Journal of Biotechnology 261 (2017) 76–84

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

RNA-bioinformatics: Tools, services and databases for the analysis of RNA-based regulation

Rolf Backofen[a,b,*], Jan Engelhardt[c], Anika Erxleben[a], Jörg Fallmann[c], Björn Grüning[a], Uwe Ohler[d], Nikolaus Rajewsky[d], Peter F. Stadler[c,e,f,g]

[a] Bioinformatics, University of Freiburg, Georges-Köhler-Allee 106, D-79110 Freiburg, Germany
[b] BIOSS Centre for Biological Signaling Studies, University of Freiburg, Schänzlestr. 18, 79104 Freiburg, Germany
[c] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[d] Max-Delbrück-Centrum (MDC), Robert-Rössle-Str. 10, D-13092 Berlin, Germany
[e] Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
[f] RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany
[g] Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

ARTICLE INFO

Keywords:
RNA
de.NBI
Galaxy
Tools
Service

ABSTRACT

The importance of RNA-based regulation is becoming more and more evident. Genome-wide sequencing efforts have shown that the majority of the DNA in eukaryotic genomes is transcribed. Advanced high-throughput techniques like CLIP for the genome-wide detection of RNA–protein interactions have shown that post-transcriptional regulation by RNA-binding proteins matches the complexity of transcriptional regulation. The need for a specialized and integrated analysis of RNA-based data has led to the foundation of the RNA Bioinformatics Center (RBC) within the German Network of Bioinformatics Infrastructure (de.NBI). This paper describes the tools, services and databases provided by the RBC, and shows example applications. Furthermore, we have setup an RNA workbench within the Galaxy framework. For an easy dissemination, we offer a virtualized version of Galaxy (via Galaxy Docker) enabling other groups to use our RNA workbench in a very simple way.

1. Motivation

Genome-wide sequencing efforts have revealed that a majority of DNA in eukaryotic genomes is pervasively transcribed. Non-coding RNAs and RNA–protein interactions are important parts of cellular regulation that were ignored at first but have received an increasing level of attention over the past decade. While the exact numbers, and even the magnitude, of functional transcripts, regulators and interactions are a matter of ongoing discussion, they reflect the current challenge for the analysis of whole transcriptome data.

The identification of new classes of regulatory RNAs such as microRNAs (miRNAs), or the genome-wide identification of RNA–protein interactions, which has been enabled by the development of new technologies such as cross-linking and immunoprecipitation (CLIP) methods, suggests that the complexity of post-transcriptional gene regulation is comparable to transcriptional gene regulation. The human genome encodes hundreds to thousands of miRNAs more than 1000 RNA binding proteins (Medenbach et al., 2011; Baltz et al., 2012; Gerstberger et al., 2014; Brannan et al., 2016; He et al., 2016; Castello et al., 2016). Along with such profiling efforts, a picture has emerged that many human diseases are caused or linked to post-transcriptional gene regulation. Examples include not only rare genetic disorders but cover the entire spectrum of cardio-vascular diseases, cancer, and neurodegenerative disorders (for recent reviews see Kapeli and Yeo, 2012; Darnell and Richter, 2012; Kong and Lasko, 2012; Ibrahim et al., 2012; Rosina and Hurst, 2017; Schmiedel et al., 2016). With increasing evidence that non-coding RNAs are also involved in epigenetic regulatory control, it is clear that RNA biology is of vital, newly emerging importance for research not only in basic molecular biology but also for medical and disease research. Consequently, many of the existing or newly founded centres for common diseases have great need to develop or get access to computational tools and databases that capture and predict regulation by RNA or RNA–protein interactions.

2. Overview over the RNA Bioinformatics Center

Non-coding regulatory RNAs have many possible functions, which require specialized approaches for their detection and analysis (see

* Corresponding author.
E-mail address: backofen@informatik.uni-freiburg.de (R. Backofen).

http://dx.doi.org/10.1016/j.jbiotec.2017.05.019
Received 21 February 2017; Received in revised form 20 May 2017; Accepted 23 May 2017
Available online 26 May 2017

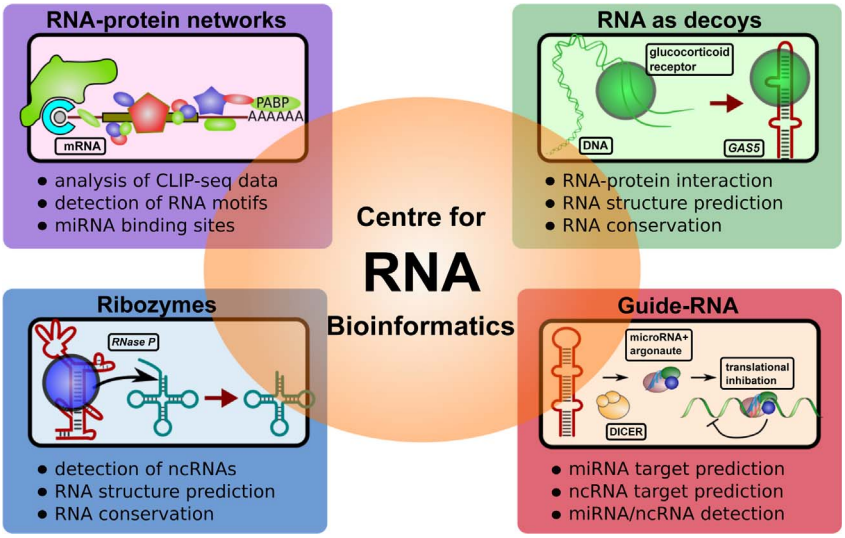

Fig. 1. Example functions of RNA and associated bioinformatics services. A comprehensive analysis and annotation of RNA function requires the integration of many different services that are provided by our centre.

