NCBI News

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

August 1997

Vice President Launches PubMed, Lauds Free MEDLINE Access

"MEDLINE...will henceforth be available free to the American people." With those words, Vice President Al Gore inaugurated the PubMed search system at a Capitol Hill press conference on June 26. PubMed, which provides Web access to the National Library of Medicine's (NLM) database of the biomedical journal literature, MEDLINE, was heralded by Senator Tom Harkin (IA) as "...the model of a smart, creative government initiative." The Vice President viewed free access to MEDLINE as consistent with the Clinton administration's other "empowerment" initiatives stating, "This development...may do more to reform and improve the quality of health care in the United States than anything else we've done in a long time."

Searching PubMed

PubMed grew out of NCBI's Entrez project which, since 1992, has offered a subset of MEDLINE records related to molecular biology. In addition to encompassing all of MEDLINE and PreMEDLINE, PubMed retains Entrez's ability to use one article as a "seed" to find other similar articles. By traversing the **See Related Articles'** links, a user can find articles similar in concept with speed and precision. PubMed expands upon Entrez by linking MEDLINE articles to full-text Web sites maintained by publishers. Currently, 95 journals are linked to PubMed, including Cell, Journal of Biological Chemistry, Journal of Cell Biology, New England Journal of Medicine, and Science. Access to publishers' Web sites may require subscriptions or registration.

NCBI Director David Lipman (far left) coaches Vice President Gore (seated) as he searches PubMed. NIH Director Harold Varmus (center) and NLM Director Donald Lindberg (far right) look on.

PubMed Options

PubMed offers the option to search MEDLINE or any of NCBI's molecular biology databases. Users can select from a variety of search fields, including but not limited to: text words, author names, and journal titles. A MEDLINE citation for which there is a corresponding online, full-text article will have a button at the top of the abstract page that links to the publisher's Web site. Additional links point to

Continued on page 2

IN THIS ISSUE

NCBI News

NCBI News is distributed two to three times a year. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence and suggestions to *NCBI News* at the address below.

NCBI News
National Library of Medicine
Bldg. 38A, Room 8N-803
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 496-2475
Fax: (301) 480-9241
E-mail: info@ncbi.nlm.nih.gov

Editors
Dennis Benson
Barbara Rapp

Design Consultant
Troy M. Hill

Photography
Karlton Jackson

Writing, Editing, Graphics, and Production
Veronica Johnson
Donna Roscoe

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 to perform research in computational molecular biology.

NIH Publication No. 97-3272

ISSN 1060-8788

PubMed, *continued from page 1*

other NCBI databases, including sequences, 3D structures or OMIM. Advanced query options allow for the creation of more complex Boolean search expressions, and a special clinical query page is optimized to perform searches for studies relating to the etiology, diagnosis, prognosis, or treatment of human diseases.

PubMed is available from the NCBI World Wide Web home page (http://www.ncbi.nlm.nih.gov). Comments and questions about PubMed are welcome. Send e-mail to info@ncbi.nlm.nih.gov or call (301) 496-2475. ■

❖ ❖ ❖ ❖ ❖

Using Sequin to Submit Sets of Related Sequences

Sequin is a program developed at the NCBI for submitting DNA sequences to GenBank, EMBL, or DDBJ. Both Sequin and BankIt, NCBI's Web-based sequence submission tool, can be used to submit simple mRNA or genomic sequences along with associated coding sequences. However, Sequin has been outfitted with a number of advanced sequence analysis capabilities. Unlike other sequence submission tools, Sequin can process sets of related sequences such as segmented sets and those generated by phylogenetic, population, or mutation studies.

Like other World Wide Web submission tools, Sequin can be used to annotate single sequences. However, it is usually easiest to annotate related sequences when they are part of a multiple sequence alignment. Sequin can import the individual sequences, as well as the alignment itself, from alignments that have been saved in FASTA+GAPs, PHYLIP, or NEXUS format. If the sequences are related, but not yet aligned, Sequin will generate an alignment from a file of FASTA-formatted sequences. Each new sequence in the alignment will receive its own accession number.

After the alignment is generated, annotation features, such as a coding sequence or rRNA, can be marked just once on a single master sequence. These features can then be propagated from the master sequence to other sequences in the alignment. The proper location of the feature will be calculated by Sequin for each sequence individually after taking into account any gaps or insertions. The Entrez nucleotide database is now accessible from Sequin, allowing sequences from GenBank to be directly downloaded into Sequin from Entrez. If the GenBank sequence is related to the sequences in the alignment, it can be brought into the alignment as the master sequence. Features can then be copied from this master sequence onto the new sequences. The GenBank record will not receive a new accession number, but rather serves only to facilitate annotation of the newly submitted sequences.

