Published online 27 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D13–D21***
doi:10.1093/nar/gkm1000

Database resources of the National Center for Biotechnology Information

David L. Wheeler*, 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, Oleg Khovayko, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Martin Shumway, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L. Tatusov, Tatiana A. Tatusova, Lukas Wagner and Eugene Yaschenko

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

Received September 18, 2007; Revised October 19, 2007; Accepted October 22, 2007

ABSTRACT

In addition to maintaining the GenBank(R) 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 available through NCBI's web site. NCBI resources include Entrez, the Entrez Programming Utilities, My NCBI, PubMed, PubMed Central, Entrez Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link, Electronic PCR, OrfFinder, Spidey, Splign, RefSeq, UniGene, HomoloGene, ProtEST, dbMHC, dbSNP, Cancer Chromosomes, Entrez Genome, Genome Project and related tools, the Trace, Assembly, and Short Read Archives, the Map Viewer, Model Maker, Evidence Viewer, Clusters of Orthologous Groups, Influenza Viral Resources, HIV-1/Human Protein Interaction Database, Gene Expression Omnibus, Entrez Probe, GENSAT, Database of Genotype and Phenotype, Online Mendelian Inheritance in Man, Online Mendelian Inheritance in Animals, the Molecular Modeling Database, the Conserved Domain Database, the Conserved Domain Architecture Retrieval Tool and the PubChem suite of small molecule databases. Augmenting the web applications are custom implementations of the BLAST program optimized to search specialized data sets. 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(R) (1) nucleic acid sequence database, to which data is submitted by the scientific community, NCBI provides data retrieval systems and computational resources for the analysis of GenBank data as well as a variety of other biological data. For the purposes of this article, the NCBI suite of database resources is grouped into three broad categories; recent developments, resource highlights and a synopsis of the remaining NCBI resources. All resources discussed are available from the NCBI home page at (www.ncbi.nlm.nih.gov). In most cases, the data underlying these resources is available for download at (ftp.ncbi.nih.gov), a link from the NCBI home page.

RECENT DEVELOPMENTS

The Database of Genotype and Phenotype (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 non-clinical traits. The Database of Genotypes and Phenotypes (dbGaP) (2) (www.ncbi.nlm.nih.gov/sites/entrez?db=gap) was recently created at NCBI to archive, distribute and to

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

support submission of data that correlates genomic characteristics with observable traits. This database is an approved NIH repository for genome wide association study (GWAS) results (grants.nih.gov/grants/gwas/index.htm).

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. Summary metrics for phenotype measures and genotype frequencies, as well as 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. The results of several studies, including the National Eye Institute Age-Related Eye Disease Study (3), the NINDS Parkinsonism Study (4), the NHLBI Framingham SHARe and GAIN (2) were released by dbGaP in 2007.

New BLAST databases

Two new Basic Local Alignment Search Tool (BLAST) databases, one for human and one for mouse, were launched over the past year containing a combination of RefSeq transcript and RefSeq genomic sequences arising from NCBI annotations. Searches of the two databases generate a new, interactive tabular display that partitions the BLAST hits by sequence type—genomic or transcript—and allows sorting by BLAST score, percent of query sequence in the alignment, or percent identity within the alignment. Human and mouse 'genomic + transcript' MegaBLAST searches use a faster, indexed algorithm that typically reduces run time by two-thirds. The pre-indexed database has been filtered to eliminate matches to low-complexity and repeat sequences.

BLAST home page redesign

The BLAST homepage has been redesigned to provide easier navigation and simplified BLAST program selection. The new page highlights options for genomic searches, features automatic parameter optimization for searches with short queries and uses an auto-complete input box for specifying organism limitations. Using the new homepage, users can assign titles to their BLAST searches, review recent BLAST search results and save BLAST forms with custom parameters for indefinite periods via My NCBI. As part of the redesign, BLAST Request Ids (RIDs) have been shortened from 36 to 11 characters.

Short Read Archive

The past year has seen a massive increase in sequencing data generated from a new generation of sequencers, including those from Roche-454 Life Sciences, Illumina Solexa and Applied Biosystems SOLiD. This motivated development of the Short Read Archive (SRA) to accommodate deposits from sequencing experiments using these platforms. The SRA recently entered service and currently holds data from 44 studies.

