NCBI News

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

Summer/Fall 2004

Entrez Programming Utilities

In dealing with specialized datasets, researchers are often restricted to one of two unattractive choices: either to download an ftp archive containing far more than the data of interest, followed by a round of local parsing, or to access the data interactively, even though the volume of data may render this method cumbersome. To help with the latter method, NCBI provides a suite of programs called the Entrez Programming Utilities (E-Utilities) that allow automated access to the Entrez databases.

What Are the E-Utilities?

The E-Utilities are a set of seven server-side programs that provide a stable interface to the search, retrieval, and linking functions of the Entrez system, using a fixed URL syntax. The output provided by the E-Utilities is in XML format, with the notable exception of the EFetch utility, which returns data in a variety of formats. The E-Utilities are designed to be called from within a computer program that can process their output. Calling an E-Utility from any of the common programming languages— including Perl, Python, and Java—is a simple matter of posting a URL.

The E-Utilities Implement Entrez Functions

Each of the E-Utilities performs a basic task within the Entrez system, and six of the E-Utilities have a direct equivalent in interactive Entrez (Box 1 on page 3). For instance, typing a text query into the NCBI home page and clicking "Go" causes Entrez to search for matches across all Entrez databases and list the number of matching records for each. This "Global Query" function is implemented by EGQuery. If a single database is queried, Entrez first maps the query to a set of integers, or unique identifiers (UIDs), for matching records in the selected database. Entrez UIDs are sometimes referred to as GI numbers for nucleotide and protein, PMIDs for PubMed, and MMDB-IDs for Structures. Entrez queries and the subsequent list of matching UIDs are implemented using ESearch. On the web, Entrez searches are automatically followed by displays of brief record listings, called Document Summaries (DocSums), for matching records. This functionality is implemented by ESummary. Access to full records in an Entrez database on the Web is provided by clicking on the accession of a displayed DocSum. These functions are implemented by EFetch. Accessing records linked to a given record on the Web is as simple as clicking on a link in the Links menu to the right of a DocSum. This linking function is provided by ELink. On the web, Batch Entrez is used to upload a list of UIDs; this function is provided by EPost. EPost places UIDs on the Entrez History server

continued on page 3

PubChem: An Entrez Database of Small Molecules

The NCBI has released three new Entrez databases that link small organic molecules to bioactivity assays, PubMed abstracts, and protein sequences and structures. The new databases constitute the PubChem project at NCBI, a part of the NIH Roadmap Initiative. They are PubChem Substance, PubChem Compound, and PubChem Bioassay.

PubChem Substance currently contains over 800,000 chemical samples imported from 14 public sources including ChemIDplus, the Developmental Therapeutics Program at NCI, KEGG, NCBI

continued on page 4

In this issue

NCBI News

NCBI News is distributed four times a year. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence to *NCBI News* at the address below. To subscribe to NCBI News, send your name and address to either the street or E-mail address below.

NCBI News
National Library of Medicine
Bldg. 38A, Room 3S-308
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 496-2475
Fax: (301) 480-9241
E-mail: info@ncbi.nlm.nih.gov

Editors
Dennis Benson
David Wheeler

Contributors
Monica Romiti
Eric Sayers

Writers
Vyvy Pham
David Wheeler

Editing and Production
Robert Yates

Graphic Design
Robert Yates

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

NIH Publication No. 05-3272

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

GenePlot

Entrez Genome offers a new pairwise comparison tool called GenePlot to visualize similarities among bacterial genomes. Support for fungal genomic comparisons is also planned. To construct a GenePlot, genes are numbered sequentially along the genomic sequences of two organisms and the two corresponding sets of predicted proteins are compared using BLAST. For every case in which a pair or proteins, one from each genome, are mutual best matches, a point is plotted using the indices of the equivalent gene in the two genomes as the X and Y coordinates. Use the GenePlot link from an organism's genome record to see a GenePlot against the organism with which it shares the highest number of reciprocal best hits. Comprisons between other organisms can be made using pull-down menus.

