Towards simultaneous analysis of morphological and molecular data in Hymenoptera

James M. Carpenter & Ward C. Wheeler

Accepted 5 January 1999

Carpenter, J. M. & W. C. Wheeler. (1999). Towards simultaneous analysis of molecular and morphological data in Hymenoptera. — *Zoologica Scripta 28*, 251–260.
Principles and methods of simultaneous analysis in cladistics are reviewed, and the first, preliminary, analysis of combined molecular and morphological data on higher level relationships in Hymenoptera is presented to exemplify these principles. The morphological data from Ronquist *et al*. (in press) matrix, derived from the character diagnoses of the phylogenetic tree of Rasnitsyn (1988), are combined with new molecular data for representatives of 10 superfamilies of Hymenoptera by means of optimization alignment. The resulting cladogram supports Apocrita and Aculeata as groups, and the superfamly Chrysidoidea, but not Chalcidoidea, Evanioidea, Vespoidea and Apoidea.
James M. Carpenter, Department of Entomology, and Ward C. Wheeler, Department of Invertebrates, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, U SA. E-mail: carpente@amnh.org

Introduction

Investigation of the higher-level phylogeny of Hymenoptera is at a very early stage. Although cladistic analysis was first applied more than 30 years ago, in an investigation of the ovipositor by Oeser (1961), a comprehensive analysis of all the major lineages remains to be done. The phylogenetic trees from the literature survey by Königsmann (1976–1978), and the noncladistic, fossilized scenario by Rasnitsyn (1980, 1988), have substituted for an hypothesis based on a comprehensive cladistic analysis in recent discussions of evolutionary trends within the order. There has been no attempt to score an extensive morphological character matrix for the entire order until now (Ronquist *et al*. 1999). Likewise, although molecular sequence data recently began to be adduced in investigation of relationships among major lineages of Apocrita (Derr *et al*. 1992a, b), a comprehensive analysis has not yet been attempted. Studies have focused primarily on the parasitic, 'lower' Apocrita, with the most extensive analysis published thus far being that of Dowton *et al*. (1997), based on 37 sequences.

Given this situation, it is to be expected that simultaneous analyses of morphological and molecular data would also be lacking. There have been just two published, by Chavarría & Carpenter (1994), which dealt with social bees, and Carpenter (1997), which dealt with social wasps. Simultaneous analysis is clearly the method of choice for handling disparate data sets: as reviewed by Nixon & Carpenter (1996a), it best applies parsimony. Application of consensus techniques to the results of independent analysis of multiple data sets, as for example in so-called 'phylogenetic supertrees' (Sanderson *et al*. 1998), does not measure the strength of evidence supporting results from the different data sources — in addition to other drawbacks, such as common use of compromise techniques to calculate 'more resolved' 'consensus' trees, leading to semi-strictly supported trees (see Nixon & Carpenter 1996b).

The scoring of an extensive morphological character matrix for Hymenoptera by Ronquist *et al*. (1999) presents the opportunity for the first extensive simultaneous analysis within the order. We previously applied 23 hymenopteran sequences to the study of relationships among orders of Holometabola (Whiting *et al*. 1997). Recent work in the Molecular Systematics Laboratory at the American Museum of Natural History has been aimed at extending the breadth of that sample of Hymenoptera. In the present paper we combine a taxonomically broad sample of sequence data with the morphological characters scored by Ronquist *et al*. (1999). The results, albeit preliminary, are the first higher-level simultaneous analysis in Hymenoptera, and allows us to illustrate principles of simultaneous analysis in cladistics.

Background: previous phylogenetic analyses

In this section, we provide some detail on previous analyses of morphological and molecular data, by way of background to the present study.

Morphological data

The first numerical cladistic analysis of morphological data in Hymenoptera was by Brothers (1975), who studied 92 characters in 25 family group taxa of Aculeata. The taxa represented most of the families of Aculeata, and the results clearly established the paraphyly of the commonly recognized superfamily Scolioidea, in terms of three other superfamilies (Pompiloidea, Formicoidea and Vespoidea). Brothers reclassified the Aculeata, recognizing just three superfamilies where it had been common to use seven. This system of three superfamilies, Chrysidoidea, Apoidea and Vespoidea, is now widely accepted in general texts on Hymenoptera (e.g. Gauld & Bolton 1988; Naumann 1991; Goulet & Huber 1993; Hanson & Gauld 1995).

Brothers did not publish the matrix for his study. Carpenter (1990) scored it from Brothers' publication, and subjected it to analysis with microcomputer cladistic programs, obtaining results that differed in some respects from Brothers (1975). Brothers & Carpenter (1993) provided a revised, expanded matrix, consisting of 219 variables scored for 34 taxa, representing all the families of Aculeata. The results also differed in some details of family relationships from those of Brothers (1975), but relationships at the superfamily level were identical, with a sister-group relationship between Apoidea and Vespoidea, and Chrysidoidea in turn the sister-group to this clade. This scheme of relationships may be considered one of the most firmly established among superfamilies of Hymenoptera, and with the exception of a few problematic taxa, family group relationships within Aculeata are among the best understood in the order.

Aside from the relatively apical Aculeata, the group where relationships may be considered well established is the sawflies. Gibson's (1985) detailed study of thoracic structure established a sister-group relationship between Orussidae and Apocrita, a possibility previously suspected based on the parasitoid habits of Orussidae. Gibson's study also supported the closer relationship of Siricidae and Xiphydriidae to Apocrita than the family Cephidae — previously considered a candidate for sister-group to Apocrita (for example, by Königsmann (1977). Gibson's characters provided some data on relationships among families of Apocrita, and in particular Gibson questioned the scheme proposed by Rasnitsyn (1980), but he did not attempt a comprehensive analysis within Apocrita. Vilhelmsen (1996) made a detailed study of the morphology of the preoral cavity in Symphyta and three apocritan families, and provided a matrix of 25 characters for 18 taxa and root, analysis of which supported closer relationship of Orussidae to the Apocrita than any other sawfly. Vilhelmsen (1997) then provided a matrix of 98 characters scored for 21 taxa and root (15 family group taxa of Symphyta and six Apocrita). The analysis again supported the sister-group relationship between Orussidae and Apocrita, and a closer relationship of Siricidae and Xiphydriidae to Apocrita than the family Cephidae. Relationships with a few superfamilies of Symphyta remain uncertain, but the paraphyly of the suborder is not in doubt. Thus, the abandonment of the traditional division of Hymenoptera into the two suborders Symphyta and Apocrita is justified.

