**BIOINFORMATICS: AN EMERGING**

 **TOOL FOR BIOLOGICAL SCIENCE**

**\*RABIYA PARVEEN1, MANKESH KUMAR1, ZAFAR IMAM1, DIGVIJAY SINGH1, SWAPNIL2, ABHINAV KUMAR1**

1. **DEPARTMENT OF PLANT BREEDING AND GENETICS, BIHAR AGRICULTURAL UNIVERSITY, SABOUR, BHAGALPUR, 813210, INDIA**
2. **DEPARTMENT OF GENETICS AND PLANT BREEDING**

**CENTURION UNIVERSITY OF TECHNOLOGY AND MANAGEMENT, PARALAKHEMUNDI, ODISHA, 761211, INDIA**

**\*Email-** **rparveen181@gmail.com**

**ABSTRACT**

Recent advances in technology have accelerated the understanding of the genetic basis of phenotypes. With these developments, genomics has altered the way that biological challenges are thought about on a genome-wide (genome-wide) scale, revealing a large amount of information and creating a myriad of opportunities. One of these recently developed fields is bioinformatics, which uses molecular biology, computer science, mathematics, and statistics to store, retrieve, and analyze biological data. Even though it is still in its infancy, it has quickly emerged as one of the field with the quickest growth rates and established itself as a crucial part of any biological research program. It is becoming more well-known because of its capacity to quickly and affordably analyze vast amounts of biological data. A biologist can use a variety of web- and/or computer-based tools provided by bioinformatics to extract valuable information from biological data, the majority of which are free to use. This introductory chapter aims to provide overall picture on basics and advancement in the field of bioinformatics benefitting readers in various fields of Biological sciences.

**Keywords :** Genomics, Genome-wide, Bioinformatics, database, molecular biology

1. **INTRODUCTION**

Due to enormous advancements in the domains of molecular biology and genomics, the amount of biological information has greatly increased during the genomic era. The field of bioinformatics has rapidly grown since its inception in the 1980s, keeping up with the growth of genome sequence data. An interdisciplinary field of study known as bioinformatics was created by the integration of several diverse fields, including biology, mathematics, computer science, and statistics, in order to provide techniques for the storage, retrieval, and analysis of biological data [30]. The rapid use of omics techniques and its growing power and more affordable costs has greatly boosted the amount of molecular data collection from various levels of organization of an organism or environmental sample. The sequencing of nucleic acids underwent yet another revolution with the recent introduction of Next-Generation Sequencing (NGS) technologies, ushering in a new age in omics methods.

 The word bioinformatics was coined by Paulien Hogeweg (1979). This discipline has grown significantly since the introduction of user-friendly interactive automated modelling and the development of SWISS-MODEL server around 18 years ago [35]. Since then, it has become an essential part of biological sciences to process biological data at a much faster rate with the databases and informatics working at the backend. Large-scale biological data management, analysis, and manipulation fall under the purview of the field, which includes all computational tools and techniques [26]. Electronic databases are needed to store, organize, and index the enormous amount of sequence data produced by the various sequencing programs. Additionally, the databases require specific tools so that researchers can access, evaluate, and contribute new or updated sequence data. Bioinformatics technologies can be used to perform reconstruction, pattern recognition, folding, simulation, and molecular modelling to find structural quirks and interactions of molecular sequences that are crucial for structural biology and the development of pharmaceuticals [38]. It is difficult to manually examine all of these enormous, genome-derived, molecular sequence studies of raw "Big Data." [28].

