Discord among the gene trees of multilocus data has motivated the development of phylogenetic approaches that account for gene-tree heterogeneity in the estimation procedure. Rather than equating a gene tree with the phylogenetic history, the new approaches explicitly consider the relationships between gene trees and the underlying history of species divergence, providing direct estimates of species trees (Fig. 1). The inherent appeal of these approaches is 2-fold. Incorporating information contained in the distribution of gene trees not only extracts phylogenetic signal, but modeling the relationship between the gene trees embedded in a species tree also reveals the biological processes that have influenced the diversification history and shaped organismal genomes. In contrast, ignoring the variance in genealogical histories (e.g., concatenating loci into a single supermatrix) disregards an inescapable biological reality—gene trees differ for a variety of reasons (reviewed in Maddison 1997; Degnan and Rosenberg 2009). As such, when the natural variation in gene trees is not taken into account during phylogenetic estimation, the reliability of inferences from such approaches is drawn into question (Degnan and Rosenberg 2006; Kubatko and Degnan 2007; Huang and Knowles 2009), historical scenarios with recently diverged taxa that have not reached reciprocal monophyly has become intractable (Carstens and Knowles 2007), and interpretation of the support for bipartitions across taxa is problematic (Mossel and Vigoda 2005).

CONCEPTUAL BASIS FOR SPECIES-TREE ESTIMATION

The direct estimation of species trees, as opposed to equating reconstructed gene trees with the phylogenetic history of species, expands the role of explicit models of evolutionary change. The most basic approaches to species-tree estimation encompass the processes of both nucleotide substitution within gene lineages (e.g., Felsenstein 1981; random sorting among gene lineages (e.g., Takahata 1989); other processes that also may be modeled include gene flow or hybridization (e.g., Thang et al. 2007; Meng and Kubatko 2009). Under the basic framework for species-tree estimation, coalescent theory is used to relate a sample of gene trees to the species tree (Degnan and Salter 2005), whereas nucleotide substitution is modeled separately (i.e., stochastic models of nucleotide substitution and gene-lineage sorting are separate and independent processes; Maddison 1997; Rosenberg 2002). For any single species tree (i.e., specific topology and branch lengths), coalescent theory gives the probability distribution of gene trees (i.e., the probability for each gene-tree topology, including those topologies that do and do not match the species tree).

The degree of gene-tree heterogeneity, and hence discord from the underlying species tree, will differ among taxa because the probability distribution of gene trees depends on the specific details of the divergence history. Gene-tree discord becomes increasing unlikely as the time between successive speciation events becomes larger (i.e., longer internal branches in the species tree) and species origins are older. With rapid species diversification and/or when species have originated recently (i.e., when the time separating speciation events or the species origin is less than \(4N_e\) generations, where \(N_e\) is the effective population size), there is insufficient time for the fixation of gene lineages by genetic drift. Consequently, gene lineages that persist into deeper portions of the species trees may coalesce with gene lineages that are not from the most closely related species (i.e., deep coalescence causes a mismatch between the gene tree and the species tree; Fig. 1).

Species-Tree Inference Procedures

The new approaches for species-tree inference (reviewed in Degnan and Rosenberg 2009) fall into 2 general classes: summary-statistic-based approaches versus maximum-likelihood (ML) and Bayesian methods. They differ in 1) the extent that the full information content of the data is considered, 2) how coalescent times are summarized or whether a full probabilistic model of the coalescent is considered, and 3) if uncertainty in the estimated gene trees is incorporated (e.g., Arvestad et al. 2003; Maddison and Knowles 2006; Ané et al. 2007; Liu 2008; Cranston et al. 2009; Kubatko 2009; Kubatko et al. 2009; Liu et al. 2009; McCormack et al. 2009).

