Bioinformatics: Concepts and Applications

Chapter · December 2019

2 authors, including:

Shailesh Kumar Singh
ITM University
97 PUBLICATIONS   **189** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   INM for Indian Goose Berry + Amorphophallus cropping system View project

Project   Post Harvest View project

Chapter 12
Bioinformatics: Concepts and Applications

Sonam[1] and Shailesh Kumar Singh[2]*
[1]Ph.D. Student, Department of Horticulture, Lovely Professional University, Punjab, 144411
[2]Associate Professor, Department of Horticulture, Lovely Professional University, Punjab, 144411`
*Corresponding Author: lpushailesh@gmail.com

Abstract: Bioinformatics is an approach which includes creation of database, development of algorithm or software and data handling for analysis and interpretation. The plant bioinformatics has potential to bring magnificent change in crop productivity specially when genetic engineering is used in conjugation with omics approaches. Genome sequencing of voluminous plant species has resulted creation of giant databases which can be used for genetic modification and reorganisation for development of noble products. This has further opened the concepts of genomics, proteomics and metabolomics which are the current field of study and are being progressively explored and applied. Thus, bioinformatics can provide greater scope for crop improvement and agricultural sustainability.

Keywords: Agriculture, Bioinformatics, Crop improvement, Genome, Genomics, Proteomics.

Introduction

Bioinformatics is an interdisciplinary approach integrating information technology with biological sciences. Bioinformatics is considered as a tool for interpretation of biological data by using information technology and computer science (Raza, 2012)**.** Plant Bioinformatics deals with application of computational study in various aspects of plant species. Bioinformatics is being used for various aspects of crop improvement in agriculture which can be varietal information system, plant genetic resources data base, studies on genomics, studies on proteomics, studies on metabolomics, studies on plant modelling, pedigree analysis, biometrical analysis, forecasting models etc. The application of bioinformatics and its components are illustrated in Figure-1. In field of agriculture, bioinformatics is useful tool for identification of gene of importance, study of metabolic pathways of various biological processes to understand the molecular movement, gene transfer for creating trait specific varieties, understanding mode of actions of pathogen and chemicals, identifying suitable techniques for plant protection measures, climate resilience agriculture, variable rate application of agricultural inputs, phylogeny prediction, germplasm characterization and conservation, many more and ultimately the crop productivity. Although, exploitation of bioinformatics and its branches for crop improvement is very low its application in allied sciences like physical, chemical, biological and medical sciences confirms the possible scope in agriculture as well.
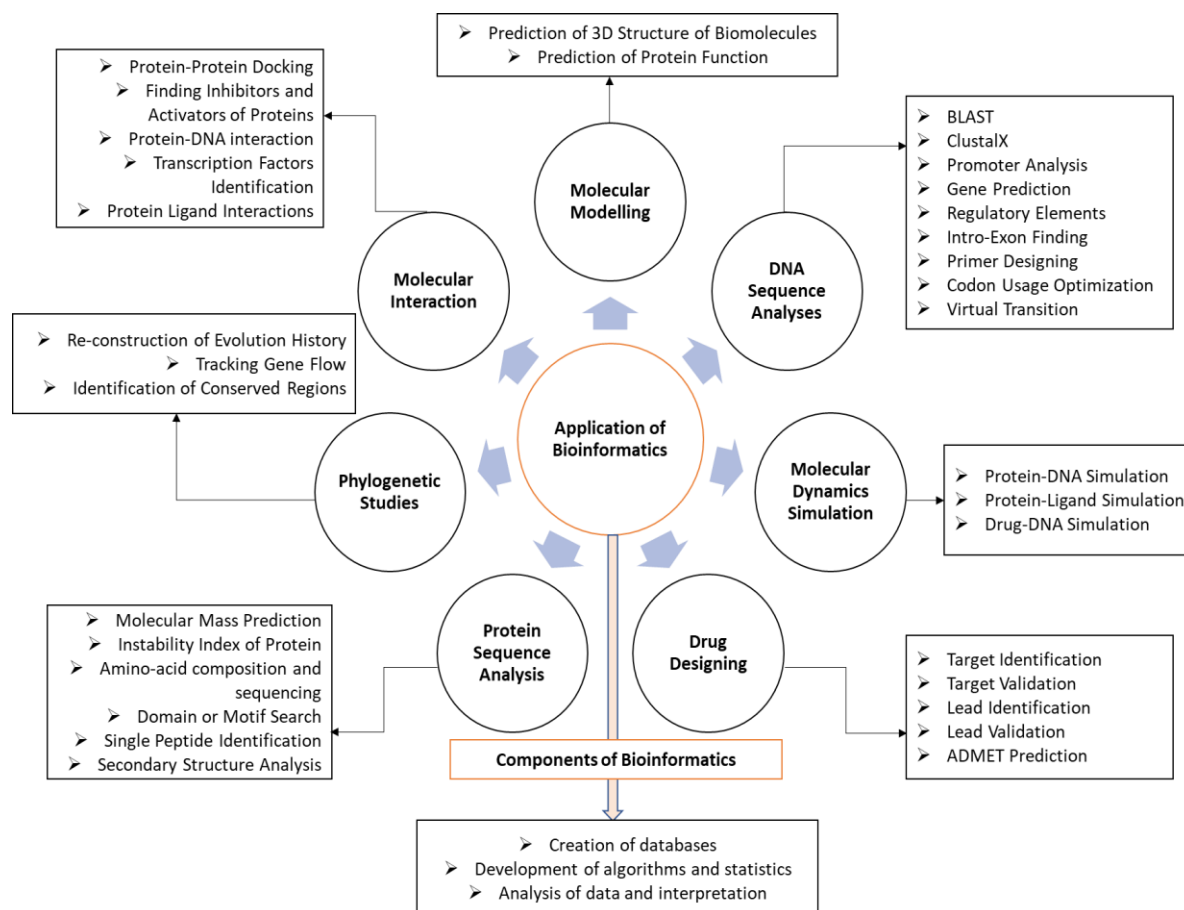
Fig. 1: Applications of bioinformatics in crop improvement

