© 1989 Elsevier Science Publishers B.V. (Biomedical Division)
The Hierarchy of Life
B. Fernholm, K. Bremer and H. Jörnvall, editors

CHAPTER 23

The systematics of insect ribosomal DNA

WARD CHRISTOPHER WHEELER*

Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, U.S.A.

Systematic background

This paper is concerned with the phylogenies of three insect groups: the Paleoptera, Eumetabola, and Holometabola. The evolutionary status of these groups has not been firmly established; in each case, not only monophyly but various paraphyletic arrangements have been proposed. The lack of consensus is due, no doubt, to the age of the taxa. Since the insect orders are ancient, much of the morphological evidence of kinship has been diluted through eons of unique evolution. In such circumstances, the homologies of slowly evolving molecules may help to resolve the monophyly versus paraphyly arguments. This paper uses the character evolution of ribosomal DNA to examine these genealogies fogged by time.

The three possible arrangements of the extant paleopteran orders each have their adherents (Fig. 1). Hennig [1] initially divided this group and placed the Odonata with the Neoptera, but later changed his view to one of monophyly [2,3]. His revised opinion is based on four putative synapomorphies: the aquatic larval stage, the fusion of the inner lobes of the maxillae, the presence of intercalary veins between longitudinal veins, and the short 'bristle-like' antennal flagellae. By contrast, Kristensen [4,5] supports the monophyly of Odonata + Neoptera, breaking up the Paleoptera. He finds the paleontological evidence unconvincing, and criticizes Hennig's characters as prone to convergence. To explain his own grouping, Kristensen cites – among other characters – the lack of imaginal molts, tracheization pattern in the wings and pterothoracic legs, muscle insertions, R and Rs veins with a common stem, and single female gonophore. Boudreaux offers yet another hypothesis [6]. He creates a new group, the Opisthopterata, to unite Ephemerida and Neoptera. His configuration is

*Present address: Department of Invertebrates, American Museum of Natural History, New York, NY 10024-5192, U.S.A.

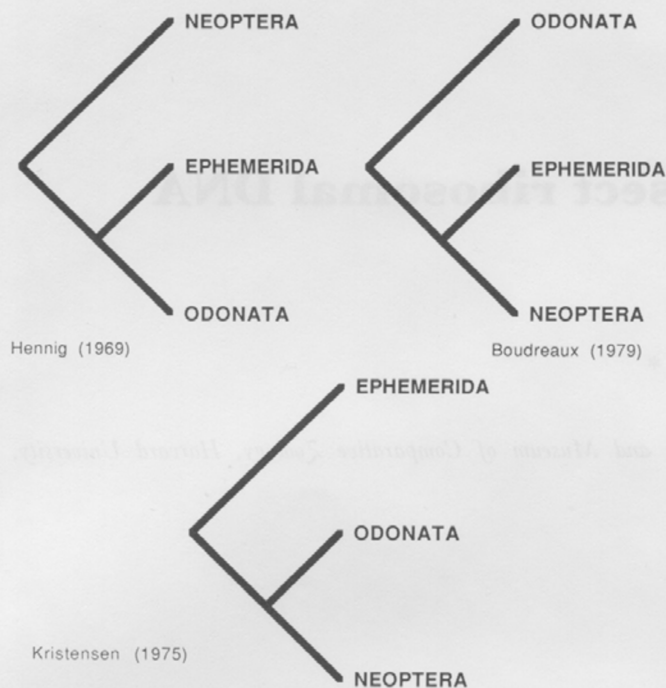


Fig. 1. Phylogenetic hypotheses proposed for the Paleoptera.

based on similarities in the folding of the nymphal wing pads, the means of sperm transfer, and on functional morphological criteria. Thus, as these competing claims demonstrate, the issue of paleopteran phylogeny is far from settled.

There is somewhat less controversy over the Eumetabola (Paraneoptera + Holometabola), or as Boudreaux calls them, the Phalloneoptera (Fig. 2). Hennig cites the Holometabola and Paraneoptera as two of the four primary Neopteran groups [2,3]. Kristensen and Boudreaux both support this grouping but differ with respect to the taxa included and the characters employed. Boudreaux unites the Phalloneoptera based on the formation of the male gonopod from 'phallic rudiments', hence the name. He places the Zoraptera with the orthopteroid orders [6]. According to Kristensen, however, the characters cited do not appear to be uniformly present [5]. Kristensen includes the Zoraptera in his Paraneoptera, discarding one of Hennig's eumetabolan characters – the absence of larval ocelli – since these structures are present in Zoraptera. By Kristensen's analysis, the 'jugal bar' of Hamilton [7] is the only possible eumetabolan synapomorphy.

