# Comparative Sequence Analysis and Patterns of Covariation in RNA Secondary Structures

**John Parsch,[1] John M. Braverman[2] and Wolfgang Stephan**

*Department of Biology, University of Rochester, Rochester, New York 14627-0211*

Manuscript received June 23, 1999
Accepted for publication October 4, 1999

ABSTRACT

A novel method of RNA secondary structure prediction based on a comparison of nucleotide sequences is described. This method correctly predicts nearly all evolutionarily conserved secondary structures of five different RNAs: tRNA, 5S rRNA, bacterial ribonuclease P (RNase P) RNA, eukaryotic small subunit rRNA, and the 3′ untranslated region (UTR) of the Drosophila *bicoid* (*bcd*) mRNA. Furthermore, covariations occurring in the helices of these conserved RNA structures are analyzed. Two physical parameters are found to be important determinants of the evolution of compensatory mutations: the length of a helix and the distance between base-pairing nucleotides. For the helices of *bcd* 3′ UTR mRNA and RNase P RNA, a positive correlation between the rate of compensatory evolution and helix length is found. The analysis of Drosophila *bcd* 3′ UTR mRNA further revealed that the rate of compensatory evolution decreases with the physical distance between base-pairing residues. This result is in qualitative agreement with Kimura's model of compensatory fitness interactions, which assumes that mutations occurring in RNA helices are individually deleterious but become neutral in appropriate combinations.

MOLECULES of RNA have a variety of important functions in biological systems, many of which depend on the RNA folding into a precise structure. For example, protein synthesis requires the participation of tRNAs and rRNAs that have highly conserved structures (Woese and Pace 1993; Dirheimer *et al.* 1995). mRNAs are known to contain important structural elements that affect localization, stability, and translational regulation (Macdonald and Struhl 1988; Mullner and Kuhn 1988; Macdonald 1990; Pandey *et al.* 1994). In addition, catalytic RNAs have been identified in both prokaryotic and eukaryotic systems (Krüger *et al.* 1982; Guerrier-Takada *et al.* 1983; Pace and Smith 1990) and have also been engineered *in vitro* (Odai *et al.* 1990; Ekland and Bartel 1996; Unrau and Bartel 1998). The activity of these RNAs is, not surprisingly, highly dependent on proper folding (Pley *et al.* 1994; Scott *et al.* 1995; Ekland and Bartel 1996). A detailed knowledge of RNA structure is therefore essential for a complete understanding of many aspects of molecular and cell biology. RNA structures are also of great interest in molecular evolution. The structures of tRNAs and rRNAs are highly conserved among all kingdoms of life and these sequences, particularly rRNAs, have been widely used to determine phylogenetic relationships among diverse taxa (Woese and Fox 1977; Gouy and Li 1989; Kumar and Rzhetsky 1996). The relatively simple pattern of intramolecular Watson-Crick (WC) base-pairing involved in RNA structures has made them a suitable model for the study of compensatory evolution and epistatic selection at the molecular level (Stephan and Kirby 1993; Schöniger and von Haeseler 1994; Kirby *et al.* 1995; Rzhetsky 1995; Tillier and Collins 1995; Stephan 1996). RNA molecules are also thought to have been among the first catalytic replicators in prebiotic evolution under the "RNA world" hypothesis (Gilbert 1986; Joyce and Orgel 1993).

Presently, the most reliable method for predicting secondary structures of large RNAs from primary DNA sequence data is through phylogenetic-comparative analysis of aligned nucleotide sequences (Fox and Woese 1975; James *et al.* 1988; Pace *et al.* 1989). The major assumption underlying this approach is that mutations that disrupt the WC base-pairing of a functionally important RNA stem have a deleterious effect, but that deleterious effect may be overcome by a second, compensatory mutation in the other half of the stem that restores the potential for base-pairing. Compensatory evolution, as mediated by RNA secondary structure, results in a detectable pattern of nucleotide substitutions ("covariations") in the phylogenetic alignment of homologous RNA sequences from different species. Covarying sites are defined as those that differ between two or more species but retain the potential for WC base-pairing in each species (for example, a GC pair in species 1 is replaced by an AU pair in species 2). The phylogenetic-comparative method has been an effective

*Corresponding author:* Wolfgang Stephan, Department of Biology, University of Rochester, Rochester, NY 14627-0211.
E-mail: stephan@troi.cc.rochester.edu

[1] *Present address:* Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138-2020.

[2] *Present address:* Le Moyne College, Syracuse, NY 13214-1499.

approach to identifying conserved RNA secondary structures and has been used to generate consensus structures for tRNAs, rRNAs, and ribozymes (Pace *et al.* 1989; Woese and Pace 1993; Dirheimer *et al.* 1995). A drawback to this method, however, is that the rules for identifying potential pairing stems are somewhat *ad hoc*. For example, an RNA stem is considered "proven" if two or more covariations are present in the aligned sequences (Fox and Woese 1975; James *et al.* 1988). The phylogenetic relationship of the aligned sequences and the level of sequence divergence are not considered. In addition, the phylogenetic-comparative method does not account for mismatches or noncanonical base pairs such as GU wobble pairs.

Muse (1995) proposed an alternate method to detect RNA secondary structures from aligned sequence data that relies on a likelihood-ratio test (LRT) to identify potential pairing regions showing constraints for WC interactions. The advantage of this approach is that it does not rely on the *ad hoc* rules used in the phylogenetic-comparative method. The pattern of nucleotide substitution at paired sites is compared to that of unpaired sites and a pairing parameter, λ, is estimated. The relative evolutionary conservation of each predicted pairing is quantified by calculating an LRT statistic. A drawback to the LRT approach has been the requirement to specify the coordinates of potential pairing stems before application of the test. Thus, so far, this method has primarily been used to test structures previously predicted by phylogenetic-comparative analysis (Kirby *et al.* 1995; Muse 1995; Parsch *et al.* 1997). Muse (1995), however, presents an example of how the likelihood method may, in principle, be used to predict secondary structures without *a priori* knowledge of the location of paired regions.

Here we present a novel method of RNA secondary structure prediction that integrates Muse's (1995) LRT approach. The method is applied to five different RNA molecules that are known to have conserved, functionally important secondary structures: tRNA, 5S rRNA, bacterial ribonuclease P (RNase P) RNA, the 3′ untranslated region (UTR) of the Drosophila *bicoid* (*bcd*) mRNA, and eukaryotic small subunit (SSU) rRNA. Furthermore, we analyze the covariations occurring in the helices of these RNA structures. The goal of this analysis is to identify physical parameters that determine the evolution of compensatory mutations. Two parameters are found to be important: the length of a helix and the physical distance between base-pairing nucleotides.

## MATERIALS AND METHODS

**Sequence collection and alignment:** Aligned, mitochondrial tRNA sequences (UGC anticodon) were downloaded from the tRNA Database (Sprinzl *et al.* 1998; http://www.uni-bayreuth.de/departments/biochemie/trna/index.html). Sequences used (followed by their sequence code) were *Aedes albopictus* (DA4800), *Drosophila melanogaster* (DA4840), *Strongylocentrotus purpuratus* (DA5080), *Xenopus laevis* (DA5120), *Rana catesbeiana* (DA5160), *Gallus gallus* (DA5220), *Mus musculus* (DA5320), *Homo sapiens* (DA5880), and *Bos taurus* (DA5360).

Aligned 5S rRNA sequences were downloaded from the Berlin RNA Databank (Specht *et al.* 1991; ftp://ftp.embl-heidelberg.de:/pub/databases/berlin/). Sequences used for analysis were *H. sapiens*, *G. gallus*, *D. melanogaster*, *Bombyx mori*, *X. laevis*, *Caenorhabditis elegans*, *Notophthalmus viridescens*, *Terrapene carolina*, and *Brachionus plicatilis*.

Aligned RNase P RNA sequences were downloaded from the RNase P Database (Brown 1998; http://jwbrown.mbio.ncsu.edu/RNaseP/home.html). The following sequences were used for analysis: *Escherichia coli*, *Salmonella typhimurium*, *Klebsiella pneumoniae*, *Erwinia agglomerulans*, *Serratia marcescens*, *Pseudomonas fluorescens*, *Bacillus brevis*, *Bacillus stearothermophilus*, and *Bacillus megaterium*.

