Published online 12 November 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D5–D16***
doi:10.1093/nar/gkp967

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], 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], 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 15, 2009; Revised October 6, 2009; Accepted October 13, 2009

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, Entrez Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link (BLink), Electronic PCR, OrfFinder, Spidey, Splign, Reference Sequence, UniGene, HomoloGene, ProtEST, dbMHC, dbSNP, 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, Entrez Probe, GENSAT, Online Mendelian Inheritance in Man, Online Mendelian Inheritance in Animals, the Molecular Modeling Database, the Conserved Domain Database, the Conserved Domain Architecture Retrieval Tool, Biosystems, Peptidome, 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 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 other biological data. For the purposes of this article, the NCBI suite of resources is grouped into nine broad categories that are discussed after a summary of recent developments. All resources discussed are available from the NCBI home page at www.ncbi.nlm.nih.gov and can be located using the Entrez *Site Search* database. In most cases, the data underlying these resources and executables for the software described are available for download at ftp.ncbi.nih.gov.

RECENT DEVELOPMENTS

Biosystems

NCBI Biosystems (www.ncbi.nlm.nih.gov/biosystems/) is a new database within Entrez that collects together

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

molecules that interact in a biological system, such as a biochemical pathway or disease. Currently, Biosystems receives data from two sources: the Kyoto Encyclopedia of Genes and Genomes (2–4) and the EcoCyc subset of the BioCyc database (5). 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.

BLAST improvements and updates

There have been three main improvements to the NCBI BLAST web site this year. The first is the addition of Sequence Read Archive (SRA) transcript libraries as a new search set, which includes all public sequences from 454 sequencing systems. These sequences can be searched using the 'Search SRA transcript libraries' link in the 'Specialized BLAST' section of the BLAST web site. NCBI has also reorganized the page for aligning two sequences using BLAST (bl2seq), which now has a search page consistent with the other BLAST pages. On this page, users can enter multiple query sequences and multiple subject sequences, instead of one each as on the older page. The report for the new page is also a standard BLAST report, although a 'Dot Matrix View' is available if only one query and one subject sequence are entered. Finally, the BLASTP report now offers a new 'Multiple Alignment' option that uses COBALT (6) to perform a multiple alignment of the query sequence and any subject sequences listed in the BLAST report. If the user selects this link, a separate multiple alignment search is started and displayed in a separate browser window.

COBALT

COBALT (6) is a new multiple alignment algorithm that finds a collection of pairwise constraints derived from both the NCBI Conserved Domain database (CDD) and the sequence similarity programs RPS-BLAST, BLASTP and PHI-BLAST. These pairwise 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. A COBALT report will then be displayed with the input protein titles at the top and the multiple alignment at the bottom. From this page, it also possible to get a tree view for the multiple alignment or to launch a modified search using the 'Edit and Resubmit' link. In the near future, the tool will provide additional display and download options such as gapped FASTA.

Discovery components within the Entrez system

Underlying and connecting the several databases within the Entrez system is an extensive network of links and precalculated similarity data that have been relatively inaccessible to users. In an effort to assist researchers in finding these links and using them to discover interesting relationships within the NCBI databases, NCBI is developing three types of 'discovery components' on Entrez web pages: sensors, which analyze search queries and display data potentially related to the query terms; database 'ads', which promote links to highly relevant data in a different database; and analysis tools, which provide further insight on the record being viewed. Examples of such components released so far include the citation and gene sensors in PubMed that, respectively, activate when citation elements or gene symbols appear in a query; the PubMed Central (PMC) and three-dimensional (3D) structure ads on PubMed abstract pages that provide links to free full-text articles or 3D structures reported by the paper; and BLAST and Primer-BLAST links provided on nucleotide sequence records. As part of this effort, the nucleotide and protein record pages were redesigned to highlight numerous links from sequences to related data including literature, Reference Sequences (RefSeqs), genes, gene homologs, transcript clusters, clones and conserved domains.

