The following E-mail was sent to Joe Gillespie and Karen Ober as instructions on how to align with secondary structure, following an invitation to collaborate on this project in 2003. Both were invited based on their familiarity with the use of POY, and were asked to use POY as rigorously as expected for publication. Neither were directly associated with my lab (as I did not think my students or associates could provide an independent test of the methods).

Dear Joe and Karen,

The object of this exercise is to compare structural alignment with computer alignment, and to address whether or not objective criteria can followed with either method, resulting in similar conclusions among investigators, using the same datasets. I selected mitochondrial mammalian 12S and 16S rRNA. This dataset was selected primarily because the relationships among these taxa is accepted among mammologists as non-controversial. The dataset is AT rich, and a quick Clustal alignment leads to a tree that is very close to the accepted relationships when using the combined 12S and 16S, but when these datasets are analyzed separately, neither performs as well. It is reasonable to assume that mitochondrial genes share the same history, so the lack of agreement between the partitioned and combined data is due to the properties of the data and/or the analysis. Since the combined data recover close to the accepted tree in a crude Clustal alignment, there is a reasonable possibility that the accepted tree can be recovered from the partitioned data. However, the length heterogeneity and the compositional bias will present a challenge. I have not performed such an analysis, and will be exploring this question with you. I am sending you the taxa without labels, in no particular order. One exception is that the cow sequence is labeled so that we have a common reference taxon for which the secondary structure has been published by Gutell. It is important that you do not attempt to figure out what the other taxa are, and align them without any reference to perceived phylogenetic structure. It may be justified to use phylogenetic information to construct a manual alignment, but such information would invalidate this test. It is also important that you follow the instructions in this letter, which will be published along with the paper on a website. You may ask for help and advice, but only direct reference to published works will be provided. Depending on your familiarity with structural alignment, you may wish to read Gutell et al., 1994, Kjer 1995, Kjer 1997, and Hickson et al., 1996. The standard of success is not perfection; the structural alignment need only be defended as more justified than an alignment based on nucleotides alone. Before you start the alignment, please comment on the rules below with any suggestions on how to improve them to make them more objective. The goal is to make them objective and repeatable without being arbitrary. An example of being arbitrary would be to say; "retain all nucleotides that are identical among all taxa for 10 nucleotides or more". Although it makes good common sense, and results in a repeatable criterion, the selection of 10 is arbitrary. In my opinion, it is better to admit to ambiguity than to pretend. Use common sense. The goal is not to think like a computer. I have no doubt that each manual alignment will differ at a few positions. Computer alignments will also differ depending on the parameters that are selected. If there exists some underlying organization of the rRNA into structurally conserved stems, and this conservation is to a higher degree than the conservation among nucleotides, then structural alignments may more consistently recover phylogenetic patterns than systems that less accurately define homology, even if the alternative systems are more explicit.

Here are the rules:
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 devise, as any manual adjustment changes a computer alignment to a manual alignment.

2. Go to the Gutell website and download the most recent 16S *Bos taurus* secondary structure.

3. Apply structural symbols as in Kjer (1995) to the cow sequence, using Gutell's cow model, 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 Gutell is correct, but if the data contradict Gutell, then do what the data tell you. The Gutell model was originally 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 the Gutell model does not fit mammals, 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 of the Gutell model 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.

4. Consult Hickson et al. (1996) and Gutell's "Phylogenetic conservation superimposed onto the *E. coli* SSU rRNA" for conserved motifs.

5. 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, use common sense with regards to how heavily to consider one or a few aberrant taxa in an otherwise length homogeneous region. Using common sense, 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.

7. Once the alignment is finished, there will be no further adjustment, except to incorporate Lutzoni et als.(2000) method of coding for ambiguously aligned regions, which were defined in step 5. Send the PAUP file back to me, with clear indication of the ambiguously aligned regions. Lutzoni et al. discuss a problem with INAASE, and I agree with them. The problem is that some step matrices call for an extreme number of changes in order to move from one sequence to another. Some of the step matrices call for up to 30 or more steps. Is it justified to put such a heavy weight on a region that is fast evolving and ambiguously aligned, that could have come about from insertions and deletions of whole strings of nucleotides? I think not. We can say that 30 is too high. But what about 29? 20? 15? 10? 3? The cutoff is arbitrary. So I ask you to pick an arbitrary maximum value for the step matrices produced by INAASE.

Definitions:
Comparative evidence comes from covarying Watson-Crick pairs. See works of Gutell and Woese. Contradiction of a covarying position is as follows: AA, AG, CC, UU, and GG cause disruptive bulges. Gutell and others have observed that some of these pairs actually do co-vary, but for alignment purposes, consider them forbidden. Remember also that GU is a permitted hydrogen bonding pair in RNA. CA pairs do not appear to be as disruptive as the five pairs listed above, and therefore if there is comparative evidence of a site, CA pairs should not contradict the site. Contradiction of a stem proposed by Gutell cannot come from a single taxon because sequencing errors can happen, just as evolution can happen. Some substitutions do not result in the immediate extinction of the lineage, so single bulges cannot contradict a stem. However, if you see multiple bulges, reject the site. The entire stem need not be rejected however, particularly if there are covarying sites at other positions.

