

Elision: A Method for Accommodating Multiple Molecular Sequence Alignments with Alignment-Ambiguous Sites

WARD C. WHEELER,* JOHN GATESY,† AND ROB DE SALLE‡

Departments of *Invertebrates, †Herpetology, and ‡Entomology, American Museum of Natural History,
Central Park West at 79th Street, New York, New York 10024-5192

Received August 12, 1993; revised March 25, 1994

The process of multiple sequence alignment provides homology statements for the phylogenetic analysis of molecular data. Unfortunately, multiple alignments are frequently nonunique. Two sources of these multiple alignments are analysis based on different sets of alignment parameter values (gap:change cost ratios) and nonunique equally costly alignments based on a single set of alignment parameters. By "eliding" these individual alignments into a single grand alignment, phylogeny that is weighted toward those positions that align more consistently can be reconstructed. Positions that show greater variation among alignments will be relatively downweighted. The technique results in a weighting procedure that is *a posteriori* and based on the evidence established from the original sequence alignments. © 1995 Academic Press, Inc.

INTRODUCTION

The problems of multiple sequence alignment press increasingly on molecular systematics. The difficulties presented by nonunique alignments (Fitch and Smith, 1983; Waterman *et al.*, 1992; Wheeler, 1994) have only recently disclosed themselves to investigators examining the dependency of putative homology on analysis parameters such as gap and change cost ratios (the relative cost of insertion-deletion events and nucleotide change). Furthermore, even in the absence of multiple alignment cost scenarios, there may be several possible alignment solutions to any single combination of parameter values (since we choose multiple alignments on the basis of parsimony, these equally costly alignments in fact yield equally parsimonious phylogenetic scenarios, hence called equally parsimonious alignments).

The process of incorporating these multiple putative homology statements (no matter how they have been generated) into phylogenetic analysis is not well delineated. Consensually supported methods for accommodating these situations do not yet exist. Here, we propose agglomerating several multiple alignments into

a single grand alignment, which can then be analyzed phylogenetically. By the accumulation of more than one alignment, positions which vary among alignments are effectively downweighted, while consistent positions are weighted more heavily. This discussion does not concern itself with how the initial alignments are created; we are not proposing any new alignment algorithm. The procedure proposed here deals solely with the problem of multiple solutions to the alignment problem presented by a set of sequences. Furthermore, we are not advocating any particular means of establishing "appropriate" gap costs or other alignment parameters. We simply seek a means to accommodate several alignments of the same sequence data in a phylogenetic framework.

Although many parameters must be specified to create an alignment (transition-transversion ratio and gap length function, among others), the most common cause of multiple multiple alignments (a set of unique alignment solutions derived from one or several sets of alignment parameters) comes from variation in gap: change cost ratios. For example, when sequences are aligned, some gap penalty must be assessed. This penalty is set in relation to the relative cost of mismatches (or benefit of matches) between bases; this is the gap cost ratio. Without this penalty, a trivial "alignment" is created with each aligned sequence consisting of nucleotide bases and apposing huge strings of gaps (Fig. 1). No mismatches would occur, for each base of one alignment would correspond to gaps in all of the others. The appropriate value for the cost is unknown, however. There is no empirical way of measuring "gap cost." The great variety of possible parameter sets has been noted previously (Fitch and Smith, 1983; Waterman *et al.*, 1992; Wheeler, 1994), but the problem of the essentially arbitrary choice of values remains. Since these alignments are the basis for phylogeny in molecular systematic studies, the extraction of this information from the universe of possible parameter values is daunting.

Gatesy *et al.* (1994) have discussed (and decried) the behavior of a method which seeks to objectively remove

Taxon 1 -----ACGCTGTGTG---TAGCATC--ACTCATGCAGT----
 Taxon 2 AAAGTCAGTCTGCAGTAGCT-----TGC-----GT-----GTCA

FIG. 1. A "trivial" alignment derived from gaps with zero cost.

alignment-ambiguous (those which vary among alignments) nucleotide positions (Fig. 2). Their method—"culling" or deleting all alignment-ambiguous sites—tended to result in robust but grossly unresolved hypotheses of relationship among the taxa whose sequences were compared. This is the result of a ruthless, but nonarbitrary removal of data. Any disagreement among alignments at a nucleotide position results in the excision of that position (Fig. 3). The "elision" process proposed here is less extreme. This method should yield a more precise (resolved) phylogenetic result since it would in essence apply weights in a continuous fashion (rather than zero or one) to nucleotide positions.

THE METHOD

The elision procedure is quite simple. Individual multiple alignments are first generated by some dynamic programming procedure, such as an n -dimensional or sequentially pairwise Needleman-Wunsch algorithm (Needleman and Wunsch, 1970; Sankoff and Cedergren, 1983; Feng and Doolittle, 1987, 1990; Hein, 1989, 1990; Higgins and Sharp, 1988, 1989). The individual alignments may be based on different base change and insertion-deletion event cost ratios or may be nonunique solutions to the same set of analysis parameter values. These multiple-alignments are then strung together into a single combined alignment, keeping track of the gap:change cost ratios used to construct each individual alignment. Parsimony-based phylogenetic analysis can then be performed on this "grand alignment" with the cost parameters used in the alignments as the cost parameters for the phylogenetic analysis (i.e., same transversion-transition cost ratio for each section of the grand alignment).

AN EXAMPLE

In their examination of the removal of discordant alignment derived sequence homologies (removing the

Alignment 1
 Taxon 1 AAAG-GTTAA
 Taxon 2 AAAGA-TTTG
 Taxon 3 AAAC-GTTAG

Alignment 2
 Taxon 1 AAAG-GTTAA
 Taxon 2 AAAG-ATTTG
 Taxon 3 AAA-CGTTAG
 ...

FIG. 2. Alignment-ambiguous nucleotide sites. The sites which vary between alignments one and two are marked by asterisks.

a Taxon 1 AAATTAA
 Taxon 2 AAATTTG
 Taxon 3 AAATTAG

b Taxon 1 AAAG-GTAAAAAG-GTTAA
 Taxon 2 AAAGA-TTTGAAAG-ATTTG
 Taxon 3 AAAC-GTTAGAAA-CGTTAG

FIG. 3. The method of "culling" and "eliding" alignment-ambiguous sites. The new alignments created from the original alignments of Fig. 2 by (a) culling and (b) eliding the alignment-ambiguous bases.

alignment-ambiguous positions), Gatesy *et al.* (1994) analyzed sequences from the mitochondrial 12S rDNA of crocodiles and 16S rDNA of insects. We use the same sequences in the present study to demonstrate the elision process. The data consist of 10 243- to 256-bp sequences from various crocodylians and 9 insect sequences of length 232–243 bp (Table 1).