The SRA offers more extensive associations than can be tracked within the Entrez system by separating the representation of study, experiment and sample parameters from actual instrument data. Indexing of these objects will allow for the presentation of a complete pipeline of scientific results going from instrument data all the way through publication. Auxiliary tools for searching short-read data and for visualizing multiple and pair-wise reference alignments are expected to appear in the coming year.

Entrez Nucleotide database is split to become CoreNucleotide, EST and GSS

An important change in Entrez over the past year is the split of the Nucleotide database into three subset databases called 'CoreNucleotide', 'EST' and 'GSS' (specified as 'nuccore', 'nucest' and 'nucgss', respectively, within the E-Utilities). The CoreNucleotide database contains records for all Entrez Nucleotide sequences that are not found within the Expressed Sequence Tag (EST) or Genome Survey Sequence (GSS) divisions of GenBank. These include sequences from all remaining divisions of GenBank, NCBI Reference Sequences (RefSeqs), Whole Genome Shotgun (WGS) sequences, Third Party Annotation (TPA) sequences and sequences imported from the Entrez Structure database. The EST database contains all records found within the EST division of GenBank. EST records contain first-pass single-read cDNA sequences and include no annotated biological features. The GSS database contains all records found within the GSS division of GenBank. GSS records contain first-pass single-read genomic sequences and rarely include annotated biological features. The partitioning of the Nucleotide database makes it easier for researchers to focus on the segment of interest by separating the most richly annotated sequences from those that are sparsely annotated. During a transition period, searches of the Nucleotide database on the Web will return links to search results in the three subsets. However, the Nucleotide database on the web will eventually be phased out entirely in favor of the three subset-databases. The Nucleotide database will be retained for E-Utility use.

Protein Clusters

The new Entrez Protein Clusters database (www.ncbi.nlm.nih.gov/sites/entrez?db = proteinclusters), contains over 222 000 sets of almost identical RefSeq proteins encoded by complete prokaryotic or chloroplast genomes and organized in a taxonomic hierarchy. These clusters are used as a basis for genome-wide comparison at NCBI as well as to provide simplified BLAST access, via Concise Microbial Protein BLAST (www.ncbi.nlm.nih.gov/genomes/prokhits.cgi). Protein Clusters provides annotation information, publications, domains, structures and external links and analysis tools including multiple alignments. Protein Clusters are also linked to genomic neighborhoods via Genome ProtMap

(www.ncbi.nlm.nih.gov/sutils/protmap.cgi?), which maps each protein from a COG (5) or VOG (Viral Orthologous Groups) (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.

HIGHLIGHTED RESOURCES

PubChem

PubChem is the informatics backbone for the NIH Roadmap Initiative on molecular libraries and focuses on the chemical, structural and biological properties of small molecules, particularly their application 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 comprise records for over 19.6 million compounds with over 11 million unique structures. 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, in addition to text queries, structural queries based on chemical Smiles, formulas or 3D 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.

Gene Expression Omnibus (GEO)

GEO (6) is a data repository and retrieval system for microarray and other forms of high-throughput molecular abundance data generated by the scientific community. In addition to gene expression data, GEO accepts array comparative genomic hybridization (aCGH) data, chromatin immunoprecipitation on array (ChIP-chip) data, SNP array data and some proteomic data types. The GEO repository accepts Minimum Information About a Microarray Experiment (MIAME)-compliant data submissions. Several data deposit options and formats are supported, including web forms, spreadsheets, XML and Simple Omnibus Format in Text (SOFT). Data may be queried and visualized from both experiment (Entrez GEO DataSets) and gene-centric (Entrez GEO Profiles) perspectives. At the time of writing, the repository contains data from over 200 000 hybridization experiments, representing ~10 billion individual measurements, derived from about 4000 array definitions, and spanning over 400 organisms.

Influenza Genome Resources

The Influenza Genome Sequencing Project (IGSP) (7) 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 almost 24 000 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 almost 50 000 influenza sequences in GenBank and NCBI's RefSeq database. Using the tools of the Influenza Virus Resource, researchers can extend their analyses to the 56 000 influenza protein sequences, 111 influenza protein structures, and 269 influenza population studies accessible within the biological databases covered by NCBI's Entrez system. 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' (8).

