Bioinformatics: The foundation of present and future biotechnology

*K. K. Tripathi**

Bioinformatics is the science of data management system in genomics and proteomics of life forms. It is a comparatively young discipline in information technology and has progressed very fast in the last few years. Bioinformatics is practised worldwide by biotechnologists to access various databases for research and to exchange information for comparison, confirmation, storage and analysis. As on date, there are a number of databases on specific genes and proteins pertaining to human, animals, plants, bacteria, and other life forms. These are being enriched and updated through research in modern biology with the practice of bioinformatics. These databases help in new inventions in biotechnology that are useful to mankind. Bioinformatics is enabling life sciences to invent novel drug discovery as well as drug delivery systems for greater progress in the field of biotechnology. Such inventions attain importance in the present scenario of patents and WTO regime. For the future development of biotechnology, bioinformatics will have to play a vital role with the involvement of internet tools and the World Wide Web (WWW). The future rDNA research would be guided largely by the databases available in generic or specific forms. Thus bioinformatics and biotechnology have to move hand in hand for their progress. However, bioinformatics can now be branded as a bonafide discipline within information technology.

WITH the advent of modern techniques in biology, biotechnology has progressed exponentially in the last decade. Genetic engineering techniques have divided the discipline into 'old' and 'new' biotechnology, where the former deals with conventional and classical methods, and the latter involves highly specific and targetted rDNA technology tools. This has led to the generation of enormous databases with information overload in the life-sciences disciplines. If this information were to be used to speed up the pace of scientific research, scientists would need to know the methods and tools for their effective use.

Bioinformatics is a science typically associated with databases in genomics and proteomics and structure and function information for genes and proteins, of all forms of life on earth. In the past decade there has been a 'cyber-war', with the introduction of a number of databases on genomics and proteomics. The comparative genomics in biotechnology has provided a path of evolution to researchers through accumulated data on genomics and proteomics. The earlier questions of evolution within the domain of anthropologists are now within the reach of scientists practising bioinformatics and biotechnology. Bioinformatics has made it possible to trace the migration patterns of ancient humans from the traces of chromosomal sequences left in the genomic patterns of modern-day society descendants. Today, maternity and paternity can be traced, assessed and certified through modern biotechnology and bioinformatics tools with chromosomal sequences of X and Y-chromosomes as well as the evolutionary process involved in the polity of sexes[1].

Apart from human genomics, the evolutionary process, horizontal gene transfer in evolution and the genetic modification of plants are major areas of research today. rDNA technology has so revolutionized the plant and agricultural sciences research that colossal data have been generated in agriculture biotechnology and transgenic plants. The benefits of plant engineering through biotechnology have started reaching the consumers of the developed world and are in the offing for the poor of the developing countries. Bioinformatics provides data access to such developments. There are increasing evidences that transgenic plants would produce healthier, storable foods with desired nutritional value. Genetically modified plants are considered as chemical factories which are capable of producing desired proteins, antigens, energy, vitamins, desired enzymes, etc. with the practice of biotechnological tools. Bioinformatics, on the other hand, manages all the information and data on such aspects of transgenic plants. It has been reported that a number of US companies active in transgenic plant research would be spending more than US $ 1.5 billion, which they would be able to accumulate from sale of modified seeds in the available market of US $ 500 billion[2].

K. K. Tripathi is in the Department of Biotechnology, Block-2, C.G.O. Complex, Lodhi Road, New Delhi 110 003, India.
e-mail: kkt@dbt.delhi.nic.in
*The views expressed in the paper are those of the author and they do not necessarily express the views of the organization to which the author belongs.

