NCBI News

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

Volume 15, Issue 1
Summer 2006

New Databases and Tools Target Influenza

Influenza virus infection is a major threat to public health in the United States, resulting in over 200,000 hospitalizations and 30,000 deaths each year. The Influenza Virus Genome Project[1] is providing researchers with a growing collection of virus sequences essential to the identification of the genetic determinants of influenza pathogenicity. NCBI provides online tools for the analysis of these and other influenza sequences in GenBank that allow researchers to:

Retrieve—viral genomic, gene encoding, or protein sequences and download them in a number of formats

Align—locally stored sequences with those in NCBI databases

Cluster—sequences for phylogenetic analysis using a variety of algorithms and weight matrices, constructing dendrograms from the result

Download—complete genomic sequences

Search—influenza sequences using BLAST®

An Example

The analysis of the coding region (CDS) of the hemagglutinin ('HA'), sequence for influenza virus A, GenBank® accession **AY653200**, serves as an example of the use of these tools to classify a new sequence. Prior to the analysis, the CDS portion of the sequence was

continued on page 6

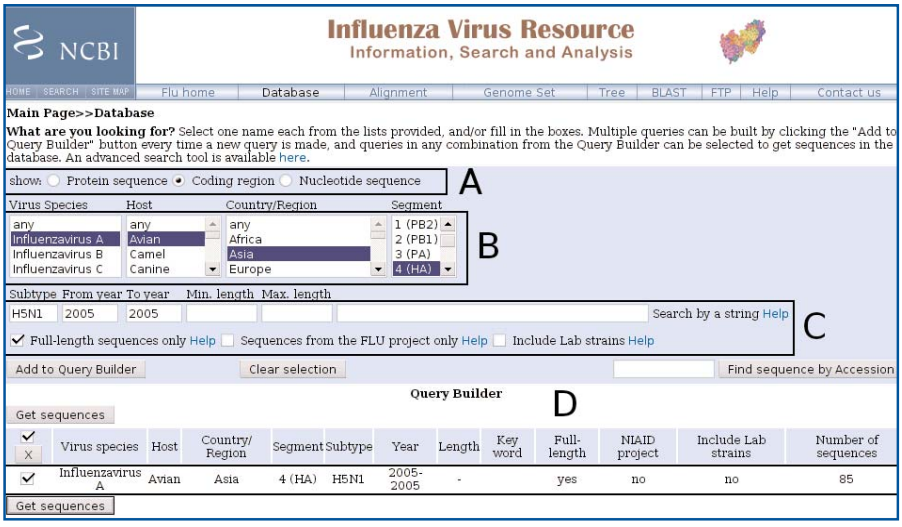

Figure 1. Query Builder for influenza sequences. Queries are built by making selections in three different sections of the form, labeled A, B, and C.

Trace Archives Tops 1 Billion Records

NCBI's Trace Archive now contains over 1.2 billion entries, making it one of the largest publicly accessible biological databases in the world. The database also ranks as one of the most important to the medical research community because it contains the genetic blueprints of hundreds of organisms important to biomedical research.

A Collaborative Effort

The Trace Archive was established in 2001 as a collaborative effort between NCBI and the European Molecular Biology Laboratory (EMBL/ENSEMBL) to collect raw

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
Medha Bhagwat
Rana Morris
Monica Romiti
Tao Tao

Writers
Peter Cooper
Rana Morris
Eric Sayers
Robert Yates

Editing and Production
Robert Yates

Print & Web 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. 06-3272

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

Entrez Nucleotide Split Facilitates Focused Searches

The Entrez Nucleotide database is now partitioned into two specialized components containing Expressed Sequence Tag (EST), and Genome Survey Sequence (GSS) records, respectively, and a third component containing the rest of the nucleotide records, called 'CoreNucleotide' (November 2005 issue of the NCBI News). About half of the 80 million Entrez Nucleotide records fall into the new EST component, with the remainder falling into the GSS and CoreNucleotide components as shown in the chart of Fig. 1 on page 3. The split facilitates searches using specialized field limitations in each of the component databases. The component databases are displayed on the database selection pull-down list on any of the NCBI Entrez pages and searches can still be performed against the combined nucleotide database through the global query or through the 'nucleotide' option. When a search is performed in the Nucleotide database, the counts for matching records in each component database are shown in a statistics line and linked to Document Summary displays, Fig. 1A. This display allows the search to be narrowed immediately to a specific component database. Within the EST and GSS component databases, the dbGSS and dbEST format is the current default display format.

As a consequence of the database split, the 'Preview/Index' tab on the 'Nucleotide' page now displays an intermediate page, shown in Fig. 1B, through which separate 'Preview/Index' forms for the CoreNucleotide, GSS and EST component databases may be selected. The 'Preview/Index pages for the EST and GSS databases allow searches using fields corresponding to sections and identifiers present in the dbGSS and dbEST format that were not available prior to the division of the database.

The new fields for these components include Citation Title, Clone ID, Library Name, GSS/ EST Name, GSS / EST ID and Submitter Name. The GSS component also has a separate Library Class field that allows the selection of different types of genomic DNA libraries. Using these component specific fields to limit the search can result in more precise retrieval than was possible before the division of the nucleotide system. For example, one can retrieve all sequences from a specific cDNA library using the Library Name field limitation. Searching with the phrase Atlantic salmon spleen in the EST component database retrieves over 9,000 records. These are from several different cDNA libraries; all contain the word Atlantic salmon or its taxonomic translation, Salmo salar, and 'spleen' in one or more of the indexed fields. Finding all of these records may be desirable. However the Library Name field can be used to retrieve sequences only from specific libraries. Library names are unique and only those records with the exact phrase as the Library Name field in the native dbEST format are retrieved. The following query retrieves 5,551 records from just two libraries.

atlantic salmon spleen[Library Name] OR atlantic salmon spleen cdna library[Library Name]

Another useful feature of the separate index is the ability to search for clone identifiers. This is especially

continued on page 5

Third Party Annotation Database

NCBI and its collaborating databases, DDBJ and EMBL, have established the inferential portion of the Third Party Annotation (TPA) database to accommodate a wider range of submission types. The TPA database was established in 2002 to allow researchers to submit their own analyses of existing GenBank sequences. TPA submissions may include genomic or transcript sequences assembled from primary data, the annotation of features such as genes, coding regions, and transcripts, or functional annotations of protein sequences. The analysis of the primary sequence data will become increasingly important as unannotated data from genome sequencing and EST projects accumulates. Prior to establishing guidlines for the inferential portion of the TPA database, direct experimental evidence was required for TPA submissions.

