Corresponding Author:
dr.syedrizwan@kiu.edu.pk

Journal of natural sciences [ISSN: 2308-5061]

BIOINFORMATICS LANDSCAPES FROM SEQUENCE TO DRUG DISCOVERY

FATIMA BANO[1], SYED RIZWAN ABBAS[2] AND GULISTAN RANI[3]

[1]Department of Animal Sciences; Zoology; Karakoram International University; Gilgit.
[2]Department of Plant Science; Karakoram International University; Gilgit.
[3]Clinical laboratory (Hemate and transfusion medicine) scientist/technologist AKUH.

ABSTRACT
Drug discovery has seen a radical transformation because to bioinformatics, which offers robust tools for discovering possible targets for drugs and evaluating large amounts of data. With an emphasis on significant advancements and potential paths forward, this study seeks to provide an overview of the state of bioinformatics in the context of drug discovery It also discusses the need for more sophisticated analysis methods as well as the difficulties and restrictions associated with data and software mining. The significance of bioinformatics in changing the clinical landscape and the possibility of future advances in precision medicine are emphasized in the review's conclusion. In addition to speeding up the identification of therapeutic targets and the screening and improvement of drug candidates, bioinformatic analysis can also help characterize adverse effects and anticipate drug resistance. Pharmaceutical repurposing and mechanism-based drug development have benefited greatly from high-throughput data, including ribosome profiling, transcriptomic, proteomic, genome architecture, genomic, and epigenetic data. More realistic protein-ligand docking experiments and more insightful virtual screening were made possible by the accumulation of protein and RNA structures, the development of homology modeling and protein structure simulation, and the availability of large structure databases of small molecules and metabolites.

KEYWORDS: Bioinformatics landscapes; High-throughput data generation; Virtual screening; Bioinformatics analysis.

INTRODUCTION
In order to find possible drug targets and improve drug candidates, bioinformatics has become a crucial component of modern drug discovery. This review looks at the state of bioinformatics in drug discovery today, emphasizing important developments and future

directions. It uses sequence information from metagenome sequencing to efficiently correct errors in single-cell amplified genome (SAG) reads, even in areas with ultra-low coverage **(Bremges, 2016).** diseases Alternative splicing is a critical process of post-transcriptional gene control, and its misregulation is connected to a wide array of human disorders **(Christinat, 2016).**

Current synteny visualization tools have challenges - they either focus on narrow genomic regions without demonstrating whole-genome trends, or they are complex to use and provide visuals that are hard to interpret. Two web-based tools that allow researchers to explore synteny across entire genomes have been developed by the Comparative Genomics Platform (CoGe) in response to these obstacles **(A Haug-Baltzell, 2017)**. Three basic methods are used by nature to control how proteins and genes are expressed or function: genetic events, which include copy number variation, translocation, and mutation. Natural substrates, which include both big and tiny molecules, are used for modification. Post-translational modifications (PTMs), primarily acetylation and methylation of DNA and histones, are the primary means of epigenetic regulation **(Schaduangrat, 2017).**

Human molecular genetic testing is being completely transformed by next-generation sequencing (NGS), a cutting-edge technology that makes it possible to parallelize sequencing reactions in a way never before possible. This results in highly multiplexed testing approaches with drastically shorter turnaround periods and lower costs **(Gavin R Oliver, 2015).**

The last ten years have seen an exponential growth in the amount of genomics and chemical informatics data, which has resulted in the creation of numerous new systematic approaches to drug discovery. In order to find new targets linked to diseases, some of these techniques concentrate on examining the structural relationships between medications and their targets. Others predict the biochemical interactions between small molecules and their respective targets, such as the Connectivity Map (CMap) approach **(Aliyu Musa, 2018).** During oncogenesis, various populations of genetically diverse cancer cells emerge, proliferate, and intera ct with the surrounding tumor microenvironment. This interplay results in metabolism hijacking, immune evasion, metastasis, and resistance to therapeutic drugs **(Diletta Rosati a, 2022).**

