D38–D51 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 20 November 2010*
doi:10.1093/nar/gkq1172

Database resources of the National Center for Biotechnology Information

Eric W. Sayers[1,*], Tanya Barrett[1], Dennis A. Benson[1], Evan Bolton[1], Stephen H. Bryant[1], Kathi Canese[1], Vyacheslav Chetvernin[1], Deanna M. Church[1], Michael DiCuccio[1], Scott Federhen[1], Michael Feolo[1], Ian M. Fingerman[1], Lewis Y. Geer[1], Wolfgang Helmberg[2], Yuri Kapustin[1], David Landsman[1], David J. Lipman[1], Zhiyong Lu[1], Thomas L. Madden[1], Tom Madej[1], Donna R. Maglott[1], Aron Marchler-Bauer[1], Vadim Miller[1], Ilene Mizrachi[1], James Ostell[1], Anna Panchenko[1], Lon Phan[1], Kim D. Pruitt[1], Gregory D. Schuler[1], Edwin Sequeira[1], Stephen T. Sherry[1], Martin Shumway[1], Karl Sirotkin[1], Douglas Slotta[1], Alexandre Souvorov[1], Grigory Starchenko[1], Tatiana A. Tatusova[1], Lukas Wagner[1], Yanli Wang[1], W. John Wilbur[1], Eugene Yaschenko[1] and Jian Ye[1]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA and [2]University Clinic of Blood Group Serology and Transfusion Medicine, Medical University of Graz, Auenbruggerplatz 3, A-8036 Graz, Austria

Received September 16, 2010; Revised October 29, 2010; Accepted November 1, 2010

ABSTRACT

In addition to maintaining the GenBank® nucleic acid sequence database, the National Center for Biotechnology Information (NCBI) provides analysis and retrieval resources for the data in GenBank and other biological data made available through the NCBI Web site. NCBI resources include Entrez, the Entrez Programming Utilities, MyNCBI, PubMed, PubMed Central (PMC), Entrez Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link (BLink), Primer-BLAST, COBALT, Electronic PCR, OrfFinder, Splign, ProSplign, RefSeq, UniGene, HomoloGene, ProtEST, dbMHC, dbSNP, dbVar, Epigenomics, Cancer Chromosomes, Entrez Genomes and related tools, the Map Viewer, Model Maker, Evidence Viewer, Trace Archive, Sequence Read Archive, Retroviral Genotyping Tools, HIV-1/Human Protein Interaction Database, Gene Expression Omnibus (GEO), Entrez Probe, GENSAT, Online Mendelian Inheritance in Man (OMIM), Online Mendelian Inheritance in Animals (OMIA), the Molecular Modeling Database (MMDB), the Conserved Domain Database (CDD), the Conserved Domain Architecture Retrieval Tool (CDART), IBIS, Biosystems, Peptidome, OMSSA, Protein Clusters and the PubChem suite of small molecule databases. Augmenting many of the Web applications are custom implementations of the BLAST program optimized to search specialized data sets. All of these resources can be accessed through the NCBI home page at 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® (1) nucleic acid sequence database, which receives data through the international collaboration with DDBJ and EMBL as well as from the scientific community, NCBI provides data retrieval systems and computational resources for the analysis of GenBank data and many other kinds of biological data. For the purposes of this article, after a summary of recent developments and an introduction to the Entrez system, the NCBI suite of resources is grouped into 10 broad categories based on those in the new NCBI Guide. All resources discussed are available from the NCBI Guide at www.ncbi.nlm.nih.gov and can also be located using the Entrez 'Site Search' database. In most cases, the data underlying these resources and

[*]To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov

executables for the software described are available for download at ftp.ncbi.nih.gov.

RECENT DEVELOPMENTS

NCBI site redesign

In late 2009, NCBI launched a long-term project of redesigning and standardizing the NCBI website. Containing more than 4000 pages, the NCBI website is a complex system of interconnected resources, many of which have unique design aspects that can make navigating the NCBI site challenging. To alleviate this, we have adopted a new set of web design standards and have applied them to several resources so far including PubMed, Nuccore, EST, GSS, Protein, Gene, dbVar and Epigenomics. The new pages have four standard elements: (i) the page header, which contains links to the NCBI home page and MyNCBI as well as two pull-down menus that provide navigation to NCBI resources and how-to guides; (ii) the search bar, which contains a pull-down menu of all Entrez databases along with links to search tools and help documentation; (iii) the page body, containing the page content such as search results or data records; and (iv) the page footer, containing five lists of links to information about NCBI, lists of categorized resources and several popular or featured resources. In the coming months, more resources will be adopting this new design that we expect will make the NCBI site more consistent and easier to navigate.

Common elements in the new Entrez page designs

In addition to the standard header and footer, resources that have been updated to conform to the new Entrez design share several common elements: a home page, search tools, display controls and download controls. The home page of a data resource (e.g. www.ncbi.nlm.nih.gov/protein/) contains links to documentation and other information for new users, to relevant tools and to related resources at NCBI. On pages containing search results and data records, new ‘Display Settings’ and ‘Send to’ controls appear on the left and right sides of the display, respectively. These new and simplified controls replace sets of pull-down menus and allow users to select multiple settings at once.

The NCBI Guide

In conjunction with the new web standards discussed above, we replaced the old NCBI home page with the NCBI Guide, an application that serves as an interactive directory of the NCBI site. On the main page of the NCBI Guide, the categories in the Resource pull-down menu in the standard header are duplicated in a list on the left of the page. Clicking on any category displays a list of relevant resources sorted into four groups: databases, downloads, submissions and tools. Popular resources are listed on the right under a ‘Quick Links’ heading. A list of how-to guides is also available via the ‘How-To’ tab on these pages. A list of the most heavily used resources is provided on the main Guide page in the ‘Popular Resources’ box and also as a list in the standard footer.

Epigenomics

The Epigenomics database (www.ncbi.nlm.nih.gov/epigenomics/) is a new information resource at NCBI specifically aimed at highlighting epigenomics data. Epigenomics is an emerging field of research that studies how, despite sharing a common genomic sequence, different cell types and cell lineages acquire distinct patterns of gene expression. Epigenetic features examined include post-translational modifications of histone proteins, genomic DNA methylation, chromatin organization and the expression of non-coding regulatory RNA. Raw data from these experiments, together with extensive meta-data, are stored in the GEO (Gene Expression Omnibus) and SRA (Sequence Read Archive) databases. The new Epigenomics resource provides a higher-level view, allowing users to search and browse the data based on biological attributes such as cell type, tissue type, differentiation stage and heath status, among many others. Data have been pre-mapped to genomic coordinates (to make ‘genome tracks’), so users are not required to be familiar with or manipulate the raw data. Tracks may be visualized in either the NCBI or UCSC genome viewers or may be downloaded to the user’s computer for local analysis. Data from the Roadmap Epigenomics project, which are currently being hosted at GEO (www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/), are being mirrored and are available for viewing and downloading from this new resource.

Database of Genomic Structural Variation

In 2010, NCBI launched the Database of Genomic Structural Variation (dbVar), an archive of large-scale genomic variants such as insertions, deletions, translocations and inversions (www.ncbi.nlm.nih.gov/dbvar/). Currently, dbVar (2) contains over 50 studies from human, rhesus macaque, chimpanzee, mouse, dog, fruit fly and pig, and accepts data derived from several methods including computational sequence analysis and microarray experiments. Each variant is linked to a graphical view showing its genomic context.

