American Journal of Bioinformatics 1 (1): 10-19, 2012
ISSN 1948-9862

Bioinformatics-Assisted Microbiological Research: Tasks, Developments and Upcoming Challenges

[1]Dhananjaya P. Singh,
[1]Ratna Prabha, [2]Anil Rai and [1]Dilip K. Arora
[1]National Bureau of Agriculturally Important Microorganisms,
Indian Council of Agricultural Research, Kushmaur, Maunath Bhanjan 275101, India
[2]Indian Agricultural Statistical Research Institute,
Indian Council of Agricultural Research, Pusa, New Delhi 110 012, India

Abstract: Problem statement: Bioinformatics in the present day microbiological research is an inevitable subject area that encompasses biological resources and high end computational skills to unravel the coded and encrypted information within the life. We have produced a brief account of the developments and tasks in the subject and upcoming challenges in the subject. The area has seen tremendous developmental pattern in the last few decades due to the emerging computational technologies dedicated for uncovering the complex but vital biological information that not only essentially constitute the basis of life but entails about the evolutionary diversification and multiphasic interaction among the organisms with their own environment. Now, with the technological advancements, bioinformatics has completely changed microbiological domain for researchers. **Conclusion/Recommendations:** In Future, the ultimate goal of bioinformatics will be such kind of integration of the biological databases and genomic resources that can result in a computer representation of living cells and organisms whereby any aspect of biology can be examined computationally.

Key words: Bioinformatics, computational biology, microbes, databases, genomes, leading edge research, evolutionary diversification, complex nature, biological system

INTRODUCTION

With the leading edge research in the science of life which is now largely being realized as 'coded information' between and within lives organisms, it remains a matter of hard task to uncover the codes and thereby, make critical understanding of biological mysteries. This task is not only confined to the identity of phenotypic traits in the organisms but it also encompasses origin of life on the earth, evolutionary diversification, basic principles of survival adaptations, habitat-wise distribution of life-forms, characterization of valuable genetic traits leading to sustenance within the species, multitrophic interactions within communities, adoptive strategies of organisms in response to biotic and biotic stresses, sustainable crop productivity, challenges of the climate change in environment and even more complex but often least deciphered characters of the unknown microbial communities in the environment. The organized and self-explanatory `coded information` in the organisms travels across the families, genera, species, strains and races to make an organism what it is? This is why the genomic Deoxyribose Nucleic Acid (DNA) is a digital master plan for all living entities and if computed properly it can uncover the biological mysteries to know all about any organism.

With the advent of massive genome sequencing projects of microbes and others molecular biology has now become a heavily "data-driven" science and wide spread, fast growing, very complex and often interdependent biological research data are coming from all across the world. This has created the problem of misleading results and inconclusive interpretations. Therefore, the re-introduction of biologically inspired computational methods in biology was needed to enhance the understanding of biological systems as information processing systems (Hogeweg, 2011). Computational, mathematical, statistical and informatics technologies developed parallel to the

Corresponding Author: Dhananjaya P. Singh, National Bureau of Agriculturally Important Microorganisms, Indian Council of Agricultural Research, Kushmaur, Maunath Bhanjan 275101, India
Tel: 91547-2530080 Fax. 91547-2530358

biological research enabled scientists to interconnect, integrate and interpret the complex nature of any biological system e.g., their phenotypic, genotypic and metabolic characteristics, cellular processes, growth and development, regulatory networks for metabolism and catabolism, protein structure, function and conformation, changes and expressions at the genetic level, whole genome, proteome and metabolite constitution, post-transcriptional and post-translational changes and their impacts, responses of organisms to the environment, pathogens and abiotic stresses and interactions with other organisms. In fact, information extraction from complex data is a great problem in biological research where computational systems, biostatistics and information technologies are finding their increasing applications. The assemblage and integration of all these technologies in solving the problems related to the biological systems has been termed as "bioinformatics" in mid 1980s (Hagen, 2000). Since then it has not only been referred to the computational methods for comparative analysis of genomic data but has been defined as the study of the informatic processes in the biotic systems (Hogeweg, 2011).

The information flow across the life at various levels include information accumulation in organisms during evolution and information flow from genetic material to regulate intra- and intercellular processes and need interpretation at multiple levels (Hogeweg, 2011). With the help of bioinformatics and computational biology, a greater understanding of complex metabolic network transformations within the biological processes can be obtained (Andreas, 2007). The central research theme of the molecular biology of life always encouraged to understand how living systems accumulate, process and use biological information within molecules (Nurse, 2008). Bioinformatics applies principles of informatics to make vast, diverse and complex biological data more reproducible, reliable, consistent, understandable and usable while, computational biology uses mathematical models and computing approaches to address experimental and theoretical queries (Gilbert, 2004). In this way, apart from being distinct in functions and approaches, there are significant overlaps in their activities to bridge the interface of the science of biological information (Fig. 1).

The encoding of the complex but highly structured data in a genome of an organism is the greatest challenges of all time. The prospects of sequencing several microbial and other genomes, being central to the bioinformatics, are great opportunities for theorists interested in structural information, design, makeup, conformation and analysis.

