**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. Since its inception in the 1980s, the field of bioinformatics has expanded quickly in tandem with the expansion of genome sequence data. In order to develop methods for the storage, retrieval, and analysis of biological data, multiple diverse fields of study—including biology, mathematics, computer science, and statistics—were combined to form the interdisciplinary field of study known as bioinformatics [30]. The amount of molecular data collection from multiple levels of organisation of an organism or environmental sample has significantly increased as a result of the quick adoption of omics techniques, their expanding power, and more affordable costs. With the recent development of Next-Generation Sequencing (NGS) technologies, the sequencing of nucleic acids underwent yet another revolution, ushering in a new era for omics techniques.

 Paulien Hogeweg is credited with creating the term "bioinformatics" (1979). Since the establishment of SWISS-MODEL server and the introduction of user-friendly interactive automated modelling about 18 years ago [35], this field has expanded tremendously. Since then, using databases and informatics at the backend to process biological data at a much faster rate has been crucial to the biological sciences. Large-scale biological data management, analysis, and manipulation fall under the purview of the field, which includes all computational tools and techniques [26]. To store, organize, and index the massive amount of sequence data generated by various sequencing techniques, electronic databases are required. Additionally, particular tools are needed by the databases so that researchers may access, evaluate, and add new or updated sequencing 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. David J. Lipman, a former director of the National Center for Biotechnology Information (NCBI), referred to Dayhoff as "the mother and father of bioinformatics" because of her significance to this field [30]. This is because of her contribution to the creation of algorithms that can recognize and display structures for use in X-ray crystallography, as well as computational approaches for protein sequence comparison that allow us to deduce the evolutionary relationships between kingdoms [20,48]. Dayhoff also created the one-letter amino acid code that is still in use today, as well as three-letter abbreviations (e.g., Lys for lysine, Ser for serine) (Table 1). For the first time this one-letter code was used in the Dayhoff and Eck's ‘Atlas of Protein Sequence and Structure’ (1965) [9], the first 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.

 George Bell and associates began compiling DNA sequences into GenBank in 1974 in order to provide immunology research with a theoretical foundation. Between 1982 and 1992, Walter Goad's team [30] produced the first version of GenBank. As a result of their efforts, the most widely used DNA sequence databases GenBank [16], "The European Molecular Biology Laboratory (EMBL) [43], and DNA DataBank of Japan (DDBJ) [8] were made in 1979, 1980, and 1984, respectively. The most significant development in DNA sequence databases was the introduction of web-based search engines, which allow researchers to locate and evaluate the target DNA sequences. These pioneering developments were made by David Benson, David Lipman, and associates, who also developed the "GENEINFO" and "Entrez" pieces of software [30]. Researchers may rapidly search database-indexed sequences and compare them to the sequence they were querying by utilizing this application. Software is now easily accessible through web-based interface of the National Center for Biotechnology Information (NCBI) database's web-based interface [44]. Comparison, analysis, and visualization of molecular sequences have become more sophisticated, and a variety of techniques have helped to advance bioinformatics in this field.

1. **Branches of Bioinformatics**

 Gene products have come into emphasis since the first draught of the human genome was finished [22, 47] instead of genes themselves. Genetic information is given a functional relevance in functional genomics. When analysing and interpreting biological data, information is taken into account at different levels, including the genome, proteome, and transcriptome. Transcriptomics is the study of the messenger RNA transcripts generated by a cell, whereas proteomics is the study of the total number of proteins (proteome) expressed 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. 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. Understanding and extracting information from the data generated by these instruments requires the use of bioinformatics techniques. The functional study entails metabolic pathway reconstruction and simulation, protein-protein interaction prediction, protein sub-cellular localization prediction, and gene expression profiling [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 that serve as repositories for raw sequence data and may be freely accessed via the Internet contain information on the sequences or structure alone (WWW). For example, the DNA Databank of Japan (DDBJ) for genome sequence, the Nucleotide Sequence Database operated by the European Molecular Biology Laboratory (EMBL), and GenBank, maintained by the National Center for Biotechnology Information (NCBI).

It gathers distinctive data from the lab and makes these data available to the 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)** consists of information that was obtained via the study of data found in primary databases, such as conserved sequences, a protein family's active sites, or conserved secondary motifs [13,18]. The primary database is subjected to computational algorithms, and the secondary database contains useful and instructive data. Some databases, including SCOP made at Cambridge University, CATH created at University College London, eMOTIF created at Stanford, etc., are built and hosted by individual researchers in their own labs.
2. **Composite database(s)** contains a range of primary database sources, eliminating the need to look up information at several places. Nucleotide and protein databases are provided by theNational Center for Biotechnology Information (NCBI), on its vast, highly accessible network of computer servers.

