Bioinformatics solutions for big data analysis in life sciences presented by the German network for bioinformatics infrastructure

Over the last decade, the amount of data generated by life science research has grown at a rapid rate and an increased complexity rendering the tasks of data analysis, management, and storage a unique challenge. To meet these rising demands, the German Network for Bioinformatics Infrastructure (de.NBI) was established in March 2015 as a national bioinformatics consortium funded by the German Ministry of Education and Research (BMBF). The de.NBI network provides bioinformatics services, training, and cloud resources to the research community in Germany.

This summary aims to brief the reader on the research applications of de.NBI services and tools in the field of big data analysis. It offers an overview on the special issue: "Bioinformatics Solutions for Big Data Analysis in Life Sciences Presented by the German Network for Bioinformatics Infrastructure". Special issue of *Journal of Biotechnology*, edited by Alfred Pühler, vol. 261, 2017, pp. 1-238.
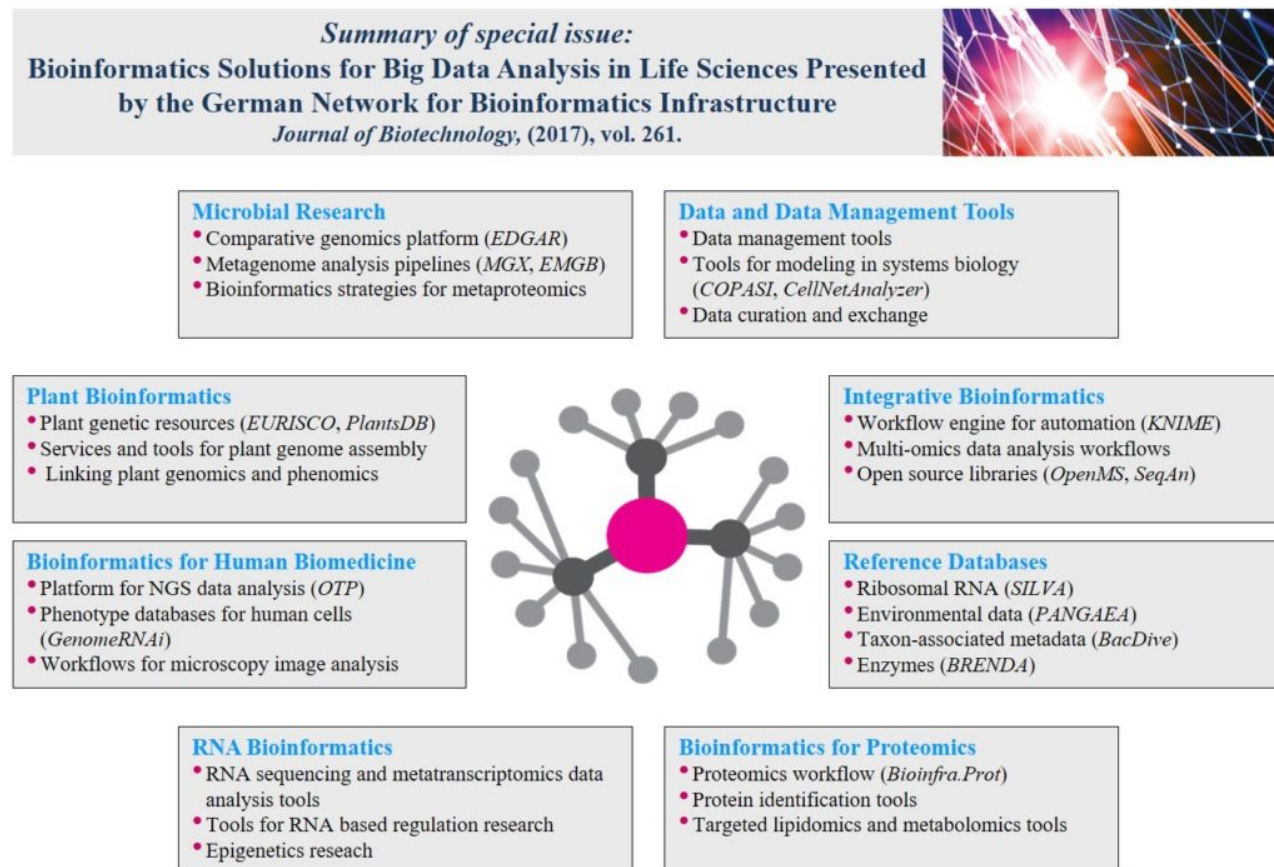

Summary of special issue:
Bioinformatics Solutions for Big Data Analysis in Life Sciences Presented by the German Network for Bioinformatics Infrastructure
Journal of Biotechnology, (2017), vol. 261.

