www.sciencevision.in

Colloquium

Sci Vis 11 (3), 137-140
July-September, 2011
ISSN (print) 0975-6175
ISSN (online) 2229-6026

Biological databases with emphasis on biodiversity and conservation

P. Borah

Department of Microbiology, College of Veterinary Science, Assam Agricultural University, Guwahati 781022, India

ABSTRACT

Bioinformatics implements the use of biology, computational mathematics, computer science and information technology. Through the combination of these methods, scientists are able to store and compare the information from all kinds of species and how they evolve. The complex and voluminous data of biodiversity can be digitalised for easy accession, analysis and interpretation. It makes easy survey, documentation and measurement of biodiversity data. The data bases, management and their applications in general as well as in relation to biodiversity and conservation are discussed.

Key words: Databases; biodiversity; conservation.

INTRODUCTION

A Database Management System (DBMS) is a software system designed to store, manage, and facilitate access to databases. It is a collection of information, usually stored in an electronic format searchable by a computer.[1]

BIOLOGICAL DATABASES: SOME STATISTICS

Biological databases are libraries of life science information collected from scientific experiments, published literature, high throughput experiment technology, and computational analyses. They contain information on genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

There are more than 1000 different databases reported so far. A total of 968 databases were reported in *The Molecular Biology Database Collection: 2007 Update*.[2] Database sizes vary from <100 kB to >100 GB (e.g. EMBL >500 GB). There are DNA databases of size >100 GB, protein databases of 1 GB and 3D structure databases of 5 GB. Updating (adding new data) frequency of various online databases varies from daily to annually. As a rule, all these databases are freely accessible.

WHY SEARCH BIOLOGICAL DATABASES?

1. To generate new sequence, that is to

Proceedings of the "National Level Workshop on Random Amplified Polymorphic DNA (RAPD) Markers and It's Applications" organized on 20-21 May 2011 by the Departments of Biotechnology & Zoology, Bioinformatics Infrastructure Facility & State Biotech Hub, Mizoram University, in association with MIPOGRASS, and sponsored by DBT.

know if it is already in the data bank or to search for reported homologous sequences.

2. To find out about the gene annotation, literature, similar non-coding sequences like repetitive elements and regulatory regions.

3. To search for homologous proteins or protein families.

4. To identify and verify PCR priming sites.

WHAT IS ANNOTATION?

Annotation means extraction, definition and interpretation of features on the genome sequence. It is derived by integrating computational tools and biological knowledge for example, known and predicted genes. Some databases are referred to as "annotated databases" meaning that they contain sequence, comments, literature references, notes on experiments, etc.

TYPES OF BIOLOGICAL DATABASES

Biological databases can broadly be categorized into:

1. Generalized databases (DNA, proteins and carbohydrates, 3D-structures)

2. Specialized databases (EST, STS, SNP, RNA, genomes, protein families, pathways, microarray data, etc.)

There are two main classes of generalized databases: DNA (nucleotide) and protein databases.[3] The large DNA (nucleotide) databases are GenBank at National Centre for Biotechnology Information, NCBI (US); European Molecular Biology Laboratory (EMBL) at European Bioinformatics Institute, EBI (UK); DNA Databank of Japan, DDBJ (Japan); and GO-Gene Ontology (by The Gene Ontology Consortium).

Leading protein databases include SWISS-PROT- Curated protein sequence database by Swiss Institute of Bioinformatics; TrEMBL (high level of annotation)- Translated EMBL: computer annotated protein sequence database at EBI, UK; and PIR (protein identification resource)- Protein Information Resource at Georgetown University Medical Center (GUMC).

Important specialized databases are ESTs (Expressed Sequence Tags); STSs (Sequence-Tagged Sites); SNPs (Single Nucleotide Polymorphisms); Organismal Genomic databases: Human (GDB), mouse (MGB), yeast (SGB), fly; HTGS (High Throughput Genomic Sequences); RNA; tRNAs, rRNAs, small RNA's & others; Protein families: PROSITE, PRINTS, BLOCKS; Pathways: metabolic, regulatory etc. - EMP, PathDB; Microarray data: expression data: GeneX, ArrayExpress, Stanford, Gene Expression Omnibus (GEO).

Biological databases can also be categorized into primary: archival - experimental data with some annotation (interpretation) and secondary: curated. In curated databases, records are added only after they have been through a curation process, i.e. checked for accuracy, and additional information (annotation), scientific judgments are made as data are cleaned up and merged. Examples of curated databases: SWISS-PROT, OMIM, RefSeq, LocusLink.

IMPORTANT BIOLOGICAL DATABASES

GenBank was created in response to the explosion in sequence information resulting from panoply of scientific efforts such as the Human Genome Project. It is an annotated collection of all publicly available DNA and protein sequences and is maintained by the National Center for Biotechnology Information (NCBI). It contains 7 million sequence records covering almost 9 billion nucleotide bases. Sequences are most often directly submitted by individual investigators through tools such as Sequin or through ''direct deposit'' by large genome sequencing centers.

GenBank, or any other biological database for that matter, serves little purpose unless the database can be easily searched and entries retrieved in a usable, meaningful format. Otherwise, sequencing efforts have no useful end,

since the biological community as a whole cannot make use of the information hidden within these millions of bases and amino acids. Much effort has gone into making such data accessible to the average user, and the programs and interfaces resulting from these efforts are the focus of this chapter. The discussion centers on querying the NCBI databases because these more ''general'' repositories are far and away the ones most often accessed by biologists, but attention is also given to a number of smaller, specialized databases that provide information not necessarily found in GenBank.[4]

DNA Data Bank of Japan (DDBJ) is located at http://www.ddbj.nig.ac.jp and contains DNA data, retrieval & analysis tools.

