NCBI News

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

Fall/Winter 2000

The Human Genome Sequence: NCBI's First Annotated Edition

The NCBI recently released its first assembled and annotated view of the human genome sequence. The assembly is based not only on the finished and draft sequence deposited in GenBank by the public sequencing centers, but also on the thousands of sequences contributed to GenBank over the years by individual scientists around the world. Hence, this resource represents a true international public effort to sequence the human genome.

Updated assemblies—incorporating new data, filling in existing gaps and increasing overall accuracy—will be released to the public on a regular basis. The human genome data can be viewed on the Web with NCBI's human genome Map Viewer or downloaded in bulk via FTP.

Assembly

NCBI's assembly process starts with the entire complement of human genomic sequence in GenBank, both draft and finished. Assembling and ordering the individual sequence units is a critical phase of the Human Genome Project. It involves many different steps, including screening for vector and other sequence contamination, before merging the input data into ordered segments of DNA referred to as contigs. This first build presents more than 6,000 contigs, representing roughly 2.8 billion base pairs. Nearly 700 contigs are longer than 1 MB. Over 75 percent of the bases in the contigs are in unbroken segments of greater than 30Kb, the size of a typical human gene.

Annotation

NCBI is also engaged in the essential process of annotating, or labeling the biologically important areas, of the human genomic sequence. Human gene annotation falls into two major tasks: the correct placement of known human genes into their proper genomic context; and the prediction of new, previously unknown genes, from the genomic sequence.

For the first task, the mRNAs from the NCBI RefSeq collection are placed on the genome primarily by alignment, with compensation

continued on page 3

BLink Enhances Entrez Exploration of Protein Similarities

A popular feature of the Entrez search system is the ability to view similar sequences with a click of the Related Sequences button. A new service called BLAST Link (or BLink) transforms this feature from a simple one-dimensional listing of similarities into a panoramic display of graphical alignments, editable taxonomic trees, conserved protein classes, protein domains, and 3D structures.

Reached through **BLink** links at the head of protein records, the

continued on page 4

In this issue

NCBI News

NCBI News is distributed four times a year. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence 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

Contributors
Colleen Broder
Donna Maglott
Scott McGinnis
Jim Ostell

Writer
David Wheeler

Editing, Graphics, and Production
Marla Fogelman
Jennifer Vyskocil

Design Consultants
Tim Cripps
Gary Mosteller

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 perform research in computational molecular biology.

NIH Publication No. 01-3272

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

Standardized Human Gene Nomenclature

Independent discovery of the same gene by more than one group of investigators is not uncommon, resulting in different names for a single gene. In order to prevent nomenclature chaos, the information arising from these independent discoveries must be standardized.

The use of systematic, consistent and hierarchical terminology not only provides a foundation for generating integrated databases and linking to related resources, but also ensures efficiency and accuracy when retrieving information. Some scientific journals now insist upon an approved gene nomenclature prior to publication of an article, a reflection of the benefit that can be gained by establishing and adopting standardized terminology.

Human Gene Nomenclature Committee

For over twenty years, the international Human Gene Nomenclature Committee (HGNC) has been recognized as the authority for assigning human gene names and symbols. The HGNC is authorized by the Human Genome Organization (HUGO) and is funded in part by several U.S. governmental agencies, including the National Institutes of Health. HGNC objectives include the promotion of universal acceptance and utilization of standardized nomenclature and coordination of terminology across species, particularly among mammals. HGNC consults with other nomenclature committees, expert groups, the international scientific community, and individual researchers when considering nomenclature schemes to avoid duplication and to indicate evolutionary relationships. Names are also chosen to reflect normal gene function and sequence relationships between genes.

The HGNC database contains over 11,700 unique symbols for human genes, with an expected 7 to 15-fold increase over the next few years. New genes requiring nomenclature are identified through three principal mechanisms: direct queries from investigators; queries from other databases and collaborating curation groups; and literature scans of major scientific journals. This last mechanism is complemented by the large-scale literature scan carried out by the Online Mendelian Inheritance in Man (OMIM) database. Public access to OMIM, a catalog of human genes and genetic diseases, is available from the NCBI Web site. The HGNC also provides access to nomenclature guidelines, an online submission form, and a nomenclature database.

