CHAPTER 7

Structural and Evolutionary Considerations for Multiple Sequence Alignment of RNA, and the Challenges for Algorithms That Ignore Them

KARL M. KJER

Rutgers University

USMAN ROSHAN

New Jersey Institute of Technology

JOSEPH J. GILLESPIE

University of Maryland, Baltimore County; Virginia
Bioinformatics Institute, Virginia Tech

IDENTIFICATION OF GOALS

What Is It You are Trying to Accomplish with an Alignment? Some of the disagreement over alignment approaches comes from differences in objectives among investigators. Are the data merely meant to distinguish target DNA from contaminants in a BLAST search? Or is there a specific node on a cladogram you wish to test? Are you aligning genomes or genes? Are the data protein-coding, structural RNAs or noncoding sequences? Do you consider phylogenetics to be a process of inference or estimation? Would you rather be more consistent or more accurate? Are you studying the performance of your selected programs or the relationships among your taxa? Different answers to each of these questions could likely lead to legitimate alternate alignment approaches. Morrison (2006) reviewed the many uses of alignment programs, and distinguished phylogenetic alignments as a special subset that requires attention to biological processes. Hypša (2006) reached a similar conclusion and emphasized the importance of adding complexity to multiple sequence alignments and phylogeny estimation. This chapter is devoted to discussing the alignment of structural RNAs, or ribosomal RNA (rRNA) and transfer RNA (tRNA) sequences, for phylogenetic analysis (although our thesis applies to other smaller RNAs, such as tmRNAs, RNase Ps, and group I and group II introns). We seek to have our phylogenetic hypotheses be predictive and accurate, even if accuracy is difficult (or impossible) to demonstrate. By *accurate* we mean that a hypothesis coincides with the true history of branching events.

If You Knew You could Improve Your Alignment, Would You Do It? Figure 7.1 shows an example of a fragment of a computer-generated alignment of the 12S rRNA from a variety of primates with that of murine rodent outgroups. It follows an optimality criterion based on minimizing costs from a Needleman–Wunsch (1970) algorithm and a guide tree.

We hope that this example, almost like the first couplet of a dichotomous key, will indicate where you stand on the issue of adjustment.

```
Mouse      GCTACATTTTCTTA--TAAAAGAACAT-TACTATACCCTTTATGA
Rat        GCTACATTTTCTTTTCCCAGAGAACAT-TACGAAACCCTTTATGA
Gibbon     GCTACATTTTCTA--TGCC-AGAAAAC-CACGATAACCCTCATGA
Baboon     GCTACATTTTCTA--CTTCAGAAAACCCCACGATAGCTCTTATGA
Orangutan  GCTACATTTTCTA---CTTCAGAAAAC-TACGATAGCCCTCATGA
Human      GCTACATTTTCTA---CCCCAGAAAAC-TACGATAGCCCTTATGA
Bonobo     GCTACATTTTCTA--CCCC-AGAAAAT-TACGATAACCCTTATGA
Chimp      GCTACATTTTCTA--CCCC-AGAAAAT-TACGATAACCCTTATGA
```

Figure 7.1. A fragment of an alignment of complete 12S rRNA, generated by ClustalX (Jeanmougin et al. 1998; Thompson et al. 1997).

Notice the "CTTCAGAAAAC" in the middle of the figure for both the baboon and the orangutan. If these were your data, would you adjust the sequences to *correct* the *errors* made by the program, or would you leave it alone and let the program make all of the decisions about homology, even when it appears to have erred? Would you adjust the nearly identical sequences between the human and the chimps? Would you make decisions about which nucleotides to exclude? Would it bother you if the baboon grouped inside the rodents? There are no correct answers, and the choices you make have implications that relate to what you wish to discover from your data, and tell a lot about both your background and your objectives. If you adjust the alignment, even in this one instance, you have converted to a manual alignment, with all its strengths and limitations. Adjusting the alignment is an attempt to improve the accuracy of an alignment. Here we define "accurate" as representing true (unknowable) homology, and also propose that accurate homology estimations will probably improve the accuracy of the phylogenetic hypothesis. But how do you know what "accurate" is, and where do you draw the line? Is manual alignment an art form subject to the whims and biases of the aligner, or can we identify a repeatable methodology? Similarly, if we criticize manual alignments as subjective and inconsistent, might these same criticisms apply to computer-generated alignments? If accuracy is a concern, where do current algorithms fail?

Many workers could legitimately state that we cannot objectively define errors made by the computer, and, in fact, the whole concept would be counter to an optimality-based study. If you are looking for the shortest tree, you should favor an alignment that reduces the number of steps. Others might assume that a few errors, even if they

```
Mouse     GCTACATT(TTCT TATA--AA AGAA)CAT--TACTATACCCTTTATGA
Rat       GCTACATT(TTCT TTTCCCAG AGAA)CAT--TACGAAACCCTTTATGA
Gibbon    GCTACATT(TTCT -ATGCC-- AGAA)AAC--CACGATAACCCTCATGA
Baboon    GCTACATT(TTCT -ACTTC-- AGAA)AACCCCACGATAGCTCTTATGA
Orangutan GCTACATT(TTCT -ACTTC-- AGAA)AAC--TACGATAGCCCTCATGA
Human     GCTACATT(TTCT -ACCCC-- AGAA)AAC--TACGATAGCCCTTATGA
Bonobo    GCTACATT(TTCT -ACCCC-- AGAA)AAT--TACGATAACCCTTATGA
Chimp     GCTACATT(TTCT -ACCCC-- AGAA)AAT--TACGATAACCCTTATGA
```

Figure 7.2. A structurally adjusted alignment of the same data as shown in Figure 7.1. Parentheses indicate the bounds of a hairpin-stem loop, with hydrogen-bonded nucleotides indicated with underlines (Kjer et al. 1994). An unaligned region (the "loop" portion of the hairpin-stem) is delimited with spaces.

could be defined, would be better left alone, because the mass of the data should counterbalance a few random errors. But what if the errors are not random, creating a potential for linking together unrelated groups that share the same systematic biases? What if there were some higher order of conservation that we could examine in making decisions about homology that does not necessarily result in shorter trees? Figure 7.2 shows the same region of rRNA as in Figure 7.1, but has been adjusted to minimize secondary structural changes predicted for this region of the molecule. Minimizing structural change is also an optimality criterion for homology assessment that we support and will explore in this chapter. Structural homology is based on position and connection, and assumes that the same structure existed in a common ancestor. Strict adherence to nucleotide homology may require that the same nucleotide state exists in a common ancestor as in its descendants, and, by this definition, structural homology and nucleotide homology may support different alignments.

ALIGNMENT AND ITS RELATION TO DATA EXCLUSION

One of the things that are not clear from the above comparisons (Figures 7.1 and 7.2) is what we should do with the nucleotides in the "loop" portion of the hairpin-stem loop, between the "TTCT" and the "AGAA." This is an extremely important issue, but somewhat outside the debate about alignment. Frequently, these regions are excluded from the analysis on the grounds that they are too variable to align. Some systematists find any form of data exclusion to be unacceptable,

```
          A      B            C
                              REC     RAA     REC'
Mouse     ?????  [TATA--AA] 1 [T-] 1 [ATAA-] 1 [-A] 1
Rat       ?????  [TTTCCCAG] 2 [TT] 2 [TCCC-] 2 [AG] 2
Gibbon    ATGCC  [-ATGCC--] 3 [--] 3 [ATGCC] 3 [--] 3
Baboon    ACTTC  [-ACTTC--] 4 [--] 3 [ACTTC] 4 [--] 3
Orangutan ACTTC  [-ACTTC--] 4 [--] 3 [ACTTC] 4 [--] 3
Human     ACCCC  [-ACCCC--] 5 [--] 3 [ACCCC] 5 [--] 3
Bonobo    ACCCC  [-ACCCC--] 5 [--] 3 [ACCCC] 5 [--] 3
Chimp     ACCCC  [-ACCCC--] 5 [--] 3 [ACCCC] 5 [--] 3
```

Figure 7.3. Three suggestions for dealing with the unaligned loop from Figure 7.2. (A) If the in-group is alignable, but the out-group cannot be aligned to the in-group, consider the out-group data as missing. (B) Eliminate the entire region, but recode it as multistate characters. Taxa sharing identical sequences are coded with the same state. (C) Gillespie's (2004) method of finding the unpaired middle, to break up the region into the ambiguously aligned loop, and flanking regions of slippage.

and those are typically the same researchers who would not adjust the alignment given in Figure 7.1. It is possible that these regions contain some potentially informative characters, yet in our observations the ability to objectively retrieve signal from these regions quickly becomes confounded with increased sequence divergence across an alignment. There is a wide variety of treatments for these data. One option would be to transform the out-group states to "missing data" (Kjer et al. 2001), with the idea that if you do not have any reasonable confidence of homology between the rodents and the primates, the data really are "missing," and since you are interested in the relationships among the in-group taxa, and in-group motifs are of uniform length, they could be treated as in Figure 7.3A. Alternatively, you could exclude the unaligned nucleotides, and recode them as multistate characters (Figure 7.3B), or as fixed state characters as in Giribet and Wheeler (2001). Taken a step further, a step matrix that calculates the minimum number of steps to transform one state to another could be applied to these multistate characters (Lutzoni et al. 2000). For example, the fewest number of steps it would take to transform ACCCC (state 5) into ACTTC (state 4) is two, just as ACCCC is two steps from ATGCC, but ACTTC is three steps from ATGCC (Lutzoni et al. 2000; Wheeler 1999). Gillespie (2004) suggested that these regions are difficult to align due to the expansion and contraction of the more variable hairpin-stem loops of rRNA. He proposed a method that defines regions of ambiguous alignment, slippage, and regions of expansion and contraction (called RAA/RSC/REC coding), which subdivides these ambiguous regions based on their structural properties and is directly applicable to the methods of Kjer et al. (2001) and Lutzoni et al. (2000). A demonstration of three alternative treatments is shown in Figure 7.3C.

DIFFERENTIATION OF MOLECULES

It is obvious that the selective processes involved in the effects of insertions or deletions (indels) on the function of a gene (and thus the probability of observing such a change in a living organism) are completely different for structural RNAs and protein-coding genes. An indel of one or two nucleotides in a protein-coding gene results in a reading frame shift, whereas an indel of even three nucleotides adds or subtracts a codon. Thus, a single indel will most likely have a major effect on protein structure and function. Indels in structural RNA genes are very different. The effect an indel has on structural RNA is variable across sites of the gene. For instance, some regions of rRNA are highly conserved in length across phylogenetic domains whose common ancestors stretch back for billions of years (Gutell 1996), implying that there is little or no tolerance for length variation in these regions of a functional ribosome. Other regions freely tolerate insertions and deletions, as observed among the most recently divergent species (Schnare et al. 1996). So the location of indels, their frequency, and their length in ribosomal RNAs are determined by the affects they have on rRNA structure and hence function. Indels in rRNA are not randomly distributed, but typically highly clustered into regions called expansion segments (or variable regions) that are much reduced, or nonexistent, in prokaryotes and lower eukaryotes. These expansion segments are located on the surface of the ribosome in regions not considered critical for ribosome function (Ban et al. 2000; Cate et al. 1999; Schluenzen et al. 2000; Spahn et al. 2001; Wimberly et al. 2000; Yusupov et al. 2001); thus their evolution can be considered less constrained than that of the core rRNA. AUQ1

There are other differences in alignment protocols that are dependent on the kinds of questions an investigator is attempting to answer. Researchers who study the evolution of genes need to look at structural variation. Information about how a protein evolves across kingdoms includes major rearrangements, missing amino acids, and large insertions, which may make alignment more difficult. At the level where we observe major substitution of codons, there is often a coincident saturation of nucleotides, typically at the first and second codon positions. On the other end of the spectrum, if you are a population geneticist looking for patterns among recently diverged populations, you may encounter variation in noncoding regions of the genome. So investigators at both the deepest and the shallowest levels of divergence may confront serious alignment problems that we do not address in this chapter. An alignment of a protein-coding gene with so many indels that render homology assignment ambiguous is probably not an ideal marker for phylogenetic studies (note, for example, how many phylogenetic papers state that their protein-coding genes were length invariant, or that alignment was trivial). Similarly, noncoding regions, such as introns, are relatively rare sources for estimating phylogenies. So, for phylogenetic systematists, alignment problems are most frequently encountered with rRNAs or tRNAs. Those who design alignment algorithms are often interested in serving all investigators, however, and many programs are specifically designed to align proteins, with the default parameters set for protein-coding genes. All of these statements seem intuitively obvious. Yet, how many times in the literature have we seen phylogenetic studies state in their methods sections that rRNAs were aligned with *default parameters*, that is, using a program whose defaults were set to align proteins? There seems to be a basic misunderstanding, or at least a lack of concern, about differentiating alignment processes according to the effect that indels have on the kinds of genes that are being aligned (but see Benavides et al. 2007). We find a disconnection between alignment philosophy and biological and evolutionary constraints. Does constructing an alignment based on maximizing nucleotide identity make sense for rRNA?

rRNA Sequences Evolve under Structural Constraints

That nucleotides in rRNA do not evolve parsimoniously can be unambiguously demonstrated. Put another way, rRNA structures change more slowly than do the nucleotides that they comprise. The Gutell laboratory (http://www.rna.icmb.utexas.edu/) maintains a database, the Comparative RNA Web Site (Cannone et al. 2002), from which secondary structural diagrams can be downloaded. To demonstrate the nonconservation of nucleotides, relative to structural features, we suggest you download and print any two structural diagrams

