Phylogenetic Inference Using Whole Genomes

Bruce Rannala[1] and Ziheng Yang[2]

[1]Genome Center and Department of Evolution and Ecology, University of California, Davis, California 95616; email: bhrannala@ucdavis.edu

[2]Department of Biology, University College London, London WC1E 6BT United Kingdom; Laboratory of Biometrics, Graduate School of Agriculture and Life Sciences, University of Tokyo, Tokyo, Japan; email: z.yang@ucl.ac.uk

Annu. Rev. Genomics Hum. Genet. 2008. 9:217–31

First published online as a Review in Advance on June 3, 2008

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev.genom.9.081307.164407

1527-8204/08/0922-0217$20.00

Key Words

genomic sequences, coalescent process, gene tree

Abstract

The availability of genome-wide data provides unprecedented opportunities for resolving difficult phylogenetic relationships and for studying population genetic processes of mutation, selection, and recombination on a genomic scale. The use of appropriate statistical models becomes increasingly important when we are faced with very large datasets, which can lead to improved precision but not necessarily improved accuracy if the analytical methods have systematic biases. This review provides a critical examination of methods for analyzing genomic datasets from multiple loci, including concatenation, separate gene-by-gene analyses, and statistical models that accommodate heterogeneity in different aspects of the evolutionary process among data partitions. We discuss factors that may cause the gene tree to differ from the species tree, as well as strategies for estimating species phylogenies in the presence of gene tree conflicts. Genomic datasets provide computational and statistical challenges that are likely to be a focus of research for years to come.

INTRODUCTION

Species tree: the unobserved tree of genealogical relationships among the species from which the genes are sampled

Gene tree: the unobserved tree of genealogical relationships among genes through time

The genome sequencing era began three decades ago with the sequencing, in 1977, of the 5368-bp DNA genome of the bacteriophage virus ϕX174 (53). Automated sequencing technologies enabled the first bacterial genome, the 1830-kb genome of *Haemophilus influenzae* (21), to be sequenced in 1995, followed by the first eukaryotic genome, the 12.5-Mb genome of the budding yeast *Saccharomyces cerevisiae* (23), in 1997. During the current decade, the number of sequenced genomes has grown exponentially. As of October 2007, Entrez (**http://ncbi.nlm.nih.gov/**) lists 543 completed genomes for eubacterial species and 47 for archaeal species. There are 23 completed eukaryotic genomes and 129 draft genomes. These include two completed mammalian genomes (human and mouse), 21 draft assembly mammalian genomes, and 26 mammalian genomes in progress. These numbers are likely to increase by as much as tenfold by the close of the current decade.

During the last several years the potential value of comparative genomics for the identification of genes, regulatory regions, and other genome features has shifted sequencing efforts away from model organisms such as mouse and *Drosophila* to include other related species. Numerous completed genome sequences are now available for evolutionarily related species, opening up the possibility of using whole genomes to infer phylogenetic relationships and divergence times among species. Moreover, new sequencing technologies have enabled resequencing of genomes for multiple individuals of a single species, strain, or population. The newly emergent fields of phylogenomics and population genomics are one consequence of these technological advances. Although the availability of whole genome sequences is quite new, the basic principles of multilocus inference in phylogenetics and population genetics, developed and refined over the last two decades, are relatively well established.

The objective of this review is to describe how existing tools for phylogenetic inference can be applied to whole genomes. The use of robust methods of analysis is clearly extremely important when such large amounts of data are analyzed; errors induced by phylogenetic inference techniques known to be prone to large-sample problems such as statistical inconsistency can be greatly magnified by the use of whole genome data (43, 50). Moreover, new problems arise, such as how to account for the effects of recombination, gene conversion, and horizontal gene transfer. In this review, we do not deal with details of genome sequencing such as sequence assembly. Nor do we consider the important related problem of sequence alignment. Rather, we focus exclusively on the problem of accurately inferring phylogenetic trees and species divergence times using a sample of aligned orthologous sequences for regions that span an entire genome. Even this relatively focused endeavor can become quite complicated in many situations of biological interest.

SPECIES TREES AND GENE TREES

Amino acid sequences, cross-reactivity of antibodies, and other measures of evolutionary divergence among multiple proteins (and species) first became widely available during the 1960s (13). Researchers developing measures of genetic distance intended for use with such data noted early on the distinction between organismal phylogenies and molecular phylogenies. Fitch (20), for example, proposed new terms to clarify the distinction between orthologs (genes descending from a shared ancestral gene owing to a shared species divergence event) and paralogs (genes descending from a shared ancestral gene owing to a gene duplication event). Nei (41), considering the problem of dating species divergence events using immunological distances, recognized that the objective was to "reconstruct or estimate the evolutionary tree of *the organisms used* rather than that of a protein." Tateno and coworkers (62) made a similar distinction, noting that "the primary objective of molecular taxonomy or phylogenetics is to construct a species tree rather than a gene tree"

and speculating that "the only way to reduce the errors involved in an estimated tree is to increase the number of genes used." In the last two decades, the importance of distinguishing between gene and species trees has been widely recognized and researchers have identified new sources of gene and species tree conflict that were previously unknown (or highly speculative). Recent genome sequencing efforts provide the opportunity for an almost limitless number of genes to be employed in a phylogenetic analysis aimed at identifying a species tree.

