2

BIOINFORMATICS: AN OVERVIEW

Pramod Tandon* and Pallavi Bhattacharjee

Bioinformatics Centre, North Eastern Hill University, Shillong–793 022
Email: tandon1@sancharnet.in, profptandon@yahoo.com

ABSTRACT

Bioinformatics is the symbiotic relationship between computational and biological sciences. With all areas of research, bioinformatics has become a complex field that stretches across vast domains of scientific investigation. No longer it is simply tied to the sequencing efforts, but instead has become a vibrant discipline that has impacted almost every aspect of biological research from algorithm development to knowledge management and everything in between. Biologists have finally caught up with the computer age and are finding the value in getting meaningful data quickly and integrating disparate data types to gain new insights. There are several efforts throughout the bioinformatics community to standardize the way biologists describe genes, proteins, biological pathways and processes through the use of ontologies and structured vocabularies. The development of new methods for analyzing single and multiple sequences within a single or multiple experiments and the integration of data through various database architectures that facilitate portability and sharing are all hot topics. (Fenstermachar, 2005) Knowledge management has become a central theme of bioinformatics requiring the need for information theorists and specialists to assist with the overwhelming tasks of keeping the data useful and meaningful for biological researchers.

Keywords: Sandwich beam, free vibration, frequency parameter, loss parameter

INTRODUCTION

"Understanding nature's mute but elegant language of living cells is the quest of modern molecular biology. From an alphabet of only four letters representing the chemical subunits of DNA, emerges a syntax of life processes whose most complex expression is man. The unraveling and use of this 'alphabet' to form new 'words and phrases' is a central focus of the field of molecular biology. The staggering volume of molecular data and its cryptic and subtle patterns have led to an absolute requirement for computerized databases and analysis

is a complex, dynamic, three-dimensional molecule. And yet we represent all of this as a simple string of the characters A, C, G and T. This is a remarkable abstraction. Most of the processes involving genes that we know about have been discovered using this grossly simplified representation of reality. It is the perfect representation for computer analysis, and without it we could never have approached a project on the scale of the Human Genome.

- ***Concept of Similarity.*** Evolution has operated on every sequence that we see today. It conserves genes that encode important proteins and sequences that are involved in gene regulation. Sequences that encode useful functions are transferred, like code modules, from one organism to another. Because of evolution, similar sequences have similar functions. Algorithms for comparing sequences and finding similar regions are at the heart of bioinformatics. At many different levels, they are used to find genes, determine their functions, study their regulation and assess how they, and entire genomes, have evolved over time.
- ***Bioinformatics is not a Theoretical Science.*** Bioinformatics is driven by real time data, which, in turn, are driven by the needs of biology. Relatively few researchers have the luxury to develop algorithms and theories in the traditional academic sense. Most people are fully consumed in the day-to-day management and analysis of data. [*http://www.macdevcenter.com*]

We have lot of data in biosciences. The introduction of automated DNA sequencing in the early 1990s created what was, at that time, 'a torrent of sequence data'. But it was the Human Genome Project, with its massive automation, production lines, and money that really opened the floodgates in the past few years. Compare the rate of growth of sequenced data in GenBank and the NIH (National Institute of Health) sequence database to Moore's Law, the well-known measure of technical advancement, and one can well-appreciate the challenge facing biology. The challenge just does not stop at sequences! Microarray technologies that are capable of measuring expression of thousands of genes in a single experiment, have developed over the past decade and now produce huge amounts of data. New techniques for looking at genetic variations in large human populations, and for identifying interactions between sets of proteins in cells, are pouring data onto file servers around the world. Bioinformatics is charged with managing and making sense of all of these data, keeping pace with both data production and technology development. Hence, there is plenty of work to go around.

BIOINFORMATICS AND ITS SCOPE

Bioinformatics uses advances in the area of computer science, information science, computer and information technology, communication technology to solve complex problems in life sciences and particularly in biotechnology. Data capture, data warehousing and data mining have become major issues for biotechnologists and biological scientists due to sudden growth in quantitative data in biology. Advancements in information technology, particularly the Internet, are being used to gather and access an ever-increasing information in biology and

Gene Therapy

Very soon the potential for using genes themselves to treat disease may become a reality. Gene therapy is the approach used to treat, cure or even prevent disease by changing the expression of an individual's genes.

