Technical Reports inTaxonomy 00-01

On The Dangers Of Aligning RNA Sequences Using "Conserved" Motifs

Roderic D. M. Page

Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow, United Kingdom

RH: Aligning RNA sequences

Keywords:

Address for correspondence and reprints: Roderic D. M. Page, DEEB, IBLS, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, UK. E-mail: r.page@bio.gla.ac.uk
Tel: +44 141 330 4778
Fax: +44 141 330 5971

Aligning RNA sequences can be a challenging task. Automatic sequence alignment programs typically align sequences only with respect to primary sequence, and as a result may yield spurious alignments. Incorporating information on RNA secondary structure can improve the alignment (Kjer, 1995; Titus and Frost, 1996), but this must usually be done by hand. Various algorithms and programs exist that incorporate RNA secondary structure, but these are either limited to pairwise alignment of one sequence with respect to a reference sequence and structure (Bafna et al., 1996; Corpet and Michot, 1994; Lenhof et al., 1998; Notredame et al., 1997), or are too computationally intensive to be applicable to sequences longer than about 150 nucleotides (Eddy and Durbin, 1994).

Given the current lack of automatic methods for aligning RNA sequences, we could ask how well standard alignment programs perform. Hickson et al. (2000) addressed this question using a suite of ten conserved motifs to score the alignments produced by five different programs. They employed a reference alignment for 10 mitochondrial 12S rRNA sequences constructed manually using conserved motifs (Hickson et al., 1996). In their discussion the authors noted (p. 535) that the honeybee caused the five alignment programs the most difficulties. By comparing their reference alignment (their fig. 1) to alignments in the small subunit RNA database (van de Peer et al., 2000), and an alignment of insect 12S rRNA secondary structure (unpublished data), it is clear that Hickson et al. have incorrectly aligned the honeybee sequence between motifs 7 and 10. They identify the five bases UGAAA at position 14866-14870 in the honeybee mitochondrial genome (Crozier and Crozier, 1993) as motif 8. Doing this results in a lengthy insertion in the honeybee sequence upstream from motif 8, and a corresponding deletion upstream of motif 10 (Figure 1). Their alignment also shows a single gap in motif 9, which includes helix 33' in Hickson et al.'s (1996) secondary structure model. No other sequence in their alignment (or the larger one they published in 1996) has a gap in this highly conserved helix. They also postulate a large deletion in helix 48, which removes the loop and part of the 3' stem from this helix (Figure 1). These violations of conserved structures in 12S rRNA casts serious doubt on the alignment. I

suggest that the authors have placed too much reliance on motif 8 ("yrgrr") being conserved across all taxa.

To investigate the apparent misalignment further, I used the program RAGA (Notredame et al., 1997) to align the honeybee sequence to the Drosophila sequence. RAGA uses a genetic algorithm to align a RNA sequence for which the secondary structure is unknown (the "slave") to a "master" sequence which has a known secondary structure. The alignment score for a pair of sequences is equal to:

$$\text{alignment score} = \text{Pr} + (\lambda \times \text{Se}) - \text{gap penalty}$$

where Pr is the primary score, Se is the secondary structure score, and λ is the relative contribution of Se to the alignment score. If $\lambda = 0$ then the sequences are aligned solely with respect to their primary structure (i.e., the nucleotide sequence); if $\lambda > 0$ then both primary and secondary structure are considered simultaneously — I used the default value of $\lambda = 3$. The primary score is simply a function of the number of sites at which the two sequences share the same nucleotide (a match scores 10, a mismatch scores 0). The secondary score is the sum of scores for each base that is paired (part of a stem), and is 30 for GC pairs, 20 for UA and UG pairs, 10 for AG pairs, otherwise it is 0. The gap penalty score was 50 for opening a gap in a loop, 30 for opening a gap in a stem, and 0.3 for extending a gap. To evaluate Hickson et al.'s alignment the Drosophila sequence (the only other insect in Hickson et al.'s alignment) was chosen as the master sequence, and the honeybee sequence was the slave. The secondary structure model for Drosophila was taken from Hickson et al. (1996, fig. 2).