In both cases, the sequences were aligned with 10 different gap:change cost ratios. These ratios varied in a logarithmic fashion (base 2) from a low with gaps costing one-half base changes to a high where gaps cost 256 times base changes ($\log_2 \text{gap/change cost} = -1, 0, 1, 2, 3, 4, 5, 6, 7, 8$). The lowest value is the logical lower bound for gap cost (Wheeler, 1993). While there is no logical upper bound for this ratio, the costs of 128 and 256 to 1 yielded very similar (insect sequences) or identical (crocodylian) results. The alignments were performed using the program MALIGN (Wheeler and Gladstein, 1992, 1994). Multiple equally parsimonious alignments were sought and retained using the MALIGN command "build," branch swapping was performed both during and after alignment construction using the MALIGN commands "alignadditionswap" and "alignswap," gaps were treated as independent events (the MALIGN default), and phylo-

TABLE 1
 List of Sequences Analyzed

Data set	Taxon	Source
Insect	<i>Heptagenia</i> sp.	Gatesy <i>et al.</i> (1994)
	<i>Dorocordulia lepida</i>	Gatesy <i>et al.</i> (1994)
	<i>Blaberus cranifer</i>	Gatesy <i>et al.</i> (1994)
	<i>Schistocerca americana</i>	Gatesy <i>et al.</i> (1994)
	<i>Cerastipsoocus venosus</i>	Gatesy <i>et al.</i> (1994)
	<i>Drosophila yakuba</i>	Clary and Wolstenholme (1985)
	<i>Aedes subpictus</i>	Hsu Chen <i>et al.</i> (1984)
	<i>Cicindela dorsalis</i>	Vogler <i>et al.</i> (1993)
	<i>Apis mellifera</i>	Vlasek <i>et al.</i> (1987)
	Crocodylian	<i>Gavialis gangeticus</i>
<i>Tomistoma schlegelii</i>		Gatesy and Amato (1992)
<i>Crocodylus rhombifer</i>		Gatesy and Amato (1992)
<i>Caiman latirostris</i>		Gatesy and Amato (1992)
<i>Caiman crocodilus</i>		Gatesy <i>et al.</i> (1994)
<i>Melanosuchus niger</i>		Gatesy <i>et al.</i> (1994)
<i>Paleosuchus palpebrosus</i>		Gatesy <i>et al.</i> (1994)
<i>Paleosuchus trigonatus</i>		Gatesy <i>et al.</i> (1994)
<i>Alligator sinensis</i>		Gatesy <i>et al.</i> (1994)
<i>Alligator mississippiensis</i>	Gatesy <i>et al.</i> (1994)	

genetic scoring was enabled with branch swapping during and after cladogram construction. With these options, multiple equally costly (= parsimonious) alignments can be found.

After these alignments were performed, cladograms were constructed for the alignments (10 for each of the 2 data sets). Hennig86 (Farris, 1988) and PAUP (Swofford, 1990) were used to find the most parsimonious phylogenetic solution(s) of the data using exact methods ("ie*" and "branch and bound"). The alignments were also elided into each other to form a single grand alignment. After compression, the removal of uninformative and repetitive character distributions, this grand alignment object was submitted to phylogenetic analysis in the same manner as the individual data. When multiple equally parsimonious cladograms were derived from individual or multiple multiple alignments, the strict consensus was constructed.

The constructed alignments were also subjected to Gatesy *et al.*'s (1994) culling procedure. Alignment-ambiguous positions were removed before submission to phylogenetic analysis. This was performed to give a reference point from which to compare the more fine grained approach presented by elision.

In all cases, the components presented by the various analyses were compared to morphologically derived schemes of relationship (Fig. 4). Although these hypotheses are by no means necessarily "true," they offer a gauge of congruence features of these methods.

RESULTS

Insect 16S mtrDNA

The insect sample of 9 sequences was aligned 10 times. In each case, a different ratio of gap to change cost was used. Of these 10 alignment scenarios, 2 (gap to change cost ratios of 4:1 and 16:1) yielded non-unique (two, both times) alignments (this does not mean that these solutions are the only ones possible, merely that the heuristic solutions employed found these two). These 12 alignments yielded 11 unique topologies after phylogenetic reconstruction (Fig. 5, Table 2). Only gap:change ratios of 128:1 and 256:1 bore the identical cladograms. The 2 topologies derived from the 2 equally parsimonious alignments at gap:

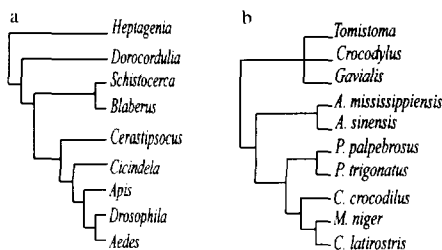


FIG. 4. Morphologically derived phylogenetic schemes. (a) Insects. (b) Crocodylians.

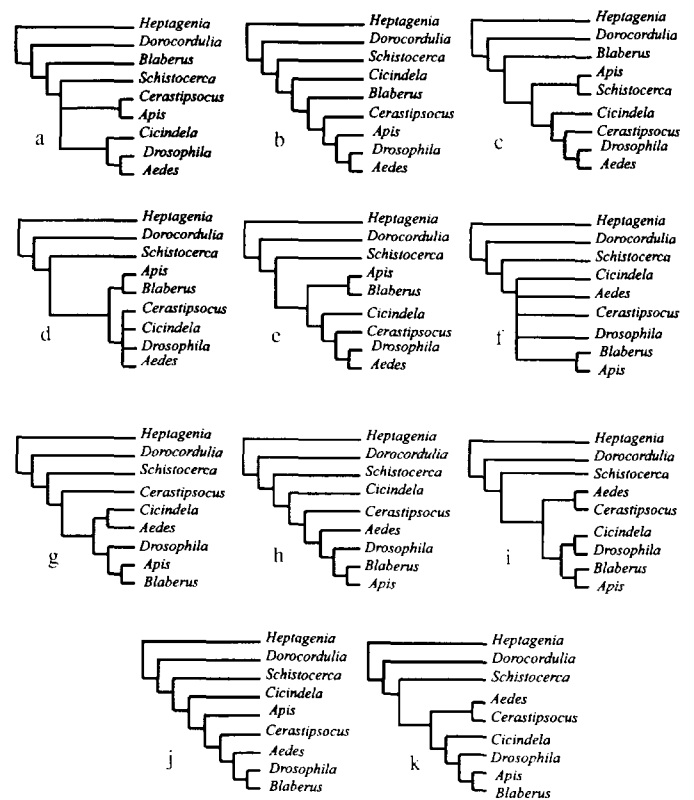


FIG. 5. Phylogenetic hypotheses generated by individual insect alignments with gap:change cost ratio of (a) 1:2, (b) 1:1, (c) 2:1, (d) 4:1 (first of two equally parsimonious alignments), (e) 4:1 (second of two equally parsimonious alignments), (f) 8:1, (g) 16:1 (first of two equally parsimonious alignments), (h) 16:1 (second of two equally parsimonious alignments), (i) 32:1, (j) 64:1, and (k) 128:1 and 256:1. For cost ratios of 1:2, 4:1 (first), and 8:1, the topology is that of the strict consensus of the two (1:2) or three (4:1 and 8:1) equally parsimonious cladograms derived from the alignments. In all of the other cases, only a single most parsimonious cladogram existed.

change ratios of 4:1 and 16:1 were very similar. The 4:1 topologies differed only in degree of resolution.