Sources of Gene Tree Conflict

In discussing gene tree conflict, it is essential to distinguish between an estimated gene tree and the true gene tree. The true gene tree is the unobserved tree of genealogical relationships among genes through time. The estimated gene tree is the current best estimate of that tree based on DNA sequence data. Even when true gene trees are identical between genes, estimated gene trees may differ owing to random and systematic errors in phylogenetic tree reconstruction. In analyses that use well-behaved statistical methods known to produce consistent estimates, such as maximum likelihood or Bayesian inference, such errors can be reduced by adding sites to a gene. However, underlying biological processes can cause the true gene trees to differ, in which case the estimated gene trees can differ regardless of the number of sites examined (see discussion below). Here, we use the term "gene tree conflict" to refer only to cases in which true gene trees differ. True gene trees can differ either in divergence times or tree topology. The species tree is the unobserved tree of genealogical relationships among the species from which the genes are sampled.

The two major types of biological process that can lead to conflicts among gene trees are, first, population genetic processes such as drift operating in an ancestral species, and second, genomic recombination, either within a single species (e.g., gene conversion, transposition, or meiotic crossovers) or between species (horizontal/lateral gene transfer). Because the effects of genomic recombination in causing conflicts between gene and species trees may depend on whether ancestral polymorphisms are present, we consider ancestral population effects first. We then go on to discuss the role of genomic recombination in generating conflicts and the influence of processes such as meiotic crossovers within populations on the probability of gene and species tree conflicts.

Ancestral Polymorphisms

Species divergence times and ages of most recent common ancestors. To illustrate the effect of ancestral polymorphism, consider a homologous segment of DNA in the human and chimpanzee genomes. Suppose we sample a single human and chimpanzee sequence; we are interested in using the pattern of DNA substitutions between the pair of aligned sequences to infer the age of the human-chimpanzee speciation event, t_{HC}. Assuming that no gene flow occurs subsequent to the speciation event (which indeed may be taken as the definition of speciation), it is impossible for the sequences to share a most recent common ancestor (MRCA) that is younger than the age of the speciation event. The time until the MRCA, T_{HC}, is then determined by population genetic processes operating in the human-chimpanzee ancestral population, such as genetic drift (**Figure 1**). Coalescent theory (25, 30, 60) can be used to calculate the probability density of the discrepancy $T_{HC} - t_{HC}$:

$$f(T_{HC}) = \frac{1}{2N_e} e^{-(T_{HC} - t_{HC})/(2N_e)}, \quad \text{for} \quad T_{HC} > t_{HC},$$

where time is measured in units of generations. This is an exponential distribution with mean $E(T_{HC} - t_{HC}) = 2N_e$ and variance $\text{var}(T_{HC}) = 4N_e^2$. The degree to which such discrepancies can be detected (and thus influence resulting estimates) when using sequence data depends on the relative accuracy of the branch length estimates (in units of expected numbers of substitutions), the mutation rate, the effective population size, and the generation time. The

Maximum likelihood: a statistical method for estimating parameters in a statistical model by maximizing the probability of the observed data

Bayesian inference: an approach to statistical inference that uses probability distributions to describe uncertainties in model parameters

Genomic recombination: the exchange of sequence information between distinct DNA molecules

Ancestral polymorphism: polymorphism or sequence differences in an extinct ancestral species

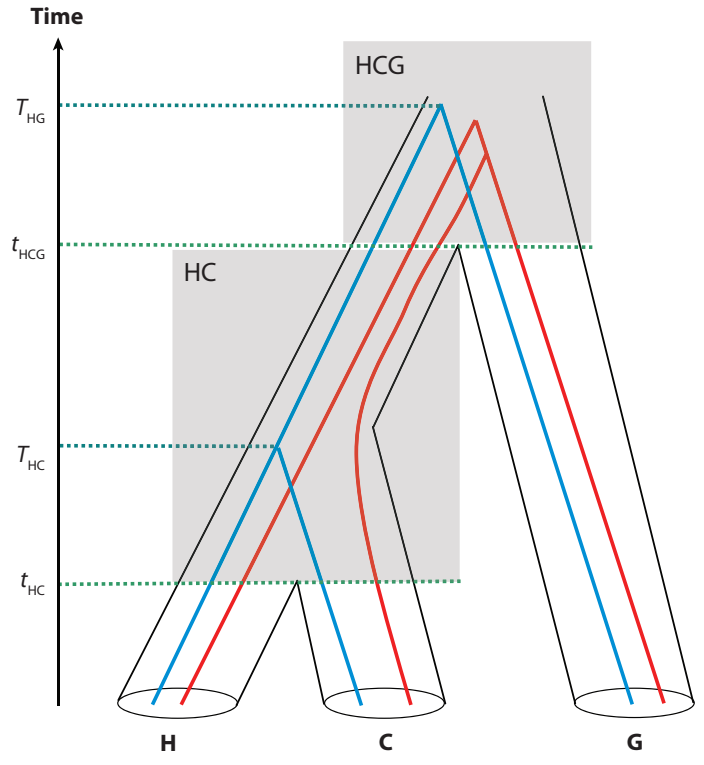

Figure 1

A species tree for three species [human (H), chimpanzee (C), and gorilla (G)] used to illustrate lineage sorting, which may generate a species tree–gene tree conflict. The species tree is ((H, C), G), with the species divergence times t_{HC} and t_{HCG}. The two ancestral species (populations) are represented as HC and HCG. Two gene trees are shown. In the blue gene tree, the H and C sequences coalesce in population HC, and the resulting gene tree matches the species tree. In the red gene tree, both coalescent events occur in population HCG, and the resulting gene tree differs from the species tree.

