Applications of Bioinformatics in Era of Genomics and Proteomics: Future Prospect and Challenges

Rohit Kumar Sharma[a], Jasleen Saini[b], Jaspreet Kaur Boparai[b], Ramanpreet Kaur[b] and , Pushpender Kumar Sharma*[b]

[a]*Department of Plant Science Crop Technology, Center University of Manitoba,Winnipeg-R3T2N2, Canada.*

[b]*Department of Biotechnology, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.*

*pushpg_78@rediffmail.com.

Abstract

Bioinformatics has played important role in better understanding the structure and function of biological macromolecules that primarily include the DNA and the proteins. Using bioinformatics tools, it has even become possible to get better insights about the important regulatory network of the cells, it has allowed the researchers in designing better drugs, crops, medicine and the therapeutic agents. The field of molecular biology together with bioinformatics can successfully address the cause of several dreadful diseases like cancer, diabetes, protein mis-folding and other metabolic disorders. However, despite these breakthroughs, it has several other challenges ahead; this mini review will provide insights about important applications, challenges and future prospects of the bioinformatics in various fields.

Keywords - Bioinformatics, Molecular biology, Dreadful diseases, Metabolic disorders.

Introduction

In recent years, the tremendous development made in the field of molecular biology has conduit the researchers towards better understanding of living organisms. It has provided in depth insights about the functioning of biological molecules. Interestingly, bioinformatics has further added towards the better understanding of the structure and functioning of these macromolecules. Each day basis, thousands of genes and proteins sequences are being added to the various databases, owing to advances in various sequencing technologies. For handling this large set of data, bioinformatics, a rapidly evolving and interdisciplinary field, is playing an important role. Bioinformatics grew in late nineties; however the more development in this field attained the pace during the Second World War, when Frederick Sanger and his colleagues at Cambridge University in 1945, successfully sequenced a complete protein, Insulin. Public accessibility to bioinformatics resources were started after the establishment of the US National Center for Biotechnology Information (NCBI) in 1988 (Smith, 2013). Later on two landmark initiatives *"Mapping and Sequencing of Human Genome"* by the National Research Council in 1988 and *"Mapping Our Genes-The Genome Project: How Big, How Fast?"* by the US Congress in 1988, brought revolution, and a new discipline- bioinformatics became inevitable. In 1992, a database of nucleotide sequences, Genbank was launched by NCBI in collaboration with European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ), and thus new era of biology emerged to fill up all data landscapes (Zou *et al.,* 2015). Before completion of human genome project (Lander *et al.,* 2001) and HapMap project (International HapMap Consortium, 2003) in different human backgrounds, sequencing of other plants (Arabidopsis genome initiative, 2000), microbial and insect genome were completed. After 2005, due to initiation of number of sequencing projects by technological advancement, and collection of data from other omics field, there has been a huge flow of data from DNA to RNA to protein was observed (Zhao *et al.,* 2012; Chen *et al.,* 2015). Therefore, there is comprehensive demand for large number of bioinformatics resources that can provide storage of data, information, curation and

analysis tool in an easy and user-friendly way. Fast accumulation of gene sequences from variety of organisms further enriched these databases with huge information that could be used in designing novel experiments (Tang *et al.*, 2015). In this review, insights would be provided about some of the important application of bioinformatics in various fields.

