**Tools, Applications and Future Scope of Bioinformatics**

Chaitramayee Pradhan1, Sohini Bhattacharyya2, Ajay Kumar Harit3 and Manoj Singh4

1,2Department of Zoology, Kalinga University, Raipur-492101, Chhattisgarh, India

chaitramayee.pradhan@kalingauniversity.ac.in

**ABSTRACT**

Bioinformatics has developed into an essential tool in a number of biological research fields, including genomics, proteomics, structural biology, and evolutionary biology, over the past few decades. In order to extract useful information from biological data, the interdisciplinary field of bioinformatics combines biology, computer science, and data analysis. The term "bioinformatics tools" refers to a broad category of computer programmes and algorithms used to handle, examine, and interpret biological data and with the help of these technologies, researchers may easily manage enormous datasets and gain previously unthinkable insights. Software for sequence alignment, genome assembly algorithms, programmes that predict protein structures, and phylogenetic analysis tools are a few examples of common bioinformatics tools. Additionally, a lot of biological data is stored and accessible through databases like GenBank, UniProt, and the Protein Data Bank (PDB).The development of high-throughput technologies, which produce vast quantities of biological data, is what propels the ongoing expansion of bioinformatics' application domain.

**INTRODUCTION**

To examine and gain insight into biological data, the multidisciplinary area of bioinformatics integrates biology, computer science, mathematics, and statistics. It entails the creation and use of computational methods and tools to analyze genetic data, comprehend biological processes, and resolve challenging biological issues. Bioinformatics was created as a result of the exponential growth of biological data, notably with the introduction of high-throughput procedures like DNA sequencing, which produce enormous quantities of data that call for specific processing and analysis approaches of future which contributes to the functional knowledge of the human genome, which will improve the identification of drug targets and enable personalized treatment, are among the opportunities in the discipline of bioinformatics (Bayat.et al., 2002).

**TOOLS USED IN BIOINFORMATICS**

Bioinformatics relies on a diverse range of tools and software for various tasks, including sequence analysis, structural prediction, functional annotation, data visualization, and more.

Followings are the some names of software and tools used in the field of bioinformatics frequently.

1. **Sequence Analysis:**

For sequence analysis only various software are used such as

* 1. **BLAST:**Basic Local Alignment Search Tool is used for finding the sequence similarity searches in databases (McGinnis & Madden., 2004).
	2. **FASTQC:** FASTQC is a Quality control tool for high-throughput sequence data. It has basic and ubiquitous formats for storing nucleotide and protein sequences (Leggett et al., 2013).
	3. **Seqtk:**  It is a Toolkit for processing sequences in FASTA/FASTQ format (Shen et al., 2016).
1. **Genome Assembly and Annotation:**

For assembly of genome and annotation tools which are used are

* 1. **SPAdes:** It used for genome assembly from Illumina reads (Bankevich et al., 2012).
	2. **PASA:** PASA is a Genome annotation tool and provides a robust computational platform for the analysis and integration of data obtained from proteomics of serum antibodies (Avram et al., 2021).
1. **Multiple Sequence Alignment:**

For Multiple Sequence Alignment the software used are

* 1. **ClustalW:** It is a multiple sequence alignment tool. (Thompson et., 2002)
	2. **MAFFT:** Multiple Alignment using Fast Fourier Transform is an accurate multiple sequence alignment tool ( Katoh et al., 2002).
	3. **Muscle:** MUSCLE stands for Multiple Sequence Comparison by Log- Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed (Edgar R. C. 2004).
1. **Protein Structure Prediction:**

For protein related analysis the software used are

* 1. **I-TASSER:** Predicts protein structures by threading and ab initio methods. Starting from the amino acid sequence of target proteins, I-TASSER first generates full-length atomic structural models from multiple threading alignments and iterative structural assembly simulations followed by atomic-level structure refinement. (Yang & Zhang 2015), (Roy et al ., 2010)
	2. **Rosetta:** The Rosetta software is for macromolecular modeling, docking, and design is extensively used in laboratories worldwide (Leman et al .,2022)
	3. **SWISS-MODEL:** Automated protein homology modeling. SWISS-MODEL server, which pioneered the field of automated modelling 25 years ago and been continuously further developed day by day. (Waterhouse et al.,2018)
1. **Structural Analysis:**

Structural analysis in bioinformatics involves the study of three-dimensional structures of biological macromolecules, such as proteins, nucleic acids, and complex assemblies.

