NCBI News

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION
National Library of Medicine
National Institutes of Health

July 1998

Accessing Complete Genomes in GenBank Via Entrez and FTP

There are currently over 400 completely sequenced organisms in GenBank, including 378 viruses, 16 bacteria, *Saccharomyces cerevisiae*, and a number of plasmids and organelles. NCBI makes these available in a variety of ways to accommodate the needs of the scientific community.

Genomes can be downloaded as single complete sequence records. Larger genomes can also be downloaded as sets of smaller overlapping sequence records if that is preferred.

In addition, the Entrez Genomes division provides graphical views of genome maps at varying levels of detail. Entrez Genomes also includes "TaxTable" summaries of analyses that have been done using BLAST to identify the possible protein functions and phylogenetic relationships.

Downloading Complete Sequence Records

Complete genome records falling within the 350-kb limit for a single sequence record (as agreed upon by the International Nucleotide Sequence Database Collaboration) can be downloaded in several formats from the Entrez Nucleotide division.

However, genome records exceeding 350 kb are not available as single, complete sequences in GenBank or in the Entrez Nucleotide division. Instead, these larger records can be downloaded from the NCBI FTP site (ncbi.nlm.nih.gov) under the genbank/genomes directory. Subdirectories exist for bacteria, *S. cerevisiae*, and *C. elegans*, and contain complete sequence records in a variety of formats as explained in the README file.

For example, the *Escherichia coli* genome is available under the bacteria/Ecoli subdirectory. There you will find a GenBank flat file record (ecoli.gbk) for the complete 4.6-Mb sequence, with corresponding biological annotations and the accession number U00096. In

Continued on page 2

Searching Unfinished Microbial Genomes

NCBI has collected preliminary sequencing data on 18 microbial genomes and has made these data available for BLAST sequence similarity searching as part of the BLAST service at http://www.ncbi.nlm.nih.gov/BLAST/. Following the link **Microbial Genomes** near the bottom of the BLAST search page leads to a search form in which any or all of the 18 partial genomes listed in the table (see page 4) can be selected as the database to search. In addition, the service has recently incorporated the 16 finished microbial genomes that are in GenBank.

The search service employs the Gapped BLAST 2.0 program and supports the TBLASTN, BLASTN, and TBLASTX search modes. Four BLAST parameters can be adjusted directly on the Web page. These include the Expect value, the Filter method, the number of Descriptions to be displayed, and the number of Alignments shown. Additional parameters can be specified via a command line switch window labeled **Other Advanced Options**. Query sequences, in FASTA format, are pasted into a sequence window as in the main BLAST page, or the accession number of a GenBank entry can be specified.

Note that the partial genomic sequences that make up this database have *not* yet been deposited in GenBank and are *not* accessible through Entrez. However, they can be retrieved by FTP from their associated sequencing centers. The link **About the Databases** on the search page provides a gateway to Web pages maintained by these centers. Because these unfinished

Continued on page 4

IN THIS ISSUE

Complete Genomes in GenBank .. 1
Unfinished Microbial Genomes ... 1
Tracking Human Sequencing 3
BLAST 2 Sequences 3
GeneMap '98 3
New Network Clients 3
Recent Publications 4
Frequently Asked Questions 5
NCBI Data by FTP 6
Malaria Genetics Web Site 6
News In Brief 7

NCBI News

NCBI News is distributed two to three times a year. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence and suggestions to *NCBI News* at the address below.

NCBI News
National Library of Medicine
Bldg. 38A, Room 8N-803
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 496-2475
Fax: (301) 480-9241
E-mail: info@ncbi.nlm.nih.gov

Editors
Dennis Benson
Barbara Rapp

Writers
Renata McCarthy
Donna Roscoe
David Wheeler

Managing Editor
Roseanne Price

Graphics and Production
Veronica Johnson

Design Consultant
Troy M. Hill

In 1988, Congress established the National Center for Biotechnology Information as part of the National Library of Medicine; its charge is to create information systems for molecular biology and genetics data, and to perform research in computational molecular biology.

NIH Publication No. 98-3272

ISSN 1060-8788
ISSN 1098-8408 (Online Version)

Complete Genomes, *continued from page 1*

the GenBank database itself, and in Entrez, this full 4.6-Mb sequence is represented as 400 smaller overlapping records, each approximately 10 kb in length, and overlapping by about 50 base pairs. Each of these smaller records has its own accession number, followed by a secondary accession number of U00096 for the full genome. The whole set of 400 can therefore be retrieved by searching for U00096.