When the multiple alignments were analyzed together by removing alignment-ambiguous areas or eliding them together, the results varied. The 4:1 and 16:1 alignment pairs were each subjected to culling and eliding. The cladograms derived from these analyses yielded unique topologies which differed not only from those produced by the multiple alignments from which the new data sets were created, but also from those derived from other gap:change ratios (Fig. 6).

When all of the 12 insect alignments were combined, this grand alignment created by removing alignment-ambiguous sites yielded two phylogenetic topologies whose strict consensus (Rohlf, 1982) was entirely unresolved except for a single questionable component (*Apis* + *Cerastipsocus*), whereas the elided result was much more resolved. These groups are also difficult to reconcile with other data. Clearly, these are not the best of data for resolving insect relationships.

TABLE 2
Alignment Cost Scenarios and the Generation of
Equally Parsimonious Alignments

Data set	Gap:cost ratio	Number of equally parsimonious alignments
Insect	1:2	1
	1:1	1
	2:1	1
	4:1	2
	8:1	1
	16:1	2
	32:1	1
	64:1	1
	128:1	1
	256:1	1
Crocodilian	1:2	6
	1:1	2
	2:1	2
	4:1	3
	8:1	2
	16:1	4
	32:1	1
	64:1	1
	128:1	1
	256:1	1

Crocodile 12S *mtrDNA*

The crocodilian sequence alignments yielded more alignments but were somewhat more consistent than those of the insect data. The 10 alignment cost scenarios yielded 23 alignments, yet only 4 different phylogenetic hypotheses were generated in total (Fig. 7). Only gap:change cost ratios of 32:1 and higher offered unique alignments. When the ratios were lower, multiple equally parsimonious alignments were generated. The lowest cost ratio (gap:change 1:2) yielded 6 equally parsimonious alignments, each of which produced the same topology. The alignment based on a cost ratio of 16:1 yielded 4 alignments, a ratio of 4:1 yielded 3, and 1:1, 2:1, and 8:1 each yielded 2. In no case of multiple equally parsimonious alignments did the phylogenetic reconstructions vary among the alignments.

These 6 (of 10) cost scenarios which yielded multiple multiple alignments were subjected to the removal and combination of data procedures. For two of these situations (gap:changes cost ratios of 2:1 and 4:1), the removal and combination of data differed. In both cases, the removal of data, culling, yielded phylogenetic reconstructions with unique topologies (the 2 cases broke up and interdigitated the genera *Paleosuchus* and *Caiman*), which the combination of data, eliding, maintained the coherence of these genera.

The cull and elision processes applied to all 23 of these alignments together yielded grand alignments whose phylogenetic repercussions were more in line with the cladograms derived from the individual align-

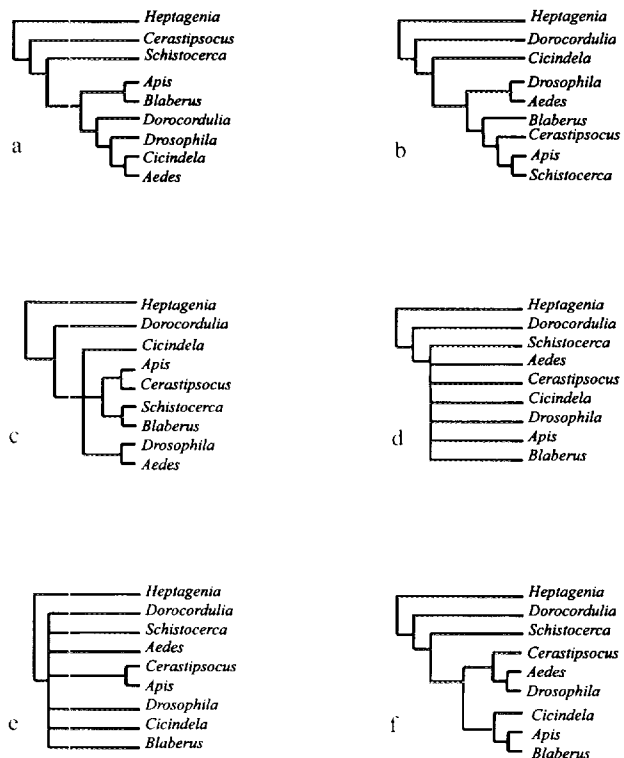


FIG. 6. Phylogenetic hypotheses generated by the “cull” and “elision” procedures applied to the insect sequences. Topologies (a) cull for a gap:change cost ratio of 4:1, (b) elision for cost ratio of 4:1, (c) cull for cost ratio of 16:1, (d) elision for cost ratio 16:1, (e) cull for all alignments taken together, and (f) elision for all alignments combined. Topologies a, b, and f are the single most parsimonious reconstructions for the combined alignments, whereas topologies c, d, and e are strict consensus cladograms of the two, six, and two (respectively) equally parsimonious cladograms derived from combined alignments.

ments relative to the insect data (Fig. 8). The combined (elided) data yielded a fully resolved cladogram identical to that produced by the majority of the original alignments. The removal of alignment ambiguous sites yielded an unresolved version of the elision scheme with three fewer resolved components.

