International Journal of Information Science and Computing
Citation: *IJISC*: **08**(01): 23-30, June 2021
DOI: 10.30954/2348-7437.2.2021.3

INFORMATION SCIENCE AND TECHNOLOGY
FOUNDATION

A Survey on Bioinformatics: Working Process Methodology

Dona Bhattacharya

Department of Computer Science and Engineering, St. Mary's Technical Campus, Kolkata, India

Corresponding author: dona.bhattacharya420@gmail.com

Received: 12-03-2021 **Revised:** 21-05-2021 **Accepted:** 02-06-2021

ABSTRACT

An uncountable number of data of Biology has reached such an extent that now it is creating a challenge to the Computer Science world. At this point in time, the introduction of this subject 'Bioinformatics' has played an immense role in both the fields of Biology and Information Technology. Bioinformatics is an interdisciplinary field that has already developed various types of methods and software tools which has been used for understanding different types of large and complicated biological data in a more analytical manner.

Keywords: Bioinformatics, Information Technology, software tools

Overview

Computational Biology or Bioinformatics is a subject where biological data are getting interpreted by the application of Computer Science. As the data on protein sequence, genomics, three-dimensional modeling of biomolecules and biological systems has developed so vastly so the Data Mining technique here acts as a base to analyze these datasets.

The tools used in this subject are used to analyze the biological data, exploring approaches, emerging methodologies, and clear up the meaning of the generated biological data.

Introduction

The subject Bioinformatics may sound like a recent days invention but it is not well known that this

How to cite this article: Bhattacharya, D. (2021). A Survey on Bioinformatics: Working Process Methodology. *IJISC.,* **08**(01): 23-30.

Source of Support: None; **Conflict of Interest:** None

subject has got introduced around 50 years back. In the year 1970s, Paulien Hogeweg and Ben Hesper has coined the term 'Bioinformatics' and has referred to this process of information study to learn the biotique system more deeply.

The study of this subject has helped in the application section of computational techniques for better understanding and organizing the information associated with biological macromolecules in a more analytical manner. With the advancement of technologies in recent years, a remarkable evolution has been observed in the field of science which has provided all the science workers a huge amount of data, and now this uncountable number of data are creating a challenging condition to the computer science field.

In Biology, it is a very challenging work to make some sense out of that enormous amount of biological structural data and sequences generated in a biological system at a multi-layered process. In this topic of interest, it is very necessary to develop Statistical and Computational tools that are capable of understanding the mechanism of biological sequence. But here comes the question of data arrangement and management, to keep all these huge counts of information in an organized manner Databases are created. These Databases are created in such a way that they can be easily accessed by the researchers and also the members of the scientific community to gather information as per their requirement. Now slowly this huge amount of data has started getting the company of an increasing number of biological databases. To compile, disseminate and update all databases along with data, 'Nucleic Acid Research Journal' has got appointed and as per their recently updated information presently there is an existence of more than 1739 biological databases.

Bioinformatics uses raw DNA sequence, protein sequence, macromolecular structure, and Genome Sequence as the source of their information and all this information is present therein in Public databases. So for easy detection process, these public database has been classified into three divisions-

1. Primary or Sequential Database
2. Secondary or Structural Database
3. Functional Database

1. Primary or Sequential Database – In Bioinformatics a Sequential Database is a type of biological database which is composed of a large amount of computerized nucleic acid sequences and polymer sequences stored in a Computer. This database contains the results of the experiment's unrevised data related to previous publications.

e.g. Green Bank at the National Centre for Biotechnology Information (NCBI), DNA Database of Japan (DDBJ), standout as the main database of Nucleotide sequence and proteins.

2. Secondary or Structural Database - In Biology a protein structural database is modeled around the various experimentally determined protein structures. These data's interpretation and compilation are done which are known as Content Curation Process.

e.g. Protein Information Resources (PIR), Swiss-Prot Protein Database (PDB), Structural Classification of Proteins (SCOP), and Prestige.