accuracy of branch length estimation depends on the number of sites analyzed and the substitution model used. If $v = v_1 - v_0$ is the smallest discrepancy that can be detected (in units of expected mutations) then the probability that a detectable discrepancy is observed between the branch length expected under the true species divergence time, $v_0 = t_{HC}G\mu$, and that of the MRCA, $v_1 = T_{HC}G\mu$, is

$$p = \int_{v/\mu}^{\infty} \frac{e^{-y/(2N_eG)}}{2N_eG} dy = e^{-v/(2N_eG\mu)},$$

where G is the generation time (in years) and μ is the mutation rate (per year). Considering the human-chimpanzee divergence, for example, if the human-chimpanzee ancestral effective population size is $N_e = 100{,}000$ (61), the generation time is $G = 20$ years, and the mutation rate is $\mu = 10^{-9}$ mutations per site per year, then a branch length discrepancy of $v = 0.01$ or larger will occur with probability $p = 0.08$ and a branch length discrepancy of $v = 0.005$ or larger will occur with probability $p = 0.29$. Given that potentially thousands of independent genes can be examined in a whole-genome phylogenetic analysis, discrepancies larger than $v = 0.01$ can be expected to commonly occur. If such discrepancies occur and are not accounted for it is evident that using node ages on the inferred phylogeny will tend to overestimate species divergence times. Indeed, if the human-chimpanzee speciation event occurred 5 million years ago, the average sequence divergence between the two species will be $E(T_{HC})G\mu = t_{HC}G\mu + 2N_eG\mu = 0.005 + 0.004$, so that $0.004/0.009 = 44\%$ of the divergence is due to ancestral polymorphism. To accurately estimate species divergence times in such cases population genetic parameters, such as ancestral effective population sizes, should be jointly estimated with species divergence times using sequence data under a population genetic model (e.g., see 61).

Lineage sorting and conflicting gene and species trees. If three or more species are examined, the ancestral coalescent process may generate discrepancies between species tree and gene tree topologies (26). Consider the species tree of human (H), chimpanzee (C), and gorilla (G), shown in **Figure 1**. If the H and C sequences coalesce in the ancestral species HC, the gene tree will be topologically the same as the species tree. However, if the H and C sequences do not coalesce in the HC population, they will enter the ancestral population HCG. Then all three sequences will coalesce in random order, and only one of the three possible resulting trees matches the species tree. Thus the probability that the gene tree differs from the species tree equals 2/3 the probability that the H and C sequences do not coalesce

in the HC ancestral population

$$P_{SG} = \frac{2}{3} e^{-(t_{HCG} - t_{HC})/(2N_e)},$$

where N_e is the effective population size of population HC (26). Note that the mismatch probability P_{SG} is greater when the two speciation events are closer in time and when the ancestral HC population is larger.

In real data analysis, the gene tree is unknown and is inferred from sequence data at the locus. Errors in phylogeny reconstruction will then inflate the mismatch probability; that is, the probability that the species tree differs from the estimated gene tree, P_{SE}, is always greater than P_{SG} (69). A commonly used approach to estimating ancestral population sizes in the case of three species is to equate the observed mismatch probability, P_{SE}, with the expected probability, P_{SG}, ignoring errors in phylogeny reconstruction (8). This so-called tree-mismatch method can seriously overestimate ancestral population sizes. **Figure 2** illustrates the relationship between the probability of a gene and species tree mismatch for either estimated or true gene trees as a function of sequence length.

Similar to the case of three species, ancestral polymorphism can cause species tree and gene tree topologies to differ when the number of species is greater than three. Several authors have derived the gene and species tree mismatch probabilities for various numbers of species when the species tree is fixed (10, 25, 45).

Another question of relevance for phylogenetic inference is how often the most common gene tree differs from the species tree (9, 33). Recently, the finding that under some conditions the most common gene tree topology does not match the species tree has attracted much attention. This occurs in situations where the species tree is highly asymmetrical and arises from the fact that the coalescent process places a uniform prior on labeled histories, rather than on topologies. The labeled history (14) takes the rank ordering of the nodes in a tree into account as well as the cladogenic relationships

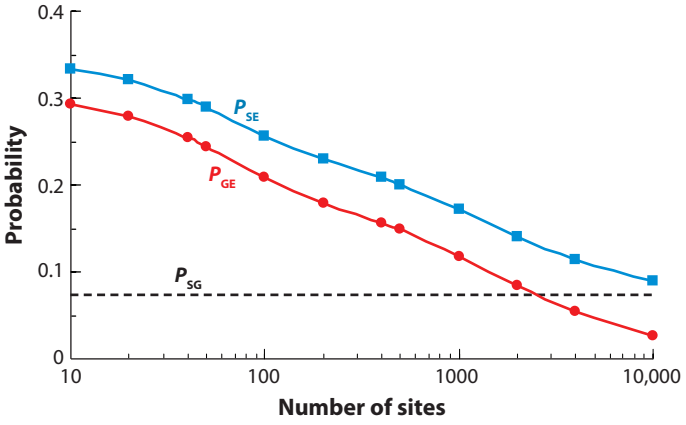

Figure 2

Tree mismatch probabilities plotted against locus size (in number of sites), calculated using computer simulation. P_{SG} is the mismatch probability between the species tree and the gene tree (0.0739 for the parameter values used). P_{SE} is the mismatch probability between the species tree and the estimated gene tree. This is much greater than P_{SG} for short loci but converges to P_{SG} with increasing sequence length. P_{GE} is the probability of a mismatch between the true gene tree and the estimated gene tree, and is the probability of error in phylogeny reconstruction. The simulation was conducted using the species tree of **Figure 1** and the following parameter values: $\theta_{HC} = 4N_{HC}G\mu = 0.0010$ for population HC, $\theta_{HCG} = 0.0031$ for population HCG, $t_{HC}G\mu = 0.0052$ for the H-C divergence, and $t_{HCG}G\mu = 0.0063$ for the HC-G divergence. Redrawn according to 69.

