Chapter 22

Bioinformatics in Biodiversity Management

Pramod Tandon and Pallavi Bhattacharjee
Bioinformatics Centre, North-Eastern Hill University, Shillong 793 022

Biodiversity is very important for human survival and is the most precious and fragile resource in the world. Yet, in an increasingly modern and technological world, people often forget how fundamental biodiversity is to daily life and are unaware of the impact of its loss. We are eliminating populations and species faster than we can discover new ones. Nations can prosper only if their biodiversity is preserved. While the Earth's biotic inventory is far from complete, our records and data from specimens, fieldwork and other research have enormous value for building our knowledge of biodiversity on our planet and for applied uses in solving real-world problems. There are two areas in biology where enormous amounts of information are generated. One is in molecular biology, which deals with base sequences in DNA and amino acid sequences in proteins and the other is the enormity of biodiversity related data. Information Technology is being used to tackle these problems with procedures, which come under the label of Bioinformatics. Bioinformatics harnesses the power of computational and information technologies to organise and analyse biological data and deliver them to users throughout the world. In the present day world, the Internet and World Wide Web (WWW) have proved to be powerful tools for linking and utilising the data of biodiversity and at the same time, help devicing methods to maintain and exploit it in a sustainable manner.

INTRODUCTION

Biodiversity or Biological Diversity is the term for the variability among living organisms from all sources. These include diversity within species, between species and of ecosystems. Diversity is the key for ensuring the continuance of life on earth. It is also a fundamental requirement for the adaptation, survival and continued evolution of species. The concept of biodiversity represents how life is organised and how it interacts on our planet. These interactions can take place on scales ranging from the smallest, at the gene/chromosome level, to

organisms, ecosystems and even to entire landscapes. The entire gamut of biodiversity is often referred to as *'ecosystem complexity'* (Sala *et al.*, 1999).

Bioinformatics is a discipline that generates computational tools, databases and methods to organise and analyse biological data. These data are derived from research collections, experiments, remote sensing, modelling, database searches and instrumentation – and delivered to users throughout the world. It comprises the study of DNA structure and functions, gene and protein expression; protein production, structure and functions; genetic regulatory systems and clinical applications.

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. 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.

REQUIREMENT OF BIODIVERSITY CONSERVATION

Biodiversity conservation is of prime importance to the Earth's health and would ultimately determine our fate and existence. However, our ignorance of most of the world's biodiversity, particularly at the level of individual species, poses a major threat. Under the current scenario, about 25% of the higher-plant species are expected to be lost in the next few decades and another 25% by the end of the 21st century (Tandon, 2004b, Singh and Khurana, 2002). There is a heavy loss of species, populations, and ecosystems due to the loss of habitat and destruction. Alterations in ecosystem composition, such as the loss or decline of a species, can lead to a loss of biodiversity. The threats to biodiversity, posed by the destruction of the natural environment by urbanisation, global warming and other factors, have given rise to a high level of public awareness in this regard (Tandon and Kumaria, 2005).

Although biodiversity conservation is not a new field in itself, it has realigned itself and brought the existing sciences of systemics, ecology and evolution closer. It has also given a platform for a unified cry for the conservation of our ecosystems. Conservation is not an end in itself, but a means of ensuring that plant and animal genetic resources are available for use by present and future generations. Efforts for conservation and management of our natural resources must derive from a set of clear objectives, mechanisms for action and commitment from all stakeholders. Apart from this, halting the process of degradation and species loss requires specialised solutions and an understanding of ecological processes. Protecting biodiversity does not merely involve setting aside chunks of area as reserves. Instead, all ecological processes that maintains the area's biodiversity, involving complex interactions between several species of plants

and animal, need to be conserved (Terborgh, 1999). "Ecologically destructive economic activities" are inefficient, not merely because of the resulting resource misallocation, but also because of the excessive scale of activity levels. In order to bring about sustainable resource conservation and management, it is essential to adopt several different approaches for managing our forests and biodiversity.

ROLE OF BIOINFORMATICS IN BIODIVERSITY CONSERVATION

Bioinformatics is a modern science that applies database and computer modelling technologies for the study of biodiversity and habitats. It has been realised that the first step towards biodiversity conservation would be to establish an elaborate catalogue that would describe and classify all the surviving species of the world and simultaneously, take appropriate steps for conservation. This would undoubtedly be an overwhelming task and would need sustained and dedicated effort, as revealed by the researchers and biotechnicians of the Human Genome Project. This is where the role of informatics has gained importance. In the current scenario, bioinformatics is placed at the core of contemporary biology, in fields ranging from structural biology to genomics to biomedical imaging. Ready access to data and analytical tools has fundamentally changed an investigator's approach to problems. Today, Internet presents state-of-the-art tools for linking and utilising the extraordinary assets of natural history institutions.