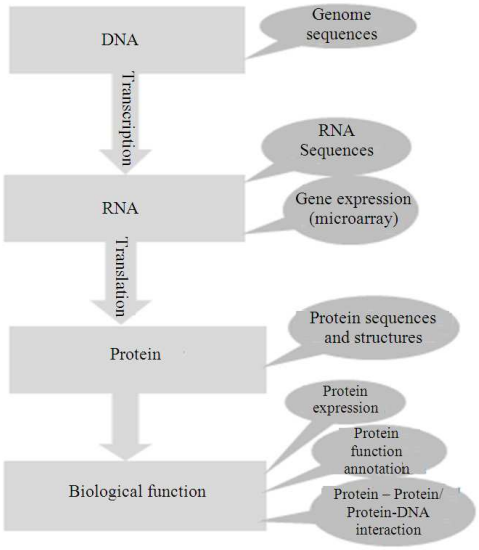

Fig. 1: Interconnecting central dogma of life with biological tasks

The problems related to interactive behavior, functions and interpretations lying with the structural biology has encouraged scientists to develop mathematical models for the integrated analysis of genomic data to facilitate appropriate and proper interpretations (Fenstermacher, 2005). Model-based analysis of microorganisms has begun to understand functional modules in metabolic and transcriptional networks for predicting cellular behavior from genome-scale physicochemical constraints and to suggest novel design principles for well-studied bacterial subsystems such as chemotaxis (Stelling, 2004). Microbes constitute an efficient model to study such problems. Since the sequencing and analysis of the first genome of free-living bacterium *Haemophilus influenza* (Fleischmann *et al*., 1995), a huge data has been generated on prokaryotic genomes (Table 2).

The underlying complexities and dynamics of data is a basic driving force to understand and substract specific biological data sets and identifying the back-end theoretical problems for elucidating, representing and analyzing the inherent structure hidden within the biological systems (Altman and Klein, 2002). This translates a biological query into the language of information and brings in the computational skills to solve the problems. The role of the bioinformatics is, thus defined as a center interconnecting the biological data (microcosm) to the underlying computational methods of structural elucidation for the abstracting of the information lying behind (Oro *et al*., 1990).

Bioinformatics: Computing for biology: Life, at its beginning is supposed to be started with only one

nucleotide sequence (Neerincx and Leunissen, 2005) and with the progression, became more complex due to the development of organisms and their biotic and biotic interactions in nature. Later on the complexity became so interwoven that the analysis of any single component seems unjustified as this might be an interrelated phenomenon in the living world (Fig. 2). The development of computational methods based on the organized algorithms, interpretational skills and high storage capacities facilitated comparison of entire genomes and thus permit biologists to study more complex evolutionary trends like gene duplication, horizontal gene transfer and prediction of factors important in speciation (Nakashima *et al*., 2005). The ultimate aim of such studies lies in deciphering the evolutionary lineages among the group of organisms in a quest to determine the tree of life and the last universal common ancestor. The segregation of biology with computers largely reflects that life itself is a kind of integrated and organized but coded information and computers are required to 1) perform repetitive biological tasks (e.g., alignment or comparison of sequences within and across the genomes) and ii) manage high-end analytical skills with reproducibility (e.g., interpretation and integration of huge datasets, deducing complex structures of proteins, interpretation of biochemical pathways and regulatory networks and RNA expression profiles) (Bansal, 2005; Gabaldon, 2008).

The quality, quantity and variety of the information dynamics in a single experiment that includes study of gene expression involves analysis of genes, determination of protein structures encoded by the genes and details of how these products interact with one another. The ease with which computers handle large quantities of diverse data at a time and probe the complex dynamics observed in nature make them indispensable to assist biological research (Heng, 2011). Algorithms, computable set of steps to achieve desired results are probably at the heart of bioinformatics (Nakashima *et al*., 2005). At various levels, algorithms are used to compare genome sequences, o find similar regions for genes, determine their functions, study their regulation and assess how they and entire genomes have evolved over the time. Over and above, the aim is to provide biologically important predictions from annotated data and transformations/manipulations of these datasets to find out most appropriately predicted results. Although, many people are engaged in the application and analytical aspects of the database management, only a handful of researchers have the privilege and skills to develop algorithms and theories in the traditional research (Nakashima *et al*., 2005).

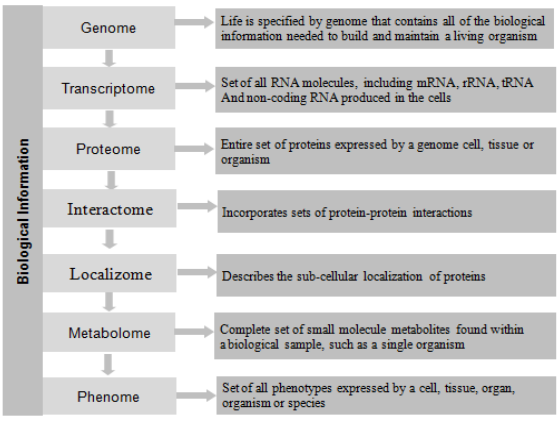

Fig. 2: Bioinformatics for integrated biological functions and their theoretical interpretations

The development of a large number of databases, softwares, tools and web-resources has been witnessed in the past few decades to facilitate bioinformatics-related tasks and this has facilitated wide applications for researchers.

In nature, widely dispersed, complex and interdependent biological systems follow structured arrays to deliver a particular function (Fenstermacher, 2005). The artificial intelligence machine can run parallel to the science of life due to the systematic organizational structures and the characters and functions that follow them (Mochida and Shinozaki, 2010). Computational skills based on the principles of logic and algorithms are able to decode the biological information that can be interpreted to understand the functional complexities (Kanehisa and Bork, 2003). Bioinformatics is basically elaborative analysis of the biological information encrypted in the form of either a coded genetic language within the living cells or organized mechanics found in the cellular processes. This is why for the analysis and integration, analytical capacity of computing technologies both for prima-face understanding and management of biological information is essentially required (Adnan, 2010).

