Knowledge Management in Genomics: The Role of Data Provenance

Vinicius W. Salazar[1], Kary Ocaña[3], Fabiano L. Thompson[2], Marta Mattoso[1]
[1]Department of Computer Science/COPPE, Federal University of Rio de Janeiro, Brazil
[2]Institute of Biology, Federal University of Rio de Janeiro, Brazil
LNCC, National Laboratory of Scientific Computing, Petrópolis, Brazil[3]

Genomics as a discipline has grown considerably in recent years and its methods have proven to be helpful for solving diverse problems across various domains of the life sciences. It has been constantly driven by data life cycles, with scientists relying on public, curated data repositories to perform their own experiments. The consequences of this are that the issue of managing data correctly and efficiently have been central to the field, and with the exponential growth in publicly available data, this has turned into a "big data issue" which concerns individual scientists, research groups, institutions, and international consortia. We discuss how knowledge management (KM) can address some of the big data issues in genomics. The KM approach has proven to be valuable in case studies, due to its systematic process of defining workflows that improves efficiency and reproducibility of experiments. However, combining a KM workflow with genomics has challenges related to the genomics data life cycle like *provenance*. Tracking the *provenance* of data in genomics is becoming part of its data life cycle, but provenance data is not part of the KM workflow. Effective data management, good practices in scientific computing, FAIR data sharing, and collaborative work between experts in different disciplines can all be facilitated by provenance tracking. Provenance tracking is a key aspect in establishing the data-information-knowledge cycle, which can be used as a framework for planning, executing scientific and monitoring experiments. Because provenance is information providing context to data, it plays a fundamental role in generating and sharing knowledge at the end of the cycle. Provenance tracking needs data capture and storage, which may affect the performance of the genomics workflow. Our proposal is focused on bioinformatics workflows and how they should be planned from a KM provenance-based perspective. We show how to adopt a KM provenance-based perspective of experiments in genomics with negligible computational costs by adopting a dataflow analysis system, in this case, the DfAnalyzer tool, which adopts the W3C PROV standard. By coupling DfAnalyzer with a genomic bioinformatics pipeline, it is possible to capture provenance data which when queried generates information about trade-offs, for example, of performance and sensivity, providing the research with an enriched knowledge of his analysis. The different DfAnalyzer components will align with those of knowledge management: the Provenance Data Extractor, Raw Data Extractor and Raw Data Index will act on data capture and storage, the Query Interface and Dataflow Viewer on information summarizing and analysing, allowing a synthesis of knowledge around the experiment which can support decision making. This is particularly useful when scaling experiments with a large number of samples or parameter sweeping. For demonstration, we executed DfAnalyzer with a bacterial genome annotation and phylogenetic analysis workflow, using the DfA Python library. We considered the steps of data and metadata collection, sequence statistics, gene call predictions, annotation with the NCBI COG database, and pairwise Average Nucleotide Identity between genomes, integrating each step to the DfA Python library with a custom Python package. In our workflow case study we show the benefits of monitoring, querying and reuse of methods and results in genomics experiments, showcasing how a KM approach to

research in the field may as well be a part of the new paradigm of data-intensive discovery in biology.