Nucleic Acids Research, 2005, Vol. 33, Database issue ***D39–D45***
doi:10.1093/nar/gki062

Database resources of the National Center for Biotechnology Information

David L. Wheeler*, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Wolfgang Helmberg, David L. Kenton, Oleg Khovayko, David J. Lipman, Thomas L. Madden, Donna R. Maglott, James Ostell, Joan U. Pontius, Kim D. Pruitt, Gregory D. Schuler, Lynn M. Schriml, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Grigory Starchenko, Tugba O. Suzek, Roman Tatusov, Tatiana A. Tatusova, Lukas Wagner and Eugene Yaschenko

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 15, 2004; Revised and Accepted October 5, 2004

ABSTRACT

In addition to maintaining the GenBank(R) nucleic acid sequence database, the National Center for Biotechnology Information (NCBI) provides data retrieval systems and computational resources for the analysis of data in GenBank and other biological data made available through NCBI's website. NCBI resources include Entrez, Entrez Programming Utilities, PubMed, PubMed Central, Entrez Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link (BLink), Electronic PCR, OrfFinder, Spidey, RefSeq, UniGene, HomoloGene, ProtEST, dbMHC, dbSNP, Cancer Chromosomes, Entrez Genomes and related tools, the Map Viewer, Model Maker, Evidence Viewer, Clusters of Orthologous Groups (COGs), Retroviral Genotyping Tools, HIV-1/Human Protein Interaction Database, SAGEmap, Gene Expression Omnibus (GEO), Online Mendelian Inheritance in Man (OMIM), the Molecular Modeling Database (MMDB), the Conserved Domain Database (CDD) and the Conserved Domain Architecture Retrieval Tool (CDART). Augmenting many of the Web applications are custom implementations of the BLAST program optimized to search specialized datasets. All of the resources can be accessed through the NCBI home page at http://www.ncbi.nlm.nih.gov.

INTRODUCTION

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health was created in 1988 to develop information systems for molecular biology. In addition to maintaining the GenBank(R) (1) nucleic acid sequence database, to which data are submitted by the scientific community, NCBI provides data retrieval systems and computational resources for the analysis of GenBank data and a variety of other biological data. For the purposes of this study, the NCBI suite of database resources is grouped into the six categories given below. All resources discussed are available from the NCBI home page at http://www.ncbi.nlm.nih.gov. In most cases, the data underlying these resources are available for bulk download at ftp.ncbi.nih.gov, a link from the NCBI home page.

DATABASE RETRIEVAL TOOLS

Entrez

Entrez (2) is an integrated database retrieval system that enables text searching, using simple Boolean queries, of a diverse set of over 20 databases, several added during the past year. A newly implemented Global Query, the default search on the NCBI homepage, now allows simultaneous searches across all the Entrez databases at speeds comparable to a single database search. On retrieving the counts of database matches, the user may then display and further refine searches in any individual database. The Entrez databases include DNA and protein sequences derived from several sources (1,3–6), the NCBI taxonomy, genomes, population sets, gene expression data, gene-oriented sequence clusters in UniGene, sequence-tagged sites in UniSTS, genetic variations in dbSNP, protein structures from the Molecular Modeling Database (MMDB) (7), three-dimensional (3D) and alignment-based protein domains, and the biomedical

*To whom correspondence should be addressed. Tel: +1 301 435 5950; Fax: +1 301 480 9241; Email: wheeler@ncbi.nlm.nih.gov

literature via PubMed, PubmedCentral, Online Mendelian Inheritance in Man (OMIM) and online books. PubMed includes primarily, the 12.8 million references and abstracts in MEDLINE(R), with links to the full text of more than 4400 journals available on the Web. The Books database contains more than 35 online scientific textbooks, including the NCBI Handbook, a comprehensive guide to NCBI resources. The NCBI website itself is among the Entrez databases, allowing users to employ the Entrez search engine to quickly find NCBI Web pages of interest.

