I J B R
International Journal of Bioinformatics Research
ISSN: 0975–3087, E-ISSN: 0975–9115, Vol. 3, Issue 2, 2011, pp-221-226
Available online at http://www.bioinfo.in/contents.php?id=21

ROLE OF BIOINFORMATICS IN AGRICULTURE AND SUSTAINABLE DEVELOPMENT

SINGH V.K.[1*], SINGH A.K.[2], CHAND R.[3] AND KUSHWAHA C.[3]
[1]Centre for Bioinformatics, Faculty of Science
[2]Department of Genetics and Plant Breeding, Institute of Agricultural Sciences
[3]Department of Mycology and Plant Pathology, Institute of Agricultural Sciences
Banaras Hindu University, Varanasi- 221 005, India
Corresponding Author: Email- vinaysingh@bhu.ac.in

Received: September 29, 2011; Accepted: October 29, 2011

Abstract- Bioinformatics is an interdisciplinary area of the science composed of biology, mathematics and computer science. Bioinformatics is the application of information technology to manage biological data that helps in decoding plant genomes. During the last two decades enormous data has been generated in biological science, firstly, with the onset of sequencing the genomes of model organisms and, secondly, rapid application of high throughput experimental techniques in laboratory research. Biological research that earlier used to start in laboratories, fields and plant clinics is now starts at the computational level using computers (*In-silico*) for analysis of the data, experiment planning and hypothesis development. Bioinformatics develops algorithms and suitable data analysis tools to infer the information and make discoveries. Application of various bioinformatics tools in biological research enables storage, retrieval, analysis, annotation and visualization of results and promotes better understanding of biological system in fullness. This will help in plant health care based disease diagnosis to improve the quality of Plant.
Key words: bioinformatics, genomics, agriculture, stress, sustainable development.

Introduction: the genomic era in 21st century

The genomic era has seen a massive explosion in the amount of biological information due to huge advances in the fields of molecular biology and genomics. Bioinformatics is generally referred to as the application of computer technology to the processing and managing the data generated in biological experiments. The term bioinformatics was originally coined for the application of information technology to large volumes of biological, particularly, genomic data. The field of bioinformatics has come to be intermingled with traditional computational biology and biostatistics, which are strictly concerned not only with how to handle the information itself, but rather how to extract biological meaning from it. This new knowledge could have profound impacts on different fields, such as human health, agriculture, environment, energy and biotechnology.

Why is bioinformatics important?

The greatest challenge facing the molecular biology community today is to make sense of the wealth of data that has been produced by the genome sequencing projects. Traditionally, molecular biology research was carried out entirely at the experimental laboratory bench but the huge increase in the scale of data being produced in this genomic era has realized a need to incorporate computers into the research process. With the advent of new tools and databases in molecular biology we are now enable to carry out the research not only at genome level but also at proteome, transcriptome and metabalome levels. The challenges faced by the bioinformatics community today are the intelligent and efficient storage of huge amount of data generated, and to provide easy and reliable access to this data. Therefore, incisive computer tools must be developed to allow the extraction of meaningful biological information. Emerging trend in pharma industry is to apply bioinformatics tools to reduce time and cost in molecular marker development, drug development [1, 2].

Genomics in Agriculture

The sequencing of the genomes of plants and animals will provide enormous benefits for the agricultural community. Bioinformatics tools can be used to search for the genes within those genomes that are useful for the agricultural community and to elucidate their functions. This specific genetic knowledge could then be used to produce stronger, more drought, disease and insect resistant crops and improve the quality of livestock making them healthier, more disease resistant and more productive.

Comparative genetics of the plant genomes has shown that the organization of genes has remained more conserved over evolutionary time than was previously

believed [3, 4, 5, 6]. These findings suggest that information obtained from the model crop systems can be used to suggest improvements to other food crops. *Arabidopsis thaliana* (water cress) *Oryza sativa* (rice), *Triticum aestivum* (wheat) and *Zea mays* (Maize) are examples of available complete land plant genomes [7, 8].

What can genome sequence tell us?