After identifying a gene that requires naming, HGNC assigns a unique identification number that links directly to descriptive information and a nucleotide sequence, if available. The descriptors used to differentiate genes include the MIM number, literature citations, one or more sequence accession numbers, and cytogenetic data. This information is obtained via various databases maintained

continued on back page

Human Genome Sequence

continued from page 1

for various problems in both the genomic and mRNA sequences, and reconciliation of close paralogs and pseudogenes. In this first release on the NCBI Web site, 8,800 of the 10,500 RefSeq mRNAs were placed on the genome.

For the second task, multiple lines of evidence including EST alignments, splice junctions, protein similarities, and other methods are combined to predict new genes.

Model Sequences Get New Accession Numbers

The NCBI assembly process produces a new kind of sequence record termed a "model sequence." Model mRNA records are created *de novo* from human genomic sequence, and aligned to mRNA reference sequences from RefSeq. Since such alignments may contain some mismatches, model sequences are assigned their own accession numbers, in the format XM_12345 for mRNA and XP_12345 for the corresponding model protein sequence.

The alignment-based evidence for the model sequences is provided through AceView, a new service currently accessed from LocusLink and the Map Viewer. AceView shows a predicted gene, its intron/exon structure, and its alignment to the corresponding RefSeq mRNA sequence.

The predicted mRNAs and proteins will be subject to change with improved data and better algorithms. Nonetheless, NCBI will do its best to keep the same accession numbers with the same predicted genes from build to build. A new release containing both known gene placements and predicted gene models was in process as this article went to press.

Additional biological features are also being annotated on the genomic sequence. This first release includes more than 1.3 million SNPs and 111,851 STS markers.

Public Access

NCBI's human genome Map Viewer may be used to view the contigs used to assemble the sequence by selecting **Contig** map. SNP data may be viewed on the **SNP** map. The Map Viewer may be used to further explore the human genome data by viewing up to 7 parallel maps selected from a pallet of nineteen—including 6 sequence maps, 5 cytogenetic maps, 2 genetic maps, and 6 radiation hybrid maps.

The data is also available for downloading from the "genomes/H_sapiens" directory of the NCBI FTP site.

The FTP site includes the contigs produced by the NCBI assembly, RefSeq and model mRNA sequences annotated on the genome, and information used by the Map Viewer to generate and display the palette of nineteen maps mentioned above.

—DW, CB, JO

What is Draft Sequence?

Two-thirds of the human genomic sequence in GenBank is termed "draft" or "unfinished." These sequences can be comprised of many unordered pieces and are of lower quality than a typical "finished" GenBank sequence. The finishing process involves closure of sequence gaps, determination of proper order and orientation, and resolution of any sequencing ambiguities and errors. This is an ongoing process in the sequencing centers of the Human Genome Project, and NCBI updates draft sequence on a daily basis.

Draft sequence is placed in the HTG (High Throughput Genomic) division of GenBank. A typical HTG record consists of all sequence data generated from a single cosmid, BAC, YAC, or P1 clone. A single accession number is assigned to this collection of HTG sequences. Each record includes a clear indication of its status—Phase 1 or Phase 2—and a prominent warning that the sequence data is "unfinished" and may contain errors. Phase 1 indicates an unfinished sequence with gaps and unknown order and orientation of the pieces. In Phase 2, the order and orientation of the pieces is known, but the length of the gaps may still be unknown. Finished sequence data, consisting of one continuous piece of high-quality DNA sequence, is moved out of the HTG division and placed in the Mammalian division of GenBank. Contigs from the NCBI human genome assembly contain finished as well as draft sequence.

BLink

continued from page 1

BLink summary page for a given protein sequence is centered around color-coded, graphical BLAST alignments to other proteins. This alignment view can be limited by taxonomic class or by BLAST score. Alternatives to the graphical alignment view include the standard BLAST taxonomy report and a **Common Tree** view. The Common Tree view allows a phylogenetic tree of the organisms represented in the alignments to be collapsed or expanded, and serves as an analysis aid.