Tools in bioinformatics are indispensable in life science (Vassilev *et al*., 2005) and computation biology is a key component to support experimental genomic studies (e.g., molecular and metabolic mapping, gene expression, genetic variation and protein interactions) to answer unanswered scientific questions (Searls, 2000). Cumulative collection, storage and concurrent analysis of integrated biological and genetic information can be applied to gene-based drug discoveries (Ouzounis, 2002), microbial identification and community analysis (Pearson and Lipman, 1988), bioremediation and biodegradation processes+, crop improvement

programs (Sugden and Pennisi, 2000; McDaniel *et al*., 2005) and environmental and agricultural development (Varshney *et al*., 2005).

Analytical and computing capability in bioinformatics facilitated the processing of huge data generated by genome sequencing projects and quickened the elucidation of systemic behavior of cellular processes, regulatory networks, changes in the metabolic efficiency and regulation of genetic profile that really control a cell (Bansal, 2005; Rao *et al*., 2008). It may remain unmanageable and un-interpreted due to the limitations of human analytical capacity and lack of expert manpower for novel operations (Pawlowski *et al*., 2001). Use of computers has emerged as a bridge that filled the gaps and evolved as a cost effective, reproducible, accurate and high-use efficiency data interpreter in biological sciences (Patterson, 2003; Patil *et al*., 2005; Mochida and Shinozaki, 2010). Technique to identify information existing with the cell processes, their components and products in the form of genes, proteins, primary and secondary metabolites, pathways and regulatory networks have been identified not only in the normal cells, but in the treated, stressed or metabolically enhanced cells so that a comparative basis of the stimuli can be known (Fig. 3). The upcoming years will witness the developments in understanding mechanisms and manipulations at cellular level using the integration of bioinformatics, wet lab research and bioimaging and cell simulation techniques. Such studies have been started by various laboratories all around the world and it is anticipated that the semi-automated study of cellular behavior at systemic level will accelerate the existing capability (Fig. 4) (Moret *et al*., 2002).

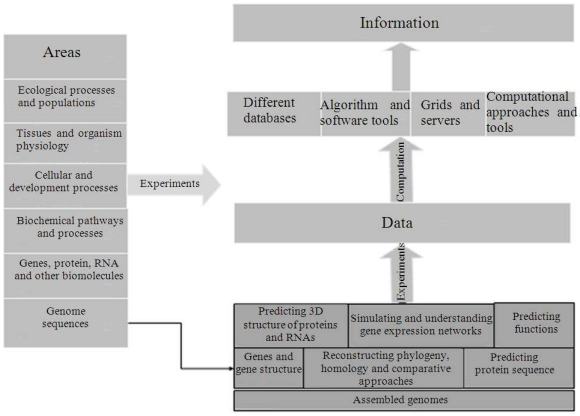

Fig. 3: Organizational characteristics for information extraction in bioinformatics

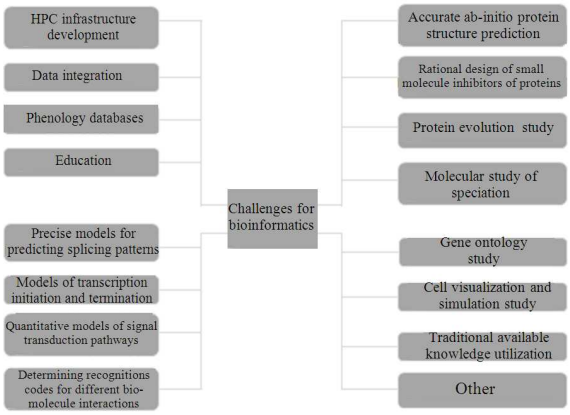

Fig. 4: Bioinformatics-trends and upcoming challenges in the future

Developments in bioinformatics: Bioinformatics helped to generate, integrate and analyze huge genomic and proteomic data and to extract the desirable and interpretable information as result of large-scale data processing (Juretic *et al*., 2005). Growth of bioinformatics and computational biology as distinct and interrelated disciplines in the recent years is unprecedented and has economic and social impacts on the life in many applied fields like pharmaceutical discovery, drug designing, disease diagnostics, environmental protection, ecological succession and agricultural implications (Huynen *et al*., 2000; Primetta *et al*., 2009). The development of integrated computer-based biotechnological systems has facilitated high-throughput screening and high-content detection systems to generate high-end data from biological system for interpretation and analysis with great precision and reproducibility (Marcotte *et al*., 1999).

Computational methods of scoring the coding DNA regions have been developed to help identify the annotation of new genes from the whole genomes of prokaryotic organisms (Salzberg *et al*., 1998).

