The potential and peril of the supertree approach: A response to van der Linde and Houle

PATRICK M. O’GRADY, RICHARD T. LAPOINT and GORDON M. BENNETT

Introduction

The family Drosophilidae currently contains over 3500 described species distributed in most habitat types (Bächli 2008). Members of the genus Drosophila have been the subject of genetic and developmental research for the past century and this group also serves as a model system for evolutionary and comparative studies (Markow & O’Grady 2007). Phylogenetic relationships within Drosophilidae, with particular emphasis on the largest genus, Drosophila, have been examined since Sturtevant’s taxonomic work in the 1920s (reviewed in Markow and O’Grady 2006). Throckmorton (1962, 1966, 1975, 1982) formalized morphological relationships in this family in a series of landmark papers. The first molecular studies on this group (Hubby 1963; Hubby and Lewontin 1966; Hubby and Throckmorton 1965; Lewontin and Hubby 1966; Throckmorton 1962) led the way to extensive allozyme work in the 1960s and 1970s (reviewed in Powell 1997). This work was followed by restriction fragment polymorphism studies in the early 1980s (e.g., DeSalle et al. 1987), widespread use of DNA sequences in the 1990s (DeSalle 1992; Pelendakis et al.1991; Pelendakis and Solignac 1993; Remsen and O’Grady 2002; Tatarenkov et al. 1999) and full genome sequences in the early 21st century (Adams et al. 2000; Pollard et al. 2006; Drosophila 12 genome Consortium 2007). These studies have continually refined our notions of phylogenetic relationships within this group. Early molecular phylogenetic work (Pelendakis et al. 1991; and Solignac 1993) suggested that the genus Drosophila was not a monophyletic group and subsequent studies have corroborated these findings (e.g., Remsen and O’Grady 2002).

Coincident with these studies was the widespread acceptance of phylogenetic methods and their applications to evolutionary biology. Systematists began revising para- and polyphyletic groups in favor of monophyletic assemblages in order to generate taxonomic hierarchies that are concordant with evolutionary relationships. The bulk of these revisionary efforts have been driven by primary character information (i.e., morphological or molecular characters) and follow traditional taxonomic and nomenclatural process. Recently supertree methods, the most common of which codes source trees into a presence-absence matrix represented using parsimony (MRP), have been introduced to summarize the results of primary data and generate hypotheses based on secondary analyses (phylogenetic trees). While this may be acceptable for character analysis in a comparative framework, it has some serious drawbacks includ-
ing lack of a statistical framework (but see Moore et al 2006), reliance on secondary sources of information, and non-independence between studies that partially overlap in taxon and character sampling.

Our critique of the van der Linde and Houle (2008) paper is based on three main points: (1) issues with supertree methods in general, (2) issues with the specific methodology employed by van der Linde and Houle, and (3) their approach of using of supertree data as the basis of taxonomic and nomenclatural revisions.

**Supertrees vs. supermatrices**

Supertrees are generated by coding multiple input tree topologies into a single character matrix and analyzing these data using a parsimony algorithm (Sanderson et al. 1998; Bininda-Emonds et al. 2002). The strength of such an approach is that a single composite phylogeny can be generated from many studies with different gene and taxon sampling strategies. The supermatrix approach (Gatesy et al. 2002) is an alternate method to summarize multiple phylogenetic studies. In a supermatrix analysis primary character information from different studies is combined into a large matrix and analyzed using phylogenetic methods. The benefits of this approach are that analyses can be placed in a statistical context, analytical methods that accommodate rate heterogeneity and different rates of evolution can be used, and conflicts between characters can be resolved in a total evidence framework. If researchers do decide to pursue a supertree approach, an understanding of the basic assumptions, and potential pitfalls of these methods is required.

**Non-independence.** – Supertree methods, while they provide an adequate summary of previous studies may not constitute independent corroboration of research findings. This is particularly true if the input trees are based on overlapping data matrices. This is an issue common to many studies and can lead to bias in resultant supertrees topology due to the effective overweighting of some data and underweighting of others. Gatesy and Springer (2004) state “all redundant source data are unacceptable and must be eliminated to avoid the non-independence problem.” While others (e.g., Bininda-Emonds et al. 2004) consider non-independence to be a major issue in tree combination, they point out that it is usually impossible to eliminate. Bininda-Emonds et al. (2004) present an explicit set of guidelines for minimizing the impact of overlapping data including selecting different genes that do not share a common evolutionary history, including non-overlapping taxa sets derived from the same gene, coding total evidence analyses generated by unique or partially overlapping data sets, picking the most recent study or most taxonomically inclusive analysis for a given locus, or using the hypothesis explicitly preferred by the authors. While not all authors may agree with these criteria, it is important that they be explicitly stated in their methods so that non-independence can readily be assessed and supertree analyses can be repeated by subsequent workers.

**Statistical support.** – Statistical support in phylogenetic analyses, including supertree analyses, is critical in expressing confidence in tree topology and in subsequent inference. Recent studies describe ways in which uncertainty can be expressed on a supertree using bootstrapping procedures (Burleigh et al. 2006; Moore et al. 2006). Three approaches have been described, each differing in the way in which they deal with issues of supertree reconstruction. The first, referred to as bootstrap-weighted MRP, weights data columns in the matrix proportionally to nodal support in source trees. This method, while it does not provide a direct estimate of nodal support, incorporates support in analyses of primary data into the supertree analysis (Moore et al. 2006; Ronquist 1996; Bininda-Emonds and Bryant 1998). This is preferable, because treating all nodes as equally supported, regardless of whether support at that node in a phylogenetic study was 100% or 51%, may have the effect of downweighting well supported nodes and upweighting those that are not statistically supported.

Two other methods allow for direct approximation of nodal support from the MRP matrix using source-tree bootstrapping and hierarchical bootstrapping (Moore et al 2006; Burleigh et al 2006; Page 2004). Source tree bootstrapping is similar in spirit to standard bootstrapping procedures, where resampled data columns are representative of input trees rather than amino acid of nucleotide sites. Hierarchical bootstrapping draws trees from original bootstrap sets derived from primary data in the creation of a supertree bootstrap matrix (Page