**Biodiversity and Environmental Biotechnology**
*Padmanabh Dwivedi, S.K. Dwivedi & M.C. Kalita*
Scientific Publishers (India), Jodhpur (2007) pp. 225-250.
E-mail: info@scientificpub.com www.scientificpub.com

## Chapter-10

# Bioinformatics in Biotechnological Research

*Pramod Tandon* and Pallavi Bhattacharjee*

Bioinformatics Centre, North Eastern Hill University, Shillong 793 022
*Corresponding author: tandon1@sancharnet.in

Bioinformatics is defined as a discipline that generates computational tools, databases and methods to support genomic and post genomic research. It comprises the study of DNA structure and function; gene and protein expression; protein production, protein structure and function; genetic regulatory systems and clinical applications. Biotechnology can be broadly termed as "using living" organisms or their products for commercial purposes (Sharma, 2003). Biological research has generated a huge quantity of data, which requires high-throughput and large-scale technologies to manipulate this data. This trend is growing exponentially with a shift in emphasis from individual biomolecules to analysis of how they interact in complex networks, which control the developmental and physiological processes of whole biological systems. Biotechnology is a global reality and essential for the objective of developing, dynamic, and innovative knowledge-based economies. The success of any knowledge-based economy rests upon the generation, diffusion and application of new knowledge (Chakraborty *et al.,* 2005). The ultimate aim of this research is to relate its results directly or indirectly to human health. This transition has increased the importance of bioinformatics and raises key challenges, which make it imperative that computer scientists work closely with biologists to refine existing bioinformatics tools and develop new ones.

**Importance of bioinformatics in biotechnological research**

In the last decade, numerous innovations have seen light and as a consequence a new biological research paradigm has evolved. This new paradigm is information-heavy and computer-driven. Now genetic information is computerized as databases and their sizes are steadily growing. Biologists need effective and efficient computational tools to store and retrieve information such as bibliographic or biological information from the databases, to analyze the sequence patterns they contain and to extract the biological knowledge the sequences have. On the other hand, there is a strong need for mathematical methods and computational techniques for challenging computational tasks and to construct evolutionary trees from the sequence data. These tools are also used to learn basic facts about biology. For example, sequences of DNA are used to code proteins for greater understanding of genes and how they influence diseases.

Another example elaborates the use of biotechnology in biological research. Biology employs a digital language for representing information using four basic alphabets (A, C, G, and T). All chromosomes in a cell of an organism are represented and identified using these alphabets. The demanding challenge is to determine how digital representation of chromosomes is converted into three-dimensional and sometimes four-dimensional languages of living and breathing organisms. Here use of bioinformatics simplifies the entire challenge. It was found that manual completion of the above-mentioned tasks was nearly impossible due to the massive volumes of data and the preciseness involved. Hence use of computers became mandatory to achieve results with acceptable perfection. Bioinformatics deals with designing and deploying efficient software tools for accomplishing complex computable tasks in a fast and precise manner. Hence, to bridge the gap between the real world of biology and precise logical nature of computers, an interdisciplinary perspective is essential.

Bioinformatics as a field of study is becoming increasingly important due to the interest of the pharmaceutical industry in genome sequencing projects, which comprise a key segment of the discipline. There is a vital need to harness this information for medical diagnostic and therapeutic uses and there are opportunities for other industrial applications also. This field is evolving rapidly, which makes it challenging for biotechnology professionals to keep up with recent advancements. Other applications include sequence alignment, protein structure prediction, metabolic networks, morphometrics and virtual evolution. Protein structure prediction is a very promising and important application of bioinformatics. The amino acid sequence of a protein, so-called primary structure, can be easily determined from sequence on the gene that codes for it. There are many successful examples of structure based drug design. Development of anti-retroviral therapy for the treatment

of AIDS or the development of anti-cancer drugs have been significantly accelerated by the knowledge of the three dimensional structure of the target proteins. Today, structure based drug design is an undertaking in every major pharmaceutical set up. Bioinformatics can help in characterizing targets (classification and sub-classification of protein families), understanding targets (analysis of their behaviour in a larger biochemical and/or cellular context), developing targets (production, detoxification, the stratification on patient population) and other gene-based variations. Advances in genomics and DNA analysis are leading to new developments in molecular electronics and bio-computing. Biomolecules or macromolecules like carbohydrates, lipids, nucleic acids and proteins form an important area of biotechnological research with application in Biosensors and DNA chips and Biochips.

