Published online 21 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D5–D15***
doi:10.1093/nar/gkn741

Database resources of the National Center for Biotechnology Information

Eric W. Sayers*, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Eugene Yaschenko and Jian Ye

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

Received September 16, 2008; Revised October 1, 2008; Accepted October 2, 2008

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, RefSeq, UniGene, HomoloGene, ProtEST, dbMHC, dbSNP, Cancer Chromosomes, Entrez Genomes and related tools, the Map Viewer, Model Maker, Evidence Viewer, Clusters of Orthologous Groups (COGs), Retroviral Genotyping Tools, HIV-1/Human Protein Interaction Database, 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) and the PubChem suite of small molecule databases. Augmenting many of the web applications is custom implementation of the BLAST program optimized to search specialized data sets. All of the 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 DNA Databank of Japan (DDBJ) and European Molecular Biology Laboratory (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, 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 (all of which are in the public domain) are available for download at ftp.ncbi.nih.gov.

RECENT DEVELOPMENTS

Enhancements to PubMed searches

Over the past year, NCBI has introduced several enhancements to the PubMed search interface and Abstract Plus views. A new *Advanced Search* option is available that reorganizes the current *Limits* and *Preview/Index*

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

interfaces, and adds a new window that by default allows for searching by author name, journal title and publication date. The author and journal fields also have an autocomplete feature. Text queries in PubMed are now immediately analyzed by two content sensors. One searches for citation data, such as author names, journal titles or abbreviations, publication dates and volume/issue numbers, and then displays matching citations at the top of the result set. A second sensor examines queries for drug names, and then displays links to information about the drug from a new addition to the NCBI Bookshelf, *PubMed Clinical Q&A*.

Primer-BLAST

In 2008, NCBI introduced Primer-BLAST, a new tool for designing and analyzing PCR primers. The primer design function of Primer-BLAST is based on the existing program Primer3 (2) 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.

BLAST improvements and updates

As part of the ongoing redesign to the NCBI BLAST web pages (3), a new version of the web BLAST report is now available. This new design allows the sections of the report, such as the Graphic Summary, Descriptions and Alignments, to be collapsed. Formatting and download options are also now available directly on the report. A new CSV download format allows easy imports of BLAST results into spreadsheet software. In addition, the Alignments section of the BLAST report now includes information from Entrez Gene below the definition line for each applicable database hit. This information includes the gene name, organism and the number of PubMed links to the Gene record. The BLAST tree view of results also has new features, including new distance measures and automatic collapsing of sub-trees containing hits from a single taxon. The trees can now be downloaded in Newick or Nexus formats, and can be re-rooted at any node. Finally, the NCBI BLAST services now have a new URL: blast.ncbi.nlm.nih.gov. NCBI recommends that all users switch to this URL, as it provides access to a larger number of BLAST machines and greater fault tolerance.

Entrez Gene improvements and updates

One of the practical issues of genome annotation is that genes are often positioned on contig sequences, which are small portions of a larger genomic unit such as a chromosome. These contigs are then assembled together, generally with gaps between them, to represent the full chromosome. This process directly produces coordinates of a gene on a contig, but the coordinates on the chromosome, which are of greater biological interest, must be calculated indirectly as determined by the contig assembly. To obviate such calculations when searching for genes, Entrez Gene now features a search by chromosomal region in the *Limits* tab. After selecting an organism, the feature allows users to choose a specific range in chromosomal coordinates for any chromosome, thereby returning all genes in that region. The resulting gene list may now be sorted in three ways: by gene symbol; by relevance, which gives priority to gene records matching query terms in important fields such as gene name; and by gene weight, which ranks genes by how well each has been characterized as measured by the quality of links to PubMed, the information content of the gene symbol and name and the inclusion of the gene in Homologene, Protein Clusters, Online Mendelian Inheritance in Man (OMIM) or the Bookshelf.

THE ENTREZ SEARCH AND RETRIEVAL SYSTEM

Entrez databases

Entrez (4) is an integrated database retrieval system that provides access to a diverse set of 35 databases that together contain over 350 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.

MyNCBI

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. Features on the BLAST pages similarly allow users to save BLAST parameter sets in MyNCBI so that searches can be repeated reliably.

