Molecular Ecology and Evolution: Approaches and Applications
ed. by B. Schierwater, B. Streit, G.P. Wagner & R. DeSalle

Sources of ambiguity in nucleic acid sequence alignment

W.C. Wheeler

Department of Invertebrates, American Museum of Natural History, Central Park West at 79th St., New York, NY 10024-5192, USA

Summary. The discussion of molecular sequence alignment is becoming more prominent in studies of molecular systematics and evolution. As the basis for initial homology statements, alignment is crucial to comparative molecular biology. Although fundamental, alignment is not a process which necessarily yields objective, precise results. Ambiguities can appear in alignment due to a number of factors. Three such sources of ambiguity are discussed here. These are ambiguity in the establishment of alignment parameters, pair-wise order and individual "path" variation. The first arises from the necessary but empirically untestable assumptions of gap costs and other factors which are required to align sequences objectively. The second is due to the possible existence of non-unique solutions to the same alignment parameters in heuristic and exact solutions. The third is a result of multiple optimal paths within single alignments, potentially generating huge numbers of equally costly but unique alignments. Some of the problems with and several possible solutions to the difficult situation of non-unique alignments are discussed.

Introduction

Alignment of sequence data is the fulcrum of molecular systematics. All evolutionary analyses of molecular sequence rely on the series of correspondences that comprise an alignment. In systematics and evolution, these correspondences are the putative homologies tested by cladograms. Cladistic analysis then gives us the ability to discern historical relationships based on the discrimination of homology and non-homology or homoplasy. However, these higher level operations are dependent entirely on the foundation alignment. Even when alignments are generated automatically, as opposed to "by eye," untested assumptions and ambiguities can creep into the alignment process. Since alignment is so basic, these same assumptions and ambiguities perfuse subsequent operations, holding them subject to the same forces encountered by earlier phases of analysis.

Sources of ambiguity

Three sources of ambiguity are discussed here: "parameter" variation, "order" variation, and "path" variation. Each of these factors can affect

the correspondence between and among sequence positions. Most of the discussion here will be limited to nucleic acid sequence data, but the arguments apply equally to amino acid data. There is an additional source of uncertainty in the alignment of phylogenetically interesting numbers of taxa and that is the operational necessity to deal with heuristic solutions. As with phylogenetic analysis, the algorithmic method to find the "best" or optimal alignment is known. The method is an extension of the two-dimensional Needleman and Wunsch procedure (Needleman and Wunsch, 1970). This well defined algorithm is almost entirely intractable for large numbers of sequences requiring storage and computation increasing by a factor of the length of each successively added sequence (m^n, where m is the length of the sequences and n the number of sequences). This extreme level of computational complexity necessitates the use of heuristic shortcuts to find workable solutions, thus adding an extra degree of fog to the analysis. This problem is not fatal, however. Phylogenetic analysis has faced this situation for some time and has been able to progress. The ambiguities discussed here apply to both heuristic and exact solutions.

The first source of ambiguity in sequence alignments comes from the necessity of specifying the relative costs of the events required to transform one sequence into another and create the alignment correspondences. The most obvious of these are insertion-deletion and base change costs. The insertion-deletion or "gap" value is the cost (in some optimality currency) of the insertion or deletion of a gap or base. This must occur when sequences are unequal in length. The base change cost is the debit incurred in the transformation of one nucleotide (amino acid) into another. The sum of base changes and gap events multiplied by their individual costs is the total "cost" of the alignment. The set of correspondences which minimizes this cost is the best alignment based on those parameter values. The ambiguity is derived from need to specify, or assume, these costs *a priori*. Different cost regimes can yield different alignments (Fitch and Smith, 1983; Waterman, 1992) which can in turn imply different schemes of phylogenetic relationship (Wheeler, in press). Since there is no way to determine directly the appropriate gap or change values through measurement, more or less arbitrary assumptions must be made (although they can be limited to some extent – Wheeler, 1993).

When multiple alignments are created, whether by exact or heuristic means, the notion of alignment order comes into play. Heuristic multiple alignment solutions are built typically from a series of ever larger pairwise alignments. Initially, two sequences are aligned and this result aligned to a third, maintaining the relative alignment between the first two ("once a gap, always a gap" – Feng and Doolittle, 1987, 1990) and so on. Any such alignment is obviously order dependent. A different order of alignment might well yield a different alignment. Even when some explicit optimality criterion, such as parsimony has been chosen, multiple orders may yield unique yet equally optimal multiple solutions. This situation gives rise to the multiple-alignment problem discussed by Gatesy et al. (1993) and Wheeler et al. (in press). The same data and parameter assumptions yield non-unique solutions. This is analogous to the problem of equally parsimonious cladograms found in almost all phylogenetic analyses. Most systematic data imply more than one most parsimonious scheme of relationships. The set of solutions may contain topologies which are almost entirely identical or highly divergent. The same situation can occur in alignment. One solution to the multiple order problem might seem to be simultaneous alignment of all sequences, but even if possible, this would be no escape. Exact methods such as that of Sankoff and Cedergren (1983) rely on a parsimony based alignment cost function. In order to calculate the cost the scheme of phylogenetic relationships among the sequences must be "known" or at least specified *a priori*. If the phylogenetic relationships are unknown or non-unique, the cost function would vary with each potential topology, potentially yielding non-unique multiple alignments.

The third source of alignment uncertainty discussed here is path variation. Path variation occurs when the alignment algorithm can follow multiple paths through the alignment space, yielding the potential for huge numbers of equally costly (or optimal) solutions. Any NW based alignment, irrespective of dimension, follows a similar procedure. A space is created bordered by the sequences to be aligned. If only two sequences are compared, they constitute a plane; three sequences are a cube, and so on to higher dimensions. The cells of this space are created by the matching and mismatching of the input sequences. Alignment starts at one corner, top left, and proceeds through the space attempting to find the path of lowest cost (or highest benefit). The "best" path is the alignment. Path variation occurs when the road diverges, so to speak. When the alignment can either insert a gap or match the bases with equal cost the number of solutions doubles. This can happen repeatedly, resulting in a large number of equally costly, but different alignments.

Each of these sources of ambiguity may result in multiple solutions to the same alignment problem. In the following sections, the NW algorithm and some of its elaborations will be described, especially within the framework of parsimony. These methods will then be applied to a situation likely to generate multiple solutions. In the final section, several as yet untested methods that attempt to accommodate multiple solutions will be discussed.

The basic algorithm

The algorithm initially proposed by Needleman and Wunsch (1970) is a dynamic programming string match algorithm applied to biological

sequences, but it can be used more generally to determine correspondences between any ordered series of objects. There are five parts to the algorithm: 1) specifying the cost function, 2) laying out the alignment matrix, 3) initializing the matrix elements, 4) "wave front" updating the matrix elements, and 5) tracing back. As an example, we may be presented with two short sequences, "ACGT" and "AGTT." The first step is to establish the cost parameters. Initially, they may be set to a gap cost of 10 and a mismatch cost of 1. (The description which follows is based on a minimal cost optimum as opposed to the maximum benefit described by Needleman and Wunsch. The details will differ from their description, but with suitable modulation of the cost function the results are identical.) A 5×5 matrix is then created (Fig. 1). These axes are one more than the lenghts of the sequences to allow for initial or leading gaps. The first row and column are labeled 0^{th} and each element is assigned an initial value of 0 (Fig. 2). The remaining elements from row and column 1 until the end (4) are given initial values of 0 if the corresponding bases from the sequences match and 1 (the mismatch cost) if they do not (Fig. 3). The matrix is now entirely initialized and the heart of the algorithm begins.

The algorithm returns to element (0, 0) in the upper left corner and updates each row and column in turn by realizing that in order to get to element (i, j) only three other cells need be examined (if they exist): those above $(i, j-1)$, to the left $(i-1, j)$, and diagonally above and to the left $(i-1, j-1)$ (Fig. 4). A path from the diagonal cell implies a correspondence between the two bases whether they match or not, whereas the left and upper cells imply gaps in either sequences II or I, respectively. This simple dependency is due to the assumption of the method if base i of sequence I corresponds to base j of sequence II, then

		A	C	G	T
	(0,0)				
A					
G					
T					
T					

Figure 1. Initial comparison matrix for two four-base-pair sequences.

		A	C	G	T
	0	0	0	0	0
A	0				
G	0				
T	0				
T	0				

Figure 2. Comparison matrix with initialization of first rwo and column complete.

		A	C	G	T
	0	0	0	0	0
A	0	0	1	1	1
G	0	1	1	0	1
T	0	1	1	1	0
T	0	1	1	1	0

Figure 3. Comparison matrix with initialization complete.

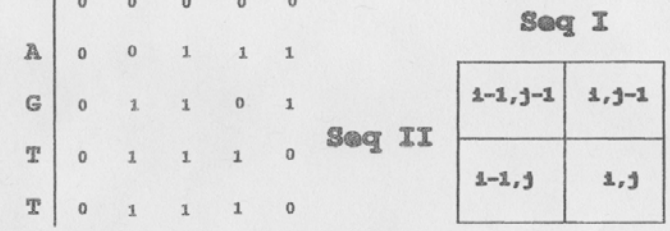


Figure 4. Cells which may lead to cell (i, j) during alignment.

$i + k (k \geq 0)$ must correspond to a base $j + l (l \geq 0)$. The first elements are those in row 0, the topmost. In these cases, only the row to the left will exist, hence, the algorithm proceeds from cell (0, 0) to the right (0, 1) *et seq.*, until all the elements have been updated. The first cell to undergo this process is cell (0, 1). Following the alignment logic, the only way to get to this cell is via a gap in sequence I. In this case, it is a leading gap which precedes any of the bases in sequence I. The new value of this cell, (1, 0) is calculated by adding two numbers. The first is the value of the preceding cell (in this case [0, 0]) and the second is the cost of inserting the gap to get from (0, 0) to (0, 1). Hence the new value is $0 + 10 = 10$. In addition to noting the new cost of the cell, the direction from which it came (here, left) is also noted (Fig. 5). The process is repeated for the entire row, in each case basing the new cell value on the one immediately to its left plus the gap cost (Fig. 6). This same operation is performed on the first column. Here, the gaps implied are in sequence I (leading gaps before sequence II; Fig. 7).

With the first row and column completed, the procedure moves down and over to row and column 1 beginning with cell (1, 1). Now that there are cells above, left and diagonally, the calculation of the new cell values are more complex. Three types of events must be considered: 1) a gap in sequence I, 2) a gap in sequence II, and 3) a correspondence between sequences I and II. The cost of the new cell will be the lowest of the three possible paths to the cell. Hence, the cost for each of the three paths is first calculated. The cost for a gap in sequence I is the sum of the value in cell (1, 0) and the gap cost for a total of 20. The cost of a gap in sequence II is in the same way the sum of cell (0, 1) and the gap cost for 20 again. In calculating the cost of the diagonal path, there is no gap penalty. The only costs are those in cell (0, 0) and (1, 1), summing to 0. The three paths, then, offer costs of 20, 20, and 0 with

		A	C	G	T
	0	→10	0	0	0
A	0	0	1	1	1
G	0	1	1	0	1
T	0	1	1	1	0
T	0	1	1	1	0

		A	C	G	T
	0	→10	→20	→30	→40
A	0	0	1	1	1
G	0	1	1	0	1
T	0	1	1	1	0
T	0	1	1	1	0

Figure 5. Beginning of "wavefront" update of comparison matrix. The second cell of the first column has been updated to reflect its origination from cell (0, 0) and the incurred gap cost. The arrow signifies that the cell to its left (0, 0) was the lowest (and only in this case) source for cell (1, 0).

Figure 6. Progression of "wavefront" update. The entire first row has been updated with each cell now containing both its cost and origin (arrow from previous cell).

the diagonal as the lowest. The diagonal path is chosen, noted in the matrix, and cell (1, 1) given its updated cost (Fig. 8). This is repeated for all cells until the mitrix is completely updated (Fig. 9). This is called a "wavefront" process because the update follows like a wave over the matrix. (A procedure similar to that for the first row and column can be employed to account for trailing gaps in the row and column.)

		A	C	G	T
	0	→10	→20	→30	→40
A	↓10	0	1	1	1
G	↓20	1	1	0	1
T	↓30	1	1	1	0
T	↓40	1	1	1	0

		A	C	G	T
	0	→10	→20	→30	→40
A	↓10	↘0	1	1	1
G	↓20	1	1	0	1
T	↓30	1	1	1	0
T	↓40	1	1	1	0

Figure 7. Update of the first column. The first column (signifying leading gaps in the top sequence) is completely updated with costs and directions.

Figure 8. Update of matrix cell (1, 1). This is the first cell which required operations involving all three possible originating cells, left, above and diagonally above and left. The least cost operation was to match from cell (0, 0) to (1, 1), resulting in a cost of zero. The costs from the cells to the left and above would both have been 20.

		A	C	G	T
	0	→10	→20	→30	→40
A	↓10	↘0	→10	→20	→30
G	↓20	↓10	↘1	↘10	→20
T	↓30	↓20	↘11	↘2	↘10
T	↓40	↓30	↓21	↓12	↘2

Figure 9. Matrix after completed "wavefront" update. The matrix comparisons have been completed and the trace-back process can begin.

