The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search Supplementary material

Jakob H. Havgaard, Rune B. Lyngsø, and Jan Gorodkin

Results

The performances of the *Scan* and *Local* types of comparisons have recently been investigated [1]. It was found that the scan method could locate known RNA structures with a sensitivity of 0.8 and a positive prediction value of 0.9. The local structure prediction performance was found to have an average Matthews Correlation Coefficient (CC) of 0.7 [2].

To find the best set of parameters for global comparison a low sequence similarity dataset was made using the 5S rRNA, tRNA, and U1 databases [3, 4, 5]. Sequences containing non-A, C, G, U nucleotides were removed. The databases were redundancy reduced to 90 percent identity [6], and sequence pairs with more than 40% identity were removed. As a global alignment reaches from one end of the sequences to the other, it is necessary that the length difference between the two sequences is less than or equal to the *Maximum length difference (delta)* parameter. For the global type of comparison this parameter has a maximum of 25 nucleotides on the web server. Sequence pairs with a length difference larger than this were therefore discarded. The global dataset contains 8 5S rRNA pairs, 1044 tRNA pairs, and 5 U1 pairs. Several sets of score matrices were tested. The average correlation coefficient (CC) for the best score matrix is 0.69. For the individual families it is: 0.69 for 5S rRNA, 0.78 for tRNA, and 0.60 for U1. The gap opening cost for 5S rRNA and U1 was -30 and elongation -15. For tRNA the gap opening cost was -60 and elongation -30. That the optimal gap penalty for predicting 5S rRNA structures is smaller than the gap penalty optimal for tRNA structures was also found by [7]. Figure S1 shows the average CC as a function of gap opening cost.

To further test the performance of the server the dataset was extended to include sequence pairs with a maximum pairwise sequence identity up to 70%, and additional independent test sequences from the SRP database (version 151) were included as well [8]. Only sequences with structures consisting of one stem-loop were included (all belonging to the bacterial subdivision). The performance as a function of the gap opening cost for the SRP data is also shown in Figure S1. The performance as function of pairwise sequence identity can be seen in Figure S2. The combined CC is 0.72, with 0.69 for 5S rRNA, 0.81 for SRP, 0.79 for tRNA, and 0.57 for U1. The fluctuations in performance for 5S rRNA, SRP, and U1 are due to the very limited amount of data. This is also true for fluctuations in the low and high identity parts of the tRNA performances.

References

[1] J.H. Havgaard, R. Lyngsø, G.D. Stormo, and J. Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 10.1093/bioinformatics/bti279, 2005.

[2] B.W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451, 1975.

[3] M. Szymanski, M.Z. Barciszewska, V.A. Erdmann, and J. Barciszewski. 5S Ribosomal RNA Database. *Nucleic Acids Research*, 30(1):176–8, 2002.

[4] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, 26(1):148–53, 1998.

[5] C. Zwieb. The uRNA database. *Nucleic Acids Research*, 24(1):76–9, 1996.

[6] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of representative protein data sets. *Protein Science*, 1(3):409–17, 1992.

[7] D.H. Mathews and D.H. Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, 317(2):191–203, 2002.

[8] M.A. Rosenblad, J. Gorodkin, B. Knudsen, C. Zwieb, and T. Samuelsson. SRPDB: Signal Recognition Particle Database. *Nucleic Acids Research*, 31(1):363–4, 2003.

