Nucleic Acids Research, 2000, Vol. 28, No. 20 **3839–3845**

Comparative analysis of secondary structure of insect mitochondrial small subunit ribosomal RNA using maximum weighted matching

Roderic D. M. Page*

Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, UK

Received July 27, 2000; Revised and Accepted August 21, 2000

DDBJ/EMBL/GenBank accession no. DS43718

ABSTRACT

Comparative analysis is the preferred method of inferring RNA secondary structure, but its use requires considerable expertise and manual effort. As the importance of secondary structure for accurate sequence alignment and phylogenetic analysis becomes increasingly realised, the need for secondary structure models for diverse taxonomic groups becomes more pressing. The number of available structures bears little relation to the relative diversity or importance of the different taxonomic groups. Insects, for example, comprise the largest group of animals and yet are very poorly represented in secondary structure databases. This paper explores the utility of maximum weighted matching (MWM) to help automate the process of comparative analysis by inferring secondary structure for insect mitochondrial small subunit (12S) rRNA sequences. By combining information on correlated changes in substitutions and helix dot plots, MWM can rapidly generate plausible models of secondary structure. These models can be further refined using standard comparative techniques. This paper presents a secondary structure model for insect 12S rRNA based on an alignment of 225 insect sequences and an alignment for 16 exemplar insect sequences. This alignment is used as a template for a web server that automatically generates secondary structures for insect sequences.

INTRODUCTION

Knowledge of RNA secondary structure is becoming increasingly important in molecular phylogenetic studies, particularly in assisting accurate sequence alignment (1,2). Automatic alignment methods that use only primary sequences may misalign RNA sequences (3) and alignments that take secondary structure into consideration can generate improved phylogenetic trees (1,4). To be able to use secondary structure to align RNA sequences we need good models for the sequences from these organisms or their relatives. Detailed models exist for well-studied organisms, such as *Escherichia coli* and key vertebrate species, but for many taxonomically important groups there are few, if any, available models. In the case of insects, which are the largest single group of animals, there is only a single structure (*Drosophila virilis*) for mitochondrial small subunit rRNA in the Comparative RNA web site (http://www.rna.icmb.utexas.edu). The European Small Subunit Ribosomal RNA database (5) has 22 structures, although nearly half are for *Drosophila* species.

As the number and taxonomic scope of RNA sequences expands there is an increasing need for secondary structures for diverse taxonomic groups. While much RNA secondary structure is conserved across large evolutionary distances, particular taxonomic groups may have unique features that may cause difficulties for alignments based on secondary structures obtained from other taxa, and may also be of structural and phylogenetic interest (6). Comparative analysis is the preferred and most successful method of inferring RNA secondary structure (7,8), however, a major practical limitation of this technique is that it requires substantial manual effort. Readily available software programs simply output listings of possible pairings based on mutual information or other criteria (9,10) or display pairwise graphs of mutual information values (11). Converting this information into a structure is not a trivial task (7). In contrast, software packages such as RNADraw (12) and MFOLD (13), which predict secondary structure based on minimising free energy, can generate and display structures automatically. Although some recent software packages (14,15) combine comparative analysis with thermal energy methods, they use evidence from compensatory mutations to help select amongst minimal free energy structures, rather than generate structures directly from comparative data. Hence, comparative analysis is difficult for non-specialists to employ, at a time when the importance of secondary structure in phylogenetic analysis is becoming more widely appreciated (1,3,16). Consequently, workers tend to rely on existing models for other taxa. The validity of these models and their applicability to other taxonomic groups is often difficult to assess. Furthermore, using a model to construct an alignment is often done manually by searching for 'conserved' motifs. There is a clear

*Tel: +44 141 330 4778; Fax: +44 141 330 5971; Email: r.page@bio.gla.ac.uk

need for objective techniques for constructing secondary structure models and for constructing structural alignments based on those models.

Maximum weight matching (MWM) (17,18) is a graph-theoretic approach to inferring RNA secondary structure that shows considerable promise. It takes as input a set of base pairing scores that can be derived from a range of sources, such as free energy considerations, mutual information or experimental data. These data can be represented as a folding graph where the vertices are alignment positions and the lines or edges connecting a pair of vertices are given a weight proportional to the amount of evidence for that pairing. A matching is a subgraph of the folding graph in which no vertex is connected to more than one other vertex. The matching with the greatest total edge weight is taken to be the best estimate of the RNA structure. This matching can include information on both secondary and tertiary structure, such as pseudoknots. MWM can greatly speed up the process of comparative structure analysis of RNA sequences and has been tested successfully on tRNA, SRP RNA and 16S rRNA sequences (18).

This paper uses MWM to develop a secondary structure model for domain III of insect mitochondrial small subunit (12S) rRNA. This portion of 12S rRNA is often sequenced in insect molecular phylogenetic studies due to the availability of conserved PCR primers (19). Sequences were obtained from the EMBL database and a secondary structure model inferred using the MWM algorithm. This model is used as a template to automatically generate structures for other insect sequences using the RNAlign program (20).

