NCBI News

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

May 2005
Volume 14, Issue 1

GENSAT Project Data Now in Entrez

The Gene Expression Nervous Systems ATlas (GENSAT) provides the anatomical location of gene expression in the mouse brain GENSAT is based on BAC transgenic vectors in which endogenous protein coding sequences have been replaced by sequences encoding a reporter gene (EGFP). The EGFP is visualized in these sections by staining with an anti-EGFP antibody, or by confocal microscopy of unstained tissue sections. The images show the relative rates of transcription for each target gene. Images are available for mice at embryonic day 15.5 (E15.5), postnatal day 7 (P7) and adult developmental stages. In addition, GENSAT contains images taken at various developmental stages of tissue sections from non-transgenic mice lines in which the expression of a given gene is visualized using radiolabelled riboprobes with *in situ* hybridization.

GENSAT can be queried by gene name, alias or symbol, region of brain, imaging methods such as *in situ* hybridization, or confocal microscopy, section types such as saggital, horizontal, or coronal, and the age and sex of the mice. As with all other Entrez databases search terms can field-limited and combined using Boolean logic. For example, we can use the following query to get a

continued on page 4

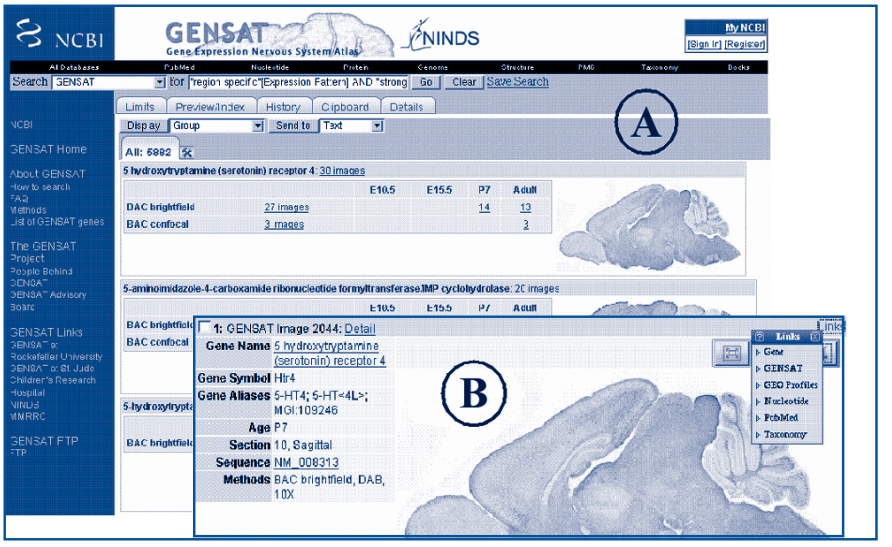

Figure 1. Entrez summary of GENSAT records matching the query given in the text. Clicking on the thumbnail image for the first record in "A" generates the view in "B" showing details of the set of images for serotonin receptor 4 and links to other Entrez databases in the "Links" menu.

Have it your way with My NCBI!

Would you like email alerts when new sequences or literature references for your favorite organism appear in Entrez? Would you like to archive your Entrez search strategies? Would you like to automatically partition your nucleotide search results by source database, species of origin, or sequence type? Would you like Entrez links displayed right on the web page rather than as a menu? My NCBI and a few simple configuration panels are all it takes to have it your way.

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. To subscribe to NCBI News, send your name and address to either the street or E-mail address below.

NCBI News
National Library of Medicine
Bldg. 38A, Room 3S-308
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
Peter Cooper
Susan Dombrowski
Monica Romiti
Tao Tao
Steve Pechous

Writers
Vyvy Pham
David Wheeler

Editing and Production
Robert Yates

Graphic Design
Robert Yates

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. 05-3272

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

Influenza Virus Resource Debuts

Each year in the USA, more than 200,000 patients are admitted to hospitals with influenza infections, and influenza-related deaths approach 36,000. Effective vaccines can reduce the numbers of hospitalizations and deaths, and the swift identification of new flu virus variants is an essential component in the development of such vaccines. The Influenza Genome Sequencing Project, funded by The National Institute of Allergy and Infectious Diseases (NIAID), aims to rapidly sequence flu viruses from samples collected throughout the world from humans and a variety of animals.

The viral nucleotide sequences resulting from this project, available in GenBank, along with the sequences of their encoded proteins, form the supporting database for NCBI's new Influenza Virus Resource (IVR). The IVR will enable scientists to quickly compare influenza virus strains so that emergent variants can be rapidly identified. As the library of viral sequences grows, this database will act as a reference to help further our understanding of the spread of animal viruses to humans, and the spread of influenza worldwide. Access the IVR home page at:

www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html

The Flu sequence database link offers an interface to the viral sequences that allows searches by virus type, subtype, host organism, genome segment and country of origin. Restrictions on year of virus isolation and sequence length may also be applied and the results of the search can be sorted, sequentially, by up to three of the search fields.

Individual nucleotide or protein sequences can also be retrieved by GenBank accession number.

Similarly, the Flu genome viewer tool can display viral nucleotide or protein sequences ordered by genome segments for each virus. All segments of the same virus are grouped together in the same background color, alternating in light blue and white, providing a convenient way to check the completeness of genome segments for viruses of interest. Database searches can be performed similarily as described above, and nucleotide or protein sequences can also be searched by adding a complete or partial virus name (e.g. Influenza A virus (A/New York/19/2003(H3N2)) or New York) in the box after "Name" and selecting "Find Nucleotides" or "Find Proteins".

