BOHR Journal of Biocomputing and Nano Technology
2024, Vol. 2, No. 1, pp. 1–5
DOI: 10.54646/bjbnt.2024.12
www.bohrpub.com

REVIEW

Study on advancement in bioinformatics for crop improvement: integrating genomics, phenomics, and machine learning

Avil Alva and N. Hemalatha*

Department of IT, Aloysius Institute of Management and Information Technology (AIMIT), St Aloysius College, Mangaluru, India

***Correspondence:**
N. Hemalatha,
hemalatha@staloysius.ac.in

Received: 06 November 2023; **Accepted:** 03 January 2024; **Published:** 24 January 2024

Bioinformatics has emerged as a vital tool in agriculture, performing an important role in crop improvement and precision agriculture. This review study explores the applications of bioinformatics in these areas, highlighting the importance of next-generation sequencing (NGS) and third-generation sequencing technologies like PacBio and ONT. The integration of bioinformatics and genomics has revolutionized crop improvement strategies, facilitating the creation of crop varieties that are more suitable for evolving climates. The review emphasizes the creation of integrated crop databases for managing large-scale genotype and phenotype data and discusses the role of bioinformatics in precision agriculture for optimizing resource usage and minimizing environmental impact. Multi-omics technologies are highlighted for gaining insights into plant biology and facilitating informed decisions in crop improvement. Machine learning algorithms play a prominent role in interpreting diverse and complex datasets generated by imaging or sequencing, aiding in efficient plant phenotyping and the identification of associations between DNA sequences and traits. The review concludes by emphasizing the importance of collaboration, data integration, and the potential of bioinformatics in advancing agriculture.

Keywords: bioinformatics, agriculture, crop improvement, machine learning, next-generation sequencing, genomics, multi-omics, data integration, genetic variations

1. Introduction

The significant surge in population growth in the mid-20th century led to a sharp rise in the demand for agricultural produce. So, it was up to the governments of various countries and scientists that this demand is fulfilled by without using extra land or space. Bioinformatics is the application of computational and statistical techniques concerning the study of biological data. In agriculture, bioinformatics is instrumental in the development of new and improved crops and livestock through the analysis of genetic data. Area of crop improvement is among the key applications of bioinformatics in agriculture. By analyzing the genetic makeup of crops, scientists can identify desirable traits, such as resilience to pests and diseases, drought tolerance, and increased yield and it has been used recently by the application of new method i.e., Next generation sequencing (NGS) and this information then can be utilized for the creation of novel crop varieties that are better suited to specific growing conditions. Bioinformatics also participates significantly in the advancement of precision agriculture. By analyzing large amounts of data on crop and soil conditions, farmers can make knowledgeable choices regarding when and how to apply fertilizers, pesticides, and other inputs. This helps to optimize crop yields and mitigate the effects of agriculture on the environment. Additionally, this paper mentions the advancement of third-generation sequencing's role.

2. Literature review

This paper has the compilation of the reviews of different authors and scientists from the various institutions across the

globe on the subject. Further the review is given based on the approaches that can yield the great perks in the subject.

Jacqueline Bately and David Edwards have elucidated how genomics and bioinformatics can be employed to improve crops amidst climate change. They emphasized the significance of agricultural advancements for conditions like extreme weather events, temperature fluctuations and emerging pests and diseases. Highlighting genomics, it is explained how it helps the researchers in unraveling the complex genetic makeup of crops by analyzing the entire DNA sequence of plants to pinpoint the genes linked to the desirable traits like drought tolerance, disease resistance and improved yield. And hence scientists can accelerate the traditional breeding programs by selecting and crossbreeding plants with desired characteristics. Bioinformatics has played the significant role in harnessing the large amount of data by developing computational tools and algorithms for analyzing and interpreting genomic information allowing researchers to manage and analyze massive datasets, detect genetic variations and forecast gene functionalities. Integration of bioinformatics and genomics has been seen in revolutionizing the crop improvement approaches that empower researchers to develop novel crop varieties with improved resilience against the change in climate in faster and targeted manner. Identification of genetic markers associated with desired traits can be employed in marker-assisted selection to precisely breed plants with higher chance of inheriting those traits (1).

