Abstract

Tracing the evolution of long non-coding RNAs - Principles of comparative transcriptomics for splice site conservation and biological applications

Anne Nitsche
(Dissertation, 2017)

Eukaryotic cells exhibit an extensive transcriptional diversity. Only about a quarter of the total RNA in the human cell can be accounted for by messenger RNA (mRNA), which convey genetic code for protein generation. The remaining part of the transcriptome consists of rather heterogenous molecules. While some classes are well defined and have been shown to carry out distinct functions, ranging from housekeeping to complex regulatory tasks, a big fraction of the transcriptional output is categorized solely based on the lack of protein-coding capacity and transcript length. Several studies have shown, that as a group, mRNA-like long non-coding RNAs (lncRNAs), are under stabilizing selection, however at much weaker levels than mRNAs. The conservation at the level of primary sequence is even lower, blurring the contrast between exonic and intronics parts, which impedes traditional methods of genome-wide homology search. As a consequence their evolutionary history is a fairly unexplored field and apart from a few experimentally studied cases, the vast majority of them is reported to be poorly conserved. However, the pervasive transcription and the highly spatio-temporal specific expression patterns of lncRNAs suggests their functional importance and makes their evolutionary age and conservation patterns a topic of interest. By employing diverse computational methods, recent studies shed light on the common conservation of lncRNA's secondary and gene structures, highlighting the significance of structural features on functionality. Splice sites, in particular, are frequently retained over very large evolutionary time scales, as they maintain the intron-exon-structure of the transcript. Consequently, the conservation of splice sites can be utilized in a comparative genomics approach to establish homology and predict evolutionarily well-conserved transcripts, regardless of their coding capacity. Since splice site conservation cannot be directly inferred from experimental evidence, in the course of this thesis a computational pipeline was established to generate comparative maps of splice sites based on multiple sequence alignments together with transcriptomics data. Scoring schemes for splice site motifs are employed to assess the conservation of orthologs. This resource can then be used to systemically study the conservation patterns of RNAs and their gene structures. This thesis will demonstrate the versatility of this method by showcasing biological applications of three distinct studies. First, a comprehensive annotation of the human transcriptome, from RefSeq, ESTs and GENCODE, was used to trace the evolution of human lncRNAs. A large majority of human lncRNAs is found to be conserved across Eutheria, and many hundreds originated before the divergence of marsupials and placental mammals. However, they exhibit a rapid turnover of their transcript structures, indicating that they are actual ancient components of the vertebrate genome with outstanding evolutionary plasticity. Additionally, a public web server was setup, which allows the user to retrieve sets of orthologous splice sites from pre-computed comparative splice site maps and inspect visualizations of their conservation in the respective species. Second, a more specific data set of non-colinearly spliced latimerian RNAs is studied to fathom the origins of atypical transcripts. RNA-seq data from two coelacanth species are analyzed, yielding thousands of circular and trans-spliced products, with a surprising exclusivity of

the majority of their splice junctions to atypically spliced forms, that is they are not used in linear isoforms. The conservation analysis with comparative splice site maps yielded high conservation levels for both cir- cularizing and trans-connecting splice sites. This fact in combination with their abundance strongly suggests that atypical RNAs are evolutionarily old and of functional importance. Lastly, comparative splice site maps are used to investigate the role of lncRNAs in the evolution of the Alzheimer's disease (AD). The human specificity of AD clearly points out a phylogenetic aspect of the disease, which makes the evolutionary analysis a very promising field of research. Protein-coding and non-protein-coding regions, that have been identified to be differentially expressed in AD patients, are analyzed for conservation of their splice site and evolution of their exon-intron-structure. Both non-coding and protein-coding AD-associated genes are shown to have evolved more rapidly in their gene structure than the genome at large. This supports the view of AD as a consequence of the recent rapid adaptive evolution of the human brain. This phylogenetic trait might have far reaching consequences with respect to the appropriateness of animal models and the development of disease-modifying strategies.

Keywords

splice sites, conservation, evolutionary plasticity, non-coding RNA, lncRNA