Sequences of interest can be selected within the search results from either tool by checking the boxes to the left of their GenBank accession numbers. Selected sequences can be downloaded or "Saved" to be combined with results from a subsequent search. Multiple alignments of selected nucleotide or protein sequences generated by the MUSCLE[1] alignment program can also be viewed. On the multiple sequence alignment display page, any two sequences can be selected for a pairwise comparison using BLAST 2 Sequences.

For more information on the resources above, access the Help document linked from each resource's home page:

www.ncbi.nlm.nih.gov/genomes/FLU/SiteAbout.html

continued on page 6

Eight ASN.1 Utility Programs for Five Computer Platforms

The NCBI ToolKit provides source code and configuration scripts that make it easy to compile NCBI software to run on a variety of computing platforms. Some of the most familiar applications that can be built using the ToolKit are blastall, blastcl3, blastpgp, the BLAST web server, Sequin, Entrez2, and Spidey. These programs are well-known to NCBI users because they are provided in executable format for many common computing platforms. However, there are many useful but less well-known ASN.1 utility programs in the ToolKit that have previously been offered only as source code. NCBI now distributes a number of these command-line utilities for converting, validating, indexing, and creating NCBI ASN.1 records in executable form for 5 major computing platforms: Alpha, Linux, Macintosh, Solaris and MS Windows. They are run in Terminal or Command Prompt windows.

Each utility program accepts a number of command line arguments, specified using a dash and a single letter option code followed by an option value. Some values are boolean and are given as either 'T', true, or 'F', false. Others are specified using one-letter codes, such as format specifiers, or strings, such as file names or GenBank accession numbers. To see a complete list of command line parameters for any of the programs, run the program with a trailing dash and no parameter. A list of the eight programs with brief descriptions is given in Box 1, while a detailed description of one of the most versatile programs, "asn2all", follows. In many situations, the multifunctional program asn2all can be run instead of asn2fsa, asn2gb or asn2xml.

The program "asn2all" is primarily intended to generate reports from

continued on page 5

New Microbial Genomes in GenBank®

Organism	GenBank | RefSeq Accession Numbers
Gluconobacter oxydans 621H	chromosome: CP000009 | NC_006677 plasmid: CP000004—8 | NC_006672—6
Staphylococcus epidermidis RP62A	chromosome: CP000029 | NC_002976 plasmid: CP000028 | NC_006663
Staphylococcus aureus subsp. aureus COL	chromosome: CP000046 | NC_002951 plasmid: CP000045 | NC_006629
Thermococcus kodakaraensis	AP006878 | NC_006624
Dehalococcoides ethenogenes 195	CP000027 | NC_002936
Campylobacter jejuni RM1221	CP000025 | NC_003912
Ehrlichia ruminantium str. Welgevonden	CR767821 | NC_005295
Bacillus clausii sp. KSM-K16	AP006627 | NC_006582
Synechococcus elongatus PCC 6301	AP008231 | NC_006576
Francisella tularensis subsp. tularensis Schu 4	AJ749949 | NC_006570
Silicibacter pomeroyi DSS-3	chromosome: CP000031 | NC_003911 plasmid: CP000032 | NC_006569
Zymomonas mobilis subsp. mobilis ZM4	AE008629 | NC_006526
Anaplasma marginale str. St. Maries	CP000030 | NC_004842
Geobacillus kaustophilus HTA426	chromosome: AP006520 | NC_006509 plasmid: BA000043 | NC_006510
Idiomarina loihiensis L2TR	AE017340 | NC_006512
Azoarcus sp. EbN1	CR555306 | NC_006513
Lactobacillus acidophilus	CP000033 | NC_006814

For more detailed information, see the online version of the May 2005 NCBI News, or use the GenBank or RefSeq Accession Number to search the Entrez "Genome" database using the query box on the NCBI Home Page.

The Iceman Preserved in GenBank

In the fall of 1991, hikers in the Alps found the frozen body of a man at the edge of a glacier which they took to be that of a recent victim of a climbing accident. In fact, the hikers had found the frozen, mummified body of a man who lived and died in the late Neolithic. Popularly known as the "iceman", this well-preserved mummy and the artifacts found with him have provided an unprecedented and detailed view of human history, culture and biology of late stone-age Europe.

The well-preserved state of the mummy has allowed the collection of molecular sequence data that has shown the iceman's relationship to modern Europeans and has provided insight into his diet and culture. Available sequences in GenBank include the iceman's mitochondrial D-loop region as well as amplified sequences obtained in analyzing his

continued on page 8

GENSAT
continued from page 1

list of images for genes with localized expression patterns:

"region specific"[Expression Pattern] AND "strong"[Expression Level] AND "neuron"[Cell Type]

Here, we are using the standard Entrez syntax to search GENSAT for the phrase "region specific" within the field "Expression Pattern", indicated within the square brackets, combined using boolean "AND"s with two other field-limited phrases. Such queries can be built easily with the tools available in the Entrez "Preview/Index" tab visible in Figure 1A. The search returns summaries of image sets for over 5,000 genes as shown in Figure 1.