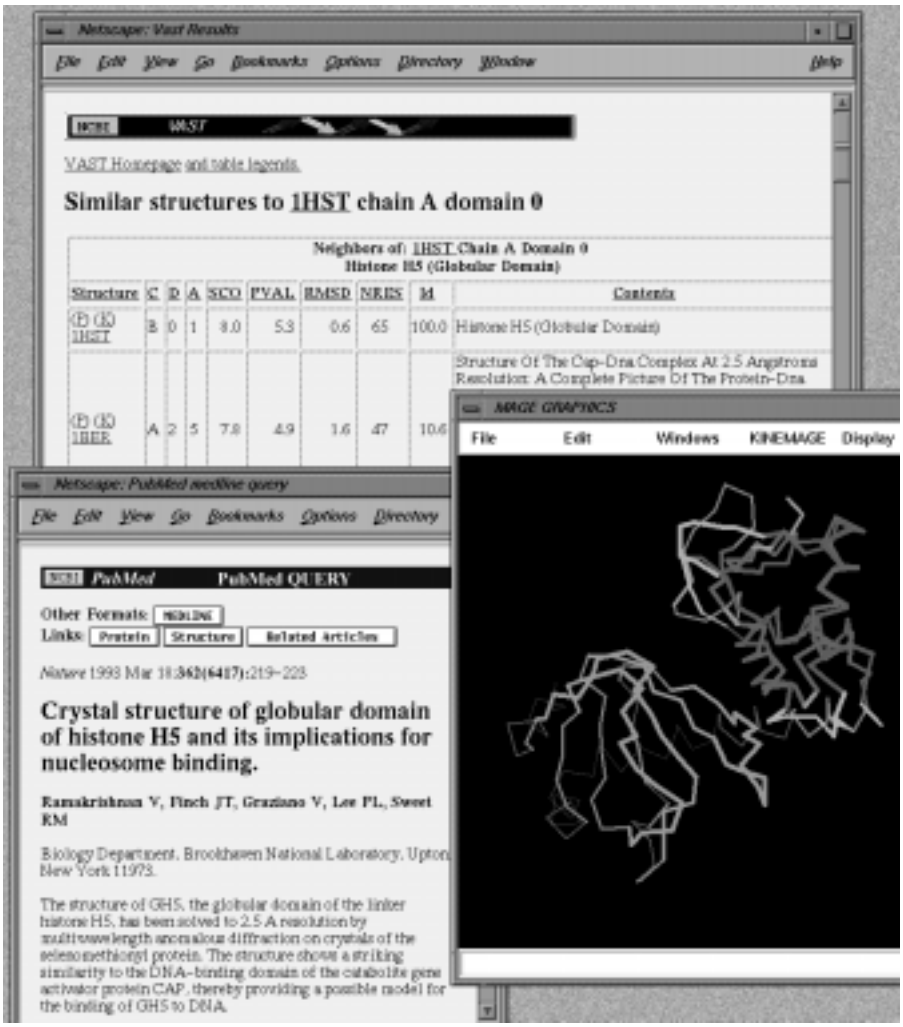
Further information about Sequin, including downloading instructions and help documentation, are available from the Sequin Web page (http://www.ncbi.nlm.nih.gov/Sequin). ■

Structure Neighbors in Web Entrez

WWW Entrez now contains "neighbors" for proteins in its 3D structure database. Structure neighbors are other proteins that have a similar 3D structure or shape. As with the protein sequence neighbors in Entrez, structure neighbors are most often homologs with similar biological functions. However, since protein evolution conserves 3D structure to a greater extent than sequence, a protein's structure neighbors may include more distant relatives not present among its sequence neighbors. These additional similarities may provide further insight into a protein's properties and biological function. By incorporating structure neighbors in Entrez, these distant relationships, detectable only by 3D structure comparison, are readily accessible to molecular biologists.

An example is provided by the globular domain of chicken histone H5 (PDB accession code 1HST). The sequence neighbors of histone H5 are all other histones from a variety of eukaryotic species. But the structure neighbors of histone H5 are diverse and include a number of DNA binding proteins from bacteria. One of these is the E. coli catabolite gene activator protein, or CAP, in complex with DNA (PDB accession code 1CGP). The structures are remarkably similar, with 46 amino acid residues of histone

Continued on page 7



Chicken histone H5

NCBI Data by FTP

The NCBI FTP site contains a variety of directories with publicly available databases and software. The available directories include 'repository', 'genbank', 'entrez', 'toolbox', 'pub', and 'sequin'.

