


Bioinformatics and omics for crop improvement

Ravi Kumar¹, Manoj Kumar Yadav¹ , Suman Lata Yadav², Mukesh Kumar³, Ankit Kumar Sharma¹, and Manoj Kumar Tripathi⁴

¹Department of Agriculture Biotechnology, Division of Plant Biotechnology, College of Agriculture; SVP University of Agriculture and Technology, Meerut INDIA

²Deputy Librarian, University Library, G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, INDIA.

³ Department of Horticulture, College of Agriculture, SVP University of Agriculture and Technology, Meerut INDIA

⁴Agro Produce Processing Division , ICAR-Central Institute of Agricultural Engineering (Govt Of India), Nabibagh, Baresia Road, Bhopal-462038 (MP) INDIA

#Corresponding author: E-mail: mkyadav711@gmail.com

Abstract: Crop improvement is a continuous process which is driven by the increasing population which needs food security with sustainable production. In the recent past bioinformatics and multiomics tools played an important role and may provide novel opportunities for improving traits by inserting the genes with high precision. In the last decade remarkable progress has been done in the discovery of useful genes for yield, quality, resistance to biotic and abiotic stress etc. With the development of new techniques such as NGS, gene editing and other omics tools with bioinformatics advancement, many new crops have been developed. Therefore, integrated omics and bioinformatics tools are having a major impact which may have a major thrust to develop high yield crops in future. The aim of the present review is to highlights the basic and applied concepts of bioinformatics and genomics by integrating them for crop improvement.

Keywords: Omics; Crop improvement; gene editing; bioinformatics tools.

Citation: Kumar et al., Bioinformatics and omics for crop improvement. Octa J. Biosci. Vol. 11 (1):24-39

Received: 14/02/2023

Revised: 1/06/2023

Accepted: 24/06/2023

Published: 30/06/2023

Publisher's Note: SPS stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In 2021, 828 million people who were 46 million more than in 2020 and 150 million more than in 2019 experienced hunger (FAO, 2019). By 2015, the United Nations Hunger Task Force wants to reduce the number of hunger people half. Since then, 70% population is poor who experience hunger and reside in rural areas. Therefore, improving agricultural production is must to feed deprived population (Sanchez *et. al.*, 2005). With the advancement of biological and agricultural research production of different crops increasing continuously but there is need to refinement of the crop traits using biotechnological strategies viz. Bioinformatics and Omics methods. Searching of novel genes for useful crop traits is a major issue for the plant scientists. Fortunately, these difficulties arise as plant scientists are making incredible strides in understanding the basic mechanisms underlying plant growth and development (Delmer *et. al.*, 2005). They're mostly concerned with the not targeted and un-biased determination of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) in a specific biological sample. Currently, the world's population is growing with a fast pace, and the biggest threat to

food security is climate change. Fast population growth, climate change, and environmental demand for accelerated breeding efforts for high output are the key issues. There are several obstacles, nevertheless, and it is widely acknowledged that increasing the potential and productivity of modern agricultural output are crucial. Researchers are looking for practical and contemporary crop production methods.

The utilisation of next-generation omics technologies such as genomes, transcriptomics, proteomics, metabolomics, and phenomics may speed up upcoming biotechnology development. During the sequencing of the human genome, Thomas Roderick was the first person who coined the term "genomics" in 1986. It is potentially an experimental approach including molecular characterisation and full genome cloning to study gene morphology, application, and synthesis, or it is a novel scientific field involving genome sequencing, analysis, and mapping. Crop improvement has been transformed by technology and scientific developments such as genome sequencing, genotyping for genomic-wide annotator studies, and genomic hypothesis (Liu and Zhang 2019). This is also known as high-dimensional biology and systems biology (Kell *et.al.*, 2007, Westerhoff *et.al.*, 2004). The fundamental belief of these methods is that a complicated system may be understood well when it seen as a whole. In systems biology and omics research, no predetermined or known hypothesis are used while these are dealt with data collection and its processing to test how a gene determines the particular traits (Kell *et.al.*, 2004). Genome research is the scientific study of the genome of an organism. The whole DNA that makes up a cell or an organism is commonly referred to as its genome. The human genome includes a total of 3.2 billion nucleotides as well as between 30 000 and 40 000 genes that code for proteins, as reported by Baltimore et al. (2001). Despite recent major advancements in microarray technology, genes have frequently been investigated separately. DNA microarrays enable the simultaneous investigation of the expression of numerous genes and measurement of DNA sequence variation within any group or species. The proteome is a database of every one of the expressed proteins in an organism, cell, or tissue (Theodorescu et al., 2007).

The term "omics" is frequently used in biology to refer to a topic under investigation or to extensive research, including whole biological collections of data including genomics, proteomics, or metabolomics. When referring to in-depth study in these areas, like the proteome, metabolome, or genome, the accompanying suffix -ome is applied. The emergence of full "omics," arrays, and high-throughput developments has allowed for the finest wide-spectrum gene analysis. Numerous cutting-edge Omics tools have been created using these breakthroughs (Deshmukh *et al.* 2014).

Proteomics seeks to describe information flow across the cell and the organism via protein pathways and networks, with the ultimate objective of understanding the functional importance of proteins (Petricoin et al., 2002; Vlahou et al., 2005). Although proteome research can provide a wealth of helpful information, it is hampered by its large domain size (>100,000 proteins) and the difficulty in successfully detecting low-abundance proteins, as proteins are most likely to be generally affected in illness and disease response. The proteome is seen to have great potential for the creation of biomarkers (Rifai et al., 2006). The proteome is a living mirror that reflects both genes and the environment. Numerous protein disease biomarkers are now available to represent this (for example, CA125 and alpha-fetoprotein).

