NCBI News

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

Winter 1999

Sequence Submissions Now May Include Alignment Data

New sequences submitted as phylogenetic studies, mutational studies, or population studies may include sequence alignments if the submission is made using Sequin. The alignments can be created within Sequin or imported into Sequin from a file in a standard alignment format, such as NEXUS or PHYLIP. Alignments may include both newly submitted sequences and sequences that are already in GenBank.

Newly submitted sequences appearing in the alignment file are interpreted as GenBank sequence submissions and will be formatted as such by Sequin. A sequence that is already in GenBank must be flagged as a "reference sequence" within the alignment file by inserting the prefix "acc", followed by the GenBank accession number of the sequence, at the beginning of its "source modifier" line, as described below. Although alignment data included within Sequin-mediated submissions are not currently visible within the Entrez retrieval system, methods for making the data accessible are under development. An example of a NEXUS format alignment file will illustrate the syntax that must be used in order to include an alignment with a Sequin submission.

The sample file shown in Figure 1 was originally generated by the MacClade program and is in the NEXUS interleaved format. This sample alignment contains three new sequences to be submitted to GenBank, as well as one reference sequence, named Rhesus_cyto, which is already in GenBank. For the purpose of the Sequin submission, the file had to be modified from the original MacClade output by adding

Continued on page 8

New Compound Accession Number Indicates Sequence Revision History

Beginning with GenBank Release 111.0 scheduled for late February, GenBank, EMBL, and DDBJ nucleic acid and protein sequence records will contain a new compound identifier that will specify both the identity of the record and the revision history of the sequence data within the record. The compound identifier will have the form "Accession.version" and may be thought of as an accession number with an extension indicating a sequence version.

For nucleotide sequences the accession number portion of the compound identifier will continue to consist of the one-letter/five-digit or two-letter/six-digit codes currently used. The compound identifier will also be applied to the translated nucleotide sequences (CDS features), which have now become the primary source of protein sequence data and have not previously been assigned accession numbers. By applying the compound identifier to protein sequences as well as to nucleotide sequences, the tracking scheme is made complete and uniform. For proteins the accession number portion will have a new format consisting of three letters followed by five digits.

Version numbers for existing as well as new records will be initialized to "1" and will be incremented with each subsequent sequence revision. The version number applies only to changes in actual sequence data, not to modifications in any other part of a GenBank record. GenBank will display the compound identifier in GenBank flat files as the first field in a new "VERSION" line.

Currently, GenBank treats accession numbers and version numbers, referred to as GI numbers, as separate entities, which are used to track data at two levels. Accession numbers provide unique and static labels for whole records,

Continued on page 4

IN THIS ISSUE

NCBI News

NCBI News is distributed four times a year. Beginning in 1999, issues are dated Winter, Spring, Summer, and Fall. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence to *NCBI News* at the address below.

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

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

NIH Publication No. 99-3272

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

HIV-1 Subtyping Tool Now on Web

HIV-1, the retrovirus responsible for the AIDS pandemic, has a highly variable RNA genome that has diversified into multiple subtypes. The genetic diversity of HIV-1 is exemplified by the fact that members of the two main genetic groups, group M (major) and group O (outlier), differ by up to 47% in the amino acid sequences of their envelope proteins.[1] Within group M, HIV-1 can be subdivided into at least nine distinct subtypes, among which there is a 25 to 35% variation in the amino acid sequence of the envelope protein.[2] The genetic variability of HIV-1 is not limited to the envelope protein but is also seen in sequences throughout the HIV-1 genome.

The genetic diversity of HIV-1 presents a problem in vaccine development. Most vaccines currently being tested are most effective against HIV-1 subtype B; however, as other subtypes develop and spread, vaccines specific to these variants will be required. The fact that a significant fraction of HIV-1 isolates have mosaic genomes resulting from intersubtype recombination further complicates the problem of vaccine development and highlights the need for subtype monitoring.

