OK INBRE Bioinformatics

October 2005

University of Oklahoma Health Sciences Center

Introduction to Biological Databases

by Edgar Scott, II, M.S.

One major desire of scientists is to understand in detail the relationship between the genome and the metabolic processes of an organism. Couple this desire with the advancing efficiency of high-throughput technology, and you have a scenario that produces a plethora of biological data that needs to be managed, organized, characterized and interpreted to extract meaningful information. To meet this need, public biological databases have arisen as vital resources to the scientific community. Because of their rising importance to science, it is important to be aware of the different databases that are available. This bulletin will introduce you different types of databases that exist and examples of some of the most popular databases.

A biological database is an electronic collection of organized biological information and can contain thousands to millions of different database records. A database record, like a file in a filing cabinet, is a single entry in the database that holds biological data (i.e. protein sequence information, coordinates of three-dimensional structures of proteins, etc.), along with related annotations. An annotation is supplementary information that describes the biological relevance of the data in the record, such as identifying the function of the gene or protein, identifying the functional domains that the resulting protein may contain, literature citations, related proteins, etc. The type of data that is stored in different databases varies greatly: nucleic or amino acid sequence data, three-dimensional coordinates of crystal structures, multiple sequence alignments of conserved protein domains, literature citations, taxonomic information, and more.

The types of sequence databases that are available can be separated into two different categories: primary sequence databases and secondary sequence databases. Primary sequence databases store records of sequence information and are archival in nature. The data is derived from sequencing a biological molecule that exists in a test tube somewhere in a lab. Sequences in primary sequence databases have not been computationally processed, nor are they a resulting consensus sequence of a population, and therefore often contain sequencing errors. Because these databases are archival, records that have been modified or edited can be traced back to their original version, which ultimately can lead to a massive collection of information to mine.

An example of a primary nucleic acid sequence database is GenBank[1] which is a member of the International Nucleic Sequence Database Collaboration (INSDC). GenBank, which is available through NCBI's website, contains an annotated collection of nucleic acid sequences as well as their protein translations. This database is populated by accepting nucleic acid sequence submissions from sequencing facilities or individuals across the world and sharing nucleic acid and amino acid sequence information with databases within INSDC and other various databases. The quality of the annotations contained in GenBank can vary depending on how well the original sequencing facility annotated the sequences they submitted to GenBank, and also on the annotations from other databases. Therefore, one must be careful when retrieving sequences from this database.

A very popular primary protein sequence database is the UniProt Knowledgebase[3] of the UniProt Consortium, which is composed of three previously independent protein sequence databases. The goal of this consortium is to "maintain a high quality protein sequence database that is comprehensive, stable, fully classified, richly and accurately annotated." UniProt consists of two types of databases: Swiss-Prot, which is a manually annotated database, and TrEMBL, which is a computationally analyzed database whose records await full manual annotation. Because of the precision and care taken by the curators of the UniProt Knowledgebase, particularly Swiss-Prot, this database is considered to be one of the most reliable resources for accurate and detailed protein sequence information.

Secondary sequence databases also store records of sequence information; however these records have been curated either computationally or manually. These databases contain mostly records that were originally stored in primary databases. After manual or computational curation these records contain more detailed and accurate annotations for the sequence in question, which help scientists better understand the biological role. These sequences will often be the best representation of that gene or protein from a population of sequences. Secondary databases are often the resource that most scientists would prefer to use because of the information stored within those files is more accurate and detailed.

An example of a secondary database is NCBI's Reference Sequence (RefSeq)[2] database. Because of the redundancy and inaccuracies of the data in GenBank, NCBI created RefSeq with the goal of creating a non-redundant, comprehensive, curated database of naturally occurring DNA, RNA, and protein sequences. All sequences are based on GenBank's sequence records. RefSeq provides an accession number for a gene that corresponds to the most stable, agreed-upon versions of the sequence, which are assigned by the NCBI staff. The staff also adds additional information such as related expressed sequence tags (ESTs), literature citations, links to conserved domains, related proteins, related structures and more.