Drosophila *bcd* sequences were obtained from GenBank (release 108.0). The sequences used (followed by their accession numbers) were *D. melanogaster* (X07870), *D. simulans* (M32123), *D. sechellia* (M32124), *D. teissieri* (M32121), *D. pseudoobscura* (X55735), *D. subobscura* (X78058), *D. virilis* (M32122), *D. picticornis* (M32126), and *D. heteroneura* (M32125). *bcd* 3′ UTR sequences (from stop codon to end of transcript) were aligned using the ClustalX program (Thompson *et al.* 1997) and then adjusted manually. Manual alignment was assisted by previously published alignments of subsets of the above sequences (Macdonald 1990; Seeger and Kaufman 1990).

Aligned SSU rRNA sequences were downloaded from the SSU rRNA Database (van de Peer *et al.* 1998; http://rrna.uia.ac.be/ssu/index.html). The following sequences were used for analysis: *Acyrthosiphon pisum*, *H. sapiens*, *C. elegans*, *Strongyloides stercoralis*, and *Saccharomyces cerevisiae*.

**Identification of potential RNA helices:** Potential RNA helices that are conserved in the aligned sequences were identified using the novel program PIRANAH. This program represents an extension of the algorithm of Han and Kim (1993). An upper triangular $n \times n$ matrix, where $n$ represents the length (in bases) of the aligned sequences, is generated in which the nucleotide state at each site of two aligned sequences is compared to that of every other site. Comparisons are performed in a pairwise fashion; thus an alignment of $N$ sequences results in $(N - 1)N/2$ comparisons. There are five possible states for each cell of the matrix: (i) a conserved WC base pair, (ii) a WC covariation, (iii) a GU wobble pair in either one or both sequences, (iv) a gap in one of the sequences, or (v) a mismatch in either one or both sequences. Potential helices are those that consist of a consecutive run of WC or wobble base pairs in both sequences (a diagonal line with slope = 1 in the $n \times n$ matrix). Wobble pairs are permitted only internally; terminal GU pairs are treated as mismatches. The user specifies the number of comparisons in which a potential helix must be conserved. This allows mismatches to be included within helices, but only as long as the number of sequences containing mismatches falls below the specified threshold. By this rule, the length of a putative helix is defined. For our examples, we required that helices be conserved in 15 out of 36 comparisons (*i.e.*, six out of the nine sequences) for tRNA, 5S rRNA, RNase P RNA, and *bcd* mRNA 3′ UTR, or 6 out of 10 comparisons (four out of the five sequences) for SSU rRNA. The user also specifies the minimum number of base pairs required per helix. Minimum helix length was set at 3 bp for tRNA, 5S rRNA, RNase P RNA, and *bcd* mRNA 3′ UTR. In the case of SSU rRNA (a much longer sequence) only helices with a minimum length of 5 bp were considered in order to keep computation time reasonable in subsequent steps.

After all helices meeting the specified criteria have been identified, LRT values (Muse 1995) are calculated for each helix. Since these calculations can be quite time consuming (depending on the number of helices, the number of sequences in the alignment, and the length of sequence), the user may choose to perform the calculations only for helices meeting a minimum "pairing score," which is a simple estimate of the level of conservation. The pairing score is defined as $(L + W)/C$, where $L$ represents the sum of the helix lengths (in bases) of every comparison, $W$ represents the sum of the number of WC covariations of every comparison, and $C$ represents the number of comparisons. Typically, the minimum score must be increased as sequence length increases in order to keep computation time reasonable. For our examples, the minimum score for LRT calculation was set at 1.0 (tRNA and 5s rRNA), 1.5 (RNase P RNA), 2.0 (*bcd* 3′ UTR), and 3.5 (SSU rRNA).

**Construction of RNA secondary structure model:** After generating a complete list of potential RNA helices and their respective LRT values with PIRANAH, the helices were assembled into a final RNA secondary structure model using another novel program, GROUPER. This program sorts through the list of helices and determines subsets that are compatible with each other. Two helices are considered compatible if either: (i) there is no overlap between their 5′ and 3′ coordinates, or (ii) the 5′ and 3′ coordinates of one helix fall between the 5′ and 3′ coordinates of a second helix. Thus, several short-range pairings may be nested within one (or more) long-range pairing. Pseudoknots are not permitted. For each compatible set of helices a total LRT value is calculated. This represents the sum of the individual LRTs for each helix in the structure. While the value of total LRT is not necessarily equal to the LRT calculated for the entire structure *in toto*, previous results suggest that this method produces a reliable estimate (Muse 1995; Parsch *et al.* 1997). The optimum secondary structure model is defined as the one with the greatest total LRT. As the number of potential structures increases rapidly with the number of helices, it soon becomes prohibitive to calculate total LRT for every possible structure. To overcome this problem, GROUPER performs iterations (up to 50,000) of random structure assembly. Structure prediction may be simplified by specifying a minimum LRT so that only helices with LRTs exceeding the cutoff are included in the final structure. This is useful in cases where there are a large number of conflicting helices.

**Significance tests of LRT values:** The distribution of the LRT statistic is $\sim\chi^2$ with one degree of freedom, and it has been demonstrated that this approximation is good for helices $\geq$10 bp in length (Muse 1995). Most of our predicted helices, however, are $<$10 bp, making the $\chi^2$ approximation questionable. In addition, there is a problem of multiple tests, as the helices that were subject to LRT calculations were previously selected to meet certain length and conservation criteria. Thus it is difficult to attach meaningful $P$ values to individual helices. To get around these problems and estimate $P$ values for the helices predicted in our analysis, we used a numerical resampling approach. A similar approach was used by Kirby *et al.* (1995). Since this procedure is very computer intensive, it was only practical to apply it to the two shortest RNA sequences (tRNA and 5S rRNA). Each observed sequence alignment was randomly shuffled 100 times. Only the linear order of nucleotides was permuted; base composition and level of sequence conservation remained unchanged. PIRANAH was then applied to each randomization using the same parameters that were used for the original alignment and a distribution of LRT values was obtained. The $P$ value of an observed helix was estimated as the frequency of obtaining an individual helix with LRT $\geq$ the observed value from the 100 randomizations. In addition, GROUPER was applied to each set of predicted helices from the randomizations to predict total structures. The $P$ value of an observed structure was estimated as the frequency of obtaining a total structure with LRT $\geq$ the observed value from the 100 randomizations.

## RESULTS

**RNA secondary structure prediction:** *tRNA:* The well-known cloverleaf structure of tRNA molecules has been established through both structural and comparative analyses (Kim *et al.* 1974; Sprinzl *et al.* 1998). To test our method of secondary structure prediction, we used the programs PIRANAH and GROUPER (see materials and methods) to identify conserved RNA helices and assemble a secondary structure model for an alignment of nine eukaryotic tRNA sequences. The final structure included four helices with LRT $>$ 15 (Table 1; Figure 1A). A representative helix is shown in Figure 2A. The four helices are in complete agreement with the tRNA consensus structure (Sprinzl *et al.* 1998). Our randomization simulations provide strong support for the four individual helices, as well as for the total structure (Table 2).

*5S rRNA:* 5S rRNA is a small RNA (typically 121 bases) that associates with 23S rRNA and ribosomal proteins to form the large ribosomal subunit (Osswald and Brimacombe 1999). 5S rRNA structure has been established through extensive comparative analysis (Specht *et al.* 1991). Application of our programs to an alignment of nine eukaryotic 5S rRNA sequences resulted in a final secondary structure of four helices with LRT $>$ 15 (Table 1; Figure 1B). A representative helix is shown in Figure 2B. All four of these helices are in agreement with the consensus structure (Specht *et al.* 1991). The randomization simulations strongly support the individual helices and the predicted structure as a whole (Table 2).

