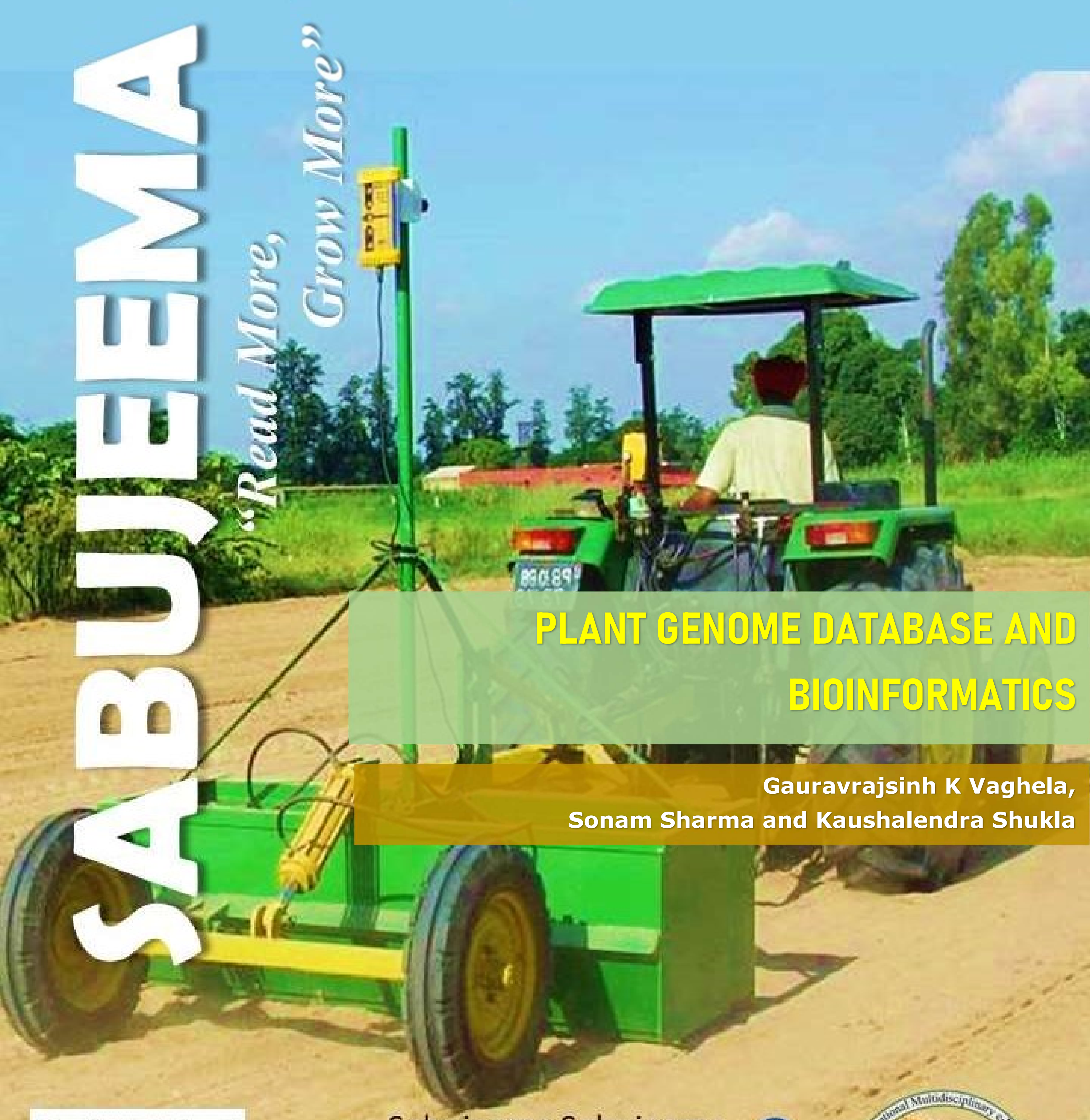
An International Multidisciplinary e-Magazine
www.sabujeema.com
Volume 2 | Issue 6 | JUNE, 2022

SABUJEEMA
"Read More, Grow More"



PLANT GENOME DATABASE AND BIOINFORMATICS

Gauravrajsinh K Vaghela,
Sonam Sharma and Kaushalendra Shukla

Sabujeema Sabujeema
editorsabujeema@gmail.com
sabujeema-international
multidisciplinary-e-magazine

Volume 2 - Issue 6 – June, 2022
SABUJEEMA
An International Multidisciplinary e-Magazine

PLANT GENOME DATABASE AND BIOINFORMATICS

[Article ID: SIMM0162]

Gauravrajsinh K Vaghela
Ph.D. Scholar,
Department of Genetics and Plant Breeding, C. P. College of Agriculture, Sardarkrushinagar Dantiwada Agricultural University, Sardarkrushinagar

Sonam Sharma
Assistant Professor, Sage University, Indore

Kaushalendra Shukla
Assistant Professor, Sage University, Indore

PLANT GENOME DATABASE

- Plant genome database (http://www.plantgdb.org/) is a database of molecular sequence data for all plant species with significant sequencing efforts. The database organizes EST sequences into contigs that represent tentative unique genes.