Instructions on how the 400 smaller records are assembled to rebuild the complete genome are contained in the GenBank summary file (ecoli.gbs) on the FTP site. That file contains no sequence data, but indicates the order for the 400 accession numbers and how to join them. This GenBank summary file can also be viewed in the Entrez Genomes division by clicking on **Text View**.

The FTP site also includes files containing the complete DNA sequence (ecoli.fna) and the full set of amino acid translations (ecoli.faa) in FASTA format. This is intended to facilitate data analysis and eliminate the need for users to process the individual GenBank flat files. A table with information on the order of proteins within the genome is also provided.

The following list indicates the types of files available on the FTP site in the subdirectory for each genome:

*.asn = ASN.1 file, print format
*.faa = FASTA amino acid file
*.fna = FASTA nucleic acid file
*.gbk = GenBank flat file format
*.gbs = GenBank summary file format
*.ptt = Protein table
*.tab = Table to assemble genome
*.val = ASN.1 binary format
*.tar.Z = UNIX tar for all files in the directory

Graphics and Analytical Data

All of the complete genomes, regardless of size, are available in the Entrez Genomes division. In addition to information supplied by submitters in the GenBank flat file, the Entrez Genomes division provides graphical views and summaries of analyses done at NCBI for each genome.

Using the *E. coli* example, you can search Entrez Genomes for *Escherichia coli* in the Organism field, or by clicking on the organism name under Prominent Organisms Complete Genomes. From the schematic overview of the complete *E. coli* map, you can zoom into a detailed view of an area of interest by searching for a specific marker(s) or clicking on the desired section of the map, and then progress to associated sequence records.

From the sidebar, you can go back to the **Overview** map, access an alphabetical list of **Markers** for the genome, or view a **ProtTable** that lists all the proteins and provides links to FASTA files and BLAST. The newest feature, **TaxTable**, summarizes the results of BLAST analyses done for the proteins, suggests the possible function of the proteins, and displays the relationship of the organism to others through a color-coded graphical summary.

Genomes in Progress

In addition to completely sequenced genomes, Entrez also contains mapping data and contiguous sequence islands for seven eukaryotic genomes whose sequencing is in progress, including human, mouse, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Leishmania major*, rice, and corn. For organisms in

Continued on page 8

Tracking Human Sequencing Progress

The Human Genome Sequencing Index (HGSI) debuted at NCBI in March 1998 and is accessible on the World Wide Web at http://www.ncbi.nlm.nih.gov/HUGO. The HGSI, a joint project of NCBI and the Human Genome Organisation (HUGO), is intended to facilitate tracking and coordination of the efforts of human genome sequencing centers worldwide. The Web interface provides forms whereby sequencing groups can register their sequencing targets and update the progress they have made toward the sequencing of previously registered targets. There are currently 17 participating centers that register their data with HGSI.

Sequencers and nonsequencers alike can monitor the progress of the Human Genome Project with the help of a graphical representation of each human chromosome provided by the HGSI. In this representation, the regions of a chromosome that have been targeted by registered sequencing centers are highlighted. A continuously updated table, listing registered sequencing centers, their sequencing targets, and the status of their sequencing efforts, can also be examined. Both the graphical and the tabular views are reached by clicking on the **Genome Map View** button, which appears on the main page.

Sequencing groups wishing to register with the HGSI can do so by first following the **Contact NCBI** link from the HGSI main page to establish a password for their group. Once a password has been established, it is possible to use the **Submit or Update Targets** button on the main page to register new targets or update previously registered targets. No password is required, however, to review the target data.

To reach the HGSI Web page from the NCBI home page, select **HGSI: Human Genome Sequencing Index** under **Other NCBI Resources.** ■

❖ ❖ ❖ ❖ ❖

BLAST 2 Sequences

BLAST 2 Sequences is a new Web service that creates an optimal alignment between two nucleotide sequences or two protein sequences. Using the BLAST 2.0 algorithm, BLAST 2 Sequences performs a Gapped BLAST search between the two sequences, allowing for the introduction of gaps due to deletions and insertions in the resulting alignment.

The query sequences can be entered directly into the Web page in FASTA format, or by specifying their GenBank accession numbers or GI numbers. At least 100 base pairs are required for optimal alignment.

The results are presented in two formats, as aligned sequences and as graphical representations of similarity. The graphical display shows schematics of the aligned portions between the two sequences, as well as a dot plot showing sequence similarity versus the query sequences on the *x* and *y* axes.

