Dominic Rose

The long and the short of computational ncRNA prediction

Abstract

Non-coding RNAs (ncRNAs) are transcripts that function directly as RNA molecule without ever being translated to protein.

The transcriptional output of eukaryotic cells is diverse, pervasive, and multi-layered. It consists of spliced as well as unspliced transcripts of both protein-coding messenger RNAs and functional ncRNAs. However, it also contains degradable non-functional by-products and artefacts - certainly a reason why ncRNAs have long been wrongly disposed as transcriptional noise.

Today, RNA-controlled regulatory processes are broadly recognized for a variety of ncRNA classes. The thermoresponsive ROSE ncRNA (repression of heat shock gene expression) is only one example of a regulatory ncRNA acting at the post-transcriptional level via conformational changes of its secondary structure.

Bioinformatics helps to identify novel ncRNAs in the bulk of genomic and transcriptomic sequence data which are produced at ever increasing rates. However, ncRNA annotation is unfortunately not part of generic genome annotation pipelines. Dedicated computational searches for particular ncRNAs are veritable research projects in their own right. Despite best efforts, ncRNAs across the animal phylogeny remain to a large extent uncharted territory.

This thesis describes a comprehensive collection of exploratory bioinformatic field studies designed to de novo predict ncRNA genes in a series of computational screens and in a multitude of newly sequenced genomes. Non-coding RNAs can be divided into subclasses (families) according to peculiar functional, structural, or compositional similarities. A simple but eligible and frequently applied criterion to classify RNA species is length. In line, the thesis is structured into two parts: We present a series of pilot-studies investigating (1) the short and (2) the long ncRNA repertoire of several model species by means of state-of-the-art bioinformatic techniques.

In the first part of the thesis, we focus on the detection of short ncRNAs exhibiting thermodynamically stable and evolutionary conserved secondary structures. We provide evidence for the presence of short structured ncRNAs in a variety of different species, ranging from bacteria to insects and higher eukaryotes. In particular, we highlight drawbacks and opportunities of RNAz-based ncRNA prediction at several hitherto scarcely investigated scenarios, as for example ncRNA prediction in the light of whole genome duplications. A recent microarray study provides experimental evidence for our approach. Differential expression of at least one-sixth of our drosophilid RNAz predictions has been reported. Beyond the means of RNAz, we moreover manually compile sophisticated annotation of short ncRNAs in schistosomes. Obviously, accumulating knowledge about the genetic material of malaria causing parasites which infect millions of humans world-wide is of utmost scientific interest.

Since the performance of any comparative genomics approach is limited by the quality of its input alignments, we introduce a novel light-weight and performant genome-wide alignment approach: NcDNAlign. Although the tool is optimized for speed rather than sensitivity and requires only a minor fraction of CPU time compared to existing programs, we demonstrate that it is basically as sensitive and specific as competing approaches when applied to genome-wide ncRNA gene finding and analysis of ultra-conserved regions.

By design, however, prediction approaches that search for regions with an excess of mutations that maintain secondary structure motifs will miss ncRNAs that are unstructured or whose structure is not well conserved in evolution.

In the second part of the thesis, we therefore overcome secondary structure prediction and, based on splice site detection, develop novel strategies specifically designed to identify long ncRNAs in genomic sequences - probably the open problem in current RNA research. We perform splice site anchored gene-finding in drosophilids, nematodes, and vertebrate genomes and, at least for a subset of obtained candidate genes, provide experimental evidence for expression and the existence of novel spliced transcripts undoubtedly confirming our approach.

In summary, we found evidence for a large number of previously undescribed RNAs which consolidates the idea of non-coding RNAs as an abundant class of regulatory active transcripts. Certainly, ncRNA prediction is a complex task. This thesis, however, rationally advises how to unveil the RNA complement of newly sequenced genomes. Since our results have already established both subsequent computational as well as experimental studies, we believe to have enduringly stimulated the field of RNA research and to have contributed to an enriched view on the subject.