Biodiversity informatics helps build tools to provide a standard format for the collection of data, offers an infrastructure for linking collections and supports predictive models for habitat expansion or destruction. A number of software packages relating to biodiversity and molecular systemics are available on the Internet. These are either downloadable shareware or are web-based applications. It is important to note that information management for biodiversity conservation is somewhat different from traditional scientific disciplines, which use mechanistic, laboratory and field studies. The vision is to create information systems for the inventory, analysis, reporting and visualisation of the vast natural resources and biological life on Earth. Currently, one of the main technologies for the inventory of biodiversity is the Geographic Information Systems, which is software for digitising geographic information and producing maps. [*http://www.usgs.gov*]

One of the major problems in the management of biodiversity information is its sheer magnitude and variety. The data is available in different formats. Moreover, biological data can be politically and commercially sensitive, leading to conflicts of interests. Because of all these complexities, it is feared that biodiversity information may not be very amenable to automatic correlation, analysis, synthesis and presentation (Schnase, 2000). Integration and analysis of such diverse data requires novel data mining tools that are either under development or still at the conceptual level. Coordinated effort is required nationally to collate, store, analyse and interpret all data related to biodiversity. Bioinformatics can play a determinable role in biodiversity conservation.

COMPUTERISATION OF BIODIVERSITY DATA

Till recently, biodiversity data were held as card indexes and taxonomic monographs, as well as physical specimens in museums and herbaria. Data storage and retrieval are vital for the research community. This was true even before gene-sequencing efforts put data storage and its analysis in the centrefold of active taxonomy research. Recorded information on specific taxonomic groups was only available to a few experts affiliated to various premium institutions of the world. In addition, information on biodiversity was only available after detailed search of a number of disjoined databases or could only be collected from paper-based libraries. Databases tailored exclusively for biodiversity conservation studies, based on field observations and scientists' notebooks, in addition to museum data, are perhaps the first step in this direction to manage the existing biodiversity information.

Estimates say that we know only 10% of the species living on earth (Lane *et al.*, 2000). In India the Ministry of Environment and Forests (MoEF) has developed a catalogue comprising 47,000 species of plants and 80,000 species of animals (MoEF, 1999). A number of species still remain undiscovered, as roughly 30% of India's geographical area still remains to be fully explored. An inventory and a comprehensive informatics system has to be developed which will help in the efforts towards the conservation of biodiversity (Gundu *et al.*, 2005).

The gradual involvement of information technology to biological science has, in effect given birth to the bioinformatics revolution, which is finally enabling researchers to communicate effectively with each other and derive the benefits of this modern science. Computerisation of large amounts of biodiversity data has developed databases, which are now available in the public domain through the Internet and are easily accessible to one and all (Sugden and Pennisi, 2000). However, just getting biodiversity related information on-line is not the only issue here, but studying this data and devising ways to protect our precious biodiversity is the ultimate aim. Specific examples of research directions and applications to the management and conservation of biodiversity are:

- Large distributed biological databases
- Biologically relevant base data
- Species distribution modelling
- Biodiversity modelling
- Ecological process modelling
- Reserve system planning
- Error estimation
- Visualisation/map making
- Interactivity

(Source: Venkatesh and Bansal, 2005)

BIODIVERSITY DATABASES

Global communication has expanded remarkably since the inception of the

internet. Technical advances have made the distribution of data from major centres to remote parts of the world possible, if those data are in digital form. Biodiversity is found around the world; however, it is not distributed evenly across the face of the planet. Then information about biodiversity (natural history collections, library materials, databases) likewise, is not distributed evenly around the globe. Three-quarters or more of the data about biodiversity are stored in the developed world. However, most of the data that may be needed can't be transferred because, either they are not digitised, or the capacity to handle digital information is lacking, or both.

Information on different aspects of biodiversity is expanding rapidly. When genome sequencing projects started to pour data for the scientific community, computer-based databases were established for processing, storing and sharing this data all around the world. However, as Maurer *et al.* (2000) rightly observed, large sophisticated databases do not just happen. They have to be planned carefully with interoperability in mind. Recent browser technology, such as extensible markup language (XML), allows a user to search several databases. Benefits of this database approach are manifold, some of them being: reduction of redundancy and inconsistency of information, fast and simultaneous sharing of data for a large number of users, standards can be maintained and last, but not the least, integrity and security restrictions of data information can be enforced.

SOME OF THE GLOBAL BIODIVERSITY DATABASES

Species 2000 is a "Federation" of database organisations working closely with users, taxonomists and sponsoring agencies. It is an ambitious undertaking, which aims to enumerate all known species of plants, animals, fungi and microbes on earth, in a comprehensive resource that ties together a variety of smaller online taxonomic indexes. [*http://www.sp2000.org/*]

Species Analyst provides access to a variety of natural history databases through a Web interface. Data includes museum collection information such as date, specimen, latitude and longitude of the collection site and the collector's name etc. This can be downloaded in a variety of formats. The site also generates a global distribution map of the collection sites for a given query and can link users to online tools for further analysis and modelling of the data at the San Diego Supercomputer Centre.
[*http://speciesanalyst.net/*]