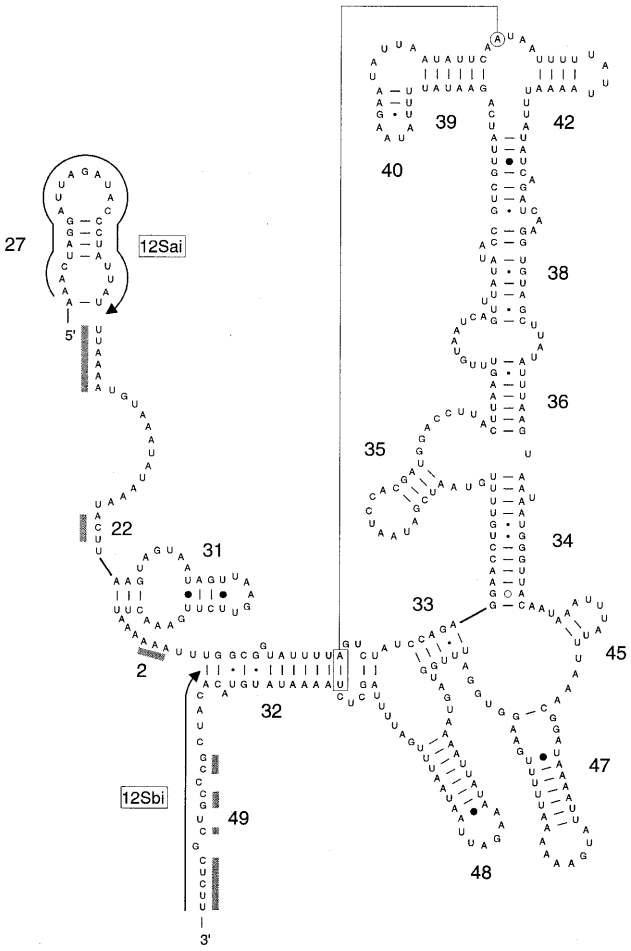

Figure 1. Secondary structure model for domain III 12S rRNA for *D.virilis*, showing the location of the two conserved PCR primers 12Sai and 12Sbi (19). Stems that pair with bases outside domain III are indicated by shaded boxes. Structure obtained from Gutell's Comparative RNA web site (http://www.rna.icmb.utexas.edu); helices numbered after van de Peer *et al.* (5).

MATERIALS AND METHODS

Sequences and alignment

A total of 225 insect domain III 12S rRNA sequences were retrieved from EMBL, representing flies (21–23), mosquitoes, (24–27), beetles (28–30), butterflies (31), the honeybee (32), cicadas (22), bugs (33,34), cockroaches (35), termites (36), grasshoppers (37), damselflies (38) and a silverfish (24). Most of these sequences were obtained with primers 12Sai and 12Sbi (19), which span domain III and part of domain II (Fig. 1). Sequences that extended beyond the location of these primers (such as those from whole mitochondrial genomes) were trimmed to just the region between the two primers. Sequences were aligned using Clustal X (39). Taxonomic groups of sequences (such as *Drosophila* or termites) were aligned separately, then these alignments were combined using the profile alignment mode of Clustal X. Alignments were assembled in the order corresponding to the taxonomic tree shown in Figure 2 (i.e. to obtain an alignment for Orthopteroidea, alignments for Termitidae and Blattaria were combined to give an alignment for Dictyptera and the latter was then aligned with the Orthoptera alignment).

Maximum weight matching

MWMs for the alignment of insect RNA sequences were computed using the programs jmixy, hlxplot, makegraf and imatch (18), available from ftp://ftp.cshl.org/pub/science/mzhanglab/tabaska/ . Mutual information (7,40) scores were computed using jmixy, with the minimum helix length parameter set to 3. The mutual information coefficient, $M(x,y)$, between sites x and y is the sum of variation at each position minus the variation between the two sites together: $M(x,y) = H(x) + H(y) - H(x,y)$. The variability of site x is described by the entropy term $H = \Sigma_b f_b \ln b$, where f_b is the frequency of the bth base [$b \in$ (A,G,U,C,–)] at site x. The value of $M(x,y)$ is greatest when sites x and y are both highly variable and the variation is correlated (e.g. if a substitution at site x is mirrored by a compensating change at site y).

Helix plot scores were computed using hlxplot with the following parameters (all are the default values): bad pair score = 2, good pair score = 1, paired gap penalty = 3, minimum helix length = 2, minimum loop length = 3 and helix length score = 2. Folding graphs for mutual information and helix plot scores were combined using makegraf and the MWM obtained using imatch. The program makegraf allows the user to specify the relative weights used to combine the mutual information and helix plot scores. Experiments with test data sets suggested that weighting the mutual information twice as much as helix plot scores gives the best results (unpublished observation).