A)

```
Austroagrion   UUAAUUAAUUUAAUUUGGUUAGUGU--UACAUAACUAUCAAU-AAUAUUUAAUUAG
Platycypha            UUAAU-AAUUUAAUUUGUUUGUUGU--GAUAUAAUUGUCAAU-AAUAUUUA-UUAG
Neoneura              UUAAU-AAUUUAAUUUAUAUAUUGU--UAUAUAAAUAUUAAU-AAUAAUUA-UUAG
Megaloprepus   CUAAU-AUUUUAUUUUAUUCAGUGU--AUCAUAAUUGUUAAU-AAUAUUAAUUAG-
Dysphaea              CUGGU-AAUUUAAUUUAUUUAUUGU--AGCAUAAAUAUUAAU-AAUAAUUAUCUG-
Uropetalura    ---CUAACAUUAUAAUUUAUUUGAUGUUACAUAAUCAUUAAA-AAUAUAAGUUAG-
Tanypteryx            ---CUAAUUAAAUAAUUUAAUUGGUGUUAUAUAACCAUUAAU-AAUAUAAAUUAG-
Oxygastra             ---CUAAAUUAAUAUUUUAUUUAUUGUUAUAUAAAUAUUAAA-AAUAUAAUUUAG-
Libellula             ---UUAAAUUAUAUUUUAGGUUAAUGGGA-AUAAUUAUUAAU-AAUAUAAUUUAG-
Chorismagrion  -CUAUCUAUUUAUUUUAUUGGUUGU--UGCAUAAACGUUAAU-AAUUUAUUGGUAG
Gomphus        ---CUAAAUUUGAAUUGGUGGUGGUGGUAUAUAAUCAUAAUU-AAUUUAAUUUUAG
Argia                 -CUAAAUUUUUAAUUUAUUUAUUGU--AAUAUAAAUAUUAAU-AAUA-AAUUUUGG
Hypopetalia    ---CUAGUUUAAUAAUUUAUUUAAUGUUUUAUAAUUAUUAGA-AAUAUAAACUAG-
Macromidia            ---CUAGAUUAUAGAUUUAUUUAAUGUGAUAAGAUUAUUAAA-GAUUUUAUUUAG-
Neogomphus            ---CUAAAUUUAUAAUUUCUUUAAUGUUUUAUAAUUAUGGAA-AAUUUAUUUUUAG
Amphiagrion    ---CUAAAACUUUAAUCUGUUUAUUGUUACAUAAAUAUCUGA-AAUAUUUUUUAG-
Progomphus            ---CUAAAACUAUAAUUUUUUUAAUGUUUCAUAAUUAUAUAU-AAUAUAGUUUUAG
Hagenius              ---CUAAAACCA-GUUAAAAUUAAUGUGGCAUAAUUAUAGUUUAACUGGGUUUUAG
Macromia              ---CUAUGUUAGUAAUUUAUUUAAUGUGGAAUAAUUAUUGAU-AAUACAUCAUAG-
Macromia              ---CUAUGUUAG-AAUUUAUUUAAUGUGGAAUAAUUAUUAAU-AAUACAUCAUAG-
Calaphaea             CUGAUUUGUUUG--AUUUGGUUAAUGUGUUAUAAUUAUCUUA-AAUAC-UCAUCUG
                                                                          ^
```

Figure 7.4. An example of how nucleotide changes should not be used to assess alignment quality. This is an example of a hairpin-stem loop structure, with hydrogen-bonded nucleotides in bold, and underlined. The first five nucleotides bind with the last five. Then there is a large bulge, followed by another four nucleotide interaction (UAGU/ACUA in the top sequence). (A) Aligned with ClustalX. (B) Structurally adjusted.

of the same rRNA sequence from distantly related taxa, and then superimpose one upon the other; hold them up to the light (or make transparencies of each, and superimpose them on a white piece of paper). If the structures between organisms are conserved, but the nucleotides within these structures are relatively less conserved, then you have proved to yourself that minimizing change among nucleotides does not make biological sense. It is not only the number of nucleotide changes that computer programs should seek to minimize, but, rather, they should seek to minimize change among structural features as a higher level of conservation, and then consider minimizing nucleotide changes after structural conservation has been optimized. Figure 7.4 shows an example. In Figure 7.4, we show a highly conserved stem of five nucleotides. The first five nucleotides in Figure 7.4 are hydrogen-bonded to the last five nucleotides in each taxon. But that is not how ClustalX, with default parameters, aligned them. If we adjust the final five nucleotides in the two *Macromia* sequences

B)

```
Austroagrion  UUAAUUAAUUUAAUUUGGUUAGUGUUACAUAACUAUCAAU-AAUAUUUAAUUAG
Platycypha           UUAAU-AAUUUAAUUUGUUUGUUGUGAUAUAAUUGUCAAU-AAUAUUU-AUUAG
Neoneura             UUAAU-AAUUUAAUUUAUAUAUUGUUAUAUAAAUAUUAAU-AAUAAUU-AUUAG
Megaloprepus  CUAAU-AUUUUAUUUUAUUCAGUGUAUCAUAAUUGUUAAU-AAUA-UUAAUUAG
Dysphaea             CUGGU-AAUUUAAUUUAUUUAUUGUAGCAUAAAUAUUAAU-AAUAAUU-AUCUG
Uropetalura   CUAAC-AUUAUAAUUUAUUUGAUGUUACAUAAUCAUUAAA-AAUAUAA-GUUAG
Tanypteryx           CUAAU-UAAAUAAUUUAAUUGGUGUUAUAUAACCAUUAAU-AAUAUAA-AUUAG
Oxygastra            CUAAA-UUAAUAUUUUAUUUAUUGUUAUAUAAAUAUUAAA-AAUAUAA-UUUAG
Libellula            UUAAA-UUAUAUUUUAGGUUAAUGGGA-AUAAUUAUUAAU-AAUAUAA-UUUAG
Chorismagrion CUAUC-UAUUUAUUUUAUUGGUUGUUGCAUAAACGUUAAU-AAUUUAUUGGUAG
Gomphus       CUAAA-UUUGAAUUGGUGGUGGUGGUAUAUAAUCAUAAUU-AAUUUAAUUUUAG
Argia                CUAAA-UUUUUAAUUUAUUUAUUGUAAUAUAAAUAUUAAU-AAUA-AAUUUUGG
Hypopetalia   CUAGU-UUAAUAAUUUAUUUAAUGUUUUAUAAUUAUUAGA-AAUAUAA-ACUAG
Macromidia           CUAGA-UUAUAGAUUUAUUUAAUGUGAUAAGAUUAUUAAA-GAUUUUA-UUUAG
Neogomphus           CUAAA-UUUAUAAUUUCUUUAAUGUUUUAUAAUUAUGGAA-AAUUUAUUUUUAG
Amphiagrion   CUAAA-ACUUUAAUCUGUUUAUUGUUACAUAAAUAUCUGA-AAUAUUU-UUUAG
Progomphus           CUAAA-ACUAUAAUUUUUUUAAUGUUUCAUAAUUAUAUAU-AAUAUAGUUUUAG
Hagenius             CUAAA-ACCA-GUUAAAAUUAAUGUGGCAUAAUUAUAGUUUAACUGGGUUUUAG
Macromia             CUAUG-UUAGUAAUUUAUUUAAUGUGGAAUAAUUAUUGAU-AAUACAU-CAUAG
Macromia             CUAUG-UUAG-AAUUUAUUUAAUGUGGAAUAAUUAUUAAU-AAUACAU-CAUAG
Calaphaea            CUGAUUUGUUUGAUUUGGUUAAUGUGUUAUAAUUAUCUUA-AAUAC-UCAUCUG
```

Figure 7.4. *(continued)*

(**CAUAG**) by inserting gaps at the arrow, the tree length increases. Whether this increase in tree length is justified is dependent on what you are trying to minimize in your algorithm: change among nucleotides or change among structures. We argue that in a structural molecule such as rRNA, secondary structure is more conserved than primary structure (nucleotides) (as we suggested above that you could prove to yourself). It is therefore unambiguous to favor the structurally aligned panel (Figure 7.4B) over the panel that was optimized to minimize change among nucleotides (Figure 7.4A). Similar conclusions have been reached even through the structural and phylogenetic analysis of internal transcribed spacer regions that interrupt subunits of rRNA (Denduangboripant and Cronk 2001; Hung et al. 1999; Hypša et al. 2005; Morgan and Blair 1998).

Gillespie, Yoder, et al. (2005, their Fig. 9A) illustrated a similar empirical example of how automated alignment failed to align nucleotides based on secondary structure in one of the most difficult-to-align regions of arthropod nuclear large subunit (LSU or 28S) rRNA. Gillespie, McKenna, et al. (2005) demonstrated the importance of structural alignments in proofreading the data. Hallmark features of rRNA, which must be present in functional rDNA genes, can be utilized as a means of checking the accuracy of generated sequences in a fashion that is no different than using translated amino acid sequences to validate the correct reading frame within protein-coding genes (Gillespie, McKenna, et al. 2005).

CHALLENGES TO EXISTING PROGRAMS

Compositional Bias Presents a Severe Challenge

One of the appealing things about DNA data is that all of the character states are discrete. With morphological characters, it often seems that as you continue to study more representatives of a taxon, your formerly “good” or “discrete” characters dissolve into a grade of continuous variation. Nucleotide characters are what they are, without intermediates. Even though this property of four discrete character states, evolving under a common mechanism, enhances the justification for models and algorithmic alignments, it also presents some new problems with homoplasy AUQ2 due to limited character-state space (Brooks and McLennan 1994; AUQ3 Lanyon 1988; Mishler et al. 1988). If a nucleotide is free to flicker back AUQ4 and forth among these four states, and if there is some nonrandom bias in the data among independent lineages, then there is the possibility for systematic error in our hypotheses of phylogeny. If life on some other planet had five nucleotides instead of four, then this problem would not be as serious as it is here on earth. If we had only two nucleotide states, this problem would be much worse. Unfortunately, there are biological systems in which there are effectively only two states. For example, arthropod mitochondrial genomes are notoriously AT rich, but this bias ranges from 65.6% in *Reticulitermes* (Isoptera) to 86.7% in *Melipona* (Hymenoptera) (Cameron and Whiting 2007). It is easy to predict that with A’s and T’s constituting nearly 87% of the genome, a particular site that can be an A or T (such as silent third and first codon sites), will be. So, taxa that have independently evolved similar compositional biases may be drawn together by rapidly evolving, meaningless sites, and this convergence is more likely with two states than it is with four (Meyer 1994). Simmons et al. (2004) discuss at length the problem of limited character-state space.

Nucleotide compositional bias is particularly problematic in the hypervariable regions of rRNA. The conserved core (the length-invariant, alignment trivial regions) may possess the four nucleotides in nearly equal proportions, whereas the hypervariable regions (which contain many if not most of the parsimony informative characters, and

```
Bug1    ATCGCTCTAGTATCGCGCTAAAAATAGAACTCGCTA
        ||||||||||||        ||||    |
Bug2    ATCGCTCTAGTAATCGCGCTAAAAATAGAACTCGCT
                    ^
```

Figure 7.5. A hypothetical alignment from Kjer et al. (2007), showing that if gap costs are too high, Needleman–Wunsch algorithms may favor phenetic solutions in regions of nucleotide compositional bias.

wherein different alignment methods produce different hypotheses) can possess extreme nucleotide compositional bias. This bias can vary a great deal among taxa. For example, analyzing the structural properties of nuclear 18S rRNA across the major lineages of insects, Gillespie, McKenna, et al. (2005) demonstrated that base compositional bias within nearly all variable regions was severe, and that the patterns of these biases were inconsistent with phylogenetic expectations. Interestingly, in instances where pairwise comparisons of base composition were not significantly different, length heterogeneity was significantly different. This suggests that variable sequence length alone is not the only problem encountered in the alignment of rRNA sequences. Base compositional bias is another confounding factor.

Homoplasy presents problems not only in phylogenetic analysis but also in the assessment of homology in alignment programs. Figure 7.5 (from Kjer et al. 2007) shows a pairwise alignment of a hypothetical region in which the top and bottom sequences are identical to one another, except for a single indel, indicated in bold. Computer alignment programs using Needleman–Wunsch (1970) algorithms function by penalizing change through setting up a ratio of costs in inserting gaps, relative to the cost of a substitution (the gap cost-to-change ratio, or "gap cost" for short). If the gap cost used in the alignment is excessively high, the program will not insert a gap in the top sequence where it "belongs," as indicated by the arrow in Figure 7.5. Rather, the algorithm will continue racking up the relatively low mismatch penalties until it reaches a region of biased nucleotide composition, where the program happily lines up A's together, despite their being offset by one base. It is important to note that even in random sequences we would expect every fourth site to "match." In regions of nucleotide compositional bias, the expectations of nonhomologous but identical states is much greater than every fourth site, and approaches 50%.

