International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 4, June 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Computational of Bioinformatics

Durgesh Raghuvanshi[1], Vivek Solanki[2], Neha Arora[3], Faiz Hashmi[4]

[1]Student of Computer Science, [2]Student of Biotechnology,
[3]Assistant Professor Department of Computer Science, [4]M.Tech Student of Biotechnology,
[1,2,3,4]IILM Academy of Higher Learning, Greater Noida, Uttar Pradesh, India

ABSTRACT

Computational methods to analyze biological data. It is a way to introduce some of the many resources available for analyzing sequence data with bioinformatics software. This paper will cover the theoretical approaches to data resources and we will get knowledge about some sequential alignments with its databases. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics, and statistics to analyze and interpret biological data. Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques. Databases are essential for bioinformatics research and applications. Many databases exist, covering various information types: for example, DNA and protein sequences, molecular structures, phenotypes, and biodiversity. Databases may contain empirical data. Conceptualizing biology in terms of molecules and then applying "informatics" techniques from math, computer science, and statistics to understand and organize the information associated with these molecules on a large scale. In this materialistic world, People are studying bioinformatics in different ways. Some people are devoted to developing new computational tools, both from software and hardware viewpoints, for the better handling and processing of biological data. They develop new models and new algorithms for existing questions and propose and tackle new questions when new experimental techniques bring in new data. Other people take the study of bioinformatics as the study of biology with the viewpoint of informatics and systems.

KEYWORDS: *algorithms, alignments, web catalogs, sequencing, software alignments*

How to cite this paper*:* Durgesh Raghuvanshi | Vivek Solanki | Neha Arora | Faiz Hashmi "Computational of Bioinformatics" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-4, June 2020, pp.128-131, URL: www.ijtsrd.com/papers/ijtsrd30891.pd

