# An Integrated Perspective on Phylogenetic Workflows

**Article** *in* Trends in Ecology & Evolution · January 2016

DOI: 10.1016/j.tree.2015.12.007



**5 authors**, including:

August Guang
Brown University
**2** PUBLICATIONS **1** CITATION
SEE PROFILE

Mark Howison
Brown University
**36** PUBLICATIONS **385** CITATIONS
SEE PROFILE

Charles Lawrence
Brown University
**136** PUBLICATIONS **9,109** CITATIONS
SEE PROFILE

Casey W Dunn
Brown University
**199** PUBLICATIONS **3,700** CITATIONS
SEE PROFILE

Trends in Ecology & Evolution

CellPress

Opinion

# An Integrated Perspective on Phylogenetic Workflows

August Guang,[1,2,4] Felipe Zapata,[2,4] Mark Howison,[3] Charles E. Lawrence,[1] and Casey W. Dunn[2,*]

**Molecular phylogenetics is the study of evolutionary relationships between biological sequences, often to infer the evolutionary relationships of organisms. These studies require many analysis components, including sequence assembly, identification of homologous sequences, gene tree inference, and species tree inference. At present, each component is usually treated as a single step in a linear analysis, where the output of each component is passed as input to the next as a point estimate. Here we outline a generative model that helps clarify assumptions that are implicit to phylogenetic workflows, focusing on the assumption of low relative entropy. This perspective unifies currently disparate advances, and will help investigators evaluate which steps would benefit the most from additional computation and future methods development.**

## Trends

Current phylogenetic analyses are implemented as multistep, linear workflows where intermediate analysis steps generate and pass on point estimates of unobserved variables. This linear structure and minimal information communication strategy embody three implicit assumptions: (i) the order of the analysis steps is biologically justified, (ii) a Markovian dependency structure, and (iii) low relative entropy of results of each analysis step.

There is evidence that these assumptions, in particular low relative entropy, are frequently violated in empirical studies with potential detrimental effects in phylogenetic analyses.

A generative model and probabilistic framework provide a unified perspective to assess the costs and benefits of relaxing these assumptions, help identify what methods and tools are missing, and provide a context for evaluating priorities for future development.

## Molecular Phylogenetics and Information Communication

Molecular phylogenetic analyses have multiple components, including **assembly** (see Glossary) of raw sequence data into gene sequences, identification of **homologous sequences** across and within species, **multiple sequence alignment**, inference of gene trees, and integration of information across gene trees to infer species trees. How each of these **analysis components** is implemented and integrated is an important decision in designing a phylogenetic study [1]. In most phylogenetic analyses, each analysis component is treated as a separate step in a linear workflow, in which results from each component are passed as input to the next component. Each result is usually communicated as a **point estimate**. For example, a single hypothesis on gene homology is estimated and passed on to indicate which gene sequences belong to the same gene tree.

There are multiple limitations, each of which reflects simplifying assumptions about the data and methods, with phylogenetic workflows that pass only a single hypothesis in a single direction between each analysis component. These workflows cannot accommodate the uncertainty present in the data or introduced in the inference process. In addition, a strictly stepwise workflow does not allow for interactions between analysis components when different stages cannot be solved independently. Biologists have long recognized these limitations in the context of particular analysis steps, including identification of homologous sequences [2], multiple sequence alignment [3], and species tree inference [4]. This has spurred important methods development to relax some of these limitations, including the use of Bayesian approaches to accommodate uncertainty when inferring phylogenetic trees from multiple sequence alignments [5] and the simultaneous estimation of gene and species trees [6–9].

While there has been much productive work on understanding and relaxing these limitations at particular points in phylogenetic analyses, there has been little work on the systematic evaluation of these limitations across the entire phylogenetic workflow. An integrated workflow perspective can make several types of important contributions. First, a better understanding of workflow

[1]Department of Applied Math, Brown University, 182 George St, Providence, RI 02906, USA
[2]Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman St, Providence, RI 02912, USA
[3]Computing and Information Services, Brown University, 3 Davol Sq., Providence, RI 02903, USA
[4]These authors contributed equally

*Correspondence: casey_dunn@brown.edu (C.W. Dunn).

assumptions will lead to better interpretation of the results of current tools. Second, this integrated perspective can provide clear criteria for prioritizing future methods development. In particular, improving communication (Box 1) between largely independent analysis components can relax some simplifying assumptions. Improved communication reduces the accumulation of errors across components [10], accommodates interactions between components, provides more information to the investigator, and enables statistically grounded interpretations of results. These improvements come at the cost of increased engineering and computational costs, and this integrated perspective provides a way to better evaluate the trade-offs between these costs and the benefits to the investigator.

Here we present a unified framework for understanding information communication in molecular phylogenetic analyses that brings together advances on particular analysis components, helps identify what methods and tools are missing, and provides a holistic context for evaluating which of these should be the highest priority for future development. This will lead to more informed decisions about how to allocate computational resources to different analysis components to maximize investigator insight. This framework is grounded in a **generative model** for raw sequence reads that reflects the biological and technical processes that underlie the observed data. This model provides the unified perspective needed to state the assumptions implicit to information communication between phylogenetic analysis components (Box 2), and places

### Box 1. Three Communication Strategies in Data Analyses

Most data analyses, including phylogenetic analyses, require multiple components that each address a subproblem in a larger analysis challenge. These analysis components often consist of separate tools that were designed to work together or were repurposed in new ways so that they could be combined into novel workflows. Data and intermediate analysis results must be communicated between these components for them to work together. While much effort has been put into the effectiveness and efficiency of each of these components, how they communicate and what information they share is a major scientific and engineering challenge that often receives less technical and theoretical attention. Here we consider three specific communication approaches between analysis components in scientific computing.