By contrast, no comprehensive analysis has been attempted of the nonaculeate Apocrita, a group sometimes recognized as an infraorder or informal group, the Terebrantes or Parasitica. Instead, there have been surveys of particular character systems focusing on this 'group': for example, Johnson (1988) on midcoxal articulations, Whitfield *et al.* (1989) on the metapostnotum, Quicke *et al.* (1992) on spermatozoa, Quicke *et al.* (1992) on ovipositor valvilli, Heraty *et al.* (1994) on the mesofurca and mesopostnotum, and Basibuyuk & Quicke (1997) on hamuli. These studies have all drawn cladistic inferences from the characters examined, and Heraty *et al.* (1994) even provided a matrix which they analysed cladistically, but there has been no integration of these diverse data sources.

The morphological character matrix for Hymenoptera of Ronquist *et al.* (1999) is the first to encompass most of the families in the order. Ronquist *et al.* scored this matrix from the character diagnoses given for the phylogenetic tree presented by Rasnitsyn (1988). They did not correct errors of interpretation nor add new characters, rather, the intent was to determine to what extent Rasnitsyn's proposed relationships were supported by actual analysis of the characters he cited. Brothers & Carpenter (1993) had performed a similar exercise for the aculeate families with Rasnitsyn's characters (and for Rasnitsyn's 1980 characters), although correcting several errors and misinterpretations. Analysis of the resulting matrices by Brothers and Carpenter differed in numerous respects from the trees drawn by Rasnitsyn, and this is also the case for the analysis by Ronquist *et al.*

It would be desirable to correct some of the interpretations in the matrix scored by Ronquist *et al.* and to include the new character systems mentioned above. But as it stands, this first extensive morphological matrix for Hymenoptera affords the first opportunity for a simultaneous analysis with molecular data. We take that opportunity here, in part to exemplify principles of simultaneous analysis in general.

Molecular data

The first DNA sequence data applied to the study of higher-level relationships within Hymenoptera were those of Derr *et al.* (1992a). They adduced sequences for the 16S mitochondrial rRNA from two sawflies (representing

Tenthredinoidea and Siricoidea), two aculeates (representing Apoidea and Vespoidea), three ichneumonoids and two chalcidoids. Unfortunately, as explained by Derr *et al.* (1992b), the Parasitica sequences were contaminated, by some vertebrate. Derr *et al.* (1992b) sequenced four ichneumonoids, and aligned these sequences to those previously obtained for the sawflies, aculeates and two dipteran outgroups, to produce a matrix of 510 base pairs, 217 of which were informative. Analysis of this matrix resulted in a single cladogram, which showed Aculeata as a group, with Ichneumonoidea its sister-group (thus, Apocrita was supported). However, the tenthredinoid and siricoid clustered together, a relationship dubious on morphological grounds (Gibson 1985; Vilhelmsen 1997).

Dowton & Austin (1994) made a more extensive study of the same gene, which was sequenced for representatives of 14 superfamilies of Hymenoptera (four sawfly, eight Parasitica, two Aculeata), for a total of 31 terminals and two dipteran outgroups, with 386 informative characters. The consensus of the resulting six cladograms resolved most of the superfamilies for which multiple representatives were included, but not Vespoidea, Ichneumonoidea or Tenthredinoidea. Aculeata was shown as a group, but Apocrita was not, with Stephanoidea the sister-group to the siricoid. Symphyta was otherwise unresolved, with no sister-group to Apocrita indicated. Relationships among the superfamilies of Apocrita were largely unresolved. Dowton and Austin also performed 'statistical' analyses, including testing for skewness, bootstrapping and T-PTP. They presented a tree (their Fig. 3) based on these analyses that was more resolved than the consensus tree. That resolution is spurious: the so-called statistical techniques are in reality pseudostatistical (on skewness, see Källersjö *et al.* 1992; on bootstrapping see Carpenter 1996; on T-PTP, see Carpenter *et al.* 1998). For example, no fewer than eight branches on the resolved tree of Dowton and Austin were 'supported' by bootstrapping with a replication frequency of less than 50% — meaning that these groups were unsupported, if not contradicted, at a frequency of more than 50%. Similarly, T-PTP can attribute significance to entirely unsupported groups, and even to both of two contradictory alternatives (examples are given in Carpenter *et al.* 1998). The lack of resolution among superfamilies is what these sequence data support.

Dowton *et al.* (1997) provided a more extensive sample of Apocrita for the same gene, including representatives of 11 superfamilies, plus two sawfly outgroups, for a total of 37 taxa, with 329 informative characters. The consensus of the three cladograms resulting after exclusion of three length polymorphic regions resolved most of the superfamilies for which multiple representatives were included, with the exception of Proctotrupoidea, Evanioidea and Ichneumonoidea. Aculeata was shown as a group, but relationships among the superfamilies were largely unresolved. Separate alignment of the length polymorphic regions, and analysis including these regions, resulted in a single cladogram, much more resolved than the consensus tree. Proctotrupoidea was the only superfamily not supported, and relationships among the included superfamilies were largely resolved. Some of the resolution depicted seems likely based on morphology, for example a sister-group relationship between Ichneumonoidea and Aculeata (Quicke *et al.* 1992). But other clades are unlikely, for example, Cynipoidea as sister-group to the remaining Apocrita. The appropriate test of these results is obviously combination and simultaneous analysis with morphological data.

Molecular sequence data in Hymenoptera have otherwise been applied to study of lower-level relationships. In particular, sequencing has been pursued in social insects: for example, in social bees (Cameron 1991, 1993; Sheppard & McPheron 1991), ants (Baur *et al.* 1993; Crozier *et al.* 1997), and social wasps (Choudhary *et al.* 1994). Large-scale sequence data sets for other Hymenoptera are only now beginning to appear (e.g. Belshaw *et al.* 1998).