Figure 1 shows a GenePlot display for the comparison of two strains of Helicobacter pylori. An area of focus can be selected by centering the grey cross-hairs with the mouse. The width of the cross-hairs is determined by the zoom level set using the zoom control prior to clicking on the rightmost "close-up" plot. Best-hit pairs visible in the close-up plot are listed below with links to Blink reports for each protein in a pair as well as BLAST2 Sequences pairwise alignments. Use the "save" link to save a table listing all the pairwise best hits.

For nearly identical genomes, the GenePlot consists of a diagonal line running from the lower left to the upper right-hand corners. Closely related genomes may have undergone rearrangements, however, and these are seen as segments displaced from, or running perpendicular to, the main diagonal. Figure 2 shows a case in which two genomic segments from strains of *Helicobacter pylori* are displaced from the main diagonal with a perpendicular orientation. This indicates both a juxtaposition of the two segments but also a reversal of their orientations between the two genomes.

—DW

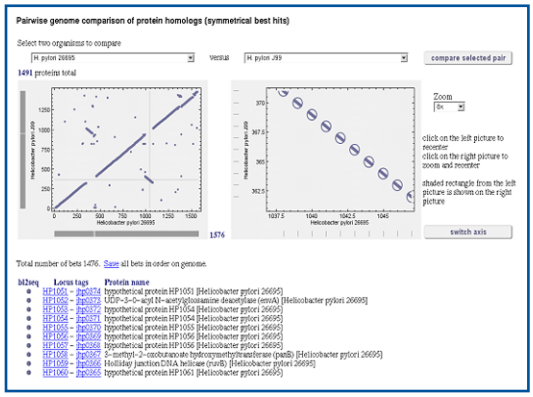

Figure 1. Full GenePlot display showing overview and close-up plots, genome selection pulldown menus, zoom control, gene list, and links to the data in tabular form and BLAST2Sequences displays of pairwise protein alignments.

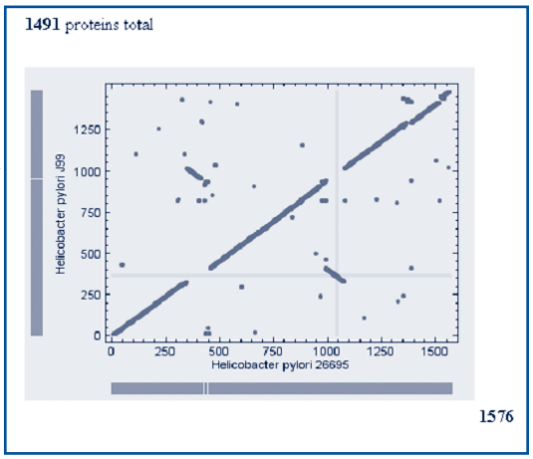

Figure 2. GenePlot comparison of proteins from two strains of Helicobacter Pylori, 26695 and J99. Each point represents a pair of proteins from the two organisms showing a symmetrical best BLAST score; the coordinates of each point correspond to the position of the protein genes in the two genomes. Note the juxtaposition and inversion of two segments of the genome between the two strains. Grey cross-hairs indicate the region displayed in the close-up plot, visible in Figure 1.

NLM Catalog Joins Entrez

The NLM Catalog provides access to NLM bibliographic data for journals, books, audiovisuals, computer software, electronic resources and other materials. Links to the library's holdings in LocatorPlus, NLM's online public access catalog, are also provided.

Entrez E-Utilities
continued from page 1

that stores the results of previous searches during an Entrez session, as can be done on the Web using the Preview/Index or History tabs. The only E-Utility that does not have a direct Web parallel is EInfo, which provides the vital statistics of Entrez databases such as the date of the last update, a list of links to other databases, and a list of indexed fields.

