NCBI News

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

Summer 1999

A Decade of Data at NCBI

Integrated Approaches to Managing the Information Explosion

Over the past 10 years the management of biological information has truly come of age, becoming increasingly integrated into the scientific process. It is now almost impossible to think of an experimental strategy in biomedicine that does not involve some online foray into scientific databases. At the core of this shift is a huge data explosion, most notably in the amount of gene sequence and mapping information.

From its inception in November 1988, NCBI was charged with providing data access and analysis tools for molecular biology information. As its 10th anniversary year draws to a close, the horizon is a familiar one—a flood of data coming in many new forms. This year-by-year tour marks highlights from NCBI's first decade.

1988 to 1989: Getting Ready

The newly established Center set up offices within the National Library of Medicine at NIH with a small staff representing a core combination of expertise in molecular biology, mathematics, and computer science.

As a research and development agenda emerged, an uppermost priority was to design tools for analyzing the growing number of sequences in GenBank, at that time managed elsewhere at NIH. Also of prime importance was to develop a flexible and robust data model to form the backbone of all data collection and data access services to come.

1990: BLAST Off!

With the development of BLAST, now a household word for many biologists, NCBI offered a program to find DNA or protein sequence similarities quickly, while also providing a statistical measure of significance that infers biological relevance. With significantly faster performance than existing algorithms, BLAST soon became the tool of choice among molecular biology researchers, and now supports more than 50,000 searches per day.

1991: Unified Entry via Entrez

A key objective of NCBI is to build integrated approaches to searching biological information. The Entrez retrieval system, first produced on CD-ROM, provided a unified interface, or entry point, for varied data types. For the first time, nucleotide sequences were linked with their protein translations. Links to "sequence neighbors" uncovered by BLAST searches were also featured, revealing previously unreported sequence similarities and even prompting some debate on what constituted a novel homology finding. Literature links facilitated follow-up on interesting relationships. Macromolecular structures, genome maps, a phylogenetic taxonomy, and even the whole of MEDLINE were later woven into the Entrez mesh.

1992: GenBank Moves to NCBI

In October 1992, NCBI assumed formal responsibility for GenBank, following a transition period during which NCBI developed the software and data infrastructure to maintain the database, enhance data quality, and provide Internet-based access.

Continued on page 2

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
Donna Maglott
Jo McEntyre
Kim Pruitt
Greg Schuler
Yanli Wang

Writer
David Wheeler

Editing, Graphics, and Production
Marla Fogelman
Veronica Johnson
Jennifer Vyskocil

Design Consultant
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. 99-3272

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

A Decade of Data at NCBI

Continued from page 1

Internet access began with e-mail servers for text searching and BLAST analysis and expanded to include client-server and Web access as the Internet evolved. GenBank has since grown from 100,000 to more than 4 million sequences.

At about this time, a technique for producing randomly initiated partial cDNA sequences known as Expressed Sequence Tags (ESTs) also came into prominent use. Following computational research that demonstrated the utility of ESTs for identifying genes in the sequence databases, NCBI created the dbEST database. A year later, a formal EST division was added to GenBank for this burgeoning class of data. EST sequences also laid the foundation for the later UniGene and GeneMap projects, which in turn contributed to progress on the Human Genome Project. Because of the utility of ESTs as short tags for whole genes, most human gene discoveries today rely heavily on EST approaches.

1993: Internet and 3-D Entrez

Network Entrez, a client-server version of the CD-ROM, brought Entrez to the Internet and paved the way for further data expansion. 3-D macromolecular structure data were also added to Entrez, enabling biologists to easily check whether the structures of proteins in the sequence databases had been determined experimentally. An outgrowth of molecular modeling research at NCBI, this enhancement is just one example of the important synergy between the research programs and the database development initiatives at NCBI. Later enhancements to Entrez would include the Cn3D structure viewer and the ability to link proteins based on structural similarity.

1994: Web Site Launched

NCBI launched its Web site (www.ncbi.nlm.nih.gov) in early 1994 with BLAST, Entrez, dbEST, and dbSTS. Recognizing the power of the Web to facilitate the level of data integration it envisioned for molecular biology, NCBI focused essentially all new development efforts on this new medium.

The dbSTS database was established in response to a growing body of new sequence data called Sequence Tagged Sites (STS), short sequences of known location in a genome used as essential markers for gene mapping and positional cloning. As the volume and importance of these sequences grew, they were consolidated into a new STS division of GenBank.