**The Human Genome Project**

The Human Genome project is a classical example reflecting the application of bioinformatics in biotechnological research. This project formally began in 1990 in United States of America. The project is a 13-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. Originally the project was planned to last for 15 years, but rapid technological advances accelerated the timeline and the project was completed in 2003. Sequence and Analysis of the human genome working draft was published in February 2001, in Nature and Science. The main goals of the project were to:

(i) Identify all the approximate 30,000 genes in human DNA.

(ii) Determine the sequences of the 3 billion chemical base pairs that make up human DNA and store this information in databases.

(iii) Improve tools for data analysis.

(iv) Transfer related technologies to the private sector.

(v) Address the ethical, legal and social issues (ELSI) that may arise from the project.

To help achieve these goals, researchers studied the genetic make up of several non-human organisms. These included the common human gut bacterium *Escherichia coli*, the fruit fly and the laboratory mouse. A unique aspect of the project was to address the ELSI implications through a first ever large-scale scientific undertaking which had arisen from the project. Another important feature of the project was the federal government's long-standing dedication to the transfer of technology to the private sector. By licensing technologies to private companies and awarding grants for innovative research, the project has catalyzed a multibillion-dollar U.S. biotechnology industry and has fostered the development of many new

medical applications. Technology and resources generated by the Human Genome Project and other genomic research already are having major impacts on research across the life sciences. [http://www.ornl.gov/sci/techresources/Human_Genome /project/journals/journals.html]

## Application of bioinformatics

Basically there are two ways of structuring the field of bioinformatics. One is *intrinsically,* by the type of problem that is under consideration. Here, the natural way of structuring is by layers of information that are complied starting from the genomic data and working out way towards various levels of the phenotype. The second is *extrinsically*, by the medical or pharmaceutical application scenario to which bioinformatics contributes and by the type of biological experimentation that it supports (Lengauer, 2002).

Some of the prominent applications of bioinformatics in biotechnological research are given below:

### Microbial Genomics

a. Rapidly detect and treat pathogens (disease-causing microbes) in clinical practice
b. Develop new energy sources (biofuels)
c. Monitor environments to detect pollutants
d. Protect citizenry from biological and chemical warfare
e. Clean up toxic waste safely and efficiently

### Bioarchaeology, Anthropology, Evolution and Human migration

a. Study evolution through germline mutations in lineages
b. Study migration of different population groups based on maternal genetic inheritance
c. Study mutations on the Y chromosome to trace lineage and migration of males
d. Compare breakpoints in the evolution of mutations with ages of populations and historical events

### DNA Identification

a. Identify potential suspects whose DNA may match evidence left at crime scenes
b. Exonerate persons wrongly accused of crimes
c. Identify crime, catastrophe and other victims

d. Establish paternity and other family relationships
e. Identify endangered and protected species as an aid to conservation strategies
f. Detect bacteria and other organisms that may pollute air, water, soil and food
g. Match organ donors with recipients in transplant programs
h. Determine pedigree for seed or livestock breeds
i. Authenticate consumables such as caviar and wine

**Agriculture, livestock breeding and Bioprocessing**

a. Grow disease-, insect- and drought-resistant crops
b. Breed healthier, more productive, disease-resistant farm animals
c. Grow more nutritious produce
d. Develop biopesticides
e. Incorporate edible vaccines into food products
f. Develop new environmental cleanup uses for plants like tobacco
g. Food safety by using modern biotechnology (Sharma, 2005)

**Gene therapy**

The use of genes themselves to treat disease or enhance particular traits has tremendous potential. This largely experimental field commonly referred to as gene transfer or gene therapy, holds potential for treating or even curing such genetic and acquired diseases as cancers, AIDS and Parkinson's disease, etc by using normal genes to supplement or replace defective genes or bolster a normal function such as immunity. Gene therapy studies also involve other multigenic and monogenic, infectious and vascular diseases. Most current protocols are aimed at establishing the safety of gene-delivery procedures rather than effectiveness.

**Gene testing**

DNA-based gene tests are among the first commercial medical applications of the new genetic discoveries. Gene tests can be used to:

- Diagnose disease
- Confirm a diagnosis
- Provide prognostic information about the course of disease
- Confirm the existence of disease in asymptomatic individuals with varying degrees of accuracy

- Predict the risk of future disease in healthy individuals or their progeny
- Detect mutations associated with rare genetic disorders
- Detect diseases like Myotonic and Duchenne muscular dystrophies, cystic fibrosis, neurofibromatosis type 1, sickle cell anemia, Huntington's disease and breast, ovarian and colon cancers

**Pharmacogenomics or molecular medicine**

As of today, DNA variants are correlated with individual responses to medical treatments; particular subgroups of patients are identified based on DNA variants drugs being customized for these patients. The discipline that blends pharmacology with genomic capabilities is called pharmacogenomics. The field of molecular medicine mainly targets to achieve:

- Improve diagnosis of disease
- Detect genetic predispositions to disease
- Create drugs based on molecular information
- Use gene therapy and control systems as drugs
- Design "custom drugs" based on individual genetic profiles.

