Role of Bioinformatics in Agriculture

M. N. V. Prasad Gajula[1*], Anuj Kumar[2], E. A. Siddiq[1] and A. K. Polumetla[1]

[1]Institute of Biotechnology, PJTSAU, Rajendra Nagar, Hyderabad (500 030), India
[2]Molecular Biology Laboratory, Ch.Charan Singh University, Meerut (200 005), India

Corresponding Author

M. N. V. Prasad Gajula
e-mail: gajula.ibt@gmail.com

Article History

Article ID: IJEP101
Received in 5th April, 2016
Received in revised form 26th April, 2016
Accepted in final form 14th May, 2016

Abstract

Bioinformatics is a practical discipline of science that employs a wide range of computational techniques including sequence analysis, data mining, gene finding, phylogenetic tree construction, prediction of protein structure and function, and interaction networks are only a few to mention. Bioinformatics is the use of computer technology, mathematical algorithms, and statistics with concepts in the life sciences to solve biological problems. The application of bioinformatics is not just limited to any particular research domain in biology. In a developing country like India, bioinformatics has a key role to play in areas like agriculture where it can be used for increasing the volume of the agricultural produce, increasing the nutritional content, and implanting disease resistance etc. The first is in terms of 'scientific' merit, where bioinformatics play a key role in the development of the underlying computational concepts and models to convert complex biological data into useful biological and chemical knowledge. The second aspect is of 'technological' accountability to manage and integrate huge amounts of heterogeneous data sources from high throughput experimentation. A most important task for bioinformatics is to make sense of the enormous quantities of sequence data as well as structural data that are generated by genome-sequencing projects, proteomics projects, and other large-scale molecular biology efforts.

Keywords: Agriculture, bioinformatics, genomics domain, proteomics, transcriptomics

1. Introduction

Bioinformatics is the use of computer technology, mathematical algorithms, and statistics with concepts in the life sciences to solve biological problems (Kothekar, 2004; Bal, 2005; Rastogi, 2008; Pevsner, 2009). The application of bioinformatics is not just limited to any particular research domain in biology. In a developing country likeIndia, bioinformatics has a key role to play in areas like agriculture where it can be used forincreasing the volume of the agricultural produce, increasing the nutritional content, and implanting disease resistance etc. (Jayaram, 2012).

The 'concept of similarity' made bioinformatics as an important aid to advance in agricultural research. Evolution has operated on every sequence that we see today and the genomes of plants remained conserved. The conserved genes that encode proteins and sequences involved in gene regulation were not very easily understood. We know similar sequences have similar functions, however, how the sequences that encode useful functions from one organism to another getting transferred are tricky to understand. Bioinformatics provide algorithms for comparing sequences, finding similar regions, finding genes and determine gene functions, regulations and much more to understand how the genes and entire genome evolved over time (Edward, 2000).

2. The Application of Bioinformatics can be Seen in Two Broader Aspects

The first is in terms of 'scientific'merit, where bioinformatics play a key role in the development of the underlying computational concepts and models to convert complex biological data into useful biological and chemical knowledge. The second aspect is of 'technological'accountability to manage and integrate huge amounts of heterogeneous data sources from high throughput experimentation.

However, the role of bioinformatics always starts with a biological question posed by either biologists or experimentalists. Figure1 shows the schematic workflow that is commonly adopted in most of the bioinformatics application scenario.

In the field of agriculture, bioinformatics can be applied in genome sequencing studies, phylogenetic analysis and in the detection of transcription factor binding sites of genes etc. (Garima et al., 2014). Whereas in the field of proteomics,

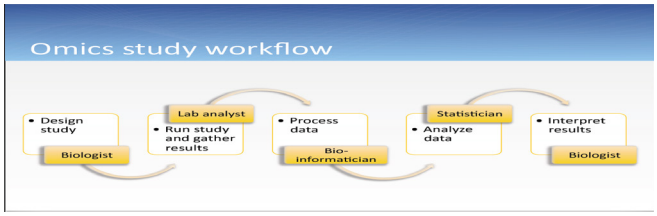

Figure 1: Workflow in bioinformatics applications

the application of bioinformatics tools is extended to find Gene expression profiling, and to determine 3D structures of protein families, predicting protein structures and functions based on various in-silico methods (Kumar, 2013; Gajulaand Steinhoff, 2013). New algorithms were being developed to find functional diversity in closely related proteins, and to analyse two-dimensional gel images etc. Involvement of bioinformatics tools is obvious in the case of metabolite identification studies.

