The Planta RBS, Vol.-5, Issue-1, April 2024: 1540 – 1544
BARTHWAL *ET AL*. (Series Chapter No. - 167)
https://www.pgrindias.in

General Article

Bioinformatics Tools in Advanced Science in Agriculture

POOJA BARTHWAL, SHIVANI TYAGI AND REDDIVARI VISHNU VARDHAN REDDY

Quantum University, Mandawar, Roorkee, Uttarakhand

Corresponding E-mail: poojabarthwal30@gmail.com

Abstract

Agricultural Bioinformatics in association with biotechnology has the potential to address longstanding issues in agriculture. Increasing applications of bioinformatics and computational biology tools for understanding genetic and epigenetic mechanisms involved in the control of economically important traits of field and horticulture crops, forest trees, and livestock species necessitated the need for a specialized policy framework to optimize the benefits of the specialized field in addressing the pertinent issues of the agricultural sector in the country. It is an urgent need for creating a pool of bioinformaticians at different levels in India to tap the benefits of this stream in Indian agriculture. It is equally important to mention that this is a fast progressing research field and hence, continuous up-gradation of skills should be an integral part of national strategy.

Introduction

Recent technologies developments and instrumentation allow large-scale as well as nano-scale biological samples probing for generation of unprecedented data in life sciences (Noman *et al.*, 2016). For human brain, it is too much difficult to process this sea of data. There is an increasing need for computational methods to process and contextualize these data. Recently, main quantitative proteomics techniques linked with nano-LC-MS/MS have specified the proteomic analysis in the model plant, *Arabidopsis thaliana* (Niehl *et al*., 2013) as well as other non-model plants, like *Zingiber zerumbet* (Mahadevan *et al*., 2014) and *Nicotiana attenuata* (Weinhold *et al.*, 2015).

Last two decades have witnessed significant contributions of bioinformatics in agriculture, specifically, after the emergence of next generation sequencing technologies, high throughput proteomics, metabolomics, phenomics tools, and precision breeding. Among all the applications, use of bioinformatics in genomics studies, e.g. NGS, omics studies, marker discovery and DNA sequence data mining has seen tremendous progress, strikingly contributing in many different aspects of agriculture. Emerging biotechnology based breeding approaches such as genome wide association studies (GWAS), SNP genotyping; genotyping-by-sequencing (GBS) and genome editing are largely dependent on computational biological tools. Bioinformatics has a critical role to play in annotation of gene function, modelling of proteins, and identification of novel metabolites, that will directly determine the success of new biotechnology tools in crop improvement programs and study of microbes, pathogens, and pests etc. Several advantages of top broadband technique comprise the gel-free handling of proteins, digestion of trypsin and use of an internal peptide, for finest quantification of protein, and thus it can be used for documentation of distinctive proteins from non-model organisms, in spite of limited genome information. Bioinformatics is the study of biological information by utilizing ideas and strategies in software engineering and statistics. It can be categorized into two classes:

1. **Management of Biological Data:** Biological data come from all fields of biology and in many formats. With the rapid advances of various high-throughput technologies, large amount of data has been generated using sequencing (nucleic acid and protein), microarray technology and macromolecule structural determination approaches, especially in efforts to understand and treat human diseases. The amount of biological data is exploding, both in size and in complexity, and to fully exploit the data, increasingly sophisticated computational techniques, efficient means for storing, searching and retrieving data, and powerful algorithms and statistical tools are required. Profacgen helps the customers to handle all sorts of data (e.g., microarray, proteomics and next-generation sequencing data) using appropriate data-management and data-analysis methods, and endeavours to transform raw data into biological knowledge. Our service covers the entire bioinformatics data lifecycle including managing and monitoring the intake, integrity, and use of diverse bioinformatics data types.

2. **Computational Biology:** Computational biology and bioinformatics is an interdisciplinary field that develops and applies computational methods to analyse large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or discover new biology.