Electronic PCR (e-PCR), an STS-hunting tool that made it possible to mine dbSTS for the map location of a sequence, was developed a short time later. E-PCR simulates conventional PCR methods for identifying unique DNA landmarks by searching for STSs. Researchers use e-PCR to assign sequence database records to map positions, test primer feasibility, and integrate and anchor genetic maps and sequence data.

Continued on page 3

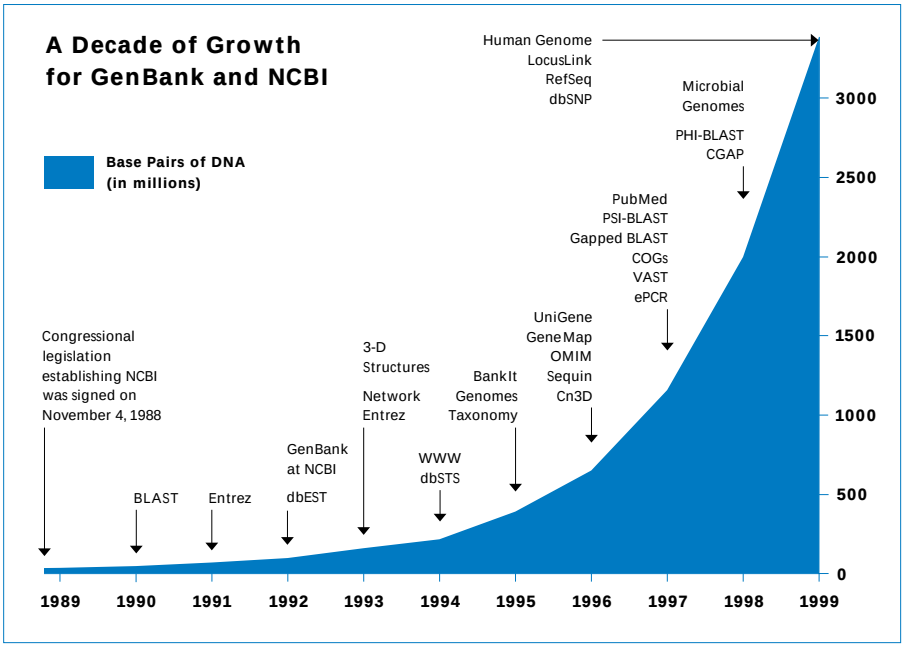
Continued from page 2

1995: Surge in Sequence and Mapping Data

What started as a trickle of sequences in the early 1980s was by now a torrent. BankIt offered a simple Web-based form for submitting DNA sequences to GenBank. Developed in response to growing interest in the Web, it quickly became the most popular submission tool.

As the Human Genome Project progressed, Entrez added a Genomes database to manage data on a genomic scale. Large-scale sequencing had by this time produced a number of completely sequenced genomes or chromosomes. The mapping initiatives had also generated many genetic and physical maps. The Genomes database provided effective ways to integrate disparate mapping and sequence data. A graphical viewer aided in the visualization of complete genomic data.

The Taxonomy Browser was developed following an NCBI initiative to create a consistent and comprehensive sequence-based taxonomy for the growing number of species represented in GenBank. Developed in collaboration with an international team of experts, this phylogenetic approach produced a classification that takes into account sequence similarities and more closely reflects evolutionary history than does classical taxonomy. Today the Taxonomy Browser provides access to more than 50,000 species in GenBank, with links into Entrez.



1996: Finding Genes Among the Sequences

Much of GenBank's growth was due to the high volume and redundancy of EST sequences, which had also started to present problems in data presentation and analysis.

The UniGene database organized matching sequences into clusters, each representing one human gene. With more than 75,000 clusters today, representing more than 75% of all human genes, UniGene serves as an important springboard for gene hunting.

GeneMap '96, the first transcript map of expressed human genes, was produced by an international radiation-hybrid-mapping consortium, which relied on UniGene as a central resource for identifying novel, non-redundant genome mapping candidates. Updated in 1998 and 1999, GeneMap integrates STS mapping data, sequence data, and UniGene clustering data and provides the mapping framework upon which to mount the complete sequence data. GeneMap '99 charts locations of more than 30,000 human genes.