- The goal of the Plant genome database website is to establish the basis for identifying sets of genes common to all plants or specific to particular species by integrating a number of bioinformatics tools that facilitate gene prediction and cross species comparisons.

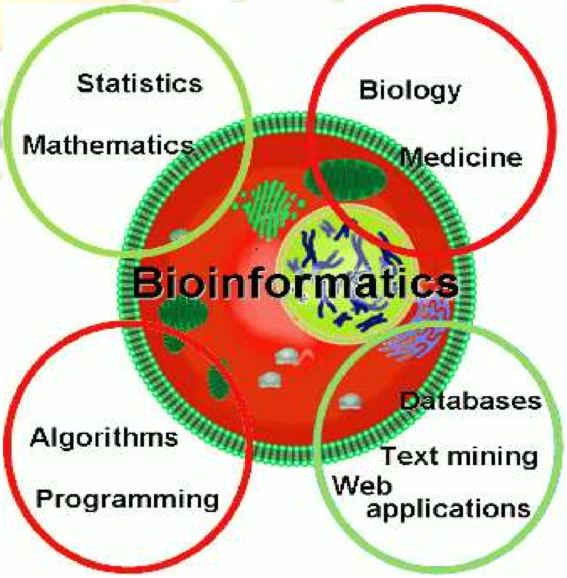
- For species with large-scale genome sequencing efforts, Plant genome database provides genome browsing capabilities that integrate all available EST and cDNA evidence for current gene models.

- There are several prominent public databases that provide access to plant genome data. These include general repositories, such as GenBank, and species-specific or datatype-specific resources, such as TAIR, MaizeGDB, Gramene and Stanford Microarray Database of course, larger database providers such as NCBI and TIGR function as both repositories and specific resources.

BIOINFORMATICS

- Bioinformatics is the science concerned with the development and application of computer hardware and software to the acquisition, storage, analysis and visualization of biological information.

- The term 'bioinformatics' is a combination of 'biology' and 'informatics'. Bioinformatics developed in the wake of generation of amino acid sequences of proteins and nucleotide sequences of DNA. In 1962, Zuckerkandl and Pauling proposed that amino acid sequences of proteins may be used to study evolutionary relationship among organisms.



THREE CHIEF COMPONENTS OF BIOINFORMATICS

1. The development of new algorithms and statistics for assessing the relationships among large sets of biological data, e.g., DNA sequence data.

2. Application of these tools for the analysis and interpretation of the various biological data, including nucleotide sequences, amino acid sequences, etc.

3. The development of database for an efficient storage, access and management of the large body of various biological information's.

HISTORY OF BIOINFORMATICS

- First comprehensive collection of amino acid sequences was compiled in the "Atlas of Protein Sequence and Structure" by The National Biomedical Research Foundation. This collection was edited by Margaret O. Dayhoff from 1965 to 1978.

- The European Molecular Biology Laboratory (EMBL) established their data library in 1980 to collect, organize and distribute nucleotide sequence data and related information. This function is now performed by the European Bioinformatics Institute (EBI), Hinxton, U.K.

- During early 1980s, the National Centre for Bioinformatics Information (NCBI) was established in U.S.A. NCBI serves as primary information databank and provider of information.

- Sometime later, DNA Data Bank was established by Japan.

- The National Biomedical Research Foundation established the Protein Information Resource (PIR) in 1984.

TYPES OF SEQUENCES IN NUCLEOTIDE SEQUENCE DATABASES

1. **cDNA Sequences:** - A cDNA molecule is obtained by reverse transcription of an RNA molecule. The cDNA sequences represent part of the genome that is transcribed into RNA. If the cDNA is obtained from mRNA, it will represent only the exon sequences of the genes expressed in the concerned cell/tissue/organism.

2. **Genomic DNA Sequences:** - These sequences represent the complete genome of the organism, irrespective of whether it is expressed or not. When the genome sequence is complete, it will contain the sequence of the entire genome of the organism. In case of prokaryotes, genome consists of, usually, a single chromosome, while in case of eukaryotes it relates to the nuclear DNA.

3. **Expressed Sequence Tag (EST) Sequences:** - These sequences are obtained by sequencing only a part of the cDNA molecules produced using mRNA. These sequences are dubbed as 'tags' because they can be used as probes for the isolation of the concerned genes from the genomic DNA.

- This approach was used by J. Craig Venter and his group for obtaining the sequence of expressed portion of human genome. The EST technique generated enormous sequence data that Permitted the construction of a preliminary transcript map of the human genome.

4. **Genome Sequence Tag (GST) Sequences:** - GSTs were developed for identifying the of Plasmodium falciparum. It was observed that the enzyme mung bean nuclease (Mnase) cuts the P. falciparum genomic DNA between genes. GSTs are developed by sequencing the DNA fragments

on either side of the points of cuts generated by Mnase.

5. **Organellar DNA Sequences:** - Organellar DNA is the DNA found in mitochondria 4, A and chloroplasts (cpDNA). The sequence of the data is compiled in databases.