Subtyping Tool Developed by NCBI

To facilitate monitoring for HIV-1 genomes, NCBI has developed a Web-based subtyping system. It can be reached at http://www.ncbi.nlm.nih.gov/retroviruses/HIV1, where you click on Subtyping HIV-1. The subtyping method employs a blastn comparison between the HIV-1 sequence to be subtyped and a panel of reference sequences taken from the principal HIV-1 variants. The subtyping panel includes complete genomic references for the A, B, C, D, E, F, G, and H group M subtypes as well as for group O and the recently described N sequence.[3] During the subtyping process, multiple blastn comparisons are made over a sliding window of a size and step value set by the user. A color-coded graph of the blastn sequence similarity score against window location is generated for comparisons between the query sequence and each reference sequence in the panel.

Submitting the Query Sequence and Performing the Analysis

To begin the subtyping process, the sequence to be subtyped is first pasted into the query window on the query page. The sequence window size and step value are entered; alternatively, the defaults (window size = 300, step value = 100) may be used. Clicking on the **Subtype** button starts the subtyping process.

The Similarity Graph

The subtyping procedure generates an output page featuring a similarity graph, which depicts the results of the blastn comparisons. The abscissa of the similarity graph is calibrated in bases. The blastn similarity scores for the sequence windows, of size and step value specified by the user, are plotted on the ordinate. A key to the right of the graph gives the trace color codes for the HIV-1 subtypes represented in the screening panel. Clicking on the color key for a particular HIV-1 subtype leads to a catenated GenBank-format display of the individual HIV-1 reference sequences used to characterize the subtype.

A similarity bar (Figure 1) at the top of the graph is composed of colored blocks, each representing a sequence window. The color coding of the blocks indicates the subtype with the highest blastn score within the corresponding

sequence window. Ambiguous regions are indicated with bicolored blocks. The blastn alignments for an individual window can be retrieved by using the similarity bar, and a global multiple sequence alignment with one or all reference sequences can also be seen by pressing the **Global Alignment** button to the left of the similarity graph.

The similarity bar resulting from the subtyping of an HIV subtype A/C mosaic (U88823) is shown in Figure 1. The lightest blocks indicate segments within the U88823 sequence that are most homologous to HIV-1 subtype A, whereas the blocks of intermediate shading indicate sequence homologies to HIV-1 subtype C. A few blocks indicating homology to HIV-1 subtypes D, G, E, and F are also visible; however, the U88823 sequence is seen here to be primarily an A/C mosaic.

The data used to generate the similarity graph may be inspected by pressing the button labeled **Print Score Table**, and may be subsequently saved as a local file.

Recognizing Recombinant Forms of HIV-1

HIV-1 sequences of a pure subtype show uniformly high scores for the reference sequences of their subtype. Recombinant forms of HIV-1, however, are easily recognized by "breakpoints" in the similarity bar. This behavior is seen in Figure 1 for the HIV A/C mosaic, U88823. Major breakpoints can be seen in two intervals: 1,100 to 2,100 and 5,300 to 7,700 bases.

General Retroviral Subtyping

A general genotyping tool is also available at http://www.ncbi.nlm.nih.gov/retroviruses/subtype/subtype.html. The general genotyping tool allows for user-specified reference sequences and performs a blastn analysis of the query sequence resulting in graphical and tabular outputs that are similar to those of the HIV-1 subtyping tool.

The HIV subtyping project was led by NCBI scientists Colombe Chappey and Uwe Plikat.

Notes

[1] Gao F, et al. Molecular cloning and analysis of functional envelope genes from human immunodeficiency virus type 1 sequence subtypes A through G: The WHO and NIAID Networks for HIV Isolation and Characterization. *J Virol* 70(3):1651–67, 1996.

[2] Ibid.