*RNase P RNA:* RNase P is an RNA-protein complex that produces mature tRNAs by cleaving the 5′ ends of precursor tRNA molecules. The RNA has been shown to be the catalytic subunit (Guerrier-Takada *et al.* 1983; Pace and Smith 1990). Due to its catalytic properties, the structure of RNase P RNA has received much attention and a model of its structure has been developed through both phylogenetic and mutational analyses (Haas *et al.* 1991). We have used the programs PIRANAH and GROUPER to identify conserved RNA helices and assemble a secondary structure model for an alignment of nine bacterial RNase P sequences. Predicted helices that are present in all nine species and have LRT $>$ 15 are presented in Table 1. An example is shown in Figure 2C. The helix 23-29/61-67, originally predicted as 26-29/61-64 (LRT = 20.15) due to an internal mismatch in six of the nine sequences, was extended by 3 bp after visual inspection and LRT was calculated for the extended helix. Seven of the eight helices are

**TABLE 1**

**Results of phylogenetic-comparative and LRT analyses**

| Pairing | Length | λ | LRT | WC | WOB |
|---|---|---|---|---|---|
| (i) tRNA | | | | | |
| 2-7/64-69 | 6 | 3.98 | 25.82 | 8 (4) | 4 (4) |
| 46-51/59-64 | 6 | 4.57 | 33.61 | 6 (5) | 2 (2) |
| 27-30/38-41 | 4 | 8.92 | 35.63 | 3 (1) | 0 (0) |
| 11-13/21-23 | 3 | 8.92 | 25.98 | 2 (0) | 0 (0) |
| (ii) 5S rRNA | | | | | |
| 1-9/110-118 | 9 | 5.20 | 40.60 | 7 (7) | 4 (4) |
| 16-21/57-62 | 6 | 8.92 | 40.02 | 5 (4) | 0 (0) |
| 67-71/104-108 | 5 | 7.65 | 28.77 | 6 (5) | 2 (2) |
| 29-32/45-48 | 4 | 8.92 | 26.87 | 5 (3) | 0 (0) |
| (iii) RNase P RNA | | | | | |
| 4-13/539-548 | 10 | 5.43 | 57.21 | 16 (14) | 3 (2) |
| 440-447/452-459 | 8 | 3.64 | 28.97 | 11 (10) | 4 (4) |
| 15-21/470-476 | 7 | 7.24 | 44.32 | 4 (4) | 1 (0) |
| 23-29/61-67 | 7 | 3.03 | 20.58 | 6 (6) | 3 (2) |
| 115-119/340-344 | 5 | 5.60 | 24.99 | 5 (2) | 1 (1) |
| 80-83/347-350 | 4 | 4.48 | 18.36 | 0 (0) | 0 (0) |
| 69-71/75-77 | 3 | 8.92 | 22.87 | 0 (0) | 0 (0) |
| 123-125/136-138 | 3 | 5.73 | 15.66 | 2 (0) | 1 (0) |
| (iv) *bcd* 3′ UTR | | | | | |
| 313-331/425-443 | 19 | 4.53 | 87.04 | 8 (8) | 3 (3) |
| 337-353/388-404 | 17 | 3.41 | 56.21 | 8 (8) | 8 (7) |
| 534-542/687-695 | 9 | 4.96 | 42.16 | 1 (1) | 2 (2) |
| 573-579/643-649 | 7 | 8.92 | 45.66 | 0 (0) | 1 (1) |
| 132-136/149-153 | 5 | 8.92 | 37.06 | 0 (0) | 0 (0) |
| 591-595/625-629 | 5 | 8.92 | 36.23 | 1 (0) | 0 (0) |
| 727-731/737-741 | 5 | 5.74 | 26.54 | 0 (0) | 1 (0) |
| 582-585/636-639 | 4 | 8.92 | 29.62 | 0 (0) | 0 (0) |
| (v) SSU rRNA | | | | | |
| 539-547/2360-2368 | 9 | 3.26 | 22.75 | 1 (1) | 2 (1) |
| 2503-2511/2516-2524 | 9 | 3.84 | 27.99 | 3 (2) | 1 (1) |
| 401-408/416-423 | 8 | 8.92 | 43.48 | 1 (0) | 0 (0) |
| 1311-1318/1340-1347 | 8 | 6.73 | 40.52 | 8 (8) | 2 (1) |
| 627-633/669-675 | 7 | 8.92 | 36.68 | 4 (4) | 0 (0) |
| 2378-2384/2469-2475 | 7 | 3.94 | 23.24 | 8 (5) | 6 (5) |
| 561-566/580-585 | 6 | 6.15 | 28.05 | 7 (5) | 2 (1) |
| 1404-1409/1440-1445 | 6 | 8.92 | 32.21 | 3 (2) | 1 (1) |
| 1469-1474/1492-1497 | 6 | 4.38 | 20.19 | 9 (6) | 2 (1) |
| 2263-2268/2279-2284 | 6 | 6.35 | 31.38 | 7 (5) | 1 (1) |
| 240-244/328-332 | 5 | 8.92 | 29.43 | 1 (0) | 0 (0) |
| 1279-1283/1390-1394 | 5 | 7.88 | 23.70 | 1 (0) | 1 (1) |
| 1528-1532/1639-1643 | 5 | 8.92 | 29.26 | 1 (1) | 0 (0) |
| 1644-1648/1683-1687 | 5 | 8.92 | 34.12 | 4 (3) | 0 (0) |
| 2096-2100/2112-2116 | 5 | 6.51 | 24.86 | 5 (2) | 1 (0) |
| 2308-2312/2320-2324 | 5 | 7.00 | 25.59 | 2 (2) | 2 (2) |

The first column gives the coordinates of the pairings, according to the gapped sequence alignment (see materials and methods). The second column gives the stem length (in base pairs). Columns 3 and 4 give values of λ and the LRT statistic (Muse 1995) for each pairing. The numbers in columns 5 and 6 represent the number of Watson-Crick (WC) and wobble (WOB) covariations observed for each pairing region, with the number of internal covariations given in parentheses.

consistent with the model of Haas *et al.* (1991; Figure 1C). The final pairing, 69-71/75-77, is not present in the model of Haas *et al.* (1991); this region forms part of a long-range pseudoknot in their model. Pseudoknot pairings are identified by PIRANAH but are not included in final structure predictions by GROUPER (Figure 1C). Overall, the consensus model of Haas *et al.* (1991) contains seven helices of length ≥3 (excluding pseudoknots and helices not present in Bacillus species), all of which were predicted by our method, with only one potential false positive.

*bcd 3′ UTR: bcd* is a maternal effect gene that plays a

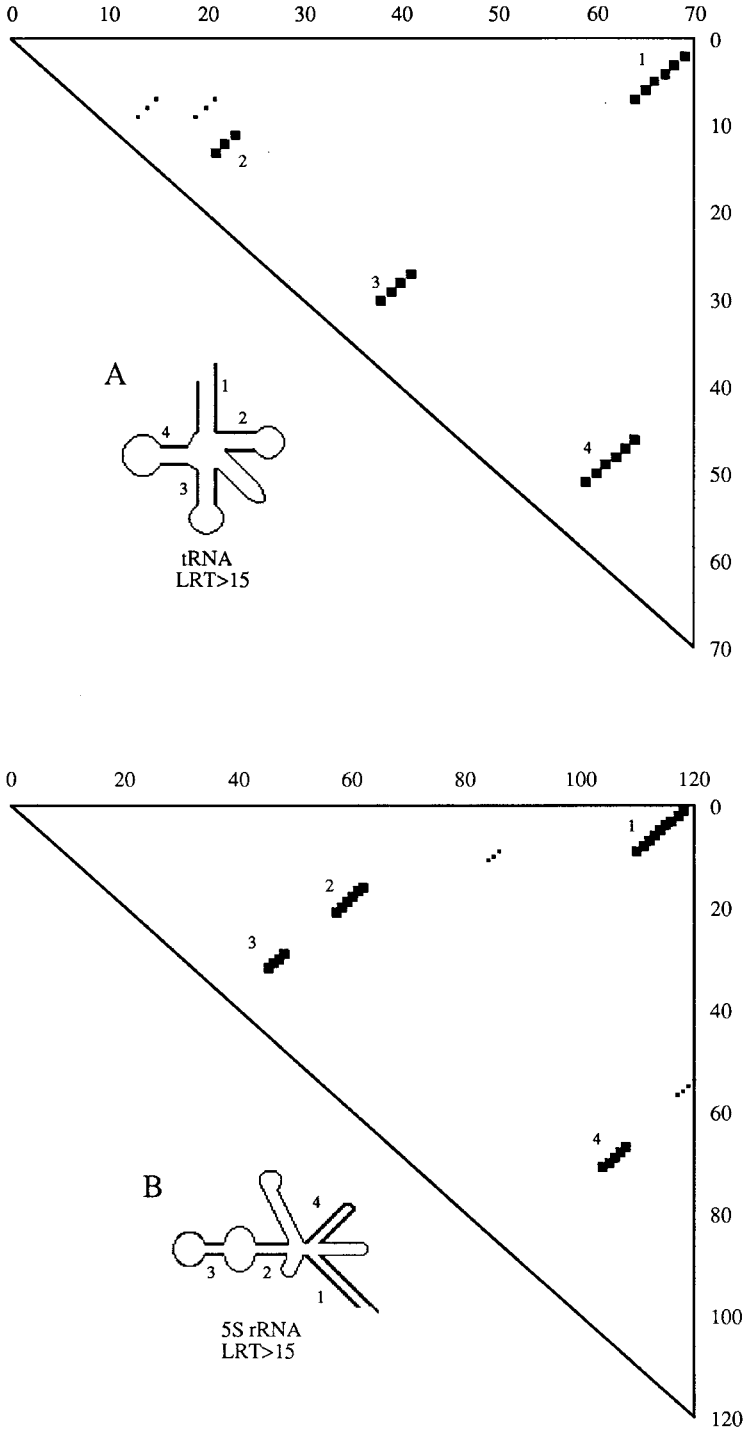



Figure 1.—Results of RNA secondary structure prediction for (A) tRNA, (B) 5S rRNA, (C) RNaseP RNA, (D) *bcd* mRNA 3′ UTR, and (E) SSU rRNA. The graphs show the $n \times n$ matrix for each RNA, where $n$ is the length of the alignment in bases. Helices identified by PIRANAH and meeting the minimum LRT requirement are plotted as diagonal lines, with the helices included in the final structure prediction by GROUPER (*i.e.*, the set of compatible helices with the greatest value of total LRT) shown in boldface. The inset shows the consensus structure for each RNA with the conserved helices shown in boldface and numbered corresponding to the above graph. Potential false positives (*i.e.*, helices included in the final structure prediction but not present in the consensus structure) are indicated by "?." In (C) the two RNase P pseudoknot pairings are indicated (pk1 and pk2).

