NCBI News

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

August 1996

See in 3D: New Entrez Release 5.0

Since September 1995, Network Entrez has included 3D structure data, based on crystallographic and NMR structure determinations. The structure data are contained in NCBI's Molecular Modeling DataBase (MMDB), which is derived from the Brookhaven Protein DataBank of more than 4,000 biomolecules. MMDB is also referred to as the Structure division of Entrez.

With the release of Entrez 5.0 in July 1996, NCBI has added a new built-in 3D-structure viewer called Cn3D ("See in 3D"). Cn3D allows one to visualize and rotate protein structure records from Entrez. Structure data can provide a wealth of information on the biological function and mechanism of action of macromolecules. By fully integrating the structure database into Entrez, we hope to make this information easily accessible to biologists.

Searching for Structures

Finding a structure in Entrez is just like any other Entrez search. A query can contain specific fields such as author names or text terms occurring anywhere in the structure description. In this way you may check for structure data on a specific protein or nucleic acid. For example, select the "structure" database from Entrez's search page, enter a search term like "copper," then press the **Retrieve Documents** button to bring up the list of 3D structure entries matching your query. To see the 3D structure, double click on the 3D icon of any record you want to display.

A more powerful search approach, however, is to select the molecule of interest in the sequence database, identify its sequence neighbors (candidate homologues), and then, by linking to the structure database, ask whether structure data is available for any of the members of this family. The structure database is smaller than the protein or nucleotide databases, but many sequenced proteins have

Continued on page 2

Advancing Genomic Research: The UniGene Collection

The UniGene collection, now accessible through NCBI's Home Page, contains more than 48,000 clusters of sequences, each representing the transcription product of a distinct human gene. With current estimates of 80,000 to 100,000 genes in the human genome, this is close to the 50% mark. The clusters are largely based on EST sequences, so most of the sequences are not complete and most of the genes have still not been characterized. But one important use of the UniGene clusters is to identify novel, nonredundant mapping candidates for generating a transcript map that identifies all coding sequences in the genome.

Although a primary goal of the Human Genome Project is to determine the complete sequence of the 3 billion base pairs in the human genome, only about 3% of the genome actually encodes protein, and the biological significance of most of the sequence that will be generated is not known. Therefore, a transcript, or expression, map is a critical resource for charting the way.

Until a few years ago, GenBank contained sequences for only 3,000 unique human genes, and developing a transcript map did not seem worthwhile based on such a small sample. But recent advancements in EST technology and the increased public availability of EST sequences have dramatically increased the numbers of genes in GenBank, so that developing a dense transcript map is now

Continued on page 10

In This Issue

NCBI News

NCBI News is distributed 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
Deborah Loer-Martin
Wendy B. Osborne

In 1988, Congress established the National Center for Biotechnology Information as part of the National Library of Medicine; its charge is to create automated systems for storing molecular biology, biochemistry, and genetics data, and to perform research in computational molecular biology.

The contents of this newsletter may be reprinted without permission. The mention of trade names, commercial products, or organizations does not imply endorsement by NCBI, NIH, or the U.S. Government.

NIH Publication No. 96-3272

ISSN 1060-8788

See in 3D*, continued from page 1*

homologues in this set, and you may often learn more about a protein by examining the 3D structure of its homologues.

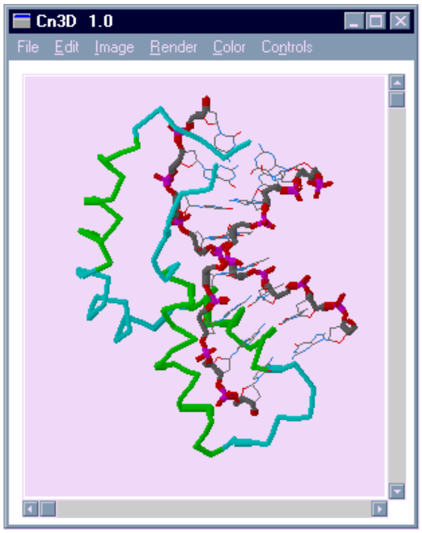
Using Cn3D From WWW Entrez

WWW users will need to download and install the Network Entrez client software and configure it as a helper application for their WWW browser. When a 3D structure is requested from WWW Entrez, the browser will automatically launch Cn3D.

Detailed instructions for installing the program, getting started, and using the viewing features are provided on the Cn3D Web page (http://www.ncbi.nlm.nih.gov/Structure/cn3d.html). If you installed your own WWW browser and your Internet connection, you can probably install Network Entrez without difficulty. For assistance, first check with a systems administrator at your institution before contacting NCBI.

Getting the Software

Entrez 5.0 with Cn3D is available for many platforms, including Mac, Windows, and UNIX. It can be downloaded from NCBI's FTP site (ncbi.nlm.nih.gov) in the 'entrez/network' directory. For installation instructions, be sure to download the README document, or see the Entrez Overview section from WWW Entrez.