Global Biodiversity Information Facility is responsible for the digitisation and global dissemination of primary biodiversity data, so that people from all countries can benefit from this information. This establishes a standard method for data exchange about specimens for researchers in different countries. [*http://www.gbif.org/*]

Tree of life (ToL) is a collaborative effort of biologists from around the world. In more than 3000 World Wide Web pages, the project provides information about the diversity of organisms on earth, their evolutionary history (phylogeny) and characteristics. Each page contains information about a particular group of organisms. ToL pages are linked to one another hierarchically, in the form of the evolutionary tree of life. Starting with the root of all life on earth and moving out along diverging branches to individual species, the structure of the ToL project thus, illustrates the genetic connections between all living things. [*http://www.tolweb.org/tree/phylogeny.html*]

Integrated Taxonomic Information System is a cooperative effort of U.S., Canadian and Mexican agencies, building "an online, scientifically credible list of biological names" for North American taxa.
[*http://www.itis.usda.gov/*]

NatureServe is an "online encyclopedia of life" sponsored by the Association for Biodiversity Information. It provides data on nomenclature, conservation, geographic distribution and the life history of more than 50,000 U.S. and Canadian plant and animal species and ecological communities. Users can search by common or scientific species name, species association and other criteria.
[*http://www.natureserve.org/*]

In addition to the above, the following sites provide links to biodiversity information:

1. http://nature.ac.uk
2. http://www.nscalliance.org/bioinformatics/index.asp
3. http://www.biodiv.org/programmes/default.shtml

SOME OF THE INDIAN BIODIVERSITY DATABASES

Biotechnology Information System (BTISNet)

It is a major Satellite and Terrestrial network on biotechnology in the country, set up by the Department of Biotechnology (DBT), Government of India. It has established and networked various bioinformatics centres in the country, that are actively involved in research and development in the field of modern biology and bioinformatics. BTISnet has now become the single largest information resource for all references to biotechnology related literature, scientific data, patent information, policy matters and related issues in India. [*http://www.btisnet.nic.in/*]

Environmental Information System (ENVIS)

It is a decentralised system with a network of distributed subject-oriented Centres, ensuring integration of national efforts in environmental information collection, collation, storage, retrieval and dissemination to all concerned (decision makers,

policy planners, scientists, engineers and research workers, etc.), all over the country. [*http://www.envis.nic.in/*]

Agricultural Databases and Information on Sacred Groves

M.S. Swaminathan Research Foundation is involved in genetically characterising coastal bioresources, in particular mangroves, identifing, isolating and characterising novel genetic combinations from mangroves. [*http://www.mssrf.org*]

The ICRISAT programme

'International Crops Research Institute for the Semi-Arid Tropics' is an international organisation devoted to science based agricultural development. Its work involves helping countries apply science to increase crop productivity and food security, reduce poverty and protect the environment. ICRISAT focusses on the farming systems of the semi-arid tropical areas of the developing world, where erratic rainfall, low soil fertility and extreme poverty are formidable constraints to agricultural development. ICRISAT conserves the seeds of 113,000 lines or types of crops and related information that are important for diet. [*http://www.icrisat.org*]

Agricultural Research Information Network (ARISNET)

It is a network of agricultural research, extension and educational institutions in India over a Satellite based Computer Communication Network of NIC called NICNET. It provides global access of information to scientists and facilitates them to build up their links with the national and international research community and promotes voluntarily exchange of information between different institutions and sharing of their resources. [*http://www.dacnet.nic.in/arisnet/*]

Biodiversity Characterisation using Remote Sensing/Geographic Information Systems

This is a multi-institutional programme on bioprospectiong of biological wealth, jointly supported by Department of Space (DOS) and DBT. The study includes integrating spatial data, like land use and land cover data, disturbance regimes and biological richness maps with non-spatial data like taxonomic and genetic information and creating landscape level information linked with a comprehensive plant species database. [*http://www.gisdevelopment.net/magazine/gisdev/2002/nov/mabc.shtml*]

BODHI

It is developed at the Supercomputer Education and Research Centre, Indian

Institute of Science, Bangalore. It is a database system, which stores plant biodiversity information. The unique feature of BODHI is that it seamlessly integrates diverse types of data, including taxonomic characteristics, spatial distributions and genetic sequences, thereby spanning the entire range from molecular to organism-level information. [*http://eprints.iisc.ernet.in/archive/00001033/*]

CDROMs on Marine Prawns, Marine Crabs, Mangroves, Lignicolous Fungi and corals of India

National Institute of Oceanography has created database that collates information on marine biology and biotechnology. The published literature about the biota in marine and estuarine environment of India is made available in an electronic format. [*http://www.nio.org*]

Endemic Trees of the Western Ghats

This database presents the geographic database of the endemic evergreen and semi-evergreen tree species of the Western Ghats, with maps, species synonyms, superimposition of forest types and species distribution areas with pictures and graphics. [*http://www.etfrn.org/etfrn/newsletter/nl22_pub.html#datta*]

