ISSN: 2663-2187

https://doi.org/10.48047/AFJBS.6.14.2024.10775-10791

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

INTEGRATING BIOINFORMATICS IN DNA SEQUENCING: CHALLENGES AND OPPORTUNITIES

Waqar Rasool Minhas[1], Aliu Olalekan olatunji[2], Almas Faryal Nizam[3], Sana Kaina[4], Andongma Esack Fonda[5], Likowsky Desir[6], Md. Rakib Rased Rana[7], M.D Issa Makdesi[8]

[1]Researcher, NIBGE-C PIEAS, Industrial Biotechnology Division (IBD), Islamabad, Pakistan
Email: waqar.minhas898@gmail.com

[2]Doctor, Department of Medical Microbiology and Immunolog, University of Toledo, USA
Email: aliu_my2004@yahoo.com

[3]Department ofZoology, Shaheed Benazir Bhutto Women University, Charsadda Road, Peshawar, Pakistan

[4]MBBS Scholar ANMCH, AL-Nafees Medical College and Hospital, Isra University, Islamabad

[5]Department of Microbiology, University of Buea, Cameroon Email: esackfonda@gmail.com

[6]MPH, MSc, Department of Surgery, Wyckoff Heights Medical Center, United States
Email: kkLikowsky@gmail.com

[7]Department of Pharmacy,Varendra University, Chandrima, Bypass Road Rajshahi-6204, Bangladesh, Gmail: rakibrana.vu@gmail.com

[8]Department of Medicine, Georgian American University, Georgia
Email: issa.makdesi@gau.edu.ge

Volume 6, Issue 14, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 29 Aug 2024

doi: 10.48047/AFJBS.6.14.2024.10775-10791

ABSTRACT:

Background: Analyzing DNA sequences involves processing vast amounts of data, necessitating advanced computing methods and data modelling techniques. The aim is to accelerate computation and inference times while improving the reliability of analyses, thereby serving as a catalyst for scientific progress in biological research.

Objective: The objective of the University of Technology's Systems Engineering and Computer Science program is to identify information technologies that can significantly advance biological science. This article focuses on examining computational methods that are instrumental in developing bioinformatics applications.

Methods: The research employed a systematic approach to evaluate various computational methods used in bioinformatics applications. This involved reviewing existing literature, analyzing case studies, and consulting experts in the field. Key criteria such as computational efficiency, accuracy, scalability, and applicability to biological data were considered in the evaluation process.

Results: Through the systematic analysis, several computational methods were identified as particularly useful for creating bioinformatics applications. These methods include machine learning algorithms, parallel computing techniques, optimization algorithms, and data mining approaches. Each method offers unique capabilities in handling largescale genomic data, improving data analysis accuracy, and accelerating research processes in biological science.

KEYWORDS: sequences, proteins, DNA, RNA, bioinformatics, biology, biological data, and technologies.

INTRODUCTION:

The scientific community conducting biological research is confronted with ever-increasing challenges daily. These challenges include managing massive volumes of data that are expanding exponentially in size and complexity due to technological advancements that enable more accurate calculations. The community seeks answers to studies on DNA sequences and molecular structure. Fortunately, significant advancements in intelligent data processing and analysis techniques have been made possible by technological growth in electronics, software development, and telecommunications. These developments have benefited scientific investigations that aim to fully comprehend the data structures of live beings [1].

Large-scale data management is difficult and thus calls for computational processes that are highly performant in terms of both response times and space.

Table 1: Challenges and Technological Advances in Biological Research

Aspect	Description

Challenges	Managing massive volumes of data, expanding exponentially in size and complexity due to technological advancements that enable more accurate
	calculations. The community seeks answers to studies on DNA sequences and molecular structure.
Technological Advancements	Significant advancements in intelligent data processing and analysis techniques are enabled by growth in electronics, software development, and telecommunications. These developments benefit scientific investigations aiming to comprehend the data structures of live beings [1].
Need for High-Performance Computing	Large-scale data management requires computational processes with high performance in terms of both response times and space.

Table 2: Computational Methods for DNA Sequence Analysis

Computational Method	Description
Machine Learning Algorithms	Techniques that enable the analysis and interpretation of complex genomic data, improving accuracy and predictive capabilities.
Parallel Computing Techniques	Approaches that utilize multiple processors simultaneously to handle large-scale genomic data efficiently.
Optimization Algorithms	Methods that focus on finding the best solutions from a set of possible solutions, enhance computational efficiency.
Data Mining Approaches	Techniques that involve extracting useful information and patterns from large datasets, aiding in the understanding of biological data structures.