1. **DEVELOPMENT OF BIOINFORMATICS**

 The networking of computers and the collection of data on genes and proteins marked the beginning of the development of bioinformatics. In 1956, the 51 amino acid residues of bovine insulin were described as the first protein sequence. A few years after the first protein sequence became accessible; the first bioinformatics database was built. Physical chemist Margaret Dayhoff, who lived in America from 1925 to 1983, was a pioneer in the use of computational techniques in the field of bioinformatics. She gathered all the sequence data that was available and created the "Atlas of Protein Sequence and Structure," the first bioinformatics database. Dayhoff’s contribution to this field is so important that David J. Lipman, former director of the National Center for Biotechnology Information (NCBI), called her ‘the mother and father of bioinformatics’ [30]. This fact is due to her role in the development of computers able to determine the peptide sequence, programs to recognize and display structures for use in X-ray crystallography and computational methods for protein sequence comparison, allowing us to infer the evolutionary connections among kingdoms [20,48]. Dayhoff developed three-letter abbreviations (e.g. Lys for lysine, Ser for serine) and later the one-letter amino acid code that is still in use today (Table 1). This one-letter code was first used in Dayhoff and Eck’s 1965 Atlas of Protein Sequence and Structure

[9], the first ever biological sequence database.

**Table 1: Symbols used to represent amino acid in Protein Sequences**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |  |  |
| --- | --- | --- |
| **Single letter code** | **Amino acid** | **Three letter code** |
| A | Alanine | Ala |
| B | Asparagine or Aaspartic acid | Asx |
| C | Cystine | Cys |
| D | Aspartic acid | Asp |
| E | Glutamic acid | Glu |
| F | Phenylalanine | Phe |
| G | Glycine | Gly |
| H | Histidine | His |
| I | Isoleucine | Lie |
| K | Lysine | Lys |
| L | Leucine | Leu |
| M | Methionine | Met |
| N | Asparagine | Asn |
| P | Proline | Pro |
| Q | Glutamine | Gln |
| R | Arginine | Arg |
| S | Serine | Ser |
| T | Threonine | Thr |
| V | Valine | Val |
| W | Tryptophan | Trp |
| X | Any amino acid | Xaa |
| Y | Tyrosine | Tyr |
| Z | Glutamine or Glutamic acid | Glx |

 |

 The foundation of the Indian Institute of Science in 1990 and the Bioinformatics Institute of India in 2002 helped the field of bioinformatics to expand in India during the 1980s. GN Ramachandran is considered as the god father of Indian bioinformatics.  The Department of Biotechnology published a plan in 2004 to make India a centre for bioinformatics worldwide.

 In order to give immunology experiments a theoretical basis, George Bell and colleagues started the collection of DNA sequences into GenBank in 1974. The first version of GenBank was created by Walter Goad's team [30] between 1982 and 1992, and their work led to the creation of the now-famous and widely used DNA sequence databases of GenBank [16], "The European Molecular Biology Laboratory (EMBL) [43], and DNA DataBank of Japan (DDBJ) [8] in 1979, 1980, and 1984, respectively. The introduction of web-based searching engines, which enables researchers to identify and compare the target DNA sequences, was the most significant breakthrough in DNA sequence databases. David Benson, David Lipman, and colleagues created such first advances and the resulting software called "GENEINFO" and its derived version, "Entrez." [30]. By using this program, researchers could quickly search database-indexed sequences and compare them to the sequence they were querying. Software became easily accessible through web-based interface of the National Center for Biotechnology Information (NCBI) database [44]. Methods for molecular sequence comparison, analysis, and visualization have improved, and numerous strategies have contributed to bioinformatics developments in this area.

1. **Branches of Bioinformatics**

 Since the first draft of the human genome was completed [22, 47] the focus has shifted from genes themselves to gene products. In functional genomics, genetic information is given a functional relevance. Analysis and interpretation of biological data considers information not only at the level of the genome but at the level of the proteome and the transcriptome. Proteomics is the analysis of the total amount of proteins (proteome) expressed by a cell, and transcriptomics refers to the analysis of the messenger RNA transcripts produced by a cell (transcriptome).

