Sven Findeiß

Expanding the repertoire of bacterial (non-)coding RNAs

The detection of non-protein-coding RNA (ncRNA) genes in bacteria and their diverse regulatory mode of action moved the experimental and bio-computational analysis of ncRNAs into the focus of attention. Regulatory ncRNA transcripts are not translated to proteins but function directly on the RNA level. These typically small RNAs have been found to be involved in diverse processes such as (post-)transcriptional regulation and modification, translation, protein translocation, protein degradation and sequestration.

Bacterial ncRNAs either arise from independent primary transcripts or their mature sequence is generated via processing from a precursor. Besides these autonomous transcripts, RNA regulators (e.g. riboswitches and RNA thermometers) also form chimera with protein-coding sequences. These structured regulatory elements are encoded within the messenger RNA and directly regulate the expression of their "host" gene.

The quality and completeness of genome annotation is essential for all subsequent analyses. In contrast to protein-coding genes ncRNAs lack clear statistical signals on the sequence level. Thus, sophisticated tools have been developed to automatically identify ncRNA genes. Unfortunately, these tools are not part of generic genome annotation pipelines and therefore computational searches for known ncRNA genes are the starting point of each study. Moreover, prokaryotic genome annotation lacks essential features of protein-coding genes. Many known ncRNAs regulate translation via base-pairing to the 5′ UTR (untranslated region) of mRNA transcripts. Eukaryotic 5′ UTRs have been routinely annotated by sequencing of ESTs (expressed sequence tags) for more than a decade. Only recently, experimental setups have been developed to systematically identify these elements on a genome-wide scale in prokaryotes.

The first part of this thesis, describes three experimental surveys of exploratory field studies to analyze transcript organization in pathogenic bacteria. To identify ncRNAs in Pseudomonas aeruginosa we used a combination of an experimental RNomics approach and ncRNA prediction. Besides already known ncRNAs we identified and validated the expression of six novel RNA genes.

Global detection of transcripts by next generation RNA sequencing techniques unraveled an unexpectedly complex transcript organization in many bacteria. These ultra high-throughput methods give us the appealing opportunity to analyze the complete RNA output of any species at once. The development of the differential RNA sequencing (dRNA-seq) approach enabled us to analyze the primary transcriptome of Helicobacter pylori and Xanthomonas campestris. For the first time we generated a comprehensive and precise transcription start site (TSS) map for both species and provide a general framework for the analysis of dRNA-seq data. Focusing on computer-aided analysis we developed new tools to annotate TSS, detect small protein-coding genes and to infer homology of newly detected transcripts. We discovered hundreds of TSS in intergenic regions, upstream of protein-coding genes, within operons and antisense to annotated genes. Analysis of 5′ UTRs (spanning from the TSS to the start codon of the adjacent protein-coding gene) revealed an unexpected size diversity ranging from zero to several hundred nucleotides. We identified and validated the expression of about 60 and about 20 ncRNA candidates in Helicobacter and Xanthomonas, respectively. Among these ncRNA candidates we found several small protein-coding genes that have previously evaded annotation in both species. We showed that the combination of dRNA-seq and computational analysis is a powerful method to examine prokaryotic transcriptomes.

Experimental setups are time consuming and often combined with huge costs. Another limitation of experimental approaches is that genes which are expressed in specific developmental stages or stress conditions are likely to be missed. Bioinformatic tools build an alternative to overcome such restraints. General approaches usually depend on comparative genomic data and evolutionary signatures are used to analyze the (non-)coding potential of multiple sequence alignments. In the second part of my thesis we present our major update of the widely used ncRNA gene finder RNAz and introduce RNAcode, an efficient tool to asses local protein-coding potential of genomic regions.

RNAz has been successfully used to identify structured RNA elements in all domains of life. However, our own experience and the user feedback not only demonstrated the applicability of the RNAz approach, but also helped us to identify limitations of the current implementation. Using a much larger training set and a new classification model we significantly improved the prediction accuracy of RNAz.

During transcriptome analysis we repeatedly identified small protein-coding genes that have not been annotated so far. Only a few of those genes are known to date and standard proteincoding gene finding tools suffer from the lack of training data. To avoid an excess of false positive predictions, gene finding software is usually run with an arbitrary cutoff of 40-50 amino acids and therefore misses the small sized protein-coding genes. We have implemented RNAcode which is optimized for emerging applications not covered by standard protein-coding gene annotation software. In addition to complementing classical protein gene annotation, a major field of application of RNAcode is the functional classification of transcribed regions. RNA sequencing analyses are likely to falsely report transcript fragments (e.g. mRNA degradation products) as non-coding. Hence, an evaluation of the protein-coding potential of these fragments is an essential task. RNAcode reports local regions of high coding potential instead of complete protein-coding genes. A training on known protein-coding sequences is not necessary and RNAcode can therefore be applied to any species. We showed this with our analysis of the Escherichia coli genome where the current annotation could be accurately reproduced. We furthermore identified novel small protein-coding genes with RNAcode in this extensively studied genome. Using transcriptome and proteome data we found compelling evidence that several of the identified candidates are bona fide proteins.

In summary, this thesis clearly demonstrates that bioinformatic methods are mandatory to analyze the huge amount of transcriptome data and to identify novel (non-)coding RNA genes. With the major update of RNAz and the implementation of RNAcode we contributed to complete the repertoire of gene finding software which will help to unearth hidden treasures of the RNA World.