Bioinformatics tools used in Agriculture

- **i) Sequence Analysis:** A biological sequence is principal biological system object at the molecular level like DNA, RNA, and protein. Genomes of *A. thaliana* (The, 2000) and rice (Goff et al., 2002) plants have been sequenced. Lotus and, a draft of genome sequences are available. Many other plants like maize, tomato *Medicago truncatula* and sorghum genome sequencing efforts are in progress (Bedell *et al*., 2005). The researcher has created expressed sequence tags (ESTs) from plants such as sorghum, wheat, cotton, beet, soybean and wheat.
- **ii) Genome Sequencing:** Sequencing technologies advances provide opportunities for processing, managing, and analyzing sequence in bioinformatics. For the sequencing of genome most common method is a shotgun. DNA pieces are randomly sheared, cloned and sequenced in parallel. There is software which can place together with the overlapping sequences which are sequenced separately (Myers, 1995). Many packages of software exist for sequence assembly (Gibbs *et al*., 2003) such as Phred/Phrap/Consed Arachne and GAP4. Package of a modular, open-source developed by TIGR has known AMOS, which can be used for assembly of the comparative genome (Pop *et al.,* 2004).
- **iii) Gene Finding and Genome Annotation:** Gene finding refers to introns and exons prediction in a DNA sequence segment. A Computer programs Dozens are available for identification of protein-coding genes Genie Many new tools of gene-finding are tailored for plant genomic sequences applications (Schlueter *et al*., 2003). Prediction of Ab initio gene remains a challenging problem for large size eukaryotic genomes. A typical gene of A. thaliana with five exons, it is expected that at least one exon have at least one of its borders predicted incorrectly by the ab initio approach (Brendel and Zhu, 2002).

Transcript evidence from full-length cDNA or EST sequences or similarity to homologs protein potential can reduce gene identification uncertainty significantly (Zhu *et al.*, 2003). In "structural annotation" of genomes, such techniques are widely used which refers to the features identification like genes and transposons in a sequence of the genome using ab initio algorithms and other information. For structural annotation, many software packages have been developed (Allen *et al.*, 2003). Genome comparison tools can be used to enhance gene identification accuracy like as SynBrowse annotation aspect is the repetitive DNAs, the analysis which is identical copies or nearly identical to the sequences present in the genome (Lewin, 2003). Repetitive sequences are present in any genome and abundant in most of the plant genomes (Jiang *et al.*, 2004). The identification and characterization of repeats are crucial to shed light on the evolution of genomes, function, and organization to enable filtering for homology searches of many types. Plant-specific repeats small library can be found at this is likely to grow as more genomes are sequenced substantially. To search genome repetitive sequences Repeat-Masker can use. Working from a known repeats library, Repeat Masker is built upon BLAST and can screen sequences of DNA for interspersed repeats and regions of low complexity. Repeats with poorly conserved patterns or short sequences are hard to identify due to the limitations of BLAST using Repeat- Masker. Various algorithms were developed to identify different repeats some widely used tools such as RECON.

iv) **Sequence Comparison:** Comparison of sequences can be vital to provide a foundation for many tools of bioinformatics and may allow genes and genomes, structure, and evolution function. For example, comparison of sequence provides a basis for a consensus gene model building like UniGene (Blueggel *et al*., 2004). For identification of homology, many computational methods have been developed (Wan and Xu, 2005). Comparison of the sequence is highly useful; it is similarity based sequence between two text strings, which may not correspond to homology especially when the result confidence level of a comparison is small. Comparison of sequence methods can be mainly grouped into pair-wise, profile sequence and profile-profile comparison. Among researchers for pair-wise comparison of the sequence are BLAST is popular. To evaluate the level of confidence for an alignment to represent a homologous relationship, a statistical measure (Expectation Value) was integrated into pair-wise sequence alignments (Karlin and Altschul, 1990). Pair-wise sequence alignment missed remote homologous relationships due to its insensitivity. For detecting remote homologs, sequence-profile alignment is more sensitive. A profile of protein sequence is generated by a closely related proteins group of multiple sequence alignment. A multiple sequence alignment builds correspondence among residues across all of the sequences simultaneously; sequences show the functional and structural relationship where it aligned in different positions. A profile of sequence is calculated using the occurrence probability for each amino acid at each of alignment. A famous example of a sequence-profile alignment tool is PSI-BLAST. Proteomics is the main innovation for the qualitative and quantitative proteins characterization and their interactions on a genome scale. The proteomics targets large-scale identification and all protein types' quantification in a cell or tissue, post-translational modification analysis and association with other proteins, and protein activities characterization and structures.