Sequin, a sequence submission tool, was also released in 1996 to handle the surge in sequence data. Sequin especially facilitates submitting large batches of sequences and sets of related sequences from phylogenetic, mutational, or population studies. Later addition of alignment capabilities furthered Sequin's ability to sort out sequence relationships.

1997: PubMed and Proteins

PubMed expanded the literature component of Entrez to encompass the entire MEDLINE database and made it available for free over the Web, with links to full-text articles on Web sites of participating publishers. In the year following its official launch by Vice President Gore in June 1997, PubMed use increased from one million to 16

Continued on page 4

million searches per month. The success of PubMed has led to the increasing involvement of NCBI in projects related to electronic access to the scientific literature.

NCBI research in protein analysis spawned three major resources this year. The Entrez Structures database was enhanced with structure-based protein neighbors, often discovered to be homologs with similar biological functions. Structure neighbors are computed by the VAST (Vector Alignment Search Tool) algorithm, which identifies proteins that exhibit a combination of strong sequence and structure similarity. Visualization of aligned structures is supported by the Cn3D structure viewer.

Gapped BLAST and PSI-BLAST (Position Specific Iterated BLAST) increased both speed and sensitivity, and ushered in a new generation of sequence similarity search tools. PSI-BLAST facilitates profile-based searches, which are potentially much more sensitive to distant relationships than are the traditional pairwise similarity searches for which BLAST was originally tailored. PSI-BLAST can be used to help delineate diverse protein families and predict function for newly sequenced proteins.

The COGs (Clusters of Orthologous Groups) project takes a different approach to analyzing biological information. COGs organizes clusters of protein sequences from completely sequenced genomes of different species. Currently there are eight genomes in the scheme, spanning the major kingdoms of life. Analysis of COGs shows the molecular similarities and differences between species, which not only can provide clues about evolution, but also may help to identify protein families, predict new protein functions, and point to potential drug targets in pathogenic species.

1998: Billions of Bases to BLAST

GenBank surpassed the two-billion base pair mark in 1998. More than half of the data comes from a single organism—*homo sapiens*—largely due to the Human Genome Project's high throughput sequencing centers. As only about 8% of the human genome sequence is currently considered "finished," and recent predictions indicate that sequencing will proceed significantly more rapidly, the flow rate is set to increase. The HTGS (High Throughput Genomic Sequences) division of GenBank was established to organize these data as they are deposited in progressive stages of completion.

For protein analysis, PHI-BLAST (Pattern Hit Initiated BLAST) complemented the profile-based searching that was previously introduced with PSI-BLAST. PHI-BLAST incorporates hypotheses as to biological function of a query sequence and restricts the analysis to a set of protein sequences that are already known to contain a specific pattern or motif.

As more than 20 complete microbial genomes and one multicellular organism, the worm, were placed in the public domain, customized BLAST services and enhancements to the Entrez Genomes database were implemented to organize, visualize, and analyze these data.

Collaborations with other NIH Institutes for disease-based services were also established. Examples include the Cancer Genome Anatomy Project for expression data from normal, precancerous, and cancerous cells; and specialized Web sites for analysis of genetic diversity in malaria and HIV.

1999: Focus on Human Genome

As the Human Genome Project nears completion, the research focus is turning from analysis of specific genes or regions to a whole genome approach. NCBI has developed a suite of genomics resources to support comprehensive analysis of the human genome. New projects such as LocusLink, a hub for integrating key descriptors of genetic loci; RefSeq, a non-redundant set of human reference sequences for known human genes; and dbSNP, a collection of data on human genetic variation, all contribute to this network of information. Online Mendelian Inheritance in Man (OMIM), the Johns Hopkins comprehensive database of human genetic disorders, supplements these resources.

The challenge for the next decade will be to keep pace with the flood of genome data, while also designing the tools and databases for the gene discoveries of the 21st century.
—JM, BR, DW

