NCBI News

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

Spring 2002

Make Your Own Gene Models with Model Maker

The Model Maker is a tool that allows the construction of an mRNA sequence model using putative exons defined by *ab initio* prediction and by aligning GenBank® transcripts (including ESTs) and NCBI RefSeqs to the NCBI human genome assembly. The ability to generate alternative transcript models using novel combinations of exons not represented in any single mRNA sequence alignment is useful in the exploration of gene model variants.

To generate a Model Maker display centered on a gene of interest, go to the Human Map Viewer page (Model Maker links will soon be available directly from LocusLink) and type the gene symbol or gene name into the query box. Select the Genes_seq map as the Master Map and click on the "mm" link to the right of the gene name. The Model Maker display for the HOXB1 gene is shown in Figure 1 on page 7. The gene is found on the RefSeq contig NT_010783.9, with the chromosome coordinates shown. You may move upstream or downstream from the contig using a pair of horizontal arrows. If you click on the sequence viewer link "sv" after this repositioning, you will be able to see the upstream or downstream sequences of the contig.

The Model Maker page is divided into several sections. The first section displays "evidence" for exons in the form of sequence records from the databases that have been aligned to human genome assembly contig NT_010783.9 to produce the NCBI gene model for HOXB1. These sequences include, from top to bottom, a GenBank mRNA sequence and the mRNA RefSeq derived from it, each contributing 2 exons; a GenomeScan-generated model, contributing an additional small exon; and a genomic RefSeq derived from the alignment of the mRNA RefSeq with the contig, contributing 2 exons.

The Model Maker gathers the unique putative exons from the various alignments used as evidence, assigns each a consecutive number, and presents them in a "graphic view". In the Figure, three unique exons are shown in the "graphic view" which may be added or removed from the nascent model by clicking them on or off. By choosing the "hits" hyperlink next to an evidence sequence, the coordinates of BLAST® hits for the evidence sequence on the human

continued on page 7

Virus Reference Sequences

The Virus Reference Genome project aims to provide molecular standards for viral genomic research. Viral RefSeqs start with one well-studied and/or best annotated full-length genomic sequence taken from GenBank. Each virus RefSeq genome is then curated by NCBI staff with the aid of outside advisors. The curation process results in the addition of relevant biological information, taken from the literature or other sequence records, as well as the correction of taxonomy and lineages. Viral RefSeqs are searchable in Entrez Genomes, which contains over 1,100 reference sequences

continued on page 3

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.

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

Editors
Dennis Benson
Barbara Rapp

Contributors
Medha Bhagwat
Eugenia Posey-Marcos

Writers
Vyvy Pham
David Wheeler

Editing and Production
Jennifer Carson Vyskocil

Graphic Design
Tim Cripps
Gary Mosteller

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. 02-3272

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

MapViewer Displays New Maps and Annotation Tracks

The Human Genome MapViewer has been enhanced with new displays and color-schemes that show additional data used by NCBI in annotating its latest build of the human genome. Some existing maps have also been revised. Several of the principal maps are described below.

The "Components" map shows the GenBank sequences that were used to construct NCBI's human genomic contigs, while the "GenBank" map shows alignments of other human genomic sequences from GenBank that were not used in the construction of the assembly. Gene Models predicted within the assembly using GenomeScan show up on the "GenomeScan" track. Dark brown models are those with BLAST hits to known vertebrate proteins; the remainder of the GenomeScan models are drawn in light brown. The Gene_Sequence map shows gene models that are color-coded by type of supporting evidence. Blue models are based on mRNA alignments to the genomic contigs with possible EST support, and are termed "Confirmed" models. Green models are based only on EST alignments. Dark brown lines indicate EST-supported GenomeScan models, while light brown lines indicate GenomeScan models lacking EST support. Orange-colored lines flag models in which a conflict exists between the model and an aligned mRNA.

When the Gene_Sequence map is made the master, two new links appear next to the familiar links to the SequenceViewer (sv), the EvidenceViewer (ev), and the Human/Mouse Homology Map (hm); these are the "seq" link and the "mm" link. The "seq" link is used to download the complete sequence for a gene model including a variable amount of 5' and 3' flanking sequence. The "mm" link invokes the Model Maker, a new tool for the construction of model transcripts. An article on the Model Maker appears on page 1 of this issue.

