Insights into the Evolution of small nucleolar RNAs

Sebastian Canzler

Abstract

Over the last decades, the formerly irrevocable believe that proteins are the only key-factors in the complex regulatory machinery of a cell was crushed by a plethora of findings in all major eukaryotic lineages. These suggested a rugged landscape in the eukaryotic genome consisting of sequential, overlapping, or even bi-directional transcripts and myriads of regulatory elements. The vast part of the genome is indeed transcribed into an RNA intermediate, but solely a small fraction is finally translated into functional proteins. The sweeping majority, however, is either degraded or functions as a non-protein coding RNA (ncRNA).

Due to continuous developments in experimental and computational research, the variety of ncRNA classes grew larger and larger, ranging from key-processes in the cellular lifespan to regulatory processes that are driven and guided by ncRNAs. The bioinformatical part primarily concentrates on the prediction, annotation, and extraction of characteristic properties of novel ncRNAs. Due to conservation of sequence and/or structure, this task is often determined by an homology-search that utilizes information about functional, and hence conserved regions, as an indicator.

This thesis focuses mainly on a special class of ncRNAs, small nucleolar RNAs (snoRNAs). These abundant molecules are mainly responsible for the guidance of 2'-O-ribose-methylations and pseudouridylations in different types of RNAs, such as ribosomal and spliceosomal RNAs. Although the relevance of single modifications is still rather unclear, the elimination of a bunch of modifications is shown to cause severe effects, including lethality.

Several *de novo* prediction programs have been published over the last years and a substantial amount of publicly available snoRNA databases has originated. Normally, these are restricted to a small amount of species and a collection of experimentally extracted snoRNA. The detection of snoRNAs by means of wet lab experiments and/or *de novo* prediction tools is generally time consuming (wet lab) and a quite tedious task (identification of snoRNA-specific characteristics).

The snoRNA annotation pipeline `snoStrip` was developed with the intention to circumvent these obstacles. It therefore utilizes a homology-based search procedure to reliably predict snoRNA genes in genomic sequences. In a subsequent step, all candidates are filtered with respect to specific sequence motifs and secondary structures. In a functional analysis, potential target sites are predicted in ribosomal and spliceosomal RNA sequences. In contrast to *de novo* prediction tools, `snoStrip` focuses on the extension of the known snoRNA world to uncharted organisms and the mapping and unification of the existing diversity of snoRNAs into functional, homologous families.

The pipeline is properly suited to analyze a manifold set of organisms in search for their snoRNAome in short timescales. This offers the opportunity to generate large scale analyses over whole eukaryotic kingdoms to gain insights into the evolutionary history of these special ncRNA molecules. A set of experimentally validated snoRNA genes in Deuterostomia and Fungi were starting points for highly comprehensive surveys searching and analyzing the snoRNA repertoire in these two major eukaryotic clades. In both cases, the `snoStrip` pipeline proved itself as a fast and reliable tool and collected

thousands of snoRNA genes in nearly 200 organisms. Additionally, the Interaction Conservation Index (ICI), which is amplified to additionally work on single lineages, provides a convenient measure to analyze and evaluate the conservation of snoRNA-targetRNA interactions across different species. The massive amount of data and the possibility to score the conservation of predicted interactions constitute the main pillars to gain an extraordinary insight into the evolutionary history of snoRNAs on both the sequence and the functional level. A substantial part of the snoRNAome is traceable down to the root of both eukaryotic lineages and might indicate an even more ancient origin of these snoRNAs. However, a plenitude of lineage specific innovation and deletion events are also discernible. Due to its automated detection of homologous and functionally related snoRNA sequences, snoStrip identified extraordinary target switches in fungi. These unveiled a coupled evolutionary history of several snoRNA families that were previously thought to be independent. Although these findings are exceedingly interesting, the broad majority of snoRNA families is found to show remarkable conservation of the sequence and the predicted target interactions.

On two occasions, this thesis will shift its focus from a genuine snoRNA inspection to an analysis of introns. Both investigations, however, are still conducted under an evolutionary viewpoint. In case of the ubiquitously present U3 snoRNA, functional genes in a notable amount of fungi are found to be disrupted by U2-dependent introns. The set of previously known U3 genes is considerably enlarged by an adapted snoStrip-search procedure. Intron-disrupted genes are found in several fungal lineages, while their precise insertion points within the snoRNA-precursor are located in a small and homologous region. A potential targetRNA of snoRNA genes, U6 snRNA, is also found to contain intronic sequences. Within this work, U6 genes are detected and annotated in nearly all fungal organisms. Although a few U6 intron-carrying genes have been known before, the widespread of these findings and the diversity regarding the particular insertion points are surprising. Those U6 genes are commonly found to contain more than just one intron. In both cases of intron-disrupted non-coding RNA genes, the detected RNA molecules seem to be functional and the intronic sequences show remarkable sequence conservation for both their splice sites and the branch site.

In summary, the snoStrip pipeline is shown to be a reliable and fast prediction tool that works on homology-based search principles. Large scale analyses on whole eukaryotic lineages become feasible on short notice. Furthermore, the automated detection of functionally related but not yet mapped snoRNA families adds a new layer of information. Based on surveys covering the evolutionary history of Fungi and Deuterostomia, profound insights into the evolutionary history of this ncRNA class are revealed suggesting ancient origin for a main part of the snoRNAome. Lineage specific innovation and deletion events are also found to occur at a large number of distinct timepoints.