This service is available only on the Web. There are no downloadable executables for the BLAST 2 Sequences programs. ■

GeneMap ’98 Coming This Fall

An updated Gene Map of the Human Genome, GeneMap ’98, will be released by NCBI in October. GeneMap ’98 is the result of the continuing efforts of an international consortium formed in 1994 for the purpose of constructing a comprehensive transcript map of the human genome.

GeneMap ’98 will consist of an ordering of 30,261 gene-based markers within a framework of 1,500 microsatellites. The new transcript map represents an enhancement of the 1996 Gene Map of the Human Genome (*Science* 274:547), which included 16,354 gene-based markers. The earlier transcript map will remain available at its current URL (http://www.ncbi.nlm.nih.gov/genemap) after the release of GeneMap ’98. ■

❖ ❖ ❖

New Network Clients Available

New versions of Network Entrez, Network BLAST, Power-BLAST, and Cn3D are now on the FTP site. The new clients are designed to run through our Web dispatcher, rather than the custom dispatcher. If you are operating behind a firewall, you may need to continue connecting to the custom dispatcher. To do this, add one line to the NCBI configuration file (“ncbi.cnf” on Mac, “ncbi.ini” on Windows, and “.ncbirc” on UNIX), in addition to following any other procedures you have for poking holes in your firewall. This line should read “SRV_CONN_MODE=DISPATCHER” and should be placed in the “[NET_SERV]” section of that file. Contact NCBI for more information if needed. ■

Selected Recent Publications by NCBI Staff

Baxevanis, AD and **D Landsman.** Homology model building of Hho1p supports its role as a yeast histone H1 protein. *In Silico Biol* 1:0002, 1998.

Chung, MS, AF Neuwald, and **WJ Wilbur.** A free energy analysis by unfolding applied to 125-mers on a cubic lattice. *Fold Des* 3:51–65, 1998.

Koonin, EV and **L Aravind.** Genomics: re-evaluation of translation machinery evolution. *Curr Biol* 8:R266–9, 1998.

Koonin, EV and **MY Galperin.** Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr Opin Genet Dev* 7(6):757–63, 1997.

Marchler-Bauer, A, M Levitt, and **SH Bryant.** A retrospective analysis of CASP2 threading predictions. *Proteins* Suppl 1:83–91, 1997.

McEntyre, J. Linking up with Entrez. *Trends Genet* 14:39–40, 1998.

Sonnhammer, EL, SR Eddy, E Birney, A Bateman, and R Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 26:320–2, 1998.

Wilbur, WJ. A comparison of group and individual performance among subject experts and untrained workers at the document retrieval task. *J Am Soc Inf Sci* 49(6):517–29, 1998.

Wolfsberg, TG, I Makalowska, and **W Makalowski.** Web alert: Oncogenes and cell proliferation. *Curr Opin Genet Dev* 8:9–10, 1998.

Wootton, JC. Evaluating the effectiveness of sequence analysis algorithms using measures of relevant information. *Comput Chem* 21:191–202, 1997.

Yedavalli, VR, **C Chappey,** E Matala, and N Ahmad. Conservation of an intact *vif* gene of human immunodeficiency virus type 1 during maternal-fetal transmission. *J Virol* 72(2):1092–102, 1998.

Unfinished Microbial Genomes, *continued from page 1*

microbial sequences represent preliminary results, they may still contain errors, and any BLAST search results must be interpreted with this in mind.

Sequencing centers wishing to make other partially sequenced genomes available at NCBI for BLAST analysis should contact NCBI to set up an FTP account to facilitate data transfer (info@ncbi.nlm.nih.gov).