Drug Development

With an improved understanding of disease mechanisms and using computational tools to identify and validate new drug targets, more specific medicines that act on the cause, not merely the symptoms, of the disease can be developed. These highly specific drugs promise of having fewer side effects than many of today's medicines.

Important and exciting potential of bioinformatics is found in the drug discovery, which is highly accelerating. Use of According to IMS health the worldwide pharmaceutical market, in 2000 totaled US $ 317 billion (approx.). By 2020 this figure is expected to be US $ 3 trillion. (Murthy, 2003)

Microbial Genome Applications

Microorganisms are ubiquitous (i.e., they are found everywhere). They have been found surviving and thriving in extremes of heat, cold, radiation, salt, acidity and pressure. They are present in the environment, our bodies, the air, food and water. Traditionally, we already use a variety of microbial properties in baking, brewing and food industries. The arrival of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities would have broad and far reaching implications for environment, health, energy and industrial applications.

Waste Cleanup

Deinococcus radiodurans is known as the world's toughest bacteria and it is the most radiation resistant organism known. Scientists are interested in this organism because of its potential usefulness in cleaning up waste sites that contain radiation and toxic chemicals.

Climate Change Studies

Increasing levels of carbon dioxide emission, mainly through the expanding use of fossil fuels for energy, have contributed to global climate change. One method of decreasing atmospheric carbon dioxide levels is to study the genomes of microbes that use carbon dioxide as their sole carbon source.

Alternative Energy Sources

Scientists are studying the genome of the microbe *Chlorobium tepidum*, which has an unusual capacity for generating energy from light. This could open an uncharted territory for tapping energy from biosources.

BIOTECHNOLOGY

The arch eon *Archaeoglobus fulgidus* and the bacterium *Thermotoga maritima* have potential for practical applications in industry and government-funded environmental remediation. These microorganisms thrive in water temperatures above the boiling point and therefore their in-depth study may provide the manufacturing industry with heat-stable enzymes suitable for use in industrial processes. Other industrially useful microbes include *Corynebacterium glutamicum*, which is of high industrial interest as a research object. This microbe is extensively used in the chemical industry for the production of amino acid lysine. *Xanthomonas campestris pv.* is grown commercially to produce the *exopolysaccharide xanthan* gum, which is used as a viscosifying and stabilizing agent in many industries. *Lactococcus lactis* is one of the most important microorganisms involved in the dairy industry. It is a non-pathogenic rod-shaped bacterium that is critical for manufacturing dairy products like buttermilk, yogurt and cheese. This bacterium is also used to prepare pickled vegetables, beer, wine, breads, sausages and other fermented foods. Researchers anticipate that understanding the physiology and genetic make-up of this bacterium will prove invaluable for food manufacturers as well as the pharmaceutical industry, which is exploring the capacity of this bacterium to serve as a vehicle for delivering drugs.

BIODIVERSITY CONSERVATION

Bioinformatics tools are also used for biodiversity conservation. For instance, the molecular marker technology and molecular diagnostics, *in vitro* technologies and cryopreservation techniques have been applied for germplasm conservation (Tandon and Kumaria, 1998, Tandon, 2004, Tandon and Kumaria, 2005). Computerization of huge biodiversity data has developed databases, which is now available in the public domain through the Internet and is easily accessible to one and all (Sugden and Pennisi, 2000).

ANTIBIOTIC RESISTANCE

Scientists have been examining the genome of *Enterococcus faecalis*—the leading cause of bacterial infection among hospital patients. They have discovered a virulence region made up of a number of antibiotic-resistant genes that may contribute to the bacterium's transformation from harmless gut bacteria to a menacing invader. The discovery of the region, known as a pathogenicity island, could provide useful markers for detecting pathogenic strains and help to establish controls to prevent the spread of infection in wards.

FORENSIC ANALYSIS OF MICROBES

Microbial forensics is a relatively new field that can help in solving cases such as bioterrorism attacks, medical negligence and outbreaks of food-borne diseases. Scientists are already using bioinformatics tools for forensic analysis in this field.

EVOLUTIONARY STUDIES

The sequencing of genomes from all three domains of life, eukaryota, bacteria and archaea means that evolutionary studies can be performed in a quest to determine the tree of life and the last universal common ancestor.

