УДК 004.522:2

BIG DATA FROM MULTI-OMICS TECHNOLOGIES IS ADVANSING MEDICINE AND HEALTHCARE

O.M. Abbasova
Lecturer of the Department of Software of information technologies,
The Institute of Telecommunications and Informatics of Turkmenistan
ogultaganabbasowa@g.mail.com

O.M. Abbasova
She graduated from the Institute of Telecommunications and Informatics of Turkmenistan. The area of scientific of interests is associated with the use and development of artificial intelligence, the organization of educational and research processes at the institute.

Annotation. The global surge in access to digital technologies has led to the collection and documentation of information in various public and private sectors. The healthcare industry has been substantially influenced by the advent of big data. Clinical records, patient lifestyle data recorded from wearable devices, IoTs, electronic health records, imaging analyses, molecular and diagnostic profiling, population studies and clinical trial information together have contributed to major advances and disruption in the healthcare sector.

Keywords: Big data, big data in biomedicine, data analytics in biomedical research, multi-omics big data, biomedical data management, big data in personalized medicine

Introduction. Big data is defined by massive amounts of data generated in the private and public sectors. In the healthcare industry, sources of big data include clinical records, health records of patients, results of medical examinations, and data collected by using diagnostic and health management devices as part of the internet of things (IoTs). Additionally, the field of biomedical research or biomedicine generates a substantial amount of data, thus contributing to data overload, which is relevant to public healthcare. Big data aims to encourage the use of new technologies to analyze large numbers of available datasets and extract new information on a variety of topics. Big data can take the form of any high volume, complex data representing a condition of interest. Collections of data are often characterized by the five “big Vs” of big data in designing computational systems for big data analytic: volume refers to the size of an overall dataset collection, velocity is the speed at which computational pipelines can load and process the datasets, veracity describes the reliability of datasets in reflecting the true nature of a condition of interest, the variety of data refers to how diverse the datasets sets are with respect to one another in the collection, and variability is a data quality representing continual changes in data and the consequent effects on the performance on a computational system. Together, these qualities contribute to the collective value of big data when applied in any discipline and must be thoroughly addressed in designing big data operational workflows.

Historically, the evolution of big data has markedly changed all aspects of personal and professional life. The collection of information followed by its documentation and analysis dates to 18,000 BCE, when people in tribes used markings, sticks and engravings to track days and food supplies. Moreover, civilizations such as the Egyptians and Romans used libraries to store information that was essential for governing and the military. In the modern era, the birth of big

data has been attributed to some of the first reports of data analysis by John Graunt during the London plague in 1663, after which the innovation and invention of several hardware processing and storage devices paved the way for managing the data generated. Subsequently, the emergence of computational language along with database management systems aided in the global evolution of information technologies and the big-technology sector. Today, technologies that use big data revolve around the gathering, processing, computing, visualizing, and storing of datasets, ideally in the form of an efficient high-speed system. The sheer volume of data available is increasing exponentially each year. Currently, the field of biomedical research is flooded with large experimental datasets obtained from patient tissue, blood, body fluids and laboratory investigations. These data include molecular expression changes obtained from high-throughput "omics"-based studies and clinical parameters from wearable devices. Hence, several large-scale cloud-based platforms have been developed to store, process, curate and manage massive amounts of biomedical data. Some of these platforms include ELIXIR, which is supported and funded by the European Molecular Biology Laboratory, the Global Alzheimer's Association Interactive Network (GAAIN) and Genomics Data Commons funded by the US National Cancer Institute, among others. As such, a growing need exists to develop new technologies that can handle such large-scale datasets while minimizing the costs and demands on computation.

Impact of big data on enterprises and the health care industry. The global surge in access to digital technologies has led to the collection and documentation of information in various public and private sectors. The healthcare industry has been substantially influenced by the advent of big data. Clinical records, patient lifestyle data recorded from wearable devices, IoTs, electronic health records, imaging analyses, molecular and diagnostic profiling, population studies and clinical trial information together have contributed to major advances and disruption in the healthcare sector.

Healthcare and medical research are shifting focus to improving disease outcomes through finding hidden associations or patterns within data derived from a wealth of available data resources. Such data can be used in prediction studies to determine highly probable outcomes of disease. These investigations are leading to advances in delivering personalized medicine tailored to individual patients' needs and the development of automated analysis to aid in the diagnosis and treatment of diseases. Big data has also contributed to better understanding of epidemiology within populations for a variety of pathogens . Immunological surveillance centers routinely process big data to identify pathogens at high risk of becoming endemic within populations. Similarly, big data from genomic libraries, such as whole-genome sequencing and whole-exome sequencing libraries, are playing key roles in accelerating data-driven biomedical discovery. Whole genome sequencing, transcriptome sequencing (RNA-seq and ribosome profiling), proteome profiling (mass spectrometry) and interactome profiling (chromosome conformation capture, ChIP-seq and hybrid assays) are enabling faster, more economical access to the biomedical information of each patient or a specific cohort. Recent advances in genomics, including single-cell genome and transcriptome sequencing of circulating tumor DNA in liquid biopsies, and metagenomics, are already markedly affecting medicine and are steadily being integrated into standard medical practice for early-stage diagnosis and various stages of disease treatment.