Microbial Genomes Available for BLAST Searches

BLAST Database	Source
Unfinished Microbial Genomes	
Actinobacillus actinomycetemcomitans	Oklahoma University Advanced Center for Genome Technology
Campylobacter jejuni	The Sanger Centre
Clostridium acetobutylicum ATCC 824	Genome Therapeutics Corporation
Deinococcus radiodurans R1	The Institute for Genomic Research (TIGR)
Enterococcus faecalis	The Institute for Genomic Research (TIGR)
Mycobacterium tuberculosis CSU#93	The Institute for Genomic Research (TIGR)
Neisseria gonorrhoeae	Oklahoma University Advanced Center for Genome Technology
Neisseria meningitidis MC58	The Institute for Genomic Research (TIGR)
Neisseria meningitidis serogroup A	The Sanger Centre
Porphyromonas gingivalis W83	The Institute for Genomic Research (TIGR)
Pseudomonas aeruginosa PAO1	Pseudomonas Genome Project, Cystic Fibrosis Foundation, University of Washington Genome Center, and PathoGenesis Corporation
Pyrococcus furiosus	Utah Genome Center
Staphylococcus aureus	Oklahoma University Advanced Center for Genome Technology
Streptococcus pneumoniae type 4	The Institute for Genomic Research (TIGR)
Streptococcus pyogenes	Oklahoma University Advanced Center for Genome Technology
Thermotoga maritima	The Institute for Genomic Research (TIGR)
Vibrio cholerae serotype O1	The Institute for Genomic Research (TIGR)
Yersinia pestis CO-92 Biovar	The Sanger Centre
Complete Genomes	
Aquifex aeolicus	Diversa Corporation
Archaeoglobus fulgidus	The Institute for Genomic Research (TIGR)
Bacillus subtilis	BSNR, Regulation de l'Expression Genetique, Institut Pasteur
Borrelia burgdorferi	The Institute for Genomic Research (TIGR)
Chlamydia trachomatis	Chlamydia Genome Project, University of California at Berkeley and Stanford University
Escherichia coli K-12	E-coli Genome Project, University of Wisconsin
Haemophilus influenzae	The Institute for Genomic Research (TIGR)
Helicobacter pylori	The Institute for Genomic Research (TIGR)
Methanobacterium thermoautotrophicum	Genome Therapeutics Corporation
Methanococcus jannaschii	The Institute for Genomic Research (TIGR)
Mycobacterium tuberculosis H37Rv	The Sanger Centre
Mycoplasma genitalium	The Institute for Genomic Research (TIGR)
Mycoplasma pneumoniae	Zentrum für Molekulare Biologie, University of Heidelberg
Pyrococcus horikoshii	National Institute of Technology and Evaluation (NITE)
Synechocystis PCC 6803	Kazusa DNA Research Institute (KDRI)
Treponema pallidum	The Institute for Genomic Research (TIGR)

Frequently Asked Questions

When I get my BLAST results, why do I see a string of X's (or N's) in my query sequence that was not there originally?

By default, BLAST filters your query sequence for low-complexity sequences to prevent artifactually high scores, or false positives. The filter substitutes any low-complexity sequence that it finds with the letter N in a nucleotide sequence (e.g., NNNNN) or X in protein sequences (e.g., XXXXX). Low-complexity regions can result in high scores that reflect compositional bias rather than any significant alignment. Filter programs eliminate these potentially confounding matches from the BLAST reports, leaving regions whose BLAST statistics reflect the specificity of the pairwise alignment. You can opt to remove filters on the BLAST query page by setting the **Filter** to **None**.

Why can I no longer retrieve records when I search UniGene for Hs.36032?

When UniGene is updated, clusters can change because UniGene clusters are continuously being reorganized and rebuilt as new sequence information is received. Some numbers are eliminated because clusters are merged, others because they contain low-quality sequence data or were found to be a contaminant. Therefore, the cluster number should not be used as a stable identifier. Instead, use the GenBank accession number assigned to the EST records within UniGene. These are stable identifiers and can be tracked.

I would like to submit mitochondrial DNA sequences data from a population survey. The populations are highly variable. I have a total of 45 different haplotypes within one species for the same locus. How should I submit these sequences?

The Sequin submission tool is very useful for submitting sequence data from population studies. Select **Population Study** from the submission type section. Import your nucleotide sequences and then choose **Haplotype** from the menu in the source modifier section. There is a self-guiding help document you can refer to as you go along. Alternatively, at the time you prepare the data files, annotate the FASTA definition line that goes along with sequence data, for example, [haplotype=x].

How do I limit a WWW Entrez search by molecule type? I would like to retrieve only mRNA records.

Select **Properties** from the pull-down Search Field menu in Entrez nucleotides, and type mRNA into the search box. Alternatively, using the List Terms mode while in the Properties field, enter biomol_rna, or simply biomol, into the search box. You will see an alphabetical list of the available molecule types. You can select one or more terms from that list to include in your query. In this case, it would be biomol_rna.

What organisms are included in the NCBI Taxonomy database?

The purpose of the taxonomy project at NCBI is to build a consistent phylogenetic taxonomy for the sequence databases in the Entrez system. Therefore the Taxonomy database contains only those organisms for which there is at least one DNA or protein sequence. Currently, approximately 40,000 species from 12,000 genera are represented.

