Bioinformatics – Supporting modern life science research, applications, and challenges

Bioinformatics – Supporting modern life science research, applications, and challenges

Bioinformática - Apoyar la investigación, las aplicaciones y los desafíos de las ciencias de la vida modernas

DOI:10.34117/bjdv10n2-011

Originals received: 01/05/2024
Acceptance for publication: 01/26/2024

Sarvendra Vikram Singh
PhD in Life Sciences
Institutio: Vigyan Bhawan, Bundelkhand University
Address: Jhansi, India
E-mail: sarvendravs@gmail.com

ABSTRACT

Bioinformatics is an interdisciplinary field that develops methods, software tools for understanding biological data and aims to investigate questions about biological composition, structure, function, and evolution of molecules, cells, tissues, and organisms using mathematics, informatics, statistics, and computer science. As we are moving towards the era of cutting-edge technologies there will be a lot of data to store, process and analyze. It offers analysis software for data studies and comparisons and provides tools for modeling, visualizing, exploring and interpreting data. It includes analysis, structural and functional characterization of biomolecules leading to the development of Genomics, Proteomics, Transcriptomics, and Metabolomics, etc. Drug discovery and development tools, supported by recent advancements in machine learning and cloud computing should shorten the time to find and produce an efficient drug compound with fewer side effects and more results emerge as a branch called Chemo-informatics. Personalized medicine where bioinformatics can help a lot to make drug molecules based on the genetic makeup of individuals for better outcomes is a prime area of research and need of the society at present. The major futures challenge of the scientific community is to create an *in-vitro* model of whole-cell or organism and further simulating a whole cell or an organism by applying *in-silico* approaches. To achieve that, reliable tools that utilize those technologies need to be developed and tested. Bioinformatics reduces the search space/size of the problem by thousand times. The main goal is to convert a multitude of complex data into useful information and knowledge. As a consequence of understanding such data, one can basically engineer longer life for society.

Keywords: bioinformatics, computational biology, genomics, proteomics, system biology.

RESUMO

Bioinformática é um campo interdisciplinar que desenvolve métodos, ferramentas de software para entender dados biológicos e visa investigar questões sobre a composição

biológica, estrutura, função e evolução de moléculas, células, tecidos e organismos usando matemática, informática, estatística e ciência da computação. À medida que avançamos para a era das tecnologias de ponta, haverá muitos dados para armazenar, processar e analisar. Oferece software de análise para estudos de dados e comparações e fornece ferramentas para modelagem, visualização, exploração e interpretação de dados. Inclui análise, caracterização estrutural e funcional de biomoléculas que levam ao desenvolvimento de Genômica, Proteômica, Transcriptomia e Metabolômica, etc. Ferramentas de descoberta e desenvolvimento de medicamentos, apoiadas por avanços recentes na aprendizagem de máquina e computação em nuvem, devem encurtar o tempo para encontrar e produzir um composto de medicamentos eficiente com menos efeitos colaterais e mais resultados emergem como uma filial chamada Quimio-informática. Medicina personalizada, onde a bioinformática pode ajudar muito a fazer moléculas de drogas com base na composição genética dos indivíduos para melhores resultados é uma área primordial de pesquisa e necessidade da sociedade no momento. O maior desafio futuro da comunidade científica é criar um modelo in vitro de célula inteira ou organismo e simular ainda mais uma célula inteira ou um organismo aplicando abordagens in silico. Para conseguir isso, é necessário desenvolver e testar ferramentas confiáveis que utilizam essas tecnologias. A bioinformática reduz o espaço de busca/tamanho do problema em mil vezes. O principal objetivo é converter uma infinidade de dados complexos em informações e conhecimentos úteis. Como consequência da compreensão de tais dados, pode-se basicamente projetar vida mais longa para a sociedade.

Palavras-chave: bioinformática, biologia computacional, genômica, proteômica, biologia de sistemas.

RESUMEN

La bioinformática es un campo interdisciplinario que desarrolla métodos, herramientas de software para la comprensión de datos biológicos y tiene como objetivo investigar preguntas sobre la composición biológica, la estructura, la función y la evolución de moléculas, células, tejidos y organismos utilizando matemáticas, informática, estadísticas y ciencias de la computación. A medida que nos acercamos a la era de las tecnologías de vanguardia, habrá una gran cantidad de datos para almacenar, procesar y analizar. Ofrece software de análisis para estudios y comparaciones de datos y proporciona herramientas para modelar, visualizar, explorar e interpretar datos. Incluye el análisis, caracterización estructural y funcional de biomoléculas que conducen al desarrollo de la Genómica, Proteómica, Transcriptómica, Metabolómica, etc. Las herramientas de descubrimiento y desarrollo de medicamentos, respaldadas por los recientes avances en el aprendizaje automático y la computación en la nube, deberían acortar el tiempo para encontrar y producir un compuesto de medicamentos eficiente con menos efectos secundarios y más resultados que surjan como una rama llamada Quimioinformática. La medicina personalizada, donde la bioinformática puede ayudar mucho a hacer moléculas de medicamentos basadas en la composición genética de los individuos para obtener mejores resultados, es un área primordial de investigación y necesidad de la sociedad en la actualidad. El principal desafío de la comunidad científica en el futuro es crear un modelo in vitro de células u organismos completos y seguir simulando una célula entera o un organismo mediante la aplicación de enfoques in silico. Para lograrlo, es necesario desarrollar y probar instrumentos fiables que utilicen esas tecnologías. La bioinformática reduce el espacio/tamaño de búsqueda del problema mil veces. El objetivo principal es convertir una multitud de datos complejos en información y conocimiento útiles. Como