**Biodiversity Conservation**

Bioinformatics tools are also used for biodiversity conservation. For instance, the molecular marker technology and molecular diagnostics, *in vitro* technologies and cryo-preservation 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).

**Bioinformatics: global efforts**

Information on different aspects of bioinformatics is expanding rapidly. The advent of Internet has brought this information directly in the hands of the biologists. When genome-sequencing projects started to pour in data in scientific community, various governments decided to establish computer based databases for processing, storing and sharing the data. Benefits of this database approach were manifold, for example:

- Redundancy and inconsistency of information was reduced
- Fast and simultaneous sharing of data for a large number of users

- Standards could be reinforced
- Integrity and security restrictions of data information could be enforced

Modern fundamental and applied research in the life sciences is critically dependent on this relatively new discipline. Molecular databases are being developed and maintained by various organizations at the global level, which can be used for research by biotechnologists. The three major organisations are:

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

In addition to these three, there are numerous organisations that are actively involved in development and maintenance of biotechnology databases in the world.

**NCBI** is an international resource for molecular biology information. NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. NCBI is conducting studies on fundamental biomedical problems at the molecular level using mathematical and computational methods. These problems include gene organization, sequence analysis and structure prediction. [*http://www.ncbi.nlm.nih.gov/*]

**EMBL** maintains the EMBL Nucleotide Sequence Database (also known as EMBL-Bank) in an international collaboration with GenBank under NCBI and the DDBJ. It is Europe's primary nucleotide sequence resource. The main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications through web based applications. [*http://www.ebi.ac.uk/embl/*]

**DDBJ** is functioning as one of the International DNA Databases in collaboration with EBI (European Bioinformatics Institute; responsible for the EMBL database) in Europe and NCBI. These three databases collaborate with each other through data exchange and information on Internet and by regularly holding two meetings, the International DNA Data Banks Advisory Meeting and the International DNA Data Banks Collaborative Meeting. DDBJ is the sole DNA data bank in Japan, which is officially certified to collect DNA sequences from researchers and to issue the internationally recognized accession number to persons submitting data. DDBJ also provides numerous tools for retrieval and analysis of data [*http://www.ddbj.nig.ac.jp/*].

Various databases available on biotechnology can be broadly divided into two categories:

- **Primary Databases**: These are databases that store sequence data derived from direct experimental characterization of a nucleic acid or protein. These databases are updated directly from laboratory based experimentally derived data. The GenBank, EMBL databank, DDBJ and PDB (Protein Data Bank) are primary databases (Scaria *et al.*, 2002)
- **Specialized (or Secondary) databases**: These are databases that derive data from primary databases and are more specialized. New data is not accepted in most of such databases.

Some of the most commonly used biotechnological databases of the world:

**PubMed (NCBI):** PubMed, available via the NCBI Entrez retrieval system, is located at the National Institutes of Health (NIH), USA. PubMed is designed to provide access to citations from biomedical literature and also provides access and links to the other Entrez molecular biology resources. [*http://www.ncbi.nlm.nih.gov/Entrez/*]

**GenBank (NCBI):** GenBank is the NIH genetic sequence database. It is an annotated collection of all publicly available DNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DDBJ, EMBL and GenBank of NCBI. Sequence data are submitted to GenBank from individual scientists from around the world, as well as from the large centers involved in the Human Genome Project. [*http://www.ncbi.nlm.nih.gov/Entrez/*]

**OMIM (NCBI):** OMIM™ (Online Mendelian Inheritance in Man) database is a catalog of human genes and genetic disorders developed for the WWW by NCBI. The database is intended for use by physicians and other professionals concerned with genetic disorders, by genetics researchers and by advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic condition are urged to consult with a qualified physician for diagnosis and for answers to personal questions [*http://www.ncbi.nlm.nih.gov/Omim/*].