However the ultimate goal of application of bioinformatics in the field of agriculture is to improve the quality and quantity of crop production. The process is multi-fold; The major steps are highlighted below (MNVP, 2014).

- Retrieve the data from experiments or any other sources.
- Convert data into knowledge
- Generate new hypotheses
- Archive the information
- Design new models and new experiments

2.1. Convert complex biological data into useful biological and chemical knowledge

A most important task for bioinformatics is to make sense of the enormous quantities of sequence data as well as structural data that are generated by genome-sequencing projects, proteomics projects, and other large-scale molecular biology efforts. (Pevsner, 2009)

2.2. Bioinformatics, in general, deals with the following important biological data

Sequences of DNA, RNA, and protein - The sequence of nucleotides in DNA or RNA, and the sequence of amino acids in a protein, can be obtained through various laboratory sequencing techniques with the help of high performance computers and tools.

Molecular Structures – High resolution structures of molecules can be obtained by combining thermodynamic data and computer modelling.

Expression Data – By using microarray technology, it is possible to determine the gene expression in certain cell types and in specific environmental conditions. Bioinformatics techniques can then be used to capture, store and analyse the data for further analysis (Prasad, 2014).

The applicability of bioinformatics in each 'omics' domain is reported by Gajula et al (MNVP, 2014; Gajula, 2012, Desai, 2014). Although the application of bioinformatics is extended to all major-omic and post genomic technologies, its focus and strengths still remain in the analysis of DNA sequences, genomes, and protein structural analysis (Babu et al., 2013; Huber, 2007).

3. Bioinformatics in Genomics domain

Ever since the completion of human genome project, the ability to produce large amounts of sequence information is growing at an ever decreasing cost. While the cost of the whole genome sequencing of the human genome project was ended around 2.7 billion USD, the modern NGS technologies projected to bring it down to the lowest of 1000USDas was announced by Illumina (Mary Ann Liebert, 2014). It doesn't make sense to argue that we should have waited till date to complete the human genome project to minimize the cost. However the completion of human genome project paved the way for latest NGS technologies and further high through put data analysis programs.

In the field of crop sciences still the EST sequences are playing a key role in identification of genes. However, reduction in the cost of sequencing enabled the scientific community to move towards whole-genome sequencing. Completion of Arabidopsis thaliana genome sequence and rice genome sequence (*Oryza sativa sp*. Japonica Nipponbare) and tomato genome sequence (*Solanum lycopersicum*) are major break throughs in plant genomics. However, the role of bioinformatics has become crucial with the availability of complete genome sequences and floods of sequence data. Without bioinformatics tools it is almost impossible to organise and analyze these huge amounts of data. The bioinformatics tools allowed us to find similarities at the genomics level between tomato and other important crop series. For example, DaniZamir et al. reported that a high quality genome sequence of domesticated tomato and a draft sequence of its closest wild relative, *Solanum pimpinellifolium* were compared to each other and also to the potato genome (*Solanum tuberosum*) (Dani Zamir). The similarity studies shows that the two tomato genomes differed by only 0.6% nucleotide divergence and clear signs of recent admixture. However, the comparison with potato shows more than 8% divergence, with nine large and several smaller inversions in the genome. The studies revealed that the Solanum lineage has undergone two consecutive genome triplications: one that is ancient and shared with rosids, and a more recent one. These triplications set the stage for the neo-functionalization of genes controlling fruit characteristics, such as colour and fleshiness (Giovanni Giuliano).

Arabidopsis also became a model system for comparison studies (Arabidopsis consortium). When Arabidopsis sequences were compared with tomato, soybean and potato, the later three small RNAs map predominantly shows gene-rich chromosomal regions, including gene promoters, in contrast to Arabidopsis sequence. When compared to the

genomes of Arabidopsis and Sorghum (Paterson, 2005), tomato has fewer high-copy, full-length long terminal repeat (LTR) retrotransposons with older average insertion ages (2.8 versus 0.8 million years (Myr) ago) and fewer high-frequency k-mers (DaniZamir).