**COMPLEXITY**

TRANSCRIPTOMICS

GENOMICS

PROTEOMICS

**Figure 1.Flow diagram representing complexity of genomic data processing (Source: Bayat A, 2002)**

 Gene is considered as the basic unit of heredity responsible for the transmission of traits from one generation to the next, and are stored in a genome as DNA molecules. Generating detailed genetic and physical maps of the genome in order to designate segments with ever higher resolution and sequentially organize the segments is the primary goal of genome sequencing.  (17). A method known as the direct shotgun methodology is also used for genome sequencing [46]. The goal of this method is to break the genome into overlapping, random fragments, and then use computer algorithms to sequence the fragments and put the sequences together. Analysis of genomic sequences reveals that each organism has an array of genes required for basic metabolic processes and genes whose products determine the specialized function of the organism. Complete genome sequencing provides a knowledge based on information can be obtained about gene and protein expression, based on the analysis of genome sequences, each organism possesses a number of genes necessary for fundamental metabolic activities as well as genes whose products dictate the organism's specific function. Complete genome sequencing offers a foundation upon which to construct knowledge about the expression of genes and proteins [25], but it is not always adequate to describe all of the protein constituents of an organism. In proteomics, the amino acid sequence of a protein is analyzed to determine its three-dimensional structure and to link it to the protein's specific function. Applications of bioinformatics in the field of proteomics included the study of amino acid sequences, the discovery of protein binding partners, the detection of polymorphism, splice variants, and post-translational modifications. The major technologies in the field of proteomics include two-dimensional gel electrophoresis, mass spectrometry, and protein microarrays. Bioinformatics tools are essential for understanding and extracting information from the data produced by these instruments. The functional analysis includes gene expression profiling, protein–protein interaction prediction, protein sub cellular localization prediction, metabolic pathway reconstruction and simulation [40, 50]. These aspects of bioinformatics analysis are not isolated, but often interact to produce integrated results.

1. **Databases**

 Biological databases are collections of biological sciences gleaned from high throughput experimentation techniques, published literature, and computational analysis. Storage and management of biological data and information in computer accessible forms is the primary goal of biological databases [2]. An integrated collection of computer software known as a "database management system (DBMS)" allows access to the information in the databases. By using this software users get access to all of the data stored in the databases. The data that is stored can be used as primary source as well as for future use. As a result of the range of information that they store, databases are divided into three categories: *primary*, *secondary*, and *composite*.

1. **Primary Sequence Database(s)** are those which contain information of the sequences or structure alone and acts as repositories of raw sequence data and can be accessed freely over the World Wide Web (WWW). For e.g. GenBank, maintained by National Center of Biotechnology Information (NCBI), the Nucleotide Sequence Database maintained by the European Molecular Biology Laboratory (EMBL) and the DNA Databank of Japan (DDBJ) for genome sequence.

 It obtains unique data obtained from the laboratory and these data are made accessible to normal users without any change. Each data when entered has their unique accession number through which data can be later retrieved.

1. **Secondary sequence database(s)** contains information that is derived from the analysis of data stored in primary databases like conserved sequences, active sites of a protein family or conserved secondary motifs of protein molecules [13,18]. Computational algorithms are applied to the primary database and meaningful and informative data is stored inside the secondary database. Some of the databases like SCOP developed at Cambridge University, CATH developed at university college of London, eMOTIF at standford etc. are created and hosted by individual researchers at their individual laboratories.
2. **Composite database(s)** includes variety of primary database sources which obviates the need to search multiple resources. The NCBI i.e. National Centre for Biotechnology Information, hosts nucleotide and protein databases in their large high available arrays of computer servers.