The E-Utilities Search for Data

Suppose that a researcher wants to find all human RefSeq protein records that have links to Online Mendelian Inheritance in Man (OMIM), and thereby have an associated phenotype. This can be done by posting the ESearch URL shown in Example 1 of ESearch in Box 1:

This URL produces XML output, a portion of which is shown below:

```
<Count>14988</Count>
<RetMax>20</RetMax>
<RetStart>0</RetStart>
<QueryKey>47</QueryKey>
<WebEnv>0hh9nVItHLfyYJGaMMIh_T
0ptRqIsaiaikdx5k_yhaM0S72qC5x-
AY</WebEnv>
```

Included is the number of records (14,988) matching the query along with the two parameters that define the location of the data set on the History server: the Query Key, with a value of 47, and the Web Environment (WebEnv), with a value of "0hh9nVItHLfyYJGaMMI . . . ." The latter is a string associated with the internet cookie for the Entrez session.

The E-Utilities Retrieve Data

Retrieving the actual records identified in the above search is performed either using ESummary to retrieve DocSums or using EFetch to retrieve formatted records, such as FASTA sequence. Other available sequence formats include GenBank, GenPept, and INSDSeq XML, which can be selected using the &rettype parameter. One consideration to bear in mind is that EFetch is limited to 500 records per URL. Therefore, to retrieve FASTA sequence for all 14,988 records, a loop within the calling program will be required to post 30 URLs, the second of which, to retrieve records 500-999, is shown in the EFetch section of Box 1.

The remaining 28 URLs would differ only in the value of the &retstart parameter, which would increment by 500 in each successive call within the loop.

The E-Utilities Limit and Link Datasets

To find the annotated genes associated with a select group of this set of RefSeq proteins, namely those that have interleukin 22 in their title, another ESearch URL can be used, as listed under Example 2 of the ESearch section of Box 1, where "%2347" is the URL encoding for "#47" and refers to our previous query key. The five resulting GIs can be extracted from the XML and used as input to ELink, shown in the ELink section of Box 1.

Since each GI was assigned in a separate &id parameter, the XML output will contain separate lists of linked GeneIDs for each protein GI. A simple analysis of the results reveals that the second, third, and fourth protein GIs are linked to the same GeneID, revealing the three transcriptional variants of the interleukin 22 binding protein, a name in turn retrieved by a single ESummary call with that common GeneID.

By using additional combinations of E-Utilities calls, a wide array of data pipelines can be constructed easily and used to process large numbers of data records.

For more information, see the following:

E-utility online documentation:

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

NCBI PowerScripting, a new NCBI course on programming with the E-utils:

www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/course.html

Building Customized Data Pipelines Using the Entrez Programming Utilities (eutils):

www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/chapter_eutils.pdf

—ES

Box 1. Helpful E-Utility URLs and E-Utility Samples — The base E-Utility URL: eutils.ncbi.nlm.nih.gov/entrez/eutils

EInfo (base/einfo.fcgi?) - provides overall statistics of a database, including a list of the indexed fields and available links to other Entrez databases
EGQuery (base/egquery.fcgi?) - equivalent to Global Entrez; provides the number of records in each Entrez database that match a text query
ESearch (base/esearch.fcgi?) - provides the number of records in a specified Entrez database that match a text query Example 1: base/esearch.fcgi?db=protein&term=srcdb+refseq[prop]+AND+human[orgn]+AND+protein+omim[filter]&usehistory=y Example 2: base/esearch.fcgi?db=protein&term=%2347+AND+interleukin+22[title]&usehistory=y&WebEnv=0hh9nVItHLfyY . . .
ESummary (base/esummary.fcgi?) - provides Document Summaries for a set of UIDs from a specified Entrez database
EFetch (base/efetch.fcgi?) - provides formatted data records for a set of UIDs from a specified Entrez database Example: base/efetch.fcgi?db=protein&retmode=text&rettype=fasta&retstart=500&retmax=500&query_key=47&WebEnv=0hh9nVItHLfyY . . .
ELink (base/elink.fcgi?) - provides a set of UIDs in a specified Entrez database that are linked to an input set of UIDs in either the same [or a different] database Example: base/elink.fcgi?dbfrom=protein&db=gene&id=31317239&id=31317243&id=31317241&id=16418459&id=10092625
EPost (base/epost.fcgi?) - posts a set of UIDs in a specified Entrez database to the Entrez History server

Selected Recent Publications by NCBI Staff

To view the citation for any article listed below, click on the PubMed link on the navigation bar at the top of the NCBI Home Page, enter the PubMed ID number (PMID) in the search query box, and click 'Go'.