Table 3: Bioinformatics Concepts, Applications, and Scope (Section 2)

Concept	Description
Bioinformatics Concepts	Fundamental ideas and principles underlying the use of computational tools in biological research.

Applications	Practical uses of bioinformatics in various areas of biological science, including genomics, proteomics, and molecular biology.
Scope	The extent and potential impact of bioinformatics in advancing scientific knowledge and research methodologies.

Table 4: DNA Sequence Alignment (Section 3)

Aspect	Description
DNA Sequence Alignment	Techniques and methods used to compare and align DNA sequences to identify similarities, differences, and patterns.
Tools and Algorithms	Specific computational tools and algorithms are utilized for DNA sequence alignment.

Table 5: Computational Tools for Bioinformatics Solutions (Section 4)

Tool	Description
Matlab Usage	The application of Matlab software in developing and implementing bioinformatics solutions.
Biological Databases	Databases that store and manage biological data, provide essential resources for bioinformatics research.
Data Warehouses	Systems that aggregate and store large volumes of biological data for analysis and retrieval.
Data Mining	Techniques for analyzing and extracting meaningful information from large datasets in bioinformatics.
Learning Machines	Machine learning models and algorithms are applied to biological data to uncover patterns and make predictions.

Table 6: Conclusions and Next Steps (Section 5)

Aspect	Description
Conclusions	Summary of findings from the analysis of computational methods for bioinformatics.

Next Steps	Future directions and recommendations for advancing bioinformatics research
	using the identified computational methods.

To create useful tools for future research, this article examines some of the most popular computational methods and approaches for analyzing DNA sequences. Section 2 of this page discusses bioinformatics's concepts, applications, and scope. The alignment of DNA sequences is discussed in Section 3. The computational tools for creating and implementing bioinformatics solutions are covered in Section 4 and include Matlab usage, biological databases, data warehouses, data mining, and learning machines. Section 5 outlines the conclusions and next steps [2].

BIOINFORMATICS

Understanding connections, structures, and patterns in biological data is one of bioinformatics' most crucial responsibilities. In recent years, several fields, including computer science, mathematics, statistics, chemistry, and non-traditional biological sciences, have been drawn to bioinformatics. This is a result of the abundance of both public and private biological data available and the urgent need to convert the data into valuable biological knowledge and information. Using sophisticated computational approaches in conjunction with these fields, initiatives including biological control, genomic analysis, and drug discovery and development are created. This entails managing and analyzing massive amounts of biological data on DNA, RNA, protein patterns, protein structures, genetic expression profiles, and protein interactions using computer technology and statistical techniques [3].

Bioinformatics' Range:

The subfields of bioinformatics consist of two complementary areas: The creation of databases and I.T. tools and the use of these to produce biological knowledge to comprehend living systems more fully. Software for recording sequences, structural and functional analysis of the sequences, and the creation and preservation of biological databases are all examples of the tools

being developed. Examining biological data frequently raises fresh issues and difficulties, propelling the advancement of more advanced computational instruments [4].

In what ways is bioinformatics applicable?

In addition to being a crucial science for fundamental studies in molecular biology and genomics, bioinformatics is also significantly influencing many fields in biotechnology and the biomedical sciences. Its applications necessitate expertise in agricultural biotechnology, forensic DNA analysis, and drug formulation. Compared to traditional trial and error, a bioinformaticsbased method greatly decreases the time and expense required to produce medications with greater potency, fewer side effects, and reduced toxicity. Molecular phylogenetic analysis results have been admitted as evidence in criminal courts in forensic medicine. Forensic identity analysis has used certain complex Bayesian statistics and plausibility-based DNA analysis techniques [5]. Genomics and bioinformatics are poised to transform healthcare systems by advancing personalized therapy. The combination of high-speed genomic sequencing and advanced information technology enables a doctor at a clinic to quickly sequence a patient's DNA. This allows them to discover potential detrimental mutations in the genome, enabling them to play a crucial role in early and successful diagnosis [6].

Medical intervention for illnesses:

Agriculture also utilizes bioinformatics techniques. Using plant genome databases and gene expression analysis has significantly contributed to advancing novel crop varieties characterized by enhanced yield and increased disease resistance. Bioinformatics encompasses the creation of databases or knowledge bases to store and retrieve biological data, algorithms for analyzing and determining these data's linkages, and statistical tools for identifying and interpreting data sets [7].