Biotechnology has been instrumental in developing databases on genomics and proteomics pertaining to such developments in agriculture as well as other forms of life in general. Scientists are now working to segregate specific databases of different life forms from the available information in different databases. These databases are the bedrock of current and future biotechnology research. Bioinformatics has played a key role in the stunning pace of change in biotechnology research through available databases, which are further being enriched and expanded. It has paved the way for the progress of biotechnology with interdisciplinary scientific research becoming essential and the barriers between traditional disciplines having crumbled. The biotechnological databases of discoveries in model organisms highlight the unity of life resulting from evolution by common descent and the data are readily applicable to human biology and other areas of research through bioinformatics. As genomics, proteomics and bioinformatics shift their focus of analysis from individual life component to complete biological systems, an informational science of the whole organism has come into being due to merger and mega-merger of disciplines like molecular, cellular, developmental, computational and physiological sciences. The biotechnological databases are being utilized for designing new experiments for future research in genomics, proteomics, metabolic regulation, transgenics, horizontal gene transfer in plants and animals, computational approaches in structural biology, protein characterization, structural genomics, etc. These databases have given leads for research in human health, disease, drug development, gene therapy and structure–function relationships in biomolecules. Thus bioinformatics provides different tools to a biotechnologist for access and compilation of data for rigorous sequence similarities search, data mining and planning for new experimental designs in all areas of biotechnology research.

What is bioinformatics?

Bioinformatics is a computer-assisted interface discipline dealing with the acquisition, storage, management, access and processing of molecular biology data. This discipline helps us to collect, compile, analyse, process and represent the information in order to understand processes of life in healthy and diseased states and find new processes or better drugs and delivery systems for genes, drugs and other genetic components for life improvement and to educate oneself[3]. It is an interdisciplinary scientific tool without barriers among various disciplines of science like biology, mathematics, computer science and information technology.

Why is bioinformatics relevant for biotechnologists?

Bioinformatics is being practised worldwide because of its great relevance in modern biology. With the advent of the Human Genome Project, more than 30,000 human genes have already been mapped[4]. A 'working draft of the human genome' has already been produced almost two years ahead of previous projections[3,4]. By the year 2003, a high quality complete genomic sequence would be available, which would identify all the estimated 100,000 or so genes in the human DNA. Thus the sequences of billions of bases making up the human DNA would be determined and the databases would be built up[5].

Apart from the human genome, information that exists for other organisms would be added, amassing huge information databases that are like tidal waves[6] of data available to active researchers in biotechnology, which are to be managed, absorbed, stored and accessed for further use. These data waves would keep coming as the information being piled up is not only on genes and proteins but it also includes continuous updating of the sequence information for genes and proteins, structure/function annotations, population variations, disease correlations as well as every bit of information that is being generated in life sciences on bacteria, viruses, fungi, protozoa, algae and various specific chromosomes of human and animals[3]. At this stage, biologists have not completely thought about how this data could be used. Bioinformatics has thus become important to find ways and tools in achieving various goals in accessing and using the accumulated information[3]. The accumulated valuable data resources can be accessed and there is immediate communication with national and international research groups all over the world at affordable costs. These utilities have made bioinformatics so relevant that biotechnologists are keen to learn about bioinformatics tools for use in their research.

The development of bioinformatics

Practically, the development of bioinformatics started with the networking of computers and accumulation of data on genes and proteins in biotechnology. This discipline is still young and may be traced back to the 1970s, when automated DNA and protein sequencing became possible and biotechnologists started generating large amounts of data on various parameters of genes and proteins. With the generation of new computer-based archives of massive databases on genomics and proteomics, structure–function relationship of new sequences and accessing and comparison of new sequences with what was already known through network computers, initiated the development of bioinformatics.

The discipline developed further during mid-1980s, when use of computers became very popular for data storage and access by users, locally or remotely. It was this period when various tools to store, manage and access

the information through computers were developed. The use of computers in biomedical sciences grew exponentially after this period when researchers started relying on them for experimental data recording, analysing, storing, browsing technical literature and modelling of proteins and genes[7]. With the accumulation of new data, new techniques were developed for the proper utilization of databases developing bioinformatics further. During the 1980s, for example, a software package called PC/GENE was introduced by a US company, Intelli-Genetics from California[8]. This package could translate a given sequence of gene to its protein end product with predictions of its secondary structure. This revolutionized the discipline of bioinformatics and paved way for a number of new software packages developed according to the requirements of biotechnologists. These software packages were integrated programmes with a graphics-oriented environment capable of storing data, comparing DNA and protein sequences, manipulating and analysing data of molecular structures and graphs with biological simulation models. The examples of early efforts in these packages are the Staden Package for DNA sequences and PROPHET for data management and analysis in life sciences[9]. In USA, these packages have been described as tailored resources for life sciences researchers working in diverse areas ranging from molecular biology to pharmacology in biomedical science.