Apart from repurposing drugs and increasing the usefulness of already-approved drugs, there's a special idea called "drug development repositioning" (sometimes called "drug rescue"). This happens when a medication candidate is developed for one indication and then repurposed for another indication after failing to meet the original goals **(Ziaurrehman Tanoli, 2021).** Recent technological developments have made it possible to quickly and affordably profile biological systems, which has resulted in the routine gathering of various forms of omics data from patient cohorts in research on human diseases. These data might lead to the creation of a new illness classification. According to Mansoor Saqi (2019), conditions that were once thought to be one homogenous entity may instead be a grouping of multiple unique disease subtypes **(Mansoor Saqi, 2019)**

High-performance computing has made it easier to conduct virtual experiments in silico by providing a means of interpolating theoretical models with laboratory results. When direct measurements are not possible, Schulten created the phrase "computational microscope" to refer to the way computational simulations can improve experimental research **(Maral Aminpour 1, 2019).** There have always been difficulties and unknowns on the path to developing new medications. Specifically, de novo drug development is an expensive, time-consuming, and dangerous process. An estimated $2 billion to $3 billion is spent on average in the entire process

of producing a new medication, and it usually takes at least 13 to 15 years from the time a treatment is first discovered to its approval for sale **(Jaswanth K. Yella 1, 2018).**

Finding a condition with distinct symptoms that profoundly affect quality of life is the first step in the drug discovery process. Traditionally, the perfect medication is one that reduces these symptoms while lowering the patient's risk of experiencing severe adverse effects. This could be a simple chemical, a complex protein, or a combination of different chemicals (**Xuhua Xia, 2017).**

Three basic mechanisms are used by nature to control the expression and activity of genes and proteins: genetic events (mutation, copy number variation, and translocation); modulation by natural substrates (small and large molecules); and epigenetic regulation (primarily through post-translational modifications, or PTMs) such as methylation and acetylation of DNA and histones **(Veda Prachayasittikul, (2017).** Several salient characteristics of the contemporary landscape become apparent: UK bioinformatics is well-established globally in several critical domains, including proteomics, crystallography, and sequence databases **(DR Mark Harvey, 2021).**

Lead discovery is now benefited by structural bioinformatics, which has historically been crucial to lead optimization and target identification. Through the use of high-throughput structure identification techniques, it provides effective fragment binding screening techniques. This topic deals with the development of peptide bonds, dihedral angles, and stabilizing forces in addition to the structural examination of different molecules, such as proteins and amino acids. It also looks into protein structure databases and secondary and tertiary structures, all of which are essential for comprehending the purposes of the basic structural components of living cells. **(Ravish Kumar, 2002).**

HIGH-THROUGHPUT DATA GENERATION

The gathering and analysis of high-throughput data has been an important catalyst of bioinformatics innovation in the field of drug development. In mechanism-based drug discovery and response analysis, data gathered from genomic, epigenetic, genome architecture, cistromic, transcriptomic, proteomic, and ribosome profiling has all been significant. In addition, there is a rapid development in the use of statistical and machine learning (ML) strategies in the field of materials science **(Rama K. Vasudevan, 2019).** Because high-throughput genomic sequencing drastically lowers prices and increases the volume and speed of data collection, it is profoundly changing the discipline of phylogenetics. Since the introduction of pyrosequencing a decade ago, there has been a significant increase in the range of techniques available for gathering genome-scale data. Researchers can now collect large amounts of genomic data because to technological advancements, which allows for a greater understanding of evolutionary interactions and mechanisms **(Ronaghi *et al.*, 1998).** High-throughput virtual screening has been used for a while in the pharmaceutical sciences, but its definition is quite dependent on what modern hardware could accomplish when the analysis was being done. Consequently, we propose that the philosophy behind the computations rather than their quantity or complexity should serve as a stronger defining characteristic for this discipline **(Pyzer- Knapp *et al.*, 2015)**