Developments in experimental technologies in molecular biology and biochemistry supported the growth of bioinformatics. Parallel advent of internet that revolutionized the information access, publication technologies and other aspects of information infrastructure have added to the efficiency, speed, memory and storage capability. The overall impact has geared up the idea and the need of using computers in understanding the complex, often huge and inter-related information resources lying behind the genetics, biochemistry and evolution of the organisms (Jones, 2001; Bansal, 2005). Thousands of full or partial sequencing of prokaryotic and eukaryotic genomes

including a cleaner draft of human genome (Human Genome Project, 2001) have been obtained and large amount of data sets are continuously being generated worldwide (Table 2 and 3). In the biological world, the expectations from these data resources in the area of life science, agriculture, food, environment (bioremediation and pollution control), medicine (animal and human health) and industry (biotechnology based) is very high. How these complex datasets can be analyzed and how the hidden information that can make a better control of any organism can be retrieved in an ordered fashion to make significance out of it? Problems related to analysis, interpretation, mining, integrating and correlating the biological data with the help of high performance computing tools are now becoming inevitable. During 1990s, the term bioinformatics was originally synonymous with the management and analysis of DNA, RNA and protein sequence data. Now, converting the analogous information lying within a linear string of four chemical groups encoding the entire blueprints for the protein machinery in the living cell into digital information was actually among the greatest tasks that unified the efforts of the biologists and the computational people (Piatnitskii *et al*., 2009).

Tasks for bioinformatics in microbial research: During early days of its emergence, bioinformatics was mainly confined to DNA and protein sequence and protein structure (3D) analysis. In the new post-genomic era that has been identified with the exponential growth and accumulation of the molecular data (Juretic *et al*., 2005) the knowledge on the-omics technologies progressed and diversified tasks have emerged to help different interrelated areas in biological systems. Bioinformatics research tasks and techniques are being listed in Table 4. It is imperative to note that all these activities may go parallel and individual projects may emerge with huge set of data where bioinformatics can be applied to address unknown correlations, corrections, interaction, integration and directional trends.

Bioinformatics has facilitated researchers to study microbial biodiversity because of its direct interventions in molecular identification, data storage and retrieval system that were the stuff and the nightmare of systematics research. The bioinformatics-driven approaches enabled people to work efficiently on microbial diversity, identification, characterization, molecular taxonomy and community analysis patterns of both culturable and unculturable organisms (metagenomics) (Rogozin *et al*., 2002). Cataloging and digitization of microbial diversity is one of the most important tasks for bioinformaticians. Description of new species, genera and even molecular taxa emerged dramatically in the literature after 1990s and these efforts are largely driven by advances in sequencing technologies. Even though microbiologists believe that 99% of the microbial world has yet to be uncovered and therefore, bioinformatics has to play an important role (Curtis *et al*., 2002).

Genomics, the science that deals with study of whole genome, largely encompasses biology of genetics at molecular level i.e., the constitution of DNA and RNA, its analysis, translating of the chemical information carried over by these materials into biological data and digitizing that huge biological data through computing (Juretic *et al*., 2005; Prabha *et al*., 2011). Because microbes possess modest-sized genomes (4-5 million bases), they represent a tractable life forms to explore and understand life processes at a single cell level. However, the whole genome of any organism is useful only when its sequences can be fully analyzed and the information within it can be interpreted. Almost every complete genome sequence from prokaryotes, when analyzed, indicated that almost half of predicted coding regions identified are of unknown biological functions (Lopez, 2008). Environmental samplings of microbes and their functional communities have led to the discovery of millions of unknown genes and proteins, thousands of species and vast variations in critical functions (Liu *et al*., 2011). This may be interesting information but, it again requires re-confirmation whether this is true, because we are still far behind in developing suitable softwares for comparative and functional genome analysis (Callister *et al*., 2008). Modern computing approaches and analytical tools will strengthen bioinformatics in near future to play a crucial role in modern microbial genomics research.

Developments in sequencing whole genome of many microbes and their analysis along with the other technologies in microbial research (Khanna, 2007) led to conclude that biological functions are genetically conserved throughout the evolutionary pathways across the species (Keller and Zengler, 2004) and this has provided a basis for the molecular microbial phylogenetic analysis on the basis of conserved sequences (Tomitani *et al*., 2006) or on whole genome phylogeny (Prabha *et al*., 2011). This information became a gateway for opening development of modern bioinformatics which is not only confined to the biological database development and management but on the extraction of useful information as much as possible from such huge databases. Now, the capabilities of bioinformatics lie in extracting and analyzing information through the tools and infrastructure to document acquired data and knowledge for high performance computing systems (Rogozin *et al*., 2002).

Microbial data resources: The biological data resources from wet-lab experimentation in myriad of whole genome, proteome, metabolome and interrelated projects are being generated at a phenomenal rate (Benson *et al*., 2000; Rhee *et al*., 2006) (Table 1). Microbes, evolved for some 3.8 billion years back and making-up most of the earth's biomass, are the most suitable organisms for such kind of studies. They inhabit virtually all environments where no other life forms can exist and survive and thrive well in extremes of heat, cold, radiation, pressure, salt, acidity and darkness. Microbial diversity under a wide range of environmental adaptations indicated that they can offer solutions for many problems for which biologists still need answers. This is why bioinformatics for microbiological systems is facing allover global attention and this has led to the problem of data explosion.

Currently over 11,364 whole genome sequences organized in three major groups of organisms i.e. eukaryota, prokaryota (archaea and bacteria) and viruses are available in Genome database of NCBI in 2011 including complete chromosomes, organelles and plasmids as well as draft genome assemblies (Table 2). In addition to these, 41 complete genome sequences are also available for viroids. Out of 11,364 whole genome sequences, 7473 genome projects running across the world belong only to microbes with 1696 completed microbial genomes projects whereas assembly is being done for 2247 organisms and 3531 genome project are still unfinished (Table 3). There are approximately 106,533,156,756 nucleotide bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,067 sequence records in the WGS division (Benson *et al*., 2000).