To enlarge the scope of third party annotations, the new inferential component of the TPA database allows submissions of sequences and features based on analysis of existing GenBank sequences without direct experimental evidence. This original TPA data and all new TPA submissions that include direct experimental support are now included in the experimental section. Inferred TPA submissions, like their experimental counterparts, must be published in a peer-reviewed journal in order to be released.

TPA records can be retrieved or combined with other Entrez queries using the search term 'tpa [Properties]', and the specific experimental or inferential records can be distinguished by the keywords 'TPA:experimental' or 'TPA:inferential'. TPA records are identified in Entrez or BLAST search results by the 'TPA_exp:' or 'TPA_inf:' labels at the beginning of the definition line visible in the example of Fig. 1.

An example of an inferred TPA record is the assembly and annotation of the complete chloroplast genome for the green alga *Chlamydomonas reinhardtii* (accession **BK000554**) shown in Fig. 1. In this example, as with all inferrential records, there is indirect experimental evidence for the sequence and new annotation including independent evidence for the individual CDSs, structural RNAs, and other features on the organelle's genome assembled from overlapping primary sequences. Genes or other features predicted by computer programs without any further evidence are not accepted in the TPA database as either experimental or inferential submissions.

For more information on TPA submissions, see:

www.ncbi.nih.gov/Genbank/TPA.html

—MB

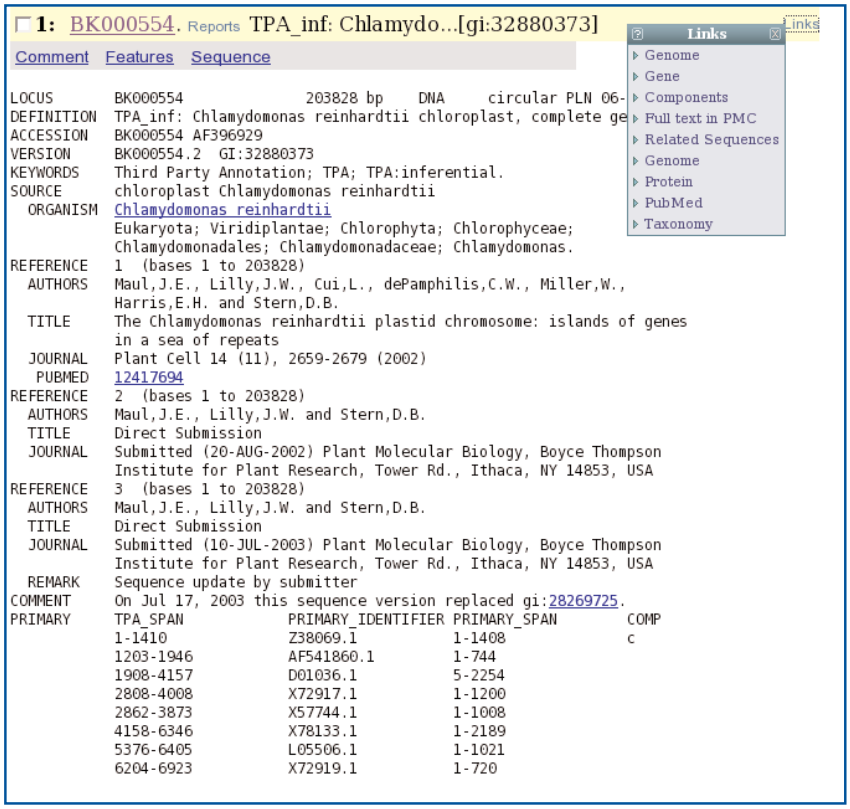

Figure 1. The *Chlamydomonas reinhardtii* chloroplast genome assembled from 101 other GenBank records that are listed in the 'Primary' field. These individual components can be retrieved by following the 'Components' link in the 'Links' menu. Inferred records are easily recognized by the abbreviation TPA_inf in the DEFINITION or TPA:inferential in the KEYWORD sections of the GenBank record.

Trace Archives
continued from page 1

data produced at sequencing centers around the world. Today, these data are submitted to one of two central processing centers—NCBI or the Wellcome Trust Sanger Centre. The amount of data in the archive has doubled every 10 months since 2001 so that it is now an overwhelming 22 trillion bytes in size, large enough to fill a stack of compact disks 10 stories high. New sequencing technologies promise an even sharper increase in data volume in the future. NCBI works closely with the groups pioneering these new techniques to develop the necessary processing , storage and retrieval technologies in advance of the anticipated data influx.

Traces are Pieces of a Puzzle

NCBI's Trace Archive provides direct access to the raw traces, typically between 300 and 1,000 DNA letters in length.

www.ncbi.nlm.nih.gov/Traces/trace.cgi?

Researchers can view and evaluate over 850 assemblies, such as that shown in Fig. 1, of trace-derived sequences for influenza virus. These assemblies are found in the Assembly Archive, a database that builds upon the sequences in the Trace Archive to provide a higher level view.

www.ncbi.nlm.nih.gov/Traces/assembly/assmbrowser.cgi?

A Vital Resource in the Fight Against Disease

Sequencing traces are vital to the hunt for polymorphisms in gene sequences that are linked to disease when they occur in human DNA or linked to virulence when they occur in the DNA of a virus. To further support studies of DNA sequence variability, NCBI maintains the core dbSNP database with detailed information for over 25 million genetic variations, predominantly single DNA letter changes called 'Single Nucleotide Polymorphisms'. The trace data, combined with that of dbSNP, is a boon to medical researchers seeking to gain greater insight into the impact of genetic variation on health. Trace sequences may be searched using MegaBLAST, or via the web-based form at

www.ncbi.nlm.nih.gov/BLAST

(see the 'Mammoth found in Trace Archive' section of the "Mammoths and Moas. . ." article on page 9 of this issue.)

