OK INBRE Bioinformatics

December 2005

University of Oklahoma Health Sciences Center

Introduction to NCBI Databases

by Edgar Scott, II, M.S.

This month's bulletin will briefly cover database searching on the National Center for Biotechnology Information (NCBI) web site (1). NCBI is dedicated to developing "new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease" (www.ncbi.nlm.nih.gov/About/glance/ourmission.html). This website is one of the major resources for a large number of databases that contain different types of biological information, ranging from nucleic acid and amino acid sequence information to taxonomic information and three-dimensional structures of proteins and conserved domains.

Biological databases arose out of a dire need to better organize, store and search tons of biological data that has been and still is being generated at a fast pace as a result of the advancement of high-throughput technology. A biological database is an electronic collection of organized biological information. The way in which biological information is organized is solely dependent on the database that is being searched, and each database can contain thousands to millions of different database records. A database record is a single entry in the database (like a file in a file cabinet) that holds some type of biological data (i.e., protein sequence information, coordinates of three-dimensional structures of a protein, etc.) along with any known annotations. An annotation is supplementary information that describes the biological relevance of the biological data in the data record such as identifying the function of the gene or protein, identifying the functional domains that the resulting protein may contain, etc. The type of data that is stored in different databases varies greatly: literature citations, nucleic or amino acid sequence data, taxonomic information, three-dimensional structures of proteins, protein domain and protein family information, gene expression information, genome specific information, expressed sequence tag (EST) information and much more.

NCBI, as mentioned earlier, has several types of databases: literature databases for literature information; nucleic databases for nucleic acid sequence information; protein databases for protein sequence information; structure databases for three-dimensional structure of proteins and domains; taxonomy databases for taxonomic information; genome databases for information for specific genomes; expression databases for gene expression data; chemical databases for small molecules. During this bulletin, I'll describe the GenBank database, RefSeq (Reference Sequence) database and the Entrez global query system, along with a few other individual Entrez query divisions: Entrez Protein, Entrez Nucleotide, Entrez Gene, PubMed, PubMed Central, and Online Mendelian Inheritance in Man (OMIM).

GenBank is the National Institutes of Health's (NIH) publicly accessible nucleotide and protein sequence database. It consists of most known public DNA and protein sequences. Sequence information in GenBank originates from the submissions of individual laboratories and large-scale sequencing centers. This database is part of the International Nucleotide Sequence Database Collaboration, where three databases (GenBank database, European Molecular Laboratory, EMBL, database, and DNA Databank of Japan, DDBJ) exchange data daily so that all three databases contain the same set of sequences. GenBank also collects protein sequence information from such online protein sequence databases as Swiss-Prot, Protein Information Resource (PIR), Protein Research Foundation (PRF) and a protein structure database called Protein Data Bank (PDB).

One major issue to be aware of is that GenBank contains a lot of redundant sequence information since there are so many different sources of information. For example, there can be more than one database record for a given gene. The Reference Sequence (RefSeq) database attempts to solve this problem by providing a biologically non-redundant collection of DNA, RNA and protein sequences. It provides one accession number (a unique ID for a database record like a social security number for a U.S. citizen) for a DNA, RNA or protein sequence that corresponds to the most stable, agreed upon version of the sequence. RefSeq database records are frequently based on GenBank database records but differ in that each RefSeq is a synthesis of information, not a piece of a primary research data in itself.

NCBI has created a query system called Entrez that offers an easy way to search its databases. Entrez is a text-based query system that integrates database data from a large number of sources, formats and databases (not just GenBank and RefSeq) into a uniform information model and retrieval system. The tightly interlinked system provides a search across all the databases maintained at NCBI. For example, if you were to go to the Entrez Global Query System (2) and perform a text search, this page would return a count for all the entries that were found in each individual database maintained at NCBI. The results for a specific database can be directly accessed by clicking that database's link.

Instead of searching through all of the databases simultaneously, you can search for one type of biological data (i.e., a nucleic acid sequence) with a specialized search system. Entrez Nucleotide is a division of the Entrez Query System that gives the public access to the nucleic acid sequence information contained in both GenBank and RefSeq via text searching. This search tool can be accessed directly from the NCBI home page

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/

Web sites

1. www.ncbi.nlm.nih.gob/
2. www.ncbi.nlm.nih.gov/gquery/gquery.fcgi

Bulletin (continued)

by selecting "nucleotide" in the database drop box. Similarly, Entrez Protein will provide the public access to both GenBank and RefSeq, but only to the protein sequence information that is contained within them.

A common difficulty that is encountered in database searches is that too much information is returned. This problem can be addressed by learning how to generate specific searches with appropriate limits. When using the different Entrez systems, you can set limits on your searches to make them more specific and to receive fewer results. For example, searching for "retinol binding protein" in Entrez Protein will return 357 results (as of March 5th, 2005). But by setting limits to search through only the "Protein Name" field, only 27 results will be returned (as of March 5th 2005).

If an investigator wanted more detailed annotated information about a specific gene, he or she could retrieve this information from Entrez Gene, which is a gene-centered resource at NCBI. It is solely responsible for providing a unique GeneID that is used to identify information for genes. The content of Entrez Gene represents the result of the curation and automated integration of data from NCBI's RefSeq, collaborating model organism databases (these could be databases at other web sites) and from many other databases available from NCBI. This query system can also be accessed from the NCBI home page by selecting "Gene" from the database drop box.

Other databases that should be mentioned are PubMed, PubMed Central (PMC) and OMIM. PubMed is a literature search service of the National Library of Medicine that gives the public access to millions of citations for biomedical articles from MEDLINE and additional life science journals. PubMed provides links to articles that are either free or require a subscription. PMC is similar to PubMed in that it is a literature search service, but all of the citations here are free! Lastly, OMIM database is a catalog of human genes and genetic disorders. It contains textual information and references to literature concerning genes and the diseases in which they play a role. All three of these databases are searched using a text-based query system and they each have their own unique set of limits that users can choose to make their search more specific.

The work of the NCBI staff to integrate sequence information between databases makes searching for information much easier. One could identify the mRNA and protein sequence for a given gene from Entrez Gene, explore the phenotypes of the diseases associated with mutations in this gene from OMIM, identify homologous sequences from a Homolog Database and identify proteins with similar 3-D structure, all by utilizing the links provided by the NCBI staff.

Related bioinformatics papers of interest…

- Benson, D.A.; Karsch-Mizrachi, I.; Lipman D.J.; Ostell, J.; Wheeler D.L. *GenBank.* Nucleic Acids Res. 2005 Jan 1; 33(Database Issue): D34-D38.
- Pruitt, K.D.; Tatusova, T.; Maglott, D.R. *NCBI Reference Sequence (RefSeq): A Curated Non-redundant Sequence Database of Genomes, Transcripts and Proteins.* Nucleic Acids Res. 2005 Jan 1; 33(Database Issue): D501-D504
- Maglott, D.; Ostell, J.; Pruitt, K.D.; Tatusova, T. *Entrez Gene: Gene-centered Information at NCBI.* Nucleic Acids Res. 2005 Jan 1; 33(Database Issue): D54-D58
- Hamosh, A.; Scott, A.F.; Amberger, J.S.; Bocchini, C.A.; McKusick, V.A. *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of Human Genes and Genetic Disorders.* Nucleic Acids Res. 2005 Jan 1; 33(Database Issue): D514-D517

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.