consecuencia de la comprensión de estos datos, básicamente se puede diseñar una vida más larga para la sociedad.

Palabras clave: bioinformática, biología computacional, genómica, proteómica, biología de sistemas.

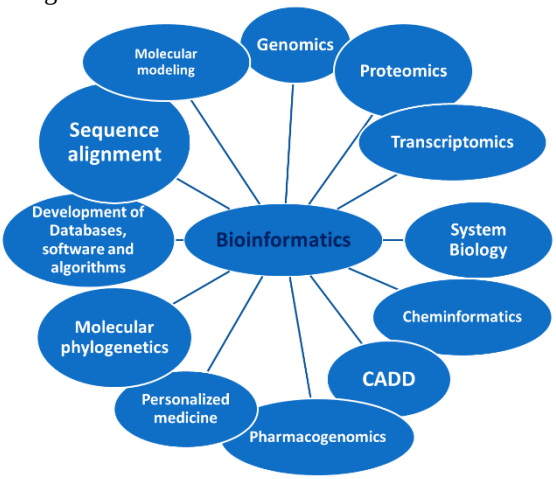
1 INTRODUCTION

In the beginning of the 1970s, Ben Hesper and Paulin Paullin Hogeweg started to use the term ‘‘Bioinformatics’’ for the research they were doing to define it as ‘‘the study of informatic processes in biotic systems’’. Gold biotechnology also known as Bioinformatics is an interdisciplinary field that addresses biological problems using computational techniques, and makes the rapid organization as well as analysis of biological data possible. While birth of molecular biology gave the thinking of how living systems store, process and use genetic information. The information which is stored in living beings can be stored and analyzed by computational means, this leads to the development of science called as Bioinformatics. Hence bioinformatics is an amalgamation of different sciences like mathematics, statistics, Molecular Biology, chemistry and computer science etc. to design algorithms for data storage, software development and further analysis of data with software to get useful information. Bioinformatics application includes applications of computer science to address biological problems and aim is to understand cell system at molecular level. Emerging branches of bioinformatics includes Genomics, Proteomics, Transcriptomics, system biology and further it can be applied to databases development, software and tools development, sequence analysis, gene expression, structural bioinformatics, comparative genomics, agro-informatics, chemo-informatics, drug discovery, personalized medicine, crop improvement, waste cleanup etc. where it is playing a significant role for advanced life science research and ultimately for Human welfare.

After completion of Human Genome Project and development of next gen sequencing technologies in post genomic era, has produced large amount of data/information, which need to be stored in digital form as databases. Starting with storing sequence information of biomolecules such as DNA, RNA and protein etc. in databases in such way so that this information can be retrieved and understand very easily for further analysis. Like information on Nucleic acids sequences in the form of genomes, genes, coding sequences, non-coding sequences, intervening sequences, single nucleotide

variation changes in the form of genome annotation. Such annotated information in databases is very helpful in understanding genomic properties and can be useful in different areas like such medical, agriculture, environmental etc. In the same way, transcriptome, which represent total content of transcription process can be sequenced and analyzed by computational methods for its various properties and its role in developmental processes. Another field called Proteomics that represents analysis of proteomes, the total content of proteins presents in an organism at a particular time or condition. Proteome analysis is very useful in characterizing protein molecules which are ultimately functional molecules of the system as most of the functions of the cell/organisms are performed by them. Hence protein annotation and characterization are an integral part of the bioinformatics. After protein sequence and structure analysis it is very easy to understand biological role of proteins in system. Further interaction of biomolecules in the cell/organism is a very important and challenging thing for the biologists to understand roles of different biomolecules in biological processes. System biology is another very important area where computational tools can be used to integrate the role of biomolecules to understand biological processes. Analysis and characterization of biomolecules in different biological processes and as an integrated approach as in system biology can give us knowledge of mechanisms of biological processes to understand the life processes. Sequence, structure and functional analysis of biomolecules can be performed by developing computer algorithms and software and hence bioinformatics has immense role in modern life science research.