Applications of Bioinformatics in Crop Improvement

There are number of applications of bioinformatics viz. sequence analysis and alignment, molecular modelling, docking, annotation and dynamic simulation to accelerate the biotechnological research (Khan, 2018). It covers many areas of biological sciences especially of modern biology viz. genomics, transcriptomics, proteomics, genetics, evolution genomics, cheminformatics, climate change studies, drug discovery and development, waste clean-up, bioenergy, crop improvement, veterinary sciences, forensic sciences and biodefense. Some important applications of bioinformatics in improvement of horticulture are described below:

Varietal Information System: Bioinformatics has useful applications in developing varietal information system. Variety refers to a genotype which has been released for commercial cultivation. State Variety Release Committee or Central Variety Release Committee and notified by the Government of India. Various types of varieties are used in plant breeding. All such terms are defined below: The detailed information about various type of varieties can be developed using highly heritable characters. Such information can be used in various ways as given below:

(i) In DUS testing for varietal identification

(ii) In grouping of varieties on the basis of various highly heritable characters.

(iii) In sorting out of cultivars for use in Pre-breeding and traditional breeding. The information can be stored in the computer memory and be retrieved as and when required.

PGR Data Base: Genetic material of plant which is used as resource by people of different generations is considered as plant genetic resource. It is also known as gene pool, genetic stock and germplasm. The germplasm is evaluated for several characters such as highly heritable morphological and other characters as given below:

- ➢ Highly heritable morphological traits
- ➢ Yield contributing traits
- ➢ Quality characters
- ➢ Resistance to biotic and abiotic stresses
- ➢ Characters of agronomic value.