In addition to above databases, numerous search tools are also available for biotechnological data, some of which are:

**SRS (Sequence Retrieval System):** SRS is a Network Browser for databanks in Molecular Biology. It allows one to search multiple databases simultaneously by entering a single text-based query. [*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. 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. This search tool is preferred for searching nucleotides. [*http://www.ebi.ac.uk/services/*]

**BLAST:** BLAST 2.0 (Basic Local Alignment Search Tool) provides a method for rapid searching of nucleotide and protein databases. This search tool is better for proteins than for nucleotides. [*http://www.ebi.ac.uk/services/*]

## 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 organisation for biotechnological research and was named as the 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 now 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/*]

A wealth of information exists on India's biodiversity resources and associated knowledge. This may be in form of specimens, gray literature such as unpublished reports of district flora projects or forest working plans and books, monographs and scientific papers. A good beginning has been made in organizing a part of this information in the form of electronic databases.

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)

- 65 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. It also coordinates linkages and cooperation with external sources and developing organization in bioinformatics including documentation and information centers abroad. The BTIS secretariat working under the DBT is situated at this Apex Centre.

Centres of Excellence (COE): The missions of the COE are 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. The COE's will have state of the art computational facilities, access to the fast network and linkages with experts in diverse areas of biology, computer science, information technology, biomaterials, statistics, etc. 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 centres 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:

- 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 area of molecular structure modeling, 3D structures, active site modeling, crystal structures, conformational analysis, protein and DNA structures and interactions, homology studies and like.

**Research activities under BTISnet programme**

**(a) Indian Institute of Science, Bangalore**

This centre is presently engaged in the following projects:

- Structural analysis of a thermostable xylanase and
- Peptide structure analysis
- Structural studies on lectins
- Structural analysis of drug binding to dopamine receptors
- Protein side chain conformation analysis
- Structural studies on Phospholipase A2 mutants

Few databases, which are under development, include databases on Peptide, Lectin, Lipase and Lysozyme. In-house software of the center are:

| S.N | SOFTWARE | URL |
|---|---|---|
| 1. | Protein Sequence Search Tool (PSST) | *http://144.16.71.10/psst/* |
| 2. | Biomolecules Segment Display Device (BSDD) | *http://144.16.71.2/bsdd/* |
| 3. | PDB Goodies | *http://144.16.71.11/pdbgoodies/* |
| 4. | Ramachandran Plot (RP) | *http://144.16.71.146/rp/* |
| 5. | Conformation Angles Package (CAP) | *http://144.16.71.146/cap/* |
| 6. | Secondary Structural Elements in Proteins (SSEP) | *http://144.16.71.148/ssep/* |
| 7. | Water Analysis Package (WAP) | *http://144.16.71.11/wap/* |
| 8. | Symmetry Equivalent Molecules (SEM) | *http://144.16.71.11/sem/* |
| 9. | Fragment Finder | *http://144.16.71.148/ff/* |

[http://physics.iisc.ernet.in/~dichome/rh.htm]

**(b) Madurai Kamaraj University, Madurai**

The major research areas of the School of Biotechnology where active work is going on are:

- Genetic analysis of *Streptomyces*
- Studies on antibiotic production
- Human and microbial genetics
- Regulation of protein expression in *E.coli*
- Structural bioinformatics of membrane proteins and protein nucleic acid interactions
- Nucleic acid molecular modeling
- Computational biology
- Microsporidiasis - Molecular biology and immunology
- Genetic engineering of crop plants for viral and fungal resistance
- Molecular and structural biology of viruses
- Molecular biology and biotechnology of the entomopathogen
- *Bacillus thuringiensis*
- Biochemistry and molecular biology of thermophilic fungi
- Immunochemistry: biodegradation of Xenobiotics

***Few Important scientific achievements of this center are:***

- Membrane protein structural analysis
- Compositional analyses of genomes
- Structure based sequence alignment of restriction endonucleases and porins
- C-H..O hydrogen bond intraction in proteins
- Analysis of plant viral sequences involved in cell to cell transport
- Analysis of sucrose hydrolyses
- Modeling of *Bacillus thuringiensis* toxin
- Modeling of antibody-epitope interactions
  [*http://www.biotechmku.org/rp.html*]

**(c) Bose Institute, Kolkata**

Achievements of the center in core areas are:

**(i) Databases**

| Name of Database | Subject Area | URL |
|---|---|---|
| CUCG | A non-redundant codon usage data base of complete genomes | *http://www.boseinst.ernet.in/dic/CUCG.html* |
| CIS_PEPTIDE | A database of features of cis peptide bonds in proteins structures | *ftp://ftp.boseinst.ernet.in/pub/pinak/cis* |

**(ii) Software**

| Name of Software | Function | URL |
|---|---|---|
| LOOP SEARCH | Extract data of loops between secondary structures from PDB files (for homology modelling) | *ftp://ftp.boseinst.ernet.in/pub/dic/loop.c* |
| CONFPLOT | Two-dimensional representation of the main-chain and side chain torsion angles | *ftp://ftp.boseinst.ernet.in/pub/pinak/confplot* |