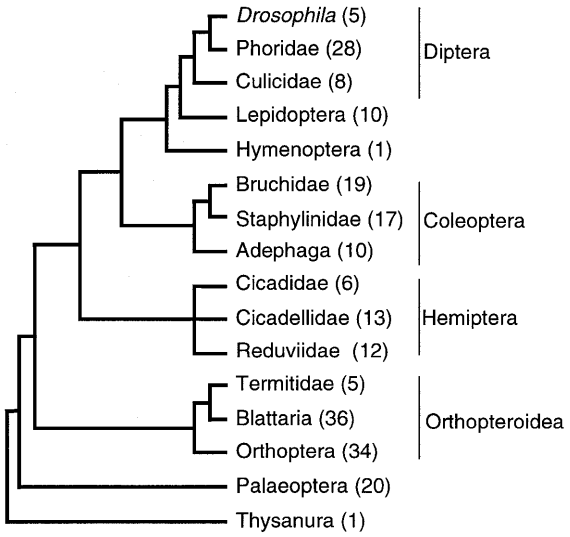

Figure 2. Phylogeny of the insect groups used to align 12S rRNA sequences. The number of sequences used for each group is given in parentheses. The topology of the tree is based on current insect taxonomy and Whiting *et al.* (53).

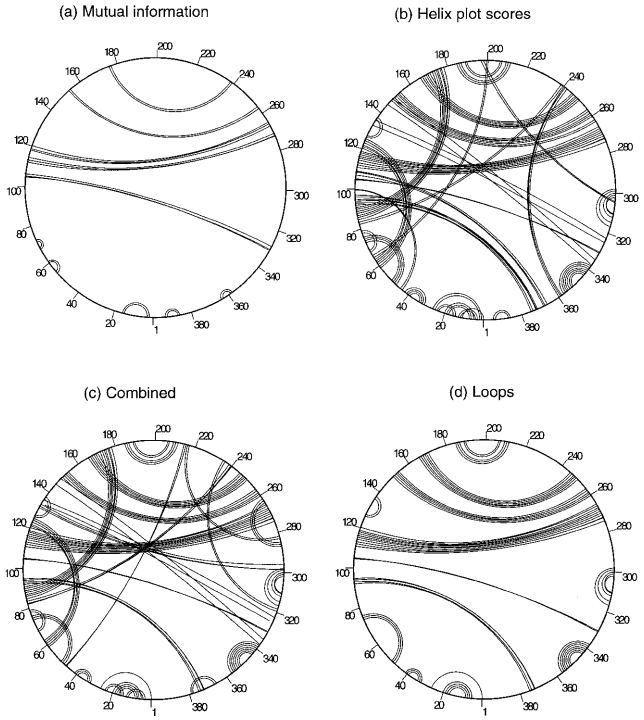

Figure 3. Circle plots for *D.yakuba* 12S rRNA, showing maximum weighted matchings for mutual information (**a**), helix plot scores (**b**) and mutual information and helix plot scores combined (**c**). In these plots only matchings corresponding to helices of length 2 or more are shown (i.e. lone pairs have been excluded). The maximum loop matching for (c) is shown in (**d**). This corresponds to the secondary structure shown in Figure 4.

Obtaining a structure

The MWM was displayed as a circle plot (41). Helices can be recognised in these plots as parallel sets of chords (40), so the MWM output was initially filtered to exclude lone pairs. RNA secondary structure is typically represented as a planar graph (i.e. a graph where no pair of edges overlap), so to generate a secondary structure from the MWM the largest planar structure was extracted using the maximum loop matching algorithm of Nussinov *et al.* (41,42). The program Circles (43) was used to create the circle plots and extract the corresponding secondary structures.

Refining the structure

To supplement the automatic comparative analysis of secondary structure, mutual information (40) values were computed for all pairs of nucleotides using the program BioEdit (http://jwbrown.mbio.ncsu.edu/RNaseP/info/programs/BIOEDIT/bioedit.html). For each site the five pairings with the highest value of $M(x,y)$ were recorded. Alternative structures that were not present in the maximum loop matching were also explored. Secondary structures were displayed using the program RnaViz (44).

Structural alignment

The secondary structure model was used to manually create an alignment for exemplar sequences from the 16 insect clades shown in Figure 2. An initial alignment was obtained using the Divide-and-Conquer (DCA) method (45), which performs well in aligning RNA sequences (3,46). The secondary structure model was then applied to the alignment and manual adjustments made where necessary to align helices. This alignment in turn suggested modifications to the model, which were incorporated into the final structure.

RESULTS

Maximum weighted matching

The alignment of 225 sequences had a length of 503 positions and is available from http://taxonomy.zoology.gla.ac.uk/rod/data/rna/ and has been deposited in the EMBL alignment database (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/) as alignment DS43718. Figure 3 shows the MWMs for the mutual information and helix plot scores for this alignment, plotted for the *Drosophila yakuba* sequence (21). The mutual information plot shows few helices, but has no overlapping edges. In contrast, the helix score plot shows more helices, many of which overlap with other helices. Although for some RNAs such a plot may indicate complicated tertiary interactions, in this case some of these matchings will be spurious. Parts of two of the prominent helices in Figure 3b (positions 111–119:266–274 and 148–154:196–203) appear in the mutual information plot, as does helix 104–105:330–331. A major difference between the two plots is the pairing of positions 85–91 with 172–178 in the helix plot. This is due to the complementarity of the motifs UGGCGGU and ACCGUCG at these two positions. Because there are few correlated changes between these two motifs this helix does not appear in the mutual information graph.

