NCBI News

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Department of Health and Human Services

Summer 2002

Cn3D 4.0: Coupling Alignments to Structure

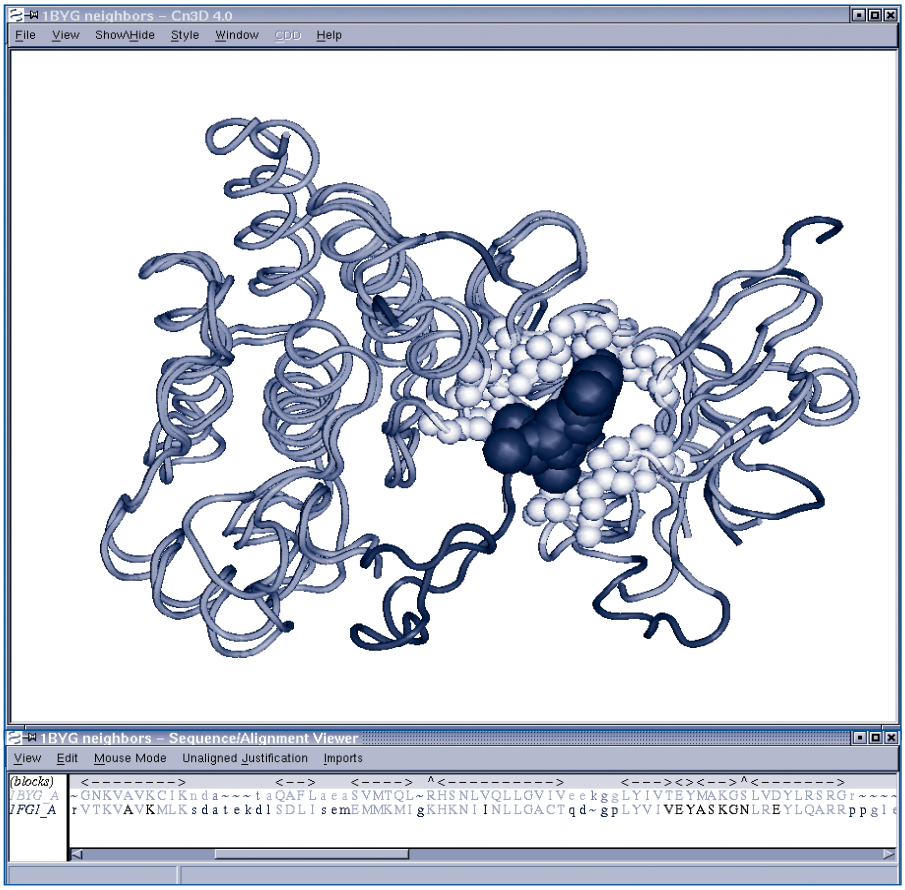

Figure 1: Cn3D 4.0 display shows a structural alignment of two human tyrosine kinases, 1BYG and 1FG1, as computed by the VAST algorithm. The structures include a bound inhibitor, shown in a space-filling representation along with all atoms within a radius of 5 angstroms in a ball and stick representation. The alignment viewer displays a portion of the structural alignment, with aligned residues in capital letters and each aligned block represented in the bar above the sequences.

NCBI has released version 4.0 of Cn3D, a program for viewing and manipulating the 3D structures of macromolecules. Cn3D 4.0 is a thorough redesign of the previous version with a growing focus on using structural information as a guide to creating protein multiple sequence alignments. Many new features are summarized below as they apply to the enhanced display of structural data and the new tools for editing and creating protein multiple sequence alignments.

continued on page 3

SNP Population Grows at NCBI

SNPs, or single nucleotide polymorphisms, are variations in genomic DNA sequences within a population of organisms. These genetic changes occur at a frequency of over 1 percent in the human genome, and are important because they are sometimes linked to heritable phenotypes. Knowledge of SNPs is useful for physical mapping, disease association, and surveys of population structure. The dbSNP database was developed at NCBI to facilitate the management of SNP data, integrate this data with other NCBI resources, and distribute the information to the scientific community.

continued on page 7

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
David Wheeler

Contributors
Medha Bhagwat
Susan Dombrowski
Scott Mcginnis
Eric Sayers

Writers
Vyvy Pham
David Wheeler

Editing and Production
Jennifer Vyskocil
Robert Yates

Graphic Design
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. 02-3272

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

1000th Viral RefSeq Unmasked!