**Table 2: LIST OF IMPORTANT DATABASES**

|  |  |  |
| --- | --- | --- |
| **DATABASE** | **DESCRIPTION** | **REFERENCES** |
| **Nucleotide Databases** |  |
| DNA Databank of Japan (DDBJ) | It is the member of International Nucleotide Sequence Databases (INSD) and is one of the biggest resources for nucleotide sequences | [8] |
| European MolecularBiology Laboratory(EMBL) | Repository of DNA and RNA sequences that is complementary to GenBank and DDBJ | [43] |
| GenBank | It is the member of International Nucleotide Sequence Databases (INSD). Main nucleotide sequence database of NCBI, USA | [16] |
| **Protein Databases** |  |
| Uniprot | One of the largest collection protein sequences | [45] |
| Protein DataBank | Major resource of proteins containing information of experimentally determined structures of nucleic acids, proteins and their complex assemblies |  [5] |
| Prosite  | Provides information on protein families, conserved domains and active sites of the proteins | [41] |
| SWISS-PROT | A section of the UniProt knowledgebase containing manually annotated protein sequences | [6] |
| InterPro | Integrated resource for protein families, conserved domains and active sites | [39] |
| SCOP | Familial and Structural Protein Relationship | [15] |
| **Genome Database** |  |
| Ensemble Plants | An integrative resource presenting genome scale information for a growing number of sequenced plant species | [14] |
| Protein Information Resource ( PIR) | Comprehensive, annotated, non-redundant protein sequence database  | [49] |
| Phytozome  | A comparative hub for plant genome and gene family data and analysis | [19] |
| **Miscellaneous databases** |  |
| TAIR | The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular data for the model plant *Arabidopsis thaliana.*  | [42] |
| KEGG | Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledgebase for systematic analysis of gene functions, linking genomic information with higher order functional information. | [24] |

1. **Gene identification and sequence analyses**

 The escalating amount of data from the genome projects has necessitated computer databases that feature rapid assimilation, usable formats and algorithm software programs for efficient management of biological data [4]. Sequence analyses refer to the understanding of different features of a biomolecules like nucleic acid or protein, which give to it its unique function. First, the sequences of corresponding molecules are retrieved from public databases. After refinement, if needed, they are subjected to various tools that enable prediction of their features related to accuracy. The analysis is useful for the identification of promoter, terminator, or un-translated regions involved in the expression regulations, recognition of a transit peptide, introns, exons or an open reading frame (ORF), and identification of certain variable regions to be used as signatures for diagnostic purposes. There are number of tools developed for this purpose, some of the important tools (enlisted in Table 3 with function).

**Table 3: Tools for Primary Sequence Analysis**

|  |  |
| --- | --- |
| **Tools** | **Description** |
| BLAST (Local Sequence Alignment) | It is a search tool, used for DNA or Protein sequence search based on identity |
| Clustal Omega (Global Sequence Alignment) | Multiple Sequence Alignments may be performed using this program |
| HMMER  | Homologous Protein sequences may be searched from the respective databases using this tool |
| Open Reading Frame (ORF) Finder | The putative genes may be subjected to this tool to find ORF |
| ProtParam | Used to predict the physic-chemical properties of proteins |
| PPP | Prokaryotic Promoter Prediction tool used to predict the promoter sequences present upstream of the gene |
| JIGSAW | To find genes, and to predict the splicing sites in the selected sequences  |
| Genscan | Used to predict the intron-exon sites in genomic sequences  |
| Softberry Tools | Several tools are specialized in annotation of animal, plant and bacterial genomes along with the structure and function prediction of RNA and Proteins |

1. **Sequence Alignment**

 For comparing related DNA or protein sequences, alignments are a potent tool. They can be used to record different information regarding the matched sequences, like shared structural function or common evolutionary ancestry. Any biological experiment that compares two or more biological sequences must perform sequence alignment as a key step, either explicitly or implicitly (whether DNA, RNA, or protein). The classification of genes and proteins, the prediction of biological function, the detection of point mutations, the construction of evolutionary trees, and other processes all rely on the sequence similarity.  If two sequences in an alignment have a common ancestor, mismatches and gaps can be read as point mutations and insertion or deletion mutations, respectively, introduced in one or both lineages since they diverged from one another, . The degree of similarity between amino acids at a certain position in a protein's sequence can be used as a rough indicator of how conserved a given region or sequence motif is across lineages. A specific section of the sequence may have structural or functional significance [32] if there are no substitutions or only highly conservative substitutions (i.e., substitution of amino acids whose side chains have similar biochemical properties).

 Sequence alignment seeks to reduce mismatches and gaps while maximising the number of matches. Three scores are given in the simplest scenario: (1) the cost of aligning a character in one sequence with a gap in the other sequence (2) The advantage of aligning a pair of sites that have the same character (state) in both sequences. (3) The cost of aligning a pair of sites that include different characters.