Protein sequence similarities to structurally determined proteins can be shown through the 3D structures option, allowing for a smooth migration from 2D sequences to 3D structures or possible structural templates. The BLink summary page also offers the results of a CDD search against NCBI's PFAM and SMART-derived library of protein domains. Links to the COGs database of protein groups, when applicable, allow access to information on the putative function and degree of conservation of similar proteins among 21 complete genomes.

BLink can also generate an Entrez report of all the proteins included in the BLAST alignment. In this manner, the protein neighbors of any protein in Entrez can be downloaded in a variety of formats for more extensive analysis.

A sample BLink report for the 54Kd human signal recognition particle

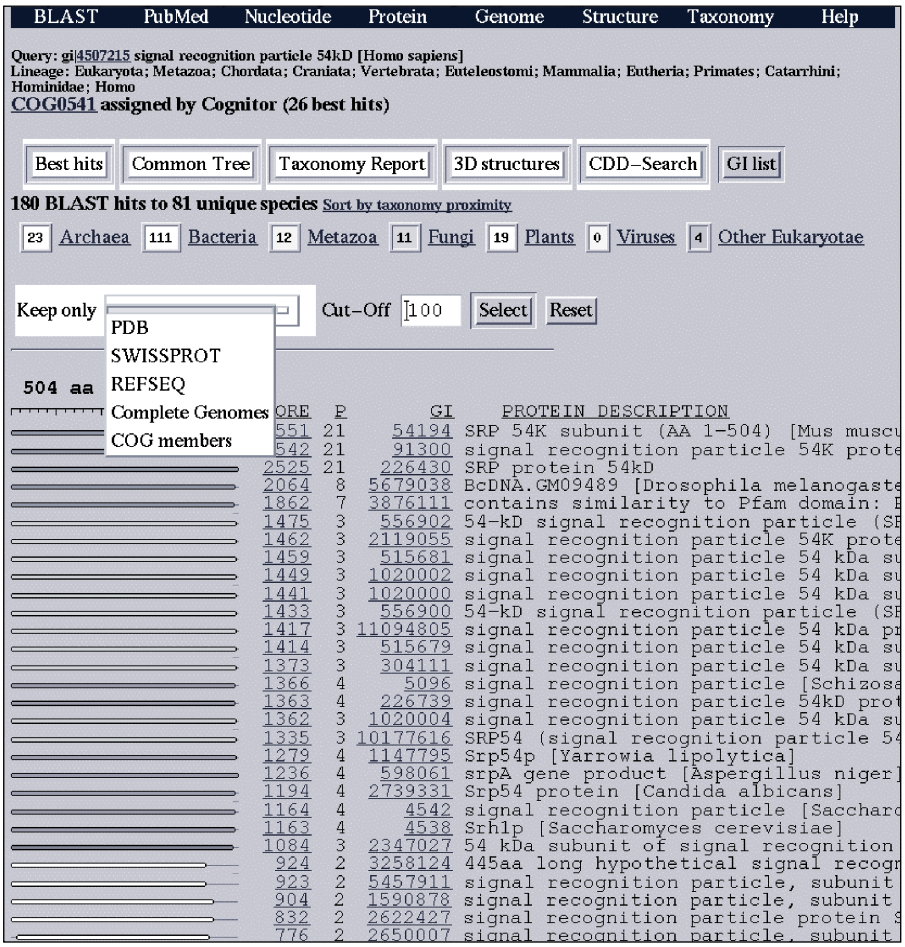

Figure 1: *BLink report for the human 54Kd signal recognition particle protein.*

(SRP) is given in Figure 1. SRP is an essential cytoplasmic ribonucleoprotein complex that targets secretory proteins to the inner cell membrane of bacteria or the endoplasmic reticulum membrane of eukaryotes. SRPs are well-conserved and are represented in all kingdoms of life. The BLink report clearly indicates that 180 proteins similar to the 54Kd SRP protein are found in 81 different species from the archaea, bacteria, metazoa, fungi, plant, and other eukaryotic groups. By clicking on a taxonomic class, the graphical BLAST alignments can be tailored to include only those proteins from the particular taxa chosen. The alignment list can be restricted to sequences associated with Protein DataBank (PDB) 3D structures, SwissProt or RefSeq records, proteins from complete genomes, or COG members. Alignments can also be limited by BLAST score using a user-defined cutoff. The BLink report for any protein within an alignment can be reached by clicking on an individual gi number.