Some organisms have multiple copies of chromosomes, diploid, triploid, tetraploid and so on. In classical genetics, in a sexually reproducing organism (typically eukaryotes) the gamete has half of the number of chromosome of the somatic cell and the genome is a full set of chromosomes in a gamete. The term genome can be applied specifically to mean that stored on a complete set of nuclear DNA (i.e., nuclear genome) but can also be applied to that stored within organelles that contain their own DNA, as with the 'mitochondrial genome' or the 'chloroplast genome'. Additionally, the genome can comprise non chromosomal genetic elements such as viruses, plasmids, and transposable elements.

Most biological entities that are more complex than a virus, sometimes or always, carry additional genetic material besides that which resides in their chromosomes. In such circumstances 'genome' describes all of the genes and information on non-coding DNA that have the potential to be present. In eukaryotes such as plants, protozoa and animals, however, 'genome' carries the typical connotation of only information on chromosomal DNA [9]. The genetic information contained by DNA within organelles i.e., chloroplast and/or mitochondria is not considered part of the genome. In fact, mitochondria are sometimes said to have their own genome often referred to as the 'mitochondrial genome'. The DNA found within the chloroplast may be referred to as the 'plastome'.

Comparative analysis within microbial genome using metabolic comparison and gene organization at metabolic reactions level with their operons using structure, pathway, reaction, compounds and gene orthologs gives better understanding of genome evolution [10, 11]. Variation in the genome size, GC content, codon usage and amino acid composition based on stains of the same species, closely related species and distantly related species. Colinearity between gene set showing their evolutionary differences in evolution of individual genes.

Improve nutritional quality and growth in poorer soils

Gene-Diet-Disease interaction of Nutritional genomics aims to study the susceptible genes and provide dietary interventions for individuals at risk of such diseases. Scientists have recently succeeded in transferring genes into rice to increase levels of Vitamin A, iron and other micronutrients. This work could have a profound impact in reducing occurrences of blindness and anaemia caused by deficiencies in Vitamin A and iron respectively. Scientists have inserted a gene from yeast into the tomato, and the result is a plant whose fruit stays longer on the vine and has an extended shelf life.

Bioinformatics play an important role to detect the metal from Metagenomic sequencing obtains from contaminated soil [12]. Soil arguably houses the most complex microbial communities because of its ancient history, complex sets of interrelating gradients, and protective, isolating and relatively resource poor and stable physical structure. This results in an incredibly diverse set of gene sequences; at least at the scale soils are normally sampled. The challenge is no longer sequence yield, but the analysis of those sequences, and especially so due to the short sequence products of current sequencing technologies. Progress has been made in developing cereal varieties that have a greater tolerance for soil alkalinity, free aluminum and iron toxicities.

Improvement for plant resistance against biotic and abiotic stresses

Application of insect genomics helps in the identification of resistance mechanisms and finding the novel target sites [13]. Genes from *Bacillus thuringiensis* that can control a number of serious pests have been successfully transferred to cotton, maize and potato. This new ability of the plants to resist insect attack means that the amount of insecticides being used can be reduced.

A plant's first line of defense against abiotic stress is in its roots. If the soil holding the plant is healthy and biologically diverse, the plant will have a higher chance of surviving stressful conditions. Plants are extremely sensitive to the changes, and do not generally adapt quickly. Plants also adapt very differently from one another, even from a plant living in the same area. When a group of different plant species was prompted by a variety of different stress signals, such as drought or cold, each plant responded uniquely. Hardly any of the responses were similar, even though the plants had become accustomed to exactly the same home environment. So, species are more likely to become population threatened, endangered, and even extinct, when and where abiotic stress is especially harsh. By using *in silico* genomics technology researcher can identify defense/ disease resistance gene-enzyme with their promoter region and transcription factor which help to enhance the immunity and defence mechanism [14, 15].

Similarity Searching Tools

The exponential growth of genomics is due to computational challenges of systematically collecting, storing, organizing, manipulating visualizing and analyzing large amounts of biological information come from the experiments carried out by the biologists.. Thus, bioinformatics, in its broad sense, can be seen as providing both the infrastructure and the scientific framework in which biologists take information and use computers to help convert it into knowledge [16]. Apart from the fact that bioinformatics is a newly recognized discipline; there is an impressive diversity of

bioinformatics resources currently available. Though a wide array of commercial resources exist, some of which are ideally suited to specific tasks, and freely available. Many of the databases and analysis tools we describe here are hosted by government or academic research centers and can be accessed via user-friendly web interfaces (Table .1).