A C G T
A G T T

Figure 10. Alignment resulting from sequences "ACGT" and "AGTT" with gap cost of 10 and nucleotide change cost of 1.

The final step is the trace-back. During the wavefront cell updates, not only the new values of the cells were noted but also the lowest cost path (direction) to those cells. In the trace-back phase, the passage through the matrix is reversed, starting at the lower right corner. The algorithm moves back up and left through the matrix, following the lowest cost path notations previously laid down. Each time the trace-back moves directly up or left, gaps are inserted in sequences I or II, respectively. Each diagonal move is a correspondence between bases. The trace-back continues until cell (0, 0) is reached and the alignment is complete (Fig. 10). In this case there were two base mismatches and no gaps required to align the sequence with a cost of 2.

Extension to multiple sequences

The two sequence or pair-wise NW algorithm can be generalized into an N-sequence algorithm quite easily. Without going into too much detail, instead of a plane matrix of two dimensions, the algorithm would encounter a cube for three sequences and higher dimension alignment spaces as the data set grows richer. Instead of requiring approximately n^2 matrix elements, n^m would be required (m sequences of length n). A data set with 100 sequences of length 1000 (which is certainly reasonable in today's empirical climate) would require (in the worst case) 10^{300} elements. When updating cells, the two-dimensional case examines three other cells. For three sequences this number jumps to seven, for four sequences 15 must be examined to account for all the possible combinations of gaps, matches, and mismatches. Generally, $2^m - 1$ cells would

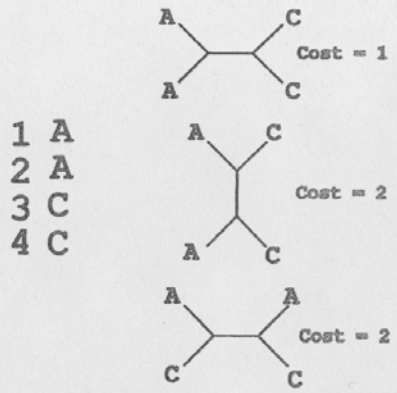


Figure 11. The impact of phylogenetic topology on alignment cost. The cost of an aligned position (left) can vary with the phylogenetic arrangement of the taxa (right).

need to be considered (m sequences) or for 100 sequences approximately 10^{30} cells would be taken into account for each of the 10^{300} elements. This is not going to happen easily.

The generalization of the cost function also presents a challenge. Whereas a change from A to C or a gap is a simple cost when only two or three sequences are examined, at data sets of four or more sequences the problem of phylogeny becomes necessarily interwoven. Four taxa may be arranged phylogenetically in three different ways (if rooting in unimportant). The distribution of bases and gaps at an aligned position can have different implied cost on different topologies (Fig. 11). Furthermore, aligned positions need not agree in their minimal cost topology favoring different taxonomic arrangements. For these reasons, Sankoff and Cedergren (1983) proposed using the principle of parsimony to determine multiple (>2 sequence) alignments. Sankoff requires, however, that the scheme of relationships be "known" or at least established *a priori*. Rarely, if ever, is this the case. In most cases, the reason for aligning the sequences in the first place is to determine these relationships. Even if there is some prior knowledge of relationships, systematic inference is always a provisional operation, awaiting more data and revision. The inability to specify evolutionary topologies before alignment can be ameliorated by performing the alignment repeatedly for the different possible topologies, but since the number of these increases combinatorially (# topologies = $\Pi[2i - 5]$; where i goes from 4 to the number of taxa and the trees are bifurcating and unrooted), this would be an exhausting procedure.

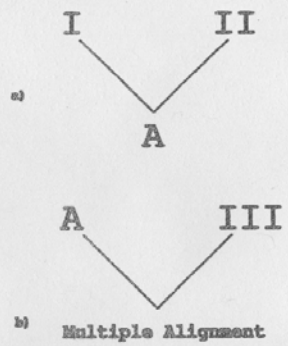


Figure 12. Tree based alignment of three sequences. Sequences I and II are aligned first via pair-wise Needleman and Wunsch (1970) forming alignment A. The alignment A and sequence III are then aligned, also via pairwise alignment, to form the multiple alignment of three sequences.

Due to these computational problems, a great diversity of heuristic methods has been proposed to achieve reasonable solutions given manageable resources. The first simplification is the reliance on pair-wise (two-dimensional) alignments. To avoid the huge complexities of true simultaneous multiple alignment an alignment "tree" is specified to allow the ordered accumulation of aligned sequences into a multiple alignment. Sequence I is aligned with sequence II and the resulting alignment with sequence III, and so on (Fig. 12). The methods differ in how this "tree" is determined and in how or whether the tree interacts with the alignment in any way.

Feng and Doolittle (1987, 1990; adapted by Higgins and Sharp, 1988, 1989) start with pairwise alignments. A distance measure is calculated from these pairwise alignments and a Fitch-Margoliash (Fitch and Margoliash, 1967) tree determined for these distances. The phylogenetic topology yielded by the Fitch-Margoliash analysis determines the order of alignment (Fig. 13). This process is taken one step further by Konings et al. (1987) and Hein (1989, 1990). In these procedures, the alignment resulting from the first round is cranked through again. This time, the distance tree is derived from the multiple alignment of the first pass. The realignments are repeated until the alignment order and derived phlogeny converge.

Absent from these methods is any criterion of optimality. The algorithms simply produce a result (Feng and Doolittle, 1987, 1990) or have a stopping rule. Furthermore, each of these methods relies at some point on distances, hence they inherit all the shortcomings of that form of analysis (Farris, 1981, 1985, 1986). Hein (1987, 1990) adds a parsi-

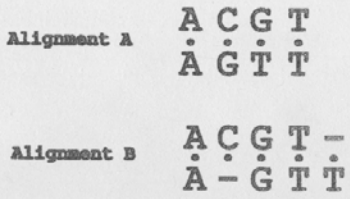


Figure 13. Alignments of two taxa based on different gap costs. Alignment A results when gaps are more costly that changes, while alignment B would result if nucleotide changes are more expensive than gaps.

mony step (with tree rearrangements) to test the convergence of phlogeny and alignment topolgy (as opposed to Konings et al., 1987), but maintains the topology = alignment stopping rule. Mindell (1991), though avoiding distance argumentation, supports this notion of phlogenetic determination of alignment order (in this case from pre-existing analyses). The necessity of this convergence is questionable (why should optimal pair-wise alignment order and phylogeny be exactly coincident?) and at best provides an endpoint for these algorithms.

Criterion of optimality

Justification of parsimony

Since sequence alignment is basically a procedure by which we can recognize and describe potential homology among nucleotide (or amino acid) positions, a logical means to assess the quality of these statements is the number of steps (or evolutionary events) required to explain or span the observed variation among the sequences. The minimum number of steps or changes required by an alignment is the most parsimonious branching diagram for these sequences. Since this is the same criterion used to determine the relative merits of the cladograms derived from these sequences, it seems only logical to extend this criterion to the alignments themselves. *The "best" alignment is that which yields the most parsimonious cladogram.* This is not a statement of reality, but one of support. The hypothesis which satisfies Occam's razor requires the fewest explanatory gymnastics. Hence, the alignment which yields the most parsimonious cladogram is the best supported (there may, of course be several such alignments).

This simple and consistent picture is complicated by the necessity of gap cots when sequences do not line up precisely. If a cost is not assigned to the insertion of gaps, a trivial (cost equal to zero) alignment will result, with one sequence having gaps at each position where the other has nucleotides. Additionally, to create an entirely logical and consistent parsimony-based alignment, the same cost function which is used in the Needleman-Wunsch prodedure must be used in the construction of cladograms from these alignments. Gaps (with the exception of those which result from the mechanics of sequencing) are then a fifth state (after A,C,G,T/U). The cost of transformation between a gap and the other states in cladogram construction is determined by the parameters of the alignment.

The MALIGN program

Software now exist to automatically align molecular sequences (Feng and Doolittle, 1987, 1990; CLUSTALL-Higgins and Sharp, 1988; Hein, 1989, 1990; MALIGN, Wheeler and Gladstein, 1993) so there is really no longer a need for "eye-ball" alignments. Through the use of these programs, the factors which affect alignment correspondences can be explored. Gap costs can be varied and multiple alignment solutions can be generated, at least with some of them. This chapter will use MALIGN (Wheeler and Gladstein, 1993; Wheeler and Gladstein, in press) to explore and demonstrate these effects. MALIGN offers a heuristic multiple alignment solution based on sequential pair-wise NW. That is, the program aligns the sequences on some "tree" where a two-dimensional alignment is performed at each node. For n taxa, n – 1 pair-wise alignments are performed. MALIGN will manipulate the alignment "tree" order through the types of operation found in phylogeny reconstruction programs (Farris, 1988; Swofford, 1993). In this way, the program can generate many (or effectively all) multiple alignments and save the single or multiple "best." MALIGN defines the "best" or optimal alignment to be that which generates the most parsimonious cladogram, whatever the topology. Hence, the alignment "tree" or trees need not conform to the evolutionary tree implied by the alignment. Facilities exist in the program (as in others) to vary alignment parameters such as gap cost and explore the generation of non-unique optimal alignments.

Sources of non-unique alignments

Parameter variation

The most important factor in determining an alignment is the cost function used to assess its quality. This model may contain any number of factors (parameters) which relate to DNA (or protein) change. Perhaps the most prominent of these is the gap cost. As described

above, this value describes the cost of the insertion or deletion of a base within the sequence. The gap cost is most frequently expressed in relation to the cost of a base change, hence the term gap cost ratio. When gap costs are low in relation to base transformations, the alignments tend to be somewhat diffuse, with many gaps and relatively few base transformations. When gaps are costly, on the other hand, the alignments are tighter, shorter, and involve much more base change. Obviously, these two situations can have vastly different phylogenetic implications. Differentiation can be made among types of gaps as well. Gaps can have lower costs if they are leading or trailing, or occur in contiguous strings. The number of gaps of any lenght can be minimized as well as gaps which violate reading frames.

As an example of the effects of gap cost, the two simple sequences aligned above, "ACGT" and "AGTT" can be aligned under different cost regimes (Fig. 13). When gaps are more expensive than base changes, alignment A is favored, while B is chosen when gaps are cheap. Interestingly, when gaps and changes are assigned equal cost, alignments A and B are equally costly as well.

Another factor which may affect alignment is the model of base change specified in the cost. All base transformations may be treated as equally informative or special weight may be given to some over others. One case of this would be the specification of transition-transversion ratios (Fig. 14). Alignment A implies (along with two gaps) one transition (C⟨->T) and one transversion (G⟨->T), whereas alignment B implies two transversions (G⟨->G and G⟨->T).

Other less commonly employed features include codon based weighting and secondary structure considerations. Specific to protein coding regions, coding weighting comes in two forms. The first of which limits, or at least heavily biases alignment gaps to groups of three. Other lengths would presumably violate reading frames. A second codon effect could come in the weighting of certain rare or unique codons which may reflect higher order protein structure or function. The correspondences

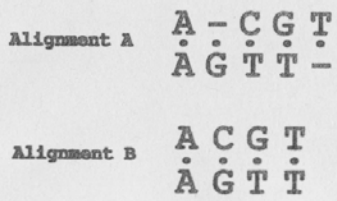


Figure 14. Alignments of two taxa showing different numbers of transitions and transversions. Alignment A implies two transversions while alignment B implies one transition, one transversion, and two gaps.

among cysteine residues are frequently given high weight in protein alignments.

Secondary structure models can function analogously to codon weights for RNA molecules. Bases changes or insertion which are non-compensatory may affect the secondary and higher order structure of the mature molecule. Alignment taking hairpin loop and fold structures may yield different results than those based solely on base-to-base comparison (assuming the secondary structure is known).

Multiple equally parsimonious alignments

When computers were first employed to find least cost cladograms, in addition to the novel (and sometimes disturbing) nature of the new solution, it was found that there were often multiple solutions. Even relatively "clean" data can generate thousands of phylogenetic solutions. It should come as no surprise, then, that this alsooccurs with automated alignment. As shown by Gatesy et al. (1993) and Wheeler et al. (in press), the same gap-to-change ratio can generate non-unique solutions. Each one of these solutions can (but not necessarily will) imply different phylogenies for the taxa.

In the same, simple example above, "ACGT" and "AGTT," when gaps and base changes are equally costly (Fig. 13), alignments A and B are equally costly. When multiple alignments are created through sequential pair-wise alignments, several alignment "trees" may yield unique and equally costly alignments. This problem is not limited to heuristic solutions, however. In the Shankoff procedure, the tree used to determine alignment costs may be unknown. When this is the case, several phylogenetic topologies may yield equally costly alignments.

Multiple equally parsimonious "paths"

As described in the section on the NW algorithm, a cell in a pairwise alignment can be reached from three other cells. The choice is based on the cost of each event, a mismatch or gaps in either of the two sequences. In many cases, two or all three may be equally costly events. This means that several alignments are equally costly at that region. The number of possible, equally costly alignments can be huge since they would multiply at each fork. This situation also pertains to true n-dimensional alignments since they also require path decision based on costs from abutting cells (in the general case there are $2^m - 1$ of them for m sequences).

