NCBI News

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

March 1996

OMIM Database Now Available From NCBI

Users of NCBI's World Wide Web (WWW) service may have noticed a recent addition to the NCBI Home Page, entitled "OMIM." The Online Mendelian Inheritance in Man database, or OMIM^{TM}, is a continuously updated catalog of human genes and genetic disorders, and is considered a phenotypic companion to the human genome project. In December 1995, responsibility for providing public access to OMIM was transferred from the Genome Data Bank (GDB) at The Johns Hopkins University (JHU) to NCBI. It will continue to be authored and edited in the Center for Medical Genetics at JHU, under the direction of Dr. Victor McKusick, with computer support from NCBI.

Web Link

NCBI's Web service introduces new features that link OMIM entries to the Entrez integrated retrieval system, thereby providing easy access to related DNA and protein sequences as well as selected MEDLINE abstracts. In addition, NCBI's powerful "neighboring" feature for locating similar records has been applied to OMIM, allowing users to expand their searches at the touch of a button. OMIM's gene map, including clinical phenotypes, is also easily accessible and can be viewed in chart form, with direct links back to the full OMIM records.

Searching OMIM

The WWW address for the OMIM Home Page is "http://www3.ncbi.nlm.nih.gov/omim". It can also be accessed from NCBI's Home Page. To initiate a keyword search or to view the OMIM gene map, go to the OMIM Home Page, and select the option "Search the OMIM Database."

For a keyword search, enter your search term in the text box, select data fields to be searched, then click on the "Submit Search" button. A list of all OMIM entries matching your search question will be displayed. Click on the OMIM identification number to see the full entry, which begins with a table of contents and list of available database links. For example, the entry for Hunting-

Continued on page 2

Exploring Network Entrez: Graphical Views of the Genomes Division

In a major new release, several features have been added to Network Entrez, including a graphical sequence viewer that presents schematics of entire chromosomes, genomes, and associated map information available through the new Genomes Division of GenBank (see September 1995 *NCBI News*). The Genomes Division currently contains more than 100 entries, including information on completely sequenced genomes and chromosomes such as *Haemophilus influenzae* and yeast, as well as contiged sequence islands from higher eukaryotes. Network Entrez presents chromosome level views of this data, then allows a chromosome subregion to be selected and enlarged to show a detailed graphical view of biological features annotated on that region.

The following step-by-step tutorial explores the use of Entrez's graphic display functions for Genomes Division searches. In this example, we will search for the location of the gene BRCA1 in the human genome. Starting with a chromosome level map view, we will use the "zoom" function to obtain an enlarged view of the specific chromosomal region

Continued on page 6

In This Issue

NCBI News

NCBI News is distributed 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

Design Consultant
Troy M. Hill

Photography
Karlton Jackson

Editing, Graphics, and Production
Veronica Johnson
Deborah Loer-Martin
Wendy B. Osborne

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

The contents of this newsletter may be reprinted without permission. The mention of trade names, commercial products, or organizations does not imply endorsement by NCBI, NIH, or the U.S. Government.

NIH Publication No. 96-3272

ISSN 1060-8788

***OMIM**, continued from page 1*

ton disease has a link to the OMIM gene map, plus 126 MEDLINE records, 2 protein sequence records, 4 DNA sequence records, 1 GDB record, and 1 NCBI UniGene location. The full OMIM records contain sections on clinical features, case studies, inheritance, animal models, and more. Within the text are links to extensive references, most of which are further linked to MEDLINE abstracts.

To search the gene map, scroll down the query page to the section “View the OMIM Gene Map,” enter a gene symbol or chromosomal location in the text box, then click the “Submit Search” button. A chart containing chromosomal location, descriptive title, associated disorders, and corresponding OMIM record numbers will be presented. To view the related OMIM text document, just click on the highlighted OMIM record numbers.

Other Access Options

OMIM is also accessible by e-mail through the NCBI Retrieve server (retrieve@ncbi.nlm.nih.gov). For information about using the Retrieve server, send the word HELP in the body of a message to the server address. The documentation will be sent to you automatically. If you have been a past user of OMIM through the IRX search system at GDB, you can obtain a new account and password to be used at NCBI. This approach is only recommended, however, for users who are unable to access the WWW version but do have telnet capability. If you are interested in downloading a full copy of the OMIM database for local use, it is available on NCBI’s Anonymous FTP site (ncbi.nlm.nih.gov) in the “repository/OMIM” directory.

