Integrative Bioinformatics and Omics Data Source Interoperability in the Next-Generation Sequencing Era - Editorial

Simona E. Rombo and Domenico Ursino

With the advent of high-throughput and Next Generation Sequencing (NGS) technologies [1], huge amounts of "omics" data (i.e., data from genomics, proteomics, pharmacogenomics, metagenomics, etc.) are continuously produced. Combining and integrating diverse omics data types is important in order to investigate the molecular machinery of complex diseases, with the hope for better disease prevention and treatment [2]. Experimental data repositories of omics data are publicly available, with the main aim of fostering the cooperation among research groups and laboratories all over the world. However, despite their openness, the effective integrated use of available public sources is hampered by the heterogeneity, complexity and large size of data stored therein.

The main issues to be addressed when approaching omics data integration are related to the difficulty in managing and analysing these data. Indeed, specific and multidisciplinary competences are required, and combining data of different types is not a simple task. Both the extensional (i.e., the real data) and intensional (i.e., the corresponding metadata) levels may be involved in this integration process, according to the specific problem under consideration. In the last few years, information systems researchers have made significant efforts in the proposal of effective methodologies for the integration of structured and semi-structured data formats [3]. However, omics data are often unstructured. This pushes towards the study of how data source integration can be successfully performed when structured, semi-structured and unstructured data sources coexist.

This themed issue provides an extensive overview of the main challenges related to omics data integration and the methods that have been recently proposed in order to address them. It comprises nine manuscripts, each dealing with one of four central key issues, as detailed below.

How cellular components impact diseases. Understanding how the direct or indirect relationships among cellular components may impact the occurrence and progress of disorders and diseases is an important issue, that requires omics data integration to be addressed. Manuscripts of this group start from the assumption, confirmed by several studies in the literature, that the occurrence and progress of many diseases have genetic causes. For example, **genetic**

variations have direct effects on individual phenotypes, possibly causing the production of partially or totally dysfunctional proteins. With this regards, Galano-Frutos, García-Cebollada and Sancho in *Molecular Dynamics Simulations for Genetic Interpretation in Protein Coding Regions: Where we Are, Where to Go and When* observe that predicting whether the replacement of one amino acid residue with another will be tolerated or cause disease is a key factor. In particular, first they review existing prediction tools based on evolutionary information and simple physical-chemical properties. Then, they describe more recent and accurate methods, such as full-atom Molecular Dynamics (MD) simulation in explicit solvent, and discuss how these methods can be used in order to interpret human genetic variations at a large scale.

Another important aspect is related to data coming from **single cell analysis**, which may be used in order to understand differences in healthy/unhealthy populations. In *Computational methods for the integrative analysis of single cell data*, Forcato, Romano and Bicciato describe the computational methods for the integrative analysis of single cell genomic data. They mainly focus on the integration of single-cell RNA sequencing datasets and on the joint analysis of multimodal signals from individual cells.

Accessing biological databases and related metadata. Omics data are represented in a wide variety of notations and formats, often with different levels of quality. This intrinsic heterogeneity in both data and repositories makes difficult their effective combination for producing new knowledge, and may hamper their correct use and exploitation. **Genomic data integration** is the topic of *The road towards data integration in human genomics: players, steps and interactions* by Bernasconi, Canakoglu, Masseroli and Ceri. In this manuscript, the authors first describe a technological pipeline from data production to data integration. Then, they propose a taxonomy of genomic data players and apply it to about 30 important players. They specifically focus on integrator players and evaluate the computational environment for data integration purposes provided by them.

The role of **conceptual models** to support the efficient management of genomic data is discussed in *Using Conceptual Modeling to Improve Genome Data Management* by Pastor, León Palacio, Reyes Román, García S. and Casamayor. The authors describe a solution that helps researchers to organize, store and process information and, at the same time, focuses only on relevant data minimizing the information overload in clinical research context.

An overview of available **patient-level datasets** containing both genotypic and phenotypic data is presented in *GenoPheno: cataloging large-scale phenotypic and next generation sequencing data within human datasets* by Gutiérrez-Sacristán, De Niz, Kothari,Won Kong, Mandl and Avillach. In this manuscript, the authors describe a dynamic, online catalogue for consultation, contribution and revision by the research community. It consists of 30 datasets and was created by them with the purpose of making it publicly available.

Learning from omics data. A survey on **machine learning** methods operating on the cloud for gene regulation studies is presented in *Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations* by Oh, Park, Kim and Chae. The authors describe these methods, categorize them according to five different goals, and summarize them in terms of multi-omics input types. They explain the positive role that the cloud can play for the analysis of multi-omics data. They also discuss some important issues to address when machine learning-based approaches operating on the cloud are adopted for the analysis of gene regulations.

Structured sparsity regularization for analyzing high-dimensional omics data by Vinga focuses on **structured regularizers** and **penalty functions**, when applied to omics data. The author analyses their potential in identifying disease's molecular signature, in order to create high-performance clinical decision support systems and, ultimately, favor personalized healthcare.

Studies on microorganisms. Microbial communities and viral populations have a crucial role in the environment and in human health. In *Comparison of Microbiome Samples: Methods and Computational Challenges*, Comin, Di Camillo, Pizzi and Vandin provide a study on **metagenomic** NGS datasets. These authors compare datasets from three different viewpoints, namely: *(i)* species identification and quantification, *(ii)* efficient computation of distances between metagenomic sample datasets, and *(iii)* identification of metagenomics features associated with a phenotype.

In *Epidemiological Data Analysis of Viral Quasispecies in the Next-Generation Sequencing Era*, Knyazev, Hughes, Skums and Zelikovsky deal with the analysis of intra-host RNA **viral populations**. In particular, they examine bioinformatics tools that: *(i)* characterize the complexity of intra-host viral population; *(ii)* support epidemiological analysis in inferring drug-resistant mutations, infection age and patient linkage; *(iii)* support surveillance systems for fast response and outbreak control.

Hopefully, this themed issue will represent a springboard for fruitful collaborations among researchers from multidisciplinary areas, which could give a significant boost to the advancement of knowledge in different fields through a more effective analysis of omics data.

The Editors are grateful to both the Editor in Chief and the Publisher for having trusted this project, and for having supported them in all their needs. Many thank also to all the authors and reviewers, whose expertise and effort allowed the realization of this themed issue.

References

[1] Jason A. Reuter, et al. High-throughput sequencing technologies. *Molecular cell*, 58(4): 586-97 (2015).

[2] H. Yang, et al. Multilevel heterogeneous omics data integration with kernel fusion. *Briefings in Bioinformatics*, 21(1): 156-170 (2020).

[3] Philip A. Bernstein, et al. Generic schema matching, ten years later. *Proceedings of the VLDB Endowment* 4(11): 695-701 (2011).

Simona E. Rombo
Dipartimento di Matematica e Informatica
Università degli Studi di Palermo
e-mail: simona.rombo@unipa.it

Domenico Ursino
Dipartimento di Ingegneria dell'Informazione
Università Politecnica delle Marche
e-mail: d.ursino@univpm.it