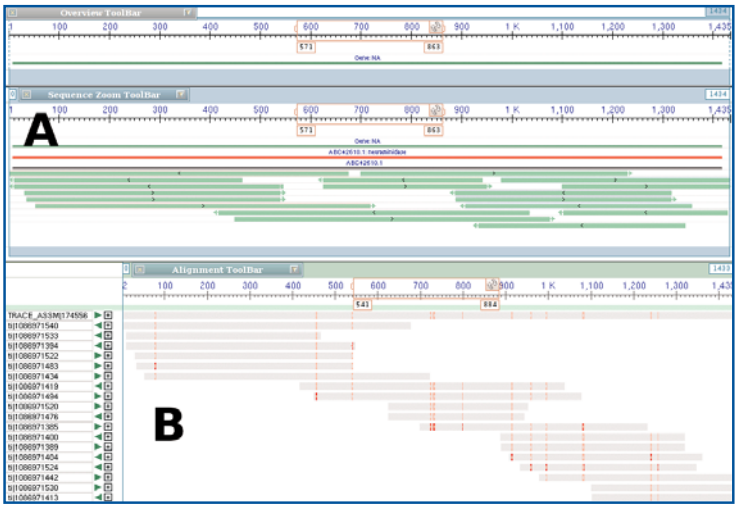

Figure 1. Assembly Viewer display for a neuraminidase assembly from influenza virus traces. The overlapping traces comprising the assembly are shown in panel A. Detailed alignments are shown in panel B with mismatches highlighted. The assembled virus sequence is one of over 850 trace assemblies available in NCBI's Assembly Archive.

RefSeq Release 18

RefSeq release 18 includes genomic, transcript, and protein sequences available as of July 1, 2006, from 3,497 organisms. The number of RefSeq accessions in Release 18 and their combined lengths is given in Table 1. RefSeq releases are posted every two months, and the next release is scheduled for September, 2006. Release notes documenting the scope and content of the database are provided at:

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

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

www.ncbi.nlm nih.gov/RefSeq

Download RefSeq by anonymous FTP at:

ftp.ncbi.nih.gov/refseq/release

Sequence Type	Number	Total Residues
Genomic	713,768	61,014,517,384
RNA	654,553	1,115,519,987
Protein	2,631,538	927,587,669

Table 1. Statistics for RefSeq release 18.

Entrez Nucleotide dbSplit
continued from page 2

helpful for retrieving EST clones from the Integrated Molecular Analysis of Genomes and their Expression (IMAGE) repository:

The query 'IMAGE 8635484[Clone ID]' finds both the 3' and 5' EST reads from that clone. Likewise matching end sequences from BAC clones can be easily be retrieved from the GSS component with a similar query, for example CH252-49B15[Clone ID]. As is true for other Entrez databases, each nucleotide component database as well as the combined nucleotide database has advanced search options and functions available through the Limits, History, and Clipboard. These are accessible from the corresponding tabs below the search box. Because the indexed fields differ for each of the component databases, the advanced features available through the Preview/Index and Details tabs can only be used after selecting a component database. Clicking on the either of these tabs from a combined database search produces an information screen, as shown in Fig. 1B, with options to select either the CoreNucleotide, EST, or GSS component database to access these features. However searches that include a field limiter valid in only one of component databases can be run without error in the combined nucleotide database. For example, the following search in the umbrella Nucleotide database that includes the GSS indexed Field Limit, [Library Class], returns BAC end sequences from the GSS database:

bac ends[Library Class] AND bos taurus[organism]

The standard options for displaying and saving records in various formats are available through the 'Send to' pull-down list in each of the component databases and on the combined results page. Batch retrieval through Batch-Entrez works as before and identifiers valid in any of the three component database can be uploaded for retrieval. If the list contains identifiers from EST, GSS, and CoreNucleotide the split result counts for each of the component database will be shown as with standard Web Entrez searches. The entire results set or results for a specific component database can be formatted and saved to file. Search tips and help for the Nucleotide split database are found at:

www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp

—MR

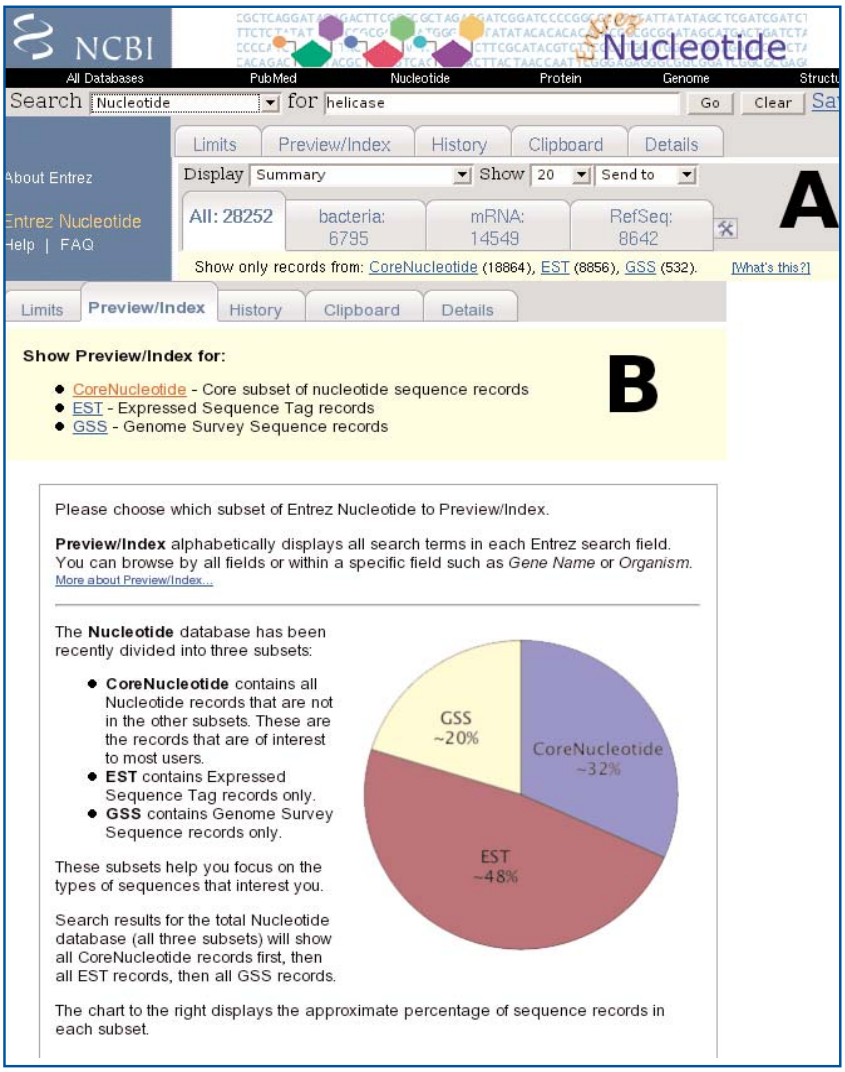

Figure 1. **A.** Result counts for searches of Entrez Nucleotide are given for each of the three component databases, with links to Document Summaries. **B.** Preview/Index pages are now maintained separately for each of the component databases. Clicking on the 'CoreNucleotide' link will display its 'Preview/Index' form, allowing the existing query of 'helicase' to be refined using advanced query-building tools.