As on date, the colossal powers that the Internet gives to the computer is mind-boggling. The 'Net' represents the transformation and evolution of the entire information age proving to be the latest 'developmental tool' for the progress of science in general and bioinformatics in particular. It unleashes the powers to communicate unlike any other system. It has made a major impact on the information technology and bioinformatics. It has proved to be an ocean of information accessible to masses around the world. During the 1990s, with the introduction of WWW the developmental work in bioinformatics got a shot in the arm with vast connectivity and accessibility to databases in a user-friendly way. The success of internet came about with the advent of various networking tools and the servers in the early and mid-1990s, which could make possible the very existence of 'WWW'.

It was Amos Bairoch, at the beginning of the last decade who established a moderate database[10] pertaining to protein sequence and structural correlations on the 'Net'. This database was known as PROSITE, which was further strengthened and complemented with a database on sequence analysis and comparison of protein sequences known as SEQANALREE (ref. 11). From 1991 scientists started depending more on the Internet for remote accessing of data, repositing their own sequences and accessing information available on the 'Net'. The early 1990s also experienced the development of proteomics-related databases on two-dimensional polyacrylamide gel electrophoresis maps of proteins pertaining to diverse groups of healthy and diseased tissues, known as SWISS-2DPAGE (ref. 12). This was the period when compilation of databases on functionality of non-coding sequences, computational protein structures, and protein–peptide characterization based on structure–function analysis started accumulating. Earlier scientists depended for the exchange of information mainly on correspondences and journals publishing such information. Then a fully organized and detailed protein sequence database[13], now famous as SWISS-PROT was introduced. This is a curated protein sequence database, which helps in providing a high level of annotation, a minimal level of redundancy, and a high degree of integration with other databases. Recent development in this database includes cross-references to additional databases. A computer-annotated supplement to SWISS-PROT has improved it significantly. This was possible with a variety of new documentation files and improvement to TrEMBL (ref. 14). The TrEMBL consists of entries in a SWISS-PROT like format derived from the translation of all coding sequences (CDS) in the EMBL nucleotide sequence database.

Biotechnology databases and tools of bioinformatics

To list the number and names of databases and software tools used to access them is beyond the scope of this article, as there are many universal resource locators (URLs) providing access to biotechnology databases either free of cost or with charges. However, some important databases which are commonly used by biotechnologists are those from Incyte, Pangea systems, PE-informatics, GCG, NCBI, PDB, SRS, UDB, SWISS-PROT, EMBLnet, ICCBnet, Medline, FlyBase, Mendelian Inheritance in Man (MIM), US Patent and Trade Mark office database and SeqWeb. There are specific databases available on the plant, animal or human genomics and proteomics; USDA maintains a crop genome database server at Cornell University at http://ars-genome.cornel.edu/. Similarly, FishNet is dedicated to the biology of Zebra fish and the URL is http://zfish.uoregon.edu/. The mosquito genomics data provides genetic data about three species of *Aedes* (*aegypti*, *albopictus* and *triseriatus*), *Anopheles gambiae* and *Culex pipiens* at http://klab.agsci.colostate.edu/. Practically a number of URLs for specific databases can be located on the web, which may run into pages if we start compiling them. However, during web search these resource locators can be easily found. There are various programmes and tools available for making searches, algorithms analysis, modelling and the computer graphics of the databases on genomics and proteomics like BLAST, FASTA, Smith-Waterman, ENTREZ, MAGE, CHIME, RasMol, CASP, CAFASP1, PDB-3D Browser, SWISS PDB-Viewer, CHROMAS, CINEMA, EDITVIEW, DNA-SEQUENCHER, FACTURA, AUTO ASSEMBLY, GCG software, Gene Explorer 1.4, etc.

Some of the tools used in bioinformatics and computational biology have been compiled by Horton[15] after obtaining information from their manufacturers.

