Scalable structural clustering of local RNA secondary structures

Steffen Heyne, Fabrizio Costa, Dominic Rose and Rolf Backofen

Bioinformatics Group, Department of Computer Science, University of Freiburg,Georges-Köhler-Allee 106, D-79110 Freiburg, Germany.

1 Introduction

It is increasingly evident that the eukaryotic genome is pervasively transcribed [4] and that non-protein coding RNAs (ncRNAs) play a key role in regulating gene expression and cell biology [3], affecting for instance (post-)transcriptional regulation, chromatin-remodelling, differentiation and development. In all kingdoms of life, genome-wide analysis are currently identifying a large numbers of potential ncRNAs (up to 450.000 only in the human genome [11]) and therefore the prediction, comparison, and functional annotation of ncRNAs are major tasks of current RNA research. Such studies identify genomic loci that are under stabilizing selection and that exhibit thermodynamically stable secondary structures and that therefore constitute prime candidates for novel functional ncRNAs. The functional annotation task is however a complex endeavor since, in contrast to protein-coding genes, ncRNAs belong to a large number of diverse classes with vastly different structures, functions, and evolutionary patterns [2].

The current classification system divides ncRNAs into (a) *families* according to functional, structural, or compositional similarities (the Rfam database lists as of today more than 2000 RNA families [6]), and (b) into *RNA classes* that group ncRNAs whose members have no discernible homology at the sequence level, but that have common structural and functional properties (e.g.snoRNAs and micro RNAs). For these reasons clustering according to sequence-structure similarity is nowadays the de-facto standard for ncRNA annotation. The quality and computational complexity of the clustering procedure depend on the underlying pairwise sequence comparison method. The most generic methods (LocARNA [17] and FOLDALIGN [16]) use derivatives of the full Sankoff algorithm [13] of simultaneous alignment and folding, but can be used only on small sets given their complexity (at least $O(n^4)$).

In order to achieve a reasonable trade-off between time and quality, many approaches use different heuristics: (a) using simplifications in the structural model, or (b) using sequence information as prior knowledge to speed up the computation. In the first case ([12] and [9]) one predicts structures for each individual sequence (a task that is known to be error prone). In the second case sequences are first clustered by sequence-alignment [15,10] and then conserved consensus structures are predicted (using RNAALIFOLD [1] or PETFOLD [14] for example). The major problem here is that ncRNA sequences evolve much faster than their structure, to the extent that often no homology on the sequence level is detectable (family assignments of sequence alignments at pairwise sequence identities below 60% are often wrong [7]).

DOI: 10.15455/CMSR.2014.0006

2 Contribution

Here, we propose an alignment-free approach for clustering RNA sequences according to sequence and structure information. We extend a fast graph kernel technique that we have developed [5] for chemoinformatics applications and we adapt it to detect similarities between RNA secondary structures. The key novelties are twofold: (1) we represent multiple folding hypothesis associated to a single RNA sequence in a flexible graph format; and (2) we efficiently convert the graph encoding into a very high dimensional sparse vectors. The first strategy allows us to compensate the inaccuracies of the minimum free energy solution. The second strategy allows us to use locality sensitive hashing methods to identify clusters with a complexity that is linear in the number of sequences N, i.e. avoiding the quadratic complexity arising from pairwise similarity computations.

We have integrated the approach in a ready-to-use pipeline for large-scale clustering of putative ncRNA. The method has been evaluated on known ncRNA classes and compared against existing approaches such as LocARNAand RNASOUP [8]. We show that not only we obtain clusters of high quality, but also we achieve striking speedups: from years to days for serial computation, down to hours when considering the parallel implementation.

We applied our method to six heterogeneous large-scale data sets containing more than 220,000 sequence fragments in total. We have analyzed predicted short ncRNAs which were lacking reliable class assignments and we have searched for local structural elements specific to experimentally validated lincRNAs. In this latter case we found enriched GO-terms for lincRNAs containing predicted local motifs that suggest a connection to vital processes of the human nervous system.

References

1. Stephan H. Bernhart, Ivo L. Hofacker, Sebastian Will, Andreas R. Gruber, and Peter F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.
2. Bompfünewerer Consortium: Backofen, R., Stephan H. Bernhart, Christoph Flamm, Claudia Fried, Guido Fritzsch, Jorg Hackermuller, Jana Hertel, Ivo L. Hofacker, Kristin Missal, Axel Mosig, Sonja J. Prohaska, Dominic Rose, Peter F. Stadler, Andrea Tanzer, Stefan Washietl, and Sebastian Will. RNAs everywhere: genome-wide annotation of structured RNAs. *J Exp Zoolog B Mol Dev Evol*, 308(1):1–25, 2007.
3. Christopher A. Brosnan and Olivier Voinnet. The long and the short of noncoding RNAs. *Current Opinion in Cell Biology*, 21(3):416–25, 2009.
4. Michael B. Clark, Paulo P. Amaral, Felix J. Schlesinger, Marcel E. Dinger, Ryan J. Taft, John L. Rinn, Chris P. Ponting, Peter F. Stadler, Kevin V. Morris, Antonin Morillon, Joel S. Rozowsky, Mark B. Gerstein, Claes Wahlestedt, Yoshihide Hayashizaki, Piero Carninci, Thomas R. Gingeras, and John S. Mattick. The reality of pervasive transcription. *PLoS Biol*, 9(7):e1000625; discussion e1001102, 2011.
5. Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings ICML*, 2010.
6. Paul P. Gardner, Jennifer Daub, John Tate, Benjamin L. Moore, Isabelle H. Osuch, Sam Griffiths-Jones, Robert D. Finn, Eric P. Nawrocki, Diana L. Kolbe, Sean R. Eddy, and Alex

Bateman. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Research*, 39(Database issue):D141–5, 2011.

7. Paul P. Gardner, Andreas Wilm, and Stefan Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, 33(8):2433–9, 2005.
8. Bogumil Kaczkowski, Elfar Torarinsson, Kristin Reiche, Jakob Hull Havgaard, Peter F. Stadler, and Jan Gorodkin. Structural profiles of human miRNA families from pairwise clustering. *Bioinformatics*, 25(3):291–4, 2009.
9. Mugdha Khaladkar, Vivian Bellofatto, Jason T. L. Wang, Bin Tian, and Bruce A. Shapiro. RADAR: a web server for RNA data analysis and research. *Nucleic Acids Research*, 35(Web Server issue):W300–4, 2007.
10. Victor Kunin, Rotem Sorek, and Philip Hugenholtz. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol*, 8(4):R61, 2007.
11. Mathieu Rederstorff, Stephan H. Bernhart, Andrea Tanzer, Marek Zywicki, Katrin Perfler, Melanie Lukasser, Ivo L. Hofacker, and Alexander Huttenhofer. RNPomics: defining the ncRNA transcriptome by cDNA library generation from ribonucleo-protein particles. *Nucleic Acids Research*, 38(10):e113, 2010.
12. William Ritchie, Matthieu Legendre, and Daniel Gautheret. RNA stem-loops: to be or not to be cleaved by RNAse III. *RNA*, 13(4):457–62, 2007.
13. David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, 45(5):810–825, 1985.
14. Stefan E. Seemann, Jan Gorodkin, and Rolf Backofen. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Research*, 36(20):6355–62, 2008.
15. Yanmei Shi, Gene W. Tyson, and Edward F. DeLong. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature*, 459(7244):266–9, 2009.
16. Elfar Torarinsson, Jakob H. Havgaard, and Jan Gorodkin. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, 23(8):926–32, 2007.
17. Sebastian Will, Kristin Reiche, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology*, 3(4):e65, 2007.