Influenza Virus Resource
continued from page 1

downloaded in FASTA format using NCBI's Entrez, and the FASTA definition line was changed from:

```
>gi-50365728:29-1735 Influenza A
virus(/chicken/Jilin/9/2004(H5N1))
segment 4, complete sequence
```

to read:

```
>local chicken
```

Selection of influenza sequences for analysis

To begin, use the Database link from the Influenza Virus Resource page at

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

to reach the Query Builder shown in Fig. 1. Check the 'Coding region' radio button, indicated in section A, to specify the type of sequence to retrieve.

From the menus in section B, select 'Influenza A', 'Avian', 'Asia', and 'HA' as the 'Virus Species', 'Host', Country/Region', and 'Segment', respectively. In addition, check 'Full-length sequences only' and restrict the search to H5N1 subtype sequences from the year 2005 using the check boxes and text fields in section C. Clicking on 'Add to Query Builder' will return the number of sequences that match, as shown in section D. Click on 'Get sequences' to generate the form shown in Fig. 2, containing a table of summaries for the 85 selected sequences. The table is sortable and the controls in section A have been used to sort the records by "Virus Name", after which 10 sequences from various hosts (3 goose, 1 quail, 2 duck, 2 chicken, 1 gull, 1 heron) have been selected for further analysis using the check boxes next to each entry-only the first two of the checked entries are visible in the figure. Using the button in section B, the FASTA sequence called "local chicken" has been uploaded, as indicated in section C.

Multiple sequence alignment

Click on 'Do multiple alignment' to align the "local chicken" sequence to the selected 85 database sequences using the multiple sequence alignment program MUSCLE[2], to generate the alignment shown in Fig. 3. The portion of the alignment displayed, indicated in section A, begins near base 950 and ends near base 1040. Two major groups of sequences, characterized by non-synonymous base changes, sections B, one synonymous base change, section C, and a three-base deletion, section D, are evident.

Clustering and Phylogenetic analysis

Click on 'Build a Tree' to invoke the setup page for phylogenetic analysis where the sequences may be selected for inclusion in the subsequent analy-

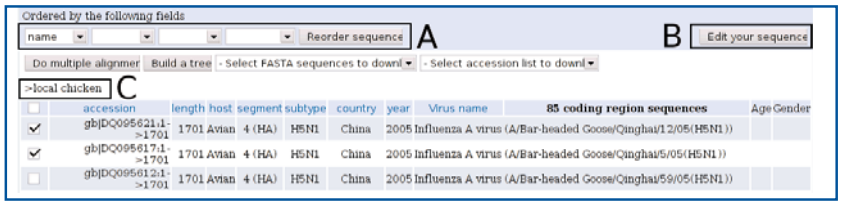

Figure 2. Selection of sequences for further analysis. For brevity, only the first three of 85 selected entries is shown.

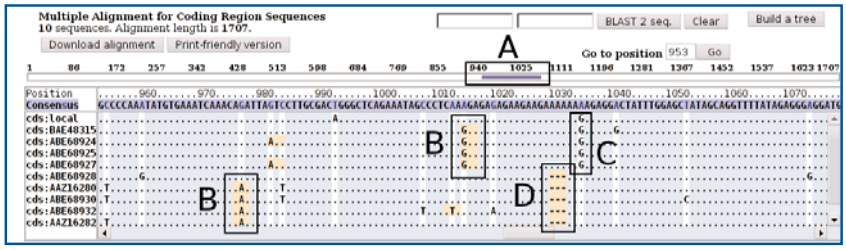

Figure 3. Multiple sequence alignment for the the "local chicken" HA sequences and 10 influenza HA coding sequences selected from the NCBI databases.

1918 Killer Flu Virus Sequence in GenBank

In the fall of 2005, scientists announced that a killer had been resurrected. The 1918 Spanish Flu virus was reconstituted using sequences in GenBank combined with newly sequenced influenza RNA polymerase. The fully reconstituted 1918 virus showed a frightening degree of pathogenicity, causing severe lung pathology and rapid death in laboratory mice unlike contemporary flu viruses that are not lethal to laboratory mice. Another finding, that the 1918 virus was likely a human-adapted avian virus, increases fears that another deadly pandemic virus arising from the H5N1 avian flu viruses in circulation today may be possible.

Sequences for the 1918 virus were obtained from the remains of victims of the deadly wave of influenza that swept across North America in the fall of 1918 killing an estimated 675,000 Americans. The most complete set of flu sequences came from a lung biopsy taken in 1997 from the frozen remains of a flu victim who died in Teller Mission (now called Brevig Mission), Alaska, in November of 1918. Other flu DNA segments were obtained from formalin-fixed paraffin-imbedded tissues archived by the Armed Forces Institute of Pathology. These were taken from two soldiers, one who died of the flu at Fort Jackson, South Carolina, and another at Camp Upton, New

sis using check boxes. Click on 'Phylogenetic Analysis' to display the next page where a clustering algorithm may be selected, and the tree built. The resulting dendrogram is shown in Fig. 4.

The dendrogram shows two clusters, as might be anticipated on the basis of the alignment of Fig. 3. Two influenza sequences from a goose host and one from a gull host lie in the first of these clusters while three from a chicken host, including our "local chicken" sequence, two from a duck and one from a heron host are in the second cluster. An outlying sequence, branching from the base of the tree, came from a goose host in Mongolia. The dendrogram may be recomputed after adjusting several parameters. A 'non-linear' two dimensional dot plot (not shown) that groups sequences to provide an overview of a large dataset may also be generated.

Phylogenetic comparisons of this type have provided valuable insight into the process of genomic reassortments in influenza that lead to influenza outbreaks[3].