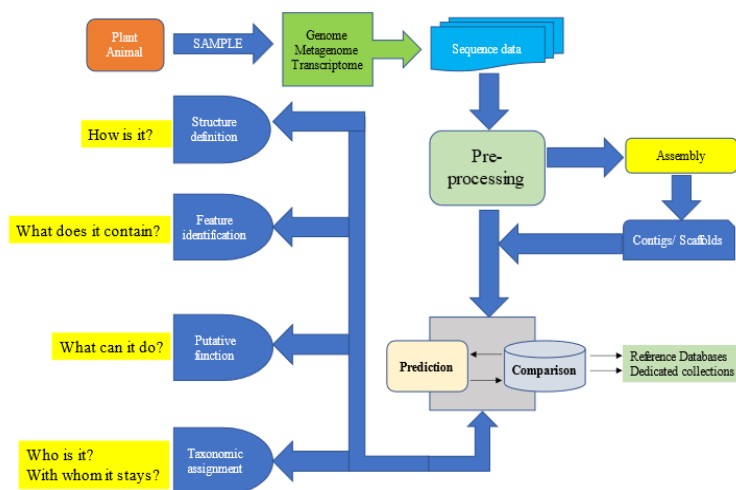


Fig. 1: The flowchart illustrating how several pathways are connected and controlled by omics tools

The use of informatics is for the organization, administration, and dissemination of biological data (Dayhoff *et al.*, 1965) and it is known as bioinformatics. It is an important tool for data analysis, interpretation, and modelling. Additionally, because of bioinformatics, it is possible to examine and comprehend the structure and function of bigger molecular collections obtained from the so-called omics techniques. These initiatives enable the representation of many elements of the biomolecular architecture of complex biological systems, from cells to ecosystems (genomics, transcriptomics, proteomics, metabolomics, etc.) (Fig 1). Due to the fast adoption of omics technologies and their increasing potency at reasonable prices, the extent of molecular data gathering from multiple levels of an organism's organisational structure or an environmental sample has rapidly increased. Due to the vast amount of data and the need for integrative efforts, this further complicated bioinformatics and encouraged an integrated perspective on the structure and function of systems (Chiusano *et al.*, 2008; Bostan *et al.*, 2015).

1. Genomics as an analysis tool

The field of genetics referred to as genomics focuses on the organisation, sequencing, and analysis of an organism's genome. Although structural genomics and functional genomics are the two basic areas of genomics, additional branches are also investigated within this field.

2.1 Structural genomics

The arrangement and sequencing of the DNA throughout the whole genome are the subjects of the scientific field known as structural genomics (Mathur *et al.*, 2021). It shows the first phases of genomic analysis, such as

1. A way enables the creation of a high resolution (HR) genomic map based on physical maps of an organism and recombination frequency (cM), indicating where genes are placed in relation to base pair distance.
2. The genome's sequence.
3. Figuring out all the proteins.

Tools for structural genomic research include genome sequencing methods and DNA-based markers. A DNA sequence that is willingly detected and whose inheritance can easily be detected is known as a molecular marker. These are not affected by any environmental factors. The

complete genome may be monitored and examined with the help of these markers. These markers are generally used to detect the molecular diversity among the germplasm or population. DNA markers also used to locate restriction sites or mutation at a PCR primer site, a DNA insertion or elimination, an alteration in the number of replicate components across two restrictions or PCR primer locations, or to detect a mutation which produces a single nucleotide polymorphism (Mathur et al., 2021).

2.2 Function-driven genomics

The area of functional genomics focuses on the roles and interfaces of proteins, gene transcription, and translation using genome data and high-throughput technology. It attempts to understand how the genome functions at various developmental junctures and in an ecological context. Its differentiating features include high-throughput or large-scale mutually evaluated procedures with computational (bioinformatics) and statistical evaluation of the outcomes (Evelien et al., 2012). Researchers in plant sciences will also use *in silico* tools to develop resistance in crops and to study many other useful traits. Expression analysis, forward and reverse genetics investigations, *in silico* gene analysis and gene sequence annotation is used to evaluate the function of a gene. Up-regulation studies (activation tagging-promoter control/enhancer control, overexpression of genes, multiple copy gene insertion or gain of function through mutagenesis), and down-regulation studies (through insertion mutagenesis, PTGS [post-transcriptional gene silencing], VIGS [virus-induced gene silencing], and chemical mutagenesis has also been done (Dwivedi and Rautela, 2018). Studies on both up-regulation (activation tagging-promoter control/enhancer control, overexpression of genes, multiple copy gene insertion or gain of function through mutagenesis), and down-regulation (insertion mutagenesis, T-DNA tagging transposon tagging, PTGS [post-transcriptional gene silencing], VIGS [virus-induced gene silencing], and chemical mutagenesis) have been described (Dwivedi and Rautela, 2018).

2.3 Genomic comparative analysis

Comparative genomics examines how different species genomes differ in terms of their structure and capabilities. Even distantly related species might benefit from the knowledge obtained on one organism. With the knowledge of species at genomic level, it enables us to comprehend evolution in detail. Identification of regulatory components and genes (both coding and noncoding) is very much helpful to know the function of a particular gene which governed a trait. Segmental duplications, rearrangements, and polyploidy all affect how the genome is organised, and comparative genomics is a strong tool for detecting these alterations. This method has been proposed to assess homologous genes and consequently identify conserved cis-regulatory motifs. Synteny between related species has been discovered through comparative genomics research (conserved gene placement within significant portions of the genome in various species). In order to comprehend how novel properties of a new gene have emerged, comparative genomics is also used (Susi et al. 2020).

2.4 Evolutionary genomics

It has been suggested to examine the genomic sequence of creatures from various species to learn how the genome has changed throughout evolution. The structure of partial or full genomes changes during evolution as a result of duplications and deletions. Once genome sequence has spread across the population, changes to genome size happen instantly. The genome sequence of a newly discovered species may be used to make assumptions about the timing of duplications and

deletions in the past as well as the relationship between changes in a genomic area and various phases of evolution (Van Tassel et al. 2020).