DISCUSSION

Removal vs Combination

The methodologies of removal (culling) and combination (elision) of alignments present the extremes of analytical procedure. When alignment-ambiguous data are culled from the analysis, robust yet unresolved hypotheses usually result. This may seem conservative and safe since components not universally supported are excluded as with the “strict” consensus, but this conservatism can be extreme. The culled insect data yielded only a single resolved component (and that not congruent with morphological information). The combination of data will most probably be bolder in its

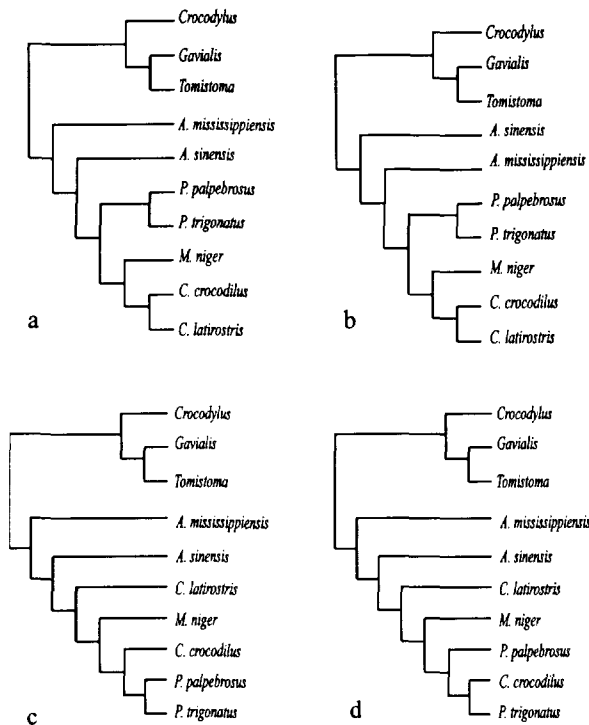


FIG. 7. Phylogenetic hypotheses generated by individual crocodylian alignments with gap:change cost ratio of (a) 1:1, 2:1, 4:1, 8:1, and 16:1, (b) 1:2, (c) 32:1 and 64:1, and (d) 128:1 and 256:1. In each case, only a single most parsimonious cladogram existed for each alignment.

assertions, that is, propose more resolved nodes. These nodes are less likely to be generally supported or "correct," however. This is due to the possibility that no single data set may support a grouping, but their agglomeration does.

The possibility that sequence data may generate nonunique solutions both for a single set of alignment parameters and for multiple parameter sets presents the possibility of relative weighting among the solutions. To avoid one parameter set dominating the final result by generating many equally costly alignments compared to a few or single results from other alignment regimes, these equally costly solutions could be elided to one another and reweighted before comparison to other solutions. In this way, each set of alignment parameters (gap costs, etc.) would contribute equally to the final result.

Overall, the combination of the nonunique alignments yielded more resolved phylogenetic hypotheses (Table 3). These resolutions are not, however, necessarily more congruent with other data. Given that the parameters used in phylogenetic reconstruction are often unmeasurable, the broad scale behavior of the process is all we may have (Wheeler, 1994). There are values which, although assignable to these parameters, yield results at variance with other, external

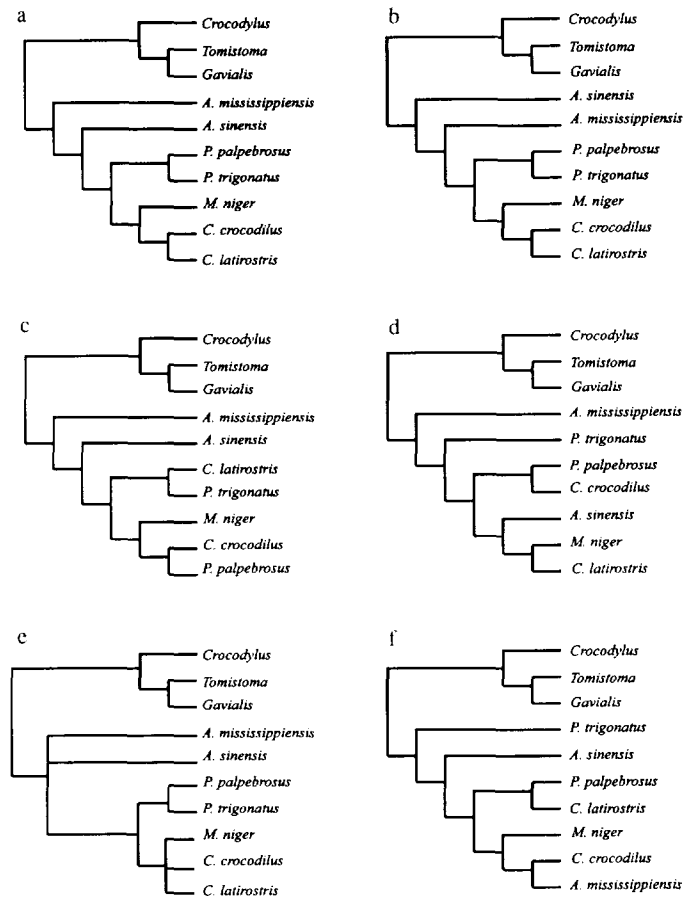


FIG. 8. Phylogenetic hypotheses generated by the "cull" and "elision" procedures applied to the crocodylian sequences. Topologies (a) cull for gap:change cost ratios, 1:1, 8:1, and 16:1; elision for gap:cost ratios 2:1, 8:1, 16:1, and for all alignments taken together; (b) cull and elision for gap:change cost ratio of 1:2; (c) cull for gap:change cost ratios of 4:1; (d) elision for gap:change cost ratio of 1:1; (e) cull for the combination of all alignments; and (f) cull for gap:change ratio of 2:1. With the exception of topology "e," each of the topologies is the single most parsimonious cladogram for the culled or elided alignments. The culled data for all of the alignments yielded four equally parsimonious reconstructions, the strict consensus of which is shown.

sources of phylogenetic information. This information may come in the guise of explicit character-based analyses of morphological features or may be based on the distributional data of biogeography or paleontology. When compared to such information, certain parameter values coincide with other analyses more than others. Such congruence information is one way to establish values for unmeasurable parameters. The monophyly of the Diptera (the flies *Aedes* and *Drosophila*) under low gap cost regimes in the insect data is one such indication.

The elided data give maximum resolution of the resultant cladograms, but have the disturbing property of assigning multiple putative homologies to the same datum. Since the alignments are repeated, a given

TABLE 3
Resolution and Congruence of Culled and Elided Alignments

Data set	Gap:change cost	No. alignments	Procedure	Components	
				Resolved	Congruent
Insect	4:1	2	Cull	6	0
			Elision	6	2
	16:1	2	Cull	5	3
			Elision	1	1
	All	12	Cull	1	0
			Elision	6	2
Crocodilian	1:2	6	Cull	6	4
			Elision	6	4
	1:1	2	Cull	6	4
			Elision	6	1
	2:1	2	Cull	6	1
			Elision	6	4
	4:1	2	Cull	6	2
			Elision	6	4
	8:1	3	Cull	6	4
			Elision	6	4
	16:1	4	Cull	6	4
			Elision	6	4
All	23	Cull	3	3	
		Elision	6	4	

base contributes more than once. Although this method will result in weighting characters based on their alignment consistency (as with successive approximations weighting—Farris, 1969), the implications for homology are unsettling, since individual bases must have individual histories, but are not treated as such.

