COMPUTATIONAL STUDIES OF NON-CODING RNAS

ROLF BACKOFEN
Institute of Computer Science, Albert-Ludwigs-University Freiburg, Germany
backofen@informatik.uni-freiburg.de

HAMIDREZA CHITSAZ
School of Computing Science, Simon Fraser University, Canada
hrc4@cs.sfu.ca

IVO HOFACKER
Institute for Theoretical Chemistry, University of Vienna, Austria
ivo@tbi.univie.ac.at

S. CENK SAHINALP
School of Computing Science, Simon Fraser University, Canada
cenk@cs.sfu.ca

PETER F. STADLER
University of Leipzig, Germany
studla@bioinf.uni-leipzig.de

1. Introduction

Until recently, RNA has been viewed as a simple "working copy" of the genomic DNA, simply transporting information from the genome into the proteins. In the 1980s, this picture changed, to certain extent, with the discovery of ribozymes and the realization that the ribosome is essentially an "RNA machine". Since the turn of the millenium, however, RNA has moved from a fringe topic to a central research topic following the discovery of RNA interference (RNAi), the post transcriptional silencing of gene expression via interactions between mRNAs and their regulatory RNAs.

More recent studies[1,2] have revealed that a large fraction of the genome sequences give rise to RNA transcripts that do not code for proteins. Those RNAs that do not code for proteins are called non-coding RNAs (ncRNAs).

A recent computational screen estimated the number of small regulatory RNAs, which form an important class of non-coding RNAs, in *Arabidopsis thaliana* to be in the order of 75,000.[3] Among small RNAs, two subclasses form the bulk of all regulatory RNAs: microRNAs (miRNAs) and small interfering RNAs (siRNAs) — which are of similar length (21 to 25 nt) and composition but different by origin. It is predicted that these two subclasses regulate at least one-third of all human genes. There are many other classes of non-coding RNAs with functionalities beyond simple regulation of gene expression: examples include snoRNAs, snRNAs, gRNAs, and stRNAs, which respectively perform ribosomal RNA (rRNA) modification, RNA editing, mRNA splicing and developmental regulation.[4] Even for these well-studied RNAs, their precise mode of function remains poorly understood.

In addition to such endogenous ncRNAs, antisense oligonucleotides have been used as exogenous inhibitors of gene expression; antisense technology is now commonly used for therapeutic purposes and as a research tool. The therapeutic objective of antisense technology is to block the production of disease-causing proteins. In principle, these artificial regulatory RNA molecules could be employed as drugs for the treatment of a variety of human diseases including various types of cancer, rheumatoid arthritis, brain diseases, and viral infections.[5] As a research tool, antisense nucleic acids may be used to study metabolic networks by controling or interfering with the dynamics and function of various modules in the network. Furthermore,

synthetic nucleic acid systems have been engineered to self-assemble into complex structures performing various dynamic mechanical motions.[6] Despite advances in computational studies of non-coding RNA, there are still many open areas and unresolved issues particularly for high-throughput applications based on the new genome sequencing technologies.

The main objective of this session is to discuss new algorithms, software tools and their applications in non-coding RNA bioinformatics. In particular, the papers in this session exemplify recent progress in computational methods that help non-coding RNA sequence prediction and identification, structure prediction and determination, and function determination. Specific problems in computational studies of ncRNAs include:

- Algorithms for modelling interactions between RNAs and other molecules, particularly RNA-RNA and RNA-protein interactions
- Learning thermodynamic parameters that are involved in the prediction of secondary and tertiary structure of non-coding RNAs
- Novel approaches to single or joint RNA structure prediction and determination
- Algorithms for exploring RNA folding pathways and kinetic traps on the energy landscape
- Alignment and comparative analysis of multiple non-coding RNA sequences
- RNA evolution
- Functional classification of non-coding RNAs
- Modelling classes of single or joint non-coding RNAs through stochastic context free grammars and their variations
- Tools that detect structural motifs in a genome sequence, especially those that could be potentially involved in the regulation of target mRNAs
- Combinatorial and heuristic tools for de novo non-coding RNA identification in a genome sequence
- Efficient algorithms for searching RNAs in a data collection with specific sequence and structural motifs.