Biased nucleotide regions are reminiscent of the *bland uniformity* that frustrates morphologists. But by throwing everything into a computer without looking at it, you would miss the fact that, under conditions of high compositional bias combined with rapid evolution and length variation, Needleman–Wunsch algorithms (and their subsequent derivations) can imitate phenetics, wherein taxa are grouped together according to the overall percentages of A's and T's, rather than by synapomorphies. The effects of compositional bias can be amplified if nonhomologous nucleotides are first aligned together with parsimony (with Needleman–Wunsch, minimizing nucleotide change), and then subject to long-branch attraction under a parsimony search. This is why we reject the assertion that alignments and analyses should logically be conducted under the same optimality criterion (Phillips et al. 2000). We believe that the goals of each endeavor (alignment and analysis), while not independent of one another, are different enough to require a different approach, with each step favoring the best option. Simmons (2004) provides a detailed discussion on the separation of homology and analysis. The regions of rRNA that most commonly accumulate extreme compositional bias are the same regions that are most length heterogeneous, and hard to align. Compositional bias presents a severe challenge to Needleman–Wunsch-based alignment algorithms.

Compositional bias is particularly problematic for the direct optimization program POY (Gladstein and Wheeler 1999) because it depends on accurate reconstruction of ancestral sequences. Collins et al. (1994) showed that under conditions of nucleotide compositional bias, or accelerated substitution rates, parsimony severely underrepresents the rare states in ancestral reconstructions. The Collins et al. (1994) study employed a series of empirical and simulation studies to show this, and the mathematical proof by Eyre-Walker (1998) confirmed their findings. Reconstructing ancestral nodes is what POY *does*, and these studies indicate that results from a POY analysis should be interpreted with caution, and with an understanding of these limitations under conditions that are characteristic of the hard-to-align regions of rRNA.

Gaps Are Not Uniformly Distributed

Not only are substitution rates elevated in the hypervariable regions of rRNA, but also these regions accumulate insertions and deletions at a much more rapid pace than does the "conserved core" (Clark et al. 1984; Hadjiolov et al. 1984; Hogan et al. 1984; Michot et al. 1984).

Anyone who has ever attempted to align rRNA data soon recognizes that gaps are clustered in regions. Kjer et al. (2007) measured the clustering of gaps by simply converting all of the nucleotides in a mammalian rRNA dataset into A's, and all of the gaps into C's. The among-site rate variation of indels from our so-altered NEXUS file was measured on the expected tree using PAUP version 4.0b10 (1998) by estimating the shape of the gamma distribution (Yang 1994). The gamma distribution is best indicated by a value called "alpha," whose values below 1 indicate serious among-site rate variation. The alpha value was 0.45, confirming that variation among sites with respect to the frequency of insertions and deletions is indeed highly variable. This clustering of indels has several important ramifications with respect to alignment. Most importantly, it means that ideal gap costs should vary among sites (Kjer 1995). Typically, computer alignments are performed with fixed gap costs. If biological gap costs vary among sites, then all analyses using fixed gap costs will underrepresent appropriate gap costs at some sites, and overrepresent gap costs at others. The "ideal" average gap cost, even if it were algorithmically and objectively defined, would be inappropriate for most sites. Kjer et al. (2007) demonstrate this with a figure, reproduced here as Figure 7.6. In Figure 7.6, structure is indicated with Kjer et al.'s (1994) notation on the nucleotides, and Gillespie's (2004) structural mask above them. The top panel contains a commonly sequenced region of mitochondrial 12S rRNA from a series of murine rodents. The lower panel contains sequences from a considerably more diverse group: a whale, an ape, an ostrich, a lizard, and a snail. Variation in length among the rodents indicates that the gap cost in those regions should be relatively low, just as invariant lengths among the different phyla should indicate the need for a high gap cost. In this region, we can see that the loop portion of stem 42 (variable region V7) should have a low gap cost, allowing for the easy introduction of gaps. Directly downstream from this hypervariable region is a region of extremely high conservation in length. Apparently, indels in strand 38' are not permitted. Even if we cannot quantify gap costs (which we cannot do because they are arbitrary, and they are arbitrary because we do not have realistic models for indels), you can scan across Figure 7.6 and apply flexible gap costs. For example, the loop between the strands of stem 40 should receive a low gap cost, and the loop between the strands of stem 42, an even lower gap cost. Contrast those low gap costs to the near infinite gap cost within strand 38' and all the undefined gap costs in between for other sites. If you wish to check your own 12S rRNA data for

```
                      H1068   H1074       H1074'          H1068'        H1113            H1113'             H1047'
                        39     40          40'               39'            42       V7        42'                 38'
                      ((..((   (((         )))              ))..))  .....  (((                )))  ....  )))).)))...))))))))
Pachyuromys duprasi   |GCCAAC| (CCT [AA-GA] AGG) [AATAAAAA] |GTAAGC| GAGAT (AAT [AAA--TTAAC--] ATT) AAAA |CGTTAGGTCAAGGTGTAGC|
Tatera afra           |GTAAAC| (CCT [AA-AA] AGG) [-ATTCAAA] |GTAAAC| AAAAG (AAT [CA-----AACA-] ATG) AAGA |CGTTAGGTCAAGGTGTAGC|
Tatera leucogaster    |GTAAAC| (CCT [AA-AA] AGG) [-ATTCAAA] |GTAAAC| AAAAG (AAT [CA-----AAC--] ATG) AAGA |CGTTAGGTCAAGGTGTAGC|
Tatera indica         |GTAAAC| (CCT [AA-AA] AGG) [ACGGTAAA] |GTGAGC| AAAAT (AAT [CA-----AAC--] ATG) AAGA |CGTTAGGTCAAGGTGTAGC|
Mus musculus          |GCAAAC| (CCT [AA-AA] AGG) [TATT-AAA] |GTAAGC| AAAAG (AAT [CA-----AAC--] ATA) AAAA |CGTTAGGTCAAGGTGTAGC|
Mus musculoides       |GCAAAC| (CCT [AA-AA] AGG) [-AGGAACA] |GTAAGC| ACAAG (AAT [AT-------CC-] ATA) AAAA |CGTTAGGTCAAGGTGTAGC|
Myospalax sp.         |GCAAAC| (CTT [AA-AA] AAG) [-AACAAAA] |GTAAGC| AAGAT (CAT [C--------CC-] ATA) AAAA |CGTTAGGTCAAGGTGTAGC|

Balaenoptera physalus |GCAAAC| (CCT [AA--A] GGG) [-AGCAAAA] |GTAAGC| ATAAC (CAT [CC-----TAC--] ATA) AAAA |CGTTAGGTCAAGGTGTAAC|
Pan troglodytes       |GCAAAC| (CCT [GATGA] AGG) [-TTACAAA] |GTAAGC| ACAAG (TAC [CC------AC--] GTA) AAGA |CGTTAGGTCAAGGTGTAGC|
Struthio camelus      |GCCCGC| (CTC [AT-GA] GAG) [----AATA] |GCGAGC| ACAAT (AGC [CC------ACCC] GCT) AACA |AGACAGGTCAAGGTATAGC|
Squalus acanthias     |GCTCAC| (CCT [GT-GA] AGG) [-ATAAGAA] |GTAAGC| AAAAA (GAA [CT-----AAC--] TCC) CATA |CGTCAGGTCGAGGTGTAGC|
Cellana tramoserica   |GTTAAC| (CTT [AT--A] AAG) [-AAAAAAA] |GTTAAC| AATAA (AGA [ATATAAAAACTT] TCT) CATA |GGTCAGATCAAGGTGCAGC|
```

Figure 7.6. A structural alignment of a 12S rRNA fragment from murine rodents (top panel) and other diverse taxa (bottom panel) from Kjer et al. (2007). Structural notation follows Kjer et al. (1994), with underlines indicating hydrogen bonds, parentheses indicating hairpin-stem loops, and vertical lines delimiting longer-range interactions. Unaligned regions are placed in brackets (and ignored in NEXUS format). Notation above the sequences follows the structural mask of Gillespie (2004). This figure shows the contrast and close proximity of relatively length-heterogeneous regions to length-invariant regions.

alignment errors, you will probably find them by lining up stem 39. Using fixed integers as gap costs and applying them across a molecule that is demonstrably length-heterogeneous as a result of regional-specific clustering of indels is a commonly used, but biologically unrealistic, approach to the "alignment problem."

Hence, the practice of exploring parameter space with *sensitivity analyses*, that is, the testing different gap costs, in an effort to select among an infinite pool of gap costs and other parameters is problematic (but supporters of sensitivity analyses would be correct in noting that this is no more problematic than performing no such tests, if you are tied to fixed gap costs). Sampling around a series of parameters, as a means of parameter selection, implies that some parameters are "good," whereas others are "bad." This is a futile endeavor. There is no ideal, single, fixed gap cost for an alignment such as this because we are dealing with a heterogeneous assortment of regions. Sensitivity analyses require that at least some of the analyses are appropriate. But when we look at rRNA sequence data, where the "gappy" regions are clustered, we can see that one-gap costs will work well for one region, and poorly for another. By changing the gap cost, other regions may be well aligned, whereas the regions previously well aligned become worse. Different gap costs may shift the appropriately aligned regions from one region to another without necessarily expanding the proportion of well-aligned sites. Of course, if homology is completely ambiguous and unknowable, one may find it useful to present alternative alignments in assessing alignment uncertainty. However, we find it unreasonable to assume that history happened in multiple ways when structural homology favors a single solution.

One proposal for selecting among parameters is to perform a sensitivity analysis on a variety of parameters, and measure each resultant tree against some external criterion: a tree based on morphological characters, for example, or minimizing ILD scores (Farris et al. 1994) AUQ5 between partitions. However, there is an infinite number of parameters to explore. Wheeler (2005) discussed the problem, and also how this infinite space might be realistically explored. Wheeler (2005) explored gap costs and transversion weights (as did Terry and Whiting 2005; Whiting et al. 1997; and others). These explorations result in a three-dimensional plot of the parameter landscape. What you would want in such a landscape is a single hump containing a distinct peak, because with such a simple distribution, if you are anywhere near the peak, then you can be assured that the combination of gap cost and some other

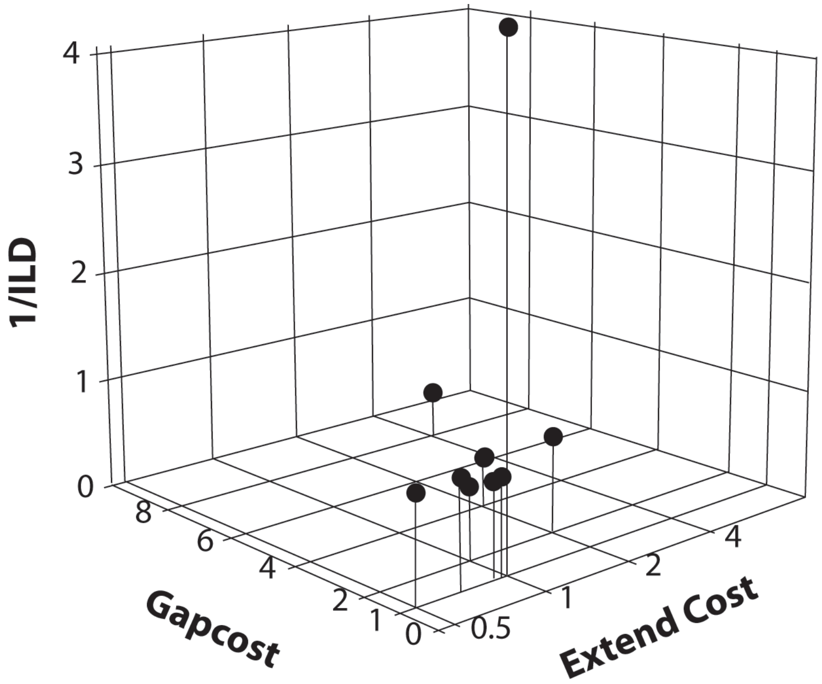


Figure 7.7. Sensitivity analysis, simultaneously exploring the cost of inserting a gap (gap cost), and the cost of extending an existing gap (extend cost), and how combinations of these costs influence ILD scores. The higher the inverse of the ILD scores, the less disagreement there is between the partitioned analyses (in this case, 12S vs. 16S topology). This figure is taken from Kjer et al. (2007).

metric (such as transversion weights) is near optimal. This is not the shape of the peak found by Kjer et al. (2007), who used an ILD test to maximize congruence between 12S rRNA and 16S rRNA datasets in optimizing gap costs with gap extension costs (the cost of inserting additional gaps, once an initial gap has been inserted). Figure 7.7 shows the shape of this "peak." The problem here is that there was a relatively flat plane of inverse ILD scores, and a single sharp spike, where gap costs = extendcosts = 1. Unlike the "single hump distribution," this kind of distribution does not instill confidence that the best gap cost, relative to the indel extension cost, has been discovered, because an even higher spike may exist among the infinite combinations of parameters that were not explored. Other studies have also found it difficult to select parameters with sensitivity analyses (Terry and Whiting 2005; Wheeler 2005), although the generality of this problem has not been explored. One thing is certain, though; it is arbitrary to perform sensitivity analyses that compare only two analytical parameters (such as gap cost vs. transversion weights, or gap costs vs. extend costs) when there are a multitude of interacting parameters that simultaneously influence phylogenetic hypotheses. Many of these parameters may not be best treated with integers, or with fixed values, and many of them, such as the gap cost, are arbitrary (Doyle and Davis 1998; Hickson et al. 2000; Kjer 1995; Phillips et al. 2000; Vingron and Waterman 1994; Wheeler 1996).

