BMC Bioinformatics

Poster presentation

Multiple alignment and structure prediction of non-coding RNA sequences

Stinus Lindgreen*[1], Paul P Gardner[2] and Anders Krogh[1]

Address: [1]Bioinformatics Centre, Institute of Molecular Biology, University of Copenhagen, Denmark and [2]Wellcome Trust Sanger Institute, Cambridge, UK

Email: Stinus Lindgreen* - stinus@binf.ku.dk

* Corresponding author

from Third International Society for Computational Biology (ISCB) Student Council Symposium at the Fifteenth Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)
Vienna, Austria. 21 July 2007

Published: 20 November 2007

BMC Bioinformatics 2007, **8**(Suppl 8):P8 doi:10.1186/1471-2105-8-S8-P8

This abstract is available from: http://www.biomedcentral.com/1471-2105/8/S8/P8

Background

As the importance of non-coding RNAs becomes more evident, the need for computational methods for 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 [1], but to date only pairwise implementations (e.g. Foldalign [2], Dynalign [3]) or heuristics for multiple sequences (e.g. FoldalignM [4], LocARNA [5], RNA Sampler [6]) exist.

Methods

We present a novel approach to the 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, and the structure is scored using basepair probabilities and a covariation term. The sampling procedure itself uses simple local moves to optimize the solution. These moves either act on the sequence alignment or the predicted structure. The input to the program can be unaligned sequences or an alignment obtained using e.g. Clustal. The structure can be constrained by indicating e.g. basepairs or unpaired positions in one of the sequences. The program *MASTR* (*M*ultiple *A*lignment of *ST*ructural *R*NAs) is implemented in C++.

Results

MASTR is compared to LocARNA, FoldalignM, RNA Sampler and Clustal+RNAalifold on various RNA families. The datasets are unaligned and of varying average pairwise identities ranging from 30% to 100%. The sequence alignments are consistently better than or comparable to all other methods, the running time is significantly faster than FoldalignM and RNA Sampler, and MASTR can handle larger datasets than both these programs. RNA Sampler is best on datasets with identities between 30% and 45%, but MASTR is better than all other programs tested from 45% ID up to 80% ID, where the structure predictions deteriorate due to the poor covariation signal.

References

1. Sankoff D: **Simultaneous solution of the RNA folding, alignment, and protosequence problems.** *SIAM J Appl Math* 1985, **45:**810-825.
2. Havgaard JH, Lyngsø RB, Stormo GD, Gorodkin J: **Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%.** *Bioinformatics* 2005, **21(9):**1815-1824.
3. Mathews DH, Turner DH: **Dynalign: an algorithm for finding the secondary structure common to two RNA sequences.** *J Mol Biol* 2002, **317(2):**191-203.
4. Torarinsson E, Havgaard JH, Gorodkin J: **Multiple structural alignment and clustering of RNA sequences.** *Bioinformatics* 2007, **23(8):**926-932.
5. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R: **Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering.** *PLoS Comput Biol* 2007, **3(4):**e65.
6. Xu X, Ji Y, Stormo GD: **RNA Sampler: A new sampling based algorithm for common RNA secondary structure prediction and structural alignment.** *Bioinformatics* 2007 in press.