3D structure of human Sry-DNA complex (PDB accession: 1HRY)

The current version, numbered 5.002, is still considered a "beta" release. There will be a series of software updates throughout the rest of the year, so check the FTP site periodically to make sure you have the most up-to-date version. We are still refining the program and welcome comments and suggestions (info@ncbi.nlm.nih.gov). ■

❖ ❖ ❖ ❖ ❖

Entrez CD-ROM Discontinued

Users are reminded that effective August 15, 1996, NCBI is discontinuing Entrez on CD-ROM. Two versions of Entrez are available free of charge over the Internet. Network Entrez is a client/server program that retains the look and feel of Entrez on CD-ROM. Client software for PC/Windows, Macintosh, and several Unix workstations can be downloaded by FTP from 'ncbi.nlm.nih.gov' in the 'entrez' directory. There is also a World Wide Web version of Entrez, accessible from NCBI's Home Page (http://www.ncbi.nlm.nih.gov). This version has essentially the same functionality as Network Entrez, but with a different search and display interface. ■

QUERY: A New E-Mail Server for Entrez

NCBI now has an e-mail server specifically designed to do text-based searches of the integrated Entrez database. As with the RETRIEVE e-mail server that has been in place for several years, users specify a data set to search, then the words or ID numbers to be used in the search. However, the new server offers a choice of output options and provides access to all the information from the various databases that make up Entrez. Some of these data, such as the molecular biology subset of MEDLINE and protein sequences entered directly from the published literature, are not available through the older RETRIEVE server.

QUERY uses the Entrez search engine so important Entrez features, such as viewing sequence neighbors or linking to associated information such as MEDLINE abstracts, are now also available through an e-mail search interface.

To use the QUERY server, send a formatted e-mail message to the address: query@ncbi.nlm.nih.gov. Your search results will be returned to you as an e-mail message.

To format a search, first specify the database (DB) to be searched: **n** for nucleotide sequences, **p** for protein sequences, **s** for both nucleotide and protein sequences, **t** for 3D structures, or **m** for the molecular biology subset of MEDLINE.

Next, specify your search term, and indicate whether it is a unique identifier for a record (UID) or a text term from elsewhere in the record (TERM). UIDs include sequence database accession numbers, sequence-specific GI numbers, and MEDLINE accession numbers. Search terms can also be restricted to specific fields such as organism, author, title, journal name, or date. In addition, you can combine search terms with Boolean logic operators.

Finally, specify a particular output format if desired, and include any other optional search specifications, such as the maximum number of records to display. Display options include such formats as FASTA or GenBank flat file, but also are used to specify that you want to see related information such as sequence neighbors or MEDLINE abstracts.

Some sample search queries are shown below. For more detailed information on formatting searches and available search options, review the QUERY server documentation. To obtain the documentation, send the word HELP as your message to the server (query@ncbi.nlm.nih.gov).

Questions or comments about the QUERY server are welcomed, and should be sent to the user support group at info@ncbi.nlm.nih.gov. ■

Sample Searches for QUERY E-Mail Server

```
DB n
UID U30150,U30153
DOPT f
```

* Retrieve the nucleotide database entries with accession numbers U30150 and U30153, and display them in FASTA format.

```
DB m
UID 88055872
```

* Display the MEDLINE record 88055872 in the default format.

```
DB n
UID U30150
DOPT m
```

* Retrieve the nucleotide database entry with accession number U30150, and display any related MEDLINE information.

```
DB p
TERM ras
```

* Search for the term “ras” in all fields of the protein database, and display in the default format.

```
DB m
TERM smith ab [auth]
DISPMAX 15
```

* Search the author field of the MEDLINE database for papers by A.B. Smith, and display the most recent 15 documents in the default report format.

```
DB n
TERM caenorhabditis elegans [ORGN] & 1996/01/28 [DATM]
DOPT g
```

* Retrieve all the *C. elegans* records added to the nucleotide database on Jan. 28, 1996, and display in GenBank format.

Selected Recent Publications by NCBI Staff

Altschul, SF, and W Gish. Local alignment statistics. *Methods Enzymol* 266:460–80, 1996.

Hogue, CWV, H Ohkawa, and **SH Bryant.** A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database. *TIBS* 21:226–9, 1996.

Koonin, EV, RL Tatusov, and **KE Rudd.** Protein sequence comparison at genome scale. *Methods Enzymol* 266:295–322, 1996.

Madden, TL, **RL Tatusov**, and J Zhang. Applications of network BLAST server. *Methods Enzymol* 266:131–41, 1996.

Schuler, GD, **JA Epstein**, **H Ohkawa**, and **JA Kans**. Entrez: molecular biology database and retrieval system. *Methods Enzymol* 266:141–62, 1996.

Silberman, JD, ML Sogin, **DD Leipe,** and CG Clark. Human parasite finds taxonomic home. *Nature* 380:398, 1996.

Wilbur, WJ, and Y Yang. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med* 26(3):209–22, 1996.

Wilbur, WJ, F Major, **J Spouge**, and **S Bryant**. The statistics of unique native states for random peptides. *Biopolymers* 38:447–59, 1996.

Wootton, JC, and **S Federhen**. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* 266:554–71, 1996.

Human/Mouse Homology Map Added to Web Site

NCBI now provides access to the Seldin/Debry Human/Mouse Homology Map through its WWW Home Page. The homology map is provided and maintained by Michael Seldin at Duke University Medical Center and Ronald Debry at the University of Cincinnati. To use the homology map, select the **Human/Mouse Homology Maps** option from the Home Page, and click on a particular human or mouse chromosome. You will then see a table comparing genes in homologous segments of DNA from human and mouse sources, sorted by position in each genome. More than 1,400 loci are presented, most of which are genes. Links to more information on using the map, table construction, and underlying assumptions are also provided. ■

Images Now Accessible Through OMIM

NCBI's WWW version of the Online Mendelian Inheritance in Man (OMIM) database now includes images of clinical phenotypes via a link to the Genetics Image Archive of the Cedars-Sinai Medical Center. If an image is available for a given OMIM record, an **Images** button is included as one of the available database links. Alternatively, from the OMIM Home Page, users can go directly to the Image Archive, where the images are organized by OMIM number. Currently more than 100 images are available. The URL for direct access to the OMIM Home Page is http://www3.ncbi.nlm.nih.gov/omim. ■

New Genome Survey Sequence Division

To keep pace with the rapidly increasing output of genomic sequence data, NCBI will be creating a new Genome Survey Sequence (GSS) division to be included in GenBank Release 96.0 (August 1996).