An excellent resource to the world of genomic databases is the annual database issue of the journal 'Nucleic Acids Research', published on the 1st of January each year (www3.oup.co.uk/nar/database/c/). In addition, Genbank [17], National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL) and the DNA Databank of Japan (DDBJ), are pioneer as the DNA and protein sequence repository. A variety of crop and model plant specific genomic databases are also accessible through UKCropNet including GrainGenes, (which holds molecular and phenotypic information on wheat, barley, oats, rye and sugarcane), and MaizeDB (for maize).

Some databases are specific to somewhat larger taxonomic assemblages e.g., the Gramene database, which aims to integrate genomic information from among all grasses using the rice genomic sequence as a focal point. Pfam, a protein sequence signature database, is a derived database [28]. Derived databases in plant genomics frequently only include those plant systems having the most abundant data. One example is the set of Gene Indices at The Institute for Genomic Research (TIGR), which is a collection of much focussed databases, each covering a different plant, animal, protist or fungal species. Each Gene Index computationally assembles the non-redundant set of gene sequences for that organism, with links to expression, homology and other information.

Plant biologists are, of course, also interested in plant symbionts and disease causing organisms. A number of plant pathogenic bacteria and fungi have either been sequenced in their entirety, including *Agrobacterium tumefaciens*, *Ralstonia solanacearum* and *Xylella fastidiosa*, or are the subject of ongoing sequencing projects, such as *Magnaporthe grisea*, *Pseudomonas syringae* pv. tomato and *Xanthomonas campestris*. In addition, a variety of plant viral genomes have been deposited in Genbank. The Genomes Online Database (GOLD) is a regularly updated online listing of prokaryotic and eukaryotic genome projects that have been completed or that are under way. TIGR offers what it calls the Comprehensive Microbial Resource database, which allows exploration and comparison of the annotated microbial sequences.

Sequence Analysis

The term "sequence analysis" in biology implies subjecting a DNA or peptide sequence to sequence alignment, sequence databases, repeated sequence searches, or other bioinformatics methods on a computer. In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Sequence analysis can be used to assign function to genes and proteins by the study of the similarities between the compared sequences [29]. Nowadays there are many tools and techniques that provide the sequence comparisons (sequence alignment) and analyze the alignment product to understand the biology.

The most commonly used similarity search method is the Basic Local Alignment Search Tool [BLAST; 30]. BLAST is a heuristic modification of the Smith–Waterman (1981) [31] algorithm and in practice it is widely used. In BLAST, statistical methods are used to determine the likelihood of a particular alignment between sequences or sequence regions arising by chance given the size and composition of the database being searched [32, 33]. The most popular online interface to BLAST is available at NCBI, where a standalone version is also available for downloading. There are several parameters controlling the behaviour of the BLAST algorithm, and these need to be carefully considered. The set of tools allows us to carry out further, more detailed analysis on our query sequence including evolutionary analysis, identification of mutations, hydropathy regions, CpG islands and compositional biases [34]. Sequence alignments are useful in bioinformatics for identifying sequence similarity, producing phylogenetic trees [35, 36] and developing homology models of protein structures [37].

Protein Function Analysis

In protein functional studies we compare the protein sequence to the secondary (or derived) protein databases that contain information on motifs, signatures and protein domains. Highly significant hits against these different pattern databases allow us to approximate the biochemical function of our query protein [38, 39]. Motif finding, also known as profile analysis, constructs global multiple sequence alignments that attempt to align short conserved sequence motifs among the sequences in the query set. This is usually done by first constructing a general global multiple sequence alignment, after which the highly conserved regions are isolated and used to construct a set of profile matrices. The profile matrix for each conserved region is arranged like a scoring matrix but its frequency counts for each amino acid or nucleotide at each position are derived from the conserved region's character distribution rather than from a more general empirical distribution [40]. The profile matrices are then used to search other sequences for occurrences of the motif they characterize.