For additional information, please phone (301) 496-2475 or send e-mail to info@ncbi.nlm.nih.gov. ■

❖ ❖ ❖ ❖ ❖

Sequin Pre-Release Available for Testing

NCBI is now making available a beta-test version of Sequin, a new stand-alone software tool for submitting and updating GenBank entries. Intended as an alternative to Authorin, Sequin is designed to simplify the sequence submission process, provide graphical viewing and editing options, and provide increased data handling capabilities. New data handling features include the ability to submit segmented entries, annotate protein features, perform data integrity checks, and accommodate bulk submissions. The graphical viewing function offers an inviting and easy-to-use interface.

Sequin also produces submission files suitable for use by the international collaborating databases (EMBL and DDBJ), and is currently being reviewed and tested by them to ensure full compatibility.

NCBI invites submitters to participate in testing Sequin and encourages feedback and comments. Although a beta-test version, the Sequin pre-release is a full-featured submission tool that you can use now for actual GenBank submissions. The beta-test version is available for Macintosh, PC/Windows, UNIX, and VMS computers and is available by Anonymous FTP from ncbi.nlm.nih.gov in the “sequin” directory. ■

Using Entrez in Secure Environments

Many users have expressed interest in using Network Entrez through institutional security systems, such as a firewall. A firewall restricts an internal network's contact with outside networks to varying degrees. This article describes some of the technical issues and solutions for implementing Network Entrez in a secure environment. You may need the cooperation and assistance of your local systems administrator to implement these solutions.

Operating Through a Firewall

NCBI has recently enhanced Network Entrez (versions 4.013 and higher) to allow users to configure it for use through a firewall. In the default mode of a client-server application such as Network Entrez, the server connects to the client. However, it is possible to reverse the direction of the client-server connection, causing the client to connect to the server. This arrangement is usually much more favorable to firewalls.

When configuring Network Entrez, you can override the default direction and reverse the connection in one of two ways: (1) select the "Outgoing connections only" checkbox on the second screen of the "netentcf" configuration program; or (2) manually set DIRECT_SVC_CON=TRUE in the [NET_SERV] section of the NCBI configuration file on your system. In addition, in the "netentcf" configuration program, you will still need to establish the two TCP/IP connections to NCBI and the Network Entrez server. To confirm that the default direction has been reversed, open Network Entrez and select the "More" button under "About Entrez." You should see the phrase "Using outgoing connection when communicating with server."

If neither of the two approaches successfully enables Network Entrez to operate through your institutional firewall, contact NCBI for information on additional technical solutions currently under development. You may also wish to compile Network Entrez for yourself from source code if you have an institutional firewall that requires special treatment. Source code is available in the NCBI toolbox (ftp ncbi.nlm.nih.gov/toolbox/ncbi_tools).

Using SOCKS Protocol

NCBI intends to continue supporting SOCKS capability, but only for the Unix platforms. SOCKS is a well-defined protocol that permits client-server applications to transparently and securely traverse firewalls by providing an additional layer between the application and transport layers. There are no plans to add any SOCKS capability for non-UNIX platforms. But note that on Windows platforms it is possible, in principle, for a SOCKS-friendly WinSock implementation to provide SOCKS-based Network Entrez access when used in combination with the DIRECT_SVC_CON mode described above.

Using Data Encryption

Another security option is to build an encrypted version of Network Entrez. You will need to build the client from scratch by using the source code from the NCBI toolbox and following directions provided in the file 'network/encrypt/README'. Note that the RSA and DES encryption technology may not be exported outside the United States and Canada, and therefore we are unable to offer encryption-enabled binaries for FTP. ■

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", and "pub".

The **repository** directory makes a number of molecular biology databases available to the scientific community. This directory includes databases such as PIR 47.00, Swiss-Prot, CarbBank, AceDB, and FlyBase.