Analysis of DNA Sequences:

DNA sequence analysis involves identifying and examining functional and structural disparities among biological sequences. This can be achieved by comparing novel (unidentified) sequences with extensively researched and annotated (established) sequences. This approach encompasses aligning sequences, examining sequence databases, identifying patterns, recreating

evolutionary relationships, and comparing genome development. Scientists have determined that two analogous sequences possess identical functional roles. The comparison can be conducted by considering either the biochemical behaviour or the protein structure [8]. Homologous sequences are referred to as such when two sequences originating from distinct organisms exhibit similarity. The process of aligning DNA sequences. The comparison of DNA sequences serves as the fundamental basis for bioinformatic analysis. This is a crucial initial stage in analyzing the newly identified sequences regarding their structure and function. Sequence alignment is the essential technique involved in this type of comparison. The approach described involves the comparison of sequences through the identification of shared character patterns and the establishment of corresponding residues across related sequences. The process of aligning sequence pairs is crucial to identify commonalities within a database and perform alignment on numerous sequences [9]. Sequence homology is a crucial notion in sequence analysis. Homologous relationship or homology refers to the connection between two sequences that share a shared evolutionary origin. Sequence similarity, although frequently used interchangeably and inaccurately, refers to the proportion of aligned residues that exhibit similar physicochemical features, including size, cost, and hydrophobicity [10].

Table of contents with references:

Topic	**Content**	**References**
Introduction to Bioinformatics	- Concepts, applications, and scope	[1]
	- Technological advancements in data processing	
Computational Methods for DNA Sequence Analysis	- Overview of bioinformatics concepts	[2]
	- Importance of DNA sequence alignment	

Tools for Bioinformatics Solutions	- Utilization of Matlab for bioinformatics	[3]
	- Role of biological databases in analysis	
	- Importance of data warehouses and data mining	
	- Implementation of learning machines in bioinformatics	
Applications of Bioinformatics	- Influence on biological control and genomic analysis	[4]
	- Impact on drug discovery and development	
	- Role in agricultural biotechnology and forensic DNA analysis	
	- Contributions to personalized therapy in healthcare systems	
Analysis of DNA Sequences	- Identifying functional and structural disparities among biological sequences	[5]
	- Techniques for aligning DNA sequences	
	- Understanding sequence homology and similarity	

Application of Computational Technologies in the Field of Bioinformatics.

Like other scientific fields, biology yields substantial amounts of data that necessitate sophisticated computing methods for real-time processing, depending on the study goals. Numerous approaches are encompassed within I.T. research and development, specifically in data storage and processing. These techniques include relational and semantic databases (D.B.), data warehouses, data mining, and various artificial intelligence methodologies [11].

Biological databases:

In contemporary biological databases, three predominant database architectures are commonly employed: flat files, relational databases, and object-oriented databases, despite the evident drawbacks associated with their utilization. The failure to effectively scale real models with the necessary data volume is the underlying cause. Biological databases can be classified into three distinct types based on their content. Primary databases are repositories that house authentic biological data. The entities above refer to raw sequence files or structural data, such as GenBank and Protein Data Bank. Secondary databases are repositories that store information that has been computationally processed or manually curated, derived from primary data sources. Protein sequence databases that have been translated encompass functional annotations that fall within this particular category, such as Swiss-Prot and PIR [12].

Specialized databases, like Flybase, cater to a specific area of research. Examples of databases that focus on a certain organism or kind of data are the Ribosomal Database Project and the HIV Sequence Database. The necessity of integrating secondary and specialized databases with main databases is the root cause of many issues that arise in scientific research. Cross-referencing and linking, or being "linked," to relevant entries in other databases with more information is advised for entries in one database. The primary obstacle to establishing connections across distinct biological databases is the lack of compatibility between the three types of database formats outlined above, which restricts their ability to communicate with one another. Using the Common Object Request Broker Architecture (CORBA) specification language, which enables database programmes in various locations to communicate over a network through a "broker interface" without needing to understand each structure independently, is one way to standardize communication between databases in distributed systems [13]. Databases are also linked using a related protocol known as eXtensible Markup Language (XML). Each biological record in this

format is broken down into smaller, fundamental parts labelled using hierarchical clustering labels. Sequences from cloning vectors and primary data redundancy brought on by repeatedly acquiring identical or matched sequences can contaminate gene sequences. The National Centre for Biotechnology Information (NCBI) developed the nonredundant database RefSeq, which merges related sequence fragments and identical sequences from the same organism into a single record to lessen this redundancy. Developing sequence cluster databases, such as UniGene, that connect expressed tag sequences (ESTs) derived from the same gene is an additional strategy to tackle redundancy. Due to many entries, the genetic sequence is frequently discovered under different names. Reannotation of genes and proteins using a common vocabulary to describe them is required to mitigate this issue. All genes and proteins have a uniform and clear naming scheme thanks to Gene Ontology [14].