—TT

[1] Ghedin E, *et al.* Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature.* 2005 Oct 20;437(7062):1162-6. Epub 2005 Oct 5. PMID: 16208317.

[2] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004 Mar 19;32(5):1792-7. Print 2004. PMID: 15034147

[3] Holmes EC, *et al.* Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol.* 2005 Sep;3(9):e300. Epub 2005 Jul 26. PMID: 16026181

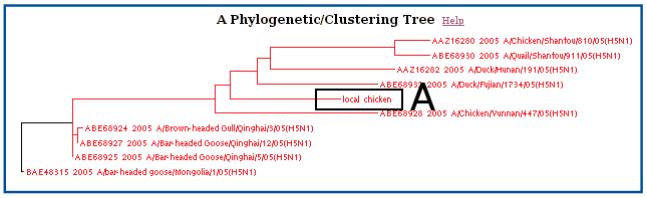

Figure 4. Dendrogram built using the Local Search Neighbor Joining method.

Figure 1. Taxonomy report for the 1918 Brevig Mission virus. The Nucleotide and Protein links retrieve the eight genomic RNA segments and the ten encoded proteins of the virus. The sequence of the 1918 virus shows unique sequence differences from contemporary H1N1 viruses, but it remains unclear how each these contribute to the pathogenicity of the strain. What is clear is that the overall effect of these differences is to produce a devastating virus, unmatched in its ability to wreak havoc.

York, both in September of 1918. Finding these 1918 virus segments at the NCBI is most easily accomplished using the taxonomy database. A global query for 'influenza A virus' finds one match in taxonomy. This link leads to the Taxonomy Browser where the different subtypes and strains are shown. The 1918 virus is of the H1N1 subtype. The three strains from 1918 are Influenza A virus (A/Brevig Mission/1/1918(H1N1)), Influenza A virus (A/South Carolina/1/18 (H1N1)), and Influenza A virus (A/New_York/1/18 (H1N1)). The NCBI taxonomy entry for the Brevig Mission virus is shown in Fig. 1.

New Organisms in UniGene

The Entrez UniGene database now offers over 1,844,162 transcript clusters, linked to nucleotide records, for over 70 animals and plants. Recent additions to UniGene include: *Aedes aegypti* (yellow fever and dengue virus mosquito) with 241,102 transcript sequences in 15,182 clusters, *Aquilegia formosa x Aquilegia pubescens* (hybrid columbine) with 72,522 transcript sequences in 7,675 clusters, *Gossypium hirsutum* (upland cotton) with 83,321 transcript sequences in 10,845 clusters, *Macaca fascicularis* (crab-eating monkey) with 62,745 transcript sequences in 7,488 clusters, *Oryctolagus cuniculus* (rabbit) with 10,827 transcript sequences in 3,766 clusters, *Pimephales promelas* (fathead minnow) with 237,026 transcript sequences in 18,541 clusters, and *Tribolium castaneum* (red flour beetle) with 27,233 transcript sequences in 6,328 clusters.

GenBank® Release 155

GenBank® Release 155 (August 2006) contains over 61 million sequence entries totaling more than 65 billion base pairs. Release 156 is expected in October. 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 155 flatfiles consume about 230 Gigabytes while the ASN.1 version consumes about 199 Gigabytes. The data can also be downloaded at a mirror site:

bio-mirror.net/biomirror/genbank

Mammoths and Moas at NCBI

The presence of ancient DNA sequences in GenBank was noted in the Spring 1999 issue of the NCBI news. Many of the ancient DNA sequences then available were sequences of extant species from ancient material such as human mummies. Some others were from long-extinct taxa such as the saber toothed cat, ground sloth, mammoth and mastodon. But all were short fragments of mainly mitochondrial sequences, a far cry from what would be needed to garner comprehensive genomic information. Six years later, at the close of 2005, the prospect of reconstructing the genomes of ancient extinct organisms has come tantalizingly close with publications of a large quantity of genomic sequence[1] and a complete mitochondrial genome[2,3] from the extinct woolly mammoth (*Mammuthus primigenius*). The groups produced the greatest quantity of sequence ever obtained (28 million base pairs) from an extinct organism and the most complete organelle genome from an extinct organism. Both the woolly mammoth genomic sequence and mitochondrial genome have been deposited at NCBI and are available for searching through NCBI's trace archive, the BLAST services, and the integrated Entrez system. This article shows how to retrieve and work with these woolly mammoth sequences. It also provides an update on the extinct organism sequences at the NCBI with tips on using the Entrez system, the taxonomy browser, the Trace Archive, and the BLAST service to access these interesting data.

Extinct Organisms in Entrez: Sequences and Taxa

The mammoth mitochondrial genome is available in GenBank as well as in the Reference Sequence database, Table 1. These sequence data are fully integrated into the Entrez nucleotide, protein, gene, and genome databases. The Entrez system and the taxonomy database can be used to retrieve these and a number of records for other extinct taxa.

A remarkable number of woolly mammoth nucleotide sequences are retrieved with the following organism query, which can be entered in the search box on the NCBI Homepage:

Woolly Mammoth [Organism]

The initial set of nucleotide sequences can be further limited to mitochondrial records by using the 'Limits' tab and limiting to 'Mitochondrion' in the 'Gene location' pull-down list. However, simply adding 'mitochondrion' as a Title search retrieves only the complete mitochondrial genomes from GenBank (accessions **DQ188829** and **DQ316067**) and RefSeq (accession **NC_007596**).

Woolly Mammoth [Organism] AND mitochondrion[Title]

An easy way to access data for all extinct taxa including the woolly mammoth is via the NCBI Taxonomy services.

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

Here, a list of extinct taxa can be retrieved using the following query:

Extinct[Properties]

An informative view of these organisms is available by selecting 'Common Tree' from the display options pull-down list. This view arranges these taxa according to taxonomic classification and makes it easy to identify the major groups of organisms present (Fig. 1). Several different ages and extinction events are represented by these organisms.

Most of the placental mammals (Eutheria) including the woolly rhinoceros (*Coelodonta antiquitatis*), mammoth (*Mammuthus primigenius*), American mastodon (*Mammut americanum*), cave bear (*Ursus spelaeus*), saber-toothed cat (*Smilodon fatalis*) and ground sloths (*Nothrotheriops shastensis*, *Mylodon darwinii*) and even *Neanderthal* man (*Homo sapiens neanderthalensis*) are from the ice age fauna of the Northern hemisphere. These organisms disappeared at the end of the Pleistocene, around 10,000 years ago, but are well represented because there are abundant well-preserved bones and even frozen soft tissues that make it possible to obtain relatively intact DNA samples. Increased hunting pressure accompanying the rise and spread of modern humans may have played a role in the disappearance of these mammals although climate and vegetation change may also have been important.