GeneReviews and GeneTests

NCBI now hosts GeneReviews and GeneTests, two resources developed by a team led by Roberta A. Pagon, 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 (7,8). 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 services 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 USA 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.

H1N1 influenza sequences

In response to the 2009 H1N1 influenza outbreak, NCBI provided a new web page as part of the NCBI Influenza Virus Resource (described below) that allows direct access to all H1N1 sequences as they are submitted

(www.ncbi.nlm.nih.go/genomes/FLU/SwineFlu.html). From this page, users can download all available sequences (currently 5000) in a single batch. In addition, NCBI has created a record in the Projects database (project ID 37813) to centralize all data related to the H1N1 influenza virus.

MyNCBI updates

MyNCBI allows users to store personal configuration options such as search filters, LinkOut preferences and document delivery providers. After logging into their MyNCBI account, a user can save searches and arrange to receive periodic emails containing updated search results. A MyNCBI feature called 'Collections' allows users to save search results and bibliographies indefinitely. Several enhancements have been made to MyNCBI in the past year, particularly regarding sharing information with other users. A new 'Shared Settings' panel provides a single interface where a user can select settings to be shared, and then by constructing a simple URL and providing it to other users, the entire group can access these common settings. In MyNCBI, both collections and bibliographies can now be set as either private or public, the latter of which can be shared with multiple users. Finally, the Recent Activity feature has been dramatically expanded to include up to 6 months of activity within MyNCBI, rather than only a user's previous five actions.

Peptidome

Peptidome (9) is a new 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 is not restricted to specific organisms, instruments or experiment types; data from any tandem mass spectrometry experiment and from any species are accepted. 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.

PubChem 3D and PC3D

PubChem now provides 3D conformers for ~70% of the 25 million records in the PubChem Compound database. Currently, only one conformer is provided for each compound, and these conformers are not necessarily at minimum energy but are low energy conformers selected from a theoretical model (for more information, see pubchem.ncbi.nlm.nih.gov/release3d.html). PubChem also provides precomputed neighboring of all 3D conformers via the 'Similar Conformers' link in Entrez. In addition, a new viewer application, PC3D, is available to view both individual conformers and overlays of similar conformers. PC3D is available both as a web application and as a downloadable executable for Windows, Macintosh and Linux platforms.

Sequence Read Archive in Entrez

In 2009, the Sequence Read Archive (SRA, see below) (10), a repository for data generated by next-generation sequencing technologies, was added to the Entrez system of databases, thereby allowing the SRA data to be searched using fielded text queries and more easily linked with related data at NCBI. Within Entrez SRA (www.ncbi.nlm.nih.gov/sra/), the data are organized into four types of records: studies (SRP accessions), experiments (SRX accessions), samples (SRS accessions) and runs (SRR accessions). Studies contain one or more experiments, each of which contains one or more runs, each of which in turn may contain data on tens of millions of individual reads. The various record types representing data from a study are all linked to one another within Entrez, allowing users to browse the data easily on the web.

dbVar—Database of genomic structural variation

In 2009, NCBI launched a new database of genomic structural variations called dbVar (www.ncbi.nlm.nih.gov/projects/dbvar/). While the site is not yet fully functional, NCBI is accepting submissions to dbVar and provides FTP access to these data. At the time of this writing, dbVar contained seven studies with >400 000 reported variants.

THE ENTREZ SEARCH AND RETRIEVAL SYSTEM

Entrez databases

Entrez (11) is an integrated database retrieval system that provides access to a diverse set of 38 databases that together contain over 400 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 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

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 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 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 found under the 'Entrez Tools' link on the NCBI home page.

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 2500 new taxa per month and indexes almost 320 000 organisms named at the genus level or lower that are represented in Entrez by at least one nucleotide or protein sequence. The Taxonomy Browser can be used to view the taxonomic position or retrieve data from any of the Entrez databases for a particular organism or group.