**(d) Jawaharlal Nehru University, New Delhi**

The Centre for Computational Biology and Bioinformatics at JNU is currently involved in the following research topics/areas:

**Computational Genomics**

- Development of GeneScan with splice site detection using Hidden Markov models.
- Automated genome annotation.
- Development of novel algorithms for motif and signal identification.
- Statistical methods in sequence analysis for Genome Fragmentation, Correlation and Entropy applied to full genomes and regulatory sites
- Development of tools for bioinformatics
  - Core Algorithms - The coding of reusable libraries of core techniques for research (Markov models and Genetic Algorithms).
  - System Bioinformatics – Development of tools for the identification of co-regulated genes from an analysis of their upstream sequences.

- Distributed Computing – Implementation of commonly available code on Linux clusters using PVM.

- Structure based Bioinformatics and *In silico* Drug Discovery
  - Virtual Screening and Drug Design
  - Prediction of ligand binding sites in proteins.
  - Active site design and functional re-engineering of enzymes.
  - The design of drugs and employ computational drug design techniques on identified drug targets.
- Database Management
  - Development of a system for integrated access to heterogeneous sources of biological data.
  - Provide access to annotated and raw data through BioGRID.

[*http://www.ccbb.jnu.ac.in/about.htm*]

**(e) University of Pune, Pune**

In-house software development activity in the center includes development of algorithms and programs for taxonomic classification, sequence analysis and structure prediction. General-purpose programs for sequence analysis and molecular visualization have been developed in collaboration with the Centre for Development of Advanced Computing (C-DAC). Few of them are:

- *PRAS (Parallel Multiple Sequence Alignment Package):* This is the first ever application of parallel processing to the multiple sequence alignment problem in the world.
- *PNAS (Protein and Nucleic Acid query System):* The PNAS allows analysis of protein and nucleic acid sequence data.
- *MG (Molecular Graphics):* MG is a molecular visualization package, which allows visualization of molecules as ball and stick, ribbon and CPK in a user-friendly environment.
- *MX:* This is a Molecular graphics and analysis software is open-ended molecular modeling software. This program displays molecules in wire frame, ball and stick, backbone and depth qued sphere representations. Graphics operations like rotation, translation, zooming, front and back clipping of the screen image is possible.
- *MxCurv*: This is an X-windows based program to view nucleic acid molecular dynamics results in 'Dials and Windows' format.
- *Genetic Algorithms for Structure Prediction:* This is used to predict three-dimensional structure of proteins and nucleic acids from their sequences.

In addition to the above few multimedia programs are also being developed in the center, which includes a teaching-aid package in life sciences on the "Immune System". Other application software under development in the center includes molecular modeling and sequence analysis.

[http://bioinfo.ernet.in/softwaremain.htm]

**(f) Indian Agriculture Research Institute, New Delhi**

Various softwares which are either developed or under development in the center are:

- Soil Information system: Gives the organic content status of Indian soils.
- Pedigree Informatics and Morphological Diversity in Wheat: This deals with information related to pedigree of established cultivars, information related to their genetic, physiological and phonological characters along with their suitability in various growing environments.
- Integrated Pest Management Information System: This provides the information about the rice pests (insects, diseases, nematodes, etc) and if user wants to know any desired information this software may answer queries and suggestions.
- Plant Protection Information System for Major Crops: This has two parts
  - Integrated Pest Management System (IPMS) which includes diagnostics, damage symptoms, management options, natural enemies.
  - Plant Disease Information System (PDIS) this has details of symptoms and characteristics, management options, geographical distribution.

  [*http://www.iari.res.in/divisions/usi/project.php*]

**(g) Centre for Cellular and Molecular Biology (CCMB), Hyderabad**

Few of the software developed or under development in collaboration from private sector are:

- Development of recombinant DNA-based Hepatitis-B vaccine for human use.
- A technology about RNasin - an enzyme-inhibitor - has been successfully transferred.
- Development of a protocol for standardization of PCR-based markers to distinguish the parental and hybrid seed varieties of rice.