International Plant Genetic Resources Institute (IPGRI), Rome, Italy has developed descriptors and descriptor states for various crop plants. Such descriptors help in uniform recording of observations on germplasm of crop plants throughout the world. Thus, huge data is collected on crop germplasm for several years. Bioinformatics plays an important role in systematic management of this huge data. Bioinformatics is useful in handling of such data in several ways as follows:

- ➢ It maintains the data of several locations and several years in a systematic way.
- ➢ It permits addition, deletion and updating of information.
- ➢ It helps in storage and retrieval of data.
- ➢ It also helps in classification of PGR based on various criteria.
- ➢ It helps m retrieval of data belonging to specific group such as early maturity, late maturity, dwarf types, tall types, resistant to biotic stresses, resistant to abiotic stresses, genotypes with superior quality, genotypes with marker genes, etc.

All such data can be easily managed by computer aided programmes and can be manipulated to get meaningful results.

Genomics: Genomics aims at the collective characterization and quantification of genes and includes the sequencing, mapping, editing and analysis of genomes to understand mechanism and path of evolution of an individual species, to assemble and analyse the structure and function of entire set of genomes, and to decipher mechanism of gene expression (Culver and Labow, 2002).

The intragenomic phenomena such as epistasis, pleiotropy, heterosis and other interactions between loci and alleles within the genome (Pevsner, 2009). The term genomics was coined by Tom Roderick in 1986 (Yadav, 2007). Plenty of data are generated during gene sequencing, analysis of structure and function of entire genomes which cannot be managed through simple statistical analysis so use of computational algorithms will be helpful to manage this vast enormous data. Thus, progressive genome sequencing of new organism the database is being increased which are helpful to provide both conceptual bases and practical approaches for detecting systemic functional behaviours of the cell and the organism (Nakagami *et al.*, 2010; Greene *et al.*, 2014).

The study of structure and function of entire genome of an organism is referred to as genomics. Thus, genomics consists of two groups, viz: (i) Structural genomics and (ii) Functional genomics. Bioinformatics plays a vital role in both of the areas, structural genomics and functional genomics. In structural genomics, the genetic structure of each chromosome of the genome or 3-dimensional structure of all macromolecular represented

in a genome (Jennifer, 2000) while functional genomics refers to transcription of RNAs from a genome and encoding of proteins by a genome.

Structural genomics emphasizes high throughput determination of 3D protein structures (Goldsmith-Fischman and Honig, 2003), determining RNA structures (Jennifer, 2000) and gene mapping (Cheema and Dicks, 2009). The structural database obtained from structural genomics is having potential to be useful for deciphering biological functions of these macromolecules which is referred as functional genomics (Jennifer, 2000).

Studies on Proteomics: Proteomics refers to the study of structures and functions of all proteins in an individual. In other words, it deals with the study of entire protein expression in an organism. Advanced techniques in biology led to accumulate enormous data of protein-protein interactions, protein profiles, protein activity pattern and organelles compositions. This vast data can be managed and access easily by using bioinformatics tools, software and databases. Till now, a bulk of algorithms in the field of proteomics viz. image analysis of 2D gels, peptide mass fingerprinting and peptide fragmentation fingerprinting have been developed (Gholizadeh and Kohnehrouz, 2010; Gholizadeh, 2011; Gholizadeh and Kohnehrouz, 2013).

Proteomics is of two types, viz: (i) Structural proteomics and (ii) Functional proteomics. The Structural Proteomics refers to the study of the structures of all proteins found in a living organism. The Functional Proteomics deals with functions of all proteins found in a living organism. In fact, proteomics is a new sub-discipline of functional genomics. It is the study of proteomes which refer to complete set of proteins encoded by a genome. A variety of techniques are used for the study of proteomics. Now computer aided programmes are available for the study of proteomics.