Hu et al. (2) have given the observations about the use of third-generation sequencing technologies like Pacific Biosciences(PacBio) and Oxford Nanopore Technologies(ONT)and the long reads generated by them improving the genome assemblies by spanning the repetitive complex regions and hence enabling the revelation of agronomically important genes and the development of genome-wide molecular markers (3–5). Combination of long-read sequencing with long-range mapping technologies and chromosome conformation capture has resulted in high contiguous chromosome-level crop genome assemblies. Discussion has been done on integrated crop databases and how they are developed to store large amount of genotype and phenotype data. Tools like KnetMiner is also been developed to facilitate data mining and uncover the connections between traits (features) and genes for crop improvement. Other approaches like mining of quantitative trait loci (QTL) studies and conducting genome-wide association studies (GWAS) to identify breeding targets. QTL (5) studies pinpoint genetic regions associated with quantitative traits, meanwhile, GWAS examine natural populations to discover genomic areas linked to variations in phenotypes. Forward and reverse genetic screenings are also used to identify and characterize genes associated with specific phenotypes. Additionally, genomic selection and targeting cis-regulatory elements offer promising avenues for crop breeding. Machine Learning (6) is applied to crop breeding for various purposes like high-throughput phenotyping, genomics research, and guide RNA (gRNA) design for genome editing (**Figure 1**). Plant phenotyping encompasses the evaluation of functional and structural characteristics at diverse levels, ranging from individual cells to complete organisms. It holds a vital significance in comprehending genomic data. Machine learning (ML) algorithms are employed in stress phenotyping and disease monitoring. A specific study focused on iron deficiency chlorosis in *Glysine max* (soybean) using linear discriminant analysis (LDA) and multiclass support vector machines as ML techniques. Additionally, an unsupervised learning mechanism was utilized to gauge the severity of foliar stresses caused by bacterial and fungal diseases in soyabean plants. ML algorithms can be used in genomics research for genome assembly, gene regulatory network inference, SNP identification, and SNP calling in error-prone long reads. Single Nucleotide Polymorphism is deletion or addition of nucleotide at unrequired places in whole genome. Clevenger et al.(7) have developed an ML – based analysis tool namely SNP-ML which utilizes neural networks and tree bagging models to filter the false positive SNPs with over 98% accuracy in data of peanut, strawberry and cotton data of SNP. In genome editing, bioinformatics tools are crucial for gRNA design to ensure efficient and specific CRISPR/Cas gene editing.

Esposito et al. (8) have shown insight of how bioinformatics intersects with agriculture in the Next Generation Sequencing (NGS) era. Advances in NGS enabling plant genomics research, enabling comprehensive genome sequencing, transcriptomics, and epigenomics studies allow identification of genetic variations, gene expression patterns, and epigenetic variations. These traits are essential for comprehending plant characteristics and how they react to environmental influences. By applying NGS into the crop breeding program they efficiently identified and utilized desirable genetic variations, enabling marker-assisted selection and accelerating the enhancement of improved crop varieties with increased yield, quality, and resistance to biotic and abiotic stresses. According to observations, agricultural Microbiology is significantly influenced by NGS (Next-Generation Sequencing) technology. Through scrutinizing the genetic composition of microbial communities associated with plants, soil, and livestock, valuable insights into their configuration and operations are obtained. This, in turn, enables the advancement of specific microbial approaches to promote sustainable agriculture, encompassing disease prevention and nutrient optimization. Integration of bioinformatics with precision agriculture where NGS data is coupled with remote sensing, climate data, and other environmental parameters. This synergy enables the optimization of agronomic practices, such as irrigation and fertilizer management, leading to increased resource-use efficiency and minimized environmental impacts (8).

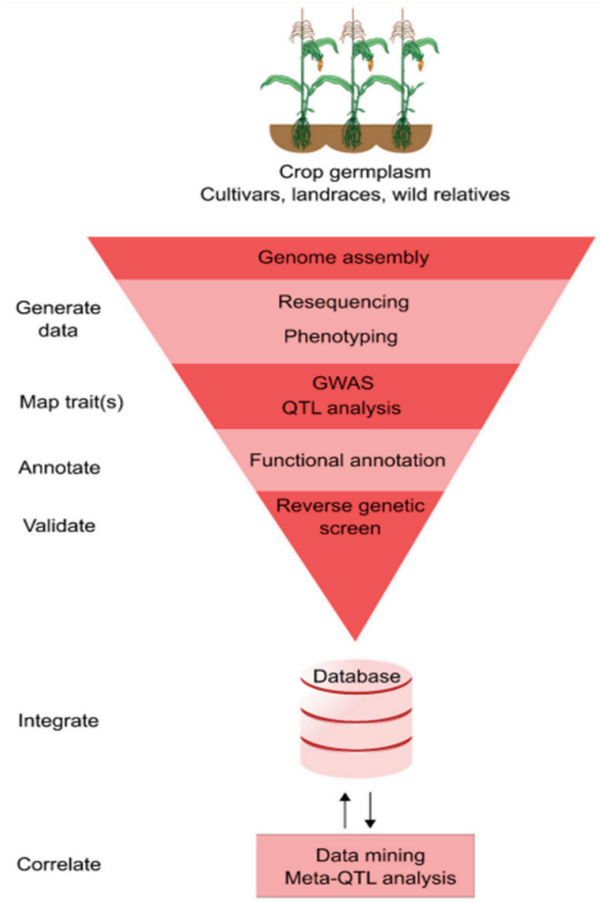

FIGURE 1 | Crop trait discovery pipeline (2).