Although the monophyly of the Holometabola is well established, the groupings within this taxon are less so (Fig. 3). With regard to the arrangements of the four major lineages (Hymenoptera, Coleoptera, Neuroptera, and Panorpida), Hennig is cautious. He states that affinities must exist, but there is no evidence of sufficient clarity to support them [2]. Boudreaux, ever bolder, suggests that the primary division is between the Coleoptera (+Strepsiptera) and the rest [6]. Telomerida share the namesake character of a male gonopod divided (unlike Gaul) into two sections, a basomere and a telomere; the absence of gastric ceca; and 'derepression of limb buds'.

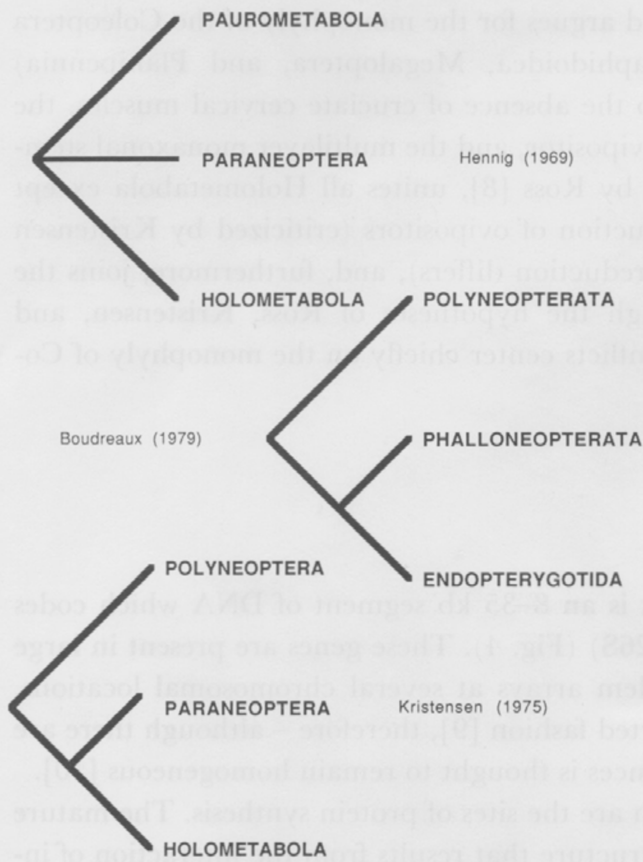


Fig. 2. Phylogenetic hypotheses proposed for the Eumetabola.

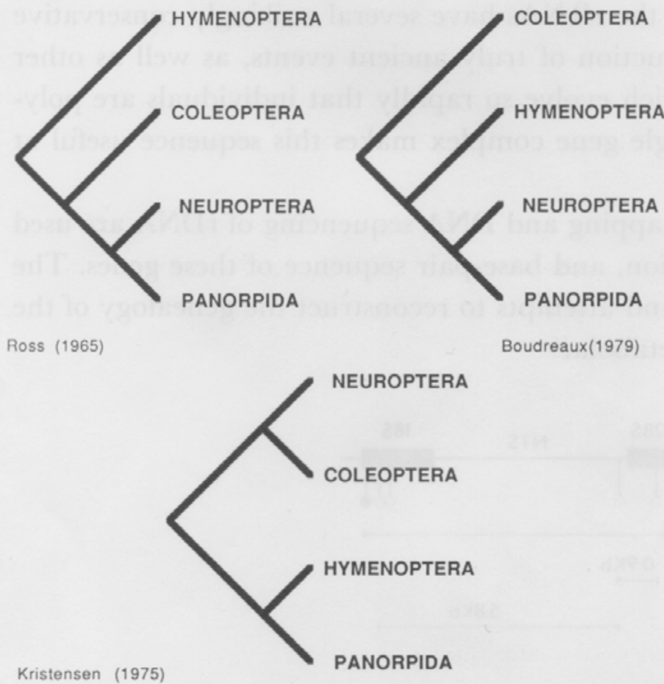


Fig. 3. Phylogenetic hypotheses proposed for the Holometabola.

Kristensen disputes these characters and argues for the monophyly of the Coleoptera (+ Strepsiptera) and Neuroptera (Raphidoidea, Megaloptera, and Planipennia) [4,5]. To define this group, he looks to the absence of cruciate cervical muscles, the reduction of the second valvule in the ovipositor, and the multilayer monaxonal stemmata. A third possibility, put forward by Ross [8], unites all Holometabola except Hymenoptera, due to their shared reduction of ovipositors (criticized by Kristensen [4] as convergent because the type of reduction differs), and, furthermore, joins the Neoptera and the Panorpida. Although the hypotheses of Ross, Kristensen, and Boudreaux are quite different, their conflicts center chiefly on the monophyly of Coleoptera + Neuroptera.