Entrez provides extensive links within and between databases to related information ranging from simple cross-references between a sequence and the abstract of the paper in which it was reported, or between a protein sequence and its corresponding DNA sequence or 3D-structure, to alignment with other sequences. Recently added are links between a genomic assembly and its components and between a master sequence and those sequences derived from its annotation. Other links based on computed similarities among sequences or PubMed abstracts, called 'neighbors', allow rapid access to groups of related records. A service called LinkOut expands the range of links from individual database records to related outside services, such as organism-specific genome databases. To accommodate the growing number of Entrez links from one record to another, a 'Links' pull down menu appears in the top, right-hand corner of Entrez displays.

The records retrieved by an Entrez search can be displayed in a wide variety of formats and downloaded individually or in batches. A redirection control allows results to be sent to a local file, formatted in the browser as plain text or sent to the clipboard. PubMed results may also be emailed directly from Entrez. Formatting options vary for records of different types. Display formats for GenBank records include the GenBank Flatfile, FASTA, XML, ASN.1 and others. Graphical display formats are offered for some types of records, including genomic records. A formatting control allows the display or download of a particular range of residues for either a nucleotide or protein record.

Access to Entrez via automated systems is facilitated using the Entrez Programming Utilities (E-Utilities), a suite of seven server-side programs which support a uniform set of parameters used to search, link between and download from the Entrez databases. A search history, available via interactive Entrez as well as via the E-Utilities, allows users to recall the results of previous searches during an Entrez session and combine them using Boolean logic. Recent additions to the E-Utilities suite include the 'einfo' utility to retrieve indexed term counts, date of last update and a list of links for an Entrez database, and 'egquery' which returns the number of matches to a query in each Entrez database; an automated system may then use E-Utilities such as 'efetch' or 'esummary', to retrieve the data. A 'Simple Object Access Protocol' (SOAP) interface to the E-Utilities has also recently been made available. Instructions for using the E-Utilities are found under the 'Entrez tools' link on the NCBI home page.

PubMed Central

PubMed Central (PMC) (8) is a digital archive of peer reviewed journals in the life sciences providing access to over 300 000 full text articles. Over 160 journals, including *Nucleic Acids Research*, deposit the full text of their articles in PMC. Participation in PMC requires a commitment to free access to full text, perhaps with some delay after publication. Some journals provide free access to their full text directly in PMC while others require a link to the journal's own site where full text is generally available free within six months to a year of publication. All PMC free articles are identified in PubMed search results and PMC itself can be searched using Entrez.

Taxonomy

The NCBI taxonomy database indexes over 165 000 named organisms that are represented in the databases with at least one nucleotide or protein sequence. The Taxonomy Browser can be used to view the taxonomic position or retrieve data from any of the principal Entrez databases for a particular organism or group. The Taxonomy Browser also displays links to the Map Viewer, Genomic BLAST services, the Trace Archive, and to model organism and taxonomic databases via LinkOut.

Searches of the NCBI taxonomy may be made on the basis of whole, partial or phonetically spelled organism names, but links to organisms commonly used in biological research are provided. The Entrez Taxonomy system adds the ability to display custom taxonomic trees representing user-defined subsets of the full NCBI taxonomy.

Entrez Gene

Entrez Gene (6), the successor to LocusLink, provides an interface to curated sequences and descriptive information about genes with links to NCBI's Map Viewer, Evidence Viewer, Model Maker, BLAST Link, protein domains from NCBI's Conserved Domain Database and other gene-related resources. Data are accumulated and maintained through several international collaborations in addition to curation by in-house staff. Links within Gene to the newest citations in PubMed are maintained by curators using Gene References into Function (GeneRIF). GeneRIF, accessible via links in Gene reports, also allows researchers using Gene to add references to a report.

THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The Basic Local Alignment Search Tool (BLAST) programs (9–11) perform sequence-similarity searches against a variety of sequence databases, returning a set of gapped alignments between the query and database sequences, and links to full database records, to UniGene, Gene, the MMDB or GEO. Sequences appearing in a BLAST alignment may be selected for bulk download. A BLAST variant, BLAST2Sequences (12), compares two DNA or protein sequences and produces a dot-plot representation of the alignments.

Each alignment returned by a BLAST search receives a score and a measure of statistical significance, called the Expectation Value (*E*-value), for judging its quality. Either an *E*-value threshold or a range can be specified to limit the alignments returned. BLAST takes into account the amino acid composition of the query sequence in its estimation of statistical significance. This composition-based statistical

treatment, used in conventional protein BLAST searches as well as PSI-BLAST (11) searches, tends to reduce the number of false-positive database hits (13).

BLAST offers several output formats including the default 'pairwise' alignment, several 'query-anchored' multiple sequence alignment formats and a tabular 'Hit Table'; an easily parsed summary of the BLAST results. Users selecting the 'new formatter' option can also view alignments in a '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 lowercase and with different colors rather than simply replacing each with an 'X' or an 'N'. In addition, BLAST can generate a taxonomically organized output that shows the distribution of BLAST hits by organism. A new 'sequence retrieval' formatting option allows database sequences to be marked for batch retrieval using check boxes appearing in the BLAST results.

The web BLAST interface allows both the initial search and the results displayed to be restricted to a database subset using the Entrez search syntax. Web BLAST uses a standard URL–API that allows complete search specifications, including BLAST parameters, such as Entrez restrictions and the search query, to be contained in a URL posted to the web page.

A BLAST variant designed to search for nearly exact matches, called MegaBLAST (14), offers a web interface that handles batch nucleotide queries and operates up to 10 times faster than standard nucleotide BLAST. MegaBLAST is the default search program for NCBI's Genomic BLAST pages that search a set of genome-specific databases and generate, where possible, genomic views of the BLAST hits using the Map Viewer. MegaBLAST is also used to search the rapidly growing Trace Archive and is available for the standard BLAST databases as well. For rapid cross-species nucleotide queries of the Trace Archive as well as the standard BLAST databases, NCBI offers Discontiguous MegaBLAST, which uses a non-contiguous word match (15) as the nucleus for its alignments. Discontiguous MegaBLAST is far more rapid than a translated search such as BLASTX, yet maintains a competitive degree of sensitivity when comparing coding regions.

Several recent additions have been made to the suite of standard BLAST databases. Environmental sample data can now be searched within the 'env_nt' or 'env_nr' databases for nucleotide and protein sequences, respectively. A 'RefSeq' database is 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.

BLink

BLAST Link (BLink) displays pre-computed protein BLAST alignments for each protein sequence in the Entrez databases. BLink can display subsets of these alignments by taxonomic criteria, by database of origin, relation to a complete genome, membership in a COG (16) or by relation to a 3D structure or conserved protein domain. BLink links are displayed for protein records in Entrez as well as within Entrez Gene reports.

RESOURCES FOR GENE-LEVEL SEQUENCES

UniGene

UniGene (17), a system for automatically partitioning GenBank sequences, including expressed sequence tags (ESTs), into a non-redundant set of gene-oriented clusters. UniGene clusters are created for all organisms for which there are 70 000 or more ESTs in GenBank and now includes ESTs for more than 25 animals and over 20 plants. Each UniGene cluster contains sequences that represent a unique gene, and is linked to related information, such as the tissue types in which the gene is expressed, model organism protein similarities, the Entrez Gene report for the gene and its map location. In the human UniGene July 2004 release (build 173), over 4.5 million human ESTs in GenBank have been reduced 42-fold in number to approximately 107 000 sequence clusters. The UniGene collection has been used as a source of unique sequences for the fabrication of microarrays for the large-scale study of gene expression (18). UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences.