CROP IMPROVEMENT

Crop plant networks provide collection of databases and bioinformatics resources for crop plants genomics. These networks have been built to harness the extensive work in genome mapping. This resource facilitates the identification of agronomically important genes by comparative analysis between crop plants and model species, allowing the genetic engineering of crop plants. New nutritional genomics biotechnology tools

HOT TOPICS IN BIOINFORMATICS

Comparative Genomics

This is a high profile field of research as of date, which has evolved from the huge data of the Human Genome Project. The first "tier" of genome sequences (human, rat, mouse, and fruit fly) is now complete and the big sequencing labs are moving on to organisms like the chimpanzee, rhesus macaque, cow, chicken, and sea urchin. This is the essence of comparative genomics. A similar approach to biology was used by Charles Darwin. By comparing the genomes of related species, we would collect a tremendous amount of information about how genomes are organized and how major evolutionary changes takes place. At the level of individual genes, we would uncover novel mechanisms for regulation that were hidden when we just had one sequence to work with.

Single Nucleotide Polymorphisms (SNPs)

Another avenue that has opened up after we have the "reference" human genome is the study of sequence differences between individuals. The genome is full of single nucleotide differences, called *polymorphisms*, or SNPs. Their distribution throughout the genome, their frequency in the human population and their patterns of inheritance make them extremely useful markers for differences between individuals. By measuring sets of SNPs in thousands of individuals and correlating them with the incidence of a disease, we can identify which regions of the genome are involved and eventually pinpoint the genes themselves. The combination of these molecular assays with large clinical studies of populations generates huge amounts of data and a whole new set of challenges for bioinformatics.

Microarrays

Microarray technologies show us which genes are turned on in different cell types in different circumstances. In response to infection, for example, certain cell types will express sets of genes and synthesize certain proteins that respond to the stress. Messenger RNA (mRNA) is like a photocopy of a blueprint that is used in the shop to build a specific type of protein. In a microarray, we can attach sequences from a range of genes to a glass slide in a series of dots, and then bind the mRNA extracted from a population of cells and measure how much binds to each dot. That gives us a snapshot of which genes are being expressed at any given time. For example, if we compare the patterns for mRNA from normal breast tissues and from breast tumor tissues, we can identify proteins that are present only in the tumor. Those proteins are potential targets for cancer treatments, vaccines, and other therapeutics.

Gene expression can be studied at a genome wide scale with the aid of modern microarray technologies. Expression profiling of tens to hundreds of individuals in a genetic population can reveal the consequences of genetic variation. (Alberts *et al.*, 2005)

Systems Biology

The genome gives us all of the genes in an organism, and microarrays tell us which subset is expressed in a particular biological process. However, the bottleneck in understanding biology is shifting to the world of proteins and the interactions between them. That is where systems biology comes in with a slew of novel technologies aimed at seeing the big picture of everything going on in a cell. New advances in mass spectrometry have allowed this established chemical analysis technology to identify the components of complex mixtures of proteins. Inventive chemical labeling techniques provide insight into the transient interactions between different proteins in the cell. This bundle of new technologies is called proteomics. The integration of all of these results with gene expression data and the collective knowledge of cell biology, contained in the scientific literature, becomes another huge challenge. This is leading to exciting work in textual analysis, pathway modeling, and network visualization.

Structural Biology

While abstraction of the DNA sequence works remarkably well in the world of proteins, the nuances of three-dimensional structure are everything. Structural biologists determine the structure of proteins using X-ray crystallography and nuclear magnetic resonance, a slew of heavy numerical methods, and a lot of computing. This field focuses on the details of structure, the dynamics of molecular motion and the specific interactions with drugs and other proteins. Bioinformatics has an uneasy interface with structural biology. However, the distinction is increasingly becoming blurred, as all of these data sources become more integrated.
[*http://www.macdevcenter.com/*]

DATABASES AND SOFTWARE TOOLS FOR BIOINFORMATICS

Web Services

(a) Web services are the programmatic interfaces for application to application communication over World Wide Web.

(b) Web services are available for most of the bioinformatics application domain.

(c) HTML web interfaces are not suitable for programmatic access.

(d) In order to overcome limitations of HTML based tools, XML–based solutions are gaining more and more momentum.