crucial role in the early development of *D. melanogaster*. Proper localization of *bcd* mRNA to the anterior pole of the developing embryo is required for formation of head and thoracic segments (Berleth *et al.* 1988; Driever and Nüsslein-Volhard 1988). Signal sequences for *bcd* mRNA localization are contained within the 3′ UTR and form part of an extensive RNA secondary structure that is conserved in the genus Drosophila (Macdonald 1990; Seeger and Kaufman 1990). Our analysis of *bcd* 3′ UTR sequences from nine Drosophila species has identified eight conserved pairings with LRT $> 25$ (Table 1; Figure 2D). Helix 313-331/425-443 (Figure 1D) was initially split into two separate helices, 313-319/437-443 (LRT = 34.92) and 324-331/425-432 (LRT = 52.47), by PIRANAH due to the presence of internal mismatches. Similarly, helix 337-353/388-404 was split into 337-348/393-404 (LRT = 36.10) and 351-353/388-390 (LRT = 15.08). In both cases, the helices were combined after visual inspection and LRT was calculated for the extended helix. Helix 534-542/687-695,

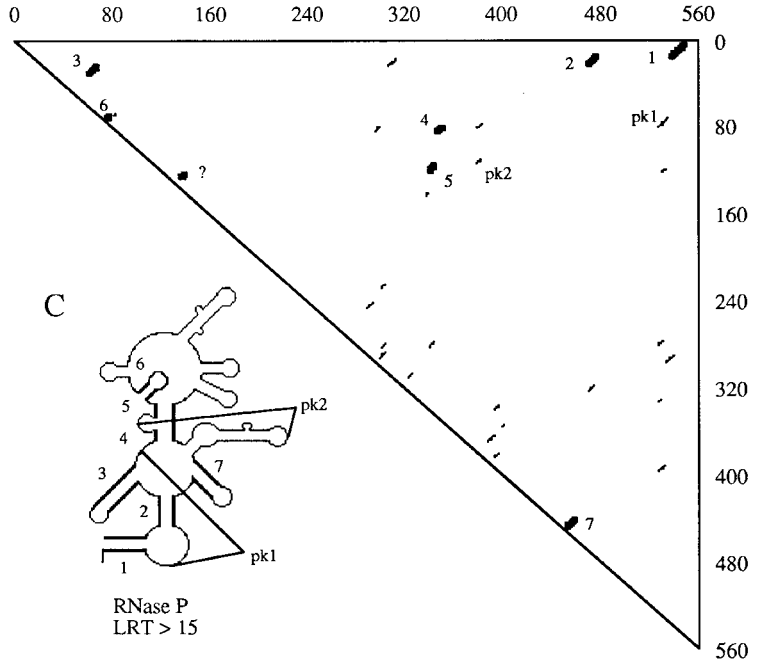
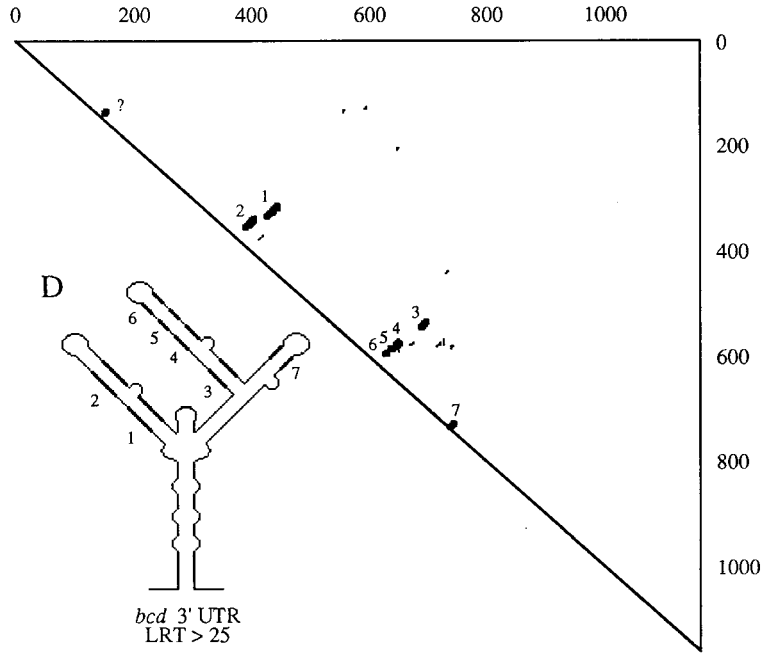



Figure 1.—*Continued.*




originally predicted as 537-542/687-692 (LRT = 30.48) due to an internal mismatch in species of the *melanogaster* subgroup, was extended by 3 bp and LRT was calculated for the extended helix. Seven of the predicted helices in our final structure are consistent with the model of Macdonald (1990). This model has been confirmed in large part by mutational analysis (Ferrandon *et al.* 1997; Macdonald and Kerr 1998). The one helix not present in the consensus model, 132-136/149-153, is a perfectly conserved short-range pairing for which there is no covariation support. We did not find support for the long-range pairings suggested by Macdonald (1990; regions I, II, and the lower portion of IV in his model). These pairings were predicted by thermodynamic folding of individual *bcd* 3′ UTR sequences, followed by comparison of the folded structures (Macdonald 1990)—not by strict phylogenetic comparison. It appears that even though similar structures are predicted for the *bcd* 3′ UTRs of several Drosophila species, the pairings are not between homologous regions. Such pairings would not be detected by our method or by the phylogenetic-comparative method. The lack of a

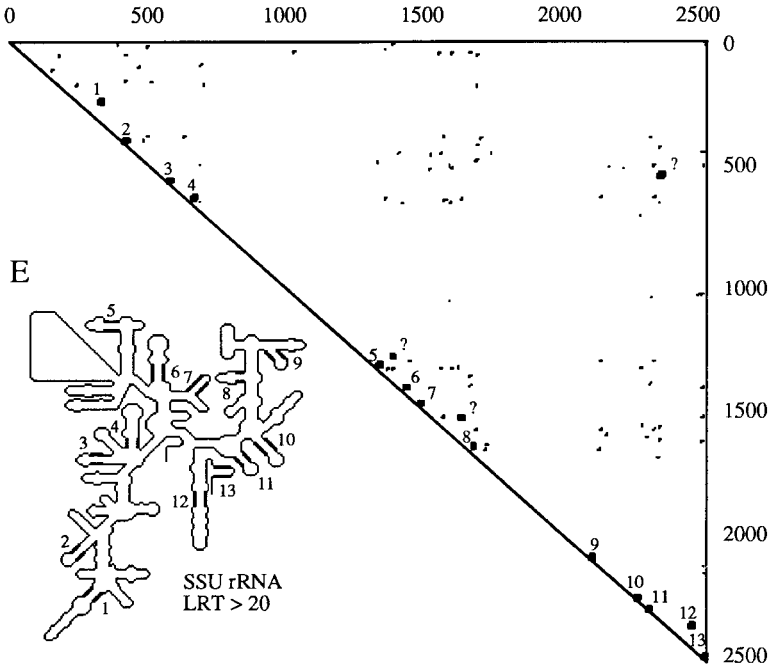



Figure 1.—*Continued.*

complete consensus structure makes it difficult to evaluate the accuracy of our predictions in this case. However, the thermodynamically predicted secondary structure of the *D. melanogaster bcd* 3′ UTR contains seven helices of length ≥3 that are conserved in the nine Drosophila species, all of which were identified in our analysis.