The Data Warehouse: A collection of integrated, subject-oriented, time-varying, non-transient data that aids in managerial decision-making is called a data warehouse (D.W.). Examining bioinformatics projects reveals that the field's needs include storing vast amounts of data with various dimensions for long periods, in various formats, and from various sources. Based on a data warehouse known as BioStar schema, the author discusses their suggested multidimensional modelling for biomedical data that can capture the complex semantics of biomedical data and offer higher extensibility and flexibility for fast-expanding biological research approaches [15].To maintain many-to-many links between the central entity and the dimensions, it is necessary to store the various measures in distinct n-tables, which can be made to support particular attributes of a measure. Ligand Depot is a comprehensive knowledge resource on nucleic acids, proteins, and tiny compounds. Its main goal is to give tiny molecules chemical and structural details. In addition, it supports keyword-based searches, offers a graphical search interface for chemical substructures, and grants access to many online resources. In subsequent work, we suggest including a more advanced graphical user interface and enhancing search capabilities [6, 16].

Bioinformatics Data Mining: The focus of data mining is the study of methods for obtaining useful information from vast amounts of biological data. Effective software tools are required for data retrieval, biological sequence comparison, pattern recognition, and knowledge discovery

visualization to accomplish this. We can highlight a few of the most popular data mining methods in bioinformatics:

- KDD, which is the entire process of extracting non-trivial, previously undiscovered, and possibly helpful knowledge from a dataset [17];

Textual mining, also known as KDT, is a process that focuses on extracting knowledge from unstructured data contained in textual databases. It involves the identification and exploration of knowledge within texts. Furthermore

- Statistics in data mining can be categorized into two groups: Supervised learning, which involves making inferences about future observations based on existing knowledge of sequence groups, and Unsupervised learning, which involves identifying previously unknown groups of similar cases in the data [18].

Bioinformatics research software tools can be categorized into four distinct classes:

Tools for recovering data:

An instance of an integrated data retrieval system is Entrez, developed by the National Centre for Biotechnology Information (NCBI). This system offers comprehensive access to several data domains, encompassing literature sequences, nucleotides and proteins, whole genomes, 3D architectures, and other relevant information. The present study aims to compare sequencing and alignment methods. One example is BLAST, known for its high speed and ability to search a complete nonredundant database quickly. GenBank and EMBL are prominent tools for managing biological databases and aligning local sequences in pairs. FASTA is a useful tool for efficiently comparing proteins or nucleotides. Employing a substitution matrix effectively enhances the sensitivity of similarity search by executing optimized searches for local alignments [19].

ClustalW is a tool that may be utilized for multiple sequence alignment. It can align DNA or protein sequences to reveal their relationships and evolutionary origin. Pattern detection techniques identify and analyze patterns or characteristics within a given dataset. Cluster analysis is a statistical technique employed to identify distinct groups within a provided dataset, wherein objects within the same group exhibit similarities to one another while displaying dissimilarities from objects in other groups. GeneQuiz is a comprehensive, integrated tool that proves to be

valuable in the exploration of expression patterns. It operates on a wide scale and employs diverse search and analysis techniques to analyze DNA and protein sequences [19].

Tools for seeing: These tools facilitate interactive and graphical visualization of genetic data. Expression Profiler and GeneQuiz, more extensive analytic programs, provide an integrated visualization tool. In addition, many visualization software packages can be accessed at no cost through online platforms. Examples of software tools that can be utilized include Protein Explorer, which offers a three-dimensional visualization of protein structure through an interactive system, and TreeView, which presents a graphical depiction of clustering outcomes and image navigation through hierarchical trees [20].