Applying the loop matching algorithm of Nussinov *et al.* to the combined MWM plot (Fig. 3c) yields the graph in Figure 3d, which corresponds to the secondary structure shown in Figure 4. This structure closely resembles the Gutell model for *D.virilis*

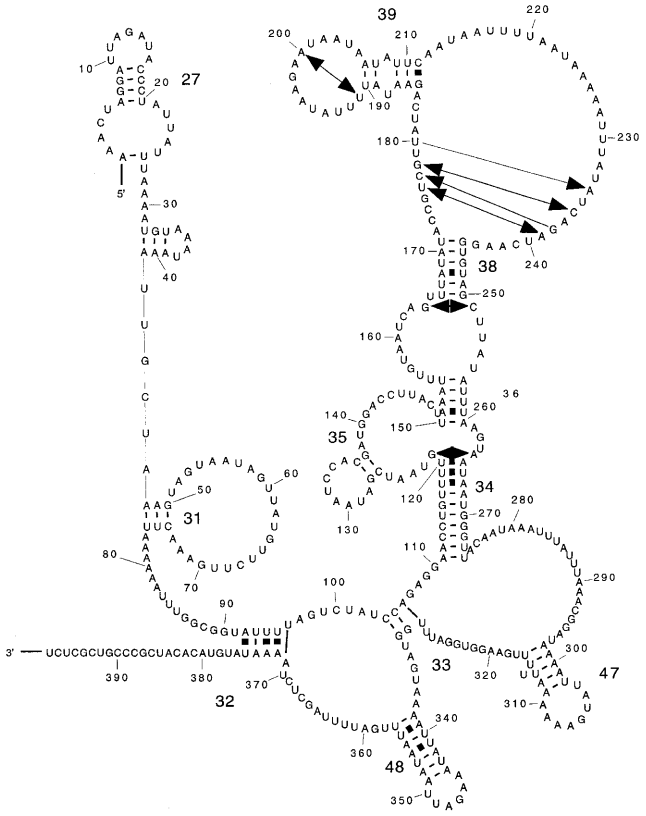

Figure 4. Initial secondary structure model for *D.yakuba* 12S rRNA, corresponding to the circle plot in Figure 3d. Base pairings marked with thick lines indicate pairings for which the mutual information coefficient, $M(x,y)$, is higher than for any other possible pairing involving either base. Possible pairings that are missed in the MWM are indicated by arrows connecting bases with the highest $M(x,y)$ values. Double arrows indicate that both of the paired bases are each others best pairing and a single arrow $x \rightarrow y$ indicates that y is the best pairing with x but some base other than x has a higher $M(x,y)$ value when paired with y.

(Fig. 1). Most of the helices receive support from mutual information scores, in that they contain base pairs for which the mutual information coefficient, $M(x,y)$, is higher than for any other possible pairing involving either base. To investigate whether the MWM helices could be extended, lone pairs that have support from mutual information and are consistent with the MWM structure were added to the model.

Refining the structure

The secondary model was further refined by inspecting the DCA alignment of 16 exemplar sequences. The initial DCA alignment was manually adjusted where the secondary structure model suggested that one or more sequences had been misaligned. In turn, the alignment suggested some structural features missed in the MWM analysis. Some helices found in Figure 4 were extended, where this was consistent with the alignment. Helix 32 was lengthened to include the UGGCGGU motif, which in the MWM analysis was mispaired with part of helix 38. This mispairing is probably because the region corresponding to helix 32′ has not been sequenced in many insects.

Helix 33 was extended to include the conserved AGA motif. Helix 34 was slightly redrawn to include a single base bulge in

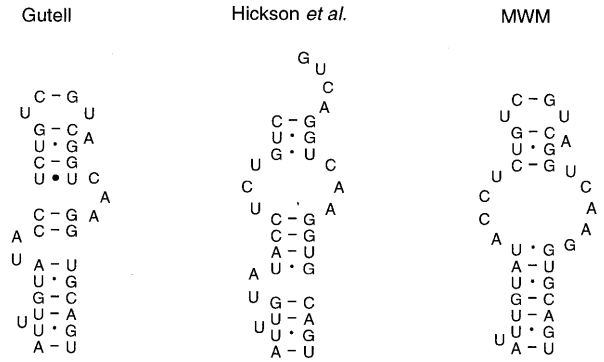

Figure 5. Two alternative published models for helix 38 in the cicada (ref. 2, fig. 6) and the MWM structure obtained in this study.

the 34′ side of the helix, which improves the alignment. Helix 38 has the most complicated structure of all helices (Fig. 5). The simple stem found in the MWM analysis can be extended with well-supported lone pairs, leaving a bulge comprising (in *D.yakuba*) 5′-ACCG-3′ on one side and 5′-UCAAG-3′ on the other. This structure resembles the Gutell model, so this model has been adopted here.

