An Analysis of Bioinformatics Applications and Overview

[1]M Manoj Kumar, [2]Naveen Kumar, [3]Marimuthu P

[1]Associate Professor in Dept. of Computer Science, GIET Gangapatna, Bhubaneswar
[2]Associate Professor in Dept. of Computer Science, GIET Gangapatna, Bhubaneswar
[3]Associate Professor in Dept. of Computer Science, GIET Gangapatna, Bhubaneswar

Abstract: Data overload has led to many biological challenges becoming computational problems. Bioinformatics, a field that employs computer methods to analyze biomolecule data at a large scale, has emerged as a significant area within molecular biology. This includes research in structural biology, genomics, and gene expression. Recent technological advancements have greatly enhanced research in phenotypic genetics. Genomic breakthroughs have revolutionized biological research on a genome-wide level, resulting in a deluge of data and numerous opportunities. However, coping with the massive volume of information created requires advancements in Moore's Law and biological information processing capabilities. Bioinformatics and computational biology have been instrumental in addressing these challenges. This review provides an introduction and summary of the field, covering bioinformatics concepts, the types of biological information and databases used, and ongoing research, particularly in transcription regulatory systems.

Keywords: Bioinformatics, Datamining, Drug designing, Genomics, Proteomics

Introduction

Recent technological advancements have led to a significant increase in "omic" data, sparking a scientific revolution. However, this influx of data presents a major challenge: making sense of the vast amount of structural data and sequences generated across different biological systems levels. Specialists from various fields have struggled with the creation of coherent crescent data and ensuring public database availability.

The field of bioinformatics plays a crucial role in addressing this challenge by employing statistical and computational methods to understand complex biological problems. Despite its reductionist approach due to the inherent complexity of science, bioinformatics and computational biology have evolved alongside molecular biology during the "new biology" era. This interdisciplinary approach has led to significant advancements and has shaped our understanding of biological systems.

This study aims to provide a concise overview of bioinformatics and related disciplines, including concepts such as biological information and databases, sequence analysis, molecular modeling, genomic analysis, and systems biology. By emphasizing key aspects and introducing new approaches, this study aims to equip researchers with tools for data analysis and interpretation, thereby enhancing our understanding of biological phenomena.

Bioinformatics goals

Bioinformatics has three goals. First, bioinformatics organises data so researchers may access it and add new entries, such the Protein Data Bank for 3D macromolecular structures [3,4]. Data curation is important, but databases are meaningless without analysis. Bioinformatics serves many more purposes. Second, create data analysis tools. Comparing a protein sequence to others is useful. FASTA [5] and PSI-BLAST [6] must examine physiologically meaningful matches for this. Such resources need computational theory and biological knowledge. The final goal is to utilise these technologies to analyse and interpret the data physiologically. Biological research often explored particular systems and compared them to related ones. Global data analysis in bioinformatics may reveal common principles and unique traits.

Bioinformatics, which is interdisciplinary, is "the use of computational tools to organise, analyse, comprehend, display and preserve information linked with biological macromolecules" [7,2]. covers bioinformatics and genomics from three angles: The cell and molecular biology's basic dogma. ii) The organism changes throughout development and bodily areas. Finally, the author stresses globalisation: iii) the tree of life, with millions of species in three evolutionary branches. Computational perspective [7]. These authors list bioinformatics aims as organising data so researchers may access it and add new entries, developing tools and resources to study data, and using these technologies to analyse and interpret data. Bioinformatics challenges fall into two categories: sequencing and biomolecular

structure (Figure 1).

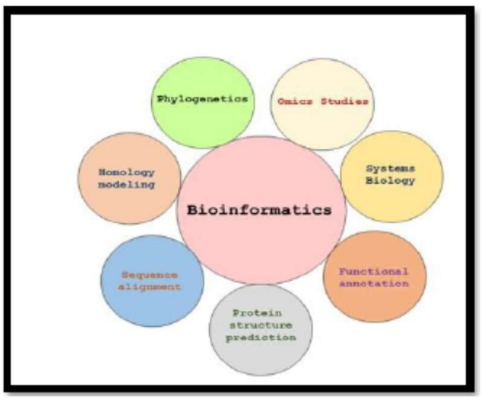

Figure 1: Bioinformatics uses.