Nonindependence of Indels

One of the reasons that gap costs are arbitrary is that we really do not have a reasonable model for insertions and deletions. One of the most unreasonable assumptions behind many of the existing algorithms is that multiple adjacent indel positions are all independent of one another, when they may have resulted from a single event. Simmons and Ochoterena (2000) discuss at length the problems with nonindependence of gaps and the problems with treating individual gaps as 5th state characters. They convincingly argue that contiguous gap positions are most parsimoniously interpreted as the result of a single event, and propose a system for coding them. Figure 7.8 offers an example of how we do not know the history of events that led to the present condition. However, we do know a number of things with certainty; first, this is a region of compositional bias (82% AT), and second, this is a region in which gaps of multiple lengths have accumulated among these and

```
                         ((((                                     ))))
Ptilocolepus        GUCAUUGAG[AAC-----------------------CGA-UAAA]CUCAGAGGC
Palaeagapetus       GUCAUUGGG[AAU------------------------CACUAAA]CCCAGAGGC
Anchitrichia        GUCAUUGGG[AAUUUUUCAAACAUA--CAAU---CAUAACUAAA]CCCAUAGGC
Brysopteryx         GUCAUUGGG[AAUAUAUGGAUAAUAAACAAUGAAUCUAACAAAA]CCCAUAGGC
Matrioptila         GUCACUGGG[AG------------------------CGAUUAAA]CCCACGGGC
Rhyacophila fuscula GUCAUUGGG[AUUUUUUUU-----------------ACACUAAA]CCCAGAGGC
Rhyacophila brunnea GUCAUUGGG[AUUUUUUU------------------ACACUAAA]CCCAGAGGC
```

Figure 7.8. A structurally aligned region of caddisfly rRNA, with ambiguously aligned nucleotides in brackets. Underlines indicate hydrogen-bonded nucleotides, as do the parentheses above the sequences. Note the compositional bias, and the extreme length variation in the unaligned region.

many other taxa. We also know that this same region is hypervariable and hard to align in a wide range of taxa. In an example taken from caddisfly (Insecta: Trichoptera) rRNA, we can see that the sequences from *Anchitrichia* and *Brysopteryx* (Hydroptilidae, Hydroptilinae) are much longer than those from *Paleagepetus* and *Ptilocolepus* (Hydroptilidae, Ptilocolepinae). If we treat each of the gaps as 5th state characters, that is, as if they were independent of one another, then that would mean that ptilocolepines share 24 independent deletions, relative to *Matrioptila*. Such a treatment of the data is nonparsimonious. It would be more parsimonious to assume that there was a large insertion of multiple nucleotides in Hydroptilinae, followed by subsequent modifications. Although the alignment of this region is ambiguous, it gives strong hints about relationships. The nucleotides between the brackets give strong support to the monophyly of *Rhyacophila*, and the large insert present in the Hydroptilines is likely homologous, since it is long and complex, and there are conserved motifs that indicate a common origin. Yet, for one to infer dozens of synapomorphies from this rapidly changing, unalignable mess, one would have to disregard common sense and assume that all characters are equally informative and independent of one another. Some studies find that 5th state coding for deletions outperform methods that treat deletions as missing data (Ogden and Rosenberg 2007b). While we agree that gaps provide important signals (Freudenstein and Chase 2001), treating them as 5th states falls short to the probability that, while many long inserts *do* indicate phylogenetic relatedness, treating all the gaps as independent characters inflates the support for these nodes (Simmons and Ochoterena 2000), whether they are due to common history, convergence, or alignment artifacts.

Long Inserts/Deletions

Whereas simultaneous deletions of five or six nucleotides at a time, occurring in independent lineages, may draw these unrelated taxa together if they are considered to be five or six independent events, much larger indels, sometimes hundreds of nucleotides long, are known to occur (e.g., Giribet and Wheeler 2001 provided a list of atypically long 18S rRNA sequences of metazoans). These long insertions have the potential to wreak havoc on a computer alignment (Benavides et al. 2007). The epitome of this problem is perhaps the bizarre insertions that interrupt both the 28S (Gillespie, unpublished) and 18S (Gillespie, McKenna, et al. 2005) rRNA sequences of the strepsipterans. Not only do inserts occur in the variable regions and expansion segments of the rRNA of these odd insects, but extraordinarily (up to 366 nts.) AUQ6 long insertions are known to occur within the hairpin-stem loop of the highly conserved pseudoknot 13/14 in the V4 region of 18S rRNA (Gillespie, Mcenna, et al. 2005). Thus, despite the earlier statement that conserved regions are less tolerant of indels, the Strepsiptera data suggest that introns can occur in conserved regions. In fact, virtually all of the introns that interrupt rRNAs occur in the most conserved regions of the tertiary structure (Jackson et al. 2002; Wuyts et al. 2001), particularly at the subunit interface or in conserved sites with known tRNA–rRNA interaction (Jackson et al. 2002). While introns are relatively rare in rRNA sequences, only a manual evaluation using a structural model would detect their presence. Still, the majority of large inserts in rRNAs are likely part of the mature molecules and are localized to the surface of the ribosome. However, structural models for the expansion segments and variable regions exposed to the surface of the ribosome are becoming more and more refined with the addition of new taxa and sequences and through refinements in the ribosome crystal structures (e.g., Alkemar and Nygård 2003; Alkemar and Nygård 2004; Buckley et al. 2000; Gillespie et al. 2004; Gillespie et al. 2006; Gillespie, McKenna, et al. 2005; Gillespie, Munro, et al. 2005; Gillespie, Yoder, et al. 2005; Hickson et al. 1996; Kjer 1997, 2004; Mears et al. 2006; Misof and Fleck 2003; Ouvrard et al. 2000; Page 2000; Schnare et al. 1996; Wuyts et al. 2000). Thus, conserved structures within even variable regions and expansion segments will be necessary to guide the assignment of nucleotide homology when high levels of length heterogeneity exist across alignments.

Lack of Recognition of Covarying Sites (A Well-Known, Seldom-Adopted Strategy)

Wheeler and Honeycutt (1988) identified a directed substitution rate within helices of the 5S rRNA of animals and plants that deviates from the neutral theory of molecular evolution explaining rRNA evolution (Kimura 1983; Ohta 1973). This slightly deleterious mode of sequence evolution in rRNA, in which noncanonical base pairings, or bulges, are replaced by compensatory base changes or reversals to the original state, has been identified in subsequent studies (e.g., Douzery and Catzeflis 1995; Gatesy et al. 1994; Kraus et al. 1992; Rousset et al. 1991; Springer and Douzery 1996; Springer et al. 1995; Vawter and

Brown 1993), and appears to be the mechanism orchestrating secondary structural conservation in rRNAs. Paramount to the findings of Wheeler and Honeycutt (1988) was not only the identification of two different selective constraints within the same molecule (pairing versus nonpairing regions), but also the realization that *nucleotides within pairing regions in rRNA datasets are not independent characters, because a change at one site influences the probability of a substitution at another site*. This poses an added difficulty when treating helices in phylogenetic analysis, as opposed to unpaired nucleotides, wherein interdependence with other positions is not easily demonstrated (although the variety of tertiary and stacking interactions could in theory be modeled). Regarding parsimony, analysis of pairing (stems) or nonpairing (loops) regions has been suggested, but not both in simultaneous analysis (Wheeler and Honeycutt 1988). Some workers have implemented a stem-loop weighting approach to accommodate the nonindependence of pairing regions (Dixon and Hillis 1993; Smith 1989; Wheeler and Honeycutt 1988). Although seemingly intuitive, down-weighting stems on the basis of their nonindependence will also relatively up-weight positions that are hypervariable, and often nonpairing and perhaps misaligned, thus inaccurately representing the information contained within pairing regions. Up-weighting compensatory mutations within pairing regions has justification (Ouvrard et al. 2000), particularly if rare substitutions define major clades; however, discerning which characters to weight within an alignment can be puzzling if the ancestral pairing cannot be immediately identified (i.e., before analysis). Simon (1991) warns against stem-loop weighting, and van de Peer et al. (1993) illustrate that stems and loops are both highly heterogeneous in terms of substitution rates. These added difficulties, coupled with the fact that assumptions of certain branch support measures such as the bootstrap (Felsenstein 1985) and Bremer support indices (Bremer 1988; Donoghue et al. 1992) are violated by the nonindependence of rRNA pairing regions, suggests that a parsimony approach to analyzing rRNA alignments may not adequately accommodate these data. There may be interacting operational vs. philosophical factors involved; if compensatory changes in paired stem sites are also relatively more conservative, parsimony may appropriately (although inadvertently) up-weight the slower-evolving characters (Kjer 2004). Similarly, standard likelihood models of DNA substitution, which are all based on a 4 × 4 rate matrix, are also insufficient for phylogeny estimation using rRNA, because of their failure to account for correlated bases forming helices.

Studies modeling the evolution of pairing regions in rRNA molecules have grown in the last decade. Although most have focused on modeling base pair evolution under likelihood, similar approaches are under development for parsimony (Yoder 2007, Yoder and Gillespie, unpublished). These studies have all centered on implementing a substitution matrix that accommodates the nonindependence of helical regions. Unlike the typical 4 × 4 substitution matrix used for modeling DNA evolution, a matrix modeling rRNA evolution consists of all possible substitutions within a pairing region. Hence, a 16 × 16 matrix is used to model pairing regions, with the most general time reversible (GTR) model allowing for 134 free parameters. A detailed explanation of the simplified families of RNA substitution models was recently provided by Gillespie (2005). Given the attention being addressed to modeling RNA evolution, two software packages have incorporated some of the above-mentioned models into their programs. MrBayes version 3.1 (and earlier versions) (Ronquist and Huelsenbeck 2003) includes model 16B (Schöniger and von Haeseler 1994) and allows for helices to be modeled independently as pairs along with other models for nonpaired sites (i.e., loops, codons, amino acids). Importantly, model 16B should be considered an F81-like model for pairing sites, and when the covarion model in MrBayes is set to REV or HKY85, model 16B becomes different for each case (Jow et al. 2005). The program PHASE version 1.1 (Jow et al. 2002) also provides a means to simultaneously model multiple partitions with different models of evolution. In addition, PHASE contains a suite of RNA models that allow for the evaluation of the performance of different RNA models on a given dataset. Most likely as a result of the study of Savill et al. (2001), those models that allow for base pair asymmetry and a nonzero rate of double substitutions, namely, models 16A, 7A, 7D, 6A, and 6B, are all included in the PHASE program. Thus, PHASE has an appeal over MrBayes 3.1 in that the user can determine the best model of evolution for an RNA dataset, rather than settle for only one RNA model (perhaps with slight modifications). The soon-to-be-released MrBayes 4.0 will contain additional rRNA models.

Doublet models are thus directly related to the alignment issue, because alignments performed within a structural context provide a template that allows for a more realistic modeling of the evolution of these complex biological molecules. Intuitively, they are more desirable to the evolutionary biologist. However, it is not our intention here to criticize the algorithmic approach to alignment just because more biologically sound methods exist. On the contrary, we fully support and prefer algorithmic methods, as long as the algorithm that is applied has some grounding in biological reality. Current methods that ignore the properties of rRNA are not biologically grounded. We hope that in pointing out the challenges faced by current methods, we will accelerate the implementation of algorithmic methods, and eventually eliminate the difficult, tedious, and nonrepeatable manual alignments. Before this can happen, however, if you favor phylogenetic hypotheses that are meaningful and predictive, then manual approaches should not be eliminated until algorithmic methods can be shown to outperform them. Replacement should not occur just because a new method remedies some of the problems and is "cool," new, and computationally expensive.

ARE STRUCTURAL INFERENCES JUSTIFIED?

One of the criticisms of rRNA structural inferences is that they are inferences, not direct observations. Despite great efforts in cryo-electron microscopy (e.g., Frank and Agrawal 2000; Frank and Agrawal 2001; Frank et al. 2000), complete ribosomal RNA secondary structures can be directly observed only through x-ray crystallography. While several atomic structures of ribosomal subunits now exist for yeast (Spahn et al. 2001), the archaean *Haloarcula marismortui* (Ban et al. 2000), and the bacteria *Thermus thermophilus* (Brodersen et al. 2002; Schluenzen et al. 2000; Wimberly et al. 2000; Yusupov et al. 2001) and *Deinococcus radiodurans* (Harms et al. 2001), most rRNA secondary structures are inferred through comparative evidence. Structural alignments identify with very high accuracy (>90%, Gutell et al. 2002) those regions involved in base pairing. Comparative evidence works under the assumption that if multiple sequences can fold into the same conserved structure, and if there is a substitution in one part of the putative stem, it is usually followed by its complementary partner. Inferential, yes, but what are the odds that structures are not real, and are you willing to take that chance? The odds that structurally superimposable structures could arise by chance are easily calculable. For example, if there is an A at one site, what is the probability that there will be a T at the position of its putative partner? Answer: 0.25 according to Jukes and Cantor (1969). So if you align two taxa together, and find 35 compensatory mutations between them, the probability of this happening by chance is 0.25^{35} (that is, a zero, followed by a decimal point, 21 zeros, and then an 8). Adding the thousands of observed compensatory substitutions among all taxa, one arrives at a number so small that the human mind (even among mathematicians who are experienced in thinking about really small numbers) cannot come close to even imagining these numbers in a meaningful way. When one considers how tenuous the whole process of phylogenetic inference is (where we never *know* anything, and the best we can do is come up with a reasonable *guess*, where our data are consistent with our hypotheses, given our assumptions), it seems absurd to argue over whether it is safe to assume whether structural constraints that have a virtually zero (but not *technically* zero) probability of being random should be abandoned on philosophical (or other) grounds. We also note that *translated* amino acids are routinely used to check DNA alignments of protein-coding sequences, even though few (if any) of these studies bother to experimentally demonstrate that the genetic code for the taxa of interest is the same as the model taxa. AUQ7

WHY ALIGN MANUALLY?