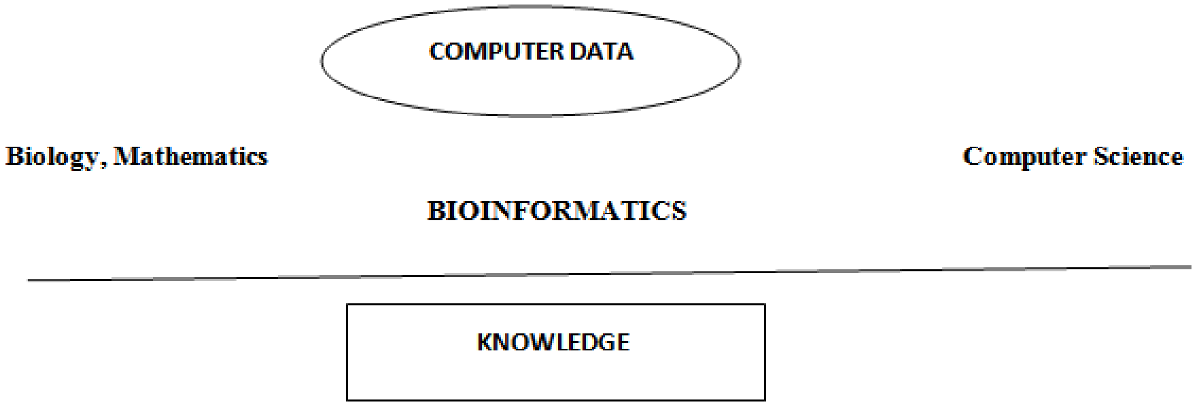
IJTSRD30891

INTRODUCTION

Bioinformatics has become a hot research topic in recent years, a hot topic in several disciplines that were not so closely linked to biology previously. Side evidence of this is the fact that the 2007 Graduate Summer School on Bioinformatics of China had received more than 800 applications from graduate students from all over the nation and a wide collection of disciplines in biological sciences, mathematics and statistics, automation and electrical engineering, computer science and engineering, medical sciences, environmental sciences, and even social sciences. It is always challenging to define a new term, especially a term like Bioinformatics that has many meanings. As an emerging discipline, it covers a lot of topics from the storage of DNA data and the mathematical modeling of biological sequences, to the analysis of possible mechanisms behind complex human diseases, to the understanding and modeling of the evolutionary history of life, etc. Another term that often goes together or closes with Bioinformatics is computational molecular biology, and also computational systems biology in recent years or computational biology as a more general term. People sometimes use these terms to mean different things, but sometimes use them in exchangeable manners. In our understanding, computational biology is a broad term, which covers all efforts of scientific investigations on or related to biology that involves mathematics and computation. Computational molecular biology, on the other hand, concentrates on the molecular aspects of biology in computational biology, which therefore has more or less the same meaning with Bioinformatics. No matter what type of Bioinformatics one is interested in, a basic understanding of existing knowledge of biology, especially molecular biology is a must. This theory was designed as the first course in the summer school to provide students with non-biological backgrounds a very basic and attractive understanding of molecular biology. It can also give biology students a clue how biology is understood by researchers from other disciplines, which may help them to better communicate with Bioinformatics. Generally, Bioinformatics is an integrative field for developing the technologies and tools of software to understand the biological data. As the name Bioinformatics applications in computer science symbolize that, this field associated with computer science, mathematics, biology, and statistics for determining and depicting the biological data. Additionally, it also holds some other fields rather than this. So it is denoted as a multidisciplinary course.

Aims of Bioinformatics

In general, the aims of bioinformatics are three-fold. First, at its simplest bioinformatics organizes data in a way that allows researchers to access existing information and to submit new entries as they are produced, e.g. the Protein Data Bank for 3D macromolecular structures. While data-curation is an essential task, the information stored in these databases is essentially useless until analyzed. Thus the purpose of bioinformatics extends much further. The second aim is to develop tools and resources that aid in the analysis of data. For example, having sequenced a particular protein, it is of interest to compare it with previously characterized sequences. Bioinformatics aims to increase the biological process of understanding. In computer science, its role is the same as for increasing the understanding of this through several fields such as statistics and mathematics. In the same way, it has three aims for the process. They are storing the biological data, developing the tools that are essential to processing the data, and the important goal of this is to exploit the computational tools for analyzing the data that simply depicts the results.



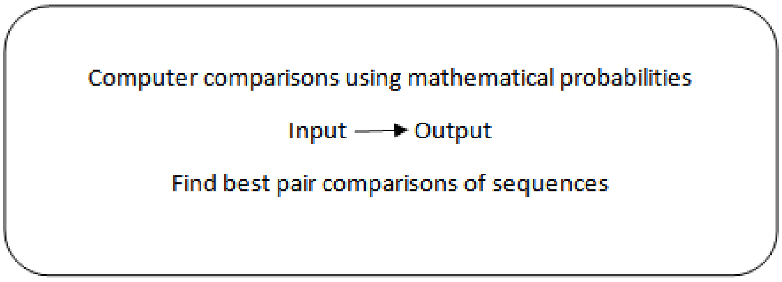
Algorithms used in bioinformatics

An algorithm is a sequence of unambiguous instructions for solving a problem, i.e., for obtaining a required output for any legitimate input in a finite amount of time. It gives an illustrative description of the relationship between problem, algorithm, and, the input and output of an algorithm. The time complexity measures the efficiency of algorithms. Euclid's algorithm is more efficient than the naive algorithm because it has a small-time complexity.

To summarize, an algorithm has the following important properties:

- Can be represented in various forms
- Unambiguity/clearness
- Effectiveness
- Finiteness/termination
- Correctness

The finiteness and correctness of an algorithm are self-clear. No one wants an algorithm to give a wrong result or an algorithm that runs forever without giving a final An algorithm is a sequence of instructions that one must perform to solve a well-formulated problem. We will specify problems in terms of their inputs and their outputs, and the algorithm will be the method of translating the inputs into the outputs. A well-formulated problem is unambiguous and precise, leaving no room for misinterpretation. To solve a problem, some entity needs to carry out the steps specified by the algorithm. A human with a pen and paper would be able to do this, but humans are generally slow, make mistakes, and prefer not to perform repetitive work. A computer is less intelligent but can perform simple steps quickly and reliably. A computer cannot understand English, so algorithms must be rephrased in a programming language such as C or Java to give specific instructions to the processor. Nature uses algorithm-like procedures to solve biological problems, for example, in the process of DNA replication. Before a cell can divide, it must first make a complete copy of all its genetic material. DNA replication proceeds in phases, each of which requires elaborate cooperation between different types of molecules. For the sake of simplicity, we describe the replication process as it occurs in bacteria, rather than the replication process in humans or other mammals, which is quite a bit more involved.