The GSS division will fill the need for a repository for genomic sequence data that is not appropriate for inclusion in the standard organism-specific divisions. Submissions to the GSS division can include sequence data generated by single pass "reads" from random genome surveys, exon trapped products, and cosmid, BAC, or YAC end clones. Creation of the new GSS division will allow users easy access to this data for use in mapping and sequencing of larger contigs, which can then be submitted to the standard GenBank divisions, while at the same time segregating this specialized type of high-volume data from the more traditional GenBank sequences. There are currently more than 7,000 sequences in this division.

There is a special data submission format for these sequences, similar to that used for EST and STS submissions. To obtain a copy of the format specifications, send a request to info@ncbi.nlm.nih.gov. ■

Sequin for Database Submissions: A Quick Guide

NCBI has recently released a new program called Sequin for submitting sequences to the GenBank, EMBL, and DDBJ databases. The advantages of Sequin over Authorin include the capacity to handle long sequences and segmented entries, easier editing and updating, and complex annotation capabilities. In addition, Sequin contains a number of built-in validation functions for enhanced quality assurance.

This overview is intended to provide a quick guide to Sequin's capabilities, including automatic annotation of coding regions, the graphical viewer, quality control features, and editing features. More detailed instructions on these and other functions can be found in Sequin's on-screen **Help** file.

Basic Sequin Organization

Sequin is organized into a series of forms for (1) entering submitting authors, (2) entering organism and sequences, (3) viewing the complete submission, and (4) editing and annotating the submission. To advance through the pages making up each form, simply click on labeled folder tabs or the **Next Page** button. After the basic information forms have been completed and the sequence data imported, Sequin provides a complete view of your submission, in your choice of text or graphic format. At this point, any of the information fields can be easily modified by double-clicking on any area of the record, and additional biological annotations can be entered by selecting from a menu.

Sequin has an on-screen **Help** file that is opened automatically when you start the program. Because it is context-sensitive, the **Help** text will change as you progress through the program.

Welcome to Sequin Form

Sequin's first window asks you to indicate the database to which the sequence will be submitted, and prompts you to start a new project or continue with an existing one. In general, each sequence submission should be entered as a separate project. However, an important new feature of Sequin is that it also accepts submissions of segmented DNA sequences, population studies, and phylogenetic studies. These entries would be submitted together as one project.

The sequence data for this example is Drosophila eukaryotic initiation factors 4E-I and 4E-II (accession number U54469).

Submitting Authors Form

The pages in this form ask you to provide the release date, a working title, names and contact information of submitting authors, and affiliation information. To create a personal template for use in future submissions, use the **File/Export** option after completing each page of the Submitting Authors form. Figure 1 shows a partially filled out page for affiliation information.

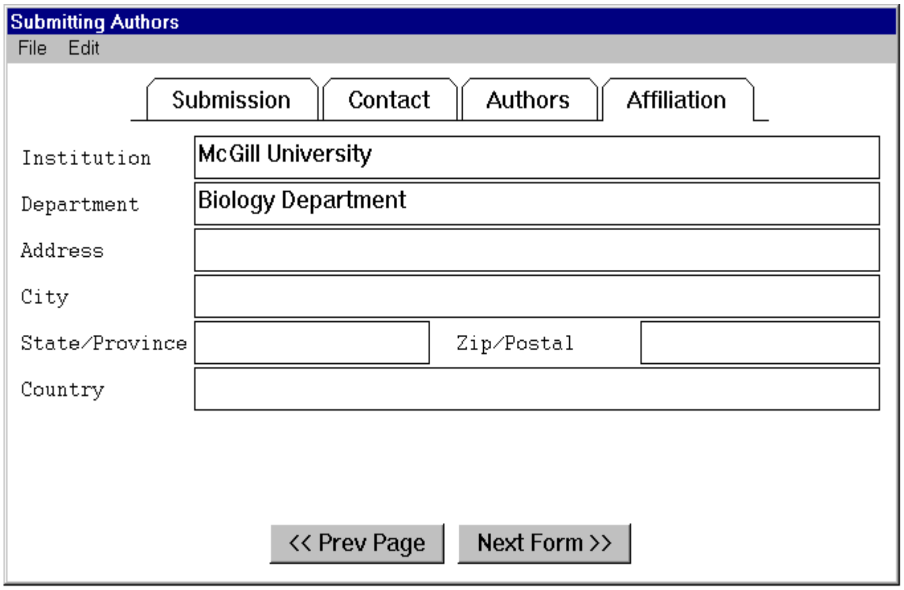
Organism and Sequences Form

The first page of this form requests information regarding the organism from which the sequence was derived. Organism information is most easily entered by selecting the appropriate organism from the scrollable list. As you begin typing the organism name, the list will jump to the right alphabetical location. Once you select an organism from the list, the corresponding scientific and common name and genetic code are filled out automatically (Figure 2). If your organism is not on the list, Sequin will simply accept what you have typed.

