NEW ERA AGRICULTURE MAGAZINE

Bioinformatics Tools and Techniques for Plant Biotechnology Research

Pradeep C[1], Mouli Paul[2], Mukesh Kumar Singh[3], Jatoth Uday Bhargav[4] and Gudipati Naveen[5]

Abstract:

In recent decades, 'bioinformatics' has emerged as a critical term across biological research disciplines. The rapid expansion of molecular biology and the accompanying surge in biological information necessitated a more sophisticated, computerized approach to collecting, storing, managing, and analyzing the extensive data generated by scientific experiments. Bioinformatics represents an innovative interdisciplinary field that has developed numerous tools and techniques for efficiently organizing biological data into comprehensive databases. It is fundamentally a computer-based scientific discipline that integrates mathematics, biology, and computer science to enable sophisticated analyses and interpretation of genomic and proteomic data. The core components of bioinformatics primarily encompass two key areas: database collection and analysis, and the development of sophisticated software tools and algorithms designed to interpret biological information. Its significance extends across multiple biological domains, providing researchers with diverse data types including nucleotide and amino acid sequences, protein structures, and expression patterns from various organisms. Particularly in plant biotechnology, bioinformatics has proven instrumental by offering complete genomic information about different plant species. This allows for more efficient exploration of plants as valuable biological resources for human use. The field continues to evolve, providing researchers with powerful computational methods to understand and manipulate biological systems. This overview aims to explore the fundamental concepts, essential tools, and practical applications of bioinformatics within plant biotechnologies, while also acknowledging the current challenges and potential areas for future development in this dynamic scientific domain.

Pradeep C[1], Mouli Paul[2], Mukesh Kumar Singh[3], Jatoth Uday Bhargav[4] and Gudipati Naveen[5]

[1]Ph.D Research Scholar, Department of Molecular biology and Biotechnology, ICAR- Indian Agricultural Research Institute, New Delhi

[2]M.Sc Scholar, Department of Genetics and Plant Breeding, Institute Ramakrishna Mission Vivekananda Educational and Research Institute, Kolkata

[3]Assistant Professor-cum-Junior Scientist, Department of Genetics & Plant Breeding, Tilka Manjhi Agriculture College, Godda (Birsa Agricultural University, Ranchi) Jharkhand

[4]Assistant professor, Department of Agriculture, Loyola academy, Osmania University, Hyderabad, Telangana

[5]Ph.D Research Scholar, Department of Entomology, Assam Agricultural University, Jorhat, Assam

Introduction:

Bioinformatics and computational biology are revolutionizing plant biology by enabling advanced scientific discoveries. Sequencing technologies have allowed scientists to uncover the genetic architecture of various plant and microorganism species, including their proteome, transcriptome, and metabolome. Sequence analysis is crucial for obtaining complete genome sequences, revealing an organism's genetic organization and potential functionality. A comprehensive genome sequence includes both coding and non-coding regions, encompassing exons, introns, regulators, and promoters, which collectively define an organism's unique traits. The advent of next-generation sequencing (NGS) and other omics technologies is progressively expanding our understanding of plant genomics. Bioinformatics has become essential in managing these large volumes of genetic data, providing systematic methods to capture, store, and organize genomic information efficiently. By enabling detailed genetic exploration, bioinformatics is transforming our ability to analyze and comprehend the complex genetic landscapes of plant and microbial species.

Bioinformatics databases and tools for plant biotechnology

Bioinformatics offers numerous databases and tools for plant biotechnology analysis. The advent of Next-generation sequencing (NGS) and bioinformatics has produced vast amounts of plant genomic data, which is stored in various publicly accessible online databases. These databases serve different purposes - some are specialized, like CottonGen, which exclusively focuses on cotton genomics and breeding data, while others are comprehensive, like the National Center for Biotechnology Information (NCBI), which housed approximately 21,000 plant genomes as of 2021. NCBI, widely recognized in the scientific community, collects and analyzes molecular biology, biochemistry, and genetic information. Users can access plant genome data through its Gene Expression Omnibus (GEO) or Sequence Read Archive (SRA) platforms by searching a plant's scientific name. For example, searching for Rosa chinensis yields datasets containing gene symbols, Ensemble IDs, and other genetic information. Researchers can then analyze this data using tools like Gene Ontology, DAVID, and BLAST. EnsemblPlants represents another significant database, focusing exclusively on plant genomes unlike NCBI's broader scope. Launched as part of the 1999 Ensembl project, it features automatic genome annotation and integration with other biological data. The platform continuously updates as new plant genomes are sequenced. EnsemblPlants offers more comprehensive data than NCBI,