BIOINFORMATIC ANALYSIS

Bioinformatic analysis is essential for speeding up the identification of drug targets and the screening of drug candidates. High-throughput data is processed using a range of algorithms and software tools to pinpoint potential drug targets and enhance drug candidates. This process often involves the application of machine learning and machine learning-based techniques to

predict drug efficacy and potential side effects. Notably, a long transcription product was discovered and identified by Okazaki in 2002 **(Okazaki *et al.,* 2003).**

When a rodent cDNA library is being sequenced. We dubbed the transcript long untranslated RNA (lncRNA). LncRNAs lack an open reading frame yet are longer than 200 nucleotides and structurally comparable to messenger RNAs. Several popular sequence analysis applications have been made available to the public since 1998 by the European Bioinformatics Laboratory (EMBL–EBI) (1,2). These consist of similarities between sequences search engines like BLAST and FASTA (3) **(Mickael Goujon, 2010).** The gathering of "big data" across various computing devices has grown more crucial in the genomics era in order to comprehend the dynamic characteristics of pancreatic ductal adenocarcinoma **(Tessa Y.S *et al.,* 2015).** Bioinformatics, with the help of the sciences of computation and biology, is becoming a vital discipline that improves our comprehension of naturally occurring data. It combines technological innovation, statistics, software engineering, mathematics, biology, and statistics into a single, coherent field **(Fatih Gurcan, 2022).**

DATA MINING AND SOFTWARE DEVELOPMENT

Software development and data mining are essential components of bioinformatics in the field of drug discovery. Developing new software tools and algorithms is crucial to keeping up with the exponential growth of high-throughput data. This includes creating large structural databases for metabolites and small compounds in order to improve virtual screening procedures and enable more precise protein-ligand docking studies.

Big Data is the term used to describe extensive, intricate, and dynamic datasets that come from numerous separate sources. large-scale data is rapidly gaining traction in all scientific and engineering domains, including the physical, biological, and biomedical sciences, thanks to developments in networking purposes, storage of data, and data collection technologies **(Xindong Wu1 *et al.,* 2013).** Establishing, investigating, and using automated techniques to find patterns in big datasets of educational data that would otherwise be challenging or even impossible to evaluate because of the massive amount of data they include is known as educational data mining, or EDM **(Cristobal Romero, 2013).**

VIRTUAL SCREENING

Even while virtual screening is now frequently utilized in drug development projects as a hit-finding strategy, there are still comparatively few accounts of it producing molecules that have made it to market or are already being used in clinical settings **(David E Clark, 2008** Computer-aided drug design (CADD) is one extensively used technique to lower the time and expense involved with the process of creating medications. Improved experiment targeting made possible by CADD can reduce the amount of time and resources spent on developing new medications. Structure-based virtual screening (SBVS) is regarded as one of the most potential in silico methods for drug creation because it is a reliable and beneficial strategy in this regard **(EHB Maia, 2020).** In the 19th century, Langley postulated that various compounds would react with particular receptors in the body which gave rise to the idea of a pharmacophore. Later on, with the discovery of Salvarsan by Paul Ehrlich, was the selectivity of drug-target interactions recognized **(Deborah Giordano, 2022).**

GENOMICS

The bulk of the functional domains found in human genes have homologs in wildly distinct organisms, according to comparative genomic analysis. But throughout time, distinct evolutionary lineages variably reorganize these shared functional regions, leading to an expanding variety of domain architectures. g evolutionary lineages, resulting in an increasing

number of domain architectures over time **(D. V. Babushok, 2007).** The study of genetic sequences for systematic identification and categorization is essential to both biodiversity monitoring and the biomedical domain. The "barcode genes," which stand for particular sections of the complete genome, are the main subject of these investigations. Alignment-free approaches have become more popular recently because they can overcome the drawbacks of conventional sequence alignment techniques **(Massimo La Rosa *et al.,* 2015).**

