Original Research Paper

Volume - 12 | Issue - 02 | February - 2022 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Bioinformatics

BIOINFORMATICS AND ITS EMERGENCE

Avantika Priya	M. Sc. Bioinformatics, University of Allahabad, Allahabad, INDIA.
Yamini Shankar*	IILM CET-AHL Greater Noida, Gautam Budhha Nagar 201306, INDIA. *Corresponding Author
Dhiresh Kumar Pathak	L.I.E.T. Greater Noida, Gautam Buddha Nagar 201306, INDIA.

ABSTRACT Bioinformatics has emerged as a field which is associated with the exclusive analysis of biological data, using computational methods. This multidisciplinary stream has made it possible to obtain, analyze and work upon cellular and sub cellular data of any living organism, with more specificity. Cell forms the basis of life, and the underlying processes going on within the cell can be extensively studied by using the tools of bioinformatics. Almost all the basic areas of biology, including various fields of molecular biology, structural biology and many more are using the techniques of bioinformatics for extracting and processing the huge amounts of raw data to derive all the possible and highly useful results. Bioinfonnatics is also playing a key role in the study of topics like signal transduction, mutations , gene and protein expression, data mining, gene finding, sequence alignment, drug discovery, prediction and modelling of bimolecular structures and their interactions. The development in all these and many related fields has given a new frame to the study of biology. Many questions related to life, which were a huge mystery in the olden days, have now been solved, hence making life much more easy and understandable for the living beings. This review intends to give an insight to how this dynamic field of Bioinfonnatics emerged, from Biology and integrated with computational techniques, in order to provide a better outlook to the study of living systems.

KEYWORDS :

INTRODUCTION:

Biology, as we all know is the stream of science that includes the study of living organisms. This area takes into account, the detailed study of the existence of life, its various physical, chemical and physiological processes. On the other hand, Computers or computational studies includes study of algorithmic processes and computational machines.
If we look approximately 2 or 3 decades back, not much could have been done by the combination of these two streams. But for today, the scenario has completely changed. Biology and computers have been linked together ever since the research on the genomes of various organisms initiated. These organisms also now include the Homo sapiens. The extensive study of the Human Genome Project the initiated in 1990, among others, brought biology and computer science, very close to each other. With the advent of new technologies, both the streams, are not only helping and influencing each other, but also gradually merging at a greater pace than ever before.

The eventual combination of these two streams is called Bioinformatics or Computational Biology. So if we want to describe Bioinformatics, we may say that is the usage of computer sciences and its related technologies to solve the problems related to Biology.

RESEARCH OBJECTIVES:

In early days the computing technology was usually applied in biology in the bio-medical field. But this new field of Bioinformatics, has its focus on the molecular and sub-cellular levels of Biology.

Although this is a comparatively recent field, a lot of research work has been already done and is still going on at a greater pace. In this review, we would attempt to study the gradual advancement in this new, challenging and highly interesting field of Bioinformatics.

A large number of software's and biological tools have been developed, which have enabled the rapid collection of massive amounts of data. We will also see how some computers applications have changed the way how data was analyzed, researched upon and used in proper way, so as to get maximum possible information of the concerned organism.

Research Methodology:

Initially, we would have a look at the areas of Biology, on which Bioinformatics actually emphasizes upon. Gradually we would have a complete overview of how basically this field involves the use of both the streams, to solve some of the real complex problems. We will also have an insight of some computers algorithms and software's, which have been helping Biologists and Bioinformaticians since long time now.

We will also try to understand how various diverse areas of Bioinformatics/ Computational Biology have opened numerous gateways, allowing the researchers to study, manipulate and analyze a simple DNA or a Protein sequence in such a way that they have can now extract various kinds of information about the concerned organism like never before.

As we have already discussed that Bioinformatics actually deals with the sub cellular levels of Biology. So we would start with giving a short explanation to the vocabulary that holds the major importance that this level.

1. The basic explanation given to a cell is that "it forms the structural and the functional unit of life". In other words we may also say that cell are the most complex, mysterious machines which are always active and which result in the activity of each and every organism that has life in it. As we go in depth, we would realize that cell is a very complex functioning and structure, but on simplification, 3 key things can be found: DNA, RNA and Proteins.