Ethnobotany

National Botanical Research Institute (NBRI) is involved in extensive and intensive ethnobotanical surveys for the collection, identification and documentation of plants and plant products of ethnobotanical value, used by the tribals of Uttar Pradesh, Uttaranchal, Madhya Pradesh, Himachal Pradesh and Andaman and Nicobar Island.
[*http://www.nbri-lko.org/randdarea/biodiversity/ethnobotany.htm*]

Sampada

It is a natural science collection computerisation initiative by the Information Division of the National Chemical Laboratory, Pune. It is an application for the storage and retrieval of data from biological collections. This includes information about museum specimens, field information and notes pertaining to biodiveristy. [*http://www.ncbi.org.in/sampada*]

Digitally Compiled Flora

Jawaharlal Nehru Centre for Advanced Scientific Research has a collation of data from over 100 Regional Flora, Monographs and other literature. [*http://www.jncasr.ac.in/bdu/*]

Medicinal plants database

Foundation for Revitalisation of Local Health Traditions's (FRLHT) contains botanical database of medicinal plants and provides correlation between over 0.1 million botanical and vernacular names of around 6000 plant species (herbs, shrubs, trees, climbers, orchids, grasses, tubers and even lichens), with 5000 images and over 460 literature references to corresponding systems of Indian medicine viz., Ayurveda, Siddha, Unani, Swa. Rigpa (Tibetan), folk medicine, homeopathy and allopathy. [*http://www.frlht-india.org/html/crg.htm*]

National Register of Green Grassroots Innovations and Traditional Knowledge

National Innovation Foundation helps grassroot innovators and traditional knowledge holders build their links with institutional scientists, technologists and designers so as to add value in their technologies. It provides support to help convert innovations into enterprises, protects their intellectual property rights and creates a culture of creativity and innovation in society. [*http://www.nifindia.org/5th_announcement.htm*]

National Wildlife Database and Zoo Database

Wildlife Institute of India maintains a National Wildlife database that deals with all the protected areas and biodiversity of wild life. It maintains an information system on zoo and wildlife. [*http://www.envfor.nic.in/wii/wii.html*]

Birds of India

Centre for Biodiversity Informatics at the National Chemical Laboratory, Pune, maintains a checklist of Indian waterbirds and wetland dependant birds. [*http://www.ncbi.org.in/wbni/final_checklist.html*]

Plants of India and Legume Database of South Asia

NBRI has created a specialised database for species diversity on Leguminosae. The premier purpose of ILDIS is to provide a global species diversity information service of over 19000 taxa of the plant family Leguminosae. [*http://www.nbri-lko.org/randdarea/bioinformatics/bioinformatics.htm*]

Sahyadri

It is a database of the Western Ghats flora and fauna and also a Biodiversity Bibliography developed and maintained by the Centre for Ecological Sciences, Indian Institute of Science. [*http://wgbis.ces.iisc.ernet.in/biodiversity/*]

Helminth Parasite Spectrum of Northeast India

The Bioinformatics Centre at North-Eastern Hill University, Shillong, maintains a database of parasites of Northeast India. Software has the facility to view scientific classification details, host, habitat, place of occurrence and images of parasites. This database will be available for public access via the Internet in the near future.

MOLECULAR DATABASES

Applied research in biodiversity is critically dependent on molecular and genetic databases. Such databases, which can be used for research on biodiversity, are being developed and maintained by various organisations at the global level. 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 the internet and by regularly holding meetings with the International DNA Data Banks Advisory Meeting and the International DNA Data Banks Collaborative Meeting. In addition to the above three, there are numerous organisations that are actively involved in the 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 analysing genome data and disseminates biomedical information — all for the better understanding of molecular processes affecting human health and diseases. [*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, USA) and the DNA Database of Japan (DDBJ). [*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 (responsible for GenBank database) in USA. DDBJ is the sole DNA data bank in Japan, which is officially certified to collect DNA sequences from researchers and to issue the internationally recognised accession number to data submitters. DDBJ also provides numerous tools for the retrieval and analysis of data. [*http://www.ddbj.nig.ac.jp/*]

Some of the most commonly used genetic databases of the world are:

PubMed (NCBI)

PubMed, available via the NCBI Entrez retrieval system, is located at the National Institute 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. In addition, PubMed provides a Batch Citation Matcher, which allows users to match their citations to PubMed citations using bibliographic information such as journal, volume, issue, page number and year. [*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. Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism and a table of features that identifies coding regions and other sites of biological significance, such as transcription units, sites of mutations or modifications and repeats. Protein translations for the coding regions are included in the feature table. Bibliographic references are included, along with a link to the Medline unique identifier for all published sequences. Sequence data are submitted to the GenBank by individual scientists from around the world, as well as by large centres involved in the Human Genome Project. [*http://www.ncbi.nlm.nih.gov/Entrez/*]

OMIM (NCBI)

OMIM™ (Online Mendelian Inheritance in Man) database is a catalogue 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 a qualified physician for diagnosis and for answers to personal questions. [*http://www.ncbi.nlm.nih.gov/Omim/*]