(e) Creating web service work flows requires adequate ontologies.

(f) Combination of centralized and distributed systems offers best of both worlds.

(g) Web services are essential steps in solving interoperability problems.
(Neerincx and Leunissen, 2005)

Molecular Databases

Applied research in bioinformatics is critically dependent on molecular and genetic databases. Such databases are being developed and maintained by various organizations at the global level, which can be used for research on biodiversity and bioscience. The three major organisations working in this field are:

- National Centre for Biotechnology Information (NCBI) in USA
- European Molecular Biology Laboratory (EMBL) in Germany
- DNA Database of Japan (DDBJ) in Japan

The above three databases collaborate with each other through data exchange and information on Internet and by regularly holding databases of interest from a large list of databases, categorized by subject such as sequence, sequence-related metabolic pathway, transcription factor, three-dimensional structure, mapping, mutation etc. SRS provides a homogenous interface to more than 80 biological databases. [*http://www.ebi.ac.uk/services*]

ENTREZ (NCBI)

Entrez is a search and retrieval system, which integrates scientific literature, DNA and protein sequence databases, 3D protein structure and protein domain data, population study datasets, expression data, assemblies of complete genomes and taxonomic information into a tightly interlinked system. [*http://www.ncbi.nlm.nih.gov/Entrez/*]

FASTA

FASTA is sequence comparison software that uses the method of Pearson and Lipman. The basic FASTA algorithm assumes a query sequence and a database over the same alphabet. It searches a DNA sequence in a DNA database or a protein sequence in a protein database. Practically, FASTA is a family of programs, allowing also queries of DNA vs. a protein database, or vice versa. In these variants there is further distinction, which regards the location of gaps: one may assume that gaps occur only in the codon frames corresponding to amino-acid insertion; alternatively, one can assume gap location to be arbitrary, accounting for insertion/deletion of nucleotides. This search tool is preferred for searching nucleotides. [*http://www.ebi.ac.uk/services/*]

BLAST 2.0

Sequenced data are compared to one another using the Basic Local Alignment Search Tool or BLAST (Altschul et al., 1990). It is BLAST that provides a method for rapid searching of nucleotide and protein databases. This search tool is better for proteins than for nucleotides. BLAST information guide is designed to assist new and veteran users in employing NCBI tools such as BLAST and PSI-BLAST in their research. Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. Both functional and evolutionary information can be inferred from well-designed queries and alignments. Since the BLAST algorithm detects local as well as global alignments, regions of similarity embedded in otherwise unrelated proteins could be detected. Both types of similarity may provide important clues to the function of uncharacterized proteins. Depending on the type of sequences to compare, there are different programs:

- blastp compares an amino acid query sequence against a protein sequence database
- blastn compares a nucleotide query sequence against a nucleotide sequence database
- blastx compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
- tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.
  [*http://www.ebi.ac.uk/services/*]

EMBOSS

EMBOSS (European Molecular Biology Open Software Suite) is a software-analysis package. It can work with data in a range of formats and also retrieve sequence data transparently from the Web. Extensive libraries are also provided with this package, allowing other scientists to release their software as open source. It provides a set of sequence-analysis programs, and also supports all UNIX platforms.

Clustalw

It is a fully automated sequence alignment tool for DNA and protein sequences. It returns the best match over a total length of input sequences, be it a protein or a nucleic acid.

RasMol

It is a powerful research tool, which displays the structure of DNA, proteins, and smaller molecules. Protein Explorer, a derivative of RasMol, is an easier to use program.

PROSPECT

PROSPECT (PROtein Structure Prediction and Evaluation Computer ToolKit) is a protein-structure prediction system that employs a computational technique called protein threading to construct a protein's 3D model.

Pattern Hunter

Pattern Hunter, based on Java, can identify all approximate repeats in a complete genome in a short time using little memory on a desktop computer. Its features are its advanced patented algorithm and data structures, and the java language used to create it.

COPIA

COPIA (COnsensus Pattern Identification and Analysis) is a protein structure analysis tool for discovering motifs (conserved regions) in a family of protein sequences. Such motifs can be then used to determine membership to the family for new protein sequences, predict secondary and tertiary structure and function of proteins and study evolution history of the sequences.