Two other studies have employed moderately extensive molecular samples of Hymenoptera, but were carried out for the purpose of examining relationships between orders of insects. Carmean *et al.* (1992) sequenced the 18S rDNA molecule for 13 Hymenoptera in their study of Holometabola. The results of several analyses including different combinations of taxa did not even unequivocally support Hymenoptera as a group. Whiting *et al.* (1997) included 23 Hymenoptera, with sequences from both 18S and 28S rDNA, in their study of holometabolan relationships. Analysis of these molecular data, separately or in combination, supported Hymenoptera as monophyletic. Relationships within Hymenoptera were largely unresolved: superfamilies with more than one representative were all supported, but Apocrita and Aculeata were not.

Combined data: examples, principles and methods

Whiting *et al.* s (1997) study also included simultaneous analysis with morphological data. However, the morphological characters pertained to interordinal relationships. The first simultaneous analysis within Hymenoptera, that by Chavarría & Carpenter (1994), dealt with social bees. They combined the 16S mitochondrial rRNA sequences from Cameron (1993) and the rRNA data from Sheppard & McPheron (1991) with several different morphological data sets. The original morphological matrices were scored for higher-level taxa, while the sequences were obtained for particular species. This is the problem of 'terminal mismatch', discussed in detail by Nixon & Carpenter

(1996a). When terminals are circumscribed differently in data sets to be combined (terminal mismatch), 'data disjunction' results when the data sets are merged, with missing values required in the combined matrix for those terminals that do not match each data set. Introduction of missing values may have a number of negative effects, beyond the obvious increase in ambiguity manifested as an increase in the number of equally parsimonious cladograms. As pointed out by Nixon (1996: 367), the parsimony criterion itself is weakened, because the test of character congruence cannot be applied to the missing values. Missing values should therefore be minimized. Deletion of terminals can sometimes accomplish this most easily, in simple cases, but as this necessarily assumes that inclusion of those terminals would not change the outcome, splicing of terminals is typically necessary so that available relevant data are included, and another means of minimizing missing values must be sought. As argued by Nixon & Carpenter (1996a: 235):

> 'Under such circumstances it is justifiable to extrapolate . . . by fusing the two terminals into a single terminal. The degree to which extrapolation is used must be determined for each case based on a trade off between ambiguity and repeatability of the results. This can also be viewed as a trade off between unnecessary ambiguity on the one hand, and specious precision (unjustified extrapolation) on the other.'

That is, extrapolation should avoid solutions that would not be parsimonious if simultaneous analyses of all component taxa were undertaken, as with decisions on inclusion or exclusion of terminals.

The resolution of this problem adopted by Chavarría and Carpenter was to treat the terminals scored in the morphological matrix as 'summary terminals' of the respective higher taxa, which they then merged with the exemplar terminals in the molecular matrices. Of scoring to summarise variation within groups of taxa, Nixon & Carpenter (1996a: 235236) stated:

> 'Summarization is often possible with morphological data for groups that are well known, because variation may either be observed from specimens or gleaned from monographic and comparative studies. . . . Examples of extrapolated character states include such well known characters as endothermy in vertebrates, double fertilization in angiosperms, and holometaboly in insects, none of which have been observed in every species to which they are attributed or denied. Our recommendation on such situations is that if counter-evidence is not present, extrapolation is justified, if clearly identified as such.'

In the study by Chavarría and Carpenter, the morphological data proved much more decisive than the molecular data, with the results from combined analysis identical to those based on the morphological characters alone, whereas the results based on the sequences alone were in conflict.

The situation was different in the second simultaneous analysis in Hymenoptera, that by Carpenter (1997). In this case, the species of social wasps sequenced for 16S mitochondrial rRNA by Choudhary *et al*. (1994) were studied for the morphological characters. Terminal mismatch was avoided, and the data sets were 'spliced' by simple concatenation of the rows of the matrices. The results from the combined analysis also differed in that, whereas separate analysis of each data set did not fully resolve relationships, the simultaneous did.

Reconsideration of the analysis by Whiting *et al*. (1997) allows illustration of another principle of simultaneous analysis, which contrasts with common practice in analysis of molecular data, as exemplified in the paper by Dowton & Austin (1997). This latter study adduced sequences from the COI and 16S rRNA mitochondrial genes from nine species of sawflies and three species of Apocrita, along with three dipteran and one lepidopteran outgroups. Separate analysis of the COI sequences 'did not recover phylogeny generally accepted from fossil and morphological evidence; e.g. the Xyeloidea and Tenthredinoidea: Pergidae should be among the most basal of the Symphyta' (Dowton & Austin 1997: 400). Analysis of the COI sequences converted to amino acids did so — however, it also showed paraphyly of Hymenoptera in terms of the lepidopteran outgroup. Of this placement of Lepidoptera, Dowton & Austin (1997: 401) stated 'Although this disrupts the Panorpida, the result could reflect that this taxon is more closely related to the Hymenoptera than are the Diptera' (in fact, Diptera and Lepidoptera are more closely related than either is to Hymenoptera; see Whiting *et al*. 1997). Turning to 16S rRNA sequences, Dowton & Austin (1997: 402) concluded that these sequences showed AT-transversion bias, and 'unweighted parsimony did not resolve certain well-accepted relationships within the Symphyta . . . e.g. the Apocrita should be monophyletic and more recently diverged than any of the Symphyta, and the Tenthredinoidea should be monophyletic.' They then downweighted transversions between A and T by a factor of four, because 'Such a model of analysis, in the presence of compositional bias, has been recently suggested', and the result was that 'the downweighted analysis placed these lineages appropriately.' They then combined the two molecular data sets by culling the xyelid and xiphidryiid, because 'sequence data were not available for both gene regions' (Dowton & Austin 1997: 402), although two dipteran outgroups and the two siricids were each merged despite this data disjunction. The results of the combined

analysis were judged to be 'in accord with morphological and fossil evidence' (Dowton & Austin 1997: 403). Moreover, 'evidence from analysis of both COI and 16 rRNA genes separately and combined indicated that the models of analysis employed were appropriate; most well-accepted relationships were recovered' (Dowton & Austin 1997: 403).