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

Sequence Retrieval System (SRS)

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. The user chooses databases of interest from a large list of databases, categorised

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 a 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 programmes, also allowing queries of DNA vs. a protein database, or vice versa. In these variants, there is a further distinction, regarding the location of gaps: one may assume that gaps occur only in the codon frames corresponding to amino-acid insertion; alternatively, one can assume the 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

It is a Basic Local Alignment Search Tool 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 characterised 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 functioning of uncharacterised proteins. [*http://www.ebi.ac.uk/services/*]

TOOLS AND SOFTWARE FOR INTEGRATING ONLINE BIODIVERSITY DATA

Alice Software

This website produces easy to use software for the creation, management and publication of biodiversity data. One who is interested in the cladistics, collation, management or publication of information about biological organisms or wishes to publish species lists or annotated checklists etc, and share with others can use this software. [*http://www.alicesoftware.com*]

Spice for Species 2000

Its approach for creating a catalogue of life has been defined in the Species 2000 Programme (refer above to the list of 'some of the global biodiversity databases'). It makes use of autonomous Global Species Databases (GSDs), created and maintained by appropriate experts. These provide the central array in a federated architecture. This project investigates how uniform authority and wide taxonomic coverage can be obtained through accurate mapping of quality GSDs from many organisations. It investigates how the federated system can be designed to overcome the inherent problems of extreme heterogeneity, scalability and stability. [*http://www.systematics.reading.ac.uk/spice/*]

Biodiversity Pro Software

This software is used for calculations based on single samples; also known as univariate diversity measures, it has the following features.
[*http://www.nhm.ac.uk/zoology/bdpro/*]

- *Alpha Calculations*: Abundance Plot- Dominance, Rank, Abundance Model-Log-Series, Broken Stick, Rarefaction, Diversity Indices-Shannon, Alpha, Caswell, Berger-Parker, Simpson, Hill, Margalef, McIntosh.
- *Beta Calculations*: SHE Analysis, Species Richness, Species Distribution
- *Multivariate Calculation*: Principal Components, Correspondence Analysis, Cluster Analysis Non-Metric MDS
- *Comparisons*: Descriptive Statistics, Kulczynski, Mann-Whitney, Rank Correlation, Correlation, Variance-Covariance. ANOSIM
- *Tools*: Send Data, Transform Data, Standardise Data, Add-Ins Options

Platypus Software

It is a comprehensive relational database programme for faunal-based taxonomy, managing taxonomic, geographic, ecological, bibliographic and graphic information, all of which are organised around a graphically displayed checklist, etc. [*http://www.environment.gov.au/abrs/platypus.html*]

Linnaeus II 2.x

Expert Centre for Taxonomic Identification's (ETI) Linnaeus II interactive software facilitates biodiversity documentation and species identification. It supports the creation of taxonomic databases, optimizes the construction of easy-to-use identification keys, expediates the display and comparison of distribution patterns and promotes the use of taxonomic data for biodiversity studies. [*http://www.eti.uva.nl/products/linnaeus.html*]

LITCHI Project

This project is belong to a collaborative group of researchers at three U.K. universities, who are seeking to write "taxonomically intelligent" software for tying together various databases on species diversity. [*http://www.litchi.biol.soton.ac.uk/*]

CODA Software

Conservation Options and Decisions Analysis (CODA) assists in the design of networks of nature reserves or protected areas. It has been used for major reserve planning studies, as a teaching resource and for research in conservation planning methods. It provides a framework and a set of tools for regional conservation planning. Using CODA, one can build and compare alternative conservation proposals in a simple and flexible way. [*http://www.members.ozemail.com.au/~mbedward/coda/coda.html*]

MultiFlora Project

This project is used for "information extraction" techniques to place descriptive data "locked" in natural-language texts into a more structured electronic database, to facilitate automated analysis. [*http://www.cs.man.ac.uk/ai/Software/MultiFlora/*]

Phylip Software

PHYLIP (the PHYLogeny Inference Package) is a package of programmes for inferring phylogenies (evolutionary trees). Methods that are available in the package include parsimony, distance matrix and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites, distance matrices and 0/1 discrete characters. [*http://evolution.genetics.washington.edu/phylip.html*]

Biota Software

Biota helps us manage specimen-based biodiversity and collected data by

providing an easy-to-use graphical interface to a fully relational database structure. For ecologists, conservation biologists, reserve managers and biogeographers, for taxonomists, systematists and collections managers, Biota offers rigorous tools for recording data and images for specimen determination, as well as for revision and evolutionary studies. [*http://viceroy.eeb.uconn.edu/Biota*]

The Specify Collections Management System

The SPECIFY software provides a standard software system for diverse disciplines of biodiversity to manage and collect information and to improve the efficiency and effectiveness of development. [*http://www.usobi.org/*]

The Biodiversity Species Workshop

Biodiversity Species Workshop is a place for the experimental analysis and mapping of the distribution data of your own species.
[*http://biodi.sdsc.edu/bsw_home.html*]