The 225 sequence alignment between helices 39 and 38′ is of relatively poor quality (Fig. 7), making it difficult to locate helices in this region. Examining $M(x,y)$ values for this region within the individual clades of insects shown in Figure 2 reveals some evidence for base pairing in the region corresponding to helix 42, particularly in the Diptera and damselflies. However, in some clades there was little evidence for this helix. Evidence for helix 40 is also rather weak; a single pair can be drawn on the basis of mutual information (Fig. 4), although this can be extended in some taxa.

Helix 47 emerges from the MWM analysis of the 225 sequence alignment, although with little mutual information support. Examination of the 16 sequence alignment provided clearer evidence of co-variation among base pairs in this helix, including the lone pair at the base in Gutell's model (Fig. 1). However, there is little consistent evidence for helix 45 and so I have not attempted to draw this helix for any taxa. The short helices 27 and 31 are in regions of poor alignment quality and the region between primer 12Sai and helix 32 also interacts with other parts of the complete 12S rRNA molecule (Fig. 1). Hence, these helices cannot be inferred with much confidence.

Figure 6 shows the final structure for the *D.yakuba* sequence, together with structure logos (47) summarising variation in the major helices. The height of each symbol in the logo is proportional to the frequency of the corresponding base in the 225 sequences. Structure logos are a convenient tool for displaying sequence variation because they also indicate the extent to which sites co-vary.

DISCUSSION

Comparison with previous models

The secondary structure model obtained using MWM closely resembles the Gutell model (Fig. 1), the structures in the European Small Subunit Ribosomal RNA database (5) and those

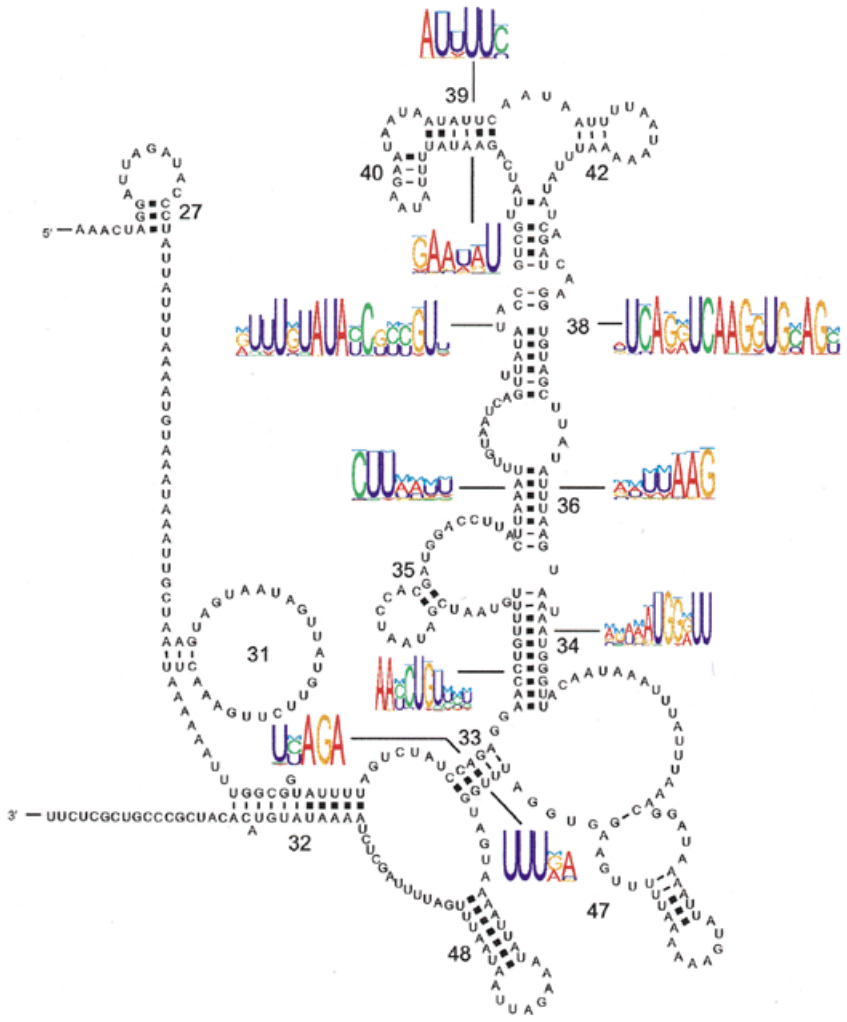

Figure 6. Refined model for *D.yakuba* 12S rRNA. This model is based on the pairings in Figure 4 (thick lines), together with additional pairings suggested by examining an alignment of 16 exemplar sequences. Pairings obtained by the MWM analysis or lone pairs supported by high values of $M(x,y)$ (Fig. 4) are indicated by thick lines. Structure logos (47) are shown for some of the major helices. The height of each symbol in the logo is proportional to the frequency of the corresponding base. The letter M indicates the amount of mutual information between corresponding base pairs.

proposed by Hickson *et al.* (2). The MWM structure has fewer helices than previously proposed models and some of the helices that are shared with previous models are shorter. In this sense, the MWM structure (Fig. 4) is fairly conservative. The major difference among the three previous models concerns helix 38. Hickson *et al.* (2) reviewed the competing models for this region and proposed their own structure. The MWM structure for helix 38 (Fig. 5) is different again, but more closely resembles the Gutell model, especially in the distal pairings, which have high values of $M(x,y)$.

