultaneously increased the chance of systematic error; particularly, heterochronous signal that is not expected to disappear with the addition of data.
- Toward achieving the goal of recovering a highly resolved tree of life, accurate establishment of character homology (i.e. multiple sequence alignment) is an essential requirement.
- Beside the known negative factors, there are possibly other hidden factors that may cause incongruence; therefore, evolutionary biologists should explore the possibility of hidden negative factors and find out their solutions.

**FUNDING**

The work was supported by the Center of Bioinformatics, University of Allahabad.

**Acknowledgements**

I thank Dr. Sudhir Kumar of Arizona State University and Dr. Dan Graur of University of Houston for useful discussions. I also thank the anonymous reviewers for their constructive comments.

**References**