In the last issue of the NCBI newsletter, we reported the release of the 1000th virus RefSeq genome but we did not identify the lucky sequence. The sequence is segment 12 of the complete genome of the Eyach virus. Clicking on the "Nucleotide" link at the top right-hand corner of the Entrez Genomes report for NC_003707.1, the accession number of segment 12, will display all of the segments for this virus genome.

The Eyach virus belongs to the family *Reoviridae* and to the genus *Coltivirus*. Viruses of this family contain dsRNA genomes that are divided into segments. The Eyach virus, which has 12 segments, has been linked to neurological disease associated with tick-borne infections in humans. It is antigenically similar to the well-known Colorado tick fever virus. The associated literature citation, accessible via the "PubMed" link, includes comparisons of genomic and morphological data of the Eyach virus to the Colorado tick fever virus, and to other reoviruses, and delineates an evolutionary relationship among these viral groups. Genomic analysis of these virus genomes could lead to breakthroughs in deciphering the mechanisms of infection and disease.

For more information on this virus, visit the Virus RefSeq Genomes Web site in Entrez Genomes at **www.ncbi.nlm.nih.gov/PMGifs/Genomes/viruses.html**.
—*VP*

New Genomes in GenBank

View the latest arrivals and learn more about the proteins they encode from Entrez Genomes.

Thermosynechococcus elongatus BP-1: BA000039: AP005369 -AP005377

Buchnera aphidicola str. Sg (Schizaphis graminum): AE013218

Chlorobium tepidum TLS: AE006470

Thermoanaerobacter tengcongensis strain MB4T: AE008691

Streptomyces coelicolor A3(2): AL645882

Methanosarcina mazei strain Goe1: AE008384

Xantomonas axonopodis pv. citri str. 306: AE008923

Xantomonas campestris pv. campestris str. ATCC 3391: AE008922

Select the Genomes database from the Entrez home page at www.ncbi.nlm.nih.gov/Entrez/.

Complete genomes are also available for downloading by FTP from ftp://ftp.ncbi.nih.gov/genomes/.

Editing Multiple Sequence Alignments

While previous versions of Cn3D allowed users to import sequence alignments into the program, Cn3D 4.0 allows users to edit these alignments. Alignments in Cn3D consist of multiple blocks of ungapped sequence, each ideally representing an element of the 3D structure such as an individual alpha-helix or beta-strand. In the new Editor mode of the Alignment Viewer, the pre-existing blocks of an imported alignment can be expanded, contracted, or deleted, and new blocks can be created using a "block editor", visible whenever editing is enabled. A given block can also be split into two blocks, or two blocks can be merged into one. One can set the coloring options so that these changes to the alignment are immediately reflected in the color of the affected residues on the displayed structure. One can also manually change the alignment of any sequence within a block by dragging the residues left or right. To assist with these tasks, the Alignment Viewer can search for a sequence pattern using Prosite syntax. Figure 1, on page 1, displays a VAST (Vector Alignment Search Tool) structural alignment of four tyrosine kinases, and the Alignment Viewer clearly indicates the block structure of the alignment.

Creating New Sequence Alignments

In addition to editing existing sequence alignments, users can now create new alignments and add new sequences to existing alignments. Both of these functions are provided in a Cn3D window named the Import Viewer. The Import Viewer serves as a workspace in which users import sequences, either over a network or from a local disk, pairwise align them to the master sequence of the currently loaded alignment, and then merge the new alignments into the Alignment Viewer. Sequences from the Alignment Viewer can also be sent to the Import Viewer for realignment if desired. Users can choose between three algorithms with which to align their imported sequences: BLAST, PSI-BLAST, or sequence-structure threading, the latter two of which are new to Cn3D 4.0. In PSI-BLAST, Cn3D 4.0 will calculate a Position Specific Scoring Matrix (PSSM) from the currently loaded alignment, and use this matrix to generate an alignment of the imported sequence to the master. Threading employs an algorithm developed at NCBI to align the imported sequence to the master using both structure and sequence information generated from a user-controlled combination of the PSSM and residue contact potentials. This combination of algorithms gives the user a powerful set of tools to correlate sequence and structure conservation.