Table 1. The Entrez databases (as of 9/30/2008)

Database	Records	Section within this article
Nucleotide	65 786 674	Genes and associated sequences
EST	56 569 180	Genes and associated sequences
SNP	51 242 511	Genotypes and phenotypes
PubChem Substance	44 576 721	Small molecules and bioassays
GEO Profiles	42 751 725	Gene expression
GSS	24 562 212	Genes and associated sequences
Protein	22 337 204	Genes and associated sequences
PubChem Compound	19 327 825	Small molecules and bioassays
PubMed	18 289 697	Literature resources
Probe	9 650 111	Gene expression
Gene	4 962 281	Genes and associated sequences
UniGene	3 488 940	Genes and associated sequences
PubMed Central	1 683 851	Literature resources
NLM Catalog	1 374 580	Literature resources
UniSTS	514 624	Genes and associated sequences
Taxonomy	460 107	Entrez search and retreival system
Protein Clusters	285 386	Genomes
3D Domains	246 719	Molecular structure and proteomics
Books	229 412	Literature resources
MeSH	205 235	Literature resources
Cancer Chromosomes	131 638	Genomes
Homologene	115 467	Genes and associated sequences
PopSet	85 977	Genes and associated sequences
GENSAT	83 553	Gene expression
Structure	53 266	Molecular structure and proteomics
dbGaP	39 617	Genotypes and phenotypes
CDD	26 660	Molecular structure and proteomics
Journals	22 762	Literature resources
OMIM	19 857	Genotypes and phenotypes
GEO Datasets	16 754	Gene expression
Genome	8792	Genomes
Site Search	4402	Introduction
Genome Project	3900	Genomes
OMIA	2577	Genotypes and phenotypes
PubChem Bioassay	1197	Small molecules and bioassays

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. The 'einfo' utility provides basic statistics on a given database, including the date last updated and lists of all search fields and available links. The utility '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 (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 found under the 'Entrez Tools' link on the NCBI home page.

Taxonomy

The NCBI taxonomy database serves as a central organizing principal 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 2200 new taxa per month and indexes almost 300 000 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 principal Entrez databases for a particular organism or group.

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
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 BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The BLAST programs (5–7) 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 (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. One variant, BLAST2Sequences (8), compares two DNA or protein sequences and produces a dot-plot representation of the alignments. 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 nonredundant 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 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.

MegaBLAST

MegaBLAST (9) is a version of BLAST designed to find alignments between nearly identical sequences, typically from the same species. It is available through a web interface that handles batch nucleotide queries and operates up to 10 times faster than standard nucleotide BLAST. MegaBLAST is the default search program for the NCBI Genomic BLAST pages, is used to search the rapidly growing Trace Archive and is available for the standard BLAST databases as well. For rapid cross-species nucleotide queries, NCBI offers Discontiguous MegaBLAST, which uses a noncontiguous word match (19) 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.

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 and sets of sequences, such as ESTs, that are mapped to the genomic sequence.

LITERATURE RESOURCES

PubMed

The PubMed database now contains more than 18 million citations dating back to the 1860s from more than 20 400 life science journals. Over 9.8 million of these citations have abstracts, the earliest from the 1880s, and 8.7 million of these abstracts 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; 10) terms and the text of titles and abstracts. The default 'AbstractPlus' display format shows the abstract of a paper along with succinct descriptions of the top five related articles, increasing the potential for the discovery of important relationships.

PubMed Central

PubMed Central (PMC; 11), a digital archive of peer reviewed journals in the life sciences, now contains over 1.6 million full-text articles, growing by 51% over the past year. More than 480 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 86 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 OMIM. Rather than treating each book as a whole that can be read sequentially, the Bookshelf represents the books as a collection of almost 230 000 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 over 1.3 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 22 000 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.

GENES AND ASSOCIATED SEQUENCES

Databases

Entrez Gene. Entrez Gene (12) provides an interface to curated sequences and descriptive information about genes with links to NCBI's Map Viewer, Evidence Viewer, Model Maker, BLAST Link (Blink), protein domains from the Conserved Domain Database (CDD) and other gene-related resources. Gene contains data for more than 4.3 million genes from some 5300 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. A tool that converts the native Gene ASN.1 format into XML, called 'gene2xml' is available at ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/gene2xml.