Figure 1: Main areas of Bioinformatics research



Source: self made based on available information

Use of high-throughput technologies to study molecular biology systems in the past decades has revolutionized biological and biomedical research, allowing researchers to systematically study the genomes of organisms (Genomics). Bioinformatic analysis can accelerate drug target identification and drug candidate screening and refinement. It can also facilitate characterization of side effects and predict drug resistance. Various types of online and offline bioinformatics resources/tools are now available for research purposes. Some of the resources are given below.

2 BIOLOGICAL DATABASES AND BIOINFORMATICS TOOLS

Several biological databases and algorithms/software have been made to store and analyze biomolecules information.

Table 1 - Following are the some of the important bioinformatics databases

Type of Databases	**Database Name**	**Database home page URLs**
Nucleotide sequence information	GenBank,	https://www.ncbi.nlm.nih.gov/genbank/
	ENA	https://www.ebi.ac.uk/ena/browser/home
	DDBJ	https://www.ddbj.nig.ac.jp/index-e.html
Protein Sequence Database	UniProtKB	https://www.uniprot.org/help/uniprotkb
2D gel databases	SWISS-2DPAGE	https://world-2dpage.expasy.org/swiss-2dpage/
Protein Structure Database	PDB	https://www.rcsb.org/
Genome Browser	UCSC	https://genome.ucsc.edu/
	Ensembl	https://asia.ensembl.org/index.html
	NCBI Genome	https://www.ncbi.nlm.nih.gov/genome/
Human genes and genetic disorders	OMIM	https://www.omim.org/
Single Nucleotide variations/Polymorphims	dbSNP	https://www.ncbi.nlm.nih.gov/snp/
Metabolic Pathways	KEGG	https://www.genome.jp/kegg/pathway.html
	MetaCyc	https://metacyc.org/
	BioCarta	https://maayanlab.cloud/Harmonizome/dataset/Biocarta+Pathways
	Reactome	https://reactome.org/
Chemistry databases	Drug Bank	http://www.drugbank.ca/
Family and domain databases	Pfam	http://pfam.xfam.org/
	PROSITE	http://prosite.expasy.org/
Protein-protein interaction databases	BioGRID	http://thebiogrid.org
	STRING	http://string-db.org

Source: Website Links for information in table

Table 2 - Following are the some of the important tools for bioinformatics:

Bioinformatics area	Tool for analysis	URL Link
Sequence alignment	BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
	FASTA	https://www.ebi.ac.uk/Tools/sss/fasta/
	CLUSTAL OMEGA	https://www.ebi.ac.uk/Tools/msa/clustalo/
Proteomic Analysis	ExPASy	https://www.expasy.org/
Gene Finding	GenScan	http://hollywood.mit.edu/GENSCAN.html

	GeneMark	http://exon.gatech.edu/GeneMark/
Protein Domain Analysis	ProDom	http://prodom.prabi.fr/prodom/current/html/home.php
Phylogenetic Analysis	PHYLIP	https://evolution.genetics.washington.edu/phylip.html
	Phylogeny.fr	https://www.phylogeny.fr/
	MEGA	https://www.megasoftware.net/
Molecular Docking	AutoDock	https://autodock.scripps.edu/
	GOLD	https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/

Source: Website Links for information in table

Working of a cell or an organism itself a major challenge for the biologist, which need to be solved as it is not very well understood. Bioinformatics is not only limited to Genomics, Proteomics, Transcriptomics and System Biology but it is also having applications in Sequence Analysis, Gene finding, Structural Biology, Protein structure prediction, Homology Search, Multiple Alignment, Phylogeny construction, Genomic full genome-genome comparisons, Rapid assessment of polymorphic genetic variations, Prediction of unknown molecular structures, protein folding, Drug Designing, Machine Learning, Advanced Algorithm for Bioinformatics, Complete construction of orthologous and paralogous groups of genes, Structure determination of large macro molecular assemblies/complexes, Investigate dynamic form and function of large macro molecular and supra molecular complexes, Rapid structural/topological clustering of proteins, Realize interactive modeling, Foster the development of bio molecular modeling, Computer simulation of membrane structure and cell as a whole etc., all of these fields have generated a lot of information in understanding cellular functions at molecular level and still more needed to be discovered with accuracy. There are a number of problems in each field of above-mentioned bioinformatics and they need to be resolved and well understood. We need to work on these problems and try to find answers. Variety of software, algorithms and tools are available which can solve these problems up to certain accuracy level. We have to develop tools and software to resolve these problems with better performance and accuracy because tools and software which work on some parameters may not necessarily work for every sequence or structure that follow their parameters. We need Bioinformaticists to narrow down the work that has to be done in a wet-lab, save time and cost, to make sense of the huge data produced, to be able to predict a lot of things like new therapeutics, genes implicated in various diseases etc. Various types of databases have been created to make better and easy understanding of life science information on biomolecules and literature to help and accelerate the biotechnological research.