As we were considering our observations about the objectives of phylogenetic alignment, and beginning to write them down for this paper, Morrison (2006) presented a review of procedures and philosophies. This excellent review thoroughly explores the differences among us, and, in fact, much of what we had thought to be intuitive but unproven could now be explained in a series of logical arguments. Morrison (2006) lays out a series of problems with current algorithms that were designed for one purpose, and then used for phylogenetics. He argues that many of the problems we face in alignments stem from a failure to recognize that the program is neither designed nor suited for phylogenetic inference. Whereas we had noticed these problems, we had assumed that some smart person out there must have some reasonable solution to phylogenetic alignment; we just had not read about it yet. Morrison (2006) presents a radical new view, stating, "Our objective should be biological plausibility rather than mathematical optimality." With respect to alignments, we are in complete agreement with this statement. Algorithms that currently align sequences with the goal of reaching a mathematical optimum may fail for phylogenetics if they do not simulate biological reality.

Perceived Advantages of Algorithms

Much about alignment has simply been assumed, without question. One's preferences, alluded to in the introduction, seem more a matter of culture and tradition than experimentally justified or even thoughtfully considered criteria. It seems intuitively obvious that computers are more objective at making alignment decisions than manual alignments. Are they? No, not if the computer requires arbitrary input parameters. The following comparison should be made. Consider a thoughtful systematist, thinking about homology under a series of structural and evolutionary constraints. Contrast this to another reasonable systematist, who believes that homology is best decided objectively with a repeatable optimality criterion implemented by a computer. The former may fail through carelessness. The latter may fail when the computer program is actually an irrational black box. Input parameters, such as gap costs, assigned by the investigator determine phylogenetic hypotheses. If these input parameters are arbitrary, then justifying algorithmic approaches over manual ones under a criterion of "objectivity" is almost impossible to argue. One must justify each of the parameters that influence the analysis. Yet the argument continues. We believe that if input parameters are arbitrary and unpredictable, then alignment methods that use them are also arbitrary and unpredictable. To submit one's data to an algorithm, with no regard for the implications of such an action, is to transfer subjective (and thoughtful) decisions about homology from the human investigator to subjective (and careless) decisions about gap cost determination. Algorithmic methods are not objective if input parameters are subjectively determined.

Another perceived advantage of algorithmic methods is that they are easier than structural alignments. In our experience, many investigators accept that structural alignments make sense, but they do not make the effort to perform them because they assume that their Clustal alignment is "good enough" and that a few alignment errors generated by the algorithm will be overridden by the mass of signal in their data. We find this cavalier attitude toward homology to be surprising, when we consider the effort and expense that goes into collecting the sequences. In our opinion, it is always worth the effort to align the data with care. As systematists, we are often are more interested in resolving controversial nodes and not so interested in re-corroborating well-established relationships. Controversial internodes are often characteristically short, and may be difficult to recover by any means with a variety of datasets. It may be that the characters we discard, because the easy method is applied, are the only ones that are informative. Or more likely, the few characters that inform us about a short internode are overwhelmed by a mass of poorly aligned noise. We could never know how a careful alignment would influence our results without the effort. Those who support sensitivity analyses to optimize parameters with POY would probably agree with us on this point, as they perform months of analysis time on parallel processors or super clusters. Careful alignment, whether performed by hand or computer, takes time, effort, and expertise. We reject the argument that carefully performed algorithmic methods are "easier," and let the reader decide whether "fast and careless" alignments are defendable.

An Example of Accuracy and Repeatability

If algorithmic methods could be shown to be more accurate than manual alignments, then we might be able to overlook the possibility that arbitrary parameter selection may sometimes lead to unpredictable hypotheses. This is not the case, however. Many empirical comparisons have shown that manual alignments tend to recover more reasonable phylogenies (Ellis and Morrison 1995; Gillespie, McKenna, et al. 2005; Hickson et al. 1996; Hickson et al. 2000; Kjer 1995; Kjer 2004; Lutzoni et al. 2000; Morrison and Ellis 1997; Mugridge et al. 2000; Schnare et al. 1996; Titus and Frost 1996; Xia et al. 2003). Phylogenies are hypotheses to be tested, accepted, or refuted by subsequent hypotheses. We never "know the truth." Such hypotheses may be accepted on the grounds that they generally equate to the recovery of expected or corroborated relationships with phylogenetic accuracy. A compelling case can be made for phylogenies generated from manually aligned datasets. Time after time, we recover "more reasonable" phylogenetic hypotheses from carefully aligned data, (while at the same time, analyses justified only on epistemological consistency continue to produce "unexpected" hypotheses). Admittedly, these empirical studies can provide only points for discussion. To demonstrate accuracy, we need either known phylogenies from experimentally manipulated systems (such as sampling evolving viruses, Hillis and Bull 1993) or simulation studies where we know the history of insertions and deletions in a simulated dataset. However, there are problems with both of these approaches, and these problems stem from the nature of rRNA. Viruses do not possess rRNA, so problems specific to rRNA alignment cannot be addressed with manipulated viral sequences. Simulation studies are only as good as the model used to simulate the data. Currently, our ability to model insertions and deletions is limited and unrealistic. Although it is possible to insert gaps into a simulated sequence, any model that assumes that gaps are independent of one another and randomly distributed is not capturing the essence of what is happening in rRNA, where insertions and deletions are frequently multiple nucleotides in lengths, and strongly clustered in variable regions. An accurate model of rRNA evolution would require a proportion of the sites to be covarying, gaps to be nonindependent, and substitution rates and length heterogeneity to be regionally variable. Without these characteristic features of rRNA built into the simulation, any generalizations drawn from these studies must be understood to be only crude approximations of biological reality.

We suggest a reasonable empirical solution to the assessment of accuracy in Kjer et al. (2006, 2007). Although accuracy cannot be fully explored with empirical data, we see at least one example where an "expected tree" is justified. For taxa whose entire mitochondrial genomes are sequenced, it can be expected that partitions of the data share the same history. We suggest that relationships that are corroborated with both nuclear genes and morphology are candidates for identifying sets of phylogenetic expectations from the mitochondrial data. If these independently corroborated nodes are also supported by the combined mitochondrial genome data, then these relationships could be used to assess alignment strategies of any partitions of the data, such as the 12S and 16S mitochondrial rRNAs. Kjer et al. (2007) used the relationships shown in Figure 7.9 to compare phylogenetic accuracy and repeatability of manual and direct optimization methods. These taxa possess complete mitochondrial genome data, and each of the nodes is supported by morphological characters (McKenna and Bell 1997; Novacek 1992; Novacek et al. 1988; Simpson 1945) and nuclear genes
AUQ8 (Amrine-Madsen et al. 2003; Delsuc et al. 2002; Waddell and Shelley AUQ9
AUQ10 2003). The phylogeny shown in Figure 7.9 is recovered from parsimony, Bayesian, and Likelihood analyses of the complete mitochondrial genomes (Gibson et al. 2005; Kjer and Honeycutt 2007; Reyes et al. 2004). One need not accept this as the "true tree," but merely a tree that is recovered by the entire dataset, and corroborated by multiple independent sources. By definition, partitions of the same linked dataset contain less data. It is therefore reasonable to use a tree derived from ten times the number of linked nucleotides in order to test alignment accuracy. There is a risk to judging an alignment method according to this sort of phylogenetic expectation. Namely, the risk would be that the "expected tree" was later shown to be inconsistent with a tree derived by some future superior method. As such, the results from

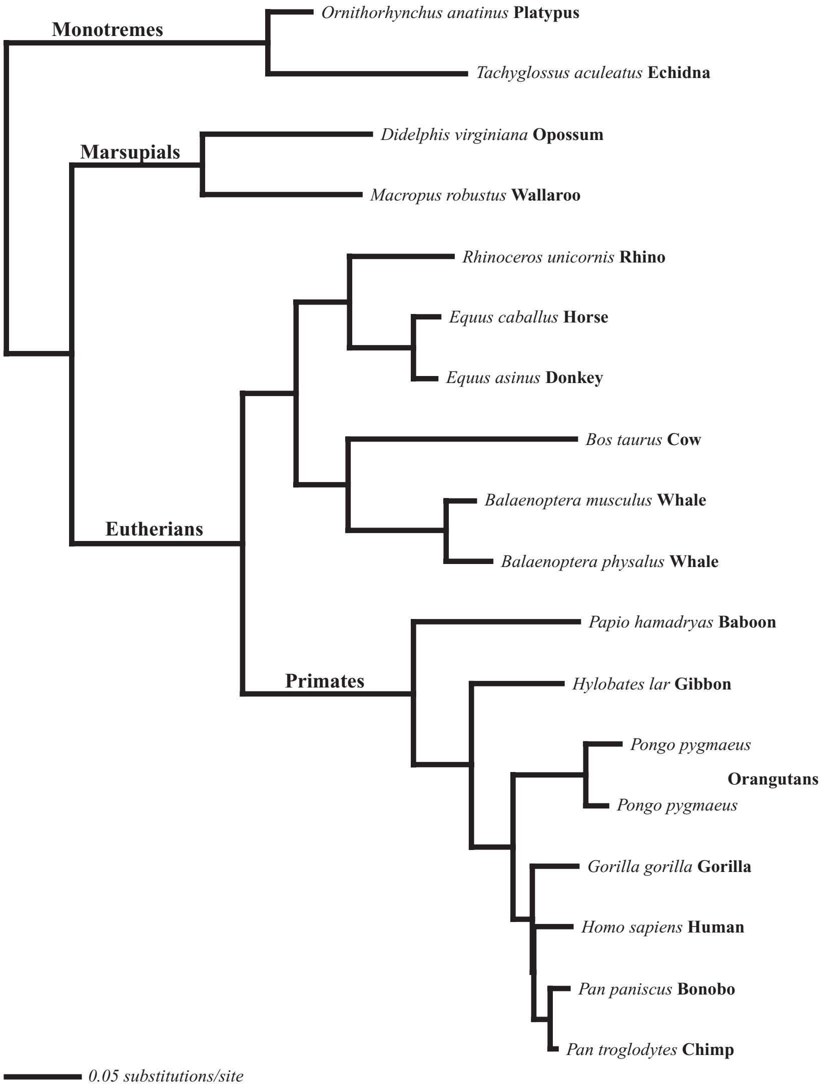
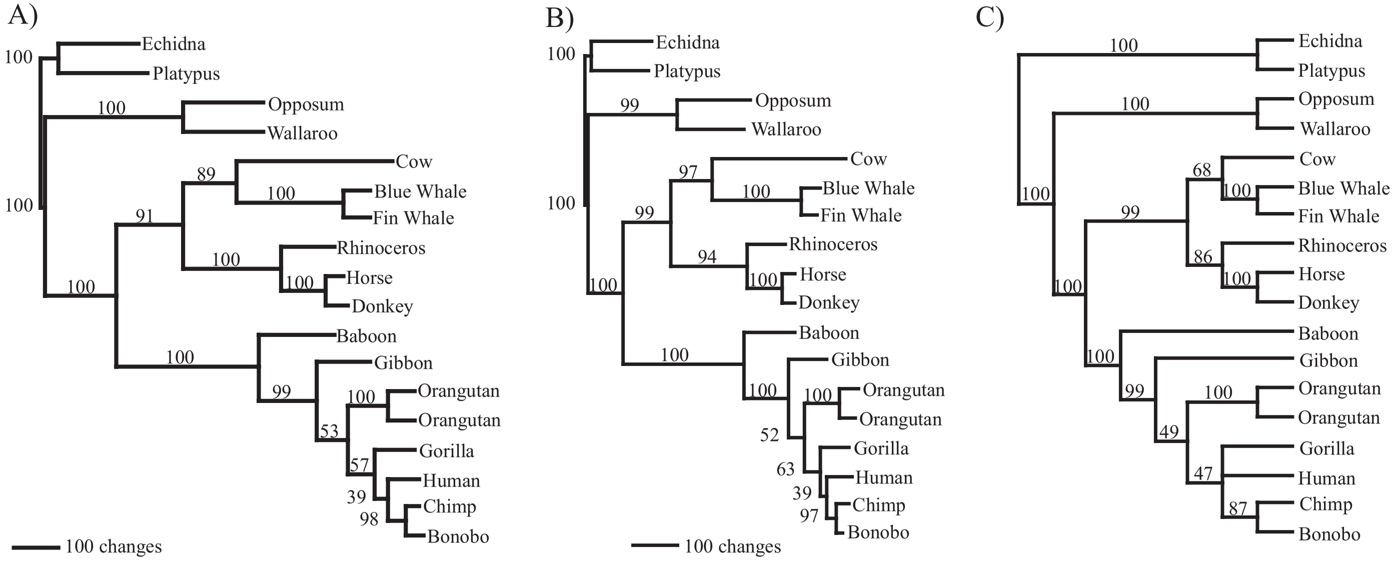

Monotremes
Ornithorhynchus anatinus **Platypus**
Tachyglossus aculeatus **Echidna**
Marsupials
Didelphis virginiana **Opossum**
Macropus robustus **Wallaroo**
Rhinoceros unicornis **Rhino**
Equus caballus **Horse**
Equus asinus **Donkey**
Bos taurus **Cow**
Balaenoptera musculus **Whale**
Eutherians
Balaenoptera physalus **Whale**
Papio hamadryas **Baboon**
Primates
Hylobates lar **Gibbon**
Pongo pygmaeus
Orangutans
Pongo pygmaeus
Gorilla gorilla **Gorilla**
Homo sapiens **Human**
Pan paniscus **Bonobo**
Pan troglodytes **Chimp**
0.05 substitutions/site

Figure 7.9. Expected tree, generated from analysis of the entire mitochondrial genome. Branch lengths calculated as likelihood, with a GTR + I + G model (Kjer et al. 2007).

the experiment could be modified by the new phylogenetic expectations. In other words, if one is transparent about how phylogenetic expectations are used to assess alignment performance, then conclusions can easily be overturned with future illumination. Science is about laying out one's assumptions, and testing hypotheses according to whether or not the data fit those assumptions.