With the help of bioinformatics tools, comparative microbial genomic studies are taking shape at a faster rate leading to the development of different types of function prediction concepts, most important of them being the gene context and gene content analysis. Gene context is the positional association of genes such as an operon in prokaryotic genomes (Huerta *et al*., 2000) while gene content analysis is a comparison of gene repertoires across different genomes (Luscombe *et al*., 2001). The postgenomic problems like protein structural determination and issues of gene function identification become more promising (Gomez *et al*., 2008) with the rapidly increasing number of completely sequenced genomes. Predicting the structures of proteins encoded by genes of interest provides subtle clues regarding the functions of these proteins (Idekar *et al*., 2001; Jones, 2000).

Table 1: Size and sources of data for bioinformatics tasks

Data source	Data size	Bioinformatics topics
Raw DNA sequence	8.2 million sequences (9.5 billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein sequence	300,000 sequences (~300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identification of conserved sequence motifs Secondary, tertiary structure prediction
Macromolecular structure	13,000 structures (~1,000 atomic coordinates each)	3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculations Intermolecular interactions Molecular simulations (force-field calculations, molecular movements, docking predictions) Characterization of repeats
Genomes	40 complete genomes (1.6 million – 3 billion bases each)	Structural assignments to genes Phylogenetic analysis Genomic-scale censuses (characterization of protein content, metabolic pathways)
Gene expression	largest: ~20 time point measurements for ~6,000 genes	Linkage analysis relating specific genes to diseases Correlating expression patterns Mapping expression data to sequence structural and biochemical data
Other data Literature	More than 15 million citations	Digital libraries for automated bibliographical searches Knowledge databases of data from literature

Table 2: Number of genome sequences according to group of organisms as available in Genome database

Organisms	Genome sequences (number)
Viruses	2683
Eukaryota	1208
Bacteria	7264
Archaea	209

Source: Genome database of NCBI, July 2011

Table 3: Number of assembled, unfinished and complete genome sequences of microorganisms

Groups	Sub-Groups	Complete	Assembly	Unfinished	Total
Archaea	Crenarchaeota	37	1	15	53
	Euryarchaeota	74	27	45	146
	Nanoarchaeota	1	0	0	1
	Others	1	2	6	9
Bacteria	Acidobacteria	5	1	4	10
	Actinobacteria	157	257	245	659
	Aquificae	10	2	7	17
	Bacteroidetes/chlorobi Verrucomicrobia	69	116	120	305
	Chlamydiae/	37	12	42	91
	Chloroflexi	15	1	5	21
	Cyanobacteria	42	24	44	110
	Deinococcus-Thermus	14	1	8	23
	Firmicutes	387	738	872	1997
	Fusobacteria	5	22	11	38
	Planctomycetes	5	4	5	14
	Proteobacteria	741	910	2005	3656
	Spirochaetes	34	97	43	174
	Thermotogae	12	1	7	20
	Others	49	31	48	128

Source: Genome database of NCBI, July 2011

In particular, the comparison of fully sequenced whole genomes allows investigating genomic context that includes chromosomal positioning of a gene relative to other genes and its evolutionary track record among the compared genomes (Prabha *et al*., 2011). This information can be exploited to find out functionally interacting partners for a protein of unknown functions and to obtain information on its role-based biological process.

Table 4: General tasks and techniques in bioinformatics

General tasks	Informatics techniques
Sequence retrieval	Databases
Similarity Search	Building, Querying
Nucleotide Vs Nucleotide	Object DB
Protein vs Protein	Text String Comparison
Translated nucleic acid vs Protein	Text Search
Unspecified Sequence Type	1D Alignment
Search for non coding DNA	Significance statistics
Functional Motif Searching	Alta Vista, grep
Restriction Mapping	Finding Patterns
Secondary and tertiary structure prediction	AI/Machine Learning
Other DNA analysis including translations	Clustering
Primer design	Data mining
ORF analysis	Geometry
Literature analysis	Robotics
Phylogenetic analysis	Graphics (Surfaces,Volumes)
Metagenomics	Comparison and 3D Matching (Vision,recognition)
Metabolomics	Physical Simulation
Proteomics	Newtonian Mechanics
	Electrostatics
	Numerical Algorithms
	Simulation

Such comparative genomics-based techniques are increasingly being used in the process of genome annotation and in the development of testable working hypothesis (Fulekar and Sharma, 2008). Growing interest in the genome sequencing lead to the generation of sequences for millions of genes but the function of majority of these genes either remains unknown or can be determined experimentally only for a few. Modern accurate and robust methods for *in silico* annotation of gene functions in comparative genomics based on computational prediction of functionally related proteins allow obtaining correct functional annotations for more than a half of all organisms' proteins (Callister *et al*., 2008; Petrosino *et al*., 2009). It is therefore, imperative that without the applications of bioinformatics, research and development in the field of crop improvement programs, microbial research, environmental biotechnology, pharmacognosy, search for new and novel molecules, drug designing and molecular modeling for research targets may come to standstill.