ProtEST

ProtEST, a tool analogous to BLASTLink, presents pre-computed BLAST alignments between protein sequences from model organisms and the six-frame translations of UniGene nucleotide sequences. Protein sequences that are derived from conceptual translations or model transcripts are excluded. ProtEST links are displayed in UniGene reports with model organism protein similarities. ProtEST reports are updated in tandem with UniGene protein similarities.

The Trace and Assembly Archives

The Trace Archive home page allows for flexible searching and download of sequencing traces from a rapidly growing database of over 500 million sequencing traces from more than 400 organisms. The 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.

HomoloGene

HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes. The genomes represented in the recent Build 37 of HomoloGene include *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Anopheles gambiae*, *Caenorhabditis elegans*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Neurospora crassa*, *Magnaporthe grisea*, *Arabidopsis thaliana* and *Plasmodium falciparum*.

NCBI has adopted a new HomoloGene build procedure which is guided by the taxonomic tree, and relies on conserved gene order and measures of DNA similarity among closely related species, while making use of protein similarity for more distantly related organisms. The new computational procedure greatly increases the reliability of the computed

homologous gene sets and the resulting HomoloGene entries now include paralogs in addition to orthologs. HomoloGene can be queried using Entrez (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene)

Among the Entrez fields unique to HomoloGene is the 'Ancestor' field, which refers to the taxonomic group of the last common ancestor of the species represented in a HomoloGene entry. Using the 'Ancestor' field it is possible to limit a search to genes conserved in one of 22 ancestral groups. HomoloGene reports include homology and phenotype information drawn from Online Mendelian Inheritance in Man (OMIM), Mouse Genome Informatics (MGI), Zebrafish Information Network (ZFIN), *Saccharomyces* Genome Database (SGD), Clusters of Orthologous Groups (COG) and FlyBase. A 'Pairwise Scores' display gives a table of pairwise statistics for members of a Homologene group that includes percent amino acid and nucleotide identities, the Jukes–Cantor genetic distance parameter, D, the ratio of non-synonymous to synonymous amino acid substitutions (K_a/K_s) for predicted proteins and the ratio of nucleotide identities within non-coding regions of the transcript to those within coding regions (K_{nr}/K_{nc}).

dbMHC

dbMHC supports clinical applications and research related to the major histocompatibility complex (MHC) and includes Reagent Database and Clinical sections. The Reagent database provides an open platform for the submission, evaluation and editing of individual DNA typing reagents as well as typing kit information. All reagents are characterized for allele specificity using an updated allele database based on IMGT/HLA. The dbMHC offers several resources for the analysis and display of the MHC and KIR region, e.g. an interactive formatting sequence retrieval tool, and a Sequencing-based typing tool, capable of aligning and interpreting heterozygote sequences. Also featured is dbMHCms, a tool to search descriptive information for known short tandem repeats within the MHC.

The Clinical section contains data generated by the 13th international HLA workshop and international HLA working group and includes sections presenting two major IHWG datasets. The first is derived from the IHWG 'Diversity/Anthropology' project to determine global HLA allele frequencies in an attempt to shed light on the evolution of HLA polymorphisms. dbMHC can display project data, such as allelic frequencies found in individuals from certain regions of the world, or frequencies for specific loci.

The second IHWG dataset is the Hematopoietic Cell Transplantation (HTC) database, containing anonymous data for selected unrelated donor transplants performed worldwide for the treatment of both malignant and non-malignant blood disorders. Online analysis tools available for the HCT data include a query interface and the ability to compute Kaplan–Meier survival plots.

Reference Sequences

The Reference Sequence (RefSeq) database (19), which provides curated references for transcripts, proteins and genomic regions, plus computationally derived nucleotide sequences and proteins. The complete RefSeq database is now being provided in the RefSeq directory on the NCBI FTP site. As of Release 6, RefSeq contained over 1.3 million sequences, including more than 1 million protein sequences, representing more than 2400 organisms. To register for the 'refseq-announce' mailing list and be informed of new releases or to read more about the RefSeq project, visit the RefSeq home page.