The **repository** directory makes a number of molecular biology databases available to the scientific community. This directory includes databases such as PIR 53.0, SwissProt, CarbBank, AceDB, and FlyBase.

The **genbank** directory contains files with the latest full release of GenBank, the daily cumulative updates, and the latest release notes.

The **entrez** directory contains the client software for Network Entrez.

The **toolbox** directory contains a set of software and data exchange specifications that are used by NCBI to produce portable software, and includes ASN.1 tools and specifications for molecular sequence data.

The **pub** directory offers public-domain software, such as BLAST (sequence similarity search program). Client software for Network BLAST and PowerBlast is also included in this directory.

The **sequin** directory contains the new Sequin submission software for Mac, PC, and UNIX platforms.

Data in these directories can be transferred through the Internet by using the Anonymous FTP program. To connect, type: **ftp ncbi.nlm.nih.gov.** Enter **anonymous** as the login name, and enter your e-mail address as the password. Then change to the appropriate directory. For example, change to the repository directory (cd repository) to download specialized databases.

Selected Recent Publications by NCBI Staff

Altschul, SF, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and **DJ Lipman**. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–402, 1997.

Baxevanis, AD and **D Landsman**. Histones and histone fold sequences and structures: a database. *Nucleic Acids Res* 25:272–3, 1997.

Galperin, MY. Sequence analysis of an exceptionally conserved operon suggests enzymes for a new link between histidine and purine biosynthesis. *Mol Microbiol* 24:443–5, 1997.

Leipe, DD. Biodiversity, genomes and DNA sequence databases. *Curr Opin Genet Dev* 6:686–91, 1996.

Makalowski, W. Mermaid: a not-so-new family of human repetitive elements. *Hum Genet* 99:696–7, 1997.

Mushegian, AR and **EV Koonin**. Sequence analysis of eukaryotic developmental proteins: ancient and novel domains. *Genetics* 144:817–28, 1996.

Neuwald, AF, DJ Liu, **DJ Lipman,** and **CE Lawrence**. Extracting protein alignment models from the sequence database. *Nucleic Acids Res* 25:1655–77, 1997.

Schuler, GD. Sequence mapping by electronic PCR. *Genome Res* 7:541–50, 1997.

Schuler, GD, MS Boguski, EA Stewart, LD Stein, G Gyapay, et al. A gene map of the human genome. *Science* 27:540–6, 1996.

Wolfsberg, TG and **D Landsman**. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res* 25:1626–32, 1997.

Zhang, Z and **TL Madden**. PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res* 7:649–56, 1997.

Hunting For Open Reading Frames With ORF Finder

Searching for open reading frames is possible with NCBI's software tool, ORF Finder, accessible from the NCBI World Wide Web home page (http://www.ncbi.nlm.nih.gov). ORF Finder is a graphical analysis tool which finds all open reading frames in a user's sequence or in a sequence retrieved from a database. ORF Finder also provides easy access to the BLAST search page and allows the deduced amino acid sequence to be compared against additional amino acid sequence databases using the BLAST options.

To use ORF Finder, enter the accession or GI number of the sequence of interest, or enter your query sequence directly into the text box in FASTA format. ORF Finder will identify all open reading frames using the standard genetic code or an alternative one for translation. Users can limit the search for open reading frames to a portion of the query sequence by specifying the positions (in base pairs) in the "From" and "To" boxes. Press the **ORF Find** button to retrieve a graphic display of ORFs and their location in the sequence in 6 reading frames. Users have the option to change the minimum ORF length to 50 or 300 nucleotides (in base pairs) and **Redraw** the query sequence. The **Six Frames** option features a graphic of all start and stop codons. Select a particular ORF by clicking on it to see the amino acid sequence with all alternative start codons. After selecting a particular ORF of interest, click on the **Accept** button and have the option to view the ORF in various formats: GenBank flat-file, FASTA nucleotide, or FASTA amino acid sequence. Selecting **View** retrieves the full GenBank record with its annotated sequence information.