The networking and current status of bioinformatics

The internet, which is an information superhighway, has practically compressed the world into a cyber colony through Local Area Networks (LAN), Wide Area Networks (WAN), Metropolitan Area Networks (MAN) and other Intranets. The genealogists form part of this colony practising bioinformatics, and sharing enormous databases. The development of the internet and the emergence of the WWW as a common vehicle for communication and instantaneous access to databases are an exciting aspect today in bioinformatics. However, searching the WWW requires various tools, software packages and strategies.

Tools for bioinformatics and computational biology include servers and computation boxes, cross-platform software, web-based interfaces to tools and databases as well as proprietary and public data sets. At present there are many bioinformatics software and data access tools available on the internet free of cost, for various computational environments. It is beyond the scope of this article to list them all. These software are mainly developed by students and biologists with a knowledge of software development languages, who constantly update the existing tools as well as produce new applications.

Strategies for searching the 'Net' can be categorized broadly as useful for cataloguing the WWW directories and search engines. The search engines are non-specific and produce voluminous results while directories are likely 'yellow pages' which are browsed by the search engines[3]. Efforts are on to create standardized bioinformatics tools so that it is easy for a user to implement them with a standard operating system. However, as on date, biologists, and biotechnologists also skilled in computer science and information technology are in short supply. The development of bioinformatics has been limited due to this fact[16], keeping in view the pace of advancements in biological and biomedical sciences and generation of more biomedical information databases. Now biologists who compute have become hot properties in bioinformatics.

Since bioinformatics can be practised anywhere in the world with an internet connection and access, it helps the users with theoretical research where data complement experimental biology. Funds for massive projects in bioinformatics are increasing day by day. There are a number of companies who have invested in developing various tools and adaptation of databases in terms of existing and new standards as well as education and training along with the European Commission[12]. In the present scenario, bioinformatics has become an essential tool and the backbone for research in biotechnology, involving researchers in population biology, environment and ecology as well as various simulations and analytical approaches through various databases on biotechnology.

Future challenges in bioinformatics

Keeping in view the present pace of investment and developments, it is impossible to predict the future of bioinformatics. The searching of biological databases via the WWW is becoming increasingly difficult. Differences in database structures and nomenclature hinder research efforts where standardizations have met with much resistance. However, researchers are optimistic that web tools developed for other purposes may help bioinformatics[17]. New software are being produced for different applications in biotechnology. Such developments are important for the future of bioinformatics and development of biotechnology.

The other emerging challenges for the future are the audit and control of databases, which are increasingly becoming larger and larger. The complex computing environment and resource crunches would make it vital for information technology auditors to have practical guidance on conducting audits and also ensuring security in today's stretched and quickly changing computing environment. Further, it has been apprehended that such fast developments in the internet and use of information available in databases may prove to be 'library killers' as subscriptions to science journals will reduce[18]. The future challenges of sequence analysis are pushing bioinformatics in a time when the demand of bioinformaticians outnumbers supply. Thus, in future more biotechnologists with computer knowledge are required. This can be made possible through various training programmes. Various network systems like EMBLnet and ICCBnet, initiated by UNESCO are playing a vital role in this regard imparting training to biotechnologists in bioinformatics through their training programmes[3].

The other challenges in the practice of bioinformatics are development of various strategies to surf WWW, and reaching a consensus of coining uniform definitions and adopting uniform platforms and technologies. The bigger the data available on the WWW, the more interesting and useful it is considered. Information technology talks in terms of terabytes (trillion bytes), petabytes (1000 terabytes) and exabytes (1000 petabytes). It is stated that all the words ever spoken by human beings amount to about five exabytes. Now a new, zettabyte term has been coined for such huge databases involving 1000 exabytes and 1000 zettabytes as a yottabyte[6] or 10^{24} bytes.

The internet information superhighway which is being flooded with data poses a question as to how biologists and biotechnologists would be able to use them. It is envisaged that to meet the future challenges in bioinformatics,

internet-II will be launched to bring focus, save energy and resources for the development of the new family of bioinformaticians capable of using advanced applications to meet emerging academic requirements in biotechnological research, teaching and training. Internet-II is a collaborative effort by more then 120 US universities trying to establish 'GIGAPOPS' (gigabyte per second point of presence) that would provide regional connectivity among universities and other organizations[3].