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, **Anantharaman V**, **Aravind L**, Kapur V. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*. 2004 Apr 16;304 (5669):441-5. Epub 2004 Mar 25. PMID: 15044751

Bazykin GA, **Kondrashov FA**, **Ogurtsov AY**, Sunyaev S, **Kondrashov AS**. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature*. 2004 Jun 3;429 (6991):558-62. PMID: 15175752

Budanov AV, Sablina AA, Feinstein E, **Koonin EV**, Chumakov PM. Regeneration of peroxiredoxins by p53-regulated sestrins, homologs of bacterial AhpD. *Science*. 2004 Apr 23;304 (5670):596-600. PMID: 15105503

Marchler-Bauer A, **Bryant SH**. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res*. 2004 Jul 1;32 (Web Server issue): W327-31. PMID: 15215404

Gibbs RA, et al (incl. **Sherry ST**). The International HapMap Project. *Nature*. 2004 Dec 18;426(6968):789-96. PMID: 14685227

Mills R, **Rozanov M**, Lomsadze A, **Tatusova T**, Borodovsky M. Improving gene annotation of complete viral genomes. *Nucleic Acids Res*. 2003 Dec 1;31(23):7041-55. PMID: 14627837

Yu YK, **Wootton JC**, **Altschul SF**. The compositional adjustment of amino acid substitution matrices. *Proc Natl Acad Sci U S A*. 2003 Dec 23;100(26):15688-93. Epub 2003 Dec 08. PMID: 14663142

PubChem

continued from page 1

MMDB, and the NIST Chemistry WebBook. Chemical entities in PubChem Substance records that have known structures are validated, converted to a standardized form, and imported into PubChem Compound. This standardizing allows NCBI to compute chemical parameters and similarity relationships between compounds. The compounds are grouped into levels of chemical similarity from most general to most specific: same bonding connectivity and any tautomer; same bonding connectivity; same stereochemistry; same isotopes; and same stereochemistry and isotopes. PubChem Compound also indexes these chemicals using 34 fields, many of which represent computed chemical properties such as the number of chiral centers, the number of hydrogen bond donors/acceptors, molecular formula and weight, total formal charge, and octanol-water partition coefficients (XlogP). These groups are provided as Entrez links that allow similar compounds to be retrieved quickly. The third database, PubChem Bioassay, currently includes 173 bioactivity studies from the Develop-mental Therapeutics Program at NCI, and each of these studies is linked to records in PubChem Substance. The PubChem Bioassay interface allows users to view substances that meet certain activity and/or chemical criteria, and the matching records can either be viewed in PubChem Substance or downloaded in several formats.

As part of the Entrez system, the three PubChem databases are linked to several related Entrez databases, including PubMed, Protein, and Structure. PubMed links are derived either from citations provided by submitters or by matching substance names to the MeSH medical thesaurus, which often provide extensive information about the biological activity of a substance. The Protein and Structure links reveal proteins known to interact with a compound and protein structures that contain the compound as a bound ligand. The reverse links also provide new functionalities. Now ligands within structures can be identified instantly by the link to PubChem Compound, as can chemicals described in PubMed abstracts.