Table 1. The Entrez Databases (as of 8/14/2009)

Database	Records	Section within this article
Nucleotide	78 783 103	Genes and Associated Sequences
EST	62 838 170	Genes and Associated Sequences
PubChem Substance	61 056 228	Small Molecules and Bioassays
SNP	59 806 469	Genotypes and Phenotypes
GEO Profiles	42 751 725	Gene Expression
Protein	28 475 324	Genes and Associated Sequences
GSS	25 787 403	Genes and Associated Sequences
PubChem Compound	25 668 433	Small Molecules and Bioassays
PubMed	19 076 621	Literature Resources
Probe	10 187 129	Gene Expression
Gene	6 261 420	Genes and Associated Sequences
UniGene	3 645 645	Genes and Associated Sequences
PubMed Central	1 834 865	Literature Resources
NLM Catalog	1 394 522	Literature Resources
Taxonomy	525 252	Entrez Search and Retreival System
UniSTS	524 629	Genes and Associated Sequences
Protein Clusters	413 052	Genomes
3D Domains	280 897	Molecular Structure and Proteomics
Books	237 535	Literature Resources
MeSH	211 794	Literature Resources
Cancer Chromosomes	134 570	Genomes
Homologene	123 767	Genes and Associated Sequences
PopSet	101 569	Genes and Associated Sequences
Biosystems	96 559	Recent Developments
GENSAT	91 458	Gene Expression
dbGaP	62 335	Genotypes and Phenotypes
Structure	59 329	Molecular Structure and Proteomics
CDD	34 735	Molecular Structure and Proteomics
Journals	23 939	Literature Resources
GEO Datasets	21 358	Gene Expression
OMIM	20 548	Genotypes and Phenotypes
Site Search	25 070	Introduction
Genome	10 777	Genomes
SRA	6562	Recent Developments
Projects[a]	5234	Genomes
OMIA	2599	Genotypes and Phenotypes
PubChem Bioassay	1691	Small Molecules and Bioassays
Peptidome	79	Recent Developments

[a]Formerly known as Genome Project.

THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The BLAST programs (12–14) 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 [Molecular Modeling Database (MMDB)] or microarray studies [Gene Expression Omnibus (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

Table 2. Selected NCBI software available for download

Software	Available binaries	Category within this article
BLAST (stand alone)	Win, Mac, LINUX, Solaris	BLAST
BLAST (network client)	Win, Mac, LINUX, Solaris	BLAST
BLAST (web server)	Mac, LINUX, Solaris	BLAST
CD-Tree	Win, Mac	Molecular Structure and Proteomics
Cn3D	Win, Mac, LINUX, Solaris	Molecular Structure and Proteomics
PC3D	Win, Mac, LINUX	Recent Developments
e-PCR	Win, LINUX	Genes and Associated Sequences
gene2xml	Win, Mac, LINUX, Solaris	Genes and Associated Sequences
OMSSA	Win, Mac, LINUX	Molecular Structure and Proteomics
splign	LINUX, Solaris	Genes and Associated Sequences
tbl2asn	Win, Mac, LINUX, Solaris	Genomes

the 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 nonredundant set of all coding sequence translations translations from GenBank along with all RefSeq, Swiss-Prot, Protein Data Bank (PDB), Protein Information Resource (PIR) and Protein Research Foundation (PRF) proteins. Subsets of this database are also available, such as PDB or Swiss-Prot sequences, along with separate databases for sequences from patents and 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 >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 Expressed Sequence Tags (ESTs) that are mapped to the genomic sequence. The default search program for the NCBI Genomic BLAST pages is MegaBLAST (15), a faster version of standard nucleotide BLAST designed to find alignments between nearly identical sequences, typically from the same species. It is available through a separate web interface that handles batch nucleotide queries and can be used to search the rapidly growing Trace Archive as well as the standard BLAST databases. For rapid cross-species nucleotide queries, NCBI offers Discontiguous MegaBLAST, which uses a noncontiguous word match (16) 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.