Open reading frame (ORF) Finder and Spidey

ORF Finder performs a six-frame translation of a nucleotide sequence and returns the location of each ORF within a specified size range. Translations of the ORFs detected can be submitted directly for similarity searching against the standard BLAST or COGs databases.

Spidey is an alignment tool for eukaryotic genomic sequences that takes as an input a set of mRNA accessions or FASTA sequences and aligns each to a single genomic sequence. Spidey takes into account predicted splice sites in constructing its alignments and can use one of four splice-site models (Vertebrate, *Drosophila*, *C.elegans* and Plant). Spidey returns exon alignments, protein translations and a summary showing the alignment quality and goodness of match to splice junction patterns for each putative exon. ORF Finder and Spidey are available via the 'Tools' link on the NCBI home page.

Electronic PCR (e-PCR)

Two types of e-PCR can now be performed from the e-PCR home page (www.ncbi.nlm.nih.gov/sutils/e-pcr). Forward e-PCR searches for matches to STS primer pairs in the UniSTS database of over 450 000 markers. Reverse e-PCR is used to estimate the genomic binding site, amplicon size and specificity for sets of primer pairs by searching against the genomic and transcript databases of *A.gambiae*, *A.thaliana*, *C.elegans*, *D.melanogaster*, *H.sapiens*, *M.musculus* and *R.norvegicus*.

To increase sensitivity, Forward e-PCR allows the size of the primer segment to be matched, number of mismatches, number of gaps and the size of the STS to be adjusted. Windows, Linux and Unix e-PCR binaries, along with the source code, are available via FTP (ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR)

A database of single nucleotide polymorphisms (dbSNP)

The dbSNP (20) is a repository for single base nucleotide substitutions and short deletion and insertion polymorphisms that contains 9.8 million human SNPs as well as about 5 million from a variety of other organisms. Now an Entrez database, dbSNP can be queried from the NCBI home page. Searches for SNPs lying between two markers and batch downloads via Entrez are supported. SNP reports link to 3D visualizations of structures from the MMDB, via NCBI's interactive macromolecular viewer Cn3D (21), that highlight amino acid changes implied by SNPs in coding regions. dbSNP provides additional information about the validation status, population-specific allele frequencies and individual genotypes for dbSNP submission. These data are available on the dbSNP FTP site in XML-structured genotype reports that include information about cell lines, pedigree IDs and error flags for genotype inconsistencies and incompatibilities. Haplotype and linkage disequilibrium data are being incorporated in dbSNP as data are released from the International HapMap project. Functional variants are identified when

dbSNP submissions can be matched to OMIM records and mutation reports in the biomedical literature.

RESOURCES FOR GENOME-SCALE ANALYSIS

Entrez Genomes

Entrez Genomes (22) provide access to genomic data contributed by the scientific community for species whose sequencing and mapping is complete or in progress. Entrez Genomes now includes over 180 complete microbial genomes, more than 1600 viral genomes, and over 550 reference sequences for eukaryotic organelles. Higher eukaryotic genomes are also included within Entrez Genomes such as the recent arrival, *Apis mellifera*. The Plant Genomes Central web page serves as a focal point for access to completed plant genomes, to information on plant genome sequencing projects, or to plant-related resources at the NCBI such as plant Genomic BLAST pages or Map Viewer. Similar resources, including specialized viewers and BLAST pages, are also available for eukaryotic organelles and viruses. In Entrez Genomes, complete genomes can be accessed hierarchically starting from either an alphabetical listing or a phylogenetic tree for each of six principal taxonomic groups. One can follow the hierarchy to a graphical overview for the genome of a single organism, onto the level of a single chromosome and, finally, down to the level of a single gene. At each level are one or more views, pre-computed summaries and links to analyses. At the level of a genome or a chromosome, a Coding Regions view displays the location of each coding region, length of the product, GenBank identification number for the protein sequence and name of the protein product. An RNA Genes view lists the location and gene names for ribosomal and transfer RNA genes. At the level of a single gene, links are provided to pre-computed sequence neighbors for the implied protein with links to the COGs database if possible. A summary of COG functional groups is presented in both tabular and graphical formats at the genome level.