3. Functional Database- provides information on the physiological role of gene products and allows analysis of interpretation of metabolic maps.

e.g. enzyme activities, mutation phenotypes or biological pathway, Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactions.

As it has come to know that the subject Bioinformatics is an interdisciplinary field of science, so for detail analysing the provided biological data this field do requires a combined knowledge of (Fig.1):

- ❖ Biology
- ❖ Mathematics
- ❖ Statistics
- ❖ Computer Science
- ❖ Information Technology
- ❖ Chemistry
- ❖ Biochemistry
- ❖ Biophysics

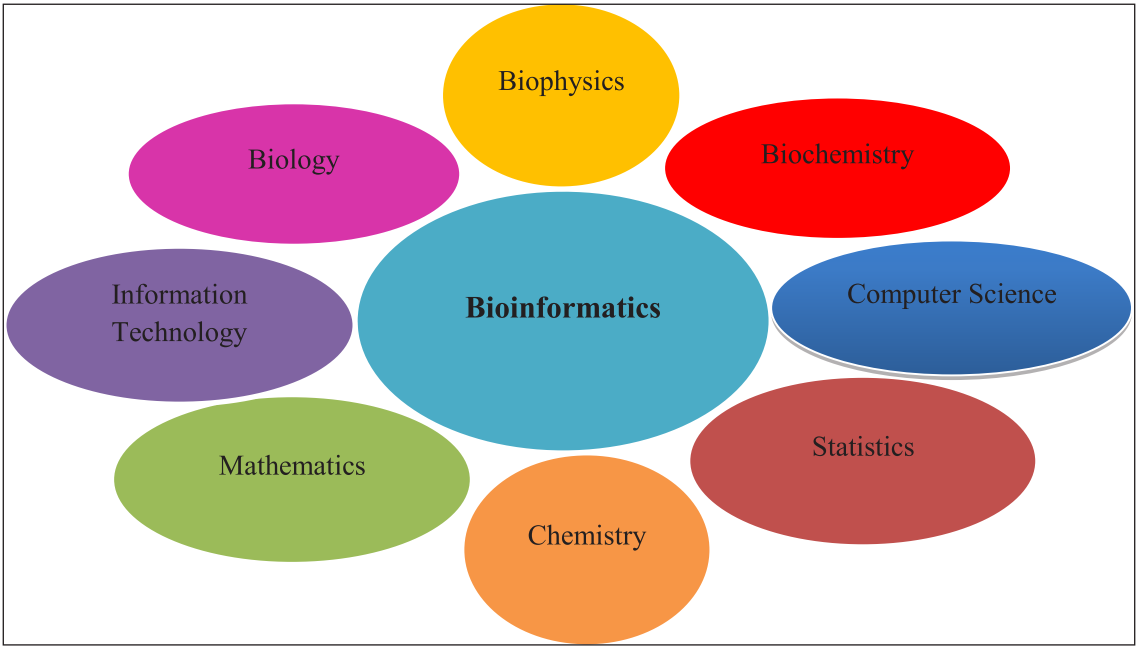


Fig. 1: Structure of combined subjects

Structural Development

The structural development of this study of Bioinformatics is totally dependent upon three necessary components-database creation, algorithm and statistical development and deep analysis of data and interpretation (Fig. 2).

Step 1- Database Creation: Database creation involves the storing and managing process of biological data sets in an organized way so that the researchers can continue their further work process of implementing some new data or information by applying their innovative thought process.

Step 2- Development of Algorithm and Statistics: Algorithm and Statistical development involve some tools which are used to develop the resources which are used to determine the relationship with the members of large datasets.

Step 3- Analysis of Data along with and Interpretation: To analyze the data and to interpret the results in a biologically meaningful manner the appropriate use of the earlier mentioned two components are necessary. This includes DNA, RNA, and protein sequence, protein expression profiles, and biological pathways.