These differences have a number of important ramifications. For example, large data sets can exceed the computational limits of the parameter-rich methods that employ sophisticated algorithms for estimating the species trees as well as divergence times and population sizes (e.g., the program BEST; Liu 2008). In contrast, methods based on summary statistics take little computational time and can be applied to phylogenomic scale data sets, with the trade-off that by not fully utilizing the content in the sequence data they may require more
data to achieve the same level of accuracy. Conversely, if there is insufficient information among the loci (i.e., limited variation), the performance of parameter-rich methods may be compromised (e.g., lack of information for accurately estimating branch lengths, which can cause problems with convergence). Depending on how information about the coalescent is incorporated into the method (e.g., using estimates of coalescent times, ranks of coalescent times, or minimum divergence times), the different approaches may exhibit differing levels of sensitivity to properties of the data, such as limited levels of variation or biological causes of discord that are not considered by the model. In summary, it appears that the best method for species-tree inference (at least among those currently available) is very much data set dependent and will be determined by the features of the underlying biological processes of the divergence history as well as sampling attributes associated with the number and choice of loci.

**Need for a Paradigm Shift?**

The vast majority of phylogenetic methods focus on the estimation of gene trees (Felsenstein 2004), even though it is the species trees—trees of populations and species in which gene trees are embedded—that are of primary interest. Is this merger of phylogenetic and population genetic perspectives really necessary? And now that we have discovered that many empirical data sets are characterized by gene-tree discord, should we really be working in a species tree (i.e., bifurcating) framework?

Evidence of incongruence among gene trees is not sufficient for abandoning a species-tree framework. In fact, the species-tree framework is required to interpret the processes underlying gene-tree heterogeneity and therefore, identify events of biological interest and understand the processes shaping the genome (e.g., distinguishing between coalescent stochasticity vs. gene flow or horizontal gene transfer as the source of discord; see Degnan and Rosenberg 2009). Even for groups typified by high levels of gene-tree heterogeneity, such as bacteria, the utility of a species-tree framework is still a meaningful concept (Galtier and Daubin 2008). Whether extensive gene-tree incongruence has erased any bifurcating phylogenetic signal is an empirical question that has to (and can) be evaluated. For example, even for very recent radiations where widespread incomplete lineage sorting would seem to obfuscate any phylogenetic signal, genetic correlation due to co-ancestry is still detectable (Maddison and Knowles 2006). However, phylogenetic signal can only be extracted with methods where the processes generating the variance in gene trees are explicitly modeled, irrespective of whether the discord arises from horizontal gene transfer, as in the case of bacteria, or extensive deep coalescence of gene lineages, as in the case of evolutionary radiations.

Notwithstanding the problems with ignoring gene-tree incongruence (e.g., unreliability of phylogenetic inferences and inflated nodal support values with analysis of concatenated data; Mossel and Vigoda 2005; Kubatko and Degnan 2007; Cranston et al. 2009), there is the undeniable and simple matter that the species tree—not the trees of genes—are the focus of systematic analyses and therefore, identify events of biological interest and understand the processes shaping the genome (e.g., distinguishing between coalescent stochasticity vs. gene flow or horizontal gene transfer as the source of discord; see Degnan and Rosenberg 2009). So the answer to the questions of 1) the utility of a species-tree framework, even with pervasive gene-tree discord, and 2) the need for a shift in the use and interpretation of gene trees, appears to be a resounding “yes.”

**The Symposium**

At the 2008 evolution meetings in Minneapolis, we held the first symposium dedicated to this promising area of phylogenetic research—the direct estimation of species trees. The papers presented here provide a sampling of some of the diversity of approaches for species-tree estimation. These works also highlight some of the important conceptual issues surrounding species-tree estimation that have yet to be answered. Other presenters at the symposium, but whose work is not published here, include Noah Rosenberg, Cecile Ané, and Luay Nakhleh.