[3] Simon F, et al. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat Med* 4(9):1032-7,1998. ■

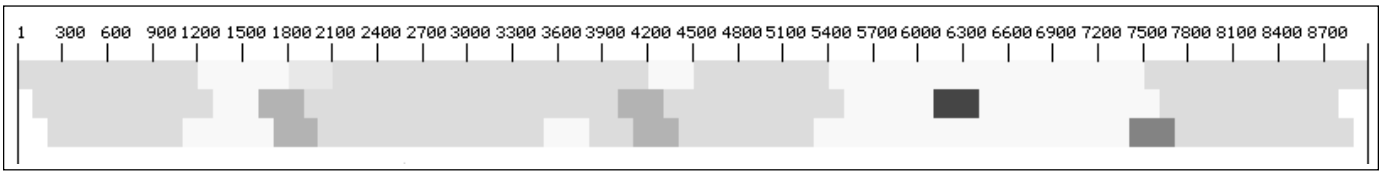

Figure 1. Blastn similarity bar for the HIV-1 query sequence. The query sequence is an HIV-1 subtype A/C mosaic.

Minor Changes to BLAST 2.0 Output

Three minor changes have been made to the output formats of some members of the BLAST 2.0 program family. In response to user requests, "strand" information has been added to the blastn alignment reports in the form of a new "Strand=" line. As an example, "Strand=Plus/Minus" now indicates a BLAST alignment between the Plus strand of the query and the Minus strand of the subject sequence.

A second change to blastn output is the removal of the number of "Positives" within each alignment. This information is redundant because "Positives" and "Identities" are indistinguishable for nucleotide searches. This is so because blastn uses a simple identity matrix for nucleotide comparisons. The number of "Identities" will be retained.

The third change involves the addition of reading frame information to blastx, tblastn, and tblastx outputs. For tblastn a new line of the form "Frame=frame#" indicates the translation frame, frame#, of the subject sequence. For tblastx a similar line now indicates the translation frame of the query sequence. For tblastx outputs the new line will indicate the reading frames of both query and subject using the form "Frame= queryframe#/subjectframe#."

There are no changes to the blastp output. ■

UniGene Includes *Rattus Norvegicus*

Rattus norvegicus has joined *Homo sapiens* and *Mus musculus* to become the third organism to enter the UniGene collection of databases. The popular laboratory rat contributes 79,028 nucleotide sequences, forming 23,518 clusters, to UniGene. Of the 23,518 clusters, 3,002 contain at least one known gene. As in the Human and Mouse UniGene databases, the Rat UniGene database can be searched by keyword, chromosome number, or library number. Rat UniGene can be accessed from the main UniGene page at http://www.ncbi.nlm.nih.gov/UniGene/. ■

Accession, *continued from page 1*

including sequence data, literature citations, and other annotations, whereas GI numbers provide identifiers for various versions of the sequences within the records. Under the current scheme, a new GI is issued to a sequence each time it is revised, while the accession number for the record in which the sequence appears is never changed. GIs now appear in the GenBank flat file on lines labeled NID and also appear within the "/db_xref" qualifiers found in CDS feature tables.

Under the new compound accession system, new version extensions will be issued to sequences whenever the sequence is revised. To maintain compatibility with existing software, a GI field will also be included after the compound accession.version field on the VERSION line, and new GIs will continue to be assigned whenever a sequence change occurs. During a period of transition, GenBank records will retain the NID line-type as well as the new VERSION line-type. Thereafter, the NID line will be dropped, leaving only the version line. The sidebar illustrates the transition.