2. Deoxy ribonucleic acid (DNA): This molecule is found within the cell and has the core information of carrying the genetic information from one generation to the other. It basically has a double helical structure. And it is made up of alternating sugar and phosphate groups. Attached to each sugar is one of four bases--adenine (A), cytosine (C), guanine (G), and thymine (T). The two strands are held together by bonds between these bases. Here A bonds with T and C bonds with G. In almost all eukaryote organisms, DNA is packed into chromosomes, and the chromosomes are together called the genome of the organism. Various specific regions on the genome are called as genes.

3. Ribonucleic acid (RNA): This is also found inside the cell and it is a polymeric molecule, which has some of the very important functions in the cell, including the cellular protein synthesis and the even the replacement of the DNA molecule as a carrier of genetic codes in some viruses. It also contains the 4 molecules, just like that in DNA, A, G, C, but the T (Thymine) is replaced by U (Uracil). Also, RNA has a single stranded structure and many times it can also assume certain unique shapes.

4. Proteins: These are the macromolecules made up of long chains of amino acids. They are large biomolecules which are essential for the most of the processes taking place within the cell. Proteins form the major component of the cell which are required for the **structure,** proper functioning, maintenance and regulation of the tissues and organs of the living organism. They do not have a linear structure. In

fact they can twist and fold to form structure that may be either simple or as complex as a 3 dimensional structure.

These work a few basic things working within the cell and also the mam focus of the bioinformatics study. Now we would shed some light on how the study of bioinformatics, takes into account the data coming from the DNA, RNA and Proteins of living cells, for extracting the information that can actually be used for solving many biological problems.

Genomics is the study of the whole genome of living organisms. It is actually a multidisciplinary field of Biology that includes the study of the structural, functional, and evolutional aspect of the genome. This also focusses on the editing and mapping of the genome.

Proteomics actually involves the study of the entire study of set of proteins (called the proteome) of an organism. Proteins form a very important part of the living beings. This area of study, has helped in the identification of a large number of proteins.

These fields of Bioinformatics basically start the study with the simple DNA/RNA/Protein sequence and try to answer some of the common questions. While answering these questions, they actually provide the complete insight of the complete organism.

It would become clearer if we try to understand with a real practical example.**Pair wise Sequence Alignment:** Now we would be taking a small part of the complete genome of *Musca domestica* **(house fly),** from the National Centre for Biotechnology Information (NCBI), which would just look like a common sequence of the letters, A,T,G and C.

```
>NW_004754939.1 Musca domestica strain aabys unplaced genomic scaffold,
Musca_domestica-2 .0.2 Scaffold0, whole genome shotgun sequence
ATGAGGTTGTGGTTGAAAATAATGAACTTAATAATAATGTTGTTTATTGG
TGGCTGTA    GGTaactcgcgtggcgtgtattggttcatggtaaaaatccgcttggactgacgcttcaaacgcggt
atgtcattaagcgatgtgatatctctgtcaaaagttatggcgttgccagatgggtccgtcgaatatggacaaccttCTAT
GGGTCCTTCATAGAACTAGGGTTGAAActttcc
```

Fig 1: A small part of the genome of the *Musca domestica* strain.

NCBI was founded in 1988 and is a part of United States National Library of Medicine. It has created and is maintaining approximately 40 databases and has become the leading source of biomedical database and various software tools for the analysis of molecular and genomic data. NCBI has a big role in the fast growth in the field of Bioinformatics. The genome of any organism, which has already been sequenced, can be easily obtained.

Now to understand how this simple sequence can be used to extract some information of the organism, we did the analysis of the gene using the Basic Local Analysis Search Tool (BLAST).

BLAST is a tool that searches for the similarity within the sequence. It is a local search alignment tool, which uses its algorithm to find the sequences, from its data base, where we just need to give the genome sequence as an input to the blast program. As the search completes, you get the details and the regions of the genome, which actually match with the input sequence.

The small portion of the sequence that we had taken gave the result, a part of which has been shown in figure 2.

Query	82 GTTCATGGTAAAAATCCGCTTGGACTGACGCTTC AAACGCGGTATGTCATTAAGC	136
Sbjct	I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I 17 GTTCATGGTAAAAATCCGCTTGGAC TGACGCTTCAAACGCGGTATGTCATTAAGC	71

Fig 2: The BLAST analysis of a small part of the genome of the *Musca domestica* strain.