Computer comparisons using mathematical probabilities
Input → Output
Find best pair comparisons of sequences

Global and local alignments

The technique of dynamic programming can be applied to produce Global alignments via the Needleman-Wunsch algorithm and local alignments via the Smith-Waterman algorithm. There are two general models to view alignments. The first model considers similarity across the full extent of the sequences (Global alignment). The second focuses on the regions of similarity in parts of the sequence only. (it is local alignment). A search for local similarity may produce more biologically meaningful and sensitive results than a global alignment.

- Global alignment: Needleman Wunsch algorithms

Global alignments attempt to align every residue in every sequence and they are most useful when the sequences in the query set are similar and of roughly equal size. Needleman and Wunsch's algorithm is used for computing a global alignment between two sequences and it is based on dynamic programming. The algorithm proposed a maximum match pathway that can be obtained computationally by applying some rules. Here cells representing identities are scored 1 and cells representing mismatches are scored 0. This process examines each cell in the matrix and finally, a summation of cells is started. When this process is completed, the maximum match pathway is constructed. Thus in global alignment comparison of the two sequences over the entire length is done. The Needleman Wunsch algorithm for global alignment is time-consuming to run if the sequences are long. This is a general algorithm for sequence comparison. It maximizes a similarity score to give a maximum score. The maximum match is the largest number of residues of one sequence that can be matched with another allowing for all possible deletions.

- Local alignments: Smith-Waterman algorithm

Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs. Local alignment searches for regions of local similarity and need not include the entire length of the sequences. Local alignment methods are very useful for scanning databases. The smith-Waterman algorithm is used for local alignments. Even if the two given sequences are dissimilar, there will be some local similarity between sequences. The smith-Waterman algorithm is used to find out this local similarity. The key feature of the Smith-Waterman algorithm is that each cell in the matrix defines the endpoint of a potential arrangement. The algorithm thus begins by filling the edge elements with 0.0 (floating point) values. Now the remaining cells in the matrix are compared. Three functions are compared at a time and the maximum of these three is chosen. Once the matrix is complete, the highest score is located. It represents the endpoints of alignment with maximum local similarity.

Software for pairwise alignments

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural, and/or evolutionary relationships between two biological sequences (protein or nucleic acid). By contrast, Multiple Sequence Alignment (MSA) is the alignment of three or more biological sequences of similar length.

Dot-matrix methods. Self-comparison of a part of a mouse strain genome. The dot-plot shows a patchwork of lines, demonstrating duplicated segments of DNA. The technique of dynamic programming can be applied to produce global alignments via the Needleman-Wunsch algorithm, and local alignments via the Smith-Waterman algorithm. Word methods, also known as k -tuple methods, are heuristic methods that are not guaranteed to find an optimal alignment solution but are significantly more efficient than dynamic programming.

Multiple sequence alignments

- Makemultiplesequencealignmentfortheproteinsequence ofhemoglobinalpha chain from 7 vertebrates [FASTA].
- Make multiple sequence alignment for the protein sequence of 12 human globins [FASTA].
- Make multiple sequence alignment for the protein sequence of 15 Arabidopsis SBP transcription factors [FASTA]; use different programs (ClustalW, T-Coffee, and DIALIGN) and compare the results.
- Makemultiplesequencealignmentforthe9repeatsequence sofhumanubiquitin C protein (NP 066289).
- Make multiple sequence alignment for spider toxin peptides [FASTA]; use manual editing to improve the results.

Sequence, Structure, and Function Analysis

Hemoglobin is one of the most well-studied proteins in the last century. The sequence, structure, and function of several vertebrates have been investigated during the past 50 years. More than 200 hundreds of hemoglobin protein sequences have been deposited into the Swiss-Prot database. Three-dimensional structure wild type and mutants from dozens of species have been solved. This provides us a good opportunity to study the relationship between sequence, structure, and function of hemoglobin. Bar-headed go sees special species of migration birds. They live in the Qinghai Lake during summertime and fly to India along over the Tibetan plateau in autumn and come back in spring. Interestingly, a close relative of bar-headed goose, the graylag goose, lives in the low land of India all year round and do not migrate. Sequence alignment of bar-headed goose hemoglobin with that of graylag goose shows that there are only 4 substitutions.