LocusLink: Cross-Referencing Across Databases

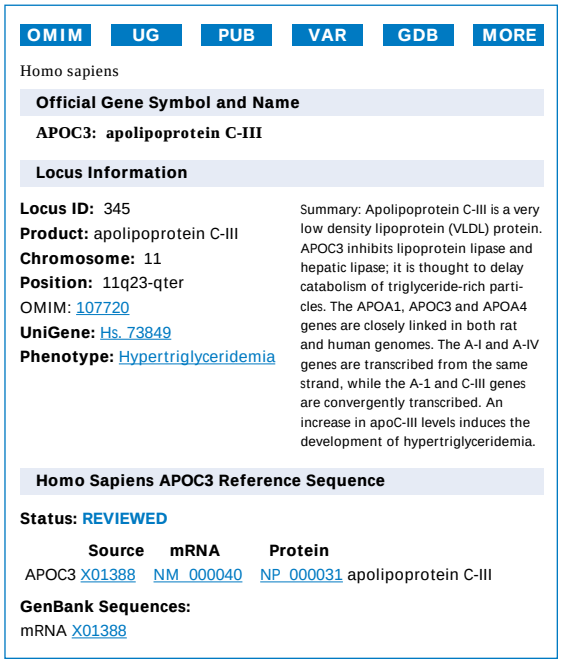
Data for a particular genetic locus, such as a gene, may exist in several qualitatively different resources. A phenotype and cytogenetic map position may be registered for a locus in OMIM, official and alternate names may be listed on the Human Gene Nomenclature page, and representative sequence data may be presented in a UniGene cluster. Additionally, GenBank itself may contain multiple sequences for a single locus. LocusLink (www.ncbi.nlm.nih.gov/LocusLink), developed by NCBI's Donna Maglott, provides an integrated querying and cross-referencing system to facilitate movement from one source to another.

LocusLink anchors an official gene name, gene aliases, database IDs, phenotypes, map positions, sequence accession numbers, and other identifiers to a stable LocusID number. The cross-referencing allows locus searches of various types to reliably converge on the same data. Currently limited to human gene loci, the service will add other organisms in the future.

Formulating LocusLink Queries

One way to search LocusLink is to select from an alphabetical list of official gene symbols. In addition, a search box supports diverse queries consisting of official gene names and aliases, accession numbers or other database identifiers, protein names, phenotypes, EC numbers, OMIM numbers, UniGene clusters, or map positions. For example, the accession number "AF053356," the UniGene cluster number "Hs.74561," or the EC number "4.1.2.13" is an acceptable search term.

The query syntax supports field restrictions and a wildcard symbol. Two particularly useful fields are **chr** for chromosome and **mim** for OMIM number. The query "2[chr]" returns a list of all loci found on human chromosome two. The query "A2*" retrieves a list of all records containing a word beginning with "A2." A multiword query such as "apolipoprotein hypertriglyceridemia" finds reports containing both words.


OMIM UG PUB VAR GDB MORE

Homo sapiens

Official Gene Symbol and Name

APOC3: apolipoprotein C-III

Locus Information

Locus ID: 345
Product: apolipoprotein C-III
Chromosome: 11
Position: 11q23-qter
OMIM: 107720
UniGene: Hs. 73849
Phenotype: Hypertriglyceridemia

Summary: Apolipoprotein C-III is a very low density lipoprotein (VLDL) protein. APOC3 inhibits lipoprotein lipase and hepatic lipase; it is thought to delay catabolism of triglyceride-rich particles. The APOA1, APOC3 and APOA4 genes are closely linked in both rat and human genomes. The A-I and A-IV genes are transcribed from the same strand, while the A-1 and C-III genes are convergently transcribed. An increase in apoC-III levels induces the development of hypertriglyceridemia.

Homo Sapiens APOC3 Reference Sequence

Status: REVIEWED

	Source	mRNA	Protein	
APOC3	X01388	NM_000040	NP_000031	apolipoprotein C-III

GenBank Sequences:
mRNA X01388

Figure 1: *LocusLink Report for APOC3.*

LocusLink Reports

Search results are summarized in a browsable list showing, for each entry, the LocusID, locus symbol, gene name, cytogenetic position, and a color-coded array of links to other resources that cite the locus, such as GenBank, RefSeq, UniGene, PubMed, dbSNP, or OMIM.

A full LocusLink report, illustrated in Figure 1 for APOC3, begins with a row of database buttons for resources that contain data on the locus, followed by the organism name and the official gene symbol assigned by the HUGO Human Nomenclature Committee.

Additional information includes the NCBI-assigned LocusID, any known alternative names or symbols, the gene product, cytogenetic location, associated OMIM records and UniGene clusters, phenotype, and a brief summary of gene function if available. Next comes a list of sequences for the locus, including a reference sequence if available. (See RefSeq article on page 7.) The reference sequence in the APOC3 example lists the accession numbers of the GenBank source sequence, the RefSeq mRNA, and the translated protein reference sequence. A second GenBank sequence for APOC3 is also reported.