Although there may be a huge number of these alignments, their divergence from the diagonal, perfect correspondence is bounded. If the

insertion of gaps is favored (when a gap or base change can take place with equal cost), then an ellipse which contains all possible alignments can be described and the "bounding" alignments produced. Within this ellipse, alignments which favor contiguous, or random choices exist.

The data of Whiting and Wheeler (1994) were subjected to this sort of path variation. In each case, the gap:change cost ratio was 2 to 1, that is gaps cost twice as much as base changes, Here, 23 800-base-pair insect rDNA sequence were aligned repeatedly. In each case, arbitrary decisions were made when equally costly paths through the matrix were found. In some cases, different types of gaps were favored, in others random choice was made. The ten alignments yielded different phylogenetic conclusions (Fig. 15). Although the different paths yield identical costs in the two-dimensional case, when these alignments are aligned to other sequences the results are not necessarily equally costly. This can be most readily seen in the placement of gaps. When gaps are favored over base changes when equally costly, the gaps in the multiple alignments tend to line up better than otherwise (Fig. 16), and yield more resolved cladograms. This may be due to the lack of "scattered" gaps when contiguous gaps are favored. Interestingly, although five different runs of "random" path were made, they each yielded the same phylogenetic conclusions.

Solutions

This chapter has presented some of the factors which can cause sequence alignment to yield ambiguous, non-unique, results. As more data sets are investigated more thoroughly, these multiple alignment solutions will press even more heavily upon us. What can be done? In recent years, much space has been devoted to dealing with ambiguity in phylogenetic analysis in systematics journals. Most methods have been proposed to either accommodate or choose among multiple equally optimal solutions. Methods will now be needed to deal with the same issue in sequence alignment.

The analysis of these multiple equally optimal alignments depends heavily on the interpretation given the alignments. Some seem to feel

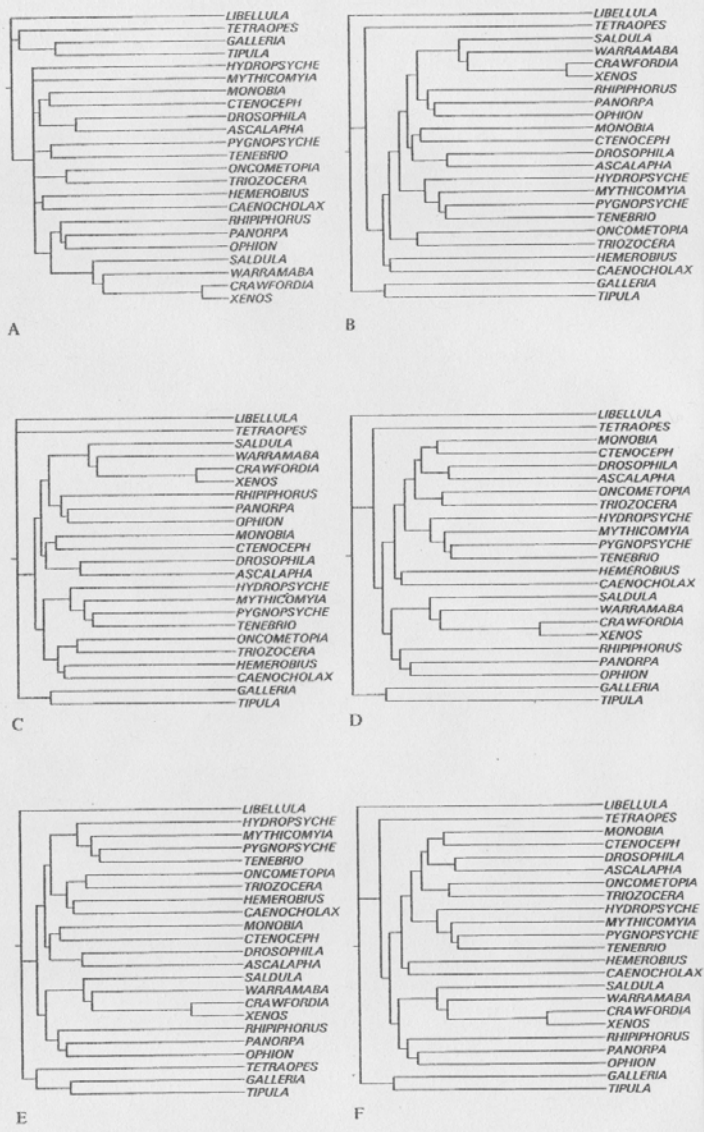
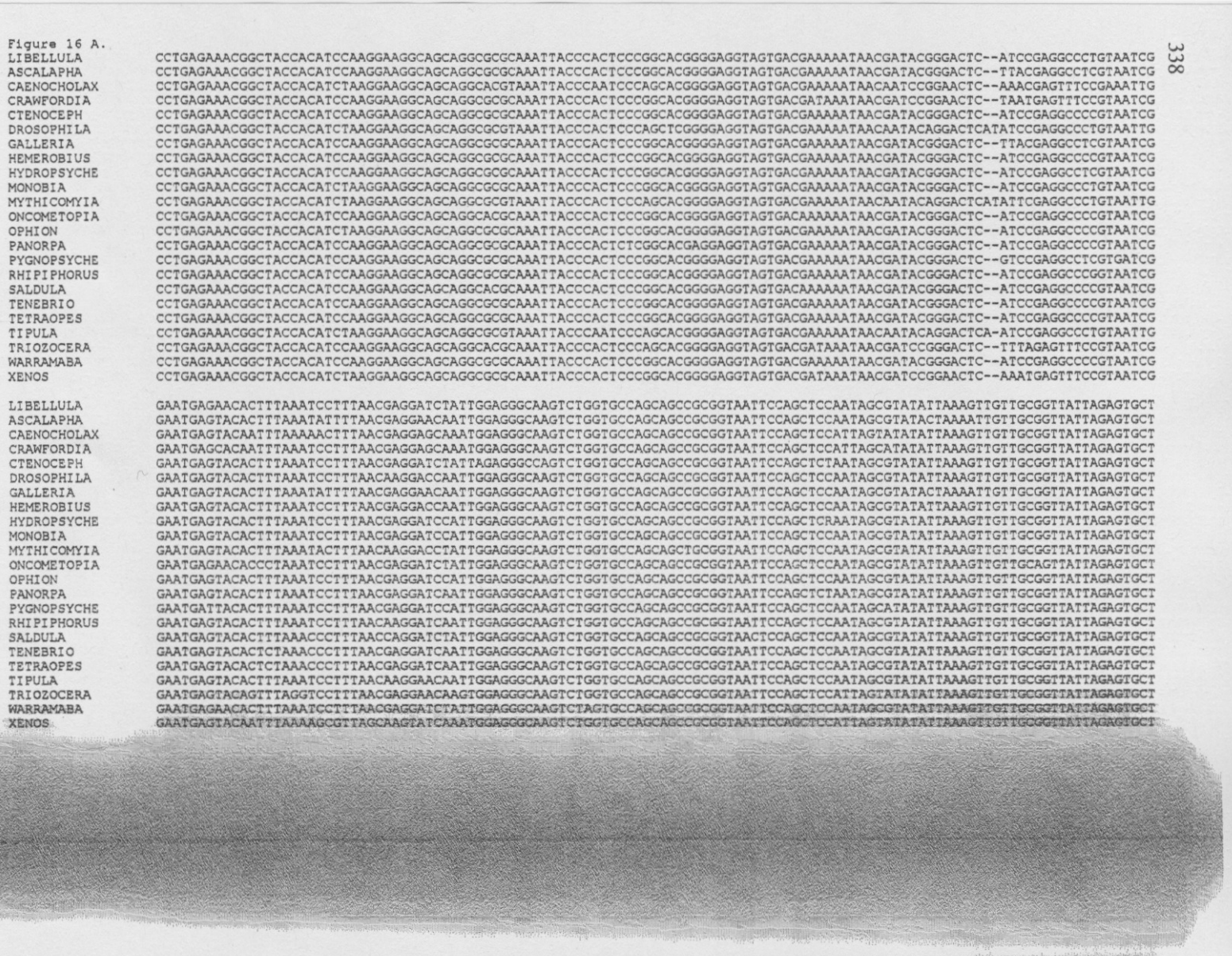


Figure 15. Phylogenetic results of alignments of insect 18S rDNA when different equally costly "paths" are favored. Phylogenetic reconstructions were performed using Farris' Hennig86 (Farris, 1988). Alignments favored: (A) matches (there were four unique alignments each of which yielded the same strict consenus cladogram), (B) contiguous gaps, (C) discontinuous gaps, (D) gaps in the Shorter sequence, (E) gaps in the longer sequence, (F) random choice among equally costly alternatives (this topology resulted from each of five alignments). In each case the gap cost was 8 and the change cost 4. (These arrangements are not meant to be definitive, and some of them are absurd. They are merly shown to demonstrate the variety of phylogenetic conclusions which can occur from different alignments.)

Figure 16 A.