However, none of Dowton and Austin's analyses 'recovered' Orussidae as the sister-group to Apocrita — surely a relationship 'in accord with morphological and fossil evidence' (see Gibson 1985; Vilhelmsen 1997). Nevertheless, Dowton and Austin proceeded to the inference that parasitism did not have a single origin within Hymenoptera. There are logical deficiencies to this conclusion. If harmony with morphological evidence is the optimality criterion, then it scarcely follows that the part of the results in conflict with morphology demonstrate that morphology is incorrect. The conclusion is rather that 'the models of analysis employed' are incorrect. Different models should therefore be sought — but, of course, this raises the question as to why, if results from morphology are the optimality criterion, should anyone bother with sequencing in the first place? An 'appropriate' model might after all be found in a particular case, but it would be easier, and certainly less expensive, to accept the morphological results, and not bother with other data.

Contrast this approach to that of Whiting *et al.* (1997). The most 'surprising' result of this study was that Strepsiptera and Diptera are sister-groups. This had been previously proposed by Whiting & Wheeler (1994), based upon data from the nuclear 18S rDNA gene for 23 holometabolan taxa and outgroups. The 'phylogeny generally accepted' placed Strepsiptera as sister-group to Coleoptera, if not a subgroup. Accordingly, Carmean & Crespi (1995) and Huelsenbeck (1997) declared that the sister-group relationship to Diptera was an artifact of 'long branch attraction' in the dreaded 'Felsenstein Zone.' This was based on their own more limited sample (13 taxa, which did not include all Holometabola). Huelsenbeck (1997: 69) in fact declared this to be the first empirical demonstration of long branch attraction, and argued maximum likelihood analysis be employed instead of parsimony, because it is 'less sensitive to the long branch problem.' What Huelsenbeck's study really revealed is that maximum likelihood is extraordinarily sensitive to alignment ambiguity (Siddall & Whiting 1997; see Siddall 1998): removal of a single site leads maximum likelihood to the same results as parsimony. The results are no more stable to taxon inclusion than to character inclusion (Siddall & Whiting 1997). This is a general argument against the use of maximum likelihood methods, but the 'phenomenon' of long branch attraction is no argument in favour of maximum likelihood. As pointed out by Whiting *et al.* (1997: 38) this amounts to the claim 'that the clades best supported by character data are the ones we should be most suspicious of.' In all of these data sets, the branch grouping Strepsiptera and Diptera is indeed long — but others are longer, for example, the branch supporting Diptera in Carmean & Crespi's (1995) data. Is Diptera then suspicious? The branch gets longer still when the more extensive molecular data from Whiting *et al.* (1997) are added — do we now become yet more suspicious? Maximum likelihood may break up this branch — but what if Strepsiptera and Diptera are really sister-groups? Then maximum likelihood has 'failed', as proponents of the method like to claim for parsimony. Significantly, the simulations upon which claims of failure of parsimony are based never included examination of the possibility of long branch repulsion by maximum likelihood, but that has now been demonstrated (Siddall 1998).

This fact highlights once again the general failure of model-dependent approaches to phylogenetic inference: As Carpenter (1992: 151) put it 'general models might not apply in specific cases, while specific models cannot be general.' One might be in either the Felsenstein Zone or the Farris Zone (Siddall 1998), but unless which is the case is known in advance, the wrong 'model of analysis' might be used. And how could it be known in advance? For example, it was 'generally accepted' that Strepsiptera and Coleoptera are closely related. Whiting *et al.* (1997) could have sought an ad hoc weighting scheme for their molecular data that obtained this result. But instead of elevating match to expectation to an optimality criterion, Whiting *et al.* adopted the position that expectations based on morphology are scientific hypotheses, and as such, must be amenable to testing. Hypotheses cannot be tested by treating them as criteria. Rather, the evidence supporting the morphological hypotheses should be tested, which is best done by simultaneous analysis of combined data (Nixon & Carpenter 1996a). In the case of Strepsiptera and Coleoptera, the sole putative synapomorphy is flight by the hind wings. Whiting *et al.* scored this character and analysed it along with other morphological features. Contrary to what was 'generally accepted', Strepsiptera did not group with Coleoptera, rather within the Panorpida, indeed as sister-group to Antliophora (i.e. Diptera, Mecoptera and Siphonaptera). Thus, analysis of morphology actually supports a closer relationship to Diptera than had been 'accepted.' Simultaneous analysis of the morphological and molecular data placed Strepsiptera as sister-group to Diptera, as resulted from analysis of the molecular data alone — but this only entailed one node change from the result of analysis of morphology alone.

Thus, simultaneous analysis of morphological and molecular data in Hymenoptera should be pursued as exempli-

fied in the studies by Chavarría & Carpenter (1994), Carpenter (1997) and Whiting *et al*. (1997). The data sources should be combined by 'splice and merge coding' (Nixon & Carpenter 1996a), and analysed together. We will present an example combining the morphological characters from Ronquist *et al*. (1999) with new molecular data. In this example, the approach taken to simultaneous analysis is the optimization alignment for sequence data proposed by Wheeler (1996). (This procedure was also implemented by Chavarría & Carpenter (1994), who obtained results similar to analysis of aligned sequences.) In optimization alignment, instead of an overall alignment *per se* among the sequences, hypothetical taxonomic intermediates (HTUs) are constructed among the sequences, seeking to minimize the cost function over the cladogram. This procedure thus combines alignment with cladogram construction. This obviates the necessity of creating gap characters to align the sequences in a matrix, and appears to generate more parsimonious cladograms (Wheeler 1996). For simultaneous analysis, different data sets, such as morphology, may be taken into account during the construction of HTUs, but these data are not 'aligned' between the HTU and descendants, as the sequences are.

Combination and a simultaneous analysis of morphological and molecular data in Hymenoptera

Molecular techniques

DNA isolation. The taxa sampled are listed in Table 1, and represent 10 of the 18 superfamilies recognized for example by Hanson & Gauld (1995). Total genomic DNA was isolated from fresh, EtOH preserved, and dried specimens by homogenization in an extraction buffer (10 mm Tris, 25 mm EDTA, 0.5% SDS, 100 mm NaCl, 0.1 mg/mL proteinase K). After 12 + hours of incubation with agitation at 55°C, the DNAs were cleaned with a standard series of phenol/chloroform extractions followed by ethanol precipitation and resuspension in water. If tissues were rare, the precipitation was replaced by purifying the supernatant in separation columns (Centricon 100) to increase the total DNA yield and quality.