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 Entrez executable programs for accessing CD-ROM data on a variety of platforms. It also contains 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), MACAW (multiple sequence alignment program), and Authorin submission software for Mac and PC systems. Client software for Network BLAST is also included in this directory.

Data in these directories can be transferred through the Internet by using the Anonymous FTP program. To connect, type: **ftp ncbi.nlm.nih.gov** or **ftp 130.14.25.1**. Enter **anonymous** for 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.

Selected Recent Publications by NCBI Staff

Ahmad, N, BM Baroudy, RC Baker, and **C Chappey**. Genetic analysis of immunodeficiency virus type 1 envelope V3 region isolates from mother and infant isolates after perinatal transmission. *J Virol* 69(2):1001–12, 1995.

Bassett DE, **MS Boguski**, F Spencer, R Reeves, M Goebl, and P Hieter. Comparative genomics, genome cross-referencing and XREFdb. *Trends Genet* 11:372–3, 1995.

Baxevanis, **AD**, and **D Landsman**. Histone sequence database: a compilation of highly-conserved nucleoprotein sequences. *Nucleic Acids Res* 24(1):245–7, 1995.

Gunderson, J, G Hinkle, **D Leipe**, HG Morrison, SK Stickel, DA Odelson, JA Breznak,TA Nerad, M Mueller, and ML Sogin. Phylogeny of trichomonads inferred from small subunit rRNA sequences. *J Eukaryot Microbiol* 42:411–5, 1995.

Koonin, EV, **RL Tatusov**, and **KE Rudd**. Sequence similarity analysis of *Escherichia coli* proteins—functional and evolutionary implications. *Proc Natl Acad Sci USA* 92:11921–5, 1995.

Kulaeva, OI, **JC Wootton**, AS Levine, and R Woodgate. Characterization of the umu-complementing operon from R391. *J Bacteriol* 177:2737–43, 1995.

Liu, JS, **AF Neuwald**, and **CE Lawrence.** Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J Am Stat Assoc* 90:1156–70, 1995.

Madei, T, **MS Boguski**, and **SH Bryant**. Threading analysis suggests that the obese gene product may be a helical cytokine. *FEBS Lett* 373:13–8, 1995.

Mushegian, **AR**, and RJ Shepherd. Genetic elements of plant viruses as tools for genetic engineering. *Microbiol Rev* 59:548–78,1995.

Neuwald, AF, and P Green. Detecting patterns in protein sequences. *J Mol Biol* 239:698–712, 1995.

Wu, SC, **JL Spouge**, SL Conley, WP Tsai, and PL Na. Human plasma enhances the infectivity of primary human immunodeficiency virus type 1 isolates in peripheral blood mononuclear cells and monocyte-derived macrophages. *J Virol* 69:6054–62, 1995.

BLAST Service Update

Several changes in the BLAST service went into effect on March 11. These include reorganization of databases, new client software, changes in sequence identifier format, and new databases available by FTP.

Database Reorganization

The BLAST databases have been reorganized for more efficient searching and better synchronization with Entrez. In order to give users more control over their BLAST searches, the EST and STS sequences have been eliminated from the “nr” (nonredundant) database, and now exist as separate data sets that can be selected for searching. In addition, a new database called “month” contains new or revised sequences released in the last 30 days, providing a rolling 30-day window of the newest sequences.

Sequence Identifiers

The one-line description used in BLAST output for GenBank conceptual translations has changed with respect to the sequence identifier number. In

Continued on page 8

Entrez CD-ROM To Be Discontinued

After careful consideration, NCBI has decided to discontinue the Entrez CD-ROM effective August 15, 1996, with Release 24. Due to the continuing rapid growth of GenBank and other sequence databases, the five-disc version of Entrez has become increasingly inconvenient to use. For some time now, the CD-ROM version has lagged behind the two Internet versions, Network Entrez and World Wide Web Entrez, in several aspects. Differences include update frequency; number of MEDLINE citations available; incorporation of new data sets such as genetic maps and 3D structures; and links to online journals. The Internet versions are updated daily, while the CD-ROM version is only updated every other month. The considerable advantages of the Internet versions of Entrez have resulted in a sustained increase in use, with simultaneous decline in CD-ROM subscriptions, from a peak of more than 2,200 subscribers to less than 1,400 today.