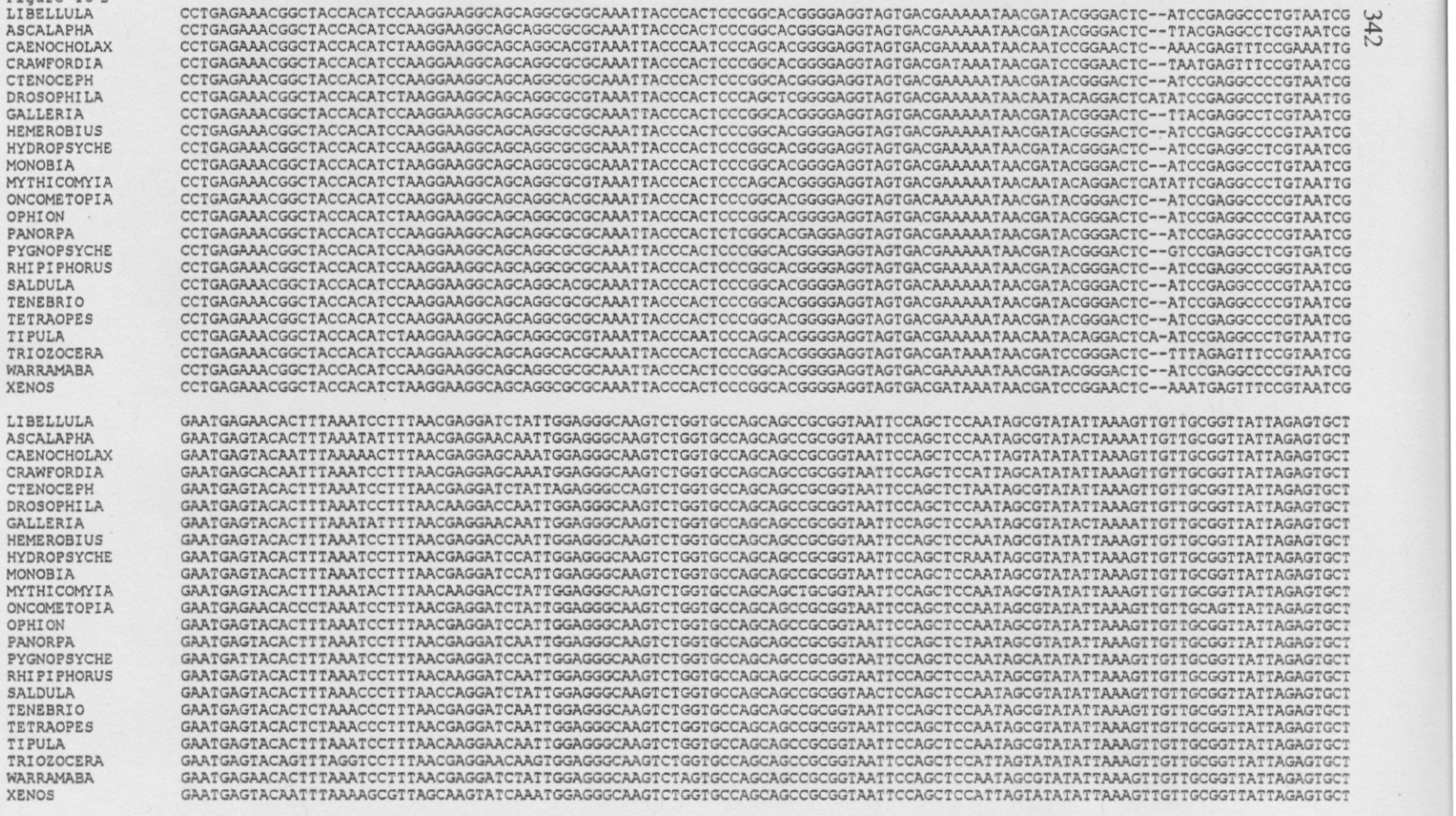
```
LIBELLULA       GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCCAGCGATCCGCCGAAGTTC-CTCCGA--TGACTCGGCGGGCAGCTT-CC-GG-AACC--A-A----------
ASCALAPHA       GAAGGCGATTAGATACCCGCCTAGTTCTAACCGTAAATNATGTCATCTAGCGATCCGCCGACGTTA-CTACAA--TGGCTCGGCGGGCAGCTT-CCGGGAAACCAAAGATTTTGGTCTCC
CAENOCHOLAX     GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACCAGCGATCCGYCGATGTTCATTYAAA-TTGACTCGACGGGCAGCTT-CCGGGAAACCAAAGTTTTTGGGTTCC
CRAWFORDIA      GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACCAGCGATCCGTCGATGTTCATTTAAA-TTGACTCGACGGGCAGCTT-CCGGGAAACCAAAGTTTTTGGGTTCC
CTENOCEPH       GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTC-CTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCC
DROSOPHILA      GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCAATTGGGTGTAGCTA-CTTTTA--TGGCTCTCTCAGTCGCTT-CCGGGAAACCAAAGCTTTTGGGCTCC
GALLERIA        GAAGGCGATTAGATACCGCCCTAGTTCTAACCGTAAATNATGTCATCTAGCGATCCGCCGACGTTA-CTACAA--TGGCTCGGCGGGCAGCTT-CCGGGAAACCAAAGATTTTGGTCTCC
HEMEROBIUS      GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTC-CTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCGTTTGGGTTCC
HYDROPSYCHE     GAAGGCGATTAGATACCGCTGTAGTTCTAACCATAAATGATGCCAGCTAGCGATCCGCCGAAGTTC-CTTCGATTNNNNNNGGTGGGCAGCTC-CCGGGAAACCAAAGCTTTCGGGCTCC
MONOBIA         GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTC-CTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCC
MYTHICOMYIA     GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCAATTGGGTGTAGCTA-CTTTTA--TGGCTTTCTCAGTCGCTTCCCGGGAAACCAAAGCTTTTGGGCTCC
ONCOMETOPIA     GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCCAGCGATCCGCCGAAGTTC-CTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCC
OPHION          GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTC-CTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCC
PANORPA         GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCAAAGTTC-CTCCGA--TGACTTGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCC
PYGNOPSYCHE     GAAGGCGATTAGATACCGCTGTAGTTCTAACCATAAATGATGCCAGCCAGCGATCCGCCGAAGTTT-CTTAATTTTAACCCGGTGGGCAGCTC-CCGGGAAACCAAAGCTTTCGGGCTCC
RHIPIPHORUS     GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGACGTTC-CTCATA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCC
SALDULA         GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGTCAGCCAGCGATCCGCCGACGTTC-ATTGAA--TGGCTCGGCGGGCAGCTT-CCCGGAAACGAAAGCTTTCGGGTTCC
TENEBRIO        GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGACGTTC-CTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCC
TETRAOPES       GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGACGTTC-CTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCC
TIPULA          GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCAATTGGATGGAGCTA-CTTATA--TGGCTCTTTCAGTCGCTTTCCGGGAAACCAAAGCTTTTGGGCTCC
TRIOZOCERA      GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACTAGCGATCCGTCGATGTTCATTTAAA-TTGACTCGACGGGCAGCTT-CCGGGAAACCAAAGTTTTTGGGTTCC
WARRAMABA       GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCCAGCGATCCGTCGACGTTC-CTTTTA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCC
XENOS           GAAGGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACCAGCGATCCGTCGATCTTCATTCAAA-TTGACTCGACGGGCAGCTT-CCGGGAAACCAAAGTTTTTGGGTTCC

LIBELLULA       ------AG-ATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCT-C--CTTAATTTGACTCAACACGGGAAACCTTACCAGGCCCAGACACC-GA
ASCALAPHA       GGGGG-AGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGGA
CAENOCHOLAX     GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGTACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGTCAGGACACTGAT
CRAWFORDIA      GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGAA
CTENOCEPH       GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACTACAAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGTCCGGACACCGGA
DROSOPHILA      GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTTACCAGGT-CGAACATAAGT
GALLERIA        GGGGG-AGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGGA
HEMEROBIUS      GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCAGACACCGGA
HYDROPSYCHE     GGGGGCAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTNACCAGGCCCGGACACTGGA
MONOBIA         GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGA
MYTHICOMYIA     GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTTACCAGGTCCGAACATAAAT
ONCOMETOPIA     GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACCTCACCAGGCCCGGACACCGGA
OPHION          GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGA
PANORPA         GGGGGAAGTATGGTTGCAAAGTTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACTGGA
PYGNOPSYCHE     GGGGGCAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTCACCAGGCCCGGACACTGGA
RHIPIPHORUS     GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGC-GCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCAGACACCGGA
SALDULA         GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACATTGGA
TENEBRIO        GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGA
TETRAOPES       GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGA
TIPULA          GGGGGAAGTATGGTTGCAAAGCTTAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGGAAACTTACCAGGTCCGAACATAAAT
TRIOZOCERA       GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGAA
WARRAMABA       GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGA
XENOS           GGGGGAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGAA
```

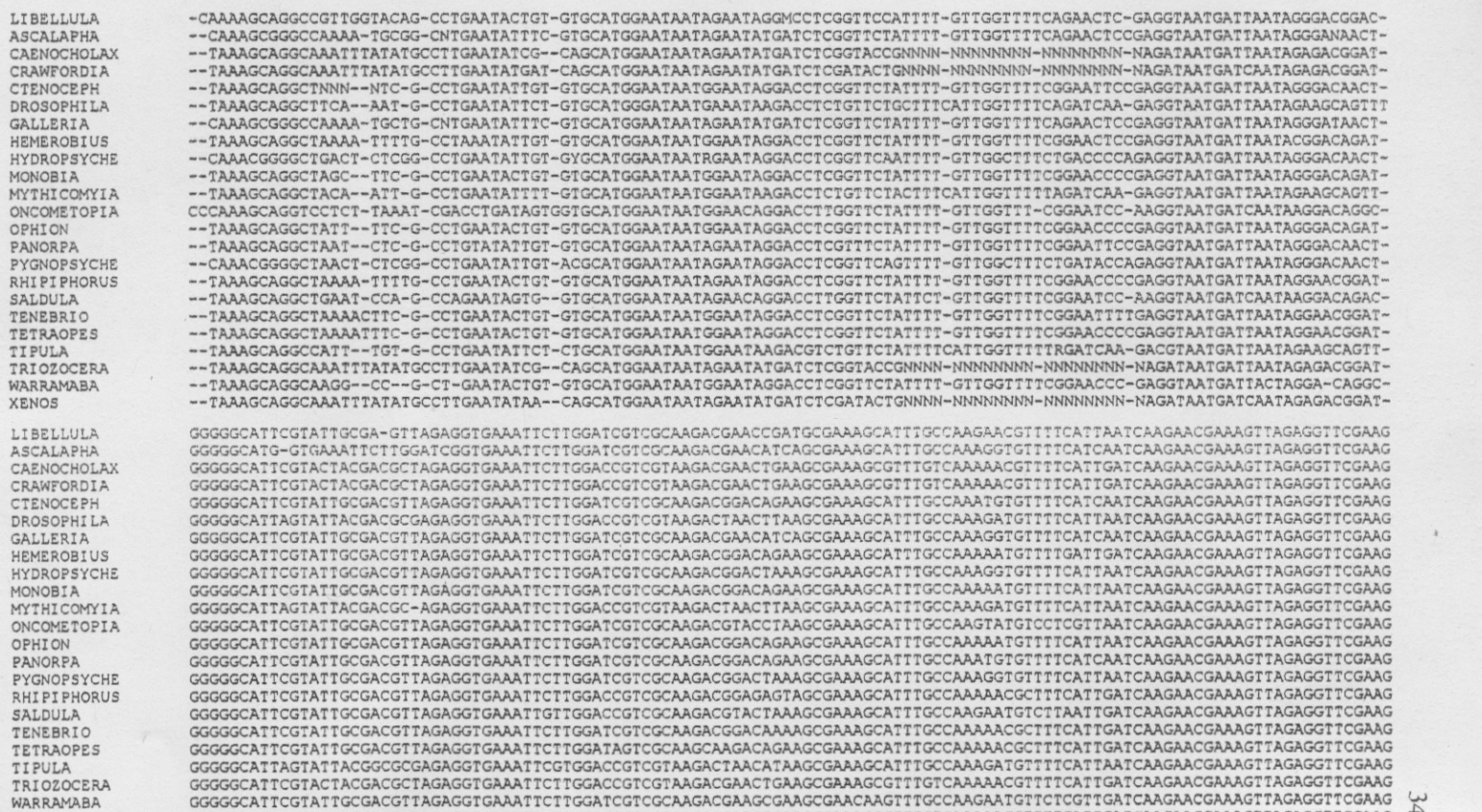
```
LIBELLULA       A-GATTGAC-GATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGG-GCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
ASCALAPHA       AGGATTGACAGATTAACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
CAENOCHOLAX     AGGATTGACAGATTGATAGCTCTTTCTTGATTCAGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
CRAWFORDIA      AGGATTGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
CTENOCEPH       AGGATTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
DROSOPHILA      GTGTAAGACAGATTGATAGCTCTTTCTCGAATCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTCGTGGAGTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
GALLERIA        AGGATTGACAGATTAACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
HEMEROBIUS      AGGATTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
HYDROPSYCHE     AGGATTGACAGATTGACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGYGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
MONOBIA         AGGATTGACAGATTGATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
MYTHICOMYIA     GAGTAAGACAGATTGATAGCTCTTTCTCGAATCTATGGGTGGTGGTGCATGGCCGTTGTTAGTTCGTGGAGTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
ONCOMETOPIA     AGGATTGACAGATCGATAGCTCTTTCTTGATTCAGTGGGTAGTGGTGCATGGCCGTTCTTAGTTGGTGGACTGATTTGTCTGGTTTATTCCGATAACGAACGAGACTC
OPHION          AGGATTGACAGATTGACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
PANORPA         AGGATTGACAGATTGAGAGCTCTTTCTTGATTCAGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
PYGNOPSYCHE     AGGATTGACGGATTAACAGCTCTTTCTCGATTCGGTGGGTGGTGGTGCATGGCCGCTCTTAGTTGGTGGAGCGATTTGTCAGGTTAATTCCGATAACGAACGAGACTC
RHIPIPHORUS     AGGATTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
SALDULA         AGGATTGACAGATTGATAGCTCTTTCTTGATTCAGTGGGTAGTGGTGCATGGCCGTTCTTAGTTGGTGGAACGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TENEBRIO        AGGATTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TETRAOPES       AGGATTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TIPULA          GAGTAAGACAGATTAATAGCTCTTTCTCGAATCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTCGTGGAGTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TRIOZOCERA      AGGATTGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
WARRAMABA       AGGATTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGACTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
XENOS           AGGATTGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
```

Figure 16. (A)

Figure 16 B


LIBELLULA CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCTGTAATCG
ASCALAPHA CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--TTACGAGGCCTCGTAATCG
CAENOCHOLAX CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCACGTAAATTACCCAATCCCAGCACGGGGAGGTAGTGACGAAAAATAACAATCCGGAACTC--AAACGAGTTTCCGAAATTG
CRAWFORDIA CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGATAAATAACGATCCGGAACTC--TAATGAGTTTCCGTAATCG
CTENOCEPH CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
DROSOPHILA CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGTAAATTACCCACTCCCAGCTCGGGGAGGTAGTGACGAAAAATAACAATACAGGACTCATATCCGAGGCCCTGTAATTG
GALLERIA CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--TTACGAGGCCTCGTAATCG
HEMEROBIUS CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
HYDROPSYCHE CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCTCGTAATCG
MONOBIA CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCTGTAATCG
MYTHICOMYIA CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGTAAATTACCCACTCCCAGCACGGGGAGGTAGTGACGAAAAATAACAATACAGGACTCATATTCGAGGCCCTGTAATTG
ONCOMETOPIA CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCACGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACAAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
OPHION CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
PANORPA CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCTCGGCACGAGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
PYGNOPSYCHE CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--GTCCGAGGCCTCGTGATCG
RHIPIPHORUS CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCGGTAATCG
SALDULA CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCACGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACAAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
TENEBRIO CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
TETRAOPES CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
TIPULA CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGTAAATTACCCAATCCCAGCACGGGGAGGTAGTGACGAAAAATAACAATACAGGACTCA-ATCCGAGGCCCTGTAATTG
TRIOZOCERA CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCACGCAAATTACCCACTCCCAGCACGGGGAGGTAGTGACGATAAATAACGATCCGGGACTC--TTTAGAGTTTCCGTAATCG
WARRAMABA CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
XENOS CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGATAAATAACGATCCGGAACTC--AAATGAGTTTCCGTAATCG