Amplification and sequencing. Double-stranded template suitable for sequencing was prepared via polymerase chain reaction (PCR) amplification with conserved primers (Whiting *et al*. 1997). For most sequences, the entire region was amplified and sequenced with internal primers. Sequencing was carried out with the PRISM cycle sequencing kit (ABI) and run on the ABI 373 A automated sequencer. In all cases, complementary strands of all fragments were independently amplified and sequenced to assure accurate results. If complementary strands disagreed, the product was re-amplified and sequenced to resolve any discrepancies. The primer sequences for the nuclear small and large subunit sequences are those used in Whiting *et al*. (1997) and the mtCOI from Folmer *et al*. (1994), with an additional primer made for a gene region external to that for the primers of Folmer et al. (sequences labelled 'COIex' in Table 1). The sequence of that primer (5′ to 3′) is: CCAGGTAAAATTAAAATATAAACTTC; the relative position in the locus is 650.

Data analysis

For the separate analysis of the morphological characters, the pertinent family scores (see Table 1) from the matrix of Ronquist *et al*. (1999) were extracted, and used as summary scores for the sequenced exemplars. The resulting matrix was analysed phylogenetically with the program NONA (Goloboff 1997), by TBR branch swapping on 10 random addition sequences, followed by extended branch-swapping. For the separate analysis of molecular data, and the combined analysis, the data were analysed phylogenetically using the optimization procedure of Wheeler (1996) for the molecular data, and standard techniques for the additive and nonadditive morphological characters in the combined analysis. These analyses were implemented using the program POY (Gladstein & Wheeler 1997; *ftp.amnh.org/pub/molecular*). For all the molecular data, insertion/deletion event (indel) costs were set to twice that of base substitutions and transitions weighted equally with transversions. Leading and trailing gaps were ignored. For inclusion of the morphological characters, the pertinent family scores were combined with the respective sequenced exemplars with the morphological data weighted equal to indel events. This set of parameter values has been shown to maximize character congruence among data sets in several studies (Wheeler 1995; Wheeler & Hayashi 1998; Wheeler *et al*. submitted), and is thus appropriate for a preliminary investigation. The cladograms were rooted at the xyelid (note that in optimization alignment, position of the root can affect cost; Wheeler 1996). For the combined analysis TBR branch swapping was performed on 10 random addition sequences, with internal self-checking.

Results

Two cladograms resulted from separate analysis of the morphological characters, with length 306, consistency index 0.65 and retention index 0.85. The consensus of these trees, with Bremer support values, is given in Fig. 1. Six cladograms resulted from separate analysis of the molecular data, at cost 3124. The consensus of these trees, with Bremer support values, is given in Fig. 2. A single cladogram resulted from the simultaneous analysis at cost 3826 (Fig. 3). The Bremer support values, in terms of cost, are given on the figure.

The morphological results conform largely as expected

Table 1 Hymenoptera taxa in the sequence data set. The genes sequenced are the 18Sd and 28S subunits of rDNA, mtCOI and a region external to the latter, denoted COIex. An X indicates which genes were sequenced for each taxon. GenBank accession numbers are AF142502-AF142552, AF146652-AF146686.

Superfamily	Family	Subfamily: Tribe	Genus, species	18Sd	28S	COI	COIex
Xyeloidea	Xyelidae	Xyelinae	*Xyela julii*		X		
Tenthredinoidea	Argidae		*Arge nigripes*		X	X	
Megalodontoidea	Pamphiliidae		*Cephalcia arvensis*		X		
Evanioidea	Evaniidae		*Evaniella* sp.		X		
	Gasteruptiidae		*Gasteruption* sp.	X	X	X	X
Chalcidoidea	Chalcididae		*Brachymeria* sp.	X	X		X
	Leucospidae		*Leucospis* sp.		X		X
Trigonalyoodea	Trigonalyidae		*Labidogonalos* sp.		X	X	X
Ichneumonoidea	Ichneumonidae	Pimplinae; Ephialtini	*Clistopuga recurva*		X		
		Pimplinae; Pimplini	*Pimpla aequalis*		X	X	
		Tryphoninae; Phytodietini	*Netelia* sp.		X	X	
		Ichneumoninae; Ichneumonini	*Limonethe maurator*		X		
		Cryptinae; Cryptini	*Lymeon orbum*		X		
			Mesostenus thoracicus		X		
		Campopleginae	*Dusona egregia*		X	X	
			Campoplex sp.		X		
		Anomaloninae Gravenhorstiini	*Habronyx* sp.		X	X	
Chrysidoidea	Bethylidae	Epyrinae			X		
	Chrysididae	Chrysidinae	*Chrysis* sp.	X	X	X	X
Apoidea	Sphecidae	Sphecinae; Ammophilini	*Ammophila* sp.		X	X	
		Sphecinae; Sceliphrini	*Chlorion* sp.		X		
	Apidae	Nomadinae	*Nomada* sp.		X		
Vespoidea	Tiphiidae	Myzininae	*Myzinum* sp.	X	X	X	X
	Mutillidae	Sphaeropthalminae; Sphaerophthalmini	*Dasymutilla* sp.	X	X	X	X
	Pompilidae	Pepsinae	*Hemipepsis* sp.	X	X		
	Rhopalosomatidae		*Rhopalosoma* sp.	X	X		
	Scoliidae	Scoliinae; Campsomeridini	*Campsomeris* sp.	X	X		X
		Scoliinae; Scoliini	*Scolia sp.*	X	X	X	X
	Vespidae	Eumeninae	*Eumenes tripunctatus*	X	X	X	
		Vespinae	*Vespula maculifrons*	X	X	X	X
		Polistinae; Polistini	*Polistes tenebricosus*		X	X	X
		Polistinae; Ropalidiini	*Belonogaster juncea colonialis*		X		X
			Ropalidia romandi cabeti		X	X	
		Polistinae; Epiponini	*Apoica pallida*	X	X	X	X
			Polybia (Myrapetra) occidentalis nigratella		X	X	X
			Polybia (Trichinothorax) affinis		X	X	X