Consider Gleevec, a potent tyrosine kinase inhibitor used to treat leukemia. In PubChem Substance, the query "gleevec" retrieves one record for Imatinib mesylate from ChemIDplus. Clicking on the SID (substance ID) number or the thumbnail structure loads a Substance Summary showing a view of the structure, other information including chemical properties and synonyms, and links to PubChem Substance, PubChem Compound, PubMed, and records of identical compounds. This record contains both Imatinib mesylate and methanesulfonic acid; a link to identical compounds leads to substances that also contain the acid. In this case, one additional substance is found that was not retrieved by the query "gleevec", showing how similarity neighboring is able to overcome differing nomenclatures. As part of the standardizing process, substances that have multiple components give rise to several records in PubChem Compound to allow more powerful searching for similar compounds. In the present case, if the Compound Displayed pulldown menu is changed from Standardized to Component1, a different Com-pound record is

continued on page 5

PubChem
continued from page 4

shown that contains Imatinib mesylate without the acid, and this compound is linked to seven identical compounds, including itself (Figure 1). Clicking the link to the right of Same Connectivity loads these identical compounds into PubChem Compound, and then choosing Protein Struc-ture from the Display pulldown menu and clicking Display reveals three crystal structures of tyrosine kinase domains containing bound Gleevec. Only one of these structures would have been found by the text query "gleevec" in Entrez Structure, illustrating the advantage of the precomputed chemical similarities provided by PubChem Compound.

PubChem Bioassay allows one to search for bioactivity. For instance, the query "leukemia AND lc50[tid description]" in PubChem Bioassay retrieves eight growth inhibition assays with measured LC50 values in various leukemia cell lines. Links are then provided to PubChem Substance and PubChem Compound for these chemicals so that they may be further explored.

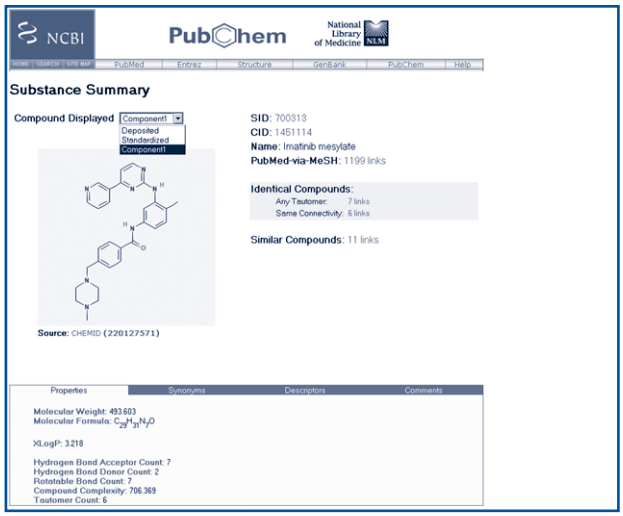

Figure 1. Substance Summary page for Compound ID 1451114 in PubChem Compound, corresponding to Imatinib mesylate (gleevec). The structure displayed is "Component1" of PubChem Substance ID 700313, the originally submitted substance that contains both imatinib mesylate and methanesulfonic acid. The standardized version of the submitted substance, containing the acid, is indexed as Compound ID 1451113 and is viewed by choosing "Standardized" from the pulldown menu.

Access PubChem at
pubchem.ncbi.nlm.nih.gov

—ES

New Genome Builds and Annotations at NCBI

Human Reference Genome Build 35 Version 1

The NCBI Map Viewer now shows Version 1 of NCBI's annotation on the reference human genome build 35. Sequences for two alternatives to the reference assembly's Major Histocompatibility Complex DR51 haplotype, DR52 and DR53, are included in this build.

The Map Viewer also displays, on the "Celera" assembly tracks, the December 2001 whole genome shotgun assembly (WGSA) generated by Celera from 27 million reads of Celera's 5.3X whole genome shotgun data and 104,000 BAC end sequence pairs from GenBank. The Map Viewer shows the alignment of RefSeq RNAs to the Celera genome assembly on the "gene_seq" track, however, GenBank mRNAs and ESTs were not aligned to genomic contigs from Celera, nor were Gnomon predictions made.