Fig. 1 for an overview of functions and associated bioinformatics services). To name only a few, they can regulate imprinting by modulating chromatin structures, act as guide RNAs for protein complexes, form scaffolds for protein–RNA complexes, regulate other RNAs by RNA–RNA-interaction (Busch et al., 2008; Mückstein et al., 2006), function as decoys for proteins and other non-coding RNAs (Memczak et al., 2013) or act as cis-regulatory elements such as riboswitches (Wachsmuth et al., 2013). MiRNAs (Rajewsky, 2006) are an abundant class of small RNAs, each of which can regulate up to hundreds of transcripts. In total, it is estimated that 60% of all human proteins are regulated by miRNAs. With the advances in high-throughput approaches to detect binding sites of RNA-binding proteins (RBP) such as CLIP-Seq, a plethora of new RNA regulatory mechanisms has been detected when analysing the RBPome (i.e., the network of protein–RNA interactions) (Rinn and Ule, 2014). Finally, ribozymes are an important class of ncRNAs that are often involved in the maturation of other RNA or DNA molecules.

With all these potential roles, it has become clear that the analysis of epigenetic and expression data is incomplete if RNA-based regulation is not taken into account. As consequence, the analysis of RNA has to be integrative, combining sequencing datasets with sequence and structure analysis of RNA elements, and allowing for integration with other regulatory mechanisms such as transcription. High-throughput techniques to analyse RNA-based regulation are rapidly evolving, which give rise to a large amount of information but also to the need to constantly adapt databases, annotations and tools.

To overcome these problems and limitations, the RNA Bioinformatics Centre (RBC) was founded within the German Network for Bioinformatics Infrastructure (de.NBI) with the following priorities:

1 To establish an integrated, easily accessible RNA analysis workbench which can be used on our own cluster or downloaded and installed on every HPC environment.
2 To work with other Bioinformatics Centers and relevant scientific communities to allow for maximal usefulness, interconnectivity, and added value of the developed infrastructure.
3 To use this infrastructure as foundation for a learning and teaching environment that fosters an awareness for the importance of RNA analysis.

In consequence, our goal in RBC is to serve as contact point for all RNA bioinformatic questions in Germany, ranging from initial study design, over providing protocols and infrastructure, up to developing specialised solutions for individual problems. In addition, the RBC provides specialized curated RNA-related information resources such as databases for protein–RNA interaction or tRNAs, which will be fully integrated into our workbench. Across the three locations Berlin, Freiburg and Leipzig, the joint expertise covers many if not all aspects of RNA biology of current interest, ranging from structure prediction and genome-wide annotations of conserved secondary structures via the detection of members of specific classes of regulatory RNAs, and the interaction of RNA binding proteins and regulatory RNAs with their targets.

3. Individual tools provided and maintained by the RBC

In this section, we will give an overview of different services and tools that are required to analyse RNA-related data. We will take our emphasis on tools that are provided by the RBC and only shortly mention other related tools. Tools and databases maintained by the RBC will be written in italics. The complete list of tools can be found under https://github.com/bgruening/galaxy-rna-workbench.

3.1. Prediction of RNA structure and detection of conserved RNA structure

Many functional RNAs require a specific structure to be formed. Very often, the so-called secondary structure (i.e., the set of Watson-Crick and G—U bonds) is well-conserved and characteristic for the function of the RNA. Prediction of the secondary structure is a well-established area in RNA-bioinformatics. The *ViennaRNA Package* consists of a C code library and several stand-alone programs for the prediction and comparison of RNA secondary structures. It is also the de-facto standard library for the development of RNA based methods (Lorenz and Bernhart, 2011).

However, the prediction of the secondary structure is usually only a first step in a whole pipeline for the analysis of RNA-related data. Often it is required to determine the conserved secondary structure, or whether a structure is conserved at all. *MARNA* (Siebert and Backofen, 2007) is an early approach that solved the problem of generating multiple alignments using well-defined pairwise RNA-alignment approaches by using Tcoffee (Notredame et al., 2000) to combine the pairwise alignments. It computes multiple sequence-structure alignments considering a single fixed structure for each sequence only. *ExpaRNA* is a fast, motif-based comparison and alignment tool for RNA molecules. Instead of computing a full sequence-structure alignment, it computes the best arrangement of sequence-structure motifs common to two RNAs (Smith et al., 2010). The gold standard here is the Sankoff (1985) approach (and its variants) of performing a sequence-structure alignment of RNAs. One approach that is provided by the RBC is *LocARNA* (Will et al., 2007, 2012). It is an efficient variant of the Sankoff