Ribosomal DNA

The ribosomal DNA (rDNA) complex is an 8–35 kb segment of DNA which codes for three ribosomal RNAs (18S, 5.8S, 26S) (Fig. 4). These genes are present in large numbers (> 100 per genome), in tandem arrays at several chromosomal locations. They are believed to evolve in a concerted fashion [9], therefore – although there are many copies – the pool of coding sequences is thought to remain homogeneous [10].

These rDNAs code for rRNAs, which are the sites of protein synthesis. The mature ribosomes have a complex secondary structure that results from the interaction of internal base-pairs. The base-paired regions seem to evolve in a different manner and at a different rate than the single-stranded areas, and these disparities have been shown to affect phylogenetic reconstruction [11].

As shown by bacterial systematics, the rRNAs have several strikingly conservative regions, which allow for the reconstruction of truly ancient events, as well as other areas – in the intergenic spacer – which evolve so rapidly that individuals are polymorphic. Such diversity within a single gene complex makes this sequence useful at a whole host of systematic levels [12].

In this paper, restriction enzyme mapping and DNA sequencing of rDNA are used to determine the structure, organization, and base-pair sequence of these genes. The analysis garners new character data and attempts to reconstruct the genealogy of the Paleoptera, Eumetabola, and Holometabola.

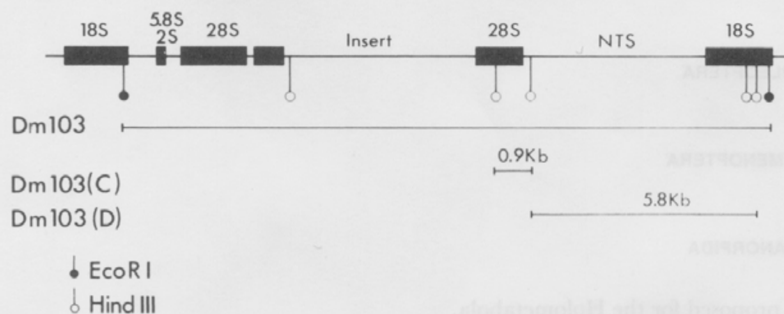


Fig. 4. Structure of the *Drosophila melanogaster* rDNA clone cDM103.

Methods

Specimens of *Podura aquatica*, *Argia fummipennis*, *Heptagenia diabasia*, *Oncopeltus fasciatus*, *Dolichovespula maculata*, *Chrysopa* sp., and *Pieris rapae* were collected in Middlesex County, Massachusetts, U.S.A. *Drosophila melanogaster* was donated by Dr. James Birchler, *Gromphaderina portentosa* by Dr. James Carpenter, *Formica excectoides* by Dr. Norman Carlin, and *Warramaba picta* by Dr. Rodney Honeycutt. Additional *Oncopeltus fasciatus* and *Tenebrio molitor* were obtained from Carolina Biological Supply.

Using the method of Bingham et al. [13] for *Drosophila*, nuclear DNA was extracted from aggregate samples of each species and purified on a CsCl gradient. To conduct restriction enzyme analysis, DNAs were digested with *Apa*I, *Bam*HI, *Bgl*II, *Cla*I, *Eco*RI, *Hin*dIII, *Pst*I, *Sac*I, and *Xba*I according to manufacturers' specifications. These reactions were performed in various combinations to determine the relative locations of recognition sites.

Digested DNAs were electrophoresed on 0.8% to 2.0% agarose slab gels, and then transferred to Gene Screen Plus charged nylon hybridization membrane according to the method of Southern [14], with the modifications advised by the supplier.

Using the method of Rigby et al. [15] and Maniatis et al. [16], the *Drosophila melanogaster* rDNA clones, pDM238 and cDM103 [17], were radioactively labelled by nick-translation (Fig. 4). They were then used to hybridize and visualize rDNA from the genomic digests.

To isolate rDNA fragments, the polymerase chain reaction procedure [18] was used according to manufacturers' specifications with the primers 5′-GAGGTAGTGACGAAAAA-3′ and 5′-CAGACTGTTATTGCTCA-3′. These primers defined a 1200 bp target sequence which was selectively amplified. Additional sequence upstream was isolated by screening genomic libraries constructed in M13, as in Benton et al. [19]. Using oligonucleotide primers, these fragments were cloned into M13 and sequenced by the method of Sanger et al. [20].

The sequences were aligned with a Needleman–Wunsch [21] type algorithm, implemented on a personal computer. Phylogenetic analysis was performed using Swofford's PAUP version 2.4.1. In the sequence analysis, gaps inserted by alignment were treated as missing data, and thus played no role in phylogenetic reconstruction. The *Drosophila* sequence was taken from [22], the *Tenebrio* from [23], and the *Artemia* from [24]. Secondary structure models used to separate the data into single- and double-stranded positions were taken from Hendriks et al. [23].

Results

Three kinds of data were collected in this study: restriction fragment length variation, gene size polymorphism, and direct sequence variation. Each type of data addressed different aspects of the insect phylogenies in question.