Transcript models produced by alignments of mRNA to the human genome are shown on the "Transcript" map while the "UniGene_Human" and "UniGene_mouse" maps show the alignment of mRNA and EST sequences from UniGene clusters to the human genome assembly. SAGE tag mappings are shown on a new "SAGE_tag" track.

Other Genomic MapViews

Several large eukaryotic genomes now have their own MapViewer displays. These include displays for the rat genome, which debuts with three linked maps: two genetic and one radiation-hybrid map. Also available are MapViewers for *Arabidopsis,* rice, yeast, zebra fish, fruit fly, and mouse. Access these from the Genomic Biology page, via the NCBI home page, or through Entrez Genomes.

Virus Reference Sequences

continued from page 1

for over 800 viral genomes. Access Viral Genomes using the link from the Related Resources section on the Entrez Genomes home page.

The Viral Genomes page includes a general introduction to viruses, a depiction of a scheme for the replication of influenza virus A and a variety of genome retrieval tools. Additionally, there are links to a complete alphabetical list of all reference genomes and to lists in which viruses are grouped by families/genera or genome type (e.g. dsDNA viruses, ssRNA positive strand viruses). In the list, each virus name shown in green italics is hyperlinked to a graphical representation of the genome — the link to the actual genome record is accessible from the genome accession number (the number that begins with the letters "NC"). The rightmost column "Nbrs" of the viruses' list leads to Genome Neighbors, which are related complete genomes found in GenBank. When available, links to the protein table for each virus genome are also displayed.

The Viral Genomes Finder, located on the main page and linked from the Tools section, allows the selection of all viruses, or viral genomes, that fall within a particular taxonomic node. The Query Tips offer help on how to formulate the search terms. For example, to retrieve all virus genomes within a particular taxonomic node, one must capitalize the first letter in the search word (i.e., "Flaviviridae"). To go to a Taxonomy Browser display of all viruses in a node, make sure the initial letter of the query is in lowercase.

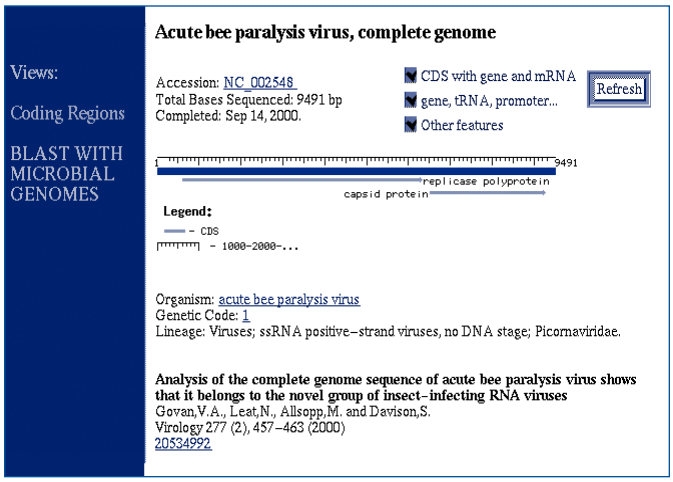


Figure 1: Entrez Genomes graphical view of the viral RefSeq for the acute bee paralysis virus.

Links from the blue sidebar menu include general overviews of viruses, statistics regarding virus reference genomes, FAQs, and help documentation on searching and viewing viral genome records. There are also links to external sites dedicated to virus biology, taxonomy and nomenclature, and sequences. The Help document presents samples of graphical formats and data layouts of the virus records in the database.

The sample record of Figure 1 shows the graphical view of a virus genome, which can have a linear or circular, monopartite or segmented architecture. The record gives a link to the GenBank flatfile record, the total number of nucleotides in the genome, and one or more literature citations. Also included are taxonomic information, additional comments about the reference sequence record, a reference to the source record from which it was derived, and the status of the record; either provisional or reviewed. Links in the blue sidebar lead to a table that lists the genes and protein products of the genome (Coding Regions), and to the Microbial Genomes BLAST program, which allows users to BLAST query sequences against the finished and unfinished microbial genomes database.

