Computational Identification and Annotation of non-coding RNAs - Summary

Within recent years compelling evidence has emerged that the majority of non-protein-coding transcripts of metazoan genomes comprise a "new" functional class of regulatory molecules, namely non-coding RNAs (ncRNAs). By generating RNA-protein, RNA-RNA or RNA-DNA complexes they control a variety of cellular processes including, for example, the control of messenger RNA splicing, transcription initiation, translation inhibition, or chromosome maintenance. While a wealth of annotation exists for protein-coding genes and their putative function, annotation of ncRNA genes has been almost non-existent but has recently become a topic of utmost interest.

One topic is the computational identification of novel ncRNA genes. Unlike protein-coding genes, ncRNA genes do not exhibit a strong common statistical signal in their sequence. However, most ncRNA families do depend on a well-defined secondary structure as a consequence of evolutionary selection acting predominantly on the secondary structure of an ncRNA molecule in order to preserve their function. Therefore, current promising approaches for de novo predictions of ncRNA genes are based on comparative approaches. `RNAz`, for example, assesses the conservation and stability of the secondary structure the sequences of a multiple sequence alignment fold into. `RNAz` has been applied to surveying the human genome and provided evidence for tens of thousands of genomic loci with evolutionary conserved secondary structures.

This thesis describes two independent computational surveys for structured ncRNAs in urochordates and nematodes utilizing `RNAz`. Urochordates are a sister group of vertebrates, which do not share the whole-genome duplications observed in vertebrates. Nevertheless, vertebrates show considerable conservation in morphology and gene function during early embryogenesis with urochordates. The nematodes, on the other hand, are representatives of protostomes following a different embryogenesis. Both genome-wide computational screens for ncRNAs contribute to the current knowledge of ncRNA evolution in bilaterian genomes and their expansion in mammals.

These investigations have produced extensive lists of candidates for functional ncRNA, but only a minority could be assigned to known ncRNAs in related organisms by sequence similarity searches. We thus present an alignment method, namely (`m`)`LocARNA`, which detects homologous secondary structure motifs in a set of sequences and define a distance measure based on the alignment score in order to detect clusters of predicted ncRNA loci sharing structural homology. The performance of the structural clustering approach was evaluated by clustering a comprehensive set of known ncRNA sequences. We further applied the clustering approach to the data set of predicted ncRNAs in urochordates and nematodes. In some cases we find that additional sequences are identified as structural relatives of known RNA families, but also predict several candidates for novel ncRNA classes.

A prerequisite for functional annotation of ncRNA candidates is to determine their reading direction with high precision. While folding energies of an RNA and its reverse complement are similar, the differences are sufficient at least in conjunction with substitution patterns to discriminate between structured RNAs and their complements. We present `RNAstrand`, which reliably classifies the reading direction of a structured RNA from a multiple sequence alignment and provides a considerable improvement in classification accuracy over previous approaches.

In summary, this thesis describes three contributions to the identification and annotation of ncRNA genes in eukaryotic genomes. Genome-wide computational screens in urochordates and nematodes provide an insight into the evolution of ncRNAs in Bilateria. A structural clustering approach is presented, which assigns ncRNA candidates to known families or, more interestingly, obtains novel ncRNA families. Lastly, the reliable prediction of the reading direction of an ncRNA candidate is achieved by the `RNAstrand` software tool.