The summaries include a zoomable, thumbnail size image of the brain section, seen in Figure 1A, the total number of images that are available for each image type, and the stage of development. Double-clicking on the thumbnail image for the first image set, that for 5-hydroxytryptamine receptor 4 (serotonin receptor 4), generates a more detailed report shown in Figure 1B. Links to other Entrez databases, including Entrez Gene, Nucleotide and PubMed, are provided within the "Links" menu. To open the image browser, click on the image of Figure 1B or the cropping icon, visible under the Links menu. The image browser, Figure 2A, can be used to select, magnify and simultaneously display up to 30 subsections of the original image.

The GENSAT project, is an ongoing, collaborative study between the Rockefeller University and the St. Jude Children's Research Hospital, and aims to map the expression of genes in the central nervous system of the mouse during the normal development. GENSAT is supported by a grant from the National Institute of Neurological Disorders and Stroke (NINDS) to the Rockefeller University. The data from this study will shed light on the genetics of disorders that affect the central nervous system and also provide insight into the brain's response to both natural and foreign chemicals.

A list of the genes examined in this study can be obtained by searching Entrez Gene with the query:

"gene_gensat" [Filter]

The modified mouse lines are being deposited in the Mutant Mouse Regional Resource Center (MMRRC).

The image data and supplemental information from the GENSAT project can be downloaded from

ftp.ncbi.nih.gov/pub/gensat

Details on the experimental methods and results of the GENSAT project are published in *Nature*. 2003 Oct 30;425(6961):917-25

The data in the GENSAT database can be accessed by searching from the GENSAT home page at:

www.ncbi.nlm.nih.gov/projects/gensat

or from within Entrez at:

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gensat

Questions regarding GENSAT should be sent to:

info@ncbi.nlm.nih.gov

—SD

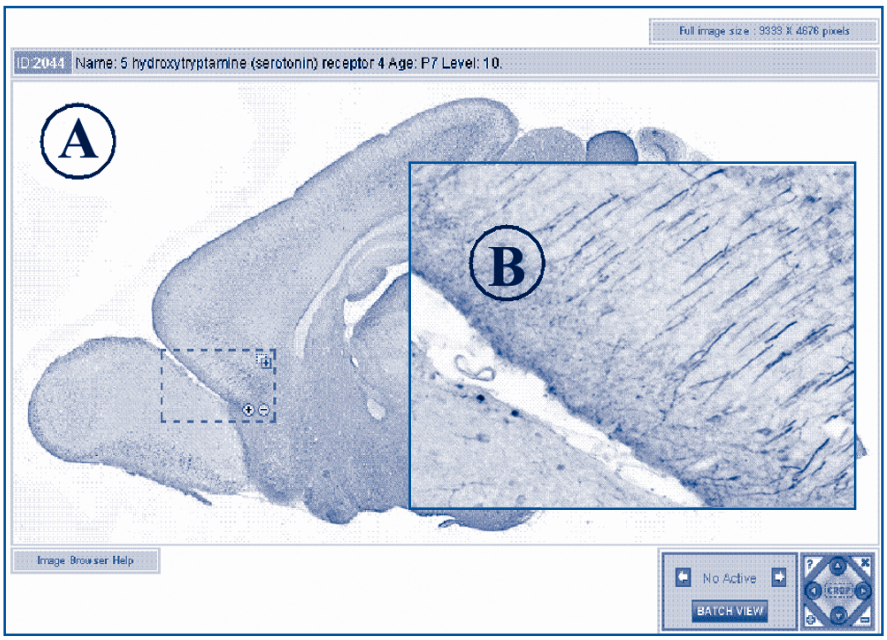

Figure 2. GENSAT image browsing tool. The rectangular area in "A" is used to zoom to the detailed view in "B" showing the expression of the dopamine 4 receptor as dark EGFP density in the elongated axons of neurons. Controls for viewing successive images in the set are visible in the lower right of the figure.

NCBI ToolKit Utility Programs

continued from page 3

the binary ASN.1 Bioseq-set GenBank release files that are available at:

ftp.ncbi.nih.gov/ncbi-asn1

Depending on the "f" argument, the program can produce GenBank and GenPept flatfiles, FASTA sequence files, INSDSet structured XML, TinySeq XML, and 5-column feature table formats. Prior to running asn2all, the GenBank release files, which have an ".aso.gz" suffix, should be uncompressed using a program such as "gunzip", resulting in files with suffix ".aso". For example, gbpri1.aso is the first file in the primate division, and the command:

gunzip gbpri1.aso.gz

will produce "gbpri1.aso"

Using asn2all, the name of the file to process is specified with the "-i" command line argument. Use "-a t" to indicate batch processing of a GenBank release file and "-b T" to indicate that it is binary ASN.1. A text ASN.1 record, such as one obtained on the web from Entrez, can be processed by using "-a a -b F" instead of "-a t -b T".

Nucleotide and protein records within ASN.1 records can be processed simultaneously. Use the "-o" argument to indicate the nucleotide output file and the "-v" argument for the protein output file.

The "-f" argument determines the format to be generated. Legal values of "-f" and the resulting formats are:

- g GenBank (nucleotide) or GenPept (protein)
- f FASTA
- t 5-column feature table
- y TinySet XML
- s INSDSet XML
- a ASN.1 of entire record
- x XML version of entire record

The command:

asn2all -i gbpri1.aso -a t -b T -f g -o gbpri1.nuc -v gbpri1.prt

will generate both GenBank reports for nucleotide sequences and GenPept reports for protein sequences from gbpri1.aso in the files "gbpri1.nuc" and "gbpri1.prt", respectively.