The elision approach offers a fine grained weighting scheme. Since these weights will reflect the ambiguities of alignment, the characters are assigned importance on the basis of consistency, and phylogenetic inference can progress in the face of sequence alignment ambiguities.

REFERENCES

- Boudreaux, H. B. (1979). "Arthropod Phylogeny with Special Reference to Insects," Wiley, New York.
- Clary, D., and Wolstenholme, D. (1985). The mitochondrial DNA molecule of *Drosophila yakuba*: Nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* **22**: 252–271.
- Densmore, L. (1983). Biochemical and immunological systematics of the order Crocodylia. In "Evolutionary Biology," Vol. 16. (M. Hecht, B. Wallace, and G. Prance, Eds.), pp. 397–465, Plenum, New York.
- Farris, S. J. (1969). A successive approximations approach to character weighting. *Syst. Zool.* **18**: 374–385.
- Farris, S. J. (1988). HENNIG86 Ver. 1.5, Program and Documentation, Port Jefferson, NY.
- Feng, D., and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**: 351–360.
- Feng, D., and Doolittle, R. F. (1990). Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol.* **183**: 375–387.
- Fitch, W. M., and Smith, T. F. (1983). Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA* **80**: 1382–1386.
- Gatesy, J., and Amato, G. (1992). Sequence similarity of 12S ribosomal segment of mitochondrial DNAs of gharial and false gharial. *Copeia* **1**: 241–243.
- Gatesy, J., DeSalle, R., and Wheeler, W. C. (1994). Alignment-ambiguous nucleotide sites and the exclusion of data. *Mol. Phylogenet.* **2**: 152–157.
- Hein, J. (1989). A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when a phylogeny is given. *Mol. Biol. Evol.* **6**: 649–668.
- Hein, J. (1990). Unified approach to alignment and phylogenies. *Methods Enzymol.* **183**: 626–644.
- Higgins, D. G., and Sharp, P. M. (1988). CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* **73**: 237–244.
- Higgins, D. G., and Sharp, P. M. (1989). Fast and sensitive multiple sequence alignments on a microcomputer. *Comput. Appl. Biosci.* **5**: 151–153.
- Hsu Chen, C. C., Koten, R. M., and Dubin, D. T. (1984). Sequences of the coding and flanking regions of the large ribosomal subunit RNA gene of the mosquito mitochondria. *Nucleic Acids Res.* **12**: 7771–7785.
- Kristensen, N. P. (1975). The phylogeny of the insect "orders": A critical review of recent accounts. *Z. Zool. Evol. Forsch.* **13**: 1–44.
- Kristensen, N. P. (1981). Phylogeny of insect orders. *Annu. Rev. Entomol.* **26**: 135–157.
- Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**: 443–453.
- Norell, M. (1988). Cladistic approaches to evolution and paleobiology as applied to the phylogeny of alligatorids. Ph.D. Thesis, Yale University, New Haven, CT.
- Rohlf, F. J. (1982). Consensus indices for comparing classifications. *Math. Biosci.* **59**: 131–144.
- Sankoff, D. D., and Cedergren, R. J. (1983). Simultaneous comparison of three or more sequences related by a tree. In "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison." (D. Sankoff and J. B. Kruskal, Eds.), pp. 253–264, Addison-Wesley, Reading, MA.
- Swofford, D. M. (1990). PAUP ver. 3.0s. Program and Documentation, Ill. Nat. Hist. Surv., Champaign-Urbana, IL.
- Vogler, A. P., DeSalle, R., Assmann, T., Knisely, C. B., and Schultz, T. D. (1993). Molecular population genetics of the endangered tiger beetle, *Cincindela dorsalis* Say (Coleoptera: Cincindeleidae). *Ann. Entomol. Soc. Am.* **86**: 142–152.
- Vlasek, I., Bungschaiger, S., and Krell, G. (1987). Honey bee mitochondrial large ribosomal RNA. *Nucleic Acids Res.* **15**: 2388.
- Waterman, M. S., Eggert, M., and Lander, E. (1992). Parametric sequence comparisons. *Proc. Natl. Acad. Sci. USA* **89**: 6090–6093.
- Wheeler, W. C. (1993). The triangle inequality and character analysis. *Mol. Biol. Evol.* **10**: 707–712.
- Wheeler, W. C. (1994). Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Syst. Biol.*, in press.
- Wheeler, W. C., and Gladstein, D. S. (1992). MALIGN Ver. 1.8, Program and Documentation, New York.
- Wheeler, W. C., and Gladstein, D. S. (1994). MALIGN: A multiple sequence alignment program. *J. Hered.*, in press.

APPENDIX

Representative Individual Alignments of Insect and Crocodile Taxa (Taxon Name Abbreviations as in Table 1)

Crocodylians: Gap:Change cost ratio 1:2--six alignments found

Tomistoma GACTTGACGGTACTTCGCACCCACCTAGAGGAGCCTGTCCCTATAATCG-A-AAACACTCGATAC--ACCCCTACCAACT-
CrocodylusT...A.....C.T.A.....C...A...C...
GavialisC.GT.....T.....T.....
A. mississippi.TAA...C.....C.G.T.A...-TAC...G...C...
C. latirostrisA.C...AA.T.....A..G.T.A...-T.C...TA...C.-C
A. sinensisCG...A.....CGGT...A...-T.C...G...C...
P. trigonatusTAA.....T...G.T.A...-C...TA...-C.C
P. palpebrosusA.....A.....T...G.T.A...-T.C...TA...-C.C
C. crocodilusA.....AA.T.....A..G.T.A...-T.C...TA...-C.C
M. nigerA.C...AA.A.....A..G.T.A...-C...TG...C.-C

Tomistoma TTTCCTTAAAC-TACA-TAA-CAGCCGTATACCGCCGTCG-CAAC-TA--ACCC--CTGAGGAC--GAAC--A--
CrocodylusG.....G.....G.T.G.....A.....-GA.C
GavialisG.....A.T.....
A. mississippi.A.....T.....G.CCGT--ATT...AAC--A...
C. latirostrisC.A.-T.....C...TT.....G.CTGT--T.G...A.AAC--A...
A. sinensisC.A.....C...G.CCGT.....GAC--A...
P. trigonatusC.A.....T.....TGT--A...AAC--A...
P. palpebrosusC.....T.....C...CGT--A...AAC--A...
C. crocodilusA.-T.....-TC--TT.C.....G...TGT--T.G...A.AAC--A...
M. nigerC.G.....-TC--T.....G...TGT--T.G...A.AAC--A...