The paper by Liu et al. (2009) presents 2 methods for species-tree estimation based on summary statistics of
coalescent times. These approaches complement other more computationally intensive methods that they have developed (see Liu 2008) and add to existing approaches that rely on summary statistics for inferring the underlying species tree (e.g., Maddison and Knowles 2006; Mossel and Roch 2007). With all these methods, discord between gene trees and a species tree is modeled as arising exclusively from the deep coalescence of gene lineages. In comparing the performance of the 2 proposed methods, species-tree estimation using average ranks of coalescences (STAR) and species-tree estimation using average coalescence times (STEAC), STAR outperforms other summary-statistic-based approaches when there is substantial variation in substitution rates among gene lineages. Because the method relies on the rank of coalescence times rather than the distances between gene lineages (i.e., alleles), the STAR method is robust to violations of molecular clock assumptions (to the extent that the topology of gene trees is accurately reconstructed). Additional comparisons of the accuracy of species trees estimated with STAR and STEAC as well as with the shallowest coalescent (Maddison and Knowles 2006) and global latest split (Mossel and Roch 2007) show how the performance of the methods can differ (i.e., STAR is not superior for all histories of species divergence). Because these 4 methods differ in how the coalescent is summarized, Liu et al. (2009) show that the best performing method will vary depending on whether the primary contributing factor to gene-tree–species-tree discord arises from the coalescent versus substitution process. Likewise, the accuracy of species-tree estimates may be seriously compromised when processes other than the coalescent (e.g., gene flow) contribute to incongruence between the gene trees and the species tree. However, as highlighted in the paper of Liu et al. (2009) the accuracy of a method when processes other than the sorting of gene lineages have generated gene-tree heterogeneity will again depend on how the procedure incorporates information from the coalescent (e.g., relying on average coalescence times as opposed to minimal coalescence times should make the procedure more robust to low levels of gene flow; see also Maddison and Knowles 2006; Eckert and Carstens 2008).

Given that processes other than the coalescence of gene lineages can contribute to discord between gene trees and species tree, there is a critical need for consideration of additional sources of gene-tree incongruence. The paper by Kubatko (2009) introduces a method for phylogenetic analysis where not only the process of gene-lineage coalescence but also hybridization might contribute to gene-tree incongruence. The simultaneous consideration of these processes in a likelihood framework provides a much needed advance, taking into account an additional biological process that may be a predominant cause of incongruence for groups in which the taxa originate via hybridization rather than a bifurcating trajectory from a common ancestor, as in plants and bacteria. With Kubatko’s (2009) method, a model-selection procedure based on standard information criteria is used to determine whether a hybridization event has occurred (i.e., identifying species of hybrid origin). Likelihood functions also describe the percentage of genes sharing a most recent common ancestor with one or the other of the two parental species (via the ML estimates of hybridization parameters), despite the presence of incomplete lineage sorting. The promising results from a simulation study, both in terms of model selection and estimating hybridization parameters, suggest that the approach could be expanded beyond the restrictive conditions modeled here—evidence for species of hybrid origin with a known phylogeny. As discussed by Kubatko (2009), further developments could include inferring species phylogenies with a search of tree space that includes hybrid species trees as well as incorporating increasingly sophisticated models within the common framework of the species tree, such as accommodating introgressive hybridization (i.e., gene flow following speciation without the formation of a distinct hybrid species).

The paper by Cranston et al. (2009) presents a phylogenetic analysis of empirical data using an approach that does not assume any specific biological process underlying the gene-tree incongruence (i.e., a Bayesian concordance analysis, BCA; Ané et al. 2007). The study showcases the ability of this species-tree approach to handle phylogenomic data sets (i.e., thousands as opposed to tens to hundreds of genes), where independent genes show significant topological incongruence, and there is evidence of multiple processes contributing to gene-tree discord (e.g., gene flow and coalescent stochasticity). It also draws attention to the limitations of computationally intensive methods (i.e., the program BEST; Liu 2008). In this case, in addition to problems with the scalability to phylogenomic data sets, the high degrees of gene-tree incongruence and other biological realities associated with the empirical data (i.e., limited levels of sequence divergence and processes other than gene lineage coalescence contributing to gene-tree heterogeneity) resulted in slow (or lack of) convergence of the Markov chain Monte Carlo used in Bayesian methods of species-tree inference. Cranston et al. (2009) also compare the results from BCA with an analysis of the concatenated data. These results emphasize the problems with interpreting nodal support values from analyses that do not explicitly consider gene-tree incongruence in the estimation procedure. In contrast with the results from BCA, the analysis of the concatenated data suffered from a disconnect from the biology of the species (i.e., failed to recognize known instances of gene flow of biological interest that was separate from the discord due to the stochasticity of the coalescent) and was subject to problems of inflated support values caused by the calculation of posterior probabilities under an assumed single tree when the reality was a mixture of gene-tree topologies (see also Mossel and Vigoda 2005).