APPLICATION OF PROGRAMMES IN BIOINFORMATICS

JAVA

Since research centers are scattered all around the globe ranging from private to academic settings, and a range of hardware and operating systems are being used, Java is emerging as a key player in bioinformatics. 'Physiome Sciences' computer-based biological simulation technologies and 'Bioinformatics Solutions' Pattern Hunter are two examples of the growing adoption of Java in bioinformatics.

Perl

String manipulation, regular expression matching, file parsing, data format interconversion, etc. are the common text-processing tasks performed in bioinformatics. Perl excels in such tasks and is being used by many developers. Yet, there are no standard modules designed in Perl specifically for the field of bioinformatics. However, developers have designed several of their own future, some bioinformatics tasks may prove to be more effectively implemented in java or python.

BIOINFORMATICS—THE NATIONAL SCENARIO

Major organisations dealing with bioinformatics research are:

(a) Biotechnology Information System (BTIS)
(b) European Molecular Biology Network (EMBnet) India node
(c) Biotech Consortium India Limited (BCIL)

BIOTECHNOLOGY INFORMATION SYSTEM (BTIS)

Recognizing the importance of information technology for pursuing advanced research in modern biology and biotechnology, the Department of Biotechnology (DBT, India) launched a bioinformatics programme during 1986 – 87. This program was envisaged to build a distributed database and network organization for biotechnological research and was named as BTISnet.

The entire network of BTIS has emerged as a very sophisticated scientific infrastructure for bioinformatics involving state-of-the-art computational and communication facilities. The computer communication network, linking all the bioinformatics centers under the BTIS, plays a vital role in the success of the bioinformatics programme. Database development, R & D activities in bioinformatics, human resource development and a variety of services in support of biotechnology R & D programme and projects, has made this network very popular and useful to the scientific community. BTIS enjoys excellent cooperation from various Government agencies, like the National Informatics Centre (NIC). It has made it possible for the network to assume the role of a closed user group representing a scientific grid in various inter-disciplinary subjects of biotechnology encompassing, agriculture, health and environment, besides other related subjects of scientific importance.

The contributions made by the scientists and academicians at the University departments of the UGC and national laboratories and institutions of the Council of Scientific and Industrial Research (CSIR) and Indian Council for Agricultural Research (ICAR) provides a variety of information resources on the Internet. More than 100 databases dealing with different aspects of R & D efforts in biotechnology are now available on the network. Several major international databases for application to genomics and proteomics have been established in the form of mirror sites under the National Jai Vigyan Mission.

Four mirror sites for mirroring important biological databases are being established at Indian Institutes of Science, Jawaharlal Nehru University (JNU), Pune University and Institute of Microbial Technology to promote and support R & D activities in Genomics and Proteomics, the two emerging fields of biotechnology requiring critical support of genomic databases. With these resources now available on the BTISnet, it has become a single largest information resource for all references to biotechnology related literature, scientific data, patent information, policy matters and related issues in India.

BTISnet is the first major Satellite and Terrestrial network on Biotechnology in the country, networking 65 Bioinformatics Centers through satellite and terrestrial links provided by NICNET. Three major network service providers in the country viz., NICNET, ERNET and VSNL, provide Internet access. The BTISnet permits remote login, file transfer, e-mail, etc. as well as connecting to various international networks, which are providing updated information support on all aspects in biotechnology ranging from bibliographic information to sequence analysis and management information. A Biotechnology Patent Facilitating Cell has been established, which uses the facilities of the BTIC to provide full-scale patent search services. [*http://www.btisnet.nic.in/*]

Following centers of the BTIS are engaged in biotechnological research and development:

- One Apex Centre
- 5 Centres of Excellence (COE)
- 10 Distributed Information Centres (DICs)
- 46 Sub-Distributed Information Centres (Sub-DICs)

Apex Center: The Apex Center called the Biotechnology Information Center (BTIC) at DBT, New Delhi, is coordinating the activities of other DICs and Sub-DICs. The BTIS secretariat working under the DBT is situated at this Apex Centre.

Centres of Excellence (COE): The missions of the COE will be to carry out advance research in bioinformatics, provide doctoral and post-doctoral training, develop new solutions to complex biological problems and provide highly trained manpower to the bioinformatics industry in India. It is envisaged that these centres will utilize the advancements in biotechnology and biological sciences to help India to become the leader in bioinformatics.