The decision to adopt the new compound accession format stems from discussions among representatives of an international database collaboration, which includes the sequence databases GenBank, EMBL, and DDBJ. The compound accession number combines the two properties of a general record identifier and a specific tracking ID for sequence modifications into a single consistent entity, which simplifies its use considerably without loss of utility in either component. ■

The Accession.Version Transition

Current: *The upper portion of a current GenBank flat file is shown below:*

```
LOCUS       AAU36846    568 bp     DNA     PRI      26-OCT-1995
DEFINITION  Aotus azarai cytochrome c oxidase subunit II (COII)
            gene, mitochondrial gene encoding mitochondrial
            protein, partial cds.
ACCESSION   U36846
NID         g1040987
...
     CDS                <1..>568
                        /gene="COII"
                        /codon_start=1
                        /product="cytochrome c oxidase subunit II"
                        /db_xref="PID:g1040988"
```

In this case, the record is identified on the ACCESSION line with the accession number U36846. The NID line, directly below, gives an index to the current version of the sequence contained within the record in the form of the GI number g1040987. A CDS feature qualifier, /db_xref="PID:g1040988," gives an ID for the protein product of the CDS.

Transition: *During the transition period, the flat file will appear as below:*

```
LOCUS       AAU36846    568 bp     DNA     PRI      26-OCT-1995
DEFINITION  Aotus azarai cytochrome c oxidase subunit II (COII)
            gene, mitochondrial gene encoding mitochondrial
            protein, partial cds.
ACCESSION   U36846
NID         g1040987
VERSION     U36846.1    GI:1040987
...
     CDS                <1..>568
                        /gene="COII"
                        /codon_start=1
                        /product="cytochrome c oxidase subunit II"
                        /protein_id="AAA12345.1"
                        /db_xref="PID:g1040988"
                        /db_xref="GI:1040988"
```

Note that the ACCESSION and NID lines remain but the new VERSION line now appears following the NID line. The version line begins with the compound accession number, U36846.1, which unambiguously identifies both the record and the version number of the sequence it contains. The GI for sequence U36846.1 is also included on this line, but without the leading g, which will no longer be used. The version line encapsulates the information contained in both the accession and the NID/PID lines, while adding information on the revision history of the sequence data. Also note the addition of a new CDS qualifier, /protein_id="AAA12345.1," which includes an example of a compound protein accession/version number.

Final: *At the end of the transition period, GenBank flat files will have the following form:*

```
LOCUS       AAU36846    568 bp     DNA     PRI      26-OCT-1995
DEFINITION  Aotus azarai cytochrome c oxidase subunit II (COII)
            gene, mitochondrial gene encoding mitochondrial
            protein, partial cds.
ACCESSION   U36846
VERSION     U36846.1    GI:1040987
...
     CDS                <1..>568
                        /gene="COII"
                        /codon_start=1
                        /product="cytochrome c oxidase subunit II"
                        /protein_id="AAA12345.1"
                        /db_xref="GI:1040988"
```

Note that the ACCESSION and VERSION lines remain, while the NID/PID line disappears. The /db_ref="PID: g1040988" CDS qualifier has also disappeared, but the new protein_id qualifier remains.

Frequently Asked Questions

Can I perform a search using Gapped BLAST via the BLAST e-mail server?

Yes, the BLAST e-mail server has recently been upgraded to perform searches using Gapped BLAST by default. To force a search using ungapped BLAST, include the directive "new FALSE" in the e-mail query message.

How can I find the sequence for the cloning vector pBR322 without having to wade through every other record in GenBank that refers to pBR322?

Begin at the initial Entrez Nucleotide database search page. Restrict the search field to **Properties** and enter "gbdiv syn" into the query box. Pressing the **Search** button reveals that about 3,100 records have been selected. These belong to the GenBank "Synthetic" division. Into the **Add Terms to Query** box that now appears, type "pbr322 complete," and change the search field to **Title Word**. Pressing the **Search** button now yields a single sequence, which is that of the pBR322 cloning vector. Follow an analogous procedure for other vector sequences.

I would like to determine if there is an STS marker within my DNA sequence. Can this sort of search be performed easily?

Yes. Choose **Electronic PCR** from the NCBI home page and paste the sequence to be probed into the input box. It will be searched against the dbSTS database. By default, a search for STSs from all organisms will be done; however, a list box below the input box can be used to restrict the search to particular organisms. Press the **Submit Query** button to perform the search.