Bioinformatics tools also enable the researchers to annotate sequences and to mine complex biological data to extract valuable biological knowledge. However, the applicability of bioinformatics of plant genomics can be seen in five major perspectives (Pevsner, 2008).

• Catalogue genomic information to find out the basic features of genome

• Catalogue comparative genomic information to find divergence, synteny, phylogeny etc.

• Understanding biological principle in order to understand the functions of the organism with respect to development and other major characteristics.

• Plant disease relevance: to understand the mechanisms by which organisms such as viruses or protozoan pathogens cause disease in plants and the plant response to defend the attacks etc.

• Finally, bioinformatics aspects to understand the overallgenomic picture by means of capturing, storing, archiving, analysing and visualising the biologically meaningful information.

4. Transcriptomics

The terms 'transcriptome' refers to the full complement of activated genes; mRNAs in a particular tissue at a particular instance of time. The application of bioinformatics in the field of transcriptomics aided in adding extra dimension to current genomic data by helping researchers to determine more accurately quantitative levels of gene expression and in genome-wide association studies. With the advances in new sequencing technologies and various bioinformatics tools now it became possible to identify novel genes or to assess gene expression in uncharacterized plants. The rapid implementation of microarrays has been followed by a growth in the bioinformatics of microarray data analysis (Moreau et al., 2003; Goodman et al., 2002; Prasad, 2014)

5. Proteomics

Proteomics involves identification, characterisation, and quantification of proteins in cells, and tissues. The applications of bioinformatics in the field of proteomics extended to amino acid sequence analaysis, determination of splice variant, polymorphism, and post translationalmodifications, identification of protein binding partners etc. Two-dimensional gel electrophoresis, mass spectrometry and protein microarrays are the major technologies in in the field of proteomics and bioinformatics tools plays crucial role in interpreting and getting meaningful information from the data coming out of these respective instruments.

Bioinformatics tools playing major role in protein structure prediction as well in relation to their sequence while establishing a link between the genome and the proteome. This application is very much important for plant biotechnology research especially to understand the relation between structure and function of the proteins in a plant cell.

With the development of more accurate algorithms for predicting protein structures it is possible to translate complete-genome DNA sequence data into protein structures and predict corresponding functions: such an advancement can provide the vital link between the genetics of an organism and its expressed phenotype. A comparison of the numbers of current plant protein sequences with predicted structures suggests that there is much scope for research in this area. The potential for bioinformatics to structure and integrate -omic data relies on an ability to model both the proteome and its interactions(Edwards &Batly, 2004). There are several successful case studies being reported on protein structure prediction, modelling and simulation that helps in understanding protein functional mechanism and protein interactions. (Goswami-Giri, 2014; Gajula, 2014; Kumar, 2013; Bharadwaj, 2013)

6. Metabolomics

Metabolomics deals with the analysis (typically high throughput or broad scale) of small-molecule metabolites and polymers. At the application level, metabolomics involves identification and characterization of a broad range of metabolites through reference to quantitative biochemical analysis. The importance lies in quantitative identification of metabolites that are direct gauge of desired phenotype (Gajula, 2012). The bioinformatics tools are essential at each step from screening to saving the data as the metabolites ultimately represent the dynamics of a cell. The challenge for bioinformatics will be the structuring and integration of these diverse types of data to study the 'system as a whole' approach in a cell, and thus to simulate the cellular interactions.

7. Limitations of Bioinformatics

• The results are as good as the data

• Errors in sequences

• Hypothesis independent

• Bioinformatics does not replace traditional hypothesis-driven approaches

• It complements and identifies new questions

• Integrate gene expression and protein functions in the cell

• Analysis at the level of systems: systems biology

• Description of a cell as a mathematical model

• Predictive value

8. Conclusion

Bioinformatics employs a wide range of computational techniques including sequence analysis, data mining, gene finding, phylogenetic tree construction, prediction of protein structure and function, and interaction networks are only a few to mention. The emphasis is on approaches integrating a variety of computational methods and heterogeneous data sources. However, the objective of this paper was to give an overview on the application of bioinformatics in the field of agriculture in general. The potential applications and data management are described elsewhere.

9. References

Arabidopsis Consortium, 2000. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815

Bal, H.P., 2007. Bioinformatics principles and applications (Eds.) Tata McGraw-Hill Publications Ltd.