PROTEOMICS

In 1995, the terms "protein" and "genomics" were combined to create the phrase "proteomics." However, Norman G. suggested much earlier in 1979 that the proteome of human being should be thoroughly characterized **(Ian Craig Lawrance, 2005).**

In microbial research, quantitative proteomics has become a vital analytical tool. A wide range of topics in both basic and applied research are covered by contemporary microbial proteomics, from characterizing individual organisms in vitro to investigating the physiological consequences of anxiety and dehydration to describing the proteome that makes up a cell at a particular time. **(Andreas Otto *et al.,* 2014).**

Mass spectrometry-based quantitative proteomics is becoming more and more important in the field of life sciences against a number of obstacles. These include the advancement of protein biomarker assay systems, the investigation of signaling networks, and expression profiling, among other areas **(Marcus Bantscheff, 2012).**

TRANSCRIPTOMICS

The transcriptome, or collection of all an organism's RNA transcripts, is studied using transcriptomics methods. An organism's genetic makeup can be expressed by transcription from the DNA that makes up its genome. In this scenario, non-coding RNAs play a number of extra functions while mRNA serves as a transient bridge in the information network. The transcriptome gives an overview of all the transcripts that are present in a cell at any one time **(Rohan Lowe *et al.,* 2017)**

CHALLENGES AND LIMITATIONS

Bioinformatics has come a long way, but there are still a number of obstacles and restrictions that need to be overcome. These include the requirement for increasingly sophisticated analytic methods, the incorporation of data from many sources, and the creation of software applications that are more resilient and expandable.

FUTURE DIRECTIONS

Numerous significant themes are expected to influence the direction of bioinformatics in drug discovery in the future. These include the growing application of artificial intelligence and machine learning, the combining of data from many sources, and the creation of increasingly sophisticated analytic methods. Furthermore, the desire for more individualized care and the advancement of precision medicine will keep bioinformatics innovation moving forward.

CONCLUSION

Drug discovery has seen a radical transformation because to bioinformatics, which offers robust tools for discovering possible targets for drugs and evaluating large amounts of data. Fast progress in data gathering, processing, and software development describe the state of bioinformatics in drug discovery today. Bioinformatics is expected to become more significant in changing the clinical landscape and providing more individualized and effective therapies as the field develops. I outline certain inherent drawbacks in data and software mining, provide the conceptual framework that guides the collection of these high-throughput data, summarize the value and potential of mining these data for drug discovery, point out new approaches to fine-

tune analysis of these varied types of data, and highlight frequently used software and databases relevant to drug discovery

REFERENCES

Bremges, A., Singer, E., Woyke, T., & Sczyrba, A. (2016). MeCorS: Metagenome-enabled error correction of single cell sequencing reads. Bioinformatics, 32(14), 2199-2201

Christinat, Y., Pawłowski, R., & Krek, W. (2016). jSplice: a high-performance method for accurate prediction of alternative splicing events and its application to large-scale renal cancer transcriptome data. Bioinformatics, 32(14), 2111-2119.

Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E., & Lyons, E. (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics, 33(14), 2197-2198.

Prachayasittikul, V., Prathipati, P., Pratiwi, R., Phanus-Umporn, C., Malik, A. A., Schaduangrat, N., ... & Nantasenamat, C. (2017). Exploring the epigenetic drug discovery landscape. Expert opinion on drug discovery, 12(4), 345-362.

Musa, A., Ghoraie, L. S., Zhang, S. D., Glazko, G., Yli-Harja, O., Dehmer, M., ... & Emmert-Streib, F. (2018). A review of connectivity map and computational approaches in pharmacogenomics. Briefings in bioinformatics, 19(3), 506-523.

Rosati, D., & Giordano, A. (2022). Single-cell RNA sequencing and bioinformatics as tools to decipher cancer heterogenicity and mechanisms of drug resistance. Biochemical Pharmacology, 195, 114811.

Tanoli, Z., Seemab, U., Scherer, A., Wennerberg, K., Tang, J., & Vähä-Koskela, M. (2021). Exploration of databases and methods supporting drug repurposing: a comprehensive survey. Briefings in bioinformatics, 22(2), 1656-1678.