(47). A completely asymmetrical (unbalanced) species tree is compatible with only one labeled history (because there is only one ordering of the coalescent times). However, if the species tree is not completely asymmetrical, it can correspond to several labeled histories and thus receives more weight in the prior than a completely asymmetrical tree. This is only an issue for particular tree shapes, but in such situations it could cause phylogenetic procedures based on unweighted summaries of gene trees (such as supertree methods or gene concatenation–based methods) to be inconsistent (9, 33). In practice, such situations are likely to be rare but the results do suggest that it is advisable to use a multigene phylogenetic inference method with independent gene trees under a coalescent prior conditional on the species tree (see below). The coalescent prior gives appropriate prior weight to different possible gene trees for a particular species tree, leading to a consistent estimator of

the species tree and avoiding the problem identified by Degnan & Rosenberg (9) and Kubatko & Degnan (33).

Effects of Genomic Recombination

Correlated gene trees and recombination. Genes that are located on the same linear segment of DNA (e.g., on a single chromosome) will, in the absence of processes such as gene conversion or meiotic recombination, have identical gene trees. This is the case for genes of the mitochondrial genome, for example. Although the gene trees are identical in this case, they may still differ from the species tree. In a phylogenetic analysis of genes that are expected to have identical trees, it is appropriate to model the process assuming a common gene tree. Substitution rates and other parameters may still vary across genes, however, so separate parameters for these processes may still be needed (see discussion below). For regions that have undergone meiotic recombination the gene trees may differ but are often highly correlated. Hudson (25) described the joint probability distribution of the gene trees for a pair of linked loci under the ancestral recombination and coalescence process within a single population.

The probability that, in a diploid species, the first crossover event on the interval between the linked genes occurs prior to time T (in units of generations) is approximately

$$\int_0^T \mathrm{e}^{-4N_e r t} dt = \frac{1}{4N_e r}\left(1 - \mathrm{e}^{-4N_e r T}\right),$$

where r is the recombination fraction between the genes per generation (e.g., the linkage distance in units of Morgans). The probability that the first coalescence occurs prior to time T is approximately

$$\int_0^T \frac{1}{2N_e} \mathrm{e}^{-t/2N_e} dt = 1 - \mathrm{e}^{-T/2N_e}.$$

If ancestral polymorphism exists, the recombination and coalescence processes compete to determine the gene tree correlations. If a coalescence event occurs first, there is an identical gene and species tree at the node for the linked genes. If a recombination occurs first, there are independent gene and species trees at the node (**Figure 3**).

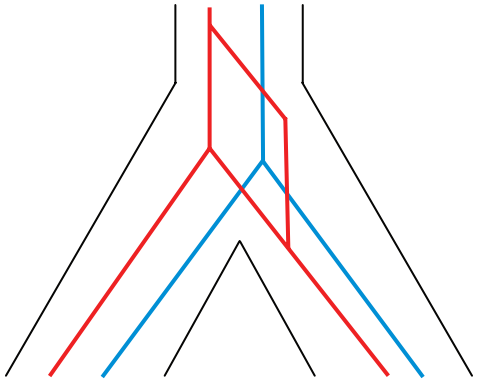

Figure 3

Node of a species tree illustrating the effects of coalescence and meiotic recombination in determining whether two genes have identical or different ages for a most recent common ancestor (MRCA) in the gene tree. The lineages in blue represent a case where a coalescence event occurs before a recombination. In that case, there is a shared MRCA for the two genes. The lineages in red represent a case where a recombination occurs before a coalescence event. Two MRCAs then exist and the two genes have independent histories at the node.

Recombinations and coalescences have the same cumulative probability distribution if $r = 1/(4N_e)$, whereas if $r \gg 1/(4N_e)$ the recombination process will dominate (and gene trees will tend to be independent) and if $r \ll 1/(4N_e)$ the coalescent process will dominate (and gene trees will tend to be identical). For example, if the human-chimpanzee ancestor had an effective population size of $N_e = 100{,}000$ (61), then $1/(4N_e) = 1/400000 = 0.0000025$ is the critical value for r. On average, in the human genome 1 Mb $\approx$ 1 cM and the critical value therefore corresponds to a physical interval approximately 250 bp in size. Thus, gene trees for human autosomal genes separated by a distance greater than $\approx$2.5 kb might be treated as independent for phylogenetic inference. Of course, if the ancestral population size is small

then all gene trees will agree with the species tree and the recombination process is irrelevant; this is probably not true for the human-chimpanzee divergence mentioned earlier.

The theory presented above provides, at best, a rough guide for choosing whether to model gene trees for linked genes as dependent or independent in a phylogenetic analysis. Both N_e and the relationship between physical distance and genetic distance can be expected to vary considerably even among closely related species. Moreover, the relationship between physical distance and genetic distance can also vary greatly across a genome even within species; this is evident from recent studies in humans and other species for which relatively precise meiotic recombination rates as well as a completed genome sequence are available (31, 37). The effective population size can also vary across a genome owing to past evolutionary forces such as directional selective sweeps affecting particular loci (which leads to younger genealogies than expected under a neutral coalescent and smaller N_e), or overdominance (which leads to older genealogies and larger N_e).

