Chapter 2

Genomic Databases and Resources at the National Center for Biotechnology Information

Tatiana Tatusova

Abstract

The National Center for Biotechnology Information (NCBI), as a primary public repository of genomic sequence data, collects and maintains enormous amounts of heterogeneous data. Data for genomes, genes, gene expressions, gene variation, gene families, proteins, and protein domains are integrated with the analytical, search, and retrieval resources through the NCBI Web site. Entrez, a text-based search and retrieval system, provides a fast and easy way to navigate across diverse biological databases.

Customized genomic BLAST enables sequence similarity searches against a special collection of organism-specific sequence data and viewing the resulting alignments within a genomic context using NCBI's genome browser, Map Viewer.

Comparative genome analysis tools lead to further understanding of evolutionary processes, quickening the pace of discovery.

Key words: bioinformatics, genome, metagenome, database, data management system, sequence analysis.

1. Introduction

Recent advances in biotechnology and bioinformatics led to a flood of genomic data and tremendous growth in the number of associated databases. As of February 2008, NCBI Genome Project collection describes more than 2,000 genome sequencing projects: 1,500 Bacteria and Archaea (631 complete genomes, 462 draft assemblies, and 507 in progress) as listed at the NCBI Genome Project site: http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi and almost 500 eukaryotic genomes (23 complete, 195 draft assemblies, and 221 in progress) as listed at http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi.

O. Carugo, F. Eisenhaber (eds.), *Data Mining Techniques for the Life Sciences*, Methods in Molecular Biology 609,
DOI 10.1007/978-1-60327-241-4_2,

Information on complete and ongoing genome projects is also available in Genomes OnLine Database (GOLD) (1), a community-supported World Wide Web resource. Hundreds of thousands of genomic sequences for viruses, organelles, and plasmids are available in the three public databases of the International Nucleotide Sequence Database Collaboration [INSDC, www.insdc.org] – EMBL (2), GenBank (3), and the DNA Data Bank of Japan (4). Additional information on biomedical data is stored in an increasing number of various databases. As published in the 15th annual edition of the journal *Nucleic Acid Research* (NAR), also known as Database Issue, the number of databases in 2008 crossed the 1,000 landmark. This issue listed 1,078 databases, 110 more than in the previous year (5). Navigating through the large number of genomic and other related "omic" resources becomes a great challenge to the average researcher. Understanding the basics of data management systems developed for the maintenance, search, and retrieval of the large volume of genomic sequences will provide necessary assistance in traveling through the information space.

This chapter is focused on the infrastructure developed by the National Center for Biotechnology Information over the last 20 years. NCBI, as a primary public repository of genomic sequence data, collects and maintains enormous amounts of heterogeneous data. The databases vary in size, data types, design, and implementation. They cover most of the genomic biology data types including the project description, project sequence data (genomic, transcript, protein sequences), raw sequence reads, and related bibliographical data (6). More recently, NCBI started to collect the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and nonclinical traits (7). All these databases are integrated in a single Entrez system and use a common engine for data search and retrieval. This provides researchers with a common interface and simplifies navigation through the large information space.

There are many different ways of accessing genomic data at NCBI. Depending on the focus and the goal of the research project or the level of interest, the user would select a particular route for accessing the genomic databases and resources. These are (1) text searches, (2) direct genome browsing, and (3) searches by sequence similarity. All of these search types enable navigation through precomputed links to other NCBI resources.

This chapter describes the details of text searching and the retrieval system of three major genomic databases, Entrez Genome and Entrez Genome Project and Entrez Protein Clusters, and also illustrates two other methods of accessing the genomic data.

2. Data Flow and Processing

The National Center for Biotechnology Information was established on November 4, 1988, as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH) in order to develop computerized processing methods for biomedical research data. As a national resource for molecular biology information, NCBI's mission is to develop automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitate the use of such databases and software by the research and medical community; coordinate efforts to gather biotechnology information both nationally and internationally; and perform research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules.

The fundamental sequence data resources at NCBI consist of both primary databases and derived or curated databases. Primary databases such as GenBank (3) archive the original submissions that come from large sequencing centers or individual experimentalists. The database staff organizes the data but do not add additional information. Curated databases such as Reference Sequence Collection (8) provide a curated/expert view by compilation and correction of the data. Records in the primary database are analogous to research articles in a journal, and curated databases to review articles. This difference is not always well understood by the users of NCBI sequence data. In response to the users' inquiries, and more specifically to a request from attendees at a 2006 workshop on microbial genomes held at NCBI, the differences between GenBank, RefSeq, and TPA databases have been recently described in the May 2007 issue of the American Society for Microbiology's journal *Microbe* (http://www.asm.org/microbe/index.asp?bid=50523).

In the same way as a review article can present an expert view or provide a result of computational analysis, the databases can be manually curated and/or computationally derived (**Table 2.1**). For more detailed information on all NCBI and database resources see also (6).

The biological sequence information that builds the foundation of NCBI primary databases and curated resources comes from many sources (**Fig. 2.1**).

This section discusses the flow of sequence data, from the management of data submission to the generation of publicly available data products. An information management system that consists of two major components, the ID database and the IQ database, underlies the submission, storage, and access of

Table 2.1
Primary and derived databases at NCBI

Database type	Database name	Database description
Primary databases	GenBank/EMBL/DDBJ (core nucleotide)	Author submissions of nucleotide (genomic and cDNA) sequence with conceptual translations as appropriate
Primary	GEO	Gene expression experimental data sets
Primary	dbGSS	Genome Survey Sequences
Primary	dbEST	Expressed Sequence Tags
Primary	dbMHC	DNA and clinical data related to the human major histocompatibility complex
Primary	dbRBC	A resource dedicated to the genetics of red blood cell antigens
Primary	dbSNP	Single nucleotide polymorphism
Primary	dbSTS	Sequence tagged sites
Primary	ProbeDB	Registry of nucleic acid reagents
Primary	Trace Archive	Raw trace data from sequencers
Primary	SRA	Short Read Archive
Primary	GenSAT	Gene expression atlas of mouse central nervous system
Primary	CancerChromosomes	Molecular cytogenetic data in cancer
Primary	dbGAP	Phenotype and genometype database
Primary	ProjectDB	
Derived	RefSeq	Curated representative sequence for major molecules of the central dogma
Derived	Genome	Complete and near-complete genomes, chromosomes, and plasmids
Derived	Gene	Gene-centered information from curated RefSeq transcripts, genome annotation
Derived	Homologene	Clusters of related genes from eukaryotic genomes
Derived	Protein Clusters	A collection of related protein sequences
Derived	Protein Neighbors	Database of precalculated protein BLAST hits
Derived	CDD	Conserved protein domains database
Derived	UniGene	Gene-oriented clusters of transcript sequences
Derived	UniSTS	Markers and mapping data

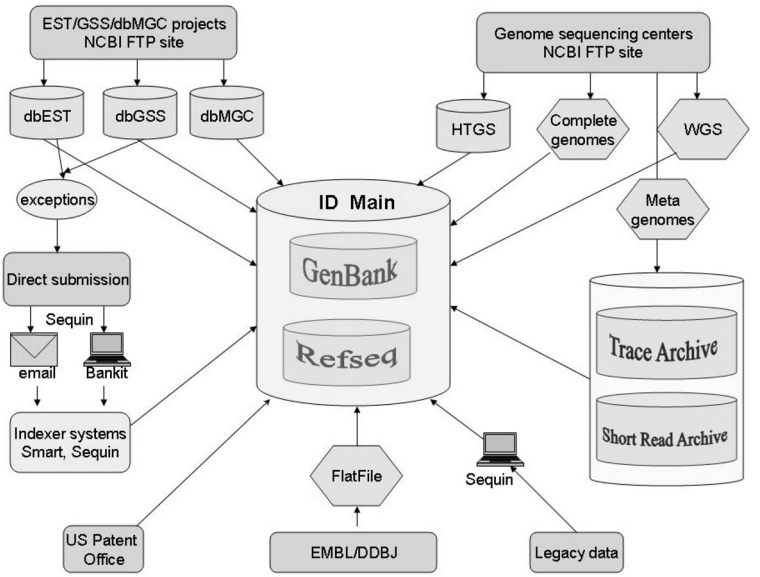

Fig. 2.1. Sources of primary sequence data available at NCBI. Rectangles represent data providers; cylinders represent primary NCBI databases.

GenBank (3), BLAST (9), and other curated data resources [such as the Reference Sequences (8) or Entrez Gene (10)]. Whereas ID handles incoming sequences and feeds other databases with subsets to suit different needs, IQ holds links between sequences stored in ID and between these sequences and other resources.

The data in ID system are stored in Abstract Syntax Notation (ASN.1) format, a standard descriptive language for describing structured information. NCBI has adopted ASN.1 language to describe the biological sequence and all related information (taxonomical, bibliographical) in a structured way. Many NCBI users think of the GenBank flatfile as the archetypal sequence data format. However, within NCBI and especially within the ID internal data flow system, ASN.1 is considered the original format from which reports such as the GenBank flatfile can be generated. As an object-oriented structured language, ASN.1 is easily transformed to other high-level programming languages such as XML, C, and C++. The NCBI Toolkit provides the converters between the data structures. Entrez display options allow to view the data in various text formats including ASN.1, XML, and GenBank flatfiles.