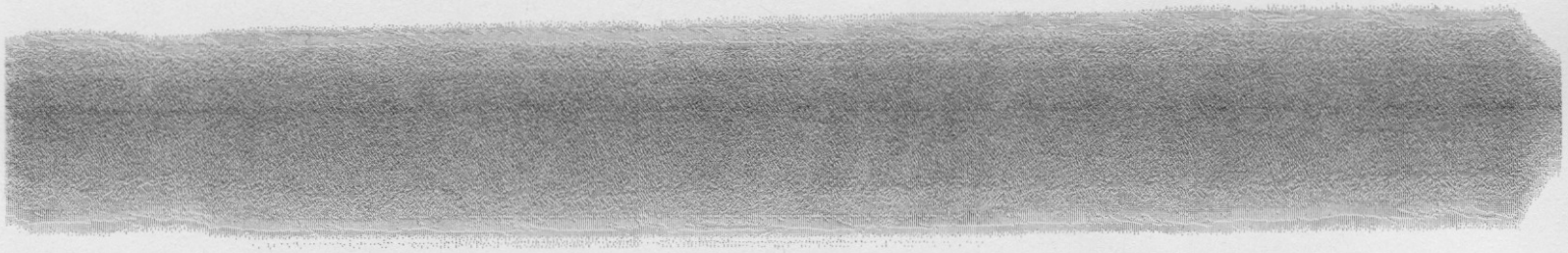
LIBELLULA GAATGAGAACACTTTAAATCCTTTAACGAGGATCTATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
ASCALAPHA GAATGAGTACACTTTAAATATTTTAACGAGGAACAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATACTAAAATTGTTGCGGTTATTAGAGTGCT
CAENOCHOLAX GAATGAGTACAATTTAAAAACTTTAACGAGGAGCAAATGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCATTAGTATATATTAAAGTTGTTGCGGTTATTAGAGTGCT
CRAWFORDIA GAATGAGCACAATTTAAATCCTTTAACGAGGAGCAAATGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCATTAGCATATATTAAAGTTGTTGCGGTTATTAGAGTGCT
CTENOCEPH GAATGAGTACACTTTAAATCCTTTAACGAGGATCTATTAGAGGGCCAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCTAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
DROSOPHILA GAATGAGTACACTTTAAATCCTTTAACGAAGGACCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
GALLERIA GAATGAGTACACTTTAAATATTTTAACGAGGAACAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATACTAAAATTGTTGCGGTTATTAGAGTGCT
HEMEROBIUS GAATGAGTACACTTTAAATCCTTTAACGAGGACCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
HYDROPSYCHE GAATGAGTACACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCRAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
MONOBIA GAATGAGTACACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
MYTHICOMYIA GAATGAGTACACTTTAAATACTTTAACAAGGACCTATTGGAGGGCAAGTCTGGTGCCAGCAGCTGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
ONCOMETOPIA GAATGAGAACACCCTAAATCCTTTAACGAGGATCTATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTATTAGAGTGCT
OPHION GAATGAGTACACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
PANORPA GAATGAGTACACTTTAAATCCTTTAACGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCTAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
PYGNOPSYCHE GAATGATTACACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCATATATTAAAGTTGTTGCGGTTATTAGAGTGCT
RHIPIPHORUS GAATGAGTACACTTTAAATCCTTTAACAAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
SALDULA GAATGAGTACACTTTAAACCCTTTAACCAGGATCTATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAACTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
TENEBRIO GAATGAGTACACTCTAAACCCTTTAACGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
TETRAOPES GAATGAGTACACTCTAAACCCTTTAACGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
TIPULA GAATGAGTACACTTTAAATCCTTTAACAAGGAACAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
TRIOZOCERA GAATGAGTACAGTTTAGGTCCTTTAACGAGGAACAAGTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCATTAGTATATATTAAAGTTGTTGCGGTTATTAGAGTGCT
WARRAMABA GAATGAGAACACTTTAAATCCTTTAACGAGGATCTATTGGAGGGCAAGTCTAGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
XENOS GAATGAGTACAATTTAAAAGCGTTAGCAAGTATCAAATGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCATTAGTATATATTAAAGTTGTTGCGGTTATTAGAGTGCT


LIBELLULA -CAAAAGCAGGCCGTTGGTACAG-CCTGAATACTGT-GTGCATGGAATAATAGAATAGGMCCTCGGTTCCATTTT-GTTGGTTTTCAGAACTC-GAGGTAATGATTAATAGGGACGGAC-
ASCALAPHA --CAAAGCGGGCCAAAA-TGCGG-CNTGAATATTTC-GTGCATGGAATAATAGAATATGATCTCGGTTCTATTTT-GTTGGTTTTCAGAACTCCGAGGTAATGATTAATAGGGANAACT-
CAENOCHOLAX --TAAAGCAGGCAAATTTATATGCCTTGAATATCG--CAGCATGGAATAATAGAATATGATCTCGGTACCGNNNN-NNNNNNNN-NNNNNNNN-NAGATAATGATTAATAGAGACGGAT-
CRAWFORDIA --TAAAGCAGGCAAATTTATATGCCTTGAATATGAT-CAGCATGGAATAATAGAATATGATCTCGATACTGNNNN-NNNNNNNN-NNNNNNNN-NAGATAATGATCAATAGAGACGGAT-
CTENOCEPH --TAAAGCAGGCTNNN--NTC-G-CCTGAATATTGT-GTGCATGGAATAATGGAATAGGACCTCGGTTCTATTTT-GTTGGTTTTCGGAATTCCGAGGTAATGATTAATAGGGACAACT-
DROSOPHILA --TAAAGCAGGCTTCA--AAT-G-CCTGAATATTCT-GTGCATGGGATAATGAAATAAGACCTCTGTTCTGCTTTCATTGGTTTTCAGATCAA-GAGGTAATGATTAATAGAAGCAGTTT
GALLERIA --CAAAGCGGGCCAAAA-TGCTG-CNTGAATATTTC-GTGCATGGAATAATAGAATATGATCTCGGTTCTATTTT-GTTGGTTTTCAGAACTCCGAGGTAATGATTAATAGGGATAACT-
HEMEROBIUS --TAAAGCAGGCTAAAA-TTTTG-CCTAAATATTGT-GTGCATGGAATAATGGAATAGGACCTCGGTTCTATTTT-GTTGGTTTTCGGAACTCCGAGGTAATGATTAATACGGACAGAT-
HYDROPSYCHE --CAAACGGGGCTGACT-CTCGG-CCTGAATATTGT-GYGCATGGAATAATRGAATAGGACCTCGGTTCAATTTT-GTTGGCTTTCTGACCCCAGAGGTAATGATTAATAGGGACAACT-
MONOBIA --TAAAGCAGGCTAGC--TTC-G-CCTGAATACTGT-GTGCATGGAATAATGGAATAGGACCTCGGTTCTATTTT-GTTGGTTTTCGGAACCCCGAGGTAATGATTAATAGGGACAGAT-
MYTHICOMYIA --TAAAGCAGGCTACA--ATT-G-CCTGAATATTTT-GTGCATGGAATAATGGAATAAGACCTCTGTTCTACTTTCATTGGTTTTTAGATCAA-GAGGTAATGATTAATAGAAGCAGTT-
ONCOMETOPIA CCCAAAGCAGGTCCTCT-TAAAT-CGACCTGATAGTGGTGCATGGAATAATGGAACAGGACCTTGGTTCTATTTT-GTTGGTTT-CGGAATCC-AAGGTAATGATCAATAAGGACAGGC-
OPHION --TAAAGCAGGCTATT--TTC-G-CCTGAATACTGT-GTGCATGGAATAATGGAATAGGACCTCGGTTCTATTTT-GTTGGTTTTCGGAACCCCGAGGTAATGATTAATAGGGACAGAT-
PANORPA --TAAAGCAGGCTAAT--CTC-G-CCTGTATATTGT-GTGCATGGAATAATAGAATAGGACCTCGTTTCTATTTT-GTTGGTTTTCGGAATTCCGAGGTAATGATTAATAGGGACAACT-
PYGNOPSYCHE --CAAACGGGGCTAACT-CTCGG-CCTGAATATTGT-ACGCATGGAATAATAGAATAGGACCTCGGTTCAGTTTT-GTTGGCTTTCTGATACCAGAGGTAATGATTAATAGGGACAACT-
RHIPIPHORUS --TAAAGCAGGCTAAAA-TTTTG-CCTGAATACTGT-GTGCATGGAATAATAGAATAGGACCTCGGTTCTATTTT-GTTGGTTTTCGGAACCCCGAGGTAATGATTAATAGGAACGGAT-
SALDULA --TAAAGCAGGCTGAAT-CCA-G-CCAGAATAGTG--GTGCATGGAATAATAGAACAGGACCTTGGTTCTATTCT-GTTGGTTTTCGGAATCC-AAGGTAATGATCAATAAGGACAGAC-
TENEBRIO --TAAAGCAGGCTAAAACTTC-G-CCTGAATACTGT-GTGCATGGAATAATGGAATAGGACCTCGGTTCTATTTT-GTTGGTTTTCGGAATTTTGAGGTAATGATTAATAGGAACGGAT-
TETRAOPES --TAAAGCAGGCTAAAATTTC-G-CCTGAATACTGT-GTGCATGGAATAATGGAATAGGACCTCGGTTCTATTTT-GTTGGTTTTCGGAACCCCGAGGTAATGATTAATAGGAACGGAT-
TIPULA --TAAAGCAGGCCATT--TGT-G-CCTGAATATTCT-CTGCATGGAATAATGGAATAAGACGTCTGTTCTATTTTCATTGGTTTTTRGATCAA-GACGTAATGATTAATAGAAGCAGTT-
TRIOZOCERA --TAAAGCAGGCAAATTTATATGCCTTGAATATCG--CAGCATGGAATAATAGAATATGATCTCGGTACCGNNNN-NNNNNNNN-NNNNNNNN-NAGATAATGATTAATAGAGACGGAT-
WARRAMABA --TAAAGCAGGCAAGG--CC--G-CT-GAATACTGT-GTGCATGGAATAATGGAATAGGACCTCGGTTCTATTTT-GTTGGTTTTCGGAACCC-GAGGTAATGATTACTAGGA-CAGGC-
XENOS --TAAAGCAGGCAAATTTATATGCCTTGAATATAA--CAGCATGGAATAATAGAATATGATCTCGATACTGNNNN-NNNNNNNN-NNNNNNNN-NAGATAATGATCAATAGAGACGGAT-