UniGene and ProtEST. UniGene (13) is a system for partitioning transcript sequences from GenBank, including ESTs, 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 58 animals, 43 plants and fungi and another 6 eukaryotes. The UniGene collection has been used as a source of unique sequences in the fabrication of microarrays for the large-scale study of gene expression (14). UniGene databases are updated weekly with new Expressed Sequence Tag (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 for automated detection of homologs, including paralogs and orthologs, among the genes of 20 completely sequenced eukaryotic genomes. HomoloGene reports include homology and phenotype information drawn from OMIM (15), Mouse Genome Informatics (MGI; 16), Zebrafish Information Network (ZFIN; 17), Saccharomyces Genome Database (SGD; 18), Clusters of Orthologous Groups (COG; 19) and FlyBase (20). 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 Reference Sequence (RefSeq) database (21) is a nonredundant set of curated and computationally derived sequences for transcripts, proteins and genomic regions. The RefSeq collection has grown by 40% over the past year so that Release 30 (July, 2008) contains 3.0 million nucleotide and 5.6 million protein sequences representing almost 5400 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 (TPA) 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 (http://www.ncbi.nlm.nih.gov/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, Caenorhabditis elegans;* Plant).

Splign (22; 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 (23) that accounts for splice signals in combination with a compartmentization 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 STS primer pairs in the UniSTS database of over 510 000 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. Binaries along with

the source code are available at ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR.

The Conserved CDS database. When separate research groups, each using their own methods, predict the locations of genes in the genomes of model organisms, the resulting annotations are often similar but not always identical. These differences make it difficult for researchers to relate sequence information for a gene between the various databases. Among the model organisms, the human and mouse genome sequences are now sufficiently stable so that it is feasible to identify of a set of 'consensus' gene annotations. 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 20 000 human and 17 500 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 (24) provides access to complete genomic sequences from the RefSeq collection for more than 850 microbes and 3100 viruses, as well as genomic sequences for over 1600 eukaryotic organelles. Over 700 chromosomes from almost 50 animals, green plants and fungi are also included, bringing the total to some 6200 sequences, 882 of which were added in the last 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 Project. The Entrez Genome Project database provides an overview of the status of complete and in-progress large-scale sequencing, assembly, annotation and mapping projects. Currently the database tracks over 2200 projects, of which over 750 are complete and another 700 are draft assemblies. The database continues to expand in scope so that it now includes multi-isolate sequencing projects such as viral population projects, targeted locus sequencing projects (such as 16S ribosomal RNA metagenomic sequencing studies) and transcriptome projects. Genome Project 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 Project 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 sequencing centers to register their project early in the sequencing process 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 1.9 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.

Short Read Archive. The Short Read Archive (SRA) is a repository for sequencing data generated from the new generation of sequencers, including the Roche-454 Life Sciences GS FLX, Illumina Genome Analyzer and Applied Biosystems SOLiD System platforms. Since its inception in 2007, the SRA has accumulated 1.3 Tbp of sequencing data in 18 billion reads (85% human).

The SRA offers more extensive opportunities for data mining by separating the representation of study, experiment and sample metadata from instrumentation data. Indexing of these objects will allow for the presentation of a complete stack of scientific information going from published results down to raw instrument data. Auxiliary methods and tools for searching short-read data and for representing multiple and pair-wise reference alignments continue to be developed.

Analysis tools and resources

Map Viewer. The NCBI Map Viewer displays genome assemblies, genetic and physical markers and the results of annotation and other analyses using sets of aligned maps. The Map Viewer home page (www.ncbi.nlm.nih.gov/mapview/) provides genomic data displays for over 100 organisms including *Homo sapiens, Mus musculus* and *Rattus norvegicus*. 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 (MM) is used to construct transcript models using combinations of putative exons derived from *ab initio* predictions or from the alignment of GenBank transcripts, including ESTs and RefSeqs, to the NCBI human genome assembly. The Evidence Viewer (EV) summarizes the sequence evidence supporting a gene annotation by displaying alignments of RefSeq and GenBank transcripts, along with ESTs, to genomic contigs. EV shows detailed

alignments for each exon, and highlights mismatches between the transcript and genomic sequences.

Cancer Chromosomes. Entrez Cancer Chromosomes contains data on human chromosomal aberrations, such as deletions and translocations, that 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 (25) and the NCI Recurrent Chromosome Aberrations in Cancer database. Graphical schematics of each aberration are available along with clinical case information and links to relevant literature. Also available are similarity reports that list terms common to a group of records returned by a search.

TaxPlot, GenePlot and gMap. TaxPlot simultaneously plots similarities between the proteins of two organisms and those 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 pre-computed segments of similarity.

Influenza Genome Resources. The Influenza Genome Sequencing Project (IGSP; 26) is providing researchers with a growing collection of virus sequences essential to the identification of the genetic determinants of influenza pathogenicity. To date, the project has generated over 33 000 influenza sequences. 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. Using the tools of the Influenza Virus Resource, researchers can extend their analyses to over 83 000 influenza protein sequences, more than 100 influenza protein structures and over 350 influenza population studies accessible within the biological databases in Entrez. 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).