**Sequence Alignment Types**

|  |  |
| --- | --- |
| Based on Number | Based on Length |
|  | Pairwise Sequence Alignment | Global Sequence Alignment |
|  | Multiple Sequence Alignment | Local Sequence Alignment |

1. **Pairwise Sequence Alignment (PSA)**

 **Pairwise sequence alignment** is a form of sequence alignment, where we compare only two sequences. A type of sequence alignment called pairwise comparison compares just two sequences. This procedure entails determining the best alignment between the two sequences, scoring according to how similar or distinct they are, and determining the importance of this score. Each pair of sequences is aligned by PSA once. The rapid search for matches to a query sequence in large DNA and protein databases is one particular application of pairwise sequence alignment .Algorithms based on dynamic programming are substantially slower than well-known heuristic algorithms as those from the FASTA [34] or BLAST families. It is the most basic type of alignment and can be carried out using either a *global* or *local* sequence alignment approach.

1. **Global Sequence Alignment**

 Sequences which are closey related  and having same length are ideal for global alignment. To determine the best possible alignment, the alignment is performed here from the beginning to the end of the sequence. Saul B. Needleman and Christian D. Wunsch created the dynamic programming approach for sequence alignment (36) known as the Needleman-Wunsch algorithm in 1970 (31). The dynamic programming solves the original problem by dividing the problem into smaller independent sub problems.  The algorithm for aligning nucleotide or protein sequences describes global sequence alignment.

TARGET SEQUENCE

5’ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3’

 5’ ACTACTAGATT­\_\_\_\_\_ACGGATC\_ \_ GTACTTTAGAGGCTAGCAACCA 3’

QUERY SEQUENCE

1. **Local Sequence Alignment**

 Smith and Waterman were the one who first proposed the fundamental local alignment method (1981b). Similar to the Needleman-Wunsch algorithm, this method is likewise based on dynamic programming, but with more options to start and terminate at any location (36). It identifies nearby areas with a high degree of resemblance. Any two sequences can be locally aligned because local alignment identifies sequence stretches with a high level of similarity without taking the alignment of the other sequence regions into account.

TARGET SEQUENCE

 5 ’ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA3’

 5’TACTCACGGATGAGGTACTTTAGAGGC 3’

QUERY SEQUENCE

1. **Multiple Sequence Alignment**

 The alignment of three or more biological sequences (protein or nucleic acid) of comparable length is performed by using multiple sequence alignment, or MSA. The results allow for the study of the evolutionary links between the sequences and the inference of homology. A known phylogenetic tree was required for early alternate methods for multiple alignments. Progressive alignment is the method for multiple sequence alignment that eventually gained a lot of traction [11,12]. In progressive alignment, one typically begins by creating all feasible pairwise alignments (there are n (n-1)/2 pairs for n sequences). Using a distance-based procedure like the unweighted pair group method with arithmetic mean (UPGMA) or neighbor joining, these pairwise alignments are utilized to estimate a phylogenetic tree. A pairwise approach is used to match the most comparable sequences to one another using the tree as a guide. On the basis of the phylogenetic tree's structure, one then gradually adds sequences to the alignment, one sequence at a time.

 Higgins created CLUSTAL series of programmes, which employs a progressive algorithm, is one of the most effective MSA solutions because it uses heuristic methods with approximate approaches [11].

1. **DYNAMIC PROGRAMMING**

 Dynamic programming is applied for obtaining the best alignment of two sequences. Instead of only applying dots, it discovers the alignment in a more quantitative manner by providing certain scores for matches and mismatches (Scoring matrices). Alignment can be precisely found by looking through the matrix's highest scores.

1. **Substitution Matrices**

 Substitution matrices are used to score the sequence alignment by pairwise and multiple sequence alignment techniques. Since all bases experience equal amounts of mutation, the score matrices used for nucleotide sequence alignment are quite simple. A match is given a positive or higher value, whereas a mismatch is given a negative or lower value. The matrices can be scored using these scores based on assumptions. Point Accepted Mutation (PAM - Dayhoff 1978) and Blocks Substitution Matrix are two common protein substitution matrix models (BLOSUM - Henikoff and Henikoff 1992).