The first is to communicate a minimum amount of information by propagating a point estimate in a single direction. Sometimes this hypothesis is chosen according to *ad hoc* criteria, and other times it is statistically explicit (such as the selection of a phylogenetic hypothesis by maximum likelihood). This is the most common communication strategy currently implemented in phylogenetic workflows.

The second approach is to communicate more information by propagating multiple hypotheses in a single direction. This allows uncertainty and error to be communicated from one analysis component to another. In phylogenetics, this approach has been used for inferring a species tree given a sample of gene trees [80]. This strategy has also been explored in other fields, such as cognitive psychology and hydrological modeling [81,82].

The third approach is to combine analysis components into a single component that simultaneously infers multiple hypotheses that were otherwise estimated independently and sequentially. This approach fully accommodates the uncertainty present in the data and generated during the inference process, as well as the non-independence of solutions between analysis components. Some phylogenetic species tree methods use this approach [8].

Changing the information communicated between analysis components can require that the analysis components themselves are re-engineered. Many tools that are critical to phylogenetic analysis are built to input and output only point estimates. Most assemblers output single assembly hypotheses. Multiple sequence aligners accept only a single estimate of each gene sequence and output only one possible multiple sequence alignment. The simplest approach to implementation is to iteratively run these existing tools on a distribution of inputs, and then propagating the distribution of outputs to later steps. This comes with minimal implementation costs since the tools can be used as-is, but is computationally very expensive. To make improved communication tractable, tools will need to be re-engineered to intrinsically accommodate multiple hypotheses or to simultaneously infer multiple steps.

The way that hypotheses are chosen has important implications for their interpretation, regardless of how those hypotheses are communicated to other analysis steps. Hypotheses are often chosen according to *ad hoc* criteria, such as minimizing gap penalties in multiple sequence alignment. If they are instead selected according to statistically explicit criteria, such as an approximation of the posterior distribution under an explicit probabilistic model, their interpretation is much clearer.

## Glossary

**Analysis components:** specific inference processes within an analysis workflow to estimate random variables. In a phylogenetic analysis, these components include assembly, homology evaluation, multiple sequence alignment, gene tree inference, and species tree inference.
**Assembly:** the technical process of aligning and merging fragments of DNA sequences to estimate the original DNA sequence.
**Generative model:** a hypothesis of how the observed data are generated through a joint probability distribution of all random variables of interest.
**Homologous sequences:** genes, or stretches of DNA, that are descendants from a common ancestral sequence. Homologous sequences are drawn from the same gene tree.
**Homology evaluation:** in this context, the technical process of identifying homologous sequences, typically through phenetic comparisons of sequence similarity as a proxy for evolutionary descent from a common ancestor.
**Incomplete lineage sorting:** if the interval between speciation events is shorter than the time it takes for gene lineages to go to fixation in the population, then the phylogenetic relationships between gene lineages will not necessarily reflect species relationships.
**Lineage sorting:** the biological process by which lineages are sorted between populations. In a speciation event, for example, an original population is split into two, and only a subset of lineages from the original population are sorted into the populations of each descendent species. In this way, a speciation event can impose structure on the lineages in a gene phylogeny.
**Markovian dependence:** a dependency structure such that inference for the current state depends directly only on neighboring states.
**Multiple sequence alignment:** the inference of site homology within sets of homologous sequences by partitioning sequence variation into indels (insertions and deletions) and site substitutions.
**Point estimate:** a single hypothesis that is used to summarize the distribution of the results of an analysis component.

these assumptions in the context of a general probabilistic framework that facilitates statistical interpretation of results (Box 3).

## A Molecular Phylogenetic Generative Model Describes How Unobserved Processes Generate Observed Sequence Data

A generative model is an explicit hypothesis of the natural and technical processes that produce observed data. Because generative models are based in probability theory (Box 3), they provide principled means to describe the entities and processes investigators seek to understand. An explicit generative model therefore clarifies the goals and role of each process in an analysis workflow. Generative models can also be used to guide simulation, which is helpful to evaluate methods and make decisions about project design before data collection. Generative models have been employed with considerable success in many fields [11–17].

A generative model can be represented graphically (Figure 1, orange). The nodes are **random variables** that describe the observed data and unobserved biological entities that the

**Random variable:** variable that takes on values randomly according to a probability distribution.
**Relative entropy:** measure of the amount of information lost when approximating the probability distribution $P$ with the distribution $Q$. When $Q$ is a good approximation of $P$, relative entropy is low; when $Q$ is not a good approximation of $P$, relative entropy is high. Here, $Q$ is often a single point estimate.
**Sensitivity analysis:** an analysis of the impact on the results from changing data, models, methods, or other features of analysis.

