Analysıs Of Structured And Unstructured Bıg Data In Bıomedıcal Research And Systems Bıology

Hasanova Afag, Rahimova Nazila

Abstract

The article examines the analysis and use of structured and unstructured big data in biomedical research and healthcare, studies on the application of big data in almost all areas of healthcare, biomedical research and population health, molecular databases, bioinformatics, systems biology and personalized medicine. The methods of using big data have been studied in areas such as preparation. At the same time, the methods of using cluster analysis have been reflected in order to analyze asthma and allergy diseases and in general to match one phenotype with another endotype.

Keywords. Big data, biomedical research, bioinformatics, systems biology, asthma and allergy, omics, proteomics, genome

The Health Informatics Society Of Australia (HISA), a member of the IMIA (International Medical Informatics Association), organized A "Big Data" Conference İn Mel Bourne, Australia İn April 2013 And 2014. The conference addressed research, industry, government, and clinical practice, introducing more than 200 healthcare professionals, healthcare leaders and managers, data and information professionals, and health information professionals to the exploding world of big data in health and healthcare**.** During these two conferences, it became clear that big data is a hot topic in healthcare and biomedical research. The increasing use of the term "big data" in the biomedical literature is indicative of the importance of large-scale datasets in healthcare and biomedicine. It has also been observed that the term "Big Data" can mean different things to different groups of people. Nevertheless, it is becoming widely accepted that healthcare, biomedical research, and population health generate massive, complex, distributed, and often dynamic datasets, and that the size and complexity of these data will present both challenges and new opportunities for healthcare organizations. [1]

Structured and unstructured big data in biomedical research and healthcare. Since the publication of the first human genome sequence in 2003, the field of genomics has represented the main driving force behind the generation of big data in biomedicine. Advances in laboratory analytical techniques (eg, dna sequencing) and mobile technologies (eg, data from physical activity monitors and apps) are now primarily responsible for the ever-increasing real-time production of high volumes of data. However, the use of big data has now reached almost all areas of healthcare, biomedical research and population health. Health services researchers can integrate administrative and clinical databases to improve health policy and develop predictive models. In addition, the pharmaceutical industry also includes large repositories of clinical and molecular data for rational drug design and pharmacogenomic approaches. [2]Since the early days of computing, the term "data" has been used primarily to refer to structured data. However, in the last few years, there has been a dramatic increase in the production of unstructured data, far exceeding the amount of structured data available. Experts at the international data corporation estimate that unstructured data currently accounts for more than 80 percent of existing data.[2]Although there is no formal definition, in general the term "structured data" refers to data with a defined schema or data model (ie, open semantics). Data stored in a database is usually structured. Measurements and signals are examples of structured data. In contrast, unstructured data refers to data that is not readily accessible to computational data management systems—the information it contains is not presented in a form with a clear data schema that allows for direct computational interpretation and analysis. Special analytical techniques are usually required to extract the information contained in this type of data and convert it into a computable form. Natural language text, images, and audio streams are examples of unstructured data.[3]There is a wealth of data to understand human health from unstructured resources. Texts, images, and audio and video streams are all commonly produced in a clinical context. These resources require the development of strategies for extracting and summarizing the information contained in them in order to apply structure and meaning