Table of contents with references:

Topic	**Content**	**References**
Biological Databases	- Overview of prevalent database architectures	[11]
	- Classification of biological databases based on content	
	- Integration challenges and strategies for biological databases	
The Data Warehouse	- Definition and characteristics of a data warehouse	[12]
	- Implementation of data warehousing in bioinformatics	
	- Utilization of multidimensional modelling for biomedical data	
Bioinformatics Data Mining	- Introduction to data mining in bioinformatics	[13]

	- Overview of popular data mining methods	
	- Categorization of bioinformatics research software tools	
Tools for Recovering Data	- Description of integrated data retrieval systems	[14]
	- Overview of tools for sequence alignment and comparison	
	- Explanation of pattern detection techniques	
Tools for Visualization	- Introduction to tools facilitating graphical visualization of genetic data	[15]
	- Examples of software tools for interactive visualization	
	- Availability of visualization software packages online	

Machine learning:

A learning machine is a dynamic mechanism that enables computers to acquire knowledge through experience, exemplification, and analogy. The neural network is a machine learning technique effectively used to address various bioinformatics challenges. An instance of applying a neural network system, which relies on brain knowledge, can be observed in the analysis of DNA sequences. The emergence of artificial neural networks can be attributed to the prevalence of two prominent genetic search engines. GRAIL is a pioneering gene search programme that utilizes a neural network integration of predictive coding algorithms to detect genes, exons, and other properties in DNA sequences. It recognizes coding potential inside fixedlength windows without extra features [21].

Gene Parser is an additional gene search method specifically developed to ascertain and analyze the intricate composition of protein genes within genomic DNA sequences. GenCANS, an artificial neural system, was created to analyze and handle substantial molecular sequencing data obtained from the Human Genome Project. The genetic algorithm has demonstrated significant efficacy in addressing several practical challenges across multiple fields, with a special emphasis on bioinformatics. Notably, it has been effectively employed in resolving multiple sequence alignment difficulties. SAGA, a widely recognized method, involves the random production of an initial population of forms, which then undergoes evolutionary changes over several generations, resulting in a progressive enhancement of the population's fitness [22].

Soft computing within the field of bioinformatics:

Expert systems, a prevalent soft computing technology, are constructed by aggregating knowledge from specialized human experts. Experts frequently encounter challenges in determining the specific guidelines they employ. The resolution of problems involves the extraction of a comprehensive depiction of the concealed circumstances, encompassing the various aspects and rules that align with the conduct of the human expert. The rapid progress in biotechnology has led to the generation of vast quantities of biological data. Moreover, the data may contain significant linkages and correlations that are not readily apparent. Certain soft computing techniques are specifically intended to process extensive data sets effectively and can also be employed to extract such correlations. Fuzzy systems are crucial in bioinformatics for constructing knowledge-based systems [19].

Fuzzy logic approaches can be effectively employed in various biomedical science and bioinformatics domains. Several significant applications of fuzzy logic include enhancing the flexibility of protein motifs, investigating variations among polynucleotides, analyzing experimental expression data through the application of the fuzzy theory of adaptive resonance, aligning sequences using a fuzzy recast of a dynamic programming algorithm, employing fuzzy systems for genetic sequencing of DNA, analyzing gene expression data, examining gene relationships, deciphering genetic networks, and classifying amino acid sequences into distinct superfamilies [23].

MATLAB is utilized in the field of Bioinformatics:

MATLAB is an abbreviated form of the acronym "MATrix Laboratory". The software is a comprehensive application processing and development environment designed to facilitate the execution of projects that require complex mathematical computations and their corresponding graphical representation. Additionally, the organization presently offers a diverse array of specialized support programmes called Toolbox. Bioinformatics offers a flexible and adaptable platform for molecular biologists and other scientific researchers to investigate concepts, develop prototype algorithms, and establish applications in medicinal research, genetic engineering, and genomic and proteomic endeavours. The Toolbox allows users to access various genomic and proteomic data formats, analysis methodologies, specialized visualizations specifically designed for genomic and proteomic sequences, and microarray research [1, 24].

CONCLUSION:

Data storage plays a crucial role in bioinformatics by enabling the exploration of biological knowledge and facilitating information interchange for research purposes. The advent of collaborative biological web apps is widely regarded as a transformative force in biological research. The assistance of an I.T. professional is crucial for scientists to effectively utilize and fully leverage the technological features of the B.D. The current body of research predominantly focuses on the life and health sciences, thereby necessitating the adoption of a research methodology rooted in computer science within this domain. It is important to establish clear definitions for the technology platforms and implementation procedures linked to the suggested solution to advance the field. During this phase, researchers see that their interests and needs are being acknowledged, which is a crucial prerequisite for the effectiveness of the proposed system.

