

# Special Focus Bioinformatics

Guest Editors: Eric P. Nawrocki<sup>1</sup> and Sarah W. Burge<sup>2</sup>

<sup>1</sup>Eddy Lab; HHMI Janelia Farm Research Campus; Ashburn, VA USA; <sup>2</sup>Sanger Institute; Cambridge, UK

The development of RNA bioinformatic tools began more than 30 y ago with the description of the Nussinov and Zuker dynamic programming algorithms for single sequence RNA secondary structure prediction. Since then, many tools have been developed for various RNA sequence analysis problems such as homology search, multiple sequence alignment, de novo RNA discovery, read-mapping, and many more. In this issue, we have collected a sampling of reviews and original research that demonstrate some of the many ways bioinformatics is integrated with current RNA biology research.

The predicted structure of an RNA can sometimes offer functional insight, and similar predicted structures for different sequences can hint at homology and/or functional similarity. Rivas provides an overview of many different methods and tools for RNA structure prediction. She offers a unifying perspective by identifying four common ingredients to all of the approaches and discusses future directions in this important area.

The computational identification of RNAs by homology in sequence data sets is another longstanding problem. Pairwise sequence similarity-based BLAST is commonly used to find some well-conserved structural RNAs in newly published genomes. More powerful methods for RNA homology search based on probabilistic models of the conserved sequence and secondary structure of RNA families called covariance models (CMs) have been around for nearly 20 y but until recently they have been too computationally expensive for routine, practical use. Nawrocki and Eddy detail the use of the CM-based Infernal software package and the Rfam database of RNA families for annotating known noncoding RNAs in metagenomic data sets. The construction of CMs from hand-curated alignments can be helpful in identifying novel or divergent classes of noncoding RNAs, as exemplified by Hafez et al., who report here the discovery of mitochondrion-encoded circular permuted tmRNA genes in oomycetes (water mold). The authors provide a training alignment and corresponding CM that will be added to the Rfam database for use by the wider RNA sequence analysis community.

A goal of homology search tools is to enable researchers to identify how phylogenetically widespread or narrow an RNA family is. Deiorio-Haggar et al. (p. 1180–4) describe several regulatory RNAs of ribosomal proteins found exclusively in *Bacilli* bacteria. Their findings suggest that many other regulatory RNA elements specific to other classes of bacteria remain undiscovered.

The rapid decrease in sequencing costs over the past two decades has led to an explosion in the amount of sequence data in publicly available databases. The new data has both

underscored the importance of rigorous and efficient methods for sequence analysis such as RNA structure prediction and homology search and has also created the need for new classes of tools, such as read-mapping programs for ultra-rapid mapping of sequence reads to their reference genomes. RNA-seq experiments apply next-generation sequencing to a sample of an organism's RNA and utilize read-mappers to quantify abundance of RNAs and alternatively spliced transcripts and reveal RNA editing. We include here a paper by Toffano-Nioche et al. describing an RNA-seq experiment and bioinformatic analysis of the transcriptome of the hyperthermophilic archaeon *Pyrococcus abyssi*. The authors combine data from previous studies, promoter analysis, and comparative analysis with other archaeal species to identify putative ORFs, noncoding RNAs, and UTRs of coding regions.

Some RNA-seq analyses specifically target the non-coding RNA repertoire of an organism. Mohorianu et al. introduce a novel bioinformatic tool for analyzing data from such experiments called CoLide with the aim of identifying small noncoding RNAs. CoLide (Co-expression based sRNA Loci Identification) incorporates expression level, genomic location, and size distribution in its predictions.

Most RNA-seq analyses discard reads that don't map directly back to the genome in one uninterrupted interval. Doose et al. reexamine such discarded reads from several published bacterial and/or archaeal RNA-seq data sets. The authors screened for noncoding RNAs that have circularized products, which are not expected to directly map to the genome. The screen revealed some previously known examples, including self-splicing introns, but also revealed several interesting novel candidates.

Other investigations of an organism's transcriptome utilize microarrays to gauge RNA expression. This issue includes one example by Petazzi et al., who provide an analysis of the long non-coding RNA transcriptome of a mouse model of Rett syndrome. Finally, Barquist et al. review methods for transposon-insertion sequencing, which takes advantage of high-throughput sequencing to identify functional genomic regions in bacteria through monitoring of a library of random transposon insertion mutants.

The collection of papers in this special issue of *RNA Biology* demonstrates that bioinformatics is integral to the many types of experiments that help us understand non-coding RNAs. As experimental techniques and technologies become increasingly high-throughput, biological data sets will continue to grow and so will the need for computational approaches toward their analyses.