LITERATURE RESOURCES

PubMed

The PubMed database now contains >19 million citations dating back to the 1860s from >21 000 life science journals. Over 10.5 million of these citations have abstracts, the earliest from the 1880s, and 10 million of these citations have links to their full-text articles. PubMed is heavily linked to other core Entrez databases, where it provides 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 articles' on the basis of computationally detected similarities using indexed Medical Subject Heading (MeSH) (17) 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

PubMed Central (PMC) (18), a digital archive of peer-reviewed journals in the life sciences, now contains over 1.8 million full-text articles, growing by 12% over the past year. More than 635 journals, including *Nucleic Acids Research*, deposit the full text of their articles in PMC. 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 a collection of over 150 online textbooks and biomedical books made available in collaboration with authors and publishers. As a separate Entrez database, the content of the Bookshelf can be searched using text queries or can be found through links from other Entrez databases, particularly PubMed, PMC, Gene and Online Mendelian Inheritance in Man (OMIM). Rather than treating each book as a whole that can be read sequentially, the Bookshelf represents the books as a collection of over 235 000 units of content, such as sections, subsections and chapters. Once within one of these content units, users can navigate to other areas of the book or search for specific content within the book.

The NLM Catalog provides bibliographic data for almost 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 almost 24 000 records, the database indexes for each journal the title abbreviation, the International Organization for Standardization abbreviation, publication data and links

to the NLM catalog and all Entrez records associated with articles from that journal.

GENES AND ASSOCIATED SEQUENCES

Databases

Entrez Gene. Entrez Gene (19) 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 CDD and other gene-related resources. Gene contains data for >5.4 million genes from over 6200 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.

UniGene and ProtEST. UniGene (20) is a system for partitioning transcript sequences (including ESTs) from GenBank into a nonredundant 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 56 animals, 43 plants and fungi and another 6 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 OMIM (21), Mouse Genome Informatics (22), Zebrafish Information Network (23), Saccharomyces Genome Database (24), Clusters of Orthologous Groups (COGs) (25) and FlyBase (26). 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.

Reference sequences. The NCBI RefSeq database (27) is a nonredundant set of curated and computationally derived sequences for transcripts, proteins and genomic regions. The number of records in the RefSeq collection has grown by 42% over the past year so that Release 36 (July 2009) contains 4.0 million nucleotide and 8.2 million protein sequences representing over 8600 organisms. RefSeq sequences can be searched and retrieved from the Entrez Nucleotide and Protein databases, 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 Genome Survey Sequence (GSS) (specified as nuccore, nucest and nucgss within the E-utilities). Entrez Nucleotide contains all GenBank sequences except those within the EST or GSS GenBank divisions. The database also contains whole genome shotgun (WGS) sequences, Third Party Annotation sequences and sequences imported from the Entrez Structure database. Conceptual translations of any coding sequences in these records are placed in the Entrez Protein database. The EST database contains all records within the EST division of GenBank, a collection of first-pass single-read cDNA sequences that include no annotated biological features. Similarly, the GSS database corresponds to the GSS division of GenBank, which contains first-pass single-read genomic sequences that rarely include annotated biological features.

Analysis tools

Open Reading Frame Finder, Spidey and Splign. 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. Spidey aligns a set of eukaryotic mRNA sequences to a single genomic sequence taking into account predicted splice sites and using one of four splice-site models (Vertebrate, *Drosophila*, *C. elegans*, Plant).

Splign (28) (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 (29) 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.

Electronic PCR. Forward electronic PCR (e-PCR) searches for matches to Sequence Tagged Site (STS) primer pairs in the UniSTS database of over 520 000 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.

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 contains over 20 100 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/.