Role of Bio Informatics in crop Sciences

Bioinformatics resources and web databases plays an important role in the most effective use of genetic, proteomic, metabolomics and phenotypic information in increasing agricultural crop productivity. Comparative analysis of the plant genomes has shown that information obtained from the model crops can be used for the better improvements of other food crops. At present, the complete genomes of Arabidopsis thaliana (Arabidopsis genome initiative, 2000), Oryza sativa (rice) (International rice genome sequencing project, 2005) and wheat draft genome (International Wheat Genome Sequencing Consortium, 2014) are available. Innovations in web based platforms for omics based research, and application of such information, has provided the necessary platform to promote molecular based research in model plants, as well as important crop plants. Bioinformatics databases have become an important tool for crop scientists in mining the gene data, and linking this knowledge to its biological significance (Mochida and Shinozaki, 2010). On the basis of comparative genome analysis, species-specific nucleotide can provide information related to phenotypic characters. These tools have been explored in better designing of the crops that include the development of cereal varieties with greater tolerance towards the soil alkalinity, crop free of aluminium and iron toxicities. These varieties will allow agriculturists to grow these varieties in soil conditions, lacking appropriate condition. Research is also in progress to develop crop varieties that can resist reduced water conditions. Various molecular markers databases have been developed to relate genotype and phenotype, for the quality improvements of crops. Some of the important web based databases for crop improvement are provided in Table1.

Table 1: Important databases related to the crop improvement

Database name	URL	References
Brassica rapagenome database?	http://brassicadb.org/	(Cheng *et al.*, 2011)
DNA Data Bank of Japan (DDBJ)	http://ddbj.sakura.ne.jp/	(Sugawara *et al.*, 2008)
EnsEMBL plants	http://plants.ensembl.org/	(Flicek *et al.*, 2011; Kersey *et al.*, 2010)
EMBL nucleotide sequence database	http://www.ebi.ac.uk/embl/	(Kulikova *et al.*, 2007; Sterk *et al.*, 2007)
GenBank	http://www.ncbi.nlm.nih.gov/genbank/	(Benson *et al.*, 2009)
Graingenes	http://wheat.pw.usda.gov/	(Matthews *et al.*, 2003)
MaizeGDB	http://www.maizegdb.org/	(Lawrence *et al.*, 2007)
Rice Genome Annotation Project	http://rice.plantbiology.msu.edu/	(Ouyang *et al.*, 2007)
CR-EST	http://pgrc.ipk-gatersleben.de/cr-est/	(Kunne *et al.*, 2005)
Wheat genome information	http://www.wheatgenome.info	(Lai *et al.*, 2011)

Bioinformatics in health and development

The creation of innovative methods for handling medically important data is considered as a priority for scientific, coordinated and implementation oriented programmes of multidisciplinary character. The Sequencing of human genome will have profound effects on the fields of biomedical research and clinical medicine. It heralds a new era, where whole genome sequencing will become routine clinical practice for diagnosis and prognosis for personalized healthcare (Baker, 2012). There are number of inherited diseases like Cystic Fibrosis and Huntington's disease. Also the environmental stress causes alterations in the genome e.g. cancers, heart disease and diabetes which needs to be cured. This would only be possible if we are able to capture, curate and analyze clinical data with 'omics' datasets using novel informatics tools to establish correlations with high level of confidence. The new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed. Genome wide association studies have been carried out to establish correlation between gene to phenotype. The field of genotype-to-phenotype association study requires new efforts to integrate standardized genotype and phenotype information, so as to enable their easy access and robust analyses by the researchers. The GEN2PHEN, was one of such endeavors that aims to efficiently gather and organize the web-based genetic information that fundamentally impacts human health and disease prognosis (Thorisson *et al.,* 2009). While organizing existing datasets still remains a challenge, it is evident that the next generation of biosensors and electronic gadgets are going to revolutionize the personalized healthcare and will generate voluminous data on individuals. The systematic organization of this data requires generation of new tools and softwares. Some important databases for medical applications are given in Table 2.

Table 2: New and important databases related to medical application.