DICs*:* Ten DICs have been established to provide subject-oriented information to other institutions and individual users interested in a particular field related to biotechnology.

Sub-DICs: Sub-DICs have been setup in a large number of R & D institutions and universities. While the DICs act as repository of information in their respective fields, Sub-DICs particularly serve these facts to the scientists working in R & D centers and universities.

All these centers are interlinked through satellite communication system, each providing information support in specific areas of biotechnology and helping in the diffusion of scientific information across the network.

Functions of BTISnet Centers are as follows:

- To provide a computer-based information storage and retrieval system of database that collects structured information generated by research and industrial institutions in the identified fields of biotechnology, continually update the databases and make the information available to the users.
- To function as an active network node, where the scientists can communicate with each other in an interactive and discussive mode and actively initiate dialogue among groups with common interest.
- To provide online or offline retrieval service and communication link with international databases.
- To develop software packages and databases specific to user needs.
- To conduct training courses in the specialized areas.

Six national facilities for Interactive Graphics based computational requirements for molecular modeling and other biocomputational needs and four long-term educational programmes started during 8th Plan are additional components of the programme. The national biocomputing facilities under the umbrella of "National Facilities on Interactive Graphics and Molecular Modeling" have been established with the task of providing discipline-wise facilities to the scientists working in the areas of molecular structure modeling, 3D structures, active site modeling, crystal structures, conformational analysis, protein and DNA structures and interactions, homology studies and like.

BIOINFORMATICS IN INDIA

India has been very fast in reacting to the global upsurge in interest on computational biology and bioinformatics. Sequencing experiments and Drug discovery is a very costly and intensive process for which developing nations like India doesn't have the capacity to be in the forefront of experimental research or Industrial R & D in this field. However, Indian Research scholars in collaboration with other groups from different parts of the world are very active in this field of research. Indian Pharma Companies have also significantly increased their R & D Budget and are actively pursuing the field of Drug Discovery in a scaled down way. Bioinformatics have proved to be of immense help in this cost intensive research domain. India with its vast genomic diversity provides a very fertile experimental ground for doing genomic experiments. Indian Industry is concentrating on the Business Process Outsourcing Model for encompassing this area of research and a substantial amount of work is expected to be carried out in India by the various Pharma Giants of the world. India with its vast talent pool is all poised to repeat the Information Technology success story in the field of bioinformatics.

OUTSOURCING

India is gradually combining its strength in biotechnology and IT to attract outsourcing contracts in bioinformatics. But the challenge for Indian bioinformatics companies is to create proprietary products that will generate high-margin revenues and therefore contribute to the economic growth of the country. Already some companies in India have made forays into the local and global market. For example, Strand Genomics from Bangalore has licensed its micro array gene expression analysis software to antibody company Abgenix in Fremont, California and Lion Bioscience Research in Cambridge, Massachusetts, has licensed NetPro, a proprietary protein interaction database of Molecular Connections, in Bangalore, to conduct drug target identification research exclusively for Bayer, Leverkusen, Germany.

Would biotech companies choose outsourcing to India for cost reduction? James Featherstone, head of European consulting at WoodMackenzie's life science practice in London explains that biotech companies outsourcing in India fear that their IP may not be adequately protected. Companies have still to be convinced that the Indian government will enforce the new patent protection regime, as part of the trade related aspects of Intellectual Property rights (TRIPS) requirements from the World Trade Organization. Until now, uncertainty concerning IP protection has prompted companies to outsource tasks in the drug discovery process up to a certain level and analyze in-house data that are more proprietary. Another concern, according to Roy Drucker, general manager of professionals has not, it says, "Companies on an average have to go through 100 short-listed resumes to finally pick one qualified person." Human capital in India is available in abundance and this could prove to be a boon if adequate efforts were made to train, preserve and retain the available manpower. The industry strongly feels that the available manpower is quite sufficient in numbers but it seriously lacks in terms of skills, the report says.

Since the need is for experienced personnel, the educational institutions should include practical training in their curriculum so that the students get hands-on experience. It said most institutes and colleges were "misusing" the hype surrounding the subject. The mushrooming of fly-by-night training institutes had compounded the problem, the report says.