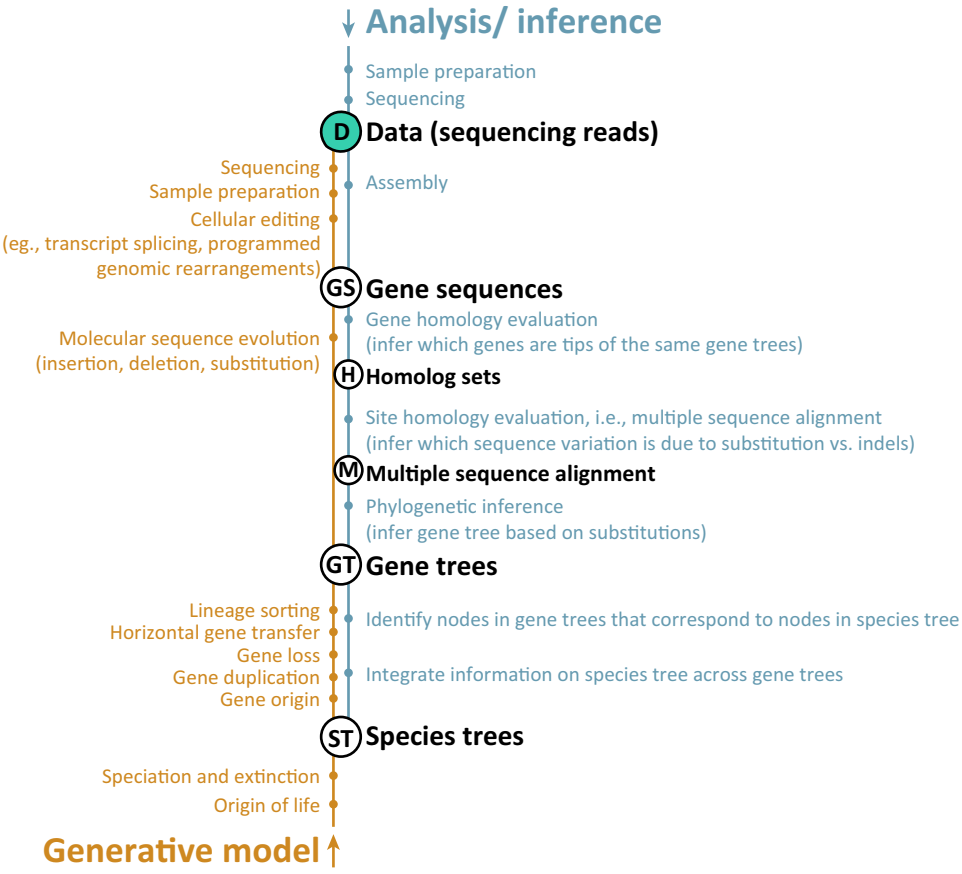


Figure 1. Graphical Representations of a Generative Model and Phylogenetic Analyses, under the Assumption of Markovian Dependence. Large open circles indicate biological entities, which can be described as random variables in the inference process. Edges (lines) represent processes. In the case of the generative model (orange), these are biological and technical processes that give rise to the entities. In the case of the inference process (blue), they are analysis processes that consist of one or more analysis components. Random variables that represent intermediate analysis products that are specific to inference are shown with small circles. The assumption of Markovian dependence specifies that variables depend only on neighboring variables, relaxing it would give a more general model wherein there are more connections in addition to those between adjacent variables.

### Box 2. Implicit Assumptions of Phylogenetic Workflows

All phylogenetic analyses make assumptions to simplify implementation and reduce computational costs. Often these assumptions are implicit. Making them explicit helps the investigator interpret results, understand the limitations of different approaches, better understand the differences between methods, and prioritize methods development. Here we describe three assumptions that are central to phylogenetic workflows.

**Assumption 1. Biologically Sensible Ordering of Steps**

The generative model is a hypothesis of the technical and biological processes of data generation. The phylogenetic analysis workflow attempts to reverse this generative process to reconstruct biological unknowns from the data. To be biologically justified, the ordering of analysis steps should therefore be the reverse of the processes in the generative model. The ordering of analysis steps therefore reflects assumptions about the generative model. The ordering of steps in the generative model presented here (Figure 1) is based on extensive understanding of the processes at play, so the ordering of analysis steps is well justified.

**Assumption 2. Markovian Dependency**

In probability theory, a Markov process is a process such that inference of the current state depends directly only on its neighboring states. A linear phylogenetic analysis workflow assumes a Markovian dependency structure because each analysis component only needs information from the previous component and/or the next component. For example, under Markovian dependency homology can theoretically be inferred from only the assembly and/or the multiple sequence alignment, and the multiple sequence alignment can be inferred from only the homology and/or the gene tree.

**Assumption 3. Low Relative Entropy**

Passing only a single hypothesis from one analysis component to another assumes that the result of an analysis component can be reasonably summarized by a single point estimate. This depends critically on the probability distribution of the result. Some probability distributions can be adequately summarized by a single estimate, others cannot. If the resulting distribution is unimodal and has low variance, then a single estimate such as the mean is a good approximation of other results that would be drawn from the distribution if it were resampled (Figure 2). If instead the resulting distribution is multimodal and has high variance, then repeat sampling would generate widely different results and none of them is a good description of the distribution as a whole. This has been shown for phylogenetic trees and for local sequence alignment [83,84].

In information theory, the concept of relative entropy [50,85] describes these differences in terms of how well one distribution is represented by another. It helps us understand when a point estimate is sufficient to describe a distribution and when more information is required. A low relative entropy means that the representation distribution is close to the target distribution (Figure 2). Current phylogenetic analysis workflows assume low relative entropy for most analysis components because they pass only a point estimate (e.g., a single multiple sequence alignment) from one component to another.

investigator would like to know about. The edges are natural and technical processes that connect these entities and describe their dependence.

A generative model for molecular phylogenetics must incorporate multiple random variables as well as evolutionary, cellular, and technical processes (Figure 1, orange). Explicit generative models have already been described for some subcomponents of a phylogenetic workflow, such as the generation of gene sequences given a phylogenetic tree and an explicit model of molecular evolution [18], or the generation of gene trees given a species tree [9,10,19]. A generative model can be extended to an entire phylogenetic analysis, from species tree through to the generation of raw sequence data. To simplify presentation, we here assume **Markovian dependence** (Box 2) in the generative model, which gives it a linear structure. The origin of life and speciation–extinction result in species trees that describe the structure of populations of individuals (*ST*, Figure 1). The origin of new genes (e.g., gene duplication, gene loss, and horizontal gene transfer between species) results in gene trees (*GT*, Figure 1) [20]. The species trees impose structure on these gene trees through **lineage sorting** [19]. If lineage sorting is complete and there is no horizontal gene transfer, then nodes in the gene trees correspond directly to nodes on the species tree. Otherwise there is not a one-to-one correspondence [21]. Processes of molecular evolution, including insertion, deletion, and substitution, result in the diversity of gene sequences (*GS*, Figure 1) present at the tips of these gene trees. These sequences can be further edited within the lifespan of an organism through cellular processes such as programmed genome rearrangements [22,23] and RNA splicing in transcriptomes.