Edit Distance and Alignments

In this materialistic world, we have been vague about what we mean by "sequence similarity" or "distance" between DNA sequences. Hamming distance (introduced in chapter4), while important in computer science, is not typically used to compare DNA or protein sequences. The Hamming distance calculation rigidly assumes that the symbol of one sequence is already aligned against the symbol of the other. However, it is often the case that the ith symbol in one sequence corresponds to a symbol at a different—and unknown—position in the other. For example, the mutation in DNA is an evolutionary process: DNA replication errors cause substitutions, insertions, and deletions of nucleotides, leading to "edited" DNA texts. Since DNA sequences are subject to insertions and deletions, biologists rarely have the luxury of knowing in advance whether the ith symbol in one DNA sequence corresponds to the ith symbol in the other.

For example

Spliced Alignment Problem: Find a chain of candidate exons in a genomic sequence that best fits a target sequence.

Input: Genomic sequence G, target sequence T, and a set of candidate exons (blocks) B.

Output: A chain of candidate exons Γ such that the global alignment score s (Γ∗, T) is maximum among all chains of exons from B.

Bioinformatics databases

There are huge amounts of online bioinformatics databases available on the Internet.

- NAR databases–the most extensive list of biological databases being maintained by the international journal

Nucleic Acids Research which publishes a special issue for molecular biology databases in the first issue of each year since 1996. All these database papers can be accessed freely. You may find links to the website of the databases described in the chapter.

- NCBI databases – the molecular databases maintained by NCBI. A Flash flowchart for 24 databases connected by lines shows the relationships and internal links among all these databases. These databases are divided into 6 major groups: nucleotide, protein, structure, taxonomy, genome, and expression. It also provides links to the individual database description page.
- EBI databases – the main portal to all EBI databases divided into several groups, such as literature, microarray, nucleotide, protein, structure, pathway, and ontology. Links to database query and retrieval systems can be found in this portal.

Open source for bioinformatics

Many free and open-source software tools have existed and continued to grow since the 1980s.[38] The combination of a continued need for new algorithms for the analysis of emerging types of biological readouts, the potential for innovative in silico experiments, and freely available open code bases have helped to create opportunities for all research groups to contribute to both bioinformatics and the range of open-source software available, regardless of their funding arrangements. The open-source tools often act as incubators of ideas, or community-supported plug-ins in commercial applications. They may also provide de facto standards and shared object models for assisting with the challenge of bio-information integration. The range of open-source software packages includes titles such as Bioconductor, BioPerl, Biopython, BioJava, BioJS, BioRuby, Bioclipse, EMBOSS, .NET Bio, Orange with its bioinformatics add-on, Apache Taverna, UGENE and GenoCAD. To maintain this tradition and create further opportunities, the non-profit Open Bioinformatics Foundation.

Conclusion

With the confluence of biology and computer science, the computer applications it becomes imperative for biologists to seek the help of information technology professionals to accomplish the ever-growing computational requirements of a host of exciting and needy biological problems, the synergy between modern biology and computer science is to blossom in the days to come. Thus the research scope for all the mathematical techniques and algorithms coupled with software programming languages, software development, and deployment tools is to get a real boost. Computational biology, which includes many aspects of bioinformatics, is the science of using biological data to develop algorithms or models to understand biological systems and relationships. Until recently, biologists did not have access to very large amounts of data. This data has now become commonplace, particularly in molecular biology and genomics. Researchers were able to develop analytical methods for interpreting biological information but were unable to share them quickly among colleagues.

References

[1] Wren, J. D. (2004), The stability and persistence of URLs published in MEDLINE', Bioinformatics, Vol. 20, pp. 668– 672 (DOI: 10.1093/bioinformatics/btg465).

[2] Baxevanis, A. D., Bader, G. D., & Wishart, D. S. (Eds.). (2020). Bioinformatics. John Wiley & Sons.

[3] Mount, D. W., & Mount, D. W. (2001). Bioinformatics: sequence and genome analysis (Vol. 1). New York:: Cold spring harbor laboratory press.

[4] Gu, J., & Bourne, P. E. (Eds.). (2009). *Structural bioinformatics* (Vol. 44). John Wiley & Sons.

[5] Lesk, A. (2019). Introduction to bioinformatics. Oxford university press.

[6] Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422-1423.

[7] Lio, P. (2003). Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics, 19(1), 2-9.