World Map

World Map is a Software for exploring geographical patterns in diversity, rarity and conservation priorities from large biological datasets. [*http://www.nhm.ac.uk/science/projects/worldmap/*]

BIOINFORMATICS IMPLICATIONS OF THE INTERNATIONAL BIODIVERSITY CONVENTIONS

Over the past 23 years, nations have agreed to a number of international treaties intended to ensure the on-going sustainability of the biota of this planet. Five of these are completely or closely related to biodiversity:

- The Convention on Biological Diversity (CBD)
- The Convention on International Trade in Endangered Species (CITES)
- The Convention on Migratory Species (CMS or the "Bonn" Convention)
- The Ramsar Convention on Wetlands of International Importance
- The World Heritage Convention

India is one of the countries which has confirmed its commitment to the principles of the above treaties. These treaties, in conjunction with others, aimed at the stabilisation of the atmosphere (the Framework Convention of Climatic Change and Montreal Protocol), or at specific aspects of environmental degradation (Convention to Combat Desertification), may mark the beginning of a new era of international co-operation, wherein nations act in consort to ensure the sustainability of the biosphere. It is still a beginning. The next few

decades will tell us if the enthusiasm and sense of urgency, which surrounded the Earth Summit in Rio in 1992, will itself be sustainable. If the hope and impetus engendered by the Earth Summit is to bear fruit, this new wave of major treaties must be fully and efficiently implemented, and must increase participation so as to be truly global.

Firstly, although not all countries have joined, participation is extensive and mostly includes a broad spectrum of industrialised, emerging and developing nations. Secondly, all of them require, in fact, significant information reporting and exchange.

Moreover, these treaties were formed in the new "Information Age", an age which has in its command information technology and infrastructure with a capacity to organise, communicate and exchange information rapidly and easily. The result is that "Biodiversity Information", formerly the domain of the notebook of the "naturalist", or the specimen in the museum, botanical garden or zoo, is now openly available to the full international scientific community, as well as the general public. One of the factors which will determine whether the treaties can, in the long run, achieve their purpose, is the ability of party nations, separately and collectively to take advantage of the appropriate tools of modern information technology ("informatics"). Moving towards this partnership will cause direct consequences to the way in which biological information is gathered, organised, maintained and disseminated within countries, that is, national "bioinformatics".

Bioinformatics implications in principle

Integration into National Strategies

The treaties insist or imply that consideration of biodiversity information must become integral to all national strategy developments and decision-making, with regard to the utilisation of biological resources. [*Implications:* National information systems and information flow must crosscut traditional discipline boundaries and bridges between the scientific and socio-economic sectors.]

External Impacts

The treaties require knowledge of the resources at risk, beyond national borders, and the potential consequences of actions and decisions on ecosystems. [*Implications:* Information must be exchanged and shared on a regional basis to complement national information. In addition, decision support systems, which are capable of forecasting consequences, are needed.]

Reporting Requirements

The treaties require nations to report, through a coordinating secretariat to a Conference of Parties, on actions taken and on summary statistics or indicators

reflecting the level of implementation of the treaty. [*Implications:* Bioinformatics systems must be available to consistently summarise national actions, progress and the status of biodiversity.]

It is abundantly clear that to meet the above demands effectively and efficiently, countries must develop a national biodiversity information system, which must be closely linked with national economic and social information systems.

Bioinformatics implications in practice

Decisions should be taken well in time so as to avert adverse consequences, or when proposed, the measures can have the best effect. The process of biodiversity data collection, integration and conversion into "information products" suitable for decision makers can be very time-consuming (even if assisted by modern computer systems). Just as it is necessary to have "water management infrastructure" in place to avert floods, before they occur, there must be an "information management infrastructure" in place, before particular instances of decision making become critical. This means:

- having available in advance, essential "core" datasets likely to be needed for a range of decision making purposes
- having the information systems in place with the processing capabilities to be able to quickly produce the specific information products required
- information exchange agreements and facilities already established.

This basic information management infrastructure must be developed to support a range or classes of biodiversity issues, in anticipation of likely decision making scenarios and requirements. The information systems capabilities would normally be required by specific expert institutions, which are the key custodians of the essential core datasets.
[*http://www.iupac.org/symposia/proceedings/phuket97/crain.html*]

BIODIVERSITY INFORMATION SYSTEM (BIS)

The DBT and DOS have joined hands to undertake studies on biodiversity characterisation. The main objectives of this project are to:

- identify the biorich areas
- evaluate the forest types for their value
- provide geographical information on the economically important species for bioprospecting.

The outcome of this project will assist forest managers, decision makers and the leading national institutes involved in genetic diversity and bioprospecting. BIS targets to provide user–level information regarding the hotspots of biodiversity.

However, it was felt that in addition to characterising information on biodiversity, it should also cater to decision-making, and monitoring and management of various resources. It is for this reason that the BIS was conceptualised as a 5-component unit.