LIBELLULA GGGGGCATTCGTATTGCGA-GTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGAACCGATGCGAAAGCATTTGCCAAGAACGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAAG
ASCALAPHA GGGGGCATG-GTGAAATTCTTGGATCGGTGAAATTCTTGGACCGTCGCAAGACGAACATCAGCGAAAGCATTTGCCAAAGGTGTTTTCATCAATCAAGAACGAAAGTTAGAGGTTCGAAG
CAENOCHOLAX GGGGGCATTCGTACTACGACGCTAGAGGTGAAATTCTTGGACCGTCGTAAGACGAACTGAAGCGAAAGCGTTTGTCAAAAACGTTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAAG
CRAWFORDIA GGGGGCATTCGTACTACGACGCTAGAGGTGAAATTCTTGGACCGTCGTAAGACGAACTGAAGCGAAAGCGTTTGTCAAAAACGTTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAAG
CTENOCEPH GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGACCGTCGCGAAGACGGACAGAAGCGAAAGCATTTGCCAAATGTGTTTTCATCAATCAAGAACGAAAGTTAGAGGTTCGAAG
DROSOPHILA GGGGGCATTAGTATTACGACGCGAGAGGTGAAATTCTTGGACCGTCGTAAGACTAACTTAAGCGAAAGCATTTGCCAAAGATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAAG
GALLERIA GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGAACATCAGCGAAAGCATTTGCCAAAGGTGTTTTCATCAATCAAGAACGAAAGTTAGAGGTTCGAAG
HEMEROBIUS GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACTAAAGCGAAAGCATTTGCCAAAGGTGTTTTGATTGATCAAGAACGAAAGTTAGAGGTTCGAAG
HYDROPSYCHE GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACTAAAGCGAAAGCATTTGCCAAAGGTGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAAG
MONOBIA GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAGAAGCGAAAGCATTTGCCAAAAATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAAG
MYTHICOMYIA GGGGGCATTAGTATTACGACGC-AGAGGTGAAATTCTTGGACCGTCGTAAGACTAACTTAAGCGAAAGCATTTGCCAAAGATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAAG
ONCOMETOPIA GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGTACCTAAGCGAAAGCATTTGCCAAGTATGTCCTCGTTAATCAAGAACGAAAGTTAGAGGTTCGAAG
OPHION GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAGAAGCGAAAGCATTTGCCAAAAATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAAG
PANORPA GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAGAAGCGAAAGCATTTGCCAAAATGTTTTCATCAATCAAGAACGAAAGTTAGAGGTTCGAAG
PYGNOPSYCHE GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACTAAAGCGAAAGCATTTGCCAAAGGTGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAAG
RHIPIPHORUS GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGACCGTCGCAAGACGGAGAGTAGCGAAAGCATTTGCCAAAAACGCTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAAG
SALDULA GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTGTTGGACCGTCGCAAGACGTACTAAAGCGAAAGCATTTGCCAAGAATGTCTTAATTGATCAAGAACGAAAGTTAGAGGTTCGAAG
TENEBRIO GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAAAAGCGAAAGCATTTGCCAAAAACGCTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAAG
TETRAOPES GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATAGTCGCAAGCAAGACAGAAGCGAAAGCATTTGCCAAAAACGCTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAAG
TIPULA GGGGGCATTAGTATTACGGCGCGAGAGGTGAAATTCGTGGACCGTCGTAAGACTAACATAAGCGAAAGCATTTGCCAAAGATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAAG
TRIOZOCERA GGGGGCATTCGTACTACGACGCTAGAGGTGAAATTCTTGGACCGTCGTAAGACGAACTGAAGCGAAAGCGTTTGTCAAAAACGTTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAAG
WARRAMABA GGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGAAGCGAAGCGAACAAGTTTGCCAAGAACGTTTTCGTTGATCAAGAACGAAAGTTAGAGGTTCGAAG
XENOS GGGGGCATTCGTACTACGAGGCTAGAGGTGAAATTGTTGGACCGTCGTAAGACGAACTGAAGCGAAAGCGTTTGTCAAAAACGTTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAAG

```
LIBELLULA       GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCCAGCGATCCGCCGAAGTTCCTCCGATG-A-CTC-GGCGGGCAGCTTCCGG-AA-CCAAAG----------
ASCALAPHA       GCGATTAGATACCCGCCTAGTTCTAACCGTAAATNATGTCATCTAGCGATCCGCCGACGTTACTACAATG-G-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGATTTTGGTCTCCGGGGG
CAENOCHOLAX     GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACCAGCGATCCGYCGATGTTCATTYAAATTGACTC-GACGGGCAGCTTCCGGGAAACCAAAGTTTTTGGGTTCCGGGGG
CRAWFORDIA      GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACCAGCGATCCGTCGATGTTCATTTAAATTGACTC-GACGGGCAGCTTCCGGGAAACCAAAGTTTTTGGGTTCCGGGGG
CTENOCEPH       GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTCCTCCGATG-A-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGCTTTTGGGTTCCGGGGG
DROSOPHILA      GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCAATTGGGTGTAGCTACTTTTATG-G-CTCTCTCAGTCGCTT-CCGGGAAACCAAAGCTTTTGGGCTCCGGGGG
GALLERIA        GCGATTAGATACCGCCCTAGTTCTAACCGTAAATNATGTCATCTAGCGATCCGCCGACGTTACTACAATG-G-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGATTTTGGTCTCCGGGGG
HEMEROBIUS      GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTCCTCCGATG-A-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGCGTTTGGGTTCCGGGGG
HYDROPSYCHE     GCGATTAGATACCGCTGTAGTTCTAACCATAAATGATGCCAGCTAGCGATCCGCCGAAGTTCCTTCGATTNNNNNN-GGTGGGCAGCTCCCGGGAAACCAAAGCTTTCGGGCTCCGGGGG
MONOBIA         GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTCCTCCGATG-A-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGCTTTTGGGTTCCGGGGG
MYTHICOMYIA     GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCAATTGGGTGTAGCTACTTTTATG-G-CTTTCTCAGTCGCTTCCCGGGAAACCAAAGCTTTTGGGCTCCGGGGG
ONCOMETOPIA     GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCCAGCGATCCGCCGAAGTTCCTCCGATG-A-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGCTTTTGGGTTCCGGGGG
OPHION          GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTCCTCCGATG-A-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGCTTTTGGGTTCCGGGGG
PANORPA         GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCAAAGTTCCTCCGATG-A-CTT-GGCGGGCAGCTTCCGGGAAACCAAAGCTTTTGGGTTCCGGGGG
PYGNOPSYCHE     GCGATTAGATACCGCTGTAGTTCTAACCATAAATGATGCCAGCCAGCGATCCGCCGAAGTTTCTTAATTTTAACCC-GGTGGGCAGCTCCCGGGAAACCAAAGCTTTCGGGCTCCGGGGG
RHIPIPHORUS     GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGACGTTCCTCATATG-A-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGCTTTTGGGTTCCGGGGG
SALDULA         GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGTCAGCCAGCGATCCGCCGACGTTCATTGAATG-G-CTC-GGCGGGCAGCTTCCCGGAAACGAAAGCTTTCGGGTTCCGGGGG
TENEBRIO        GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGACGTTCCTCCGATG-A-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGCTTTTGGGTTCCGGGGG
TETRAOPES       GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGACGTTCCTCCGATG-A-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGCTTTTGGGTTCCGGGGG
TIPULA          GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCAATTGGATGGAGCTACTTATATG-G-CTCTTTCAGTCGCTTTCCGGGAAACCAAAGCTTTTGGGCTCCGGGGG
TRIOZOCERA      GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACTAGCGATCCGTCGATGTTCATTTAAATTGACTC-GACGGGCAGCTTCCGGGAAACCAAAGTTTTTGGGTTCCGGGGG
WARRAMABA       GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCCAGCGATCCGTCGACGTTCCTTTTATG-A-CTC-GGCGGGCAGCTTCCGGGAAACCAAAGCTTTTGGGTTCCGGGGG
XENOS           GCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACCAGCGATCCGTCGATCTTCATTCAAATTGACTC-GACGGGCAGCTTCCGGGAAACCAAAGTTTTTGGGTTCCGGGGG

LIBELLULA       A--T--GGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCT-C--CTTAATTTGACTCAACACGGGAAACCTTACCAGGCCCAGACACCG-AAG--A
ASCALAPHA       A-GTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGGAAG-GA
CAENOCHOLAX     AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGTACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGTCAGGACACTGATAG-GA
CRAWFORDIA      AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGTCAGGACACTGAAAG-GA
CTENOCEPH       AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACTACAAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGTCCGGACACCGGAAG-GA
DROSOPHILA      AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTTACCAGGT-CGAACATAA-GTGTGT
GALLERIA        A-GTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAG-GA
HEMEROBIUS      AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCAGACACCGGAAG-GA
HYDROPSYCHE     CAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTNACCAGGCCCAGACACYGGAAG-GA
MONOBIA         AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAG-GA
MYTHICOMYIA     AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTTACCAGGTCCGAACATAA-ATGAGT
ONCOMETOPIA     AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAG-GA
OPHION          AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAG-GA
PANORPA         AAGTATGGTTGCAAAGTTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACTGGAAG-GA
PYGNOPSYCHE     CAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTCGACTCAACACGGGAAAACTCACCAGGCCCGGACACTGGAAG-GA
RHIPIPHORUS     AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGC-GCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCAGACACCGGAAG-GA
SALDULA         AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACATTGGAAG-GA
TENEBRIO        AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAG-GA
TETRAOPES       AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAG-GA
TIPULA          AAGTATGGTTGCAAAGCTTAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGGAAACTTACCAGGTCCGAACATAA-ATGAGT
TRIOZOCERA      AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGAAAG-GA
WARRAMABA       AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAG-GA
XENOS           AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGAAAG-GA
```



```
LIBELLULA       TTGAC-GATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGG-GCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
ASCALAPHA       TTGACAGATTAACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
CAENOCHOLAX     TTGACAGATTGATAGCTCTTTCTTGATTCAGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
CRAWFORDIA      TTGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
CTENOCEPH       TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
DROSOPHILA      AAGACAGATTGATAGCTCTTTCTCGAATCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTCGTGGAGTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
GALLERIA        TTGACAGATTAACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
HEMEROBIUS      TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
HYDROPSYCHE     TTGACAGATTGACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGYGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
MONOBIA         TTGACAGATTGATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
MYTHICOMYIA     AAGACAGATTGATAGCTCTTTCTCGAATCTATGGGTGGTGGTGCATGGCCGTTGTTAGTTCGTGGAGTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
ONCOMETOPIA     TTGACAGATCGATAGCTCTTTCTTGATTCAGTGGGTAGTGGTGCATGGCCGTTCTTAGTTGGTGGACTGATTTGTCTGGTTTATTCCGATAACGAACGAGACTC
OPHION          TTGACAGATTGACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
PANORPA         TTGACAGATTGAGAGCTCTTTCTTGATTCAGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
PYGNOPSYCHE     TTGACGGATTAACAGCTCTTTCTCGATTCGGTGGGTGGTGGTGCATGGCCGCTCTTAGTTGGTGGAGCGATTTGTCAGGTTAATTCCGATAACGAACGAGACTC
RHIPIPHORUS     TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
SALDULA         TTGACAGATTGATAGCTCTTTCTTGATTCAGTGGGTAGTGGTGCATGGCCGTTCTTAGTTGGTGGAACGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TENEBRIO        TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TETRAOPES       TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TIPULA          AAGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTCGTGGAGTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TRIOZOCERA      TTGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
WARRAMABA       TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGACTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
XENOS           TTGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
```

Figure 16. (B)