Previous models contain some small helices, such as 40, 42 and 45, for which there is little supporting evidence within insects as a whole. Part of the difference in structures may be due to inaccuracies in the initial alignment of the 225 insect sequences (see below). Examining alignments for individual clades yielded some support for helices 40 and 42, but I found no convincing evidence for helix 45. This poor support for some helices suggests that some secondary structure elements will need to be inferred at more local levels within the phylogenetic tree. It also cautions against over-reliance on a single general model of secondary structure. In practice, structures for a new sequence are often inferred by applying a previously obtained model for another sequence onto the new sequence. There is a danger that if the existing model does not apply to the new sequence, erroneous pairings may be proposed in order to force the new sequence to fit the model. Some helices in the European Small Subunit Ribosomal RNA structures for insect 12S rRNA have unconventional base pairings, such as A-A and U-U in helices 41 and 45 (neither helix appears in the MWM analysis), that may be due to applying a too general secondary structure model. This emphasises the need for objective methods for inferring secondary structure, so that taxon-specific features can be readily identified.

Performance of MWM

In this study I used a rather crude initial alignment as input to the MWM algorithm. This alignment made no use of secondary structure and, although Clustal X does rather well in

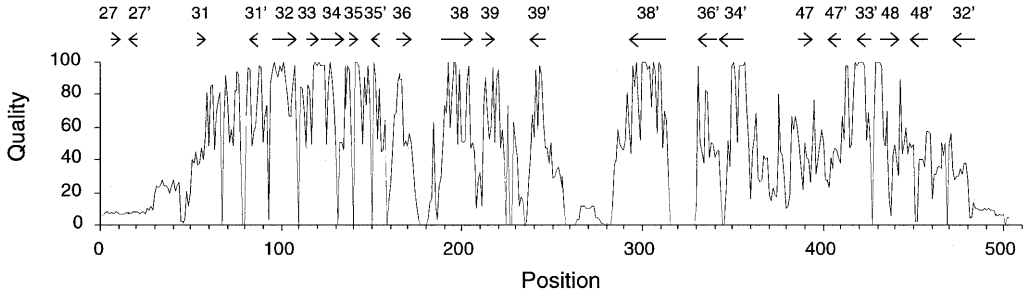

Figure 7. Plot of quality scores (39) for the Clustal X alignment of 225 insect 12S rRNA sequences, showing the locations of the stems shown in Figure 6. Note that positions are numbered with respect to the 225 sequence alignment, whereas in Figure 6 the bases are numbered with respect to the *D.yakuba* sequence.

aligning animal 12S rRNA sequences (3), refining the alignment in the light of the secondary structure would likely improve the performance of MWM. Indeed, a better starting point for MWM might be obtained using alignment methods that took potential secondary structure into consideration (48). When applying MWM to manually constructed alignments that explicitly incorporate secondary structure considerations, such as those for 12S rRNA (2) and ribonuclease P (49), MWM generated very clean matchings that closely agree with the accepted structures for these molecules (unpublished observation). This reflects the dependence of comparative analysis on accurate alignments. However, Tabaska *et al.* (18) found that even with 10% of the sequences unreliably aligned, MWM could generate previously accepted structures.

Figure 7 shows the location of stems in the 12S rRNA model with respect to alignment quality scores (39) for the 225 insect sequence alignment. The 5′- and 3′-ends of the alignment are of low quality because for many sequences one or other end is not available. Many of the core stems (33, 34, 36, 38 and 39) are in well-aligned regions. In contrast, the previously proposed helices which MWM did not confirm are in regions that are difficult to align (e.g. helix 40 between 39 and 39′, helix 42 between 39′and 38′ and helix 45 between 34′ and 47′). However, MWM still recovered helices in areas that are relatively poorly aligned, such as helices 47 and 48.

The other factor affecting the performance of any comparative method that uses mutual information is whether there is enough sequence variation to compute meaningful values of $M(x,y)$. Due to both sequence saturation and the non-uniform distribution of substitutions along RNA sequences (22), it is difficult to make global recommendations about the necessary level of sequence divergence for successful MWM analysis. The 225 insect sequences considered here had a mean sequence difference of 29%, with the most divergent pairs of sequences differing at 49% of their sites. When MWM was applied to subsets of the data (corresponding to the clades in Fig. 2), it detected the core stems in 12S rRNA domain III (i.e. 32, 33, 34, 36 and 38) when mean sequence difference was at least 15–20% (such as the Diptera and the Orthopteroidea). In smaller, less divergent clades (e.g. Palaeoptera, mean divergence 7.9%) or clades with some poorly aligned sequences (e.g. Blattaria) MWM was less successful. When given adequate taxonomic and sequence variation, MWM obtained a well-supported structure that agrees in many respects with previous, manually obtained models.