Database name	URL	Brief description
ADReCS	http://bioinf.xmu.edu.cn/ADReCS	Adverse Drug Reaction Classification System
AHTPdb	http://crdd.osdd.net/raghava/ahtpdb/	Antihypertensive peptides database
Cancer3D	http://www.cancer3d.org	Mapping of cancer mutations to protein structures
CancerPPD	http://crdd.osdd.net/raghava/cancerppd/	Experimentally validated anticancer peptides
CeCaFDB	http://www.cecafdb.org	Carbon flux data of central metabolism in various organisms
CMPD	http://cgbc.cgu.edu.tw/hmpd	Cancer Mutant Proteome Database
DDMGD	http://www.cbrc.kaust.edu.sa/ddmgd/	Associations between gene methylation and disease
Digital Ageing Atlas	http://ageing- map.org/	Human ageing-related data
EHFPI	http://biotech.bmi.ac.cn/ehfpi/	Essential Host Factors for Pathogenic Infection
MethBank	http://dnamethylome.org	Nucleotide methylomes of gametes and early embryos
ValidatorDB	http://ncbr.muni.cz/ValidatorDB	Validation results for ligands and residues in the PDB

Gene therapy has been employed in treatment of several diseases by modification of genes or gene expression. This field is in its infantile stage, however drug development is a growing field with an improved understanding of disease mechanisms and using computational tools to identify and validate

new drug targets, developing more specific medicines for the diseases. Infectious diseases are now the world's biggest killers of children and young adults. Most deaths from infectious diseases occur in developing countries. The cause for this has been attributed to the unavailability of efficient drugs and if at all available, the high cost associated with those drugs. Development of cheap and efficient drugs for a disease is one of the major problems faced by mankind. The solution to this problem could be from rational drug design using Bioinformatics. There are many examples where targeted pharmacogenetic dosing algorithm is used e.g. warfarin (International Warfarin Pharmacogenetics Consortium, 2009; Sagreiya *et al.,* 2010) and the incidence of adverse events is reduced by checking for susceptible genotypes for drugs like abacavir, carbamazepine and clozapine (Dettling *et al.,* 2007; Ferrell and McLeod, 2008).

Role of bioinformatics in microbial Biotechnology

Microbes are ubiquitous, and can survive in extreme environmental conditions. Complete sequencing of genome genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far-reaching implications for environment, health, energy and industrial applications. Since the sequencing of the first complete microbial genome of Haemophilusinfluenzaein 1995 (Fleischmann *et al.,* 1995), hundreds of microbial genomes have been sequenced and archived for public research.The major impact of bioinformatics research has been to automate the genome sequencing, automated development of integrated genomics and proteomics databases, automated genome comparisons to identify the genome function, automated derivation of microbial metabolic pathways, gene expression analysis to derive regulatory pathways, the development of statistical techniques, clustering techniques and data mining techniques to derive protein-protein and protein-DNA interactions, and modeling of 3D structure of proteins and 3D docking between proteins and biochemicals for rational drug design, difference analysis between pathogenic and non-pathogenic strains to identify candidate genes for vaccines and anti-microbial agents, and the whole genome comparison to understand the microbial evolution. The development of bioinformatics techniques has enhanced the pace of biological discovery by automated analysis of large number of microbial genomes. This has quickened the pace of discoveries, the drug and vaccine design (Robinson *et al.,* 2003) and the design of anti-microbial agents (Liu *et al.,* 2004). Bioinformatics can help to understand cellular mechanism and cellular manipulation using the integration of sequence data, wet lab, and cell simulation techniques.

Application in other relevant area of Biotechnology

With the advent of modern techniques in biology,biotechnology has progressed exponentially.This field has been divided into 'old' and 'new' biotechnology, where theformer deals with conventional and classical methods, and the latter involves highly specific and targeted r-DNAtechnology tools. Apart from human genomics, the evolutionary processes, genetic modification of plants, genome editing, biofortification, bioremediation, micro-RNAs and marker-assisted breeding are major areas of biotechnological research today.The benefits of plant engineering through biotechnology have started reaching the consumers of the developed world and are in the offing for the poor of thedeveloping countries. Bioinformatics provides data access to such developments. There are increasing evidences that transgenic plants would produce healthier, storable foods with desired nutritional value. Genetically modified plants are considered as chemical factories, which are capable of producing desired proteins, antigens, energy, vitamins,desired enzymes, etc. with the practice of biotechnological tools. Bioinformatics manages all the information and data on such aspects. This has led to the generation of enormous databases with information overload in the lifesciences disciplines. If this information were to be used tospeed up the pace of scientific research, scientists wouldneed to know the methods and tools for their effective use.Some of the important databases which are commonly used by biotechnologists are thosefrom Incyte, Pangea systems, PEinformatics, PDB, SRS, UDB, SWISS-PROT, EMBLnet, ICCBnet, Medline, FlyBase, MendelianInheritance in Man (MIM) etc. Some important tools,