NCBI Data by FTP

The NCBI FTP site contains a variety of directories with publicly available databases and software. The available directories include 'repository,' 'genbank,' 'entrez,' 'toolbox,' 'pub,' and 'sequin.'

The **repository** directory makes a number of molecular biology databases available to the scientific community. This directory includes databases such as PIR, REBASE, CarbBank, Repbase, ACeDB, and OMIM.

The **genbank** directory contains files with the latest full release of GenBank, the daily cumulative updates, and the latest release notes.

The **entrez** directory contains the client software for Network Entrez.

The **toolbox** directory contains a set of software and data exchange specifications that are used by NCBI to produce portable software, and includes ASN.1 tools and specifications for molecular sequence data.

The **pub** directory offers public-domain software, such as BLAST (sequence similarity search program). Client software for Network BLAST and PowerBLAST is also included in this directory.

The **sequin** directory contains the new Sequin submission software for Mac, PC, and UNIX platforms.

Data in these directories can be transferred through the Internet by using the Anonymous FTP program. To connect, type: **ftp ncbi.nlm.nih.gov.** Enter **anonymous** as the login name, and enter your e-mail address as the password. Then change to the appropriate directory. For example, change to the repository directory (cd repository) to download specialized databases.

Malaria Genetics Web Site

The Malaria Genetics and Genomics Web resource provides data and information relevant to *Plasmodium falciparum*, the apicomplexan parasite responsible for the most serious form of human malaria. This is a joint service of NCBI and the Malaria Genetics Section of the National Institute of Allergy and Infectious Diseases (NIAID). The URL is http://www.ncbi.nlm.nih.gov/Malaria.

Resources available through this Web site include *P. falciparum* sequence databases, a specialized BLAST service, genome maps, linkage markers, epidemiologic and taxonomic data, and information about genetic studies and research projects in the NIAID Malaria Genetics Section. Links are also provided for other malaria-related Web sites and selected references.

BLAST Service for *P. falciparum*

The specialized BLAST service allows searching against *P. falciparum* sequences in GenBank, including Expressed Sequence Tags (ESTs) and Genome Sequence Tags (GSTs), as well as ESTs and GSTs for *Toxoplasma gondii*, a related apicomplexan parasite. In addition, as a collaborative effort, NCBI is compiling several BLAST databases corresponding to individual chromosomes currently being sequenced as part of the international *P. falciparum* genome sequencing effort. These chromosome-specific databases, listed below, are provided only for BLAST purposes and are not redistributed by NCBI.

Chromosome 1: Short Contigs	The Sanger Centre
Chromosome 2: Long Contigs	The Institute for Genomic Research
Chromosome 3: Long Contigs	The Sanger Centre
Chromosome 3: Short Contigs	The Sanger Centre
Chromosome 4: Short Contigs	The Sanger Centre
Chromosome 12: Long Contigs	Stanford DNA Sequence and Technology Center
Chromosome 14: Short Contigs	The Institute for Genomic Research

Much of the data is preliminary in nature and should be used as such. Links to the sequencing centers' Web sites are also provided.

Linkage Maps and Microsatellite Markers

The effort to develop the first genetic map of the *P. falciparum* genome is currently under way. The Malaria Genetics Web site provides mapping data that include sequenced markers, location, and segregation data for each of the 14 chromosomes. There is also a separate list for markers that have not been assigned to a chromosome. Graphic displays of the genetic maps are expected in the near future. Links to physical maps will be added as they become available.

Epidemiologic and Taxonomic Data

Information concerning the *P. falciparum* taxonomy is provided through the NCBI Taxonomy database, with links to Entrez for retrieving sequences. Epidemiologic data are illustrated on two world maps. Malaria is endemic in 91 countries, confined to the tropics with extensions into the subtropics in

Continued on page 8

News In Brief

❐ Updated NCBI Toolkit

A new version of the NCBI toolkit is on the NCBI FTP site at ftp://ncbi.nlm.nih.gov/toolbox/ncbi_tools/. This release includes source code for new client/server libraries, client software for Gapped BLAST version 2.0, Sequin software updates, and updates to the BLAST software (version 2.0.5).

The new client/server libraries are designed to run through HTTP instead of our custom dispatcher to connect the client to the service. This arrangement gives more reliable support and is now the default. By simply using the new clients, or linking with these new libraries, you will automatically use the new libraries and the HTTPD server at NCBI.