There is another series of extinctions more closely linked to the arrival of humans that is documented in the extinct birds (Aves). Two of the birds listed, the great auk (*Pinguinus impennis*) and the dodo (*Raphus cucullatus*), are by now standard icons of the devastating consequences of human predation in historic times. Many of the other extinct bird species are testimony to the arrival of modern humans in New Zealand less than 1,000 years ago. Prior to the arrival of the Maori settlers, there were no terrestrial mammals on the islands. In New Zealand many of the ecological roles occupied by mammals in other places were occupied by the flightless moas (Dinornithiformes). All of the moas became extinct within 200 years of the arrival of mammals in the form of man, his domestic animals, and camp followers such as rats. The giant eagle, *Harpagornis moorei*, that likely preyed on moas, also disappeared, along with a number of other New Zealand endemics. Twenty one of the 25 extinct birds with sequences in GenBank were former residents of the New Zealand archipelago. While many of the sequences for the extinct birds are short fragments, nearly complete mitochondrial

genomes are available for three moas; the little bush moa (*Anomalopteryx didiformis*)[4], the eastern moa (*Emeus crassus*)[5] and the giant moa (*Dinornis giganteus*)[5] (Table 1). A similar set of extinctions to those on the New Zealand archipelago followed the arrival of Polynesians on the Hawaiian Islands 1,600 years ago and resulted in the loss of perhaps half of the endemic species of birds. This extinction event is represented at NCBI by sequences from two of the giant flightless waterfowl species once present on the islands—*Thambetochen chauliodous* and the Giant Hawaiian Goose (*Branta sp.*).

Woolly Mammoth found in Trace Archive

The woolly mammoth genomic sequence is made up of short unassembled environmental sequence reads that are accessible through a Trace Archive query or through the Trace Archive BLAST services:

www.ncbi.nlm.nih.gov/Traces/

(see the "Trace Archives" article on page 1 of this issue.)

The woolly mammoth sequences can be displayed and downloaded by entering the following query in the Trace Archive search box:

species_code='MAMMUTHUS PRIMIGENIUS'

There are 302, 692 traces for woolly mammoth. Only 40 thousand traces at a time can be obtained from the Trace Archive Web service, so a better option for getting all of the mammoth data is the species ftp directory:

ftp.ncbi.nlm.nih.gov/pub/TraceDB/mammuthus_primigenius/

The Trace Archive sequences can be searched using the specialized trace archive megablast services linked to the nucleotide section of BLAST homepage:

www.ncbi.nlm.nih.gov/BLAST

as well as on the Trace Archive homepage. Both a discontiguous megablast service for cross-species comparisons and a standard megablast service for intra-species comparisons are available. The woolly mammoth traces are available in the 'Mammuthus primigenius-other' database through the trace archive megablast services.

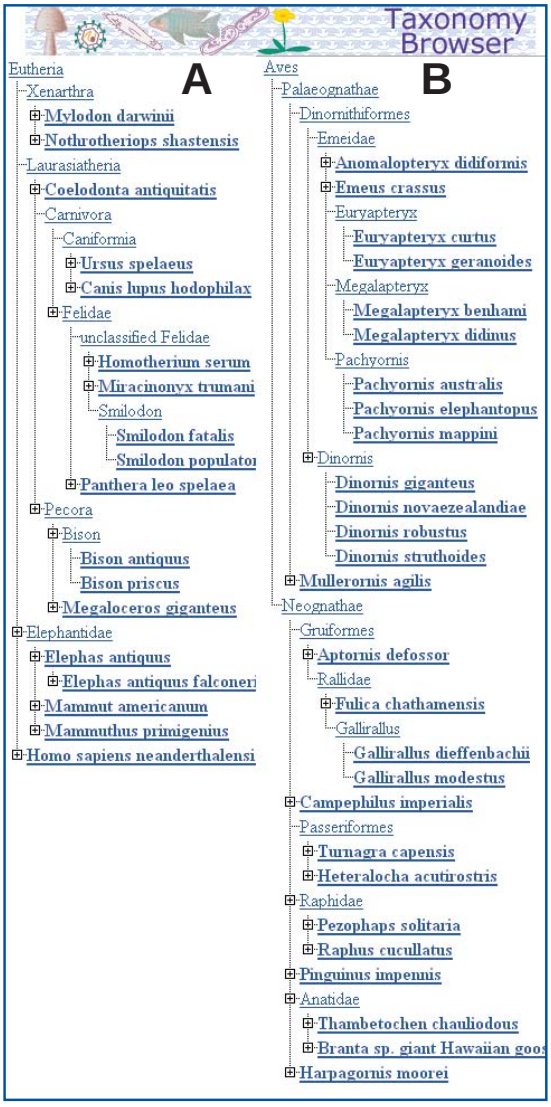

Figure 1. Common trees generated from the Entrez taxonomy service showing the extinct mammals and birds. A. Selected placental mammals from the Pleistocene. B. Extinct birds. All bird species except the great auk (*Pinguinus impennis*), the dodo (*Raphus cucullatus*), the Rodrigues Solitaire (*Pezophaps solitaria*), the elephant bird (*Mullerornis agilis*), and the imperial woodpecker (*Campephilus imperialis*) were endemic to the New Zealand or Hawaiian archipelagos.

Species	GenBank Acc.	RefSeq Acc.	Age of Source
Woolly Mammoth	DQ188829	NC_007596	12, 000 yrs
Woolly Mammoth	DQ316067	-	33, 000 yrs
Little Bush Moa	AF338714	NC_002779	Not reported
Eastern Moa	AY016015	NC_002673	1,200 yrs
Giant Moa	AY016013	NC_002672	600 yrs

Table 1. GenBank and RefSeq accession numbers of mitochondrial genomes of extinct organisms at NCBI.

Improvements in techniques for handling ancient biological samples and improvements in DNA sequencing technologies will increase the availability of molecular data from the recent as well as the distant past. These data will provide important insights into the phylogenetic relationships of extinct taxa, and details of the make-up of past biological communities. The Entrez system and the NCBI BLAST services will continue to provide rapid and powerfully integrated access to these important data.