The next generation of bioinformatics database access system has been called as NGI, i.e. Next Generation Internet, which has been approved by the US government at a cost of about US $102 million for the next two fiscal years. The NGI initiative is a multi-disciplinary federal research and development programme that is being developed for advanced networking technologies, which is going to revolutionize applications that require advance networking. This would demonstrate the capability of test beds that are 100 to 1000 times faster end-to-end than today's internet. May be internet-III would be the new challenge and course of future for further developments in bioinformatics[3].

Indian scenario

The whole paradigm shift in molecular biology towards data-intensive research in search of useful genes is basically due to the fact that the genetic data is becoming the major driving force in drug discovery, protein engineering, design of new molecules and other related areas. The impact of bioinformatics on Indian bioscience and biotechnology can be seen both in tangible and non-tangible terms. Research and development activities in these fields grew in quantity as well as quality, as can be seen from research papers published from India. The Department of Biotechnology, Ministry of Science and Technology, has played a key role in the advent of bioinformatics in the country by establishing bioinformatics centres. The initiatives launched by the government in liberalizing the access to internet and deciding to establish national networks are expected to benefit the programme significantly in its attempt to disseminate bioinformatics resources to a large number of scientists in universities and R&D institutions. Presently, the network comprises fifty-two bioinformatics centres with a main centre in New Delhi. The Biotechnology Information Centre (BTIC) is responsible for coordinating, organizing and providing information services at a national level covering a wide range of subjects on large sectors of national endeavours in biotechnology. The BTIC has the mandate to continuously assess information requirements in biotechnology and organize creation of necessary computer and communication infrastructure to provide bioinformatics support to the national community of users spread across the country. It is coordinating the activities of other centres and plans to provide a nationwide communication network between the Distributed Informatics Centres (DICs) and other Sub-centres. Ten DICs have been established with the task of providing subject/discipline-oriented information to all institutions belonging to the branch and other institutions and individual users interested in any particular subject of information related to biotechnology. These centres are located at Indian Institute of Science, Bangalore; Jawaharlal Nehru University, New Delhi; Madurai Kamaraj University, Madurai; Bose Institute, Calcutta; University of Poona, Pune; Indian Agricultural Research Institute, New Delhi; Centre for Cellular and Molecular Biology, Hyderabad; National Institute of Immunology, New Delhi; Institute of Microbial Technology, Chandigarh, and National Brain Research Centre, New Delhi. Thirty-eight R&D institutions and universities form chains of Distributed Sub-centres set up in the country. While the DICs act as repository of information in their respective specialized disciplines, the Distributed Information Sub-centres provide an access mechanism for the information to be available at the universities, R&D and manufacturing institutions. Thus, the distributed sub-centres provide an added dimension of access and diffusion of information across the network. The main centre also co-ordinates linkages and cooperation with external sources in bioinformatics, including documentation and information centres abroad[19]. The centres are networked through a satellite communication system. The network approach has been useful in the successful implementation of the project, as it has established a link between diverse groups of scientists working in various interdisciplinary areas of biotechnology. The network encouraged sharing of knowledge and greater interaction amongst the scientific community irrespective of their geographical locations. It has endeavoured in establishing national databases in the country in collaboration with several national and international agencies and has encouraged bilateral and international collaboration in bioinformatics, e.g. with EMBLnet, ICCBnet, etc. The University of Pune has developed several databanks on animal viruses, agricultural pests, biological and medical research, etc. and provides an update and accurate information in the area of biotechnology. The URLs for this important site, which has recently been selected to feature on the Web, 'Pick of the Day' for searching databanks on genome related sites in India, are http://bioinformo.ernet.in and http://bioinfo.ernet.in/~sunita/main.html. The other important Indian site on the Web is 'Gatew
maintained by the Molecular Biology Group of Tata Institute of Fundamental Research, Mumbai. Similarly MKU, Madurai and IISc, Bangalore have their own databases, which can be accessed by biotechnologists. The important URLs can be found at 'Resource of Biology Pages' at http://neehow.ym.edu.tw/wonderful/biosites/bioinforg97.html, which is updated periodically

and any centre can add its URL for access by researchers. Similarly there are a number of such resource sites available providing Indian links. Despite the fact that a number of private organizations have excelled in information technology, none has entered in the area of bioinformatics so far, like in developed countries. Soon multinational biopharmaceutical giants may open this arena in India, like e-commerce for industry, as a business for developing new drugs, delivery systems, geneticeuticals and immunobiologicals.