Importing Nucleotide and Protein FASTA Files

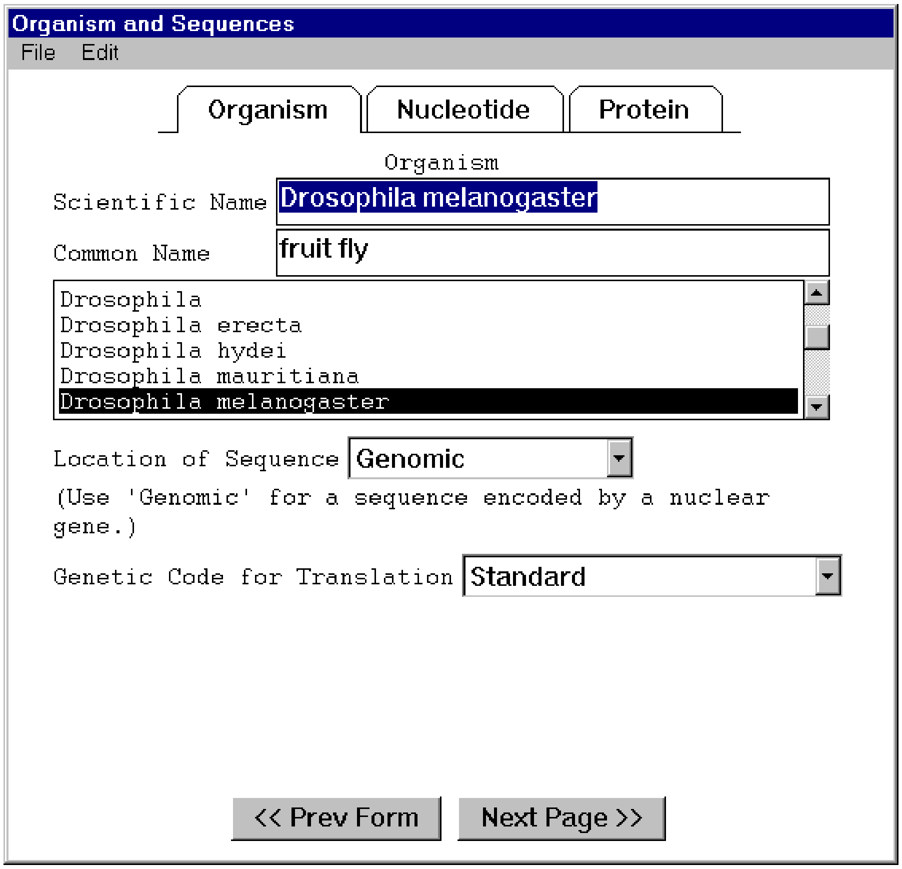
With Sequin, the actual sequence data are imported from an outside data file. So before you begin, prepare your sequence data files using a word processor or perhaps a text editor associated with your laboratory sequence analysis software. One

Continued on page 6



Institution McGill University
Department Biology Department
Address
City
State/Province Zip/Postal
Country

Figure 1



Scientific Name: Drosophila melanogaster

Common Name: fruit fly

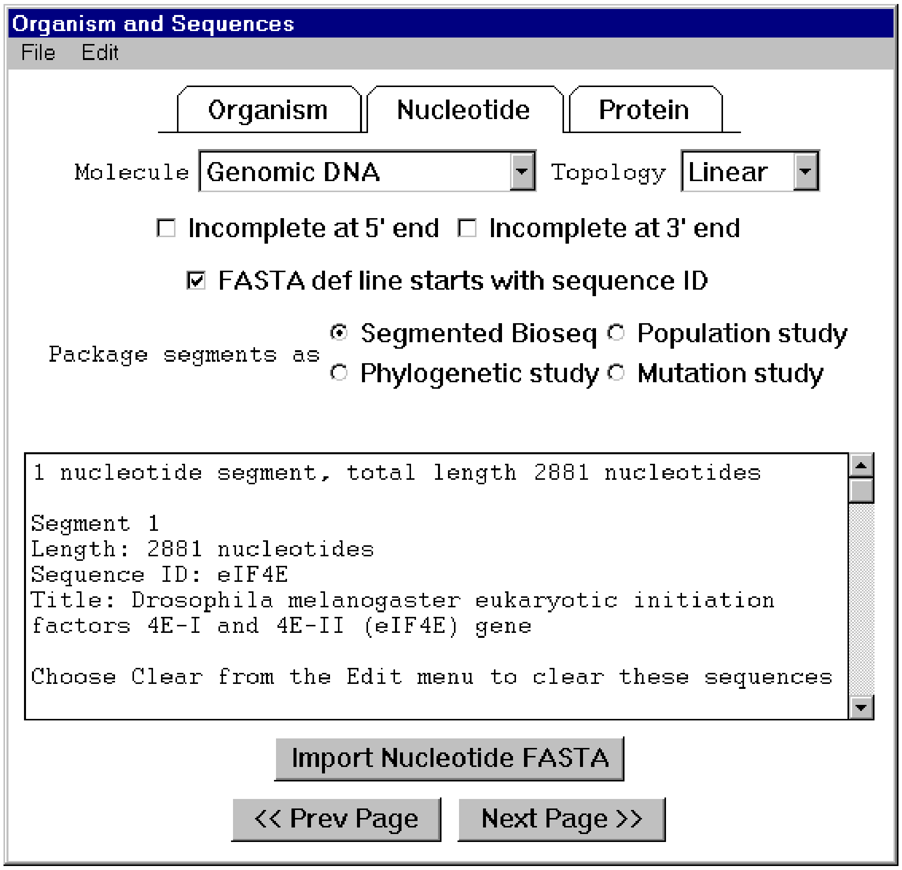
```
Drosophila
Drosophila erecta
Drosophila hydei
Drosophila mauritiana
Drosophila melanogaster
```

Location of Sequence: Genomic ▾

(Use 'Genomic' for a sequence encoded by a nuclear gene.)