For those scientists submitting sequence data, ORF Finder is also packaged with the Sequin sequence submission software. ORF Finder can be used in conjunction with Sequin's Sequence Editor to annotate new coding regions on the record, perform basic editing, and translate nucleotide sequences. The Sequin program can be downloaded from NCBI's FTP site accessible from the NCBI WWW home page. ■

❖ ❖ ❖ ❖ ❖

Mapping Unique Genome Sites by Electronic PCR

It is possible to determine the gene map location of a new sequence using NCBI's software tool, Electronic PCR (e-PCR), located on the NCBI World Wide Web home page. Electronic PCR simulates conventional PCR methods for identifying sequence tagged sites (STSs) by searching for sites in a query sequence which match the sequence and orientation of a set of primers. STSs are unique DNA landmarks used in the construction of genetic and physical maps of the human genome. The e-PCR tool searches for matches between a user's query sequence and STS primer sequences in the STS database (dbSTS). Researchers can use e-PCR to assign

Continued on page 7

CGAP Revolutionizes Cancer Research

The knowledge that genetic mutations are central to the development of cancerous cells has prompted the National Cancer Institute, in partnership with NCBI, and other government, academic and industry leaders, to initiate the Cancer Genome Anatomy Project, or CGAP.

CGAP merges state-of-the-art technologies in pathology, molecular biology and bioinformatics, to catapult a new strategic attack on cancer. It is an unprecedented assemblage of sequence information characterizing the genetic constitution of cells at various stages: normal, precancerous, and tumor.

Worldwide, participants in CGAP are collecting a variety of tissue samples from cells at different stages; generating cDNA libraries from the tissue samples; and sequencing the cDNA libraries. NCBI uses powerful sequence similarity searching tools, such as BLAST, to make electronic comparisons between the libraries of a given tissue type at different stages, and generate discrete lists of genetic candidates as causative components of the carcinogenic process.

CGAP has collected over 40,000 DNA sequences so far in its trek over the next few years toward a complete index of genes expressed in tumors—referred to as the Tumor Gene Index. Initially, this index will be compiled from five major cancers: prostate, breast, lung, colon, and ovarian. NCBI will continue to map new index sequences to the Human Genome, building upon NCBI's unique collection of human gene sequences (UniGene) used to construct the Human Transcript Map.

The focal point of the CGAP project is its Web site, located at www.ncbi.nlm.nih.gov/ncicgap. Managed and supported by NCBI members Mark Boguski, Ken Katz, Greg Schuler, and Carolyn Tolstoshev, the CGAP Web site is the central repository for all of the information generated by the project. This includes tissues, libraries, sequences, and links to additional value-added information, such as related DNA and protein sequences, genome mapping data, and biomedical references.

Ultimately, the resourceful use of information housed in the CGAP Web site is expected to lead to innovative diagnostic, preventative, and curative technologies which will forever alter the way scientists conduct cancer research. ■

NCI **CGAP** Cancer Genome Anatomy Project **NCBI**

Comparison of Normal versus Tumor Prostate Cell Gene Expression

Normal	Precancerous	Malignant	Gene index	Gene description
0.0163	0.0330	0.0078	Hs.1548	Prostate specific antigen (APS)
0.0163	0.0024	0.0104	Hs.73487	Beta-microseminoprotein (prostate secreted) (MSMB)
0.0000	0.0000	0.0156	Hs.82186	V-erb-b2 avian erythroblastic leukemia viral oncogene homolog 3 {alternative products} (ERBB3)
0.0000	0.0000	0.0130	Hs.5417	ESTs, Weakly similar to F43E2.7 [C.elegans]
0.0069	0.0071	0.0000	Hs.62954	Ferritin heavy chain (FTH1)
0.0050	0.0071	0.0000	Hs.38972	ESTs, Weakly similar to CD63 ANTIGEN [H.sapiens]
0.0000	0.0047	0.0000	Hs.18910	ESTs

CGAP displays a list of genes with statistically significant expression differences. Dot intensity is proportional to relative frequency of EST expression.

Frequently Asked Questions

I submitted a sequence using BankIt one month ago, however I have not yet received a GenBank accession number. Why?

Since BankIt submissions normally receive GenBank accession numbers within 24-48 hours, an error in the submission process most likely occurred. Submitting sequence information by BankIt involves completing, reviewing, and submitting your BankIt file. Once you have finished entering your data and information, and have reviewed it for accuracy and completeness, switch the selection from "Modify Submission" to "Submit to GenBank" on the final BankIt page and click on the BankIt button one last time. Users will receive a return message indicating receipt of the submission and a GenBank accession number soon thereafter.

I have a list of interesting titles I retrieved using PubMed. How do I obtain the full documents?

The PubMed system provides access to bibliographic citations and corresponding abstracts, but does not contain the full-text of articles. However, PubMed offers links to a number of publishers who provide access to full-text journal articles from their Web sites. PubMed displays the link at the top of the Display title/abstract page when available. Currently the number of participating journals is small and the journals may require a user to subscribe before being able to view the full-text. A list of PubMed journals that offer full-text can be retrieved from the URL: http://www.ncbi.nlm.nih.gov/PubMed/fulltext. If the journal you are interested in is not on the list, contact the nearest library for information about obtaining articles.