Concluding remarks

The advent of the relatively new discipline, information technology, has helped in the development of bioinformatics as a foundation of biotechnology. Its main focus is on the biological information management, independent of the origin or representation of the biotechnological data. The computational and mathematical approaches in analysing various biological data supplemented with other methodologies of laser mass spectrometry and X-ray techniques, generating data about the structure and function inter-relationships of biomolecules would further strengthen bioinformatics. This discipline is enabling life sciences to invent novel drug discovery as well as drug delivery systems to make biotechnological progress much faster. Such inventions attain importance in the present scenario of patents and WTO regime. There is no doubt that the advent of bioinformatics will revolutionalize biotechnology. The success of biosciences would depend upon the databases. The involvement of industry has placed bioinformatics in a post-genomic age and now it has formed its own society, the International Society for Computational Biology[20]. It is expected that soon a theoretical biologist of the post-genomic period with a understanding of networks, pathways and cascades, signal transduction, metabolism and genetic regulation would be able to guide rDNA research in a manner like 'reverse transcriptase'. The accelerating developments in information technology would make essential the availability of compatible hardware too, which can cope with the available bioinformatics tools. There is no doubt that bioinformatics has come to an age within past few years to become a 'bonafide discipline'[12]. The advent of the internet and WWW has developed bioinformatics enormously, making it capable to form a new society of biotechnology and biological scientists, who may call themselves 'bioinformaticians' and the discipline itself may be coined bio-information technology (Bio-IT) at par with information technology (IT).

1. Barbara, R. J. and Hines, P. J., *Science*, 1999, **286**, 443.
2. Philip, H. A. and Hines, P. J., *Science*, 1999, **285**, 367–368.
3. Edelman, M., The ICCBnet Bioinformatics Training Workshop, International Centre for Co-operation in Bioinformatics Network, Weizmann Institute of Science, Israel, July 11–20, 1999.
4. Deloukas, P. *et al.*, *Science*, 1998, **282**, 744–746.
5. Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. and Walters, L., *Science*, 1998, **282**, 682–689.
6. Reichardt, T., *Nature*, 1999, **399**, 517–520.
7. Malakoff, D., *Nature,* 1999, **284,** 1742.
8. Moore, J., Engelberg, A. and Bairoch, A., *BioTechnique*, 1988, **6**, 566–572.
9. Hollister, C., *Nucleic Acids Res.,* 1986, **14**, 21–24.
10. Bairoch, A., *Nucleic Acids Res.*, 1991, **19** (suppl.), 2241–2245.
11. Bairoch, A., *Comput. Appl. Biosci.*, 1991, **7**, 268.
12. Persidis, A., *Nature Biotechnol.*, 1999, **17**, 828–830.
13. Bairoch, A. and Boeckmann, B., *Nucleic Acids Res.*, 1991, **19** (suppl.), 2247–2249.
14. Bairoch, A. and Apweiler, R., *Nucleic Acids Res.*, 1999, **27**, 49–54.
15. Horton, B., *Nature*, 1998, **393**, 603.
16. Marshall, E., *Science*, 1998, **272**, 1731–1740.
17. Sobral, B. W. S., *Nature*, 1997, **389**, 418.
18. Butler, D., *Nature*, 1999, **397**, 195–2000.
19. Annual Report 1999–2000, Department of Biotechnology, Ministry of Science and Technology, Government of India, Biotechnology Information System Network (BTISnet), Chapter 8.
20. Gershon, D., *Nature*, 1997, **389**, 417–418.

Received 7 December 1999, revised accepted 2 June 2000