Tomistoma GTTAAGTCAACAG--CT---CA-TT-TGA-GCTAATACGTCAGGTCAAGGTGCAGCCAAT-AAG-TTGGAAGAGATGG
Crocodylus -C...CA...T.-A.....C.C.....G.....T.....
GavialisA...T.....T.....G.....
A. mississippi.-CG-C.C.....CAAC--...C.....C...G.....
C. latirostris-G-G.CA.....C...C.CT...A.....C...A.T...GGG.TG-...T
A. sinensis-CG...CA...T.C...C.C...GC.....G...G.....A
P. trigonatus-CG...A.....C...C.C...A.T...CGG...C...T
P. palpebrosus-CA...A.....C...C.C...T...GG...C...T
C. crocodilus-G...CA...T.C...C.C...-C...A.....C.C...T...GGG.TG-...T
M. niger-A...CA...C...C.C...-C...A.....C...GGG.TG-...T

Tomistoma GCTACATTTTCT--AC-CTCATAGAAATAT--GTC-ACGGAGA-G-CCCTGTGAAA
CrocodylusA.....G.....G.....
GavialisAC.A.....
A. mississippi.CA.A.-G.....T.A.....A.....
C. latirostrisA.A.-G.....C...G.....C.T.....
A. sinensisA.A.....-A.....
P. trigonatusA.A.....-C...A.A...-C...A.....
P. palpebrosusA.A.....-C...A.A...-C...C.....
C. crocodilusA.A.....-C...G.....C.T.....
M. nigerA.A.....-G...-A.....-C.T...-A.....

Gap:Change cost ratio 8:1--two alignments foundTomistoma

GACTTGACGGTACTTCGCACCCACCTAGAGGAGCCTGTCCCTATAATCGAAAACACTCGATACCCCTACCAACTTTTGC
CrocodylusT...A.....C.T.A.....C...A...C...
GavialisC.GT.....A.T.....G...C...A...
A. mississippi.C...TAA...C.....C.GT.A.T...G...C...A...
C. latirostrisA.C...AA.T.....GT.A.T...T...G...C...C.C.A.T
A. sinensisCG...A.....CGGT...A...T...G...C...C.A...
P. trigonatusTAA.....T...G.T.A...C...TA...CCTCC.A...
P. palpebrosusA.....A.....T...G.T.A...T...TA...CCTCC...
C. crocodilusA.....AA.T.....GT.A.T...T...TA...C.C.A.T
M. nigerA.C...AA.A.....GT.A...C...TG...C.C.C.G...

Tomistoma CTTAAACTACATAACAGCCTGTATACCGCCTCGCAAACTAACCCCTGAGGGACGAACAGTTAAGTGCACACCTATT
Crocodylus .C.--AG--GC.T.TA.-ACCG...T...AAGCTTAGC..ATGA.G.AC.-AG.ACC.AGCACAAT.ACT.A.T.C
GavialisG-T.C.....G.....A...A.T...A...T...T...
A. mississippi.-GT.T.TA.-ACCG...T...AAGCC.GTC..ATT...A.CA.AACGC.C...T...AC
C. latirostris-C...-GTTT.TA.-ACCG...T...-CA.GC..GT.T.G...A...A.CA.AA.GG.CA...TCC
A. sinensis-C...-GC.T.TA.-ACCG...T...-CA.GC.CGT...A...A.CA.AACG..CA...T...TCC
P. trigonatus-T...-GT.T.TA.-ACCG...T...-CA..CT.GT...A...A...A.CA.AACG...A...TCC
P. palpebrosus-T...-GT.T.TA.-ACCG...T...-CA..CTCGT...A...A.CA.AAC...A...TCC
C. crocodilusT...-TC--GTTT.A.-ACCG...T...-CA.GCT.GT.T.G...A...A.CA.AA.G..CA...T...-CC
M. niger-C-TC--GT.T.TA.-ACCG...T...-CA.GCT.GT.T.G...A...A.CA.AA...CA...T...-TCC

Tomistoma TGAGCTAATACGTCAGGTCAAGGTGCAGCCAATAGTTGGAAGAGATGGGCTACATTTTCT-ACCTCATAGAAATATGTC
CrocodylusG.....T.....AC.....TG...
GavialisG.....A.....
A. mississippi.C...C.....C...G.....C.A.A.TG.....T.CA
C. latirostrisAG...A.....C...A.T...GG.G...G.AG...T.....A.A.....C.C.G
A. sinensisCAG...GC.....G...G.....A.....A.A.....C.A
P. trigonatusCAG.....A.T.CGGAGC.....T.....A-A.....C.A
P. palpebrosusCAG.....T...GGAGC.....T.....A-A.....C.A
C. crocodilusC-C...A.....C.C...T...GG.G...G.AG...T.....A.A.....C.C.G
M. nigerCAG...A.....C.....GG.G...G.AG...T.....A.A.....G.A

Tomistoma ACGGAGGCCCTGTGAAA
CrocodylusG.....
GavialisA.....
A. mississippi.AC.T.C.....
C. latirostrisA.....
A. sinensisA.A.C.....A.....
P. trigonatusA.A.C.....C.....
C. crocodilusA.A.C.....T.C.....
M. nigerC.T.CA.....

APPENDIX—Continued

Gap:Change cost ratio 256:1--one alignment found

Tomistoma GACTTGACGGTACTTCGCCACCCACCTAGAGGAGCTGTCTTATAATCGAAACACTCGATACACCCACCAACTTTTGC
Crocodylus C . T . A C A C
Gavialis C . GT T T
A. mississippi C TAA C C . GT A . T G C A
C. latirostris A . C AA . T GT A T TA C . CC . A . T
A. sinensis CG A CGGT A T G C . C . A
P. trigonatus TAA T G . T A C TA CCTCC . A
P. palpebrosus A A T G . T A T TA CCTCC
C. crocodilus A AA . T GT A TA C . C . A . T
M. niger A . C AA . A GT A C TG C . CC . G