Information types and databases

Data organization and storage are needed due to its huge volume. Thus, databases were built to store and interpret vast amounts of biological data for scientific use [7,8]. Nucleic Acids Research has compiled, updated, and disseminated biological databases as data has grown. 1739 biological databases were updated in January 2017. Bioinformatics uses raw DNA, protein, macromolecular, and genomic sequences.

Primary and secondary public databases contain large volumes of data. The main databases include experimental data without comprehensive examination of earlier publications. Content curation in secondary databases compiles and interprets data [8]. KEGG, Reactome, and other functional databases provide metabolic map analysis and interpretation [9].

GenBank @ NCBI, DNA Database of Japan (DDBJ), and European Molecular Biology Laboratory (EMBL) are key databases of nucleotide sequences and proteins [2]. The International Nucleotide Sequence Database Collaboration (INSDC) participants contribute data regularly [9]. PIR, UniProtKB / Swiss-Prot, PDB, SCOP, and Prosite are secondary databases. These curated databases provide solely protein structure, domains, function, and categorization information.

Biological sequence analysis

NGS data has made alignment, important for biological sequence comparison, easier [10]. This method involves comparing two or more nucleotide sequences (DNA or RNA) or amino acid sequences (peptides or proteins) for a set of characters or patterns [11,12]. Why compare sequences? This technique provides evolutionary information on animals, people, genes, prediction functions, and structures [12]. Protein alignment is another bioinformatics technique. Alignment determines whether amino acids are equal across structures, whereas comparison analyses similarities and contrasts [12]. Sequence similarity analysis seems simple, but the algorithm calculates a "cost" to align sequences to minimize differences and get the "best possible result" [11,12].

Simple alignment

This technique emphasises dynamic programming methods, dot matrix analysis, and k-tuple method. The dynamic programming approach uses Bellman's optimality principle to tackle complicated problems by solving their subproblems [12]. This approach produces global and local alignments using Needleman-Wunsch and Smith-Waterman algorithms [11]. Alignment requires a scoring mechanism for matches, mismatches, amino acids, and gaps. Thus, the method determines the best sequence alignment. The dot matrix method detects indels and repeats easily [11]. An identity matrix may visually display similarities [12].

Multiple alignments

This technique emphasises dynamic programming methods, dot matrix analysis, and k-tuple method. The dynamic programming approach uses Bellman's optimality principle to tackle complicated problems by solving their subproblems [12]. This approach produces global and local alignments using Needleman-Wunsch and Smith-Waterman algorithms [11]. Alignment requires a scoring mechanism for matches, mismatches, amino acids, and gaps. Thus, the method determines the best sequence alignment. The dot matrix method detects indels and repeats easily [11]. An identity matrix may visually display similarities [12].

BLAST

BLAST is a Smith-Waterman-derived local alignment method that gives two sequences the highest alignment score [13]. BLAST searches the database using a k-tuple heuristic and dynamic programming from the algorithm [12]. The k-tuple technique restricts the search to more important terms, such as amino acids and nucleotides, which are 3 and 11 characters, respectively [13].

Comparative molecular modeling

Homology modelling involves modelling a protein's 3-D structure from the structure of a homologous protein [14]. Evolutionarily related sequences have the same tertiary structure folding pattern [15]. The 3-D structure helps comprehend function, protein dynamics and interaction, functional prediction, and therapeutic targets [16].

X- ray diffraction crystallography and nuclear magnetic resonance (NMR) may determine structure, although they have limits. Experimental approaches like ab initio modelling or homology may be used [16]. Ab initio protein modelling calculates the best structure using physical and chemical concepts. Homology modelling yields more accurate findings [17]. However, its accuracy depends on target-template similarity [14]. 25–30% identity values are acceptable, while larger values improve projected model quality [15,14]. Prediction involves five processes (Figure 7): Reference identification, template selection, alignment, construction, and model validation.