Li and Yan (9) have directed attention toward the application of multi omics technologies such as genomics, transcriptomics, proteomics, and metabolomics in promoting sustainable agriculture. It is observed that omics technologies have immense potential unravel the complex facets of to plant biology. Through comprehensive analyses of genes, proteins, and metabolites involved in agronomic traits, a knowledge-driven approach enables breeders to make informed decisions in crop improvement, fostering higher yields, improved quality, and enhanced resilience, all while mitigating negative environmental impacts. Genomics lies at the core of omics technology. By leveraging state-of-the-art sequencing techniques, breeders can identify essential genetic variations, which includes single nucleotide polymorphisms (SNPs) and structural variations, that underpin important traits. These genetic variations serve as valuable markers for targeted selection and breeding. Proteomics and metabolomics complement genomics and transcriptomics by revealing the performance and biochemical aspects of plant physiology. By analyzing protein profiles and metabolite compositions, breeders gain a thorough perception of the underlying mechanisms governing plant growth, stress responses, and nutrient utilization. This knowledge facilitates the selection of superior crop genotypes capable of thriving under diverse environmental conditions (9).

Beiki et al. (10) have given picture of how integration of advanced technologies, such as high-throughput sequencing and bioinformatics, has significantly enhanced the speed and accuracy of data analysis. This has enabled researchers to unravel complex gene regulatory networks, epigenetic modifications, and interactions i.e., interplay of genes and environment. Such comprehensive understanding lays the basis for formulating inventive approaches to enhance crop and livestock production, improve disease resistance, and mitigate the outcomes of climate change (10).

Chen et al. (11) and colleagues have emphasized on how bioinformatics tools and algorithms have greatly contributed to the extraction of valuable information from vast datasets. Authors have shown their anticipation on advancement of sequencing technologies, leading to even more cost-effective and efficient genome sequencing (11).

Ma et al. (12). provides an overview of the application of bioinformatics in studying oxidative stress biomarkers in *Oryza sativa* (rice) using the "omics" approach. The authors focus on the advancements made in bioinformatics to identify and analyze biomarkers linked to oxidative stress, along with biotic stress (**Figures 2**, **3**) in rice plants. Oxidative stress occurs in plants when there is an inequilibrium between the production of reactive oxygen species (ROS) and the plant's antioxidant defense mechanisms. *Oryza sativa*, being a model plant species, has been extensively investigated to comprehend the mechanisms governing oxidative stress and its effects on the growth and development of the plant. The authors discuss how bioinformatics tools and algorithms are utilized in the examination of genomic data to identify genes involved in oxidative stress response. They highlight the importance of gene expression profiling using transcriptomics to detect genes exhibiting differential expression under oxidative stress conditions. The analysis of proteomics and metabolomics data employing bioinformatics tools help in identifying protein and metabolite biomarkers associated with oxidative stress. Furthermore, the authors emphasize the integration of multiple "omics" datasets through bioinformatics approaches. This integration enables a more holistic understanding of the complex regulatory networks involved in oxidative stress response in rice. They highlight the significance of network analysis and pathway enrichment analysis in uncovering key molecular pathways and biological processes affected by oxidative stress (12).

Li et al. (13) has provided an overview of how crops deal with the stress and the mechanism involve in stress responses. The integration of genomics and bioinformatics with transcriptomics has been instrumental in unraveling the elaborate gene regulatory networks implicated in stress responses. Authors discuss how transcriptomic data analysis using bioinformatics approaches, such as differential gene expression analysis and co-expression