Restriction enzyme analysis

Two types of polymorphism were observed in this analysis – restriction site changes and insertion/deletion events involving large segments of DNA.

rDNA maps. The nine enzyme digestions yielded 49 sites in the taxa (Fig. 5). Of these, nine were invariant and 11 were autapomorphic. Of the 29 remaining sites the 22 sites which fell in coding regions were used to test the phylogenetic hypotheses of Ross, Hennig, Kristensen, and Boudreaux. The other sites occurred in spacer regions which undergo expansion and contraction in their evolution in addition to simple nucleotide transformation [12], hence it was not possible to homologize these sites. Phylogenetic analysis of the informative restriction sites yielded 17 equally parsimonious trees at the minimum length of 51 steps, with a consistency index of 0.43 (via branch-and-bound algorithm). The strict (Nelson) consensus tree that was constructed from these cladograms is shown in Fig. 6. Even though the consistency is quite low, the resolved groups are supported by all the most parsimonious cladograms.

Two groups are supported by this consensus tree: Odonata + Neoptera, and Coleoptera + Neuroptera. The results affirm Kristensen's groupings [4,5]. Phylogenetic hypotheses were constructed de novo from the restriction site maps; in addition, the models of paleopteran and holometabolan genealogy were tested to determine their

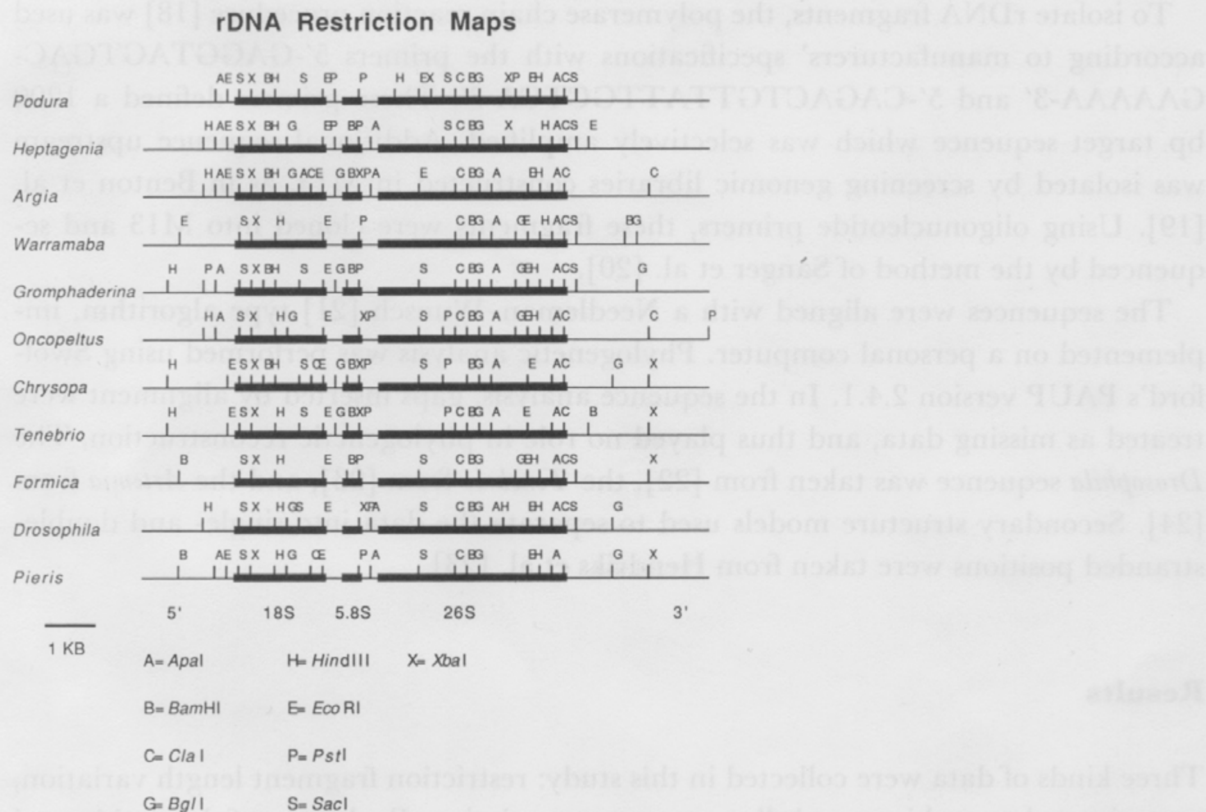


Fig. 5. Restriction enzyme maps.