### Box 3. A Probabilistic Perspective on Integrated Phylogenetic Analyses

The five random variables in the phylogenetic analysis workflow (Figure 1) - *GS* (gene sequences), *H* (homology), *M* (multiple sequence alignment), *GT* (gene trees), *ST* (species tree) - all concern high dimensional unknowns. A fully general model assumes that these five random variables are all directly related to one another. For example, to compute the probability of any random variable given *D* (data), say *ST*, we have to sum over all the values of the other unknowns:

$$P(ST|D) = \sum_{GS}\sum_{H}\sum_{M}\sum_{GT} P(GS,H,M,GT,ST|D)$$

$$= \sum_{GT} P(ST|GT,M,H,GS,D)\sum_{M} P(GT|M,H,GS,D)\sum_{H} P(M|H,GS,D)\sum_{GS} P(H|GS,D)P(GS|D)$$

Current linear phylogenetic workflows make the assumption that these random variables form a Markov chain (Box 2). Specifically, this model specifies that

$$P(GS,H,M,GT,ST|D) = P(ST|GT)P(GT|M)P(M|H)P(H|GS)P(GS|D)$$

and that for any particular random variable, for example *M*,

$$P(M|H,GS,D) = P(M|H)$$

and

$$P(M|GT,ST) = P(M|GT)$$

Estimating the probability of *ST*|*D* then becomes

$$P(ST|D) = P(ST|GT)\sum_{GT} P(GT|M)\sum_{M} P(M|H)\sum_{H} P(H|GS)\sum_{GS} P(GS|D)$$

A Markov model is specified in the forward direction, for example, the model specifies the dependence of homology on the gene sequence, $P(H|GS, D) = P(H|GS)$. However, this specification also produces a dependence in the backward direction as can be seen using Bayes rule: $\mathbf{P(GS|H)} = \frac{\mathbf{P(H|GS)P(GS)}}{\mathbf{P(H)}}$. Generative inference procedures account for this bidirectional dependence using a forward algorithm to account for upstream effects and a backward algorithm to account for downstream effects [86]. The fact that current phylogenetic workflows employ only backward steps (e.g., gene trees depend on assembled gene sequences, but sequence assembly does not use any information about gene trees) imposes important limitations. For example, one of the hardest problems in sequence assembly is correctly deconvolving repeats. One biological source of repeats is gene duplication. Since a phylogenetic workflow includes gene tree inference, this analysis component could help the assembler correctly assemble regions that repeat due to gene duplication, but this strategy is not currently employed.

In addition, the typical phylogenetic workflow makes the strong assumption of low relative entropy (Box 2), so that summing over all possible hypotheses is not required, as the probability of the point estimate is close to 1. Given these two assumptions, estimating *ST*|*D* becomes

$$P(ST|D) = P(ST|GT)P(GT|M)P(M|H)P(H|GS)P(GS|D) \approx 1$$

This equation embodies the linear workflow (Figure 1): the simplest possible inference process of species trees from the data that passes minimal information, that is, a point estimate, from one analysis component to the next (Box 1) under the assumptions that there is a biologically relevant ordering of steps, the steps form a Markov chain, and that relative entropy between the point estimate and the true distribution is low (Box 2).

When the investigator isolates molecules from the organism and prepares them for sequencing they are often further modified through fragmentation, ligation, and errors introduced during amplification. Finally, sequencing produces raw reads (*D*, Figure 1) that are estimates of the sequences in these prepared samples. Sequencing instruments introduce errors during this process and have ascertainment biases that make it more difficult to observe some sequences than others [24]. The raw reads (*D*) are the observed data, and the other random variables (*GS*, *GT*, and *ST*) are unknowns that the investigator seeks to estimate from the observed data.

Simulation tools have already been developed for most of the steps in this generative process, including raw sequence reads from gene sequences [25–27], gene sequences given gene trees [28–31], and gene trees given species trees [32–34]. Even so, there is not yet an integrated tool for simulating raw data under all the processes in a phylogenetic analysis that biologists are regularly interested in.

