Pak. J. Biochem. Mol. Biol., 2020, 53(1):11-13

Opinion

Needs of enhancing the usability and archival stability of Bioinformatics analysis tools in "Multi-OMICS" Era

Waqasuddin Khan*, Javairia Khalid, M. Imran Nisar, Fyezah Jehan

[1]Department of Peadiatrics and Child Health, Faculty of Health Sciences, Medical College, The Aga Khan University, Stadium Road, Karachi- 74800, Pakistan.
*Corresponding Author: Email: waqasuddin.khan@aku.edu
Research group page: https://www.aku.edu/mcpk/research/Pages/biorepository-and-omics.aspx

Abstract

Huge omics data have been generated from high-throughput profiling platforms. Several databases contain thousands of multiple-omics public repositories generated from diverse platforms, across many experimental conditions and sample types. The exponential growth in the number of available datasets has created many challenges to data analysis and integration, e.g., substantial coding skills are required to navigate between inputs and outputs from one platform to another, and from one dataset (genomics) to another (transcriptomics). This makes the whole process both difficult and time-consuming. In this opinion, we urge to make these workflows more accessible and useful to interpret for biologists and clinicians. Consequently, we put up some suggestions to alleviate the need for researchers to invest time and effort to strengthen the necessary computational tools advancement.

Key points:

1. Uniformity of -omics data types and its subsequent analysis,
2. Proper ease of usability and archival stability should be indispensable for any –omic related tool, and
3. Dedicated resources should be utilized for the development of Bioinformatics software.

Since the sequencing of the human genome, there has been an explosion of biological data, both in size and complexity. High-throughput sequencing technologies are increasingly being used in attempts to further understand disease pathophysiology and treatment. Democratization of genome-scale technologies means that the generated 'big genomics data' is stored as large heterogeneous databases, offering advancement in the analysis by software development. Novelties in -omics technologies have added another layer of complexity to drive innovations in biomedical research as these databases and tools are mostly used by bench biologists making it possible to study simultaneously genomics, transcriptomics, proteomics, metabolomics and epigenomics, a task which was at a previously unimaginable level [1]. However, integrating and analyzing the significant volumes of data generated from different high-throughput omics technology platforms remain a significant challenge to basic and clinical research scientists without bioinformatics expertise or access to bioinformatics collaboration.

While integrated analysis is desirable to understand complex biological phenomena,

there are inherent differences in the datasets generated, and therefore integrating multiple omics platforms remains a challenge for researchers worldwide [2]. As a consequence, there is a gap between data yield and analysis resulting in difficulty to extrapolate information from large datasets to generate knowledge of clinical relevance. The fortunate development of powerful algorithmic techniques has led to the production of numerous software and tools that practically address important biological questions, and subsequently lay the foundation for novel clinical translations.

With the increasing popularity of integrating these omics platforms-enabled approaches into the research workflow, anecdotal usability and archival stability issues have been brought to the forefront of discussions among life sciences community. Software usability circles around the degree of ease and satisfaction of targeted users to execute and achieve the quantified objectives (the desired output). Archival stability of software is to keep the repository, or backup of version numbers, track changes history, and any other activity that helps the developer to go back and find the debugging lines. While many peer-reviewed scientific journals demand to share both the (test) data and algorithm for reproducibility — the proficiency of the tool to analyze and reproduce results as reported in the original study depends on numerous factors, such as, software default setting options and parameters used. Current guidelines are not efficient enough to promote usability and long-term archival stability of such software tools [3], e.g., there is no effective rule regarding the availability, installability and executing the code on different platforms/operating systems, or even at different file formats/versions.

The harmony between computational and wet-lab researchers and the reproducibility of published results is especially advantageous when software developers introduce and distribute sophisticated computational tools and packages with user-friendly interfaces for installing and executing the tools [4]. For instance, many medical and life sciences researchers lack formal training in computer science and may struggle to perform manual tweaking, such as, editing software codes and installing software dependencies [4]. In addition to that, published software tools are often made accessible through the uniform resource tool locators (URLs) that are typically mentioned in the articles and assumed to be permanent. However, it is not uncommon for URLs to become inactive due to reconfiguration or removal of web content leading to the loss of many digital resources.

Rapid emergence of omics technologies has outpaced software development and capacity to handle large and complex datasets. An ideal software could be adaptable to add new features and data types, be efficient and easy-to-use, and be easily accessible on the web through a web tool or web server. Additionally, it could perform automated analysis in synchronization with the biological and clinical information and allow a meta-analysis. This is only possible with more time and resource allocation to the development and maintenance of bioinformatics software to progress the science of multi-omics.

References

1. Sandhu C, Qureshi A, Emili A. Panomics for precision medicine. Trends Mol Med. 2018;24(1):85-101.
2. Berger B, Peng J, Singh M. Computational solutions for omics data. Nat Rev Genet. 2013;14(5):333-346.

3. Stodden V, Seiler J, Ma Z. An empirical analysis of journal policy effectiveness for computational reproducibility. Proc Natl Acad Sci. 2018;115(11):2584-2589.

4. List M, Ebert P, Albrecht F. Ten simple rules for developing usable software in computational biology. PLoS Comput Biol. 2017;13(1):e1005265.