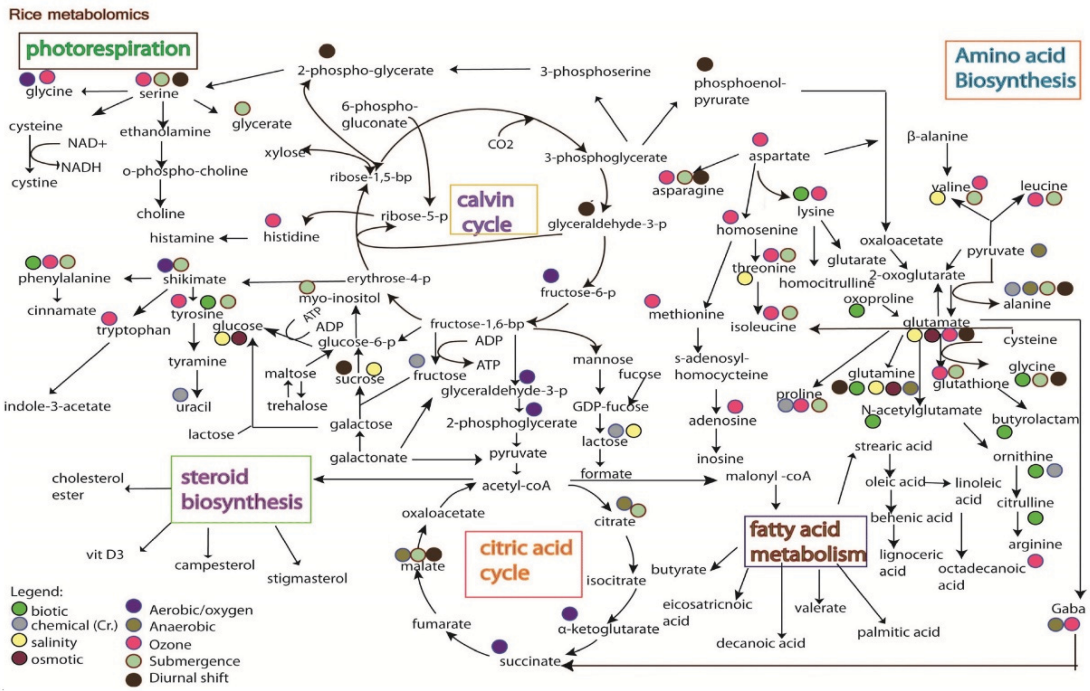

FIGURE 2 | Alterations in rice metabolites under different biotic stress, chemical, ozone, anaerobic, aerobic, submergence and metabolite changes under diurnal cycle (12).

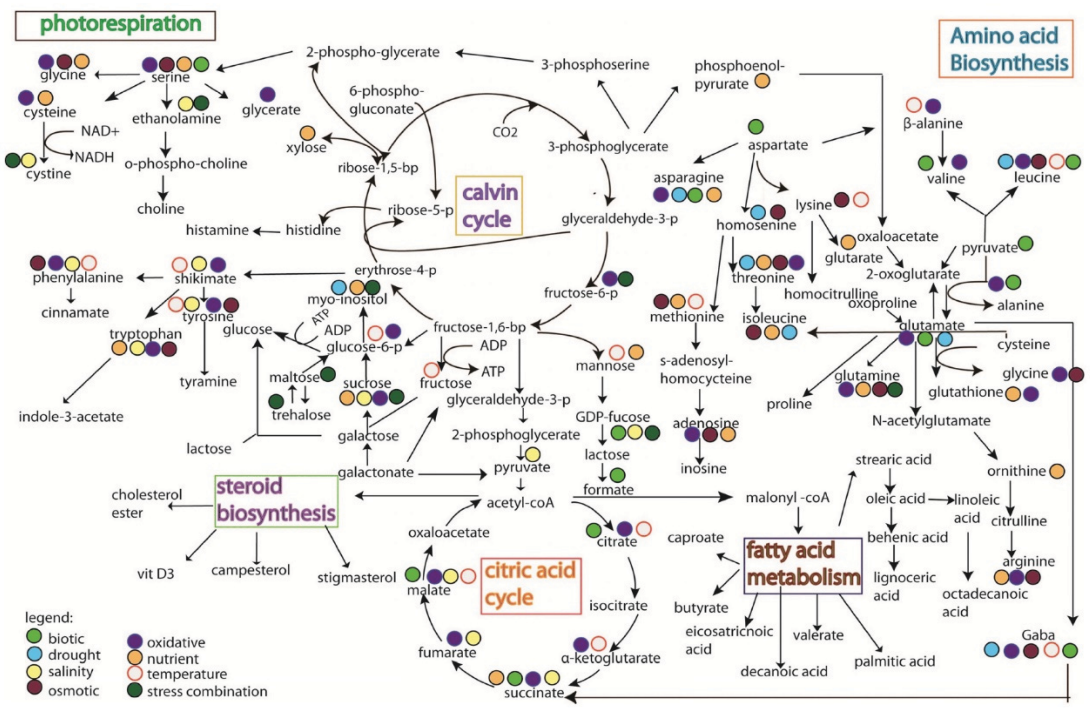

FIGURE 3 | Alteration in rice plant metabolism under different stress treatments like biotic stress, drought, salinity, oxidative stress, temperature, osmotic stress and combination of stresses (12).