Using the optimality criterion implemented in RAGA, Hickson et al.'s alignment between Drosophila and honeybee has a score of 48200. The best alignment found with RAGA's genetic algorithm had a score of 61190. Figure 1 shows these two alignments for the region downstream of helix 47. The bases in the honeybee sequence identified as motif 8 by Hickson et al. are offset by 13 positions in the RAGA alignment, so that the sequence corresponding to motif 8 in the honeybee is "UUAAU" instead of "UGAAA". Similarly, motif 9 has been misidentified in the honeybee sequence, and the RAGA alignment preserves the conserved helices 33 and 48. Hickson et al.'s (1996, p. 166) suggestion that the original honeybee sequence (Crozier and Crozier, 1993) might be in error, despite the gels being rechecked, also is a direct result of their misaligning the honeybee. They noted (p. 166) that most taxa have "GrA" just prior to helix 33', whereas in their 1996 alignment helix 33' in the honeybee is preceded by "AAA." In the RAGA alignment (Figure 1), helix 33' is preceded by "GAA", which is consistent with the "GrA" motif. Given that the RAGA alignment has a better score than the Hickson et al. alignment, and does not violate accepted secondary structural models for 12S rRNA (including their own in Hickson et al., 1996), I think it is clearly the better alignment. Interestingly, for these two sequences ClustalW (Thompson et al., 1994) produces an alignment with a higher score (53780) than the manual Hickson et al. alignment. Hence, for these two sequences, automatic alignment ignoring secondary structure outperformed a manual structural alignment.

In drawing attention to the misalignment of the honeybee sequence in Hickson et al.'s study, I do not wish to suggest that secondary structure is not an important consideration in RNA sequence alignment, nor to challenge Hickson et al.'s general

conclusions about the performance of the alignment programs. Hickson et al. (p. 535) note even if the honeybee sequence was removed, there were differences in the accuracy among the five sequence alignment programs they investigated. The important point is that relying on conserved motifs to align RNA sequences may lead to gross errors of alignment if, in fact, those motifs are not conserved. This highlights the pressing need for computationally feasible methods for aligning multiple RNA sequences with respect to both primary and secondary structure.

Acknowledgments

This work was partly supported by NERC grant GR3/11075, and by an EMBO Visiting Fellowship while the author was at the DKFZ, Heidelberg.

References

Bafna, V., S. Muthukrishnan, and R. Ravi. 1996. Computing similarity between RNA strings. DIMACS Technical Report **96-30**.

Corpet, F., and B. Michot. 1994. RNAlign program: alignment of RNA sequences using both primary and secondary structures. Comput. Applic. Biosci. **10**:389-399.

Crozier, R. H., and Y. C. Crozier. 1993. The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. Genetics **113**:97-117.

Eddy, S. R., and R. Durbin. 1994. RNA sequence analysis using covariance models. Nucleic Acids Res. **222**:2079-2088.

Hickson, R. E., C. Simon, A. Cooper, G. S. Spicer, J. Sullivan, and D. Penny. 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. Mol. Biol. Evol. **13**:150-169.

Hickson, R. E., C. Simon, and S. W. Perrey. 2000. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. Mol. Biol. Evol. **17**:530-539.

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. Phylog. Evol. **4**:314-330.

Lenhof, H.-P., K. Reinert, and M. Vingron. 1998. A polyhedral approach to RNA sequence structure alignment. J. Comput. Biol. **5**:517-530.

Notredame, C., E. A. O'Brien, and D. G. Higgins. 1997. RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Res. **25**:4570-4580.

Thompson, J. D., D. G. Higgns, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673-80.

Titus, T. A., and D. R. Frost. 1996. Molecular homology assessment and phylogeny in the lizard family Opluridae (Squamata: Iguania). Mol. Phylog. Evol. **6**:49-62.

van de Peer, Y., P. de Rijk, J. Wuyts, T. Winkelmans, and R. de Wachte. 2000. The European Small Subunit Ribosomal RNA database. Nucleic Acids Res. **28**:175-176.

Figure caption

Figure 1

Comparison of the alignment of part of domain III of 12S rRNA for Drosophila and the honeybee by Hickson et al. (Hickson et al., 2000, fig. 1) with that produced using the program RAGA (Notredame et al., 1997). In each alignment the location of stems are indicated by '>' and '<' and are numbered following van de Peer et al. (2000). The location of the conserved motifs used by Hickson et al. are indicated by black boxes. Note that Hickson et al.'s alignment interrupts two helices (33' and 48'). The RAGA alignment restores these helices, and suggests that motifs 8 and 9 in the honeybee sequence have been misidentified, leading the two sequences to be misaligned.

Figure 1