Figure 16 C.

```
LIBELLULA     CGGGGGCATTCGTATTGCGA-GTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGAACCGATGCGAAAGCATTTGCCAAGAACGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAA
ASCALAPHA     TGGGGGCA-TGGTGAAATTCTTGGATCGGTGAAATTCTTGGATCGTCGCAAGACGAACATCAGCGAAAGCATTTGCCAAAGGTGTTTTCATCAATCAAGAACGAAAGTTAGAGGTTCGAA
CAENOCHOLAX   TGGGGGCATTCGTACTACGACGCTAGAGGTGAAATTCTTGGACCGTCGTAAGACGAACTGAAGCGAAAGCGTTTGTCAAAAACGTTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAA
CRAWFORDIA    TGGGGGCATTCGTACTACGACGCTAGAGGTGAAATTCTTGGACCGTCGTAAGACGAACTGAAGCGAAAGCGTTTGTCAAAAACGTTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAA
CTENOCEPH     TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAGAAGCGAAAGCATTTGCCAAATGTGTTTTCATCAATCAAGAACGAAAGTTAGAGGTTCGAA
DROSOPHILA    TGGGGGCATTAGTATTACGACGCGAGAGGTGAAATTCTTGGACCGTCGTAAGACTAACTTAAGCGAAAGCATTTGCCAAAGATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAA
GALLERIA      TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGAACATCAGCGAAAGCATTTGCCAAAGGTGTTTTCATCAATCAAGAACGAAAGTTAGAGGTTCGAA
HEMEROBIUS    TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAGAAGCGAAAGCATTTGCCAAAAATGTTTTGATTGATCAAGAACGAAAGTTAGAGGTTCGAA
HYDROPSYCHE   TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACTAAAGCGAAAGCATTTGCCAAAGGTGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAA
MONOBIA       TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAGAAGCGAAAGCATTTGCCAAAAATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAA
MYTHICOMYIA   TGGGGGCATTAGTATTACGACGC-AGAGGTGAAATTCTTGGACCGTCGTAAGACTAACTTAAGCGAAAGCATTTGCCAAAGATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAA
ONCOMETOPIA   CGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGTACCTAAGCGAAAGCATTTGCCAAGTATGTCCTCGTTAATCAAGAACGAAAGTTAGAGGTTCGAA
OPHION        TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAGAAGCGAAAGCATTTGCCAAAAATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAA
PANORPA       TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAGAAGCGAAAGCATTTGCCAAATGTGTTTTCATCAATCAAGAACGAAAGTTAGAGGTTCGAA
PYGNOPSYCHE   TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACTAAAGCGAAAGCATTTGCCAAAGGTGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAA
RHIPIPHORUS   TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGACCGTCGCAAGACGGAGAGTAGCGAAAGCATTTGCCAAAAACGCTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAA
SALDULA       CGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTGTTGGACCGTCGCAAGACGTACTAAAGCGAAAGCATTTGCCAAGAATGTCTTAATTGATCAAGAACGAAAGTTAGAGGTTCGAA
TENEBRIO      TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGGACAAAAGCGAAAGCATTTGCCAAAAACGCTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAA
TETRAOPES     TGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATAGTCGCAAGCAAGACAGAAGCGAAAGCATTTGCCAAAAACGCTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAA
TIPULA        TGGGGGCATTAGTATTACGGCGCGAGAGGTGAAATTCGTGGACCGTCGTAAGACTAACATAAGCGAAAGCATTTGCCAAAGATGTTTTCATTAATCAAGAACGAAAGTTAGAGGTTCGAA
TRIOZOCERA    TGGGGGCATTCGTACTACGACGCTAGAGGTGAAATTCTTGGACCGTCGTAAGACGAACTGAAGCGAAAGCGTTTGTCAAAAACGTTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAA
WARRAMABA     CGGGGGCATTCGTATTGCGACGTTAGAGGTGAAATTCTTGGATCGTCGCAAGACGAAGCGAAGCGAACAAGTTTGCCAAGAATGTTCTCGTTGATCAAGAACGAAAGTTAGAGGTTCGAA
XENOS         TGGGGGCATTCGTACTACGAGGCTAGAGGTGAAATTGTTGGACCGTCGTAAGACGAACTGAAGCGAAAGCGTTTGTCAAAAACGTTTTCATTGATCAAGAACGAAAGTTAGAGGTTCGAA
```

```
LIBELLULA     GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCCAGCGATCCGCCGAAGTTCCTCCGA--TGACTCGGCGGGCAGCTT-CC-GG-AACC--A----------------
ASCALAPHA     GGCGATTAGATACCCGCCTAGTTCTAACCGTAAATNATGTCATCTAGCGATCCGCCGACGTTACTACAA--TGGCTCGGCGGGCAGCTT-CCGGGAAACCAAAGATTTTGGTCTCCGGGG
CAENOCHOLAX   GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACCAGCGATCCGYCGATGTTCATTYAAATTGACTCGACGGGCAGCTT-CCGGGAAACCAAAGTTTTTGGGTTCCGGGG
CRAWFORDIA    GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACCAGCGATCCGTCGATGTTCATTTAAATTGACTCGACGGGCAGCTT-CCGGGAAACCAAAGTTTTTGGGTTCCGGGG
CTENOCEPH     GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTCCTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCCGGGG
DROSOPHILA    GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCAATTGGGTGTAGCTACTTTTA--TGGCTCTCTCAGTCGCTT-CCGGGAAACCAAAGCTTTTGGGCTCCGGGG
GALLERIA      GGCGATTAGATACCGCCCTAGTTCTAACCGTAAATNATGTCATCTAGCGATCCGCCGACGTTACTACAA--TGGCTCGGCGGGCAGCTT-CCGGGAAACCAAAGATTTTGGTCTCCGGGG
HEMEROBIUS    GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTCCTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCGTTTGGGTTCCGGGG
HYDROPSYCHE   GGCGATTAGATACCGCTGTAGTTCTAACCATAAATGATGCCAGCTAGCGATCCGCCGAAGTTCCTTCGATTNNNNNNGGTGGGCAGCTC-CCGGGAAACCAAAGCTTTCGGGCTCCGGGG
MONOBIA       GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTCCTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCCGGGG
MYTHICOMYIA   GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCAATTGGGTGTAGCTACTTTTA--TGGCTTTCTCAGTCGCTTCCCGGGAAACCAAAGCTTTTGGGCTCCGGGG
ONCOMETOPIA   GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCCAGCGATCCGCCGAAGTTCCTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCCGGGG
OPHION        GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGAAGTTCCTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCCGGGG
PANORPA       GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCAAAGTTCCTCCGA--TGACTTGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCCGGGG
PYGNOPSYCHE   GGCGATTAGATACCGCTGTAGTTCTAACCATAAATGATGCCAGCCAGCGATCCGCCGAAGTTTCTTAATTTTAACCCGGTGGGCAGCTC-CCGGGAAACCAAAGCTTTCGGGCTCCGGGG
RHIPIPHORUS   GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGACGTTCCTCATA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCCGGGG
SALDULA       GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGTCAGCCAGCGATCCGCCGACGTTCATTGAA--TGGCTCGGCGGGCAGCTT-CCCGGAAACGAAAGCTTTCGGGTTCCGGGG
TENEBRIO      GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGACGTTCCTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCCGGGG
TETRAOPES     GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCGATCCGCCGACGTTCCTCCGA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCCGGGG
TIPULA        GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCTAGCAATTGGATGGAGCTACTTATA--TGGCTCTTTCAGTCGCTTTCCGGGAAACCAAAGCTTTTGGGCTCCGGGG
TRIOZOCERA    GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACTAGCGATCCGTCGATGTTCATTTAAATTGACTCGACGGGCAGCTT-CCGGGAAACCAAAGTTTTTGGGTTCCGGGG
WARRAMABA     GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCAGCCAGCGATCCGTCGACGTTCCTTTTA--TGACTCGGCGGGCAGCTT-CCGGGAAACCAAAGCTTTTGGGTTCCGGGG
XENOS         GGCGATCAGATACCGCCCTAGTTCTAACCATAAACGATGCCGACCAGCGATCCGTCGATCTTCATTCAAATTGACTCGACGGGCAGCTT-CCGGGAAACCAAAGTTTTTGGGTTCCGGGG
```

```
LIBELLULA     -AAG-ATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCT-C--CTTAATTTGACTCAACACGGGAAACCTTACCAGGCCCAGACACC-GAA-GA
ASCALAPHA     G-AGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGGAAGGA
CAENOCHOLAX   GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGTACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGTCAGGACACTGATAGGA
CRAWFORDIA    GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGAAAGGA
CTENOCEPH     GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACTACAAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGTCCGGACACCGGAAGGA
DROSOPHILA    GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTTACCAGGT-CGAACATAAGTGTGT
GALLERIA      G-AGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGGAAGGA
HEMEROBIUS    GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCAGACACCGGAAGGA
HYDROPSYCHE   GCAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTNACCAGGCCCAGACACYGGAAGGA
MONOBIA       GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAGGA
MYTHICOMYIA   GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTTACCAGGTCCGAACATAAATGAGT
ONCOMETOPIA   GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAGGA
OPHION        GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAGGA
PANORPA       GAAGTATGGTTGCAAAGTTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACTGGAAGGA
PYGNOPSYCHE   GCAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTCGACTCAACACGGGAAAACTCACCAGGCCCGGACACTGGAAGGA
RHIPIPHORUS   GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGC-GCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCAGACACCGGAAGGA
SALDULA       GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACATTGGAAGGA
TENEBRIO      GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAGGA
TETRAOPES     GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAGGA
TIPULA        GAAGTATGGTTGCAAAGCTTAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGGAAACTTACCAGGTCCGAACATAAATGAGT
TRIOZOCERA    GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGAAAGGA
WARRAMABA     GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCAGGCCCGGACACCGGAAGGA
XENOS         GAAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTTAATTTGACTCAACACGGGAAATCTCACCAGGCCCGGACACCGAAAGGA
```

```
LIBELLULA     CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCTGTAATCG
ASCALAPHA     CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--TTACGAGGCCTCGTAATCG
CAENOCHOLAX   CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCACGTAAATTACCCAATCCCAGCACGGGGAGGTAGTGACGAAAAATAACAATCCGGAACTC--AAACGAGTTTCCGAAATTG
CRAWFORDIA    CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGATAAATAACGATCCGGAACTC--TAATGAGTTTCCGTAATCG
CTENOCEPH     CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
DROSOPHILA    CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGTAAATTACCCACTCCCAGCTCGGGGAGGTAGTGACGAAAAATAACAATACAGGACTCATATCCGAGGCCCTGTAATTG
GALLERIA      CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--TTACGAGGCCTCGTAATCG
HEMEROBIUS    CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
HYDROPSYCHE   CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCTCGTAATCG
MONOBIA       CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCTGTAATCG
MYTHICOMYIA   CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGTAAATTACCCACTCCCAGCACGGGGAGGTAGTGACGAAAAATAACAATACAGGACTCATATTCGAGGCCCTGTAATTG
ONCOMETOPIA   CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCACGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACAAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
OPHION        CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
PANORPA       CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCTCGGCACGAGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
PYGNOPSYCHE   CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--GTCCGAGGCCTCGTGATCG
RHIPIPHORUS   CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCGGTAATCG
SALDULA       CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCACGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACAAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
TENEBRIO      CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
TETRAOPES     CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
TIPULA        CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGTAAATTACCCAATCCCAGCACGGGGAGGTAGTGACGAAAAATAACAATACAGGACTCA-ATCCGAGGCCCTGTAATTG
TRIOZOCERA    CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCACGCAAATTACCCACTCCCAGCACGGGGAGGTAGTGACGATAAATAACGATCCGGGACTC--TTTAGAGTTTCCGTAATCG
WARRAMABA     CCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGAAAAATAACGATACGGGACTC--ATCCGAGGCCCCGTAATCG
XENOS         CCTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGCAAATTACCCACTCCCGGCACGGGGAGGTAGTGACGATAAATAACGATCCGGAACTC--AAATGAGTTTCCGTAATCG
```

```
LIBELLULA       GAATGAGAACACTTTAAATCCTTTAACGAGGATCTATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
ASCALAPHA       GAATGAGTACACTTTAAATATTTTAACGAGGAACAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATACTAAAATTGTTGCGGTTATTAGAGTGCT
CAENOCHOLAX     GAATGAGTACAATTTAAAAACTTTAACGAGGAGCAAATGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCATTAGTATATATTAAAGTTGTTGCGGTTATTAGAGTGCT
CRAWFORDIA      GAATGAGCACAATTTAAATCCTTTAACGAGGAGCAAATGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCATTAGCATATATTAAAGTTGTTGCGGTTATTAGAGTGCT
CTENOCEPH       GAATGAGTACACTTTAAATCCTTTAACGAGGATCTATTAGAGGGCCAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCTAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
DROSOPHILA      GAATGAGTACACTTTAAATCCTTTAACAAGGACCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
GALLERIA        GAATGAGTACACTTTAAATATTTTAACGAGGAACAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATACTAAAATTGTTGCGGTTATTAGAGTGCT
HEMEROBIUS      GAATGAGTACACTTTAAATCCTTTAACGAGGACCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
HYDROPSYCHE     GAATGAGTACACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCRAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
MONOBIA         GAATGAGTACACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
MYTHICOMYIA     GAATGAGTACACTTTAAATACTTTAACAAGGACCTATTGGAGGGCAAGTCTGGTGCCAGCAGCTGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
ONCOMETOPIA     GAATGAGAACACCCTAAATCCTTTAACGAGGATCTATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTATTAGAGTGCT
OPHION          GAATGAGTACACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
PANORPA         GAATGAGTACACTTTAAATCCTTTAACGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCTAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
PYGNOPSYCHE     GAATGATTACACTTTAAATCCTTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCATATATTAAAGTTGTTGCGGTTATTAGAGTGCT
RHIPIPHORUS     GAATGAGTACACTTTAAATCCTTTAACAAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
SALDULA         GAATGAGTACACTCTAAACCCTTTAACGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAACTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
TENEBRIO        GAATGAGTACACTCTAAACCCTTTAACGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
TETRAOPES       GAATGAGTACACTCTAAACCCTTTAACGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
TIPULA          GAATGAGTACACTTTAAATCCTTTAACAAGGAACAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
TRIOZOCERA      GAATGAGTACAGTTTAGGTCCTTTAACGAGGAACAAGTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCATTAGTATATATTAAAGTTGTTGCGGTTATTAGAGTGCT
WARRAMABA       GAATGAGAACACTTTAAATCCTTTAACGAGGATCTATTGGAGGGCAAGTCTAGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCGGTTATTAGAGTGCT
XENOS           GAATGAGTACAATTTAAAAGCGTTAGCAAGTATCAAATGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCATTAGTATATATTAAAGTTGTTGCGGTTATTAGAGTGCT

LIBELLULA       -CAAAAGCAGG-CCGTT-GGTACAGCC-TGAATACTGTGTGCATGGAATAATAGAATAGGMCCTCGGTTCCA-TTTTGTTGGTTTTCAGAA-CTCGAGGTAATGATTAATAGGGACGG-A
ASCALAPHA       --CAAAGCGGG--CCAA-AATGCGGCN-TGAATATTTCGTGCATGGAATAATAGAATATGATCTCGGTTCTA-TTTTGTTGGTTTTCAGAACTCCGAGGTAATGATTAATAGGGANAA-C
CAENOCHOLAX     --TAAAGCAGG--CAAATTTATATGCCTTGAATATCG-CAGCATGGAATAATAGAATATGATCTCGGTACCG-NNNNNNNNNNNNNNNNNNNNNAGA--TAATGATTAATAGAGACGG-A
CRAWFORDIA      --TAAAGCAGG--CAAATTTATATGCCTTGAATATGATCAGCATGGAATAATAGAATATGATCTCGATACTG-NNNNNNNNNNNNNNNNNNNNNAGA--TAATGATCAATAGAGACGG-A
CTENOCEPH       --TAAAGCAGG--CTNN-NNT-C-GCC-TGAATATTGTGTGCATGGAATAATGGAATAGGACCTCGGTTCTA-TTTTGTTGGTTTTCGGAATTCCGAGGTAATGATTAATAGGGACAA-C
DROSOPHILA      --TAAAGCAGG--CTTC-AAA-T-GCC-TGAATATTCTGTGCATGGGATAATGAAATAAGACCTCTGTTCTGCTTTCATTGGTTTTCAGAT-CAAGAGGTAATGATTAATAGGAAGCAGTT
GALLERIA        --CAAAGCGGG--CCAA-AATGCTGCN-TGAATATTTCGTGCATGGAATAATAGAATATGATCTCGGTTCTA-TTTTGTTGGTTTTCAGAACTCCGAGGTAATGATTAATAGGGATAA-C
HEMEROBIUS      --TAAAGCAGG--CTAA-AATTTTGCC-TAAATATTGTGTGCATGGAATAATGGAATAGGACCTCGGTTCTA-TTTTGTTGGTTTTCGGAACTCCGAGGTAATGATTAATACGGACAG-A
HYDROPSYCHE     --CAAACGGGG--CTGA-CTCTCGGCC-TGAATATTGTGYGCATGGAATAATRGAATAGGACCTCGGTTCAA-TTTTGTTGGCTTTCTGACCCCAGAGGTAATGATTAATAGGGACAA-C
MONOBIA         --TAAAGCAGG--CTAG-CTT-C-GCC-TGAATACTGTGTGCATGGAATAATGGAATAGGACCTCGGTTCTA-TTTTGTTGGTTTTCGGAACCCCGAGGTAATGATTAATAGGGACAG-A
MYTHICOMYIA     --TAAAGCAGG--CTAC-AAT-T-GCC-TGAATATTTTGTGCATGGAATAATGGAATAAGACCTCTGTTCTACTTTCATTGGTTTTTAGAT-CAAGAGGTAATGATTAATAGAAGCAG-T
ONCOMETOPIA     CCCAAAGCAGGTCCTCT-TAAATCGAC-CTGATAGTG-GTGCATGGAATAATGGAACAGGACCTTGGTTCTA-TTTTGTTGG-TTTCGGAA-TCCAAGGTAATGATCAATAAGGACAG-G
OPHION          --TAAAGCAGG--CTAT-TTT-C-GCC-TGAATACTGTGTGCATGGAATAATGGAATAGGACCTCGGTTCTA-TTTTGTTGGTTTTCGGAACCCCGAGGTAATGATTAATAGGGACAG-A
PANORPA         --TAAAGCAGG--CTAA-TCT-C-GCC-TGTATATTGTGTGCATGGAATAATAGAATAGGACCTCGTTTCTA-TTTTGTTGGTTTTCGGAATTCCGAGGTAATGATTAATAGGGACAA-C
PYGNOPSYCHE     --CAAACGGGG--CTAA-CTCTCGGCC-TGAATATTGTACGCATGGAATAATAGAATAGGACCTCGGTTCAG-TTTTGTTGGCTTTCTGATACCAGAGGTAATGATTAATAGGGACAA-C
RHIPIPHORUS     --TAAAGCAGG--CTAA-AATTTTGCC-TGAATACTGTGTGCATGGAATAATAGAATAGGACCTCGGTTCTA-TTTTGTTGGTTTTCGGAACCCCGAGGTAATGATTAATAGGAACGG-A
SALDULA         --TAAAGCAGG--CTGA-ATC-CAGCC-AGAATAGTG-GTGCATGGAATAATAGAACAGGACCTTGGTTCTA-TTCTGTTGGTTTTCGGAA-TCCAAGGTAATGATCAATAAGGACAG-A
TENEBRIO        --TAAAGCAGG--CTAA-AACTTCGCC-TGAATACTGTGTGCATGGAATAATGGAATAGGACCTCGGTTCTA-TTTTGTTGGTTTTCGGAATTTTGAGGTAATGATTAATAGGAACGG-A
TETRAOPES       --TAAAGCAGG--CTAA-AATTTCGCC-TGAATACTGTGTGCATGGAATAATGGAATAGGACCTCGGTTCTA-TTTTGTTGGTTTTCGGAACCCCGAGGTAATGATTAATAGGAACGG-A
TIPULA          --TAAAGCAGG--CCAT-TTG-T-GCC-TGAATATTCTCTGCATGGAATAATGGAATAAGACGTCTGTTCTATTTTCATTGGTTTTTRGAT-CAAGACGTAATGATTAATAGAAGCAG-T
TRIOZOCERA      --TAAAGCAGG--CAAATTTATATGCCTTGAATATCG-CAGCATGGAATAATAGAATATGATCTCGGTACCG-NNNNNNNNNNNNNNNNNNNNNAGA--TAATGATTAATAGAGACGG-A
WARRAMABA       --TAAAGCAGG--C-AA-GGC-C-G-C-TGAATACTGTGTGCATGGAATAATGGAATAGGACCTCGGTTCTA-TTTTGTTGGTTTTCGGAA-CCCGAGGTAATGATTACTA-GGACAG-G
XENOS           --TAAAGCAGG--CAAATTTATATGCCTTGAATATAA-CAGCATGGAATAATAGAATATGATCTCGATACTG-NNNNNNNNNNNNNNNNNNNNNAGA--TAATGATCAATAGAGACGG-A
```

```
LIBELLULA       TTGAC-GATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGG-GCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
ASCALAPHA       TTGACAGATTAACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
CAENOCHOLAX     TTGACAGATTGATAGCTCTTTCTTGATTCAGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
CRAWFORDIA      TTGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
CTENOCEPH       TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
DROSOPHILA      AAGACAGATTGATAGCTCTTTCTCGAATCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTCGTGGAGTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
GALLERIA        TTGACAGATTAACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
HEMEROBIUS      TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
HYDROPSYCHE     TTGACAGATTGACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGYGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
MONOBIA         TTGACAGATTGATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
MYTHICOMYIA     AAGACAGATTGATAGCTCTTTCTCGAATCTATGGGTGGTGGTGCATGGCCGTTGTTAGTTCGTGGAGTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
ONCOMETOPIA     TTGACAGATCGATAGCTCTTTCTTGATTCAGTGGGTAGTGGTGCATGGCCGTTCTTAGTTGGTGGACTGATTTGTCTGGTTTATTCCGATAACGAACGAGACTC
OPHION          TTGACAGATTGACAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
PANORPA         TTGACAGATTGAGAGCTCTTTCTTGATTCAGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
PYGNOPSYCHE     TTGACGGATTAACAGCTCTTTCTCGATTCGGTGGGTGGTGGTGCATGGCCGCTCTTAGTTGGTGGAGCGATTTGTCAGGTTAATTCCGATAACGAACGAGACTC
RHIPIPHORUS     TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
SALDULA         TTGACAGATTGATAGCTCTTTCTTGATTCAGTGGGTAGTGGTGCATGGCCGTTCTTAGTTGGTGGAACGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TENEBRIO        TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TETRAOPES       TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TIPULA          AAGACAGATTAATAGCTCTTTCTCGAATCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTCGTGGAGTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
TRIOZOCERA      TTGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
WARRAMABA       TTGACAGATTGAGAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGACTGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
XENOS           TTGACAGATTAATAGCTCTTTCTTGATTCGGTGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTCCGATAACGAACGAGACTC
```