REFERENCES:

1. Maljkovic Berry, I., et al., Next generation sequencing and bioinformatics methodologies for infectious disease research and public health: approaches, applications, and considerations for laboratory capacity development. The Journal of Infectious Diseases, 2020. **221**(Supplement_3): p. S292-S307.
2. Baxevanis, A.D., G.D. Bader, and D.S. Wishart, Bioinformatics. 2020: John Wiley & Sons.

3. Tyagi, A., A. Sharma, and M. Bhardwaj, Future of bioinformatics in India: A survey. International Journal of Health Sciences, 2022(II): p. 431187.
4. Wood, A., K. Najarian, and D. Kahrobaei, Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Computing Surveys (CSUR), 2020. **53**(4): p. 1-35.
5. Liu, M., et al., A practical guide to DNA metabarcoding for entomological ecologists. Ecological entomology, 2020. **45**(3): p. 373-385.
6. Chen, C., et al., TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. Molecular Plant, 2023. **16**(11): p. 1733-1742.
7. Vaisvila, R. et al., Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome research, 2021. **31**(7): p. 1280-1289.
8. Jovic, D., et al., Single-cell RNA sequencing technologies and applications: A brief overview. Clinical and Translational Medicine, 2022. **12**(3): p. e694.
9. van der Loos, L.M. and R. Nijland, Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Molecular Ecology, 2021. **30**(13): p. 3270-3288.
10. Miller, S. et al., Point-counterpoint: should we be performing metagenomic next-generation sequencing for infectious disease diagnosis in the clinical laboratory? Journal of Clinical Microbiology, 2020. **58**(3): p. 10.1128/jcm. 01739-19.
11. Urban, L., et al., Freshwater monitoring by nanopore sequencing. Elife, 2021. **10**: p. e61504.
12. Beck, D., M. Ben Maamar, and M.K. Skinner, Genome-wide CpG density and DNA methylation analysis method (MeDIP, RRBS, and WGBS) comparisons. Epigenetics, 2022. **17**(5): p. 518530.
13. Sigsgaard, E.E., et al., Population-level inferences from environmental DNA—Current status and future perspectives. Evolutionary Applications, 2020. **13**(2): p. 245-262.
14. Qian, X.-B., et al., A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. Chinese Medical Journal, 2020. **133**(15): p. 1844-1855.
15. Novák, P., P. Neumann, and J. Macas, Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nature Protocols, 2020. **15**(11): p. 3745-3776.
16. Thareja, P. and R.S. Chhillar, A detailed survey on data mining based optimization schemes for bioinformatics applications. ECS Transactions, 2022. **107**(1): p. 4689.
17. Shen, W., et al., Sangerbox: a comprehensive, interaction-friendly clinical bioinformatics analysis platform. Imeta, 2022. **1**(3): p. e36.

18. Phan, J.H., C.-F. Quo, and M.D. Wang, Functional genomics and proteomics in the clinical neurosciences: data mining and bioinformatics. Progress in brain research, 2006. **158**: p. 83-108.
19. Ge, H., et al., Potential role of LINC00996 in colorectal cancer: a study based on data mining and bioinformatics. OncoTargets and therapy, 2018: p. 4845-4855.
20. Jung, Y., et al., Clinical validation of colorectal cancer biomarkers identified from bioinformatics analysis of public expression data. Clinical Cancer Research, 2011. **17**(4): pp. 700-709.
21. Shanbehzadeh, M., R. Nopour, and H. Kazemi-Arpanahi, Comparison of four data mining algorithms for predicting colorectal cancer risk. Journal of Advances in Medical and Biomedical Research, 2021. **29**(133): p. 100-108.
22. Qi, Y., et al., Bioinformatics analysis of key genes and pathways in colorectal cancer. Journal of Computational Biology, 2019. **26**(4): p. 364-375.
23. Yue, C., et al., DUXAP8 a pan-cancer prognostic marker involved in the molecular regulatory mechanism in hepatocellular carcinoma: a comprehensive study based on data mining, bioinformatics, and in vitro validation. OncoTargets and therapy, 2019: p. 11637-11650.
24. Li, Y. et al., Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. Briefings in Bioinformatics, 2021. **22**(5): p. bbab024.