including polymorphic loci, population structure, genotype information, and comparative genomics capabilities. This extensive functionality helps plant bioinformatics researchers save time, though they retain the ability to reanalyze data based on their specific requirements.

Biotechnology and bioinformatics for plant breeding

Plant breeding involves modifying and enhancing specific plant traits to develop improved crop varieties that benefit humanity. Genetically engineered plants offer several advantages, including better quality, enhanced nutrition, and increased yields. Advances in molecular biology and genomics have significantly accelerated plant breeding progress by applying genomic research findings to crop development. In contemporary agriculture, transgenic technology involves genetic modification through gene alteration or the introduction of foreign genes to improve plant characteristics and productivity. The emergence of next-generation sequencing (NGS) has generated extensive biological data requiring specialized database storage. These databases provide open access to complete genome sequences, enabling researchers to analyze gene sequences, potential functions, and genetic mapping positions across different genomes. Using specialized software, scientists can develop predictive models and incorporate desired traits into plants by identifying reliable genetic markers for breeding. Beyond genomic data, metabolite databases play a vital role in studying how proteomics and genomics interact to influence organism phenotypes and functions. Popular plant metabolomics databases include Metlin, which contains approximately 240,000 metabolites and 72,000 high-resolution MS/MS spectra, and PlantCyc, which documents biochemical pathways, catalytic enzymes, and plant genes. Additionally, NGS technology has enhanced the identification of single-nucleotide polymorphism markers through RNA sequencing (RNA-seq), which directly measures mRNA profiles.

Current challenges of bioinformatics applications in plant biotechnology

Despite the promising potential of bioinformatics in plant biotechnology, several challenges and limitations must be addressed to maximize its benefits. As plant genome data mining and database development continue to expand rapidly, bioinformaticians and scientists encounter various obstacles. These challenges can be categorized into several key areas, as outlined in the following subsections.

Bioinformatic data management and organization and synchronize update resources

Since its commercial debut in 2004, next-generation sequencing (NGS) has

generated massive amounts of plant genome data, with thousands of gigabytes of plant sequences being added to public databases each month. The continuous sequencing and resequencing of plant genomes has led to an exponential growth in new genome sequences across public databases. While technological advances have enabled more plant genome sequencing, this has created challenges in data storage and maintenance. Updates must be implemented across all comparative databases, not just individual genome repositories. This synchronized updating across different plant genomic platforms is essential to maintain a robust, current, and dependable database ecosystem that researchers can rely on.

Complexity of plant genetic content

The plant research community faces challenges not only from the vast amount of genomic data but also from the complexity of plant genetic material. Despite NGS technologies enabling quick DNA sequencing of non-model and orphan plant species, plant sequencing lags behind that of animals and microorganisms. This delay occurs primarily because plant genomes can be up to hundreds of times larger than animal and microbial genomes. Further complicating matters is polyploidy—genome duplication—which occurs in approximately 80% of plant species. The assembly of large plant genomes with numerous repetitive sequences has been compared to solving a puzzle of blue sky broken up by nearly identical white cloud patches representing small genes. This challenge stems from NGS producing shorter sequence lengths than traditional Sanger sequencing, requiring specialized assembly algorithms. As a result, plant genomes sequenced through NGS are primarily limited to creating gene catalogs, analyzing repeat content, studying evolutionary mechanisms, and conducting basic comparative genomics research.