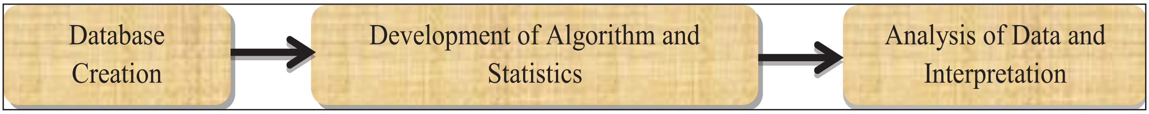
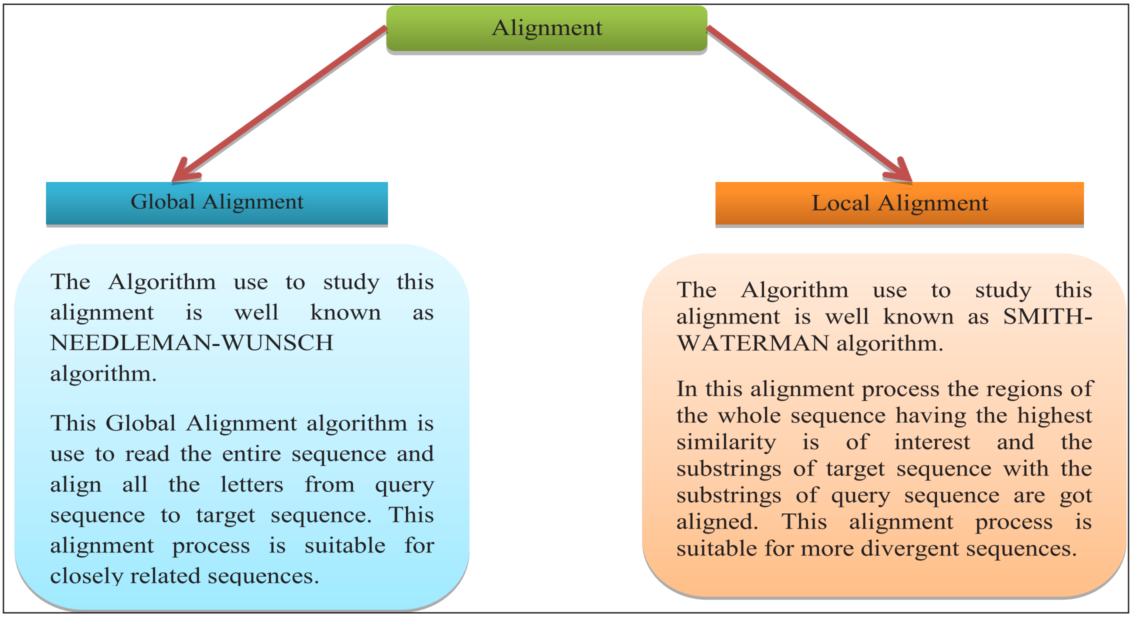

Database Creation
Development of Algorithm and Statistics
Analysis of Data and Interpretation

Fig. 2: Pictorial format of steps involved on data processing

As the information are been collected from the Public Database then here comes the development of the Algorithmic section. For this development process, checking of sequential alignment is necessary and for that, the researchers do use Nucleotide Base Sequence of DNA/RNA and Amino Acid Sequence. By the application of the sequential alignment process, the researchers came to know the efficiency of the functional, structural and evolutionary relationship within the selected sequences.

Now this alignment process has been divided into two sections (i) Global Alignment and (ii) Local Alignment (Fig. 3).



The Algorithm use to study this alignment is well known as NEEDLEMAN-WUNSCH algorithm.

This Global Alignment algorithm is use to read the entire sequence and align all the letters from query sequence to target sequence. This alignment process is suitable for closely related sequences.

The Algorithm use to study this alignment is well known as SMITH-WATERMAN algorithm.

In this alignment process the regions of the whole sequence having the highest similarity is of interest and the substrings of target sequence with the substrings of query sequence are got aligned. This alignment process is suitable for more divergent sequences.