Applying the structural model to other sequences

One of the motivations for developing secondary structure models is to improve the quality of sequence alignments based solely on primary sequences. Often authors of secondary structure models provide guidelines for aligning new sequences to an existing model that consist of locating conserved motifs and then manually adjusting the alignment (1,2). The reliance on conserved motifs may result in mistakes in alignment if the motif is not conserved in all taxa. A case in point is the misalignment of the honeybee sequence by Hickson *et al.* (2). Hickson *et al.* noted that the bee sequence lacks the GGA motif before helix 33′ and in their alignment of the bee sequence the conserved helix 33 is only four bases long, as opposed to five in every other taxon they looked at. They suggested that absence of the GGA motif might be due to a sequencing error, although none was found on the sequencing gels. Comparison of their alignment with the 225 and 16 sequence alignments obtained here shows that the honeybee has GAA instead of the GGA motif and that the region Hickson *et al.* have identified as helix 33′ in the bee is actually part of helix 48. In a subsequent study Hickson *et al.* (3) employed this alignment as a benchmark to evaluate the accuracy of multiple sequence alignment programs. Given the error in alignment it is no surprise that the honeybee sequence caused these programs the most difficulties.

Structure logos (47) are a more sophisticated tool for representing conserved motifs, and are used in Figure 6. Each of the core helices comprises a mix of highly conserved bases and variable sites that co-vary. There are very few sites that are completely invariant. Rather than rely on manually locating ‘conserved’ motifs to align RNA sequences, an alternative approach is to use automated techniques that simultaneously align sequences using both primary and secondary structure (20,50–52). Corpet and Michot (20) have developed a program (RNAlign) that uses a reference alignment and secondary structure model as a template for inferring the secondary structure of a new sequence. I have created a web server (http://taxonomy.zoology.gla.ac.uk/cgi-bin/rna.cgi) that uses RNAlign and the secondary structure model shown in Figure 6. The server takes as input a single 12S rRNA sequence and returns an inferred secondary structure in MFOLD ct format (13). The server enables the results of the MWM comparative analysis to be readily applied to any insect sequence. This approach can be easily generalised to any RNA sequence and taxonomic group for which secondary structure models are available.

ACKNOWLEDGEMENTS

I thank Jack Tabaska for help with his maximum weighted matching programs, Florence Corpet for her assistance with RNAlign and the helpful comments of the reviewers. This work was partially supported by NERC grant GR3/11075 and was completed while the author was a University of Auckland 2000 Foundation Visitor.