The method used to convert sequence coordinates to cytogenetic bands has been changed for Build 35.1 and is described in Furey and Haussler D.[1]

New maps include a Repeats track showing the alignment of repetitive elements detected by RepeatMasker[2]. Also new are the *Gallus gallus* UniGene map showing the alignment of mRNAs from UniGene clusters and the EST map showing the alignment of mRNAs and ESTs.

Mouse Reference Genome, Build 33 Version 1

The Map Viewer now shows mouse Build 33 Version 1 that is based on data available as of June 18, 2004, and includes 25,874 mapped genes. Build 33 was assembled using HTGS phase 3 sequence, single fragment HTGS phase 2 sequence and the †MGSCv3. A tiling path was hand-curated by combining data from clone based Tiling Path Files (TPFs) and the MGSCv3. The Sanger Center's nearly-complete assembly for chromosome 11 is also presented.

New Tracks appearing in Map Viewer include a Repeats track similar to that described above, a Rat UniGene track showing alignment of rat mRNAs, labeled according to the UniGene cluster to which they belong, a Rat EST track, and a track showing CpG islands identified using the algorithm described in Takai and Jones.[3]

Chicken Build 1, Version 1 in Map Viewer

The NCBI Map Viewer now displays NCBI's annotation of the chicken genome assembly from Washington University School of Medicine Genome Sequencing Center, along with genetic and physical maps. Assemblies for chromosomes 25, and 29 through 31, and 33 through 38 are not yet available. Four linkage groups that have not yet been mapped to any chromosome are also represented.

†*Mouse Genome Sequencing Consortium*
[1]*Hum Mol Genet.* 2003 May 1;12(9):1037-44. PMID: 12700172
[2]Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996-2004 <http://www.repeatmasker.org>
[3]*Proc Natl Acad Sci U S A.* 2002 Mar 19;99(6):3740-5. PMID: 11891299

New Microbial Genomes in GenBank

Organism	GenBank | RefSeq Accession Numbers
Staphylococcus aureus subsp. aureus MRSA252	BX571856 | NC_002952
Staphylococcus aureus subsp. aureus MSSA476	BX571857 | NC_002953
Bartonella quintana str. Toulouse	BX897700 | NC_005955
Bartonella henselae str. Houston-1	BX897699 | NC_005056
Bacillus thuringensis serovar Konkukian str. 97-27	AE017355 | NC_005957
Erwinia carotovora subsp. atroseptica SCRI1043	BX950851 | NC_004547
Acinetobacter sp. ADP1	CR543861 | NC_005966
Bacillus anthracis str. 'Ames Ancestor'	chromosome: AE017334 | NC_007530 plasmid pX01: AE017336 | NC_007322 plasmid pX02: AE017335 | NC_007323
Yarrowia lipolytica CLIB99	CR382127—32 | NC_006067—72
Kluyveromyces lactis NRRL Y-1140	CR382121—26 | NC_006037—42
Candida glabrata CBS138	CR380947—59 | NC_005967—68 NC_006026—36
Debaryomyces hansenii CBS767	CR382133—39 | NC_006043—49
Mesoplasma florum L1	AE017263 | NC_006055
Propionibacterium acnes KPA171202	AE017283 | NC_006085
Streptococcus pyogenes MGAS10394	CP000003 | NC_006086
Leifsonia xyli subsp. xyli str. CTCB07	AE016822 | NC_006087
Desulfotalea psychrophila strain LSv54	chromosome: CR522870 | NC_006138 large plasmid: CR522871 | NC_006139 small plasmid: CR522872 | NC_006140
Rickettsia typhi strain wilmington	AE017197 | NC_006142
Borrelia garinii PBi	CP000013 | NC_006156
Yersinia pseudotuberculosis IP 32953	chromosome: BX936398 | NC_006155 plasmid pYptb32953: BX936399 | NC_006153 pYV: BX936400 | NC_006154

For more detailed information, see the online version of the Summer/Fall 2004 NCBI News, or use the GenBank or RefSeq Accession Number to search the Entrez "Genome" database using the query box on the NCBI Home Page.