*SSU rRNA:* SSU rRNA is an integral part of the translational machinery of the cell and has a highly conserved structure in Bacteria, Eukarya, and Archaea (van de Peer *et al.* 1998). The SSU rRNA secondary structure has been predicted from extensive phylogenetic comparison; there are over 2800 aligned sequences presently available (van de Peer *et al.* 1998). Due to the relatively large sequence length, the high level of conservation, and the vast number of represented taxa, we chose to focus our analysis on five diverse eukaryotic species for which the phylogenetic relationship is unambiguous. Species included represent arthropods, nematodes, mammals, and fungi. The same five species were previously used in a study of SSU rRNA nucleotide substitution rates (Rzhetsky 1995). Because application of the PIRANAH program to the aligned sequences resulted in a large number of potential helices (over 650 with length ≥5), we used a hierarchical approach for final structure prediction by GROUPER. An original structure was generated using helices of length ≥6, LRT > 25, and at least one WC covariation. Helices with length ≥5, LRT > 20, and at least one WC covariation that were compatible with the helices determined above were then added to the structure. The final structure was composed of 16 helices (Table 1), 13 of which are consistent with previous models of eukaryotic SSU rRNA structure (Figure 1E; Maidak *et al.* 1997; van de Peer *et al.* 1998). An example is shown in Figure 2E. Three of these helices (627-633/669-675, 561-566/580-585, and 2096-2100/2112-2116), which contained either a terminal mismatch or terminal GU wobble pair, were extended by 1 bp after visual inspection and LRT was recalculated for the extended helix. The SSU rRNA consensus structure contains 15 helices of length ≥5 that meet the conservation criteria used in our analysis. Thus our method detected 87% of the consensus helices, with three potential false positives. One consensus helix that was not included in our final structure is a 5-bp helix, 1588-1592/2295-2299, which contains an internal GU wobble pair in all five species. The LRT value for this helix was 13.72; thus it did not meet our condition of LRT > 20. The reason for this low LRT value is that in Muse's (1995) algorithm GU wobble pairs are considered as mismatches. The other consensus helix that was not included in our final structure is a 7-bp helix, 1575-1581/2339-2345, which has a 1-base gap in the *S. stercoralis* sequence, but is otherwise perfectly conserved. This helix did have a relatively high LRT (40.52) but was not included in the final structure due to its lack of WC covariations. Each of the three potential false positives (helices 539-547/2360-2368, 1528-1532/1639-1643, and 1279-1283/1390-1394) is supported by only a single WC covariation occurring in one out of the five species.

**Analysis of covariations:** In this section we investigate whether the patterns of covariations observed in the inferred helices can be described by simple parameters. One interesting parameter that may affect the rate of compensatory molecular evolution is the length of helices in which compensatory substitutions occur. Thus, for each helix in the final secondary structure models, we determined the number of WC and wobble covariations present in the aligned sequences (Table 1). As terminal pairings at either end of a helix may be under

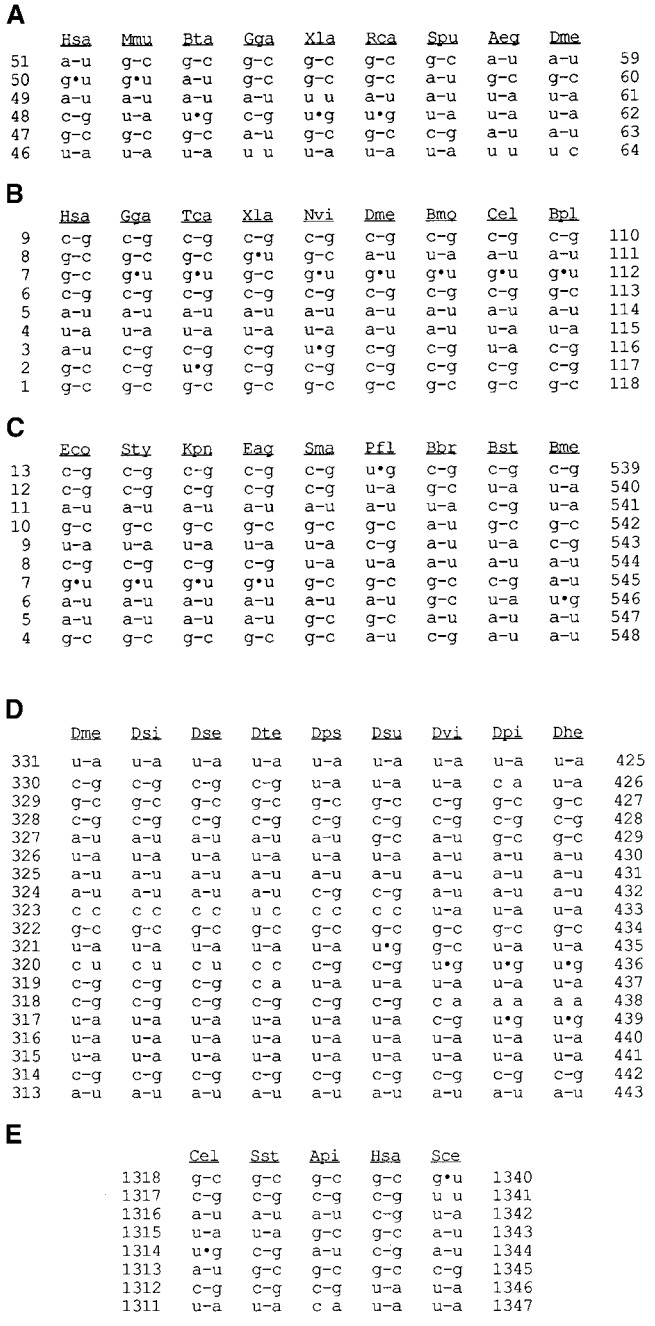



Figure 2.—Examples of RNA secondary structures predicted by PIRANAH/GROUPER. Representitive helices from (A) tRNA, (B) 5S rRNA, (C) RNaseP RNA, (D) *bcd* mRNA 3′ UTR, and (E) SSU rRNA are shown. The 5′ coordinates (according to the gapped alignment) are indicated at the far left; the 3′ coordinates at the far right. Species abbreviations are given at the top of each structure; complete species names are given in materials and methods. Watson-Crick pairs are shown connected by straight lines. GU wobble pairs are indicated by dots.

different selective constraints than internal pairings, we also determined the number of internal covariations.

Table 3 summarizes the results of our regression analyses. Highly significant correlations between the number of WC covariations per pair and stem length were found for the Drosophila *bcd* 3′ UTR mRNA and for the bacterial RNase P RNA, although in both cases only eight helices were identified. Both correlations are tighter for the internal covariations. A significant correlation was also observed for the internal covariations occurring in tRNA (which consists of four helices), but not for all covariations. For the ribosomal RNAs (both 5S and SSU), however, correlations between stem length and the number of covariations per pair were not found.

In Figure 3A, the number of internal covariations per pair is plotted against stem length for the *bcd* 3′ UTR mRNA helices. To increase the data set, two *Adh* pre-mRNA helices were included. The latter two helices were identified in the adult intron and in intron 1 of Drosophila *Adh* and are well supported statistically (Kirby *et al.* 1995) based on a similar set of sequences as the *bcd* analysis. Similarly, the number of internal covariations (per pair) of the RNase P RNA helices as a function of stem length is shown in Figure 3B. In both cases, the regression line describes the relationship

**TABLE 2**

**Significance of LRT values**

| Type of RNA | LRT | $P^a$ |
|---|---|---|
| (i) tRNA | | |
| Individual helices | 35.63 | 0.00 |
| | 33.61 | 0.00 |
| | 25.98 | 0.04 |
| | 25.82 | 0.04 |
| Total structure[b] | 121.04 | 0.00 |
| (ii) 5S rRNA | | |
| Individual helices | 40.60 | 0.00 |
| | 40.02 | 0.00 |
| | 28.77 | 0.02 |
| | 26.87 | 0.02 |
| Total structure[b] | 136.26 | 0.00 |

[a] The frequency of obtaining an LRT statistic ≥ the observed value from 100 random permutations of the original alignment.

[b] Total structures were predicted using a minimum LRT of 15. Total structure prediction using a minimum LRT of 10 produced identical results.

**TABLE 3**

**Results of regression analysis**

| Type of RNA | $R^2$ | $P$ |
|---|---|---|
| tRNA | 0.78 (0.95) | 0.12 (0.02) |
| 5S rRNA | 0.75 (0.59) | 0.14 (0.23) |
| RNase P RNA | 0.66 (0.92) | 0.01 (0.0002) |
| *bcd* 3′ UTR | 0.82 (0.95) | 0.002 (0.00003) |
| SSU rRNA | 0.01 (0.02) | 0.68 (0.51) |

The results of linear regression analyses for the numbers of covariations per pair *vs.* stem length are presented for total covariations and, in parentheses, for internal covariations.