Tomistoma CTTAAACTACATAACAGCCTGTATACCGCCGTGCAAACTAACCCCTGAGGGACGAACGTTAAGTGCAACAGCTCATT
Crocodylus G GCC-TGTATACC---G.CGTC-GCA.GCT.G...ATGAG.ACAAG.ACC.AGCACAAT.ACT.A.T.C
Gavialis -GTC G A A T T
A. mississippiACTCA-GTC.-GT.TA.C.----.T.-.AAGCC.GTC.ATT.....AAC.A.ACCGCCGA.CAGCTCA.CC
C. latirostrisACCC---GT.-TGTATACC---G.CGTC-GCA.GC.GT.T.G...A.AAC.-AA.GG.CA...C.CC.A.T
A. sinensisCCTC.-G.C.-GT.TA.C.----.T.-.CA.GC.CGT.....GAC.-A.ACG.CA...T...CTCCC
P. trigonatusCTC.-GTC.-GT-ATACC---G.CGTC-GCA.CT.GT...A...AAC.-A.ACG.A-C.A.CAGC.TCC
P. palpebrosusCTC.-GTC.-GT-ATACC---G.CGTC-GC.....CET...ACTGA..GAA.C.-AAAC.AGTGC.ACAGC.TCC
C. crocodilus TA.CC.-GTT.-G.-ATACC---G.CGTC-GCA.GCT.GI.T.G...A.AAC.-AA.G.TA...C.ATAGC.CCC
M. nigerC.CC.-GTC.-GT-ATACC---G.CGTC-GCA.GCT.GI.T.G...A.AAC.-AA...CA.....CTCCC

Tomistoma TGAGCTAATACGTCAGGTCAAGGTGCGCCAATAAGTTGGAGAGATGGGCTACATTTCTACCTCATAGAAATATGTCA
Crocodylus G T AC TG
Gavialis G G
A. mississippi GAGCTA.CACGTCAG.TCA.G.TGCAGC.A.CA.G..G.A.CAGATG.CTACAT...CTC.A.ATG.....TCA.
C. latirostris A.....A.....C.....A.T...GG.G...G.AG...T.....A.A.....C.G.
A. sinensis A.....GC.....G.....G.....G.....A.....A.....A.....CA.
P. trigonatus CA-.GCT.ATACGTCA.GTC.A.GTGCAA.T.ACG.AGC.G.A-.GAT.TGCT.CA.T.CTAAA.....C.A.
P. palpebrosus CA-.GCT.ATACGTCA.GTC.A.GTGCA.G.T.ATG.AGC.G.A-.GAT.TGCT.CA.T.CTAAA.....C.A.
C. crocodilus C.-CTA..ACGTCAG.TCA.C.CGCAGCTA.TGGG..G.....T.....A.....C.G.
M. niger A.....A.....C.....GG.G...G.AG...T.....A.A.....G.A.

Tomistoma CCGAGAGCCCTGTGAAA
Crocodylus G
Gavialis
A. mississippi A
C. latirostris AC . T . C
A. sinensis
P. trigonatus A . AC
P. palpebrosus A . AC
C. crocodilus AC . T . C
M. niger C . T . CA

Insects: Gap:Change cost ratio 1:2--one alignment found

Heptagenia TATTAATCCAACATCGAGGTCCGCAAGCCTTCTGTAAATTTGAACCTGAAAGAAGATTACGCTGTATCCCTAAGGTA
Dorocordulia A
Blaberus C T C AA C A
Schistocerca C T . TGCT CG . A T C A . CA
Cerastipsocus C T T . T . CG . A G . A GT C . A
Aedes C T . T . T CG . AA A A
Drosophila C T . T . T . A . CG . A CC . A A
Cicindela T A . T . T . A . CG . AA . G CT . A . A
Apis C T . C . G A . A . CT . TACT . A . GTACA . C AG . T

Heptagenia ACTT-ATCTAATGATC-AATAATAATGGATC-CAATATTCA'AAATTAAT-GGITATTAAGTTATA-AAAGCTTAAGTTA
Dorocordulia A
Blaberus T AG . T . A C A . AC . AT . C ATG . AA T . A . A G TTA . T
Schistocerca T A T TA T A ACA . A A . C AAA . G AT . A . A GA . G TA . T
Cerastipsocus T A TT . CT . TA . A . TATT . A AA . A ATT . A G . AC . A . A . AT . T T T
Aedes A T . T . A A . T A A T A . AC . A ATA A
Drosophila A T . T . A T G . A . T T T . A . A TA T
Cicindela A TT . A T AT . A . C C A . A . GT . A . AG . G . A AG . T
Apis T AT TT . A CA . T . T . AAT A . A . ATC . TT . ATCAAAA AATCT . A . AT TA

Heptagenia TATATTTTTGTCCACCCCAACAAAACAATAAATAAATTTATATTAATTACTTAAT-C-A-AATAATATAAAAAAATTAG
Dorocordulia A TA C . A
Blaberus C . CA A TC T . CT . C TA AT . A . TA A C A
Schistocerca AT . C . CA TC . ICATT . AA AG . AAGA . C . A . A A . C T A . TAA . A
Cerastipsocus T ACA . A G T TT . A . TCAT A . A . AAT A AA . TTC . T A TC . TAGG
Aedes A AA . AC GT TT . TT AC AAT . A . T TT . TT CT . T AA . A
Drosophila A AA . A T T . T . TT . ATTTA AATTA . T TT . T TA T . A . TAA . A
Cicindela A C . C . A . A . T TTCA . TTA AACTAA . T C . T T . CA T . T . A . TA
Apis A ACCAA . CT TC TTTA . TC . TT . AA ATT . - . T A . ATT . TA A . TTT . A A

Heptagenia TGTAAAGCTCA-
Dorocordulia GCTC
Blaberus
Schistocerca T
Cerastipsocus C . CTC T . T
Aedes AA A T
Drosophila A A . T
Cicindela AA A
Apis AA T T

APPENDIX—Continued

Gap:Change cost ratio 8:1--one alignment found

Heptagenia TATTAATCCAACATCGAGGTCGCAAGCCTTCTTGTA...
Dorocordulia .C.....A.A.GC.-...
Blaberus .C.....T...T...C...AA...C...A...A...
Schistocerca .C.....T.TGCT...CG...A...T...C...A.CA...
Cerastipsocus .C.....T.T.CG...A.G...A...GT...C.A...
Aedes .C.....T...T...CG...AA...A...A...A...
Drosophila .C.....T...T.A.CG...A...CC...A.A...A...
Cicindela .T.....A.T.T.A.CG...AA...G...CT...A.A...
Apis .C.....T.C.G...A.A.CT...TACT...A...GTACA.C...AG.T...