Access to BLink reports is also available from Entrez Genomes, AceView, SequenceView, and the COG pages. Links will soon appear as part of the protein BLAST output. For more detailed information, click on Help found at the top of any BLink report.

Q & A

Frequently Asked Questions

Q.

A.

How can I determine what general class of protein I have from its sequence?

Aside from running a standard BLAST search, NCBI offers two ways to do this. You can perform a Conserved Domain Database Search using your protein sequence. From the Structure page (www.ncbi.nlm.nih.gov/Structure), select CDD.

You can also perform a search against the Clusters of Orthologous Groups (COGs), a database containing 21 complete proteomes. From the COGs page (www.ncbi.nlm.nih.gov/COG), select COGnitor.

How can I see all the LocusLink entries for human chromosome 22?

From the NCBI home page, select the human **Map Viewer**, then select chromosome 22. In the sidebar of the chromosome 22 map, click on the **Chr.22 Resource** link, which leads to a summary page with additional information about the chromosome. In the sidebar of the resource page, click on **LocusLink Chr.22** Loci.

How can I see all the UniGene Clusters mapping to chromosome 22?

From the chromosome 22 resource page mentioned in the previous question, click on the **UniGene Chr.22 Clusters** link.

How can I set up a search using the PubMed "Cubby" so that I can retrieve only papers published since my last search?

You first need to register as a Cubby user. From PubMed, click on **Cubby** in the sidebar and follow the onscreen instructions.

To store a search, first run it as usual within PubMed, then click the Cubby link from the sidebar, and login to Cubby as prompted. Your current search will be displayed; press the **Store in Cubby** button to add it to your stored search list. To collect any new material, select the search of interest and press the **What's New for Selected** button.

What classes of genetic variation are included in the dbSNP database?

The database accepts several classes of genetic variation, including SNPs, microsatellite repeats, and small insertion/deletion polymorphisms. The scope of dbSNP includes disease-causing clinical mutations as well as neutral polymorphisms.

Selected Recent Publications by NCBI Staff

Aravind, L, and **EV Koonin**. Eukaryote-specific domains in translation initiation factors: implications for translation regulation and evolution of the translation system. *Genome Res* 10(8):1172-84, 2000.

Bauer, H, H Mayer, **A Marchler-Bauer,** U Salzer, and R Prohaska. Characterization of p40/GPR69A as a peripheral membrane protein related to the lantibiotic synthetase component C. *Biochem Biophys Res Commun* 275(1):69-74, 2000.

Beja, O, **L Aravind, EV Koonin,** MT Suzuki, A Hadd, LP Nguyen, SB Jovanovich, CM Gates, RA Feldman, JL Spudich, EN Spudich, and EF DeLong. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289(5486):1902-6, 2000.

Kim, W, and **WJ Wilbur.** Corpus based statistical screening for phrase identification. *J Am Med Inform Assoc* 7(5):499-511, 2000.

Lash, AE, CM Tolstoshev, L Wagner, GD Schuler, RL Strausberg, GJ Riggins, and **SF Altschul.** SAGEmap: a public gene expression resource. *Genome Res* 10(7):1051-60, 2000.

Li, CS, FM Hueber, and **CL Hotton.** A neotype for *Drepanophycus spinaeformis* Göppert 1852. *Can J Bot* 78(7):889-902, 2000.

Natale, DA, CJ Li, WH Sun, and ML DePamphilis. Selective instability of Orc1 protein accounts for the absence of functional origin recognition complexes during the M-G(1) transition in mammals. *EMBO J* 19(11):2728-38, 2000.

