Modelling the secondary structures of slippage-prone hypervariable RNA regions: the example of the tiger beetle 18S rRNA variable region V4

John M. Hancock* and Alfred P. Vogler[1,2]

Gene and Genome Evolution Group, Medical Research Council Clinical Sciences Centre, Imperial College School of Medicine, Hammersmith Hospital, London W12 0NN, UK, [1]Department of Entomology, The Natural History Museum, London SW7 5BD, UK and [2]Department of Biology, Imperial College at Silwood Park, Ascot, Berkshire SL5 7PY, UK

Received December 5, 1997; Revised and Accepted February 6, 1998

ABSTRACT

Variable regions within ribosomal RNAs frequently vary in length as a result of incorporating products of slippage. This makes constructing secondary structure models problematic because base homology is difficult or impossible to establish between species. Here, we model such a region by comparing the results of the MFOLD suboptimal folding algorithm for different species to identify conserved structures. Based on the reconstruction of base change on a phylogenetic tree of the species and comparison against null models of character change, we devise a statistical analysis to assess support of these structures from compensatory and semi-compensatory (i.e. G.C to G.U or A.U to G.U) mutations. As a model system we have used variable region V4 from cicindelid (tiger beetle) small subunit ribosomal RNAs (SSU rRNAs). This consists of a mixture of conserved and highly variable subregions and has been subject to extensive comparative analysis in the past. The model that results is similar to a previously described model of this variable region derived from a different set of species and contains a novel structure in the central, highly variable part. The method we describe may be useful in modelling other RNA regions that are subject to slippage.

INTRODUCTION

One of the great successes of modern theoretical biology has been the use of base pair covariation in the modelling of RNA secondary structure (1). The paradigm of covariation analysis is that there must be a strict relationship between bases that pair with one another within an RNA molecule, so that base pairing between the equivalent sites in a multiple alignment of sequences is always conserved. If a base at one of the pairing positions undergoes a transversion, for example from G to C, a corresponding transversion from C to G must take place to preserve the interaction. If a large enough sample of sequences is available, it will be possible to distinguish chance concerted base changes from true covariation, and the secondary structure of a molecule will emerge. Under this model, only positive replacement of a base pair, not conservation of sequence without change, is treated as informative. Because of this, the method requires a set of sequences that are sufficiently evolutionarily distant from one another to include numerous base changes at all positions within a structure. This approach has produced models of large- and small-subunit ribosomal RNAs consistent with experimental probing of the native conformation (1) and is generally regarded as being a more successful approach to modelling RNA secondary structure than the alternative method of predicting RNA secondary structure on the basis of energy minimization.

A large set of unambiguously aligned sequences is essential for successful modelling of RNA structure by covariation. However, some kinds of study may generate datasets for which either too few sequences are available for covariation to be established unambiguously or for which alignment is ambiguous, for example if sequences have undergone length variation during evolution as a result of the action of replication slippage. It may nevertheless be important to define a secondary structure model of the region under study, for example as an aid to alignment for phylogenetic analysis (2), investigating probabilities of nucleotide substitutions (3) or if the aim is to investigate the secondary structure itself either with respect to its functional role in experimental systems or from an evolutionary viewpoint. This could apply to any rapidly evolving nucleic acid sequence that adopts a secondary structure [for example the product of the mammalian X-inactivation gene Xist (4), mitochondrial control regions (5) and viral genomes (6)], but it is likely to be a particular problem in eukaryotic ribosomal RNAs, which in many cases contain long regions (known as expansion segments or variable regions) that appear to have evolved by slippage, are not unambiguously homologous between species (7–9), but which have been used in phylogenetic reconstruction because their rapid evolution provides high resolution (9).

Here, we evaluate an approach to modelling secondary structures of length-variable regions in a phylogenetically limited dataset which does not rely on a pre-existing unambiguous alignment. We do this by making use of the secondary structure prediction program MFOLD (10–13) to identify elements of secondary structure that occupy homologous regions in a set of variable region sequences. Potential structures able to form in a majority of sequences are analysed for base covariation in the normal way, i.e. by identifying compensatory changes within the sequence, and by using a statistical approach to analyse semi-

*To whom correspondence should be addressed. Tel: +44 181 383 8232; Fax: +44 181 383 8334; Email: jhancock@rpms.ac.uk compensatory mutations [defined as mutations that transform a full Watson–Crick base pair (A–U or G–C) into a wobble pair (G.U), or vice versa, by a single point mutation]. This allows for occasional non-pairing combinations of bases without setting pre-determined limits for their frequency. We examine the procedure's utility on a set of sequences of variable region V4 from the small subunit ribosomal RNAs (SSU rRNAs) of tiger beetles (Coleoptera: Cicindelidae) and relatives (9). This variable region has the advantage that it contains regions of high sequence conservation, for which a detailed secondary structure model has been proposed (14,15), interspersed with highly variable regions. This allows us to test the ability of our method to recover well-tested elements of secondary structure while at the same time investigating the more variable parts of V4 for the presence of potentially conserved structures, such as a long variable stem that we have suggested previously might form in the variable central region of V4 (9).

MATERIALS AND METHODS

DNA sequences and sequence alignment

The DNA sequences used for structure modelling are those obtained previously for taxonomic purposes (9) plus sequences for *Drosophila melanogaster* (16) and *Tenebrio molitor* (17). The alignment used is essentially identical to figure 1 of ref. 9 except in the region of helix IX, where it has been modified to be consistent with conservation of this stem in other species. The alignment contains 32 sequences and is 526 characters long.

Secondary structure modelling

Preliminary modelling by energy minimization was carried out using the MFOLD program of Zuker and Jaeger (10–13) running under version 8.1 of the GCG package (18) on the MRC Human Genome Mapping Project computer, Cambridge, UK.

Complete variable region sequences were passed through the program and the *n* structures falling within a window of stability determined automatically by the program were plotted out using the GCG program PLOTFOLD. Parameters used for PLOTFOLD analysis were: maximum size and lopsidedness of internal loop, 30; energy increment, 2.0; window size, 3. Optimal and sub-optimal structures were generated for 27 of the 32 species. Three sequences (from *Cicindela repanda*, *Megacephala klugi* and *Neocollyris* sp.) could not be modelled as their V4 sequences were incomplete, although they could be incorporated into analyses of compensatory mutation (see below).

The individual structural elements found in these global structures were identified and compiled in separate alignments for each structure to confirm their homology and structural similarity. The number of species in which each structure element appeared was determined and the most common elements (i.e. those found in most species) subjected to covariation analysis.

Analysis of mutational patterns

Because we have a phylogenetic tree for these species (19; Fig. 2), we were able to map base changes seen within putative secondary structures onto the tree and, in most cases, to ascribe directionality to them. We defined nine kinds of change between pairs of bases opposed within secondary structure models, corresponding to changes among three classes of opposed bases: Watson–Crick paired bases, wobble (G.U) base pairs and unpaired pairs of bases. Changes of state were identified using MacClade (version 3.04; 20). Base pairings were coded as unordered multistate character states: codes 1–4 corresponded to Watson–Crick pairs (A–U, U–A, C–G and G–C, respectively), and 5 and 6 to wobble pairs (G.U and U.G, respectively). The remaining four states for character coding available in MacClade were used for non-pairing combinations on an *ad hoc* basis for each stem. For the subsequent statistical analysis, changes between pairs whose direction was unambiguous in the context of the phylogenetic tree (19; Fig. 2) were then counted. To minimise effects of double mutational hits, *D.melanogaster* and *T.molitor* sequences were not included in the statistical analysis because of their great evolutionary distances from the remainder of the taxa considered.

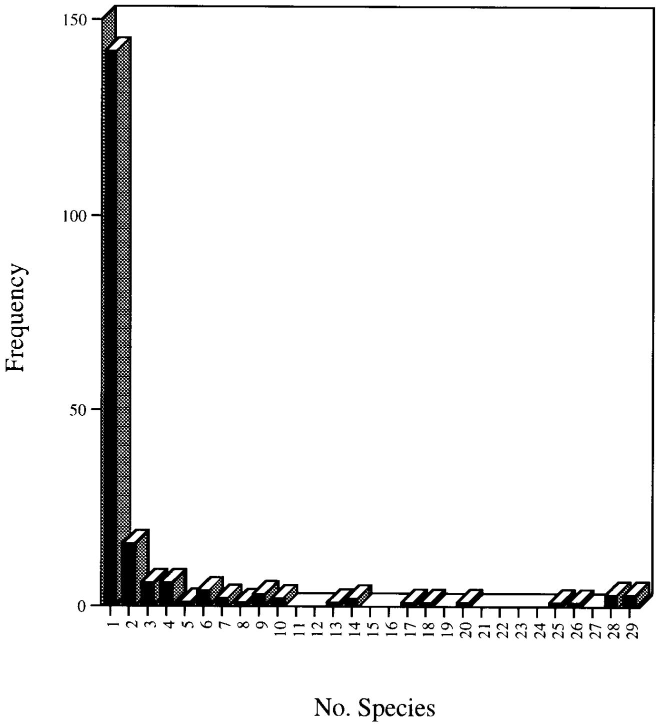


Figure 1. Numbers of stem structures appearing in MFOLD analyses in different numbers of species. Vertical axis, frequency; horizontal axis, number of species.

Statistical analysis of changes

Frequencies of different classes of change were compared with expectations using the χ^2 test (see legend to Table 3 for details). Expected values were calculated based on the same total number of changes and using three models for the frequencies of change. Under the first model used (M1), all changes at covarying sites were assumed to be the result of a single base change in one of the covarying partners. Under this model, a Watson–Crick base pair can undergo 12 possible changes by single mutations, of which 10 result in non-compensatory changes and two give rise to G.U pairs ($P = 0.833$ and 0.167, respectively). Similarly, G.U pairs can give rise to six products, of which two are Watson–Crick pairs and four are unpaired ($P = 0.333$ and 0.667, respectively). The second model (M2) was a modification of the first model which assumed that transitions were twice as likely to occur as transversions. Under this model a Watson–Crick pair has a 4/16 ($P = 0.25$)