[1]Poinar HN, *et al.* Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science.* 2006 Jan 20;311(5759):392-4. Epub 2005 Dec 20. PMID: 16368896.

[2]Rogaev EI, *et al.* Complete mitochondrial genome and phylogeny of Pleistocene Mammoth Mammuthus primigenius. *PLoS Biol.* 2006 Mar;4(3):e73. Epub 2006 Feb 7. PMID: 16448217.

[3]Krause J, *et al.* Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature.* 2006 Feb 9;439(7077):724-7. Epub 2005 Dec 18. PMID: 16362058.

[4]Haddrath O, Baker AJ. Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis. *Proc Biol Sci.* 2001 May 7;268 (1470):939-45. PMID: 11370967.

[5]Cooper A, et al. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature.* 2001 Feb 8;409 (6821):704-7. PMID: 11217857.

Two NCBI Papers most cited in 2004-05

Two NCBI Papers, "CDD: a Conserved Domain Database for protein classification" (PMID 15608175) and "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins" (PMID 15608248) are ranked 28th and 36th, respectively, out of 40 in "The Hottest Research of 2004-05" published in the March-April 2006 issue of *Science Watch*®.

The CDD[2] paper discusses the protein classification component of NCBI's Entrez query and retrieval system. The RefSeq[3] paper discusses the RefSeq database, which provides a curated, non-redundant collection of genomic, transcripts, and protein sequences.

[1]King, C. "The Hottest Research of 2004-05". *Science Watch*. 2006 March-April. 17:2. online: <<http://www.sciencewatch.com/march-april2006/sw_march-april2006_page1.htm>>.

[2]Marchler-Bauer A, et al. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res*. 2005 Jan 1;33(Database issue):D192-6. PMID: 15608175.

[3]Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 2005 Jan 1;33(Database issue):D501-4. PMID 15608248.

Selected Recent Publications by NCBI Staff

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

Chakrabarti S, **Lanczycki CJ**, **Panchenko AR**, **Przytycka TM**, **Thiessen PA**, **Bryant SH**. Refining multiple sequence alignments with conserved core regions. *Nucleic Acids Res*. 2006 May 17;34(9):2598-606. Print 2006. PMID: 16707662

Yampolsky LY, Allen C, **Shabalina SA**, **Kondrashov AS**. Persistence time of loss-of-function mutations at nonessential loci affecting eye color in *Drosophila melanogaster*. *Genetics*. 2005 Dec; 171(4):2133-8. Epub 2005 Aug 22. PMID: 16118190.

Ghedin E, Sengamalay NA, Shumway M, Zaborsky J, Feldblyum T, Subbu V, Spiro DJ, Sitz J, Koo H, **Bolotov P**, **Dernovoy D**, **Tatusova T**, **Bao Y**, St George K, Taylor J, **Lipman DJ**, Fraser CM, Taubenberger JK, Salzberg SL. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*. 2005 Oct 20;437(7062):1162-6. Epub 2005 Oct 5. PMID: 16208317.

Morgulis A, **Gertz EM**, **Schäffer AA**, **Agarwala R**. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*. 2006 Jan 15;22(2):134-41. Epub 2005 Nov15. PMID: 16287941.

Iyer LM, **Balaji S**, **Koonin EV**, **Aravind L**. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Research. 2006 Apr;117(1):156-84. Epub 2006 Feb21. PMID: 16494962.

NCBI Courses

NCBI Courses are a great way for researchers, students, librarians, and teachers to keep up with the ongoing enhancements made to NCBI's molecular biology resources.

The courses are offered free of charge at NCBI and at universities and research institutes throughout the United States. For detailed course descriptions and schedules of upcoming courses, see the NCBI Education Homepage at

www.ncbi.nlm.nih.gov/Education

NCBI Technical Workshop Series — NCBI 4-PAK

September 6-7, 2006

NCBI, Bethesda, MD

Four problem or resource-based bioinformatics modules demonstrating the practical applications of NCBI resources—BLAST, Map Viewer, and Entrez.

To register, and for more info:

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

Exploring 3D Molecular Structures Using NCBI Tools

September 21, 2006, 9 AM—5 PM

NCBI, Bethesda, MD

Lectures and hands-on computer workshops on effectively using NCBI 3D macromolecular structural databases, search services, and analysis tools.

To register, and for more info:

www.ncbi.nlm.nih.gov/Class/Structure/nlm.html

A Field Guide to GenBank and NCBI Molecular Biology Resources

October 4-5, 2006, 9 AM—5 PM

National Library of Medicine, Bethesda, MD

General introduction to NCBI molecular biology databases and tools. Lecture and hands-on workshop featuring practical examples using BLAST, Entrez, genomic, and structure resources.

To register, and for more information:

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

New TreeView Display Option in NCBI BLAST

The new Tree View option on the NCBI Web BLAST service presents a dendrogram or tree display that clusters sequences according to their distances from the query sequence. This display is helpful for recognizing the presence of aberrant or unusual sequences or potentially natural groupings of related sequences such as members of a gene families or homologs from other species in the BLAST output.

The link labeled 'Distance tree of results' that leads to the Tree View display appears on the BLAST output for all DNA-DNA or protein-protein comparisons. Trees can be rendered as rectangular, slanted (cladogram), radial or force displays by the selecting the corresponding tab from the tree output. Both the rectangular and radial outputs are scaled to show the distances between sequences. Subsets of the tree or any of the alignments can be displayed through the 'Show subtree' or 'Show alignment' links pop-up menu that appears at the internal nodes of the tree on mouse-over. The following two examples show how the tree display can produce useful trees for nucleotide or protein sequences.

A Tree based on nucleotide-level comparisons

Fig. 1 shows a radial tree display generated by searching against the 'refseq_genomic' database with the woolly mammoth complete mitochondrial genome (RefSeq accession **NC_007596**). The RefSeq genomic database was limited to the mammalian taxon 'afrotheria'. This tree reconstructs the accepted taxonomic groupings of these mammals and reinforces the proposition that the woolly mammoth is most closely related to the African and Asian elephants.