- Research project to develop a new therapeutic agent for the treatment of cancer.
- A research programme to provide rapid, accurate and reliable diagnostic service based on modern molecular methods to cancer patients.
- *Unique RNase from Cobra Snake Venom* : Development of RNase cutting enzyme, which has specificity towards cytidylic acid was isolated, purified and characterized from Cobra Venom. The enzyme has a potential to be used as a tool for sequencing RNAs and in their structural studies.
- *Gene Delivery System*: Formulation of system for efficient delivery of DNA into the cells.
- *Microbes from Antarctica: Biodiversity and Cold Adaptation*: Identification of cold-loving bacteria and yeasts, hitherto unknown from Antarctica.
- *Protein folding:* Research on Protein folding. CCMB scientists have arrested the protein (RNase-A) under a certain set of conditions in a state of partially folded active protein - the state known as molten globule.
- *Peptides to fight bacterial invasion*: Research on designing peptides consisting of 11-15 amino acid residues, which possess selective antibacterial activity.
- *Protein from seminal plasma plays multi-functional role*: Discovery of seminal plasmin protein, which is a potent microbial agent as well as anti-HIV agent.
- *Regulation of cell division*: Research on a new model explaining the non-phenomenology of regulation of cell division and malignant transformation.
- *A new universal probe for DNA fingerprinting*: Development of a new probe derived from an Indian banded krait snake. This probe is being used for forensic investigations, paternity determination and seed stock verifications and DNA fingerprinting.
- *Molecular basis of sex determination*: Isolation of highly conserved sex chromosome-specific satellite DNA, `Bkm' from the female Indian branded krait snake, which is being used for further study of sex reversal in humans.
- *Research on sex reversal*: Identification and characterization of the new gene(s) involved in the complex pathway of sex determination.

[http://www.ccmb.res.in]

**(h) National Institute of Immunology (NII), New Delhi**

The center has a number of ongoing research projects focused on molecular modeling, analysis of gene and protein sequences, 3-D structures of proteins and protein structure prediction using knowledge-based approaches. Some of the projects undertaken at NII are:

- Analysis of sequence signature defining functional specificity and structural stability in Helix-loop-Helix proteins
- Molecular modeling of protein-ligand complexes using knowledge approach
- Analysis of protein sequences from microbial genomes and structural proteomics

[http://www.btisnet.nic.in/files/dics/nii.htm]

**(i) Institute of Microbial Technology, Chandigarh**

Currently following major research projects are in progress at the Bioinformatics Center, in the field of protein structure prediction:

- Structure determination (X-ray crystallography) and molecular modeling, with emphasis on proteins involved in signal transduction and key aspects of the metabolism of pathogenic microorganisms.
- Protein structural biochemistry which utilizes a combination of wet biochemistry, separation and analytical methods, spectroscopy, and recombinant DNA technology, to deal with a variety of fundamental and applied issues relating to enzyme function, protein-protein interactions, protein folding, mis-folding, aggregation, sequence-structure relationships.
- Development of bench-scale technology for the indigenous production and purification of Natural Streptokinase, a blood clot dissolving drug.
- Benchmark in Field of Protein Modeling.
- Prediction of T Cell epitopes from their primary sequence.
- Identification of novel drug targets in Mycobacteria employing genomic informatics approaches.

*[http://imtech.res.in/raghava/www.html]*

**(j) National Brain Research Centre, New Delhi**

Main research areas in this center are Molecular and Cellular Neuroscience, Systems Neuroscience and Theoretical Neuroscience. This center manages a FTP server which hosts sharable information related to Neuroscience and Neuroinformatics. NBRC also manages the Virtual Private

Network (VPN) for the DBT to connect all its 11 DBT institutes located across the entire country [http://www.nbrc.ac.in].

**Contributions of other Indian institutions**

In addition to the above centres, other institutions that are actively contributing in the field of biotechnology are:

**The Centre for DNA Fingerprinting and Diagnostics (CDFD)**

Following projects are under development in the computational biology laboratory of the centre:

- Computational genome analysis: The project deals with an analysis of microsatellites with respect to their distribution, abundance, enrichment and mutations in the known prokaryotic genomes. A database of prokaryotic microsatellites, MICdb, and a microsatellite analysis software called MICAS have been developed.
- Computational structural genomics: The proteins implicated in multi-drug resistance and intracellular survival of the pathogens are being identified and subjected to comparative modeling.
- Computer based modeling of SNPs in target proteins: This project's goal is to integrate genetic variation data into protein structures in an attempt to discover the molecular roots defining genotype-phenotype relationships.
- Computer simulations to investigate stability of DNA with small molecular adducts: In this project the details of intermolecular interactions between DNA and a family of small molecular adducts which exhibit antitumor activity have been investigated using molecular modeling and dynamics simulations.
- Development of knowledge based computational methods for automated structure prediction and modeling of proteins: In this project the focus is to develop new knowledge-based computational procedures for reliable structure and function prediction of protein sequences. [*http://www.cdfd.org.in/rsenres.html*]

**CDFD is also a national node of the** European Molecular Biology Network (EMBnet). The EMBnet is a science-based group of 37 collaborating nodes throughout Europe and a number of nodes outside Europe. The EMBnet India Node provides bioinformatics services in the form of browsing bio-molecular sequence databanks, macro molecular structure databanks, genome and other useful databases. It provides in-house services for the comparison and analysis of sequence/structure/genome data protein 3-D modeling molecular graphics [http://www.in.embnet.org/bioinfo.html].