2.5 Epigenomics

A group of epigenetic alterations in a cell's epigenome, which is its genetic material, are the subject to the study of its molecular biology known as epigenomics. The DNA or histones in a cell that affect gene expression without altering the DNA sequence are examples of reversible epigenetic changes. Histone modification and DNA methylation are two prominent epigenetic changes which occurred frequently in genome. The expression and control of genes, as well as cellular processes including development, differentiation, and cancer, are all impacted by epigenetic changes (Kapazoglou et al. 2018).

2.6 Metagenomics

The field of biology known as "metagenomics" involves research on metagenomes or genetic material that has been recovered from ambient tissues. Environmental genomics, eco-genomics, and community genomics are other terms for it. Traditional approaches, on the other hand, rely on improved clonal cultures, firstly environmental gene sequencing, and cloned particular genes (16S rRNA gene) to generate variety in an already existing species (Kumar et al. 2020).

Application of genomics

- ✓ Gene identification and cloning.
- ✓ Gene prediction/discovery.
- ✓ Genome sequence gives the structure of the chromosome, genetic mapping and location of the genes.
- ✓ Genome sequencing identified the mutated region in sequence.
- ✓ Genome manipulation in genetic features like crop yield, disease resistance, growth abilities, nutritive qualities and drought tolerance.
- ✓ Quantitative trait locus (QTL) analysis and marker-assisted selection.
- ✓ Comparative genomics.
- ✓ Gene banks and chromosome stocks.
- ✓ Understanding expression profiles, responses and interactions.
- ✓ Oral plant vaccines against hepatitis B where transgenes create surface antigens that stimulate immunity.
- ✓ Compare the genomic sequence of species and understand how the genome has been remodelled in the course of evolution.
- ✓ Genomics also engages the research of intragenomic methodologies such as heterosis, epistasis and pleiotropy along with the interfaces between loci and alleles within the genome.

Fig 2: Various applications of Genomics

By utilizing structural, functional, and comparative information, genomics, the study of the genome has given momentum to agricultural research for improving useful traits. At the structural level, molecular markers and sequencing technology provide good knowledge. The advancement of sequencing and genotyping technology has aided in the further development of new molecular markers. This results in crop genotyping on a massive scale, which may then be utilized to create genetic and physical maps with high densities. To ascertain the gene's function, functional genomics is studied. The majority of crops' genomes are also analyzed using this method. It helps on precise genetic modification of gene which may help to increase production and survival of crop plants under adverse conditions (Fig 2). Other techniques such as SMART (Simple

Modular Architecture Research Tool) breeding or breeding by design can make advantage of the combination of genomics and other omics technologies (Mathur et al., 2021).

3. Proteomics

The scientific study of proteins, primarily their function and assembly, is known as proteomics. It is a superb method for studying metabolic changes in response to varied unfavourable conditions. To characterise the thorough characterization of the whole protein part of a cell, tissue, or organism, the term "proteomics" was first used in 1995 (Evelien et al., 2012).

3.1 Types of proteomics

3.1.1 Structural proteomics

Understanding the physical complexities of functioning proteins and their three-dimensional (3D) shape is made easier by structural proteomics. When the amino acid sequence of a protein is fixed in stone, either by sequencing or by a process known as homology modelling, it is considered that the protein exists. This approach provides detailed confirmation of the design and function of protein complexes in a given cell organelle. Knowing every protein in composite structures such as membranes, ribosomes, and cell organelles, as well as analysing every protein complex that may occur inside them, are both conceivable. For structural fortitude, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy were utilised (Beale, 2020).

3.1.2 Functional proteomics

This technique describes how proteins perform an empathetic role in the cell, as well as sly molecular tactics based on the discovery of linked protein complexes. The complexities of a peculiar protein, with complexes suited to a certain protein engaged in a specific way, will be disturbingly similar to its biological function. A full explanation of the multiple intracellular signalling processes may be required to appreciate the description of protein-protein interactions in vivo (Liu et al. 2018; Popp and Maquat 2018; Heusel et al. 2019; Schaffer et al. 2019; Bludau and Aebersold 2020; Salas et al. 2020).

3.1.3 Expression proteomics

Expression proteomics is used to show entire proteins in two contexts in a qualitative and quantitative manner. Naturally, the study of expression protein decorations in abnormal cells is related to that of expression proteomics. Two-dimensional (2D) gel electrophoresis and mass spectrometry studies have been suggested for visualising the protein expressional behaviours, whether or not a protein is expressed in a cancer cell (Mathur et al., 2021). When connected, these activities can be identified as signalling pathways, multiprotein complexes, and protein achievements (Kwok et al. 2020).

3.2 Techniques involved in proteomics

Both logical and bioinformatics techniques have been proposed in this area of biological sciences to identify protein structure and function. These tools include of 2D gel electrophoresis; MALDI-TOF-MS (matrix-assisted laser desorption/ionization time-of-flight mass spectrometry) and some other recent techniques.

3.2.1 2D gel electrophoresis

Isoelectric focusing (IEF), a technique used in 2D gel electrophoresis, is used to quantify protein tissues, followed by molecular weight. The result is a shadow made up of many tiny dots, each of which represents a protein. 1000–2000 protein spots may be seen on a large 2D gel, and these spots can be seen after staining as gel dots. This approach is especially suggested when comparing identical samples to identify specific protein variations. Additionally, a concentration of the

protein solution is created before being used for IEF. Key protein solubility, natural charge, and relative presence are some particular considerations. The isoelectric point is used to distinguish proteins (pI). The sample is loaded as a pH gradient employing sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), and various expressed proteins are loaded in accordance with size using an appropriate immobilised pH gradient (IPG) strip of a certain length. Proteins are further sorted by size using SDS-PAGE, and the best-sized proteins are then chosen. Fluorescent protein tags are used to screen proteins. The next step is the acquisition of digital photographs utilizing 2D setups and the best imaging hardware and software. After that, 2D software is used to study the expression methods. A concentrated protein is separated from the gel and digested (Mathur et al., 2021).