Inferred Biomolecular Interactions Server

Recently, NCBI introduced the Inferred Biomoleculars Interactions Server (IBIS), a research server that analyzes and predicts interaction partners and binding site locations in proteins (3). IBIS (www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi) integrates the interactions observed in structural complexes from the Molecular Modeling Database (MMDB) for different types of binding partners including proteins, chemical ligands, nucleic acids, peptides and ions. IBIS also infers binding sites and partners from homologous protein complexes. To emphasize biologically relevant binding sites, similar sites are clustered together based on their evolutionary conservation. In the future, NCBI plans to incorporate observed and inferred interactions of this kind throughout the Entrez 3D structure resources.

New outreach resources and services

NCBI recently redesigned its main Education page (www.ncbi.nlm.nih.gov/Education/) and introduced several new outreach initiatives including training webinars and a new series of courses called Discovery Workshops (4). The new page has links to documentation, educational tools, upcoming conference exhibits and news items. Also on the page are links to the new NCBI pages on Facebook and Twitter, plus YouTube pages that contain short video tutorials and videos from special events at NCBI.

BLAST and COBALT updates

The Short Read Archive (SRA) BLAST page, accessible from the 'Specialized BLAST' section of the main BLAST page (blast.ncbi.nlm.nih.gov), now has an option for searching WGS sequences from 454 Sequencing systems. The WGS sequences are grouped by genus in a pull-down menu, and if multiple species have data within a genus, a separate menu appears allowing individual species to be selected. These data sets are updated daily, so new WGS data are available for searching quickly. The standard BLAST pages now have additional options for filtering searches. If the 'Align two or more sequences' checkbox is not checked, users can either include or exclude data from any number of specified organisms or taxons, greatly increasing the range of customized data sets available. In addition, checkboxes are available that allow users to exclude 'model' sequences (RefSeq XM and XP accessions) as well as sequences from uncultured or environmental samples. Finally, COBALT (5) users can download the output multiple alignment to a file in several popular formats including gapped FASTA, ClustalW, Phylip and Nexus.

MyNCBI updates

MyNCBI allows users to store personal configuration options such as search filters, LinkOut preferences and document delivery providers. Several enhancements have been made to MyNCBI in the past year, including an update to allow users to sign in using credentials for an account with a partner organization such as Google, eRA Commons, VeriSign or a local university. My Bibliography was enhanced to allow users to add citations from books, meetings, presentations, patents and articles not found in PubMed, and also to give users the ability to manage their compliance with the NIH Public Access Policy. In addition, the number of PubMed filter selections has been expanded from five to 15, and users may now change their PubMed default settings for display format, items per page, and the method for sorting search results.

Updates to literature resources

In addition to the changes outlined above for PubMed as part of the Entrez redesign, NCBI released several enhancements for both PubMed and PubMed Central (PMC). For the first time, PubMed now includes citations for book and book chapters available on the NCBI Bookshelf. To aid in searching, an autocomplete feature was added to the PubMed search box, and the PubMed Clinical Queries page (www.ncbi.nlm.nih.gov/pubmed/clinical) was redesigned to show immediate results for clinical studies, systematic reviews and medical genetics side by side. To assist users in finding related literature, PMC full-text views now include a list of related PubMed abstracts on the right. In addition, links to PubMed abstracts cited in the text now appear to the right of the paragraph containing the citation.

New discovery components within the Entrez system

NCBI continued to add new discovery components that assist researchers in finding particular Entrez links and using them to discover interesting relationships within the NCBI databases. Two such components were introduced on protein sequence view pages: an ad that alerts users that the protein being viewed is part of a biological pathway or other system within the Biosystems database, and which provides a link to that pathway; and an ad that describes and links to a cluster of sequences in the Protein Clusters database that includes the protein being viewed. Both of these ads appear on the right column of the sequence view page. For search operations that retrieve 20 or fewer nucleotide or protein sequences, links now appear in the right column that allow users to run BLAST and/or COBALT on all or any checked subset of the sequences.

THE ENTREZ SEARCH AND RETRIEVAL SYSTEM

Entrez databases

Entrez (6) is an integrated database retrieval system that provides access to a diverse set of 38 databases that together contain over 450 million records (Table 1). Entrez supports text searching using simple Boolean queries, downloading of data in various formats and linking of records between databases based on biological relationships. In their simplest form, these links may be cross-references between a sequence and the abstract of the paper in which it is reported, or between a protein sequence and its coding DNA sequence or its three-dimensional (3D) structure. Computationally derived links between 'neighboring records', such as those based on computed similarities among sequences or among PubMed abstracts, allow rapid access to groups of related records. A service called LinkOut expands the range of links to include external services, such as organism-specific genome databases. The records retrieved in Entrez can be displayed in many formats and downloaded singly or in batches.

Entrez programming utilities (E-Utilities)

The Entrez Programming Utilities (E-Utilities) are a suite of eight server-side programs supporting a uniform set of parameters used to search, link and download data from the Entrez databases. EInfo provides basic statistics on a given database, including the last update date and lists of all search fields and available links. ESearch returns the identifiers of records that match an Entrez text query, and

Table 1. The Entrez databases (as of 1 September 2010)

Database	Records	Section within this article
Nucleotide	105 131 187	DNA and RNA
PubChem Substance	72 112 459	Chemicals and Bioassays
SNP	71 036 396	Genetics and Medicine
EST	66 693 283	DNA and RNA
GEO Profiles	63 811 486	Genes and Expression
Protein	35 020 254	Proteins
PubChem Compound	28 801 560	Chemicals and Bioassays
GSS	28 560 647	DNA and RNA
PubMed	20 139 180	Literature
Probe	10 243 420	Genes and Expression
Gene	7 578 739	Genes and Expression
UniGene	4 304 399	Genes and Expression
PubMed Central	2 041 249	Literature
NLM Catalog	1 417 314	Literature
Taxonomy	653 718	Taxonomy
UniSTS	528 865	Genomes
dbVar	510 291	Recent Developments
Protein Clusters	507 133	Proteins
PubChem Bioassay	462 678	Chemicals and Bioassays
3D Domains	313 714	Domains and Structures
Books	288 700	Literature
MeSH	219 574	Literature
Cancer Chromosomes	140 494	Genetics and Medicine
Biosystems	135 309	Genes and Expression
Homologene	123 767	Genes and Expression
PopSet	118 358	DNA and RNA
dbGaP	99 307	Genetics and Medicine
GENSAT	97 980	Genes and Expression
Structure	67 522	Domains and Structures
CDD	40 561	Domains and Structures
GEO Datasets	28 853	Genes and Expression
Journals	25 887	Literature
SRA	25 432	DNA and RNA
OMIM	21 140	Genetics and Medicine
Genome	12 399	Genomes
Genome Projects[a]	5893	Genomes
Site Search	4902	Introduction
OMIA	2658	Genetics and Medicine
Epigenomics	490	Recent Developments
Peptidome	322	Proteins

[a] Soon to be renamed 'BioProjects'.

when combined with EFetch or ESummary, provides a mechanism for downloading the corresponding data records. ELink gives users access to the vast array of links within Entrez so that data related to an input set can be retrieved. By assembling URL or Simple Object Access Protocol (SOAP) calls to the E-utilities within simple scripts, users can create powerful applications to automate Entrez functions to accomplish batch tasks that are impractical using web browsers. Instructions for using the E-Utilities are now found on the NCBI Bookshelf at www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book = helpeutils.