Enhanced Molecular Display

Upon opening a structure record in Cn3D 4.0, users will notice improvements in both the quality and speed of the graphical display. In addition, the manipulation of the structure is now more straightforward and responsive. For example, zooming the view in or out can now be performed by dragging the mouse while holding down a control key (**Ctrl** on PCs, **Cmd** on Macs), and translation of the structure is likewise performed while holding down the **Shift** key. Moreover, Cn3D makes animating molecules easier by allowing the user to step through multiple structures in the display with simple keystrokes. The latter is especially useful for NMR ensembles, structural alignments, or crystal structures with multiple sets of coordinates.

Cn3D's menus have been redesigned to provide easier access to the numerous display options. The Color and Style menus are combined into a new Style menu that collects commonly used options and provides direct access to dialog boxes controlling rendering details. In keeping with the major updates in handling alignments, a new Show/Hide menu gives convenient access to commands controlling the display of aligned versus unaligned portions of the displayed structure.

Cn3D 4.0 now features a Select by Distance option that allows users to find all atoms within a chosen radius, in Angstroms, from currently selected atoms. The residues found are highlighted both in the structure and sequence views for ease of identification. This option is especially useful for exploring contacts between a ligand and binding site residues, or between adjacent units of secondary structure. In Figure 1, the residues within 5 Angstroms of the two tyrosine kinase inhibitors are shown in a ball and stick representation.

In the near future, users should expect the release of the curated Conserved Domain records that will be viewed using Cn3D. Cn3D will not only show the alignments, but also curated annotations on both the sequences and structures, highlighted substructures and alignment columns of functional importance, and links to PubMed literature. These unique records will provide a wealth of critical biochemical data about a protein family in a single, integrated view. *—ES*

View the Mouse Genome with Map Viewer

NCBI offers access to two annotated assemblies of the mouse genome. The first is an NCBI assembly constructed from curated contigs representing finished (Phase 3) high throughput genomic sequence. The second is the Mouse Genome Sequencing Consortium's (MGSC) whole genome shotgun assembly, based on the February 1, 2002 freeze of the data. Views of the MGSC assembly include annotations by NCBI. Combined, these data greatly facilitate the comparison of the mouse and human genomes and pave the way for the study of mouse homologues of human disease-causing genes.

The two data sets can be accessed and viewed with the Mouse Map Viewer, a Web-based, interactive, genome display tool that enables one to visualize annotated features on the assembly, such as predicted genes, sequence tagged sites (STSs), expressed sequence tags (ESTs) UniGene clusters, and single nucleotide polymorphisms (SNPs or variations).

Queries can be made against the whole genome or an entire chromosome by supplying a gene name or symbol, accession number or any other valid numerical identifier in the textbox on the main search page. Checking the box next to the "Show linked entries" will display all of the mapped elements associated with the query term. Clicking on the chromosome of interest leads to view of a whole chromosome. The main search page also provides a link to the mouse genome BLAST pages. The BLAST results page provides a "Genome View" button that links to the Mouse Map Viewer and permits the visualization of BLAST hits in a genomic context.

A variety of maps are available for the two assemblies, including sequence-based maps, fingerprint clone data, and genetic, radiation hybrid and YAC mapping data. A total of 13 maps from a collection of 28 can be displayed in a single view. The "Maps and Options" link provides additional display options for the selected maps. By default, the selected maps are displayed from left to right, and the rightmost map, termed the "Master Map", gives additional descriptive information about a given map element. Any map can be promoted to the Master Map by simply clicking on the arrow above the map or by choosing this feature from Maps and Options.

Some features of the Mouse Map Viewer are illustrated in Figure 1, which shows the results of a search for the gene encoding the mouse *Lpl* gene for lipoprotein lipase. The *Lpl* locus is displayed in the context of five different maps: Contig_ MGSCv3, Component _MGSCv3, Mm._Unigene_MGSCv3, Variation_ MGSCv3, and Gene_ MGSCv3. *Lpl* maps to mouse chromosome 8