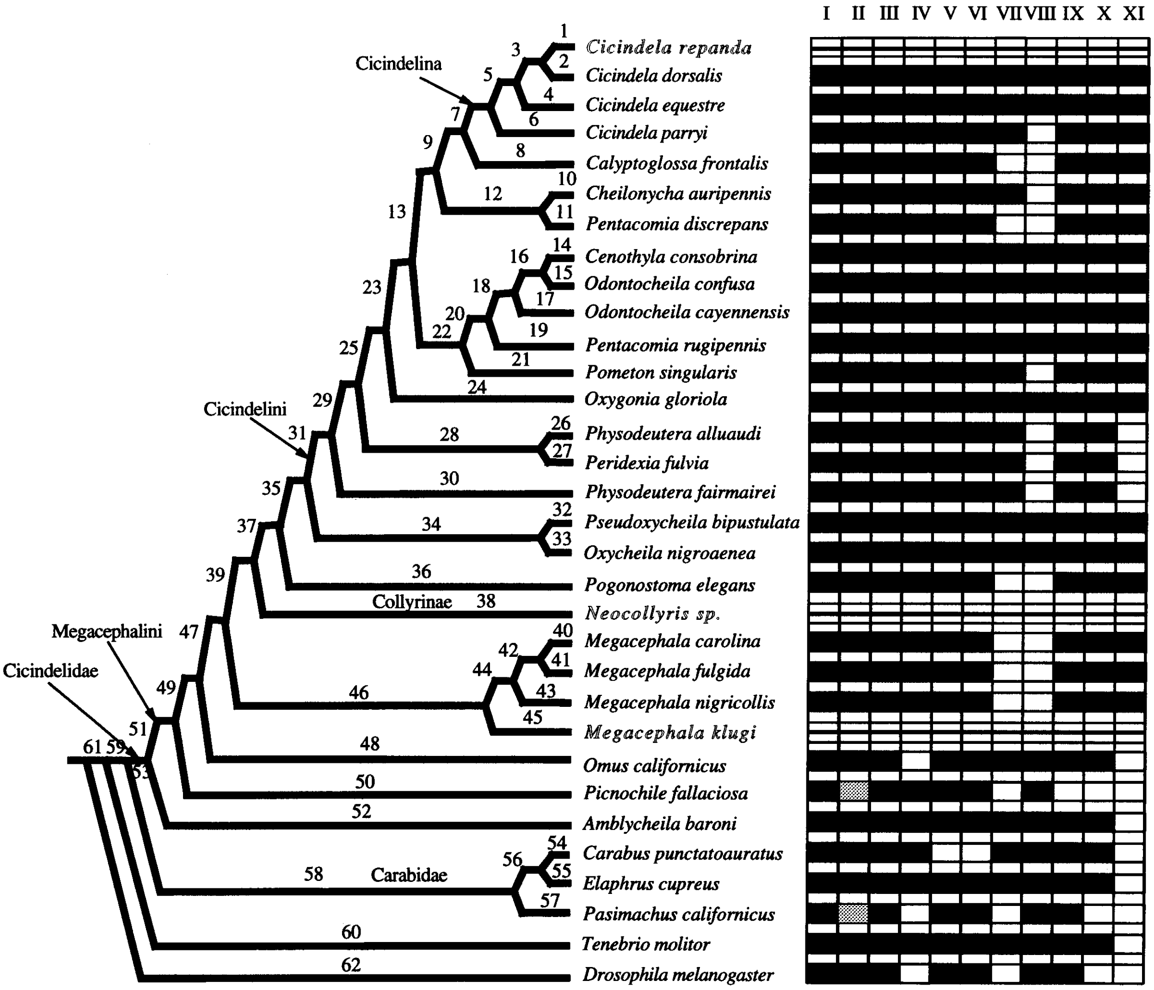


Figure 2. Phylogenetic tree of species considered in this analysis and distribution of the commonly found potential secondary structures. The phylogenetic tree is re-drawn from Vogler and Pearson (19). Nodes identifying traditionally recognized groupings are indicated and individual branches are numbered to allow their identification in Table 1. The grid on the right hand side represents whether or not a given structural element was found in MFOLD predictions for the sequence from a particular species. Each column in the grid represents a particular structure (labelled from I to XI). A dark shaded box indicates that the structure was found in MFOLD analysis, a light shaded box that it was found in modified form and a white box that it was not found. Horizontal lines are drawn through rows corresponding to species for which MFOLD analysis was not carried out because sequences were incomplete.

probability of changing to a wobble pair and a 0.75 probability of giving rise to an unpaired combination. Wobble pairs give rise to Watson–Crick and unpaired combinations at probabilities of 0.357 and 0.643, respectively.

As well as changes that appeared to have taken place by single base mutations, the data also included a number of examples of changes that had to involve more than one base change. As we could not exclude the possibility that all the changes seen here, including those appearing to result from single base changes, in reality resulted from multiple changes, we also invoked a third model (M3) that made no assumption about the process of change. In this model, each pair can give rise to all 15 other possible pairs with equal probability, so that any Watson–Crick pair can give rise to the three other Watson–Crick pairs ($P = 0.2$), two wobble pairs ($P = 0.133$) and 10 unpaired combinations ($P = 0.667$). For wobble pairs, the predicted proportions are 0.267 for Watson–Crick pairs, 0.067 for the other wobble pair and 0.667 for unpaired bases.

RESULTS

One hundred and ninety seven distinct potential structural elements were identified in the search of minimum energy structures. Figure 1 shows a histogram of the frequency of structures found in different number of species. Of the 197 structures, 11 were found in 15 or more species (i.e. the majority) and were subjected to subsequent analysis. The species for which these 11 structures were predicted are identified in Figure 2 in the context of the likely phylogeny (19). In addition, Neefs and De Wachter (21) suggested that a pseudoknot forms at the 3′ end of the region. As pseudoknots are not detected by MFOLD, we added this structure to the analysis. Putative secondary structures for 10 of the 11 frequently found secondary structures plus the pseudoknot are shown in Figure 3. The bulk of stem II contained too many insertions/deletions (presumably resulting from slippage in this region; 9) to be aligned with confidence. A sequence alignment summarising the relative positions of these 12 structures

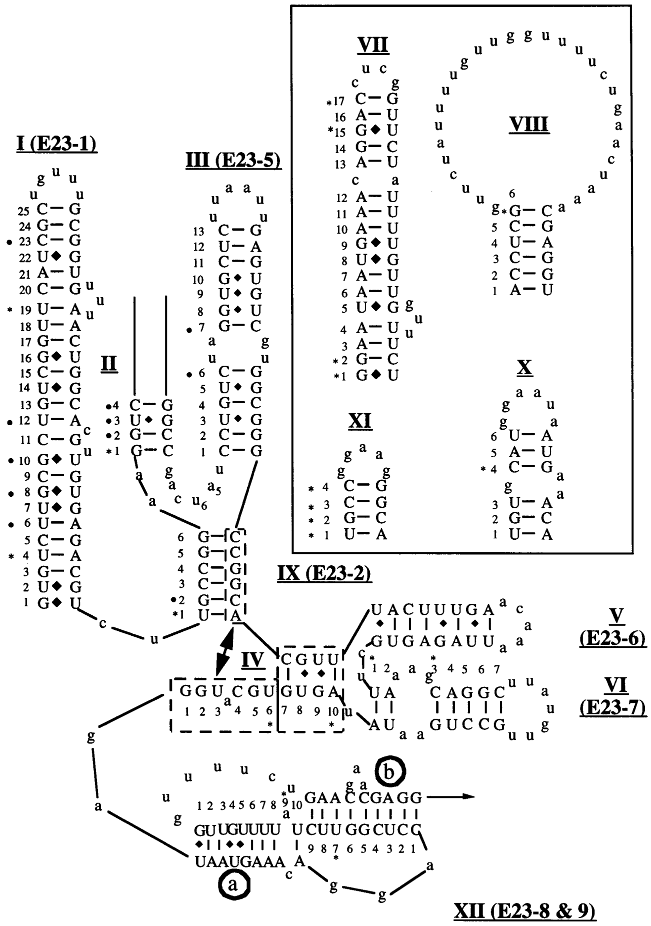


Figure 3. Summary of the secondary structure model arising from these analyses. Stems are labelled according to the numbering system used in the Discussion and according to the model of De Rijk *et al.* (15) if also present in that model. Sequences used are majority rule consensi of the 32 sequences except for stem I, which is a consensus of 29 sequences. Paired bases are presented in upper case; unpaired bases in lower case. The 5′ to 3′ direction proceeds from the left to right hand side of the figure. Positions within secondary structural elements are numbered. Stems IV and IX are long-range interactions which overlap but were not unambiguously resolved. Stem IX, which is present in De Rijk *et al.*'s model (15), is shown in its correct position; the arrows represent the possible interaction between the component parts of stem IV which overlap with stem IX, and all component parts of this stem are contained within dotted boxes. Structures in the main part of the figure represent the final model arising from these analyses; the structures within the box were seen in more than half of MFOLD analyses but were not supported otherwise. Paired positions marked with an asterisk showed only semi-compensatory changes and those marked with a black dot showed at least one fully compensatory change. A structure for stem II is not included because it varied greatly between species. Examples of stem II structures for *Odontocheila confusa*, *E.cupreus* and *P.fallaciosa* are presented in figure 2 of ref. 9.

in the V4 sequences is shown in Figure 4. The alignment used is essentially identical to figure 1 of ref. 9 except in the region of helix IX, where it has been modified to be consistent with conservation of this stem in other species.

The numbers of species in which these putative structures were found, and the result of the analysis of compensatory and non-compensatory changes for each of them, can be summarised as follows (stem numbering as in Figs 3 and 4; details of all base changes within putative structures are summarised in Table 1):

Stem I. This stem was identified in all 29 species studied. In covariation analysis, positions 1–3, 9, 11, 16–18, 20, 21 and 24 of the stem were found to be invariant. Compensatory and semi-compensatory changes predominated in this structure although positions 4, 10 and 19 showed single non-compensatory changes.

Stem II. This was the long, hypervariable stem, suggested previously to occur in these species (9). Because this region of the sequence showed little or no sequence conservation, this stem was defined by the criterion that it was enclosed by a pair of conserved, base-pairing motifs: GRRC and GYYC. By this definition the stem was seen clearly in 26 species. In addition, in *Neocollyris* sp. a region bounded by these motifs was predicted to form a discrete unit of secondary structure that bifurcated at its tip to form a Y-shaped structure. In *Pasimachus californicus* the basal motifs were not clearly distinguishable (Fig. 4) but a stem could form starting at the equivalent position. In the extremely long sequence of *Picnochile fallaciosa*, a discrete stem could form that included most of the region between the two conserved motifs, but the motifs themselves were predicted to pair with other parts of V4 in a number of additional structures. Examples of the structure of stem II for some of the species considered here are included in ref. 9.