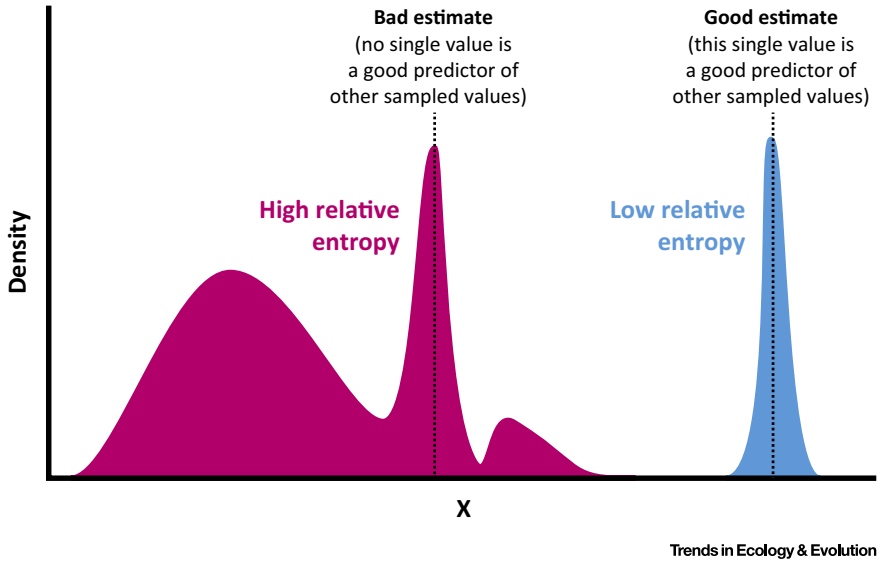


Figure 2. Two Probability Density Functions Representing Random Variables. The relative entropy between the magenta (left) distribution and the point estimate is high due to the high variance and a complex shape of the distribution. No single estimate is a good predictor of other values drawn from this distribution, but two point estimates can provide a much better approximation. The relative entropy between the cyan (right) distribution and the point estimate is low. A single estimate drawn from this distribution is a good predictor of other values.

## A Molecular Phylogenetic Workflow Makes Inferences about Unobserved Variables from Observed Sequence Data

The goal of an analysis workflow is essentially to reverse a generative model. While a generative model explains how observed data are generated from unobserved variables, an analysis workflow estimates unobserved variables from the observed data. In the case of linear models and workflows, there is an antiparallel relationship between the generative model (Figure 1, orange) and the analysis workflow (Figure 1, blue) [9].

Here, we describe a phylogenetic analysis workflow that is typical of many currently implemented, although often differences exist in goal and design details between studies. Different data acquisition approaches can simplify various analysis components. Targeted enrichment, for example, greatly simplifies assembly and **homology evaluation**, but at the cost of only providing data for preselected genes [35].

The first analysis step begins by estimating gene sequences (*GS*, Figure 1) from the observed data (*D*, Figure 1). This analysis process is referred to as assembly. It identifies which sites in different reads correspond to the same sites in the original molecules, identifies the relative location of reads to each other, corrects technical errors introduced by sample preparation and sequencing, and accommodates and annotates some aspects of cellular editing (such as splice variation) [36,37]. Changes in sequencing technology have had major impacts on assembly methods [36,38]. It is anticipated, for example, that read lengths will be much longer in the near future, even extending to the full length of the biological molecules under study [39–41]. This will make it much simpler to identify the location of reads, but does not obviate assembly as it is still necessary to identify which reads are derived from the same molecules and to correct technical error.

The next analysis challenge is to proceed from gene sequences (*GS*, Figure 1) to gene trees (*GT*, Figure 1). Most phylogenetic tools assume homology, so the first step in this process is to identify homologous sequences through phenetic comparisons of sequence similarity [42]. This is

typically implemented in a few steps. First, pairwise sequence comparisons and clustering are used to identify homologous sequences, that is, sequences that are tips in the same gene tree. Some sequence homology tools compare all sequences to each other [43], others compare new sequences to a curated set of preselected sequences [44]. Next, site homology is evaluated within each set of homologous sequences by multiple sequence alignment. Multiple sequence alignment partitions the variation observed among homologous sequences into two categories: variation due to insertion and deletion (which results in sites placed in different columns) and variation due to substitution (which is contained within a single column) [45]. Homology evaluation and multiple sequence alignment generate two random variables not present in the generative model, which do not represent estimates of entities that occur in nature (*H*, *M*, Figure 1). With the alignments in hand, the investigator can proceed to the core of a phylogenetic study – phylogenetic inference. Given the gene sequences and a model of molecular evolution, phylogenetic inference tools evaluate alternative hypotheses regarding the evolutionary relationships between the sequences [18,46,47]. For historical and computational reasons these tools typically model only site substitution, and the insertion and deletion events identified in multiple sequence alignment are not evaluated.

Inferring species trees (*ST*, Figure 1) requires both the identification of gene tree features that correspond to features of the species tree (i.e., nodes due to speciation that result in orthologs) as well as the integration of information from many gene trees to learn about their shared history constrained through speciation and lineage sorting. These steps are implemented in many different ways in different linear studies, including consensus trees [48] and matrix concatenation [49] approaches.

## The Implicit Assumptions of Phylogenetic Workflows

Several implicit assumptions (Box 2) are usually made to simplify analyses and reduce their computational cost, resulting in the simple linear type of workflow described above. Making these assumptions explicit is critical to understanding the limitations imposed on results by current methods and establishing future research priorities. First, it is assumed that the specific ordering of the analysis components sufficiently describes the processes that generated the data. Second, it is assumed that the dependence between random variables forms a Markov chain, that is, the inference process for each analysis component depends only on its two neighbors. These two assumptions justify the direction of information propagation and the consideration at each analysis component of only the results of the preceding component. In the phylogenetic analysis workflow outlined above, only a single hypothesis from each analysis component is passed on to the next component. This reflects a third implicit assumption: a single result from each analysis component is sufficient to summarize all that is needed from all the analysis components that precede it. In probability theory, this is described as an assumption of low **relative entropy** (Figure 2) [50].

Explicitly stating these assumptions raises the question of whether they are routinely violated in real-world analyses and if these violations jeopardize the interpretation of the results. These are empirical as well as theoretical questions. The ordering of the analysis steps is justified inasmuch as it is equivalent to the backward algorithm of a simplified linear generative model. The Markov chain assumption could be violated in a number of ways. For example, duplication events within a gene tree are a source of repeats encountered in assembly. Hence a dependency exists between gene trees and gene sequence assembly, and the assembly process could benefit from consideration of gene tree inference. In addition, population size can simultaneously impact multiple processes, including speciation, rates of molecular evolution, and **incomplete lineage sorting**.