CONCLUSION

Upcoming bioinformatics: The future of bioinformatics lies in the integrated ability of the computational methods, simulation and modeling to extract information or predict what exactly is going on within a cell in real time (Altman and Klein, 2002). Integration of a wide variety of data sources like genomic, proteomic, metabolomic. Will allow us to use disease symptoms to predict genetic mutations and *vice versa*. The integration of GIS data like geographical maps and weather systems with crop health and genotypic traits can predict successful outcomes of diseases and pests in agricultural systems. Another challenging research area in bioinformatics is large-scale comparative genomics. The development of practical tools to compare whole genomes of organisms can unravel the discovery rate in bioinformatics. The modeling and visualization of complex networks of cellular systems can be used in the future to predict how the system (or cell) reacts to an predicted or unpredicted stress. Technical challenges before bioinformaticians need to be addressed by fast computing power, advanced storage capabilities and increased bandwidth. The future challenge of bioinformatics will be the way of addressing how computationally complex biological observations such as gene expression patterns and protein networks, metabolic regulatory networks and their interactions can be interpreted effectively, easily and efficiently. Bioinformatics is about converting complex biological information in to a model understandable to the computer which ultimately develops a pattern in the complexities. The problem of digitizing phenotypic data such as complex behavior of microbes in different niches and correlating the same with the crop or soil health in a manner readable for computer offers exciting opportunities for future bioinformaticians.

The area is very rewarding for software developers to look into the biological side of microbial research and to develop simulated insights into how cells work. This is also parallel to developing huge datasets of genomic, proteomic and metabolomic information. Therefore, there is something overlapping for both the microbiologists to perform bioinformatics-assisted wet-lab tasks and computational people to develop and design databases, user interfaces and advanced statistical algorithms. Presently, for central dogma-based biological processes like DNA sequence to protein sequence, protein sequence to protein structure and protein structure to fonction there is massive need to develop bioinformatics tools to uncover the hidden mystery within the tiny and often unseen microbial life forms. The integration of information obtained from these key biological processes within the cells will allow us to achieve complete understanding of the biology processes of any organisms.

The future research approaches will focus on targets of pathophysiological processes arising due to microbial diseases in plants rather than only remedies. For this purpose, metabolomics, a recently emerging -omics area is finding out its applications along with the bioinformatics capabilities for data integration, management and analysis. Other area that assumes increasingly higher significance is the application of information technology to the entire agriculturally important microbiological sector in a manner similar

that introduced in the industrial sector with improved efficiency, reduced cost and wide access (Rogozin *et al*., 2002). In coming years, most impactful tasks in microbial research and development such as microbial mapping and identification of different agroecological regions using culture dependent or metagenomics approaches, molecular taxonomy, finding out potential genes and gene products for microbial management of disturbed agricultural soils, bioprospecting for novel metabolites, biotic and abiotic stress tolerance in crops (Tiwari *et al*., 2011), bioremediation, biofermentation, microbe-associated soil fertility and crop improvement programs, development of next generation microbial inoculants as biofertilizers and biopesticides (Singh *et al*., 2011) may not be completed without the applications of bioinformatics (Wollenweber *et al*., 2005).

Numerous databases and computational tools have been created in order to provide the scientific community access to a range of genomic data, as well as to the results of comparative analyses of such data. Diverse options to visualize, search, retrieve and analyse these data are offered, providing the opportunity to acquire more detailed knowledge about genomes and their respective organisms. However, this wealth of information is presently fragmented, dispersed across all these computational resources and is redundant in many circumstances clearly requiring unification in order to provide a global and integral picture of the biology of such genomes and species.

Other emerging challenges in the future are the authentication, monitoring, auditing and control of fast increasing microbe-based databases. The complex computing structure and resource crunches would make it vital for computational people to formulate a practical guideline for ensuring authenticity of the data resources in quickly changing computational environment. Future challenges of sequence analysis are pushing bioinformatics in an era when the demand of bioinformaticians is going to be fastened and thus, in coming days more microbial biotechnologists with computation skills will be needed to pursue the tasks of molecular biology.

Till now bioinformatics has been applied in almost all the fields of biological studies, starting from genome and up to phenome. More recent areas in this list are interactome, which incorporates sets of protein-protein interactions and localizome, which describes the subcellular localizations of proteins. In Future, the ultimate goal of bioinformatics will be such kind of integration of the biological databases and genomic resources that can result in a computer representation of living cells and organisms whereby any aspect of biology can be examined computationally.

ACKNOWLEDGEMENT

The financial assistance from the National Agricultural Innovation Project (NAIP), India in the form of the project entitled “Establishment of National Agricultural Bioinformatics Grid in ICAR” is gratefully acknowledged.

REFERENCES

Adnan, A., 2010. Introduction to bioinformatics: Role of mathematics and technology. biotecharticles.com.

Altman, R.B. and T.E. Klein, 2002. Challenges for biomedical informatics and pharmacogenomics. Annual Rev. Pharmacol. Toxicol., 42: 113-133. DOI: 0.1146/annurev.pharmtox.42.082401.140850

Andreas, W., 2007. From bit to it: How a complex metabolic network transforms information into living matter. BMC Syst. Biol., 1: 33-33. DOI: 10.1186/1752-0509-1-33

Bansal, A.K., 2005. Bioinformatics in microbial biotechnology – a mini review. Microb. Cell Fact. 4: 19-19. DOI: 10.1186/1475-2859-4-19

Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell and B.A. Rapp *et al*., 2000. GenBank. Nucl. Acids Res., 30: 17-20. DOI: 10.1093/nar/30.1.17