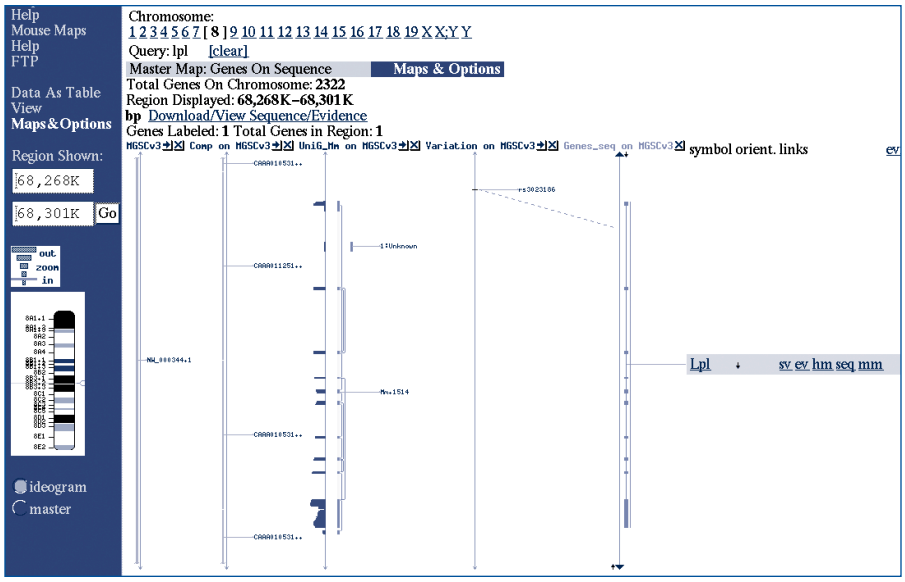

Figure 1: Results of a Mouse Map Viewer search for the mouse lipoprotein lipase gene.

continued on page 5

Mouse Genome Resources Accommodate a Diverse Array of Data

As sequence data accumulates, it is becoming clear that the genomic sequences of human and mouse share a high degree of similarity. The emerging mouse genomic sequence will strengthen the mouse as a model system for studying human disease and investigating potential disease treatments.

The Mouse Genome Resources page collects a variety of services running the gamut from information on mouse strains and annotation projects to powerful genomic display and analysis tools such as the Map Viewer and Mouse Genome BLAST. A central resource for mouse genomic data is the Mouse Genome Map Viewer, which displays NCBI-assembled sequence contigs derived from HTG sequence as well as those assembled using Whole Genome Shotgun (WGS) sequences from the Mouse Genome-Sequencing Consortium (MGSC). The Map Viewer, accessible by chromosome via the "Jump to the Genome" link, is the subject of the article beginning on page 4.

Mouse Genome BLAST can be used for similarity searches of mouse-specific databases, such as assemblies from the MGSC, ARACHNE, PHUSION, WGS traces, curated NT contigs, HTG sequences, and BAC ends. The Mouse Genome BLAST page uses MegaBLAST for nucleotide searches. Optimized for aligning sequences of high sequence similarity, MegaBLAST is up to 10 times faster than the more conventional blastn program and is able to efficiently handle much longer DNA sequences than blastn.

Other resources collected on the Mouse Genome Resources page include the Clone Registry, which contains information on BAC clones; the Human-Mouse Homology Map, which displays blocks of conserved synteny between mouse and human; LocusLink, which provides a single query interface to curated sequence and descriptive information about genetic loci; RefSeq, which provides Reference Sequences for genomic contigs, mRNAs, and proteins; the Sequencing page, which provides a summary of mouse genome sequencing progress and current genome data; the Trace Archive, a collection of raw sequence traces from sequencing projects, such as the mouse; and FTP Sites, for downloading data associated with mouse resources from NCBI.

NCBI's Mouse Genome Resources can be accessed at **www.ncbi.nlm.nih.gov/genome/guide/mouse/**. *—VP*

View the Mouse Genome

continued from page 4

and has been annotated on supercontig NW_000344.1. The Component map shows the tiling path of the GenBank records used to create the Contig map. Supporting EST evidence and a link to the corresponding UniGene cluster are displayed on the UniGene map, while observed polymorphisms within *Lpl* are shown on the Variation map. The Genes_ seq map provides links to other NCBI resources and molecular data, including LocusLink (Lpl link), Sequence Viewer (sv), Evidence Viewer (ev), *Lpl* gene sequence (seq link), Human-Mouse Homology Map (hm) and a link to NCBI's Model Maker (mm). When a ruler is displayed alongside the Genes_ seq map, nucleotide sequences flanking *Lpl* can be obtained by following the seq link or Download/View Sequence/Evidence link and specifying the region of interest. Additional information adjacent to the Genes_seq map includes an arrow that shows the direction of transcription and an evidence code that indicates the type of evidence used to construct the gene model. Try the Mouse Map Viewer from the Mouse Genomic Biology Page.