There is good reason to believe that the third assumption of low relative entropy is frequently violated in ways that negatively impact phylogenetic analyses. There is often considerable

uncertainty regarding each step in a phylogenetic analysis [2–4,51], and discarding information about other hypotheses that also find support from the data can mislead analyses and greatly complicates their biological interpretation. This concern was in part the motivation for the field of phylogenetics moving away from single point estimates of trees toward the generation and presentation of whole distributions of results, such as Bayesian posterior distributions of trees and bootstrap replicates [52]. This makes relaxing the assumption of low relative entropy a particularly high priority, and is therefore the focus of this manuscript.

Two different communication strategies (Box 1) can relax the assumption of low relative entropy. One is to propagate multiple hypotheses from one component to another, providing a distribution of hypotheses. Another is to jointly estimate multiple components, allowing for a comprehensive assessment of uncertainty contributed by the components.

## Existing Approaches that Relax the Assumption of Low Relative Entropy

The phylogenetic analysis workflow described above (Figure 1, blue) represents one extreme – it communicates the minimal amount of information in the simplest possible way. Even so, it has been widely adopted due to two key practical advantages: it is the most computationally tractable (since all but one hypothesis is discarded at each step) and it is technically the most straightforward to implement (since it can be built from existing tools). Implementing this workflow as a software pipeline has varied from completely manual analyses, where information is formatted and passed to each component by hand or semiautomated tools, to fully automated workflows that also track and summarize results [43,53–55].

Most of the recent work on improving communication has focused on the joint estimation of gene trees and a species tree [6,7,56]. This focus is motivated in part by the fact that gene trees and species trees should not necessarily be expected to be congruent [57], and that the incongruence can only be correctly accounted for by simultaneously estimating both gene and species relationships. Incomplete lineage sorting has been the primary focus of most recent work in this area, largely due to the mature mathematical and statistical foundation provided by coalescent theory [6,58–60]. Nevertheless, recent work now also accounts for other sources of incongruence such as gene duplication and loss [7,9,61–63]. An interesting recent approach to this problem is the work of de Oliveira Martins *et al.* [8], which considers a Bayesian posterior distribution of gene trees during species tree estimation.

Efforts to improve communication at earlier steps have been largely neglected. Current approaches treat homology evaluation (the identification of homologous sequences), multiple sequence alignment (the identification of homologous sites in homologous sequences), and phylogenetic tree inference (identification of phylogenetic trees that best explain the variation in homologous sites) as separate problems. Better integration of multiple sequence alignment and phylogenetic tree inference has been one focus of work [64]. Several approaches have been developed to improve communication between these two analysis components by co-estimating multiple sequence alignment and gene trees inference under parsimony [65], maximum likelihood [66,67], and Bayesian approaches [68].

## Identifying Future Priorities for Improved Communication

There are clear benefits to relaxing the assumption of low relative entropy through improved information communication between analysis components, including the ability to evaluate results according to statistically explicit criteria and compare support for alternative hypotheses. These potential benefits, however, come with costs that must be considered when setting priorities for future work. The cost of implementation can be very high, as many tools will have to be re-engineered to accommodate multiple hypotheses or information from non-neighboring analysis components. Most investigators already struggle with the computational costs of

existing tools, so understanding the relative benefit of increasing the cost of any particular step is critical to making informed decisions about which investments will have the greatest benefit.

**Sensitivity analyses** can play an important role in evaluating the benefits of relaxing or strengthening assumptions. If the investigator relaxes an assumption by a far more computationally expensive approach but this has little impact on the result, then the cost of adopting the more general approach might not be justified.

As discussed earlier, recent advances have largely focused on improving information communication between gene tree and species tree inference. Recent reviews have addressed these advances in detail [9,69,70]; hence, we here focus on opportunities to relax the assumption of low relative entropy by improving information communication between upstream analysis components.

Relaxing the assumption of low relative entropy of assembly output will likely have great benefit to researchers [51]. Most assemblers choose a single hypothesis according to *ad hoc* criteria, assuming it to be an adequate point estimate of the original gene sequence (e.g., [71–73]). This reduces computational cost, but at the price of discarding uncertainty in the results and neglecting important biological information such as variation due to heterozygosity or somatic polymorphisms. Generating multiple assembly hypotheses for each gene and propagating them to homology evaluation would create the opportunity to assess alternative assembly hypotheses from a more informed position, where they can be compared to homologous sequences from other species. Rather than only evaluate assembly hypotheses according to information available within each species, information would be borrowed across species.

For example, chimeric gene sequences, formed by the spurious fusion of sequences from two different genes, are particularly common and problematic [74]. They confound homology evaluation by creating spurious sequence similarities between disparate gene families. These spurious similarities provide signatures during homology evaluation (e.g., high degree centrality in graphs of hypothesized homologous sequences) that make it possible to identify them and remove them [75]. Retaining multiple gene sequence hypotheses and passing them all forward would allow the analysis workflow to identify and retain non-chimeric gene sequences at this later stage.

Implementing this improved communication strategy between assembly and homology evaluation requires minimal cost when analyzing transcriptome data. Transcriptome assemblers already generate multiple assembly hypotheses per gene to accommodate splice variants; however, some of these variants are assembly errors rather than true alternative splice variants [74]. In addition, homology evaluation tools will not require much re-engineering to accept multiple assembly hypotheses, as it would just be a matter of performing the same sequence similarity comparisons but on a larger dataset. Recent advances in assembly methods will also facilitate this approach with non-transcriptome data. There has been growing interest in moving to a statistically grounded approach to sequence assembly [51,76,77]. These tools provide a distribution of gene sequences, rather than a single assembly, that captures uncertainty about the original biomolecule sequence [78,79]. These are not yet computationally viable for analyses of large real-world datasets, although they could soon be.

