

# Abstract

Alignments of non-protein-coding RNAs (ncRNAs) have a wide range of applications: they are used for inference of phylogeny, for homology database searches and for finding new ncRNAs. Alignment quality is crucial for the successful application of these methods. Yet, aligning ncRNAs correctly is very difficult because paired regions evolve by compensatory base pair changes, i.e. mutual mutations which preserve base pairing but destroy sequence homology. An algorithm for the simultaneous alignment of ncRNA sequence and structure exists, but its computational complexity is exponential, making it virtually unemployable. Even simplified implementations are restricted to pairwise alignment only, due to their computational complexity. Thus sequence alignment programs are used for the alignment of ncRNAs.

In this work a benchmark of alignment programs upon ncRNAs should be performed. This benchmark including the respective database can be considered as an RNA counterpart of the protein specific database called BAliBASE. To make such a benchmark possible, appropriate accuracy measures are needed which display RNA alignment properties on sequence and structure level. Here, the measures SPS (“Sum-of-Pairs-Score”) and SCI (“Structure Conservation Index”) were used, which complement each other perfectly. Furthermore, test sets including reference solutions needed to be constructed, which vary systematically in their properties (sequence number and sequence homology), thus making it possible to quantify the effect of these properties on the programs. The initially planned approach to compile these test sets by means of CONSTRUCT was discarded as this approach would take an unreasonably long time. Instead, two different approaches were employed using large, reliable alignments of the Rfam-Database (“RNA family Database”).

In a collaboration the first systematic benchmark of alignment programs upon ncRNA sequences was carried out. On the basis of this benchmark it was possible to optimize program parameters for the RNA alignment problem. This was done for the programs MAFFT, MUSCLE and STRAL for instance. This first benchmark was complemented by a second one, which used up-to-date program versions, improved test sets and statistical rank tests. By means of these two data sets and the applied quality rating system, an objective evaluation of alignment programs was possible.

Amongst other things it was demonstrated that the so called “Twilight Zone” – the homology threshold below which alignment quality drops drastically – is at 55% sequence homology, compared to 20% for proteins. Above 75% sequence homology all programs perform equally well. Further on, it was shown that iterative alignment methods perform clearly better than non-iterative methods, particularly if sequences are divergent and the number of sequences rises. The performance of the program MAFFT (option “ginsi”) was statistically better than that of all other tested programs.