The ID database is a group of standard relational databases that holds both ASN.1 objects and sequence identifier-related information. In the ID database, blobs (binary large objects) are added into a single column of a relational database and are stored and processed as a unit.

Although the columns behave as in a relational database, the information that makes each blob, such as biological features, raw sequence data, and author information, is neither parsed nor split out. In this sense, the ID database can be considered as a hybrid database that stores complex objects.

The IQ database is a Sybase data-warehousing product that preserves its SQL language interface, but which inverts its data by storing them by column, not by row. Its strength is in its ability to increase speed of searches based on anticipated indexing. This nonrelational database holds links between many different objects.

3. Text Search and Retrieval System: Entrez

3.1. Organizing Principles

Entrez is the text-based search and retrieval system used at NCBI for all of the major databases, and it provides an organizing principle for biomedical information. Entrez integrates data from a large number of sources, formats, and databases into a uniform information model and retrieval system. The actual databases from which records are retrieved and on which the Entrez indexes are based have different designs, based on the type of data, and reside on different machines. These will be referred to as the "source databases." A common theme in the implementation of Entrez is that some functions are unique to each source database, whereas others are common to all Entrez databases.

An Entrez "node" is a collection of data that is grouped and indexed together. Some of the common routines and formats for every Entrez node include the term lists and posting files (i.e., the retrieval engine) used for Boolean queries, the links within and between nodes, and the summary format used for listing search results in which each record is called a DocSum. Generally, an Entrez query is a Boolean expression that is evaluated by the common Entrez engine and yields a list of unique ID numbers (UIDs), which identify records in an Entrez node. Given one or more UIDs, Entrez can retrieve the DocSum(s) very quickly.

3.1.1. Query Examples

Each Entrez database ("node") can be searched independently by selecting the database from the main Entrez Web page (http://www.ncbi.nlm.nih.gov/sites/gquery) (*see* **Fig. 2.2**). Typing a query into a text box provided at the top of the Web page and clicking the "Go" button will return a list of DocSum records that match the query in each Entrez category. These include nucleotides, proteins, genomes, publications (PubMed), taxonomy, and many other databases. The numbers of results returned in each category are provided on a single summary page and provide the

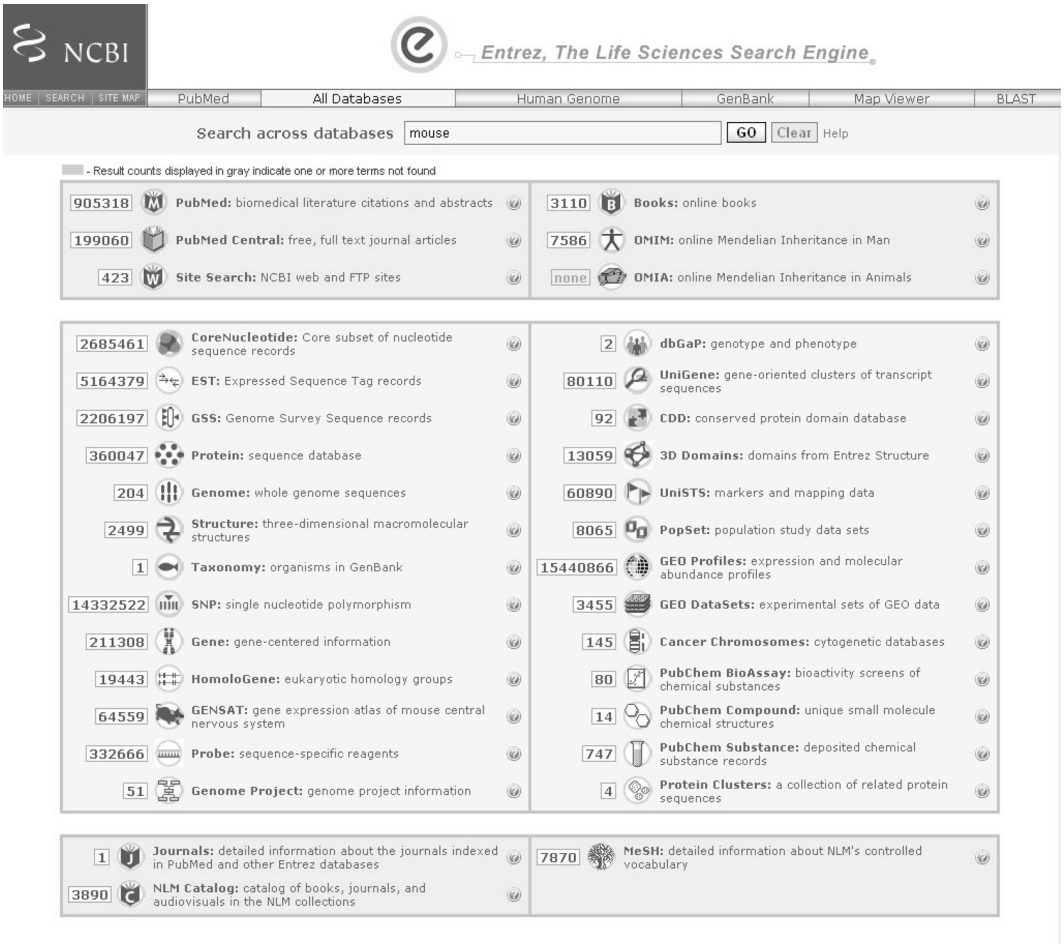

Fig. 2.2. Cross database search Web Entrez interface. The counts next to the database description show the number of the records in each database matching the simple text query "mouse."

user with an easily visible view of the results in each of ~35 databases. The results are presented differently in each database but within the same framework, which includes the common elements such as search bar, display options, page formatting, and links.

In processing a query, Entrez parses the query string into a series of tokens separated by spaces and Boolean operators (AND, NOT, OR). An independent search is performed for each term, and the results are then combined according to the Boolean operators.

Query uses the following syntax: term [field] OPERATOR term [field] where "term" refers to the search terms, "field" to the search field defined by specific Entrez database, and "OPERATOR" to the Boolean operators.

More sophisticated searches can be performed by constructing complex search strategies using Boolean operators and the various functions listed below, provided in the Feature Bar:

"Limits" restricts search terms to a specific search field.

"Preview/Index" allows users to view and select terms from search field indexes and to preview the number of search results before displaying citations.

"History" holds previous search strategies and results. The results can be combined to make new searches.

"Clipboard" allows users to save or view selected citations from one search or several searches.

"Details" displays the search strategy as it was translated by query engine, including error messages.

More information about Entrez system can be found from NCBI online Help Manual at http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp.

The main goals of the information system are reliable data storage and maintenance, and efficient access to the information. The retrieval is considered reliable if the same information that was deposited can be successfully retrieved. The Entrez system goes beyond that by providing the links between the nodes and pre-computing links within the nodes. The links made within or between Entrez nodes from one or more UIDs (Unique IDentifier) are also a function across all Entrez source databases. There are three different linking mechanisms described below.

3.1.2. Links Between the Nodes

The power of Entrez organization lies in the connections between the individual nodes that increase the information space. These links, created during indexing, are reciprocal and stored in a special database, for example, links between genome sequence records and the corresponding genome project. Links can also be provided by the original submitters, for example, links between a nucleotide sequence and a publication (PMID). Links between nucleotide and protein sequences (conceptual translation) of the annotated coding region can also be provided by the original submitters. **Figure 2.3** shows the diagram of the Entrez databases and the connections between them.