A remote fetching option, "-r T", allows the download of an ASN.1 record from NCBI over a network connection using an accession number or NCBI gi number as an identifier. For instance, to download the feature table within the Reference Sequence record, or RefSeq, for the *Escherichia coli* genome via remote fetch, use:

asn2all -r T -A NC_000913 -f t

The output of this command for the first NC_000913 feature is given below. The 5-column feature table format used is identical to that required as input to generate an ASN.1 sequence file using tbl2asn, described in Box 1.

```
>Feature ref|NC_000913.2|
190 255 gene
                gene            thrL
                gene_syn        EG11277
                locus_tag       b0001
                db_xref
                GeneID:944742
```

The eight ASN.1 utility programs may be downloaded at:

ftp.ncbi.nih.gov/asn1-converters

Box 2. Converting the Entrez Gene FTP files with gene2XML

The gene2xml program is the most recent to join the group of NCBI conversion tools. It reads the binary ASN.1 Entrezgene-Set files offered on the Entrez Gene ftp site and converts them into an easily parsable XML format. The program can accept the name of a single file as input or the path to a group of files to be converted. A option to filter the output by NCBI Taxon Id allows organism-specific XML files to be created from a single multi-species ASN.1 file. The Entrez Gene FTP ASN.1 files are found at:

ftp.ncbi.nih.gov/gene/DATA/

Box 1. Eight NCBI ToolKit utility programs now available in executable for five computer platforms

asn2all: converts GenBank release files in ASN.1 format to a variety of other formats

asn2fsa: converts binary or text ASN.1 sequence files to FASTA format

asn2gb: converts binary or text ASN.1 sequence files to GenBank or GenPept flatfile formats

asn2idx: Generates accession/file offset indices for Bioseq-set release files

asn2xml: converts binary or text ASN.1 sequence files to XML format

asnval: validates ASN.1 release files

tbl2asn: automates the creation of sequence records for submission to GenBank

gene2xml: converts text or binary ASN.1 files of Entrez Gene records into XML

RefSeq Cumulative Updates Discontinued in March 2005

Now that a reliable RefSeq release cycle has been established, the RefSeq cumulative update was discontinued on March 1, 2005. The affected files are located in the directory:

ftp.ncbi.nih.gov/refseq/cumulative

The best method of staying current with the RefSeq database is to download the RefSeq Incremental Update (RIU) products that are available daily:

ftp.ncbi.nih.gov/refseq/daily

Each RIU contains all new or updated RefSeq records processed in the preceding 24 hours or, in the event of a failure, since the most recently generated RIU. Users who wish to maintain a local copy of the RefSeq database typically start by processing a complete RefSeq release, and then downloading and processing the RIUs on a daily basis. RIU processing starts at 2:00am Eastern Time and is usually complete by 3:30am. The recommended time to check and transfer new RIU updates from the RefSeq FTP site is 4:30am. Note that completion times for the RIU can sometimes be several hours later, particularly if a substantial number of complete-chromosome RefSeq records for multiple eukaryotic genomes have been updated on a single day.

RefSeq Release 11

RefSeq Release 11 is now available by anonymous FTP at:

ftp.ncbi.nih.gov/refseq/release

Release 11 includes genomic, transcript, and protein sequences available as of May 8, 2005, from 2,928 organisms. The number of RefSeq accessions in Release 11 and their combined lengths is given in the shaded box.

RefSeq releases are posted bimonthly, and the next release is scheduled for July. Release notes documenting the scope and content of the release are provided at:

ftp.ncbi.nih.gov/refseq/release/release-notes

For more information, visit the NCBI RefSeq Web Site at:

www.ncbi.nih.gov/RefSeq

	# of Accessions	# of Basepairs/Residues
Genomic	651,418	39,048,660,749
RNA	400,504	683,041,613
Protein	1,425,971	507,980,644

Influenza Virus

continued from page 2

Additional resources are under development, including a multiple protein alignment tool that will allow users to statistically analyze and compare sets of protein sequences of the influenza virus. Protein datasets can be visually represented and analyzed using hierarchical clustering with user-selected similarity criteria and clustering algorithms.

Links to Reference Sequences (RefSeqs) of other influenza genomes, protein structures, other protein and nucleotide sequences, and the most recent influenza virus literature in PubMed, are also available from the IVR home page.

[1]Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 32(5), 1792-97.

New Organism in UniGene

UniGene now covers 50 animals and plants and can be searched using the Entrez search system where it is linked to nucleotide records. A recent addition to UniGene is the sweet orange, *Citrus Sinensis* with 49,999 transcript sequences in 5,830 clusters.

GenBank® Release 147

GenBank Release 147 (April 2005) contains over 44 million sequence entries totaling more than 48 billion base pairs. Release 148 is expected in June. GenBank is accessible via the Entrez search and retrieval system. The flatfile and ASN.1 versions of the Release are found in the "genbank" and "ncbi-asn1" directories respectively at:

ftp.ncbi.nih.gov

Uncompressed, the Release 147 flatfiles consume about 185 gigabytes while the ASN.1 version consumes about 145 gigabytes. The data can also be downloaded at the mirror site:

bio-mirror.net/biomirror/genbank

My NCBI
continued from page 1

Getting Started

To try My NCBI, click on the "My NCBI" link under "Hot Spots" on the NCBI home page. Cubby users can log into their My NCBI accounts immediately, using their Cubby user names and passwords. New users can click on the "register for an account" link to open an account. In order for My NCBI to remember your preferences and apply them whenever you log on, your web browser must be configured to accept cookies.