using the internal structure or patterns inherent in the source. Technological solutions to provide automatic interpretation of such resources, while helping the people tasked with interpreting these data sources, allow to expand the analysis and consider a wider data set (hospital, population, or scientific research community). This has resulted in some highly innovative research demonstrating the power of large-scale data analysis in medicine.[4]Molecular Databases, Bioinformatics, Systems Biology And Personalized Medicine. Science and research are changing. The fields of bioinformatics, systems biology, and personalized medicine are areas where we can most clearly see the transition from hypothesis-driven to data-driven research. Work in computational drug discovery and medicinal chemistry shows that this trend is clearly being followed. The unprecedented generation of molecular data has also led to new challenges in data visualization. Various software frameworks have been developed to facilitate data analysis and speed up time. One of these was dive (data intensive visualization engine, a software framework designed to facilitate the analysis of big data), a data visualization engine applied to the study of proteins.[5]Proteomics was also the focus of a proteomics standards initiative (human proteome organization) workshop in the uk in april 2013. İs about understanding how they interact.).under the auspices of this organization, standards for data representation in proteomics have been updated and improved to pave the way for big data approaches. The use of standards is also important for sharing big data in genomics. An article by tenenbaum et al reinforces this point and advocates the use of open data standards developed by the community[6]. Sharing genomic data between laboratories is critical to advances in infectious disease control. Iwasaki et al. Report on the application of novel bioinformatics strategies to analyze large influenza virus genome sequence files and identify potential threats. [7]Systems biology is a relatively recent development that addresses the increasing complexity of biomedical research questions. The term was used in the 1960s to describe the mathematical modeling of physiological systems. Today, it includes expertise in many fields, including biology, mathematics, statistics, informatics, and computer science. The "systems" community is diverse, and thus there is no single definition of the term "Systems Biology." However, it is presented as the study of biomedical problems involving complex systems and their interactions through the exploration and integration of high volumes of data that can span large spatio-temporal scales [8]. These “big datasets” typically arise from “omics” research fields involving high-throughput measurements of biomolecules: for example, genomics for dna, transcriptomics for rna transcripts, and proteomics for translated proteins (figure 1). Mathematical and computational expertise is then required to explore this high-volume data using techniques such as dimensionality reduction, data and text mining, modified statistical analyses, machine learning, and mathematical modeling that account for spatio-temporal complexity and multiple testing load. Further experiments can be performed in which the biological system can be disrupted (e.g., through receptor antagonists or gene knockouts) to identify functionally relevant elements of the system [8]. Therefore, systems biology is multifaceted and interdisciplinary by nature. Fig 1. "omics" and their interactions in allergy. A description of the different omics available in allergy and asthma research. The lines connecting the omics represent the various biological connections, associations, or interactions that may exist. The transcriptome is the sum of all the rna molecules expressed from the genes of an organism. The genes in the genome are expressed in certain cells and at certain times. This new concept is called "epigenetics", which means "genetics over genes". All formations involved in epigenetic structures and mechanisms are called "epigenome". Microbiome-microorganism.[9] Asthma and allergies are well suited to systems approaches as biomedical problems. These diseases have a complex pathogenesis with a lot of biological complexity, polygenicity and gene-environment interactions. Systems approaches used in asthma and allergy research may include:

1) detection of disease associations in each omic domain;
2) determination of relationships within omic fields;
3) examining the heterogeneity of disease states and phenotypes, usually by clustering or classifying

the multidimensional structure of omics data;

4) research of interaction between system components in omic data by network analysis;

5) mathematical modeling to model physiological systems or disease states, generate and test predictions (fig 2).

Fig 2. A review of systems-based approaches to address research questions in allergy and asthma. Various ways in which the systems biology of allergy can be interrogated include:

A) detection of disease associations within each research area;

B) determination of relationships within omics;

C) examining the heterogeneity of disease states or phenotypes, typically by examining the structure of omics data through clustering or classification;

D) research of interaction between system components in omic data by network analysis;

E) mathematical modeling to model physiological systems or disease states and to generate and test predictions.

Charts are for illustrative purposes only and do not convey real information.machine learning is a set of methods that use computation to learn and form solutions from data provided with or without explicit human input. It is already in common use with various biomedical and environmental applications [10]; however, it is particularly useful when dealing with complex, high-throughput, and multidimensional data, especially when pre-existing human knowledge may be unavailable or insufficient to decipher the data.