Genetic Code for Translation: Standard ▾

Figure 2



Molecule: Genomic DNA ▾ Topology: Linear ▾

□ Incomplete at 5' end □ Incomplete at 3' end

☑ FASTA def line starts with sequence ID

Package segments as ● Segmented Bioseq ○ Population study ○ Phylogenetic study ○ Mutation study

```
1 nucleotide segment, total length 2881 nucleotides

Segment 1
Length: 2881 nucleotides
Sequence ID: eIF4E
Title: Drosophila melanogaster eukaryotic initiation
factors 4E-I and 4E-II (eIF4E) gene

Choose Clear from the Edit menu to clear these sequences
```

Figure 3

***Sequin**, continued from page 5*

great feature of Sequin is that the program can automatically annotate your sequence and coding regions if you format the identifying descriptive information (known in Sequin as the FASTA definition line) in a particular structured manner. See "Before You Begin" on page 8 for format details.

To import the nucleotide sequence data, click on the **Nucleotide** folder tab to advance to the next page (Figure 3). Select molecule type and topology, check any additional boxes that apply, then click on **Import Nucleotide FASTA** and select the appropriate file. When the sequence file import is complete, a box will appear showing the number of nucleotide segments imported, the total length in nucleotides of the sequences entered, and the local ID you designated, but the actual sequence data is not shown. If any of this information is missing or incorrect, check the file containing the sequence data for proper FASTA format, choose Clear from the **Edit** menu, then reimport the sequence.

To import the amino acid sequence, click on the **Protein** folder tab and proceed in the same manner as nucleotide data. In this example, we imported two protein sequences. These are the alternative splice products of the same gene. As shown on page 8, both protein sequences are in the same data file, but each has its own definition line with local ID.

Viewing Your Submission

After you have completed importing the data files, Sequin will display your full submission information in the GenBank text format (Figure 4).

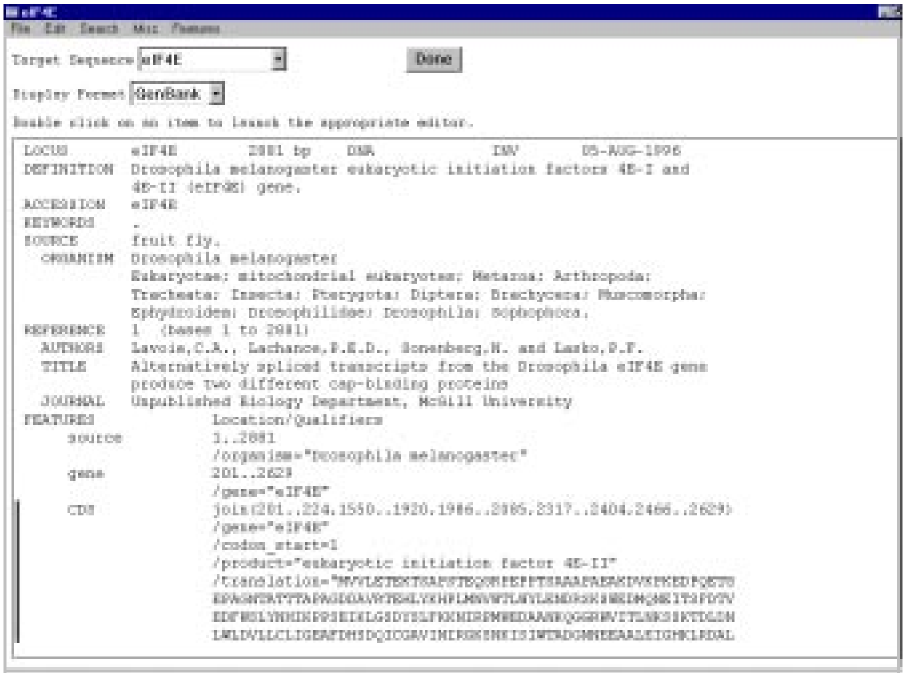
Based on information provided in your DNA and amino acid sequence files, any coding regions will be automatically identified and annotated for you. Figure 4 shows only the top portion of the GenBank record, but you can see the first of two coding region (CDS) features. There are also two mRNA features (not shown in figure) that, with minor editing, can be extended to include the 5' and 3' UTRs.

To get a graphical view, use the **Display Format** pop-up menu to change from GenBank to Graphic (Figure 5). Reviewing your submission in Graphic format allows you to visually confirm expected location of exons, introns, and other features in multiple interval coding regions. The Graphic view in our eukaryotic initiation factor example illustrates how the coding region intervals for the two protein products are spatially related to each other. This figure shows the record after the initial mRNA intervals have been edited to include the 5' and 3' UTRs.

Editing and Annotating Your Submission

At this point, Sequin could process your entry based on what you have submitted so far. However, to optimize usefulness of your entry for the scientific community, you will probably wish to provide additional information to indicate biologically significant regions of the sequence. This information may be in the form of Descriptors or Features. (Descriptors are annotations that apply to an entire sequence or set of sequences. Features are annotations that apply to a specific sequence interval.)

Sequin provides two convenient methods to modify your entry: (1) to edit existing information, double click on the text or graphic area you wish to modify, and Sequin will display forms requesting needed information, or (2) to add new information, use the **Misc** and **Feature** menus and select from the list of available annotations.