Broughton, R.E., S.E. Stanley, AND R.T. Durrett. 2000. Quantification of homoplasy for nucleotide transitions and transversions and reexamination of assumptions in weighted phylogenetic analyses. Syst. Biol. 49:617-627

Gutell, R.R., N. Larsen, AND C.R. Woese. 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol. Rev. 58:10-26.

Hickson, R. E., C. Simon, A. J. Cooper, G. Spicer, J. Sullivan, & D.Penny. 1996. Refinement of a secondary structure model for the third domain of animal 12S rRNA, with a comparison of alignment programs. Mol. Biol. and Evol. 13:150-169

Kjer, K.M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Mol. Phylogenet. Evol. 4:314-330.

Kjer, K.M. 1997. An alignment template for amphibian 12S rRNA, domain III: Conserved primary and secondary structural motifs. J. Herpetology. 31:599-604.

Lutzoni, F., P. Wagner, AND V. Reeb. (2000). Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses using unequivocal coding and optimal character-state weighting. Syst. Biol. 49:628-651.

Helpful Hints (written by K. Kjer, Aug. 14, 2006)

These instructions were NOT included to Drs. Ober and Gillespie. In fact, Joe Gillespie has his own ideas on this, and I encourage you to look at his jRNA website (http://hymenoptera.tamu.edu/rna/). Joe's materials provide the ability to convert structural alignments into a variety of masks that are read by programs such as PHASE.

Since we began working on this paper, which deals more with justifications and philosophy than with "how to", I have visited with people about structural alignments. I find it is most often the mechanics of the 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". I understand their pain, but disagree. So here is what I do that should make the whole process a little easier:

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 devise, as any manual adjustment changes a computer alignment to a manual alignment. I like Clustal because it 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" pulldown 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 you have about the right number of characters in each block to be able to look at them all without scrolling.

3. Open the interleaved .dat file you just made in Microsoft Word. 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 don't scroll over. (This document should be viewed in "landscape" format).

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 the As into green As, the Cs into blue Cs, the Us into red Us, and leave the Gs black. Now you have something that looks like this:

```
AY037172 UUAUUAGAUCAAAGCCAAUCGAACUUUCGGGUU------------------------------CGUUUUAUUGGUGACUCUGAAUAAC
U61301_L UUAUUAGAUCAAAGCCAAUCGAGUUUCGGCUC----------------------------------GUUUUGUUGGUGACUCUGAAUAAC
Z36893_C UUAUUAGAUCAAAGCCAAUCGAACUCUCGGGU------------------------------UCGUUUAAUUGGUGACUCUGAAUAAC
X89485_P UUAUUAGAUCAAAGCCAAUCGGACUCUCGGGU------------------------------UCGUAUUGUUGGUGACUCUGAAUAAC
Z26765_H 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
```

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. The palette starts and ends with brackets, so that NEXUS will ignore the contents.

```
AY037172 UUAUUAGAUCAAAGCCAAUCGAACUUUCGGGUU------------------------------CGUUUUAUUGGUGACUCUGAAUAAC[()| *-]
U61301_L UUAUUAGAUCAAAGCCAAUCGAGUUUCGGCUC----------------------------------GUUUUGUUGGUGACUCUGAAUAAC[()| *-]
Z36893_C UUAUUAGAUCAAAGCCAAUCGAACUCUCGGGU------------------------------UCGUUUAAUUGGUGACUCUGAAUAAC[()| *-]
X89485_P UUAUUAGAUCAAAGCCAAUCGGACUCUCGGGU------------------------------UCGUAUUGUUGGUGACUCUGAAUAAC[()| *-]
Z26765_H 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[()| *-]
```

5. The reason I use Microsoft word is that it permits the 3 things I want to be able to do. First, I want to see colors. Second I 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 6). 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.)

6. Pasting the structural symbols only provides 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. The way I do this is with underlines. Underlines indicate confirmed hydrogen bonds. They mean that I 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 the bulge indicated by the lack of underlines below in **Z26765 GCAAA ... UUGGU.** If you can't trust the underlines, you can't trust the alignment. This is why I don't use GCG or some of the other fancier data editors, because they don't offer the opportunity to visualize individual hydrogen bonds (or I they do, I didn't know). I would not be surprised if there is 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.

```
[              V2                                  Region 4                                    ]
AY037172 (UUAUUAGAUCAAAGCCAA U CGAA ----------------------------CUUUCGGG UUCG UUUUA-- UUGGUGACUCUGAAUAA)C[()| *-]
U61301_L (UUAUUAGAUCAAAGCCAA U CGAG U--------------------------------UUCGG CUCG UUUUG-- UUGGUGACUCUGAAUAA)C[()| *-]
Z36893_C (UUAUUAGAUCAAAGCCAA U CGAA CU-----------------------------CUCGGG UUCG UUUAA-- UUGGUGACUCUGAAUAA)C[()| *-]
X89485_P (UUAUUAGAUCAAAGCCAA U CGGA C------------------------------UCUCGGG UUCG UAUUG-- UUGGUGACUCUGAAUAA)C[()| *-]
Z26765_H (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--- UUGGUGACUC-AAACGA)A[()| *-]
```

7. Line up the stems. If the stems don't 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 these 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, or try some iterative tree-based computer alignment on these regions.

8. 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. I'm here to help if you need it; Kjer@aesop.rutgers.edu.