Int. J. Mendel, Vol. 30 (1-4), 13-15, 2013 **IN SILICO BIOLOGY** **ISSN 0970-9649**

BIOLOGICAL DATA MANAGEMENT IN MODERN RESEARCH

R. P. Upadhyaya and Rakesh Ranjan*

Key words : Biological Database, Management, Retrieval tools, Our contributions.

Biological databases are information pakages which have become an integral part of modern biological research. Computational biology is now an emerging field to store, manage and transfer biological information at an amazingly fast speed and with great precision. Some novel structural databases on some well researched biological topics have been developed and these are available for the users.

INTRODUCTION

The present century belongs to biology and the pace of progress in this area of human knowledge has simply been astonishing. The appetite towards biological research is more pronouncedly focused towards biomolecular dynamics of cells. The increasing interest and sophistications achieved in technology have resulted into accumulation of biological information with stunning speed and magnitude to an extent that the quanta of arising information are becoming difficult to analyse, interpret and manage (Jiawei Han, Micheline Kamber and Jian Pei, 2012).

As rightly been perceived by Alexis Leon and Mathews Leon (1999) information is nothing but refined data and according to Burch and Grudnitski (1989) information is data that has been put into a meaningful and useful context and communicated to the recipient who uses in decision making. In its truest sense information is a communication and reception of intelligence for knowledge. Biological information are born through the experience and observations as they apprise and notify, surprise and stimulate, reduce uncertainty and reveal alternative possibilities to perceive (Lucas, 1997). In this age of rapidly changing environment one has to manage the future and for this sake arriving information are to be managed properly. After all, this is the age of information explosion, where people are bombarded with data from which right information, with right time, in right amount are to be retrieved. Computer based data management of biological information is a need of the time. Biological data bases are shared collection of related information in such a way that a computer based programme can very quickly select the desired piece of information. It is basically an electronic filing system (Rob and Coronel, 2000; Elamasri and Navathe, 2000).

Development of biological database and their maintenance filled with retrieval programmes have witnessed an exponential growth during last few decades. This is chiefly because of practices of DNA and protein sequencing (Fig.-1). These databases accommodate adequate information about millions and millions of gene and protein sequences, structural details including three dimensional modules of proteins. In fact, biological database management has become an integral part of modem biological researchers more especially in the area of gene and protein related biology (Upadhyaya *et al.,* 2007). Growth in a number of biological databases is presented in Fig.-2.

Types of Biological Data Bases :

Nature of biological information is quite diverse and the approach and tools applied, therefore, differ. For instance, storage pattern of DNA sequences can be stored from 2D, 3D gel and image store. Data can be stored in "flat file format" or, in table, showing relational aspect, or in the form of object oriented databases. There are three major forms of biological databases (i) primary (ii) secondary and (iii) specialized.

Primary (PDB) : This is a collection of raw sequences data submitted in structural form. This can decipher sequence of nucleic acid eg., gene bank, EMBL, European Mol. Bio Lab. & DDBJ, IDMA data base of Japan and specification protein, such as PIR (Protein Informations Recourses).

Structure database: PDB (Protein Data Base) : This gives an idea of 3-D structural details of biomolecules such as nucleic acids and proteins. The information pertaining to this mode is obtained through X-ray crystallography, NMR spechoscopy and the information are freely available on coebsites of member organizations such as PDBe, PDBj & RCSB. Protein Data Bank (PDB) is maintained by an organisation called the world wide protein safe bank (WW PSB). The organisation is a key tool in the area of structural biology. The research labs submit their data to this bank. The primary data bank information are retrieved to derive and develop secondary data bases as per the need. Database SCOP & CATH deals with the structural details for the purpose of visualising and analysing phylogenetic relationship; likewise GO produces structure based on gene.

Secondary database (SDB) :

In primary database information on sequence annotation is not sufficient and just to turn these raw information into a more meaningful way through post processing exercise and for this sec. databases are developed.