LITERATURE

PubMed

The PubMed database now contains more than 20 million citations dating back to the 1860s from more than 22 000 life science journals. Over 11 million of these citations have abstracts, the earliest from the 1880s, and 11 million have links to their full-text articles, with 3 million having both an abstract and a link to full text. PubMed is heavily linked to other core Entrez databases, thereby providing a crucial bridge between the data of molecular biology and the scientific literature. PubMed records are also linked to one another within Entrez as 'related citations' on the basis of computationally detected similarities using indexed Medical Subject Heading (MeSH) (7) terms and the text of titles and abstracts. The default Abstract display format shows the abstract of a paper along with succinct descriptions of the top five related articles and numerous Discovery Components (see above), increasing the potential for the discovery of important relationships.

PubMed Central

PMC (8) is a digital archive of peer reviewed journals in the life sciences and now contains over 2 million full-text articles, growing by 11% over the past year. More than 1000 journals, including *Nucleic Acids Research*, deposit the full text of their articles in PMC, and more than 400 of these began depositing their data in the last year. Publisher participation in PMC requires a commitment to free access to full text, either immediately after publication or within a 12-month period. As a consequence of the mandatory NIH Public Access Policy that went into effect on 7 April 2008, PMC is also the repository for all final peer-reviewed manuscripts arising from research using NIH funds. All PMC articles are identified in PubMed search results and PMC itself can be searched using Entrez.

The NCBI Bookshelf, the NLM Catalog and the Journals database

The NCBI Bookshelf is an online resource of textbooks, reports and databases in the biomedical sciences. Supported by both the PMC database framework for publishing and archiving and the Entrez system for search and retrieval, the Bookshelf provides users free access to the full text of this content. Bookshelf is now home to over 600 titles, which include NIH-funded reports from the National Academies of Sciences and Clinical Guidelines from UK's National Institute for Health and Clinical Excellence. Databases such as GeneReviews and MICAD (Molecular Imaging and Contrast Agent Database) are updated regularly. Earlier this year, Bookshelf began submitting records to PubMed for a subset of books and chapters in its database. Book records in PubMed can be identified by the 'Books and Documents' label and link back to the respective book or chapter in Bookshelf.

The NLM Catalog provides bibliographic data for over 1.4 million NLM holdings including journals, books, manuscripts, computer software, audio recordings and other electronic resources. Each record is linked to the NLM LocatorPlus service as well as related catalog records with similar title words or associated MeSH terms. The Journals database contains all journals referenced in any Entrez database. Currently holding over 25 000 records, the database indexes for each journal the title abbreviation, the International Organization for Standardization (ISO) abbreviation, publication data and links to the NLM catalog and all Entrez records associated with articles from that journal.

TAXONOMY

The NCBI taxonomy database serves as a central organizing principle for the Entrez biological databases and provides links to all data for each taxonomic node, from superkingdoms to subspecies. The database is growing at the rate of 3800 new taxa per month and indexes over 380 000 organisms named at the genus level or lower that are represented in Entrez by at least one nucleotide or protein sequence. The Taxonomy Browser (www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) can be used to view the taxonomy tree or retrieve data from any of the Entrez databases for a particular organism or group.

DNA AND RNA

Reference sequences

The NCBI Reference Sequence (RefSeq) database (9) is a non-redundant set of curated and computationally derived sequences for transcripts, proteins and genomic regions. The number of nucleotide records in the RefSeq collection has grown by 10% over the past year so that Release 42 (July 2010) contains 4.4 million sequences representing over 10 700 organisms. RefSeq DNA and RNA sequences can be searched and retrieved from the Entrez Nucleotide database, and the complete RefSeq collection is available in the RefSeq directory on the NCBI FTP site.

Sequences from GenBank and other sources

Sequences from GenBank (1) can be searched in and retrieved from three Entrez databases: Nucleotide, EST and GSS (specified as nuccore, nucest and nucgss within the E-utilities). Entrez Nucleotide contains all GenBank sequences except those within the Expressed Sequence Tag (EST) or Genome Survey Sequence (GSS) GenBank divisions. The database also contains Whole Genome Shotgun (WGS) sequences, Third Party Annotation (TPA) sequences and sequences imported from the Entrez Structure database. In addition, those sequences that have been submitted as part of a population, phylogenetic or environmental study are placed in the PopSet database.

The Trace and Assembly archives

The Trace Archive contains over 2 billion traces (12% human) from gel and capillary electrophoresis sequencers. More than 10 000 species are represented. The Trace Assembly Archive links reads in the Trace Archive with genetic sequences in GenBank. An Assembly Viewer displays multiple alignments of assembled reads against consensus sequences to provide support for GenBank deposits.

Sequence Read Archive

The Sequence Read Archive (SRA; 10) is a repository for sequencing data generated from the new generation of sequencers, including the Roche-454 GS and FLX, Illumina Genome Analyzer, Applied Biosystems SOLiD System, Helicos Heliscope, and CompleteGenomics platforms. The SRA is part of the Entrez system and contains over 56 Terabasepairs (Tbp) of biological sequence data. Within Entrez SRA (www.ncbi.nlm.nih.gov/sra/), the data are organized in four types of interlinked records: studies (SRP), experiments (SRX), samples (SRS) and runs (SRR). A study is a collection of related experiments, and each experiment is a set of laboratory operations performed on one or more samples. The results of these experiments are called runs. Additional information about these SRA concepts, along with documentation on using and submitting data to the resource, is available in a new help manual at www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book = helpsra. Sequence read BLAST searches are now offered for transcript and whole-genome sequence data sets from 454 Sequencing systems, and regular expression pattern matching against short reads of all types is possible. A version of the SRA has been deployed behind dbGaP authorized access in order to provide archive services for human sequencing data under usage or privacy restrictions.

PROTEINS

Databases

Reference sequences. In addition to genomic and transcript sequences, the RefSeq database (9) contains protein sequences that are curated and computationally derived from these DNA and RNA sequences. The number of protein records in the RefSeq collection has grown by 29% over the past year so that Release 42 (July 2010) contains 10.6 million protein sequences. RefSeq protein sequences can be searched and retrieved from the Entrez Protein database, and the complete RefSeq collection is available in the RefSeq directory on the NCBI FTP site.

Sequences from GenBank and other sources. As part of standard submission procedures, NCBI produces conceptual translations for any sequence in GenBank (1) that contains a coding sequence and places these protein sequences in the Entrez Protein database. In addition to these 23 million 'GenPept' sequences, the Protein database also contains sequences from TPA, SWISS-PROT (11), the Protein Information Resource (PIR) (12), the Protein Research Foundation (PRF) and the Protein Data Bank (PDB) (13).

Protein Clusters. The Protein Clusters database (www.ncbi.nlm.nih.gov/proteinclusters/) contains over 500 000 sets of almost identical RefSeq proteins encoded by complete genomes from prokaryotes, eukaryotic organelles (mitochondria and chloroplasts), viruses and plasmids as well as from some protozoans and plants. The clusters are organized in a taxonomic hierarchy and are created based on reciprocal best-hit protein BLAST scores (14). These clusters are used as a basis for genome-wide comparison at NCBI as well as to provide simplified BLAST searches via Concise Microbial Protein BLAST (www.ncbi.nlm.nih.gov/genomes/prokhits.cgi). Protein Clusters provides annotations, publications, domains, structures, external links and analysis tools, including multiple sequence alignments and phylogenetic trees.