I would like to find the human UniGene clusters that have 10 or fewer sequences. How can I obtain this information?

The file Hs.data.Z contains data on all the clusters in the current release of the human UniGene database. For each cluster, there is an ID line, which gives the cluster name, and an SCOUNT line, which gives the cluster size. This file is available at ftp://ncbi.nlm.nih.gov/repository/unigene/.

Is information available on the gene expression profiles for tumor cell lines?

The Web page for the Cancer Genome Anatomy Project (CGAP) provides a link to a database of cDNA expression profiles for CGAP EST libraries called the xProfiler. Entering "Lib.282" into the xProfiler query box, for example, returns the expression profile for CGAP Library 282, derived from a prostate neoplasm. This profile includes data on the expression level of ESTs belonging to 1,526 UniGene clusters.

What does it mean when a UniGene cluster is retired?

UniGene clusters are retired for several reasons. A cluster is retired if the sequences in the cluster are retracted. A cluster may also be retired if it is merged with others during periodic UniGene reclustering; the new cluster gets one of the preexisting cluster IDs, while the other IDs are retired. A cluster may also be retired if it is split into two or more clusters, although in most cases such splits result in one of the new clusters retaining the old ID number. A utility that allows users to see what has become of the sequences in retired clusters is planned for the future.

BLAST Lab

atcgtacatagcgcatgtcagncaaatccgttcttg
atcgtacatggcacatgtcagtcaaatccattcttg

How to Search Custom Subsets of GenBank Using Standalone BLAST

The WWW version of BLAST provides for searches of several predefined data sets that can be selected from the Database pull-down menu, including nr, month, dbest, htgs, swissprot, E.coli, yeast, and others. In addition, the Microbial Genomes BLAST page allows searches of any combination of 42 finished and unfinished microbial genomes. However, these data sets may not be sufficient for all researchers. How can BLAST be constrained to search *any* unique subset of GenBank or Entrez? Here are five easy steps.

Step 1. Download Standalone BLAST from NCBI's FTP Site.

BLAST programs are found at
ftp://ncbi.nlm.nih.gov/blast/executables/

Archive files for Windows and several Unix platforms (sorry, no Mac) are named as follows:

PC: blastz.exe
Unix: unix-platform.tar.Z

Download the applicable file into a fresh directory, named blast for example.

Step 2. Uncompress the BLAST Archive File.

PC: blastz.exe is a self-extracting archive and should be run within the new blast directory you created.

Unix: unix-platform.tar.Z must be uncompressed and detarred within the new blast directory you created. To uncompress, use the following command:
uncompress unix-platform.tar.Z

The uncompress operation will produce a Unix Tape Archive (tar) file, called unix-platform.tar, which is detarred as follows:
tar -xvf unix-platform.tar

For both platforms, the unarchiving process will yield several files, of which we will use only two for this example:
blastall and **formatdb**

Step 3. Create or Modify the NCBI Configuration File.

You will need to create a configuration file unless another NCBI tool, such as Network Entrez or Cn3D, has previously created it (in which case you just modify the existing file by adding a single line). The configuration file specifies the path to the BLAST data directory on your local system. It is named ncbi.ini on PCs and must be placed in the c:\win or c:\wnnt directory. On Unix systems, the file is named .ncbirc and must be placed in your home directory.

You can create or modify this file with a word processor (be sure to use ASCII text format) or text editor. The file must contain the two lines shown below. If you are modifying an existing file, just insert the Data= line into the existing [NCBI] section.

PC: If your BLAST directory is c:\blast:
[NCBI]
Data=c:\blast\data

Unix: If your BLAST directory is /home/username/blast:
[NCBI]
Data=/home/username/blast/data

Step 4. Use Batch Entrez to Create a Custom Sequence File.