Entrez Protein Clusters. The Entrez Protein Clusters database (www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters), contains over 280 000 sets of almost identical RefSeq proteins encoded by complete prokaryotic or chloroplast genomes and organized in a taxonomic hierarchy (27). 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. Protein Clusters are also linked to genomic neighborhoods via Genome ProtMap (http://www.ncbi.nlm.nih.gov/sutils/protmap.cgi), which maps each protein from a COG (19) or VOG (Viral Orthologous Groups; http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/vog.html) back to its genome, and displays the genomic segments coding for members of its group of related proteins.

Clusters of Orthologous Groups. The COGs database (19) presents a compilation of orthologous groups of proteins from completely sequenced organisms. A eukaryotic version, KOGs, is available for seven organisms including *H. sapiens, C. elegans, D. melanogaster* and *Arabidopsis thaliana*. Alignments of sequence from COGs have been incorporated into CDD (described subsequently) and Genome ProtMap.

GENOTYPES AND PHENOTYPES

The Database of Genotypes and Phenotypes

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. The Database of Genotypes and Phenotypes (dbGaP; www.ncbi.nlm.nih.gov/sites/entrez?db=gap) (28) is part of the Entrez system and archives, distributes and supports submission of data that correlates genomic characteristics with observable traits. This database is an approved NIH repository for NIH-funded genome-wide association study (GWAS) results (grants.nih.gov/grants/gwas/index.htm). Currently dbGaP contains data from more than 25 studies that 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 pre-computed associations between genotype and phenotype.

A database of Single Nucleotide Polymorphisms

The Database of Single Nucleotide Polymorphisms (dbSNP; 29), a repository for single-base nucleotide substitutions and short deletion and insertion polymorphisms, contains nearly 18 million human SNPs and another 33 million from a variety of other organisms. dbSNP provides additional information about the validation status, population-specific allele frequencies and individual genotypes for each dbSNP submission. These data are available on the dbSNP FTP site in XML-structured

genotype reports that include information about cell lines, pedigree IDs and error flags for genotype inconsistencies and incompatibilities.

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

dbMHC (www.ncbi.nlm.nih.gov/mhc/MHC.fcgi?cmd=init) focuses on the major histocompatibility complex (MHC) and contains data about variations found in 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 contains over a thousand sequences for MHC alleles along with allele frequency distributions as well as human leukocyte antigen (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 and integrates the Blood Group Antigen Gene Mutation Database (30) 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 (31) 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 (15). 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 almost 20 000 entries, including data on over 12 500 established gene loci and phenotypic descriptions. These records link many important resources, such as locus-specific databases and GeneTests (www.genetests.org).

OMIA

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 (32). The database holds over 2500 records containing textual information and references, as well as links to relevant records from OMIM, PubMed and Entrez Gene.

GENE EXPRESSION

Gene Expression Omnibus

The Gene Expression Omnibus (GEO; 33) 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 genome copy number variation, genome–protein interaction surveys and methylation profiling. 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; 34,35). 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 10 000 experiments submitted by 5000 laboratories and comprising 300 000 samples and 16 billion individual abundance measurements for over 500 organisms.

GENSAT

GENSAT (36–38) is a gene expression atlas of the mouse central nervous system produced with data supplied by the National Institute of Neurological Disorders and Stroke. GENSAT catalogs images of histological sections of the mouse brain in which tags, such as Enhanced Green Fluorescence Protein, have been used to visualize the relative degree of localized expression for a wide array of genes. In addition to search tools, GENSAT provides download, zoom and comparison facilities for the more than 80 000 images in the collection.

Entrez 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 some 9.6 million probe sequences of 31 types including probes for genotyping, SNP discovery, gene expression, gene silencing and gene mapping. Submission of probe reagents can be made by contacting probe-admin@ncbi.nlm.nih.gov for submission templates. A web-submission tool for probe data is currently under development.

MOLECULAR STRUCTURE AND PROTEOMICS

Databases

MMDB. The NCBI MMDB (39) contains experimentally determined coordinate sets from the Protein Data Bank (40), augmented with domain annotations and links to relevant literature, protein and nucleotide sequences, chemicals (PDB heterogens) and conserved domains in CDD (41) as well as structural neighbors computed by the VAST algorithm (42,43) 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 (44), the NCBI structure and alignment viewer. NCBI also provides pre-computed BLAST results

against the PDB database for all proteins in Entrez through the *Related Structures* link.