Zhang, Z, S Schwartz, **L Wagner,** and W Miller. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7(1-2):203-14, 2000.

Standalone BLAST Incorporates MegaBLAST, RPS-BLAST, and BLASTClust

With release 2.1.2, NCBI's Standalone BLAST package now contains three new programs, MegaBLAST, RPS-BLAST, and BLASTClust. These programs supplement the standard distribution, including the basic BLAST program, blastall, blastpgp, PSI- and PHI-BLAST, bl2seq (a standalone version of BLAST2Sequences), and formatdb (the program used to create BLAST-ready databases).

MegaBLAST uses an algorithm developed by Webb Miller et al. [1], that is designed to swiftly compare two large sets of nucleotide sequences that differ only slightly from one another, perhaps as a result of sequencing errors. MegaBLAST is about 10 times faster than BLAST and is used by NCBI to assemble the clusters that comprise UniGene.

BLASTClust uses a single-linkage clustering process to group protein sequences based on pairwise matches found using the BLAST algorithm. The program accepts as input a file of concatenated protein sequences in FASTA format, each with a unique sequence identifer. It returns a file of sequence identifiers arranged in clusters.

RPS-BLAST uses a protein sequence query to search a library of PSI-BLAST Position Specific Score Matrices (PSSMs). Also included with the package are the programs "makemat" and "copymat". These programs are required to create the RPS-BLAST PSSM libraries. Partially processed libraries, based on the protein domains in the PFAM and SMART databases, are available from NCBI via FTP at ftp://ncbi.nlm.nih.gov/pub/mmdb/cdd/.

Additional Enhancements

In order to calculate more accurate Expect-values, BLAST and PSI-BLAST now take into account the amino acid composition of individual database sequences. The new Expect-value calculations use a scaling procedure [2,3] that creates a composition-corrected scoring system for each individual database sequence. Hence, identical alignments may receive different scores depending upon the amino acid composition of the database sequences involved.

An option to generate XML output is also provided (using the command line option "-m 7").

Pick up Standalone BLAST 2.1.2 in the "executables" directory at ftp://ncbi.nlm.nih.gov/blast.
—DW, SM

References

1. Zhang, Z, S Schwartz, L Wagner, and W Miller. *J Comput Biol* 7:203-14, 2000.

2. Altschul, SF, et al. *Nucl. Acids Res.* 25:3389-3402, 1997.

3. Schäffer, AA, et al. *Bioinformatics* 15:1000-1011, 1999.

caaatccgttcttgatcgtacatagcgcatgtcagncaaatccgttcttg
caaatccattcttgatcgtacatggcacatgtcagtcaaatccattcttg

How to Set Up a BLAST Web Server

Do you need to provide local Web access, with an NCBI-like interface, to BLAST, PSI-BLAST, PHI-BLAST, MegaBLAST, BLAST2Sequences, and RPS-BLAST? Do you need to support batch queries by sequence ID or accession number? NCBI offers a Web Server package that will support these interfaces as well as batch searching, XML output, and more.

Step 1: FTP the Web Server Package

Download the latest Web Server package by FTP from:

ftp://ncbi.nlm.nih.gov/blast/server/

Files are located in the "current_release" directory. Releases are available for Linux/Intel, OSF/DEC-Alpha, SGI and Solaris/Sun or Intel.

Step 2: Unpack the Archive

Unpack the distribution into any local document directory under the HTTPD server directory that has permission to run CGI programs with a *.cgi file extension. The BLAST Web server package is a Gzipped, tarred archive. Gunzip the archive using a command such as:

gzip -d wwwblast.platform.tar.gz
or
gunzip wwwblast.platform.tar.gz

Next, untar the archive using:
tar -xvpf wwwblast.platform.tar

Note the use of the "p", or "preserve" TAR option. This option preserves the file protections as they existed when the files were archived. This is essential to allow the proper access to all files by the server software. Uncompressing the archive will create a directory called "blast", which contains all the files needed to implement the WWW BLAST server.