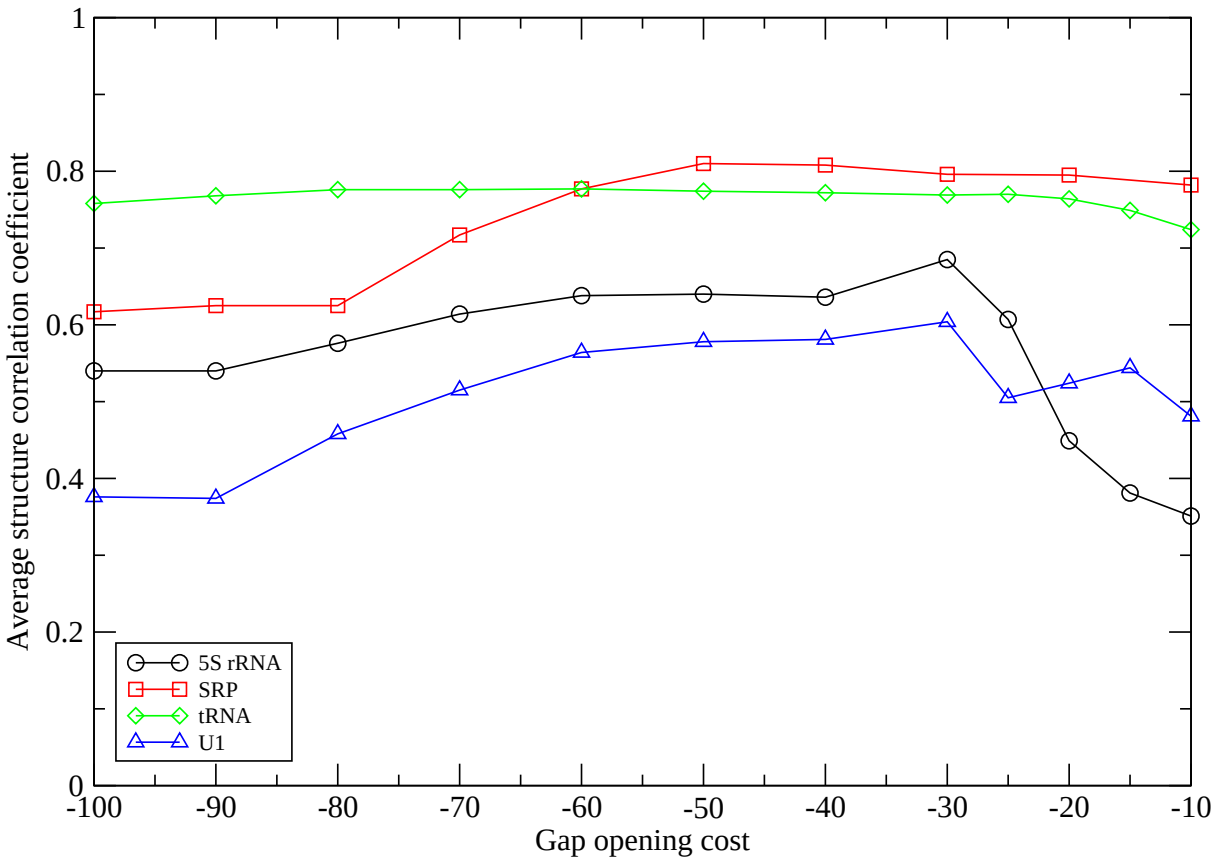

Figure S1: The average performance as function of the gap opening cost. In addition to the performance for low similarity 5S rRNA, tRNA, and U1 data the performance for the stem-loop SRP data is also shown.

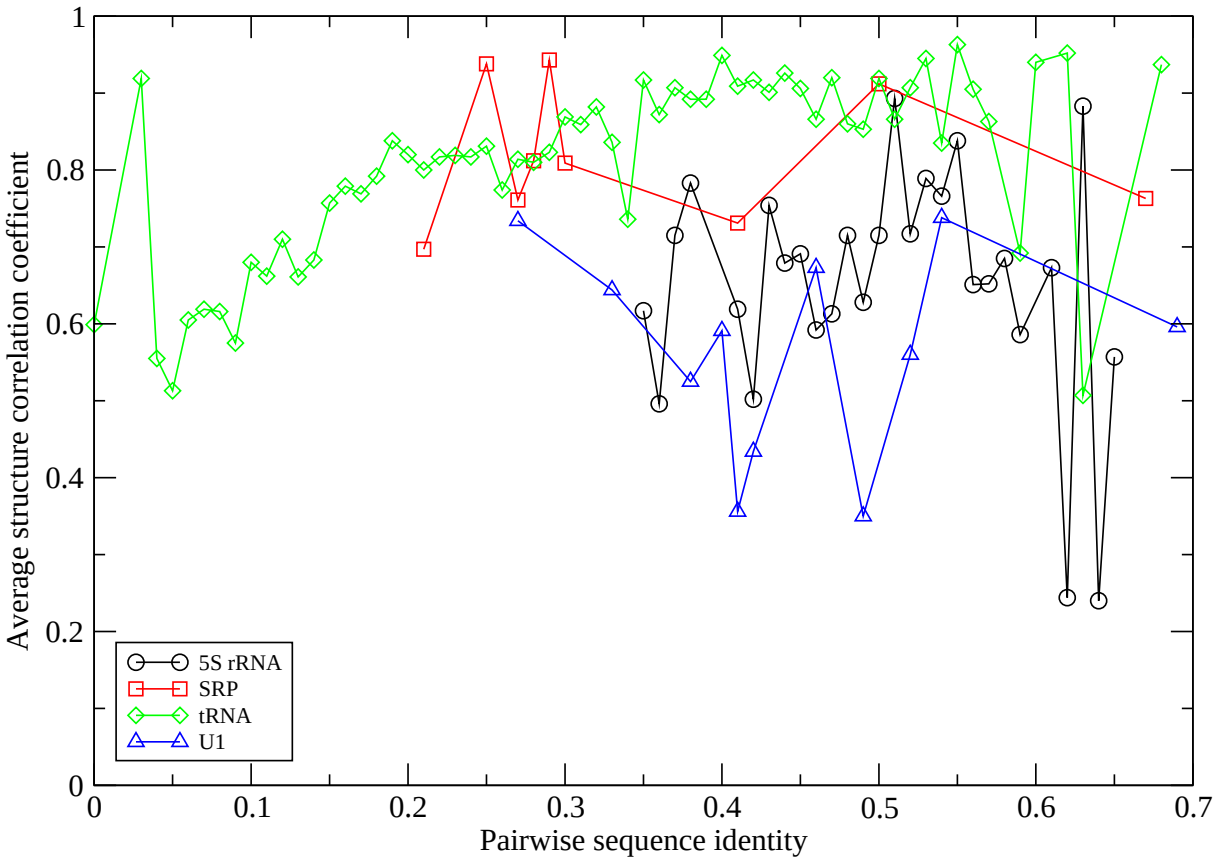

Figure S2: Performance as a function of sequence identity. The optimal gap penalties for 5S rRNA and U1 were: Gap opening -30 and gap elongation -15. For tRNA they were: Gap opening -60 and gap elongation -30. For the stem-loop SRPs the gap opening cost was -50, and the gap elongation cost was -25. The fluctuations in performance for 5S rRNA, SRP, and U1 are likely to be due to the limited amount of data at these identities. This is also true for the low and high similarity parts of the tRNA curve.