Instead of using population genetic models to predict the physical distances over which genes may be treated as independent for the purposes of a phylogenetic analysis, as we have done above, one could attempt to infer the presence of recombination using the sequence data (17, 38), so that the correlations among gene trees are inferred as part of the phylogenetic analysis. Although promising, such approaches can present significant computational challenges. A conservative approach would be to use only a small subset of the tens or hundreds of thousands of genes available for many genomes, choosing genes separated by intervals of say 1 Mb. Although this might seem wasteful, it is clear from many empirical analyses and simulation studies that species phylogenies can often be precisely inferred with far fewer genes than are available in the genome as a whole, suggesting that the gain in simplicity and reduced model complexity and assumptions may at least partially offset the loss of data in this case.

Horizontal (lateral) gene transfer. Horizontal gene transfer (HGT) between species is well documented in prokaryotes (32), and typically occurs via processes such as bacterial transformation and conjugation. HGT is also likely to occur in most eukaryotes via processes such as transduction of viral genes. Even very low rates of HGT can have a large impact on phylogeny, disrupting the usual patterns of vertical transmission of genes from parents to offspring and causing gene trees to differ from the species tree. Indeed, with high levels of HGT even the existence of a species tree may be questioned (12). In principle, one could develop a model of horizontal gene transfer among species, exploiting similarities to the population genetic process of migration among populations. However, most recent attempts to model HGT for the purposes of phylogenetic inference have used much simpler models (see below).

Horizontal gene transfer (HGT): the transfer of genetic material from one species to another species; also known as lateral gene transfer

PHYLOGENOMIC INFERENCE

Analyzing Multigene Data

Traditional approaches to phylogenetic inference make the assumption (implicitly or explicitly) that a single phylogenetic tree underlies the data. The population genetic and evolutionary processes outlined above can contradict this assumption when sequence data from multiple genes are analyzed, potentially producing erroneous conclusions in a phylogenetic analysis. Here we discuss some of the methods that have been proposed for analyzing multigene datasets, with an emphasis on parametric statistical methods.

The basic parameters of a phylogenetic analysis are the tree topology, τ, the branch lengths, ν, and the substitution model parameters, θ (e.g., nucleotide frequencies and transition/transversion rate ratios). With a single gene the maximum likelihood estimator of τ and ν is obtained by maximizing the likelihood of the sampled sequences $\mathbf{X} = \{X_i\}$ with respect to the model parameters (18) for a given tree

topology τ_k,

$$L_k = \max_{\theta,\nu} \prod_{i=1}^{n} f(X_i|\theta, \nu, \tau_k),$$

and then choosing the topology τ_j with $L_j > L_k$ for all $k \neq j$. The equivalent Bayesian formulation (47) places priors on all parameters and produces a posterior distribution of phylogenetic trees,

$$f(\tau|\mathbf{X}) = \int_{\theta}\int_{\nu} f(\mathbf{X}|\theta, \nu, \tau) f(\tau, \nu) f(\theta) d\theta d\nu / f(\mathbf{X}).$$

Several strategies for analyzing data from multiple loci or genome regions are available. The simplest is to concatenate the sequences, replacing the missing data with question marks, and analyze the data as one "supergene." This approach is commonly used (e.g., see 40). In likelihood and Bayesian methods, such an analysis uses the same set of parameters for all genes and ignores possible heterogeneities among the genes. It is known that different genes may evolve at very different rates and could have different base compositions or different transition-transversion rate biases, etc., but such differences are ignored by the concatenation approach. If factors leading to gene and species tree conflicts exist, such as ancestral polymorphism or HGT, concatenation can also lead to inconsistency of the phylogenetic inference method (e.g., see 33).

An alternative method is to analyze the different genes separately, and then sum log likelihoods across genes. In a modeling framework, the concatenation approach assumes the simplest model, in which one set of parameters applies to all genes, whereas the separate analysis uses the most general model, in which each gene has its own set of parameters. Proposed methods of multigene analysis correspond to various ways of partitioning the genome, allowing model parameters to vary across genes. In this review, we distinguish between partitioning of the genome according to the substitution model and partitioning according to the underlying gene trees. Often, both types of partitions are needed for accurate phylogenetic inference.

Partitioning the genome according to substitution process. An important factor to accommodate in all multigene analyses is variation in substitution rates (expected branch lengths) and parameters of the substitution model across genes. The Bayesian formulation is

$$f(\tau, \underline{\nu}|\mathbf{X}) = \int_{\underline{\theta}} f(\mathbf{X}|\underline{\theta}, \underline{\nu}, \tau) f(\tau, \underline{\nu}) f(\underline{\theta})\, \mathrm{d}\underline{\theta} / f(\mathbf{X}),$$

where, if there are k genes, $\underline{\theta} = \{\theta_1, \ldots, \theta_k\}$, $\underline{\nu} = \{\nu_1, \ldots, \nu_k\}$, θ_j are the parameters of the substitution model for the jth gene, ν_j are the branch lengths for the jth gene, and so on.

Yang (68) discussed strategies for partitioning the data according to the substitution model and used maximum likelihood to implement a number of models that lie between the two extremes mentioned above (i.e., one model for all genes versus a separate model for each gene), which allow some aspects of the evolutionary process to be the same among genes while other aspects are different. All models implemented by Yang (68) assume that the branch lengths are proportional. Biologically, this model assumes that either a molecular clock exists (with different genes evolving at different overall rates) or lineage-specific rate changes apply similarly across all genes. Pupko and colleagues (46) implemented additional models, and used the Akaike information criterion (AIC) (1) to evaluate their fit to real data sets.