**Table 2: LIST OF IMPORTANT DATABASES**

|  |  |  |
| --- | --- | --- |
| **DATABASE** | **DESCRIPTION** | **REFERENCES** |
| **Nucleotide Databases** |  |
| DNA Databank of Japan (DDBJ) | It is one of the largest databases for nucleotide sequences and a member of the International Nucleotide Sequence Databases (INSD). | [8] |
| European MolecularBiology Laboratory(EMBL) | Repository of DNA and RNA sequences that is complementary to GenBank and DDBJ | [43] |
| GenBank | It belongs to the international nucleotide sequence databases (INSD). NCBI's primary nucleotide sequencing database in the USA | [16] |
| **Protein Databases** |  |
| Uniprot | One of the largest collection protein sequences | [45] |
| Protein DataBank | A significant source of information on proteins, giving details on the complex assemblages of nucleic acids, proteins, and their empirically confirmed structures |  [5] |
| Prosite  | Gives details on the active sites, conserved domains, and protein families. | [41] |
| SWISS-PROT | A section of the UniProt knowledge repository where protein sequences have been manually annotated. | [6] |
| InterPro | Resource that includes information on protein families, conserved domains, and active sites. | [39] |
| SCOP | Relationship between Familial and Structural Proteins | [15] |
| **Genome Database** |  |
| Ensemble Plants | A comprehensive database that provides genome-scale data for an increasing number of sequenced plant species. | [14] |
| Protein Information Resource ( PIR) | Non-redundant, comprehensive, and annotated protein sequence database | [49] |
| Phytozome  | A comparative hub for comparison and research of plant genome and gene family information. | [19] |
| **Miscellaneous databases** |  |
| TAIR | For the model plant Arabidopsis thaliana, The Arabidopsis Information Resource (TAIR) keeps a collection of genetic and molecular information. | [42] |
| KEGG | The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledge repository that connects genomic data with higher order functional data to analyse genes' activities systematically. | [24] |

1. **Gene identification and sequence analyses**

 Understanding the various characteristics of a biomolecule, such as a protein or nucleic acid, that give it its particular function is referred to as a sequence analysis. The sequences of the associated compounds are first acquired from open databases. After refinement, if needed, they are subjected to various tools that enable prediction of their features related to accuracy. The analysis can be used to recognise a transit peptide, introns, exons, or an open reading frame (ORF), as well as to identify specific variable regions that can be used as signatures for diagnostic purposes. It can also be used to identify promoter, terminator, or untranslated regions involved in the expression regulations. There are many tools designed for this purpose, some of which are crucial (enlisted in Table 3 with function).

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

|  |  |
| --- | --- |
| **Tools** | **Description** |
| BLAST (Local Sequence Alignment) | It is a search tool that is used to identify DNA or protein sequences. |
| Clustal Omega (Global Sequence Alignment) | This program can be used to conduct multiple sequence alignments. |
| HMMER  | With the use of this tool, homologous protein sequences can be searched in the relevant databases |
| Open Reading Frame (ORF) Finder | This tool can be used to find ORF for putative genes. |
| ProtParam | used to forecast proteins' physicochemical qualities |
| PPP | Prokaryotic Promoter Prediction tool used to determine the promoter sequences that are present upstream of the gene |
| JIGSAW | To find genes, and to predict the splicing sites in the selected sequences  |
| Genscan | makes use of genomic sequences to determine the intron-exon locations |
| Softberry Tools | Along with the structure and function prediction of RNA and proteins, some technologies are specialized in the annotation of animal, plant, and bacterial genomes. |

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 identification of point mutations, the creation of evolutionary trees, the categorization of genes and proteins, biological function prediction, the classification of genes and proteins, and other activities all rely on sequence similarity. If two sequences in an alignment share an ancestor, mismatches and gaps can be interpreted as point mutations and insertion or deletion mutations, respectively, introduced in one or both lineages since they diverged from one another. Indicators of how conserved a given section or sequence motif is throughout lineages can be made using the degree of similarity between amino acids at a specific place in a protein's sequence.

 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**

The ideal sequences for global alignment are those that are closely related and have the same length. Here, the alignment is carried out from the start of the sequence to the end in order to find the optimal alignment. The Needleman-Wunsch algorithm, often known as dynamic programming for sequence alignment, was developed by Saul B. Needleman and Christian D. Wunsch in 1970. (31). By breaking the original problem down into smaller, independent sub-problems, dynamic programming is able to solve it. The global sequence alignment is described in the algorithm for aligning nucleotide or protein sequences.