If you are operating behind a firewall and require a state-based connection, you should still take the new clients but configure to use the old connection functionality. Do this by adding the line "SRV_CONN_MODE=DISPATCHER" to the "[NET_SERV]" section of your NCBI configuration file.

❐ Batch Entrez

Downloading very large sets of sequence records from Entrez is possible through Batch Entrez, accessible from the main Entrez page. In June, NCBI added the capability to do text searches using the search statement format for regular Entrez. Previously, the service was limited to downloading an entire set of records for a given organism or downloading a set of records defined by a list of accession numbers or GI numbers.

❐ NCBI Exhibits at Conferences

A list of conferences where NCBI will be exhibiting is available from our home page. Under Welcome to NCBI, select **Upcoming Conferences**. For this fall, the list includes Human Genome Sequencing and Analysis in Miami, American Society for Tropical Medicine and Hygiene in San Juan, American Society for Human Genetics in Denver, and American Society for Cell Biology in San Francisco.

❐ New Sequin

Edit, annotate, and PowerBLAST your sequences with new Sequin 2.60. The Sequin home page at http://www.ncbi.nlm.nih.gov/Sequin/ describes the latest developments and has a new Frequently Asked Questions section and the most recent version of the help documentation. A new Sequin Quick Guide offers a step-by-step description of how to prepare your Sequin submission. Find out how Sequin can make it easy to submit multiple sequences as well as phylogenetic or population studies.

❐ Cn3D 2.0 Available

Version 2.0 of Cn3D, NCBI's molecular structure viewer, is now available. This release offers linked structure and sequence views to facilitate visualization and analysis of the VAST structural alignments. A sequence alignment function also allows the mapping of a protein sequence to a three-dimensional structure. Follow the **Structure** link from the NCBI home page for more information.

❐ Retrieve Server Replaced

The Retrieve e-mail server has been retired. The Query e-mail server (query@ncbi.nlm.nih.gov) is now the only NCBI system for text searches by e-mail. For a limited time, Query will continue to accept searches in the Retrieve server format.

The Query server provides access to the full suite of Entrez databases, including full MEDLINE via PubMed. It also offers a variety of output formats. For instructions, send the word HELP in a message to query@ncbi.nlm.nih.gov, and the documentation will be sent to you by return e-mail.

❐ Mouse Unigene

A Mouse UniGene collection is available and contains more than 8,600 clusters of sequences, each representing a unique gene. The murine set is expected to help expedite gene discovery. This service is included as part of NCBI's UniGene service.

❐ New Feature Table

Ever wonder what "misc_signal" or "specimen_voucher" means in the GenBank record? The Feature Table defined by the International Nucleotide Sequence Database Collaboration describes the format and definitions of the biological annotations used in the creation and maintenance of GenBank, DDBJ, and EMBL sequence records. Version 2.0 of the Feature Table is available though the Web at http://www.ncbi.nlm.nih.gov/collab/FT/index.html and via FTP at ftp://ncbi.nlm.nih.gov/genbank/docs/.

Complete Genomes, *continued from page 2*

which multiple mapping efforts are under way, such as human, the chromosomes are presented as co-ordinate map systems. These are integrated genetic and physical maps of higher eukaryotes for which only small sections of a chromosome have been sequenced.

The Genomes help documentation provides further information about the scope and features of that division and is accessible from the sidebar of the first Entrez Genomes page or by clicking on the question mark in the title bar of any subsequent page.

Unfinished Microbial Genomes

While Entrez contains schematics of genomes in progress based on sequence data deposited in GenBank, there is another category of genomes in progress that are not available in Entrez—unfinished microbial genomes whose sequences have not yet been formally deposited into GenBank. Those are accessible through a specialized BLAST Web page called **Microbial Genomes** (see article on page 1).

As genome sequencing and mapping continue to progress, NCBI will continue to be a comprehensive online source for public genome data. ■

Malaria, *continued from page 6*

Africa, Asia, and Latin America. Molecular data of *P. falciparum* are known for 18 countries. The DNA World View section provides information on more than 100 DNA sequences where the country of origin of the parasite is known. A second map gives a more detailed illustration of the worldwide spread of chloroquine-resistant strains of *P. falciparum* since the late 1950s. ■

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service, National Institutes of Health
National Library of Medicine
National Center for Biotechnology Information
Bldg. 38A, Room 8N-803
8600 Rockville Pike
Bethesda, Maryland 20894