Figure 16. Three alignments of 18S rDNA. The sequence data of Whiting and Wheeler (1994) are used to demonstrate variation in alignment due to "path" selection. In the first alignment (A), matches are favored over gaps when equal cost would result from either. The second (B) favors contiguous gaps, that is inserts a gap if one already exists. The third alignment (C) favors discontinuous gaps, favoring gaps when gaps do not already exist.

that alignment ambiguities (positional correspondences which vary among alignments) are a statement of error. This error could be in the assumptions of alignment or in the data themselves in that they are "too variable" for the question at hand. The logical direction of this thinking is to remove or at least downweight nucleotide positions which do not align consistently.

In their chapter on computer packages which aid evolutionary sequences analysis, DeSalle et al. (this volume) discuss two procedures which lessen the impact of alignment-ambiguous (*sensu* Gatesy et al., 1993) sites: CULL and ELISION. Briefly, if alignment ambiguous sites are thought to be totally unreliable, they can be removed. This is what CULL does. Investigators have reported removing "unalignable" regions of molecules many times. Unfortunately, in removing all disagreeing positions, most of the signal seems to be removed as well. The properties of such harsh weighting (0 or 1) are discussed more fully by Gatesy et al. (1993).

A second procedure along the same lines as CULL is ELISION (Wheeler et al., in press). Like CULL, ELISION distrusts alignment-ambiguous nucleotide positions, but to a less severe extent. In this method, the several alignments are combined into a single "grand" alignment. This is then analyzed phylogenetically. By combining the alignments, positions which align consistently are given relatively higher weights than are those which are less stable. ELISION seems to maintain much more of the original signal of the alignments than does CULL. ELISION is in its continuous weighting based on internal consistence, like Successive Approximations Weighting (SAW), originated by Farris (1969) as a method for character weighting.

An entirely different approach to multiple solutions to alignment problems is found in congruence based procedures. One such procedure (Wheeler, in press), uses the congruence between external data (morphology, biogeography) to choose the "best" among several multiple alignments. An alignment would be chosen which maximized some congruence measure between the sequences at hand and other, preexisting data. On the other hand, from a purely phylogenetic perspective, one might accept only those groups implied by all alignments (akin to CULL).

Conclusions

The intense discussion of the methods by which solutions are generated which has characterized cladogram construction for so long is now coming to sequence alignment. As with phylogenetic analysis, the results can sometimes not be entirely resolved. Only by embracing this phenomenon, however, can we hope to understand it and its implications for phylogeny.

Acknowledgements

I would like to acknowledge the assistance of Rob DeSalle for his forbearance and assistance through all phases of this work. This work was supported by National Science Foundation Grants BSR 92-07335 and BSR 92-07624.

References

Farris, J.S. (1969) A successive approximations approach to character weighting. *Syst. Zool.* 18: 374–385.

Farris, J.S. (1981) Distance data in phylogenetic analysis. *In*: V.A. Funk and D.R. Brooks (eds.): *Advances in Cladistics: Proceedings of the First Meeting of the Willi Hennig Society*. New York Botanical Garden, New York, pp. 3–22.

Farris, J.S. (1985) Distance data revisited. *Cladistics* 1: 67–85.

Farris, J.S. (1986) Distances and statistics. *Cladistics* 2: 144–157.

Farris, J.S. (1988) Hennig86 version 1.5, Program and documentation. Port Jervis, New York.

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. *In*: R.F. Doolittle (ed): *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Method. Enzymol.* 183: 375–387.

Fitch, W.M. and Margoliash, E. (1967) The construction of phylogenetic trees. *Science* 155: 279–284.

Fitch, W.M. and Smith, T.F. (1983) Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA* 80: 1382–1386.

Gatesy, J., DeSalle, R. and Wheeler, W. (1993) Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phyl. Evol.* 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. *In*: R.F. Doolittle (ed.): *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Method Enzymol.* 183: 626–644.

Hendy, M.D. and Penny, D. (1982) Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.* 59: 277–290.

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. *CABIOS* 5: 151–153.

Konings, D.A.M., Hogeweg, P. and Hesper, B. (1987) Evolution of the primary and secondary structures of the E1a mRNAs of the adenovirus. *Mol. Biol. Evol.* 4: 300–314.

Mindell, D. (1991) Aligning DNA sequences: homology and phylogenetic weighting. *In*: M.J. Miyamoto and J. Cracraft (eds): *Phylogenetic Analysis of DNA Sequences*. Oxford University Press, New York, pp. 73–89.

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.

Shankoff, D.D. and Cedergren, R.J. (1983) Simultaneous comparison of three or more sequences related by a tree. *In*: D. Shankoff and J.B. Kruskal (eds): *Time Warps, String Edits, and Macromolecules: the Theory and Practise of Sequence Comparison*. Addison-Wesley, Reading, Massachusetts, pp. 253–264.

Swofford, D.L. (1993) PAUP version 3.1, Program and documentation. Champaign, IL.

Waterman, M.S., Eggert, M. and Lander, E. (1992) Parametric sequence comparison. *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. and Gladstein, D.M. (1993) Malign: A Multiple Sequence Alignment Program. program and documentation, version 1.92. New York.

Wheeler, W.C. (1994) Sequence alignment, parameter sensitivity, and the phylogentic analysis of molecular data. *Syst. Biol.; in press.*
Wheeler, W.C. and Gladstein, D.G. (1994) Malign: a multiple nucleic acid sequence alignment program. *J. Hered.; in press.*
Wheeler, W.C., Gatesy, J. and DeSalle, R. (1994) Elision: a method for accomodating multiple molecular sequence alignments with alignment-ambiguous sites. *Mol. Phyl. Evol.: in press.*
Whiting, M., and Wheeler, W.C. (1994) Phylogenetic position of the Strepsiptera: evidence for a homeotic reciprocal thoracic transformation. *Nature*; 368: 696.

Molecular Ecology and Evolution: Approaches and Applications
ed. by B. Schierwater, B. Streit, G.P. Wagner & R. DeSalle

Computational problems in molecular systematics

R. DeSalle[1], C. Wray[2] and R. Absher[3]

[1]*Department of Entomology, American Museum of Natural History, New York, NY 10024, USA*
[2]*Department of Geology and Geophysics, Yale University, New Haven, CT 06511, USA*
[3]*Department of Biology, Yale University, New Haven, CT 06511, USA*

Summary. The development of extremely powerful computer programs and the ready availability of microcomputers has revealed several computational problems with data analysis. These problems occur in the handling of systematic data in general and molecular systematic data in particular. This paper examines three areas of controversy in molecular systematics resulting from increased computer power. We start by examining the first step in DNA sequence analysis, the establishment of homology via sequence alignment. Next we examine several problems in phylogenetic analysis that have arisen in the last few years due to use of the PAUP (Swofford, 1991), HENNIG86 (Farris, 1988), and PHYLIP programs. These problems include limitations on the number of taxa examined in a given analysis and the accuracy of the parsimony trees in such analyses. The final subject is an examination of programs used for assessing tree robustness. We concentrate on certain programs (such as MALIGN (Wheeler and Gladstein, 1993), PAUP (Swofford, 1991), HENNIG86 (Farris, 1988), PHYLIP (Felsenstein, 1990), CLADOS (Nixon, 1993), MacClade (Maddison and Maddison, 1993), etc.), but similar comments about other programs could also be made.

Sequence alignment

The establishment of homology of DNA sequence positions has been recognized as a central problem in modern systematics (Fitch and Smith, 1983; Feng and Doolittle, 1986; Mindell, 1991). The basic approach used in sequence alignment is the dynamic programming approach of Needleman and Wunsch (1970). Several computer algorithms have been developed using various criteria to accomplish the alignment (Sankoff et al., 1973; Feng and Doolittle, 1987; Higgins and Sharp, 1988, 1989; Konings et al., 1987; Hein, 1989, 1990; Mindell, 1991; Wheeler and Gladstein, 1993). These approaches are highly dependent on the initial parameters used in the alignment procedure. As Wheeler describes in this volume, most alignment procedures for sequences are based on the costs assigned for assuming a base or amino acid change versus the insertion of a gap. These are referred to as change and gap costs respectively. It has been demonstrated that these parameters (Fitch and Smith, 1983; Waterman et al., 1992; Gatesy et al., 1993) as well as the order in which sequences are read into a sequence alignment run (Lake, 1991) affect the final alignment in a substantial manner.