GENOMES

Databases

Entrez Genome. Entrez Genome (30) 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 >1000 microbes and 3400 viruses, as well as for over 2000 eukaryotic organelles. For higher eukaryotes, the Genome database includes genomes for 32 species, as well as data from almost 550 WGS projects for 80 species. More than 20% of the almost 11 000 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 Projects. The Entrez Projects database, formerly known as Entrez Genome Project, provides an overview of the status of a variety of genomic projects, ranging from large-scale sequencing and assembly projects to projects focused on a particular locus, such as 16S ribosomal RNA or a notable medical event, such as the 2009 H1N1 flu outbreak. While over 90% of the >5200 projects are traditional single-organism sequencing projects, the scope of the database continues to expand so that it now includes viral population projects, metagenome and environmental sampling projects, comparative genomics projects and transcriptome projects. Entrez 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, Entrez 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.

The Trace and Assembly Archives. The Trace Archive contains over 2 billion traces (12% human) from gel and capillary electrophoresis sequencers. More than 4500 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 (SRA). The Sequence Read Archive (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. Since its inception in 2007, the SRA has accumulated over 10 Tbp of biological sequence data. The SRA is part of the Entrez search system, and SRA data sets are linked to PubMed, WGS, GEO, Database of Genotypes and Phenotypes (dbGaP) and Projects databases. 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 now 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.

Analysis tools and resources

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.

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.

Cancer Chromosomes. Cancer Chromosomes (31) 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 (32) 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.

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 precomputed 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.

Protein Clusters. The Protein Clusters database (www.ncbi.nlm.nih.gov/proteinclusters/) contains over 280 000 sets of almost identical RefSeq proteins encoded by complete prokaryotic, mitochondrial or chloroplast genomes and organized in a taxonomic hierarchy (33). 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.

Influenza genome resources. The Influenza Genome Sequencing Project (IGSP) (34) is providing researchers with a growing collection of over 40 000 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 70 000 influenza sequences in the GenBank and RefSeq databases, as well as other Entrez databases containing 113 000 influenza protein sequences, 140 influenza protein structures and 490 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).

GENEOTYPES AND PHENOTYPES

Database of Genotypes and Phenotypes (dbGaP)

The correlation of genetic and environmental factors with human disease is vital to the development of diagnostic and therapeutic techniques. Large-scale genotype studies that provide the data for such analysis run the gamut from genome-wide association surveys, medical sequencing, molecular diagnostic assays and surveys of association between genotype and nonclinical traits. Within Entrez, dbGaP (35) (www.ncbi.nlm.nih.gov/gap/) archives, distributes and supports submission of data that correlate genomic characteristics with observable traits. This database is an approved NIH repository for NIH-funded genome-wide association study results (grants.nih.gov/grants/gwas/index.htm). The dbGaP collection has grown rapidly in the past year, from ~25 studies to now >160, 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 (dbSNP)

dbSNP (36), a repository for single-base nucleotide substitutions and short deletion and insertion polymorphisms, has nearly 18 million human records and 35 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, literature citations (PubMed) and individual genotypes for clustered reference records (rs numbers). 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.

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

Database cluster for routine clinical applications: dbMHC, dbLRC, 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 (37) 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 (38) and tools for DNA probe alignments.

OMIM

NCBI provides as part of Entrez the online version of the OMIM catalog of human genes and genetic disorders authored and edited by the late Victor A. McKusick and his staff at The Johns Hopkins University (21). 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 20 500 entries, including data on over 12 900 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 (39). The database holds 2600 records containing textual information and references, as well as links to relevant records from OMIM, PubMed and Entrez Gene.

GENE EXPRESSION

GEO

GEO (40) 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' (41,42). 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 data sets, which contain entire experiments. Currently, the GEO database hosts over 13 000 experiments submitted by 6500 laboratories and comprising 330 000 samples and 23 billion individual abundance measurements for over 700 organisms.

GENSAT

GENSAT (43–45) 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 >90 000 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, SNP discovery, gene expression, gene silencing and gene mapping. The probe database now provides submission templates to simplify the process of depositing data (www.ncbi.nlm.nih.gov/genome/probe/doc/Submitting.shtml).