The experimental design of Kjer et al. (2007) was simple. The 16S rRNA sequences from the taxa shown in Figure 7.9 were assembled, the taxon names were disguised, and the taxon order was shuffled within the matrix. The masked data were then sent to three investigators with simple instructions: "Align these data with secondary structure, and also with POY." It was predicted that if secondary structure could provide a reasonable means of homology assessment, then different investigators would come to similar decisions about structurally influenced homology. More simply stated, if the structures were real, we would all find them, and structurally aligned data would lead to similar phylogenetic conclusions among investigators, because they would be using a nonarbitrary means of homology assessment, even if the alignments themselves were not identical. The second prediction was that if parameter decisions were arbitrary (Doyle and Davis 1998; Hickson et al. 2000; Kjer 1995; Phillips et al. 2000; Vingron and Waterman 1994; Wheeler 1996), and these parameters had a strong influence on phylogenetic conclusions, then different investigators using an algorithmic approach to a phylogenetic problem would arrive at different phylogenetic hypotheses, given the same data.

Figure 7.10 shows the results of the experiment. All three of the structural alignments yielded nearly identical results, with the only difference being on the Chimp/Human/Gorilla branch (which is a reasonable reflection of reality, because this branch has no perceivable length). The structurally aligned data also recovered the expected tree. The hypotheses generated from the independent POY analyses (not shown; see Kjer et al. 2007, Fig. 5) resulted in each investigator proposing a different phylogenetic hypothesis, none of which were the expected tree. Each of the POY analyses resulted in different opinions on how to present the confusing array of trees that were generated from the many explorations of alternative parameters. We have already discussed the ambiguity of sensitivity analyses when each of the explorations of parameter space is biologically unrealistic. Alarmingly, even when the same parameters (but not necessarily the same search heuristics) were used, each of the investigators recovered different trees with POY. In all probability, this



AUQ15 Figure 7.10. Trees generated from a manual alignment of the 16S rRNA data. Branch lengths calculated with parsimony. (A), (B), and (C) generated by Kjer, Ober, and Gillespie, respectively. (C) is a consensus of two trees. Numerals near the nodes are bootstrap proportions.

was the result of an insufficient search strategy. All heuristic exercises, including routine tree searches, can suffer from this problem, and if you start from the same random seed, you will get the same tree. However, direct optimization is more complex than simple tree searches, and figuring out how long to run the analysis is another decision that needs to be made. In this example, structurally aligned data resulted in phylogenetically identical hypotheses that conformed to the corroborated tree. Is that not what we *want* in terms of repeatability? Consider the scenario of baking a cake. We follow a recipe: 2 eggs, a cup of flour, a cup of sugar, and so forth, . . . mix well, and then bake in the oven at 375 °F for 30 minutes. Perhaps one person uses 5% more flour than another, or one person stirs with greater vigor. Regardless, the end product is a cake. With manual alignments guided by secondary structure, we all get cake in the end (Figure 7.10). Repeatability in science has always been defined in this way. We describe the methods, and then see if others can repeat it. Taken even further, if the alignment is presented, the analyses can be precisely repeated, and the decisions that went into it can be assessed and changed. Kjer et al. (2007) show that when you follow the cake recipe, and pop the data into POY, you do not know what will come out; it might be a loaf of bread. Of course, our scenario is one of exaggeration, but it serves to make a point. When you are cooking (or applying mathematics), you can tell the difference between the end results of a cake and a loaf of bread. In phylogenetics, the end products cannot be so easily distinguished from one another with respect to which is correct.

Not everyone agrees with the generalizations we reported in Kjer et al. (2007), and as with any work, there are, no doubt, legitimate criticisms that we would like people to consider. This work (Kjer et al. 2007) was presented as an opinion piece to foster some discussions about the ambiguities of alignments, both manual and computer generated. We asked one of our critics, Gonzalo Giribet, to summarize the basic weakness of our work here.

> I think that we both agree that secondary structure information is valuable for refining homology hypotheses but we differ in the way we incorporate such ancillary information into our homology-assignment techniques, being those multiple alignments or simply putative synapomorphies in "direct optimization" techniques (what I have called "single-step phylogenetics"). We have different understanding of what reproducibility may mean, and although I see an alignment as a pure topology-dependant hypothesis you may view it as something that is fundamentally knowable, i.e., that there is "one" alignment. This is what causes that you may search for

> "the" alignment while I am more interested in exploring what alternative parameter sets may have to do with my homology hypothesis, i.e., assessing the stability of my results to alternative parameter sets.

Naturally we do not agree with everything Giribet has to say about sensitivity analysis. There seem to be two purposes for sensitivity analyses; one would be to find the most justified set of parameters, and therefore select a favored hypotheses derived from the preferred input parameters. The other purpose would be to explore the stability of the data to a variety of parameters, and present a phylogenetic hypothesis that includes measures of alignment uncertainty. Giribet supports the second of these in his statement above, and there may have been instances where we have confused these two justifications for sensitivity analyses. We appreciate his effort to clarify this difference, but still find both uses of sensitivity analyses problematic.

We think that phylogenetics is a near impossible enterprise, and the best we can do is to do our best. We should not be ashamed of pursuing accuracy, even if it is impossible to assess. We agree with Giribet that sensitivity to alignment uncertainty should be explored, and that uncertainty in an alignment should eventually be part of our estimates of support. We think that this would probably require a Bayesian method, similar to that proposed by Redelings and Suchard (2005).

If one is interested in a "best estimate" of phylogeny, this estimate will likely come from an analysis in which homology (alignment) is optimized. Another area of ambiguity is amplified when multiple alignment parameters offer many different trees; it becomes difficult to informatively select among them. The utility of a phylogenetic hypothesis is drastically reduced when no hypothesis can be considered better than another, because all of the trees are devoted to explorations of the data under different alignment input parameters. For example, what can we take from Wheeler et al. (2001) as a phylogenetic hypothesis? Surely not the "discussion tree," but if not that, should we favor the myriad of other trees that collapse to a near meaningless polytomy? We believe that it is the responsibility on an investigator to clearly present his or her hypothesis. The best way to do so would be to state "the best estimate of phylogeny we could make comes from analysis 'X,' shown in Fig. 'Y.'" It is the responsibility of the reader to then evaluate the results and either agree or disagree with the findings. For this dynamic process to occur, one must present the data, present the alignment, and justify the decisions that were made. The current feasibility of justifying one's decisions with an algorithmic approach is not as straightforward as is the case with a manual approach.

COMPARISON TO PROTEIN ALIGNMENT— PROGRAMS AND BENCHMARKS

Protein alignment programs have seen much development in recent years, beginning with ClustalW (Thompson et al. 1994) to current state-of-the-art approaches, such as Probalign (Roshan and Livesay 2006), Probcons (Do et al. 2005), and MAFFT (Katoh et al. 2005). These programs use a variety of techniques, such as hidden Markov models (Durbin et al. 1998), maximal expected accuracy (Durbin et al. 1998), fast Fourier transforms (Katoh et al. 2005), profile alignment (Edgar 2004a), and consistency alignment (Do et al. 2005). Most protein alignment programs aim to align parts of proteins conserved in sequence or structure. This is facilitated by amino acid substitution-scoring matrices estimated from real data (such as PAM, Dayhoff and Eck 1968; and BLOSUM, Henikoff and Henikoff 1992) and manually created protein alignment benchmarks, also based on real data (such as HOMSTRAD, Mizuguchi et al. 1998b; and BAliBASE, Thompson et al. 2005b). These benchmarks, which are primarily structure-based alignments, not only allow for comparison of different programs, but also enable optimization of gap penalty parameters on real data. As a result, protein alignment programs have shown a steady increase in accuracy over the years. The most accurate programs use a combination of techniques, such as PSI-BLAST profiles, and predicted secondary structures as found in the PROMALS program (Pei and Grishin 2007).

Most protein sequence alignment programs can be used for RNA alignment in principle. However, substitution scoring matrices and alignment benchmarks (analogous to BLOSUM and BAliBASE, for example) were not developed for RNA until recently. The BRaliBASE RNA alignment benchmark (Gardner et al. 2005), similar to BAliBASE for proteins, is the first RNA alignment benchmark produced by aligning sequence while taking into consideration secondary structure. Subsequent efforts have expanded BRaliBASE (see Wilm et al. 2006). Yet, BRaliBASE still lacks the size and diversity of its protein counterparts. The RIBOSUM scoring matrices (Klein and Eddy 2003) for RNA are comparable to the BLOSUM matrices for proteins; however, recent studies show that they perform more poorly than simpler scoring matrices when used in the ClustalW program on enhanced BRaliBASE benchmarks (Wilm et al. 2006).

Further development of RNA alignment benchmarks, better substitution scoring matrices, and adaptation of techniques used in state-of-the-art protein alignment programs will eventually lead to better algorithms for aligning RNA. Alignment of conserved regions (in sequence or structure) is accepted as a measure of correctness in the protein domain. In light of our discussion in the preceding sections, the same criteria should apply when aligning RNA.

CONCLUSION

When initially asked to contribute to this book, I thought that we would provide a chapter on the problems with algorithmic methods. However, we find this to be an overly negative approach. We all have different backgrounds—and experience. Sometimes we see things in different ways, and our experience differs greatly from that of many of the other contributors to this book. These differences are a good thing, as differing points of view should be openly discussed and debated. Thus, our science progresses. It is all too easy in science to take an adversarial approach to those who disagree with us. It is not our intention in this chapter to be overly critical. We do support properly invoked algorithmic methods, and clearly stated optimality criteria. It is our hope that, by our pointing out some of the problems we have experienced in the alignment of rRNA data, program developers can incorporate solutions to these issues in their algorithms. Biologically realistic algorithms could make manual alignment less and less relevant. Here are some of the issues we see as most important toward the improvement of alignment programs.

In molecules whose function is dependent on structure, the conservation of the structure should be part of the optimality criterion. Minimizing structural change is as justified as minimizing change among nucleotides. Perhaps a program could be developed that could locate covarying sites in a multiple alignment. Some multistepped combination of calculating minimum free energy structures that are shared among multiple taxa, and then confirming those hypothesized structural interactions based on compensatory base changes, seems possible. Sites containing such compensatory base changes should be aligned together. A research group led by P. Stadler at the University of Leipzig and a research group led by B. Misof at the University of Bonn have developed a promising program called RNAsalsa that promises to do these things (B. Misof, personal communication), but we have not had a chance to evaluate it.

Gap costs should vary among regions. Manual alignments contain flexible gap costs, in that when a person comes to a hypervariable region with a lot of variation in length among sequences, gaps are more freely inserted. With an algorithm, there should be a way to locate conserved regions by some criteria, and then measure the range of length variation among taxa in the regions between them. Gap costs could then be regionally assigned based on how much length variation was observed, giving the lowest gap costs to regions that contain the most length heterogeneity. The standard deviation of the lengths could also play a role in gap cost determination, and in data exclusion criteria.

The iterative process of moving from guide trees to multiple alignments should be improved upon. Perhaps the initial guide tree should be developed from unambiguously aligned regions; we could then iteratively move through more difficult regions with guide trees developed only from data whose homology reaches some confidence threshold. It is fairly easy to see by eye when one lineage cannot be aligned with another. Reconstructing ancestral states to the root of a particular lineage may yield sequences so different from other lineages that it would be foolish to attempt to homologize them. If it is that easy to see for the human eye, there should be a way for a computer to measure this incompatibility as well, and reach objective criteria for data exclusion. Gblocks (Castresana 2000) provides a conservative means of data exclusion. Another interesting program called “Aliscore,” based on the identification of randomness in sequence alignments using a Monte Carlo approach, is being developed by Misof and Misof (personal communication). We need more information about how gaps accumulate and evolve in rRNA to model these characters. The greatest challenge is that gaps are not independent of one another, and are not randomly distributed across sites. Alignment programs must recognize both of these properties.

An ideal program would have a means to assess alignment uncertainty (as in Redelings and Suchard 2005). But alignment uncertainty is linked with the model, so it is important to remember that if the model for gaps is biologically unrealistic, then the “uncertainty” cannot be disentangled from those limitations. It is our impression that the differences among trees that are attributed to alignment methods are more often associated with different data exclusion criteria and philosophies. It should be possible to produce a program that incorporates some data exclusion criterion with alignment uncertainty.