Peptidome. Peptidome (15) is a data repository for tandem mass spectrometry peptide and protein identification data generated by the scientific community. Data from all stages of a mass spectrometry experiment are captured, including original mass spectra files, experimental metadata and conclusion-level results. The submission process is facilitated through acceptance of data in commonly used open formats, and all submissions undergo syntactic validation and curation in an effort to uphold data integrity and quality. Peptidome accepts data from any tandem mass spectrometry experiment and from any species. In addition to data storage, web-based interfaces are available to help users query, browse and explore individual peptides, proteins, or entire Samples and Studies. Metadata for all public Samples and Studies along with that for the associated proteins in each Sample are loaded into Entrez Peptidome.

HIV-1/Human Protein Interaction Database. The Division of Acquired Immunodeficiency Syndrome of the National Institute of Allergy and Infectious Diseases, in collaboration with the Southern Research Institute and NCBI, maintains a comprehensive HIV Protein-Interaction Database of documented interactions between HIV-1 proteins, host cell proteins, other HIV-1 proteins or proteins from disease organisms associated with HIV or AIDS (16). Summaries, including protein RefSeq accession numbers, Entrez Gene IDs, lists of interacting amino acids, brief descriptions of interactions, keywords and PubMed IDs for supporting journal articles, are presented at www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/ . All protein–protein interactions documented in the HIV Protein-Interaction Database are listed in Entrez Gene reports in the HIV-1 protein interactions section.

Analysis Tools

COBALT. COBALT (5) is a multiple alignment algorithm that finds a collection of pair-wise constraints derived from both the NCBI Conserved Domain database and the sequence similarity programs RPS-BLAST, BLASTP and PHI-BLAST. These pair-wise constraints are then incorporated into a progressive multiple alignment. COBALT searches can be launched either from a BLASTP result page or from the main COBALT search page (http://www.ncbi.nlm.nih.gov/tools/cobalt/), where either FASTA sequences or accessions (or a combination thereof) may be entered into the query sequence box. Links at the top of the COBALT report provide access to a phylogenetic tree view of the multiple alignment and allow users either to launch a modified search or download the alignment in several popular formats.

BLink. BLAST Link (BLink) displays pre-computed BLAST alignments of similar sequences for each protein sequence in Entrez Protein. BLink can display alignment subsets limited by either taxonomic criteria or the database of origin, and provides links to a COBALT multiple sequence alignment of the resulting sequences or a BLAST search with the query protein. BLink links are presented on protein records in Entrez as well as within Entrez Gene reports.

The Open Mass Spectrometry Search Algorithm. The Open Mass Spectrometry Search Algorithm (OMSSA) (21) analyzes MS/MS peptide spectra by searching libraries of known protein sequences, assigning significant hits an expectation value computed in the same way as the E-value of BLAST. The web interface to OMSSA allows up to 2000 spectra to be analyzed in a single session using either the BLAST nr, RefSeq or Swiss-Prot sequence libraries for comparison. Standalone versions of OMSSA that accept larger batches of spectra and allow searches of custom sequence libraries can be downloaded at pubchem.ncbi.nlm.nih.gov/omssa/download.htm.

BLAST SEQUENCE ANALYSIS

BLAST

The BLAST programs (17–19) perform sequence-similarity searches against a variety of nucleotide and protein databases, returning a set of gapped alignments with links to full sequence records as well as to related transcript clusters (UniGene), annotated gene loci (Gene), 3D structures (MMDB) or microarray studies (GEO). The NCBI web interface for BLAST allows users to assign titles to searches, to review recent search results and to save parameter sets in MyNCBI for future use. The basic BLAST programs are also available as standalone command line programs, as network clients and as a local Web-server package at ftp.ncbi.nih.gov/blast/executables/LATEST/ (Table 2).

BLAST databases

The default database for nucleotide BLAST searches (‘Human genomic plus transcript’) contains human RefSeq transcript and genomic sequences arising from the NCBI annotation of the human genome. Searches of this database generate a tabular display that partitions the BLAST hits by sequence type (genomic or transcript) and allows sorting by BLAST score, percent identity within the alignment and the percent of the query sequence contained in the alignment. A similar database is available for mouse. Several other databases are also available and are described in links from the BLAST input form. Each of these databases can be limited to an arbitrary taxonomic node or those records satisfying any Entrez query.

For proteins the default database (nr) is a non-redundant set of all CDS translations from GenBank along with all RefSeq, Swiss-Prot, PDB, PIR and PRF proteins. Subsets of this database are also available, such as the PDB or Swiss-Prot sequences, along with separate databases for sequences from patents and

Table 2. Selected NCBI software available for download

Software	Available binaries	Category within this article
BLAST (stand alone)	Win, Mac, LINUX, Solaris	BLAST Sequence Analysis
BLAST (network client)	Win, Mac, LINUX, Solaris	BLAST Sequence Analysis
BLAST (web server)	Mac, LINUX, Solaris	BLAST Sequence Analysis
CD-Tree	Win, Mac	Domains and Structures
Cn3D	Win, Mac, LINUX, Solaris	Domains and Structures
PC3D	Win, Mac, LINUX	Chemicals and Bioassays
e-PCR	Win, LINUX	Genomes
gene2xml	Win, Mac, LINUX, Solaris	Genes and Expression
Genome Workbench	Win, Mac, LINUX	Genomes
OMSSA	Win, Mac, LINUX	Domains and Structures
splign	LINUX, Solaris	Genomes
prosplign	LINUX	Genomes
tbl2asn	Win, Mac, LINUX, Solaris	Genomes

environmental samples. Like the nucleotide databases, these collections can be limited by taxonomy or an arbitrary Entrez query.

BLAST output formats

Standard BLAST output formats include the default pairwise alignment, several query-anchored multiple sequence alignment formats, an easily parsable Hit Table and a report that organizes the BLAST hits by taxonomy. A 'pairwise with identities' mode better highlights differences between the query and a target sequence. A Tree View option for the Web BLAST service creates a dendrogram that clusters sequences according to their distances from the query sequence. Each alignment returned by BLAST is scored and assigned a measure of statistical significance, called the Expectation Value (E-value). The alignments returned can be limited by an E-value threshold or range.

Genomic BLAST

NCBI maintains Genomic BLAST pages for more than 100 organisms shown in the Map Viewer. By default, genomic BLAST searches the genomic sequence of an organism, but additional databases are also available, such as the nucleotide and protein RefSeqs annotated on the genomic sequence, as well as sets of sequences such as ESTs that are mapped to the genomic sequence. The default search program for the NCBI Genomic BLAST pages is MegaBLAST (20), a faster version of standard nucleotide BLAST designed to find alignments between nearly identical sequences, typically from the same species. For rapid cross-species nucleotide queries, NCBI offers Dis-contiguous MegaBLAST, which uses a non-contiguous word match (21) as the nucleus for its alignments. Dis-contiguous MegaBLAST is far more rapid than a translated search such as blastx, yet maintains a competitive degree of sensitivity when comparing coding regions.