Fig. 3: Tabular form of Alignment

After the completion of this section of Algorithmic and Statistical development, here comes the deep analysis of all these sequences. By the application of Deep Learning, this analytical section gets studied. Deep learning which is a subset of Machine Learning is used to make predictions in the field of bioinformatics which deals with the mathematical and computational approach for understanding and processing biological data.

Software programs that are designed for extracting meaningful information from biological databases are known as Bioinformatics Tools. This software is data mining software that used to collect genomic sequences from the database and visualization tools are used to analyze and retrieve this information from proteomic databases.

These tools can be classified as:

1. Homology and Similarity tools
2. Protein Functional Analysis Tools
3. Sequence Analysis of Miscellaneous Tools.

To sum up the process of organizing and understanding the biological data for Bioinformatics, it can be done in a two-dimensional process- (i) Depth and (ii) Breadth.

If the example of Protein Synthesis is considered then firstly we have to minimize the function ability of it from the gene sequence to its final structure and its ligands. Secondly, the aim of this topic is the comparison between genes, determination of protein structures, and the evolutionary mechanism between species to species.

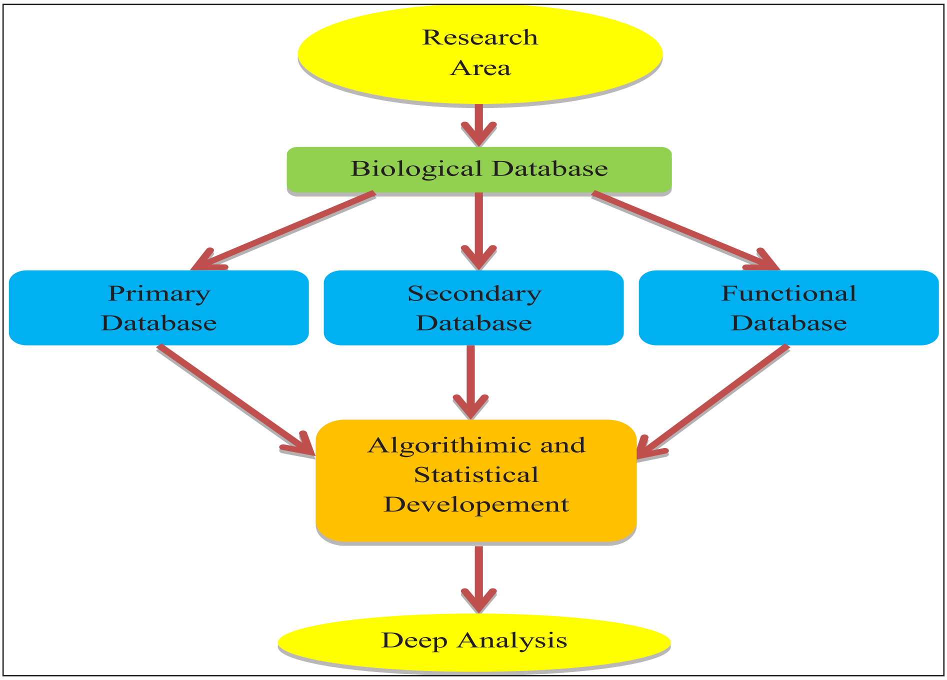


Fig. 4: Tabular Representation of the whole working process

Identification Purpose

- Generally, Bioinformatics is used to identify:
- Candidates gene
- Single nucleotide polymorphism
- Genetic basis of diseases
- Unique adaptation
- Proteomics (difference between population and organisational principles within nucleic acid and protein sequence)

Application Area

- The study of Bioinformatics is used in diversified fields:
- In Genetical Field- this subject is used to study sequence and annotating genomes along with their observation in Mutation.
- In Text Mining of Biological literature- Bioinformatics helps in the development of Biological, Genetical, ontologies to organize and Query biological data.
- In Structural Biology- the study of Bioinformatics is used in the simulation and modeling of DNA, RNA, Protein, and biomolecular interaction.
- Bioinformatics plays an immense role in analyzing gene, protein expression, and regulation.
- Drug designing
- Forensic DNA analysis
- Architectural Biotechnology
- Computational studies of protein legends
- Knowledge of three-dimensional structure of the protein

CONCLUSION AND FUTURE SCOPE

Bioinformatics has played an important role in many areas of Biology. Some of the Bioinformatics techniques such as signal processing and image processing allow the extraction of useful results from a large amount of raw data. The use of computer information technologies to collect, organize, maintain, access, and analyse the data. For the further betterment process, the development of tools which are used to gather data generation, capture and annotation, and databases for comprehensive functional studies. Improvement of contents and utility of Databases is also required.