Studies on Metabolomics: Metabolomics refers to the study of all metabolic pathways in a living organism. In other words, it is the computer aided information of all metabolic pathways of a living organism. Main points related to metabolomics are listed below:

➢ It deals with the study of all metabolic pathways in a living organism.
➢ It is computer-based information about metabolic pathways in a living organism.
➢ It helps in identification and correction of metabolic disorders in an organism.
➢ It helps in selection of individuals with normal metabolic pathways.
➢ It helps early detection of genetic disorders associated with metabolic pathways.

Modelling of Plants: Bioinformatics plays an important role in modelling of crop plants. Such computer aided studies have already been made in field pea and several other field crops. First the plant model is conceptualized using various plant traits and then efforts are made to develop such model by using appropriate breeding procedures.

First donor sources for these traits are identified from the available germplasm. Then efforts are made to combine these traits in one genotype particularly in a popular variety. Such computer-based studies help in developing plant ideotype suitable for machine picking and used in multiple cropping system.

Pedigree Analysis: Computer aided studies are useful in pedigree analysis of various cultivars and hybrids. Information about the parentage of cultivars and hybrids is entered into the computer memory which can be retrieved any time. The list of parents that are

common in the pedigree of various cultivars and hybrids can be sorted out easily. It helps in the pedigree analysis which in turn can be used in planning plant breeding programmes especially in the selection of parents for use in hybridization programmes. Through study of protein structures, it helps in pedigree analysis.

Biometrical Analysis: In plant breeding and genetics, various types of biometrical analyses such as correlation, path coefficient, discriminant function, diallel, partial diallel, triallel, quadriallel, generation means, line x tester, triple test cross, stability parameters, D statistics, metroglyph etc. are carried out. Computer aided programmes are very much useful in carrying out such biometrical analyses. The information obtained from such biometrical analysis is used in better planning of plant breeding programmes for achieving specific goal.

Forecasting Models: Computer aided programmes have wide applications in developing various types of forecasting models especially useful for predicting crop production and productivity and in forecasting incidence of insects and diseases in crop plants. Weather parameters are used in making such predictions. Computer aided remote sensing techniques are used for such predictions.

Other Applications: Besides agricultural applications, bioinformatics has several other useful applications. Such applications include use of bioinformatics in: (i) Medical science, (ii) Forensic science, (iii) Pharmaceutical and biotech industry.

In medical science, computer aided studies are useful in detection of genetic diseases at an early stage of life. It can help in cure of genetic diseases in some cases. The pedigree analysis helps in advising future parents to prevent certain genetic diseases. In forensic science, bioinformatics is useful in settling disputed cases of children and detecting criminal cases. In pharmaceutical industry, computer aided programmes help in detecting various metabolic pathways involved in the production of a medicine. Thus, it can help in mass production of such chemicals.

> *Drug Discovery***:** Bioinformatics is playing an important role in drug discovery, drug assessment and drug development through its utility for prediction, analysis and interpretation in clinical and preclinical findings (Livak and Schmittgen, 2001). Tradition approach of drug discovery was time consuming and cumbersome practice while the bioinformatics-based approach i.e. computer-aided drug design (CADD) has made the task rapid, cost effective and easier to handle the increasing and voluminous demand of drugs with low risk. Bioinformatics provides a range of drug-related databases and algorithms which are used for drug designing and drug development (Rabbani *et al*., 2003).
>
> *Transcriptomics:* The study of sets of all messenger RNA molecules in the cell is called as transcriptomics. This can also be called as Expression Profiling where DNA microarray is used to determine the expression level of mRNA in a given cell population. The microarray technique generates vast amount of data, single run generates thousands of data value and one experiment requires hundreds of runs. Analysis of such vast data is done by numerous software packages. In this way, bioinformatics is used for transcriptome analysis where mRNA expression levels can be determined (Gholizadeh, 2014). RNA sequencing (RNAseq) also has been included under transcriptomics. It is carried out using next generation sequencing to

determine the presence and quantity of RNA in a sample at a given time. It is used to analyze the continuously changing cellular transcriptome.