Another issue is that some companies face shortage of funds and infrastructure. The turn around time for an average biotech industry to breakeven would be around three to five years. Most of the venture capitals and other sources of funding would not be very supportive, especially if the company is not part of a larger group venture. Hence, the active participation of the government in building infrastructure and funding small and medium entrepreneurs becomes critical in the overall success of any biotech entrepreneurship.

Other barriers include expensive and unreliable power, and high prices for transportation and real estate. While low wages more than compensate for high infrastructure costs, as the country becomes more integrated into the global economy, skilled manpower costs are likely to increase. [*http://www.hindu.com/2005/07/25/stories/2005072502180900.htm*]

INBIOS

Inbios, or Bioinformatics Society of India, is a registered non-profit society that has been set up to promote the rapidly emerging field of bioinformatics in India. The goals of this society are to promote awareness of bioinformatics and related disciplines in India and to serve as a resource for those aspiring to pursue this field. In addition, Inbios strives to serve as a bridge between the industry and the academia to ensure that India becomes a major hub for bioinformatics in the near future. Personnel involved in running this ambitious effort are either eminent scientists, bioinformatics researchers or professionals in the field of life sciences and information technology. They all volunteer their time to help lay a strong foundation for the future of bioinformatics in India.

This society is aimed at making bioinformatics a long term success in India. The current growing interests and trends have serious repercussions in the years to come. This society basically aims at being a bridge between the educational and corporate sector.

MAJOR KEY PLAYERS IN INDIAN BIOTECHNOLOGY INDUSTRY

Studies by various independent agencies of the world indicate that India will be a potential superpower in bioscience in the next decade. This prediction has been arrived at, after due consideration of factors like biodiversity, human resources, infrastructure facilities and government's initiatives. Pharmaceutical firms and research institutes of India are looking forward for cost-effective and high-quality research, development and manufacturing of drugs with more speed. This sector is the quickest growing field in the country. The promising start-ups are already there in Bangalore, Hyderabad, Pune, Chennai, and Delhi. There are over 200 companies functioning in these places. IT majors such as Intel, IBM, and Wipro are getting into this segment spurred by the promises in technological developments and huge profit margins. Some of the leading private companies related to bioinformatics research are:

Bharat Biotech International Ltd. and *Biological E Ltd.* based at Hyderabad, acquired competency in recombinant protein R & D and created their own manufacturing technologies. In addition to these startups, India's fifth largest pharmaceuticals company, Wockhardt Ltd. (Mumbai), manufactures and sells hepatitis B vaccine based on licenses from Rhein Biotech NV (NMarkt:RBO, Maastricht, the Netherlands).

Shantha Biotechnics Pvt. Ltd., manufactures interferon IIA and has five other proteins in the pipeline. Pfizer Inc. distributes Shantha's hepatitis B vaccine in India and has right of first refusal on the company's new products. This company is thus engaged in the process of developing low priced hepatitis B vaccines to compete with imported products that most Indians could not afford.

Dr. Reddy's Laboratory, secured lab space and technical assistance from the Centre for Cellular and Molecular Biology (CCMB, Hyderabad) and funding from the Bank of Oman for construction of a manufacturing plant. Two years later, India's first genetically engineered vaccine was on the market, selling for 20 times less than the imported product.

Biocon Ltd is the first and largest biotech company in India. Using a proprietary solid surface fermentation technology, Biocon retains its roots in the world markets for food and industrial enzymes. It has also turned its fermentation technology to drug manufacturing, developing a pipeline of statins and winning U.S. FDA approval to market generic lovastatin for cholesterol reduction. Biocon has also created contract drug discovery, clinical trials, genomics and chemistry research units and sister companies.

Ranbaxy, India's largest pharma company with also view innovation as key to its future. The company has branched out from creating new formulations of existing drugs and has half a dozen molecules under development. Ranbaxy has collaborations with several US and European companies to develop new formulations and delivery technologies.

COLLABORATION EFFORTS BETWEEN THE PUBLIC AND PRIVATE SECTOR

Traditional barriers between government funded research laboratories and industry are crumbling, enabling Indian biotech companies to tap into expertise, resources and manpower. CSIR operates a network of government laboratories with mandates to collaborate with industry. CSIR's Centre for Cellular and Molecular Biology incubated India's first recombinant protein product, hepatitis B vaccine from Shantha Biotechnics, and it has numerous industrial relationships, including a joint venture with Biological E and Amersham Pharmacia to build DNA microarrays.