programmes and algorithms available for analysis, like BLAST, FASTA, SmithWaterman, ENTREZ, MAGE, CHIME, RasMol, CASP, CAFASP1,PDB-3D Browser, SWISS PDB-Viewer,MG-RAST (Meyer *et al.*, 2008), Gene investigator, Trinity (Manfred *et al.*, 2011) etc.

Applications of Bioinformatics in Evolutionary Studies

Evolutionary studies help in exploring the relationships of organisms with their ancestors. Organisms can be classified on the basis of their evolutionary relationships. This classification could be depicted using a phylogenetic tree or evolutionary tree which could be rooted or unrooted (Huson and Bryant, 2005). A phylogenetic tree is composed of nodes and branches in which nodes represent taxonomic units and branches connect two nodes. Phylogenetic trees are constructed using computational methods like Distance based methods and Sequence based methods. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) is a distance based method that uses sequential clustering approach in order to build a phylogenetic tree. Neighbour Joining method (NJ) is another distance based method that build a tree by joining closest neighbours. Sequence based methods make use of the substitutions among the sequences in order to build a tree. Maximum Parsimony (MP) and Maximum Likelihood (ML) are the two approaches that are used to evaluate evolutionary history using the sequence changes. Bioinformatics play a very important role in determining the evolutionary relationships among the organisms through phylogenetic analysis. Various softwares are available for the phylogenetic analysis like Dendroscope (Huson *et al.*, 2007), ProtTest (Abascal *et al.*, 2005), MEGA (Molecular Evolutionary Genetics Analysis) (Kumar *et al.*, 2008; Kumar *et al.*, 2004), PANTHER (Mi *et al.*, 2007), PhyML (Guindon *et al.*, 2005) etc. Databases that are available for studying the evolutionary relationship are provided in Table 3.

Table 3: Important databases related to evolutionary analysis

Database Name	URL	References
COG Database	http://www.ncbi.nlm.nih.gov/COG/	(Tatusov *et al.*, 2003)
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	(Bauer *et al.*, 2007)
FUSARIUM-ID	http://fusarium.cbio.psu.edu/	(Geiser *et al.*, 2004)
UNITE	http://unite.zbi.ee	(Kõljalg *et al.*, 2004)
Comparative RNA	http://www.rna.icmb.utexas.edu/	(Cannone *et al.*, 2002)
		Web(CRW) Site
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	(Conte *et al.*, 2000)
EzTaxon	http://eztaxon-e.ezbiocloud.net	(Kim *et al.*, 2012)
MIRU-VNTRplus	www.miru-vntrplus.org/	(Allix-Be´guec *et al.*, 2008)

Application in Forensics sciences

In the modern era, bioinformatics has emerged as the science to deal with the various major problems in the field of biotechnology. Bioinformatics has made it possible to identify the genetic differences in the DNA sequences which arise due to copy number variation (Redon *et al.*, 2006), Single nucleotide polymorphisms (SNPs), Variable number of tandem repeats (VNTRs) (Nakamura *et al.*, 1987) or Short Tandem repeats (STRs). The detection of these variations in the DNA sequences has proved to be a boon for the forensic scientists. DNA profiling has been used as a major tool to detect the variable number of tandem repeats and short tandem repeats. Forensic science combines bioinformatics in order to solve various crime investigations, paternity suites and family relationships. Information related to the repeats has been stored in the form of databases. Some of the databases used in forensic science are given in Table 4.