Archiving Search Strategies

Searches are archived by performing an Entrez search and clicking on the "Save search" link that appears to the right of the query box when the results are returned. To see a table of your saved searches, click on the "Saved searches" link in the My NCBI sidebar. Searches that are no longer needed can be selected using checkboxes in the table and deleted. Saved search strategies remain within My NCBI indefinitely, until changed or deleted by you.

Requesting Email Alerts

Search results are emailed on a regular schedule to users who specify 'Schedule' options in the "Saved searches" table. A flexible scheduler accommodates email intervals of days to months and allows the specification of particular days of the week so that results always arrive when you expect them. Report styles include full text, summary, and several other Entrez formats and may be emailed in HTML or plain text. A Document Delivery service to which PubMed will send your orders when you use the "Send to Order" option may be specified by clicking on the "Document delivery" link in the My NCBI sidebar and choosing a service from the list. By default, PubMed sends document delivery orders to Loansome Doc, NLM's document delivery service.

Configuring Filter Tabs

Each of the Entrez databases in My NCBI offers a rich selection of filters from which you can select up to five that automatically partition your search results. To configure filters, click on the "Filters" link on the My NCBI page and select the desired Entrez database. A default set of filters is used for each database in new My NCBI accounts and these can be seen under the "My Selections" tab. To delete a filter, simply uncheck it. The full selection of filters can be examined using the "Browse" tab where they are organized as three hierarchical groups, "Linkouts," "Links," and "Properties". The "Linkouts" group offers tabs for records with links to resources provided by outside organizations. The "Links" group provides filters for records that link to other Entrez databases. Use the "properties" group to create tabs for records that fall into various categories, such as PubMed publication type, Nucleotide molecule type, or database source.

Figure 1 shows filter tabs in use during a search of the Entrez Protein database for "superoxide dismutase." Of the 4410 records returned, accessible via the "All" tab, 626 are RefSeqs, 3075 have links to the Entrez Nucleotide database, 229 have links to the Map Viewer, 456 have links to the Protein Data Bank, and 113 have linkouts to an external organism-specific database, Saccharomyces Genome Database. The latter set is shown in the active tab; the inset shows the result of clicking on the "linkout" link for the first record. To add a filter to your original Entrez query, click on the push-pin shown on the active tab.

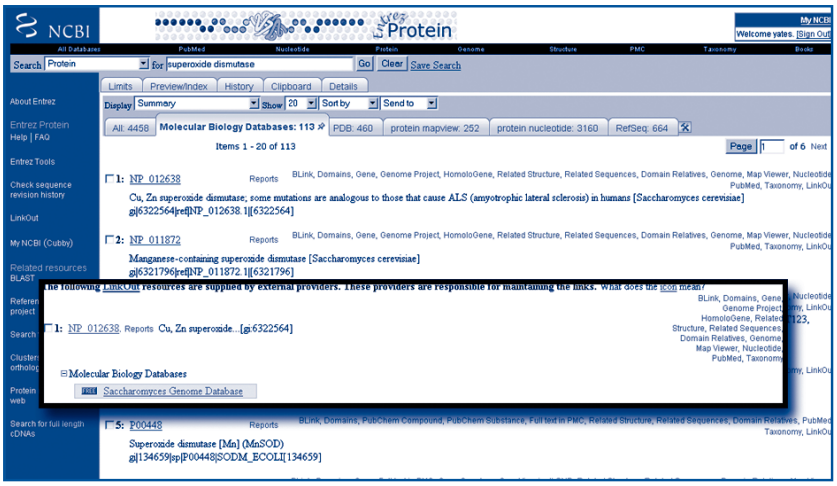

Figure 1. Document summaries returned by a query of "superoxide dismutase" in the Entrez Protein database. The records returned are partitioned into 5 tabbed sets; those with linkouts to organism-specific databases, those derived from the Protein Data Bank, those linked to the Map Viewer, those linked to records in Entrez Nucleotide, and those that are RefSeqs. The linkout target of the first record is Saccharomyces Genome Database. Note that the Entrez links are displayed as "Plain links" rather than as components of a pulldown "links" menu. This configuration option is available under "User preferences" on the My NCBI page.

Configuring Entrez Link Formats

To configure the Entrez link display format, click on "User Preferences"

continued on page 10

ICEMAN
continued from page 3

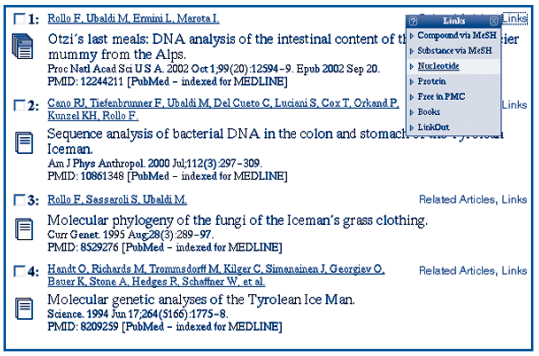

Figure 1. Display of four PubMed abstracts resulting from a query using the phrase 'neolithic mummy AND "pubmed nucleotide"[Filter]'. Each record contains links to corresponding data in other Entrez datasets, as shown from the Links menu.

artifacts, skin surface and intestinal content. Molecular data from the iceman are available in GenBank and integrated into the NCBI's Entrez system.