Data from the two genome assemblies can be downloaded via FTP at **ftp://ftp.ncbi.nih.gov/genomes/M_musculus/**. For additional information, click on the 'help' link in the Map Viewer. *—SD*

Selected Recent Publications by NCBI Staff

Anantharaman V, EV Koonin, and **L Aravind.** Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.* 2002 Apr 1; 30(7):1427-64.

Galperin MY, TA Gaidenko, AY Mulkidjanian, M Nakano, and CW Price. MHYT, a new integral membrane sensor domain. *FEMS Microbiol Lett* 2001; 205(1):17-23.

Kondrashov AS, and **SA Shabalina.** Classification of common conserved sequences in mammalian intergenic regions. *Hum Mol Genet* 2002; 11(6): 669-74.

Makarova KS, L Aravind, NV Grishin, IB Rogozin, and **EV Koonin.** A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* 2002; 30(2).

McCune AR, RC Fuller, AA Aquilina, RM Dawley, JM Fadool, D Houle, J Travis, and **AS Kondrashov.** A low genomic number of recessive lethals in natural populations of bluefin killifish and zebrafish. *Science.* 2002 Jun 28; 296(5577):2398-401.

Nikolskaya AN, and **MY Galperin.** A novel type of conserved DNA-binding domain in the transcriptional regulators of the AlgR/AgrA/LytR family. *Nucleic Acids Res.* 2002 Jun 1;30(11):2453-9.

Panchenko AR, and **SH Bryant.** A comparison of position-specific score matrices based on sequence and structure alignments. *Protein Sci* 2002; 11(2).

Rogozin IB, KS Makarova, J Murvai, E Czabarka, YI Wolf, RL Tatusov, LA Szekely, and **EV Koonin.** Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.* 2002 May 15; 30(10):2212-23.

Webb CT, **SA Shabalina, AY Ogurtsov,** and **AS Kondrashov.** Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae. Nucleic Acids Res* 2002; 30(5):1233-9.

Wilbur WJ. A thematic analysis of the AIDS literature. *Pac Symp Biocomput.* 2002; 386-97.

Submitting Large Sequin Files

A new web-based tool, called SequinMacroSend, is available to avoid problems associated with large file transfers via e-mail delivery systems. To use the system, a Sequin submission file is created in the usual manner and uploaded into a form from a local disk. SequinMacroSend submissions are immediately given a temporary "DSub" number. The submitter automatically receives an e-mail with the DSub number in the subject line and confirmation of receipt of the submission. Note that the DSub number is not the accession number, and therefore should not be used in publication. The GenBank accession number will be assigned when the submission is processed, normally after two business days.

Try SequinMacroSend at **www.ncbi.nlm.nih.gov/LargeDir Subs/dir_submit.cgi/**.

GenBank Release 131 is Available

GenBank release 131 (August 2002) contains over 18 million sequence entries totaling more than 22 billion base pairs. GenBank may be searched using NCBI's Entrez search and retrieval system or may be downloaded from the NCBI FTP site. GenBank flatfiles can be downloaded from the "genbank" directory; the ASN.1 version of GenBank is found in the "ncbi-asn1" directory. Links to GenBank mirror sites are available on the GenBank FTP site.

BLAST Version 2.2.4 Released: Web BLAST Linkouts

The latest version of the BLAST programs, version 2.2.4, supports discontiguous word matching in MegaBLAST. In addition, standalone BLAST supports an out-of-frame gapping option for blastx searches, allowing one or two bases, rather than the usual in-frame multiples of three, to be inserted or deleted from an alignment. The expect values are still calculated assuming in-frame gapping.

Fastacmd, a program to retrieve sequences from a formatted BLAST database, can now retrieve partial sequences (using the -L option) and print taxonomic information for a sequence.

BLAST version 2.2.4 is available at **ftp://ftp.ncbi.nih.gov/blast/executables/**.

Web BLAST Linkouts

Web BLAST output has been enhanced to include links to UniGene and LocusLink, from BLAST alignments. In addition, all database hits resulting from a BLAST search may be downloaded using a new button appearing in the Web BLAST output, and individual sequences from the output may be selected for download using checkboxes in the alignments section. *—SM*

SNP Population Grows
continued from page 1

Composition of the Database

The data in dbSNP includes SNPs, microsatellite repeats, and small insertion/deletion polymorphisms. There is no minimum allele frequency or requirement that a SNP result in a measurable phenotype for submission to dbSNP, and a large portion of the polymorphisms in the database are neutral polymorphisms. Currently, dbSNP contains predominantly human data, but variation information for several other organisms can be found in the database. Release 106 of dbSNP contained 4.5 million SNPs, and the database is growing at a rate of 90 SNPs per month.