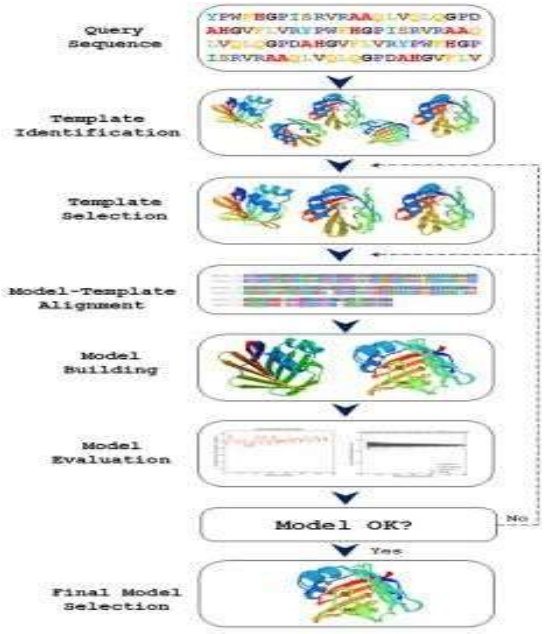

Figure 7: Comparative modelling 3-D structure prediction

Genome-wide analyses from genome to proteome

DNA sequencing transforms genomic designs and opens new avenues in molecular biology [18]. NGS technology have several uses. Bioinformatics will study the genome, transcriptome, and proteome to face infinity.

Genome

Due to cheaper sequencing, several genomes have been published. The new methods' read size and quality (150 to 300 bp) restrict assembly software [19]. They generate more sequences [20].

Transcriptomics

DNA sequencing or hybridization can estimate the transcriptome [17]. Real-time PCR (qPCR) and DNA microarray methods have advanced yet have limitations[21,17]. NGS systems may evaluate global expression instead [22].

Proteomics

Understanding cellular physiology requires identifying, quantifying, and characterising all cell proteins [23]. To systematise the study of protein structure, function, relationships, and dynamics in space and time, proteomics has grown fast [24].

Applications of bioinformatics Finding Homologues

Bioinformatics seeks biomolecule similarities, as mentioned before. Protein homologue identification has practical applications beyond data organisation. The most evident is protein-protein communication. For example, a poorly characterised protein may be searched for homologues that are more understood and cautiously applied to the former. Protein structural models are mainly based on experimentally solved structures of near homologues [25]. Fold recognition uses similar methods to anticipate tertiary structures based on distant homologous structures and energetic viability [26]. When biochemical or structural data are inadequate, yeast may be used to study homologues in higher-level species like humans, where experiments are more difficult. Genomic methods are similar. Homologue finding and functional data are used to validate coding areas in freshly sequenced genomes. Early structural genomics efforts focused on *Mycoplasma genitalium* because it reduces the difficulty of comprehending complex genomes by analysing simple species first and then applying the same concepts to more sophisticated ones [27].

Rational Drug Design

Bioinformatics was first used in rational medication design. Figure 2 shows the usual drug target method using the MLH1 gene product. MLH1 encodes a mismatch repair protein (mmr) on the short arm of chromosome 3 [28]. Linkage studies and its similarity to mouse mmr genes linked the gene in nonpolyposis colorectal cancer [29]. Translation software can predict the protein's amino acid sequence from the nucleotide sequence. After finding homologues in model animals, sequence similarity may be utilised to model the human protein's structure on experimentally characterised structures. Finally, docking algorithms might build compounds that bind the model structure, allowing biochemical experiments to assess their biological activity on the protein.

Large-scale censuses

Although databases can easily store genome, structure, and expression dataset data, it is helpful to compress this data into user-friendly patterns and facts. Broad generalisations highlight fascinating topics for deeper research and contextualise fresh discoveries. This shows whether they are uncommon.

Considerations and perspectives

Data collection, processing, and interpretation have improved, suggesting a bright future. New analytical methods are emerging due to widespread scientific advances. We can improve our knowledge of complicated biological systems by using molecular information in systemic ways. Data integration is not the end. New theories and findings start a feedback mechanism. Genomic innovations in gene therapy and customised medicine will enhance health. This suggests that diverse research organisations and scientists with expertise in several fields are needed to make significant scientific discoveries.

Conclusions

Computational approaches are essential to biological research due of the data flood. Bioinformatics now covers structural biology, genomics, and gene expression investigations in addition to sequence analysis. This review introduced and summarised the field. We focused

on biological information and databases, transcription regulatory system investigations, and practical applications. All bioinformatics research use two methods. First is comparing and categorising data by biologically relevant similarities, and second is using one kind of data to infer and comprehend another type. The field's key goals are to interpret and organise biological molecular data on a big scale. Bioinformatics has deepened and broadened biological research. Thus, we may study individual systems and compare them to related ones to find common principles and uncommon traits.