So if we have a sequence of genome with us, we can easily find out of which organism that gene sequence, most probably, is. This procedure of lining up two sequences to attain the maximum levels of identity for the purpose of assessing the degree of similarity and the possible homology is known as pair wise sequence alignment.

Gene Finding: Just like sequence alignment uses a simple sequence to extract the homology, the same sequence can be also used to encode where on the entire genome, actually the genes are present. Genes are the functional and physical entity of heredity. The identification of their location on the genome, can provide some of the majorly useful information related to the organism under observation.

GENEID a program based on the algorithm that helps researchers to predict genes, exons, splice sites and other signals along a DNA sequence. Similarly, JIGSAW a program that allow us to predict gene models using the output from other annotation software.

Active site and Transcription factor binding site identification: This involves the identification of the sites on a sequence, where another biomolecule can bind. Binding site on a protein sequence can be understood as an area on it, which can form bonds, with some other molecule, with specificity. This binding can be a protein binding to another protein or any enzyme substrate etc. This is also sometimes referred to as a substrate binding to its specific ligand, which in most of the cases, is responsible for triggering a conformational change within the protein molecule because of which the related cellular functions are also altered.

Active sites are those areas on the protein, which are chemically active and are involved in most of the protein-protein interactions. Protein-protein interactions (PPis) are physical contacts between two or more protein molecules, having very high specificity, which can also occur due to the hydrogen bonding or electrostatic bonding's.

A term widely used in this aspect, is Docking, which actually involves the prediction of who a protein or enzyme binds with other smaller ligands. It is a molecular modelling technique, in which two or more molecules interact, to give a stable molecule, while also predicting the orientation and position of the ligands, within the specific sites.

In many of the docking software's, we can actually see the chemical process going on between the proteins and its ligands, and gradually the most stable molecule possible after the reaction, is obtained.

Software's like 3DLigandSite and many more, provide a helping hands for predicting these binding sites, within the protein, by using its sequence or structure.

Protein Structure Determination: If we have a protein sequence, and want to determine the proper structure of the same, Bioinformatics has the tools for this too. With just a sequence of protein, we can understand the primary, secondary and even tertiary structure of the protein.Protein Data Bank is a perfect example of such a database, which takes the protein sequence an input, and gives the most probable 3 dimensional structure of the same.

RESULTS AND CONCLUSION:

In this review, we have tried to understand the emergence of the very complex though exciting field of Bioinformatics. We have given an idea through which one can understand how the merging of the two fields of Biology and computers have provided some really powerful, helpful and realistic tools to the researchers who are enthusiastic to work in this field.

The sequences of the DNA/RNA/Proteins, are the key to almost all the functions that make it possible for a cell of a living organism to work in the proper way. We all know that the proper functioning of the underlying cell is responsible for the proper physical and mental growth of any living organism. By this we get an idea of how important these sequences and the information hidden within them are of living being.

We have seen how Bioinformatics tools are lending help right from the initial stages of sequence extraction, evolution, identification, and analysis to the complex ones including the structural and functional studies and chemical interactions, to form the most stable products. It is the action of these stable products that are responsible for the effects that are seen to bring differences in the cellular processes.

The topics discussed here are just the beginning of this vastly diverse and challenging field ofBioinformatics. Many more advanced areas are associated with this field which also includeComputer aided drug designing(the discovery and development of a new drug),

Microarray(study of the gene expression of a large number of gene by the use of a microchip) etc.

This areas are much more complex and allow a much better use of data available. The recent advancement in the field is also moving at faster pace today, than ever before. This would definitely lead to the better understanding to the life, from its basics. The vision of this field can unfold many secrets and provide many answers to questions related to the living beings.

REFERENCES:

1. Computational Biology & Bioinformatics: A Gentle Overview. Achuthsankar S Nair University of Kerala. India.
2. What is bioinformatics? An Introduction and overview. November 2000. Yearbook of Medical Informatics. I0(01). Nicholas M Luscombe, Dov Greenbaum, Mark Gerstein.
3. Database resources of the National Center for Biotechnology Information. January 2016. Nucleic Acids Research 44(Dl):D7-Dl9. R. Agarwala, T. Barrett. J. Beck et al.
4. BLAST: An introductory tool for students to Bioinformatics Application. December 2013. Gareth Syngai, Prajan Barman, Rupjyoti Bharali, Sudip Dey.
5. Bioinformatics and its applications. Alla L Lapidus.