Advance in sequencing technologies

Genome assembly employs two main approaches: comparative and de novo assembly. Comparative assembly is a reference-based method that uses existing genome or transcriptome data as a guide, while de novo assembly involves reconstructing genomes from previously unsequenced organisms. Though various assembly tools and NGS technologies are available for genome sequencing, these approaches aren't entirely separate due to limited bioinformatics tools capable of handling the unique complexities of plant genomes. Algorithm development represents one of the biggest challenges in bioinformatics software creation. Current bioinformatics programs are computationally intensive and typically rely on single assembly methods. To create more reliable final assemblies, there's a crucial need to develop

more efficient algorithms that can combine different assemblers using varied underlying approaches.

Database accessibility

Among the approximately 374,000 known plant species worldwide, Arabidopsis thaliana was the first to have its complete genome sequenced in 2000 using Sanger sequencing. While molecular biology has helped with species identification, obtaining complete plant genomic data remains difficult due to genome complexity. Despite NGS platforms advancing plant genome sequencing, database repositories contain relatively few sequenced datasets. Currently, PlantGDB genome browser provides access to only 29 plant genome databases, where researchers can find information about gene structure, GSS contigs, protein similarities, and spliced EST alignments. The PlaD database, developed by China Agricultural University, focuses on plant defense-related microarray data but covers only Arabidopsis, rice, maize, and wheat. Another resource, the Plant Omics Data Center, offers web-based access to omics data including co-expressed profiles, regulatory networks, and plant ontology information, but is limited to eleven specific plants including Arabidopsis, tobacco, earthmoss, and others. These public databases require regular updates with new and resequenced data to ensure researchers have access to the most current genome datasets for their studies.

Conclusion

The integration of bioinformatics into plant biotechnology marks a transformative approach to studying living organisms. This field plays a pivotal role in advancing agriculture by facilitating the study of stress resistance and plant pathogens, both essential for crop breeding advancements. Next-generation sequencing (NGS) and related technologies are expected to expand the availability of plant genome data in public databases, enabling the identification of genomic variants and predicting protein structures and functions. Additionally, genome-wide association studies (GWAS) have streamlined the identification of loci and allelic variations associated with valuable traits, thereby simplifying crop modification and improvement. In summary, the progress in bioinformatics applications in plant biotechnology has allowed researchers to gain a deeper and more systematic understanding of economically significant plants. However, challenges remain, such as achieving automated, low-cost full genome sequencing and assembly. Effective bioinformatics tools capable of generating longer reads with unbiased coverage are crucial to addressing the complexity of plant genomes. To this end,

advancements in algorithm development are needed to enhance data mining, analysis, and comparison capabilities. The role of bioinformaticians and experts with strong mathematical and programming skills is critical in introducing innovative methods and knowledge to the field. Such contributions will not only propel plant biotechnology and agriculture forward but also influence the future of humanity.

References

1. Babu P, Baranwal DK, Harikrishna PD, Bharti H, Joshi P et al (2020) Application of genomics tools in wheat breeding to attain durable rust resistance. Front Plant Sci 11:567147.
2. Bolser D, Staines DM, Pritchard E, Kersey P (2016) Ensembl plants: integrating tools for visualizing, mining and analyzing plant genomics data. Methods Mol Biol 1374:115–140.
3. Genome Research on Wheat Consortium. Genetics. 168(2):1087–1096.
4. Genomics Proteomics Bioinformatics 18(3):221–229.
5. Gill BS, Appels R, Borta-Oberholster AM, Buell CR, Bennetzen JL, Chalhoub B et al (2004) A workshop report on wheat genome sequencing: International
6. Guan J, Garcia DF, Zhou Y, Appels R, Li A, Mao L (2020) The battle to sequence the bread wheat genome: a tale of the three kingdoms.
7. Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, Fuks G et al (2005) Structure and architecture of the maize genome. Plant Physiol 139(4):1612–1624
8. Li C, Song W, Luo Y, Gao S, Zhang R, Shi Z et al (2019) The HuangZaoSi maize genome provides insights into genomic variation and improvement history of maize. Mol Plant 12(3):402–409.