3.1.3. Links Within the Nodes

Entrez data can be also integrated by calculating the relationships between the records in a single database. For example, nucleotide and protein sequences can be linked by sequence similarity. The similarity is calculated using BLAST (9), stored in a special database, and made readily available in Entrez via the "Related Sequences" link. In PubMed, the inter-database links are calculated by comparing the frequency terms of the document. The similarity between two documents is based on the number of the

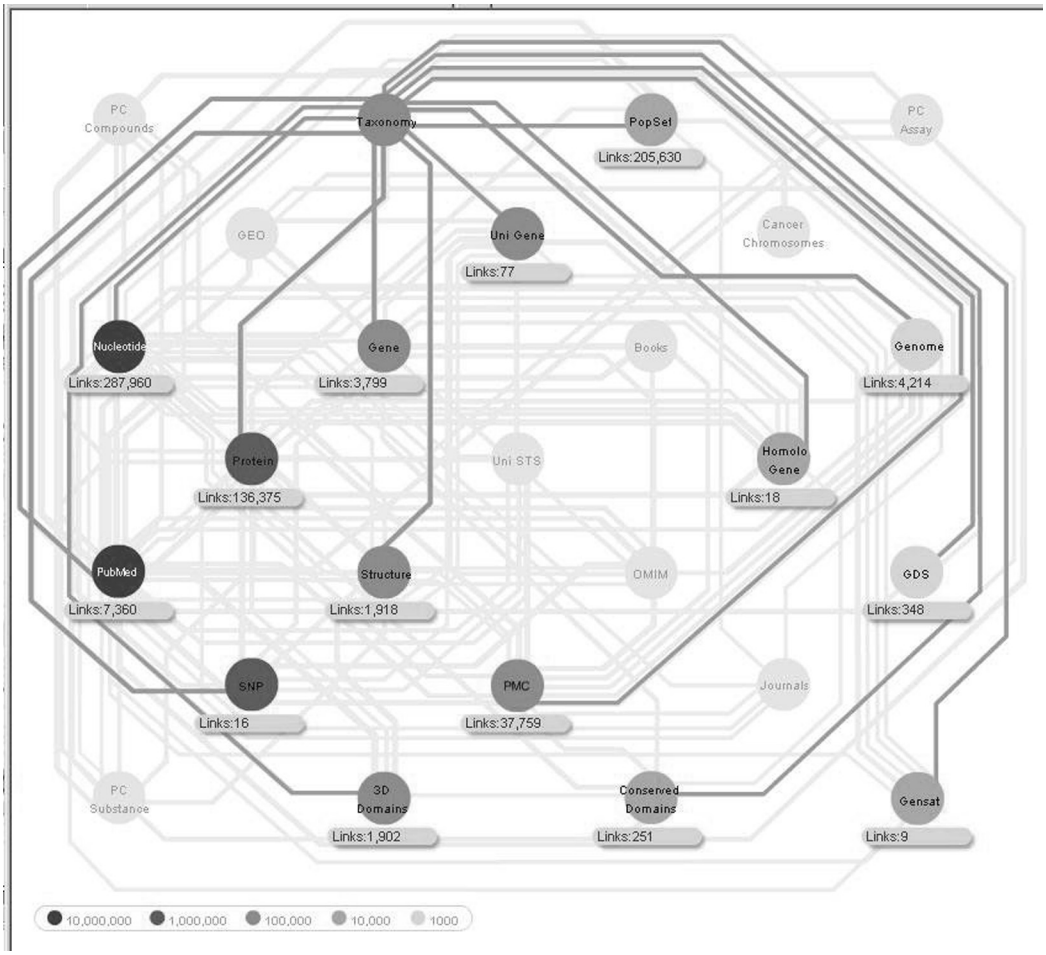

Fig. 2.3. The diagram of the Entrez databases and the connections between them. Each database is represented by a colored circle, where the color indicates the approximate number of records in the database.

weighted terms the two documents have in common. The highest scoring documents can be viewed for each document by selecting Related Articles from the Links menu.

3.1.4. Links Outside the Nodes

Links to outside resources are available through LinkOut, a special service of the Entrez system. It allows relevant outside online resources to link directly to the records in Entrez system. The outside users provide a URL, a resource name, the UID of the record they wish to link to, and a brief description of their Web site in a simple XML format. The request is processed automatically and links are added to the corresponding records in Entrez. This resource gives the end user a central place to look for the information available at NCBI and easily explore the relevant resources.

3.2. Tools for Advanced Users

The Entrez Programming Utilities (eUtils) are a set of eight server-side programs that provide a stable interface to the Entrez query and database system. The eUtils use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve data and represent a structured interface to the Entrez system databases.

To access these data, a piece of software first posts an eUtils URL to NCBI, then retrieves the results of this posting, after which it processes the data as required. The software can thus use any computer language that can send a URL to the eUtils server and interpret the XML response, such as Perl, Python, Java, and C++. Combining eUtils components to form customized data pipelines within these applications is a powerful approach to data manipulation. More information and training on this process are available through a course on NCBI Powerscripting: http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/course.html.

4. Genomic Databases

The genome sequencing era that started about 20 years ago has brought into being a range of genome resources. Genomic studies of model organisms give insights into understanding of the biology of humans enabling better prevention and treatment of human diseases. Comparative genome analysis leads to further understanding of fundamental concepts of evolutionary biology and genetics. A review on genome resources (11) reports on a selection of genomes of model species – from microbes to human. Species-specific genomic databases comprise a lot of invaluable information on genome biology, phenotype, and genetics. However, primary genomic sequences for all the species are archived in public repositories that provide reliable, free, and stable access to sequence information. In addition, NCBI provides several genomic biology tools and online resources, including group-specific and organism-specific pages that contain links to many relevant Web sites and databases (see **Table 2.2** for the list of available resources and URLs).

4.1. Trace Repositories

Most of the data generated in genome sequencing projects is produced by whole genome shotgun sequencing, resulting in random short fragments (traces).

For many years, the traces (raw sequence reads) remained out of the public domain because the scientific community has focused its attention primarily on the end product: the fully assembled final genome sequence. As the analysis of genomic data progressed, it became necessary to go back to the experimental evidence that underlies the genome sequence to see if there is any ambiguity or uncertainty about the sequence.

Table 2.2
Web genome resources at NCBI

Trace Archive	http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?
Assembly Archive	http://www.ncbi.nlm.nih.gov/Traces/assembly/assmbrowser.cgi?
Short Read Archive (SRA)	http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?
Entrez (cross-database search)	http://www.ncbi.nlm.nih.gov/sites/gquery
Genomic Biology	http://www.ncbi.nlm.nih.gov/Genomes/
Fungal Genome Central	http://www.ncbi.nlm.nih.gov/projects/genome/guide/fungi/
Microbial genomes	http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html
Organelles	http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/organelles.html
Plant Genome Central	http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html
Influenza Virus Resource	http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html
Retrovirus Genomes	http://www.ncbi.nlm.nih.gov/retroviruses/
Viral genomes	http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html
Genomic BLAST	http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi
Concise BLAST	http://www.ncbi.nlm.nih.gov/genomes/prokhits.cgi
gMap	http://www.ncbi.nlm.nih.gov/sutils/gmap.cgi
Map Viewer	http://www.ncbi.nlm.nih.gov/projects/mapview/
ProtMap	http://www.ncbi.nlm.nih.gov/sutils/protmap.cgi
TaxPlot	http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi

4.1.1. Trace Archive

To meet these needs, NCBI and The Wellcome Trust Genome Campus in Hinxton, United Kingdom, created in 2001 a repository of the raw sequence traces generated by large sequencing projects that allows retrieval of both the sequence file and the underlying data that generated the file, including the quality scores. The Assembly Archive (12) created at NCBI in 2004 links the raw sequence information found in the Trace Archive with consensus genomic sequence.

4.1.2. Short Read Archive (SRA)

Trace Archive has successfully served as a repository for the data produced by capillary-based sequencing technologies for many years. New parallel sequencing technologies (e.g., 454, Solexa, Illumina, ABI Solid, Helicos) have started to produce massive amounts of short sequence reads (20–100 kb). Due to the

structure and volume of this data, it is clear that it does not efficiently and effectively fit in the current Trace Archive design, so NCBI has constructed a more appropriate repository, the Short Read Archive. The SRA project is well underway and is being built in collaboration with Ensembl, sequencing centers, and the vendors themselves. SRA Web site has been launched in January 2008: http://www.ncbi.nlm.nih.gov/Traces/sra.

4.1.3. GenBank – Primary Sequence Archive

GenBank is the NIH genetic sequence database, an archival collection of all publicly available DNA sequences (3). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ) (4), the European Molecular Biology Laboratory (EMBL) (2), and GenBank at NCBI. These three organizations exchange data on a daily basis. Many journals require submission of sequence information to a database prior to publication to ensure an accession number will be available to appear in the paper. As of February 2008 GenBank release 164.0 (ftp://ftp.ncbi.nih.gov/genbank/release.notes/gb164.release.notes) contains more than 83 billion bases in over 80 million sequence entries. The data come from the large sequencing centers as well as from small experimentalists. These sequences are accessible via Web interface by text queries using Entrez or by sequence queries using BLAST. Quarterly GenBank releases are also downloadable via FTP (see Section 8).

4.2. Entrez Databases

A family of Entrez databases comprise an integrated information system that links together heterogeneous information on biomedical and bibliographical data. The major concepts of Entrez information system are described in Section 3. Below are three examples of Entrez databases containing information on genome projects, genomic sequences, and protein sequence encoded by complete microbial genomes.

4.2.1. Entrez Genome

Entrez Genome (13), the integrated database of genomic information at the NCBI, includes the types of records and formats for major taxonomic groups, as well as the precomputed data and online analytical programs developed to aid investigation. The database was created as part of Entrez in September 1995 for large-scale genome sequencing projects. It was motivated by the release of the first complete microbial genome of *Haemophilus influenzae* sequenced at TIGR (14).