The simplest way to find the sequences associated with the iceman is by linking to them from a PubMed search. Searching with 'neolithic mummy' finds a dozen articles at the time of this writing. Four of these articles link to sequence records in the nucleotide database. A more precise set of results can be obtained by using the filter that shows only articles with links to nucleotide sequences. The following query finds four articles reporting iceman sequences:

neolithic mummy AND "pubmed nucleotide"[Filter]

In PubMed reference (4), Handt *et al.* report that the iceman's mitochondrial D-loop (hypervariable control region) is consistent with the genetic variation of modern Europeans and is most similar to central and northern European populations. The nucleotide link from the PubMed summary leads to the iceman's D-loop sequence in GenBank record S69989. The Web BLAST service can be used to compare the iceman sequence to selected D-loop sequences from GenBank as shown in Figure 2 in which the iceman sequence is aligned to sequences from a modern European, AY041019, and from a Neanderthal human, AF254446.

The remaining three articles report analyses and identification of non-human DNA associated with the iceman's body and clothing. A diverse taxonomic assemblage of sequences is linked to the PubMed reference (1) article, in which Rollo *et al.* reported the analysis of the iceman's intestinal contents. These sequences provide a partial menu for the iceman's last two meals and show the types of pollen present in his environment.

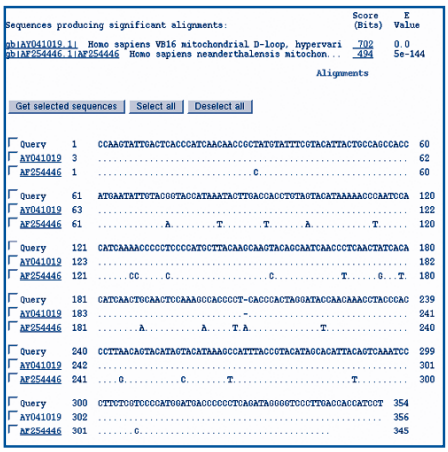

Figure 2. BLAST alignment between the iceman's mitochondrial D-loop sequence, the Query sequence, and those of a modern human, AY041019, and a Neanderthal human, AF254446, respectively. The output shown was created by using web BLAST with an Entrez query to limit the BLAST database to the two target sequences. The results are presented in flat query-anchored format. The iceman sequence is identical to the modern and distinctly different from the Neanderthal sequence.

Following the nucleotide link from the PubMed summary of this article, displays the 15 sequence records reported. These are all conserved sequences of mitochondrial and chloroplast genes amplified by PCR. The biological origin of these sequences was inferred by sequence similarity. In most cases, the authors were able to assign organism sources only to higher taxonomic categories. The red deer (*Cervus elaphus*) and ibex (*Capra ibex*) sequences were confirmed by amplifying and sequencing them both from the iceman and from modern samples of the animals themselves. An interesting view of the taxonomic range of the organisms associated with the iceman with sequence data can be obtained by displaying all taxonomy links and creating a common tree within the Entrez taxonomy page.

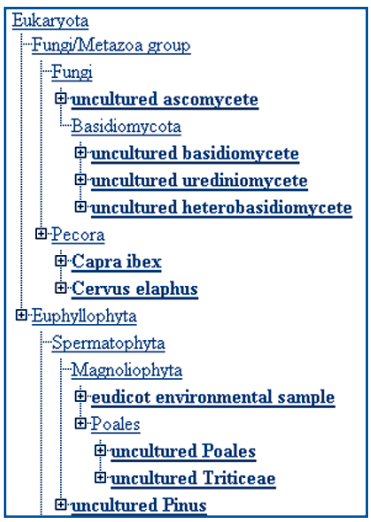

Figure 3. A dendrogram showing the taxonomic classification of the DNA source organisms associated with the iceman.

The authors show that some of the sequence signatures found, such as those for the deer and ibex meat and cereal grain (*Triticeae*), came from food items while others such as the pine (*Pinus*) and fern signatures were likely from unintentionally ingested pollen and spores.

The iceman has been touted by some as one of the most important discoveries of the century. Even though the iceman died over 5,000 years ago, his molecular legacy lives on at the NCBI.

—PC

New Genome Build and Map Viewer Display

Dog Reference Genome, Build 1 Version 1

The Map Viewer now shows dog genome Build 1 Version 1, a 7.6-fold whole-genome shotgun (WGS) assembly of a female boxer produced by the Broad Institute.[1] The dog genome consists of 38 pairs of autosomes and a pair of sex chromosomes, designated X and Y. The mitochondrial genome presented in Build 1.1, with RefSeq accession number NC_002008, is not derived from the boxer used for the WGS assembly but was obtained from a dog of the Sapsaree breed.

Graphical displays of features on the dog genome assembly, as well as radiation hybrid and physical maps, are shown. Displayable map features include NCBI contigs on the 'Contig' track, the WGS contigs on the 'Component' track, and genes, STSs, ESTs, Gnomon predicted gene models, and a radiation hybrid map.[2] Each of the markers on the radiation hybrid map has been integrated into NCBI's UniSTS resource which provides regular updates of positioning of these markers on sequences available in GenBank.

[1]Broad Institute: ww.broad.mit.edu/media/2003/pr_03_tasha.html
[2]Guyon et al. A 1-Mb resolution radiation hybrid map of the canine genome. *Proc Nat Acad Sci USA.* 2003 Apr 29;100(9):5296-301 PMID: 12700351

Towards a Uniform Human Genome Annotation: the Consensus CDS Database

Annotations of genes on the human genome are displayed within several public resources. These annotations are made using different methods, resulting in gene coordinates and sequences that are similar but not always identical. The human genome sequence is now sufficiently stable to begin to compile a standard set of gene annotations on the human genome by identifying those gene placements that are identical. The Consensus CDS (CCDS) project is a collaborative effort to identify a core set of human protein coding regions that are consistently annotated and are of high quality.