Alignments involving moderate to extreme length heterogeneity across sampled sequences will undoubtedly invoke some degree of subjectivity from the investigator, regardless of the methodological approach (Kjer et al. 2007). The legitimate disagreements about the kinds of subjectivity that are justified will likely continue. Here we state our beliefs. Phylogenies are hypotheses only. We think that even though we can never prove a phylogeny to be true, phylogenies that are wrong are worse than worthless because they promote further inaccurate predictions. For phylogenies to be predictive, they must be accurate, and even if we cannot prove accuracy, none of us should be embarrassed to pursue it. Given that, we think it is imperative to intervene when it can be demonstrated that existing methods are failing. Reasoned subjectivity, with all assumptions defined and the alignment made public, is far more accessible than black box analysis justified under some philosophical principle. We suggest that you should do the best you can today with what you have, because if something better comes along later, the data are still available in GenBank for reanalysis. And the addition of new sequences to existing alignment templates is likely the better approach, not only for the re-estimation of phylogenies, but also for the evaluation of structural and functional predictions derived from said alignments (see Morrison 2006). Thus, we disagree with favoring purely algorithmic approaches, such as POY and others based only on the perceived future of direct optimization, simply because no algorithms to date match the level of empiricism ingrained within the biological (manual) method.

TERMINOLOGY

Comparative evidence for secondary structure base pairing comes predominantly from the observation of covarying Watson–Crick pairs (see the early works of Gutell, Noller, and Woese). Typically, contradiction of a AUQ11 covarying position is as follows: AA, CC, UU, GG; AC, AG, and CU (and their symmetrical equivalents) cause disruptive bulges. Gutell and others have observed that some of these pairs actually do covary, mostly within highly conserved regions of rRNA (e.g., see Lee and Gutell 2004), wherein selection favors noncanonical base pairs to foster a variety of tertiary interactions (reviewed in Noller 2005); however, for the alignment of variable regions of rRNA, consider them forbidden. Remember also that G↔U is a permitted hydrogen bonding pair in RNA. C↔A pairs do not appear to be as disruptive as the other noncanonical pairs listed above, and therefore, if there is comparative evidence of a site, C↔A pairs should not contradict the site. Contradiction of core helices and other strongly supported features of rRNA cannot come from a single taxon because sequencing errors can happen, just as evolution can happen. Some deleterious substitutions do not result in the immediate extinction of the lineage, so a single bulge or even a few bulges across an alignment cannot contradict a stem. AUQ12 However, if you see multiple bulges, treat the site as suspect and evaluate the support for the helix by a variety of means. Base pair frequency tables provide a means to measure the degree of covariation for a given base pair. Other statistics, such as mutual information and Kramer's statistic, further illustrate the manner of dependence within sites of a base pair. The entire stem need not be rejected, however, particularly if there are covarying sites at other positions. Ideally, an objective approach would be to accept only base pairs that have less than a certain percentage of noncanonical base pairs within proposed helices.

- **Helix (stem):** A right-handed double helix composed of a succession of complementary hydrogen-bonded nucleotides between paired strands.
- **Single strand (loop):** Unpaired nucleotides separating helices.
- **Hairpin-stem loop:** Helix closed distally by a loop of unpaired nucleotides (terminal bulge).
- **Terminal bulge:** Succession of unpaired nucleotides at the distal end of a hairpin-stem.
- **Lateral bulge:** Succession of unpaired nucleotides on one strand of a helix.
- **Internal bulge:** Group of nucleotides from two antiparallel strands unable to form canonical pairs.
- **Compensatory base change:** Subsequent mutation on one strand of a helix to maintain structure following initial mutation of a complementary base (aka CBC).
- **Insertion:** A single insertion of a nucleotide relative to the rest of the multiple sequence alignment (dependent on frequency and determination of direction of event relative to out-group).
- **Deletion:** A single deletion of a nucleotide relative to the rest of the multiple sequence alignment (dependent on frequency and determination of direction of event relative to out-group).
- **Indel:** An ambiguous position (column) within a multiple sequence alignment that cannot be described as an insertion or deletion.
- **Region of ambiguous alignment (RAA):** Two or more adjacent, nonpairing positions within a sequence wherein positional homology cannot be confidently assigned due to the high occurrence of indels in other sequences.
- **Region of slipped-strand compensation (RSC):** Region involved in base pairing wherein positional homology cannot be defended

across a multiple sequence alignment; inconsistency in pairing likely due to slipped-strand mispairing.

- **Region of expansion and contraction (REC):** Variable helical region flanked by conserved base pairs at the 5' and 3' ends, and an unpaired terminal bulge of at least three nucleotides; characteristic of RNA hairpin-stem loops.

Acknowledgments

JJG acknowledges support from NIAID contract HHSN266200400035C awarded to Bruno Sobral (VBI).

APPENDIX: INSTRUCTIONS ON PERFORMING A STRUCTURAL ALIGNMENT

(NOTE: *A conceptual approach for structural alignment of rRNA sequences and further preparation of the data for RNA maximum likelihood models of evolution is available at http://hymenoptera.tamu.edu/rna/methods.php. Below we provide a didactic example that will help the reader begin manually adjusting his or her own data*).

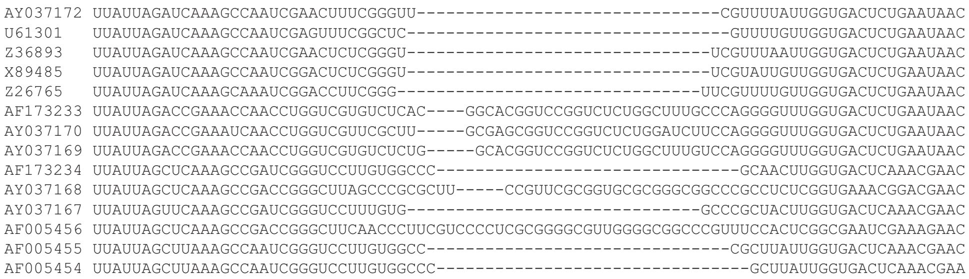
It is most often the mechanics of manual alignment that trip people up. It is not hard to convince people that doing structural alignments is a good idea. It is just that it becomes "too hard," and if people can get "close enough" with Clustal, they call it "good enough." We understand their pain, but disagree. The following suggestions should make the whole process a little easier:

The goal of a manual structural alignment is to make objective and repeatable decisions, using minimizing structural changes without being arbitrary. An example of being arbitrary would be to say: "retain all nucleotides that are identical among all taxa for ten nucleotides or more." Although it makes good sense, and results in a repeatable criterion, the selection of ten is arbitrary. In our opinion, it is better to admit to ambiguity than to hide from it and pretend you are being objective, in the hope that nobody will notice.

1. Align the sequences with Clustal or any other computer alignment program as a starting point. It works best to avoid a "gappy"-looking alignment, because you will need to manually adjust the gaps. The computer alignment is simply a timesaving

device, as any manual adjustment changes a computer alignment to a manual alignment. Clustal is easy to use, is available for multiple platforms, and permits multiple export formats. Export the alignment in a NEXUS format.

2. Open the Clustal NEXUS alignment with PAUP. From PAUP, go to the "file" pull down menus and "export data," using "file format, NEXUS," and clicking the box "interleave" with 130 characters per line. Close the original PAUP file, and reopen the new ".dat" file, which is now interleaved with 130 characters per line. This is just so that you have about the right number of characters in each block to be able to look at them all without scrolling back and forth on the screen.
3. Open the interleaved .dat file you just made with Microsoft Word (of course, there are other text editors; yet, not all of them allow for easy manipulation of the data). Format with courier bold 9 point font. You may have to format the document in "landscape" view, and reduce to 25% so that the lines do not wrap over. Setting a custom page size (22" by 22") may help. Now color the nucleotides by going to the "Edit" menu; go to "Replace" (select "more options" to find the "font" option, under "format"). Change all of the A's into green A's, the C's into blue C's, the U's into red U's, and leave the G's black. Now you have something that looks like Figure 7.11, except that yours AUQ13 would be in color (this example can be found on Kjer's website,):
4. Add a *palette* to each of the rows. A palette contains a variety of symbols that you may wish to insert, as a column, into the data matrix. As shown in Figure 7.12, the palette starts and ends with



```
AY037172 UUAUUAGAUCAAAGCCAAUCGAACUUUCGGUU--------------------------------CGUUUUAUUGGUGACUCUGAAUAAC
U61301   UUAUUAGAUCAAAGCCAAUCGAGUUUCGGCUC---------------------------------GUUUUGUUGGUGACUCUGAAUAAC
Z36893   UUAUUAGAUCAAAGCCAAUCGAACUCUCGGGU--------------------------------UCGUUUAAUUGGUGACUCUGAAUAAC
X89485   UUAUUAGAUCAAAGCCAAUCGGACUCUCGGGU--------------------------------UCGUAUUGUUGGUGACUCUGAAUAAC
Z26765   UUAUUAGAUCAAAGCAAAUCGGACCUUCGGG-------------------------------UUCGUUUUGUUGGUGACUCUGAAUAAC
AF173233 UUAUUAGACCGAAACCAACCUGGUCGUGUCUCAC----GGCACGGUCCGGUCUCUGGCUUUGCCCAGGGGUUUGGUGACUCUGAAUAAC
AY037170 UUAUUAGACCGAAAUCAACCUGGUCGUUCGCUU-----GCGAGCGGUCCGGUCUCUGGAUCUUCCAGGGGUUUGGUGACUCUGAAUAAC
AY037169 UUAUUAGACCGAAACCAACCUGGUCGUGUCUCUG-----GCACGGUCCGGUCUCUGGCUUUGUCCAGGGGUUUGGUGACUCUGAAUAAC
AF173234 UUAUUAGCUCAAAGCCGAUCGGGUCCUUGUGGCCC-------------------------------GCAACUUGGUGACUCAAACGAAC
AY037168 UUAUUAGCUCAAAGCCGACCGGGCUUAGCCCGCGCUU-----CCGUUCGCGGUGCGCGGGCGGCCCGCCUCUCGGUGAAACGGACGAAC
AY037167 UUAUUAGUUCAAAGCCGAUCGGGUCCUUUGUG------------------------------GCCCGCUACUUGGUGACUCAAACGAAC
AF005456 UUAUUAGCUCAAAGCCGACCGGGCUUCAACCCUUCGUCCCCUCGCGGGGCGUUGGGGCGGCCCGUUUCCACUCGGCGAAUCGAAAGAAC
AF005455 UUAUUAGCUUAAAGCCAAUCGGGUCCUUGUGGCC-------------------------------CGCUUAUUGGUGACUCAAACGAAC
AF005454 UUAUUAGCUUAAAGCCAAUCGGGUCCUUGUGGCCC--------------------------------GCUUAUUGGUGACUCAAACGAA
```

Figure 7.11. Algorithmically aligned sequences imported into Word, awaiting manual refinement.

```
AY037172 UUAUUAGAUCAAAGCCAAUCGAACUUUCGGGUU-------------------------------CGUUUUAUUGGUGACUCUGAAUAAC[()| *-]
U61301   UUAUUAGAUCAAAGCCAAUCGAGUUUCGGCUC---------------------------------GUUUUGUUGGUGACUCUGAAUAAC[()| *-]
Z36893   UUAUUAGAUCAAAGCCAAUCGAACUCUCGGGU-------------------------------UCGUUUAAUUGGUGACUCUGAAUAAC[()| *-]
X89485   UUAUUAGAUCAAAGCCAAUCGGACUCUCGGGU-------------------------------UCGUAUUGUUGGUGACUCUGAAUAAC[()| *-]
Z26765   UUAUUAGAUCAAAGCAAAUCGGACCUUCGGG-------------------------------UUCGUUUUGUUGGUGACUCUGAAUAAC[()| *-]
AF173233 UUAUUAGACCGAAACCAACCUGGUCGUGUCUCAC----GGCACGGUCCGGUCUCUGGCUUUGCCCAGGGGUUUGGUGACUCUGAAUAAC[()| *-]
AY037170 UUAUUAGACCGAAAUCAACCUGGUCGUUCGCUU-----GCGAGCGGUCCGGUCUCUGGAUCUUCCAGGGGUUUGGUGACUCUGAAUAAC[()| *-]
AY037169 UUAUUAGACCGAAACCAACCUGGUCGUGUCUCUG-----GCACGGUCCGGUCUCUGGCUUUGUCCAGGGGUUUGGUGACUCUGAAUAAC[()| *-]
AF173234 UUAUUAGCUCAAAGCCGAUCGGGUCCUUGUGGCCC-------------------------------GCAACUUGGUGACUCAAACGAAC[()| *-]
AY037168 UUAUUAGCUCAAAGCCGACCGGGCUUAGCCCGCGCUU-----CCGUUCGCGGUGCGCGGGCGGCCCGCCUCUCGGUGAAACGGACGAAC[()| *-]
AY037167 UUAUUAGUUCAAAGCCGAUCGGGUCCUUUGUG------------------------------GCCCGCUACUUGGUGACUCAAACGAAC[()| *-]
AF005456 UUAUUAGCUCAAAGCCGACCGGGCUUCAACCCUUCGUCCCCUCGCGGGGCGUUGGGGCGGCCCGUUUCCACUCGGCGAAUCGAAAGAAC[()| *-]
AF005455 UUAUUAGCUUAAAGCCAAUCGGGUCCUUGUGGCC-------------------------------CGCUUAUUGGUGACUCAAACGAAC[()| *-]
AF005454 UUAUUAGCUUAAAGCCAAUCGGGUCCUUGUGGCCC--------------------------------GCUUAUUGGUGACUCAAACGAA[()| *-]
```

Figure 7.12. The sequences with a symbol palette added onto the end of each row. The symbols will be used to describe structural elements. The palette should be contained within square brackets [] so as to keep the sequences in valid NEXUS format.

brackets, so that NEXUS will ignore the contents after they are eventually reimported into PAUP.