LocusLink by FTP

The LocusLink database itself, updated weekly, is available by FTP through a **Download** link. A README file details the content of each data file. One file contains the basic LocusLink data, and three others contain cross-references from LocusIDs to GenBank, RefSeq, or OMIM numbers. *—DM, DW*

Selected Recent Publications by NCBI Staff

Gabrielian, A, D Landsman, and A Bolshoy. Curved DNA in promoter sequences (www.bioinfo.de/isb/1999/01/0017/). *In Silico Biol* 1:0017, 1999.

Hahn, T, E Matala, **C Chappey,** and N Ahmad. Characterization of mother-infant HIV type 1 gag p17 sequences associated with perinatal transmission. *AIDS Res Hum Retroviruses* 15(10): 875–88, 1999.

Jang, W, HC Chen, H Sicotte, and **GD Schuler.** Making effective use of human genomic sequence data. *Trends Genet* 15(7):284–6, 1999.

Makarova, KS, L Aravind, MY Galperin, NV Grishin, RL Tatusov, YI Wolf, and **EV Koonin.** Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res* 9(7):608-28, 1999.

Matsuo, Y, and **SH Bryant.** Identification of homologous core structures. *Proteins* 35(1):70–9, 1999.

Ponting, CP, L Aravind, J Schultz, P Bork, and **EV Koonin.** Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J Mol Biol* 289(4): 729–45, 1999.

Spouge, JL, and SP Layne. A practical method for simultaneously determining the effective burst sizes and cycle times of viruses. *Proc Natl Acad Sci USA* 96(12):7017–22, 1999.

Tatusova, TA, and **TL Madden.** BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 174 (2):247–50, 1999.

Wan, H, and **JC Wootton.** Axiomatic foundations of complexity functions of biological sequences. *Annals of Combinatorics* 3:77–99, 1999.

QBLAST Provides Quick Reformat Feature

A queuing system has been implemented for both Basic and Advanced BLAST. The new "QBLAST" system offers rapid reformatting of search results and enhances server performance by reducing the connection time with each user. The BLAST algorithm has not changed; QBLAST simply offers a modular approach that separates the search step from the output formatting step.

Previously, an output format had to be specified prior to running a BLAST search. To change formats, it was necessary to completely rerun the BLAST search. QBLAST saves the results under a Request ID number, which is then used to retrieve them in the desired formats. For secure data retrieval, Request IDs are unique numbers, containing a random component, and are not issued in sequential order. Therefore it is not possible for users to change any digit in their Request ID and receive the results of another person's search.

To initiate a search, enter a sequence query as usual and press **Search**. A Formatting page reporting your Request ID and showing all the display options will be returned. At this point, you can wait at your browser for the search to finish, or jot down the Request ID and use it to call up the results later. To view results immediately, press the **Format** button and they will be displayed as soon as the search is completed. To view results later, go to "Click here to retrieve results if you already have a Request ID", and enter your Request ID number. Although a few very large files will be deleted after 30 minutes, most results will be held for up to 24 hours.

From the Formatting page, you can choose from six alignment formats, adjust the number of alignments and descriptions, and toggle the display of both NCBI gi numbers and the graphical overview. To change output formats, simply select your entry and press the **Format** button again. *—DW*

Cn3D 2.5 Enhances Editing and Display

Version 2.5 of the Cn3D sequence/structure viewer is now ready for downloading through the NCBI Structure site. Cn3D works as a helper-application for Entrez, providing 3-D visualization and comparison of macromolecular structures, and is available for PCs, MACs, and most Unix platforms. Cn3D 2.5 offers several enhancements over previous versions:

- Features consisting of selected residues may be defined, annotated, and given unique display styles.
- Cn3D renderings, complete with feature definitions and aligned structures, can be saved as a single file for later retrieval.
- The PC version can be installed as a helper-application for Netscape or Internet Explorer with a single click of the mouse.
- The controls for the show/hide and display-style settings are now presented within easy-to-use control panels. *—DW, YW*

RefSeq: A Database of Reference Sequences

RefSeq is a new database, distinct from GenBank, which currently comprises a non-redundant set of human reference sequences for mRNAs and proteins. RefSeq is designed to simplify the analysis of a burgeoning sequence databank by providing reference sequence standards representing the transcripts and proteins encoded by loci. RefSeq records incorporate data gleaned from multiple GenBank records and literature sources. Although currently limited to human mRNA and protein sequences, RefSeq will eventually expand to include sequences from other organisms represented in GenBank. Additional sequence types, such as constructed genomic contigs and entire chromosomal reference sequences, will also be included.