CDD and CDART. The CDD (41) contains over 26 000 PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the Simple Modular Architecture Research Tool (Smart; 45), Pfam (46) and from domain alignments constructed by CDD curators or derived from COGs and Entrez Protein Clusters. The NCBI Conserved Domain Search (CD-Search) service can be used to locate conserved domains within a protein sequence, and the results of this search are pre-computed for all proteins in Entrez and available 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 (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 and can also be edited or created *de novo* using a new standalone tool called CDTree (http://www.ncbi.nlm.nih.gov/Structure/cdtree/cdtree.shtml). 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. In 2008 NCBI released CD-Tree 3.1, the first version available for both Windows and Mac OS X platforms.

Analysis tools

Blink. BLink displays pre-computed 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 displayed for protein records in Entrez as well as within Entrez Gene reports.

Open Mass Spectrometry Search Algorithm. The Open Mass Spectrometry Search Algorithm (OMSSA; 47) analyzes MS/MS peptide spectra by searching libraries of known protein sequences, assigning significant hits an Expect-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' or 'refseq' 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 (48). 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/index.html. 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 (49) 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 nearly 41 million substances containing over 19 million unique structures. More than 750 000 of these substances have bioactivity data in at least one of the nearly 1200 PubChem BioAssays. The PubChem databases link not only to other Entrez databases such as PubMed and PMC 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 Handbook, available in the NCBI Bookshelf, describes 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 comprehensive 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/About/newsletter.html). In addition, a number of mailing lists provide updates on a variety of NCBI resources (www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). RSS feeds for some NCBI resources (www.ncbi.nlm.nih.gov/feed/)

are also now available, including a new RSS feed, 'ncbi-announce' that broadcasts a variety of NCBI updates.

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. (2009) GenBank. *Nucleic Acids Res.,* this issue.
2. Rozen,S. and Skalestsky,H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In Krawetz,S. and Misener,S. (eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp. 365–386.
3. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–9.
4. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Meth. enzymol.*, **266**, 141–162.
5. 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.
6. 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.
7. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–9.
8. Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.
9. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
10. Sewell,W. (1964) Medical Subject Headings in Medlars. *Bull. Med. Libr. Assoc.*, **52**, 164–170.
11. Sequeira,E. (2003) PubMed Central - three years old and growing stronger. *ARL*, **228**, 5–9.
12. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–31.
13. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
14. Ermolaeva,O., Rastogi,M., Pruitt,K.D., Schuler,G.D., Bittner,M.L., Chen,Y., Simon,R., Meltzer,P., Trent,J.M. and Boguski,M.S. (1998) Data management and analysis for gene expression arrays. *Nat. genet.*, **20**, 19–23.
15. Hamosh,A. (2009) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.,* this issue.
16. 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**, D630–637.
17. 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**, D581–585.
18. 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**, D577–581.
19. 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.
20. 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**, D486–491.
21. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2009) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.,* this issue.
22. Kapustin,Y., Souvorov,A., Tatusova,T. and Lipman,D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct*, **3**, 20.
23. 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.
24. Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
25. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat. Genet.*, **15** (Spec No), 417–474.
26. 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.
27. Klimke,W. (2009) Protein Clusters. *Nucleic Acids Res.,* this issue.
28. 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.
29. 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.
30. 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. *Human Mut.*, **23**, 8–16.
31. Helmberg,W., Dunivin,R. and Feolo,M. (2004) The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences. *Nucleic Acids Res.*, **32**, W173–175.
32. 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**, D599–601.
33. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2009) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.,* this issue.
34. 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.
35. 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.
36. Geschwind,D. (2004) GENSAT: a genomic resource for neuroscience research. *Lancet Neurol.*, **3**, 82.
37. 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.
38. Heintz,N. (2004) Gene expression nervous system atlas (GENSAT). *Nat. Neurosci.*, **7**, 483.
39. 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**, D298–300.
40. 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**, D301–303.
41. Marchler-Bauer,A., Anderson,J.B., Derbyshire,M.K., DeWeese-Scott,C., Gonzales,N.R., Gwadz,M., Hao,L., He,S., Hurwitz,D.I., Jackson,J.D. *et al.* (2009) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.,* this issue.
42. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
43. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
44. 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.
45. 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**, D257–260.
46. 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**, D247–251.
47. 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.
48. Fu,W. (2009) HIV Protein Interaction Database. *Nucleic Acids Res.,* this issue.
49. Wang,Y. (2009) PubChem BioAssay. *Nucleic Acids Res.,* this issue.