For complete microbial genomes, pre-computed BLAST neighbors for protein sequences, including their taxonomic distribution and links to 3D structures, are given in TaxTables and PDBTables, respectively. Pairwise sequence alignments are presented graphically and linked to the Cn3D macromolecular viewer (21), which provides interactive display of 3D structures and sequence alignments. The TaxPlot tool plots similarities in the proteomes of two organisms to that of a third, reference organism, and is available for both prokaryotic and eukaryotic genomes. A new GenePlot tool, available from within Entrez Genome reports for microbial genomes, allows genome-wide comparisons of protein homologies to be visualized in a configurable graph. Using GenePlot, genomic inversions, deletions and insertions between bacterial strains and closely related species are easily highlighted. Resources for the genomes of higher eukaryotes are discussed below.

Clusters of Orthologous Groups

The rapid progress in sequencing has produced sequences for over 180 prokaryotic genomes comprising 155 species included in 95 different taxonomic genera. This avalanche of genomic sequence presents a challenge to researchers attempting to identify orthologous genes and to visualize protein clusters. The COGs database (16) presents a compilation of orthologous groups of proteins from 66 completely sequenced organisms. A eukaryotic version, KOGs, is available for seven eukaryotes, including *H.sapiens*, *C.elegans*, *D.melanogaster* and *A.thaliana*. Alignments of sequence from COGS have been incorporated into the Conserved Domain Database described below.

Retroviral Genotyping Tools

NCBI offers a web-based genotyping tool that employs a BLASTN comparison between a retroviral sequence to be subtyped and either a default panel of reference sequences or a panel provided by the user. A HIV-1-specific subtyping tool uses a set of reference sequences taken from the principle HIV-1 variants.

Map Viewer

The NCBI Map Viewer displays genome assemblies, genetic and physical markers, and the results of annotation and other analyses using sets of aligned maps. The Map Viewer home page organizes the available organisms by taxonomic group and provides links to both Map Viewer and Genomic BLAST pages. Map Viewer displays are available for the genomes of 29 organisms including *H.sapiens*, *M.musculus* and *R.norvegicus*. The genomic maps displayed by the Map Viewer vary according to the data available for the subject organism and are selected from a set of cytogenetic maps, physical maps, maps showing predicted gene models, EST alignments with links to UniGene clusters and mRNA alignments used to construct gene models. Maps from multiple organisms or multiple assemblies from the same organism can now be displayed in the same view. Map Viewer displays links to related resources such as Entrez Gene, or tools such as the Evidence Viewer and Model Maker. The Map Viewer can generate a tabular view of the current display that is convenient for export to other programs. Segments of a genomic assembly may be downloaded using the Map Viewer's 'Download/View Sequence' link in either GenBank or FASTA formats.

Queries can be made in the Map Viewer using gene names or symbols, marker names, SNP identifiers, accession numbers and other identifiers. The plant genomes in the Map Viewer can be searched together as a group using a special cross-species query page to generate a Map Viewer display composed of the chromosome maps from the different species on which the query was matched. A 'Map Viewer' Link in the Entrez 'Links' menu for nucleotide or protein sequences shown in the Map Viewer provides a convenient route to a Map Viewer display for a region of interest.

Model Maker

The Model Maker (MM) is used to construct transcript models using combinations of putative exons derived from *ab initio* predictions or from the alignment of GenBank transcripts, including ESTs and NCBI RefSeqs, to the NCBI human genome assembly. The MM displays an overview of transcript alignments to a genomic contig collecting each unique block of alignments as a putative exon. Transcript models are constructed by selecting from this collection. As a transcript is created, the implied protein translation is given in each reading

frame with any internal stop codons indicated. Previously observed exon splice patterns are indicated as guides to model building. Completed models may be saved locally or analyzed with ORF Finder.