Primer-BLAST

Primer-BLAST is a tool for designing and analyzing polymerase chain reaction (PCR) primers based on the existing program Primer3 (22) that designs PCR primers given a template DNA sequence. Primer-BLAST extends this functionality by running a BLAST search against a chosen database with the designed primers as queries, and then returns only those primer pairs specific to the input template DNA, in that they do not generate valid PCR products on sequences other than the template. Users can also specify a forward or reverse primer in addition to a DNA template, in which case the other primer will be designed and analyzed. If both primers are specified along with a template, the tool performs only the final BLAST analysis. Users may also enter two primers without a template, in which case the BLAST analysis will display those templates in the chosen database that best match the primer pair. The available databases range from RefSeq mRNA or genomic sets for one of twelve model organisms to the entire BLAST *nr* database.

GENES AND EXPRESSION

Entrez Gene

Entrez Gene (23) provides an interface to curated sequences and descriptive information about genes with links to NCBI's Map Viewer, Evidence Viewer, Model Maker, BLink, protein domains from the Conserved Domain Database (CDD), and other gene-related resources. Gene contains data for almost 6.7 million genes from over 6700 organisms. These 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 and provided as Gene References into Function (GeneRIF). The complete Entrez Gene data set, as well as organism-specific subsets, is available in the compact NCBI ASN.1 format on the NCBI FTP site. The gene2xml tool converts the native Gene ASN.1 format into XML and is available at ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/gene2xml/.

RefSeqGene

In collaboration with Locus Reference Genomic (LRG) (www.lrg-sequence.org), RefSeqGene provides stable, standard genomic sequences annotated with standard mRNAs for well-characterized human genes (9). RefSeqGene records are part of the RefSeq collection and are created in consultation with authoritative locus-specific databases or other experts on particular loci and provide a stable genomic sequence for establishing numbering systems for exons and introns and for reporting and identifying genomic variants, especially those of clinical importance (24). By default, a RefSeqGene record begins 5 kb upstream of the first exon of the gene and ends 2 kb downstream of the final exon, but those positions will be adjusted on request. A RefSeqGene sequence may differ from the current genomic build so as to reflect standard alleles. RefSeqGene records can be retrieved from Entrez Nucleotide using the query 'refseqgene[keyword]', are available on corresponding Entrez Gene reports and can be downloaded from ftp.ncbi.nih.gov/refseq/H_sapiens/RefSeqGene.

The Conserved CDS database

The Conserved CDS database (CCDS) project (www.ncbi.nlm.nih.gov/CCDS/) is a collaborative effort among NCBI, the European Bioinformatics Institute, the Wellcome Trust Sanger Institute and University of California, Santa Cruz (UCSC) to identify a set of human and mouse protein coding regions that are consistently annotated and of high quality. To date, the CCDS database contains over 23 700 human and 17 700 mouse CDS annotations. The web interface to the CCDS allows searches by gene or sequence identifiers and provides links to Entrez Gene, record revisions histories, transcript and proteins sequences and gene views in Map Viewer, the Ensemble Genome Browser, the UCSC Genome Browser and the Sanger Institute Vega Browser. The CCDS sequence data are available at ftp.ncbi.nlm.nih.gov/pub/CCDS/.

Gene Expression Omnibus

Gene Expression Omnibus (GEO) (25) is a data repository and retrieval system for high-throughput functional genomic data generated by microarray and next-generation sequencing technologies. In addition to gene expression data, GEO accepts other categories of experiments including studies of genome copy number variation, genome-protein interaction surveys and methylation profiling studies. The repository can capture fully annotated raw and processed data, enabling compliance with major community-derived scientific reporting standards such as 'Minimum Information About a Microarray Experiment' (MIAME) (26,27). Several data deposit options and formats are supported, including web forms, spreadsheets, XML and plain text. GEO data are housed in two Entrez databases: GEO Profiles, which contains quantitative gene expression measurements for one gene across an experiment, and GEO Datasets, which contains entire experiments. Currently, the GEO database hosts over 18 000 studies submitted by 8000 laboratories and comprising 460 000 samples and 33 billion individual abundance measurements for over 1300 organisms.

UniGene and ProtEST

UniGene (28) is a system for partitioning transcript sequences (including ESTs) from GenBank into a non-redundant set of clusters, each of which represents a potential gene locus. UniGene clusters are created for all organisms for which there are 70 000 or more ESTs in GenBank and includes ESTs for 68 animals, 54 plants and fungi and another six eukaryotes. UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences. As an aid to identifying a UniGene cluster, ProtEST presents precomputed BLAST alignments between protein sequences from model organisms and the six-frame translations of nucleotide sequences in UniGene.

Homologene

HomoloGene is a system that automatically detects homologs, including paralogs and orthologs, among the genes of 20 completely sequenced eukaryotic genomes. HomoloGene reports include homology and phenotype information drawn from Online Mendelian Inheritance in Man (OMIM) (29), Mouse Genome Informatics (MGI) (30), Zebrafish Information Network (ZFIN) (31), Saccharomyces Genome Database (SGD) (32), Clusters of Orthologous Groups (COG) (33) and FlyBase (34). The HomoloGene Downloader, appearing under the 'Download' link in HomoloGene displays, retrieves transcript, protein, or genomic sequences for the genes in a HomoloGene group; in the case of genomic sequence, upstream and downstream regions may be specified.

GENSAT

GENSAT (35–37) is a gene expression atlas of the mouse central nervous system produced with data supplied by the Rockefeller University and the St. Jude Children's Research Hospital. GENSAT (www.ncbi.nlm.nih.gov/projects/gensat/) catalogs images of histological sections of the mouse brain in which biochemical tags have been used to visualize local gene expression. In addition to search tools, GENSAT provides download, zoom and comparison facilities for the more than 97 000 images in the collection.

Probe

The NCBI Probe database is a public registry of nucleic acid reagents designed for use in a wide variety of biomedical research applications, together with information on reagent distributors, probe effectiveness and computed sequence similarities. The Probe database archives 10.2 million probe sequences, among them probes for genotyping, single-nucleotide polymorphism (SNP) discovery, gene expression, gene silencing and gene mapping. The probe database also provides submission templates to simplify the process of depositing data (www.ncbi.nlm.nih.gov/genome/probe/doc/Submitting.shtml).

Biosystems

NCBI Biosystems (www.ncbi.nlm.nih.gov/biosystems/) collects together molecules that interact in a biological system, such as a biochemical pathway or disease. Currently, Biosystems receives data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (38–40), BioCyc (41), Reactome (42) and the Pathway Interaction Database (43). These source databases provide diagrams of pathways that display the various components with their substrates and products, as well as links to relevant literature. In addition to being linked to such literature in PubMed, each component within a Biosystem record is also linked to the corresponding records in Entrez Gene and Protein, while the substrates and products are linked to records in PubChem (see below) so that the Biosystem record centralizes NCBI data related to the pathway, greatly facilitating computation on such systems.

GENOMES

Databases

Entrez Genome. Entrez Genome (44) provides access to genomic sequences from the RefSeq collection and is a convenient portal both for retrieving such sequences from multiple organisms and for viewing small genomes, such as those from prokaryotes. Currently, the database contains complete genomes for more than 1200 microbes and 3600 viruses, as well as for over 2400 eukaryotic organelles. For higher eukaryotes, the Genome database includes complete genomes for 39 species, as well as data from over 800 other genome sequencing projects. More than 11% of the 12 400 total sequences were added in the past year. For higher eukaryotes, Entrez Genome provides direct links to the NCBI Map Viewer; for prokaryotes, viruses and eukaryotic organelles, specialized viewers and BLAST pages are available. The Plant Genomes Central Web page serves as a portal to completed plant genomes, to information on plant genome sequencing projects or to other resources at NCBI such as the plant Genomic BLAST pages or Map Viewer.