Entrez Genome displays data from small viral and organelle genomes, complete and nearly complete genomes from bacteria, and eukaryotes. An entry in Genomes database represents a single replicon such as a chromosome, organelle, or plasmid. As of February 2008 Entrez Genome houses a collection of 7,850 entries organized in six large taxonomic groups: Archaea, Bacteria, Eukaryota, Viroids, Viruses, and Plasmids. It presents the tools and views

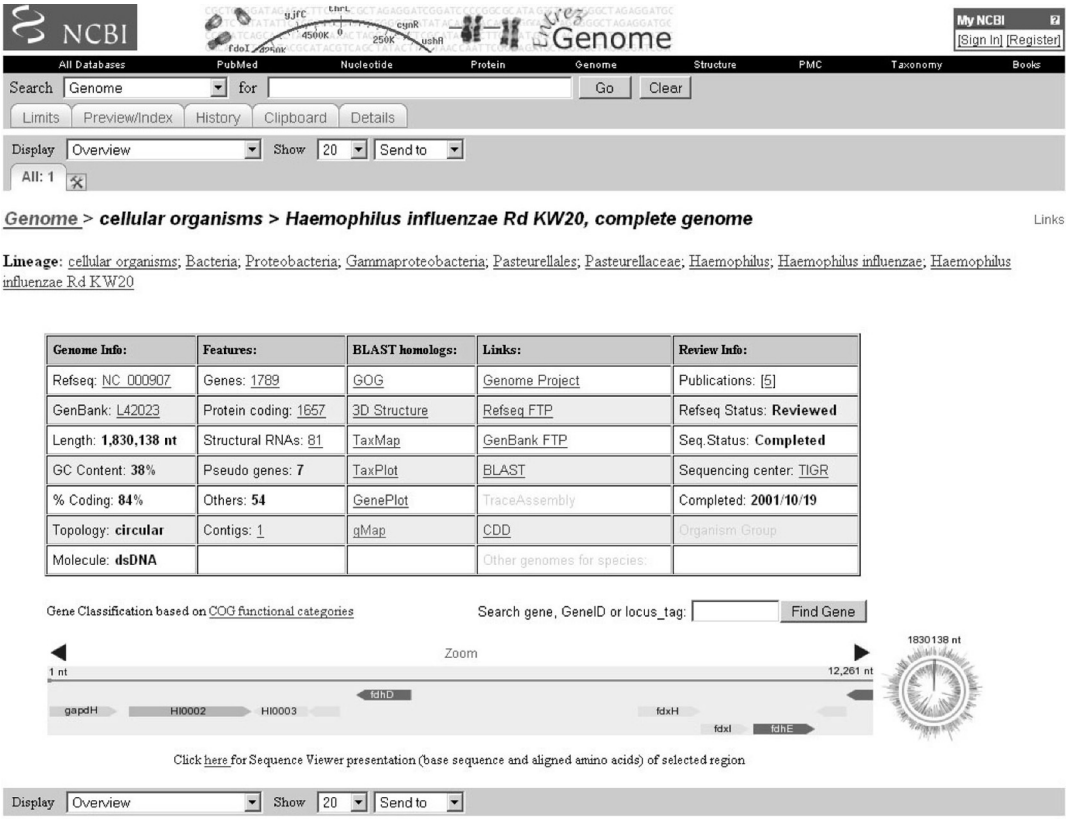

Fig. 2.4. *Haemophilus influenzae* complete genome: single circular chromosome overview. Entrez provides a graphical view of the chromosome with genes color-coded by COG functional categories.

at various levels of detail. For each record, Entrez Genome provides a graphical overview of the chromosome with genes color-coded by COG (clusters of orthologous groups) (15) functional categories (**Fig. 2.4**) as well as other types of text views including flat file, ASN.1, XML, and many others that can be user-selected from a menu. The table provides additional genome information and access to analysis tools.

The available tools include multiple alignments of complete genomes for viruses, precomputed protein clusters from microbial genomes, GenePlot (a genome-scale dotplot generator), TaxPlot (for three-way genome comparisons), gMap, and many others. Some of these tools are described in Section 5 of this chapter. More detailed description of microbial genome resources at NCBI can be found in "In Silico Genomics and Proteomics" (16). Plant genome resources at NCBI have been recently published in a chapter of "Plant Bioinformatics" (17).

Microbial genome sequencing has come a long way since the first *H. influenzae* project. As of February 2008 public collection contains more than 600 complete genomes and close to 500 draft

genome assemblies. The collection represents a very diverse set of organisms; ranging from small (160 kb) endosymbiont Carsonella (18) to the 13-Mb genome of myxobacterium *Sorangium cellulosum* (19). There are organisms isolated from extreme environments such as *Hyperthermus butylicus* (20), an extreme hyperthermophilic, anaerobic archeon, and bacterial species representing deeply branching taxa such as *Rhodopirellula baltica* (21). On the other hand, many projects are aimed toward the comparative analysis of pathogenic bacteria and sequencing multiple strains and isolates of the same organism. For example, *H. influenzae* bacterium is represented in the database by 16 entries including chromosomes and plasmids from different isolated strains. Entrez provides tools that facilitate comparative genome analysis leading into new insights to be gained from genome sequences.

Query examples

Find all the chromosomes of *Haemophilus influenzae:*
***Haemophilus influenzae*[organism] AND chromosome[replicon type]**

4.2.2. Entrez Genome Project

The NCBI Genome Project database is a collection of complete and incomplete (in-progress) large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms. A project is defined by a unique combination of organism name (or metagenomic project name), sequencing center, and sequencing method.

Currently, the database is comprised of projects that have submitted data to NCBI, intend to submit data, or have received public funding. A large eukaryotic genome project usually consists of several components. In the database, projects are organized in a hierarchical, parent–child relationship. A top-level project represents an organism-specific overview and links together all relevant child projects. Each project has its own unique identifier, the Project ID.

The International Nucleotide Sequence Databases Consortium (INSDC) has acknowledged the need to organize genomic and metagenomic data and to capture project metadata. Starting from 2006, the submitters of genome sequence data are required to register their project and obtain a unique project ID. As presented at EMBL guidelines Web site, http://www.ebi.ac.uk/embl/Documentation/project_guidelines.html,

"A project is defined as a collection of INSDC database records originating from a single organization, or from a consortium of coordinated organizations. The collective database records from a project make up a complete genome or metagenome and may contain genomic sequence, EST libraries and any other sequences that contribute to the assembly and annotation of the genome or metagenome. Projects group records either from single organism studies or from metagenomic studies comprising communities of organisms."

NCBI has developed a SOAP (simple object access protocol) compliant Web service, supporting the functions of inserting, updating, deleting, and retrieving of the documents which are used by INSDC collaborators to access/edit the Genome Project database, which in turn controls ProjectIDs and Locus-tag prefixes as well as other project information.

The NCBI Entrez Genome Project database (GenomePrj) is organized into organism-specific overviews that function as portals from which all projects pertaining to that organism can be browsed and retrieved. **Figure 2.5** shows a schematic diagram of a generic eukaryotic genome project.

GenomePrj is integrated into the Entrez search and retrieval system, enabling the use of the same search terms and query structure used in other Entrez databases.

GenomePrj is a companion database to Entrez Genome. Sequence data are stored in Entrez Genome (as complete chromosomes, plasmids, organelles, and viruses) and Entrez Nucleotide (as chromosome or genomic fragments such as contigs). While

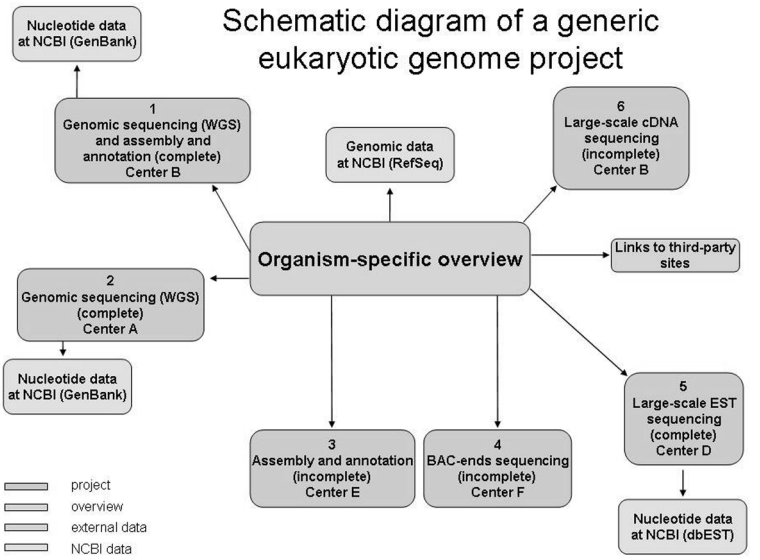

Fig. 2.5. Schematic diagram of a generic eukaryotic genome project. The main overview page shows links to all subprojects, numbered 1–6. Various sequencing centers are associated with each subproject (A–F). These various centers could actually be consortiums consisting of multiple centers. A given center could produce more than one type of project, and data for a given project type could be generated from multiple independent centers. Some of the projects are complete with associated data present in various forms in different Entrez databases at NCBI, while other projects are in progress with no publicly available data at NCBI. A project could be converted over time from containing preliminary data (e.g., WGS) to one where a complete data set is present. RefSeq genomic data are associated with the overview project. Links to third-party sites which contain information of interest regarding the organism are provided.