***In silico Cis*-acting elements and Transcription factor studies**

Eukaryotic transcription is more complex than prokaryotic transcription. In eukaryotes the genetic material (DNA), and therefore transcription, is primarily localized to the nucleus, where it is separated from the cytoplasm by the nuclear membrane. DNA is also present in mitochondria and mitochondria utilize a specialized RNA polymerase for transcription. This

allows for the temporal regulation of gene expression through the sequestration of the RNA in the nucleus, and allows for selective transport of RNAs to the cytoplasm, where the ribosomes reside.

Among eukaryotes, the core promoter of protein-encoding gene contains binding sites for the basal transcription complex and RNA polymerase II, and is normally within about 50 bases upstream of the transcription initiation site. Further transcriptional regulation is provided by upstream control elements (UCEs), usually present within about 200 bases upstream of the initiation site [22]. The core promoter for Pol II sometimes contains a TATA box, the highly conserved DNA recognition sequence for the TATA box binding protein (TBP) whose binding initiates transcription complex assembly at the promoter [23, 18]. Some genes also have enhancer elements that can be thousands of bases upstream or downstream of the transcription initiation site [41]. Combinations of these upstream control elements and enhancers regulate and amplify the formation of the basal transcription complex. The transcription of a gene can be regulated by *cis*-acting elements within the regulatory regions of the DNA, and trans-acting factors that include transcription factors and the basal transcription complex.

FUTURE PROSPECTS

Harnessing the plant- pathogen genomics

Organisms that cause infectious disease include fungi, oomycetes, bacteria, viruses, viroids, virus-like organisms, phytoplasmas, protozoa, nematodes and parasitic plants [42]. Plant pathology involves the study of pathogen identification, disease etiology, disease cycles, economic impact, plant disease epidemiology, plant disease resistance, how plant diseases affect humans and animals, pathosystem genetics, and management of plant diseases. Genome sequenced from fungi, oomycetes, bacteria, viruses, viroids, virus-like organisms, phytoplasmas, protozoa, nematodes and parasitic plants gives opportunities to understand plant-pathogen interaction which helps to management, diagnose the disease and make disease resistance transgenic plant [43]. In its most simple form, the gene-for-gene hypothesis states that plants contain single dominant resistance R genes that specifically recognize pathogens that contain complementary avirulence genes. Avirulence genes can be defined as genes in the pathogen that encode a protein product that is conditionally recognized directly or indirectly only by those plants that contain the complementary R gene.

To survive, plants must defend themselves from numerous pathogens. Some defenses are constitutive, such as various pre-formed anti-microbial compounds, whereas others are activated by pathogen recognition. The recognition process includes the product of a dominant or semi-dominant resistance R gene present in the plant and the corresponding dominant avirulence (*Avr*) factor encoded by or derived from the pathogen. The recognition of the *Avr* factor by the host plant starts one or more signal transduction pathways that activate several of the plant's defenses, thus compromising the ability of the pathogen to colonize the plant. The interactions between plants and pathogens are specific, complex and dynamic [44]. The identification of resistant genes in the germplasm of wild species of field crops and their subsequent introgression into commercial cultivars has been the main approach of many plant breeders. Several strategies for the identification, characterization and functional analysis of plant genes involved in the triggering, signaling and response to biotic and abiotic factors have been recently envisaged. *In-silico* biology plays an important role to understand the plant pathogen interaction at gene and genome of plants and pathogens [45].

References

[1] Untergasser A., Nijveen H., Rao X., Bisseling T., Geurts R. and Leunissen J.A.M. (2007) *Nucleic Acids Research*, 35, W71-W74.

[2] Kumari N., Singh V.K., Narayan O.P., Rai L.C. (2011) *Online Journal of Bioinformatics*, 12, 289-303.

[3] Mahalakshmi V. and Ortiz R. (2001) *Electronic Journal of Biotechnology*, 3.

[4] Matthews D.E., Carollo V.L., Lazo G.R. and Anderson O.D. (2003). *Nucleic Acids Research*, 31, 183-186.