Impact on domestic and non-U.S. CD-ROM users was assessed through a survey conducted by NCBI in December 1994. Results showed that more than 70 percent of U.S. users and more than 50 percent of non-U.S. subscribers were willing to switch to an Internet version. An additional 20 percent indicated they would switch to Internet alternatives as the number of discs per Entrez release approached five. Entrez now uses five discs and the August 1996 release is expected to require six or more.

Access to the Internet versions of Entrez is simple. Many users have free high-speed Internet access through their academic institutions or companies, and dial-up access is available almost anywhere in the United States for $15–$25 per month. If you have a World Wide Web browser such as Netscape or Mosaic, simply point your browser to “http://www.ncbi.nlm.nih.gov/.” The Web version of Entrez has all the capabilities of the CD-ROM version, but

Continued on page 7

Frequently Asked Questions

I have noticed a new NID (nucleotide id) field in GenBank. What does this number indicate?

The NID field is used specifically for NCBI's nucleotide sequence identifier, called an "NID gi." Previously, this number was included in the COMMENT field, but beginning with Release 94.0, it will appear only in the new NID field.

An NID gi number is assigned to every nucleotide sequence in GenBank. A new number is issued whenever a revision is made in the DNA sequence data. In this way it differs from the accession number, which is a stable identifier that does not change when modifications are made to the record.

I was making a submission using BankIt and my system crashed. Can I get a copy of the data I have prepared so far?

If you clicked the BankIt button at least once when you were entering your data, the information was transmitted to our logs and a BankIt processing number was assigned to your submission. In this case, we can retrieve a copy of your file in HTML format and send it to you by e-mail. You can then load it into your WWW browser and continue your submission.

I have a number of submissions to make and would like my accession numbers in consecutive order. Is this possible?

If you are using BankIt, the following prompt appears at the beginning of the BankIt form: "This submission is number __ of a set of __ submission(s)." Indicate the number of sequences you intend to submit, and the GenBank staff will wait for the entire batch before assigning a block of accession numbers.

If you are using Authorin to submit by e-mail, include a message at the beginning of your submission stating how many .sbt files you are sending, and indicate that you would like sequential accession numbers. If you are submitting more than 20 sequences, please divide up the submission into separate messages containing no more than 20 files each. This will facilitate our processing, and you will still receive sequential accession numbers.

What is the difference between dbEST and the EST Division of GenBank? Are the sequence records different?

The sequences and accession numbers in dbEST are identical to those in the EST Division of GenBank. The dbEST database, however, contains additional annotation not found in GenBank, such as information on the clone library, sequencing method, map location if known, sources for obtaining the physical clone, and results of BLAST sequence similarity searches.

Using OMIM, if I know the chromosome location of a disorder, is it possible to see diseases and disorders in this same region?

Yes, this is possible by using the Gene Map section of OMIM. From the OMIM Home Page, select "Search the OMIM Database," then scroll down to the section "View the OMIM Gene Map" on the search page. Enter the chromosome location, and click on "View." A table including map location, gene symbols, OMIM number, and associated disorder information will be presented.

Once I've located an OMIM record I am interested in, how do I find out if there is a known map location?

Each OMIM record starts with a Table of Contents, under which is an array of buttons indicating cross-database links. The "Gene Map" button indicates that mapping information is available, and the location will also be displayed under the line of buttons. To see the Gene Map, click on the button.

***Graphical Views**, continued from page 1*

containing BRCA1, then link to the corresponding GenBank record.

1. To access the Genomes Division, begin at the initial **Query** window of Network Entrez, and select **Genome** in the Database menu.

2. To search for the BRCA1 gene, select **Gene Symbol** in the Field menu, and **Selection** in the Mode menu. Then enter “BRCA1” (without quotes) in the **Term Entry Box** and press return. It should now appear alphabetically in the **Term Selection Box**. Double click on BRCA1 to include it in the **Query Refinement Box**.

3. Click on the **Retrieve** button to obtain the record for the chromosome containing the BRCA1, in this case *Homo sapiens* Chromosome 17.