Ren and coworkers (49), Shapiro and coworkers (55), and Bofkin & Goldman (5) evaluate different strategies for analyzing protein-coding genes and conclude that it is important to account for the differences in the evolutionary process (such as rates, base compositions, and transition/transversion rate ratios) among the three codon positions. Several studies also evaluated the utility of codon models (24, 42) for phylogenetic analysis of protein-coding genes, and found that although computationally expensive, they were effective in recovering difficult phylogenies (43, 49). Nucleotide-based models that account for the differences in the three codon positions offer

a computationally feasible substitute (55). Differences among the codon positions, likely a reflection of how purifying natural selection acting on the protein interacts with the genetic code, are a major feature of the evolutionary process of protein-coding genes, and it is important to accommodate them in a phylogenetic analysis of such data (6, 56, 64). Rate variation among sites within a partition (codon position) can be accommodated via the use of a random rate distribution such as the gamma (65, 66).

Programs that implement partitioned substitution models. Models for phylogenetic analysis of data from multiple genes or genomic partitions using maximum likelihood are not well developed in currently available programs. The BASEML and CODEML programs in the PAML package implement the models of Yang (68), but these programs do not include efficient tree search algorithms, and are not usable in phylogeny reconstruction with more than a dozen species. The PAUP* program (59) includes the site-specific rates model (invoked by setting up partitions and then specifying the option rates = sitespec in the lset command). This is the same as the proportional-branch model (68) and allows different partitions to have different rates—the option might thus be more appropriately called a partition-specific rates model. The program does not include models that allow other features of the evolutionary process (such as base compositions, transition/transversion rate ratio, and among-site rate variation) to differ among partitions.

The likelihood models discussed above can be used in Bayesian phylogenetic inference, as demonstrated by Suchard and colleagues (58) and Nylander and coworkers (44). MrBayes (51) is currently the only Bayesian tree-reconstruction program that has implemented a variety of models for combined analysis of multipartition datasets (invoked by setting up partitions and then using the link and unlink commands).

Partitioning the genome across local gene trees. To allow for processes such as lineage sorting and HGT that can cause gene trees and species trees to differ, one must partition the genome so that different genes may have potentially different underlying (true) gene trees. Such processes lead to different priors on the set of gene trees for the sampled genes (possibly conditional on a species tree). For example, Rannala & Yang (48) used a prior on gene trees derived on the basis of a coalescent process operating within the context of a fixed species tree topology (48). Under the molecular clock or relaxed clock models the species divergence times (rather than the branch lengths) are considered parameters of the model and may be estimated if some independent information is available concerning rates of substitution, for example from fossil-based age calibrations on one or more nodes (48). This approach effectively allows for variable gene trees owing to lineage sorting in estimating ancestral population sizes and species divergence times by integrating over the probability distribution of unobserved gene trees. Although the probability density of gene trees was not of specific interest, this information is also generated as a byproduct of the analysis. By integrating over unobserved gene trees under a coalescent prior the lineage sorting process is explicitly accounted for in the model, and the problems with inconsistency of phylogenetic inference methods described by Degnan & Rosenberg (9) are no longer an issue.

Lineage sorting: random coalescent events in ancestral populations that generate gene trees for the extant species that differ from the species tree

The Bayesian approach has been extended to integrate across the unobserved species tree as well as gene trees, effectively allowing a common species tree to be estimated while accounting for the effects of lineage sorting (15). The difficulty of performing the numerical calculations for this model led to the use of several ad hoc approximations, whose effect on accuracy needs to be studied further; the development of fully Bayesian methods is desirable.

Researchers have also developed phylogenetic inference methods that are intended to accommodate gene tree variations that arise

as a result of horizontal gene transfer. These methods include a Bayesian approach that integrates over gene trees using a heuristic model of recombination based on the SPR algorithm. The SPR algorithm is commonly used to search among phylogenetic trees—however, in this case, the SPR is used as a physical model of HGT (57). The use of SPR in this context has been criticized because it does not impose the obvious time constraint that lineages involved in a horizontal gene transfer event must be contemporary (22). Approximate maximum likelihood methods have also been proposed that aim to estimate the extent of HGT (36) or to estimate phylogenetic networks (29). The likelihood method of Linz (36) is practical with only a very small number of taxa and a composite-likelihood approximation is used to deal with larger numbers of taxa. It is not currently clear how well these approximations perform. Many methods have been proposed for constructing networks from data comprised of genes with potentially different underlying gene trees. However, such networks have no clear biological interpretations. Bayesian phylogenetic models of host-parasite cospeciation (28) in which host switching occurs share many features with HGT processes and could potentially be adapted for use in modeling HGT for phylogenetic inference, with the host tree playing the same role as the species tree and the parasite trees playing the role of gene trees.

The recent likelihood method of Ané and colleagues (2) allows an independent gene tree and substitution model to underlie each gene. Genes are analyzed separately using Bayesian analysis and posterior distributions of gene trees are then combined through the use of a gene-to-tree map that is, in turn, used to estimate the proportion of genes for which any given clade is true (the sample-wide concordance factor). A drawback of this method is that the prior does not explicitly model biological processes such as lineage sorting or HGT and therefore the clades with high concordance factors are not necessarily present in the species tree.