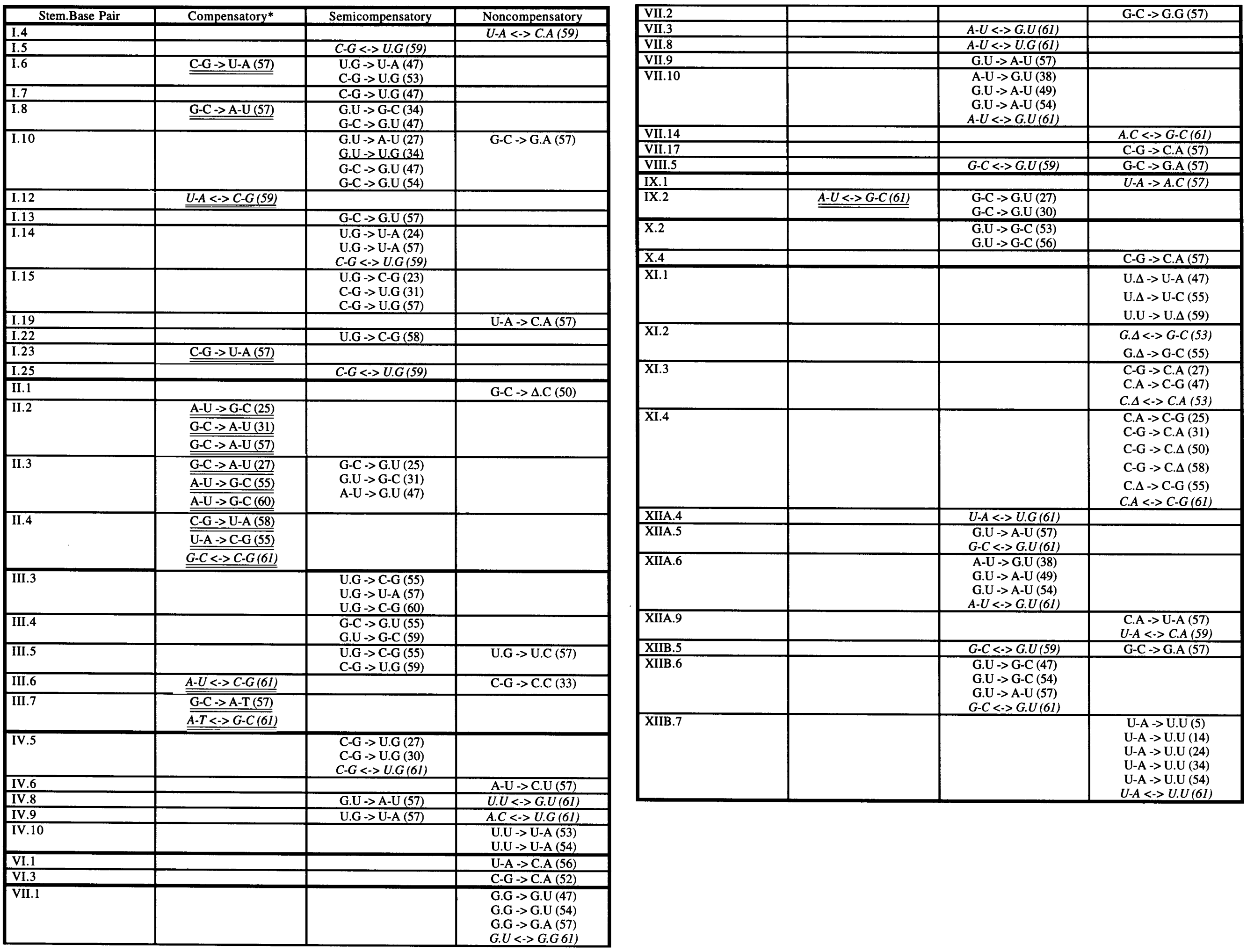
Because of the extreme divergence and multiple indels in this region, it was not possible to identify compensatory mutations in most of it. However, in the stem formed by the basal motifs, three of the four positions showed compensatory mutations. Only in *P.fallaciosa* (which showed additional potential secondary structures involving these sequences) was an unambiguous non-compensatory change seen, at position 1.

Stem III. This is the highly conserved stem included in both of the current models of SSU-rRNA structure (14,22). This stem was found in folds from all 29 species investigated. In this dataset positions 1, 2 and 8–13 were invariant, while positions 3–5 showed semi-compensatory changes in a number of species. Position 6 showed a compensatory G–C to A–U change in *D.melanogaster* and a G–C to G.U change in *Elaphrus cupreus*, but in *Pseudooxychelia bipustulata*, *Oxychelia nigroaenea* and *Carabus punctatoauratus* this pairing could not form. The absence of this pairing could be accommodated in all three of these species by alternative pairing arrangements differing only subtly from the model presented in Figure 3. Position 7 showed two compensatory changes, a G–C to A–U change in the *P.californicus* lineage and either a G–C to A–U or A–U to G–C change at the root of the tree.

Stems IV–VI. The individual elements of this potential Y-shaped structure showed different degrees of conservation between species. The stem forming the base of this structure (stem IV) was predicted in 26 species, the exceptions being *O.californicus*, *P.californicus* and *D.melanogaster.* It was shortened by one base pair in *E.cupreus* and *T.molitor.* Stems V and VI, the two bifurcating stems, were both predicted in 28 species. In *E.cupreus* and *C.punctatoauratus* stem V was slightly reorganized as a result of changes in flanking bases. Some variation could also be seen in the terminal loop of stem VI. This varied in length from three bases in *C.punctatoauratus* to 10 in *P.californicus.*

Co-variation analysis of these stems showed that for stem IV, positions 1–4 and 7 were invariant and positions 5, 8 and 9 showed semi-compensatory changes. Positions 5 and 6 showed single non-compensatory changes, while positions 8 and 9 could

Table 1. Summary of changes in stems I–XII



Stem.Base Pair	Compensatory*	Semicompensatory	Noncompensatory
I.4			*U-A <-> C.A (59)*
I.5		*C-G <-> U.G (59)*	
I.6	C-G -> U-A (57)	U.G -> U-A (47) C-G -> U.G (53)	
I.7		C-G -> U.G (47)	
I.8	G-C -> A-U (57)	G.U -> G-C (34) G-C -> G.U (47)	
I.10		G.U -> A-U (27) G.U -> U.G (34) G-C -> G.U (47) G-C -> G.U (54)	G-C -> G.A (57)
I.12	*U-A <-> C-G (59)*		
I.13		G-C -> G.U (57)	
I.14		U.G -> U-A (24) U.G -> U-A (57) *C-G <-> U.G (59)*	
I.15		U.G -> C-G (23) C-G -> U.G (31) C-G -> U.G (57)	
I.19			U-A -> C.A (57)
I.22		U.G -> C-G (58)	
I.23	C-G -> U-A (57)		
I.25		*C-G <-> U.G (59)*	
II.1			G-C -> Δ.C (50)
II.2	A-U -> G-C (25) G-C -> A-U (31) G-C -> A-U (57)		
II.3	G-C -> A-U (27) A-U -> G-C (55) A-U -> G-C (60)	G-C -> G.U (25) G.U -> G-C (31) A-U -> G.U (47)	
II.4	C-G -> U-A (58) U-A -> C-G (55) *G-C <-> C-G (61)*		
III.3		U.G -> C-G (55) U.G -> U-A (57) U.G -> C-G (60)	
III.4		G-C -> G.U (55) G.U -> G-C (59)	
III.5		U.G -> C-G (55) C-G -> U.G (59)	U.G -> U.C (57)
III.6	*A-U <-> C-G (61)*		C-G -> C.C (33)
III.7	G-C -> A-T (57) *A-T <-> G-C (61)*		
IV.5		C-G -> U.G (27) C-G -> U.G (30) *C-G <-> U.G (61)*	
IV.6			A-U -> C.U (57)
IV.8		G.U -> A-U (57)	*U.U <-> G.U (61)*
IV.9		U.G -> U-A (57)	*A.C <-> U.G (61)*
IV.10			U.U -> U-A (53) U.U -> U-A (54)
VI.1			U-A -> C.A (56)
VI.3			C-G -> C.A (52)
VII.1			G.G -> G.U (47) G.G -> G.U (54) G.G -> G.A (57) *G.U <-> G.G 61)*
VII.2			G-C -> G.G (57)
VII.3		*A-U <-> G.U (61)*	
VII.8		*A-U <-> U.G (61)*	
VII.9		G.U -> A-U (57)	
VII.10		A-U -> G.U (38) G.U -> A-U (49) G.U -> A-U (54) *A-U <-> G.U (61)*	
VII.14			*A.C <-> G-C (61)*
VII.17			C-G -> C.A (57)
VIII.5		*G-C <-> G.U (59)*	G-C -> G.A (57)
IX.1			*U-A -> A.C (57)*
IX.2	*A-U <-> G-C (61)*	G-C -> G.U (27) G-C -> G.U (30)	
X.2		G.U -> G-C (53) G.U -> G-C (56)	
X.4			C-G -> C.A (57)
XI.1			U.Δ -> U-A (47) U.Δ -> U-C (55) U.U -> U.Δ (59)
XI.2			*G.Δ <-> G-C (53)* G.Δ -> G-C (55)
XI.3			C-G -> C.A (27) C.A -> C-G (47) *C.Δ <-> C.A (53)*
XI.4			C.A -> C-G (25) C-G -> C.A (31) C-G -> C.Δ (50) C-G -> C.Δ (58) C.Δ -> C-G (55) *C.A <-> C-G (61)*
XIIA.4		*U-A <-> U.G (61)*	
XIIA.5		G.U -> A-U (57) *G-C <-> G.U (61)*	
XIIA.6		A-U -> G.U (38) G.U -> A-U (49) G.U -> A-U (54) *A-U <-> G.U (61)*	
XIIA.9			C.A -> U-A (57) *U-A <-> C.A (59)*
XIIB.5		*G-C <-> G.U (59)*	G-C -> G.A (57)
XIIB.6		G.U -> G-C (47) G.U -> G-C (54) G.U -> A-U (57) *G-C <-> G.U (61)*	
XIIB.7			U-A -> U.U (5) U-A -> U.U (14) U-A -> U.U (24) U-A -> U.U (34) U-A -> U.U (54) *U-A <-> U.U (61)*

*Numbers in brackets represent the branch of the tree in Figure 2 on which the change is predicted to have taken place. Changes in italics represent changes whose direction could not be defined unambiguously; double underlined changes represent fully compensatory changes; single underline represents a wobble→wobble change involving more than one base change.

not form in *D.melanogaster.* Position 10 showed two changes from non-paired to potentially paired states (U.U→U–A) in the *C.punctatoauratus* lineage and after the divergence of the *P.californicus* lineage from the rest of the tree. All positions corresponding to stem V were invariant. In stem VI positions 2 and 4–7 were invariant. Positions 1 and 3 both showed single non-compensatory mutations: from U–A to C.A at position 1 in the lineage leading to *E.cupreus* and *C.punctatoauratus,* and from C–G to C.A at position 3 in *Amblychelia baroni.*

Stem VII. This was the most frequently found structure 3′ to stem VI during energy modelling. It was found in 20 species, and was missing mostly in taxa basal on the tree (Fig. 2). The structure presented in Figure 3 is a structure with a long stem found in 10 of these 20 species. Other species showed versions truncated by generally one or two base pairs. Positions 1, 2, 14 and 17 of the long form of this stem (Fig. 3) showed non-compensatory changes. Positions 4–7, 11–13 and 16 were invariant, while positions 3 and 8–10 showed semi-compensatory changes.

Stem VIII. This structure comprised a variable length stem enclosing a variable length, but generally long, terminal loop (14 nucleotides in *D.melanogaster*, 28 in seven species). It was observed in 17 species but was lacking from many of the taxa in the Cicindelini. Position 5 showed one semi-compensatory mutation and one non-compensatory mutation.