***Cheminformatics*:** Cheminformatics (chemical informatics) focuses on storing, indexing, searching, retrieving, and applying information about chemical compounds. It involves organization of chemical data in a logical form to facilitate the retrieval of chemical properties, structures and their relationships. Using bioinformatics, it is possible through computer algorithm to identify and structurally modify a natural product, to design a compound with the desired properties and to assess its therapeutic effects, theoretically. Cheminformatics analysis includes analyses such as similarity searching, clustering, QSAR modelling, virtual screening, etc. (Ausubel *et al*., 1999).

***Evolutionary Studies/Phylogenetics*:** The study of evolutionary relationship among individuals or group of organisms is defined as phylogenetics. Taxonomists find the evolutionary relationship using various anatomical methods that takes too much time. Using Bioinformatics, phylogenetic trees are constructed based on the sequence alignment using various methods. Various algorithmic methods are developed for the construction of phylogenetic tree that are used depending on the various evolutionary lineages (Shinozaki et al., 2003).

Crop Improvement: Sustainability in agricultural production in response to global climate change and population pressure is the issue which needs to be addressed on urgent basis. The integrated 'omics' approach is helpful to clarify the molecular system of the plant which are applied to improve the crop productivity. The application of comparative genomics improves the understanding the structure, biological properties and functioning of genes. The available databases are applied for design and development of new techniques and experiments for ensuring plant production (Shinozaki and Yamaguchi, 2007).

***Veterinary Science*:** Food production from livestock can meet demand of human population for food. For better bio-economy, there is a need of efficient animal production and reproduction. This is achieved with better understanding of livestock species. Current and new methods in livestock species using data from experimental or field studies with bioinformatics are helping in understanding the systems genetics of complex traits and provide biologically meaningful and accurate predictions. Finally, almost all of the next generations-omics tools and methods that are used in other fields of biological sciences, can also be used in veterinary sciences (Yamaguchi and Shinozaki, 2005; Bartels *et al*., 2005).

Forensic Science: Forensic science includes the study regarding identification and relatedness of individuals. It is inherently interdisciplinary with bioinformatics as both are dependent on computer science and statistics. This field is based on the molecular data and many databases are being developed to store the DNA profiles of known offenders. This field is being pushed due to technological and statistical advances in microarray, Bayesian networks, machine learning algorithms, TFT biosensors and others. This provides the effective way of evidence organization and inference (Seki et al., 2001).

***Biodefense*:** Biodefense includes measures to restore biosecurity to a group of organisms who are subjected to biological threats or infectious diseases. Although, there is significant development in bioinformatics application in various field but its application in forensic studies, medical intelligence and mitigation of biothreats needs special attention to develop advanced algorithms for better interoperability (Schimid et al., 2005; Valdivia-Granda, 2010).

Waste Clean-up: Today, the major concern all over the Globe is environmental pollutants. The main concern of the environmentalists is waste generated from the industries. These pollutants progressively deteriorate the environment which in turn affects human health. There are few microorganisms that are considered to remediate the pollutants into the natural biogeochemical cycle. Bioremediation is the recent technology which explores the microbial potentiality for biodegradation. This technology can be further improved by using bioinformatics. Genomic and bioinformatics data provide a wealth of information that would be greatly enhanced by structural characterisation of some protein. Bioinformatics provides data of microbial genomics, proteomics, systems biology, computational biology, and bioinformatics tools for understanding of the mechanisms of biodegradative pathways (Raza, 2012; Valdivia-Granda, 2010).

***Climate Change Studies*:** Another Global concern is the Climate change because of loss of sea ice, accelerated sea level rise and longer and more intense heat waves. To solve this issue, bioinformatics may help by way of sequencing microbial genome which can reduce levels of carbon dioxide and other greenhouse gases. This plays an important role in stabilizing the global climate change. Not much work has been done in this area in bioinformatics domain, and more region-specific work must be conducted considering microbes of that region and their capability in CO_2 reduction (Khan, 2018).