1. **PAM Matrices:** The PAM matrix, which stands for Point Accepted Mutations, was created for the first time by Margaret Dayhoff. The discrepancies in closely related proteins are used to calculate PAM matrices. One PAM unit (PAM1) specifies one permitted point mutation for every 100 amino acid residues, meaning that only 1% of the original structure is altered.
2. **BLOSUM:**Henikoff and Henikoff's created BLOcks SUbstitution Matrix in 1992, which utilizes conserved regions. Actual percent identity values make up these matrices. Blosum 62 means there is 62 % similarity.
3. **Gap (-)** representing one or more nucleotide indel events.

 About the algorithm used, it may be classified as optimal or heuristic [36]. The optimum result is the best alignment possible, while the heuristic, although not presenting an optimal result, presents the best alignment for a given period of analysis.

1. **BLAST**

The most commonly used similarity search method is the Basic Local Alignment Search Tool (BLAST). BLAST finds regions of similarity between sequences. The program compares nucleotide or protein sequences and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

 BLAST is a specific local alignment algorithm derived from the Smith-Waterman algorithm (1981) that presents a maximum alignment score of two sequences [1]. In addition to the dynamic programming arising from the algorithm mentioned above, BLAST employs a heuristic based on the *k*-tuple method to search the sequences in the database [23]. The *k*-tuple method limits the search to those words that are more significant, being the size of 3 and 11 characters for amino acids and nucleotides, respectively [1]. BLAST is a family of programs used for different purposes according to the type of sequence of interest and the database to be searched [37]. Several applications available by BLAST include those listed in Table 4

.

**Table 4: Description of BLAST family programs**

|  |  |  |
| --- | --- | --- |
| Program | Query | Subject |
| BLASTx | nt\* | aa |
| BLASTn | nt | nt |
| BLASTp | aa | aa |
| tBLASTn | aa | nt\* |
| tBLASTx | nt\* | nt\* |

nt: nucleotide ; aa: amino acid ; \*Translated for all possible sequences Source: Amaral et al (2007)

 The BLAST results are presented according to two parameters: the value of the score (Score bits) and the E-value. The E-value represents the statistical value that indicates the probability that the alignment did not occur at random, considering the alignment score and the database size [37, 1]. On the other hand, the score is attributed by the algorithm based on the matches and mismatches between the input sequences and database [1].

1. **Phylogenetic analyses**

Phylogenetics and sequence alignment are closely related fields due to the shared necessity of evaluating sequence relatedness [[33]](https://en.wikipedia.org/wiki/Sequence_alignment#cite_note-ortet-25). The field of [phylogenetics](https://en.wikipedia.org/wiki/Phylogenetics) makes extensive use of sequence alignments in the construction and interpretation of [phylogenetic trees](https://en.wikipedia.org/wiki/Phylogenetic_tree), which are used to classify the evolutionary relationships between homologous [genes](https://en.wikipedia.org/wiki/Gene) represented in the [genomes](https://en.wikipedia.org/wiki/Genome) of divergent species. The degree to which sequences in a query set differ is qualitatively related to the sequences' evolutionary distance from one another. The distance-matrix methods such as Neighbour Joining (NJ) or Unweighted Pair Group Method with Arithmetic mean (UPGMA) are the simplest.

**Table 5: Description of Tools to study Phylogenetic Relationship**

|  |  |
| --- | --- |
| Tools | Description  |
| MEGA (Molecular Evolutionary Genetics Analysis) | Builds phylogenetic tress to study the evolutionary closeness |
| PAML | A package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood. |
| PHYLIP | A package for phylogenetic studies |
| TreeView  | Software to view the phylogenetic trees, with the provision of changing view |
| itol (Interactive Tree of Life) | Online tool for display, annotation and management of phylogenetic tree |