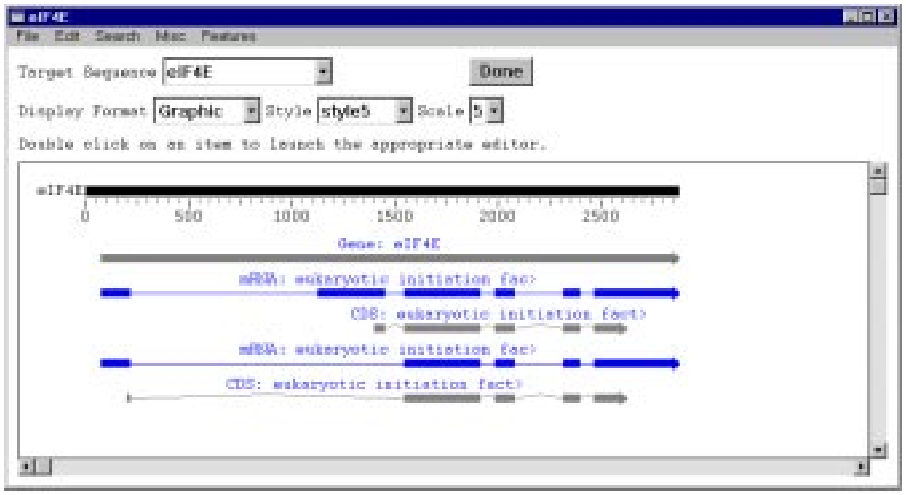
Target Sequence eIF4E

Display Format GenBank

Double click on an item to launch the appropriate editor.

```
LOCUS       eIF4E        2881 bp    DNA             INV       05-AUG-1996
DEFINITION  Drosophila melanogaster eukaryotic initiation factors 4E-I and
            4E-II (eIF4E) gene.
ACCESSION   eIF4E
KEYWORDS    .
SOURCE      fruit fly.
  ORGANISM  Drosophila melanogaster
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Arthropoda;
            Tracheata; Insecta; Pterygota; Diptera; Brachycera; Muscomorpha;
            Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1  (bases 1 to 2881)
  AUTHORS   Lavoie,C.A., Lachance,P.E.D., Sonenberg,N. and Lasko,P.F.
  TITLE     Alternatively spliced transcripts from the Drosophila eIF4E gene
            produce two different cap-binding proteins
  JOURNAL   Unpublished Biology Department, McGill University
FEATURES             Location/Qualifiers
     source          1..2881
                     /organism="Drosophila melanogaster"
     gene            201..2629
                     /gene="eIF4E"
     CDS             join(201..224,1550..1920,1986..2085,2317..2404,2466..2629)
                     /gene="eIF4E"
                     /codon_start=1
                     /product="eukaryotic initiation factor 4E-II"
                     /translation="MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETG
                     EPAGNTATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLENDRSKSWEDMQNEITSFDTV
                     EDFWSLYNHIKPPSEIKLGSDYSLFKKNIRPMWEDAANKQGGRWVITLNKSSKTDLDN
                     LWLDVLLCLIGEAFDHSDQICGAVINIRGKSNKISIWTADGNNEEAALEIGHKLRDAL
```

Figure 4


eIF4E
File Edit Search Misc Features
Done
eIF4E
0
500
1000
1500
2000
2500
Gene: eIF4E
mRNA: eukaryotic initiation fac>
CDS: eukaryotic initiation fact>
mRNA: eukaryotic initiation fac>
CDS: eukaryotic initiation fact>

Target Sequence eIF4E

Display Format Graphic Style style5 Scale 5

Double click on an item to launch the appropriate editor.

Figure 5

Additional sequence data can also be added using Sequin's powerful sequence editor. Sequin will automatically adjust feature intervals when editing the sequence. But first, save the entry so that if you make any unwanted changes during the editing process you can revert to the original copy.

In this example, there are two RNA sequences transcribed from the same region, and we have additional information about their 5' and 3' UTRs. With minor editing, we can extend the two mRNA features to include these untranslated intervals. Just double-click on an mRNA feature, then click on the **Location** tab, and you will see a small spreadsheet showing the existing intervals. Edit the locations in the spreadsheet to extend the mRNA. The interval of the appropriate gene feature will automatically be adjusted as well.

Publication information can also be added at this point. To change the publication status from Unpublished to published in the *Journal of Biological Chemistry*, just double-click on the Reference section, and fill in the citation form that is presented.

Continued on page 8

Before You Begin: Preparing Nucleotide and Amino Acid Data

Prepare your sequence data files using a word processor or some other text editor, and save in ASCII text format. The data should be arranged in FASTA format, which simply requires that line 1 begin with a > sign, followed by identifying descriptive text. The sequence begins in line 2. Note that many sequence analysis software packages include FASTA as one of the available output formats.

For the DNA sequence, the definition line should contain your own local ID code for the sequence and a working title. During the submission process, NCBI staff will change your local ID to a GenBank accession number.

If you have an amino acid translation, create a separate sequence file in the same manner as above. Multiple amino acid sequences can be included in a single file. Our eukaryotic initiation factor example has two protein products, which are contained in the same file, but with separate definition lines.

In order to take advantage of Sequin's automatic annotation feature, the definition line for amino acid sequences must be in the structured format illustrated below. Additional information can also be provided for other features, but we are only showing the minimum information required.

Segmented Nucleotide Sets—A segmented nucleotide entry is a set of noncontiguous genomic DNA sequences, for example, encoding exons along with fragments of their flanking introns. Segmented sets apply only to incomplete genomic DNA sequences, not complete genomic DNA sequences or mRNA sequences. In order to import nucleotides in a segmented set, each individual sequence must be in FASTA format with an appropriate definition line, and all sequences may be in the same file. The file containing the sequences is imported into Sequin as described.

Population or Phylogenetic Studies—For phylogenetic studies, the scientific or common name of each organism should be encoded in each FASTA definition line, e.g., [org=mouse]. In this case, the organism page should not be filled out. For population studies, you can encode strain, clone, and isolate information in the definition line, e.g., [strain=BALB/c].

