As the importance of non-coding RNAs becomes more evident, the need of computational methods for analyzing ncRNAs grows. Predicting the secondary structure is of great importance, and combining this with multiple alignment yields a useful tool for researchers. The exact solution to the problem of simultaneous multiple alignment and structure prediction for RNA sequences was described by Sankoff in 1985, but to date only limited pairwise implementations (e.g. Foldalign, Dynalign) or heuristics for multiple sequences (e.g. PMmulti, FoldalignM) exist. We present a novel approach to the simultaneous multiple alignment and structure prediction problem. Using Markov chain Monte Carlo in a simulated annealing framework, we sample multiple alignments and secondary structures. The sampling is based on a scoring system that combines a sequence measure with a structure measure: The sequence alignment is scored using the log-likelihood of the alignment, and the structure is scored using both the basepair probabilities and a covariation term. The sampling procedure itself uses simple local moves to optimize the alignment. These moves either act on the sequence alignment or the predicted structure. The program can initialize the sampling procedure using either unaligned sequences or an alignment obtained using e.g. Clustal or a partial alignment using trusted data from e.g. Rfam. The structure can be constrained by indicating e.g. specific basepairs or unpaired positions in one of the sequences. We compare the results of the developed algorithm with other programs and on various RNA datasets.

Stinus Lindgreen
University of Copenhagen
Bioinformatics Centre
stinus@binf.ku.dk