***Bioenergy/Biofuels*:** Biofuels has greater potential to serve as alternative and renewable energy source. Bioinformatics is important in understanding and analysis of biofuel producing pathways. Recent progress in algal genomics, in association with various "omics" approaches, has enabled the identification of metabolic pathways and genes for development of genetically modified micro-algal strains for optimum lipid production (National Human Genome Research Institute, 2010).

Advantages of Bioinformatics

Bioinformatics has several practical applications in genetics and plant breeding as discussed above. Its main advantages in crop improvement are given below:

1. It provides systematic information about genomics, proteomics and metabolomics of living organisms. This information is useful in planning various breeding and genetical programmes.

2. It helps in finding evolutionary relationship between two species. Studies of nucleotide and protein sequences help in such matter. The closely related organisms have similar sequences and distantly related organisms have dissimilar sequence. The time of divergence between two species can also be estimated from such studies. Thus, bioinformatics helps in the study of evolutionary biology. It helps in drawing phylogenic trees (trees of relatedness).

3. Rapid Method. Is a rapid method of gene mapping and sequencing. Earlier methods of gene mapping were time consuming and pains taking. Bioinformatics has made this task very simple. Now gene hunting has become faster, cheaper and systematic.

4. Identification of similar genes. Computer aided studies help in identification of similar genes in two species. For example, genes similar for biotic and abiotic stresses in two species can be easily detected.

5. High Accuracy. The computer-based information has very high level of accuracy and is highly reliable.

6. Bioinformatics has led to advances in understanding basic biological processes which in turn have helped in diagnosis, treatment and prevention of many genetic diseases:

7. It has become possible to reconstruct genes from Expressed Sequence Tags (EST). The EST is nothing but short pieces of genes which can express.

8. Computer aided programmes have made it possible to group proteins into families based on their relatedness.

9. Computer aided programmes are useful in designing primers for PCR. Such primers can be designed with little efforts. Such primers are used to sequence unknown genes or genes of interest.

10. In life science, computer aided programmes are useful in storing, organizing and indexing huge databases.

Limitations of Bioinformatics:

Computer based programmes have helped in better understanding of various processes of life science. However, there are some limitations of bioinformatics which are listed below:

1. Bioinformatics requires sophisticated laboratory of molecular biology for in depth study of biomolecules. Establishment of such laboratories requires lot of funds.

2. Computer based study of life science requires some training about various computer programmes applicable for the study of different processes of life science. Thus, special training is required for handling of computer based biological data.

3. There should be uninterrupted electricity (power) supply for computer aided biological investigations. Interruption of power may sometimes lead to loss of huge data from the computer memory.

4. There should be regular checking of computer viruses because viruses may pose several problems such as deletion of data and corruption of the programmes.

5. The maintenance and up keeping of molecular laboratories involve lot of expenditure which sometimes becomes a limiting factor for computer based molecular studies.

Conclusion

The increasing population questions about the food and nutritional security, increasing pollution questions for environmental security, biodiversity degradation questions for conservational security and growing database questions for database security but the answer is single i.e. Bioinformatics.

References

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., ... & Schaechter, M. (1999). *Short protocols in molecular biology: a compendium of methods from current protocols in molecular biology* (No. 574.88 S559). Universidade de Sao Paulo, Piracicaba, SP (Brasil). Faculdade de Medicina Veterinaria e Zootecnia.

Bartels, D., & Sunkar, R. (2005). Drought and salt tolerance in plants. *Critical reviews in plant sciences*, 24(1), 23-58.

Cheema, J., & Dicks, J. (2009). Computational approaches and software tools for genetic linkage map estimation in plants. Briefings in bioinformatics, 10(6), 595-608.

Culver, K. W., & Labow, M. A. (2002). "Genomics". In: Robinson, R. Genetics. Macmillan Science Library. Macmillan Reference USA.