3.2.2 MALDI-TOF

The investigation of biomolecules including proteins, peptides, and DNA is the sole application of the revolutionary method known as MALDI-TOF, which makes use of soft ionisation technology in spectrometry. The minor instability and thermal shakiness of these synthetic polymers and biologically active chemicals restrict the use of MS as a tool for identifying a specific protein of interest. These kinks have been ironed out thanks to advancements of MALDI-TOF-MS, which is often employed for the molecular weight evaluation of bioactive compounds by vaporisation and ionisation. Ionising the sample with a laser beam has been suggested; it does not render the chemical inactive and it continues to exist in its natural condition thereafter (Hou et al. 2019).

3.3 Advances in proteomics methods

3.3.1 ICAT, or isotope-coded affinity tag

It is used to quantify proteomics that depends on synthetic labelling agents and is free from the gel. These probes have three main components: an isotopically coded linker, an identified side chain of amino acids, and a docking site for the empathy separation of tagged proteins and peptides. Each sample is marked with isotopic light for the quantitative analysis of two proteomes, while the other sample is marked with a large report. Both samples had isotope-coded tagging mixes applied to them. LC-MS is used to research these peptides. Typically, deuterium tags are employed. The method is specifically used to quantify proteins virtually in two or more biological samples. is that can be seen can be used in this approach (Mathur et al., 2021).

3.3.2 Isobaric tags for relative and absolute quantification (iTRAQ)

For counting proteins, the non-gel-based approach of utilising iTRAQ is also proposed. It is used in proteomics to analyse quantitative changes in the proteome. The reagents 4-plex and 8-plex can be used to tag all peptides from various samples by covalently attaching the N-terminus and side-chain amines of peptides from protein cleavage with tags of varied weight. MS/MS analysis of additional samples is possible. Various tools, such as j-Tracker and j-TraqX 20, should be used to evaluate the MS/MS spectra (Morales et al., 2017).

3.3.3 Absolute quantification (AQUA)

For research on the unqualified measurement of proteins and their various forms, the AQUA approach is advised. To create in vitro proteins, covalent modifications are used. These changes share chemical similarities with posttranslational processes that are logically present. With the use of a tandem MS, these kinds of peptides are used to count the post-translational changed proteins after full digestion (Schnatbaum *et al.* 2020).

3.3.4 Electrospray ionisation-quadrupole-ion trap-mass spectroscopy (ESI-Q-IT-MS)

Proteomics makes great use of ESI-Q-IT-MS. Ionization proteins in ESI are charged differently and are ionized in solution. The advantage of using this approach for investigations on the mass of proteins is that the TOF detector has great mass precision in this examination region and that proteins' substantial charge state causes their m/z dimensions to generally be less than 2000. The results are more accurate ESI-Q-TOF mass measurements for proteins (Bian *et al.* 2020).

3.3.5 Surface-enhanced laser desorption/ionization (SELDI) TOF-MS

Protein mixtures are determined using the (SELDI) TOF-MS method, an ionization assay used in MS. To screen proteins in clinical trials and to compare the number of proteins with and without a disease that may be indicated for biomarker research, SELDI is notably employed with TOF mass spectrometers (Hill *et al.* 2020).

4. Applications of proteomics

The first step is to examine the peptide sequence after receiving fragmentation spectra data from MS. Two of these techniques are de novo peptide sequencing and probing against databases of fragmentation spectra (Chen *et al.* 2020). The pattern of all expressed or expected protein sequences that have been in silico cleaved is used to create a target database in the latter procedures. A peptide spectrum match (PSM) score is then computed for each fragmentation spectrum and each piece of experimental fragmentation spectrum evidence from the target database. The query peptide can be given top priority by preserving the peptide with the greatest PSM score. It is a never-ending problem to select the finest probing algorithms that deliver the best peptide spectrum matching findings from databases. The basis for recording purposes in traditional protein database search engines like SEQUEST is standardised cross-correlation of the mass-to-charge ratio assumed from the identified sequence of amino acids in databases and the fragment ions recognised from the tandem mass spectrum (Timp and Timp 2020). MASCOT (Timp and Timp 2020), another well-known programme developed later, packs a probability-based score for protein identification utilising frame capacity, protein sequence data, peptide molecular weights from protein absorption, and tandem MS data.

4.1 Gene ontology (GO)

Gene Ontology (GO) is the method that is most typically used to advance analysis. In order to decrease the severance in expressions, it is characterised as a set of preset clusters to which certain genes have been assigned according to their functional properties. The GO keywords are made up of three major word groups: "molecular function," "the biological process," and "cellular component." Each sentence has a connection to the others and can be recognised differently. The AmiGO database (Munoz-Torres and Carbon 2017) has GO word footnotes for numerous species, however not every protein has a precise and comprehensive annotation. The use of informative GO keywords from other proteins in the same database can help proteins with incomplete interpretations. GO term assumption techniques such ProLoc-GO (Lande *et al.* 2020), PFP (Wang *et al.* 2020), and IGNA (Piovesan *et al.* 2015) might be used to avoid these time-consuming activities.