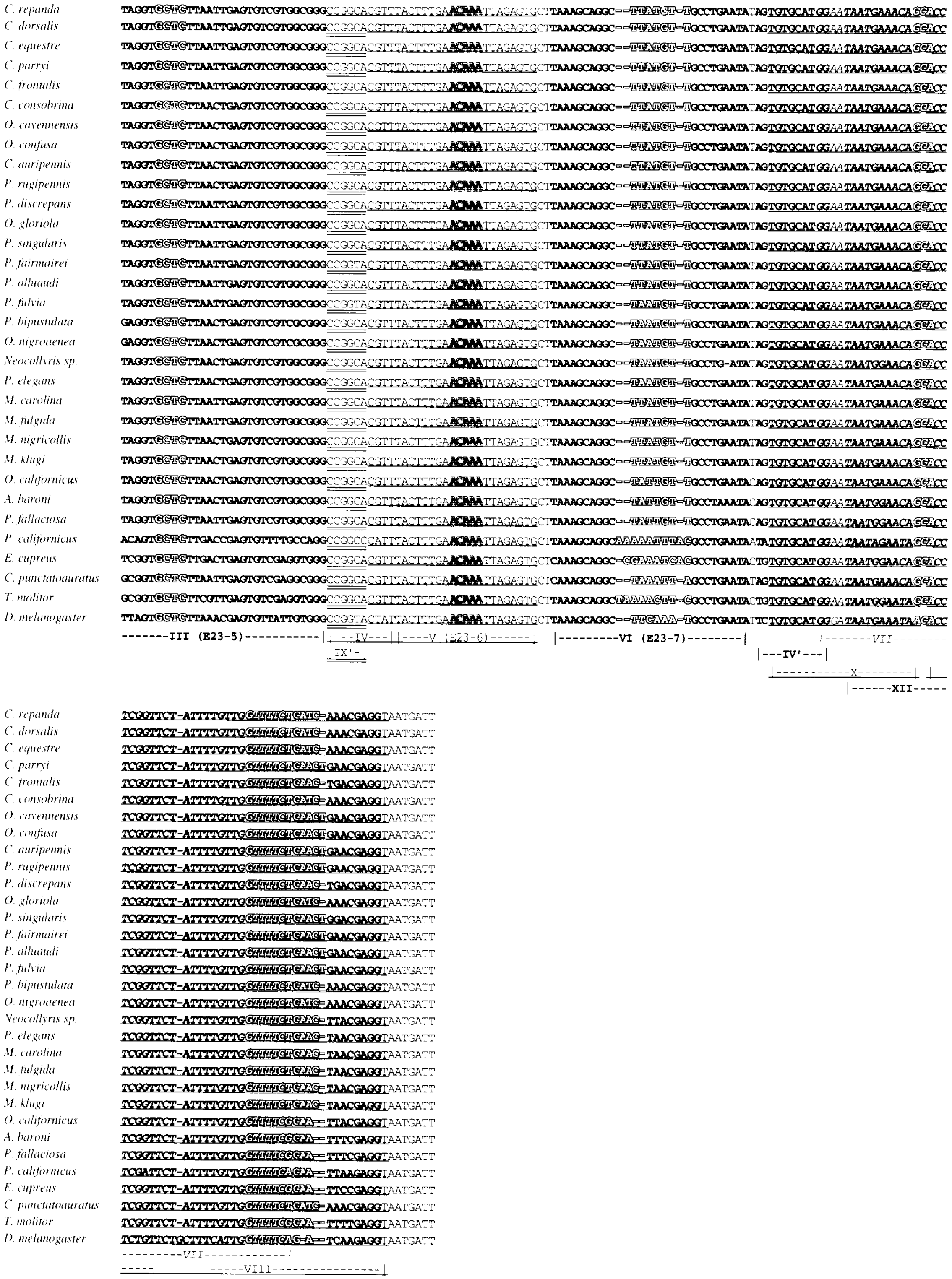
Stem IX. This was a long range interaction between sequences immediately 5′ to stem II and 3′ to stem III and overlapping stem IV. It was detected in folds from 28 species. Of the six base pairs making up this stem, four, positions 3–6, were invariant. Position 1 showed a non-compensatory A–U→A.C change in *P.californicus,* while position 2 showed semi-compensatory G–C→G.U changes in *Peridexia fulvia* and *Physodeutera fairmaiei,* and an A–U pair in *D.melanogaster.*