Trace Archive Grows; Assembly Archive Links Traces to Assemblies

The Trace Archive of sequencing traces has grown to over 500 million sequencing traces from more than 400 organisms. A new Assembly Archive links the raw sequence information found in the Trace Archive with assembly information found in GenBank. An Assembly Viewer allows displays of multiple sequence alignments as well as the sequence chromatograms for traces that are part of assemblies. Both the Trace Archive and Assembly Archive are accessible via links on the NCBI home page.

New Organisms in UniGene

The Entrez UniGene database now offers over 715,000 transcript clusters, linked to nucleotide records, for 48 animals and plants. Recent additions to UniGene include:

Neurospora crassa (a fungus from the Ascomycota) with 28,217 transcript sequences in 3,217 clusters, *Magnaporthe grisea* (rice blast fungus), with 46,488 transcript sequences in 5,721 clusters, and *Ciona savignyi*, with 56,284 transcript sequences in 6,315 clusters.

Whole Genome Shotgun Sequence (WGS) Project Page

WGS sequences appear in GenBank as sets of WGS contigs, many bearing annotations, originating from a single sequencing project. These sequences bear accession numbers consisting of a 4-letter project ID, followed by a two-digit version number, and a 6-digit contig ID. Hence, the WGS accession number "AAAA01072744" is assigned to contig number "072744" of the first version of project "AAAA". Whole Genome Shotgun (WGS) sequencing projects have contributed over 4,000,000 contigs to GenBank and these primary sequences have been used to construct some 237,000 large-scale assemblies of scaffolds and chromosomes. WGS project contigs for *Homo sapiens, Canis familiaris, Pan trodlodytes, Drosophila, Saccharomyces* , and more than 100 other organisms and environmental samples are available. For a complete list of WGS projects with links to the data, see:

www.ncbi.nlm.nih.gov/Genbank/WGSprojectlist.html

New Formatting Options, new Databases in Web BLAST

Selecting the "new formatter" option allows BLAST results to be formatted in a new "Pairwise with identities" mode that highlights differences between the query and a target sequence. The new formatter also offers an option to display masked characters in lower-case and with different colors rather than simply replacing each with an "X" or an "N". A "sequence retrieval" formatting option allows database sequences to be marked for batch retrieval using check boxes appearing in the BLAST results.

A "refseq" database is now available for protein searches and "refseq_rna" and "refseq_genomic" databases are available for nucleotide searches. Also available for nucleotide searches are the "wgs", and "chromosome" databases for Whole Genome Shotgun project sequences and complete genomes, chromosomes, or contigs from RefSeq, respectively. The RefSeq and WGS databases are also available via FTP at:

ftp.ncbi.nlm.nih.gov/blast/db

Submissions Corner

Sequin Enhancements

Sequin is NCBI's primary GenBank® submissions and update tool that can handle large genomes, population and phylogenetic sets, single sequence submissions, Third Party Annotation (TPA) submissions, and updates to records already in GenBank.

A number of enhancements have been made recently to Sequin that make it easier for a user to submit sequences to GenBank.

TPA submissions require several elements that non-TPA submissions do not, such as an explanation of the experimental evidence for the annotation and the primary accession numbers of GenBank records on which the TPA is based. Sequin's self-guided tabs direct the TPA submitter to indicate that the submission is a TPA and a pop-up frame reminds the submitter that in order to be released, TPA records require a publication which describes the biological experiments used as evidence for the annotation. A box beneath the reminder is provided for the submitter to list the evidence and the experiments supporting the submission. The submitter is then returned to the main menu to continue the submissions process. After the sequence file has been imported into Sequin, an "Assembly Tracking" menu appears which allows entry of the primary accession numbers of the sequences used in the TPA.

The Annotate menu offers an updated definition line generator that places the organelle in which a sequence is located at the end of the definition line.

Sequin also has an enhanced alignment reader. Submitters can now specify which characters within an alignment are meant to designate gaps, ambiguities, and match (identical) characters. Different characters can be specified for the Beginning, Middle, and End gap characters. If the alignment used for submission is not valid, errors will be reported to the submitter indicating the specific problem and suggesting possible solutions.