The application of big data in genomics is rapidly changing therapeutic development. In the past, drugs were developed according to knowledge of the biological pathways involved in disease, often derived from experiments performed in cells or. Very often, these experimental findings fail to be replicated in humans. Furthermore, drug development is limited to biological processes that are known or believed to be known. The ability to access and analyze the whole genome has allowed scientists and clinicians to identify novel pathways involved in disease, and to develop therapies relevant to humans. This approach supersedes the classical method of drug development, which is conducted according to knowledge of biological pathways involved in disease, based on results obtained from *in vitro* or *in vivo* experiments. However, the major

challenge in the classical method has been the poor rate of success, given the species differences and biological and technical variables. Because genetic variants are often rare, or have small effect sizes, large datasets are required to make valid inferences about the roles of these variants in disease. The collective large amount of genetic information (genomics big data) is in turn guiding the development of therapeutic modalities that are, or could be, individualized to patients according to their genetic profiles.

The sequence of the human genome is the foundation for the understanding of how genetic (DNA and RNA) instructions lead to biological function in humans. In-depth understanding of the genetic landscape of Homo sapiens has been gained from the Human Genome Project. Continued genomic research using this genome map has increased understanding of the functions of genes and the molecular factors that regulate them; variations in DNA and RNA sequences among individuals, and their roles in diseases; and nucleic acid-protein interactions and their regulation in an environment specific manner. Specifically, several international initiatives have been undertaken to evaluate the genetic profiles of patient populations and identify patterns associated with disease risk and therapeutic approaches and responses. For example, The Cancer Genome Atlas (TCGA) is a project initiated by the National Institutes of Health and the National Human Genome Research Institute to sequence specimens from multiple disease sites and gain in-depth understanding of the genetic alterations associated with cancer. Similarly, the genome-wide association study (GWAS) approach allows the whole genomes of individuals to be compared to identify genetic variations called single-nucleotide polymorphisms that are associated with disease. The International HapMap Project was established to determine the patterns in human genome variation and determine their effects on health and disease, including responses to drugs or environmental factors . The Encyclopedia of DNA Elements project was designed to identify the functional and regulatory elements in the human genome sequence, including proteins and noncoding RNA molecules that affect genetic function and outcomes in health and disease. Beyond genomic DNA regions, substantial evidence indicates that RNAs, including messenger RNAs and noncoding RNAs, such as microRNA (miRNA), may also enable effective disease diagnostics in early and late stages of various diseases, including cancer, neurological disorders, and infectious diseases.

With ongoing improvements in the field of bioinformatics and the combination of basic and clinical science with machine learning and high-power computing, major advances have been made in disease diagnosis and the monitoring of molecular changes in organ regions, tissues and circulating body fluids, such as the blood, plasma/serum, urine and cerebrospinal fluid. Collectively, the use of big data in biomedicine involves extracting the essential information from large amounts of research and clinical data to assist in decisions advancing the field toward effective patient-centric healthcare and disease management.

Data integration across different “omics” branches advances biomedical research.

Omics technology generates big data that require sophisticated bioinformatic analysis and has enabled the detailed study of many cellular components in their roles and behaviors across many cells and tissues. The availability of high-quality omics data has substantially contributed to understanding across the many tiers of biological conditions in health and disease. Integration of big data across different branches through multi-omics has accelerated outcomes in biomedical research; this integration is relevant for pathogen detection, disease diagnosis and timely therapeutic intervention.

Analytical pipelines for typical data integration across different omics branches advanc**e** biomedical research. The analytical procedures, including commonly used analysis packages for various omics data, are summarized in this schematic. Overall, all omics data analysis requires raw data acquisition (gray boxes); quality control (green boxes); identification, quantification, and statistical inference pipelines (yellow boxes); and functional enrichment analysis (red boxes). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Data sources and applications of multi-omics. Multi-omics data generated from various experimental pipelines (genomics, epigenomics, transcriptomics, proteomics and metabolomics) in pre-clinical and clinical investigations provide useful insights into the flow of biological information at multiple levels (genotype to phenotype), thus aiding in clarifying the mechanisms and factors that contribute to disease pathology.