Stem X. This stem could form immediately 3′ to stem VI in 25 species. It showed four invariant pairing positions (1, 3, 5 and 6). Position 2 showed two semi-compensatory G.C→G–U

```
C. repanda          NNNNNNNNNNNNNNNNNNNNNNNNNNNTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TA-TG-TT-GTTCG
C. dorsalis         GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TATGTTTT-GTCA-
C. equestre         GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TA-TG-TT-GTTC-
C. parryi           GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTGTTA-TATGTTTTATTTA-
C. frontalis        GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCATG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TA-AGTTT-GTTCT
C. consobrina       GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TT-TG--T-G-TC-
O. cayennensis      GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TT-TG--T-G-TC-
O. confusa          GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TT-TG--T-G-TC-
C. auripennis       GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TT-TG--T-G-TC-
P. rugipennis       GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TT-TG--T-G-TC-
P. discrepans       GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA--T-T-TGT-G-TC-
O. gloriola         GAATTTGTGTCTTGCGCTGTTGGTTCATCGC-T-GTTTGCGGTGTTTAACTGACACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA--T-T---T-G-TC-
P. singularis       GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA-TTTGTGTCGTTCAT
P. fairmairei       GAATTTGTGTCTTGCGCTGTNGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGAGCATGGTATATGTTTTA--T-T-TGC-G-TC-
P. alluaudi         GAATTTGTGTCTTGCGCTGTTGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGAGCATGGTATATGTTTTA-TT-TG--T-G-TC-
P. fulvia           GAATTTGTGTCTTGCACTGTTGGTTCATCGC-T-GTTTGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGAACATGGTATATGTTTTA-TT-TTTGC-GTCAT
P. bipustulata      GAATTTGTGTCTTGCTCTGTCGGTTCATCGC-T-GACTGCGGTGTTTAACTGGCATG-G-------TGCGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA--T-T---T-G-TGT
O. nigroaenea       GAATTTGTGTCTTGCTCTGTCGGTTCATCGC-T-GACTGCGGTGTTTAACTGGCATG-G-------TGCGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA--T-T---T-G-TGT
Neocollyris sp.     NNNNNNNNNNNNNNNNNNNNTCGGTTCATCGC-TAGGCATGCGGTGTTTAACTGGCACG-T-------CGTGAGACGTCTTGCCGGAAGGGCATGGTATGTGTTTTG--T-T-T-T-G-TG-
P. elegans          GAATTTGTGTCTTGCGCTGTCGGTTCATCGC-TAGGCATGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATGTTTTA--T-T-TGT-G-TG-
M. carolina         GAATTTGTGTCTTGCGCTGTCGGTTCATCGCTT-GAATGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATATTTTG-TTTAGTTTAATTT-
M. fulgida          GAATTTGTGTCTTGCGCTGTCGGTTCATCGCTT-GAATGCGGTGTTTAACTGGCATG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATATA-TTTT-ATTATGTT-ATTT-
M. nigricollis      GAATTTGTGTCTTGCGCTGTCGGTTCATCGCTT-GAATGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATGTA-TTTG-GTTATGTT-ATTT-
M. klugi            GAATTTGTGTCTTGCGCTGTCGGTTCATCGCTT-GAATGCGGTGTTTAACTGGCACG-T-------TGTGAGACGTCTTGCCGGAAGGGCATGGTATGTA-TTTG-GTTATG-T-TTTA-
O. californicus     GAATTTGTGTCTCGCGCTGTCGGTTCATCGCGT-GCGTGCGGTG-TTAACTGGCACG-C-------CGCGGGACGTCTTGCCGGAAGGAC-CGGTATGTG-TTTGCTTCGTGTCGTTCTT
A. baroni           GAATTTGTGTCTCGCGCTGTCGGTTCATCGCGT-GCGTGCGGTG-CTAACTGGCACG-C-------CGCGGGACGTCTTGCCGGAAGGAC-CGGTATGTGTTTCGCTTCGTGTCGTTCTT
P. fallaciosa       GAATTTGTGTCTCGCGCTGTCGGTTCATCGCGT-GCGTGCGGTG-TTAACTGGCACG-C-------CGCGGGACGTCTTGCCGGAA-GAC-CGGTATGTGTTTTTCAACGTGTCATTCTT
P. californicus     TGATTTGTGTCTCACGCTGTTGGTCCA---C-T-GCGCGCAGTG-TTAACTGATATG-A-------CGTGAGACGTCTAGCCGG-TGAATTCAACAAGGTTGCAGTTTTTCATTTTCTAT
E. cupreus          GAATCTGTGTCCCGCGCTGTCGGTTCA---C-G-GCGCGCGGTGTTTAACTGGCACGTC-------CGCGGGACGTCCTGCCGG-TGGCCCGTCCTGTGCGTTTCGGCG----------
C. punctatoauratus  GAATCTGTGTCCCGCGCTGTCGGTTCA---C-G-GCGCGCGGTGTTTAACTGGCACGAT-------CGCGGGACGTCCTGCCGG-TGGATGTGTGTGTCAGTGCATTTCGGTG-------
T. molitor          GAATCTGTGTCCCGCGCCGCCGGTTCA---T-G-GTTGGCGGT-GTTAACTGGCGTG-C-------CGCGGGACGTCCTGCCGG-TGGGCTTAGCTCGTGA------------------
D. melanogaster     GAACTTGTGCTTCATACGGGTAGTACAACTTACAATTGTGGTTAGTACTATACCTTTATGTATGTAAG-----CGTATTACCGG-TGGAGTTCTTATATGTGATTAAATAC---------
                          |-------------------------I (E23-1)--------------------------------|  |-IX-|  |---------------II---------------
                                                                                                |-----XI----|

C. repanda          A----ACA---TAA-----T--GTACA-T--TA---ATC-A--T-A-A--T--A-A---TTTATAT--TTAA-AT-ATAGAGTGT-AAAAACT---TTAT----ATTTTAT-A-T-TA-T
C. dorsalis         CA-A-ACA--TAAT-T-------TGTGC--ATA---ATC-A-TA-T-A--AT-G-A---TATATAT--TTAA-AT-AT-AAAG-TGTAAAAGC--TTTATAT---TT---T---T--TA-
C. equestre         A----ACA---TAA-----T--GTACA-T-AGA---TTC-A--T-A-A--T--A-A---TTTATAT--ATTA-AT-ATAGAGT-T-AATAGCT---TTAT----ATTTTAT-A-T-TA-T
C. parryi           CA-A-GCATTTTAT-T----A-ATATGC--ATA---ATC-A-TA-A-T--AT-A-T---TAT-TAT--TCAA-AT-AT-AGAG-TGTAAAAGC--TTTATAT---TTGG-T-A-T-GTA-
C. frontalis        C----GCA-T-TAA-----T-GGCACATTGGTC---ATA-A-CA-A-A--A--A-A---TTTATAT--TTAA-AT-ATAGAGTGT-AAAAGCT---TTGT----ATTTTAT-ATTATATT
C. consobrina       T----TCA---TAA---------TATA-T--CA---CT--A--T-A-A--T--G-A---TTGAT-T--TTAA-AT-AT--AGT-T-AATAGCT---GTAT-----TTTGAT---T-GA-T
O. cayennensis      T----TCA---TAA---------TATA----CA---CT--A--T-A-A--T--G-A---TTGAT-T--TTAA-AT-AT--AGT-T-AATAGCT---GTAT-----TTTAAT---T-GA-T
O. confusa          T----TCA---TAA---------TATA----CG---CT--A--T-A-A--T--G-A---TTGAT-T--TTAA-AT-AT--AGT-T-AATAGCT---GTAT-----TTTAAT---T-GA-T
C. auripennis       T----GCA---TAA-----TATGCACTAT-ATA---ATG-A--T-T-G--A--T-T---TTT-TAT--TTTT-TT--CATAGT-T-AATAGCTA-ATTGTG---ATA-TAT-A-TAAATT
P. rugipennis       T----TCA---TAA---------TATG----CA---CT--A--T-A-A--T--G-A---TTG---A--TTTT-AT-AT--AGT-T-AATAGCT---ATA------TT-AAT---T-GA-T
P. discrepans       T----GCA---TAA-----TATGCACTAT-ATA---ATG-A--T-T-G--A--T-T---TTT---T--ATAT-AT--TATAGT-T-AATAGCT---TAA------TA-TAT--TTGAATT
O. gloriola         T----TCA---TAA---------TATG----CA---CT--A--T-A-A--T--G-A---TTG---T--ATGT-AT--T--GTT-T-AATAGCT---TTA------TA-TAT---T-GA-T
P. singularis       AATATGCACTATAA-T------ATATGG--ATAT--TTT-T-TA-TGT--AT-G-TG--TGT-TGT--GTTA-GT-TC-CTTT-TACTAGGGC-TTTTACAT--ACACA-C-A-A-TTA-
P. fairmairei       T----GCA---TGA-----TATGCACT---ATA---AT--A--T-G-A--T--T-A---TTT-T-T--ATAT-AT-AT-TGTT-T-AATAGCG---GTA------TA-TAT-ATTATA-T
P. alluaudi         T----GCA---TGA---------TATG----CA---CT--A--T-A-A--T--A-T---GTTAT-T--TTGT-AT-AT-TAGT-T-AATAGCT---GTAT----ATATTTT-A-T-GA-T
P. fulvia           T----TCACTCTAA-----TATTCACTATGATA---TG--A--T-T-A--T--T-T---TGTGT-A--TTTT-AT-AT--AGTGT-AATAGCTA-TATAT----ATA-TAT-ATTATA-T
P. bipustulata      A----AAA---TAT---------AATA----TA---TGC-A--C-G-G--T--A-T---TTA---A--TATT-AT-AT--AGT-T-AATAGCC---ATA------TA-AAT-A-T-TA-T
O. nigroaenea       A----AAA---TAT---------AATA----TA---TAC-A-CT-G-G--T--A-A---TTA---A--TATT-AT-AT--AGT-T-AATAGCT---ATA------TT-TAT-A-T-TT-T
Neocollyris sp.     T----TC-----AG---------TGCA----TA---TT--A--A-T-T--T--A-T---ATA---T--ATAT-AT--G--GTT-T-AACAGCC---TA-------TT-G-----T-TA--
P. elegans          T----CCA---TAG---------TAAT----TA---CT--A--T-T-T--A--T-T---ATA-T-T--ATAT-TG--G-TGTG-T-AATAGCC---CT-------TT-A-----T-AA--
M. carolina         TA-ATTAATGTATA-T-------GAGTG--TTA---AAA-A-TA-T-T--AT-T-A---TTAATAA--GGGT-GT-AT-ACAT-CATTGTGTG--TGTGCAT---TT---C---T--TA-
M. fulgida          TA-ATTAATGTATA-T------GTTGTG--TTA---AAA-A-TA-TTA--AT-A-T---AATGTAT--TGTT-GAGTG-GTTG-TGTAAAAAC--TCCGC-T---CA---C---T--TA-
M. nigricollis      TA-A-TTA--ATGA-T------ATGTGG--A-A---AAA-T-TT-A-T--AT-A-T---TTA-T-A--TTAT-AA-AT-AGGT-TGTAAAAGG--CCTAT-T---TA---T---A--TA-
M. klugi            T----TTT--AATT-T-------AATGG--T-G---ATG-T-TG-A-A--A--A-A---TAT-T-A--ATGT-AT-AT-GGTG-TG-TAAAAG--CCTGT-A---TA---T---A--T--
O. californicus     GATGTCCATGTTAC-T----CTAATGTC---CAT--AAT-T-TT-A-TG-AT-T-TG-TTGT-TGTG-GTAT-AT-TT-ATAT-T-TTAAACG-GTATAT-T--ATATA-T-T-A--CA-
A. baroni           AACGTCCATGTACTCTAAATATATATAC-GTTA-AATTTATGTTATGTGCATGCACGTGTGTACGT-TTCATCGA-TC-ATATTAAAAAAATGATTCGATGTTCGTTTATCGTGTGTCAT
P. fallaciosa       AACGTCCATATTACATAAA-ATATTTTCATTTATAAAATAT-TTGATTG-TTGTATGTTTGTGTGTGTATAT-ATGTT-ATTTGTGTTTGGTG-TTTTGTGT@CCACACAT-T-A-CCA-
P. californicus     ------------------------------------------------------------------------------------------------------------------------
E. cupreus          ------------------------------------------------------------------------------------------------------------------------
C. punctatoauratus  ------------------------------------------------------------------------------------------------------------------------
T. molitor          ------------------------------------------------------------------------------------------------------------------------
D. melanogaster     ------------------------------------------------------------------------------------------------------------------------
                    ------------------------------------------------------------II----------------------------------------------------------

C. repanda          ----TGTTA---TTG-AT-T--TTA-TG-T--TA---CATTATGTT-TT-A-TG-GA--C-CAT-ATAAT-A-ACGTATACGTAGTCCGAC----TTTTTTT----AA--AAATCCTGTC
C. dorsalis         -TA--ATTG---TTA-TCAT--GA-T-TTTTGTAA--C-AATTATG-TTTAATT-GAC-C--AT-ATAAT-A-ACATATACGTAGTCCGAC-----TTTTTTA-A-ATAAAAATCCTGTC
C. equestre         ----TGTTA---TTGAAC-T--TTA-TGTT--TA---CATTATG-T-TA-A-TG-GA--C-CAT-ATAAT-A-ACGTATACGTAGTCCGAC----TTTTTTT----AA--AAATCCTGTC
C. parryi           -TATTATTG--ATTT-TTGT--TA-TATTATTTAATGT-AAAAATT-CGTAAGT-GACGCATAT-ATAATAATACATGTACGTAGTCCGACTTTTTTTTTTTATAGAAAAAAATCCTGTC
C. frontalis        G-A-TGTTA---TTG-AT-T--CTT-TG-T--AA---TATTATGTG-AA-C-AG-GA--C-CTT-ATATTAA-ATGTATACGTAGTCCGAC----TTTTTTT----AA--AAATCCTGTC
C. consobrina       ----TGTTA---TT--A--G--TT--TG-T--TA---TATTATG---TT---TG-G---C-CTG-ATAAT-A-ACGTGTACGTAGTCCGAC------TTTTT-----A--AAATCCTGTC
O. cayennensis      ----TGTTA---TT--A--G--TT--TG-T--TA---TATTATG---TT---TG-G---C-CTG-ATAAT-A-ACGTGTACGTAGTCCGAC------TTTTT-----A--AAATCCTGTC
O. confusa          ----TGTTA---TT--A--G---TT-TG-T--TA---TATTATG---TT---TG-G---C-CTG-ATAAT-A-ACGTGTACGTAGTCCGAC------TTTTT-----A--AAATCCTGTC
C. auripennis       G-ATTGTTA---TT--A--G---TT-TG-T--AA---TATTATGTT-TT---TG-G---C-CTT-ATAAT-A-ACGTGTACGTAGTCCGAC----TTTTTGA----AA--AAATCCTGTC
P. rugipennis       ----TGTTA---TT--T--G---T--TG-T--TT---TATTATG---TA---TG-G---C-CTG-ATAAT-A-ACGTATACGTAGTCCGAC------TTTTT-----A--AAATCCTGTC
P. discrepans       G-ATTGTTT---AT--A--G---GT-TA-G--TA---TATTATGTT-TT---CG-G---C-CTT-ATAAT-A-ACGTGTACGTAGTCCGAC----TTTTTGA----AA--AAATCCTGTC
O. gloriola         ----TGTTA---TT--A--G---T--TG-T--TA---TATTGTG---TA---TG-G-----CAG-GTAAT-A-ACGTGTACGTAGTCCGAC------TTTTT-----A--AAATCCTGTC
P. singularis       -TA-TATAA--TTAA-TTGA--AATTATTATTAGTTGT-TTATATTATGTAATTAGAC-C--TT-ATAAT-A-ACGTGTACGTAGTCCGAC------TTTTT-----T-AAAATCCTGTC
P. fairmairei       G-ATTATTA---TT--A--G---T--TG-T--TA---TATTGTG---TT---TT-GA--C-CTG-ATAAT-A-ATGTGTACGTAGCTCGAC------TTTTT-----A--AAATCCTGTC
P. alluaudi         ----TATTA---TT--A--G---T--TG-T--TA---TATTGTG---TT---TT-GA--C-CTG-ATAAT-A-ACGTGTACGTAGCTCGAC------TTTTT-----A--AAATCCTGTC
P. fulvia           G-ATTATTA---TT--A--G---T--TG-C--TT---TATTGTGTG-TT-A-TT-GA--C-CATAATAAT-A-ATGTGTACGTAGTTCGAC------TTTTT-----A--AAATCCTGTC
P. bipustulata      ----T-TTA---TTG-T--T---G--TT-T--TA---TATTATA---TT---GC-G---C-CGG-ATAAT-A-ACGTGTACGTAGTCCGAC--------TTT-----A--TAATCCTGTC
O. nigroaenea       --A-T-TAT---TTG-T--T---G--TT-T--TA---TATTATA---TT---GC-G---C-CAG-ATAAT-A-ACGTGTACGTAGTCCGAC--------TTT-----A--TAATCCTGTC
Neocollyris sp.     ------TTT---A---T--T---T--TATA--TG---T--TA-C---TG---AC-------CTG-GC-AA-G-ACGTGTACGTAGTCCGAC---------TT-----A--TAATCCTGTC
P. elegans          ------TAA---A---T--T---TA-TGTG--TG---TACTATG---TG---GC-------CGG-AT-TT-A-ACGTATACGTAGTCCGAC---------TT-----A--TAATCCTGTC
M. carolina         -TTTTATAT---TAT-ATTA--TA-TAAGCTCATA--A-TATCATTATTAAAGT-TA--C--TA-ACAAT-A-ATGTGTACGTAGTCCGAC--------TTT-----T--CAATCCTGTC
M. fulgida          -CATTTTTA--TTAT-ATTA--TA-TAAGC-TCAT--A-TATCATT-ATTAAGA-TA--C--TA-ATAAT-A-ATGTATACGTAGTCCGAC--------TTA-T---T--TAATCCTGTC
M. nigricollis      --T--ATTA---TAT-TTAT--TA-T-ATA-AACA--T-ATTCATT-ATTAAGA-TA--C--TA-ACAAT-A-ATGCGTACGTAGTCCGAC--------ATT-----T--CAATCCTGTC
M. klugi            --T--ATTT---ATT-AC-A--TA---TGC-ATTA--T-CATTGTTAAGATATT-T---C--TA-AC-TT-G-ATGCGTACGTAGTCCGAC---------TT-----C--TAATCCTGTC
O. californicus     -CA-CATTTC--TTA-TTAT--GGGT-TCA--TATTGT-CATGGCTCCATCATGCTTCGG-CCG-GC-TT-A-GTGCGTACGTAGTCCGAC---------TT-----T---AATCCTGTC
A. baroni           GTATCATTTCATTATTTTATACGTTTTATAAATATTGT-CATGGCTCCGTTTCGCTTCGG-CCG-GC-TT-A-GTGCATACGTAGTCCGAC---------TT-----T--CAATCCTGTC
P. fallaciosa       -CA-CACATC-ATTATTTATAAGAAT-GGATTTATTGTATTTGGCTCCGTTATGCTTTGG-CCG-GC-TT-A-GTGCATACGTAGTCCGAC---------TT-----T--CAATCCTGTC
P. californicus     ----------------------------------------ATATTTAAATAAATTTTTAATGATAACTTGTAAACGGTTGAATTCAAC------TTTATGTTTATAAAGATCCTGTC
E. cupreus          ------------------------------------------------------------------------TACGGCGGCCCGAC---------TC-----A--CAATCCCGCT
C. punctatoauratus  ---------------------------------------------------------------------TATCTGCTCCGCATCCGAC---------TC-----AACAAATCCTGT-
T. molitor          ---------------------------------------------------------------------------GGGCGGCCCAAC---------TC---------AATCCCGCC
D. melanogaster     --------------------------------------------------------------TTGTATTTTTTCATATGTTCCTCCTAT---------TT-------AAAAACCTGCA
                    -------------------------------------II-------------------------------------------------|                  |--III
```