**Anna University, Chennai**

Following projects are under development in the computational biology laboratory of the centre:

- Research on causes and pathogenesis of tropical pulmonary eosinophilia.
- Research on Pathogenesis of EPEC diarrhea.
- Bioscreening and isolation of medicinal properties from plant sources towards new drug development.
- Immunopathology of Filariasis a. T cell and macrophage function in filarial patients.
- Research on *Wolbachia* genes.
- Identification of new genes from *W. bancrofti.*
- Immunodiagnosis and Immunophylaxis of Canine Distemper Virus infection in Dogs.
- Production of recombinant proteins to White Spot Syndrome Virus in Shrimps.
- Spermicidal effects of neem oil and its fractions.

*[http://www.annauniv.edu/biotech/main.htm]*

**Biotech Consortium India Limited (BCIL)**

BCIL, New Delhi was set up in the year 1990 with the objective of providing linkage between the R&D and industry to facilitate the commercialization of biotechnology products and processes. BCIL has been engaged in technology development, technology transfer, project consultancy, fund syndication, information dissemination, manpower development, placement and training related to biotechnology. It has so far assisted more than 150 clients belonging to diverse background including scientists, technologist, research institutions, universities, first generation entrepreneurs, the corporate sector, government, banks and financial institutions. [http://www.biotech.co.in].

**Institute of Genomics and Integrative Biology (IGIB)**

IGIB under the CSIR setup is working on major projects related to immunology and molecular genetics of respiratory disorders including allergy, fungal infections and predisposition to asthma, molecular genetics of neuro-psycho disorders and functional significance of repetitive sequences in the genome, genome informatics and drug target identification. IGIB is also working on the development of molecular markers for pathogenic organisms,

including *Mycobacterium tuberculosis,* molecular recognition/interaction studies, design and synthesis of modified oligonucleotides for antisense and gene targets and design, synthesis and structural studies of peptides with a role in neurological function and dysfunctions. (Suresh, 2003)

**Major key players in Indian biotechnology industry**

Studies by various independent agencies of the world indicate that India will be a potential super-power in bioscience in the next decade. This prediction has been arrived at, after due consideration of factors like bio-diversity, 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:

*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 also has 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 also has created contract drug discovery, clinical trials, genomics and chemistry research units and sister companies.

*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).The Indian vaccine startups are now turning to indigenous versions of other recombinant proteins, including granulocyte colony stimulating factor (GCSF) and other growth factors, erythropoietin (EPO) and insulin.

*Ranbaxy*, India's largest pharma company also views 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 U.S. and European companies to develop new formulations and delivery technologies.

*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.

*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.

**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 Biotechni*, and it has numerous industrial relationships, including a joint venture with *Biological E and Amersham Pharmacia* to build DNA microarrays.

U.S. 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 few 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 the 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 the potential to lead if it invests in this area. However, to achieve this biotech and pharma companies would need tremendous software support. Software expertise would be required to write algorithms, develop software for existing algorithms, manage databases, and in the final process of drug discovery.

Major pharmaceutical and genome-based biotech companies are investing heavily in software. Pure cost benefits for the biotech companies will definitely drive the bioinformatics industry in the country. The biotech companies would be forced to outsource software rather than developing propriety software like in the past. Since the cost of programs for handling this data is extremely high in the west, Indian IT companies have a great

business opportunity to offer complete database solutions to major pharmaceutical and genome-based biotech companies in the world.

India grew its computer software and services industry from zero to more than $7 billion in exports over the last decade. This fact would be big motivation to India's biotech aspirations, as do the strong business and cultural connections between the subcontinent and Silicon Valley.