The Conserved CDS database (CCDS)

Model organism gene predictions made by various groups using different methods result in annotations that are similar but not always identical. These differences often make it difficult for researchers to relate sequence information obtained for a gene in one database with information in another. Among the model organisms, the human and mouse genome sequences are now sufficiently stable that the identification of a set of 'consensus' gene annotations is feasible. The 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 some 18 000 human and 13 000 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 is available at (ftp.ncbi.nlm.nih.gov/pub/CCDS/).

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

dbMHC (www.ncbi.nlm.nih.gov/mhc/MHC.fcgi? cmd = init) focuses on the Major Histocompatibility Complex (MHC) and contains information and data about variations found in alleles of the MHC, a highly variable array of genes playing a vital role in the success of organ transplants and susceptibility to infectious diseases. dbMHC contains over a thousand sequences for MHC alleles and data on allele frequency distributions as well as data from a project to collect 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 (9) 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, and analysis tools to interpret homozygous or heterozygous sequencing results (10) and tools for DNA probe alignments.

SYNOPSIS OF NCBI REMAINING RESOURCES

Database retrieval tools

Entrez, My NCBI and the Entrez Programming Utilities. Entrez (11) is an integrated database retrieval system that supports text searching, using simple Boolean queries, of a diverse set of 35 databases that together comprise well over a quarter of a billion records. 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.

'My NCBI' allows users to store personal configuration options such as search filters, LinkOut preferences and document delivery providers. My NCBI also saves searches and can automatically e-mail updated search results. A My NCBI feature called 'Collections' allows users to save search results and bibliographies indefinitely. BLAST parameter sets may also be saved indefinitely using an option on the newly redesigned BLAST pages, described subsequently.

Scripted access to Entrez is provided by the Entrez Programming Utilities (E-Utilities), a suite of eight server-side programs supporting a uniform set of parameters used to search, link between and download from, the Entrez databases. The 'einfo' utility can be used to retrieve lists of supported databases and search fields. The 'egquery' utility returns the number of matches to a query in each Entrez database. E-Utilities such as 'efetch' or 'esummary', are used to retrieve full records or summaries, respectively. Espell checks spelling within Entrez queries and offers corrections. A Simple Object Access Protocol (SOAP) interface to the E-Utilities is supported. Instructions for using the E-Utilities are found under the 'Entrez Tools' link on the NCBI home page.

PubMed and PubMed Central. The PubMed database passed a milestone over the past year, indexing its 17 millionth citation and providing full-text links to some 8.7 million articles. PubMed covers more than 19 600 life science journals for biomedical articles back to the 1950s, most with abstracts and many with links to the full-text article. 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) (12) terms and the text of titles and abstracts. The default 'AbstractPlus' display format shows, in addition to the abstract of a paper, succinct descriptions of the top five related articles, increasing the potential for the discovery of important relationships.

PubMed Central (PMC) (13), a digital archive of peer reviewed journals in the life sciences, also recently passed a milestone by adding its 1 millionth full-text article, growing by 47% over the past year. More than 340 journals, including *Nucleic Acids Research*, deposit the full text of their articles in PMC. Participation in PMC requires a commitment to free access to full text, either immediately after publication or within a 12-month period. All PMC free articles are identified in PubMed search results and PMC itself can be searched using Entrez.

Taxonomy. The NCBI taxonomy database, growing at the rate of 1700 new taxa a month, indexes over 260 000 named organisms that are represented in the databases with at least one nucleotide or protein sequence. The Taxonomy Browser can be used to view the taxonomic position or retrieve data from any of the principal Entrez databases for a particular organism or group.

THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The BLAST programs (14–16) perform sequence-similarity searches against a variety of databases, returning a set of gapped alignments with links to full database records, to UniGene, Gene, the MMDB or GEO. One variant, BLAST2Sequences (17), 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/).

BLAST output formats

Standard output formats include the default pairwise alignment, several query-anchored multiple sequence alignment formats, an easily-parsable Hit Table and a taxonomically organized output. 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 (18), designed to find nearly exact matches, 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 NCBI's 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 non-contiguous 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 76 organisms shown in the Map Viewer. Genomic BLAST may be used to search the genomic sequence of an organism, the nucleotide and protein RefSeqs used in, and resulting from, the annotation of the genomic sequence, or sets of sequences, such as ESTs, that are mapped to the genomic sequence.