Virus genome reference sequences are added to the database daily and more analytical programs will be introduced in the near future. *—VP*

Release of the 1,000th Virus Reference Genome!

In April 2002, NCBI released the 1,000th virus reference genome into the public databases. The release of the 1,000th viral RefSeq is the fruit of a successful collaborative effort between the NCBI RefSeq team and numerous external experts who have expended much effort in annotating the virus reference genomes. These genomes, a part of NCBI's Reference Sequence collection, are available from Entrez Genomes.

What is the 1,000th viral genome RefSeq? Read about this interesting virus in the summer issue!

Mouse Genome BLAST: Scan NCBI Contigs and Three Draft Assemblies

BLASTing against the growing body of mouse genomic data is easier than ever using the Mouse Genome BLAST page, which supports both untranslated and translated searches of a variety of mouse genome databases, including NCBI contigs and three draft assemblies of the genome.

The default database for Mouse Genome BLAST is the database of curated NCBI contigs, called "NT_Contigs". These are assemblies of finished mouse BAC clones that are annotated with SNPs and STSs. Other databases include HTGS, Traces, BAC-Ends, Reference mRNAs and proteins from the NCBI RefSeq project, and ESTs.

The HTGS database is comprised of draft and finished sequence as submitted by the sequencing centers. The "Traces" database contains the mouse Whole Genome Shotgun (WGS) traces. The WGS Traces database may be searched with MegaBLAST, while the other databases may be queried with either nucleotide or protein sequences using the suite of blastn, blastp, tblastn, and blastx sequence-similarity search programs.

In addition, three assemblies of the mouse genome are available for searching. The first two, Phusion™ and Arachne, are preliminary assemblies of the mouse WGS reads based on a February data freeze and were created by the Sanger Center and the Whitehead Institute respectively. The third, the MGSCv3 assembly, produced by the Mouse Genome Sequencing Consortium, is comprised of WGS scaffolds generated using the end pairing information from the WGS reads. All three assemblies should be considered to be preliminary and will change as new data is added.

More Organism-Specific BLAST Pages

NCBI offers a number of specialized BLAST pages for model organisms that allow queries ranging from MegaBlast searches for nearly exact nucleotide matches to tblastn protein queries of translated nucleotide databases. The databases for each organism differ, depending upon the availability of sequence data, however, the spectrum of nucleotide databases includes Whole Genome Shotgun traces, HTG sequences, ESTs, genome assemblies, and mRNA sequences from annotated genes. Protein databases comprised of the protein translations of annotated genes are also available for many organisms. Links to organism-specific BLAST pages are found within the "Genomic BLAST Pages" section of the main BLAST page. In addition to human, the organisms covered include mouse, rat, and microbes, *Arabidopsis*, puffer fish, rice, mosquito, zebrafish, and other eukaryotes.

New Genomes in GenBank

Visit the Entrez Genomes Web page to see the latest batch of new genomes to enter GenBank. There are now over 70 complete bacterial and more than 15 complete archaeal genomes in GenBank. A couple of recent arrivals are:

Pyrococcus furiosus: An anerobic, hyperthermophilic member of the archaea that cannot live below 70 degrees Fahrenheit. *Pyrococcus furiosus* dwells near undersea thermal vents, where it uses peptides as a carbon source and sulfur as a final electron acceptor. The genome of *Pyrococcus furiosus* features 2,065 annotated genes and 105 annotated structural RNAs. The Entrez Genomes report indicates that 631 *Pyrococcus furiosus* proteins bear significant sequence similarity to proteins of known 3-D structure.