- BioSPATIAL (Biodiversity characterisation at the landscape level)
- BIOSPEC (Bioprospecting and molecular taxonomy programme)
- FRIS (Forest resource information system)
- PhytoSIS (Species information system)
- Biocon SDSS (Spatial decision support system for biodiversity conservation prioritisation)

(Source: Roy, 2005)

Networking of Biodiversity Databases

For hundreds of years, explorers and scientists have collected plant and animal specimens from every region and habitat of the earth. Millions of these specimens are preserved with the great natural historical collections of the world, along with detailed notes and logs describing locations and collecting events. The collected data include records spanning many decades, showing where specific species of plants and animals lived, how they lived and how they responded to changes in the environment. This vast global archive of collection data is a priceless source of information on the earth's biosystems; however, most of this data are catalogued through incompatible, roll-your-own data-management applications, developed for individual museums. Scientists therefore, must study each collection in isolation, without the aid of advanced modelling and data-management techniques, for identifying patterns across multiple collections. In some cases, an individual collection may have only one or two specimens of a specific plant or animal—not enough for any meaningful predictions regarding the species. Other museums, however, may have hundreds of examples of this same plant or animal. Combining several collections into one study leads to more accurate conclusions and, in some cases, it leads to new conclusions that, otherwise would have been unobservable or statistically insignificant.

At present, there is inconsistency in biodiversity and its information that is distributed globally (Lane *et al.*, 2000). Biodiversity is mostly distributed in the developing world, whereas the information on biodiversity exists in the developed countries. For example, it is more likely that information pertaining to a plant in an African region may be found in some herbarium in Europe. Due to this, information is not readily available in the region where the biodiversity is mainly distributed and this in turn, is severely affecting initiatives for proper policy making by governmental and non-governmental agencies towards biodiversity conservation (Gundu *et al.*, 2005). By linking the world's natural history collections into a single virtual database, scientists can build a better and deeper understanding of questions such as:

- Where on the earth are you most likely to find a certain species (for example, a species of hawk or an endangered wildflower)?
- What climatic conditions are necessary for this species to thrive?
- How will the range of this species be affected by changes in the environment, such as changes due to global warming?
- What will happen if other competing species are introduced into the environment through human error or intervention?

Bioinformatics helps in accomplishing the task of moving natural history collection data from the gilded age to the information age by developing programmes and techniques that apply to all forms of habitat modelling, regardless of whether the supporting data originates from natural history collections.

India possesses a wealth of knowledge associated with biodiversity, be it the orally held knowledge of folk healers or herders, or the traditional knowledge codified in Ayurvedic, Sidha or Yunani texts. Information also exists in the form of specimens, grey literature, such as unpublished reports of the District Flora Project or Forest Working Plans and books, monographs and scientific papers. Potential exists in building upon the country's biodiversity resources and its associated knowledge; upon promoting biodiversity-based enterprises in the modern and traditional sectors; upon developing biotechnology industries at the cutting edge of new technologies, as well as encouraging local-level value-addition to biodiversity resources.

The Biological Diversity Act visualises the establishment of a National Biodiversity Authority (NBA), State Biodiversity Boards (SBB) and Biodiversity Management Committees (BMC) at the level of all local bodies, namely, Gram, Taluk and Zilla Panchayats, as well as Municipalities and Corporations. The NBA working with SBBs and BMCs will have the responsibility of handling all issues related to conservation and the sustainable use of biodiversity and bioresources of our country. This would need a well-organised in place information system on India's biodiversity resources and associated modern as well as traditional, codified as well as oral knowledge, so as to do justice to achieve their objectives. Such a system will have to deal with a whole range of spatial scales from local to national, as well as link properly with global databases. It will also have to address issues such as linking to information on the large holdings of biological specimens of Indian origin located in herbaria and museums abroad.

FUTURE APPROACHES

India's strength in Information Technology and Biotechnology is in an excellent position to turn the above array of significant challenges into welcome opportunities. This calls for the networking of the country's existing biodiversity databases so as to take advantage of synergies and to link all of these to activities leading to value-addition. As a part of this process, the existing biodiversity databases will need to be considerably augmented and strengthened and new

ones created. We will have to come up with novel ways of bringing on board, the substantial knowledge base of the country's barefoot ecologists and grass-roots innovators. We also need to devise a country-wide decentralised system of monitoring biodiversity. Such a decentralised system could serve to enhance the quality of education by engaging teachers and students in the first hand understanding of biodiversity and its associated knowledge and in creating, using and managing electronic databases, including those employing Indian languages. Computer-aided storage and retrieval systems of plant and animal genetic resources will form an important tool for developing technology packages for conservation and ensuring the exchange of information (Tandon, 2004a)

A good beginning has been made in organising a part of this information in the form of electronic databases. There are, however, many lacunae and we may summarise the situation as follows:

- Databases are in heterogeneous formats.
- Few are on the web, while many are available offline.
- Some of these are well structured, others are largely project /species specific and/or unstructured.
- These databases exist independently.
- There is no framework to link the scattered data so as to facilitate exchange of data amongst the different databases.
- There is no meta-data.
- The gap between data managers and data producers is widening.