* 1. **PyMOL:** It visualize and analyze protein structures. (Kagami et al 2020), Mooers ., 2021).
	2. **Chimera:** It shows theinteractive visualization and analysis of molecular structures. (Meng et al .,2006)
1. **Gene Expression Analysis:**

Gene expression analysis plays a pivotal role in understanding biological processes, unraveling disease mechanisms, and identifying potential therapeutic targets.

* 1. **DESeq2:** Differential gene expression analysis for RNA-sequence data is been done by DESeq2.  DESeq2 is a method for differential analysis of count data, using shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates. This enables a more quantitative analysis focused on the strength rather than the mere presence of differential expression. (Love et al., 2022).
	2. **edgeR:** Differential expression analysis for RNA-sequence and other count data are been done byedgeR .( Mark et al., 2009)
	3. **limma:** limma does linear modeling for microarray data analysis.(Matthew et al ., 2015)
1. **Pathway and Functional Analysis:**

It involves identifying enriched biological functions, pathways, and interactions within datasets, helping researchers gain insights into the underlying mechanisms of various biological processes.

* 1. **KEGG:** Kyoto Encyclopedia of Genes and Genomes helps in pathway mapping and functional annotation (Hiroyuki et al., 1999).
	2. **Reactome:** Reactome used for pathway analysis and visualization. (Croft et al., 2011).
	3. **GO tools:** Gene Ontology is afunctional annotation and enrichment analysis tools. The Gene Ontology Consortium's objective is to create a dynamic, regulated language that can be used to describe all eukaryotes even while knowledge about the functions of genes and proteins in cells continues to grow and change (Ashburner et al ., 2000).
1. **Metagenomics Analysis:**

Metagenomics analysis involves studying the genetic material collected directly from environmental samples, such as soil, water, human gut, or other microbial habitats. It provides insights into the diversity, composition, functional potential, and interactions of microbial communities.

* 1. **QIIME:** Quantitative insights into microbial ecology analyze and visualize microbiome data. (Kuczynskiet al 2011)
	2. **MetaPhlAn:** Profiling of microbial communities from metagenomics data.( Blanco-Míguez, et al .,2023)
1. **Network Analysis:**
	1. **Cytoscape:** Visualize and analyze biological networks. (Shannon et al ., 2003)
	2. **STRING:** Predict protein-protein interactions and functional associations. (Damian et al .,2021)
2. **Data Visualization:**
	1. **ggplot2:** It is a data visualization package in R. (Valero-Mora, P. M. 2010).
	2. **matplotlib:** 2D plotting library for Python can be done by this. (Barrett et al 2005)
	3. **Tableau:** Tableau performs data visualization and exploration platform. (Ko & Chang, 2017).
3. **Genome Browsers:**

It involves creating visual representations of biological data to facilitate understanding, interpretation, and communication of complex information. It helps researchers uncover patterns, trends, and insights that might be challenging to discern from raw data alone.

* 1. **UCSC Genome Browser:** Explore genomic data and annotations. (Karolchik et al., 2009).
	2. **Ensembl Genome Browser:** Genome annotation and comparative genomics. (Birney et al., 2004).

**APPLICATIONS IN BIOINFORMATICS**

There are numerous uses for bioinformatics in the fields of biology, medicine, and biotechnology. The area is still evolving and has a bright future ahead of it as technology improves and the amount of biological data increases. Some bioinformatics uses and potential futures are

1. **Genomics and Personalized Medicine:** For identification of disease-related to genes and variants for diagnostics and treatment and also to predict of individual drug responses based on genetic makeup. (Brittain et al., 2017)
2. **Drug Discovery and Design:** Virtual screening of compounds against drug targets can be done as well as insilico drug design and optimization using molecular simulations. (Zhou & Zhong., 2017).
3. **Cancer Genomics:** Characterization of tumor genomes for targeted therapies and identification of biomarkers for early detection and prognosis is possible through it. (Berger & Mardis., 2018).
4. **Metagenomics and Microbiome Analysis:** Understanding microbial communities and their roles in health and disease and exploration of potential probiotics and therapeutic interventions (Wang et al., 2015)
5. **Functional Genomics And Structural Biology:** Elucidation of gene function and regulation using high-throughput data. Network analysis to understand biological processes and interactions. Prediction of protein structures and interactions for drug targeting. Rational design of proteins with desired functions. (Bunnik & Le Roch et al., 2013).