Evidence Viewer

The Evidence Viewer (EV) displays the alignments to a genomic contig of RefSeq transcripts, GenBank mRNAs, known or potential transcripts, and ESTs supporting a gene model. The EV uses a graphical summary of the alignments to indicate the coordinate range of the gene model on the genomic contig, the areas of alignment to the transcripts and EST alignment density along the contig. Areas of disagreement between transcript sequences and the genomic sequence are highlighted. Exon-by-exon alignments of all the transcript sequences against the genomic contig, including flanking genomic sequence for each exon, are given along with protein translations. Any protein annotated on the transcript sequences are shown and mismatches between transcripts and the genomic contig or between proteins annotated on the aligned transcripts are highlighted.

Cancer Chromosomes

Three databases, the NCI/NCBI SKY (Spectral Karyotyping)/M-FISH (Multiplex-FISH) and CGH (Comparative Genomic Hybridization) Database, the NCI Mitelman Database of Chromosome Aberrations in Cancer (23) and the NCI Recurrent Chromosome Aberrations in Cancer databases comprise the new 'Cancer Chromosomes' Entrez database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Cancer Chromosomes).

Three search formats are available: a conventional Entrez query, a Quick/Simple Search and an Advanced Search. The Simple Search offers a set of menus to select a disease site or diagnosis that can be combined with specifications for a particular chromosomal location and anomaly. The Advanced Search offers a combination of forms for more complex queries. Search results may list all cases matching the query terms, a 'case-based report', or list each clone or cell separately, the 'clone/cell report'. Similarity reports show terms common to a group or records within several term categories, such as diagnosis or disease site and cytogenetic abnormalities, among the selected cases or 'clones/cells'.

RESOURCES FOR THE ANALYSIS OF PATTERNS OF GENE EXPRESSION AND PHENOTYPES

SAGEmap

NCBI's SAGEmap (24) provides a two-way mapping between regular (10 base) and LongSAGE (17 base) SAGE tags and UniGene clusters. The SAGEmap repository presently contains 381 SAGE experiments from 11 organisms. SAGEmap can also construct a user-configurable table of data comparing one group of SAGE libraries with another. SAGEmap is updated weekly, immediately following the update of UniGene the data appears in the human, mouse and rat genome Map Viewer as the SAGE track.

Gene Expression Omnibus (GEO)

GEO (25) is a data repository and retrieval system for any high-throughput gene expression or molecular abundance data. GEO contains microarray-based experiments measuring the abundance of mRNA, genomic DNA and protein molecules, as well as non-array-based technologies such as SAGE and mass spectrometry peptide profiling. The GEO repository accepts data via Web or in batch. The repository can be browsed from the GEO home page, and may be queried from both experiment- (Entrez GEO DataSets) and gene-centric (Entrez GEO Profiles) perspectives. At the time of writing, the repository contains high-throughput gene expression data from about 30 000 hybridization experiments, has about 1000 array definitions, and approximately half a billion individual spot measurement data derived from over 100 organisms.

OMIM

NCBI provides the online version of the OMIM catalog of human genes and genetic disorders authored and edited by Victor A. McKusick at The Johns Hopkins University (26). The database contains information on disease phenotypes and genes, including extensive descriptions, gene names, inheritance patterns, map locations, gene polymorphisms and detailed bibliographies. The OMIM Entrez database contains about 16 000 entries, including data on over 10 000 established gene loci and phenotypic descriptions. These records anchor links to many important resources, such as locus-specific databases and GeneTests.