REFERENCES

1. Altelaar, A.F.M., Munoz, J. and Heck, A.J.R. 2013. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.,* **14**: 35-48.

2. 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.

3. Amaral, A.M., Reis, M.S. and Silva, F.R. 2007. O programa BLAST: guia prático de utilização. 1st edn. Embrapa Recursos Genéticos e Biotecnologia. EMBRAPA, Brasília.

4. Calixto, P.H.M. 2013. Aspectos gerais sobre a modelagem comparativa de proteínas. *Cienc. Equat.,* **3**: 10-16.

5. Capriles, P.V.S.Z., Trevizani, R., Rocha, G.K. and Dardenne, L.E. 2014. Modelos tridimensionais. *In:* Bioinformática da biologia à flexibilidade molecular (Verli H, ed.). SBBq, São Paulo, pp. 147-171.

1. Chaisson, M.J.P., Wilson, R.K. and Eichler, E.E. 2015. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.,* **16**: 627-640.

2. Daugelaite, J., O' Driscoll, A. and Sleator, R.D. 2013. An overview of multiple sequence alignments and cloud computing in bioinformatics. *Int. Sch. Res. Not.,* e615630.

3. Dayhoff, M.O., Schwartz, R. and Orcutt, B.C. 1978. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure (volume 5, supplement 3 ed.). *Nat. Biomed. Res. Found.,* Washington, D.C.

4. Goff, L.A., Trapnell, C. and Kelley, D. 2012. CummeRbund : visualization and exploration of Cufflinks high-throughput sequencing data. R package version 2.16.0.

5. Hagen, J.B. 2000. The origins of bioinformatics. *Nat. Rev. Genet.,* **1**: 231-236.

6. Hawkins, R.D., Hon, G.C. and Ren, B. 2010. Next-generation genomics: an integrative approach. *Nat. Rev. Genet.,* **11**: 476-486.

7. Hong, S., Chen, X., Jin, L. and Xiong, M. 2013. Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res.,* **41**: e95.

8. Hunt, L.T. 1984. Margaret Oakley Dayhoff 1925-1983. *Bull. Math. Biol.,* **46**: 467-472.

9. Institute for Systems Biology (2016). What is a systems biology. Institute for systems biology. Available at [https://www. systemsbiology.org/about/what-is-systems-biology/].

10. Jensen, O.N. 2006. Interpreting the protein language using proteomics. *Nat. Rev. Mol. Cell Biol.,* **7**: 391-403.

11. Junqueira, D.M., Braun, R.L. and Verli, H. 2014. Alinhamentos. *In:* Bioinformática da biologia à flexibilidade molecular (Verli H, ed.). SBBq, São Paulo, pp. 38-61.

12. Kitano, H. 2002. Systems biology: a brief overview. *Science,* **295**: 1662-1664.

13. Kogelman, L.J.A., Cirera, S., Zhernakova, D.V., Fredholm, M. *et al.* 2014. Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. *BMC Med. Genomics,* **7**: 57.

14. Luscombe, N.M., Greenbaum, D. and Gerstein, M. 2001. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.,* **40**: 346-358.

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

16. Malone, J.H. and Oliver, B. 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.,* **9**: 34.

17. Manohar, P. and Shailendra, S. 2012. Protein sequence alignment: A review. *World Appl. Program,* **2**: 141-145.

18. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. *et al.* 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.,* **18**: 1509-1517.