Table 4: Important databases related to forensic science

Database name	URL	References
STRBase	http://www.str-base.org/index.php	(Ruitberg *et al.,* 2001)
ENFSI	http://www.enfsi.org/	(Gill *et al.,* 2003)
EMPOP	http://www.empop.org	(Parson and Dur, 2007)
MITOMASTER	http://mammag.web.uci.edu/twiki/bin/view/	(Brandon *et al.,* 2009)

Application in understanding the Structure and Function.

Proteomics is the systematic study that deals with the determination of structure, function and expression profile of proteins in the living system whose data elucidation and data mining can be done with the help of bioinformatics. Prediction of secondary structure of the proteins can be done using the stastical methods like Chou-Fasman algorithm (Chen *et al.,* 2006) and Garnier-Osguthorpe Robson (GOR) method (Garnier and Robson, 1989) that are based on the probability parameters derived from studies of known protein structures by X-ray crystallography. The tertiary structure prediction of proteins has also become possible using the computational methods like Ab-initio and Homology Modelling (Aszodi and Taylor, 1996). The concept of proteomics can be well understood after the separation of proteins and their analysis. The separation of proteins can be achieved by 2-D gel electrophoresis techniques and further analysis using Mass spectrometry (MS) approaches. The identification of differentially expressed proteins can be done using software like XCMS that compares these differentially expressed peptides (Smith *et al.,* 2006). Development of various softwares and databases has made the study of proteomics more convenient. Table 5 further discusses the databases being used in proteomic studies.

Table 5: Important databases for Proteomics

Database name	URL	References
wwPDB	http://www.wwpdb.org/	(Berman *et al.,* 2007)
ExPASy	http://www.expasy.org	(Gasteiger *et al.,* 2003)
PHOSIDA	http://www.phosida.com	(Gnad *et al.,* 2007)
PRIDE	http://www.ebi.ac.uk/pride	(Martens *et al.,* 2005)
PPDB	http://ppdb.tc.cornell.edu	(Sun *et al.,* 2009)
PhosphoSite	www.phosphosite.org/	(Hornbeck *et al.,* 2004)
COMET	http://comet-ms.sourceforge.net/	(Eng *et al.,* 2013)
AT_CHLORO	http://at-chloro.prabi.fr/at_chloro/	(Ferro *et al.,* 2010)

Conclusion

Keeping in view, the present pace of developments and investments, the future of bioinformatics is bright. New software's are being developed for easy and fast analysis of the data. Such developments are important for the future of bioinformatics. Emergence of Bioinformatics at fast pace is appreciable; however, some challenges are associated with this development. First, the integration of databases with each other, because they are increasingly becoming larger and larger and it will become very difficult to use individually for analysis. Efforts in this direction have been started like the development of MG-RAST tool, which is comprehensive pipeline for analysis of microbial genomes. Further, for using these resources we need trained biotechnologists with good programming knowledge. This can be made possible through various training programmes at various levels. Other major problem is associated with the storage and curation of data because it needs Exabyte of the space or high speed computing systems. Efforts in this field were also started. But in the future we cannot avoid the use of databases and we need a

database structure that can handle complex data types and be expandable, tools for querying, visualizing, and analyzing the data, and more standardized ways of not only designing and performing experiments, but also describing and analyzing the data.

References

Abascal, F., Zardoya, R., Posada, D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics,* 21, 2104–2105.

Allix-Be´guec, C., Harmsen, D., Weniger, T., Supply, P., Niemann, S. 2008. Evaluation and Strategy for Use of MIRU-VNTRplus, a Multifunctional Database for Online Analysis of Genotyping Data and Phylogenetic Identification of Mycobacterium tuberculosis Complex Isolates. *Journal Of Clinical Microbiology,* 46, 2692-2699.