2. Session papers

Recent improvements in sequencing methods have introduced high-throughput, low-cost, and cloning-free (thus less labor-intensive) technologies. The revolution in DNA sequencing will shortly result in an enormous collection of sequence data pertaining to the genomes and transcriptomes of various human individuals from different populations and also various species. Several papers in this session try to address the increased demand for this type of data analysis.

RNA secondary structure and folding kinetics have always been a central research topic in computational studies of RNA. In the first paper of this session, Thachuk *et al.* make two new contributions to the problem of calculating pseudoknot-free folding pathways with minimum energy barrier between pairs (A, B) of RNA secondary structures. Their first contribution is an exact algorithm to find a minimum barrier direct folding pathway for a simple energy model in which each base pair contributes equally to the structures stability. In a direct (minimum length) folding pathway, intermediate structures contain only base pairs in A and B and are of length $|A| + |B|$. The problem is NP-hard, therefore their algorithm requires exponential time in the worst case. Their second contribution proves that for the simple energy model, repeatedly adding or removing a base pair from A or B along a pathway does not lower the energy barrier.

Dotu *et al.* describe dynamic programming segmentation algorithms to segment RNA secondary and tertiary structures into distinct domains in the second paper. A possible application is to determine the boundaries of predicted ncRNAs. Under the assumption that microRNA precursors are less than 100nt long, their method predicts the precursors embedded in a genomic context of up to 1000nt with an accuracy around 90%. They also compare their algorithm to the manual segmentation of 16S rRNAs reported in the literature.

As a solution to a fundamental problem in computational studies of ncRNAs, `RNAz` has been used for de novo prediction of structured non-coding RNAs in comparative genomics data. In the third paper, `RNAz`

2.0 is presented which improves the previous version in several aspects: It uses a dinucleotide background models to increase accuracy, and an entropy measure to represent sequence similarities. In addition, it has been trained on a larger data set, using either sequence-based or structural alignments. As a result `RNAz` 2.0 has a significantly lower false positive rate than the previous version.

Recent advances in high-throughput sequencing for the first time provide the opportunity to study the entire transcriptome sequences and concentrations. In many cases, a significant part of the transcriptome consists of non-coding RNAs. Some of these ncRNAs are processed by the cell post-transcriptional machinery to yield shorter RNA products. It is believed that these splicing patterns depend on the secondary structure. This leads to specific patterns of short reads that can be detected after mapping the read sequences to the reference genome. In the fourth paper, Langenberger et al. suggest that these read patterns are characteristic for the spliced RNA transcripts. Therefore, Langenberger *et al.* explore the potential of short read sequence data in the classification and identification of non-coding RNAs.

Okada *et al.* report on an improved version of the Structure Conservation Index, used in RNAz. based on centroid estimators rather than minimum free energy structures. Poolsap *et al.* present a dynamic programming approach to compute RNA-RNA interactions for the case where the sequences motifs involved in the binding are known in advance.

Acknowledgements

We thank all the authors who submitted papers to the session, and we gratefully acknowledge the reviewers who contributed their time and expertise to the peer review process.

References

1. The FANTOM Consortium, *Science* **309**, 1159 (2005).
2. ENCODE Project Consortium, *Nature* **447**, 799 (2007).
3. M. W. Vaughn and R. Martienssen, *Science* **309**, 1525 (2005).
4. R. F. Gesteland, T. R. Cech and J. F. Atkins (eds.), *The RNA World*, 3rd edn. (Cold Spring Harbor Laboratory Press, Plainview, NY, 2006).
5. A. Fjose and O. Drivenes, *Birth Defects Res C Embryo Today* **78**, 150 (2006).
6. P. Guo, *J Nanosci Nanotechnol* **5**, 1964 (2005).