Reference

[1] Ritchie MD, Holzinger ER, Li R, Pendergrass SA, et al. (2015). Methods of integrating data to uncover genotypephenotype interactions. Nat. Rev. Genet. 16: 85- 97

[2] Pevsner J (2015). Bioinformatics and functional genomics, 3rd ed. John Wiley & Sons Inc, Chichester Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, et al. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem 1977;80(2):319-24.

[3] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-42.

[4] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85(8):2444-2448.

[5] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-3402

[6] Luscombe NM, Greenbaum D and Gerstein M (2001). What is bioinformatics? A proposed definition and overview of the field. Methods Inf. Med. 40: 346-358 10.1053/j.ro.2009.03.010.

[7] Prosdocimi F (2010). Introdução à bioinformática. Curso Online. Available at [http://www2.bioqmed.ufrj.br/prosdocimi/ FProsdocimi07_CursoBioinfo.pdf].

[8] Prosdocimi F, Cerqueira GC, Binneck E and Silva AF (2002). Bioinformática: Manual do usuário. Biotec. Cienc. Des. 12-25.

[9] Daugelaite J, O' Driscoll A and Sleator RD (2013). An overview of multiple sequence alignments and cloud computing in bioinformatics. Int. Sch. Res. Not. e615630.

[10] Manohar P and Shailendra S (2012). Protein sequence alignment: A review. World Appl. Program. 2: 141-145

[11] Junqueira DM, Braun RL and Verli H (2014). Alinhamentos. In: Bioinformática da biologia à flexibilidade molecular (Verli H, ed.). SBBq, São Paulo, 38-61.

[12] Amaral AM, Reis MS and Silva FR (2007). O programa BLAST: guia prático de utilização. 1st edn. Embrapa Recursos Genéticos e Biotecnologia. EMBRAPA, Brasília.

[13] Capriles PVSZ, Trevizani R, Rocha GK and Dardenne LE (2014). Modelos tridimensionais. In: Bioinformática da biologia à flexibilidade molecular (Verli H, ed.). SBBq, São Paulo, 147-171.

[14] Calixto PHM (2013). Aspectos gerais sobre a modelagem comparativa de proteínas. Cienc. Equat. 3: 10-16

[15] Madhusudhan MS, Marti-Renom MA and Eswar N (2005). Comparative protein structure modeling. In: The proteomics protocols handbook (Walker, J.M., ed.). Human Press, New Jersey, 831-860.

[16] Wang J (2009). Protein structure prediction by comparative modeling: an analysis of methodology. Comp. Gen. Pharmacol. 218: 1-13

[17] Zhou X, Ren L, Meng Q, Li Y, et al. (2010). The next-generation sequencing technology and application. Protein Cell 1: 520-536

[18] Miller JR, Koren S and Sutton G (2010). Assembly algorithms for next-generation sequencing data. Genomics 95: 315-327

[19] Altmann A, Weber P, Bader D, Preuss M, et al. (2012). A beginners guide to SNP calling from high-throughput DNA sequencing data. Hum. Genet. 131: 1541-1554

[20] Marioni JC, Mason CE, Mane SM, Stephens M, et al. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18: 1509-1517

[21] Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, et al. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773-777

[22] Schmidt A, Forne I and Imhof A (2014). Bioinformatic analysis of proteomics data. BMC Syst. Biol. 8: 2, S3. doi:10.1186/1752-0509-8-S2-S3

[23] Jensen ON (2006). Interpreting the protein language using proteomics. Nat. Rev. Mol. Cell Biol. 7: 391-403

[24] Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234(3):779- 815.

[25] Jones DT, Taylor WR, Thornton JM. A new

approach to protein fold recognition. Nature 1992;358(6381):86-9

[26] Teichmann SA, Chothia C, Gerstein M. Advances in structural genomics. Curr Opin Struct Biol 1999;9(3):390-9

[27] Kok K, Naylor SL, Buys CH. Deletions of the short arm of chromosome 3 in solid tumors and the search for suppressor genes. Adv Cancer Res 1997;71:27-92.

[28] Syngal S, Fox EA, Eng C, Kolodner RD, Garber JE. Sensitivity and specificity of clinical criteria for hereditary nonpolyposis colorectal cancer associated mutations in MSH2 and MLH1. J Med Gen 2000;37(9):641-645