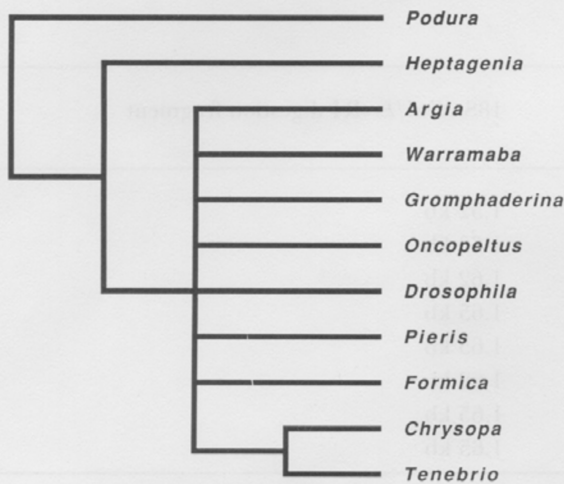


Fig. 6. Strict consensus tree of the seven most parsimonious rDNA restriction map data cladograms.

consistency with these data. Using PAUP, the restriction fragment characters were fitted to the hypotheses of Hennig [2,3], Kristensen [4,5], Boudreaux [6], and Ross [8]. Overall, Kristensen's hypothesis has the greatest explanatory power. Nonetheless, with 60 steps, it is less parsimonious than the optimal but unresolved version of the same topology. Interestingly, Hennig's proposal of monophyly for the Paleoptera, at 61 steps, is almost as well supported as Kristensen's. However, modifying the shortest tree to accommodate the group Odonata + Ephemerida would require an additional two restriction site changes. Neither Ross' proposal (62 steps) nor Boudreaux's (68 steps) are as parsimonious as Kristensen's. The clear support for Kristensen's genealogy is due to the presence of two to six site changes on the branch leading to Coleoptera + Neoptera, and six to seven changes on the branch joining Odonata + Neoptera. Variation in the number of synapomorphies stems from alternate equally parsimonious character state reconstructions.

Size variation

Eukaryotes possess an *Xba*I site at the 5′ end of the 18S gene and an *Eco*RI site near its 3′ terminus [24]. These sites are separated by approximately 1500 bases. To define the fragment, genomic DNAs of various hexapods were digested with these two enzymes. As depicted in Table 1 and Fig. 7, the collembolan, *Podura aquatica*, has a slightly smaller 18S fragment. Thus the addition of approximately 100 bases to the insect 18S gene may define a group perhaps as restrictive as Pterygota or as all-encompassing as Insecta sensu stricto.

A second insertion event appears to support the monophyly of the Eumetabola. As shown in Table 2 and Fig. 7, when digested with restriction enzymes that cut the repeat unit only once, *Oncopeltus*, and *Tenebrio*, *Chrysopa*, *Drosophila*, and *Pieris* among the

TABLE 1

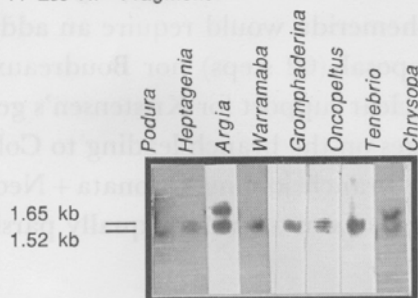
Size variation in 18S gene.

	18S *Xba*I/*Eco*RI digestion fragment
Podura	1.52 kb
Heptagenia	1.63 kb
Argia	1.62 kb
Warramaba	1.65 kb
Gromphaderina	1.65 kb
Oncopeltus	1.62 kb
Tenebrio	1.65 kb
Chrysopa	1.65 kb

holometabolan taxa share a large – 5 kb – insertion in the central portion of the 26S gene. *Formica* and *Dolichovespula* do not appear to have this insertion. Moreover, when present, this insertion appears to be polymorphic: two bands corresponding to com-

Gene Region Size Variation

18S *Xba* I / *Eco* RI Fragment



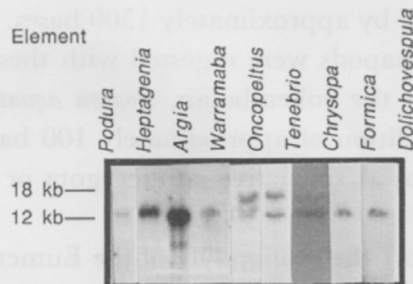


Fig. 7. Size variation in the 18S and 26S sequences.

TABLE 2

Repeat length polymorphism.

	Repeat length
Podura	13 kb
Heptagenia	14 kb
Argia	13.5 kb
Warramaba	14 kb
Oncopeltus	12/17 kb
Tenebrio	13/18 kb
Chrysopa	13.5/18 kb
Drosophila	12/17 kb
Formica	12 kb
Dolichovespula	13 kb

plete repeats are observed. The size and position of the insertion correspond to those of the intervening sequence in the rDNA repeats on the X-chromosome of *Drosophila* [17]. They may be involved in a 'bobbed' type of expression.

Sequence analysis

Of the 1083 aligned sequence positions 398 are variable and 276 informative. Of these, 147 contribute to the creation of secondary structures through internal double helix formation, while the remaining 129 lie in the single-stranded areas of the molecule. (The aligned sequences are available on request.)