RESOURCES FOR GENE-LEVEL SEQUENCES

Databases

Entrez Gene. Entrez Gene (20) provides an interface to curated sequences and descriptive information about genes with links to NCBI's Map Viewer, Evidence Viewer, Model Maker, BLink, protein domains from the Conserved Domain Database (CDD), and other gene-related resources. Gene contains data for more than 3.2 million genes from some 4500 organisms. Data is 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 for several popular computer platforms at: (ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/gene2xml).

UniGene and ProtEST. UniGene (21) is a system for partitioning GenBank sequences, including ESTs, into a non-redundant set of gene-oriented clusters. UniGene clusters are created for all organisms for which there are 70 000 or more ESTs in GenBank and includes ESTs for some 44 animals and another 41 plants and fungi. 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 (22). UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences.

ProtEST, tightly coupled to UniGene, presents precomputed BLAST alignments between protein sequences from model organisms and the 6-frame translations of nucleotide sequences in UniGene.

HomoloGene. HomoloGene is a system for automated detection of homologs among the genes of 18 completely sequenced eukaryotic genomes including those of *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Anopheles gambiae*, *Caenorhabditis elegans*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Eremothecium gossypii*, *Neurospora crassa*, *Magnaporthe grisea*, *Arabidopsis thaliana* and *Oryza sativa*. HomoloGene entries include paralogs in addition to orthologs. HomoloGene reports include homology and phenotype information drawn from Online Mendelian Inheritance in Man (OMIM) (23), Mouse Genome Informatics (MGI) (24), Zebrafish Information Network (ZFIN) (25), Saccharomyces Genome Database (SGD) (26), Clusters of Orthologous Groups (COG) (5) and FlyBase (27). The new HomoloGene Downloader, appearing under the 'Download' link in HomoloGene displays, allows the retrieval of any or all 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.

A database of Single Nucleotide Polymorphisms (dbSNP). The dbSNP (28), a repository for single-base nucleotide substitutions and short deletion and insertion polymorphisms, contains over 12 million human SNPs and another 39 million from a variety of other organisms, with 17 million of these added over the past year. The dbSNP database provides additional information about the validation status, population-specific allele frequencies and individual genotypes for dbSNP submission. These data are available on the dbSNP FTP site in XML-structured genotype reports that include information about cell lines, pedigree IDs and error flags for genotype inconsistencies and incompatibilities.

Reference Sequences. The RefSeq database (29), provides curated references for transcripts, proteins and genomic regions, plus computationally derived nucleotide sequences and proteins. The complete RefSeq database is provided in the RefSeq directory on the NCBI FTP site. The number of sequences in RefSeq has grown by 33% over the past year. As of Release 24, RefSeq contained over 6.1 million sequences, including more than 3.9 million protein sequences, representing 4500 organisms.

Tools for gene-level analysis

Open Reading Frame (ORF) Finder, Splign and Spidey. ORF Finder performs a six-frame translation of a nucleotide sequence and returns the location of each ORF within a specified size range.

Splign (30) 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 (31) that accounts for splice signals in combination with a compartmentization algorithm to identify possible locations of genes and their copies. The Splign Web site can be found at (www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi). A link is

provided to download a standalone version that is instrumental for large-scale processing.

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).

Electronic PCR (e-PCR). Forward e-PCR searches for matches to STS primer pairs in the UniSTS database of almost 500 600 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 for several computer platforms, along with the source code, are available at (ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR).

RESOURCES FOR GENOME-SCALE ANALYSIS

Databases for genomic analysis

Entrez Genome. Entrez Genome (32) provides access to over 570 complete microbial genomic sequences (200 added over the past year), more than 2840 viral genomic sequences (390 added) and over 1300 RefSeqs for eukaryotic organelles (300 added). Over 25 higher eukaryotic genomes are also included, such as the recent arrival, *Equus caballus*, the horse. 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. Specialized viewers and BLAST pages are also available for eukaryotic organelles and viruses.