Well-planned institutional arrangements and legal provisions at the national level, as well as in terms of links with international agencies to address these manifold concerns regarding biodiversity conservation are required. DBT with its National Bioresources Development Board, the MoEF and Governmental agencies such as Ministries of AYUSH (Ayurveda, Siddha, Unani and Homeopathy), Health and Commerce, Council of Scientific and Industrial Research, Indian Council for Agricultural Research and DOS will play a significant role. The challenge before us is to set standards and make technological choices that would facilitate the networking of databases and add real value to the information being brought together, and at the same time, maintain the autonomy of the various databases, so as to ensure the security of the data and protect all legitimate intellectual property rights. This would obviously have to be worked out as a group exercise by all concerned institutions and individuals, to involving an inventory of the on-going Indian efforts and a review of the various pertinent standards. These surveys would have to address issues of data characterisation and classification, validation and authentication, organisation and structuring, storage, archival, warehousing, retrieval, dissemination, sharing and interoperability, access, security, visualisation, analysis and value addition, use of multiple languages and capacity building needs. [*http://wgbis.ces.iisc.ernet.in/biodiversity/Network*]

Conservation, sustainable utilisation and management of biodiversity should become an important agenda, as it will be the key to the survival and economic

well-being of man kind in the 21st century (Tandon, 2003). Sciences dealing with biodiversity management involve analysis, maintenance, cataloguing and conservation of all different life forms. Bioinformatics, being an applied discipline, it utilises computational tools to conceptualise biology, and would be capable of assisting in biodiversity conservation and its sustainable utilisation. Bioinformatics would provide methods for collecting and storing genes, detecting and eliminating diseases in gene bank collections, identifying useful genes, improving techniques for long-term storage, distributing germplasm to users in a safer and more efficient manner. Optimised use of advanced information technology tools for the management of biological resources is the need of the hour. Government agencies, institutions and NGOs will require working as a group to network the biodiversity databases of the country. Advanced computation and data archives would produce quantitative improvements in the analysis of biodiversity, through more accurate and efficient solutions.

ACKNOWLEDGEMENTS

The authors thank the Department of Biotechnology, Government of India for providing the facilities at the Bioinformatics Centre, North-Eastern Hill University, Shillong, India.

REFERENCES

Gundu RK, Bose B, Swamynathan S, Sreenu VB, Pavan N, Acharya S and Nagarajaram HA (2005) Frontiers in Bioinformatics Research: The Biodiversity Issues. In: Biodiversity: Status and Prospects (Tandon P, Sharma M and Swarup R eds), Narosa Publishing House, India p 197-201

Lane MA, Edwards JL and Nielsen ES (2000) Biodiversity Informatics: The challenge of rapid development, large databases and complex data. In: Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt

Maurer SM, Firestone RB and Scriver CR (2000) Science's neglected legacy. *Nature* 405 (6783): 117-120

MoEF (1999) National policy and macro-level action strategy on biodiversity. Ministry of Environment and Forests, Government of India, New Delhi

Roy PS (2005) Biodiversity Conservation: Perspective from Space. In: Biodiversity: Status and Prospects (Tandon P, Sharma M and Swarup R eds), Narosa Publishing House, India p 170-196

Sala OE, Chapin FS, Gardner RH, Lauenroth WK, Mooney HA and Ramakrishnan PS (1999) Global change, biodiversity and ecological complexity. In: The Terrestrial Biosphere and Global Change: Implications for Natural and Managed Ecosystems (Walker B, Steffen W, Canadell J and Ingram J eds), Cambridge: Cambridge Univ. Press

Schnase JL (2000) Research directions in biodiversity informatics. In: Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt

Singh JS and Khurana E (2002) Paradigms of Biodiversity: An Overview (Biodiversity, Ecology and Conservation- Reviews and Tracts). Proceedings of the Indian National Science Academy. Part B, 68(3):273-296

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

Tandon P (2003) Biodiversity – A scientific approach: agenda for the 21st century. National Academy Science Letters. 26 (5&6): 111-118

Tandon P (2004a) Conservation and sustainable development of plant resources of North-East India. Man and Society 1 (1): 49-59

Tandon P (2004b) Role of biotechnology in the 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 (Banerjee SP and Mukherjee SP eds), Auto Print and Publicity House, Kolkata, India, p 40-67

Tandon P and Kumaria S (2005) Prospects of plant conservation biotechnology in India with special reference to the northeastern region. In: Biodiversity: Status and Prospects (Tandon P, Sharma M and Swarup R eds), Narosa Publishing House, India p 79-92

Terborgh J (1999) Requiem for Nature. Washington, DC, Island Press, USA 234 p

Venkatesh S and Bansal M (2005) Role of Informatics in Biodiversity Conservation: Overview of Available Resources. In: Biodiversity: Status and Prospects (Tandon P, Sharma M and Swarup R eds), Narosa Publishing House, India p 134-145