**CONCLUSION**

The field of bioinformatics is broad and is always changing. It includes genome annotation and analysis, protein structure prediction, the investigation of evolutionary links, metagenomics research, and the advancement of personalized medicine. Bioinformatics will continue to be crucial in expanding our understanding of the biological world and tackling difficult problems in environmental science, healthcare, and other fields as technology develops. The capabilities of bioinformatics will be considerably improved in the upcoming years as a result of the combination of artificial intelligence, machine learning, and big data analytics. This interdisciplinary discipline will continue to be at the vanguard of biological research, providing ground-breaking answers to some of the most important life science problems and fostering developments that benefit society as a whole. As a result, the field of bioinformatics is a witness to the efficacy of interdisciplinary cooperation and the revolutionary potential of data-driven approaches in biology.

**REFERENCES**

1. Bunnik, E. M., & Le Roch, K. G. (2013). An Introduction to Functional Genomics and Systems Biology. *Advances in wound care*, *2*(9), 490–498. https://doi.org/10.1089/wound.2012.0379
2. Berger, M. F., & Mardis, E. R. (2018). The emerging clinical relevance of genomics in cancer medicine. *Nature reviews. Clinical oncology*, *15*(6), 353–365. <https://doi.org/10.1038/s41571-018-0002-6>
3. Wang, W. L., Xu, S. Y., Ren, Z. G., Tao, L., Jiang, J. W., & Zheng, S. S. (2015). Application of metagenomics in the human gut microbiome. *World journal of gastroenterology*, *21*(3), 803–814. <https://doi.org/10.3748/wjg.v21.i3.803>
4. Brittain, H. K., Scott, R., & Thomas, E. (2017). The rise of the genome and personalised medicine. *Clinical medicine (London, England)*, *17*(6), 545–551. <https://doi.org/10.7861/clinmedicine.17-6-545>
5. Zhou, S. F., & Zhong, W. Z. (2017). Drug Design and Discovery: Principles and Applications. *Molecules (Basel, Switzerland)*, *22*(2), 279. <https://doi.org/10.3390/molecules22020279>
6. Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002 Apr 27;324(7344):1018-22. doi: 10.1136/bmj.324.7344.1018. PMID: 11976246; PMCID: PMC1122955.A. Science, medicine, and the future: Bioinformatics. BMJ. 2002 Apr 27;324(7344):1018-22. doi: 10.1136/bmj.324.7344.1018. PMID: 11976246; PMCID: PMC1122955.
7. Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H. R., Iyer, V., Jekosch, K., … Clamp, M. (2004). An overview of Ensembl. *Genome research*, *14*(5), 925–928. <https://doi.org/10.1101/gr.1860604>
8. Ko, I., & Chang, H. (2017). Interactive Visualization of Healthcare Data Using Tableau. *Healthcare informatics research*, *23*(4), 349–354. <https://doi.org/10.4258/hir.2017.23.4.349>
9. McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, *32*(Web Server issue), W20–W25. <https://doi.org/10.1093/nar/gkh435>
10. Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J Jensen, Christian von Mering, The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D605–D612, <https://doi.org/10.1093/nar/gkaa1074>
11. Barrett, Paul & Hunter, J. & Miller, J.T. & Hsu, J.-C & Greenfield, P.. (2005). matplotlib -- A Portable Python Plotting Package.
12. Blanco-Míguez, A., Beghini, F., Cumbo, F. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* (2023). <https://doi.org/10.1038/s41587-023-01688-w>
13. Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite, D., & Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Frontiers in genetics*, *4*, 288. <https://doi.org/10.3389/fgene.2013.00288>
14. Shen W, Le S, Li Y, Hu F (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS ONE 11(10): e0163962. <https://doi.org/10.1371/journal.pone.0163962>
15. Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics*, *Chapter 2*., <https://doi.org/10.1002/0471250953.bi0203s00>
16. Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of personality and social psychology*, *122*(3), 523–553. <https://doi.org/10.1037/pspp0000386>