5. Go to Gutell's (http://www.rna.ccbb.utexas.edu) or Gillespie's website and download the most recent secondary structure diagram. AUQ14

6. Microsoft Word permits three essential things you need to be able to do. First, you want to see colors. Second, you must be able to move columns. To move columns in Word, you simply depress the "option" key as you drag down a column with the mouse. Finally, underlines are essential (more about them in step 9). So the next step is to find a stem from a structural model, and paste in the structural symbols from Kjer (1995) to indicate the putative boundaries of the stems. (Note: since 1995, I have replaced the bracket symbols with the "|" symbol to indicate long range stems, because the brackets have meaning in NEXUS that I had not considered in 1995.)

7. Apply structural symbols as in Kjer (1995) to the reference sequence, and fit them, one by one, onto each of the other sequences. Attempt to subdivide long single-stranded regions by looking for covariation, as in Kjer (1995). As a first pass, assume the structural model you have is correct, but if the data contradict it, then do what the data tell you. Structural models are inferred from comparative evidence, which is exactly what you have before you for a more specific set of taxa. These structures may evolve. If you see that your model does not fit your taxa, then alter it to a model that is supported by the evidence presented by the sequences. The signal in these regions comes from universal and covariable inferred hydrogen bonds (compensatory base changes). If all of the taxa can bond in a thermodynamically stable stem that is supported by compensatory base changes, and would also be unlikely to exist by chance, then this stem should be inferred and used in an alignment. You may propose modifications to structural models this way. It is not your task to construct a perfect secondary structural model, but, rather, to use the structure to infer homology. A portion of the stem for which the structure is ambiguous from the data cannot be used to define homology beyond what you can infer from the nucleotides (primary structure). So you should freely contract stems to the minimum common supported size, and let others

whose primary goals are to develop structural models worry about the differences.

8. Consult Hickson et al. (1996) and “Phylogenetic conservation superimposed onto the *E. coli* SSU rRNA” on Gutell’s website for conserved motifs.

9. Pasting the structural symbols provides only an initial rough hypothesis of base pairing. The next step is to confirm the hydrogen bonds. Since this is an iterative process, you MUST be able to trace what you have looked at, and differentiate those regions from the regions you have not yet finished. One way to do this is with underlines. Underlines indicate confirmed hydrogen bonds. They mean that you have looked at those individual nucleotides, and their partners, and a Watson–Crick, or G-U, base pair is possible. Laziness is the biggest problem at this point, because it is easy to drag entire columns of nucleotides, and simply underline them all without checking. A sloppy alignment is full of non-Watson–Crick pairs that are mistakenly underlined. Note in Figure 7.13 the bulge indicated by the lack of underlines in Z26765 GCAAA...UUGGU. If you cannot trust the underlines, you cannot trust the alignment. This is why we do not use some of the fancier phylogenetic data editors—because they do not offer the opportunity to visualize individual hydrogen bonds (or if they do, Kjer did not know). There may be a better way to do this. But any system must have confirmation of bonding at each site, as opposed to a mask applied to the top sequence.

10. Line up the stems. If the stems do not line up because there are alternative lengths, slippage, or a lack of structural conservation, pull back on the stems, and consider them *unaligned*. Put an empty space to mark the unaligned regions, as above. Use empty spaces to help you break up the alignment, so that you can get a better look at it. Think carefully about data exclusion. For example, can you justify aligning the above UUUUG with CAAC? If not, then eliminate this region, and code it as you see fit with some other method. The structure will define the aligned regions, and delimit the unaligned regions. If there is no length variation in the single-stranded region, keep it in, as in the region below “V2”: AGAUCAAA. If there is length variation, without conserved nucleotide motifs, throw it out, put it into INAASE,

```
[               V2                                Region 4                                           ]
AY037172 (UUAUUAGAUCAAAGCCAA U CGAA -----------------------------CUUUCGGG UUCG UUUUA-- UUGGUGACUCUGAAUAA)C[()| *-]
U61301   (UUAUUAGAUCAAAGCCAA U CGAG U------------------------------UUCGG- CUCG UUUUG-- UUGGUGACUCUGAAUAA)C[()| *-]
Z36893   (UUAUUAGAUCAAAGCCAA U CGAA CU-----------------------------CUCGGG UUCG UUUAA-- UUGGUGACUCUGAAUAA)C[()| *-]
X89485   (UUAUUAGAUCAAAGCCAA U CGGA C-----------------------------UCUCGGG UUCG UAUUG-- UUGGUGACUCUGAAUAA)C[()| *-]
Z26765   (UUAUUAGAUCAAAGCAAA U CGGA CC-----------------------------UUCGGG UUCG UUUUG-- UUGGUGACUCUGAAUAA)C[()| *-]
AF173233 (UUAUUAGACCGAAACCAA C CUGG UCGUGUCUCACGGCA-CGGUCCGGUCUCUGGCUUUGC CCAG GGGU--- UUGGUGACUCUGAAUAA)C[()| *-]
AY037170 (UUAUUAGACCGAAAUCAA C CUGG UCGUUCGCUUGCGAG-CGGUCCGGUCUCUGGAUCUU- CCAG GGGU--- UUGGUGACUCUGAAUAA)C[()| *-]
AY037169 (UUAUUAGACCGAAACCAA C CUGG UCGUGUCUC-UGGCA-CGGUCCGGUCUCUGGCUUUGU CCAG GGGU--- UUGGUGACUCUGAAUAA)C[()| *-]
AF173234 (UUAUUAGCUCAAAGCCGA U CGGG UCCUU----------------------------GUGG CCCG CAAC--- UUGGUGACUCAAACGAA)C[()| *-]
AY037168 (UUAUUAGCUCAAAGCCGA C CGGG CUUAGCCCGCGCUUC--CGUUCGCGGUGCGCGGGCGG CCCG CCUC--- UCGGUGAAACGGACGAA)C[()| *-]
AY037167 (UUAUUAGUUCAAAGCCGA U CGGG UCCU---------------------------UUGUGG CCCG CUAC--- UUGGUGACUCAAACGAA)C[()| *-]
AF005456 (UUAUUAGCUCAAAGCCGA C CGGG CUUCAACCCUUCGUCCCCUCGCGGGGCGUUGGGGCGG CCCG UUUCCAC UCGGCGAAUCGAAAGAA)C[()| *-]
AF005455 (UUAUUAGCUUAAAGCCAA U CGGG UCCUUGU----------------------------GG CCCG CUUA--- UUGGUGACUCAAACGAA)C[()| *-]
AF005454 (UUAUUAGCUUAAAGCCAA U CGGG UCCUUGU----------------------------GG CCCG CUUA--- UUGGUGACUCAAACGAA)-[()| *-]
```

Figure 7.13. Underlined text is used to indicate confirmed hydrogen bonds (Watson–Crick and G-U base pairings) and allows one to keep track of which sites have been examined.

or try some iterative tree-based computer alignment on these regions.

11. Think about dots. When applied to sequences to indicate identity with a reference sequence (on top in Figure 7.14), they help you to visualize compensatory base changes, as well as synapomorphies. They will also make misalignments stand out so that you can see them. You can insert them by moving a section into MacClade, and selecting the "matchar" option, but in doing so, you would lose your structural symbols.

12. Define regions of ambiguous alignment. A candidate for a region of sequence that may be considered as "ambiguously aligned" is initially any region containing length variation among taxa. Objectively subdividing this assignment becomes the more important task because the initial definition applies to the whole sequence. There are three types of information that help to designate regions into aligned and ambiguously aligned classes. First, an ambiguously aligned region is any region containing length variation among taxa that is flanked by hydrogen-bonded stems, in which there is more than one equally plausible alignment. This assignment alone will subdivide the whole gene into multiple fragments. Once the secondary structure has defined the boundaries of ambiguity, additional information comes from the nucleotides. Attempt to manually align the region. Consult both Gutell and Hickson et al. for conserved motifs, and if all taxa have them, align the conserved motifs together to further subdivide the region. Ask yourself if a panel of judges were to look at every gap in your alignment, whether or not you could defend your decisions to the point where no other placement would be equally parsimonious. Consider transitions to be more likely than transversions, and also, make consistent decisions about how heavily to consider one or a few aberrant taxa in an otherwise length-homogeneous region. Decide the degree of nucleotide similarity among taxa that is required to expand into the regions defined by the flanking hydrogen bonds. Remember, each decision you make is a hypothesis of homology that can be reviewed and overturned. Therefore, you do not need to be perfect, because if you publish your hypotheses, they can be repeated and or contested.

```
[                V2                                   Region 4                                      ]
AY037172 (UUAUUAGAUCAAAGCCAA U CGAA -----------------------------CUUUCGGG UUCG UUUUA-- UUGGUGACUCUGAAUAA)C[()| *-]
U61301   (.................. . ...G U------------------------------UUCGG- C... UUUUG-- ...G.............).[()| *-]
Z36893   (.................. . .... CU-----------------------------CUCGGG .... UUUAA-- .................).[()| *-]
X89485   (.................. . ..G. C-----------------------------UCUCGGG .... UAUUG-- .................).[()| *-]
Z26765   (...............A.. . ..G. CC-----------------------------UUCGGG .... UUUUG-- .................).[()| *-]
AF173233 (........C.G..A.... C .UGG UCGUGUCUCACGGCA-CGGUCCGGUCUCUGGCUUUGC CCA. GGGU--- .................).[()| *-]
AY037170 (........C.G..AU... C .UGG UCGUUCGCUUGCGAG-CGGUCCGGUCUCUGGAUCUU- CCA. GGGU--- .................).[()| *-]
AY037169 (........C.G..A.... C .UGG UCGUGUCUC-UGGCA-CGGUCCGGUCUCUGGCUUUGU CCA. GGGU--- .................).[()| *-]
AF173234 (.......C........G. . ..GG UCCUU----------------------------GUGG CC.. CAAC--- ..........AA.CG..).[()| *-]
AY037168 (.......C........G. C ..GG CUUAGCCCGCGCUUC--CGUUCGCGGUGCGCGGGCGG CC.. CCUC--- .C.....AA.G..CG..).[()| *-]
AY037167 (.......U........G. . ..GG UCCU---------------------------UUGUGG CC.. CUAC--- ..........AA.CG..).[()| *-]
AF005456 (.......C........G. C ..GG CUUCAACCCUUCGUCCCCUCGCGGGGCGUUGGGGCGG CC.. UUUCCAC .C..C..A..GA..G..).[()| *-]
AF005455 (.......C.U........ . ..GG UCCUUGU----------------------------GG CC.. CUUA--- ..........AA.CG..).[()| *-]
AF005454 (.......C.U........ . ..GG UCCUUGU----------------------------GG CC.. CUUA--- ..........AA.CG..)-[()| *-]
```

Figure 7.14. The use of dots to indicate identity with a reference sequence can ease the identification of compensatory base changes, as well as synapomorphies.

13. Once you have finished the alignment in Word, import the whole thing back into PAUP, and use the "Replace" option in the Edit menu to change all the parentheses, and lines "(" , ")" and " | " into blank spaces. The NEXUS file should be an exact match to the Word file, except that it will lack color, and the structural symbols.

AUQ1: (Differentiation of Molecules, 1st paragraph) Please supply reference entry for Cate et al. 1999.
AUQ2: (Compositional Bias Presents a Severe Challenge, 1st paragraph) Please supply reference entry for Brooks and McLennan 1994.
AUQ3: (Compositional Bias Presents a Severe Challenge, 1st paragraph) Please supply reference entry for Lanyon 1988.
AUQ4: (Compositional Bias Presents a Severe Challenge, 1st paragraph) Please supply reference entry for Mishler et al. 1988.
AUQ5: (Gaps Are Not Uniformly Distributed, 3rd paragraph) Suggest define "ILD"
AUQ6: (Long Inserts/Deletions, 1st paragraph) Is "nts" an abbreviation for nucleotides? If so, probably better to write out the word
AUQ7: (Are Structural Inferences Justified?, last line) Would "taxa" in both places be better as "taxon"?
AUQ8: (An Example of Accuracy and Repeatability, 2nd paragraph) Please supply reference entry for Amrine-Madsen et al. 2003.
AUQ9: (An Example of Accuracy and Repeatability, 2nd paragraph) Please supply reference entry for Delsuc et al. 2002.
AUQ10: (An Example of Accuracy and Repeatability, 2nd paragraph) Please supply reference entry for Waddell and Shelley 2003.
AUQ11: (Terminology, 1st paragraph) Please provide reference entries for the "early works of Gutell, Noller, and Woese."
AUQ12: (Terminology, 1st paragraph) "a single bulge" correct?
AUQ13: (Appendix, item 3) Please provide the URL for Kjer's website.
AUQ14: (Appendix, item 5) Please provide the URL for Gillespie's website.
AUQ15: (Figure 7.10 caption) Please supply reference entries for Kjer, Ober, and Gillespie.