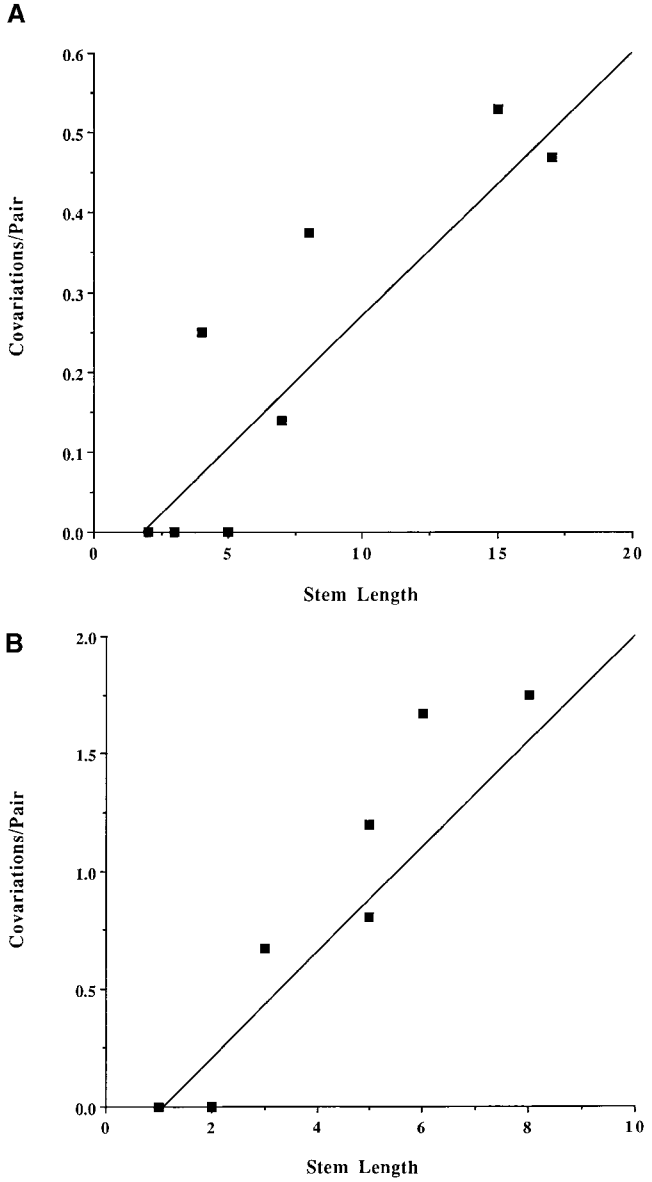



Figure 3.—Number of covariations per pair *vs.* stem length. The data are from (A) eight helices of Drosophila *bcd* mRNA and two helices of *Adh* pre-mRNA (see text) and (B) eight helices of bacterial RNase P RNA. The numbers of internal covariations are plotted.

between the number of covariations per pair and stem length well. For longer helices, the number of covariations is roughly proportional to stem length. Short helices (with total length $\leq 4$) show no internal covariations, however, so that the regression lines of Figure 3, A and B, intersect with the *x*-axis at positive values.

Our results suggest that compensatory evolution in the Drosophila mRNA and bacterial RNase P RNA structures occurs faster in longer helices. The most likely explanation for this observation is that selective constraints are relaxed in longer helices, because mutations occurring in longer stems result in less helix destabilization than those occurring in shorter stems. This hypothesis can be investigated further by defining a normalized number of covariations per pair for longer helices such that the number of covariations in a helix is scaled by the square of the stem length (instead of the stem length, as above). This definition is suggested by the proportionality between the number of covariations per pair and stem length for longer helices (see Figure 3, A and B). Thus, the normalized number of covariations per pair is expected to be nearly independent of differences in selective pressure for helices of different lengths. For pairings in RNA helices that are subjected to similar selection pressure, Kimura's (1985) model of compensatory evolution predicts that the rate of compensatory changes depends critically on the physical distance between the interacting nucleotides. If selection against mutations that destabilize a helix is much stronger than genetic drift, the rate of compensatory evolution is expected to decrease with physical distance (Kimura 1985; Stephan 1996).

We explored this prediction for the Drosophila mRNA and the bacterial RNase P RNA structures. In Figure 4, A and B, we plotted the total number of covariations (divided by the square of the stem length) for the longer helices of the Drosophila mRNA (both *bcd* 3′ UTR and *Adh*) and the bacterial RNase P RNA structures, respectively. To increase the number of helices containing covariations, the total number of covariations was considered (instead of internal covariations). In both cases, the six helices that exhibit covariations are shown. One longer helix (length = 7 bp) was removed from the *bcd* data because it did not have any covariations and may thus be under stronger selective constraints. For the Drosophila mRNA helices, a significantly negative correlation between physical distance and the normalized number of covariations per pair was found ($R^2 = 0.89$, $P < 0.005$); for the bacterial structures, no correlation was observed ($R^2 = 0.201$, not significant). For internal covariations, qualitatively similar results were obtained.

Based on predictions of Kimura's model (Stephan 1996) and estimates of Drosophila recombination rates (Lindsley and Sandler 1977), these results suggest that selection pressure on individual WC pairs in the longer Drosophila mRNA helices is relatively strong. The results for bacterial RNase P RNA are harder to interpret because the rate and pattern of recombination in bacteria are not well understood. Assuming that selection pressures are comparable in Drosophila and bacteria, the results may be explained by a lack of recombination in bacteria (see discussion).

## DISCUSSION

**Method of RNA secondary structure prediction:** Our approach to RNA secondary structure prediction has proven effective at identifying conserved pairing regions in five types of RNA: tRNA, 5S rRNA, RNase P RNA,

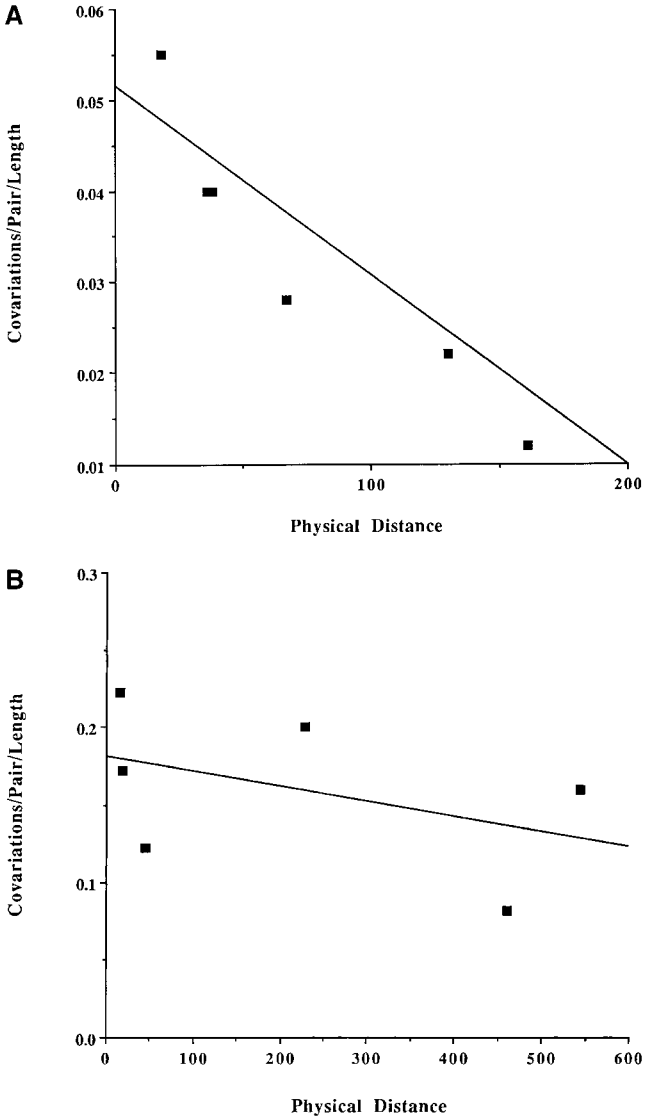



Figure 4.—Number of covariations per pair (scaled by stem length) *vs.* physical distance (in nucleotides). The data are from (A) four helices of Drosophila *bcd* mRNA and two helices of *Adh* pre-mRNA and (B) six helices of bacterial RNase P RNA. Only helices that are longer (≥5 bp) and have covariations are shown (see text). Here the numbers of total covariations are plotted.