**Building expertise in Biotechnology**

Estimations and market study show that the biotech industry of India is expected to cross a billion dollar mark during current financial year. Last year the growth rate was nearly 40% and this year it is expected that the same would accelerate. (Sangal *et al.*, 2004). The growth of this industry would in turn require trained manpower in this field. In India, biotechnology industry has received great impetus and support from the National Biotechnology Development Board and DBT. These organizations are actively involved in imparting quality education and also have provided the required infrastructure to cater for the needs in research areas. Few of top ranking institutes, which are imparting education at undergraduate and/or postgraduate level in biotechnology are:

- University of Hyderabad, Hyderabad
- Anna University, Chennai
- University of Pune, Pune
- Indian Institute of Technology, New Delhi
- Jawaharlal Nehru University, Delhi

In order to provide exposure for Indian scientists to newer trends in R&D a programme of Biotechnology Overseas Associateship has been launched. This envisages the award of long-term (one year) and short-term (three to six months) associateship to Indian scientists to work in foreign research laboratories and upgrade their knowledge. In addition to this, a scheme has been started to promote visits of eminent foreign scientists to Indian academic and research institutions engaged in this field. This is expected to help Indian scientists share the current state of knowledge in this area of research.

Inspite of these endeavours by various government and private organizations, gap still remains. Adequate trained manpower is absolutely essential for the development of the industry. But today since the field is still nascent so it does not have a large absorption capacity. Later when the industry matures, there will be an acute lack of qualified professionals. Therefore the field of molecular biology is in a typical "catch 22" situation. There is a need of comprehensive HRD planning in view of the demand and employment opportunities in the country.

## Problem Area

Building biotech companies is far more complex than establishing IT Companies. And to achieve its potential in biotech, India will have to overcome some significant barriers, including a confused regulatory environment, uncertainties about intellectual property protection and the slow pace of integration between academic and commercial science. A strong patent protection system is immediately required in the country.

The industry people, meanwhile, say that the mushrooming of bioinformatics institutes is creating a problem of finding talented and trained individuals in this industry. While many of them have a superficial knowledge and a certificate, India lacks true professionals in this area. Most people who opt for bioinformatics are from the life sciences areas that do not have exposure to the IT side of bioinformatics, which is very important. 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.

## Conclusion

Bioinformatics is a segment created by the merger of two hot areas: information technology and biotechnology. Without bioinformatics, new research in most fields of medicine and biology would come to standstill. The technology is versatile and can be applied whenever gene, protein and cell research are used for the discovery of a new drug or a new herbicide/ herbicide-resistant crop combination. Drug toxicology, pharmacogenetics and clinical trial studies can also benefit from this technology which can even be used to genetically engineer crops and livestock that have enhanced nutritional qualities and the ability to produce pharmaceuticals. To foster biotech education and research in the country, the culture of DBT supported courses gave a huge impetus to the human resource generation. Organizations like CSIR and other CoEs have also highly contributed to this field. Indian IT sector also zeroed in on life sciences and healthcare as the new opportunity areas. As the Indian Bioinformatics industry is still in its

nascent stage, problem areas including qualified man-power shortage, patent and Intellectual property rights issues, adequate funding for specialized research, etc need to be addressed. The fact that this field has tremendous growth potential, dictates a series of actions, which Government and the private sector companies should initiate in collaboration and mutual harmony. Bioinformatics is an indispensable ally of biotechnology researchers. Bioinformatics and biotechnology have come to stay and will be a major technologies of the future.

## References

Chakraborty, N., Chakraborty, S. and Datta, A. (2005). Nutritional genomics: commitment to society. In: Tandon, P., Sharma, M. and Swarup, R. (Eds.) *Biodiversity: Status and Prospects.* Narosa Publishing House, New Delhi, India, pp 35-41.

Lengauer, T. (2002). From genomes to drugs with bioinformatics. In: Mannhold R., Kubinyi, H. and Timmerman, H. (Eds.) *Bioinformatics – From Genomes to Drugs.* Wiley-VCH, Weinheim, (Federal Republic of Germany), pp 3-25.

Sangal, R. Anand, S. and Sangal, N. (2004). India's top biotech schools. *Biospectrum,* Dec edition, p. 6-16.

Scaria, J., Raveender, V. and Verma, S.K. (2002). Genome and molecular databases. In: Khan, A. and Khanum, A. (Eds.) *Emerging Trends in Bioinformatics.* Ukaaz Publications, Hyderabad, India, pp 26-80.

Sharma, M. (2003). Potential of Biotechnology. *Kerala Calling,* Dec, 2003, p. 13-18.

Sharma, M. (2005). Biodiversity regulations in India. In: Tandon, P., Sharma, M. and Swarup, R. (Eds.) *Biodiversity: Status and Prospects.* Narosa Publishing House, New Delhi, India, pp 119-123.

Sugden, A. and Pennisi, E. (2000). Diversity digitised. *Science,* 289, 2305.

Suresh, N. (2003). A versatile gene lab. *Biospectrum,* June edition, p. 47-49.

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.* Allied Publishers Ltd., India, pp. 140-147.

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

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