Integration of data in multi-omics approaches provides a multi-faceted readout by combining results from multiple data layers (e.g., genome, transcriptome, epigenome, and metabolome) of a biological condition. Indeed, several research groups have successfully used multi-omics data to profile biological phenotypes and gain a better understanding of the mechanisms underlying various diseases. Biological data integration can be performed in several ways, depending on the biological question to be answered. Several R software packages have been developed to facilitate easier integration of several types of omics data. One example includes the R package HiCeekR, which enables the integration of Hi-C, RNA-seq and ChIP-seq datasets to explore chromatin configuration changes and their effects on the transcriptome of a host. These packages are usually characterized by each of the different data types (e.g., Hi-C, RNA-seq or ChIP-seq) analyzed in parallel with their results being subjected to a post-integration process. A key downstream analysis is functional annotation of the genes of interest identified by various omics data. The mix Omics R package uses a variety of multivariate methods to account for the increased complexity of integrated datasets, owing to the accumulation of more experimental variables from each of the datasets. Multivariate approaches have been shown to be perform well in the integration of large complex datasets. Currently, increasing omics profiling endeavors have focused on the single-cell level, to uncover the heterogeneity of multi-cellular tissues. Simultaneous profiling of transcriptomes and epigenomes can be performed at the single-cell, or a given spatial location. In this regard, Seurat, a powerful R package with versions (V1 to V3) developed by an interdisciplinary team of investigators and data scientists, can integrate single-cell RNA-seq and single cell ATAC-seq data. More recently, a multimodal reference mapping approach, which can integrate the datasets of single-cell RNA and protein profiles, has been included in Seurat V4.

Other methods use a variety of statistical measurements to infer connections between different datatypes such as Bayesian inference, correlation coefficients and similarity indices. These data integration tools have successfully identified new disease subtypes, such as those in cancer, but are also applicable in the study of other diseases. In viral research, the availability of omics data has contributed to the formation of several sequence databases, such as GISAID, to facilitate the tracking of viral evolution and the emergence of new influenza and SARS-CoV-2 strains. Omics big data has also influenced the understanding of how viral infections develop over. Multi-omics approaches have been successful in characterizing changes in cell-type populations during SARS-CoV-2 infection. Because most diseases are characterized beyond the findings from a single omics data type, combining omics datatypes helps to provide more comprehensive details regarding the development and progression of diseases and related pathological features.

The key downstream analysis involves annotation of the functions of the genes of interest identified by various omics data. The popularity and performance of several gene set analyses have been summarized in a recent review. Of note, one of the most used and powerful algorithms is GSEA. DAVID, a pathway enrichment tool, remains very popular, although it appears to be out of date, given that the most updated version (DAVID 6.8) was released 5 years ago. In contrast, the R package clusterProfiler (version 4.0), which has recently been substantially updated, is an outstanding enrichment tool in this regard. In addition to gene set enrichment analysis and the comparison of enrichment results from multiple gene lists, clusterProfiler provides interfaces for data operation and visualization.

Conclusion. In conclusion, big data has revolutionized the healthcare industry, significantly impacting various sectors from clinical practice to biomedical research. With its vast sources and diverse types such as a genomic data, clinical records, and we are able device outputs, big data

allows for a deeper understanding all of disease mechanisms and the development of personalized medicine. The integration of multiple “omics” data types has facilitated comprehensive insides into disease pathology, aiding in more accurate diagnostics, treatment options, and epidemiological surveillance. As computational technologies continue to evolve, the efficient management and analysis of large-scale data are poised to further enhance healthcare outcomes and accelerate the development of tailored therapeutic approaches. Ultimately, the ongoing integration of big data into biomedical research promises to transform patient care, enabling more precise, individualized, and effective healthcare solutions.

List of references

[1] PC Magazine Encyclopedia. http://www.pcmag.com/encyclopedia/term/62849/big-data.

[2] O’Driscoll A, Daugelaite J, Sleator RD. ‘Big data’, Hadoop and cloud computing in genomics. J Biomed Inform 2013

БОЛЬШИЕ ДАННЫЕ В MULTI – OMICS ТЕХНОЛОГИЯ – ЭТО ПРЕИМУЩЕСТВО В МЕДИЦИНЕ И ЗДРАВООХРАНЕНИЯ

О. Аббасова

Преподаватель кафедры «Программное обеспечение информационных технологий»
Институт Телекоммуникаций и информатики Туркменистана

Аннотация.. Глобальный всплеск доступа к цифровым технологиям привел к сбору и документированию информации в различных государственных и частных секторах. На индустрию здравоохранения существенно повлияло появление больших данных. Клинические записи, данные об образе жизни пациентов, записанные с носимых устройств, IoTs, электронных медицинских карт, анализов изображений, молекулярного и диагностического профилирования. Популяционные исследования и информация о клинических испытаниях вместе способствовали крупным достижениям и прорывам в секторе здравоохранения.

Ключевые слова: Большие данные, большие данные в биомедицине, аналитика данных в биомедицинских исследованиях, мулти – омикс большие данные, управление биомедицинскими данными, большие данные в персонализированной медицине.