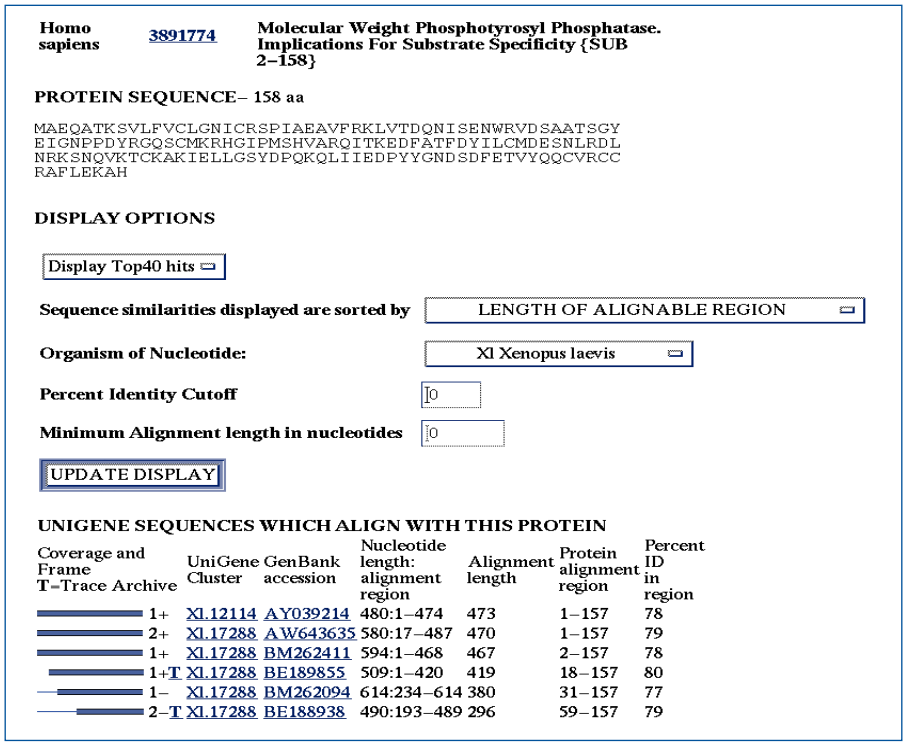
Schizosaccharomyces pombe: The popular fission-yeast so vital to molecular biologists and first isolated from an East African beer, called "pombe". This genome can be viewed in the NCBI MapViewer, which displays a Gene, Genetic, and Clone map for each of the three chromosomes and integrates a total of 5,095 genes, 169 genetic markers, and 582 clones. Other features annotated on this genome include promoters, introns, and repeat regions.

ProtEST: A Window on Protein Matches to ESTs

ProtEST is a new NCBI tool, analogous to BLASTLink, that presents a graphical view of pre-computed alignments between protein sequences and the translations of UniGene nucleotide sequences.

To generate the alignments, the 6-frame translations of mRNA and EST sequences in UniGene are compared to protein sequences using BLAST. The proteins compared for ProtEST are limited to those from eight model organisms: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Escherichia coli*. In order to exclude protein sequences that are derived from conceptual translations or models, the protein sequences from these model organisms are further limited to those derived from the structural databases, Swissprot, PIR, PDB, or PRF. ProtEST reports are updated in tandem with UniGene protein similarities.

ProtEST is accessible from any UniGene cluster page via links from the model organism protein similarities. For each nucleotide sequence match, the report shows the UniGene cluster ID, the GenBank accession number of the sequence, and the percent identity between the protein and nucleotide translation in the aligned region. When trace data is available, a link is also provided to the sequence trace in the NCBI Trace Archive. A schematic of the aligned region is linked to a BLAST2Sequences alignment display. Entries in the report can be sorted on the basis of percent identity, alignment length, alignment origin, and alignment end-point, UniGene cluster ID, and GenBank accession code. The entries shown can then be filtered by percent identity or by the length of the alignment.



Homo sapiens 3891774 **Molecular Weight Phosphotyrosyl Phosphatase. Implications For Substrate Specificity {SUB 2–158}**

PROTEIN SEQUENCE– 158 aa

MAEQATKSVLFVCLGNICRSPIAEAVFRKLVTDQNISENWRVDSAATSGY
EIGNPPDYRGQSCMKRHGIPMSHVARQITKEDFATFDYILCMDESNLRDL
NRKSNQVKTCKAKIELLGSYDPQKQLIIEDPYYGNDSDFETVYQQCVRCC
RAFLEKAH

