Projektleiter(in)

Name:	**Prof.** **Peter Stadler**
Förderkennzeichen:	**031A538B**
Projketlaufzeit:	**01.03.2015 - 28.02.2018**
Abteilung/AG:	**Fak. Mathematik und Informatik** **Bioinformatik/IZBI**
Institut/Universität:	**Universität Leipzig** **Härtelstrasse 16-18** **04107 Leipzig**

Summary: Overview on the activities of the individual project

Leipzig is part of the de.NBI RNA bioinformatics center (RBC). The RBCs main objectives within de.NBI are as follows:

1. Tool interface descriptions
2. Galaxy integration
3. Virtualisation
4. Quality Management
5. Genome Annotation
6. Maintenance & Optimisation
7. Coordination & Training
8. Community & User Support

This tasks are tackled in an cooperative effort involving all three RBC service centers. The RBC-Leipzig contributed as follows:

Spearheaded by RBC-Freiburg, the `Galaxy` framework was established as central workflow system within the RBC. A major objective therefore, is the integration of existing tools in the Galaxy Tool Shed, making them accessible to researchers, enabling reproducibility and long time availability. To this purpose, many of the tools (co-)developed at Leipzig have been or are currently being integrated into the Galaxy workflow system. An overview of this tools can be found in section . As most of the provided tools are under constant development, this is a continuous process. For integration into the galaxy framework, these tools are being made available via the `BioConda` package manager, further increasing visibility and usability.

Besides development, maintenance and `Galaxy` integration of tools, another major topic that concerns the RBC-Leipzig is quality management of RNA-seq and other RNA centric high-throughput sequencing (HTS) data. Initial efforts were directed at an exhaustive survey of current state-of-the-art sequencing technologies and established quality control pipelines, with focus on non-standard applications and their specific needs. The fast evolving field of RNA-HTS raised our awareness that instead of formulating constraint and inflexible standard operating procedures (SOPs) for a snapshot of technologies and methods available at a certain timepoint, a more flexible, adaptale approach has to be found. Especially recent advances in the analysis of RNA interaction or epigenomics speed up this development. To this end, we decided to implement a collection of "best-practice"documents in favor if SOPs. The collection is hosted on `Github` (https://github.com/galaxyproject/training-material), together with a series of how-to's and training material for `Galaxy` and NGS analysis in general. This guarantees that the material is publicly available, and can be brought up-to-date by contributors in a revision controlled, community driven and reviewed way, while simultaneously visibility is increased. This allows to keep up with developments in the field easier, as not only the RBC members, but also scientist who use the tools we provide as well as other experts can contribute their knowledge to the collection in an easy and convenient way, while the community can review changes and stays in control of what is integrated in the collection.

Consequently, adaption of analysis and quality control tools like e. g., `FastQC`, which are necessary in response to recent advances, is an ongoing process. This also includes the development of new tools, as e. g., `Sierra platinum`, which is specifically designed for multi-replicate peak calling and quality control of NGS data. Providing adequate documentation and the integration of tools into the ELIXIER registry are also ongoing efforts.

A main pillar of de.NBI's mission is training. In that context, the RBC Leipzig was part of or hosting summerschools, training events, dedicated lectures and meetings, as can be seen in section **??**. Some of these trainings were conducted in cooperation with other RBC units and our industry partner ecSeq. Furthermore, the Leipzig group has successfully assisted wet-lab groups, e. g., in

the design of Riboswitches (Prof. Narberhaus, Prof. Dersch) und genomewide structureprobing (Prof. Moerl).

Tools and services developed and provided

The RBC-Leipzig is focused on the development and maintenance of tools for RNA (coding and non-coding) structure prediction, conservation analysis, motif detection, target prediction and annotation. These tools are made available to users as stand alone versions as well as an integrated part of the RBC `Galaxy` workbench and /textttConda package manager.

The tools provided by the RBC-Leipzig accumulated more then 4,000 citations over the last years. More than 200 users per year are actively contacting us by mail for support or advice regarding our tools. The latter are being downloaded more than 14,000 times per year. The web-services provided us have been visited more than 50,000 times, leading to more than 40,000 jobs being processed and 4TB of data stored per year. Database entries are requested more then 16,000 times per year, with more than 2,000 unique visitors, this includes requests via REST interfaces. Although a lot of users contact us for support, user satisfaction evaluation is so far unproductive.

In the following a list of tools and databases provided by the RBC-Leipzig, most of which have already been integrated into the `Galaxy` workframe and `Conda` package manager repositories or have been made available via wrapper scripts.

- Vienna RNA package – RNA-Strukturvorhersage mit Softconstraints
- LocARNA – multiple structure-based RNA alignment and folding
- SPARSE – highly efficient structure-based RNA alignment
- MEA – maximum expected accuracy prediction
- ExpaRNA-P –matching and folding of RNA alignments
- SparseMFEFold – space-efficient RNA folding
- CARNA 1.3.1 – constraint-based alignment of RNAs
- ViennaNGS – a perl suite for NGS analysis
- RNAcop – context-optimization with probability
- IntaRNA – prediction of RNA-RNA interactions
- Sierra platinum – multi-replicate peak-calling and quality control of NGS data
- AREsite2 – a database for AU-/GU-/U- rich elements in human and model organisms
- tRNAdb – a comprehensive database of tRNAs
- MITOS – annotation of metazoan mitochondrial genomes
- snoSTRIP – analysis of small nucleolar RNAs (snoRNAs) in fungi
- Kinwalker – co-transcriptional folding kinetics of large RNAs
- RNAsnoop – efficient target prediction for H/ACA snoRNAs

Together with the de.Stairs group in Leipzig, the following tools were integrated into the `Galaxy` and `Conda` frameworks.