These three data sets – all variable bases, double-stranded positions, and single-stranded positions – were analyzed separately (Table 3). The genealogies supported by the different data sets are at large variance with one another. When all 276 informative sites are used, there are two shortest cladograms: one breaks up the Holometabola and the other supports the Eumetabola (Fig. 8A). The base-paired positions also refute the monophyly of Holometabola, by uniting the mayfly, *Heptagenia*, with the grasshopper, *Warramaba* (Fig. 8B). Single-stranded base positions produce a single most parsimonious hypothesis at 247 steps (Fig. 8C). This is the Eumetabola hypothesis.

In order to provide a more statistical framework for cladistic analyses of sequence data, I have developed a method (Wheeler, in preparation) which can distinguish among equally parsimonious cladograms based on sequence data, and place a confidence interval about the derived tree length statistics. Briefly, the procedure involves the creation of a matrix of transformation frequencies based on the relative frequency of the concurrence of nucleotides throughout the molecule (i.e., whether As and Gs occur together at the same position more frequently than As and Ts). This forms the foundation of a weighting schedule. The relative transformation frequencies are then

TABLE 3

Cladogram lengths for 18S rDNA data.

Data set	Hypothesis	Number of characters	Length (no autapomorphies)
All bases	Eumetabola		483
	Lower Neoptera	276	488
	Holometabola paraphyly		483
Stem bases	Eumetabola		236
	Lower Neoptera	147	238
	Holometabola paraphyly		229
Loop bases	Eumetabola		247
	Lower Neoptera	129	250
	Holometabola paraphyly		254

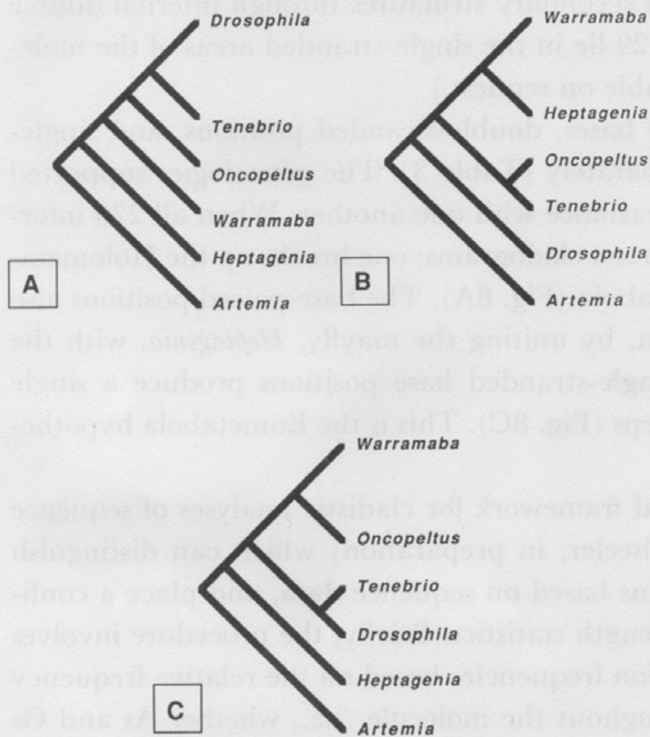


Fig. 8. Sequence-based phylogenies of the Eumetabola.

translated into character weights by taking the absolute value of their natural logarithms.

Thus the length of the tree is the sum of the costs of all the steps required to explain the data. In this way, two trees which require the same number of events with different likelihoods of occurrence (such as many transversions versus many transitions) are distinguished from one another. Since the weights are determined from a matrix, which is itself calculated from a finite number of variable bases in the sequence, there is an associated sampling error. Through simulation and analysis of these values, called 'lnprobs' for 'natural logarithmic probabilities', an estimate of the variance in tree length is easily calculable. Through further computer simulation, these values are found to be normally distributed; thus, with knowledge of means and variances, the statistical significance of the difference in the lengths of two cladograms can be estimated. The lnprob procedure was applied to the two hypotheses of eumetabolan relationships (Fig. 8A,C) using the loop data alone.

As is seen in Table 4, the eumetabolan phylogeny is favored, if marginally, over the monophyly of the Paraneoptera + Polyneoptera ($P = 0.20$).

Conclusions

The molecular information presented here supports monophyletic groups at several levels: Ectognatha- Odonata + Neoptera, Eumetabola, and Coleoptera + Neuroptera (Fig. 9). But this hierarchy is not limited to the taxa; for the data themselves, the types of synapomorphy, are also inherently hierarchical.

Even at the broadest level, large insertions and deletions are conspicuously informative. For example, the 100 bp insertion that defines the ectognath insects is a large-scale event, perhaps not observable at lower levels. If the only information available were DNA or RNA sequence data, such a large change might go unnoticed.