Heptagenia ACTT-ATCTAATGATCAATAATAATGGATCCAATATTCATAAA...
Dorocordulia ...A.....T.AGTTA...T
Blaberus .T.AG...T.A...C.....A.AC.AT.C...ATG.-AA.TA.T.A...A...GAG...TTA.-T.-AT
Schistocerca ...A...T...TA...T...A.ACAA.A.C.A...-AAATGAT.TAATA.TG.AGAGTTA.T...T
Cerastipsocus .T.A...TT.CT.T.A...T.TAGAT...A.AATTC...T.AT.A-A.GG.A.CAAA.A.TT.AG...TTT.-T.-T
Aedes ...A.T.TT.A...A.T...A...A...T...A.AC...A.A...T.A.TAA...-T
Drosophila ...A.T.TT.A...T...T...GA.T...T...T.AAAT...T.TTAA...-T
Cicindela ...A...TT.A...T...AT.A.C...C...A.A.GTAA.G.G...TTAAGTTA...-CT
Apis .T.AT...TT.A...T.CA.T.T.AAT...A.A.ATC.TT...ATCA-AAAT.A...TC.TA...T.AAG.TTA.-TA-A.

Heptagenia TTTTTGTCACCCCAACAAAACAATAAATAAATTTATATTAAT...
Dorocordulia ...G.CA.C...AAC.A...C.AT.A.T.A.TT.A...TAA.T.ATTAATCCA...
Blaberus C.CA...A...-C.ATC...AT.ACT.CT...A...-A...TAATTA.TCA...C...T...AT...AAC.C-TA
Schistocerca ...CA...-C.AC...AC...C.TC.TTAAATATAG...AG...AA.CAA...ACTATAT...AT.AA.TGT...GC
Cerastipsocus .ACAA.CG.C...-A.TT...TTA...C.TA.T...AA.A...TTAATAAA.C.TCTT...A...TC...TAGG...C.CTC.
Aedes .AA.ACCA.C...-AGTA...TTTT.T.TA...C.ATA.A.TTAATTA.TC.TTTT.C.TAT...TTA...AA.AATA...GA
Drosophila .AA.A.CA.C...-A.TA...TATT.T...TT...A-A-A.T.AATTAATC.TT.TA.T.A...T.A.TA.AA.TATA...GA
Cicindela C.A.A.TA.C...-A.TA...TTC.ATT.A...T...A.C...A.TAA.CC.A...T...TC.A...T...AATA...GA
Apis ...ACCAATC.T...CCA.TC.A.T.T...CTTAAA...ATT.ATTAA...T.T.A.TA.ATTT...A...A.AATA...TT

Heptagenia CTCTA
Dorocordulia TC...
Blaberus ----
Schistocerca TCTA-
Cerastipsocus ..T.T
Aedes TCTAT
Drosophila T.TA-
Cicindela TCTA-
Apis ..A.-

Gap:Change cost ratio 256:1--one alignment found

Heptagenia TATTAATCCAACATCGAGGTCGCAAGCCTTCTTGTA...
Dorocordulia .C.....CATCGAG.ACGAAGCCTC.TG.AA.TT.GAACT...GAA.G.AG.TTACG.TG.TA.CC.TAAGGTAAC
Blaberus .C.....T...T.TGTC.A.AAG.ACTCTC...A.AGA...ACGCTG.TAT...CT.A.GT
Schistocerca .C.....T.TGCT...CG...A...T...C...A.CA...
Cerastipsocus .C.....T.CATCGAG.TCGCA.TCCT-CG...A.G...A...GT...C.A...
Aedes .C.....T.T...T...CG...AA...A...A...A...
Drosophila .C.....T.T...T.A.CG...A...CC...A.A...A...
Cicindela .T.....A.T...T.TATCGA.AAG.GCTCTCT.A...A.A.TACGCTG.TAT...CT.A.GT
Apis .C.....T.C.G...A.A...T...CTA.ATGGTA.ATC...A.GATA.TACGCTG.TAT...CT.A.GT

Heptagenia ACTTATCTAATGATCAATAATAATGGATCCAATATTCATAAA...
Dorocordulia T-.A.....T.AGTTA...T
Blaberus -.A.T.AGTC.T...A.TC.C...AATGGATCA.AA...T.CA...-A.G.A.TTA.TA.TT.AAAAG...T.-AT.
Schistocerca -.C.TAA.CT.ATGATCAT.A.T.ATGGATC.A.ACA.C...CAT.AATAA...G.TA.A...TGAAGAGTT.AT...
Cerastipsocus ...TAA.CT.TTCITTA...A.TATTAGATCA.AAA...TC.A.T...AAA.GG...CAA.A...TTAAG.TTT...T.-T.
Aedes -.C.TAA.TT.TTAATCA.TA...ATTGGATC.A...AT.C.T...ATTAATG...AAC.AA...AT...AG...TAA...-T.
Drosophila -.C.TAA.TT.TTAATCATTATTAATGGATG.AT.AT-.C.TAA.T.-AA.G...TTT.A.A.TT.AAAGTTT...A-AT.
Cicindela -.ACT.AATC.TT.A.TC...ATTGGATCATAAA-CTC...A...CAA.G.A.TAAGTA...G...AA.T.AG.TA-ATC
Apis -.A.T.A.TC.TT.A.T...CA.T.TATAATTC.A.AA...TC...T...-A.C.AA.TTAA...CTT.AA.A...TA-.TA

Heptagenia TTTTTGTCACCCCAACAAAACAATAAATAAATTTATATTAAT...
Dorocordulia ...G.CA.C...AAC.A...C.AT.A.T.A.TT.A...TAA.T.ATTAATCCA...
Blaberus .A.C.CATTAT-.C.CC...TC.A.T.ACT.-C...ATA.T.A...TAATTAT.TCA...ATA.TCT.AT.A...TAT...CT
Schistocerca .C...CA.GT.A...CCA.C...A.CATC...C.T.AA...AG.AAGACA...CA...A.C...T...A.TAA.A...T...
Cerastipsocus ACAA.CG.C...AATTA.TT.A.TC.T.AT.TAAA.AT...T.A...AAACT...TTTA.TA.A...TCTAT.GGGTCTCTCGCT
Aedes AA.ACCA.C...AGT.A...TTTTT...T.A...C.A...A...T.A.TA...CT.TTT.ACT...TT.A...AA.A.AA.G.T
Drosophila ..AA.A...T...T.T.TT.ATTI-A.TA.AATTAA.T.A.TCTTT...TA...T.A.TAA.A.A...A
Cicindela .C.A.A.T...T...TTCA.TTA...TA.ACT.AATTAACTCT...T.CA...T...T.A.TAAA...A
Apis AA...AC...AT...TC.C...T...ATTTA.TC...A.ATA...T...T.A...ATT.TA...A.TTT.A...A.AAA...TT

Heptagenia TCTA
Dorocordulia C...
Blaberus CTA-
Schistocerca ...
Cerastipsocus .T.-
Aedes .A.-
Drosophila .T...
Cicindela ...
Apis CTAT