Format forDNA Sequence Definition Line

```
>local ID [org=organism] title
```

DNA Sequence File

```
>eIF4E Drosophila melanogaster eukaryotic
initiation factors 4E-I and 4E-II (eIF4E) gene
CGGTTGCTTGGGTTTTATAACATCAGTCAGTGACAGGCATTTCCAGA
GTTGCCCTGTTCAACAATCGATAGCTGCCTTTGGCCACCAAAATCCC
AAACTTAATTAAAGAATTAAATAATTCGAAT.....
```

Format for Protein Sequence Definition Line

```
> local ID [gene=locus; optional description] [prot=name;
optional description] optional title
```

Protein Sequence File

```
>4E-I [gene=eIF4E] [prot=eukaryotic initiation factor 4E-I]
MQSDFHRMKNFANPKSMFKTSAPSTEQGRPEPPTSAAAPAEAKDVKP
KEDPQETGEPAGNTATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLEN
DRSKSWEDMQNEITSFDTVEDFWSLYNHIKP.....
>4E-II [gene=eIF4E] [prot=eukaryotic initiation factor 4E-II]
MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGEPA
GNTATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLENDRSKSWEDMQN
EITSFDTVEDFWSLYNHIKPPSEIKLGSDYS.....
```

***Sequin**, continued from page 7*

Validation

Once you are satisfied that you have entered all the relevant information, save your file! Then select **Validate** under the **Search** menu. You will either receive a message that the validation test succeeded or see a screen listing the validation errors. Just double click on an error item to launch the appropriate editor for making corrections. See the Sequin **Help** text for more information on correcting errors. The validator includes checks for such things as missing organism information, correct coding region length, internal stop codons in coding regions, mismatched amino acids, or nonconsensus splice sites.

Submitting the Entry

When the entry is properly formatted and error-free, click the **Done** button or select **Prepare Submission** under the **File** menu. You will be prompted to save your entry and e-mail it to the database you selected. The address for GenBank is gb-sub@ncbi.nlm.nih.gov. ■

Frequently Asked Questions

Since the yeast genome has now been completely sequenced, how can I now retrieve these records? Can I search it with the BLAST servers at NCBI?

Yes, a single copy of the complete *Saccharomyces cerevisiae* genome is now available from the Entrez retrieval system (using the genomes database) and for BLAST searches. NCBI has a searchable database called "yeast" for either the nucleotide or protein sequences, using blastn, blastp, blastx, tblastn, or tblastx search engines. The sequences are also available from the NCBI anonymous FTP (ncbi.nlm.nih.gov) site in the '/genbank/genomes/S_cerevisiae' directory. See the README file in the '/genbank/genomes' directory for a description of the files present in this directory.

What is the difference between the GenBank accession number and the GI number?

The accession number is assigned to every GenBank record when it is submitted. It applies to the full record and does not change if parts of the record are modified, such as the publication information, feature annotations, or even sequence corrections.

The GI identification numbers are assigned specifically to the sequence components of the record in order to track changes in the sequence itself. The nucleotide sequence gets a GI number (called an NID), plus each protein sequence gets an individual GI number (called a PID). Any time the sequence is modified by the submitter, a new GI number (NID or PID) is assigned. But the older numbers are still retained in the system, and can be retrieved if needed.

How does your BLAST queuing system work? How can one get bumped from position 3 to 7, or from 12 to 13, for example?

You can fall back in line if others come in with jobs that take up fewer resources. For example, a tblastn job, which is very computing-intensive, could be bumped back by blastn or blastp jobs that take only seconds to run. Priority is also given to queries against small databases. Note that about 8,500 BLAST queries are performed each day through the Web page, and queues tend to be shorter in the early morning or at night, eastern time. Also, the Web BLAST service now allows for results to be returned by e-mail (and also in HTML format for viewing in a Web browser).

When I do a BLAST search, I am only interested in matches to human sequences. Can I limit my results that way?

Yes. If you are using Network BLAST (server/client version), there is now a new client available, PowerBlast, which permits filtering searches by organism, among several other features. See the BLAST article on page 11 for details.

Does the nr database already include the sequences for genomes, like the E. coli *genome or other available genome sequences?*

With the exception of EST and STS sequences, the nr database includes all the sequences that are in GenBank, including sequences from complete genomes. (For EST and STS database searches, you need to explicitly specify those databases.)

Unigene Collection, *continued from page 2*

feasible. The Merck-funded EST project at Washington University alone has produced 320,000 EST sequences so far, with new data being submitted at the rate of 4,500 sequences per week. Mark Boguski, who leads NCBI's EST database project, says, "The transcript map will provide needed reality checks for the large-scale sequencing efforts ahead," and adds that "the disease gene hunting community has long had a desire to develop a transcript map."

Organizing the UniGene Clusters

When EST sequence data started rolling into GenBank by the thousands earlier this year, NCBI's Greg Schuler began investigating ways to use them to identify unique human genes. The problem was to organize the data in such a way that all representations of a single gene were collected in a single cluster.

As a comprehensive collection of publicly available sequence data, GenBank is also a historical archive with a large degree of internal redundancy. A sequence for the same gene may have been submitted by multiple labs, and a given gene may have separate entries from different types of sequence (e.g., contiguous and noncontiguous genomic sequences, mRNA sequences with alternative splicing, and EST sequences). For EST sequences, redundancy and overlap are especially prevalent. This data redundancy makes it difficult to identify unique markers for mapping, thus the need for the UniGene project.