Strategies and Difficulties in Site Partitioning

Thus far, we have used the term gene to refer operationally to a particular set of sites in a sequence alignment. Although it is often natural to partition sites according to genes, other strategies may be more appropriate for particular data sets. The main consideration in partitioning sites should be to accommodate the most important types of large-scale heterogeneity among sites. Features to be considered may include the evolutionary rate, the base composition, or the local genealogical tree topology induced by coalescent and/or HGT processes. For example, in vertebrate mitochondrial protein-coding genes, the three codon positions have very different evolutionary rates as well as different base compositions and transition/transversion rate ratios, but the differences among the genes are not so great (e.g., see 35). In this case it is better to partition the data by codon position than by gene (e.g., see 54, 67).

In the maximum likelihood method, allowing separate rates for each gene implies a great number of parameters if thousands of genes are analyzed jointly. Estimating so many parameters by maximum likelihood may pose computational problems. Furthermore, the statistical performance of the method may be affected as well, especially if some genes are small or otherwise uninformative (19). A standard statistical practice for dealing with the problem of too many parameters is to use a statistical distribution to describe the among-partition rate variation, in the same way that the gamma model is used to describe within-partition substitution rate variation among sites (65). The likelihood function then involves integrating over the among-partition distribution, and may be expensive to calculate.

In comparison, such multiparameter models are relatively easy to implement in a hierarchical Bayesian framework. The rate (or other parameters reflecting features of the evolutionary process) for a partition is assigned a prior, and integration over the prior is carried out

in the Markov chain Monte Carlo algorithm in a straightforward manner. As in the case of likelihood methods, careful thought is needed to choose an appropriate partitioning of the sites.

Supermatrix and Supertree Controversies

Debate is ongoing in the literature concerning two particular strategies for phylogenetic analysis of multigene data sets, especially when some of the genes to be analyzed are not yet sequenced in some species. The supermatrix method concatenates sequences from multiple loci into a supersequence, with missing data represented as question marks, and uses the resulting data supermatrix to perform phylogenetic analysis. The supertree method instead conducts phylogenetic analysis on individual genes separately, and then uses one of several heuristic algorithms to combine the subtrees from the individual genes into a supertree for all species (for a summary, see 4). Several reviews have been published that either support the use of supermatrix methods (e.g., see 11) or instead advocate the use of supertree methods (3, 4, 52).

From a statistical modeling perspective, the debate is moot because both methods have serious drawbacks when used to analyze multigene data. The supermatrix method uses a simplistic substitution model that ignores heterogeneities in the evolutionary process among genes. Numerous simulation studies suggest that ignoring among-site or among-partition heterogeneities in the model can adversely affect phylogenetic analysis, sometimes causing systematic biases in the estimated tree (e.g., see 27, 34, 63). Moreover, as mentioned previously, simulation studies (33) show that lineage sorting may lead to inconsistent estimates of the species tree when concatenation is used, although the circumstances in which this may occur are probably rare.

The supertree method estimates an independent set of parameters for every gene and may overfit the data, inflating the variances of the estimates. Most supertree methods for constructing composite species trees use heuristic algorithms that lack a statistical basis and ignore uncertainties in the estimated subtrees (such as bootstrap support values, Bayesian posterior clade probabilities, or estimated branch lengths). Although ad hoc approaches have been suggested to remedy this (7, 39), their statistical performance has not been adequately studied. Computer simulations (e.g., see 16) tend to suggest that supertree algorithms can perform poorly even in the best-case scenario where the multiple genes are of the same length and evolve at the same rate under the same evolutionary model so that the information content in each gene is roughly equal. For example, the performance of some supertree methods can even deteriorate with the inclusion of more genes in the dataset.

In summary, although supertree methods can be useful as a tool for generating empirical summaries of the phylogenetic trees obtained in different studies from different types of characters, they are not statistically efficient for analyzing genomic data from multiple loci. Much of the theoretical research on supertree algorithms, mostly using the parsimony method of phylogenetic tree reconstruction, has emphasized combinatorial properties and computational algorithms but has neglected to examine basic statistical properties of the methods.

FUTURE ISSUES

Phylogenetic inference using whole genome data poses tremendous statistical and computational challenges. There is a profound need to develop new models for the analysis of multigene or multipartition datasets that can accommodate factors such as the heterogeneity of the evolutionary process among genes, or partitions, in whole genome phylogenetic analyses. Improved statistical methods are needed that account for genomic variation in evolutionary rates, transition/transversion rate ratios, and local gene trees. Moreover, there is an urgent need to develop efficient

computer programs for combined analysis of multipartition datasets, particularly those suitable for parallel computer systems. Genome-wide phylogenetic inference will likely continue to present challenging problems for computational biologists for years to come.

SUMMARY POINTS

1. Genome-wide datasets offer opportunities for resolving difficult phylogenetic problems but also pose computational and statistical challenges for data analysis.
2. Multigene datasets should ideally be analyzed jointly, with the heterogeneity among data partitions appropriately accounted for in the model.
3. Neither supermatrix nor supertree methods are adequate for analysis of multipartition genome-wide data sets.
4. It is important to develop new statistical models and computational algorithms for efficient analysis of multigene data sets.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This paper was written while the authors were guests of the Laboratory of Biometrics, University of Tokyo. We express our gratitude for the generous hospitality of our host, Professor Hiro Kishino. Z.Y. is supported by the Natural Environmental Sciences Research Council (NERC, UK).