Entrez Genome does not collect all data for a given organism, GenomePrj provides an umbrella view of the status of each genome project, links to project data in the other Entrez databases and a variety of other NCBI and external resources associated with a given genome project. Sequences associated with a given organism can also be retrieved in the taxonomy browser. However, no distinction is made between GenBank (non-curated) and RefSeq (curated) sequences. There is also no distinction based on which sequencing center submitted the data. Entrez Genome Project also lists projects that are in progress or for which NCBI has not yet received any data. See **Table 2.3** for a comparison of all three databases.

As of January 2008 Genome Project database contains 80 metagenomics project. As shown in **Fig. 2.6**, the database entry contains brief project description, listing of all related subprojects, and project data which include links to genomic data, publication, and Trace data. NCBI Resource Links include an option to BLAST against this particular collection as well as an option to BLAST against all available environmental sequences.

Table 2.3
Comparison of entrez databases

Entrez databases	Organism-specific sequences	Project-specific sequences	Submitter-specific sequences	Complete and in progress	GenBank and RefSeq sequences
Genome	Yes	No	Yes	No	Separated
Taxonomy	Yes	No	No	No	Together
Genome project	Yes	No	Yes	Yes	Separated

Query examples

Find all complete fungal genome projects.

fungi[ORGN] AND complete[SEQSTAT]

Find all projects that correspond to pathogens that can infect humans.

human[HOST]

Find all metagenomic projects

type_environmental[All Fields]

4.2.3. Entrez Protein Clusters

Protein Clusters database is a rich collection of related protein sequences from complete prokaryotic and organelle Reference Sequence (RefSeq) genomes.

Fig. 2.6. Mouse gut metagenome project in Entrez Genome Project database: comparisons of the distal gut microbiota of genetically obese mice and their lean littermates.

Proteins from all complete microbial genomes and plasmids (and separately all chloroplasts) are compared using BLAST all-against-all. Protein clusters are created using a modified BLAST score that takes into account the length of the hit (alignment) versus both the query and the subject. The modified score is then sorted, and all proteins that are contained within the top hits are clustered together. Automatically constructed clusters are then evaluated manually by curators. Based on the sequence alignment information and biological expertise, curators can join or split clusters and add annotation information (protein name, gene name, description) and publication links.

As of January 2008, the database contains 1.4 million proteins that compose 6,043 curated clusters and more than 200,000 automatic clusters. The Entrez Protein Clusters database uses all of the features of other Entrez databases. There are numerous ways to query protein clusters, either with search terms in Entrez or with

a protein or nucleotide sequence. The display for each cluster provides information on cluster accession, cluster name, and gene name, as well as links to protein display tools, external databases, and publications (**Fig. 2.7**). Protein Clusters database can be queried with a protein or nucleotide sequence by using Concise Protein BLAST, a new Web resource developed at NCBI. Concise BLAST is an efficient alternative to standard BLAST. The searchable database is comprised of only one randomly chosen protein from each cluster, and also proteins which are not included in any cluster to assure completeness. This allows rapid searching of the smaller database, but still assures an accurate identification of the query while providing a broader taxonomic view.

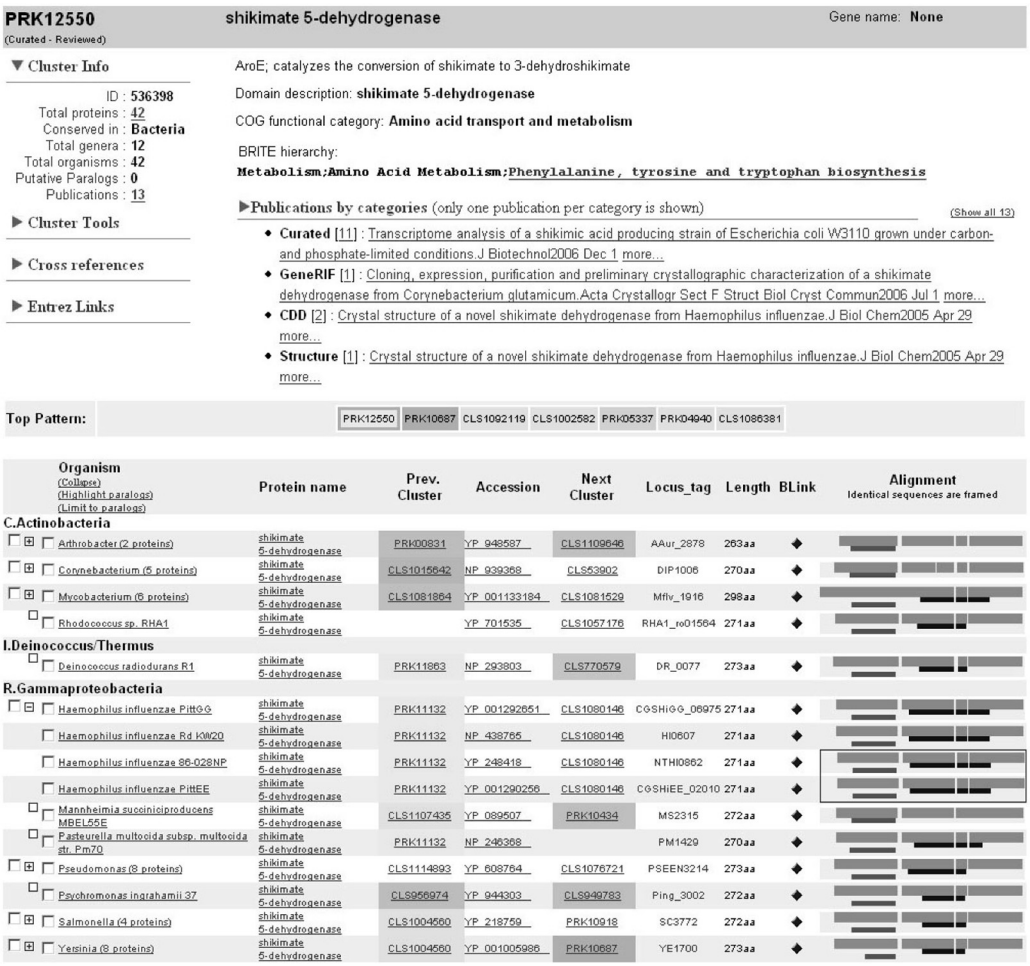

Fig. 2.7. Shikimate 5-dehydrogenase overview in Entrez Protein Clusters database. The top part of the page presents text description, some statistics (Cluster Info), direct access to Cluster Tools, cross-references to outside resources, and links to other Entrez databases. The bottom part presents a colored table: clusters are organized into taxonomic groups; cluster position neighbors are shown as well as a summary of alignments and conserved domains. Clicking on alignment summary will open a detailed multiple alignment view (not shown).

Query examples

Retrieve all clusters containing the protein beta galactosidase:

beta galactosidase [Protein Name]

Find all clusters associated with *Escherichia coli*:

Escherichia coli[Organism]

5. Analysis of Prokaryotic Genome Data

5.1. gMap – Compare Genomes by Genomic Sequence Similarity

gMap is one of the tools available in Entrez Genome that allows to view and analyze the regions of similarity in closely related genomes. **Figure 2.8** shows closely related strains of *H. influenzae*.

Genomic sequences are compared using BLAST and the resultant hits are filtered out to find the largest syntenic regions. Similar regions are shown color-coded and numbered in each genome with an arrow denoting the 5′–3′ direction of the hit with respect to similar segments in other genomes. Additional sequences can be added by inputting the accession number.