v) **Ontologies Applications:** Ontology is a set of vocabulary terms whose relations and meanings with other terms are stated explicitly and which are used to annotate data (Ashburner *et al.*, 2000). Ontologies are used for description of gene and protein function (Harris, 2004), types of cell (Bard *et al.*, 2005), organisms anatomies and stages of development (Garcia-Hernandez, 2002), metabolic pathways (Mao *et al*., 2005) and microarray experiments (Stoeckert *et al*., 2002). Toannotate data ontologies are used such as sequences, experiments and strains cluster of gene expression.

vi) **Biological Databases:** There are three types of biological databases that have been established and developed: community-specific databases, large-scale public repositories and project-specific databases. Nucleic Acids Research publishes an issue of the database in every year January, and Plant Physiology has started publishing databases describing articles (Rhee and Crosby, 2005). Government agencies or international consortia developed and maintained large-scale public repositories and places for long-term data storage. Examples include sequences Gene Bank (Wheeler *et al*., 2005), UniProt (Schneider *et al*., 2005) for information on protein, Protein Data Bank (PDB) (Deshpande *et al*., 2005), for structure information of protein and Array Express (Parkinson et al., 2005) and Gene Expression Omnibus (GEO) (Edgar *et al*., 2002) microarray data. There are many community-specific databases, which typically contain high standards information and address the particular researchers' community needs. Prominent community-specific databases are an example of those that Cater to researchers focused on model organisms study (Lawrence et al., 2005) or clade-oriented comparative databases (Gonzales *et al*., 2005). Databases focused on specific types of data such as metabolism (Zhang *et al*., 2005) modification of protein (Tchieu *et al*., 2003) are examples of community-specific databases. The community-specific databases concept is subject to change as researchers are widening their research scope. For example, databases focused on genome sequences comparing have recently emerged. Smaller-scale and short-lived are the third category of databases that are developed for management of data project during the funding period. These databases and web resources are reassured through the project funding period, and currently, there is no depositing or archiving standard way of these databases after the period of funding. Some issues are observed in database management of the database. The major aim of the projects is to a generic organism database tool kit to allow researchers to a genome database "off the shelf" set up. There is a general infrastructure for supporting, managing and using digital data archived in databases and websites in the long term (Lord and Macdonald, 2003). Several projects are building systems of the digital repository that can be models for a repository such as DSpace and the CalTech Open Digital Archives Collection. Some additional challenges in long-term data archiving were articulated in a recent National Science Board report.

Recent Advancement

- ✓ **Genome Editing:** Tools like CRISPR-Cas9 have enabled precise genetic modifications in crops, leading to the development of novel traits and improved crop varieties.
- ✓ **Micro biome Analysis:** Analyzing the plant micro biome has led to improved soil health and sustainable farming practices.
- ✓ **Phenomics:** High-throughput phenotyping tools provide insights into plant traits, aiding in the selection of superior crop varieties.
- ✓ **Gene Expression Analysis:** Understanding gene expression patterns helps in optimizing crop development and response to environmental factors.

Future Aspects

Bioinformatics plays an important role in agriculture science. As the data amount grows exponentially, there is a parallel growth in tools and methods demand in visualization, integration, analysis, prediction and management of data. At the same time, many researchers in the field of plant sciences are unfamiliar with available methods, databases, and tools of bioinformatics which could lead to missed information opportunities or misinterpretation.

- ✓ **Precision Agriculture:** Bioinformatics will enable even more precise and data-driven decision-making in agriculture, leading to optimal resource use and minimal environmental impact.
- ✓ **Multi-Omics Integration:** Combining genomics, transcriptomics, proteomics, and metabolomics data will provide a holistic view of plant biology, enhancing crop development.
- ✓ **Biological Pest Control:** Bioinformatics will support the development of eco-friendly pest control methods using natural predators and biological agents.
- ✓ **Customized Nutritional Crops:** Bioinformatics will help design crops with specific nutritional profiles to combat malnutrition and enhance food security.
- ✓ **Global Collaboration:** Collaboration among researchers worldwide will accelerate progress in agriculture by sharing genomic data and best practices.

References

Ash burner, M., Ball, C., Blake, J., Botstein, D., Butler, H., 2000. Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat. Genet*. 25:25–29.

Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids. Res.*, 30:207–10.

Gonzales, M.D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., 2005. The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res.*, 33:660–65.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., 32: 258–261.

Lister, R., Gregory, B.D., Ecker, J.R. 2009. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr.Opin. Plant Biol*., 12: 107–118.