- metilene – differentielle Methylierungsanalyse
- segemehl – a mapper for split read alignment of NGS reads

Most of the tools listed here are under constant development, such that new releases are being ported to `Galaxy` and `Conda` on a regular basis. Furthermore, most of the new releases have been linked to a continuous integration framework, which guarantees high quality and error-free packages for delivery to the end user.

Contribution to de.NBI activities

Contributions to workshops; summer schools and symposia Contributions to the de.NBI cloud (if applicable) Further plans Bezüglich Workpackages 7.3 und 8.1 wurde eine Summer School zu RNA-Seq für SPP 1738 geplant, die in Zusammenarbeit mit dem Bioinformatik-Dienstleister ecSeq abgehalten wird. Die darüberhinausgehende Kooperation mit ecSeq ist angebahnt. Bezüglich Workpackage 8.2 wurde die Dokumentation unserer Tools weitergepflegt; insbesondere wurden eine Reihe von Tools in die ELIXIR-Registry, die zentrale Katalogisierung und standardisierte Beschreibung von Bioinformatik-Tools ermöglicht, eingetragen.

Contributions to ELIXIR activities

Contributions to implementations projects and core data resources Other contributions

Together with members of the group of Jan Gorodkin (ELIXIR) and the other RBC centers, we worked on the ELIXIR registry, updated existing and added new tools in a series of workshops and meetings.

General information on the project

Composition of the project group Up to three important publications

Prof. Dr. Peter F. Stadler is full professor of Bioinformatics at Leipzig University, External Scientific Member at the MPI-MIS, where he directs a research group on discrete biomathematics, senior scientific advisor to the RNomics group at FH IZI in Leipzig, and an external professor at the Santa Fe Institute. He has pioneered RNA bioinformatics with the ViennaRNA package since the early 1990s. Beyond core algorithms for RNA folding his group has developed scanning algorithms for large genomes, facilities to compute consensus structures, and methods to deal with RNA-RNA interactions. Complementary, the group develops methods for the analysis of high throughput transcriptomics data. In particular, focusing on read mapping, functional RNA recognition from short read patterns, and detection of potential chemical modifications in RNA-seq data. In addition to the applicant, the following researchers from AG Stadler are or have been part of the RBC-Leipzig. Until 11.2016, Dr. Sebastian Will, since 01.2016 Dr. Jörg Fallmann, since 11.2016 Dipl. Bioinf. Jan Engelhardt. Together they have strong experience with development and maintenance of several RNA related software suites and databases concerning structure prediction, interaction, folding dynamics, and comparative analysis of structural RNA, as well as genome-wide prediction of non-coding RNA (CARNA, LocARNA, MEA, AREsite, etc.)

[P1] Washietl S, Hofacker IL, **Stadler, PF**. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **102**: 2454-2459 (2005).

[P2] Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, **Stadler, PF**. Mapping of conserved RNA Secondary Structures predicts Thousands of functional Non-Coding RNAs in the Human Genome. *Nature Biotech.* **23**: 1383-1390 (2005).

[P3] Washietl S, Pedersen JS, Korbel JO, Gruber A, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Stocsits C, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde

J, Drenkow J, Kapranov P, Gingeras TR, Guigó R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, **Stadler PF**. Structured RNAs in the ENCODE Selected Regions of the Human Genome. *Genome Res.* **17**: 852-864 (2007).

[P4] Rose D, Hackermüller J, Washietl S, Findeiß S, Reiche K, Hertel J, **Stadler PF**, Prohaska, SJ. Computational RNomics of Drosophilids. *BMC Genomics* **8**: 406 (2007).

[P5] Gruber AR, Findeiß S, Washietl S, Hofacker IL, **Stadler PF**. `RNAz 2.0`: improved noncoding RNA detection. Pac. Symp. Biocomput. **15**: 69-79 (2010).

[P6] Findeiß S, Engelhardt J, Prohaska SJ, **Stadler PF**. Protein-Coding Structured RNAs: A Computational Survey of Conserved RNA Secondary Structures Overlapping Coding Regions in Drosophilids. *Biochimie* **93**: 2019-2023 (2011).

[P7] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, **Stadler PF**, Hofacker IL. `ViennaRNA` Package 2.0. *Alg. Mol. Biol.* **6**: 26 (2011).

[P8] Lorenz R, Bernhart SH, Qin J, Höner zu Siederdissen C, Tanzer A, Amman F, Hofacker IL, **Stadler PF**. 2D meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction. *IEEE Trans. Comp. Biol. Bioinf.* **10**: 832-844 (2013).

[P9] Smith MA, Gesell T, **Stadler PF**, Mattick, JS. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* **41**: 8220-8236 (2013).

[P10] Lorenz R, Hofacker IL, **Stadler PF**. RNA Folding with Hard and Soft Constraints. *Alg. Mol. Biol.* **11**: 8 (2016).