1. **Application of Bioinformatics**

**Figure 2 Basic bioinformatics tools of biological sciences. (Source: Mehmood et al 2014)**

**Table 6: Application of basics Bioinformatic tools in various areas of Biological Sciences**

|  |  |
| --- | --- |
| **DNA Sequence Analyses** | * BLAST
* Clustal X
* Promoter Analyses
* Gene Prediction
* Regulatory Elements
* Intron - Exon finding
* Primer Designing
* Codon Usage Optimization
* Virtual Translation
 |
| **Molecular Dynamic Simulations** | * Protein-DNA Simulations
* Drug-DNA Simulation
* Protein-Ligand Simulation
 |
| **Drug / Pesticide Designing** | * Target identification
* Target validation
* Lead identification
* Lead optimization
* ADMET prediction
 |
| **Protein Sequence Analyses** | * Molecular mass, pI, amino acid composition
* Domain and Motifs search
* Single peptide identification
* Secondary structure analyses
 |
| **Phylogenetic Analyses** | * Reconstruction of evolutionary history
* Tracking gene flow
* Identification of conserved regions
 |
| **Molecular Interactions** | * Protein-Protein docking
* Finding inhibitors and activators of proteins
* Protein-DNA interactions
* Transcriptional factors identification
* Protein-Ligand interactions
 |

1. **Conclusion**

 The emerging discipline “Bioinformatics” is the solution for the current demand in the every field of plant research. The discipline implements a wide range of computational techniques including sequence analysis, data mining, gene finding, phylogenetic tree construction, prediction of protein structure and function, and interaction networks etc. Bioinformatics is an approach that will be a major role in the research of plants. On the off chance that plant science can be summed into a single word, it would be "integration." Bioinformatics will give the glue with which all of these types of integration will occur. With the advancement of big data era in plant science, a highly powerful technology like deep learning facilitates efficient information storage and management as well as the extraction of useful information from these data.