THE MOLECULAR MODELING DATABASE, THE CONSERVED DOMAIN DATABASE SEARCH, CDART AND PROTEIN INTERACTIONS

The NCBI Molecular Modeling Database (MMDB), built by processing entries from the Protein Data Bank (5), is described previously (7). The structures in the MMDB are linked to sequences in Entrez and to the Conserved Domain Database (CDD). The CDD contains over 10 000 PSI–BLAST-derived position-specific score matrices representing domains taken from the Simple Modular Architecture Research Tool (Smart) (27), Pfam (28) and from domain alignments derived from COGs. NCBI's Conserved Domain Search (CD-Search) service can be used to search a protein sequence for conserved domains in the CDD. Wherever possible, CDD hits are linked to structures which, coupled with a multiple sequence alignment of representatives of the domain hit, can be viewed with NCBI's 3D molecular structure viewer, Cn3D (21), now in version 4.1 and enhanced with advanced alignment-building tools that use the PSI-BLAST and threading algorithms. The Conserved Domain Architecture Retrieval Tool (CDART) allows search of protein databases on the basis of a conserved domain and returns the domain architectures of database proteins containing the query domain. Alignment-based protein domain information from the CDD and 3D domains from the MMDB can be searched via the Entrez interface.

HIV-1/Human Protein Interaction Database

The Division of Acquired Immunodeficiency Syndrome (DAIDS) of the National Institute of Allergy and Infectious Diseases (NIAID), in collaboration with the Southern Research Institute and NCBI, has begun compiling a comprehensive 'HIV Protein Interaction Database' to provide a concise summary of documented interactions between HIV-1 proteins and host cell proteins, other HIV-1 proteins or proteins from disease organisms associated with HIV or AIDS. Summaries, including protein RefSeq accession numbers, Entrez Gene ID numbers, lists of interacting amino acids, brief description of interactions, keywords and PubMed IDs for supporting journal articles are presented (www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html). Interaction summaries can be selected for viewing using the pull down phrase lists to apply filters, and batches of summaries may be downloaded. All protein–protein interactions documented in the HIV Protein Interaction Database are listed in Entrez Gene reports in the 'HIV-1 protein interactions' section.

FURTHER INFORMATION

The resources described here include documentation, other explanatory material and references to collaborators and data sources on the respective websites. The NCBI Handbook, available in the Books database, describes the principal NCBI resources in detail. Several tutorials are also offered under the Education link from NCBI's home page. A Site Map provides a comprehensive table of NCBI resources, and the About NCBI feature provides bioinformatics primers and other supplementary information. A user support staff is available to answer questions at info@ncbi.nlm.nih.gov.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
2. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
3. Wu,C.H., Yeh,L.S.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
4. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
5. Bourne,P.E., Addess,K.J., Bluhm,WF, Chen,L., Deshpande,N., Feng,Z., Fleri,W., Green,R., Merino-Ott,J.C., Townsend-Merino,W. *et al.* (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.*, **32**, D223–D225.
6. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58
7. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D., Ke,Z. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.*, **33**, D192–D196.
8. Sequeira,E. (2003) PubMed Central—three years old and growing stronger. *ARL*, **228**, 5–9.
9. Altschul,S.E., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Mcginnis,S. and Madden,T. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
12. Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.
13. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
14. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
15. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
16. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
17. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
18. Ermolaeva,O., Rastogi,M., Pruitt,K.D., Schuler,G.D., Bittner,M.L., Chen,Y., Simon,R., Meltzer,P., Trent,J.M. and Boguski,M.S. (1998) Data management and analysis for gene expression arrays. *Nature Genet.*, **20**, 19–23.
19. Pruitt,K., Tatusov,T. and Maglott,D. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
20. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Pham,L., Smigielski,E. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
21. Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci.*, **25**, 300–302.
22. Tatusova,T., Karsch-Mizrachi,I. and Ostell,J. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
23. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.*, **15**, 417–474.
24. Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **7**, 1051–1060.
25. Barrett,T., Suzek,T., Troup,D., Wilhite,S., Ngau,W., Ledoux,P., Rudnev,D., Lash,A., Fujibuchi,W. and Edgar,R. (2005) NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res.*, **33**, D562–D566.
26. McKusick,V.A. (1998) *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders, 12th edn*. The Johns Hopkins University Press, Baltimore, MD.
27. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
28. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.