[5] Caetano-Anolles. (2005) *Crop Science*, 45, 1809-1816.

[6] Jaiswal P., Ni J., Yap I., Ware D., Spooner W., Youens-Clark K., Ren L., Liang C., Zhao W., Ratnapu K., Faga B., Canaran P., Fogleman M., Hebbard C., Avraham S., Schmidt S., Casstevens T.M., Buckler E.S., Stein L. and McCouch S. (2006) *Nucleic Acids Research*, 34, D717-D723.

[7] Paterson A.H., Freeling M. and Sasaki, T. (2005) *Genome Research*, 15, 1643-1650.

[8] Varshney R.K., Hoisington A.D. and Tyagi K.A. (2006) *Trends in Biotechnology*, 24, 1-10.

[9] Angellotti M.C., Bhuiyan S.B., Chen G. and Wan Xiu-Feng (2007) *Nucleic Acids Research*, 35, W132-W136.

[10] Kale U.K., Bhosle S.G., Manjari G.S., Joshi M., Bansode S. and Kolaskar A.S. (2006) *BMC Bioinformatics*, S12-S27.

[11] Tsuru T. and Kobayashi I. (2008 *Molecular Biology Evolution*, 25, 2457-2473.

[12] Handelsman J. (2004) *Microbiology and Molecular Biology Reviews*, 68, 669-685.

[13] Cory J.S. and Hoover K. (2006) *Trends in Ecology and Evolution*, 21, 278-286.

[14] Kummerfeld S.K. and Teichmann S.A. (2006) *Nucleic Acids Research*, 34, D74-–D81.

[15] Pandey S.P. and Somssich I.E. (2009) *Plant Physiology*, 150, 1648-1655.

[16] Todd J. Vision and Aoife McLysaght (2003) *Handbook of Plant Biotechnology (Ed. Paul Christou and Harry Klee). John Wiley and Sons Ltd.*