**REFERENCES**

1. Amaral AM, Reis MS and Silva FR (2007). O programa BLAST: guia prático de utilização. 1st edn. Embrapa Recursos Genéticos e Biotecnologia. EMBRAPA, Brasília.
2. Attwood TK, Gisel A, Eriksson, NE, Bongcam-Rudloff E (2011) Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective". *Bioinformatics - Trends and Methodologies*. InTech.
3. Bayat A. (2002) Bioinformatics: Science, Medicine and the Future. British Medical Journal.; 324; 1018-22
4. Benton D. (1996) Bioinformatics-principles and potential of a new multidisciplinary tool. *Trends Biotech*; 14:261­312.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.
6. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365-370.
7. Bostan H, Chiusano ML. (2015) NexGenEx-Tom: a gene expression platform to investigate the functionalities of the tomato genome. BMC Plant Biol.;15(1):48.
8. DataBank of Japan [Internet].(2016). http://www.ddbj.nig.ac.jp. Accessed: 2016-03-10
9. Dayhoff MO. (1965) National Biomedical Research Foundation. *Atlas of Protein Sequence and Structure*, Vol. 1. Silver Spring, MD: National Biomedical Research Foundation.
10. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., (1978). A model of evolutionary change in proteins. In: Dayhoff, M.O. (Ed.), Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington D.C., pp. 345-352.
11. Feng D-F, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25: 351–360
12. Feng D-F, Doolittle RF (1990) Progressive alignment of protein sequences. Methods Enzymol 183:375–387
13. Finn RD, Bateman A, Clements J, Marco Punta, Penny C Coggill, et al. (2014) Pfam: the protein families database. Nucl Acids Res 42: D222-D230.
14. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2012) Ensembl 2012. Nucleic Acids Res 40: D84-90.
15. Fox NK, Brenner SE, Chandonia JM (2014) SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42: D304-309.
16. GenBank [Internet]. 2016. http://www.ncbi.nlm.nih.gov/genbank. Accessed: 2016-03-10
17. Gibson G and Muse SV (2002) A Primer of Genome Science, Sinauer, Associates, Sunderland, MA, USA.
18. Gonzalez S, Binato R, Guida L, Mencalha AL, Abdelhay E4 (2014) Conserved transcription factor binding sites suggest an activator basal promoter and a distal inhibitor in the galanin gene promoter in mouse ES cells. Gene 538: 228-234.
19. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012 Jan;40(Database issue):D1178-86.
20. Hagen JB (2000). The origins of bioinformatics. *Nat. Rev. Genet.* 1: 231-236
21. Henikoff, S., Henikoff, J.G., (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U S A 89, 10915-10919.
22. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature (*2001) 409:860­921.
23. Junqueira DM, Braun RL and Verli H (2014). Alinhamentos. In: Bioinformática da biologia à flexibilidade molecular (Verli H, ed.). SBBq, São Paulo, 38-61.
24. Kanehisa M (2002) The KEGG database. Silico Simulation of Biological Processes 247: 91-103.
25. Kerr MK, Martin M and Churchill G (2000) Analysis of variance for gene expression in microarray data. J Computer Biology 7: 819-837.
26. Luscombe NM, Greenbaum D, Gerstein M, (2001) What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med.; 40(4):346-58.
27. Mehmood MA, Sehar U, Ahmad N (2014) Use of Bioinformatics Tools in Different Spheres of Life Sciences. J Data Mining Genomics Proteomics 5: 158.
28. Molecular Modelling [Internet]. 2016://en.wikipedia.org/wiki/Molecular modelling. Accessed: 2016-03-10
29. Moody G. Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business. London: Wiley, 2004.
30. Mount WD.)2004) Bioinformatics: sequence and genome analysis. 2nd ed. New York: Cold Spring Harbor Laboratory Press;. 692 p.
31. Needleman Saul B and Wunsch Christian D. (1970)  “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. Journal of Molecular Biology  48 (3): 443–53.
32. Ng PC; Henikoff S (May 2001). ["Predicting deleterious amino acid substitutions"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC311071). *Genome Res*. 11 (5): 863–74
33. Ortet P, Bastien O. Where Does the Alignment Score Distribution Shape Come from? *Evolutionary Bioinformatics*. January 2010.
34. Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of National Academy* *of Sciences U S A* **85**(8): 2444-2448. DOI:10.1073/pnas.85.8.2444
35. Peitsch MC (1996) ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. Biochem Soc Trans 24: 274-279.
36. Polyanovsky, V. O.; Roytberg, M. A.; Tumanyan, V. G. (2011). ["Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3223492). *Algorithms for Molecular Biology*. 6(1): 25.
37. Prosdocimi F, Cerqueira GC, Binneck E and Silva AF (2002). Bioinformática: Manual do usuário. *Biotec. Cienc. Des.* 12-25.
38. Protein Folding [Internet]. 2016. <https://en.wikipedia.org/wiki/Protein_folding>. Accessed: 2016-03-10
39. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, et al. (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33: W116-120.
40. Rao, V. S., Das, S.K.., Rao, V.J and Srinubabu, G. (2008). Recent developments in life sciences research: Role of Bioinformatics. African Journal of Biotechnology7 (5): 495-503.
41. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, et al. (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41: D344-347.
42. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids*;36: D1009–D1014.
43. The European Molecular Biology Laboratory [Internet]. 2016. <http://www.embl.org>. Accessed: 2016-03-10
44. The National Center of Biotechnology Information (NCBI) [Internet]. (2016). http://www.ncbi.nlm.nih.gov. Accessed: 2016-03-10
45. UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36: D190-195.
46. Venter JC (2001) The sequence of human genome. Science 291: 1304 – 1351.
47. Venter JC, Adams MD,Myers EW, Li PW, Mural RJ, Sutton GG. (2001) The sequence of the human genome. *Science*; 291:1304­51.
48. Verli H (2014). O que é Bioinformática? In: Bioinformática da biologia à flexibilidade molecular (Verli H ed.). SBBq, São Paulo, 1-12.
49. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J,. (2003) The Protein Information Resource. Nucleic Acids Res 31: 345-347.
50. Xiong, J. (2009). Essential bioinformatics: introduction and biological databases. Cambridge University press, USA. http://www.cambidge.org /catalogue.