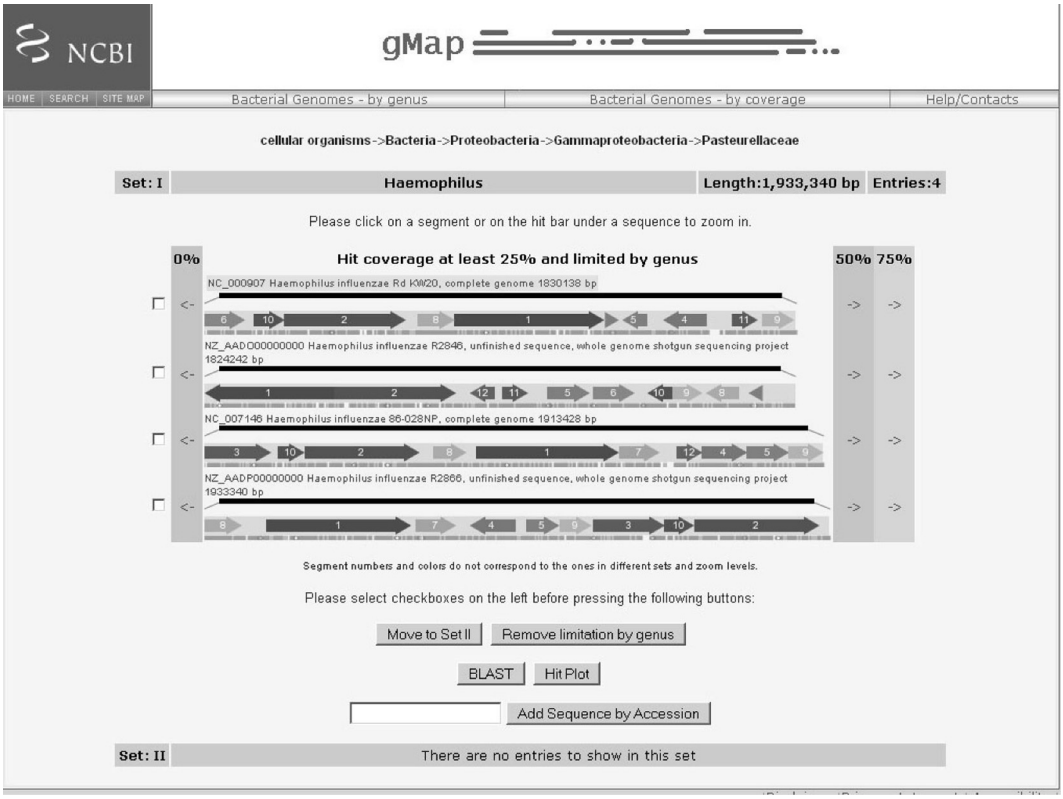

Fig. 2.8. gMap results for four closely related strains of *Haemophilus influenzae* at 25% coverage.

The tool allows navigating from the general overview of domains of life (e.g., bacteria or viruses) down to genome sets with different degrees of mutual similarity. It allows more detailed views of every similarity segment, including the ability to view sequence alignment of two selected similarity regions. Zooming in can be accomplished by clicking on a syntenic group to expand all similar segments. Alternately, a user can click on a hit bar just below the segments to zoom into the surrounding region of the current sequence; this action also displays homologous syntenies from other organisms. After zooming in, all segments are recalculated, recolored, and renumbered, providing a truly dynamic and interactive system with each calculated view presented as a standalone display which is visually easy to comprehend. Pairs of genomic sequences can be selected for output to BLAST, GenePlot, or HitPlot and any number of sequences can be removed from the list by the user to customize the final view to be most appropriate for the user's project. HitPlot shows a dotplot of the two genomes selected based on the magnification level. Precomputed results are available for two categories, one for genomes from the same genus and one for genomes based on the coverage of BLAST hits. Genomes of two or more species from the same genus may not display high levels of synteny, but similar segments in their two genomes can be found at different levels of hit coverage. An example of this would be the *Mycoplasma* genomes. The converse is that organisms from different genera have large syntenic blocks in their genomes such as is found in *Escherichia*, *Salmonella*, and *Shigella*, which are all members of the Enterobacteriaceae family (22). Genomes in both categories are grouped together based on single linkage clustering of coverage level. For example, if genome A has 75% coverage to genome B and genome B has 75% coverage to genome C, then they will all be included in a cluster at the 75% level even though the coverage between A and C may not reach the 75% level.

5.2. Genome ProtMap – Compare Genomes by Protein Sequence Similarity

Genome ProtMap is a comparative display of the genome neighborhoods linked by the orthologous protein sequences. It displays a 10-kb region surrounding either all the proteins in the cluster or, alternately, all the proteins that have the same Cluster of Orthologous Group – COG (15) – or in the case of viruses, VOGs. In the Genome ProtMap display (**Fig. 2.9**), the organism groups are collapsed; clicking the + will expand the group. Clicking the accession number will link to the RefSeq nucleotide record. Mouse over the proteins gives detailed information such as name, cluster ID, and genome location. Clicking on any protein brings up a pop-up menu with links to protein, gene, or cluster. The list of taxa in the ProtMap can be collapsed or expanded by clicking the + or – next to the taxon. "Show Legends" gives the color-coded functional category for the

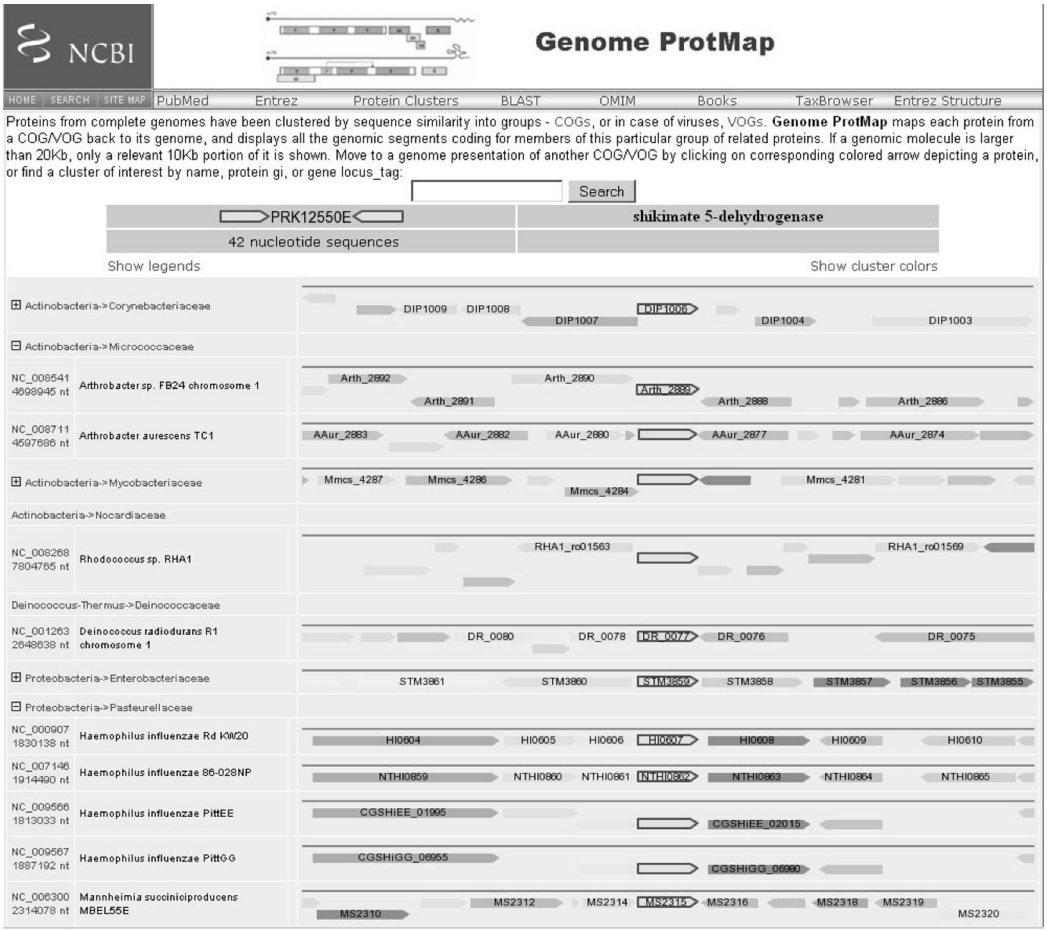

Fig. 2.9. Genome ProtMap shows local genomic neighborhood centered on a set of related genes (via the proteins encoded by them). Related genes are defined by protein clustering algorithms (COG, VOG, or PRK). All proteins in the surrounding area are color-coded by COG category (if applicable) or gray (proteins that do not belong to a COG). ProtMap for PRK12550 – shikimate 5-dehydrogenase is displayed.

proteins while "Show Cluster Colors" lists all the clusters in the ProtMap colored by COG functional category and the name of the cluster.

5.3. Concise BLAST

Concise protein BLAST uses the BLAST engine to allow searching of protein clusters' data sets with a protein or nucleotide sequence query. The database represents protein sequences from complete microbial (prokaryotic) genomes. It uses precalculated clusters of similar proteins at the genus level to represent proteins by groups of related sequences. One representative from each cluster is chosen in order to reduce the data set. The result is reduced search times through the elimination of redundant proteins while providing a broader taxonomic view.

6. Browsing Eukaryotic Genome Data

The main NCBI genome browser Map Viewer provides special browsing capabilities for a subset of organisms in Entrez Genome. The list of organisms available for Map Viewer browsing can be found on the Map Viewer home page (http://www.ncbi.nlm.nih.gov/projects/mapview/).

Map Viewer can display a collection of aligned genetic, physical, or sequence-based maps, with an adjustable focus rangingfrom that of a complete chromosome to that of a portion of a gene. The maps displayed in Map Viewer may be derived from a single organism or from multiple organisms; map alignments are performed on the basis of shared markers. The availability of whole genome sequences means that objects such as genes, markers, clones, sites of variation, and clone boundaries can be positioned by aligning defining sequences from these objects against the genomic sequence. This positional information can then be compared to information on order obtained by other means, such as genetic or physical mapping. The results of sequence-based queries (e.g., BLAST) can also be viewed in genomic context as described in the next section.