Callister, S.J., L.A. McCue, J.E. Turse, M.E. Monroe and K.J. Auberry *et al*., 2008. Comparative bacterial proteomics: analysis of the core genome concept. PLoS ONE, 3: e1542-e1542. DOI: 10.1371/journal.pone.0001542

Curtis, T.P., W.T. Sloan and J.W. Scannell, 2002. Estimating prokaryotic diversity and its limits. Proc. Nat. Acad. Sci., 99: 10494-10499. DOI: 10.1073/pnas.142680199

Fenstermacher, D., 2005. Introduction to bioinformatics. J. Am. Soc. for Information Sci. Technol., 56: 440-446.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton and E.F. Kirkness *et al*., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science, 269: 496-512. DOI: 10.1126/science.7542800

Fulekar, M.H. and J. Sharma, 2008. Bioinformatics applied in bioremediation. Innov. Rom. Food Biotechnol., 2: 28-36.

Gabaldon, T., 2008. Comparative genomics-based prediction of protein function. Methods Mol. Biol., 439: 387-401. DOI: 10.1007/978-1-59745-188-8_26

Gilbert, D., 2004. Bioinformatics software resources. Brief. Bioinform., 5: 300-304. DOI: 10.1093/bib/5.3.300

Gomez, S.M., K. Choi and Y. Wu, 2008. UNIT 8.2 prediction of protein-protein interaction networks. Curr. Protocols Bioinform. DOI: 10.1002/0471250953.bi0802s22

Hagen, J.B., 2000. The origins of bioinformatics. Nat. Rev. Genet., 1: 231-236. PMID: 11252753

Heng, L.S., 2011. The digital side of biology. Asia research news.

Hogeweg, P., 2011. The roots of bioinformatics in theoretical biology. PLoS Comput. Biol., 7: e1002021-e1002021. DOI: 10.1371/journal.pcbi.1002021

Huerta, M., F. Haseltine, Y. Liu, G. Downing and B. Seto, 2000. NIH working definition of bioinformatics and computational biology. Bioinformatics Definition Committee.

Huynen, M., B. Snel, W. Lathe and P. Bork, 2000. Exploitation of gene context. Curr. Opin. Struct. Biol., 10: 366-370. DOI: 10.1016/S0959-440X(00)00098-1

Idekar, T., T. Galitski and L. Hood, 2001. A new approach to decoding life: Systems biology. Annu. Rev. Genom Hum. Gen., 2: 343-372. DOI: 10.1146/annurev.genom.2.1.343

Jones, D.T., 2000. Protein structure prediction in the postgenomic era. Curr. Opin. Struct. Biol., 10: 371-379. DOI: 10.1016/S0959-440X(00)00099-3

Jones, D.T., 2001. Protein structure prediction in bioinformatics. Brief Bioinform., 2: 111-125.

Juretic, D., B. Lucic and N. Trinajstic, 2005. Why focusing on bioinformatics? Period. Biol., 107: 379-383.

Kanehisa, M. and P. Bork, 2003. Bioinformatics in the post-sequence era. Nat. Genet. Suppl., 33: 305-310. DOI: 10.1038/ng1109

Keller, M. and K. Zengler, 2004. Tapping into microbial diversity. Nat. Rev Microbiol., 2: 141-150. DOI: 10.1038/nrmicro819

Khanna, V.K., 2007. Existing and emerging detection technologies for DNA (deoxyribonucleic acid) finger printing, sequencing, bio- and analytical chips: a multidisciplinary development unifying molecular biology, chemical and electronics engineering. Biotechnol. Adv., 25: 85-98. DOI: 10.1016/j.biotechadv.2006.10.003

Liu, M.Y., S. Kjelleberg and T. Thomas, 2011. Functional genomic analysis of an uncultured δ-proteobacterium in the sponge *Cymbastela concentrica*. ISME J., 5: 427-435. DOI: 10.1038/ismej.2010.139

Lopez, R., 2008. Biological data resources at the EMBL-EBI. Rev. Colomb. Biotecnol., 10: 120-128.

Luscombe, N.M., D. Greenbaum and M. Gerstein, 2001. What is bioinformatics? A proposed definition and overview of the field. Method Inform. Med., 40: 346-58.

Marcotte, E.M., M. Pellegrini, H.L Ng, D.W. Rice, T.O. Yeates and D. Eisenberg, 1999. Detecting protein function and protein-protein interactions from genome sequences. Science, 285: 751-753. DOI: 10.1126/science.285.5428.751

McDaniel, A.M., P.L. Benson, G.H. Roesener and J. Martindale, 2005. An integrated computer-based system to support nicotine dependence treatment in primary care. Nicotin Tob. Res., 7: S57-S66. DOI: 10.1080/14622200500078139

Mochida, K. and K. Shinozaki, 2010. Genomics and bioinformatics resources for crop improvement. Plant Cell Physiol., 51: 497-523. DOI: 10.1093/pcp/pcq027

Moret, B.M.E., D.A. Bader and T. Warnow, 2002. High-performance algorithm engineering for computational phylogenetics. J. Supercomput.., 22: 99-111. DOI: 10.1023/A:1014362705613

Nakashima, N., Y. Mitani and T. Tamura, 2005. Actinomycetes as host cells for production of recombinant proteins. Microb. Cell Fact., 4: 7-7. DOI: 10.1186/1475-2859-4-7

Neerincx, P.B.T. and J.A.M. Leunissen, 2005. Evolution of web services in bioinformatics. Brief. Bioinform., 6: 178-188. DOI: 10.1093/bib/6.2.178