The Trace and Assembly Archives. The Trace Archive is a rapidly growing database of over 1.8 billion sequencing traces. More than 4400 organisms are represented, an increase of 3600 over the past year. The Assembly Archive links the raw sequence information found in the Trace Archive with assembly information found in GenBank. An Assembly Viewer allows displays of multiple sequence alignments as well as the sequence chromatograms for traces that are part of assemblies.

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. Genome Project links to project data in the other Entrez databases, such as the Entrez Nucleotide databases 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; environmental requirements, such as salinity, temperature and pH range; oxygen requirements and pathogenicity. The database allows genome-sequencing centers to register their project early in the sequencing process so that project data can be linked to other NCBI-hosted data at the earliest opportunity.

Other Resources for Genomic Analysis

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 links to both Map Viewer and Genomic BLAST pages for some 76 organisms including *H. sapiens*, *M. musculus* and *R. norvegicus*. Maps available for display in the Map Viewer 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 has the ability to show previous genome builds. The Map Viewer can generate a tabular display for convenient export to other programs and segments of a genomic assembly may be downloaded using a Download/View Sequence link.

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) displays the alignments to genomic contigs of RefSeq and GenBank transcripts, and ESTs supporting gene models. Mismatches between transcript and genomic sequences are highlighted. Both MM and the EV have been extended to cover many new organisms over the past year.

Cancer Chromosomes. 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 (33), and the NCI Recurrent Chromosome Aberrations in Cancer databases comprise the Cancer Chromosomes Entrez database. Simple and advanced interfaces are offered for searches and 'similarity reports' can be generated, showing terms common to a group or records returend by a search.

TaxPlot, GenePlot and gMap. TaxPlot plots similarities in the proteomes of two organisms to that of a reference organism for more than 700 prokaryotic and almost 45 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.

Clusters of Orthologous Groups. The COGs database (5), 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 *A. thaliana*. Alignments of sequence from COGs have been

incorporated into the CDD and Genome ProtMap, both described subsequently.

RESOURCES FOR THE ANALYSIS OF PATTERNS OF GENE EXPRESSION

Resources for the analysis of gene expression

GENSAT. GENSAT 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.

Probe. The Entrez Probe database archives some 8.8 million probe sequences of 68 types, along with data on their experimental utility. Probe entries indicate the intended experimental application and include the experimental results generated using the probe.

RESOURCES SUPPORTING CORRELATIONS BETWEEN GENOTYPES AND PHENOTYPES

OMIM

NCBI provides the online version of the OMIM catalog of human genes and genetic disorders authored and edited by Victor A. McKusick at The Johns Hopkins University (23). The database contains information on disease phenotypes and genes, including extensive descriptions, gene names, inheritance patterns, map locations, gene polymorphisms and detailed bibliographies. The OMIM Entrez database contains about 18 000 entries, including data on over 12 000 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, authored by Professor Frank Nicholas of the University of Sydney, Australia and colleagues. The database contains textual information and references, as well as links to relevant records from OMIM, PubMed and Entrez Gene.

RESOURCES FOR MOLECULAR STRUCTURE AND PROTEOMICS

Structure databases

The Molecular Modeling Database. The NCBI Molecular Modeling Database (MMDB), built by processing entries from the Protein Data Bank (34), is described in Ref. (35). The structures in the MMDB are linked to sequences in Entrez and to the (36) CDD. Results lists generated by searches of the Structure database now display thumbnail images of structures. Clicking on a thumbnail launches Cn3D, described subsequently, to allow interactive viewing of the structure.

The CDD and CDART. The CDD contains over 23 500 PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the Simple Modular Architecture Research Tool (Smart) (37), Pfam (38), and from domain alignments derived from COGs. NCBI's Conserved Domain Search (CD-Search) service can be used to search a protein sequence for conserved domains in the CDD. Wherever possible CDD hits are linked to structures which, coupled with a multiple sequence alignment of representatives of the domain hit, can be viewed with NCBI's 3D molecular structure viewer, Cn3D (39) (www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml), equipped with advanced alignment-building tools that use the 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 or edited using a new standalone tool called CDTree (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 building hierarchies of alignment-based protein domains.

Tools supporting Proteomics

BLink. BLAST Link (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, by database of origin, relation to a complete genome, membership in a COG (5) or by relation to a 3D structure or conserved protein domain. BLink links are displayed for protein records in Entrez as well as within Entrez Gene reports.