Any text search term can be used as a query at the top of the Map Viewer home page. These include, but are not limited to, a GenBank accession number or other sequence-based identifier, a gene symbol or alias, or the name of a genetic marker. For more complex queries, any query can be combined with one of three Boolean operator terms (AND, OR, and NOT). Wild cards, which are denoted by placing a * to the right of the search term, are also supported. Map Viewer uses the Entrez query search engine, described in section 3, to analyze a complex query and perform a search.

Another way of getting to a particular section of a genome is to use a range of positions as a query. First it is necessary to select a particular chromosome for display from a genome-specific Map Viewer page. Once a single chromosome is displayed, position-based queries can be defined by (1) entering a value into the Region Shown box. This could be a numerical range (base pairs are the default if no units are entered), the names of clones, genes, markers, SNPs, or any combination. The screen will be refreshed with only that region shown.

Map Viewer provides an option to simultaneously search physical, genetic, and sequence maps for multiple organisms. This option is currently available for plant and fungal genomes. Multi-organism plant searching is available at http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=33090.

Since the early 1990s several researchers have shown that large-scale genome structure is conserved in blocks across the grasses (23–26). Locus nomenclature is organism-specific and is unreliable as a query method between species; however, the regular nomenclature of plasmids (27) is not influenced by how the plasmid or insert is

used. The data for the plant maps available through Map Viewer include the probe–locus relationship for each locus where the allelic state is identified by the probe. This information enables the rendering of the visual connection between those mapped loci in adjacently displayed maps that were identified by the same probe. This locus–probe relationship allows a cross-species text search using the probe name as the query string. **Figure 2.10** shows the result of the search across all plants using "cdo718" as a query.

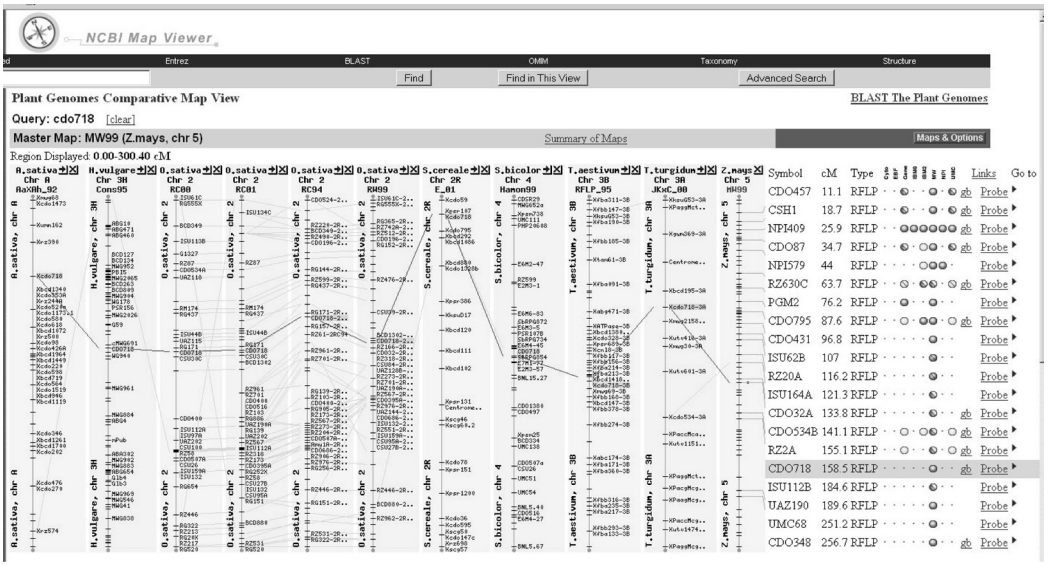

Fig. 2.10. Map Viewer Displays resulting from a search for marker "cdo718" showing aligned maps from several plants. The marker "cdo718" is highlighted on each map with lines between maps connecting the highlighted markers.

"cdo718" is the name of a plasmid with an oat cDNA insert. This probe was used to map loci in nine maps available in Map Viewer: the AaXAh-92 map in *Avena sativa*, the Cons95 map in *Hordeum vulgare*, the RC94, RW99, R, RC00, and RC01 maps in *Oryza sativa*, the E-01 map in Secale cereale, the S-0 map in *Triticum aestivum, JKxC map in Triticum turgidum*, and the RW99 map in *Zea mays*. The dark grey lines between each map connect the loci identified by the probe. The light gray lines connect the other loci in adjacent maps that have been identified by the same probe.

7. Searching Data by Sequence Similarity (BLAST)

The Basic Local Alignment Search Tool (BLAST) (28) finds regions of local similarity between sequences. By finding similarities between sequences, scientists can infer the function of newly sequenced genes, predict new members of gene families, and explore evolutionary relationships.

7.1. Organism-Specific Genomic BLAST

Genome-specific BLAST pages that restrict a search to a specific genome are provided for several organisms and allow the results of the search to be displayed in a genomic context (provided by Map Viewer).

Query sequence (protein or nucleotide) can be compared to genomic, transcript, or protein coded by the genome. **Table 2.4** provides the list of available databases. Not all databases are always available; some projects provide additional data sets such as SNP, traces, and alternative assemblies. If the reference genome (the default) is selected as the database to be searched, the Genome View button (**Fig. 2.11B**) will appear on a diagram showing the chromosomal location of the hits (**Fig. 2.11C**). Each hit links to a Map Viewer display of the region encompassing the sequence alignment.

7.2. Multi-organism Genomic BLAST

Microbial Genomic BLAST (29) provides access to complete genomes and genome assemblies of 940 Bacteria and 48 Archaea and 162 Eukaryota (as of February 2008). Genomic BLAST has been recently extended to include data sets for insects, fungi, nematodes, protozoa, and metagenomes. The genomes can be viewed

Table 2.4
Customized project-specific BLAST databases

DB name	Description
Genome (all assemblies)	Sequences from all available genome assemblies
Genome (reference only)	Sequences from the reference assembly only
RefSeq RNA	RefSeq transcript sequences (NM + XM)
RefSeq protein	RefSeq protein sequences (NP + XP)
Non-RefSeq RNA	GenBank transcript sequence
Non-RefSeq protein	GenBank protein sequences
Build RNA	Proteins generated in the annotation run
Build protein	Proteins generated in the annotation run
Ab initio RNA	Transcripts generated in the annotation run by Gnomon only
Ab initio protein	Proteins generated in the annotation run by Gnomon only
EST	EST sequences by organism
Clone end sequences	Clone end sequences by organism
Traces WGS	Raw sequence reads for genomic assemblies
Traces EST	Raw sequence reads for EST
SNP	Custom database of Single Nucleotide Polymorphism database

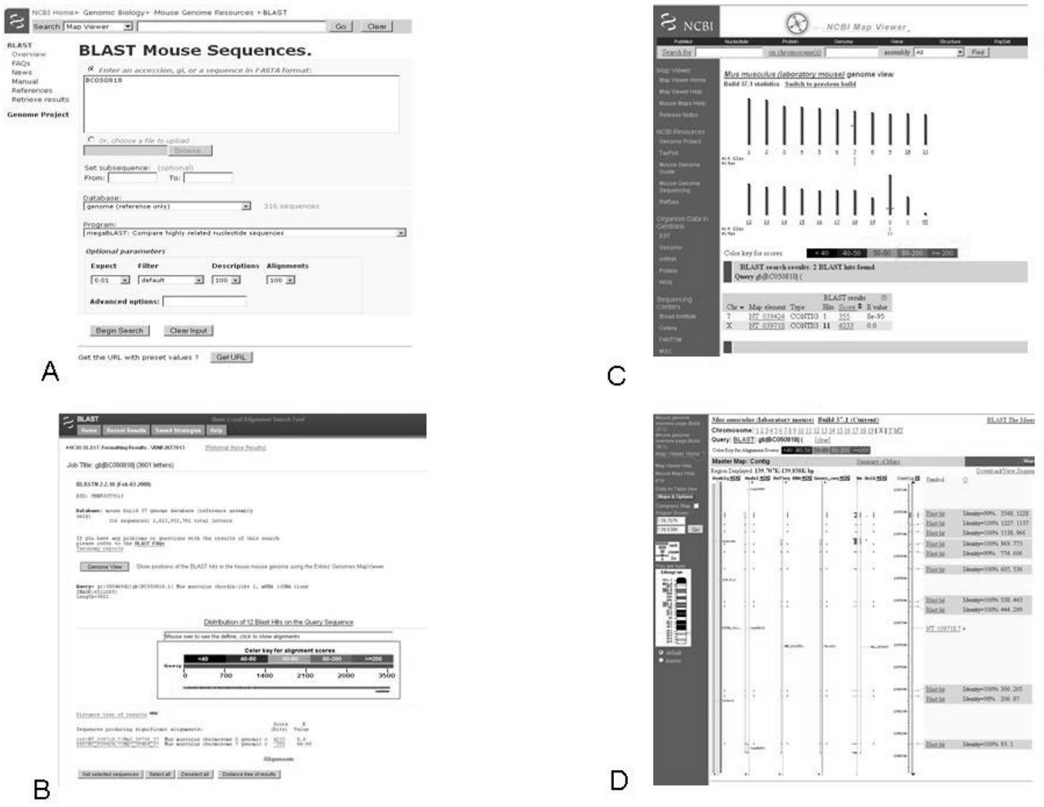

Fig. 2.11. Genomic BLAST: BLAST Mouse Sequences. A Query mouse reference genome with mouse cDNA clone, accession BC050818. B BLAST output page has an additional "Genome View" button that provides an option to show the hits in genome environment via Map Viewer. C Genome overview of BLAST hits. The hits are represented by colored ticks providing the links to zoomed-in view of the chromosome. D Positions of the BLAST hits on the chromosome. The maps shown include Model (NCBI annotation pipeline prediction), RefSeq transcript, and mouse UniGene.Interesting to note that the first exon (hit 3548..1228) is not included in the RefSeq model, although it is supported by UniGene and predicted by NCBI annotation pipeline.

in taxonomic groups or in alphabetical order. A flexible user-friendly interface allows to construct virtual blast databases for the specific searches.