17. Valero-Mora, P. M. (2010). ggplot2: Elegant Graphics for Data Analysis. *Journal of Statistical Software, Book Reviews*, *35*(1), 1–3. <https://doi.org/10.18637/jss.v035.b01>
18. Karolchik, D., Hinrichs, A. S., & Kent, W. J. (2009). The UCSC Genome Browser. *Current protocols in bioinformatics*, *Chapter 1*, Unit1.4. <https://doi.org/10.1002/0471250953.bi0104s28>
19. Kagami, L. P., das Neves, G. M., Timmers, L. F. S. M., Caceres, R. A., & Eifler-Lima, V. L. (2020). Geo-Measures: A PyMOL plugin for protein structure ensembles analysis. *Computational biology and chemistry*, *87*, 107322. Advance online publication. <https://doi.org/10.1016/j.compbiolchem.2020.107322>
20. Mooers B. H. M. (2021). A PyMOL snippet library for Jupyter to boost researcher productivity. *Computing in science & engineering*, *23*(2), 47–53. <https://doi.org/10.1109/mcse.2021.3059536>
21. Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., … Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, *39*(Database issue), D691–D697. <https://doi.org/10.1093/nar/gkq1018>
22. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, *19*(5), 455–477. <https://doi.org/10.1089/cmb.2012.0021>
23. Avram, O., Kigel, A., Vaisman-Mentesh, A., Kligsberg, S., Rosenstein, S., Dror, Y., Pupko, T., & Wine, Y. (2021). PASA: Proteomic analysis of serum antibodies web server. *PLoS computational biology*, *17*(1), e1008607. <https://doi.org/10.1371/journal.pcbi.1008607>
24. Kazutaka Katoh, Kazuharu Misawa, Kei‐ichi Kuma, Takashi Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Research, Volume 30, Issue 14, 15 July 2002, Pages 3059–3066, <https://doi.org/10.1093/nar/gkf436>
25. Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, *5*(4), 725–738. <https://doi.org/10.1038/nprot.2010.5>
26. Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., & Knight, R. (2011). Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Current protocols in bioinformatics*, *Chapter 10*, 10.7.1–10.7.20. <https://doi.org/10.1002/0471250953.bi1007s36>
27. Meng, E.C., Pettersen, E.F., Couch, G.S. *et al.* Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* **7**, 339 (2006). <https://doi.org/10.1186/1471-2105-7-339>
28. Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, Minoru Kanehisa KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Research, Volume 27, Issue 1, 1 January 1999, Pages 29–34, <https://doi.org/10.1093/nar/27.1.29>
29. Edgar R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, *5*, 113. <https://doi.org/10.1186/1471-2105-5-113>
30. Yang, J., & Zhang, Y. (2015). Protein Structure and Function Prediction Using I-TASSER. *Current protocols in bioinformatics*, *52*, 5.8.1–5.8.15. <https://doi.org/10.1002/0471250953.bi0508s52>
31. Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., Aprahamian, M., Baker, D., Barlow, K. A., Barth, P., Basanta, B., Bender, B. J., Blacklock, K., Bonet, J., Boyken, S. E., Bradley, P., Bystroff, C., Conway, P., Cooper, S., Correia, B. E., … Bonneau, R. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature methods*, *17*(7), 665–680. <https://doi.org/10.1038/s41592-020-0848-2>
32. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*(W1), W296–W303. <https://doi.org/10.1093/nar/gky427>
33. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014). <https://doi.org/10.1186/s13059-014-0550-8>
34. Mark D. Robinson  Davis J. McCarthy, Gordon K. Smyth  ,edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics, Volume 26, Issue 1, January 2010, Pages 139–140, <https://doi.org/10.1093/bioinformatics/btp616>
35. Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, Gordon K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Research, Volume 43, Issue 7, 20 April 2015, Page e47, <https://doi.org/10.1093/nar/gkv007>
36. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, *25*(1), 25–29. <https://doi.org/10.1038/75556>
37. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498–2504. <https://doi.org/10.1101/gr.1239303>