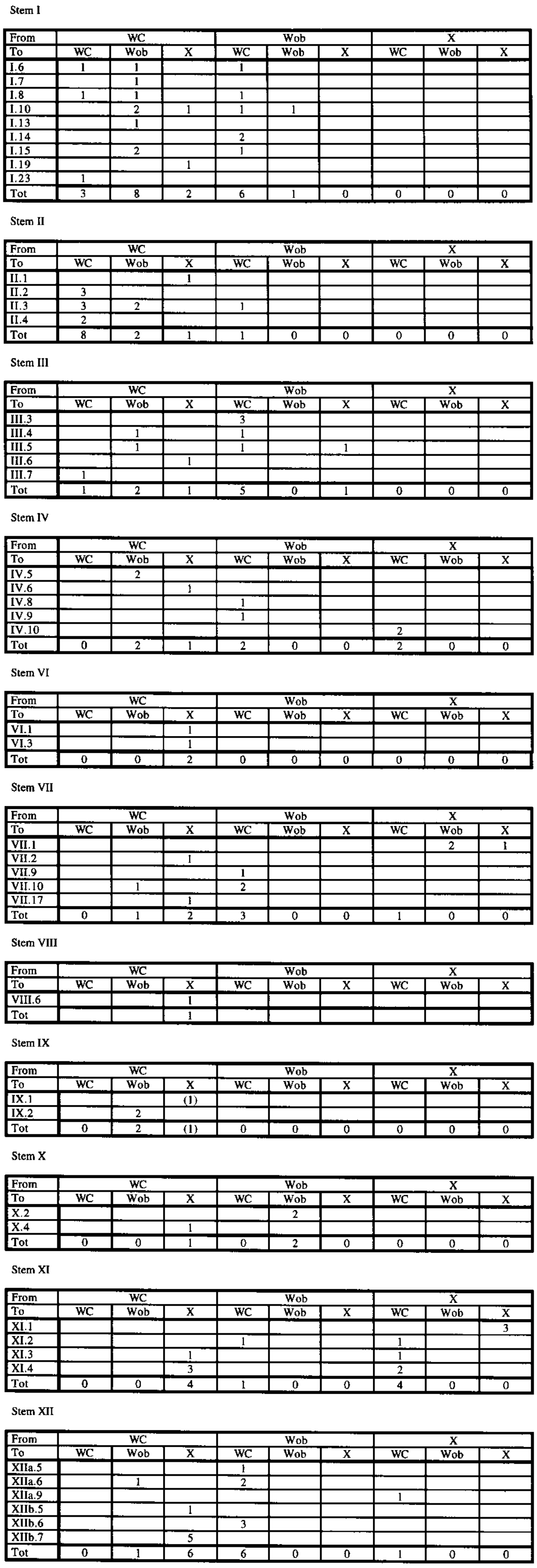


```
C. repanda          TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
C. dorsalis         TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
C. equestre         TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
C. parryi           TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
C. frontalis        TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
C. consobrina       TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
O. cayennensis      TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
O. confusa          TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
C. auripennis       TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
P. rugipennis       TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
P. discrepans       TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
O. gloriola         TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
P. singularis       TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
P. fairmairei       TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGTACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
P. alluaudi         TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
P. fulvia           TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGTACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TAATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
P. bipustulata      GAGGTGCTCGTTAACTGAGTGTCGTCGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TAATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
O. nigroaenea       GAGGTGCTCGTTAACTGAGTGTCGTCGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TAATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
Neocollyris sp.     TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TAATGT-TGCCTG-ATATAGTGTGCATGGAATAATGGAACAGCACC
P. elegans          TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TAATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
M. carolina         TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
M. fulgida          TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
M. nigricollis      TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
M. klugi            TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTATGT-TGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
O. californicus     TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TATTGT-TGCCTGAATACAGTGTGCATGGAATAATGAAACAGCACC
A. baroni           TAGGTGCTCGTTAACTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TATTGT-TGCCTAAATACAGTGTGCATGGAATAATGGAACAGCACC
P. fallaciosa       TAGGTGCTCGTTAATTGAGTGTCGTGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TATTGT-TGCCTGAATACAGTGTGCATGGAATAATGGAACAGCACC
P. californicus     ACAGTGCTCGTTGACCGAGTGTTTTGCCAGGCCGGCCCATTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGCAAAAATTTACGCCTGAATAATATGTGCATGGAATAATAGAATAGCACC
E. cupreus          TCGGTGCTCGTTGACTGAGTGTCGAGGTGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTCAAAGCAGGC-CGAAATCACGCCTGAATACTGTGTGCATGGAATAATGGAACAGCACC
C. punctatoauratus  GCGGTGCTCGTTAATTGAGTGTCGAGGCGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTCAAAGCAGGC--TAAATT-AGCCTGAATATAGTGTGCATGGAATAATGAAACAGCACC
T. molitor          GCGGTGCTCGTTCGTTGAGTGTCGAGGTGGGCCGGCACGTTTACTTTGAACAAATTAGAGTGCTTAAAGCAGGCTAAAACTT-CGCCTGAATACTGTGTGCATGGAATAATGGAATAGCACC
D. melanogaster     TTAGTGCTCGTTAAACGAGTGTTATTGTGGGCCGGTACTATTACTTTGAACAAATTAGAGTGCTTAAAGCAGGC--TTCAAA-TGCCTGAATATTCTGTGCATGGGATAATGAAATAACACC
                    -------III (E23-5)-----------|.---IV---||----V (E23-6)------.  |---------VI (E23-7)--------|       |-------VII--------
                                                  IX'-
                                                                                                            |---IV'---|
                                                                                                               |-----------X--------| |-
                                                                                                                          |------XII-----

C. repanda          TCGGTTCT-ATTTTGTTGGTTTTGCTCATC-AAACGAGGTAATGATT
C. dorsalis         TCGGTTCT-ATTTTGTTGGTTTTGCTCATC-AAACGAGGTAATGATT
C. equestre         TCGGTTCT-ATTTTGTTGGTTTTGCTCATC-AAACGAGGTAATGATT
C. parryi           TCGGTTCT-ATTTTGTTGGTTTTGCTCAACTGAACGAGGTAATGATT
C. frontalis        TCGGTTCT-ATTTTGTTGGTTTTGCTCAAC-TGACGAGGTAATGATT
C. consobrina       TCGGTTCT-ATTTTGTTGGTTTTGCTCATC-AAACGAGGTAATGATT
O. cayennensis      TCGGTTCT-ATTTTGTTGGTTTTGCTCAACTGAACGAGGTAATGATT
O. confusa          TCGGTTCT-ATTTTGTTGGTTTTGCTCAACTGAACGAGGTAATGATT
C. auripennis       TCGGTTCT-ATTTTGTTGGTTTTGCTCAACTGAACGAGGTAATGATT
P. rugipennis       TCGGTTCT-ATTTTGTTGGTTTTGCTCAACTGAACGAGGTAATGATT
P. discrepans       TCGGTTCT-ATTTTGTTGGTTTTGCTCAAC-TGACGAGGTAATGATT
O. gloriola         TCGGTTCT-ATTTTGTTGGTTTTGCTCATC-AAACGAGGTAATGATT
P. singularis       TCGGTTCT-ATTTTGTTGGTTTTGCTCAACTGGACGAGGTAATGATT
P. fairmairei       TCGGTTCT-ATTTTGTTGGTTTTGCTCAACTGAACGAGGTAATGATT
P. alluaudi         TCGGTTCT-ATTTTGTTGGTTTTGCTCAACTGAACGAGGTAATGATT
P. fulvia           TCGGTTCT-ATTTTGTTGGTTTTGCTCAACTGAACGAGGTAATGATT
P. bipustulata      TCGGTTCT-ATTTTGTTGGTTTTGCTCATC-AAACGAGGTAATGATT
O. nigroaenea       TCGGTTCT-ATTTTGTTGGTTTTGCTCATC-AAACGAGGTAATGATT
Neocollyris sp.     TCGGTTCT-ATTTTGTTGGTTTTGCTCAAC-TTACGAGGTAATGATT
P. elegans          TCGGTTCT-ATTTTGTTGGTTTTGCTCAAC-TAACGAGGTAATGATT
M. carolina         TCGGTTCT-ATTTTGTTGGTTTTGCTCAAC-TAACGAGGTAATGATT
M. fulgida          TCGGTTCT-ATTTTGTTGGTTTTGCTCAAC-TAACGAGGTAATGATT
M. nigricollis      TCGGTTCT-ATTTTGTTGGTTTTGCTCAAC-TAACGAGGTAATGATT
M. klugi            TCGGTTCT-ATTTTGTTGGTTTTGCTCAAC-TAACGAGGTAATGATT
O. californicus     TCGGTTCT-ATTTTGTTGGTTTTCGGCAA--TTACGAGGTAATGATT
A. baroni           TCGGTTCT-ATTTTGTTGGTTTTCGGCAA--TTTCGAGGTAATGATT
P. fallaciosa       TCGGTTCT-ATTTTGTTGGTTTTCGGCAA--TTTCGAGGTAATGATT
P. californicus     TCGATTCT-ATTTTGTTGGTTTTCACAA--TTAAGAGGTAATGATT
E. cupreus          TCGGTTCT-ATTTTGTTGGTTTTCGGCAA--TTCCGAGGTAATGATT
C. punctatoauratus  TCGGTTCT-ATTTTGTTGGTTTTGCTCATC-AAACGAGGTAATGATT
T. molitor          TCGGTTCT-ATTTTGTTGGTTTTCGGCAA--TTTTGAGGTAATGATT
D. melanogaster     TCTGTTCTGCTTTCATTGGTTTTCAG-A--TCAAGAGGTAATGATT
                    ---------VII------------|
                    ------------------VIII-----------------|
                    ----------XII (E23-8/9)--------------|
```

Figure 4 (above and opposite). Multiple alignment of cicindelid SSU rRNA V4 sequences plus *D.melanogaster* and *T.molitor* sequences (9). The positions of the structures identified in Figure 3 are indicated below the alignment. Because these potential structures overlap, individual structures have been coded using bold text, italics and single or double underlining. Loop regions of structures included in the final model (Fig. 3) are shown in outline text. The type of coding used for each potential structure is indicated on the stem notation below the alignment. Hyphens (-) indicate gaps introduced to optimize the alignment, except in the lines indicating structures, where they indicate the extent of the structure concerned.