3 CHALLENGES IN BIOINFORMATICS

The biotic system which is most complex and mysterious system on this planet always an area of research interest for scientific community. The whole scientific community is trying to understand working of cell at molecular level which is a greatest challenge for them. Understanding genetic diversity, how each species are different with each other at genetic level is another herculean task. Understanding speciation, Structure determination of large complexes, folding of proteins, precise structure prediction of unknown molecules, simulation of cell dynamics, development of non-redundant databases with annotation, development of better algorithms and tools, understanding working of molecules involved in metabolic pathways, genome, transcriptome and proteome analysis and their regulation, effective drug designing etc. are major areas of challenges. Designed software work on some parameters may not necessary that every sequence or structure follow these parameters and overall accuracy of all these tools and software will be a great concern and challenge.

4 CONCLUSION

Bioinformatics now is so integral to all aspects of biological research, we can not imagine any research in life sciences without some kind of bioinformatics analyses being used such as it may be sequence analysis, structure prediction and simulations, NGS analysis, evolutionary relationships, and phylogeny etc. Because now it is very much an integral and indispensable part of all biological sciences related research. In the past time focus of scientific community was towards *in-vivo* to *in-vitro* studies but now the paradigm has shifted from *in-vitro* to *in-silico*. This is one way where we might be able to accurately model biology at a molecular level and use this knowledge to test hypothesis on a much larger scale. Bioinformatics is developing at a much faster pace than ever before and helping a lot in understanding the biology of cell or an organism and decoding the mysteries of life. Bioinformatics has evolved a lot; the field is now focusing on improvement of existing algorithms or development of new algorithms for better performance and accuracy. Bioinformatics can be best utilized in understanding the language of four alphabets that is DNA nucleotides which is converted into language of twenty amino acids with respect to their structure and functions. As the whole scientific community is trying to solve mysteries of life, Bioinformatics together with latest cutting-edge technologies may solve some of them.

REFERENCES

1. Xia X. Bioinformatics and Drug Discovery. Curr Top Med Chem. 2017;17(15):1709-1726. doi: 10.2174/1568026617666161116143440. PMID: 27848897; PMCID: PMC5421137.

2. Kafarski, Pawel. (2012). Rainbow code of biotechnology. Chemik. 66. 814-816.

3. Ridley M. Genome. Harper Perennial; New York: 2006. The Roots of Bioinformatics, Searls DB (2010) The Roots of Bioinformatics. PLOS Computational Biology 6(6): e1000809

4. The Roots of Bioinformatics in Theoretical Biology, Hogeweg P (2011) The Roots of Bioinformatics in Theoretical Biology. PLOS Computational Biology 7(3): e1002021

5. Tarczy-Hornoch, Peter & Minie, Mark. (2005). Bioinformatics Challenges and Opportunities. 10.1007/0-387-25739-X_3.

6. National Research Council (US) Committee on Frontiers at the Interface of Computing and Biology; Wooley JC, Lin HS, editors. Catalyzing Inquiry at the Interface of Computing and Biology. Washington (DC): National Academies Press (US); 2005. B, Challenge Problems in Bioinformatics and Computational Biology from Other Reports. Available from: https://www.ncbi.nlm.nih.gov/books/NBK25461/

7. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015 Jan;43(Database issue): D789-98. doi: 10.1093/nar/gku1205. Epub 2014 Nov 26. PMID: 25428349; PMCID: PMC4383985.

8. Khalid Raza: Indian Journal of Computer Science and Engineering Vol 1 No 2, 114-118

9. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. and Olson, A. J. (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. J. Computational Chemistry 2009, 16: 2785-91.

10. Development and Validation of a Genetic Algorithm for Flexible Docking G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, J. Mol. Biol., 267, 727-748, 1997

11. Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle

12. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

13. Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. 2022 Apr:gkac240. DOI: 10.1093/nar/gkac240. PMID: 35412617; PMCID: PMC9252731.

14. Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539 doi:10.1038/msb.2011.75

15. Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis Nucleic Acids Res. 31:3784-3788(2003).

16. F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112: 535-542.

17. Burge, C. B. and Karlin, S. (1998) Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8, 346-354.

18. GeneMark Borodovsky M. and McIninch J. "GeneMark: parallel gene recognition for both DNA strands." Computers & Chemistry, 1993, Vol. 17, No. 19, pp. 123-133

19. Kanehisa, M.; Toward pathway engineering: a new database of genetic and molecular pathways. Science & Technology Japan, No. 59, pp. 34-38 (1996)

20. Koichiro Tamura, Glen Stecher, and Sudhir Kumar (2021) MEGA11: Molecular Evolutionary Genetics Analysis version 11. Molecular Biology and Evolution 38:3022-3027