[17] Pruitt K.D., Tatusova T. and Maglott D.R. (2007) *Nucleic Acids Research*, 35, D61–D65.
[18] Gao G., Zhong Y., Guo A., Zhu Q., Tang W., Zheng W., Gu X., Wei L. and Luo j. (2006) *Bioinformatics*, 22, 1286-1287.
[19] Tateno Y., Imanishi T., Miyazaki S., Fukami-Kobayashi K., Saitou N., Sugawara H. and Gojobori T. (2002) *Nucleic Acid Research*, 30, 27-30.
[20] Dong Q., Lawrence C.J., Schlueter S.D., Wilkerson M.D., Kurtz S., Lushbough C. and Brendel V. (2005) *Plant Physiology*, 139, 610-618.
[21] Sakata K., Nagamura Y., Numa H., Antonio B.A., Nagasaki H., Idonuma A., WatanabeW., Shimizu Y., Horiuchi I., Matsumoto T., Sasaki T. and Higo K.(2002) *Nucleic Acids Research*, 30, 98-102.
[22] Iida K., Seki M., Sakurai T., Satou M., Akiyama K., Toyoda T., Konagaya A. and Shinozaki K. (2005) *DNA Research*, 12, 247-256.
[23] Guo A., He K., Liu D., Bai S., Gu S., Wei, L. and Luo J. (2005) *Bioinformatics*, 21, 2568-2569.
[24] Riano-Pachon D.M., Ruzicic S., Dreyer I. and Mueller-Roeber B. (2007) *BMC Bioinformatics*, 8, 42-53.
[25] Marla S., Singh V.K. (2007) *In Silico Biology*, 7, 543-555.
[26] Singh V.K., Ambwani S, Marla S., Kumar A. (2009) *Bioinformation*, 4, 182-183.
[27] Sanseverino W., Roma G., Simone M.D., Faino M., Melito S., Stupka E., Frusciante L. and Ercolano M.R. (2010) *Nucleic Acids Research,* 38, D814-D821.
[28] Finn R.D., Mistry J., Schuster-Bockler B., Griffiths-Jones S., Hollich V., lassmann T., Moxon S., Marshall M., Khanna A., Durbin R., Eddy S.R., Sonnhammer E.L.L. and Bateman A. (2006) *Nucleic Acids Research*, 34, D247-D251.
[29] Stormo G.D. (2000) *Bioinformatics*, 16, 16-23.
[30] Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990) *Journal of Molecular Biology*, 215, 403-410.
[31] Smith T.F., and Waterman M.S. (1981) *Journal of Molecular Biology*, 147, 195–197.
[32] Kushwaha H., Gupta N., Singh V.K., Kumar A., Yadav D. (2008) *Online Journal of Bioinformatics*, 9, 130-143.
[33] Kushwaha H., Gupta S., Singh V.K., Rastogi S. and Yadav D. (2010) *Molecular Biology Reporter* (OI: 10.1007/s11033-010-0650-9).
[34] Puigbo P., Guzma E., Romeu A. and Garcia-Vallve A. (2007) *Nucleic Acids Research*, 35, W126-W131.
[35] Yadav P.K., Singh V.K., Yadav S., Yadav K.D.S., and Yadav D. (2009) *Biochemistry (Mosc),* 9, 1049-1055.
[36] Dubey AK, Yadav S, Kumar M, Singh VK, Sarangi BK, Yadav D. (2010) *Enzyme Research*, 19, 2010:950230.
[37] Tamura K., Dudley J., Nei M. and Kumar S. (2007) *Molecular Biology Evolution,* 24, 1596-1599.
[38] Bailey T.L. and Gribskov M. (1998) Bioinformatics, 14, 48-54.
[39] Bailey T.L., Williams N., Misleh C. and Li W.W. (2006) *Nucleic Acids Research*, 34, W369-W373.
[40] Castro E. D., Sigrist C.J.A., Gattiker A., Bulliard V., Langendijk-Genevaux P.S., Gasteiger E., Bairoch A. and Hulo N. (2006) *Nucleic Acids Research*, 34, W362-W365.
[41] Yadav D., Singh V.K., Singh N.K. (2007) *Online Journal of Bioinformatics*, 8,: 1-9.
[42] Montesinos E., Bonaterra E.A., Badosa E.E., Frances E.J., Alemany J., Llorente E.I., Moragrega E.C. (2002) *International Microbiology*, 5, 169-175.
[43] Kim E., Kosack H. and Jonathan D.G.J. (*1997)Annual Review of Plant Physiology*, 48, 575-607.
[44] Matsumura H., Reich S., Ito A., Saitoh H., Kamoun S., Winter P., Kahl G., Reuter M., Kru ger D.H. and Terauchi R. (2003) *PNAS*, 100, 15718-15723.
[45] Wan J., Dunning F.M., Bent A.F. (2002) *Functional and Integrative Genomics*, 2, 259-273.

Table1- *List of Databases used for homology/similarity search.*

Database	Description	Website	Reference
NCBI	National Center for Biotechnology Information	http://www.ncbi.nlm.nih.gov	[17]
DRTF	Database of Rice transcription factor	http://drtf.cbi.pku.edu.cn/	[18]
DDBJ	DNA Data Bank of Japan	http://www.ddbj.nig.ac.jp/Welcomee.html	[19]
PGD	Plant Genome Data base	http://www.plantgdb.org/	[20]
RICE-GAAS	Rice Genome Automated Annotation System	http://ricegaas.dna.affrc.go.jp	[21]
GRAINGENES	genome database for small-grain crops	http://www.graingenes.org	[4]
GRAMENE	cereal genome database	http://www.gramene.org/	[6]
RARTF	RIKEN Arabidopsis transcription factor	http://rarge.gsc.riken.jp/rartf/	[22]
DATF	Database of *Arabidopsis* transcription factors	http://datf.cbi.pku.edu.cn/	[23]
PlnTFDB	plant transcription factor database	http://plntfdb.bio.uni-potsdam.de.	[24]
PGV	Pathogenic Genome Viewer	http://www.insilicogenomics.in/tools-agro-bacterium.asp	[25]
Cry-Bt identifier	Database of Cry genes	http://www.insilicogenomics.in/cry-bt-search.asp	[26]
PRGdb	platform for plant resistance gene analysis	http://prgdb.cbm.fvg.it/	[27]