Entrez Genome Projects. The Entrez Genome Projects database, soon to be renamed Entrez BioProjects, provides an overview of the status of a variety of genomic and other biomedical projects, ranging from large-scale sequencing and assembly projects to projects focused on a particular locus, such as 16 S ribosomal RNA, or a viral disease, such as SARS. The scope of the database continues to expand so that only one-third of the more than 15 000 projects are traditional single-organism genome sequencing projects, while the other two-thirds are projects such as viral population projects, metagenome and environmental sampling projects, comparative genomics projects and transcriptome projects. Genome Projects links to project data in the other Entrez databases, such as Entrez Nucleotide and Genome, and to a variety of other NCBI and external resources. For prokaryotic organisms, Genome Projects indexes a number of characteristics of interest to biologists such as organism morphology and motility, pathogenicity and environmental requirements such as salinity, temperature, oxygen levels and pH range. NCBI encourages depositors to register their projects early in their development so that project data can be linked via the project ID to other NCBI-hosted data at the earliest opportunity.

Influenza Genome resources. The Influenza Genome Sequencing Project (IGSP) (45) is providing researchers with a growing collection of over 46 000 virus sequences essential to the identification of the genetic determinants of influenza pathogenicity. NCBI's Influenza Virus Resource links the IGSP project data via PubMed to the most recent scientific literature on influenza as well as to a number of online analysis tools and databases. These databases include NCBI's Influenza Virus Sequence Database, comprised of over 150 000 influenza sequences in the GenBank and RefSeq databases, as well as other Entrez databases containing 167 000 influenza protein sequences, 170 influenza protein structures and 590 influenza population studies. An online influenza genome annotation tool analyzes a novel sequence and produces output in a 'feature table' format that can be used by NCBI's GenBank submission tools such as tbl2asn (1).

NCBI now also provides the Virus Variation resource (www.ncbi.nlm.nih.gov/genomes/VirusVariation/) that extends services available for Influenza to other viruses, such as the Dengue virus. Virus Variation provides a portal for retrieving, downloading, analyzing and annotating virus sequences using pages customized to unique aspects of viral sequence data, including genotype, severity of the resulting disease and the year a sample was collected.

Analysis tools

Map Viewer. The NCBI Map Viewer (www.ncbi.nlm.nih.gov/mapview/) displays genome assemblies, genetic and physical markers and the results of annotation and other analyses using sets of aligned maps for 110 organisms. The available maps vary by organism and may include cytogenetic maps, physical maps and a variety of sequence-based maps. Maps from multiple organisms or multiple assemblies for the same organism can be displayed in a single view. Map Viewer also can display previous genome builds and can produce convenient formats for downloading data.

Genome Workbench. NCBI's Genome Workbench is a stand-alone application (Table 2) for sequence and genomic evaluation, offering tools for visualization and analysis, including integrated graphical views of sequences and alignments, text and tabular displays of annotation and common sequence analysis tools, including BLAST, MUSCLE and Splign. Genome Workbench offers the power of computation on a user's own computer, and can easily mix private data with data available for public retrieval. The NCBI Genome Workbench Team has recently released an updated version, v.2.1.2. The new version contains many critical bug fixes, and all current users are encouraged to upgrade.

Model Maker and Evidence Viewer. Model Maker 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 RefSeqs, to the NCBI human genome assembly. The Evidence Viewer summarizes the sequence evidence supporting a gene annotation by displaying alignments of RefSeq and GenBank transcripts, along with ESTs, to genomic contigs. The tool also shows detailed alignments for each exon, and highlights mismatches between the transcript and genomic sequences.

Open Reading Frame Finder, Splign and ProSplign. NCBI provides several tools that assist in identifying coding sequences in genomic DNA. The Open Reading Frame (ORF) Finder (www.ncbi.nlm.nih.gov/projects/gorf/) performs a six-frame translation of a nucleotide sequence and returns the location of each ORF within a specified size range. Splign (46) (www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi) is a utility for computing cDNA-to-genomic sequence alignments that is accurate in determining splice sites, tolerant of sequencing errors and supports cross-species alignments. Splign uses a version of the Needleman-Wunsch algorithm (47) that accounts for splice signals in combination with a compartmentalization algorithm to identify possible locations of genes and their copies. A link to download a standalone version designed for large-scale processing is provided on the Splign web page. Finally, ProSplign (www.ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html) aligns protein sequences to genomic DNA sequences using an algorithm similar to that of Splign in that it accounts for introns and splice signals to yield optimal alignments. Standalone versions of the program are also available on the ProSplign web page.

Electronic PCR. Forward electronic PCR (e-PCR) searches for matches to STS primer pairs in the UniSTS database of almost 530 000 markers. Reverse e-PCR is used to estimate the genomic binding site, amplicon size and specificity for sets of primer pairs by searching against genomic and transcript databases. Both e-PCR binaries and source code are available at ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR.

TaxPlot, GenePlot and gMap. TaxPlot plots similarities in the proteomes of two organisms to that of a reference organism for complete prokaryotic and eukaryotic genomes. A related tool, GenePlot, generates plots of protein similarity for a pair of complete microbial genomes to visualize deleted, transposed or inverted genomic segments. The gMap tool combines the results of pre-computed whole microbial genome comparisons with on-the-fly BLAST comparisons, clustering genomes with similar nucleotide sequences, and then graphically depicting the precomputed segments of similarity.

GENETICS AND MEDICINE

The Database of Genotypes and Phenotypes

Within Entrez, the Database of Genotypes and Phenotypes (dbGaP) (48) (www.ncbi.nlm.nih.gov/gap/) archives, distributes and supports submission of data that correlate genomic characteristics with observable traits. This database is a designated NIH repository for NIH-funded genome wide association study (GWAS) results (grants.nih.gov/grants/gwas/index.htm). The dbGaP collection contains over 240 studies, 33% of which were submitted in the past year, and each of which can be browsed by name or disease.

To protect the confidentiality of study subjects, dbGaP accepts only de-identified data and requires investigators to go through an authorization process in order to access individual-level data. Study documents, protocols and subject questionnaires are available without restriction. Authorized access data distributed to primary investigators for use in approved research projects includes de-identified phenotypes and genotypes for individual study subjects, pedigrees and some precomputed associations between genotype and phenotype.

Database of Single Nucleotide Polymorphisms

Database of Single Nucleotide Polymorphisms (dbSNP) (49), a repository for single-base nucleotide substitutions and short deletion and insertion polymorphisms, contains over 30 million human records and 40 million more from a variety of other organisms. In addition to archiving the sequence that defines the variant, dbSNP maintains information about the validation status, population-specific allele frequencies, PubMed citations and individual genotypes for clustered reference records (rs#). These data are available on the dbSNP FTP site (ftp://ftp.ncbi.nih.gov/snp/organisms/) in XML-structured genotype and VCF reports that include information about cell lines, pedigree IDs, allele frequency and error flags for genotype inconsistencies and incompatibilities.

In collaboration with Locus Specific Databases (LSDBs), dbSNP integrates information about rare genetic variants with clinical relevance. Two web submission forms were created to facilitate submission of LSDB/Clinical variant information and support variant descriptions using the HGVS standards with a RefSeq standard sequence. Users can search and annotate existing variations or submit novel ones, either as a single variation (http://www.ncbi.nlm.nih.gov/projects/SNP/tranSNP/tranSNP.cgi) or as a batch (http://www.ncbi.nlm.nih.gov/projects/SNP/tranSNP/VarBatchSub.cgi).