MOLECULAR STRUCTURE AND PROTEOMICS

Databases

The Molecular Modeling Database. The NCBI MMDB (46) contains experimentally determined coordinate sets from the Protein Data Bank (47), augmented with domain annotations and links to relevant literature, protein and nucleotide sequences, chemicals (PDB heterogens) and conserved domains in CDD (48) as well as structural neighbors computed by the VAST algorithm (49,50) on compact structural domains in the 3D Domains database. Structure record summaries retrieved by text searches display thumbnail images of structures that link to interactive views of the data in Cn3D (51), 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 Conserved Domain Architecture Retrieval Tool. The CDD (48) contains over 28 000 PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the Simple Modular Architecture Research Tool (Smart) (52), Pfam (53), TIGRFAM (54) 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 annotations 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 'Conserved Domains' link. Wherever possible, protein sequences with known 3D structures are included in CDD alignments, which can be viewed along with these structures using Cn3D. Cn3D is also equipped with advanced alignment-editing tools that use variants of PSI-BLAST and threading algorithms. The Conserved Domain Architecture Retrieval Tool 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, 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.

Analysis Tools

Blink. BLAST Link (BLink) displays precomputed BLAST alignments to similar sequences for each protein sequence in the Entrez databases. BLink can display alignment subsets limited by taxonomic criteria or database of origin, and can generate a multiple sequence alignment of the resulting sequences or launch 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) (55) analyzes MS/MS peptide spectra by searching libraries of known protein sequences, assigning significant hits an *E*-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 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.

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 (56). 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.

SMALL MOLECULES AND BIOASSAYS

PubChem (57) 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 substance information, compound structures and bioactivity data of the PubChem project. The databases hold records for 61 million substances containing 25 million unique structures. More than 920 000 of these substances have bioactivity data in at least one of the 1700 PubChem BioAssays. 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 web sites. 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 the Education link from the NCBI home page. A Site Map provides a table of NCBI resources, 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: 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*., this issue.
2. 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.
3. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.
4. 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.
5. 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.
6. Papadopoulos,J.S. and Agarwala,R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**, 1073–1079.
7. Pagon,R.A. (2006) GeneTests: an online genetic information resource for health care providers. *J. Med. Libr. Assoc*., **94**, 343–348.
8. Waggoner,D.J. and Pagon,R.A. (2009) Internet resources in medical genetics. *Curr. Protoc. Hum. Genet*., Chapter 9, Unit 9 12.
9. 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*., this issue.
10. Shumway,M. (2010) The Sequence Read Archive (SRA)–a worldwide resource. *Nucleic Acids Res*., this issue.
11. 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.
12. 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.
13. 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.
14. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res*., **34**, W6–W9.
15. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol*., **7**, 203–214.
16. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
17. Sewell,W. (1964) Medical Subject Headings in Medlars. *Bull. Med. Libr. Assoc*., **52**, 164–170.
18. Sequeira,E. (2003) PubMed Central–three years old and growing stronger. *ARL*, **228**, 5–9.
19. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **35**, D26–D31.
20. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med*., **75**, 694–698.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. Kapustin,Y., Souvorov,A., Tatusova,T. and Lipman,D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct*., **3**, 20.
29. 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.
30. Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
31. 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.
32. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat. Genet*., **15**, 417–474.
33. Klimke,W. (2009) Protein clusters. *Nucleic Acids Res*., this issue.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. Geschwind,D. (2004) GENSAT: a genomic resource for neuroscience research. *Lancet Neurol*., **3**, 82.
44. 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.
45. Heintz,N. (2004) Gene expression nervous system atlas (GENSAT). *Nat. Neurosci*., **7**, 483.
46. 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.
47. 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.
48. 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.
49. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
50. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
51. 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.
52. 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.
53. 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.
54. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
55. Geer,L.Y., Markey,S.P., Kowalak,J.A., Wagner,L., Xu,M., Maynard,D.M., Yang,X., Shi,W. and Bryant,S.H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.*, **3**, 958–964.
56. 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.
57. 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.