Once an alignment such as a population or phylogenetic set is loaded, users can view all the nucleotides for all the sequences by selecting Alignment from the Format option. Formerly, Sequin presented a graphical view of the alignment. Using this alignment view, and targeting the different sequences, users can study the differences among the sequences in the alignments.

Further extending Sequin's batch processing capability is a 'Batch Feature Apply' option, found under the Annotate menu that allows the annotation of a batch of sequences with global features such as coding regions or source qualifiers. When 'Batch Feature Apply' is selected, a list of various feature types is presented that can be applied to all sequences. The user may choose to have each feature span the entire sequence to which it is applied or the user may specify the left and right ends of the locations for all features.

Finally, enhancements have been made to Sequin's robust editing capabilities. The Update Sequence function has been enhanced to allow the specification of actions to be taken regarding coding regions and references when updating the sequence. If Update Proteins for Updated Sequences is selected, then Sequin will attempt to adjust the locations of coding regions on the updated sequence based on an alignment between the old translated protein and the translation of the updated sequence. Options are also available to truncate retranslated proteins at stops, extend retranslated proteins without stops or extend retranslated proteins without starts. The "Correct CDS" genes function adjusts the corresponding gene span based on the new coding region span.

—MR

RefSeq Version 8 on FTP Site

RefSeq Release 8 is now available by anonymous FTP at:

ftp.ncbi.nih.gov/refseq/release

Release 8 includes genomic, transcript, and protein sequences available as of October 31, 2004 from 2,645 organisms. The number of RefSeq accessions in Release 8 and their combined lengths is given in the shaded box.

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

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

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

www.ncbi.nih.gov/RefSeq

	# of Accessions	# of Basepairs/Residues
Genomic	180,180	26,278,669,655
RNA	311,277	535,717,003
Protein	1,218,266	430,300,369

Change in Definition Lines for Predicted Records

The label "PREDICTED" has been added to the title of RNA records with accessions beginning with "XM_" and "XR_" and to the title of protein accessions beginning with "XP_" to indicate that these sequences are derived from genomic placements and not directly from a cDNA. The label "PREDICTED" does not indicate that the gene itself is predicted, although it may be. The "PREDICTED" label appears in the definition lines seen in retrievals from Entrez , as shown below, and in BLAST results.

GenBank format in Entrez:

LOCUS XM_038604 4998 bp mRNA linear PRI 23-AUG-2004
DEFINITION PREDICTED: Homo sapiens unc-13 homolog A (C. elegans) (UNC13A), mRNA.

BLAST alignment short descriptions:

Sequences producing significant alignments:	Score (bits)	E Value
gi|51493417|ref|XM_038604.8| PREDICTED: Homo sapiens unc-13...	9610	0.0

BLAST 2.2.10 Released

BLAST 2.2.10 features more modular and extensible code as well as a new search engine that significantly enhances performance. The latest version of the BLAST programs, version 2.2.10, is available at:

ftp.ncbi.nih.gov/blast/executables

GenBank Release 144

GenBank Release 144 (October 2004) contains over 38 million sequence entries totaling more than 43 billion base pairs. Release 145 is expected in December. GenBank is accessible via the Entrez search and retrieval system. The flatfile and ASN.1 versions of the Release are found in the "genbank" and "ncbi-asn1" directories respectively at:

ftp.ncbi.nih.gov

Uncompressed, the Release 144 flatfiles consume about 147 gigabytes while the ASN.1 version consumes about 128 gigabytes. The data can also be downloaded at two mirror sites:

genbank.sdsc.edu/pub

bio-mirror.net/biomirror/genbank

Department of Health and Human Services
Public Health Service, National Institutes of Health
National Library of Medicine
National Center for Biotechnology Information
Bldg. 38A, Room 3S308
8600 Rockville Pike
Bethesda, Maryland 20894