Drosophila *bcd* mRNA, and eukaryotic SSU rRNA. The approach can be summarized as follows. A complete list of potential helices meeting specified length and conservation criteria is generated from an alignment of homologous RNA-encoding sequences. For each helix, the constraint for WC base-pairing is estimated by an LRT statistic (Muse 1995). Compatible helices are then combined to form a final secondary structure model that maximizes total LRT. The use of LRT in final structure assembly is an improvement over previous comparative methods, because LRT provides a quantitative measure of helix conservation that takes substitution patterns and phylogenetic relationships into consideration. Previous methods (*e.g.*, Han and Kim 1993) used simple measures, such as number of paired bases or number of mismatches, to rank helices for structure assembly. It is important to note that our method requires no *a priori* knowledge of the location of pairing regions, which was a limitation to most previous applications of LRT (Kirby *et al.* 1995; Muse 1995; Parsch *et al.* 1997). In addition, our method does not rely on the *ad hoc* rules used by the phylogenetic-comparative method (Fox and Woese 1975; James *et al.* 1988). In fact, covariations need not be present for a helix to be identified and included in the final structure (although helices containing covariations can be given additional weighting).

In four of our examples, tRNA, 5S rRNA, RNase P, and *bcd* 3′ UTR, we identified 100% of the conserved pairings in the established consensus structures. In the fifth example, SSU rRNA, the success rate was 87%. The number of false positives (predicted pairings that are not present in the consensus structure) was quite low for all five RNAs. Perhaps the most striking example of successful structure prediction is that of SSU rRNA. This is by far the longest of the five RNAs (1761–2487 nt depending on the species; 2533 nt in the gapped alignment) and also the one for which we used the fewest representative sequences (five). A previous study using the thermodynamic folding algorithm of Jaeger *et al.* (1990) reported a success rate of ~30% for eukaryotic SSU rRNA consensus helix prediction (Konings and Gutell 1995). It should be noted, however, that our method differs from thermodynamic prediction in that the former is designed to identify evolutionarily conserved helices from an alignment of homologous sequences, while the latter is designed to predict the secondary structure of a single RNA sequence.

An important consideration when using the above method is the choice of parameter values. Parameters must be chosen for both the initial identification of helices by PIRANAH and the assembly of the final secondary structure by GROUPER. Since the most time-consuming part of the process is LRT calculation during the initial identification step, we chose parameter values that would keep computation time reasonable by limiting the total number of potential helices. In practice, this means increasing the minimum score for LRT calculation and the minimum helix length as sequence length increases. The values we chose for these parameters were quite conservative, however, even for the longer sequences. Thus, the number of helices identified by PIRANAH far exceeded the number of helices included in the final structures of all five RNAs. For example, in the case of SSU rRNA (the longest of the five RNAs) only 4% of the helices identified by PIRANAH were included in the final structure prediction by GROUPER, and many of the helices (73%) had LRT values falling below the LRT cutoff used for final structure assembly. Thus it is very unlikely that any evolutionarily conserved

helices were overlooked due to the choice of parameter values in this initial step.

For the second step, assembly of the final secondary structure model, the major parameter value is minimum LRT. Typically, this value must be increased as the sequence length increases. In the case of short sequences, such as tRNA, 5S rRNA, or RNase P RNA, the LRT cutoff may be set relatively low (15 in our examples) because there are few conflicting helices. In these cases, nearly all of the helices in the final structure have LRT values well above the cutoff (see Table 1), so our choice of minimum LRT was a conservative one. In addition, for the two short sequences that were used for randomization simulations (tRNA and 5S rRNA; Table 2) the results were identical for LRT cutoffs of 15 and 10. This suggests that the choice of minimum LRT does not greatly affect the significance of the predicted structure. For very long RNA sequences, the hierarchical approach of the SSU rRNA example may be used. Here the first helices assembled by GROUPER were those that had a high LRT value and a long stem, *i.e.*, helices whose pairing potential was evolutionarily most conserved and that were thermodynamically most stable. In subsequent steps, helices with shorter lengths and lower values of LRT were added to the structure. Also, the number of potential helices may be reduced by requiring that at least one WC covariation be present in each helix. Such constraints can greatly simplify structure prediction in cases where there are many conflicting helices. This approach, however, may lead to structures that are incompatible with each other, depending on what the cutoff value of LRT for individual helices is. It may also overlook helices that are perfectly conserved and thus have no covariations. More work is required to explore this potential problem.

The above examples revealed another problem inherent in identifying potential helices by sequence comparisons; that is, the length of homologous pairing regions may differ among species due to internal or terminal mismatches. PIRANAH uses a set of strict rules about mismatches in searching for potential helices and may, for instance, find two helices where there would be only a single one if internal mismatches were allowed (see the *bcd* 3′ UTR example). PIRANAH may also fail to include the terminal base pair of a helix in cases where a mismatch or a GU wobble pair is present (see the SSU rRNA example). It is therefore advisable to inspect the output of PIRANAH before it is subjected to GROUPER. The rule we followed during visual inspection of the PIRANAH output was as follows: a helix was extended by including mismatches or GU wobble pairs only if the extended helix produced a greater value of LRT than the originally predicted helix.

Improvements to our method of secondary structure prediction will certainly be possible as computer processing time becomes less limiting. For example, the criteria used for initial helix identification by PIRANAH may be relaxed, allowing more potential helices to be identified and considered in final structure prediction. It may also be possible to integrate sequence alignment with secondary structure prediction. Currently, alignment and structure prediction are completely separate procedures. Alignments are typically adjusted manually so that potential pairing stems are at corresponding positions in all sequences (James *et al.* 1988). This process could be automated by calculating LRT for each potential stem under several different alignment schemes and choosing the alignment that maximizes LRT. Finally, increased computer processing power will make randomization simulations, such as those used for tRNA and 5S rRNA (Table 2), practical for longer sequences. This will allow meaningful *P* values to be assigned to the LRT statistics of individual helices and total structures.

**Effect of stem length:** Our analysis of covariations identified two parameters that are important for the evolution of compensatory mutations: the length of a helix and the physical distance between base-pairing residues. Positive correlations between the number of covariations (per pair) and stem length were observed for the Drosophila *bcd* 3′ UTR RNA, bacterial RNase P RNA, and tRNA (Table 3 and Figure 3). The observed correlations may be explained by differences in selective constraints. Selective constraints in longer helices appear to be relaxed because single mutations occurring in these helices result in less helix destabilization than those occurring in short stems.

It is noteworthy that a similar correlation was not found for the ribosomal RNAs, in particular SSU rRNA. A plot (not shown) of the helices of SSU rRNA (which includes also the helices of shorter length that were not considered in our analysis) suggests that there is an increase in the number of covariations for shorter stems and a decrease for longer ones, with a maximum rate for an intermediate stem length (of 6 bp). The increase in the number of covariations with stem length for shorter stems may be due to the relaxation of selective constraints with increasing helix length, as discussed above for the other types of RNA. However, other mechanisms, possibly related to the specific function of this type of RNA, have to be invoked to explain the decrease in the rate of compensatory evolution for longer SSU rRNA helices (Golding 1994).

**Distance effect:** For the two larger RNAs (Drosophila mRNA and bacterial RNase P RNA) that showed a positive correlation between stem length and the rate of covariation, we found that the number of covariations (per pair) scaled by stem length decreases with the physical distance between base-pairing nucleotides (Figure 4). In contrast to bacterial RNase P RNA, for the Drosophila mRNA helices this negative correlation was found to be highly significant.

Kimura's (1985) model of compensatory evolution provides a simple explanation for these results. All other

things being equal (in particular, selection pressure on base-pairing residues), this model suggests that the difference may be due to a lack of recombination in bacteria. Indeed, using reasonable estimates of Drosophila recombination rates (Lindsley and Sandler 1977) and of effective population size, $N$, and assuming sufficiently strong selection on individual WC pairs, Stephan's (1996) formula (8c) predicts a substantial decay in the rate of compensatory evolution over a distance of 100 nucleotides, as applies to these data (Figure 4A). [In formula (8c) mentioned above, $A$ may be approximated by $2/3\ N$; W. Stephan, unpublished result.]

This theory suggests that the strength of selection on individual WC pairs, measured by the parameter $2Ns$ (where $s$ is the selection coefficient), is on average much larger than one for the Drosophila data. On the other hand, the occurrence of wobble pairs and mispairings in the helices (see Figure 2D and Table 1) indicates that the strength of selection may vary substantially among base pairs and that some stem evolution proceeds through slightly deleterious intermediates. A statistical method is needed to estimate the parameter $2Ns$ directly from comparative sequence data.