GeneReviews and GeneTests

NCBI hosts GeneReviews and GeneTests, two resources developed by a team led by Roberta A. Pagon, MD at the University of Washington. GeneReviews (www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book = gene) is a compendium of continually updated, expert-authored and peer-reviewed disease descriptions that relate genetic testing to the diagnosis, management and genetic counseling of patients and families with specific inherited conditions (50,51). These reviews can be searched via the GeneReviews tab at the GeneTests home page (www.ncbi.nlm.nih.gov/sites/GeneTests/), NCBI's Bookshelf site, NCBI's All Databases interface, or major web search engines.

The GeneTests Laboratory Directory and Clinic Directory list information voluntarily provided by laboratories about their tests and by genetics clinics about their clinical genetics services. As appropriate, users can search by a disease name, gene symbol, protein name, clinical genetics service and information about a lab/clinic, such as its name, director and location. Clinics in the United States can also be found via a map-based search. Together, GeneReviews and the GeneTests directories support the integration of information on genetic disorders and genetic testing into a single resource to facilitate the care of patients and families with inherited conditions.

OMIM

NCBI provides as part of Entrez the online version of the Mendelian Inheritance in Man catalog of human genes and genetic disorders authored and edited by the late Victor A. McKusick and his staff at The Johns Hopkins University (29). The database contains information on disease phenotypes and genes, including extensive descriptions, gene names, inheritance patterns, map locations, gene polymorphisms and detailed bibliographies. Entrez OMIM contains over 21 000 entries, including data on over 13 100 established gene loci and phenotypic descriptions.

Online Mendelian Inheritance in Animals

Online Mendelian Inheritance in Animals (OMIA) is a database of genes, inherited disorders and traits in animal species other than human and mouse, and is authored by Professor Frank Nicholas of the University of Sydney, Australia and colleagues (52). The database holds 2600 records containing textual information and references, as well as links to relevant records from OMIM, PubMed and Entrez Gene.

Cancer Chromosomes

Cancer Chromosomes (53) contains data on human and mouse chromosomal aberrations, such as deletions and translocations, which are associated with cancer. Cancer Chromosomes consists of three databases: the NCI/NCBI SKY (Spectral Karyotyping)/M-FISH (Multiplex-FISH) and CGH (Comparative Genomic Hybridization) Database, the National Cancer Institute Mitelman Database of Chromosome Aberrations in Cancer (54) and the NCI Recurrent Chromosome Aberrations in Cancer database. Graphical schematics of each aberration in the SKY/M-FISH and CGH collections are available along with clinical case information and links to relevant literature. Cancer Chromosomes also provides similarity reports that list terms common to a group of records returned by a search, including similarities between CGH data and karyotypes.

Database cluster for routine clinical applications: dbMHC, dbLRC and dbRBC

dbMHC (www.ncbi.nlm.nih.gov/projects/gv/mhc/) focuses on the Major Histocompatibility Complex (MHC) and contains sequences and frequency distributions for alleles of the MHC, an array of genes that play a central role in the success of organ transplants and an individual's susceptibility to infectious diseases. dbMHC also contains HLA genotype and clinical outcome information on hematopoietic cell transplants performed worldwide. dbLRC offers a comprehensive collection of alleles of the leukocyte receptor complex with a focus on KIR genes. dbRBC represents data on genes and their sequences for red blood cell antigens or blood groups. It hosts the Blood Group Antigen Gene Mutation Database (55) and integrates it with resources at NCBI. dbRBC provides general information on individual genes and access to the ISBT allele nomenclature of blood group alleles. All three databases, dbMHC, dbLRC and dbRBC, provide multiple sequence alignments, analysis tools to interpret homozygous or heterozygous sequencing results (56) and tools for DNA probe alignments.

DOMAINS AND STRUCTURES

The Molecular Modeling Database

The NCBI Molecular Modeling Database (MMDB) (57) contains experimentally determined coordinate sets from the Protein Data Bank (13), augmented with domain annotations and links to relevant literature, protein and nucleotide sequences, chemicals (PDB heterogens) and conserved domains in CDD (58). Compact structural domains within protein structures are stored in the 3D Domains database, and structural neighbors computed by the VAST algorithm (59,60) are available for structures containing these domains. Structure record summaries retrieved by text searches display thumbnail images of structures that link to interactive views of the data in Cn3D (61), the NCBI structure and alignment viewer. NCBI also provides precomputed BLAST results against the PDB database for all proteins in Entrez through the 'Related Structures' link.

CDD and CDART

The Conserved Domain Database (CDD) (58) contains over 37 000 PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the Simple Modular Architecture Research Tool (Smart) (62), Pfam (63), TIGRFAM (64) and from domain alignments derived from COGs and Entrez Protein Clusters. In addition, CDD includes 3100 superfamily records, each of which contains a set of CDs from one or more source databases that generate overlapping annotation on the same protein sequences. The NCBI Conserved Domain Search (CD-Search) service locates conserved domains within a protein sequence, and these results are available for all proteins in Entrez through the 'Identify Conserved Domains' link in the upper right of a sequence record. Wherever possible, protein sequences with known 3D structures are included in CDD alignments, which can be viewed along with these structures and also edited within Cn3D. The Conserved Domain Architecture Retrieval Tool (CDART) allows searches of protein databases on the basis of a conserved domain and returns the domain architectures of database proteins containing the query domain. CD alignments can be viewed online, edited or created *de novo* using CDTree. CDTree uses PSI-BLAST to add new sequences to an existing CD alignment and provides an interface for exploring phylogenetic trends in domain architecture and for building hierarchies of alignment-based protein domains.

CHEMICALS AND BIOASSAYS

PubChem (65) is the informatics backbone for the NIH Roadmap Initiative on molecular libraries and focuses on the chemical, structural and biological properties of small molecules, in particular their roles as diagnostic and therapeutic agents. A suite of three Entrez databases, PCSubstance, PCCompound and PCBioAssay, contain the structural and bioactivity data of the PubChem project. The databases hold records for 72 million substances containing 29 million unique structures. Nearly 1.8 million of these substances have bioactivity data in at least one of the 460 000 PubChem BioAssays. PubChem also provides a single, low-energy 3D conformer for about 90% of the records in the PubChem Compound database. A viewing application, PC3D, is available to view both individual conformers and overlays of similar conformers. The PubChem databases link not only to other Entrez databases such as PubMed and PubMed Central but also to Entrez Structure and Protein to provide a bridge between the macromolecules of genomics and the small organic molecules of cellular metabolism. The PubChem databases are searchable using text queries as well as structural queries based on chemical SMILES, formulas or chemical structures provided in a variety of formats. An online structure-drawing tool (pubchem.ncbi.nlm.nih.gov/search/search.cgi) provides a simple way to construct a structure-based search.

FOR FURTHER INFORMATION

The resources described here include documentation, other explanatory material and references to collaborators and data sources on the respective websites. The NCBI Help Manual and the NCBI Handbook, both available in the NCBI Bookshelf, describe the principal NCBI resources in detail. Several tutorials are also offered under with the Training and Tutorials category link on the left side of the NCBI home page. An alphabetical list of NCBI resources is available from a link in the upper left of the NCBI home page, and the About NCBI pages provide bioinformatics primers and other supplementary information. A user-support staff is available to answer questions at info@ncbi.nlm.nih.gov. Updates on NCBI resources and database enhancements are described in the NCBI News newsletter (www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=newsncbi). In addition, NCBI supports several mailing lists that provide updates (www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html), as well as RSS feeds (www.ncbi.nlm.nih.gov/feed/).