The Genome Information Broker (GIB, http://gib.genes.ac.jp) includes > 50 complete microbial genome data and Arabidopsis genome data.

The Human Genome Studio (HGS, http://studio.nig.ac.jp) provides a set of sequences being as continuous as possible in any one of the 24 chromosomes.

EMBL Nucleotide Sequence Database is located at http://www.ebi.ac.uk/embl/, maintained at European Bioinformatics Institute (EBI). Major contributors are individual authors and genome project groups. DB releases are produced quarterly. Free access via ftp, email, and WWW interfaces. EBI's Sequence Retrieval System (SRS) – network browser for databanks, integrates and links the main nucleotide and protein DBs.

Ensembl is located at http://www.ensembl.org/ and is a comprehensive source of stable automatic annotation of the human genome sequence + confirmed gene predictions. It is available via interactive web site or flat files. It is also an open source SWE to develop a portable system to handle sequence analysis to data storage and visualization.

STACK is located at http://www.sanbi.ac.za/Dbases.html. It is a tool for detection and visualization of expressed transcript variation in the context of developmental and pathological states. It is generated at least four times a year. The tools for its generation are available at http://www.sanbi.ac.za/CODES

TIGR Gene Indices is located at http://www.tigr.org/tdb/tgi.shtml. It is a collection of transcribed sequences in a variety of organisms. Gene indices are constructed for selected organisms by first clustering, then assembling EST, and annotated gene sequences from GenBank.

BIODIVERSITY

The term 'biodiversity' comprehends the totality and variability of species, genes and the ecosystems they occupy. Biodiversity is usually used to refer biological diversity at three levels such as (1) genetics, (2) species and (3) ecology. Biodiversity is the variation of life forms within a given ecosystem, biome, or for the entire Earth. The term is often used as a measure of the health of biological systems. The biodiversity found on Earth today consists of many millions of distinct biological species, which is the product of nearly 3.5 billion years of evolution.

BIODIVERSITY DATABASES

The Biodiversity Databases are classified into two main categories: specimen-focused and species-focused. In between these two categories lie nomenclator databases. The distinction between these categories is sometimes somewhat arbitrary and some databases contain information in multiple category.

Specimen-focused databases contain data about individual specimens and include: catalogues of vouchered museum specimens (i.e. collections databases), collections of specimen photographs, databases of field-based specimen observations, morphological data or genetic data as well as meta-databases or consortia of databases. Example: ants of the world (http://www.antweb.org).

Nomenclators act as summaries of taxonomic revisions and are in effect a key between specimen-focused and species-focused databases. They do this because taxonomic revisions use specimen data to determine species limits. Example: Australian Plant Name Index (http://www.cpbr.gov.au/apni/index.html)

Species-focused databases – contain information summarised at the level of species. Some species-focused databases attempt to compile comprehensive data about particular species (e.g. FishBase, http://www.fishbase.org/home.htm), while others focus on particular species attributes, for example: checklists of species in a given area (e.g. FEOW), or the conservation status of species (e.g. CITES or IUCN Red List).

Other taxon-focused *databases* have genera or families as their basic unit. Example: all catfishes species inventory (http://silurus.acnatsci.org)

INDIAN BIODIVERSITY DATABASES

The Biological Diversity Act visualizes 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, municipalities and corporations. A wealth of information exists on India's biodiversity resources and associated knowledge. A good beginning has been made in organizing a part of this information in electronic databases.

Some of the key initiatives are Agricultural databases and information on sacred groves (MSSRF, 2003); Agricultural Research Information Network (ARISNET); Bibliographic and referral information on Western Ghats (CES, 2003); Biodiversity characterization using RS/GIS; Biotechnology Information System (BTISNet) (DBT, 2003); CDROMs on Marine Prawns, Marine Crabs, Mangroves, Lignicolous Fungi and corals of India (NIO, 2003); Endemic Trees of Western Ghats; Environmental Information System (ENVIS) (MoEF, 2003); Ethnobotany (NBRI); Flora of Karnataka; LIFKEY/LIFDAT (CES, 2003); Indian Medicinal Plants database (FRLHT, 2003); National and State Forest Vegetation maps and National Basic Forest Inventory (NBFIS) (FSI, 2003); National Register of Green Grassroots Innovations and Traditional Knowledge (NIF, 2002); National Wildlife Database and Zoo Database (WII, 2003); NCL Center for Biodiversity Informatics (NCL, 2003), Birds of India (SACON, 2003); People's Biodiversity Registers (CES, 2003); Plants of India and Legume Database of South Asia (NBRI, 2003); SAHYADRI: Western Ghats Biodiversity Information System (database of Western Ghats flora and Fauna (http://ces.iisc.ernet.in/biodiversity); Sasya Sahyadri (Ganeshaiah, 2003); and Traditional Knowledge Digital Library (NISCAIR, 2002).

LACUNAE

- Many institutions have information documented in the form of databases.
- 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.

REFERENCES

1. http://www.ncbi.nlm.nih.gov
2. http://www.embl-heidelberg.de/
3. http://www.expasy.ch/sprot/sprot-top.html
4. http://www.ddbj.nig.ac.jp