Saqi, M., Lysenko, A., Guo, Y. K., Tsunoda, T., & Auffray, C. (2019). Navigating the disease landscape: knowledge representations for contextualizing molecular signatures. Briefings in bioinformatics, 20(2), 609-623.

Aminpour, M., Montemagno, C., & Tuszynski, J. A. (2019). An overview of molecular modeling for drug discovery with specific illustrative examples of applications. Molecules, 24(9), 1693.

Yella, J. K., Yaddanapudi, S., Wang, Y., & Jegga, A. G. (2018). Changing trends in computational drug repositioning. Pharmaceuticals, 11(2), 57.

Xia, X. (2017). Bioinformatics and drug discovery. Current topics in medicinal chemistry, 17(15), 1709-1726.

Harvey, M., & McMeekin, A. (2002). UK Bioinformatics: current landscapes and future horizons. London, UK: Department of Trade and Industry.

Ravish Kumar, & Giordano, A. (2002). Single-cell RNA sequencing and bioinformatics as tools to decipher cancer heterogenicity and mechanisms of drug resistance. Biochemical Pharmacology, 1955.

Vasudevan, R. K., Choudhary, K., Mehta, A., Smith, R., Kusne, G., Tavazza, F., ... & Hattrick-Simpers, J. (2019). Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. MRS communications, 9(3), 821-838.

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. Annual Review of Ecology, Evolution, and Systematics, 44(1), 99-121.

Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., & Aspuru-Guzik, A. (2015). What is high-throughput virtual screening? A perspective from organic materials discovery. Annual Review of Materials Research, 45(1), 195-216.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., & Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL–EBI. Nucleic acids research, 38(suppl_2), W695-W699.

Le Large, T. Y., Mato Prado, M., Krell, J., Bijlsma, M. F., Meijer, L. L., Kazemier, G., ... & Giovannetti, E. (2016). Bioinformatic analysis reveals pancreatic cancer molecular subtypes specific to the tumor and the microenvironment. Expert review of molecular diagnostics, 16(7), 733-736.

Gurcan, F., & Cagiltay, N. E. (2022). Exploratory analysis of topic interests and their evolution in bioinformatics research using semantic text mining and probabilistic topic modeling. IEEE Access, 10, 31480-31493.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-1073.

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data mining and knowledge discovery, 3(1), 12-27.

Clark, D. E. (2008). What has virtual screening ever done for drug discovery? Expert opinion on drug discovery, 3(8), 841-851.

Maia, E. H. B., Assis, L. C., De Oliveira, T. A., Da Silva, A. M., & Taranto, A. G. (2020). Structure-based virtual screening: from classical to artificial intelligence. Frontiers in chemistry, 8, 343.

Giordano, D., Biancaniello, C., Argenio, M. A., & Facchiano, A. (2022). Drug design by pharmacophore and virtual screening approach. Pharmaceuticals, 15(5), 646.

Babushok, D. V., Ostertag, E. M., & Kazazian, H. H. (2007). Current topics in genome evolution: molecular mechanisms of new gene formation. Cellular and molecular life sciences, 64, 542-554.

La Rosa, M., Fiannaca, A., Rizzo, R., & Urso, A. (2015). Probabilistic topic modeling for the analysis and classification of genomic sequences. BMC bioinformatics, 16, 1-9.

Lawrance, I. C., Klopcic, B., & Wasinger, V. C. (2005). Proteomics: an overview. Inflammatory bowel diseases, 11(10), 927-936.

Otto, A., Becher, D., & Schmidt, F. (2014). Quantitative proteomics in the field of microbiology. Proteomics, 14(4-5), 547-565

Bantscheff, M., & Kuster, B. (2012). Quantitative mass spectrometry in proteomics. Analytical and bioanalytical chemistry, 404, 937-938.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. PLoS computational biology, 13(5), e1005457.