For example, with many closely related microbial genomes sequenced, one might want to exclude the close relatives from consideration in order to reveal more evolutionary interesting remote relationships.

8. FTP Resources for Genome Data

The source genome records can be accessed from the GenBank directory; these are the records that were initially deposited by the original submitters. The reference genomes,

assemblies, and associated genes and proteins can be downloaded from the Genomes and RefSeq directories. Information on the data content in these FTP directories is located in the README files.

Download the full release database, daily updates, or WGS files:

ftp://ftp.ncbi.nih.gov/genbank/

Download complete genomes/chromosomes, contigs and reference sequence mRNAs and proteins:

ftp://ftp.ncbi.nih.gov/genomes/

Download the curated RefSeq full release or daily updates:

ftp://ftp.ncbi.nih.gov/refseq/

Download curated and non-curated protein clusters from microbial and organelle genomes:

ftp://ftp.ncbi.nih.gov/genomes/Bacteria/CLUSTERS

9. Conclusion

The tremendous increase in genomic data in the last 20 years has greatly expanded our understanding of biology. Genome sequencing projects now span from draft assemblies, complete genomes, large-scale comparative genomic projects, to the new field of metagenomics where genetic material is recovered directly from environmental samples and the entire complement of DNA from a given ecological niche is sequenced. Although these provide an ever greater resource for studying biology, there is still a long way to go from the initial submission of sequence data to the understanding of biological processes. By integrating different types of biological and bibliographical data, NCBI is building a discovery system that enables the researcher to discover more than would be possible from just the original data. By making links between different databases and computing associations within the same database, Entrez is designed to infer relationships between different data that may suggest future experiments or assist in interpretation of the available information. In addition, NCBI is developing the tools that provide users with extra layers of information leading to further discoveries.

Genomics is a very rapidly evolving field. The advance in sequencing technologies has lead to new data types which require different approaches to data management and presentation. NCBI continues to add new databases and develop new tools to address the issue of ever-increasing amounts of information.

Acknowledgments

The authors would like to thank, in alphabetic order, Vyacheslav Chetvernin, Boris Fedorov, Andrei Kochergin, Peter Meric and Sergei Resenchuk, and Martin Shumway for their expertise and diligence in the design and maintenance of the databases highlighted in this publication and Stacy Ciufo for the helpful discussion and comments. These projects represent the efforts of many NCBI staff members along with the collective contributions of many dedicated scientists worldwide.

References

1. Liolios, K., Mavrommatis, K., Tavernarakis, N., Kyrpides, N. C. (2007) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 36(Database issue), D475–D479.
2. Cochrane, G., Akhtar, R., Aldebert, P., Althorpe, N., Baldwin, A., Bates, K., Bhattacharyya, S., Bonfield, J., Bower, L., Browne, P., Castro, M., Cox, T., Demiralp, F., Eberhardt, R., Faruque, N., Hoad, G., Jang, M., Kulikova, T., Labarga, A., Leinonen, R., Leonard, S., Lin, Q., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Plaister, S., Robinson, S., Sobhany, S., Vaughan, R., Wu, D., Zhu, W., Apweiler, R., Hubbard, T., Birney, E. (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res* 36(Database issue), D5–D12.
3. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Wheeler, D. L. (2008) GenBank. *Nucleic Acids Res* 36(Database issue), D25–D30.
4. Sugawara, H., Ogasawara, O., Okubo, K., Gojobori, T., Tateno, Y. (2008) DDBJ with new system and face. *Nucleic Acids Res* 36(Database issue), D22–D24.
5. Galperin, M. Y. (2008) The molecular biology database collection: 2008 update. *Nucleic Acids Res* 36(Database issue), D2–D4.
6. Wheeler, D. L., et al. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 36(Database issue), D13–D21.
7. Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z. Y., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J., Sherry, S. T. (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 39(10), 1181–1186.
8. Pruitt, K. D., Tatusova, T., Maglott, D. R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue), D61–D65.
9. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17), 3389–3402. Review.
10. Maglott, D. R., Ostell, J., Pruitt, K. D., Tatusova, T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 35(Database issue), D26–D31.
11. Hillary, E. S., Maria, A. S., eds. (2006) *Genomes (Cold Spring Harbor Monograph Series, 46)*. Cold Spring Harbor, New York.
12. Salzberg, S. L., Church, D., DiCuccio, M., Yaschenko, E., Ostell, J. (2004) The genome Assembly Archive: a new public resource. *PLoS Biol.* 2(9), E285.
13. Tatusova, T. A., Karsch-Mizrachi, I., Ostell, J. A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics* 15(7–8), 536–543.
14. Fleischmann, R. D., et al. Whole-genome random sequencing and assembly of Haemophilus influenza Rd. (1995) *Science* 269(5223), 496–512.

15. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., Natale, D. A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41.
16. Klimke, W., Tatusova, T. (2006) Microbial genomes at NCBI in (Mulder, N., Apweiler, R., eds.) *In Silico Genomics And Proteomics: Functional Annotation of Genomes And Proteins*, Nova Science Publishers; 1st ed., pp. 157–183.
17. Tatusova, T., Smith-White, B., Ostell, J. A. (2006) Collection of plant-specific genomic data and resources at the National Center for Biotechnology Information, in (David, E., ed.), *Plant Bioinformatics: Methods and Protocols (Methods in Molecular Biology), Humana Press*, 1st ed., pp. 61–87.
18. Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A., Hattori, M. (2006) The 160-kilobase genome of the bacterial endosymbiont. *Carsonella Sci* 314(5797), 267.
19. Schneiker, S., et al. (2007) Complete genome sequence of the myxobacterium Sorangium cellulosum. *Nat Biotechnol* 25(11), 1281–1289.
20. Brügger, K., et al. (2007) The genome of Hyperthermus butylicus: a sulfur-reducing, peptide fermenting, neutrophilic Crenarchaeote growing up to 108 degrees C. *Archaea* 2(2), 127–135.
21. Teeling, H., Lombardot, T., Bauer, M., Ludwig, W., Glockner, F. O. (2004) Evaluation of the phylogenetic position of the planctomycete 'Rhodopirellula baltica' SH 1 by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees. *Int J Syst Evol Microbiol* 54, 791–801.
22. Darling, A. C., Mau, B., Blattner, F. R., et al. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14(7), 1394–1403.
23. Ahn, S. N., Tanksley, S. D. (1993) Comparative linkage maps of the rice and maize genomes. *Proc Natl Acad Sci USA* 90, 7980–7984.
24. Devos, K. M., Chao, S., Li, Q. Y., Simonetti, M. C., Gale, M. D. (1994) Relationship between chromosome 9 of maize and wheat homeologous group 7 chromosomes. *Genetics* 138, 1287–1292.
25. Kurata, N., Moore, G., Nagamura, Y., Foote, T., Yano, M., Minobe, Y., Gale, M. D. (1994) Conservation of genome structure between rice and wheat. *Biotechnology (NY)* 12, 276–278.
26. van Deynze, A. E., Nelson, J. C., O'Donoghue, L. S., Ahn, S. N., Siripoonwiwat, W., Harrington, S. E., Yglesias, E. S., Braga, D. P., McCouch, S. R., Sorrells, M. E. (1995) Comparative mapping in grasses: oat relationships. *Mol Gen Genet* 249, 349–356.
27. Lederburg, E. M. (1986) Plasmid prefix designations registered by the Plasmid Reference Center 1977–1985. *Plasmid* 1, 57–92.
28. Altschul, S. F., Gish, W., Miller, W., et al. (1990). Basic local alignment search tool. *J Mol Biol* 215(3), 403–410.
29. Cummings, L., Riley, L., Black, L., Souvorov, A., Resenchuk, S., Dondoshansky, I., Tatusova, T. (2002) Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes. *FEMS Microbiol Lett* 216(2), 133–138.

http://www.springer.com/978-1-60327-240-7

Data Mining Techniques for the Life Sciences
Carugo, O.; Eisenhaber, F. (Eds.)
2010, XII, 408 p. 89 illus., Hardcover
ISBN: 978-1-60327-240-7
A product of Humana Press