Step 3: Format Local Databases

Format the local databases that you wish the server to use. This is accomplished in the usual manner using the program "formatdb", included in the server package. The database files should be placed in the "db" directory, which is created when you untar the archive. To format a database for BLAST, use:

formatdb -i infile -o T

Note that databases to be used with PSI- or PHI-BLAST should always be created using the "-o T" flag.

Step 4: Test the Installation

The server package comes with two BLAST databases called "test_aa_db"—a sample protein database, and "test_na_db"—a sample nucleotide database. These databases are configured to be used immediately. Perform a test search on these database using the "blast.html" Web page.

Step 5: Configure Local Databases

To add your own databases to the list used by the server, you must edit two configuration files called "blast.rc" and "psiblast.rc". These files specify the combinations of database names and BLAST program that are considered valid. Below is an excerpt from the file "blast.rc" as it appears "out of the box."

```
# Here is list of program/database,
# that allowed by BLAST service.
# Format: <program> <db> <db>  ...
#
blastn test_na_db
blastp test_aa_db
blastx test_aa_db
tblastn test_na_db
tblastx test_na_db
```

The line "blastn test_na_db," for example, specifies that the database named "test_na_db" is valid when used with the blastn program. To add another nucleotide database to the list searchable by blastn, simply add its name to this line so that it now reads:

blastn test_na_db new_database

The new database name must also be added to the html search forms in order to appear in the pull-down database menu boxes, so the blast.html file also needs to be edited. The relevant lines for blast.html are:

```
<select name = "DATALIB">
    <option VALUE = "test_na_db">
    test_na_db
    <option VALUE = "test_aa_db">
    test_aa_db
    <option VALUE = "test_aa_db pdb">
    test_aa_db & pdb
    <option VALUE = "pdbaa"> PDB
</select>
```

In this case, the first two "option" lines specify the two test databases. The third "option" line specifies that a database labeled "test_aa_db & pdb" can be selected from the pull-down list box. If a user selects this database, BLAST will search two databases, "test_aa_db" and "pdb," as though they were a single database. The fourth "option" line indicates that a database called "pdbaa" will be labeled in the pull-down list box as "PDB." Other html search forms, such as the form for MegaBLAST, will also require the same type of alteration when a new database is added.

Step 6: Celebrate

Run your first flawless test of the BLAST server!

The BLAST Lab feature is intended to provide detailed technical information on some of the more specialized uses of the BLAST family of programs. Topics are selected from the range of questions received by the BLAST Help Group.

Gene Nomenclature
continued from page 2

at NCBI. HGNC also provides links to other NCBI resources, such as PubMed and LocusLink, with a planned link to NCBI's reference sequences (RefSeq).

The NCBI Connection: Use of HGNC's Names

The official nomenclature established by HGNC, as well as the official nomenclature generated by other nomenclature committees, is an integral component of NCBI's LocusLink and RefSeq resources. When available, an official gene name and symbol are the preferred labels for a gene in LocusLink and are incorporated into the RefSeq mRNA and genome annotation records. NCBI also provides a link to the respective committee's source records for any individual requiring further information.

Access to Nomenclature Database

For more information on HGNC and to access the nomenclature database, visit their Web site at www.gene.ucl.ac.uk/nomenclature.

Other Nomenclature Groups

Nomenclature committees for zebrafish, fruit fly, mouse, rat, yeast, and human, plus additional nomenclature resources, are listed on NCBI's Web site. From the LocusLink home page, follow the Nomenclature link. *—CB, DM*

Mirror FTP Site for GenBank

The San Diego Supercomputer Center (SDSC) now offers an alternative FTP site for downloading the full releases of GenBank data as well as daily updates. Access to additional NCBI data will be added over time. The SDSC site can be reached directly at ftp://genbank. sdsc.edu/pub. There is also a link to the SDSC server from NCBI's FTP page at www.ncbi. nlm.nih.gov/Ftp/.

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

Official Business
Penalty for Private Use $300

FIRST CLASS MAIL
POSTAGE & FEES PAID
PHS/NIH/NLM
BETHESDA, MD
PERMIT NO. G-816