DISPLAY OPTIONS

Display Top40 hits

Sequence similarities displayed are sorted by LENGTH OF ALIGNABLE REGION

Organism of Nucleotide: Xl Xenopus laevis

Percent Identity Cutoff 0

Minimum Alignment length in nucleotides 0

UPDATE DISPLAY

UNIGENE SEQUENCES WHICH ALIGN WITH THIS PROTEIN

Coverage and Frame T–Trace Archive	UniGene Cluster	GenBank accession	Nucleotide length: alignment region	Alignment length	Protein alignment region	Percent ID in region
1+	Xl.12114	AY039214	480:1–474	473	1–157	78
2+	Xl.17288	AW643635	580:17–487	470	1–157	79
1+	Xl.17288	BM262411	594:1–468	467	2–157	78
1+T	Xl.17288	BE189855	509:1–420	419	18–157	80
1–	Xl.17288	BM262094	614:234–614	380	31–157	77
2–T	Xl.17288	BE188938	490:193–489	296	59–157	79

Figure 1: ProtEST report for BLASTX alignments to human acid phosphatase I.

A typical ProtEST report is shown in Figure 1 above for human acid phosphatase I. This protein hits ESTs from *Bos taurus*, *Arabidopsis thaliana*, *Rattus norvegicus*, and *Xenopus laevis*, however, the display has been limited to ESTs from *Xenopus* and sorted by alignment length. Two ESTs from *Xenopus* are associated with sequencing traces available in the Trace Archive as indicated by the "T" link next to the accession number.

Trace Archive Expands

The NCBI Trace Archive contains traces for EST, WGS, shotgun, and finishing projects from a variety of organisms, including traces from human, mouse, rat, zebrafish, worm, mosquito, frog, soybean, rice, and others. The most current listing can be found by following the "Trace Archive" link from the NCBI home page to the Trace Archive page. The trace data may be queried using a text-based search interface on the Trace Archive page, and an FTP link from this page allows for bulk downloads. The trace data may also be searched using MegaBLAST via the link provided.

Searching Finished and Unfinished Microbial Genomes

The genomes of microbes contain gene sequences required for life under a variety of conditions such as those of harsh temperature and pH, as well as those privileged conditions experienced by intracellular parasites. A comparison of the gene complements needed by organisms living under different environmental constraints is important in elucidating the mechanisms of pathogenesis and in defining the genetic diversity of life on earth.

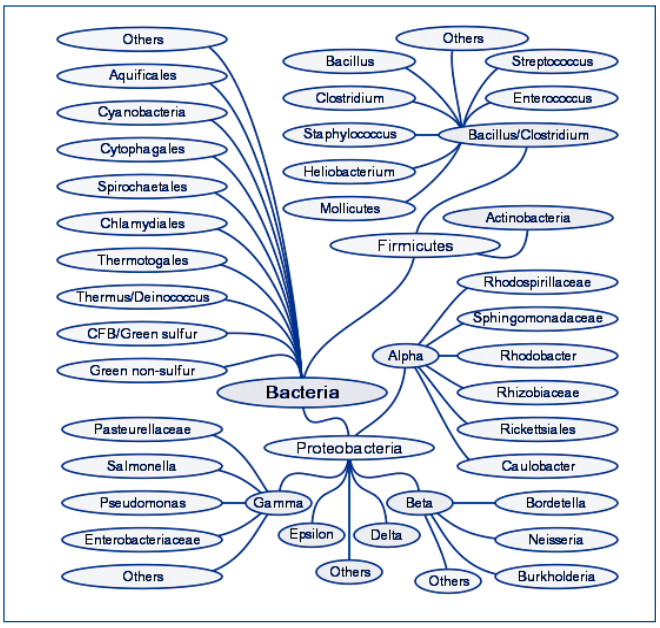


Figure 1: Genomes Tree display of the bacterial genomes offered for searching on the Microbial Genomes BLAST page. Clicking on a node shows the organisms included.

There are now over 80 complete bacterial and archaeal genomes in GenBank and about an equal number of unfinished genomes for which sequencing is in progress. The unfinished genomes have not been deposited in GenBank and are therefore not available for downloading or for conventional BLAST searching. However, because of the importance of this set of genome sequences, NCBI offers the Microbial Genomes BLAST page for similarity searches of both finished sequence now in GenBank, and unfinished microbial genomic sequence provided by sequencing centers prior to publication.