## Concluding Remarks

As the focus on phylogenetic workflow development transitions from improving individual analysis components to better integrating those components, it will be necessary to prioritize those developments that will have the most positive impact (see Outstanding Questions). A generative model and probabilistic framework provide the perspective needed to describe and

### Outstanding Questions

Phylogenetic workflows make strong assumptions about order, dependence, and relative entropy. These assumptions are likely violated in most analyses. To what extent do these violations negatively impact analysis results?

Which steps in a phylogenetic analysis would benefit most from improved communication that relaxes one or more of these assumptions? This question requires the assessment of trade-offs among implementation costs, computational costs, and improved results.

Existing approaches are already struggling to keep up with rapidly growing datasets. Are there more assumptions we can impose to speed up computation in phylogenetic analyses without negatively impacting the quality of results?

evaluate different potential improvements. At one extreme is the current approach of passing a single hypothesis between steps in linear analyses, discarding all alternative hypotheses. This approach is simple to implement and minimizes computational demands, but relies on an assumption of low relative entropy that is routinely violated in ways that negatively impact analyses. At the other extreme is the simultaneous estimation of all unknowns, including gene sequences, gene trees, and species trees. This approach makes no assumptions about Markovian dependence, ordering of analysis steps, or low relative entropy, but is computationally prohibitive. In practice, the optimal approach to molecular phylogenetic analyses will be between these extremes. Sensitivity analyses and probability theory provide an informative guide for finding where that optimal point will be, and for making well-informed decisions about the relative benefits of allocating additional engineering and computational resources to different analysis steps.

## Acknowledgments

This work was supported in part by the National Science Foundation (NSF) Alan T. Waterman Award to C.W.D. and NSF grant DEB-1256695. Thanks to Sam Church for feedback on the manuscript.

## References

1. Anisimova, M. *et al.* (2013) State of the art methodologies dictate new standards for phylogenetic analysis. *BMC Evol. Biol.* 13, 161
2. Lunter, G. *et al.* (2008) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.* 18, 298–309
3. Wong, K.M. *et al.* (2008) Alignment uncertainty and genomic analysis. *Science* 319, 473–476
4. Huelsenbeck, J.P. *et al.* (2000) Accommodating phylogenetic uncertainty in evolutionary studies. *Science* 288, 2349–2350
5. Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* 43, 304–311
6. Liu, L. and Pearl, D.K. (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* 56, 504–514
7. Boussau, B. *et al.* (2013) Genome-scale coestimation of species and gene trees. *Genome Res.* 23, 323–330
8. de Oliveira Martins, L. *et al.* (2014) A Bayesian supertree model for genome-wide species tree reconstruction. *Syst. Biol.* Published online October 3, 2014. http://dx.doi.org/10.1093/sysbio/syu082
9. Szöllő"si, G.J. *et al.* (2015) The inference of gene trees with species trees. *Syst. Biol.* 64, e42–e62
10. Boussau, B. and Daubin, V. (2010) Genomes as documents of evolutionary history. *Trends Ecol. Evol.* 25, 224–232
11. Blei, D.M. *et al.* (2003) Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022
12. Chomsky, N. (1956) Three models for the description of language. *IRE Trans. Inform. Theory* 2, 113–124
13. Collins, M. (2003) Head-driven statistical models for natural language parsing. *Comput. Linguist.* 29, 589–637
14. Lu, W. *et al.* (2008) A generative model for parsing natural language to meaning representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 783–792, Association for Computational Linguistics
15. Langmead, C.J. (2014) Generative models of conformational dynamics. *Adv. Exp. Med. Biol.* 805, 87–105
16. Rabiner, L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286
17. Fischer, A. and Igel, C. (2012) An introduction to restricted Boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 14–36, Springer
18. Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376
19. Degnan, J.H. and Salter, L.A. (2005) Gene tree distributions under the coalescent process. *Evolution* 59, 24–37
20. Szöllosi, G.J. and Daubin, V. (2012) Modeling gene family evolution and reconciling phylogenetic discord. *Methods Mol. Biol.* 856, 29–51
21. Nichols, R. (2001) Gene trees and species trees are not the same. *Trends Ecol. Evol.* 16, 358–364
22. Kloc, M. and Zagrodzinska, B. (2001) Chromatin elimination – an oddity or a common mechanism in differentiation and development? *Differentiation* 68, 84–91
23. Smith, J.J. *et al.* (2012) Genetic consequences of programmed genome rearrangement. *Curr. Biol.* 22, 1524–1529
24. Ross, M.G. *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome Biol.* 14, R51
25. Lysholm, F. *et al.* (2011) An efficient simulator of 454 data using configurable statistical models. *BMC Res. Notes* 4, 449
26. Huang, W. *et al.* (2012) ART: a next-generation sequencing read simulator. *Bioinformatics* 28, 593–594
27. Caboche, S. *et al.* (2014) Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC Genomics* 15, 264
28. Strope, C.L. *et al.* (2009) Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Mol. Biol. Evol.* 26, 2581–2593
29. Fletcher, W. and Yang, Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.* 26, 1879–1888
30. Cartwright, R.A. (2005) DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics* 21 (Suppl. 3), iii31–iii38
31. Stoye, J. *et al.* (1998) Rose: generating sequence families. *Bioinformatics* 14, 157–163
32. Sjöstrand, J. *et al.* (2013) GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics* 14, 209
33. Heled, J. *et al.* (2013) Simulating gene trees under the multispecies coalescent and time-dependent migration. *BMC Evol. Biol.* 13, 44
34. Maddison, W.P. and Maddison, D.R. (2007) *Mesquite: A Modular System for Evolutionary Analysis. Version 2.75. 2011.* ( http://mesquiteproject.org)
35. Lemmon, E.M. and Lemmon, A.R. (2013) High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 44, 99–121
36. Martin, J.A. and Wang, Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12, 671–682
37. Nagarajan, N. and Pop, M. (2013) Sequence assembly demystified. *Nat. Rev. Genet.* 14, 157–167
38. Bradnam, K.R. *et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2, 10