Batch Entrez is found at
http://www.ncbi.nlm.nih.gov/Entrez/batch.html

To create the BLAST database, you will need a source file of sequences in FASTA format. This file will be processed by the program formatdb to produce three output files that will be used together as the BLAST database.

The source FASTA file will have the form:

>First sequence description
XXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXX...
>Second sequence description
XXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXX...
>Last sequence description
XXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXX...

where the Xs are one-letter amino acid or nucleotide codes and the > sign is the symbol denoting a one-line identifying description of each sequence.

The source file can be built using Batch Entrez. For example, to create a file of only fungal, bacterial, and human helicase protein sequences, type the following into the Batch Entrez search box labeled **Entrez Search Query**, entering on one line:

(fungi[ORGN] OR bacteria[ORGN] OR human[ORGN]) AND helicase[PROT]

Make sure that the radio button that specifies **Entrez Search Query** is selected. Select **Protein** as the Sequence Type from the pull-down menu at the top of the page. Specify **FASTA** as the download format. Then press the **Submit Query** button. A file window will appear showing the default name for the output FASTA file. Assuming that you change the default name, which is "results", to "helicase", Batch Entrez will download the sequences from Entrez and deposit them in a file called helicase. This file may also be renamed *after* it is written.

Step 5. Process the Source FASTA File Using Formatdb.

Assuming that the source FASTA file is named "helicase", type the following at the command line:
formatdb -i helicase -p T

This command tells formatdb to use the source FASTA file, "helicase", as the input and to process the file as a set of protein sequences.

In the case of nucleotide sequences the final switch is given as "-p F" and the command would read:
formatdb -i helicase -p F

Three files will be produced from the source FASTA file, leaving the source file unmodified. These files together constitute a BLAST database:
helicase.phr
helicase.pin
helicase.psq

For nucleotide databases, the extensions are nhr, nin, and nsq.

BLAST Away!

To BLAST a query sequence in FASTA format against the new custom database named helicase, first put the query sequence in a file called "query.fasta." Then use the command line below, entered on one line:
blastall -p blastp -d helicase -i query.fasta -o query.blast

This command line invokes blastall to run the blastp program on the helicase database using the FASTA file query.fasta as the query. The output will be written to a file named query.blast. ■

PROFILE

NCBI Team Ranked First in CASP3 Protein Structure Prediction Contest

The NCBI team of Steve Bryant, Aron Marchler-Bauer, and Anna Panchenko was ranked first in the fold-recognition category by the CASP3 assessor, Alexei Murzin, at the recent Third Meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP3). The NCBI team was one of 98 research groups to participate in CASP3. The CASP3 entrants collectively submitted a total of 3,807 structure predictions for a set of 43 target proteins whose structures had been determined but not revealed. The predictions were made on the basis of the amino acid sequences of the targets and fell into three methodological categories: ab initio, fold recognition, and homology modeling methods.

The NCBI team submitted models for 19 target proteins, 15 of which were classified as fold-recognition targets by the CASP3 organizers on the basis of low sequence similarity to any known structure. Over 40 groups contributed 764 primary models in this prediction class.

The prediction method of the NCBI team involved the threading of a target sequence through a structural template selected from the database of known structures by using fold-recognition algorithms. Hypothetical alignments of the target sequence to the structural template were scored using a composite potential function that included both a term for a conventional residue contact potential and a term for a quantitative measure of residue conservation between the target and the template sequences. The latter term was calculated with the aid of a Position Specific Scoring Matrix (PSSM) generated from PSI-BLAST database searches using the template sequence as the query. In addition, information about structural conservation within the family of proteins represented by the template was used in the predictions.

Using this method, the NCBI team was awarded the top score in the fold-recognition category. Figure 1 shows the structure of the template, carbamoyl phosphate synthetase from *E. coli*, which was used by the NCBI team to model the structure of CASP3 target T0081. The regions of the structure that match well between the template and the target are shown with a light trace. A paper describing the methods used by the NCBI team will appear shortly in the journal *Proteins*.