A recurrent theme (albeit not the primary focus) found throughout the papers is the question of how many data are needed to accurately estimate the history
of divergence. The paper by McCormack et al. (2009) focuses explicitly on this question, and in line with the other presented papers, finds that the answer to the question is complicated. Unlike obtaining accurate gene-tree estimates, obtaining accurate species-tree estimates is more nuanced than simply collecting more data. This study demonstrates that there is not a single optimal sampling strategy—the notion that more genes are always better (as opposed to also collecting sequences from multiple individuals for a given sampling effort) is not defensible. By examining a diversity of speciation scenarios through simulations, as opposed to focusing on a single history, they show that the accuracy of the species trees differs, depending on both the specific details of the divergence history and the sampling strategy (see also Maddison and Knowles 2006). In general, the results confirmed that sampling more than 1 individual is critical to accurate species-tree estimation when the taxa have originated recently, in contrast to older divergences where sampling more loci substantially improved the accuracy of inferences. However, for any given sampling design (e.g., specific number of individuals and loci), the accuracy of species-tree estimates varied substantially across different species trees. McCormack et al. (2009) describe how the relative timing of species divergence (i.e., the relative length of the internal branches of a species tree) determines the accuracy of a species-tree estimate, not just the recency of divergence. They also compare the summary-statistic approach of minimizing the number of deep coalescents to the ML estimate of the species trees (Kubatko et al. 2009) and find that gains in accuracy were achieved under a full probabilistic model. Yet, the lack of increased accuracy when the number of individuals increased from 9 to 27 gene copies for the ML species-tree estimate hints that not only the quantity but also the quality of data may be important. Their findings also highlight the inherent challenge of reconstructing phylogenetic relationships for species that have undergone rapid radiations (see also Knowles and Chan 2008). Nevertheless, the study of McCormack et al. (2009) shows that species trees can indeed be accurately estimated in the face of widespread incomplete lineage sorting.

PROSPECTS FOR THE FUTURE

The presented papers represent just a subset of the recently developed approaches that are now available for species-tree inference (see review Degnan and Rosenberg 2009). A flurry of recent empirical papers that apply these methods is testament to the interest and promise surrounding the direct estimation of species trees (e.g., Carstens and Knowles 2007; Edwards et al. 2007; Belfiore et al. 2008; Brumfield et al. 2008; Carling and Brumfield 2008; Linnen and Farrell 2008; Liu et al. 2008; Themudo et al. 2009). Yet, this nascent area of phylogenetics requires much more work to be done and there are many outstanding questions. In contrasts to the extensive investigation of the factors affecting the accuracy of inferred gene trees (Felsenstein 2004), a range of issues from taxon sampling, missing data, to the choice of loci that could impact the accuracy of species-tree estimates have yet to be addressed. It is also unclear what properties of the species’ divergence history render estimates of the species tree more (or less) reliable. Nevertheless, the momentum generated by the increasing number of multilocus empirical data sets coupled with advances in computational tools and resources suggests that the merger of phylogenetic and population genetic perspectives will not only be enduring, but will also be inspiring many more interesting developments in the future (e.g., more sophisticated models that capture more of the biological processes that accompany species divergence). As with the field of statistical phylogeography (Knowles 2009), the move away from the tradition of using gene trees as a literal interpretation of the species history toward procedures that explicitly model the relationship between gene trees and species trees should propel the field of phylogenetics forward.

ACKNOWLEDGEMENTS

I thank the Society of Systematic Biologists for their financial support of the symposium, all the presenters of the symposium, and my co-organizer Scott Edwards. I also thank Jack Sullivan for all the helpful comments he provided on the symposium papers.

REFERENCES