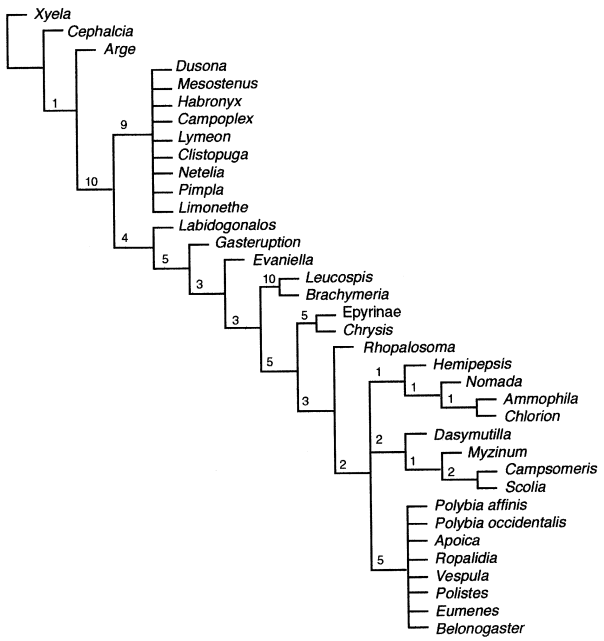

Fig. 1 Consensus tree for genera of Hymenoptera, based on analysis of the morphological data for the taxa listed in Table 1 from Ronquist *et al.* (1999). Numbers above branches represent Bremer support values, in terms of steps.

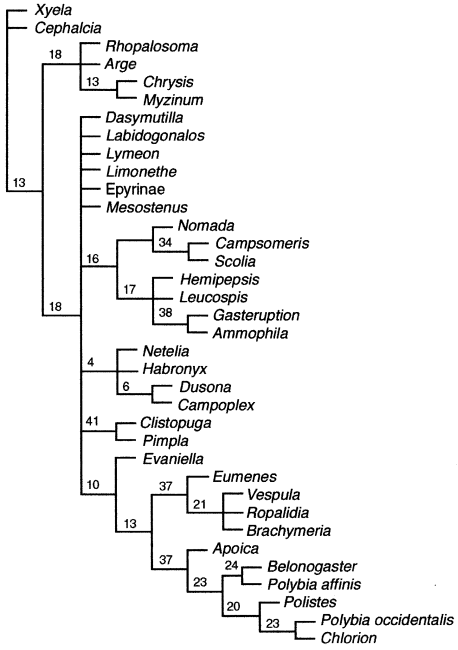

Fig. 2 Consensus tree for genera of Hymenoptera, based on analysis of the molecular data for the taxa listed in Table 1. Numbers above branches represent Bremer support values, in terms of cost.

to the analysis of the full matrix by Ronquist *et al.* (1999). However, Vespoidea is not a group, because Rhopalosomatidae is basal to its position in analysis of the full matrix, where it is part of Vespoidea. As mentioned above, thre reanalysis of the aculeate data of Rasnitsyn (1988) by Brothers & Carpenter (1993) differed from the analysis by Ronquist *et al.* because of extensive correction. Integration of the data of Brothers & Carpenter (1993) in an analysis of Aculeata is desirable, but is a task we shall take up later.

The molecular results support neither Apocrita nor Aculeata, none of the superfamilies represented by multiple exemplars, and only one family so represented (Scoliidae).

The cladogram for the simultaneous analysis shows Apocrita as a group, as well as Aculeata and the superfamily Chrysidoidea. Trigonalyoidea is basal within Apocrita. But Evanioidea, Chalcidoidea, Apoidea and Vespoidea are not groups, rather paraphyletic or polyphyletic. This is also true of Ichneumonidae, with *Clistopuga* clustering with the Chalcidoidea. The families Vespidae and Scoliidae are supported.

Overall, the results from the simultaneous analysis appear more similar to the analysis of the morphological data separately, but the Bremer supports for such clades as Apocrita and Aculeata are much larger.

Other aspects of the cladograms could be discussed, no doubt at length, but that does not seem necessary for

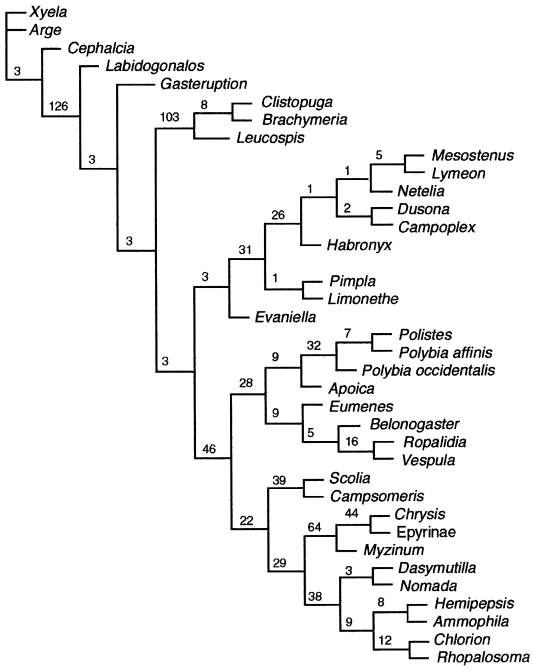

Fig. 3 Cladogram for genera of Hymenoptera, based on simultaneous analysis of the molecular data for the taxa listed in Table 1 combined with morphological characters from Ronquist *et al.* (in press) as described in the text. Numbers above branches represent Bremer support values, in terms of cost.

preliminary results. We will instead note that we did not perform a sensitivity analysis (Wheeler 1995). In sensitivity analysis, the parameters used in making the alignment are varied, and different alignments constructed, the process is repeated and the entire range of alignments analysed cladistically. The point is not to conclude that a given combination of parameters, such as gap to change ratio or transition-transversion cost, is necessarily best, as much as it is to determine how sensitive the results are to arbitrary values of these parameters. This allows one to avoid assumption-specific, unstable conclusions. The results depicted in Fig. 3 may indeed be dependent on the specific parameters used in the optimization alignment. We have not tested that possibility, because sensitivity analysis is computationally intensive, and rather than proceeding to further analysis of this particular data set, we believe it more fruitful to proceed with augmenting the sequence sample. This augmentation should be both taxonomic, to include more of the superfamilies and families of Hymenoptera, and also to include additional sequences, including those previously published. As perusal of Table 1 shows, only the large subunit rDNA molecule was sequenced for all of the exemplars; the other three molecules were each sequenced for only part of the exemplars, with just seven taxa sequenced for all four molecules. We are now pursuing the augmentation of that sample.