REFERENCES

1. Kjer,K.M. (1995) *Mol. Phylogenet. Evol.*, **4**, 314–330.
2. Hickson,R.E., Simon,C., Cooper,A., Spicer,G.S., Sullivan,J. and Penny,D. (1996) *Mol. Biol. Evol.*, **13**, 150–169.
3. Hickson,R.E., Simon,C. and Perrey,S.W. (2000) *Mol. Biol. Evol.*, **17**, 530–539.
4. Titus,T.A. and Frost,D.R. (1996) *Mol. Phylogenet. Evol.*, **6**, 49–62.
5. van de Peer,Y., de Rijk,P., Wuyts,J., Winkelmans,T. and de Wachte,R. (2000) *Nucleic Acids Res.*, **28**, 175–176.
6. Lydeard,C., Holznagel,W.E., Schnare,M.N. and Gutell,R.R. (2000) *Mol. Phylogenet. Evol.*, **15**, 83–102.
7. Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) *Nucleic Acids Res.*, **20**, 5785–5795.
8. Konings,D.A.M. and Gutell,R.R. (1995) *RNA*, **1**, 559–574.
9. Brown,J.W. (1991) *Comput. Appl. Biosci.*, **7**, 391–393.
10. Hall,T.A. (1999) *Nucleic Acids Symp. Ser.*, **41**, 95–98.
11. Gorodkin,J., Stærfeldt,H.H., Lund,O. and Brunak,S. (1999) *Bioinformatics*, **15**, 769–770.
12. Matzura,O. and Wennborg,A. (1996) *Comput. Appl. Biosci.*, **12**, 247–249.
13. Zuker,M. (1989) *Science*, **244**, 234–237.
14. Hofacker,I.L., Fekete,M., Flamm,C., Huynen,M.A., Rauscher,S., Stolorz,P.E. and Stadler,P.F. (1998) *Nucleic Acids Res.*, **26**, 3825–3836.
15. Lück,R., Gräf,S. and Steger,G. (1999) *Nucleic Acids Res.*, **27**, 4208–4217.
16. Morrison,D.A. and Ellis,J.T. (1997) *Mol. Biol. Evol.*, **14**, 428–441.
17. Cary,R.B. and Stormo,G.D. (1995) In Rawlings,C., Clark,D., Altman,R., Hunter,L., Lengauer,T. and Wodak,S. (eds), *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 75–80.
18. Tabaska,J.E., Cary,R.E., Gabow,H.N. and Stormo,G.D. (1998) *Bioinformatics*, **14**, 691–699.
19. Simon,C., Frati,F., Beckenbach,A., Crespi,B., Liu,H. and Flook,P. (1994) *Ann. Entomol. Soc. Am.*, **87**, 651–704.
20. Corpet,F. and Michot,B. (1994) *Comput. Appl. Biosci.*, **10**, 389–399.
21. Clary,D.O. and Wolstenholme,D.R. (1985) *J. Mol. Evol.*, **22**, 252–271.
22. Simon,C., Nigro,L., Sullivan,J., Holsinger,K., Martin,A., Grapputo,A., Franke,A. and McInstosh,C. (1996) *Mol. Biol. Evol.*, **13**, 923–932.
23. Austin,J.J. and Disney,R.H.L. (1999) EMBL accession nos AF126298–AF126325.
24. Ballard,J.W.O., Olsen,G.J., Faith,D.P., Odgers,W.A., Rowell,D.M. and Atkinson,P.W. (1992) *Science*, **258**, 1345–1348.
25. Beard,C.B., Hamm,D.M. and Collins,F.H. (1993) *Insect Mol. Biol.*, **2**, 103–124.
26. Mitchell,S.E., Cockburn,A.F. and Seawright,J.A. (1993) *Genome*, **36**, 1058–1073.
27. Beebe,N.W., Cooper,R.D., Morrison,D.A. and Ellis,J.T. (1999) EMBL accession nos AF121064, AF121065 and AF121072–AF121074.
28. Silvain,J.-F. and Delobel,A. (1998) *Mol. Phylogenet. Evol.*, **9**, 533–541.
29. Ballard,J.W.O., Thayer,M.K., Newton,A.F. and Grismer,E.R. (1998) *Syst. Biol.*, **47**, 367–396.
30. Duering,A. and Brueckner,M. (1999) EMBL accession nos AF190021–AF190030.
31. Sutherland,R.M. and Axton,J.M. (2000) EMBL accession nos AF232882–AF232887 and AF232889–AF232892.
32. Crozier,R.H. and Crozier,Y.C. (1993) *Genetics*, **113**, 97–117.
33. Dietrich,C.H., Fitzgerald,S.J., Holmes,J.L., Black,W.C.I. and Nault,L.R. (1998) EMBL accession nos AF051276–AF051288.
34. Garcia,B.A. and Powell,J.R. (1998) *J. Med. Entomol.*, **35**, 232–238.
35. Kambhampati,S. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 2017–2020.
36. Ohkuma,M. (1997) EMBL accession nos AB006580, AB006582, AB006584, AB006586 and AB006588.
37. Flook,P.K., Klee,S. and Rowell,C.H.F. (1999) *Syst. Biol.*, **48**, 233–253.
38. Chippindale,P.T., Dave,V.K., Whitmore,D.H. and Robinson,J.V. (1999) *Mol. Phylogenet. Evol.*, **11**, 110–121.
39. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4876–4882.
40. Chiu,D.K.Y. and Kolodziejczak,T. (1991) *Comput. Appl. Biosci.*, **7**, 347–352.
41. Nussinov,R., Pieczenik,G., Griggs,J.R. and Kleitman,D.J. (1978) *SIAM J. Appl. Math.*, **35**, 68–82.
42. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
43. Page,R.D.M. (2000) *Bioinformatics*, in press.
44. de Rijk,P. and de Wachter,R. (1997) *Nucleic Acids Res.*, **25**, 4679–4684.
45. Stoye,J., Moulton,V. and Dress,A.W.M. (1997) *Comput. Appl. Biosci.*, **13**, 625–626.
46. Perry,S.W., Stoye,J., Moulton,V. and Dress,A.W.M. (1997) *Universität Bielefeld Forschungsschwerpunkt Mathematisierung–Strukturbildungsprozesse*, Materialien/Preprint 111, pp. 1–26.
47. Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) *Comput. Appl. Biosci.*, **13**, 583–586.
48. Kim,J., Cole,J.R. and Pramanik,S. (1996) *Comput. Appl. Biosci.*, **12**, 259–267.
49. Brown,J.W. (1999) *Nucleic Acids Res.*, **27**, 314.
50. Bafna,V., Muthukrishnan,S. and Ravi,R. (1996) *DIMACS Tech. Rep.*, 96–30.
51. Lenhof,H.-P., Reinert,K. and Vingron,M. (1998) *J. Comput. Biol.*, **5**, 517–530.
52. Notredame,C., O'Brien,E.A. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4570–4580.
53. Whiting,M.F., Carpenter,J.C., Wheeler,Q.D. and Wheeler,W.C. (1997) *Syst. Biol.*, **46**, 1–68.