Gholizadeh, A. (2011). Heterologous expression of stress-responsive DUF538 domain containing protein and its morpho-biochemical consequences. *The protein journal*, 30(5), 351-358.

Gholizadeh, A. (2014). Prediction of tertiary structure homology between bactericidal/permeability increasing protein of innate immune system and hydrolase enzymes. *Int J Biosci*, 5(2), 1-6.

Gholizadeh, A., & Kohnehrouz, B. B. (2010). Identification of DUF538 cDNA clone from Celosia cristata expressed sequences of nonstressed and stressed leaves. *Russian journal of plant physiology*, 57(2), 247-252.

Gholizadeh, A., & Kohnehrouz, S. B. (2013). DUF538 protein super family is predicted to be the potential homologue of bactericidal/permeability-increasing protein in plant system. *The protein journal*, 32(3), 163-171.

Goldsmith-Fischman, S., & Honig, B. (2003). Structural genomics: computational methods for structure analysis. Protein Science, 12(9), 1813-1821.

Greene, C. S., Tan, J., Ung, M., Moore, J. H., & Cheng, C. (2014). Big data bioinformatics. Journal of cellular physiology, 229(12), 1896-1900.

Jennifer A. D. (2000). Structural genomics of RNA, In: nature structural biology, structural genomics supplement, pp. 954-956. Available on: http://doudnalab.org/Publications/nsb-7s-954.pdf

Khan, N. T. (2018). The Emerging Role of Bioinformatics in Biotechnology. *Journal of Biotechnology and Biomedical Science*, 1(3), 13.

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. *methods*, 25(4), 402-408.

Nakagami, H., Sugiyama, N., Mochida, K., Daudi, A., Yoshida, Y., Toyoda, T., ... & Shirasu, K. (2010). Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants. *Plant physiology*, 153(3), 1161-1174.

National Human Genome Research Institute (2010). A Brief Guide to Genomics. Genome.gov. Retrieved on 2011-12-03.

Pevsner, J. (2009). Bioinformatics and functional genomics (2nd ed.). Hoboken, N.J.: Wiley-Blackwell.

Rabbani, M. A., Maruyama, K., Abe, H., Khan, M. A., Katsura, K., Ito, Y., ... & Yamaguchi-Shinozaki, K. (2003). Monitoring expression profiles of rice genes

under cold, drought, and high-salinity stresses and abscisic acid application using cDNA microarray and RNA gel-blot analyses. *Plant physiology*, 133(4), 1755-1767.

Raza, K. (2012). Application of data mining in bioinformatics. arXiv preprint arXiv:1205.1125.

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., ... & Lohmann, J. U. (2005). A gene expression map of Arabidopsis thaliana development. *Nature genetics*, 37(5), 501.

Seki, M., Narusaka, M., Abe, H., Kasuga, M., Yamaguchi-Shinozaki, K., Carninci, P., ... & Shinozaki, K. (2001). Monitoring the expression pattern of 1300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray. *The Plant Cell*, 13(1), 61-72.

Shinozaki, K., & Yamaguchi-Shinozaki, K. (2007). Gene networks involved in drought stress response and tolerance. *J. Exp. Bot.*, 58(2), 221-227.

Shinozaki, K., Yamaguchi-Shinozaki, K., & Seki, M. (2003). Regulatory network of gene expression in the drought and cold stress responses. *Current opinion in plant biology*, 6(5), 410-417.

Valdivia-Granda, W. A. (2010). Bioinformatics for biodefense: challenges and opportunities. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, 8(1), 69-77.

Yadav, S. P. (2007). The wholeness in suffix -omics, -omes, and the word om. *Journal of Biomolecular Techniques*, 18 (5), 277.

Yamaguchi, S. K., & Shinozaki, K. (2005). Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. *Trends Plant Sci.*,10(2), 88-94.