The CCDS set is built by consensus among the collaborating members including the European Bioinformatics Institute (EBI), National Center for Biotechnology Information (NCBI), the Wellcome Trust Sanger Institute (WTSI), and the University of California, Santa Cruz (UCSC).

Annotated genes that are included in the CCDS set are given a unique identifier and version number (e.g., CCDS1.1, CCDS234.1) akin to the GenBank "accession.version" system. If the CDS structure changes or if the underlying genome sequence changes, then the version number will be incremented. With annotation and sequence based genome browser update cycles, the CCDS set will be mapped forward, maintaining identifiers. All changes to existing CCDS genes are made by collaboration agreement.

The CCDS set is calculated on the basis of coordinated whole genome annotation updates carried out by the NCBI and Ensembl. To be included in the CCDS set, coding regions must be annotated as full-length, with an initiating ATG and valid stop codon; must be translated from the genome without frameshifts, and must use consensus splice-sites.

continued on page 10

NCBI Courses

NCBI Technical Workshop: Programming with NCBI BLAST® May 25-27, 2005 at the National Library of Medicine, Bethesda, MD

Participants will learn how to create local BLAST databases and keep them current, script BLAST searches using the URL-API, and set up a BLAST Web Server. A working knowledge of NCBI resources and the Perl scripting language is required.

For more information and to apply for the course, see the course Web page at:

www.ncbi.nlm.nih.gov/Class/PowerTools/

NCBI FGPlus: Enhanced Field Guide June 6-7, 2005 at the National Library of Medicine, Bethesda, MD

This expanded course provides detailed coverage of NCBI molecular databases and tools, especially the Entrez system and NCBI BLAST (Web, standalone and client versions). Special emphases of the course are genomic information and molecular structures. The hands-on practical portion is more extensive and includes the "Exploring 3D Molecular Structures" and "Identification of Disease Genes" NCBI course materials.

To register, and for more information:

www.ncbi.nlm.nih.gov/Class/FieldGuide/FGPlus

PubMed® Corrects Spellling

Entrez's PubMed database now includes a spell-check feature that offers spelling corrections for words within PubMed queries that appear to be misspelled. For instance, a misspelled query such as "celular" generates a little over 4,600 hits. However, PubMed now returns a link to the results of a correctly spelled query:

Did you mean: cellular (341,678 items)

Access to the PubMed spell-checker via scripts is provided by a new Entrez Utility called "espell". Espell takes four parameters; "db", "term", "email" and "tool". For example, an Eutility call such as:

eutils.ncbi.nlm.nih.gov/entrez/eutils/espell.fcgi?db=pubmed&term=cardiiac+thromboosis+ischemia

returns the following XML formatted result:

```
<?xml version="1.0"?>
<!DOCTYPE eSpellResult PUBLIC "-//NLM//DTD eSpellResult, 23 November 2004//EN"
"http://www.ncbi.nlm.nih.gov/entrez/query/DTD/eSpell.dtd">
<eSpellResult>
        <Database>pubmed</Database>
        <Query>cardiiac thromboosis ischemia</Query>
        <CorrectedQuery>cardiac thrombosis ischemia</CorrectedQuery>
        <SpelledQuery><Replaced>cardiac</Replaced><Original>
</Original><Replaced>thrombosis</Replaced><Original>
ischemia</Original></SpelledQuery>
        <ERROR/>
</eSpellResult>
```

In the call above, the "db" parameter specifies the "PubMed" database, the only database supported at present, while the "term" parameter gives the search phrase. As with all Eutility calls, spaces within search phrases are represented with "+" signs. The parameters "email" and "tool" are optional, however you may use them to provide an email address that NCBI can use to contact you and to identify your script, respectively. The use of the "email" and "tool" parameters is helpful in cases of a script malfunction.

The result includes the original query, the complete corrected query, and a breakdown of the terms in the complete corrected query flagged as either "replaced" or "original".

Terms that are field-restricted, such as "heart[title]", are not checked since each field implies the use of a separate vocabulary that may include standard abbreviations that would be flagged as misspelled words in the context of a more general usage.

To read more about the Entrez Utilities, or to subscribe to the "utilities-announce" mailing list, see:

eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

CCDS Database
continued from page 9

Annotations are made via a mixture of manual curation and automated computational processing. Genome annotations resulting from the NCBI and Ensembl pipelines are first compared to identify annotated coding regions that have identical locations on the genome. Then, lower quality CDSs from this core set are removed pending additional review among the collaboration groups. Quality tests include analysis to identify putative pseudogenes, retrotransposed genes, consensus splice sites, supporting transcripts, and protein homology.

As of March 2005, the initial CCDS dataset contains 14,795 coding sequences and 13,142 genes, representing more than half of the human genes, according to the current gene number.

Visit the CCDS Project Web site at:

www.ncbi.nlm.nih.gov/CCDS/index.html

My NCBI
continued from page 7

in the blue side bar and choose the format you prefer from the "Links Display" menu. Choices are "javascript menu," the default, "plain links," "standard pull-down," and "pop-up menu." Figure 1 shows an Entrez display in which links have been configured as "Plain Links" using My NCBI, and are therefore visible for each record without the need to first click on a "Links" link.