Fig 2. Data-driven and hypothesis-driven machine learning for omics data integration. A) data-driven (unsupervised) cluster analysis used to generate de novo clusters reflecting common pathophysiology ("endotypes"); b) hypothesis-driven (supervised) classification to compare known phenotypes or endotypes and allow prediction of phenotype/endotype membership for additional samples.applications of machine learning in biomedicine typically involve the study of data structure or the creation of predictive or explanatory models of biological systems. Cluster analysis and classification are techniques used to divide data samples or individuals into different groups or categories, thus providing a summary of the data structure. Such methods typically use machine learning at its most fundamental level: for example, hierarchical clustering. İterative process where the "objective function" will be to minimize intra-cluster similarity and/or maximize inter-cluster dissimilarity.[11] there is usually a subtle difference between clustering and

classification: cluster analysis is a data-driven approach where omics data is used. To generate clusters in an unsupervised manner. Clusters can then be interpreted to generate and test hypotheses. In contrast, classification is a hypothesis-driven approach: known phenotypes or pre-selected categories are used to define a classification model based on training data, which can then be applied to other data sets or tested to look for additional biological associations (figure 3).

Phenotype or external structure - the reflection of the characteristics caused by genetic (genotype) and environmental factors in the external appearance of an organism.

The phenotype is mainly determined by the genes, but in some cases other factors can prevent the phenotype from fully matching the genotype. [11]endotype: a disease subgroup defined by different functional or pathobiological/pathophysiological mechanisms.

Conclusıon

The availability of electronic medical records and administrative data sets is enabling a wave of innovation in health services research projects. In these studies, patients can be identified as being at increased risk for resuscitation or their estimated length of stay in intensive care units can be modeled. Many of these big data analytics tasks require advanced computing frameworks for high data volumes

and intensive data processing. As for clinical applications, big data methods combined with mobile technologies can enable medical professionals to remotely read patients' signals and images, as well as store, deliver, retrieve and manage various medical files for teleconsultation and telediagnosis. Also cloud

computing service architecture can be used to support big data analysis in medical diagnoses with good results.

Reference

[1] F. Martin-Sanchez, K. Verspoor, "Big Data İn Medicine İs Driving Big Changes.", 2014

[2] Jongmin Lee , Hyun Myung Jung, Sook Kyung Kim, Kwang Hayoo, Ki-Suck Jung, Sang Haak Lee, Chin Kook Rhee, "Factors Associated With Chronic Obstructive Pulmonary Disease Exacerbation, Based On Big Data Analysis", 2019.

[3] Christopher Baechle , Ankur Agarwal And Xingquan Zhu, "Big Data Driven Co-Occurring Evidence Discovery İn Chronic Obstructive Pulmonary Disease Patients", 2017.

[4] Yue Zhang, Shu-Li Guo, Li-Na Han, Tie-Ling Li, "Application And Exploration Of Big Data Mining İn Clinical Medicine", 2016.

[5] Howard H.F. Tang, Peter D. Sly, Patrick G. Holt, Kathryn E. Holt, Michael Inouye, "Systems Biology And Big Data İn Asthma And Allergy: Recent Discoveries And Emerging Challenges", 2020.

[6] Tenenbaum JD, Sansone SA, Haendel M., "A Sea Of Standards For Omics Data: Sink Or Swim", 2014.

[7] Iwasaki Y, Abe T, Wada Y, Wada K, Ikemura T. Novel, "Bioinformatics Strategies For Prediction Of Directional Sequence Changes İn İnfluenza Virus Genomes And For Surveillance Of Potentially Hazardous Strains", 2013.

[8] Jane A. Leopold, Bradley A. Maron, Joseph Loscalzo, "The Application Of Big Data To Cardiovascular Disease: Paths To Precision Medicine", 2020.

[9] Https://Erj.Ersjournals.Com/Content/Erj/55/1/1900844.Full.Pdf

[10] Https://Www.Ahajournals.Org/Doi/Epub/10.1161/CIRCRESAHA.121.319969

[11] Https://Tr.Wikipedia.Org/Wiki/Proteomik