The most often utilised statistical tests suggested in GO enrichment are the Fisher's exact test and the hypergeometric test. According to statistics, significant GO keywords are those that seem to repeat in the effort protein slope more often than would be predicted under normal conditions and may have indicated extraordinary biological mechanisms that need more research. Since GO keywords often characterise ORF products rather than recognised methodologies, scientists should carefully analyse the GO words in the improved findings to verify that they have

significant ties between the correct forms and linked genes. As with GO terms, previous familiarity with stabilised mechanism networks and disorders may be used to finish augmentation research (Welzenbach et al. 2016). Biological pathways show a plausible biological process by explaining how chemicals within a cell react chemically in predictable ways. Databases like PANTHER and Reactome (Aslam et al. 2017) contain the interface mappings for multiple mechanisms together with a selection of improvement tools in order to achieve enhancement consistently on the database webpage. For instance, the R package Path view (Luo and Brouwer 2013) supports the KEGG route (Kanehisa et al. 2017) produced data aggregation and visualising. Protein set enrichment analysis (PSEA), a variant of gene set enrichment analysis (GSEA), is another popular enrichment technique. A weighted running total statistic is used to calculate the PSEA's upgrading score, and proteins that don't undergo sufficiently major modifications may have a negative effect on the score. Many various kinds of GSEA software can do PSEA, although programmes designed exclusively for protein quantification data, such PSEA-Quant (Lavallée-Adam et al. 2014), could offer a more practical and acceptable method for proteomics.

4.2 Agriculture-related applications

Promising advancements in plant proteomics have been made. Additionally, plant proteomics is utilized to identify plant-insect connections that help identify important genes involved in the defense mechanisms of plants against herbivores. The stability of agricultural crop yields is severely constrained by the rise of the population and the impact of global atmospheric changes (Chin and Tan 2018). Food technology frequently uses proteomics for resource discovery and optimisation, course innovation, revealing batch-to-batch variances, and switching the superiority of the final product. In particular, biological and microbiological securities as well as the usage of genetically modified crops are the subjects of more study on the aspects of food safety (Wu et al. 2017).

5. The most recent *in-silico* prediction tools

The prediction tool evolved into a useful tool for studying new sequencing data. Due to the availability of multiple bioinformatic tools, it is now possible to anticipate a gene's presence in a sequence as well as its protein's structure and function, hence saving time and resources during validation. Following is a discussion of some of the prediction tools:

5.1 Tools for predicting genes

Exploring the existing sequences that are stored in the databases is a crucial aspect of genome annotation. The current aim is to annotate the existing genome for study and beneficial uses utilising existing high-throughput sequencing technologies. Gene prediction tools (table 1.) might be valuable in this research. The primary techniques used in gene prediction include finding transcriptional regulatory sites, splice sites, poly-A tail sites, translation start/stop sites, ORFs, and homology searches.

Table 1. Various bioinformatics tools used for predicting genes structure and their function

Name	Description	Species	References
FINDER	RNA-Seq data and related protein sequences can be used to annotate eukaryotic genes.	Eukaryotes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-pmid33879057-1
FragGeneScan	sequencing and predicting genes in whole genomes Read	Prokaryotes, Metagenomes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-2
AUGUSTUS	Tools for eukaryote gene predictor	Eukaryotes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-pmid21216780-5
EasyGene	A hidden Markov model (HMM) that is automatically calculated for a fresh genome serves as the foundation for the gene finder.	Prokaryotes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-8 https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-9
Eugene	Integrative gene finding	Prokaryotes, Eukaryotes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-10 https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-11
FGENESH	Multiple genes, both chains, HMM-based gene structure prediction	Eukaryotes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-pmid10779491-12
GeneParser	Distinguish between introns and exons in DNA sequences.	Eukaryotes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-18
ORFfinder	All open reading frames may be found using a graphical analysis tool.	Prokaryotes, Eukaryotes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-40
GeneTack	Predicts prokaryotic genomes with frameshift genes	Prokaryotes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-pmid20556861-23
BioNIX	GRAIL, FEX, HEXON, MZEF, GENEFINDER, FGENE, BLAST, POLYAH, REPEATMASKER, and TRNASCAN are some examples of websites that mix the output of several programmes.	Prokaryotes, Eukaryotes	https://en.wikipedia.org/wiki/List_of_gene_prediction_software#cite_note-37

5.2 Software used to predict genes sequence

5.2.1 Similarity-based

It is the easiest method for determining how closely two gene sequences match those of other genomes' genes, ESTs, and proteins. For identification, these strategies make use of local and global alignments. BLAST uses local alignments to determine how similar genes, ESTs, and proteins are to one another. Global alignments use homologous proteins of translated ORFs in a genomic sequence to predict genes.

5.2.2 *Ab initio*-based

Ab initio gene predictions rely on the recognition of gene architecture rather than similarities. It frequently uses the gene model to discover any genes. This technique makes use of the signal and content sensors. Signal sensors are able to recognise splice sites, poly-A tail sites, and translation starts/stop sites whereas content sensors employ statistical detection methods to distinguish exonic codons from noncoding. This method enables the use of gene prediction tools like Gene ID and Genie (Wang et al. 2004).

5.2.3 *Combined evidence*-based

In this model, which integrates the aforementioned model methodologies, the gene model is coupled with alignment with recognised ESTs and proteins. When compared to the first two procedures, they produce the best results. Brendel et al. (2004) used GeneSequer in a combined evidential technique.

5.3 Protein prediction software

Proteins are the end product of fundamental dogma. Because of advances in sequencing technology, the quantity of protein sequences is increasing on a daily basis. For academics, determining such sequence structure and function is a time-consuming task. Using protein prediction methods, the "sequence structure/function gap" can be bridged. Protein prediction tools, like gene prediction tools, identify the protein using one of three approaches, which are described below (table 2.).

5.3.1 Homology modelling

The similarity between formerly existing proteins in databases is a key component of this similarity-based strategy, or template protein and candidate protein. In this approach, the proportion of identical residues and the amino acid sequences decide the structure, making it the simplest. The identification and initial alignment of templates, alignment corrections, backbone generation, loop modelling, side-chain modelling, model optimisation, and model validation are the seven primary steps that go into homology modelling (Krieger et al. 2003). The different tools are given in table 2 used for predicting proteins (Homology modelling).