US Silicon Valley and European collaborations are tapping into Indian expertise. CCMB recently won a contract from Onconova Therapeutics Inc. (Princeton, N.J.) to create transgenic fruit fly high throughput assay systems and use them to screen drug targets for anticancer effects.

Nicholas Piramal India Ltd., a Mumbai pharmaceutical company, has formed a partnership with the government's Centre of Biotechnology to conduct genomic research with the nation's diverse populations and to explore India's traditional medicines.

Strand, a spin-off from the Indian Institute of Science, is developing a suite of tools for genomics annotation, in silico research and macromolecular structure analysis.

AlphaGene Inc. (Woburn, Mass.) has announced a collaboration to use bioinformatics technology from Questar Bioinformatics Ltd. (Hyderabad) to mine AlphaGene's protein library. Questar will provide support for structure determination, pathway identification, and small molecule library development.

SCOPE IN BIOTECHNOLOGY FOR INDIAN PRIVATE SECTOR COMPANIES

The past two years have seen many large multinational pharmaceutical companies acquiring other small companies in the biosciences sector. Considering the fact that the local market is presently less mature than those in the US and Europe, more aggressive growth is forecasted beyond 2005. Enterprise applications including data warehousing, knowledge management and storage are being pursued by companies involved in bioscience related projects as on date.

It is expected that IT spending in biosciences in India will soar in recent future, mainly in the areas of system clusters, storage, application software, and services. Also the government's present initiative on life science focus provides a great deal of necessary backbone to develop and deliver innovative products and technologies. This focus will also help to build fast-growing and lucrative enterprises, attract international investment, and create additional high-value employment opportunities. Hence the focus of the IT sector should be on products and services that align with bioscience needs. Demonstrating a true understanding of the IT requirements of biotechnology processes is the key for IT suppliers to bridge the chasm that currently exists between IT and Science.

While advances in technology have reduced the relevance of low cost manpower, India still can apply a different kind of human resources to genome discovery. The diverse range of ethnic populations in India can be valuable in providing information about disease predisposition and susceptibility, which in turn will help in drug discovery.

However, as India lacks the records of clinical information about the patients, sequence data without clinical information will have little meaning, and hence partnership with clinicians is essential. It is essential that new drugs are developed by the Indian companies and not merely supply genetic information and data to the foreign companies, who would then use this information to discover new molecules. India is well-placed to take the global leadership in genome analysis, as is in a unique position in terms of genetic resources.

The genomic data provides information about the sequence, but it doesn't give information about the function. It is still not possible to predict the actual 3-D structure of proteins. This is a key area of work as tools to predict correct folding patterns of proteins will help drug design research substantially. India has

Rashidi HH and Buehler LK (2000). Introduction to bioinformatics; in *Bioinformatics Basics – Applications in Science and Medicine,* 2-3 (CRC Press LLC, Florida).

Neerincx PBT and Leunissen JAM (2005). Evolution of Web Services in Bioinformatics; *Briefings in Bioinformatics* 6**(2)** 178-188.

Sugden A and Pennisi E (2000). Diversity digitized; *Science* 289 (5488), 2305.

Tandon P and Kumaria S (1998). Threats to plant diversity in high altitude of North-East India and conservation of rare and endangered plants using biotechnological approaches; in *Science at High Altitude,* pp. 140-147 Eds. S Saha, PK Ray and B Sinha (Allied Publishers Ltd. India).

Tandon P (2004). Role of biotechnology in conservation of plant genetic resources in the 21st Century-an Indian perspective; in *Platinum Jubilee Lectures, 87th and 88th Session of Indian Science Congress Association,* pp. 40-67 Eds. SP Banerjee and SP Mukherjee (Auto Print and Publicity House, Kolkata, India).

Tandon P and Kumaria S (2005). Prospects of plant conservation biotechnology in India with special reference to northeastern region; in *Biodiversity: Status and Prospects,* pp 79-92 Eds. P Tandon, M Sharma and R Swarup (Narosa Publishing House, New Delhi).