Microbial Research
- Comparative genomics platform (*EDGAR*)
- Metagenome analysis pipelines (*MGX, EMGB*)
- Bioinformatics strategies for metaproteomics

Data and Data Management Tools
- Data management tools
- Tools for modeling in systems biology (*COPASI, CellNetAnalyzer*)
- Data curation and exchange

Plant Bioinformatics
- Plant genetic resources (*EURISCO, PlantsDB*)
- Services and tools for plant genome assembly
- Linking plant genomics and phenomics

Integrative Bioinformatics
- Workflow engine for automation (*KNIME*)
- Multi-omics data analysis workflows
- Open source libraries (*OpenMS, SeqAn*)

Bioinformatics for Human Biomedicine
- Platform for NGS data analysis (*OTP*)
- Phenotype databases for human cells (*GenomeRNAi*)
- Workflows for microscopy image analysis

Reference Databases
- Ribosomal RNA (*SILVA*)
- Environmental data (*PANGAEA*)
- Taxon-associated metadata (*BacDive*)
- Enzymes (*BRENDA*)

RNA Bioinformatics
- RNA sequencing and metatranscriptomics data analysis tools
- Tools for RNA based regulation research
- Epigenetics reseach

Bioinformatics for Proteomics
- Proteomics workflow (*Bioinfra.Prot*)
- Protein identification tools
- Targeted lipidomics and metabolomics tools

www.denbi.de

Fig. 1. Summary of the bioinformatics services and tools published in "Bioinformatics Solutions for Big Data Analysis in Life Sciences as Presented by the German Network for Bioinformatics Infrastructure. Special issue of Journal of Biotechnology, edited by Alfred Pühler, vol. 261, 2017, pp. 1-238.

The repertoire of software tools highlighted in the special issue addresses the bioinformatics need of the life science community at a wide range of topics. In the field of microorganism research, the enhanced software platform for comparative genomics *EDGAR*, and the comprehensive platforms for the analysis of microbial metagenomes *MGX* and *EMGB* are presented. Bioinformatics strategies for metaproteomic data analysis are further discussed. Researchers in the field of plant science are directed to a collection of plants genetics information resources, e.g., *EURISCO*, *IPK Blast Server*, *PlantsDB*, and *PlabiPD*. Tools important for plant genome assembly and analysis are also presented. Scientists dealing with high throughput datasets in the field of human biomedical research can learn about the use of the automated workflow management system *OTP*, and gain information on human cells phenotype databases such as *GenomeRNAi* and *GenomeCRISPR*. For microscopy image analysis, workflow systems for the integration of analysis techniques focusing on *KNIME* and *Galaxy* are also introduced. The bioinformatics services of the de.NBI network expand to include software for the analysis of data produced by high throughput omics technologies. RNA research is supported by an extensive portfolio of relevant tools applied in the field of RNA sequencing and metatranscriptomic data analysis. A collection of presented tools and workflows offer specialized bioinformatics solutions and cover questions related to RNA based regulation, RNA folding, and epigenetics research (e.g., *LocARNA*, *IntaRNA*, *ViennaRNA*, *PicTar*, *Galaxy RNA workbench*, *WBSA*). For the analysis of high throughput proteomics data, a comprehensive workflow for LC-MS/MS featuring multiple proteomic data analysis components, the peptide search engine *SearchGUI*, and the protein inference tool *PeptideShaker* are described in details. Targeted tools for lipidomics research such as *Skyline* and *LipidXplorer*, and the *KNIME* integrated *MetFrag* tool for metabolite identification are also added.

Papers reviewing the *OpenMS*, *KNIME*, and the open source library *SeqAn* are examples on workflows that embrace the integration of bioinformatics tools to provide solutions for the flow and reproducibility of data.

As a national research infrastructure, de.NBI importance is emphasized by its vital role in database provision in bioinformatics. These activities are highlighted providing a guided reference to the use of the databases *SILVA* (ribosomal RNA), *PANGAEA (*environmental data), *BacDive* (taxon associated metadata), and *BRENDA* (enzyme information). The two kinetic and metabolic modeling, *COPASI* and C*ellNetAnalyzer* systems biology tools are reviewed together with the structure base design workflow for chemoinformatics *ProteinsPlus*.

The task of data management is becoming increasingly important for scientists in the rapidly

evolving big data era. To address these issues, examples of curated tools for systems biology modeling such as *SABIO-RK*, and platforms for data management including *SEEK* are discussed in details.

With this summary, The German Network for Bioinformatics Infrastructure presents the scope of its bioinformatics services and hopes to further encourage their application in analyzing big experimental data.

Alfred Pühler
de.NBI - German Network for Bioinformatics Infrastructure, c/o CeBiTec, Bielefeld University,
33594 Bielefeld, Germany

Publication