Accession numbers of five types, distinguished by the initial two characters, are issued to RefSeq records, depending on the type of sequence involved:

NC_123456 Complete chromosome sequence.

NG_123456 Genomic region (e.g., immunoglobulin).

NM_123456 mRNA sequence with full-length coding information. No pseudogenes.

NP_123456 Protein sequence translation of a RefSeq nucleic acid coding region.

NT_123456 Constructed genomic contig sequence.

RefSeq Records May Be Either Provisional or Reviewed

Sequence records are incorporated into RefSeq in two stages, Provisional and Reviewed. In creating a Provisional record, the first step involves associating sequence data with named genes to select an initial input sequence. The input sequence is then used to locate a "source sequence" for the Provisional RefSeq entry. The source sequence is usually the longest mRNA sequence in GenBank that both contains the input sequence and is annotated with a complete coding region. This source sequence is then fed into an automatic process that generates the Provisional RefSeq records. A Provisional record is generated from the source GenBank record, with the addition of gene names and aliases, a stable LocusID number, the MIM number for the gene, and a statement in the Comment field that the entry is Provisional.

The date of deposition of candidate records is not considered when selecting the source sequence. Hence, the selection of a particular GenBank record as the basis of a RefSeq record does not imply primacy of publication.

In stage two, a Provisional RefSeq record is reviewed by NCBI staff or outside experts to produce a Reviewed record. During this stage, the Provisional entry may be modified and augmented considerably, incorporating data from other sequence records or from the scientific literature, in order to reflect the current state of knowledge of the locus in question. References to the literature are added along with a brief summary of the locus. The RefSeq mRNA sequence may also be extended using data from other genomic or mRNA GenBank records; however, because of their error-prone nature, EST data are not incorporated into RefSeq records. If there is strong evidence that a gene produces multiple biologically important transcripts and proteins, then individual RefSeq records are created to represent each. Such transcript variant RefSeq records are constructed after a careful review of the literature, or in collaboration with experts.

RefSeq Records Can Be Found Using Entrez and LocusLink

RefSeq records are included in the Entrez nucleotide database and may be searched in the same way as GenBank records. For example, a simple search using a RefSeq accession number, such as NM_000642, will return the corresponding record. LocusLink can be used to retrieve RefSeq records on the basis of a gene name, LocusID, or chromosomal location. (See LocusLink article on page 5.)

RefSeq Records Are BLASTable

RefSeq records are included in the BLAST nr database. The first portion of the one-line description in this BLAST output will resemble the following:

```
gi|4557284|ref|NM_000646.1|AGLf|
[4557284]
```

This line gives the unique gi number 4557284, accession number NM_000646, and LOCUS name AGLf assigned to the Reference sequence.

For more information, click on the **Reference Sequences** link on the NCBI home page. *—DM, KP, DW*

NCBI Exhibits and Workshops

NCBI will be exhibiting at the following scientific meetings; two of which also feature a BLAST workshop. The list of conferences is also available at www.ncbi.nlm.nih.gov/About/exhibitsched.html. For further information, contact NCBI at info@ncbi.nlm.nih.gov.

Genome Sequencing and Analysis Conference
Miami, Florida
September 18-21, 1999

National Institutes of Health Research Festival
Bethesda, Maryland
October 5-8, 1999

American Society of Human Genetics
San Francisco, California
October 19-23, 1999
BLAST workshop: October 20, 1999

Computational Genomics
Baltimore, Maryland
November 18-21, 1999

American Society of Tropical Medicine and Hygiene
Washington, DC
November 28-December 2, 1999

American Society for Cell Biology
Washington, DC
December 11-15, 1999
BLAST workshop: December 15, 1999

Take an NCBI Coffee Break

NCBI Coffee Break is a collection of short reports on recent biological discoveries. Each report incorporates interactive tutorials that show how bioinformatics tools are used as part of the research process. New Coffee Break stories will appear on the Web about every two weeks. A brief description of upcoming stories can be found on the Coffee Break page, as well as a link to the **archives**. Click on **Coffee Break** from the NCBI home page.

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