LITERATURE CITED

1. Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Contr AC* 19:716–23
2. Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* 24:412–26
3. Bininda-Emonds ORP. 2004. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Dordrecht, the Netherlands: Kluwer Academic
4. Bininda-Emonds ORP. 2005. Supertree construction in the genomic age. *Methods Enzymol.* 395:745–57
5. Bofkin L, Goldman N. 2007. Variation in evolutionary processes at different codon positions. *Mol. Biol. Evol.* 24:513–21
6. Buckley TR, Simon C, Chambers GK. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst. Biol.* 50:67–86
7. Burleigh JG, Driskell AC, Sanderson MJ. 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Syst. Biol.* 55:426–40
8. Chen F-C, Li W-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68:444–56
9. Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:e68
10. Degnan JH, Salter LA. 2005. Gene tree distributions under the coalescent process. *Evolution* 59:24–37
11. de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. *Trends Ecol. Evol.* 22:34–41
12. Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–29

13. Eck RV, Dayhoff MO. 1966. *Atlas of Protein Sequence and Structure*. Silver Spring, MD: National Biomedical Research Foundation
14. Edwards AWF. 1970. Estimation of the branching points of a branching diffusion process. *J. R. Stat. Soc. B* 32:155–74
15. Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci. USA* 104:5936–41
16. Eulenstein O, Chen D, Burleigh JG, Fernandez-Baca D, Sanderson MJ. 2004. Performance of flip supertree construction with a heuristic algorithm. *Syst. Biol.* 53:299–308
17. Fang F, Ding J, Minin VN, Suchard MA, Dorman KS. 2007. Brother: relaxing parental tree assumptions for Bayesian recombination detection. *Bioinformatics* 23:507–8
18. Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368–76
19. Felsenstein J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J. Mol. Evol.* 53:447–55
20. Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19:99–113
21. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. 1995. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science* 269:496–98, 507–12
22. Galtier N. 2007. A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst. Biol.* 56:633–42
23. Goffeau A, Aert R, Agostini-Carbone ML, Almed A, Aigle M, et al. 1997. The yeast genome directory. *Nature* 387(suppl.):1–105
24. Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11:725–36
25. Hudson RR. 1983. Properties of a neutral allele with intragenic recombination. *Theor. Popul. Biol.* 23:183–201
26. Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–17
27. Huelsenbeck JP. 1995. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol. Biol. Evol.* 12:843–49
28. Huelsenbeck JP, Rannala B, Larget B. 2000. A Bayesian framework for the analysis of cospeciation. *Evolution* 54:352–64
29. Jin G, Nakhleh L, Snir S, Tuller T. 2006. Maximum likelihood of phylogenetic networks. *Bioinformatics* 22:2604–11
30. Kingman JFC. 1982. The coalescent. *Stoch. Process Appl.* 13:235–48
31. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* 31:241–47
32. Koonin E, Makarova K, Aravind L. 2001. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* 55:709–42
33. Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56:17–24
34. Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11:459–68. Erratum. 1995. *Mol. Biol. Evol.* 12:525
35. Kumar S. 1996. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics* 143(1):537–48
36. Linz S, Radtke A, von Haeseler A. 2007. A likelihood framework to measure horizontal gene transfer. *Mol. Biol. Evol.* 24:1312–19
37. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–84
38. Minin VN, Dorman KS, Fang F, Suchard MA. 2007. Phylogenetic mapping of recombination hotspots in human immunodeficiency virus via spatially smoothed change-point processes. *Genetics* 175:1773–85
39. Moore BR, Smith SA, Donoghue MJ. 2006. Increasing data transparency and estimating phylogenetic uncertainty in supertrees: Approaches using nonparametric bootstrapping. *Syst. Biol.* 55:662–76

40. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, et al. 2001. Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science* 294:2348–51
41. Nei M. 1977. Standard error of immunological dating of evolutionary time. *J. Mol. Evol.* 9:203–11
42. Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–36
43. Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree—the power and pitfalls of phylogenomics. *Genome Biol.* 8:R199
44. Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53:47–67
45. Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568–83
46. Pupko T, Huchon D, Cao Y, Okada N, Hasegawa M. 2002. Combining multiple data sets in a likelihood analysis: which models are the best? *Mol. Biol. Evol.* 19:2294–307
47. Rannala B Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* 43:304–11
48. Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–56
49. Ren F, Tanaka H, Yang Z. 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Syst. Biol.* 54:808–18
50. Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* 56:389–99
51. Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–74
52. Sanderson MJ. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends Ecol. Evol.* 13:105–9
53. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, et al. 1977. Nucleotide sequence of bacteriophage ϕX174. *Nature* 265:687–95
54. Sasaki T, Nikaido M, Hamilton H, Goto M, Kato H, et al. 2005. Mitochondrial phylogenetics and evolution of mysticete whales. *Syst. Biol.* 54:77–90
55. Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* 23:7–9
56. Simon C, Buckley TR, Frati F, Stewart JB, Beckenbach AT. 2006. Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* 37:545–79
57. Suchard MA. 2005. Stochastic models for horizontal gene transfer: Taking a random walk through tree space. *Genetics* 170:419–31
58. Suchard MA, Kitchen CM, Sinsheimer JS, Weiss RE. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst. Biol.* 52:649–64
59. Swofford DL. 2003. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4*. Sunderland, Massachusetts: Sinauer Associates
60. Tajima F. 1983. Evolutionary relationship of DNA-sequences in finite populations. *Genetics* 105:437–60
61. Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.* 48:198–221
62. Tateno Y, Nei M, Tajima F. 1982. Accuracy of estimated phylogenetic trees from molecular data. *J. Mol. Evol.* 18:387–404
63. Tateno Y, Takezaki N, Nei M. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* 11:261–77
64. Whelan S, de Bakker PI, Quevillon E, Rodriguez N, Goldman N. 2006. Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Res.* 34:D327–31
65. Yang Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10:1396–401
66. Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–14

67. Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139:993–1005
68. Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42:587–96
69. Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* 162:1811–23