Nurse, P., 2008. Life, logic and information. Nature, 454: 424-426. DOI: 10.1038/454424a

Oro, J., S.L. Miller and A. Lazcano, 1990. The Origin and early evolution of life on earth. Ann. Rev. Earth Planet. Sci., 18: 317-356. DOI: 10.1146/annurev.ea.18.050190.001533

Ouzounis, C., 2002. Bioinformatics and the theoretical foundations of molecular biology. Bioinformatics, 18: 377-378. DOI: 10.1093/bioinformatics/18.3.377

Patil, S.D., D.G. Rhodes and D.J. Burgess, 2005. DNA-based therapeutics and DNA delivery systems: A comprehensive review. AAPS J., 07: E61-E77. DOI: 10.1208/aapsj070109

Patterson, S.D., 2003. Data analysis—the Achilles heel of proteomics. Nat. Biotechnol., 21: 221-222. DOI: 10.1038/nbt0303-221

Pawlowski, K., L. Rychlewski, B. Zhang and A. Godzik 2001. Fold predictions for bacterial genomes. J. Struct. Biol., 134: 219-231. DOI: 10.1006/jsbi.2001.4394

Pearson, W.R. and D.J. Lipman, 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci., 85: 2444-2448.

Petrosino, J.F., S. Highlander, R.A. Luna, R.A. Gibbs and J. Versalovic 2009. Metagenomic pyrosequencing and microbial identification. Clin. Chem., 55: 856-866. DOI: 10.1373/clinchem.2008.107565

Piatnitskii, M.A., A.V. Lisitsa and A.I. Archakov, 2009. Prediction of functionally related proteins by comparative genomics in silico. Biomed. Khim., 55: 230-246. DOI: 10.1134/S1990750809040015

Prabha, R., D. P. Singh and D.K. Arora, 2011. Whole genome phylogeny of Prochlorococcus marinus group of cyanobacteria. Proceedings of 5th Asian Young Researchers Conference on Computational and Omics Biology in Daejeon, (COBD' 11), Korea, pp: 29-29.

Primetta, F., S.M. Antonio, M. Caterina and T. Valeria, 2009. From DNA sequence to plant phenotype: bioinformatics meets crop science. Curr. Bioinform., 4: 173-176. DOI: 10.2174/157489309789071066

Rao, V.S., S.K. Das, V.J. Rao and G. Srinubabu, 2008. Recent developments in life sciences research: Role of bioinformatics. Afr. J. Biotechnol., 7: 495-503.

Rhee, S.Y., J. Dickerson and D. Xu, 2006. Bioinformatics and its applications in plant biology. Annu. Rev. Plant Biol., 57: 335-360. DOI: 10.1146/annurev.arplant.56.032604.144103

Rogozin, I.B., K.S. Makarova, D.A. Natale, A.N. Spiridonov and R.L. Tatusov *et al*., 2002. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucl. Acids Res., 30: 4264-4271. DOI: 10.1093/nar/gkf549

Salzberg, S.L., A.L. Delcher, S. Kasif and O. White, 1998. Microbial gene identification using interpolated Markov models. Nucl. Acids Res., 26: 544-548. DOI: 10.1093/nar/26.2.544

Searls, D.B., 2000. Bioinformatics tools for whole genomes. Ann. Rev. Genomics Hum. Genetics, 1: 251-279. DOI: 10.1146/annurev.genom.1.1.251

Singh, D.P., R. Prabha, M.S. Yandigeri, D.K. Arora, 2011. Cyanobacteria-mediated phenylpropanoids and phytohormones in rice (*Oryza sativa*) enhance plant growth and stress tolerance. Antonie Van Leeuwenhoek, 100: 557-568. DOI: 10.1007/s10482-011-9611-0

Stelling, J., 2004. Mathematical models in microbial systems biology. Curr. Opin. Microbiol., 7: 513-518. DOI: 10.1016/j.mib.2004.08.004

Sugden, A. and E. Pennisi, 2000. Diversity digitized. Science, 289: 2305-2305.

Tiwari, S., P. Singh, R. Tiwari, K.K. Meena and M. Yandigeri *et al*., 2011. Salt-tolerant rhizobacteria-mediated induced tolerance in wheat (*Triticum aestivum*) and chemical diversity in rhizosphere enhance plant growth. Biol. Fertil. Soils, 47: 907-916. DOI: 10.1007/s00374-011-0598-5

Tomitani, A., A.H. Knoll, C.M. Cavanaugh and T. Ohno, 2006. The evolutionary diversification of cyanobacteria: Molecular–phylogenetic and paleontological perspectives. Proc. Nat. Acad. Sci., 103: 5442-5447. DOI: 10.1073/pnas.0600999103

Varshney, R.K., A. Graner and M.E. Sorrells, 2005. Genomics-assisted breeding for crop improvement. Trends Plant Sci., 10: 621-630. DOI: 10.1016/j.tplants.2005.10.004

Vassilev, D., J. Leunissen, A. Atanassov, A. Nenov and G. Dimov, 2005. Application of bioinformatics in plant breeding. Biotechnol. Biotechnol. Equip., 19: 139-152.

Wollenweber, B., J.R. Porter and T. Lubberstedt, 2005. Need for multidisciplinary research towards a second green revolution. Curr. Opin. Plant. Biol., 8: 337-41. DOI: 10.1016/j.pbi.2005.03.001