Although dbSNP accepts submissions from any laboratory or individual, the bulk of the submissions are derived from large-scale contributors associated with the National Human Genome Research Institute's (NHGRI) grants program that aims to catalog 50,000 SNPs by 2005. SNPs are submitted to dbSNP using a special procedure that involves registering a submission "handle" with the NCBI SNP group, followed by the preparation of a set of structured submission files. Instructions on how to submit to dbSNP are located on the dbSNP home page. Each SNP in the database is given an identifier beginning with "ss", for "submitted SNP". If there are multiple submissions for the same SNP, then a reference SNP cluster is created, to incorporate information from the multiple submitters. The reference SNP clusters, given "rs" identifiers, are used in the annotation of reference genome sequences.

The SNP Record

A SNP record contains the observed alleles at a particular locus, the flanking sequence that surrounds the variation, the experimental method used to assay the variation, including protocols and conditions, and cross-references to associated GenBank records or UniGene clusters. Other types of data that can be included are genetic map locations, population-specific frequencies, individual-specific genotype information, relevant publications that document the details of the methodologies or populations, known genes in the region, synonyms for a submitter's SNP ID used in the submission, and validation information to describe the quality of the frequency data.

Searching dbSNP

Searches of dbSNP may be limited to Entrez fields such as allele variation, validation status, chromosome on which the SNP is mapped, and many others. SNP records retrieved in Entrez are displayed in a summary format tailored to the structure of a SNP record. There are, however, several additional display formats, such as a graphical summary and a chromosome report. Entrez SNP results may also be sorted by various fields including organism, SNP ID, success rate, and heterozygosity.

Special SNP query services offer pre-formulated search methods by Submitter, New Batches, Method type, Population Detail, Publication, Locus Information, and STS Markers; two Free Form Search services are also offered. In addition, the dbSNP data can be searched using a special BLAST page. These search options are linked from the blue sidebar menu of the dbSNP home page.

The integration of SNPs into other resources, such as the Map Viewer, provides a way to see them in their genomic context. When the "SNP" Master Map is viewed in the Map Viewer, a graphical summary showing mapping information, associated gene features, and marker heterozygosity is provided.

Downloading SNPs

Batch downloads of SNPs can be performed using Batch Entrez or via an e-mail-mediated query service that allows for the retrieval of a large number of SNPs by using individual submissions (ss#), submitter IDs, or dbSNP RefSNP cluster IDs (rs#). The SNP batch query service is accessible from **www.ncbi.nlm.nih.gov/SNP/batchquery.html/**.

The SNP data may also be downloaded at **ftp://ftp.ncbi.nih.gov/snp/**.

The SNP home page is found at **www.ncbi.nlm.nih.gov/SNP/**.

The Entrez SNP page can be reached from the "Hotspots" column on the NCBI home page.
—VP

Initial Assembly of the *Anopheles Gambiae* Genome

Anopheles gambiae is the primary mosquito vector for human malaria, a disease that causes an estimated 200 million clinical cases and more than one million deaths annually. Coupled with the recently released sequences of human and rodent malaria species, and that of the human host genome, the sequence of *Anopheles gambiae* will help to deepen our understanding of this disease.

The *Anopheles gambiae* genome was sequenced using a whole genome shotgun (WGS) approach, and the resulting WGS sequences were assembled into contigs for submission to GenBank under the project accession AAAB00000000. The first version of the assembly consists of 8,987 sets of ordered and oriented contigs with estimated gap sizes, called scaffolds. These scaffolds are identified with accession numbers ranging from AAAB-01 000001 to AAAB01008987. Over 100 scaffolds comprising about 82 percent of the genome have been mapped to a chromosomal location. An additional 60,737 short contigs, typically less than a thousand base pairs each, have not been uniquely placed within the genome.

An NCBI Map Viewer display for the three chromosomes of *Anopheles gambiae* is available from **www.ncbi.nlm.nih.gov/cgibin/Entrez/map_search?chr=agambiae.inf/**.

Sequence data can be found at **ftp://ncbi.nlm.nih.gov/genbank/genomes/Anopheles_gambiae/**.
—*VP*

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