39. Kasianowicz, J.J. *et al.* (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13770–13773
40. Rusk, N. (2009) Cheap third-generation sequencing. *Nat. Methods* 6, 244
41. Clarke, J. *et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4, 265–270
42. Pearson, W.R. (2013) An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinformatics* Chapter 3, Unit 3.1
43. Dunn, C.W. *et al.* (2013) Agalma: an automated phylogenomics workflow. *BMC Bioinformatics* 14, 330
44. Ebersberger, I. *et al.* (2009) HaMStR: profile hidden Markov model based search for orthologs in ESTs. *BMC Evol. Biol.* 9, 157
45. Löytynoja, A. and Goldman, N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10557–10562
46. Ronquist, F. *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542
47. Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313
48. Bryant, D. (2003) A classification of consensus methods for phylogenetics. *Discrete Math. Theoret. Comput. Sci.* 61, 163–184
49. De Queiroz, A. and Gatesy, J. (2007) The supermatrix approach to systematics. *Trends Ecol. Evol.* 22, 34–41
50. Cover, T.M. and Thomas, J.A. (2012) *Elements of Information Theory,* Wiley
51. Howison, M. *et al.* (2013) Toward a statistically explicit understanding of de novo sequence assembly. *Bioinformatics* 29, 2959–2963
52. Holder, M. and Lewis, P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* 4, 275–284
53. Grant, J.R. and Katz, L.A. (2014) Building a phylogenomic pipeline for the eukaryotic tree of life – addressing deep phylogenies with genome-scale data. *PLoS Curr.* 6, http://dx.doi.org/10.1371/currents.tol.c24b6054aebf3602748ac042ccc8f2e9
54. Oakley, T.H. *et al.* (2014) Osiris: accessible and reproducible phylogenetic and phylogenomic analyses within the Galaxy workflow management system. *BMC Bioinformatics* 15, 230
55. Szitenberg, A. *et al.* (Published online May 15 2015) ReproPhylo: an environment for reproducible phylogenomics. *bioRxiv* 2015 http://biorxiv.org/content/biorxiv/early/2015/05/18/019349.full.pdf
56. Akerborg, O. *et al.* (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5714–5719
57. Maddison, W.P. (1997) Gene trees in species trees. *Syst. Biol.* 46, 523–536
58. Degnan, J.H. and Rosenberg, N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340
59. Heled, J. and Drummond, A.J. (2010) Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27, 570–580
60. Maddison, W.P. and Knowles, L.L. (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55, 21–30
61. Steel, M. *et al.* (2013) Identifying a species tree subject to random lateral gene transfer. *J. Theor. Biol.* 322, 81–93
62. Bayzid, M.S. *et al.* (2013) Inferring optimal species trees under gene duplication and loss. *Pac. Symp. Biocomput.* 2013, 250–261
63. Rasmussen, M.D. and Kellis, M. (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* 22, 755–765
64. Wheeler, W. (1996) Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics* 12, 1–9
65. Varón, A. *et al.* (2010) POY version 4: phylogenetic analysis using dynamic homologies. *Cladistics* 26, 72–85
66. Liu, K. *et al.* (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324, 1561–1564
67. Liu, K. *et al.* (2012) SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst. Biol.* 61, 90–106
68. Redelings, B.D. and Suchard, M.A. (2005) Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.* 54, 401–418
69. Knowles, L.L. and Kubatko, L.S. (2011) *Estimating Species Trees: Practical and Theoretical Aspects,* Wiley
70. Edwards, S.V. (2009) Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1–19
71. Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829
72. Simpson, J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123
73. Grabherr, M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652
74. Yang, Y. and Smith, S.A. (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* 14, 328
75. Misner, I. *et al.* (2013) Sequence comparative analysis using networks: software for evaluating de novo transcript assembly from next-generation sequencing. *Mol. Biol. Evol.* 30, 1975–1986
76. Rahman, A. and Pachter, L. (2013) CGAL: computing genome assembly likelihoods. *Genome Biol.* 14, R8
77. Ghodsi, M. *et al.* (2013) De novo likelihood-based measures for comparing genome assemblies. *BMC Res. Notes* 6, 334
78. Howison, M. *et al.* (2014) Bayesian genome assembly and assessment by Markov chain Monte Carlo sampling. *PLoS ONE* 9, e99497
79. Maretty, L. *et al.* (2014) Bayesian transcriptome assembly. *Genome Biol.* 15, 501
80. Kubatko, L.S. *et al.* (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25, 971–973
81. Kemp, C. and Tenenbaum, J.B. (2008) The discovery of structural form. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10687–10692
82. Clark, M.P. *et al.* (2011) Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.* 47, W09301
83. Sanderson, M.J. *et al.* (2011) Terraces in phylogenetic tree space. *Science* 333, 448–450
84. Newberg, L.A. and Lawrence, C.E. (2009) Exact calculation of distributions on integers, with application to sequence alignment. *J. Comput. Biol.* 16, 1–18
85. Shannon, C.E. (2001) A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.* 5, 3–55
86. Durbin, R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids,* Cambridge University Press