How can I import references from PubMed into Endnote?

First choose the MEDLINE format on the "document summaries page" and then display. Select the appropriate save format at the bottom of the display screen and press the "Save" button.

I recently read an article that referenced a GenBank accession number, but I can't find the sequence record in the database. Why?

Sometimes authors request that a sequence be held confidential until publication. Once GenBank is informed that the sequence has been published, GenBank staff will verify the publication and release the sequence. GenBank encourages users who are unable to retrieve a record to send the accession number and complete citation in which it appeared to update@ncbi.nlm.nih.gov.

How can I find out if a particular gene has been mapped?

Conduct a GenBank search using common names for the gene. If the gene has been sequenced, its sequence record can be retrieved from GenBank. The accession number obtained from that record can be entered into the text search tool of the Human Transcript Map (available from the NCBI home page). Please note that the number of transcripts mapped in this study is estimated to represent one-fifth of the total number of genes in the human genome so the odds are that a gene has not yet been mapped.

Continued on next page

***Structure Neighbors**, continued from page 3*

H5 superimposing to 1.7 angstroms residual, even though sequence identity is only 10%. These proteins would appear to be homologs, and the protein-DNA structure of CAP suggests a model for the interaction of histone H5 with DNA.

Structure neighbors may be accessed from the "Structure Summary" of a protein in WWW Entrez's 3D structure database. Neighbor lists are displayed when one clicks the button **Protein 3D Structures** in the line reading "Protein 3D Structures similar to <Chain A> computed by VAST." Here **<Chain A>** is a pull-down menu that allows one to select the individual polypeptide chain, or compact domain within that chain, for which structure neighbors are to be retrieved. Structure neighbor lists have been computed by the VAST algorithm (Vector Alignment Search Tool[1]), and are sorted according to a VAST similarity score. VAST compares the relative orientations of helices and beta-strands in two protein domains, and if similarity is more extensive than one would expect by chance, produces a detailed structure alignment by comparison of atomic coordinates.

WWW Entrez also supports visualization of protein structures by molecular graphics. The Cn3D viewer may be started by the **View** button on the "Structure Summary" page. The 3D superpositions of structure neighbors can be viewed using the **Kinemage** button on neighbor-list pages (viewing programs such as Cn3D and Kinemage are helper applications that may be downloaded to your computer by following hotlinks on the Structure home page). 3D viewing is important for interpretation of structural similarities. In the 1HST protein, for example, one may see that histone H5 contains positively charged residues in the region that superimposes onto the DNA-binding interface of CAP as does CAP itself. 3D viewing thus supports the inference that these proteins "dock" with DNA in a similar manner. Possible functional similarities of structure neighbors may also be explored, of course, by examining the MEDLINE citations and sequence neighbors associated with each protein.

WWW Entrez's structure neighbor service is the work of NCBI researchers T. Madej, C. Hogue, J.-F. Gibrat, J. Spouge, H. Ohkawa and S. Bryant. Improvements in the visualization of structure similarities are still in progress, and comments and suggestions are welcome.

[1] Gibrat J-F, Madej T, Bryant SH: Surprising similarities in structure comparison. *Curr Opin Struct Biol* 1996, 6:377–85. ■

***Electronic PCR**, continued from page 4*

sequence database records to map positions, test primer feasibility, and integrate and anchor genetic maps and sequence data.

To map sequences by e-PCR, enter the sequence of interest in FASTA format into the text box or retrieve a sequence from GenBank using an accession or GI number. The "Retrieve STS from" setting can be used to limit the search to STSs from specific organisms. Press the **Submit Query** button to begin. The results list the STSs found with their relevant identifiers, position of the primer binding sites within the query sequence, chromosome number (if known), and the expected and observed size of the amplicon. Hypertext links to GenBank and dbSTS records (linked to Entrez) are provided for more detailed information.

The number of STS results one can expect for a typical search depend on a variety of factors, such as length of the query sequence and size of the STS database. Results that are reported are unequivocal and more reliable than those identified using the general-purpose database search tool BLAST. Chances of obtaining STS matches will improve as the number of sequences in the STS database continues to increase dramatically.

For more information contact info@ncbi.nlm.nih.gov. ■

***Frequently Asked Questions,** continued from page 6*

I am interested in performing non-redundant BLAST searches on some sequences I have. Is there a way to do this for new GenBank entries on a regular basis?

Perform one BLAST search using the nonredundant database, nr. After that, you can use BLAST to search the month database only. The month database has all new records from the last month and is obviously much smaller than nr. It was intended for this kind of surveillance blasting.

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