A Sound and Repeatable Approach to Building Integrated Repositories of Genomic Data

Anna Bernasconi[1]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133, Milan, Italy

PhD Thesis Award (Extended Abstract). The integration of genomic data and of their describing metadata is, at the same time, an important, difficult, and well-recognized challenge. It is *important* because a wealth of public data repositories is available to drive biological and clinical research; combining information from various heterogeneous and widely dispersed sources is paramount to a number of biological discoveries. It is *difficult* because the domain is complex and there is no agreement among the various data formats, data models, and metadata definitions, which refer to different vocabularies and ontologies. It is *well-recognized* in the bioinformatics community because, in the common practice, repositories are accessed one-by-one, learning their specific metadata definitions as result of long and tedious efforts, and such practice is error-prone; moreover, downloaded datasets need considerable efforts prior to insertion in analysis pipelines.

Within the context of the European project data-driven Genomic Computing (GeCo), which supports genomic research by proposing bioinformatics abstractions and tools, this PhD thesis has focused on the data integration problem, sharing the motivations and methodologies of the project and addressing one of its objectives.

First, we thoroughly analyze the players involved in the genomic data context and proposed a conceptual model of metadata (the Genomic Conceptual Model [1]) to represent in a general way the most common information attributes that describe genomic samples and experiments in the available sources. The model represents a typical genomic region data file by different perspectives (biology, technology, management and extraction) and sets the basis to query the underlying data sources for locating relevant experimental datasets. We then describe META-BASE [2], our architecture for integrating datasets, retrieved from a variety of genomic data sources, based upon a structured transformation process; we present a number of innovative techniques for data extraction, cleaning, normalization and enrichment and we show a general, open, and extensible pipeline that can easily incorporate any number of new sources. The resulting repository – already integrating several important sources such as the Encyclopedia of DNA Elements, The Cancer Genome Atlas, the Roadmap Epigenomics, and the 1000 Genomes Project – is exposed by means of user interfaces that respond to biological researchers' needs. We provide both a graph-based endpoint for expert users, who need to explore the semantic structure of metadata, and GenoSurf [3] (http://www.gmql.eu/genosurf/), a user-friendly search system providing access to the consolidated repository of metadata attributes, enriched by a

ITADATA2022: The 1st Italian Conference on Big Data and Data Science, September 20–21, 2022, Milan, Italy
0000-0001-8016-5750 (A. Bernasconi)

multi-ontology knowledge base, that allows to locate relevant genomic datasets, which can be then analyzed with off-the-shelf bioinformatics tools. This interface has been evaluated by running an extended empirical study with participants knowledgeable in both Biology and Computer Science, collecting many insights on the practices of different user profiles and on their understanding of procedures for extracting datasets relevant for research.

The models, frameworks and tools that are described in this thesis are already included in follow-up projects; they can be exploited to provide biologists and clinicians with a complete data extraction/analysis environment, equipped by a 'marketplace' of ready-to-use best practices. The process may be guided by a conversational interface, which breaks down the technological barriers that currently slow down the practical adoption of our systems. Our commitment is to continue the inclusion of relevant data sources for bioinformatics tertiary analysis, continuously improving the process from a data quality and interoperability point of view.

Inspired by our work on genomic data integration, during the outbreak of the COVID-19 pandemic we searched for effective ways to help mitigate its effects; in this direction, we successfully re-applied the model-build-search paradigm used for human genomics.

Even if the domain of viral genomics is completely new, it presents many analogies with our previous challenges. In this new context, we model viral nucleotide sequences as strings of letters, with corresponding sub-sequences – the genes – that encode for proteins composed of amino acids. To highlight differences with previously considered data, we have designed the Viral Conceptual Model [4] which accounts for their technological, biological and organizational aspects, in addition to computed annotations and mutations on both nucleotides and amino acid sequences. We then integrate sequences with their metadata from a variety of different sources and propose the powerful search interface ViruSurf [5] (http://www.gmql.eu/virusurf/), able to quickly extract sequences based on their combined mutations, to compare different conditions, and to build interesting populations for downstream analysis. When applied to SARS-CoV-2, the virus responsible for COVID-19, complex conceptual queries upon our system are able to replicate the search results of recent articles, hence demonstrating considerable potential in supporting virology research.

This work has been realized during the first spread of the SARS-CoV-2 pandemic (March–December 2020); after setting the first milestones, we are now moving forward, considering the next challenges of this new domain with growing interest. These include the development of a requirements elicitation technique for emergency times, the extension of ViruSurf to other types of data (e.g., relevant for vaccine design), and the provision of visual and statistical support to the integrated data.

The results on this thesis are part of a broad vision: the availability of conceptual models, related databases, and search systems for both human and viral genomics will provide important opportunities for genomic and clinical research, especially if virus data will be connected to its host, the human being, who is the provider of genomic and phenotype information.

[1] A. Bernasconi, et al., Proceedings International Conference ER 2017, Springer, pp. 325–339.
[2] A. Bernasconi, et al., IEEE/ACM Trans. on Comput. Biol. and Bioinf. 19 (2022) 543–557.
[3] A. Canakoglu, A. Bernasconi, et al., Database, Volume 2019, baz132.
[4] A. Bernasconi, et al., Proceedings International Conference ER 2020, Springer, pp. 388–402.
[5] A. Canakoglu, P. Pinoli, A. Bernasconi, et al., Nucleic Acids Research 49 (2021) D817–D824.