Table 2. Frequencies of changes within secondary structural elements



Stem I

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
I.6	1	1		1					
I.7		1							
I.8	1	1		1					
I.10		2	1	1	1				
I.13		1							
I.14				2					
I.15		2		1					
I.19			1						
I.23	1								
Tot	3	8	2	6	1	0	0	0	0

Stem II

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
II.1			1						
II.2	3								
II.3	3	2		1					
II.4	2								
Tot	8	2	1	1	0	0	0	0	0

Stem III

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
III.3				3					
III.4		1		1					
III.5		1		1		1			
III.6			1						
III.7	1								
Tot	1	2	1	5	0	1	0	0	0

Stem IV

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
IV.5		2							
IV.6			1						
IV.8				1					
IV.9				1					
IV.10							2		
Tot	0	2	1	2	0	0	2	0	0

Stem VI

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
VI.1			1						
VI.3			1						
Tot	0	0	2	0	0	0	0	0	0

Stem VII

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
VII.1								2	1
VII.2			1						
VII.9				1					
VII.10		1		2					
VII.17			1						
Tot	0	1	2	3	0	0	1	0	0

Stem VIII

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
VIII.6			1						
Tot			1						

Stem IX

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
IX.1			(1)						
IX.2		2							
Tot	0	2	(1)	0	0	0	0	0	0

Stem X

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
X.2					2				
X.4			1						
Tot	0	0	1	0	2	0	0	0	0

Stem XI

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
XI.1									3
XI.2				1			1		
XI.3			1				1		
XI.4			3				2		
Tot	0	0	4	1	0	0	4	0	0

Stem XII

From	WC			Wob			X		
To	WC	Wob	X	WC	Wob	X	WC	Wob	X
XIIa.5				1					
XIIa.6		1		2					
XIIa.9							1		
XIIb.5			1						
XIIb.6				3					
XIIb.7			5						
Tot	0	1	6	6	0	0	1	0	0

For each position of each stem in the secondary structure model shown in Figure 4, the number of each of the nine different kinds of change was counted, taking into account the phylogenetic relationships shown in Figure 2.

changes. Position 4 showed a single non-compensatory change in *P.californicus*.

Stem XI. This could form immediately 5′ to stem II in 18 species, replacing the base of stem II. The stem showed numerous non-compensatory changes and point deletions in species in which it was not predicted to form and was missing from all the basal cicindelids and the outgroups (Fig. 4).

Stem XII. This was the pseudoknot structure predicted by Neefs and De Wachter (14,21). The base-paired segments of this structure were very similar to those making up stems VII and VIII predicted from energy criteria. For purposes of covariation analysis, the structure was divided into its two component stems, labelled XIIa and b in Figure 3. XIIa corresponded to part of stem VII and XIIb to stem VIII. Stem XIIa was invariant at positions 1–3, 7, 8 and 10. Positions 4–6 showed numerous semi-compensatory changes. The looped out position 9 showed non-compensatory changes. In *D.melanogaster*, an additional base was present next to position 9. Only three of the nine positions in stem XIIb varied, although this region contained insertions of one to three bases in all the species considered here. Position 6 showed exclusively semi-compensatory changes, with one semi-compensatory and one non-compensatory change at position 5. Position 7 showed only non-compensatory changes with at least six changes in state.

Analysis of full compensatory mutations

Table 1 shows that 10 base pairs in three of the predicted structures (stems I, II and III) showed fully compensatory changes within the Cicindelidae, providing *prima facie* evidence for their existence. In addition, stem IX showed a fully compensatory change at position 2 on branch 61 of the tree, which separates Tenebrio from the Cicindelidae. The majority of fully compensatory changes were seen as complete transformations between base pairs (e.g. a G–C to A–U conversion at position 3 of stem II in *P.fulvia* compared with its sister taxon *Physodeutera alluaudi*). However, we also observed three compensatory changes that, according to the reconstruction of character changes, proceeded via two semi-compensatory steps (i.e. via G.U intermediates), at positions 1 and 10 of stem I, and position 3 of stem II.

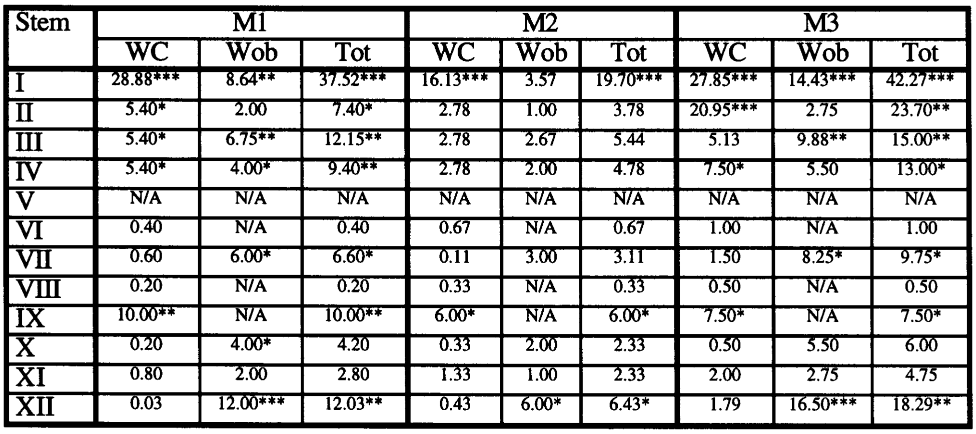
Statistical analysis of patterns of nucleotide change

A number of the commonly found potential stems did not show fully compensatory changes, and for stems III and IX, the fully compensatory changes that were observed occurred at the root of the tree only, raising the possibility that these structures were not conserved within the Cicindelidae. We therefore carried out a statistical analysis of the pattern of base changes within the Cicindelidae to investigate whether these patterns provided any additional evidence for or against the existence of any of the other structures suggested by MFOLD analysis.

The aim of the statistical analysis was to test whether the patterns of nucleotide changes observed on the phylogenetic tree deviated significantly from random expectation. We defined our random expectations according to three null models, all of which assume that the probabilities of changes occurring between different base combinations are independent between sites (see Materials and Methods).

To test the data against these models, we counted only base changes that were unambiguous on the phylogenetic tree (Fig. 2).

Table 3. Results of χ^2 analysis of frequency data from Table 2



Stem	M1			M2			M3		
	WC	Wob	Tot	WC	Wob	Tot	WC	Wob	Tot
I	28.88***	8.64**	37.52***	16.13***	3.57	19.70***	27.85***	14.43***	42.27***
II	5.40*	2.00	7.40*	2.78	1.00	3.78	20.95***	2.75	23.70**
III	5.40*	6.75**	12.15**	2.78	2.67	5.44	5.13	9.88**	15.00**
IV	5.40*	4.00*	9.40**	2.78	2.00	4.78	7.50*	5.50	13.00*
V	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
VI	0.40	N/A	0.40	0.67	N/A	0.67	1.00	N/A	1.00
VII	0.60	6.00*	6.60*	0.11	3.00	3.11	1.50	8.25*	9.75*
VIII	0.20	N/A	0.20	0.33	N/A	0.33	0.50	N/A	0.50
IX	10.00**	N/A	10.00**	6.00*	N/A	6.00*	7.50*	N/A	7.50*
X	0.20	4.00*	4.20	0.33	2.00	2.33	0.50	5.50	6.00
XI	0.80	2.00	2.80	1.33	1.00	2.33	2.00	2.75	4.75
XII	0.03	12.00***	12.03**	0.43	6.00*	6.43*	1.79	16.50***	18.29**

χ^2 values are shown individually for changes from Watson–Crick (WC) or wobble (Wob) base pairs, as the models predict relative frequencies of changes within but not between these classes of change. The value Tot is the pooled value of χ^2.

M1, Model 1; assuming single base changes. one degree of freedom for WC and Wob only analyses; two degrees of freedom for pooled value.

M2, Model 2; as Model 1, but assuming a 2:1 transition:transversion ratio.

M3, Model 3; not assuming single base changes. two degrees of freedom for WC and Wob only analyses; four degrees of freedom for pooled value.

**P* <0.05; ***P* <0.01; ****P* <0.001.

In addition we excluded *T.molitor* and *D.melanogaster* from the analyses, both to test for structures that might be specific to the Cicindelidae and because their great evolutionary distance from the Cicindelidae meant that we could not exclude multiple changes giving rise to apparent single-base changes. Table 2 summarises the numbers and types of changes observed for each of the variable stem positions. Table 3 gives the results of χ^2 analysis of these frequencies against the three null models. The analyses under models 1 and 3 provided similar, but not identical, patterns of support. Stem I showed the strongest level of support (P <0.001), while stems II, III, IV, IX and XII reached the P <0.01 level under at least one model. Stems VII and X reached only low levels of significance (P <0.05). Model 2, which represents a modification of model 1 in which the transition:transversion ratio is set to 2, supported only stem I strongly, and stems IX and XII weakly.

DISCUSSION

Our aim was to examine the utility of a method for modelling the secondary structures in highly variable regions where alignment and the determination of positional homology are unreliable. We tested the method against a well-established structural model and investigated levels of support for the resulting model using comparative phylogenetic data. We first discuss the secondary structure model that our analyses produced, then the method's usefulness for modelling variable region structure.

Secondary structure model

Of the eight elements of secondary structure contained in the current structural model of V4 (14), six were among the structures found most commonly in MFOLD analyses of cicindelid sequences (Fig. 3). E23-1, -2, -5, -6, -7 and -8/9 correspond to our structures I, IX, III, V, VI and XII, respectively. Of these, stem I was strongly supported by analysis of full compensatory mutations, as was stem II. Stems III and IX were more weakly supported by compensatory mutations.

Stems I, II, III, IV, IX and XII were also supported to at least the P <0.01 level by statistical analysis under at least one of the models. Stem V (E23-6) was not testable in this set of species as this region of the sequence showed complete conservation. Stem VI (E23-7) was supported poorly by statistical analysis as it showed two non-compensatory and no other changes. In the species in which these changes had taken place (*A.baroni*, *C.punctatoaureus*, *E.cupreus*) the potential structure reduced to a conserved core corresponding to base pairs 4–7 in Figure 3, indicating that the base of this structure is either a spurious result from the minimum energy calculations or of limited taxonomic distribution.

Our study provided support for two structures that are not included in the existing model. The first of these was stem IV, which could form from sequences flanking stems V and VI to form a Y-shaped structure. This overlapped the 3′ strand of stem IX (E23-2), which was supported by a compensatory change at the root of the tree. Both stems IV and IX were supported by statistical analysis, which was therefore unable to unambiguously distinguish between the two. A fully compensatory change appears to provide stronger evidence for stem IX than stem IV. However, as this compensatory change affects a change at the outgroup node only, whereas the statistical analysis of semi-compensatory changes supports the presence of stem IV within the ingroup, we cannot exclude the possibility that stem IV rather then stem IX occurs in the ingroup. We have therefore included both structures as alternatives in Figure 3.