Baxevanis, A.D., Fancis Ouellette, D.F., (Eds.) 2001. Bioinformatics: A practical guide to the analysis of genes and proteins. 2nd ed., Wiley-Interscience, New York.

Bharadwaj, Gajula, M.N.V.P., Soni, G., Babu, G., Rai, A., 2013. Molecular interaction studies of shrimp antiviral protein, PmAV with WSSV RING finger domain in-silico. Journal of Applied Bioinformatics Computer Biology 2 (1).

Borovykh, S., Ceola, Gajula, P., Gast, P., Steinhoff, H.J., Huber, M., 2007. Distance between a native cofactor and a spin label in the reaction centre of Rhodobactersphaeroides by a two-frequency pulsed electron paramagnetic resonance method and molecular dynamics simulations. Journal of Magnetic Resonance 180 (2), 178–185.

DaniZamir, 2012. The tomato genome sequence provides insights into fleshy fruit Evolution. Nature 485, 635–641.

Desai, S., Gajula, M.N.V.P.S., Srivastava, R., 2014. Implementation of multi-omics platform on a national grid for bioinformatics. BIOINFO Computer Engineering 6 (1), 515–518.

Edward N. Trifonov, 2000. Earliest pages of bioinformatics. Bioinformatics, 16(1), 5–9.

Edwards, Batley, 2004. Plant Bioinformatics: from genome to phenome. Trends in Biotechnology 22.

Gajula, M.N.V.P., 2012. Its Time to Integrate Multi Omics Data to understand Real Biology. International Journal of Systems, Algorithms & Applications 2 (ijsaa), 31–34.

Gajula, M.N.V.P., 2013. Molecular Modeling. E-book on Advances in Statistical Genetics, IASR.

Gajula, M.N.V.P., Vogel, K.P., Steinhoff, H.J., 2013. How far in-silico computing meets real experiments. A study on the structure and dynamics of spin labeledvinculin tail protein by molecular dynamics simulations and EPR spectroscopy. BMC Genomics 14(Suppl 2), S4.

GarimaSoni, S., Vanisree, M.B., Dastagiri, Gajula, M.N.V.P., 2014. Outlook on Application of Bioinformatics in Agriculture. World Research Journal of Bioinformatics 2(1), 33–40.

Giovanni Giuliano, 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635–641.

Goodman, 2002. Biologcal data becomes computer literate : new advances in bioinformatics. Current opinion in Biotechnology 13, 68–71.

Goswami-Giri, A.S., Oza, R., 2014. Bioinformatics Overview of lantana camara, an Environmental Weed. Research Jornal of Pharmaceutical Biology and Chemical Sciences 5(2), 1712.

Huber, M., Gajula, P., Milikisyants, S., Steinhoff, H.J., 2007. A short note on orientation selection in the DEER experiments on a native cofactor and a spin label in the reaction center of Rhodobactersphaeroides. Applied Magnetic Resonance 31(1), 99–104.

Jayaram, 2012. Bioinformatics For Better Tomorrow. Available from http//www.scfbio-iitd.org

Kothekar, V., 2004, Introduction to Bioinformatics, Dhruv Publications, Vol 1.

Kumar, A., Mishra, D.C., Rai, A., Gajula, M.N.V.P., 2013. In silico analysis of protein-protein interaction between resistance and virulence protein during leaf rust disease in wheat (*Triticum aestivum* L.). World Research Journal of Peptide and Protein 2(1), 52–58.

Mary Ann Liebert, Inc 2014., "Illumina Sequencers Enables $1000 Genome", Genetic Engineering & Biotechnology news Vol. 34 , No.4, 18

MNVP Gajula, 2014. A novel approach to study protein dynamics by EPR and MD simulations.scholars-press, Germany.

Moreau, Y.S., 2003, Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends in Genetics 19, 570–577.

Paterson, A. H., 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457, 551–556.

Pevsner, 2009. Bioinformatics and functional genomics, 2nd ed., john Wiley & Sons, Inc, USA.

Prasad, 2014. Microarray data exchange: An initiative. LAP LAMBERT Academic Publishing, Germany, 1–78.

Rastogi, 2008. Bioinformatics Methods and Applications. 3rd ed. PHI learning pvt.Ltd.