The same applies to the very large, 5 kb, insertion in the 26S gene. This event is interesting in and of itself, because of the systematic ramifications; but it is all the more interesting when we consider the nature of the synapomorphy, which is – and I believe must be – polymorphic. The 5 kb insertion maps to the site of the interruption in the 26S of *Drosophila* that causes the 'Bobbed' phenotype [12]. In *Drosophila*,

TABLE 4

Significance value for eumetabolan phylogenies (loop positions).

Phylogeny	Length	lnprob value	SD
Eumetabola	247	232.5	3.6
Lower Neoptera	250	235.5	3.6

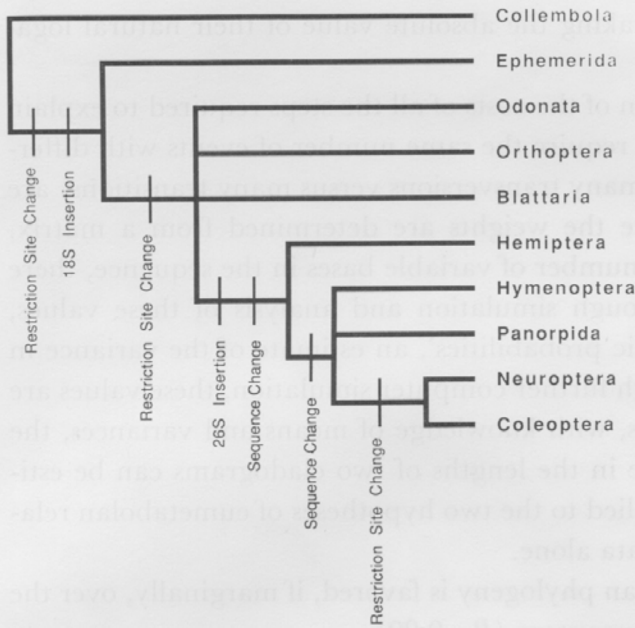


Fig. 9. Insect phylogeny as deduced from rDNA.

there are two rDNA repeat types. One has an uninterrupted 26S (Y-chromosome type), and the second has a large insertion in the 26S (X-chromosome type) which disables that gene. When the copy number of uninterrupted, fully functional rDNA repeats falls below about 50, the 'bobbed' phenotype can be lethal. Since the organism cannot survive without functional rDNAs, this polymorphism will not become fixed.

As mentioned above, the 5 kb insertion defines the Eumetabola with one exception: the Hymenoptera. The Hymenoptera examined here possess only one detectable repeat type, which is about the size of the short, uninterrupted gene complex in other species. Since the Hymenoptera are haplodiploid, the effective rDNA copy number in males may be only half that of females. This disparity may increase the selection pressure for transcriptional efficiency, hence reducing the viability of those with any 'bobbed' genotype. Such an interpretation may appear to offer little more than a nice story to justify the homoplasy in this character; the observation is testable, however, through the analysis of other haplodiploid eumetabolans such as members of the Micromalthidae and thrips.

At a finer grain, restriction enzyme analysis is better equipped than large structural variants to evaluate the level of sequence change among organisms. Since these enzymes recognize specific, known strings of bases, the polymorphisms in restriction sites are usually assumed to result from single base substitutions in one of the handful of bases that the enzyme recognizes. If the changes can be compared to a known sequence, the insights of sequence evolution from mapping can rival those from sequencing. Moreover, while the information acquired by mapping is not as direct or

exact as that gained by sequencing, restriction enzyme analysis does offer the opportunity to sample larger stretches of the genome than could readily be sequenced. In this paper, polymorphisms in restriction sites are distributed over a 10–15 kb segment. Falling outside the area that could be sequenced in a reasonable period of time are many changes that support the union of the Odonata with the Neoptera and the sister group status of the Coleoptera and Neuroptera. Because restriction site changes are the products of descent with modification, hierarchically distributed, they are potentially extremely informative – even though we may not know the details of their evolution.

But to truly understand the mechanisms of molecular evolution, sequencing is invaluable. As shown by Wheeler and Honeycutt [11], ribosomal RNAs (at least short ones) are prone to a type of molecular convergence that can impede systematic inference. The secondary structure of the mature rRNAs places constraints on the evolution of the primary structure, the sequence of bases. Observations such as these can only be made through direct sequence analysis. Likewise, although other types of molecular polymorphism are informative, homoplasy and its roots can only be examined with the sequences themselves. Given the ability to examine the causes of homoplasy, we can place greater confidence in conclusions drawn from sequence-based phylogenies.