The Open Mass Spectrometry Search Algorithm (OMSSA). OMSSA (40) 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 for several popular computer platforms 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. 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.

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 Books database, describes the principal NCBI resources in detail. Several tutorials are also offered under the Education link from NCBI's home page. A Site Map provides a comprehensive table of NCBI resources, and the About NCBI feature provides bioinformatics primers and other supplementary information. A user-support staff is available to answer questions at (info@ncbi.nlm.nih.gov). 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 including announcements of upcoming NCBI training courses.

ACKNOWLEDGEMENT

Funding to pay the Open Access publication charges for this article was provided by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, Database issue, in press.
2. Group,T.G.C.R., Manolio,T.A., Rodriguez,L.L., Brooks,L., Abecasis,G., the International Multi-Center ADHD Genetics Project, Ballinger,D., Daly,M., Donnelly,P. *et al.* (2007) New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.*, **39**, 1045–1051.
3. The Age-Related Eye Disease Study Research Group. (1999) The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. *Control Clin. Trials*, **20**, 573–600.
4. Fung,H.C., Scholz,S., Matarin,M., Simón-Sánchez,J., Hernandez,D., Britton,A., Gibbs,J.R., Langefeld,C., Stiegert,M.L. *et al.* (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.*, **5**, 911–916.
5. 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. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41–41.
6. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A. and Tomashevsky,M. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**(Database issue), 760–765.
7. Ghedin,E., Sengamalay,N.A., Shumway,M., Zaborsky,J., Feldblyum,T., Subbu,V., Spiro,D.J., Sitz,J., Koo,H. *et al.* (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**, 1162–1166.
8. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) GenBank. *Nucleic Acids Res.*, **35**(Database issue), 21–25.
9. 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.
10. Helmberg,W., Dunivin,R. and Feolo,M. (2004) The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences. *Nucleic Acids Res.*, **32**(Web Server issue), 173–175.
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. Sewell,W. (1964) Medical subject headings in MEDLARS. *Bull. Med. Libr. Assoc.*, **52**, 164–170.
13. Sequeira,E. (2003) PubMed Central – three years old and growing stronger. *ARL*, **228**, 5–9.
14. 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.
15. Altschul,S.F., Madden,T.L., Schäffer,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.
16. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**(Web Server issue), 6–9.
17. 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.
18. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
19. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
20. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T., (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, Database issue, in press.
21. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
22. Ermolaeva,O., Rastogi,M., Pruitt,K.D., Schuler,G.D., Bittner,M.L., Chen,Y., Simon,R., Meltzer,P., Trent,J.M *et al.* (1998) Data management and analysis for gene expression arrays. *Nat. Genet.*, **20**, 19–23.
23. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**(Database issue), 514–517.
24. Blake,J.A., Eppig,J.T., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2006) The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res.*, **34**(Database issue), 562–567.
25. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P. *et al.* (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**(Database issue), 581–585.
26. Nash,R., Weng,S., Hitz,B., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2007) Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res.*, **35**(Database issue), 468–471.
27. 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**(Database issue), 486–491.

28. 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.
29. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**(Database issue), 61–65.
30. Kapustin,Y. and Souvorov,A.T.T. (2004) Splign – a hybrid approach to spliced alignments. In. *RECOMB 2004 – Currents in Computational Molecular Biology,* Association for Computing Machinery, New York, p. 741.
31. 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.
32. Tatusova,T.A., Karsch-Mizrachi,I. and Ostell,J.A. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
33. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat. Genet.*, **15**, 417–474.
34. 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**(Database issue), 301–303.
35. Wang,Y., Addess,K.J., Chen,J., Geer,L.Y., He,J., He,S., Lu,S., Madej,T., Marchler-Bauer,A. *et al.* (2007) MMDB: annotating protein sequences with Entrez's 3D-structure database. *Nucleic Acids Res.*, **35**(Database issue), 298–300.
36. Marchler-Bauer,A., Anderson,J.B., Derbyshire,M.K., DeWeese-Scott,C., Gonzales,N.R., Gwadz,M., Hao,L., He,S., Hurwitz,D.I. *et al.* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.*, **35**(Database issue), 237–240.
37. 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**(Database issue), 257–260.
38. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**(Database issue), 247–251.
39. 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.
40. 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.