Table 2. Tools for predicting proteins

Name	Method	Description	Link
IntFOLD	A unified interface used for Tertiary structure prediction/3D modelling, 3D models quality assessment, Intrinsic disorder prediction, Domain prediction, Prediction of protein-ligand binding residues	Automated web server and a few programmes for download	https://www.reading.ac.uk/bioinf/IntFOLD/
RaptorX	Protein 3D modelling, distant homology discovery, and binding site prediction	a downloaded software and an automated web server	http://raptorx.uchicago.edu/
Biskit	External programmes are bundled into an automated process	T-Coffee alignment, BLAST search, and MODELLER building	http://biskit.sf.net/
ESyPred3D	Template recognition, alignment, and 3D modelling	Webserver that is automatically updated	http://www.fundp.ac.be/urbm/bioin

			fo/esympred/
FoldX	Protein design and energy calculations	Downloadable software	http://foldx.crg.es/
Phyre and Phyre2	Used to search remote template identification, alignment, 3D modelling, multi-templates, and <i>ab initio</i> study.	Webserver with task management, fold library that is automatically updated, genome searching, and other features	http://www.sbg.bio.ic.ac.uk/~phyre/
HHpred	Template recognition, alignment, and 3D modelling	Webserver that is interactive and provides assistance	http://arquivo.pt/wayback/20160514083149/http%3A://toolkit.tuebingen.mpg.de/hhpred
MODELLER	Satisfaction with spatial limitations	The standalone programme is written primarily in Fortran and Python.	https://salilab.org/modeller/
CAN FOLD	Contact satisfaction and distance constraints	The standalone programme is written primarily in Fortran and Perl.	https://github.com/multicom-toolbox/CONFOLD
MOE (Molecular Operating Environment)	Loop modelling, rotamer libraries for sidechain conformations, and utilising MM forcefields, as well as template identification, multiple template use, and accounting for additional environments (e.g., excluded ligand volumes)	Platform proprietary, available on Windows, Linux, and Mac.	http://www.chemcomp.com/
ROBETTA	Rosetta homology modelling and fragment assembly used for Ginzu domain prediction	Webserver	http://robeta.bakerlab.org/
BHAGEERA TH-H	Methods of <i>ab initio</i> folding and homology are combined.	Predictions of protein tertiary structure	http://www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp
SWISS-MODEL EL	Local resemblance/fragment assembly	Web server automation (based on ProModII)	http://swissmodel.expasy.org/

Yasara	Ligand and oligomer modelling, template detection, alignment, and model fragment hybridization	Text mode (clusters) or a graphical user interface	http://www.yasara.a.org/ http://www.yasara.a.org/casp8.htm
AWSEM-Suite	Molecular dynamics simulation used to know co-evolutionarily history, template-guided, optimised folding landscapes	Automated webserver	http://awsem.rice.edu/

5.3.2 *Ab initio* or *de novo* modelling

When the proportion between the candidate protein and the proteins in the databases is low, this modelling is carried out (Faiza 2017). Native protein will be used in this simulation, which is based on free energy. For *ab initio* modelling, tools (table 3) like ROSETTA and TOUCHSTONE-II (Zhang *et al.* 2003) are helpful.

Table 3. Various tools for *de novo* or *ab initio* modelling

Name	Method	Description	Link
trRosetta	The trRosetta method used to predicts <i>de novo</i> protein structures quickly and precisely. It design the protein structure using constrained Rosetta and direct energy minimizations. A deep residual neural network's prediction of the inter-residue distance and orientation distributions is one of the constraints.	source code and web server. Folding 300 amino acids (AAs) or less proteins take around one hour.	https://yanglab.nankai.edu.cn/trRosetta/ https://yanglab.nankai.edu.cn/trRosetta/download/
ROBETTA	Ginzu domain prediction combined with Rosetta homology modelling and <i>ab initio</i> fragment assembly	Webserver	http://new.robetta.org/
Rosetta@home	Rosetta algorithm implementation in distributed computing	downloadable application	http://boinc.bakerlab.org/rosetta/
Abalone	Folds in molecular dynamics	Program	http://www.biomolecular-modeling.com/Abalone/Protein-folding.html
C-QUARK	C-QUARK is a technique for predicting protein structures from scratch. based on contact-map predictions from simulations of fragment assembly that use deep learning.	Webserver	https://zhanggroup.org/C-QUARK/

5.3.3 Threading

The folds of the template and candidate proteins are used in threading for identification. In that a particular fold must be derived from a protein database, it is similar to homology modelling. This searches for related folds in proteins with evolutionary connections. THREADER, I-TASSER (Roy et al. 2010), COTH, and other programmes are examples of threading tools (table 4).

Table 4. The different threading tools used to elucidate protein structure and their modelling

Name	Method	Description	Link
I-TASSER	Reassembling fragmented structures using threading	The protein modelling website server	http://zhanglab.ccmb.med.umich.edu/I-TASSER/ http://zhanglab.ccmb.med.umich.edu/I-TASSER/download
THREADER	a programme called threading, which employs a thorough 3-D model of the protein structure in order to recognise folds.	The online server is used for the 3-D representation of protein structure.	https://bio.tools/threader
COTH	To find and assemble protein complex structures from both tertiary and complex structure libraries, researchers developed the multiple-chain protein threading method.	the online tool for recognising and recombining protein complex structures	https://zhanggroup.org/COTH/

6. Conclusion

This review is focussed on the applications of proteomics in diverse areas of research and provides a technique to use the produced genomics and other omics resource in a more controlled manner. It does so based on the aforementioned important results and omics applications. It is essential to combine genome and proteome data with other omics activities to develop a fresh bioinformatics prototype. Since genes and proteins may be linearly ordered across the two omics domains, judgments between proteomics and transcriptomics can be made in a well-defined manner. Other omics data can also be used in the contrast design of protein signalling grids. According to the data at hand, alternative reconstructions of omics data are usually at odds with one another and MS-based proteomics will have garnered a lot of attention when seen from a multi-omics perspective. As a result, as new data are generated by the MS-based proteomics approach, the related bioinformatics tools need to be updated. Bioinformatics and genomics-related inspection processes have been addressed in the present review that may be utilized for crop improvement using these novel tools through incorporating useful genes for high yielding and quality traits.