The sequence data presented here also show the confounding effects of secondary structure, but these can be mitigated by removing the double-stranded base positions from the analysis. An additional benefit of sequence data emerges in the opportunity to perform statistical analysis. Since the individual base positions only come in four flavors, bases may acquire the behavior of a large number of independent, semi-independent, or tightly linked variables. With different notions of the mechanisms of base transformation, phylogenetic analysis can take advantage of the serial homology of base positions. Such a method (Wheeler, in preparation) is used here to distinguish among the sequence-based phylogenies of the Eumetabola.

One of the drawbacks to the sequencing approach, however, is that it can only examine relatively short segments of DNA. Without doubt, local structural and selective pressures affect the evolution of specific sequences. In the analysis of single genes or regions of genes, the dangers of being misled by unforeseen, systematically confounding phenomena are certainly worthy of consideration.

Clearly in phylogenetic analysis there are informative data at many levels, each with its own strengths and weaknesses. The insect groupings examined here are supported by very different qualities of synapomorphy. The Insecta sensu stricto (or perhaps Pterygota) are unified by a medium-sized sequence insertion, the Odonata + Neoptera by restriction site changes, the Eumetabola by both sequence data and a large structural change, and the Coleoptera + Neuroptera again by restriction site synapomorphies. Without the simultaneous examination of molecular information at multiple levels, many of the distinctions among these genealogies would be missed. Different levels of synapomorphy require different types of synapomorphy.

Acknowledgements

I would like to thank Elise Broach for her comments and suggestions. This research was supported by a National Science Foundation Graduate Fellowship, National Science Foundation Grants BSR-8700966 and BSR-85-08479, and the Richmond Fund of Harvard University.

References

1 Hennig W. Kritische Bemerkungen zum phylogenetischen System der Insekten. Beitr Ent 1953; 3 (Sonderheft): 1–85.
2 Hennig W. Die Stammesgeschichte der Insekten. Frankfurt am Main, 1969.
3 Hennig W. Insect Phylogeny. New York: John Wiley and Sons, 1981. (Translation of Hennig, 1969)
4 Kristensen NP. The phylogeny of hexapod 'orders'. A critical review of recent accounts. Z Zool Evol Forsch 1975; 13: 1–44.
5 Kristensen NP. Phylogeny of insect orders. Annu Rev Entomol 1981; 26: 135–157.
6 Boudreaux HB. Arthropod Phylogeny with Special Reference to Insects. New York: John Wiley and Sons, 1979.
7 Hamilton KGA. The insect wing part IV. Venational trends and the phylogeny of the winged orders. J Kansas Entomol Soc 1972; 45: 295–308.
8 Ross HH. A Textbook of Entomology. New York: John Wiley and Sons, 1965.
9 Arnheim N. Concerted evolution of multigene families. In: Nei M, Koehn RK (Eds.) Evolution of Genes and Proteins. Sunderland: Sinauer, 1983; 38–61.
10 Dover GA. Molecular drive: a cohesive mode of species evolution. Nature 1982; 229: 111.
11 Wheeler WC, Honeycutt RL. Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. Mol Biol Evol 1988; 5: 90–96.
12 Appels R, Honeycutt RL. rDNA: evolution over a billion years. In: Dutta SK (Ed.) DNA Systematics. Boca Raton: CRC Press, 1986; 81–135.
13 Bingham P, Levis R, Rubin GM. Cloning of DNA sequences from the *white* locus of *D. melanogaster* by a novel and general method. Cell 1981; 25: 693–704.
14 Southern EM. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 1975; 98: 503–517.
15 Rigby PWJ, Diechmann M, Rhodes C, Berg P. Labelling deoxyribonucleic acid to high specific activity in vitro by nick-translation with DNA polymerase I. J Mol Biol 1977; 113: 237–251.
16 Maniatis T, Fritsch EK, Sambrook J. Molecular Cloning: Laboratory Manual. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 1982.
17 Glover DM, Hogness DS. A novel arrangement of the 18S and 28S sequences in a repeating unit of *Drosophila melanogaster* rDNA. Cell 1977; 10: 167–176.
18 Chien A, Edgar DB, Trela JM. Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus*. J Bacteriol 1976; 127: 1550–1557.
19 Benton WD, Davis RW. Screening lambda-gt recombinant clones by hybridization to single plaques in situ. Science 1977; 196: 180.
20 Sanger F, Nicklen S, Coulson AR. Sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977; 74: 5463–5467.
21 Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970; 48: 443–453.

22 Tautz D, Hancock JM, Webb DA, Tautz C, Dover GA. Complete sequence of the rRNA genes in *Drosophila melanogaster*. Mol Biol Evol 1988; 5: 366–376.

23 Hendriks L, De Baere R, Van Broeckhoven C, De Wachter R. Primary and secondary structure of the 18S ribosomal RNA of the insect species *Tenebrio molitor*. FEBS Lett 1988; 232: 115–120.

24 Huysmans E, De Wachter R. Compilation of small ribosomal subunit RNA sequences. Nucl Acids Res 1987; 14 (Suppl): r73–r118.