As an example, consider the protein NP_248745, a hypothetical protein from *Pseudomonas aeruginosa* that is conserved among three organisms: *Mesorhizobium loti*, *Caulobacter crescentus*, and *Pseudomonas aeruginosa*. The hierarchal taxonomic tree of Figure 1, generated by clicking on the link at the top of the Microbial Genomes BLAST page, places the first two organisms in the alpha proteobacterial lineage and the third in the gamma lineage. It might be of interest to determine if a similar protein is also found in the beta proteobacteria. This search is easily performed by using NP_248745 as a tblastn query for a search against all of the beta proteobacterial genomic nucleotide sequences. The beta proteobacteria are selected as the BLAST database by clicking on the appropriate node of this tree.

The one-line BLAST descriptions returned by this search, shown in Figure 2, indicate a number of good hits to sequences from the beta proteobacteria, as indicated by the low Expect values given. Hence, this conserved protein from the gamma and alpha proteobacteria has a potential homolog in some of the beta proteobacteria. Note that 4 of the 5 hits are to incomplete genomes, not in GenBank, and searchable at NCBI only through this interface. The Microbial Genomes BLAST page is linked from the main NCBI BLAST page.

Sequences producing significant alignments:	(bits)	Value
gnl|DOE_134537|Contig394 Burkholderia fungorum unfinished f...	145	3e-35
gnl|TIGR_13373|24 Burkholderia mallei unfinished fragment o...	136	1e-32
gnl|Sanger_28450|bpsmalle_Contig784 Burkholderia pseudomall...	136	1e-32
gnl|DOE_119219|Contig600 Ralstonia metallidurans unfinished...	135	2e-32
ref|NC_003296.1| Ralstonia solanacearum, complete genome	121	5e-28

Figure 2: Highest-scoring BLAST hits resulting from a search using NP_248745 as a tblastn query against all of the beta proteobacterial genomic nucleotide sequences.

genomic sequence are shown and intron lengths may be deduced. It is also possible to select whole "exon sets" from evidence mRNAs, using the "set" link, add EST alignments to the evidence list, extend the region shown downstream and upstream or switch to the opposite strand.

The information in the "graphic view" is also presented as a "table view" that gives the start and stop positions of each putative exon, along with the first three and last three bases of each exon and two bases immediately upstream and downstream of the exon. Exons are selected for inclusion in the model using check boxes. Numbered links at the ends of each exon facilitate the selection of adjacent exons implied by existing transcripts. As each exon is selected, the 3- frame translations of the model sequence are updated in the ORF Frames boxes. The longest ORF in each translation is shown in UPPER CASE letters. As compatible exons are added to the model, in phase, an ORF in at least one of the three reading-frames lengthens. When the stop codon is reached and the model is complete, it may be saved to a local file in FASTA format for use with other programs. *—EP, MB*