Acknowledgements

We thank Fredrik Ronquist for organizing the workshop 'Phylogeny of the Hymenoptera: The State of the Art', which provided the impetus for this paper, and also for providing a copy of the morphological matrix he and coauthors scored. We are also grateful to Paul Hanson, Lars Vilhemsen and Dave Wahl for providing some of the material for sequencing, and Hanson Liu for his effort in accomplishing the sequencing.

References

Basibuyuk, H. H. & Quicke, D. L. J. (1997). Hamuli in the Hymenoptera (Insecta) and their phylogenetic implications. *Journal of Natural History*, *31*, 1563–1585.

Baur, A., Buschinger, A. & Zimmermann, F. K. (1993). Molecular cloning and sequencing of 18S rDNA gene fragments from six different ant species. *Insectes Sociaux*, *40*, 325–335.

Belshaw, R., Herniou, E., Gimeno, C., Fitton, M. G. & Quicke, D. L. J. (1998). Molecular phylogeny of the Ichneumonoidea (Hymenoptera) based on D2: expansion region of 28S rDNA. *Systematic Entomology*, *23*, 109–123.

Brothers, D. J. (1975). Phylogeny and classification of the aculeate Hymenoptera, with special reference to Mutillidae. *University of Kansas Science Bulletin*, *50*, 483–648.

Brothers, D. J. & Carpenter, J. M. (1993). Phylogeny of Aculeata, Chrysidoidea and Vespoidea (Hymenoptera). *Journal of Hymenoptera Research*, *2*, 227–304.

Cameron, S. A. (1991). A new tribal phylogeny of the Apidae inferred from mitochondrial DNA sequences. In: Smith, D. R. (Ed.) *Diversity in the Genus Apis*. pp. 71–87, Westview Press, Boulder.

Cameron, S. A. (1993). Multiple origins of advanced eusociality in bees inferred from mitochondrial DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *90*, 8687–8691.

Carmean, D. & Crespi, B. J. (1995). Do long branches attract flies? *Nature*, *373*, 666.

Carmean, D., Kimsey, L. S. & Berbee, M. L. (1992). 18S rDNA sequences and the holometabolous insects. *Molecular Phylogenetics and Evolution*, *1*, 270–278.

Carpenter, J. M. (1990). On Brother's aculeate phylogeny. *Sphecos*, *19*, 9–10.

Carpenter, J. M. (1992). Random cladistics. *Cladistics*, *8*, 147–153.

Carpenter, J. M. (1996). Uninformative bootstrapping. *Cladistics*, *12*, 177–181.

Carpenter, J. M. (1997). Phylogenetic relationships among European *Polistes* and the evolution of social parasitism (Hymenoptera: Vespidae, Polistinae). *Mémoires Du Muséum National d'Histoire Naturelle*, *173*, 135–161.

Carpenter, J. M., Goloboff, P. A. & Farris, J. S. (1998). PTP is meaningless, T-PTP is contradictory: a reply to Trueman. *Cladistics*, *14*, 105–116.

Chavarría, G. & Carpenter, J. M. (1994). 'Total evidence' and the evolution of highly social bees. *Cladistics*, *10*, 229–258.

Choudhary, M., Strassmann, J. E., Queller, D. C., Turillazzi, S. & Cervo, R. (1994). Social parasites in polistine wasps are monophyletic: implications for sympatric speciation. *Proceedings of the Royal Society of London (Series B)*, *257*, 31–35.

Crozier, R. H., Jermiin, L. S. & Chiotis, M. (1997). Molecular evidence for a Jurassic origin of ants. *Naturwissenschaften*, *84*, 22–23.

Derr, J. N., Davis, S. K., Woolley, J. B. & Wharton, R. A. (1992a). Variation and the phylogenetic utility of the large ribosomal subunit of mitochondrial DNA from the insect order Hymenoptera. *Molecular Phylogenetics and Evolution*, *1*, 136–137.

Derr, J. N., Davis, S. K., Woolley, J. B. & Wharton, R. A. (1992b). Reassessment of the 16S rRNA nucleotide sequence from members of the parasitic Hymenoptera. *Molecular Phylogenetics and Evolution*, *1*, 338–341.

Dowton, M. & Austin, A. D. (1994). Molecular phylogeny of the insect order Hymenoptera: apocritan relationships. *Proceedings of the National Academy of Sciences of the United States of America*, *91*, 9911–9915.

Dowton, M. & Austin, A. D. (1997). Evidence for AT-transversion bias in wasp (Hymenoptera: Symphyta) mitochondrial genes and its implication for the origin of parasitism. *Journal of Molecular Evolution*, *44*, 398–405.

Dowton, M., Austin, A. D., Dillon, N. & Bartowsky, E. (1997). Molecular phylogeny of the apocritan wasps: the Proctotrupomorpha and Evaniomorpha. *Systematic Entomology*, *22*, 245–255.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome oxidase subunit 1 from diverse metazoan invertebrates. *Molecular Markers for Biology and Biotechnology*, *3*, 294–299.

Gauld, I. D. & Bolton, B. (1988). *The Hymenoptera.* British Museum (Natural History), London, and Oxford University Press, Oxford.

Gibson, G. A. P. (1985). Some prothoracic and mesothoracic structures important for phylogenetic analysis of Hymenoptera, with a review of terms used for the structures. *Canadian Entomologist*, *117*, 1395–1443.

Gladstein, D. & Wheeler, W. C. (1997). POY. *[Computer Software].* American Museum of Natural History, New York.