4. Double click on the icon for Chromosome 17 to see the “map view” of this chromosome, shown below as Figure 1. (Tip: Enlarge the window for easier viewing.) The **Map** view is the default display mode in the Genomes Division. **Graphic** and **Alignment** views, shown as tabs at the top of the screen, are used to provide more detailed presentations of selected map regions.

As is the case with most higher eukaryotes, only small parts of human Chromosome 17 have been sequenced. For partially sequenced chromosomes, NCBI has collected several genetic and physical maps, placed them onto a common coordinate system, and aligned any shared markers (shown in Entrez by green connecting lines). In this example, the **Map** view shows the alignment of the MIT physical map, the NCBI transcript map, the CHLC linkage map, the Genethon linkage map, and the GDB cytogenetic map. Note that Stanford radiation hybrid maps will also be added as they become available; currently the Chromosome 4 map is in Entrez.

5. To find BRCA1 on the map, click on the **Find by Gene or Product Name** button, which will bring up a window with an alphabetical list of available markers located on Chromosome 17. This is a composite list of all markers that are included on at least one of the chromosome 17 maps used in Entrez. Enter BRCA1 in the text box (see Figure 1 inset) and click on the **Accept** button.

6. The **Map** view will be redrawn, and BRCA1 will now appear in red letters above the maps on which the marker is present. BRCA1 appears above the NCBI transcript map and the GDB cytogenetic map (see Figure 1).