network analysis which offers a deeper understanding of the critical genes and pathways linked to stress response and tolerance. Additionally, the authors emphasize the relevance of proteomics and metabolomics data analysis in crop stress research. Bioinformatics tools enable the identification and quantification of stress-responsive proteins and metabolites, shedding light on the biochemical processes underlying stress tolerance mechanisms. The authors also highlight the emerging field of epigenomics, that study the heritable changes in gene expression that are unrelated to

modifications in the DNA sequence. Bioinformatics plays a crucial role in analyzing epigenomic data, enabling the exploration of epigenetic mechanisms involved in stress response and adaptation (13).

Öborn et al. (14) emphasize the contribution of bioinformatics in data integration and predictive modeling. By incorporating diverse data sources, including climatic data, socio-economic data, and ecological data, bioinformatics enables the development of comprehensive models to simulate future agricultural scenarios. These models can help in understanding the potential impacts of climate change, population growth, and other factors on agricultural productivity and sustainability (14).

3. Conclusion

In conclusion, it can be inferred that Multi-omics and Bioinformatics plays the major role in studying, predicting the genome sequence and assembly in crop genomes, improving the yield and environmental stress bearing mechanism in the crops. NGS plays as of gold standard in sequencing. But as of now NGS is preferred more than third generation sequencing because of its cost effectiveness. ML algorithms are not that easy to apply frequently because of non-availability of data and the results are less accurate, if test is performed in less amount of data. Data is also cost effective making it only available to the bigger firms and institutions. Maintenance of huge biological databases which comprises of genotypic and phenotypic data is a must in present condition. One standard format must be maintained universally so that it becomes easy to interpret the various data across the databases. Integration of scientists and research personnel from various field like biology, genomics, statistics and computer science will improve and catalyze the work faster.

References

1. Batley J, Edwards D. The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Curr Opin Plant Biol.* (2016) 30:78–81.
2. Hu H, Scheben A, Edwards D. Advances in integrating genomics and bioinformatics in the plant breeding pipeline. *Agriculture.* (2018) 8:75.
3. Li C, Lin F, An D, Wang W, Huang R. Genome sequencing and assembly by long reads in plants. *Genes.* (2017) 9:6.
4. Vlk D, Řepková J. Application of next-generation sequencing in plant breeding. *Czech J Genet Plant Breed.* (2017) 53:89–96.
5. Chen K, Ji F, Yuan S, Hao W, Wang W, Hu Z, et al. The performance of activated sludge exposed to arsanilic acid and amprolium hydrochloride in sequencing batch reactors. *Int Biodeterioration Biodegrad.* (2017) 116:260–5.
6. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* (2015) 16:321–32.
7. Clevenger JP, Korani W, Ozias-Akins P, Jackson S. Haplotype-based genotyping in polyploids. *Front Plant Sci.* (2018) 9:564. doi: 10.3389/fpls.2018.00564
8. Esposito A, Colantuono C, Ruggieri V, Chiusano ML. Bioinformatics for agriculture in the next-generation sequencing era. *Chem Biol Technol Agric.* (2016) 3:9.
9. Li Q, Yan J. Sustainable agriculture in the era of omics: knowledge-driven crop breeding. *Genome Biol.* (2020) 21:1–5.
10. Beiki H, Eveland AL, Tuggle CK. Recent advances in plant and animal genomics are taking agriculture to new heights. *Genome Biol.* (2018) 19:48.
11. Chen F, Song Y, Li X, Chen J, Mo L, Zhang X, et al. Genome sequences of horticultural plants: past, present, and future. *Hortic Res.* (2019) 6:112.
12. Ma NL, Rahmat Z, Lam SS. A review of the "omics" approach to biomarkers of oxidative stress in Oryza sativa. *Int J Mol Sci.* (2013) 14:7515–41.
13. Li MW, Qi X, Ni M, Lam HM. Silicon era of carbon-based life: application of genomics and bioinformatics in crop stress research. *Int J Mol Sci.* (2013) 14:11444–83.
14. Öborn I, Bengtsson J, Hedenus F, Stenström M, Vrede K, Westin C, et al. Scenario development as a basis for formulating a research program on future agriculture: a methodological approach. *Ambio.* (2013) 42:823–39.