Goloboff, P. A. (1997). NONA, Version 1.6. *[Computer software and manual].* Fundación e Instituto Miguel Lillo, Tucumán, Argentina.

Goulet, H. & Huber, J. (Eds) (1993). *Hymenoptera of the World: an Identification Guide to Families.*Agriculture Canada, Ottawa.

Hanson, P. & Gauld, I. D. (Eds). (1995). *The Hymenoptera of Costa Rica.* The Natural History Museum, London, and Oxford University Press, Oxford.

Heraty, J. M., Woolley, J. B. & Darling, D. C. (1994). Phylogenetic implications of the mesofurca and mesopostnotum in Hymenoptera. *Journal of Hymenoptera Research*, *3*, 241–277.

Huelsenbeck, J. (1997). Is the Felsenstein Zone a fly trap? *Systematic Biology*, *46*, 69–74.

Johnson, N. F. (1988). Midcoxal articulations and the phylogeny of the order Hymenoptera. *Annals of the Entomological Society of America*, *81*, 870–881.

Källersjö, M., Farris, J. S., Kluge, A. G. & Bult, C. (1992). Skewness and permutation. *Cladistics*, *8*, 275–287.

Königsmann, E. (1976). Das phylogenetische System der Hymenoptera. Teil 1: Einführung, Grundplanmerkmale, Schwestergruppe und Fossilfunde. *Deutsche Entomologische Zeitschrift (N. F.)*, *23*, 253–279.

Königsmann, E. (1977). Das phylogenetische System der Hymenoptera. Teil 2: 'Symphyta'. *Deutsche Entomologische Zeitschrift (N. F.)*, *24*, 1–40.

Königsmann, E. (1978a). Das phylogenetische System der Hymenoptera. Teil 3: 'Terebrantes' (Unterordnung Apocrita). *Deutsche Entomologische Zeitschrift (N. F.)*, *25*, 1–55.

Königsmann, E. (1978b). Das phylogenetische System der Hymenoptera. Teil 4: aculeata (Unterordnung Apocrita). *Deutsche Entomologische Zeitschrift (N. F.)*, *25*, 365–435.

Naumann, I. D. (1991). Hymenoptera (Wasps, bees, ants, sawflies). In: *CSIRO, the Insects of Australia*, *2nd edn*. pp. 916–1000, Melbourne University Press, Melbourne.

Nixon, K. C. (1996). Paleobotany in cladistics and cladistics in paleobotany: enlightenment and uncertainty. *Review of Palaeobotany and Palynolology*, *90*, 361–373.

Nixon, K. C. & Carpenter, J. M. (1996a). On simultaneous analysis. *Cladistics*, *12*, 221–241.

Nixon, K. C. & Carpenter, J. M. (1996b). On consensus, collapsibility, and clade concordance. *Cladistics*, *12*, 305–321.

Oeser, R. (1961). Vergleichend-morphologische Untersuchungen über den Ovipositor der Hymenopteren. *Mitteilungen Aus Dem Zoologische Museum in Berlin*, *37*, 3–119.

Quicke, D. L. J., Fitton, M. G. & Ingram, S. (1992). Phylogenetic implications of the structure and distribution of ovipositor valvilli in the Hymenoptera (Insecta). *Journal of Natural History*, *26*, 587–608.

Quicke, D. L. J., Ingram, S. N., Baillie, H. S. & Gaitens, P. V. (1992). Sperm structure and ultrastructure in the Hymenoptera (Insecta). *Zoological Scripta*, *21*, 381–402.

Rasnitsyn, A. P. (1980). Origin and evolution of hymenopterous insects. *Trudy Paleontologicheskogo Instituta Akademiya Nauk SSSR*, *174*, 1–191 [in Russian].

Rasnitsyn, A. P. (1988). An outline of the evolution of the hymenopterous insects (Order Vespida). *Oriental Insects*, *22*, 115–145.

Ronquist, F., Rasnitsyn, A. P., Roy, A., Eriksson, K. & Lindgren, M. (1999). Phylogeny of the Hymenoptera: A cladistic reanalysis of Rasnitsyn's 1988 data. *Zoologica Scripta*, *28*, 13–50.

Sanderson, M. J., Purvis, A. & Henze, C. (1998). Phylogenetic supertrees: assembling the tree of life. *Trends in Ecology and Evolution*, *13*, 105–109.

Sheppard, W. S. & McPheron, B. A. (1991). Ribosomal DNA diversity in Apidae. In: Smith, D. R. (Ed.). *Diversity in the Genus Apis.* pp. 89–102, Westview Press, Boulder.

Siddall, M. E. (1998). Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. *Cladistics*, *14*, 209–220.

Siddall, M. E. & Whiting, M. F. (1997). Building a better fly trap. Paper presented at the Sixteenth Meeting of the Willi Hennig Society, October 23–26, 1997. George Washington University, Washington, D. C.

Vilhelmsen, L. (1997). The phylogeny of the lower Hymenoptera (Insecta), with a summary of the early evolutionary history of the order. *Journal of Zoological Systematics and Evolutionary Research*, *35*, 49–70.

Wheeler, W. C. (1995). Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Systematic Biology*, *44*, 321–331.

Wheeler, W. C. (1996). Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics*, *12*, 1–9.

Wheeler, W. C. & Hayashi, C. Y. (1998). The phylogeny of the extant chelicerate orders. *Cladistics*, *14*, 173–192.

Wheeler, W. C., Whiting, M., Wheeler, Q. D. & Carpenter, J. M. (submitted). The phylogeny of the extant hexapod orders. *Cladistics*,

Whitfield, J. B., Johnson, N. F. & Hamerski, M. R. (1989). Identity and phylogenetic significance of the metapostnotum in nonaculeate Hymenoptera. *Annals of the Entomological Society of America*, *82*, 663–673.

Whiting, M. F., Carpenter, J. M., Wheeler, Q. D. & Wheeler, W. C. (1997). The Strepsiptera problem: Phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. *Systematic Biology*, *46*, 1–68.

Whiting, M. F. & Wheeler, W. C. (1994). Insect homeotic transformation. *Nature*, *368*, 696.