Author contributions

All the authors have the equal contribution to develop this review paper. MKY and MK conceive the idea; RK and AKS wrote the review. SLY collected and corrected the references. MKY and MKY corrected the review.

Funding

There is no funding involved to develop this manuscript.

Acknowledgement

Authors of this manuscript are thankful for Sardar Vallabhbhai Patel University of Agriculture and Technology Meerut for their support. We are also acknowledged Incharge, University Library, G.B. Pant University of Agriculture and Technology, Pantnagar (Uttarakhand) for providing facility and support for reference collection.

References

- Aslam B., Basit M., Nisar M.A., Khurshid M., Rasool M.H. (2017), Proteomics: technologies and their applications. *Journal of Chromatographic Science*, 55(2): 182–196.
- Bączor R., Waliczek M., Stefanowicz P., and Szewczuk Z. (2019), Trends in the design of new isobaric labeling reagents for quantitative proteomics. *Molecules*, 24(4): 701.
- Baltimore D. (2001), Our genome unveiled. *Nature*, 409(6822): 815-816.
- Beale J. H. (2020), Macromolecular X-ray crystallography: soon to be a road less travelled? *Acta Crystallographica Section D: Structural Biology*, 76(5): 400-405.
- Bian Y., Zheng R., Bayer F. P., Wong C., Chang Y. C., Meng C., Kuster B. (2020), Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC–MS/MS. *Nature communications*, 11(1): 157.
- Bludau I., Aebersold R., (2020), Proteomic and inter atomic insights into the molecular basis of cell functional diversity. *Nature Reviews Molecular Cell Biology*. 21(6):1–14
- Bostan H., Chiusano M. L., (2015), NexGenEx-Tom: a gene expression platform to investigate the functionalities of the tomato genome. *BMC Plant Biol.* 15 (1): 48.
- Brendel V., Xing L., and Zhu W., (2004), Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics* 20:1157–1169.
- Chen C., Hou J., Tanner J. J., Cheng J., (2020), Bioinformatics methods for mass spectrometry-based proteomics data analysis. *International Journal of Molecular Sciences*.21(8): 2873
- Chin C.F., Tan H.S. (2018), The use of proteomic tools to address challenges faced in clonal propagation of tropical crops through somatic embryogenesis. *Proteomes*. 6(2): 21.
- Chiusano M. L., D'Agostino N., Traini A., Licciardello C., Raimondo E., Aversano M. Monti L. (2008), ISOL@: an Italian SOLANaceae genomics resource. *BMC bioinformatics*. 9(2), 1-11.
- Dayhoff M. O., Eck R. V., (Eds.). (1972), *Atlas of protein sequence and structure*. National Biomedical Research Foundation.
- Delmer D. P. (2005), Agriculture in the developing world: connecting innovations in plant research to downstream applications. *Proceedings of the National Academy of Sciences*. 102(44), 15739-15746.
- Deshmukh R., Sonah H., Patil G., Chen W., Prince S., Mutava R., Nguyen H. T. (2014), Integrating omic approaches for abiotic stress tolerance in soybean. *Frontiers in Plant science*. 5, 244.
- Dwivedi M., Rautela A. (2018), Application of functional genomics in agriculture. *International Journal of Chemical Studies* 6(1): 462–465.
- Evelien M. Bunnik K. G., Le Roch (2012), An Introduction to Functional Genomics and Systems Biology, *Advances in Wound Care*, 2:9
- Faiza M. (2017) Ab-initio prediction of protein structure: an introduction. *Bioinforma Rev* 2:4
FAO, FAO, Rome 2019.
- Food and Agriculture Organization (FAO) of the United Nations; International Fund for FAO: Rome, Italy, 2019.
- Heusel M., Bludau I., Rosenberger G., Hafen R., Frank M., BanaeiEsfahani A. Aebersold, R. (2019), Complex centric proteome profiling by SECSWATH MS. *Molecular systems biology*. 15(1), e8438.
- Hill V., Kuhnert P., Erb M., Machado R. A. (2020), Identification of Photorhabdus symbionts by MALDI-TOF mass spectrometry. *bioRxiv*. 2020-01.
- Hou T.Y., Chiang-Ni C., Teng S.H. (2019), Current status of MALDI-TOF mass spectrometry in clinical microbiology. *Journal of Food and Drug Analysis*. 27(2): 404–14.