Arabidopsis genome initiative. 2000. Analysis of genome sequence of the flowering plant Arabidopsis thaliana. *Nature,* 408, 796–815.

Aszodi, A., Taylor, W R. 1996. Homology modelling by distance geometry. *National Institute of Medical Reasearch,* 1, 325-334.

Baker, M. 2012. Contest to sequence centenarians kicks off. *Nature,* 487, 417.

Bauer, A M., Anderson, J B., Derbyshire, M K., Scott, C., Gonzales, N R., Gwadz, M., et al. 2007. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Research,* 35, D237–D240.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W. 2009. GenBank. *Nucleic Acids Research,* 37, 26–31.

Berman, H., Henrick, K., Nakamura, H., Markley, J L. 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research,* 35, D301-D303.

Brandon, M C., Ruiz-Pesini, E., Mishmar, D., Procaccio, V., Lott, M T., Nguyen, K C., Spolim, S., Patil, U., Baldi, P., Wallace, D C. 2009. MITOMASTER – A Bioinformatics Tool For the Analysis of Mitochondrial DNA Sequences. *Hum Mutat,* 30, 1-6.

Cannone, J J., Subramanian, S., Schnare, M N., Collett, J R., D'Souza, L M., Du, Y., et al. 2002. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics,* 3, 1471-2105.

Chen, H., Gu, F., Huang, Z. 2006. Improved Chou-Fasman method for protein secondary structure prediction. *BMC Bioinformatics,* 7, S4-S14.

Chen, T., Zhao, J., Ma, J., Zhu, Y. 2015. Web resources for mass spectrometry-based proteomics. *Genomics Proteomics Bioinformatics,* 13, 36–39.

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., Li, P., Hua, W., Wang, X., Cheng, F., et al. 2011. BRAD, the genetics and genomics database for Brassica plants. *BMC Plant Biology,* 11,1–6.

Conte, L L., Ailey, B., Hubbard, T J.P., Brenner, S E., Murzin, A G. 2000. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Research,* 28, 257-259.

Dettling, M., Cascorbi, I., Opgen-Rhein, C., Schaub, R. 2007. Clozapine-induced agranulocytosis in schizophrenic Caucasians: confirming clues for associations with human leukocyte class I and II antigens. *Pharmacogenomics Journal,* 7, 325–332.

Eng, J K., Jahan , T A., Hoopmann , M R. 2013. Comet: an open-source MS/MS sequence database search tool. *Proteomics,* 13, 22-24.

Ferrell, P. B., McLeod, H. L. 2008. Carbamazepine, HLA-B*1502 and risk of Stevens-Johnson syndrome and toxic epidermal necrolysis: US FDA recommendations. *Pharmacogenomics,* 9,

1543–1546.

Ferro , M., Brugiere, S., Salvi, D., Seigneurin-Berny , D., Court, M., Moyet, L., Ramus, C., Miras, S., Mellal , M., Le Gall, S., Kieffer-Jaquinod, S., Bruley , C., Garin, J., Joyard, J., Masselon, C., Rolland, N. 2010. AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Mol Cell Proteomics,* 9, 1063-1084.

Fleischmann, R. D., Adam, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A.R., Bult, C. J., Tomb, J. F., et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenza. *Science,* 269, 496–512.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S. et al. 2011. Ensembl. *Nucleic Acid Ressarch,* 39, 800–806.

Garnier, J., Robson, B. 1989. The GOR Method for Predicting Secondary Structures in Proteins. *Prediction of Protein Structure and the Principles of Protein Conformation,* 417-465.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R D., Bairoch, A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research,* 31, 3784–3788.

Geiser, D M., Gasco, M J., Kang, S., Makalowska, I., Veeraraghavan, N., Ward, T J., Zhang, N., Kuldau, G A. 2004. FUSARIUM-ID v. 1.0: A DNA sequence database for identifying Fusarium. 473-479.