We are grateful to K. Han and S. Muse who kindly made their programs and C code available. Furthermore, we thank two reviewers for their critical comments and helpful suggestions. The programs and sequence alignments used in our analyses, as well as additional documentation, are available at http://maple.lemoyne.edu/~braverjm/ss.html. This research was supported in part by a National Science Foundation/Sloan Foundation postdoctoral fellowship to J.M.B., and National Institutes of Health grant GM-58405 to W.S.

## LITERATURE CITED

Berleth, T., M. Burri, G. Thoma, D. Bopp, S. Richstein *et al.*, 1988 The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo. EMBO J. **7:** 1749–1756.

Brown, J. W., 1998 The ribonuclease P database. Nucleic Acids Res. **26:** 351–352.

Dirheimer, G., G. Keith, P. Dumas and E. Westhof, 1995 Primary, secondary, and tertiary structures of tRNAs, pp. 93–126 in *RNA: Structure, Biosynthesis, and Function*, edited by D. Söll and U. Rajbhandry. American Society for Microbiology, Washington, DC.

Driever, W., and C. Nüsslein-Volhard, 1988 A gradient of *bicoid* protein in Drosophila embryos. Cell **54:** 83–93.

Ekland, E. H., and D. P. Bartel, 1996 RNA-catalysed RNA polymerization using nucleoside triphosphates. Nature **382:** 373–376.

Ferrandon, D., I. Koch, E. Westhoff and C. Nüsslein-Volhard, 1997 RNA-RNA interaction is required for the formation of specific *bicoid* mRNA 3′ UTR-STAUFEN ribonucleoprotein particles. EMBO J. **16:** 1751–1758.

Fox, G. E., and C. R. Woese, 1975 5S RNA secondary structure. Nature **256:** 505–507.

Gilbert, W., 1986 The RNA world. Nature **319:** 618.

Golding, B., 1994 Using maximum likelihood to infer selection from phylogenies, pp. 126–139 in *Non-neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

Gouy, M., and W.-H. Li, 1989 Molecular phylogeny of the kingdoms Animalia, Plantae, and Fungi. Mol. Biol. Evol. **6:** 109–122.

Guerrier-Takada, C., K. Gardiner, T. Marsh, N. Pace and S. Altman, 1983 The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell **35:** 849–857.

Haas, E. S., D. P. Morse, J. W. Brown, F. J. Schmidt and N. R. Pace, 1991 Long-range structure in ribonuclease P RNA. Science **254:** 853–856.

Han, K., and H.-J. Kim, 1993 Prediction of common folding structures of homologous RNAs. Nucleic Acids Res. **21:** 1251–1257.

Jaeger, J. A., D. H. Turner and M. Zuker, 1990 Predicting optimal and suboptimal secondary structure for RNA, pp. 281–306 in *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences* (Methods in Enzymology, Vol. 52), edited by R. F. Doolittle. Academic Pres, San Diego.

James, B. D., G. J. Olsen, J. Liu and N. R. Pace, 1988 The secondary structure of ribonuclease P RNA, the catalytic element of a ribonucleoprotein enzyme. Cell **52:** 19–26.

Joyce, G. F., and L. E. Orgel, 1993 Prospects for understanding the origin of the RNA world, pp. 1–25 in *The RNA World*, edited by R. F. Gesteland and J. F. Atkins. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Kim, S. H., F. L. Suddath, G. J. Quigley, A. McPherson, J. L. Sussman *et al.*, 1974 Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. Science **185:** 435–440.

Kimura, M., 1985 The role of compensatory neutral mutations in molecular evolution. J. Genet. **64:** 7–19.

Kirby, D. A., S. V. Muse and W. Stephan, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92:** 9047–9051.

Konings, D. A. M., and R. R. Gutell, 1995 A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. RNA **1:** 559–574.

Krüger, K., P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling *et al.*, 1982 Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell **31:** 147–157.

Kumar, S., and A. Rzhetsky, 1996 Evolutionary relationships of eukaryotic kingdoms. J. Mol. Evol. **42:** 183–193.

Lindsley, D. L., and L. Sandler, 1977 The genetic analysis of meiosis in female *Drosophila melanogaster*. Philos. Trans. R. Soc. Lond. B **277:** 295–312.

Macdonald, P. M., 1990 *bicoid* mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. Development **110:** 161–171.

Macdonald, P. M., and K. Kerr, 1998 Mutational analysis of an RNA recognition element that mediates localization of *bicoid* mRNA. Mol. Cell. Biol. **18:** 3788–3795.

Macdonald, P. M., and G. Struhl, 1988 *Cis*-acting sequences responsible for anterior localization of *bicoid* mRNA in *Drosophila* embryos. Nature **336:** 595–598.

Maidak, B. L., G. J. Olsen, N. Larsen, R. Overbeek, M. J. McCaughey *et al.*, 1997 The RDP (ribosomal database project). Nucleic Acids Res. **25:** 109–111.

Mullner, E. W., and L. C. Kuhn, 1988 A stem-loop in the 3′ untranslated region mediates iron-dependent regulation of transferrin receptor mRNA stability in the cytoplasm. Cell **53:** 815–825.

Muse, S. V., 1995 Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139:** 1429–1439.

Odai, O., H. Kodama, H. Hiroaki, T. Sakata, T. Tanaka *et al.*, 1990 Synthesis and NMR study of ribo-oligonucleotides forming a hammerhead-type RNA enzyme system. Nucleic Acids Res. **18:** 5955–5960.

Osswald, M., and R. Brimacombe, 1999 The environment of 5S rRNA in the ribosome: cross-links to 23S rRNA from sites within helices II and III of the 5S molecule. Nucleic Acids Res. **11:** 2283–2290.

Pace, N. R., and D. Smith, 1990 Ribonuclease P: function and variation. J. Biol. Chem. **265:** 3587–3590.

Pace, N. R., D. K. Smith, G. J. Olsen and B. D. James, 1989 Phylogenetic comparative analysis and the secondary structure of ribonuclease—a review. Gene **82:** 65–75.

Pandey, N. B., A. S. Williams, J. H. Sun, V. D. Brown, U. Bond *et al.*, 1994 Point mutations in the stem-loop at the 3′ end of mouse histone mRNA reduce expression by reducing the efficiency of 3′ end formation. Mol. Cell. Biol. **14:** 1709–1720.

Parsch, J., S. Tanda and W. Stephan, 1997 Site-directed mutations reveal long-range compensatory interactions in the *Adh* gene of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **94:** 928–933.

Pley, H. W., K. M. Flaherty and D. B. McKay, 1994 Three-dimensional structure of a hammerhead ribozyme. Nature **372:** 68–74.

Rzhetsky, A., 1995 Estimating substitution rates in ribosomal RNA genes. Genetics **141:** 771–783.

Schöniger, M., and A. von Haeseler, 1994 A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. **3:** 240–247.

Scott, W. G., J. T. Finch and A. Klug, 1995 The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. Cell **81:** 991–1002.

Seeger, M. A., and T. C. Kaufman, 1990 Molecular analysis of the *bicoid* gene from *Drosophila pseudoobscura*: identification of conserved domains within coding and noncoding regions of the *bicoid* mRNA. EMBO J. **9:** 2977–2987.

Specht, T., J. Wolters and V. A. Erdmann, 1991 Compilation of 5S rRNA and 5S rRNA gene sequences. Nucleic Acids Res. **19:** 2189–2191.

Sprinzl, M., C. Horn, M. Brown, A. Ioudovitch and S. Steinberg, 1998 Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. **26:** 148–153.

Stephan, W., 1996 The rate of compensatory evolution. Genetics **144:** 419–426.

Stephan, W., and D. A. Kirby, 1993 RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics **135:** 97–103.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins, 1997 The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876–4882.

Tillier, E. R. M., and R. A. Collins, 1995 Neighbor-joining and maximum likelihood with RNA sequences: addressing interdependence of sites. Mol. Biol. Evol. **12:** 7–15.

Unrau, P. J., and D. P. Bartel, 1998 RNA-catalysed nucleotide synthesis. Nature **395:** 260–263.

van de Peer, Y., A. Caers, P. de Rijk and R. de Wachter, 1998 Database on the structure of small ribosomal subunit RNA. Nucleic Acids Res. **26:** 179–182.

Woese, C. R., and G. E. Fox, 1977 Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA **74:** 5088–5090.

Woese, C. R., and N. R. Pace, 1993 Probing RNA structure, function, and history by comparative analysis, pp. 91–117 in *The RNA World*, edited by R. F. Gesteland and J. F. Atkins. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Communicating editor: G. B. Golding