6. **Sequences of other Molecules:** - In addition to the DNA sequence databases, sequences of such molecules as tRNA, small RNAs, etc.

DATABASES AND SEARCH TOOLS

- There are three public domain bioinformatics facilities:

1. National Centre for Biotechnology Information (NCBI), located at National Library of Medicine.
2. European Bioinformatics Institute (EBI), U.K.,
3. GenomeNet (Japanese Bioinformatics Service), Japan.

☐ Database: - A database is a vast collection of data pertaining to a specific topic e.g., Nucleotide Sequence, Protein Sequence etc.

☐ Databases are the heart of bioinformatics.

DATABASES AND ANALYSIS TOOLS

- The utilization of various databases requires the use of suitable search engines and analysis tools. These tools are often called database mining tools and the process of database utilization is known as database mining.

1. **BLAST (Basic Local Alignment Search Tool):** - BLAST is a family of user-friendly sequence similarity search tools on the web.

- The BLAST server is supported through NCBI, USA. This tool is designed to identify potential homologues for a given sequence.
- It can analyze both DNA and protein sequences. Identification of homologues allows the prediction of potential functions and in modelling of the 3-D structure.
- The new BLAST programmes are called BLAST 2.0.

TYPES OF BLAST PROGRAMMES

I. **BLASTp:** - It compares the submitted protein sequence against a protein database.

II. **BLASTx:** - This programme translates the submitted nucleotide sequence into amino acid sequence and compares the latter with a protein database.

III. **BLASTn:** - This is used to compare a nucleotide sequence with a nucleotide sequence database.

IV. **tBLASTn:** - The programme converts the submitted protein sequence into nucleotide sequence and compares it with a nucleotide sequence database.

V. **tBLASTx:** - This programme translates the submitted nucleotide sequence as well as the nucleotide sequence database into amino acid sequences and searches for homology between the two.

2. **ENTREZ:** - ENTREZ is one of the most popular search engines at NCBI, USA. ENTREZ searches bibliographic citations and biological data from a variety of reliable databases, viz., SWISS-PROT, PDB, GeneBank, EMBL, etc. For bibliographic or citation search, it uses PubMed's bibliographic database.

3. **TAXONOMY BROWSER**: - This search tool provides taxonomic information on various species. The TAXONOMY database of NCBI has information (including scientific and common names) about all organisms for which some sequence

information is known. The server provides genetic information and the taxonomic relationships of the species in question. TAXONOMY has links with other servers of NCBI, e.g., Structure and PubMed.

4. **LOCUS LINK**: - Locus Link contains information about genes, including their official names. In addition, it allows one to search for genes homologous to a given gene, and to obtain information about these genes. For example, one can easily obtain information about mouse genes that are homologous to given human genes. One can also search for homologues of a given gene in several other organisms.

5. **PROSITE:** - PROSITE has a collection of functional sites and sequence patterns found in many proteins. Entries in PROSITE are generally linked to SWISS-PROT and other relevant databases. The PROSITE file includes the sequence entries that share the matched sequence motif of interest. The characterized motifs are well documented to minimize redundancy.

Sr No	Software/web server	Functions
1	AIGO	A unified framework for the analysis and the inter-comparison of GO functional annotations
2	Blast	Most important tool to find out the functions through sequence comparisons
3	Blast2GO	A computational tool for functional annotation
4	CoreGenes	A computational tool for identifying and cataloging "core" genes in a set of small genomes
5	ESTree db	A computational tool for peach functional genomics
6	GeneViTo	Visualizing gene-product functional and structural features in genomic datasets
7	GPCG	The Graphical Pipeline for Computational Genomics to perform the computational steps required to analyze NGS data
8	PFGRC	The pathogen functional genomics resource Center provides scientists with centralized resources necessary to conduct functional genomics studies on a variety of pathogens for which genomic-sequence information
9	RSAT	Nucleotide sequence-based tool to predict regulatory elements of bacterial genes

Table 1: - Software and web server for functional genomics

INDIAN DATABASES

1. **GM Crops Database:** - This database, developed at NRC on Plant Biotechnology, New Delhi. It is an interactive web resource storing information on biosafety of transgenics (released in India), including over 800 publications on the subject. For example, a total of 139 transgenic lines using 4 genes (crylAb, crylAc, cry2Ab and vip3A) and a single promoter (CaMV 35S) have been developed.

2. **Vanshanudhan:** - This database has been developed by NRC on Plant Biotechnology, New Delhi. Scientists as an outcome of Indian rice genome initiative. It contains information on the 56,298 rice genes.

REFERENCES

- Singh, B. D. Biotechnology Expanding Horizons. 4th Edition, Kalyani Publishers, New Delhi. p. 733-743.
- Katara, Pramod (2014). Potential of Bioinformatics as functional genomics tool: an overview. Springer. 3(52): 1-7.
- Khan, Nida. (2018). Structural and Functional Bioinformatics. Letter in Health and Biological Science. 3(1): 7-11.
- Dong, Q.; Schlueter, S. D. and Brendel, V. (2004). Nucleic Acids Research. 32: 354-359.