—MR

Using seedtop to find patterns in protein and nucleotide sequences

Seedtop is one of the programs included within the NCBI standalone BLAST package and is used to find matches to a pattern in a protein or nucleotide sequence database.

Using seedtop to locate a pattern in protein and nucleotide sequences

Seedtop, like blastall and formatdb, is a commandline program with parameters specified with a leading dash, followed by a one-letter parameter code. To find a pattern in a protein sequence, we may use:

```
seedtop -i input -k pat -p patmchp -o pat_out
```

The file "pat" contains the pattern for a serine protease motif:

```
ID Serine Protease Motif, cd00190
PA C-[AVLS]-X(3,9)-[DSNAR]-X-[CG]-X-[GSR]-[DE]-[SAPG]-G-[GS]-[PAG]-[LFMV]
```

The file "input" contains the sequence for human kallikrein 1 pre-proprotein, NP_002248.1, then, upon completion of the search.

The parameter "-p", for "program mode", specifies that the search will be of a protein sequence; for nucleotide sequences use "-p pattern".

The result of the search will be placed in file "pat_out":

```
Name  Serine Protease Active Site, cd00190
Pattern C-[AVLS]-X(3,9)-[DSNAR]-X-[CG]-X-[GSR]-[DE]-[SAPG]-G-[GS]-[PAG]-[LFMV]
At position 219 of query sequence
```

The format of the pattern file is rigid. There must be an id line starting with the letters "ID", followed by a space and some identifying free form text. The second line must begin with the letters "PA", followed by a space and the pattern, in Prosite format.

To find pattern matches to a set of protein sequences requires an extra step, in which a file with multiple FASTA sequences is formatted with formatdb, another program from the standalone blast package. If the sequences to be searched are in the file "proteases", we begin by formatting the file using:

```
formatdb -i proteases -pT -oT
```

To perform a batch pattern match, we will call the database using "-d database" in lieu of the "-i filename" option. The following command line searches the database "proteases" and identifies a pattern specified in the file "active_site", the identified matches and their coordinates are saved in "pat_out":

```
seedtop -d proteases -k active_site -p patternp -o pat_out
```

As shown by a partial result below, the target sequence with a match is marked by the "seqno=" line. This is followed by the reiterated pattern and matching positions marked by the first and last numbers in the "HI" field. The actual coordinates should be incremented by 1 due to offset by zero. For the match below, the actual coordinates are 198 and 287.

```
seqno=254       gi|5031829|ref|NP_005542.1|
ID  Serine Protease Motif, cd00190
PA  C-[AVLS]-X(3,9)-[DSNAR]-X-[CG]-X-[GSR]-[DE]-[SAPG]-G-[GS]-[PAG]-[LFMV]
HI (197 268) (276 276) (278 278) (280 286)

seqno=257       gi|4504875|ref|NP_002248.1|
```

Since the database has been formatted using the -oT option, we can retrieve the sequence of the matched region using fastacmd from standalone blast package:

```
fastacmd -d proteases -s 45580723 -L 282,301
```

which returns the subsequence containing the pattern:

```
>gi|45580723:282-301 haptoglobin-related protein ...
CVGMSKYQEDTCYGDAGSAF
```

In addition to protein pattern matches, seedtop can search for patterns in a nucleotide sequence or database. A pattern which is specified using IUPAC codes needs to be converted to ProSite syntax. To search a single nucleotide sequence for given patterns, use:

```
seedtop -i input_seq -p patmatch -k pat_file
```

To search a nucleotide database for given patterns, use:

```
seedtop -d database -p pattern -k pat_file
```

For questions regarding the BLAST services, please contact the BLAST help desk at:

blast-help@ncbi.nlm.nih.gov

—TT

LocusLink Retired

NCBI has stopped providing the LocusLink interface as of March 1, 2005. Only one file on the LocusLink ftp site is still being updated (LL_tmpl.gz), and the update of that file will end on June 1, 2005.

During the transition period, subsets of data were removed from LL_tmpl to be reported only from Entrez Gene. These include GeneRIF data and genes from *Drosophila melanogaster* and *Caenorhabiditis elegans*. Also, new gene-specific data are being processed that are reported only in Entrez Gene. This includes (1) links to pathway information from KEGG for genomes other than human, (2) links to Reactome, and (3) access to GENSAT (Gene Expression Nervous System ATlas) (mouse only):

www.ncbi.nlm.nih.gov/projects/gensat

Records in Entrez Gene that have expression data viewable from GENSAT can be retrieved from Gene by the query:

gene_gensat[filter]

Entrez Gene maintains several files, described in the shaded box, to help support the transition from use of LocusLink.

If you are interested in being notified when changes are made to Entrez Gene, subscribe to gene-announce:

www.ncbi.nlm.nih.gov/mailman/listinfo/gene-announce

New Files Added to Entrez Gene

1. The Gene README file
   ftp.ncbi.nih.gov/gene/README
2. Gene Help
   www.ncbi.nlm.nih.gov/entrez/query/static/help/genehelp.html#gene_ftp
3. The LocusLink->Gene transition file.
   www.ncbi.nlm.nih.gov/entrez/query/static/help/LL2G.html
4. gene2xml, described in this README:
   ftp.ncbi.nih.gov/asn1-converters/documentation/gene2xml.txt
   which facilitates converting the ASN.1 extraction of Entrez Gene to XML.

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 3S308
8600 Rockville Pike
Bethesda, Maryland 20894

Official Business
Penalty for Private Use $300

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