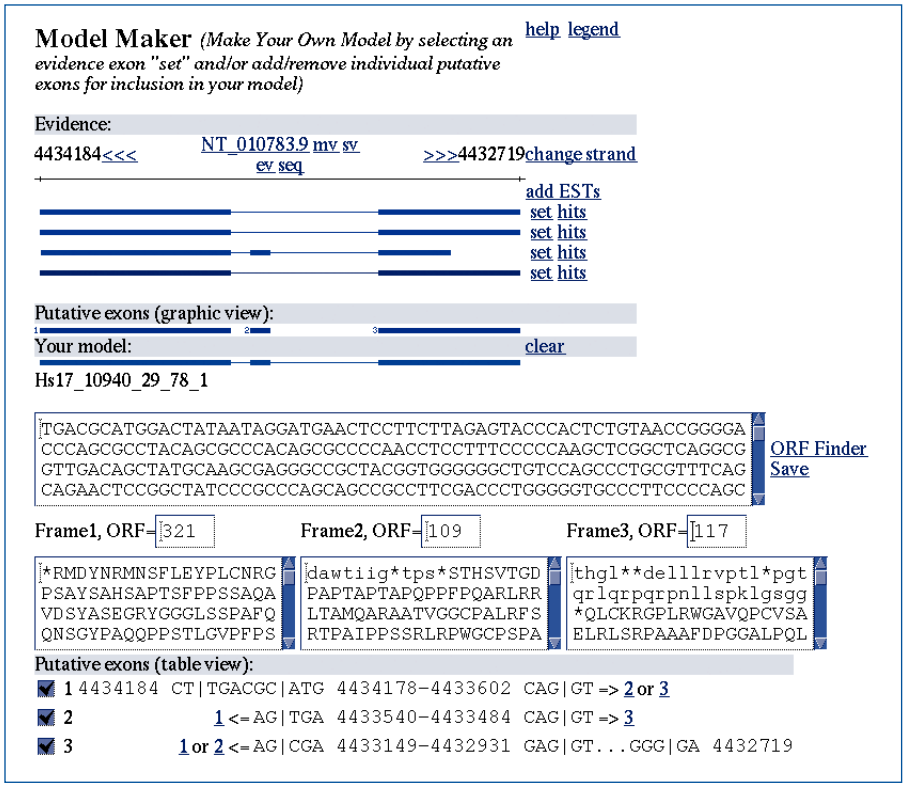


Figure 1: Model Maker display for HOX1B located on contig NT_010783.9. Access the Contig record by clicking the Accession Number hyperlink NT_010783.9. Move upstream or downstream by clicking on the arrows: <<< (upstream) or >>> (downstream). View the opposite strand by clicking on "change strand" hyperlink. Select "set" to select an evidence record as your model. Select "add ESTs" to add a graphical display of ESTs. To build a custom model, select the exon or EST of interest by clicking on the graphic view segment or select the exon from the table by using a check box. Monitor the growing model by viewing its translation in three different reading frames. The longest ORF in each frame is denoted by UPPERCASE letters. Save your results by clicking on the "Save" hyperlink.

Find Out "About NCBI"

Need an answer to the question "Who and what is NCBI?" Visit the new "About NCBI" area from the link on our home page! Here you can read and learn about NCBI databases and data mining tools, obtain an overview of NCBI's human genome resources, and access information on the model organisms for which genomic data is available. You will also find introductory material on the basic science underlying the acquisition and analysis of NCBI's molecular biology resources. To round out your tour, check out our online tutorials, outreach and education programs, and browse the *NCBI News* — the free quarterly newsletter you are now reading —available by subscription or via the Web.

New FTP Hierarchy Simplifies Data Downloads

The NCBI FTP Web site has been reorganized to provide more direct access to a number of popular datasets. To see the new site structure, click on the FTP link on the NCBI home page. Two of the many new direct links to FTP directories are: "Genome Assembly/Annotation Projects", for access to complete genomes/chromosomes, contigs and reference sequences; and "RefSeq" for the full release or daily updates of the RefSeq database. Access NCBI's FTP site by anonymous FTP at:

ftp.ncbi.nih.gov

NCBI and the *NCBI News* Extend a Fond Farewell to Dr. Barbara Rapp

After a decade of newsletters, *NCBI News* Editor, Dr. Barbara Rapp will leave NCBI to accept the position of Associate Fellows Coordinator at the National Library of Medicine (NLM).

For the past 12 years, Barbara has directed the NCBI User Services group, which provides phone and e-mail support for NCBI's databases and database-related tools. The NCBI User Services group also implements outreach and training programs for the scientific community and the general public, including workshops, presentations at scientific conferences, and the publication of the *NCBI News*. Barbara has been an indispensable asset to the overall growth in public awareness and understanding of NCBI's databases and programs, as well as a leader in various initiatives for enhancing education in bioinformatics and information science.

Prior to joining NCBI, Barbara received a doctorate in 1985 for research on document clustering based on co-occurrence of index

terms versus cited references in scientific databases. She then served on the faculty of the School of Library and Information Science at the Catholic University of America, where she managed the program in health sciences librarianship. In 1988, she returned to NLM to perform evaluation research in the Office of Planning and Evaluation before joining NCBI in 1990.

We wish Barbara the fondest farewell and best wishes in her new endeavors! —*VP*

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

Official Business
Penalty for Private Use $300

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