The second difference between our analysis and the model of Van de Peer *et al* (14,15) lay in the central, highly variable region of V4 between stems I and III. We consistently found a single, long stem to form in this region (stem II) which was supported by analyses of both fully and semi-compensatory changes. Stem II corresponds to a structure we proposed previously for this region based on a less rigorous analysis (9). Van de Peer *et al.*'s model of this region (14) contains two stems: E23-3 and E23-4. Both of these structures appear in only a small number of insect species (E23-3 is described for four species: *Acyrthosiphon pisum* (pea aphid), *D.melanogaster*, *Meloe proscarabaeus* and *T.molitor*; E23-4 only for *D.melanogaster*). These two stems are made up by the two complementary halves of our stem II in *D.melanogaster*. Stem II was observed in 26 of the species analysed here, contained numerous fully compensatory changes and reached the $P < 0.01$ level of support based on only four base pairs. As none of the 197 minimum or near-minimum energy structures we considered during modelling gave rise to a pair of stems corresponding to the two halves of stem II, we saw no evidence to support the double structure in the cicindelid rRNA. Its recognition in *D.melanogaster* may reflect a restricted taxonomic distribution; statistical analysis of covariation at lower hierarchical levels, for example within closely related drosophilids, may provide the evidence to decide between competing stem structures. However, as our stem II has homologues in *D.melanogaster* and *T.molitor*, and given the very restricted taxonomic distribution of stems E23-3 and E23-4, we believe there is strong *prima facie* evidence for a long, variable stem in the central region of insect V4.

Although stem XII was clearly homologous to the pseudoknot structure described previously (14,21), position 9 in stem XIIa and positions 5 and 7 in stem XIIb did not show strongly conserved patterns of pairing (Table 1). Position 9 of stem XIIa shows two non-compensatory changes: one in *C.repanda* and one in the lineage leading to *E.cupreus* and *C.punctatoauratus*. In *D.melanogaster* this site also neighbours a single base insertion that would have to be looped out of the structure. This position within stem XIIa is therefore not necessarily paired and can accommodate mispairing and looping out of at least one base. Position 7 of stem XIIb shows much less tendency to be paired than other positions within stem XII, being transformed to a U.U mispairing in numerous cases. This again suggests that this position is not necessarily paired. The fact that a base insertion is observed in *D.melanogaster* immediately neighbouring position 7 suggests that extrahelical bases can be tolerated in this region of the structure. Position 5, which flanks a bulged motif that also varies in length, also showed a single non-compensatory change, from G–C to G.A. However, it remains possible that these positions adopt atypical helix conformations, as C.A, U.U and G.A pairs can form under some circumstances (23).

Taking the above considerations into account, our model for the V4 region of the Cicindelids is as presented in Figure 3. We regard stems I–III and XII as well supported, but the structure of the region made up of stems V–VI is presented only for illustration, as these structures were either untestable or poorly supported by our analyses. We have included both stems IV and IX although the longer range compensatory changes support the existence of stem IX in the broader sample of V4 structures. It is noteworthy that the structures included in the model all occurred in ≥26 of the 29 MFOLD predictions (~90%) whereas stems VII, VIII and XI, which were excluded, occurred in fewer predictions (21–24). Thus the frequency of occurrence of any structure in MFOLD predictions may be a direct indicator of the likelihood of its occurrence. Stem X was intermediate, occurring in 25/29 predictions (86%). This structure is excluded by stem IV, with which it overlaps, but not by stem IX, and may occur if stem IX occurs.

Three other models of this region have been proposed: Hancock *et al.*'s original model for *D.melanogaster* (24), an alternative comparative model (25) and a model for the highly expanded V4 in *Acyrthosiphon pisum* (26). The model developed in this paper shares stems II, III and V with Hancock *et al.*'s model (24), which also identified stems VI, VII and VIII as possible structures. Nickrent and Sargent's model (25) clearly shares stems III and V and a stem similar to VI with the current model; other stems in their model resemble our stems I, VIII and X although homology is difficult to ascertain as the structures published were for species highly divergent from insects. Similarities to the Kwon *et al.* model (26) are not obvious.

Statistical analysis

Support for secondary structures has conventionally been based on the accumulation of fully compensatory changes. A problem with this approach, which is intuitively attractive, is the difficulty in handling the occasional non-compensatory change (21) and the zero-weighting accorded to semi-compensatory changes. The statistical approach we have used here allows account to be taken of these other kinds of change and, by making predictions about the expected ratio of the different kinds of change, allows testing of the nature of sequence change in putatively base-pairing regions. Our method makes use of phylogenetic trees to infer the character changes and their directions. This permits the determination of the number and types of mutational steps, which in turn can be used for inferences about underlying mutational processes such as the action of slippage-like processes. Comparison with a model of random change then allows assessment of deviations from null expectations and a calculation of statistical support for the various secondary structures suggested from the minimum energy analysis.

We investigated three different models of change here. Models 1 and 2 assumed that semi-compensatory changes resulted from single base substitutions only, with different assumptions about the relative frequencies of transitions and transversions. Model 2, which assumes a 2:1 transition to transversion ratio, produced much lower levels of support for individual structures. This is unsurprising as the base changes that convert between Watson–Crick and wobble pairs are transitions (C↔U; G↔A). This underlines the necessity of using an appropriate model of random mutation in such studies. Model 3, which makes no assumptions about the number of base changes giving rise to a change in any given base pair, provided much stronger support for stem II than models 1 and 2. This is consistent with model 3 being a more suitable model for sets of sequences that are so rapidly evolving, or so distant, that multiple base changes are likely to have occurred during their evolution. It also provides a means of testing whether mutations within a given structure are approaching saturation, which might be of value for phylogenetic analysis.

As rRNA sequences are widely used for phylogenetic reconstruction, data sets are increasingly amenable to this analysis. Consideration of the distribution of the structures on the phylogenetic tree, and of non-compensatory changes within the structure, may provide more information about the reality of a given weakly supported structure. We applied statistical tests to whole structures observed in MFOLD modelling. This approach is necessary when data (i.e. variation) is limited to best measure the support accruing to a particular structure. However, it has the disadvantage that it cannot characterise the support for a particular paired position within a structure. Such an approach might be possible for larger, and/or more variable, datasets.

Four of the six stems supported strongly by our statistical approach (i.e. at the P <0.01 level under at least one model) also showed at least one fully compensatory change, providing evidence that the two approaches measure the same properties of sequences. However the pseudoknot stem XII showed no compensatory changes in our dataset. Despite this, it was supported by statistical analysis. Given that this structure is strongly supported in a wide range of other species, this suggests that the statistical approach is more sensitive than simple measurements of compensatory changes. This is not surprising as fully compensatory changes involve at least two coordinated base changes while changes involving G.U base pairs can involve only single base changes.

Like any other method based on sequence comparison, this method relies on sufficiently high rates of nucleotide change. Cases in which only a few changes are observed can give rise to Type II error (failure to detect a true deviation from random expectation). As well as being sensitive to low rates, the method is also sensitive to high rates of change. This is seen for the basal bases of stem II, which showed many more changes than the other structures analysed here, including many more completely compensatory mutations. This indicates high rates of sequence change in this part of the expansion segment and high rates of double mutation, and coincides with the disparity observed between the χ^2 values obtained under models 1 and 3 for the four base pairs forming the base of stem II.

In conclusion, the use of minimum energy calculations combined with analyses of compensatory and semi-compensatory changes has allowed us to use a relatively small dataset to recover many (but not all) features of a secondary structure model that is based on a wider ranging set of data. It has also revealed at least one possible difference between the broader model and the cicindelid sequences which may reflect differences in structure of variable region V4 in different evolutionary lineages, while providing evidence for a structure that forms from a highly variable part of the sequences that may not have been evident from more distant sequence comparisons (stem II). This approach may be useful in other studies in which data are limited or sequences are highly variable and difficult to align unambiguously.

ACKNOWLEDGEMENTS

We thank the UK Medical Research Council and Natural Environment Research Council (grant number 3/11362) for financial support. We also thank Donald Quicke for discussions.

REFERENCES

1 Schnare, M.N., Damberger, S.H., Gray, M.W. and Gutell, R.R. (1996) *J. Mol. Biol.* **256**, 701–719.

2 Hickson, R.E., Simon, C., Cooper, A., Spicer, G.S., Sullivan, J. and Penny, D. (1996) *Mol. Biol. Evol.* **13**, 150–169.

3 Vawter, L. and Brown, W.M. (1993) *Genetics* **134**, 597–608.

4 Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S. and Rastan, S. (1992) *Cell* **71**, 515–516.

5 Hoelzel, A.R., Hancock, J.M. and Dover, G.A. (1993) *J. Mol. Evol.* **37**, 190–197.

6 Hancock, J.M., Chaleeprom, W., Chaleeprom, W., Dale, J. and Gibbs, A. (1995) *J. Gen. Virol.* **76**, 3229–3232.

7 Hancock, J.M. and Dover, G.A. (1988) *Mol. Biol. Evol.* **5**, 377–391

8 Hancock, J.M. (1995) *J. Mol. Evol.* **40**, 629–639.

9 Vogler, A.P., Welsh, A. and Hancock, J.M. (1997) *Mol. Biol. Evol.* **14**, 6–19.

10 Zuker, M. (1989) *Science* **244**, 48–52.

11 Jaeger, J.A., Turner, D.H. and Zuker, M. (1990) *Methods Enzymol.* **183**, 281–306.

12 Zuker, M., Jaeger, J.A. and Turner, D.H. (1991) *Nucleic Acids Res.* **19**, 2707–2714

13 Zuker, M. (1994) *Methods Mol. Biol.* **25**, 267–294.

14 Van de Peer, Y., Jansen, J., De Rijk, P. and De Wachter, R. (1997) *Nucleic Acids Res.* **25**, 111–116.

15 De Rijk, P., Neefs, J.M., Van de Peer, Y. and De Wachter, R. (1992) *Nucleic Acids Res.* **20**, 2075–2089.

16 Tautz, D., Hancock, J.M., Webb, D.A., Tautz, C. and Dover, G.A. (1988) *Mol. Biol. Evol.* **5**, 366–376.

17 Hendriks, L.R., De Baere, R., Van Broekhoven, C. and De Wachter, R. (1988) *FEBS Lett.* **232**, 115–120.

18 GCG (1994) *Program Manual for the Wisconsin Package.* Version 8, August 1994. Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711.

19 Vogler, A.P. and Pearson, D.L. (1996) *Mol. Phylogenet. Evol.* **6**, 321–338.

20 Maddison, W.P. and Maddison, D.R. (1992) *MacClade: Analysis of Phylogeny and Character Evolution.* Version 3.0, Sinauer Associates, Sunderland, Massachusetts.

21 Neefs, J.-M. and De Wachter, R. (1990) *Nucleic Acids Res.* **18**, 5695–5704.

22 Maidak, B.L., Olsen, G.J., Larsen, N., Overbeek, R., McCaughey, M.J. and Woese, C.R. (1997) *Nucleic Acids Res.* **25**, 109–110.

23 Gutell, R.R., Larsen, N. and Woese, C.R. (1994) *Microbiol. Rev.* **58**, 10–26.

24 Hancock, J.M., Tautz, D. and Dover, G.A. (1988) *Mol. Biol. Evol.* **5**, 393–414.

25 Nickrent, D.L. and Sargent, M.L. (1991) *Nucleic Acids Res.* **19**, 227–235.

26 Kwon, G.-Y., Ogino, K. and Ishikawa, H. (1991) *Eur. J. Biochem.* **202**, 827–833.