The NCBI Team

Steve Bryant has been a senior investigator in the Computational Biology Branch of NCBI since 1991 and heads the Protein Structure Group. He and his group conduct research and development on the comparative analysis of macromolecular 3-D structure. Steve's research interests include continued development of threading and structure-structure comparison algorithms.

Aron Marchler-Bauer earned a Ph.D. from the University of Vienna in 1996 for research on empirical potential functions for protein sequence-structure threading. He has since joined Steve Bryant's group at NCBI, where his main interest lies in the study of the significance statistics of comparative sequence-structure analysis.

(left to right) Steve Bryant, Aron Marchler-Bauer, and Anna Panchenko.

Anna Panchenko graduated from Moscow State University in 1993 and subsequently did research on protein dynamics and protein folding. Currently, she is studying various aspects of protein structure prediction and sequence analysis at NCBI.

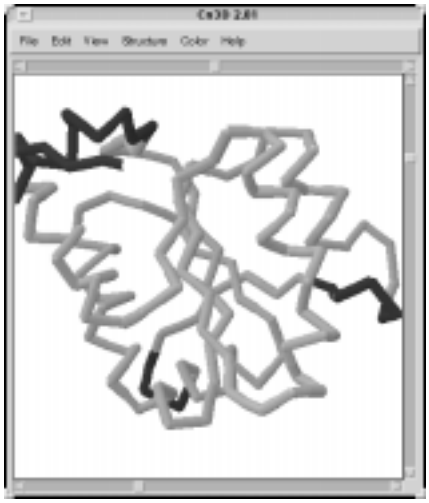

Figure 1. Cn3D backbone rendering of the structure of domain 8 of carbamoyl phosphate synthetase (PDB code 1JDB(K)) used as a template by the NCBI team to build a model for CASP3 target T0081. Light-trace regions are in agreement with VAST alignments between the template and the target structures.

Alignment, *continued from page 1*

the four lines beginning with ">" after the "endblock" line. These four lines are used to specify source modifiers required by Sequin for each sequence. In this example, the "organism" source modifier is specified.

The source modifier line for the reference sequence, which is already in GenBank, begins with "acc" and is followed, without a space, by the accession number of the reference sequence. Sequin compares the reference sequence shown in the alignment Rhesus_cyto with the existing GenBank sequence specified by the accession number. Sequin confirms that the sequences are identical and replaces the reference sequence in the alignment with a pointer to its GenBank record.

Although the sample file below specifies only a single reference sequence, a Sequin submission may contain many existing GenBank sequences as references within a sequence alignment. The alignment between these GenBank reference sequences and others in the alignment file is maintained by using a set of pointers to particular positions within the reference sequences. This scheme allows the alignment to be updated automatically, by updating the pointers, when any component reference sequence changes. ■

```
#NEXUS
[MacClade 3.07 registered to NCBI]
BEGIN DATA;
DIMENSIONS NTAX=4 NCHAR=48;
FORMAT DATATYPE=DNA MISSING=- GAP=- INTERLEAVE ;OPTIONS
MSTAXA=UNCERTAIN ;
MATRIX
[                          10      20      30      40                  ]
[                          .       .       .       .                   ]
Rhesus_cyto    gtcgacaaggtcaagtccctctccgatgttcgcttcgaaccggaggat [48]
worm           ----------------------cgaagatcgcatccaaccagagcat [48]
fly            -------------------------agatcgtattcgactagagcat [48]
pig            -------------------------aaatggcatccaactagagcat [48]
;
endblock;
>accAF079495[org=Rhesus_cyto]
>[org=worm]
>[org=fly]
>[org=pig]
```

Figure 1. NEXUS alignment file as modified for Sequin submission.

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

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

Official Business
Penalty for Private Use $300