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**

 Multiple sequence alignment, or MSA, is a technique used to align three or more biological sequences (protein or nucleic acid) of comparable length. The results allow us to study the evolutionary relationships between 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**

The application of dynamic programming results in the best alignment of two sequences. It discovers the alignment in a more quantitative way rather than just applying dots by awarding specific values for fits and mismatches (Scoring matrices). The greatest scores in the matrix can be used to precisely locate alignment.

1. **Substitution Matrices**

 Sequence alignment using pairwise and multiple sequence alignment techniques is scored using substitution matrices. 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:** Margaret Dayhoff invented the PAM matrix, also known as the Point Accepted Mutations matrix. PAM matrices are computed using differences in proteins that are closely related to one another. For every 100 amino acid residues, one PAM unit (PAM1) indicates one allowable point mutation, implying that only 1% of the original structure is changed.
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.

The algorithm that was utilised can be categorised as either optimum or heuristic [36]. The best alignment is the ideal outcome, but the heuristic, while not delivering an optimal result, displays the best alignment within a specific time window of analysis.

1. **BLAST**

The Basic Local Alignment Search Tool (BLAST) is the most popular similarity search tool. Regions of similarity between sequences are discovered via BLAST. The software compares protein or nucleotide sequences and determines the statistical importance of matches. In addition to assisting in the identification of gene family members, BLAST can be used to infer functional and evolutionary links between sequences.

 The Smith-Waterman method (1981) was the source of the specialised local alignment algorithm known as BLAST, which displays a maximum alignment score of two sequences [1]. To search the sequences in the database, BLAST uses a heuristic based on the *k*-tuple approach in addition to the dynamic programming resulting from the aforementioned algorithm [23]. The *k*-tuple method restricts the search to words that are more significant, which are words with a length of 3 or 11 characters for amino acids and nucleotides, respectively. BLAST is a group of software tools that can be used for a variety of tasks depending on the sequence type of interest and the database being searched [37]. Table 4 lists a few applications that are accessible by BLAST.

**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 [1]

 The value of the score (Score bits) and the E-value are the two parameters through which the BLAST results are displayed. When the alignment score and database size are taken into account, the E-value is a statistical value that shows the likelihood that the alignment did not happen at random [37, 1]. However, the algorithm assigns the score based on the similarities and differences between the input sequences and the database [1].

1. **Phylogenetic analyses**

Due of the shared requirement of determining sequence similarity, phylogenetics and sequence alignment are closely related fields [[33]](https://en.wikipedia.org/wiki/Sequence_alignment#cite_note-ortet-25). Sequence alignments are frequently used in the study of phylogenetics to build and understand phylogenetic trees, which are used to categorise the evolutionary relationships between homologous genes found in the genomes of different species. Sequences' evolutionary separation from one another is qualitatively connected to how much they differ from one another. The simplest approaches use distance matrices, like neighbour joining (NJ) or unweighted pair group method with arithmetic mean (UPGMA).

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

|  |  |
| --- | --- |
| Tools | Description  |
| MEGA (Molecular Evolutionary Genetics Analysis) | constructs phylogenetic trees to investigate evolutionary proximity |
| PAML | A collection of applications for  phylogenetic analyses of DNA or protein sequences using maximum likelihood. |
| PHYLIP | A package for phylogenetic studies |
| TreeView  | Software that allows to switch between different views of the phylogenetic trees. |
| itol (Interactive Tree of Life) | An online tool for managing, displaying, and annotating phylogenetic trees. |

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 protein
* Protein-DNA interactions
* Transcriptional factors identification
* Interaction between Protein and Ligand
 |

1. **Conclusion**

 The emerging discipline “Bioinformatics” is the solution for the current demand in the every field of plant research. Sequence analysis, data mining, gene discovery, the development of phylogenetic trees, the prediction of protein structure and function, interaction networks, and many more computational approaches are used in this field. The field of bioinformatics will play a significant role in plant research. If plant science could be summed up in one word, it would be "integration," barring any unforeseen circumstances. The link that will enable all of these forms of integration will come from bioinformatics. Deep learning, a very potent technique, enables effective information storage and administration as well as the extraction of meaningful information from these data, which is made possible by the development of the big data era in plant science. The outcomes of genetic research will revolutionise the medical industry. As a result of genomic status, diagnostic methods will change quickly and may now concentrate on the association between genotypes and complicated phenotypes with the use of bioinformatics. This will advance the discipline of bioinformatics since the future generation of biologists will need to be equally at ease using a computer workstation as they are using laboratory benches.

**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.