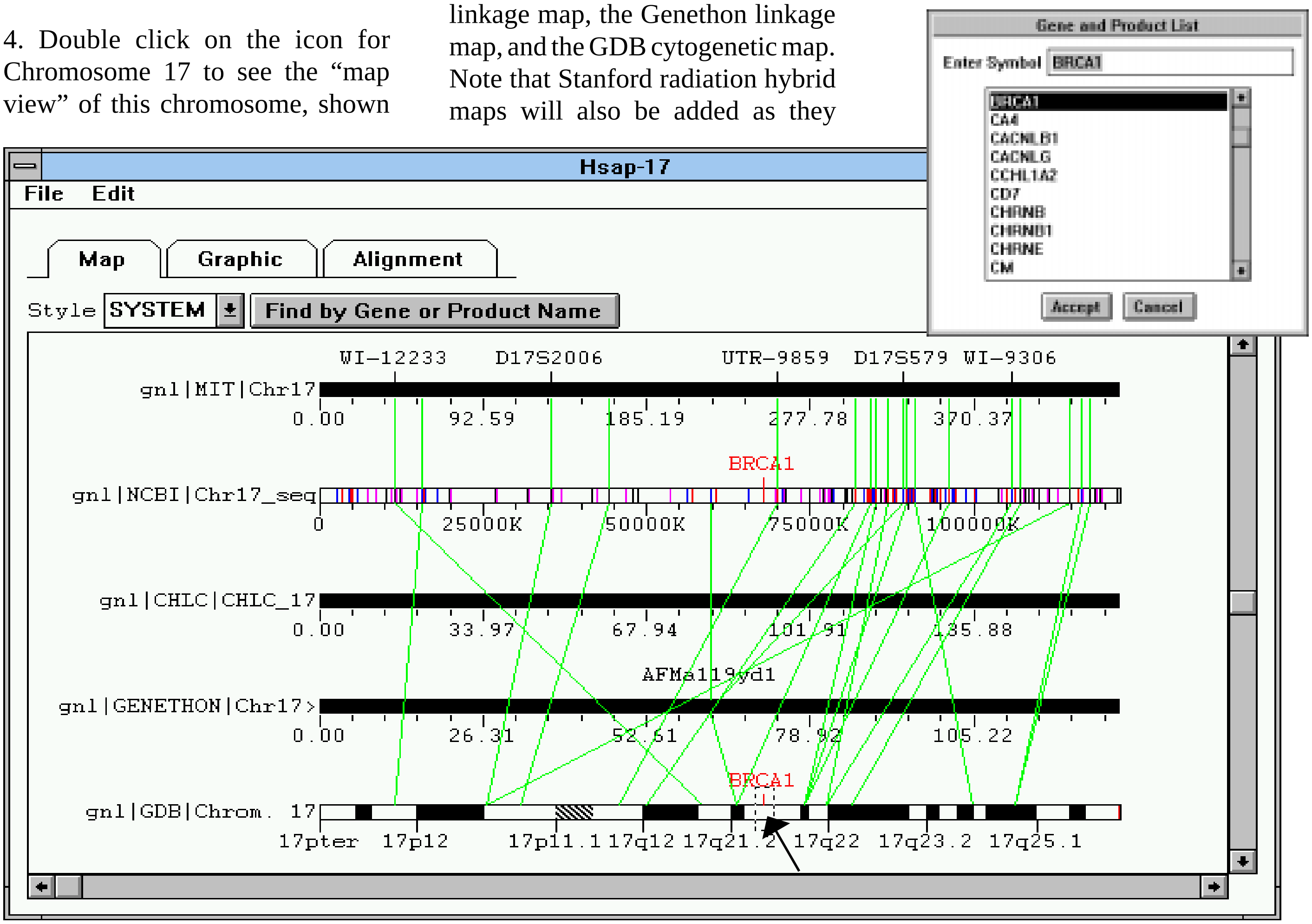
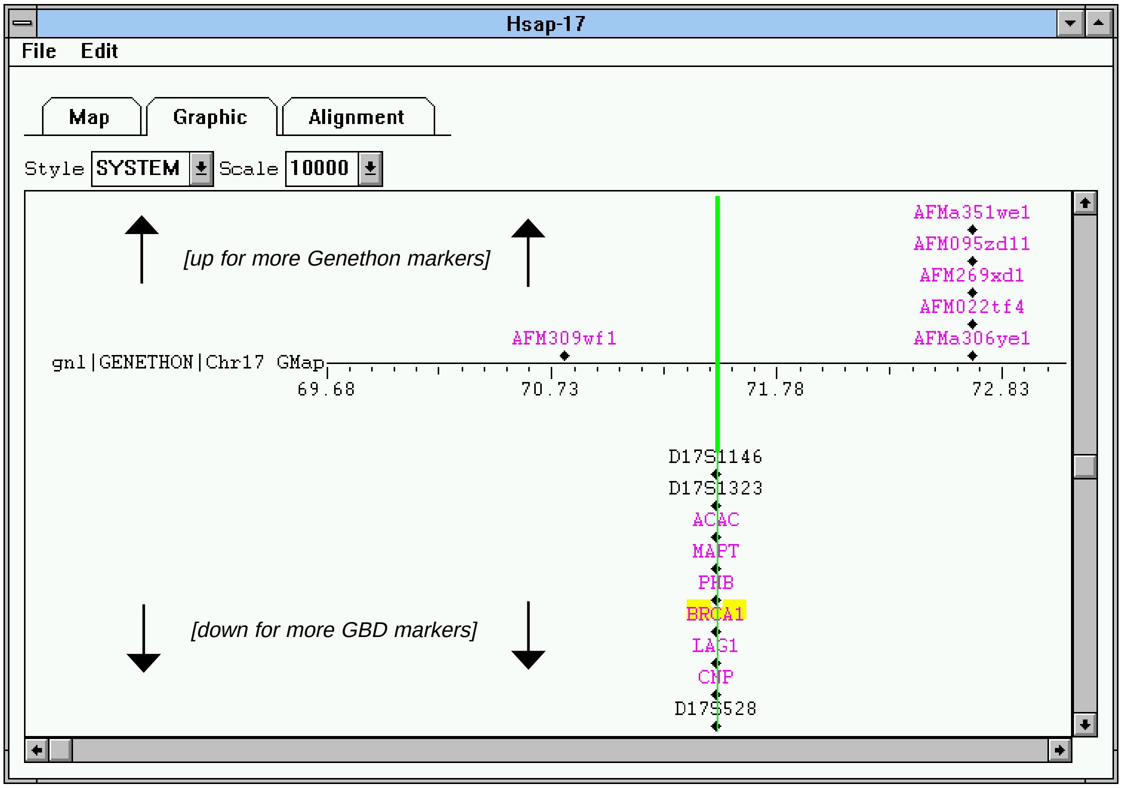


Figure 1. Map view

7. To obtain a detailed view of the subregion containing BRCA1, we will use the graphic viewer "zoom" function. First, select one of the BRCA1 regions using a mouse technique called "rubber banding." That is, drag your cursor over the desired area while depressing the mouse button (see arrow on Figure 1). Then zoom in on the selected area by clicking on the **Graphic** tab. The **Graphic** view (Figure 2) shows detailed features of the selected region and the positions of GenBank sequences that align to it.