The collection of protein sequences in databases like the ones previously mentioned provide a source for creating and maintaining protein signature databases that contain information

The University of Oklahoma Health Sciences Center

Laboratory for Genomics and Bioinformatics

Edgar Scott II, M.S.
Multicampus Bioinformatics Education Specialist

Phone:
405-271-2133 Ext. 32511
E-Mail:
edgar-scott@ouhsc.edu

We're on the Web!

See us at:

microgen.ouhsc.edu/inbre/

Websites

1. www.ncbi.nih.gov/Genbank/index.html
2. www.ncbi.nlm.nih.gov/RefSeq/
3. www.ebi.uniprot.org/index.shtml
4. us.expasy.org/prosite/
5. smart.embl-heidelberg.de/
6. www.sanger.ac.uk/Software/Pfam/
7. www.rcsb.org/pdb/
8. www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Structure
9. scop.mrc-lmb.cam.ac.uk/scop/
10. www.genome.jp/kegg/kegg2.html
11. bind.ca/
12. dip.doe-mbi.ucla.edu/

Bulletin (continued)

based on protein families and domains. Because some regions of the protein are important for its structure and function, they will tend to be more conserved within a protein family and can be represented as a pattern or a profile. A pattern is a qualitative description of the amino acids that make up the motif in question. A profile, however, is a quantitative description of the likelihood of different amino acids appearing within specific positions of the motif. For example, the EF Hand found in Calmodulin (as well as in Troponin C and other proteins) binds to calcium and has a conserved amino acid sequence that can be represented as a pattern or a profile.

There are several different signature (or secondary) protein databases that are maintained. PROSITE[4] consists of a large collection of biologically meaningful signatures described as patterns or profiles. The Single Modular Architecture Research Tool (SMART)[5] database contains extensively annotated protein domains with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. The Protein Family (Pfam)[6] database consists of multiple sequence alignments and hidden Markov models covering many common protein domains. Major strides have also been made in obtaining the three-dimensional structure of proteins. The Research Collaboratory for Structural Bioinformatics (RCSB) Protein Databank (PDB)[7] is the main public repository for the crystal structures of proteins solved by X-ray crystallography and NMR spectroscopy and can be considered a primary "structure" database because it is archival like GenBank. An example of a curated secondary "structure" database is the Molecular Modeling Database (MMDB)[8] at NCBI, which performs simple computations to ensure that there is an agreement between the atomic coordinates and the primary sequence. A second example of a secondary "structure" database is the Structural Classification of Proteins (SCOP)[9] database, which provides a comprehensive description of protein structures and evolutionary relationships based upon a hierarchical classification scheme.

In addition, there are databases that store biological pathway information. One such example is the Kyoto Encyclopedia of Genes and Genomes (KEGGs)[10] database, which is a nicely designed web interface for annotated biological pathways. The Biomolecular Interaction Network Database (BIND)[11] and the Database of Interacting Proteins (DIP)[12] are resources for information on the binding interactions between different proteins within a specific biological pathway.

With databases such as GenBank, RefSeq, UniProt, KEGGs, and others, scientists are better equipped to manage, gain access to and analyze biological data. With more genomes being sequenced and better bioinformatics tools being developed, a major issue will be keeping these databases up-to-date and accurate. Databases are also steadily changing as new ideas arise to make them more efficient, more streamline, search faster, etc. For more information on databases, you can access the websites below.

Related bioinformatics sites of interest...

- www.ebi.ac.uk/2can/databases/index.html
- www.geocities.com/bioinformaticsweb/datalink.html
- biotech.icmb.utexas.edu/pages/science/nucacid_db_intro.html
- biotech.icmb.utexas.edu/pages/science/protein.html

About Us...

The OUHSC Laboratory for Genomics and Bioinformatics is a full-service genomics facility offering DNA sequencing (small- and large-scale projects), microarray design and hybridization and other services, including bioinformatics support. Edgar Scott is the INBRE Multicampus Bioinformatics Education Specialist, responsible for fostering the development of bioinformatics education on 14 undergraduate campuses in the state of Oklahoma, and coordinating INBRE-related bioinformatics activities with the INBRE Bioinformatics Core. The University of Oklahoma is an equal opportunity institution. This publication is printed and issued by the University of Oklahoma; the cost of $150 was paid by OK INBRE.