Specialized database :

This is highly specific to particular organism. There is a provision where sequences may overlap with those retrived per PDB but will have new data submitted by the researchers. Examples - Plasmodb, Fly Base, Worm Base, FAIR.

University Deptt. of Botany, T.M. Bhagalpur University, Bhagalpur - 812 007
University Centre of Bioinformatics, T.M. Bhagalpur University, Bhagalpur - 812 007
email : ramprakash.upadhyay@rediffmail.com

Storage techniques of Biological Database :

Flat File :

This is a simple file storing information with no structural relationships; such a DB is easy to access and understand. Not much expertise is required in this regard. It is infact not a true data base by definitions and a collection of similar files not having a standard format. The model is to be issued towards formatting data for a flat file database, based on character level to a level that when retrieved data would appear as a printed page. These are utilized for ordering indexing on a computer file system, collected flat files may be ordered and stored in labelled folder like collected printed pages are ordered in a file drawer, they are made available by search index. It is something equivalent to books index where a word such as protein may be located in the concerned pages.

Because of simplicity of configuration and easy retrieval, these biological databases are popular and there is nothing wrong about this category. Data can be recognised in a meaningful way having proper indexing, these can easily be searched. As collection flat file gets larger and longer, working with them is a problem. It is not adequately placed to provide relational attribute.

Relational Database :

This is well devised and of better category. In this specific class of database information are stored in a collection of tables, thus making it more inclusive and purposeful. If someone goes for protein search, the relational based data would provide sets of information in separate tables revealing secondary structure elements, atomic positions, experimental conditions and soon each table has a label with protein identity ship of protein. Selecting a protein, for example, one can come across experimental conditions above position and further details scattered in various tables by just a click.

Object Oriented Database :

This is yet another version of biological databases which is associated with object oriented programming. This DBMS facilitates simultaneous interactions by multiple clients. One can handle complex objects beyond the limits of table of character data owing to its flexibility utility; major portion of the biological databases are object oriented DBMSs and have emerged highly useful and popular. The system handles information as objects instead of simple files and tables. These are more complete information packages providing access to practically everything from simple text format data to image and video files. For search they do not need SQL quarry language.

National Centre of Biotechnology Informatics (NCBI)

This organization was founded in 1988 and is focussed towards developing information system on Molecular Biology. NCBI maintains Gene Bank where information pertaining to gene sequence is deposited from the scientists and research organisations. Stored data can be procured using retrieval systems, information are available at NCBI home page at http://www.ncbi.nlm.nih.gov and its link (ftp.ncpi.nih.gov).

Database Retrieval system:

Following are the vital tools applied for data base retrieval.

(1) Entrez : It helps data search using simple Boolean queries comprising 30 databases, having an account of over 70 million DNA and protein sequences from phylogeny, genome and proteome research results. Some important database are Uni STS, Uni Gene, Molecular modelling databases (MMDB), on-line Mendelian inheritance in Man (OMIM) and On Line Books and Journals.

Entrez provides extensive links within and between database and the records retrieved in Entrez can be displayed in many formats and downloaded singly or in batches.

(2) PubMed Central

PubMed Central (PMC) (7) is a digital archive of peer reviewed journals in the life sciences providing access to 400,000 full-text articles, an increase of 100,000 from the past year. More than 200 journals, including Nucleic Acids Research, deposit the full text of their articles in PMC. Participation in PMC requires a commitment to free access to full text, either immediately after publication or within a 12-month period. All PMC free articles are identified in PubMed search results and PMC itself can be searched using Entrez.

(3) Taxonomy

The NCBI taxonomy database, growing at the rate of 3000 new taxa a month, indexes 205,000 named organisms that are represented in the databases with at least 1 nt or protein sequence. The Taxonomy Browser can be used to view the taxonomic position or retrieve data from any of the principal Entrez databases for a particular organism or group. The Taxonomy Browser also displays links to the Map Viewer, Genomic BLAST services, the Trace Archive, and to external model organism and taxonomic databases via Link Out. Searches of the NCBI taxonomy may be made on the basis of whole, partial or phonetically spelled organism names. Entrez Taxonomy displays include custom taxonomic trees representing user-defined subsets of the full NCBI taxonomy.

THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The Basic Local Alignment Search Tool (BLAST) programs perform sequence similarity searches against a variety of sequence databases, returning a set of gapped alignments with links to full database records, to UniGene, Gene, the MMDB or GEO. One variant, BLAST 2 Sequences, compares two DNA or protein sequences and produces a dot-plot representation of the alignments. Each alignment returned by BLAST is scored and assigned a measure of statistical significance, called the Expectation Value (E-value). BLAST takes into account the amino acid composition of the query sequence in its estimation of statistical significance. This composition-based statistical treatment, used in conventional protein BLAST searches as well as PSI-BLAST searches, tends to reduce the number of false- positive database hits. The alignments returned can be limited by an

E-value threshold or range. Standard output formats include the default pairwise alignment, several query-anchored multiple sequence alignment formats, an easily-parsable Hit Table and a taxonomically organized output. Database sequences appearing in BLAST results may be marked for batch retrieval using check boxes. A new, enhanced, formatter displays alignments against data-base sequences that are > 200,000 bp in length with links to nearby features, such as genes. A new 'Pairwise with identities' mode better highlights differences between the query and a target sequence.

Our Efforts :

Our research team has been successful in developing following structural data bases at the University Centre of Bioinformatics, T.M. Bhagalpur University, Bhagalpur.

1. Data Base on 'Pahariya' tribe :

Pahariya tribe is an ethnic community inhabiting selected few regions of the districts of Godda, Bhagalpur and Dumka. They live on hill tops, come down only in case of meeting basic needs and in emergency. They constitute a highly isolated ancient race whose population is dwindling. Data on population index, physical profile, health conditions, disease incidence, food habit economy, etc. have been taken into account.

2. Biological Data base on 'Katarni' Rice variety:

Basmati Rice variety 'Katarni' is an exclusive variety of Bhagalpur region. This variety is liked for its exceptionally pleasing aroma. The cultivation of this variety is restricted only to 10-12 blocks of Bhagalpur district. Due to low yield and high disease incidence the cultivators are losing interest. There is danger that this novel variety cherished for its flavour and taste is out of cultivation and human selection. Important information pertaining to the climate and soil profile of the area, cultivation pattern, physical and biological characters of this variety has been taken into account. Attempts are in the pipeline to sequence the genome of this variety to track down genes responsible for the rare aroma of this rice variety.

3. Database on Medicinal plants

Another novel biological database developed through our *in-silico* lab. is on the medicinal plants of Bhagalpur and Santhal Pargana area. In all, 92 plant species have been identified. Full details on the taxonomy, plant parts used in herbal medicines, active principles along with ethnobotanical perspective, are available.

ACKNOWLEDGEMENT

The authors are indebted to DBT for financial support.

References

Burch, J., and Grudnitsg. 1989 information system : Theory and Practices (5th Ed.) John Wiley & Sons.

Elumasri, R. and Navathe International Thompson hearing S. 2000 : Fundamentals of Database Systems (3rd Ed.) Pearson Education.

Jaiwei,H., Kamber, M. and Pei, J., 2012 : Data Mining Concepts and Techniques, Elsevier, Sydney, Tokyo.

Leon, A and Leon, M., 1999 : Database Management System, Leon Vikas, Chennai.

Lucas, H.C., 1997 : Information Theory of Management (6th Ed.) McGraw Hill Companies, Inc.

Rob, P. and Coronel, C, 2000. Database Systems : Design, Implementation and Management.

Upadhyaya, R.P., Singh, V. and Ranjan, R. 2007 : Emerging Fields in Bioinformatics, University Centre of Bioinformatics.