Style SYSTEM Scale 10000

Figure 2. Graphic view

8. Scroll down to the GDB cytogenetic map (just past the Genethon map) and you will see BRCA1 in pink lettering with yellow highlighting, as well as a list of the other markers that have been mapped to the same region. All the markers appearing in pink lettering have links to corresponding DNA sequence, protein sequence, or MEDLINE records (but MEDLINE links are not yet implemented). Double click on BRCA1 and the corresponding GenBank record will appear.

This example illustrates just a few of the many features available with Entrez's new graphical viewing functions, and focuses only on locating a gene within an incompletely sequenced chromosome. Note that, for completely sequenced genomes and chromosomes, **Map** and **Graphic** view presentation details will differ from this example. However, the ability to go to successively more detailed displays, using the **Map**, **Graphic**, and **Alignment** views, is a basic feature of the Genomes Division in Entrez. The **Alignment** view, not discussed in the BRCA1 example, shows the relationship of sequences aligned at base pair resolution in cases where multiple database sequences map to the same region.

Network Entrez users are encouraged to explore these new features and contact NCBI with any questions or suggestions. The new release of Network Entrez is located in the "entrez/network" directory on NCBI's Anonymous FTP site (ncbi.nlm.nih.gov). The graphical viewing functions have not yet been implemented in the World Wide Web version of Entrez. ■

***Entrez**, continued from page 4*

with the visual style of the World Wide Web. If you prefer the "look and feel" of the CD-ROM version, you may download Network Entrez from the NCBI's Anonymous FTP site (ncbi.nlm.nih.gov). Network Entrez offers all the advantages of Web Entrez. Versions are available for PC/Windows, Macintosh, and several Unix workstations in the "entrez/network" directory.

The last release of Entrez on CD-ROM will be on August 15, 1996. If you choose to renew your subscription between now and then, the Government Printing Office (GPO) will charge you the full six-issue cost. You must request a pro-rated refund from the GPO for any amount that will be due to you after the final release. ■

***BLAST**, continued from page 4*

order to identify each specific protein sequence, NCBI is now assigning a stable identifier called a "gi" number for all sequences, both protein and nucleotide. The "gi" is a unique integer that changes whenever the sequence changes, but not when the features or references of an entry are updated.

The new format for protein sequence one-line descriptions starts with the label "gi," followed by the gi number, the accession number in parens, and the text description, for example: "gi|451623 (U04987) env gene product [Simian immunodef...]."

New Blast2 Client

For users of Network BLAST, a new client called Blast2 has been introduced that provides a better interface for postprocessing of search results. Blast2 represents the foundation for NCBI's future development of the BLAST service. Although the older client, known as "Experimental Blast," will operate with the new database organization, users are encouraged to upgrade to Blast2, which is available on the FTP site (ncbi.nlm.nih.gov) in the "blast/network/blast2" directory.

Performance Improvements

Many users have noticed that, over the past several months, BLAST searches have taken much longer to complete. This is due to an increase in the number of searches and in the size of the database. We are taking several steps to improve performance in addition to the reorganization of the databases. First, we will be upgrading the processors in the BLAST servers within the next 6 weeks, which should double the speed of typical jobs. Second, the operating system and queuing software are being tuned for greater efficiency. And finally, algorithm improvements to BLAST are being tested and are expected to yield up to fivefold performance gains for protein searches. Gradual improvements in search times over the next several months are expected.

Databases on FTP Site

For those users who wish to implement the BLAST search engines in-house, all of the databases, including the nonredundant DNA and protein sequence databases, are available as FASTA files from the NCBI FTP site (ncbi.nlm.nih.gov) in the "blast/db" directory. ■

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