- Kanehisa M., Furumichi M., Tanabe M., Sato Y., Morishima K. (2017), KEGG: new perspectives on genomes, pathways, diseases, and drugs. *Nucleic Acids Research* 45(1): 353–361
- Kapazoglou A., Ganopoulos I., Tani E., Tsaftaris A. (2018), Epigenetics, epigenomics, and crop improvement. In *Advances in Botanical Research* 86:287–324. Academic Press, London
- Kell D.B., Oliver S.G. (2004), Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*. 26:99–105.
- Kell D.B. (2007), The virtual human: towards a global systems biology of multiscale, distributed biochemical network models. *Iubmb Life*; 59:689–95. doi:10.1080/15216540701694252
- Krieger E., Nabuurs S.B., Vriend G. (2003), Homology modeling. In: Bourne P, Weissig H (eds) *Structural bioinformatics*. Wiley-Liss, Hoboken. 507–521
- Kumar A., Ravindran M., Sarsaiya B., Chen S., Wainaina H., Singh S., Zhang Z. (2020), Metagenomics for taxonomy profiling: tools and approaches. *Bioengineered* 11(1): 356–374
- Kwok C.S.N., Lai K.K.Y., Lam S.W., Chan K.K., Xu S.J.L., Lee F.W.F. (2020), Production of high-quality two-dimensional gel electrophoresis profile for marine medaka samples by using trizol-based protein extraction approaches. *Proteome Science* 18: 1–13.
- Lande N.V., Barua P., Gayen D., Kumar S., Chakraborty S., Chakraborty N. (2020), Proteomic dissection of the chloroplast: Moving beyond photosynthesis. *Journal of Proteomics* 212: 103542.
- Lavallée-Adam M., Rauniyar N., McClatchy D. B., Yates III J. R. (2014), PSEA-Quant: a protein set enrichment analysis on label-free and label-based protein quantification data. *Journal of Proteome Research*. 13(12): 5496–5509.
- Liu X., Salokas K., Tamene F., Jiu Y., Weldatsadik R. G., Öhman T., Varjosalo M. (2018), An AP-MS-and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nature communications*. 9(1), 1188.
- Liu, D. D., and Zhang, L. 2019. Trends in the characteristics of human functional genomic data on the geneexpression omnibus, 2001–2017. *Laboratory Investigation* 99(1): 118–127.
- Luo W., Brouwer C. (2013), Pathview: a R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 29(14): 1830–1831.
- Mathur M., Prajapat R. K., Upadhyay T. K., Lal D., Khatik N., Sharma D. (2021), Advances in Genomics and Proteomics in Agriculture. In *Crop Improvement* (pp. 23-35). CRC Press.
- Morales A. G., Lachén-Montes M., Ibáñez-Vea M., Santamaría E., Fernández-Irigoyen J. (2017), Application of isobaric tags for relative and absolute quantitation (iTRAQ) to monitor olfactory proteomes during Alzheimer’s disease progression. In *Current Proteomic Approaches Applied to Brain Function*, 29–42. Humana Press, New York.
- Munoz-Torres M., Carbon S. (2017), Get GO! retrieving GO data using AmiGO, QuickGO, API, files, and tools. In *The Gene Ontology Handbook*, 149–160. Humana Press, New York.
- Petricoin E, Zoon K, Kohn E, Barrett J and Liotta L. (2002), Clinical proteomics: translating benchside promise into bedside reality. *Nat Rev*. 1:683–695.
- Piovesan D., Giollo M., Leonardi E., Ferrari C., Tosatto S. C. (2015), INGA: protein function prediction combining interaction networks, domain assignments, and sequence similarity. *Nucleic Acids Research*. 43(1): 134–140
- Popp M. W., Maquat L. E. (2018), Nonsense-mediated mRNA decay and cancer. *Current Opinion in Genetics & Development*. 48: 44–50.
- Rifai N, Gillette MA and Carr S.A. (2006), Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature Biotechnol*. 24:971–983.
- Roy A, Kucukural A, Zhang Y. (2010), I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protoc*.5:725–738.
- Salas D., Stacey R.G., Akinlaja M., Foster L.J. (2020), Next-generation interactomics: considerations for the use of co-elution to measure protein interaction networks. *Molecular and Cellular Proteomics*. 19(1): 1–10.
- Sanchez P. A., Swaminathan M. S. (2005), Cutting world hunger in half. *Science*. 307(5708), 357-359.

- Schaffer L. V., Millikin R. J., Miller R. M., Anderson L. C., Fellers R. T., Ge Y. Smith, L.M. (2019), Identification and quantification of proteoforms by mass spectrometry. *Proteomics*. 19(10), 1800361.
- Schnatbaum K., SolisMezarino V., Pokrovsky D., Schäfer F., Nagl D., Hornberger L., Reimer U. (2020), Front Cover: New Approaches for Absolute Quantification of Stable Isotope Labeled Peptide Standards for Targeted Proteomics Based on a UV Active Tag. *Proteomics*, 20(10), 2070081.
- Susič N., Janežič S., Rupnik M., Stare B.G. (2020), Whole genome sequencing and comparative genomics of two nematocidal *Bacillus* strains reveals a wide range of possible virulence factors. *G3: Genes, Genomes, Genetics*, 10(3), 881-890.
- Theodorescu D., Mischak H. (2007), Mass spectrometry-based proteomics in urine biomarker discovery. *World journal of urology*, 25. 435-443.
- Timp W., Timp, G. (2020), Beyond mass spectrometry, the next step in proteomics. *Science Advances*, 6(2), eaax8978.
- Van Tassel D. L., Tesdell O., Schlautman B., Rubin M. J., DeHaan L. R., Crews T. E., Streit Krug A. (2020), New food crop domestication in the age of gene editing: genetic, agronomic and cultural change remain co-evolutionarily entangled. *Frontiers in Plant Science*. 11: 789.
- Vlahou A., Fountoulakis M. (2005), Proteomic approaches in the search for disease biomarkers. *Journal of Chromatography B*. 814(1): 11-19.
- Wang Z., Chen Y., Li Y. (2004), A brief review of computational gene prediction methods. *Genomics, proteomics & bioinformatics*. 2(4), 216-221.
- Li Y. C., Wang L., Law J. N., Murali T. M., Pandey G. (2022), Integrating multimodal data through interpretable heterogeneous ensembles. *Bioinformatics advances*. 2(1), vbac065.
- Wang, L., Law, J., Murali, T.M.N., and Pandey, G. 2020. Data integration through heterogeneous ensembles for protein function prediction. bioRxiv. <https://doi.org/10.1101/2020.05.29.123497>
- Welzenbach J., Neuhoff C., Heidt H. (2016), Integrative analysis of metabolomic, proteomic, and genomic data to reveal functional pathways and candidate genes for drip loss in pigs. *International Journal of Molecular Sciences*. 17(9): 1426
- Westerhoff H.V., Palsson B.O. (2004), The evolution of molecular biology into systems biology. *Nature Biotechnol*. 22:1249–52.
- Wu F., Zhong F., He F. (2016), Microbial proteomics: approaches, advances, and applications. *Journal of Bioinformatics, Proteomics and Imaging Analysis*. 2(1): 85–91.
- Zhang Y., Kolinski A., Skolnick J. (2003), TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophysical journal*. 85(2): 1145-1164.