Gill, P., Foreman, L., Buckelton, J S. 2003. Analysis of DNA databases across Europe compiled by the ENFSI group. *Forensic Sci Int,* 131, 184-196.

Gnad, F., Ren, S., Cox, J., Olsen, J V., Macek, B., Oroshi, M., Mann, M. 2007. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biology,* 8, 250.

Guindon, S., Lethiec, F., Duroux, P., Gascuel, O. 2005. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research,* 33, W557–W559.

Hornbeck, P V., Chabra, I., Kornhauser, J M., Skrzypek, E., Zhang, B. 2004. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics,* 4, 1551-1561.

Huson, D H., Richter, D C., Rausch, C., Dezulian, T., Franz, M., Rupp, R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics,* 8, 1471-2105.

Huson, D. H., Bryant, D. 2005. Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol,* 23, 254-267.

International HapMap Consortium. 2003. The International HapMap Project. *Nature,* 426, 789–796.

International rice genome sequencing project. 2005. A map-based sequence of rice genome. *Nature,* 436, 793–800.

International Warfarin Pharmacogenetics Consortium. 2009. Estimation of thewarfarin dose with clinical and pharmacogenetic data. *The New England Journal of Medicine,* 360, 753–764.

International Wheat Genome Sequencing Consortium. 2014. A chromosome based draft sequence of the hexaploid (Triticumaestivum) bread wheat genome. *Science,* 345, 286–300.

Kersey, P., Lawson, D., Birney, E., Derwent, P. S., Haimel, M., Herrero, J., Keenan, S., et al. 2010. Ensembl Genomes: Extending Ensembl across the taxonomic space. *Nucleic Acids Research,* 38, 563–569.

Kim, O., Cho, Y., Lee, K., Yoon, S., Kim, M., Na, H., Park, S., Jeon, Y S., Lee, J., Yi, H., Won, S., Chun, J. 2012. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *International Journal of Systematic and Evolutionary Microbiology,* 62, 716–721.

Kõljalg, U., Larsson, K H., Abarenkov, K., Nilsson, R H., Alexander, I J., Eberhardt, U., Erland, S., et al. 2004. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist,* 166, 1469-8137.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., *et al.* 2007. EMBL Nucleotide Sequence Database. *Nucleic Acids Research,* 35, 16–20.

Kumar, S., Nei, M., Dudley, J., Tamura, K. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings In Bioinformatics,* 5, 150–163.

Kumar, S., Nei, M., Dudley, J., Tamura, K. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings In Bioinformatics,* 9 , 299-306.

Kunne, C., Lange, M., Funke, T., Miehe, H., Thiel, T., Grosse, I., Scholz, U. 2005. CR-EST: A resource for crop ESTs. *Nucleic Acids Research,* 33, 619–621.

Lai, K., Berkman, P.J., Lorenc, M.T., Duran, C., Smits, L., Manoli, S., Stiller, J., Edwards, D. 2011. WheatGenome.info: An integrated database and portal for wheat genome information. *Plant Cell Physiology,* 53, 1–7.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., *et al*. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409, 860–921.

Lawrence, C.J., Schaeffer, M.L., Seigfried, T.E., Campbell, D.A., Harper, L.C. 2007. MaizeGDB's new data types, resources and activities. *Nucleic Acids Research,* 35, 895–900.

Liu, J., Dehbi, M., Moeck, G., et al. 2004. Antimicrobial drug discovery through bacteriophage genomics. *Nature Biotechnology,* 22,185–191.

Manfred, G. G., Brian, H. J., Moran, Y., Joshua L. Z., et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology,* 29, 644–652.

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., Apweile, R. 2005. PRIDE: The proteomics identifications database. *Proteomics,* 5, 3537-3545.