In the first phase of the UniGene project, Schuler screened all ESTs against existing functionally cloned GenBank entries to eliminate redundancies. He then developed techniques to screen the remaining ESTs against each other to determine those likely to be derived from the same gene. If sequences were found to share statistically significant DNA sequence similarity in the 3' UTR, they were assigned to the same cluster.

The first phase of the UniGene project resulted in a set of 3,125 nonredundant unique human 3' UTRs, referred to as the UniGene set. The UniGene set serves as a source of mapping candidates and as a standard to compare and screen new EST submissions. New EST submissions that do not match any sequences in the UniGene set are considered new human genes and are organized into unique clusters to provide additional mapping candidates. To date, more than 48,000 3'-anchored UniGene clusters have been generated. Some clusters contain more than 1,000 ESTs, while others consist of as few as 1 EST. As would be expected, the largest clusters correspond to well-studied genes, such as the hemoglobin subunits and the serum albumin precursor.

Developing the Transcript Map: A Collaborative Effort

Once the UniGene clusters were identified, there was an immediate use for them in developing a comprehensive transcription map of the human genome. The mapping project is a collaborative effort, involving NCBI, several genome mapping centers, and the sequence submissions of individual scientists. NCBI distributes nonoverlapping cluster sets to the various mapping centers to ensure that redundancy does not creep back into the databases and that duplication of mapping effort is kept to the minimum necessary for data accuracy checks and cross referencing. This collaborative effort has resulted in the placement of 15,000–20,000 transcripts on RH and YAC maps.

Using the UniGene Clusters

Aside from their contribution to large-scale mapping efforts and to basic research in genome organization, the UniGene collection and subsequent transcript maps are an important resource for many investigators. For example, of great interest to disease gene hunters is that 82% of the positionally cloned genes that are currently known to be mutated in human disease states are represented by exact matches with one or more ESTs in GenBank. Gene hunters can use the transcript maps to gain valuable clues to expected gene location and density in an area of interest. UniGene clusters are also being studied to find gene polymorphisms. And recently developed techniques for assessing gene expression on a genomewide scale (e.g., microarray expression systems) take advantage of the abundance of unique EST sequences that can be readily retrieved from GenBank.

The UniGene data set can be accessed through NCBI's WWW service (http://www.ncbi.nlm.nih.gov). From the Home Page, scroll to "Other NCBI Resources," and click on **Unigene**. The UniGene page displays icons for each of the 23 chromosomes. To see a list of all the UniGene clusters that have been identified for a given chromosome and the sequences comprising the cluster, just click on the chromosome. To search for clusters containing a specific word or phrase, enter the search term in the text box at the top of the UniGene page.

UniGene is updated every 2 months, approximately 1 week after a new GenBank release is produced. Files can be downloaded from NCBI's FTP site in the 'repository/unigene' directory. No search tools are provided other than the Web interface. ■

New BLAST Services Now Offered

If you have visited the Web BLAST page recently, you will have discovered that the service has undergone substantial revision and several new features have been added. Users now have the option to select either the "Basic" BLAST search using default parameters or the "Advanced" search using customized BLAST search parameters. In addition, an e-mail option has been added for convenient delivery of search results. By using this option, your BLAST output will be delivered by e-mail, and your Web browser will not be tied up while the BLAST search is being performed.

Introducing PowerBlast

NCBI has released PowerBlast, a new Network BLAST application for automated analysis of genomic sequences. PowerBlast combines BLAST searching with additional filtering for low complexity regions and repeats. In addition, PowerBlast features a one-to-many alignment output showing the alignment of the query sequence with all the matching sequences (as opposed to standard BLAST results that show the query sequence aligned individually against each matching sequence). The one-to-many presentation illustrates the differences between the query sequence and the search results, rather than the similarities, as in standard BLAST results. The multiple alignment results are displayed in both text and graphical formats. The graphic view shows the computed optimal alignment gaps, and annotated features are superimposed on the aligned sequences. PowerBlast can also generate organism-specific output—for example, searches restricted to human sequences. Versions of PowerBlast are available for Macintosh, PC, SunOS, and Solaris platforms, and can be downloaded from NCBI's FTP site in the 'pub/sim2/PowerBlast' directory.

New BLAST E-Mail Server

All BLAST e-mail queries sent to "blast@ncbi.nlm.nih.gov" after August 5 are being processed by a new e-mail server at the NCBI. The server address and query format will not change.

The most important new features of the server are—

1. Filtering of the query sequence is performed as the default. Low complexity sequence that is found by a filter program is substituted using the letter "N" in nucleotide sequences and the letter "X" in protein sequences. The program "dust" is used for BLASTN queries; "seg" is used for all others. For a description of these filtering programs, the advantages of filtering, and instructions on how to perform queries without filtering, see section 5 of the new Help document.

2. There are two new directives: NCBI_GI, which causes the GI to be displayed in the output, and HTML, which causes the output to be in HTML format, suitable for viewing by a Web viewer. Both of these command options are discussed in section 5 of the new Help document.

To receive the documentation for the new BLAST e-mail server, send a message consisting of only the word HELP to the server address. Questions and comments on the new service are welcome at blast-help@ncbi.nlm.nih.gov. ■

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 48.00, Swiss-Prot, 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 Entrez executable programs for accessing CD-ROM data on a variety of platforms. It also contains 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) and MACAW (multiple sequence alignment 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** or **ftp 130.14.25.1**. 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.

Special:
Four-page quick guide to **Sequin**,
NCBI's newest submission software,
begins on page 5.

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

NATIONAL INSTITUTES OF HEALTH • National Library of Medicine

NCBI News

August 1996