FUNDING

Funding for open access charge: the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) GenBank. *Nucleic Acids Res.*, doi:10.1093/nar/gkq1079.
2. Church,D.M., Lappalainen,I., Sneddon,T.P., Hinton,J., Maguire,M., Lopez,J., Garner,J., Paschall,J., Dicuccio,M., Yaschenko,E. *et al.* (2010) Public data archives for genomic structural variation. *Nat. Genet.*, **42**, 813–814.
3. Shoemaker,B.A., Zhang,D., Thangudu,R.R., Tyagi,M., Fong,J.H., Marchler-Bauer,A., Bryant,S.H., Madej,T. and Panchenko,A.R. (2010) Inferred Biomolecular Interaction Server–a web server to analyze and predict protein interacting partners and binding sites. *Nucleic Acids Res.*, **38**, D518–D524.
4. Cooper,P.S., Lipshultz,D., Matten,W.T., McGinnis,S.D., Pechous,S., Romiti,M.L., Tao,T., Valjavec-Gratian,M. and Sayers,E.W. (2010) Education resources of the National Center for Biotechnology Information. *Brief Bioinform.*, doi:10.1093/bib/bbq022.
5. Papadopoulos,J.S. and Agarwala,R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**, 1073–1079.
6. 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.
7. Sewell,W. (1964) Medical subject headings in Medlars. *Bull. Med. Libr. Assoc.*, **52**, 164–170.
8. Sequeira,E. (2003) PubMed Central - three years old and growing stronger. *ARL*, **228**, 5–9.
9. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
10. Leinonen,R., Sugawara,H. and Shumway,M. (2011) The Sequence Read Archive. *Nucleic Acids Res.*, **39**, D19–D21.
11. 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.
12. Wu,C.H., Yeh,L.S., 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.
13. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
14. Klimke,W., Agarwala,R., Badretdin,A., Chetvernin,S., Ciufo,S., Fedorov,B., Kiryutin,B., O'Neill,K., Resch,W., Resenchuk,S. *et al.* (2009) The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.*, **37**, D216–D223.
15. Ji,L., Barrett,T., Ayanbule,O., Troup,D.B., Rudnev,D., Muertter,R.N., Tomashevsky,M., Soboleva,A. and Slotta,D.J. (2010) NCBI Peptidome: a new repository for mass spectrometry proteomics data. *Nucleic Acids Res.*, **38**, D731–D735.
16. Fu,W., Sanders-Beer,B.E., Katz,K.S., Maglott,D.R., Pruitt,K.D. and Ptak,R.G. (2009) Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res.*, **37**, D417–D422.
17. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
18. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., 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.
19. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.

20. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
21. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
22. Rozen,S. and Skalestsky,H.J. (2000) In Krawetz,S. and Misener,S. (eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Totowa, NJ, Humana Press, pp. 365–386.
23. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
24. Gulley,M.L., Braziel,R.M., Halling,K.C., Hsi,E.D., Kant,J.A., Nikiforova,M.N., Nowak,J.A., Ogino,S., Oliveira,A., Polesky,H.F. *et al.* (2007) Clinical laboratory reports in molecular pathology. *Arch. Pathol. Lab. Med.*, **131**, 852–863.
25. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.
26. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
27. Whetzel,P.L., Parkinson,H., Causton,H.C., Fan,L., Fostel,J., Fragoso,G., Game,L., Heiskanen,M., Morrison,N., Rocca-Serra,P. *et al.* (2006) The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, **22**, 866–873.
28. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
29. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
30. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2007) The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res.*, **35**, D630–D637.
31. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P., Ramachandran,S. *et al.* (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**, D581–D585.
32. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.
33. 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.
34. Crosby,M.A., Goodman,J.L., Strelets,V.B., Zhang,P. and Gelbart,W.M. (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res.*, **35**, D486–D491.
35. Geschwind,D. (2004) GENSAT: a genomic resource for neuroscience research. *Lancet Neurol.*, **3**, 82.
36. Gong,S., Zheng,C., Doughty,M.L., Losos,K., Didkovsky,N., Schambra,U.B., Nowak,N.J., Joyner,A., Leblanc,G., Hatten,M.E. *et al.* (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature*, **425**, 917–925.
37. Heintz,N. (2004) Gene expression nervous system atlas (GENSAT). *Nat. Neurosci.*, **7**, 483.
38. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
39. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
40. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
41. Keseler,I.M., Bonavides-Martinez,C., Collado-Vides,J., Gama-Castro,S., Gunsalus,R.P., Johnson,D.A., Krummenacker,M., Nolan,L.M., Paley,S., Paulsen,I.T. *et al.* (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.*, **37**, D464–D470.
42. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
43. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
44. Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
45. Ghedin,E., Sengamalay,N.A., Shumway,M., Zaborsky,J., Feldblyum,T., Subbu,V., Spiro,D.J., Sitz,J., Koo,H., Bolotov,P. *et al.* (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**, 1162–1166.
46. Kapustin,Y., Souvorov,A., Tatusova,T. and Lipman,D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct*, **3**, 20.
47. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
48. Manolio,T.A., Rodriguez,L.L., Brooks,L., Abecasis,G., Ballinger,D., Daly,M., Donnelly,P., Faraone,S.V., Frazer,K., Gabriel,S. *et al.* (2007) New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.*, **39**, 1045–1051.
49. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
50. Pagon,R.A. (2006) GeneTests: an online genetic information resource for health care providers. *J. Med. Libr. Assoc.*, **94**, 343–348.
51. Waggoner,D.J. and Pagon,R.A. (2009) Internet resources in medical genetics. *Curr. Protoc. Hum. Genet.*, **Chapter 9**, Unit 9 12.
52. Lenffer,J., Nicholas,F.W., Castle,K., Rao,A., Gregory,S., Poidinger,M., Mailman,M.D. and Ranganathan,S. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res.*, **34**, D599–D601.
53. Knutsen,T., Gobu,V., Knaus,R., Padilla-Nash,H., Augustus,M., Strausberg,R.L., Kirsch,I.R., Sirotkin,K. and Ried,T. (2005) The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. *Genes Chromosomes Cancer*, **44**, 52–64.
54. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat. Genet.*, **15**, 417–474.
55. Blumenfeld,O.O. and Patnaik,S.K. (2004) Allelic genes of blood group antigens: a source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. *Hum. Mutat.*, **23**, 8–16.
56. Helmberg,W., Dunivin,R. and Feolo,M. (2004) The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences. *Nucleic Acids Res.*, **32**, W173–W175.
57. Wang,Y., Addess,K.J., Chen,J., Geer,L.Y., He,J., He,S., Lu,S., Madej,T., Marchler-Bauer,A., Thiessen,P.A. *et al.* (2007) MMDB: annotating protein sequences with Entrez's 3D-structure database. *Nucleic Acids Res.*, **35**, D298–D300.
58. Marchler-Bauer,A., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R., Gwadz,M. *et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.*, **37**, D205–D210.
59. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

60. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
61. 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.
62. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
63. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
64. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
65. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.