Matthews, D., Carollo, V.L., Lazo, G.R., Anderson, O.D. 2003. GrainGenes, the genome database for small-grain crops. *Nucleic Acids Research,* 31,183–186.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J. 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenome. *BMC Bioinformatics,* 9(386), 1–8.

Mi, H., Guo, N., Kejariwal, A., Thomas, P D. 2007. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Research,* 35, D247–D252.

Mochida, K., Shinozaki, K. 2010. Genomics and bioinformatics resources for crop improvement. *Plant Cell Physiology,* 51, 497–523.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., et al. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science,* 235, 1616-1622.

National Research Council (US) Committee on Mapping and Sequencing the Human Genome. 1988. Mapping and sequencing the human genome, Washington (DC), National Academies Press, http://www.ncbi.nlm.nih.gov/books/NBK218252/.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, Y., et al. 2007. The TIGR rice genome annotation resource: Improvements and new features. *Nucleic Acids Research,* 35, 883–887.

Parson, W., Dur, A. 2007. EMPOP--a forensic mtDNA database. *Forensic Sci Int Genet,* 1, 88-92.

Redon, R., Ishikawa, S., Fitch, K R. 2006. Global variation in copy number in the human genome. *Nature,* 444, 444–454.

Robinson, W. H., Fontoura, P., Lee, B. J., et al. 2003. Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis. *Nature Biotechnology,* 21, 1033–1039.

Ruitberg, C M., Reeder, D J., Butler, J M. 2001. STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research,* 29, 320-322.

Sagreiya, H., Berube, C., Wen, A., Ramakrishnan, R., Mir, A., Hamilton, A., Altman, R. B. 2010. Extending and evaluating a warfarin dosing algorithm thatincludes CYP4F2 and pooled rare variants of CYP2C9. *Pharmacogenet Genomics,* 20, 407–413.

Smith, C A., Want, E J., O'Maille, G., Abagyan, R., Siuzdak, G. 2006. XCMS:? Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal Chem,* 78, 779-787.

Smith, K. 2013. A brief history of NCBI's formation and growth. In: The NCBI handbook. Bethesda, National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/ books/NBK148949/.

Sterk, P., Kulikova, T., Kersey, P., Apweiler, R. 2007. The EMBL nucleotide sequence and genome reviews databases. *Methods in Molecular Biology,* 406, 1–21.

Sugawara, H., Ogasawara, O., Okubo, K., Gojobori, T., Tateno, Y. 2008. DDBJ with new system and face. *Nucleic Acids Research,* 36, 22–24.

Sun, Q., Zybailov, B., Majeran, W., Friso, G., Dominic, P., Olinares, B., van Wijk, K J. 2009. PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Research,* 37, D969-974.

Tang, B., Wang, Y., Zhu, J., Zhao, W. 2015. Web resources for model organism studies. *Genomics Proteomics Bioinformatics,* 13, 64–68.

Tatusov, R L., Fedorova, N D., Jackson, J D., Jacobs, A R., Kiryutin, B., Koonin, E V., Krylov, D M., et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* 4.

Thorisson, G. A., Lancaster, O., Free, R. C., Hastings, R. K., Sarmah, P., Dash, D., Brahmachari, S. K., Brookes, A. J. 2009. HGVbaseG2P: a central genetic association database. *Nucleic Acids Research,* 37, 797–802.

US Congress, Office of Technology Assessment. 1988. Mapping our genes, the Genome Project: how big, how fast, Washington (DC), US Government Printing Office, http://www.ornl. gov/sci/techresources/Human_Genome/publicat/OTAreport.pdf

Zhao, D., Wu, J., Zhou, Y., Gong, W., Xiao, J., et al. 2012. WikiCell: a unified resource platform for human transcriptomics research. *Omics,* 16, 357–362.

Zou, D., Ma, L., Yu, J., Zhang, Z. 2015. Biological databases for human research. *Genomics Proteomics Bioinformatics,* 13, 55–63.