A tree based on protein-level comparisons

A protein tree is shown in Fig. 2. This tree was generated from the results of a search using a Trypanosome arginine kinase (RefSeq accession **XP_826998**) against the Swiss-Prot database. This tree shows two distinct groups of related enzymes in the results, creatine kinases from vertebrates and arginine kinases from arthropods, a few other invertebrates, and the trypanosomes. The tree highlights the surprisingly close relationship between the trypanosome proteins and the arthropod proteins that has given rise to the hypothesis that the trypanosome arginine kinase genes were acquired by horizontal transfer from an arthropod host of these parasites.[1]

Method and caveats

The BLAST tree display is created from genetic distances calculated using standard methods from the aligned sequences—Jukes-Cantor[2] for nucleotide comparisons, Kimura's method[3] for proteins. The trees themselves are then built from these distance matrices using either the Fast Minimum Evolution (FastME)[4], or Neighbor Joining[5] methods. Since BLAST is used to create the alignments, the database sequences are only compared and aligned to the query, not to each other as they would be in a multiple sequence alignment. Therefore it is important to keep in mind that the alignments used to create the Tree View display may differ from the global multiple sequence alignments typically used to infer phylogenies. But, as the above examples show, with judicious choice of query sequence and control over the database sequences searched very informative trees can be generated.

—TT

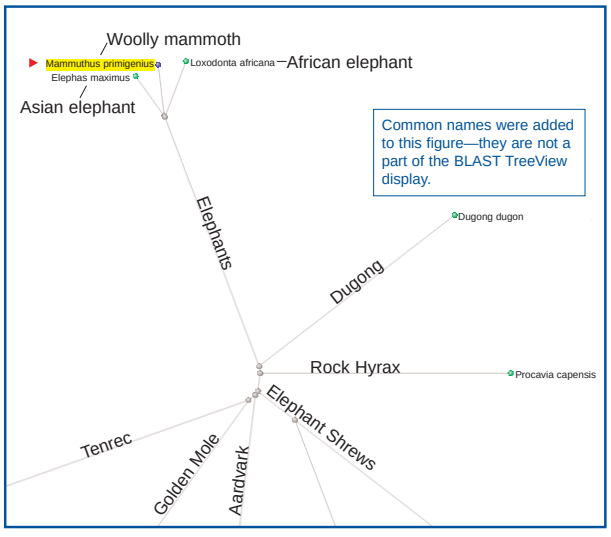

Figure 1. Radial tree display generated by searching against the refseq_genomic database with the woolly mammoth complete mitochondrial genome (RefSeq accession **NC_007596**, shaded). The mammoth sequence, highlighted, clusters closely with that of the modern elephants.

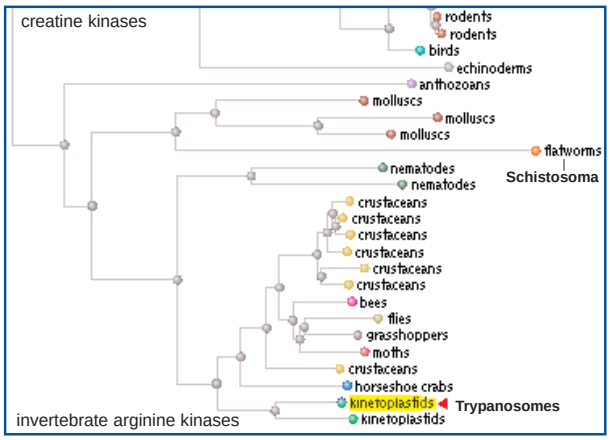

Figure 2. A portion of the rectangular tree generated from a search using a Trypanosome arginine kinase (RefSeq accession **XP_826998**) against the Swiss-Prot database showing the invertebrate arginine kinases. The terminal nodes are labeled by 'blast name' to highlight taxonomic trends. The trypanosome query sequence clusters with the arthropod sequences suggesting the possibility of horizontal gene transfer between these phylogenetically distant groups. A sequence from the flatworm *Schistosoma mansoni* clusters with four molluscan sequences suggesting another possible case of gene transfer. View the full results live in the online NCBI News:
ncbi.nlm.nih.gov/About/newsletter.html

[1]Pereira CA, Alonso GD, Paveto MC, Iribarren A, Cabanas ML, Torres HN, Flawia MM. Trypanosoma cruzi arginine kinase characterization and cloning. A novel energetic pathway in protozoan parasites. *J Biol Chem.* 2000 Jan 14;275(2):1495-501. PMID: 10625703

[2]T. Jukes, C. Cantor, in *Mammalian Protein Metabolism,* H.N. Munro, J. Allison, Eds. (Academic Press, New York, 1969, vol. 3 pp. 21-32).

[3]M Kimura. The Neutral Theory of Molecular Evolution. Cambridge University Press, 1983.

[4]Desper R, Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol.* 2002;9(5):687-705. PMID: 12487758.

[5]Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987 Jul;4(4):406-25. PMID: 3447015.

Genome Builds and Map Viewer Displays

Rhesus macaque now in Map Viewer

The genome of the rhesus macaque, *Macaca mulatta*, a widely used primate model organism, is now, available in the Map Viewer via a link on the Map Viewer home page at:

www.ncbi.nlm.nih.gov/mapview/

The macaque genome build 1.1 is NCBI's assembly and annotation of the Macaque Genome Sequencing Consortium version 1.0 whole genome shotgun (WGS) assembly. The assembly provides 5.1X coverage of the genome with a total sequence length of 2.87 gigabases and is anchored to the 20 pairs of macaque autosomes, the X, and the Y chromosomes. The mitochondrial genome given in NCBI RefSeq **NC_005943**, derived from GenBank record **AY612638**, is also displayed in the Map Viewer. Sequence maps available within Map Viewer include the NCBI contigs, the WGS sequences and the location of genes, STSs, ESTs, UniGene clusters and Gnomon predicted gene models. Maps of aligned human transcripts are available to aid in locating and evaluting macaque gene annotatons. Currently, a total of 24,974 genes and their transcripts are placed on the macaque sequence. Accompanying the sequence maps is a radiation hybrid map containing 800 markers covering all macaque chromosomes. As with all genomes with assembled sequences in the Map Viewer, a Genome BLAST page is available that allows searches against the genome and specialized sets of macaque sequences including GenBank and RefSeq mRNA and protein sequences, expressed sequence tags, high throughput genomic sequence, whole genome shotgun and trace archive sequences. Results of searches agains the assembled genome can be displayed in the macaque map viewer to provide essential genomic contextual information. Genomic BLAST may be reached via links on the Map Viewer home page.

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