Nucleic Acids Research, 2004, Vol. 32, Database issue ***D35–D40***
DOI: 10.1093/nar/gkh073

Database resources of the National Center for Biotechnology Information: update

David L. Wheeler*, Deanna M. Church, Ron Edgar, Scott Federhen, Wolfgang Helmberg, Thomas L. Madden, Joan U. Pontius, Gregory D. Schuler, Lynn M. Schriml, Edwin Sequeira, Tugba O. Suzek, Tatiana A. Tatusova and Lukas Wagner

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

Received September 17, 2003; Revised and Accepted September 30, 2003

ABSTRACT

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

INTRODUCTION

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

DATABASE RETRIEVAL TOOLS

Entrez

Entrez (2) is an integrated database retrieval system that enables text searching, using simple Boolean queries, of a diverse set of 20 databases, several added during the past year. These databases include DNA and protein sequences derived from several sources (1,3–6), the NCBI taxonomy, genomes, population sets, gene expression data, gene-oriented sequence clusters in UniGene, sequence-tagged sites in UniSTS, genetic variations in dbSNP, protein structures from the Molecular Modeling Database (MMDB) (7), 3D and alignment-based protein domains, and the biomedical literature via PubMed, PubmedCentral, Online Mendelian Inheritance in Man (OMIM) and online Books. PubMed includes primarily the 12.7 million references and abstracts in MEDLINE(R), with links to the full text of more than 4000 journals available on the web. The Books database contains more than 25 online scientific textbooks including the NCBI Handbook, a comprehensive guide to NCBI resources. Recently, the NCBI website itself has been added to the list of Entrez databases, allowing users to employ the Entrez search engine to quickly find NCBI web pages of interest.

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

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

one record to another, a new 'Links' pull-down menu now appears in the top right-hand corner of Entrez displays.

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

Access to Entrez via automated systems is facilitated using the new Entrez Programming Utilities, a suite of five server-side scripts which support a uniform set of parameters used to search, link between and download from, the Entrez databases. A search history, available via interactive Entrez as well as via the Entrez Programming Utilities, allows users to recall the results of previous searches during an Entrez session and combine them using Boolean logic.

PubMed Central

PubMed Central (PMC) (8) is a digital archive of peer-reviewed journals in the life sciences. Over 130 journals, including *Nucleic Acids Research*, deposit the full text of their articles in PMC. Participation in PMC requires a commitment to free access to full text, perhaps with some delay after publication. Some journals provide free access to their full text directly in PMC while others require a link to the journal's own site where full text is generally available free within 6 months to a year of publication. All PMC free articles are identified in PubMed search results and PMC itself is now searchable using Entrez.

Taxonomy

The NCBI taxonomy database indexes over 150 000 organisms that are represented in the databases with at least one nucleotide or protein sequence. The Taxonomy Browser can be used to view the taxonomic position or retrieve data from any of the principal Entrez databases for a particular organism or group. The Taxonomy Browser also displays links to Map Viewer, Genomic BLAST services, the Trace Archive, and to model organism and taxonomic databases via LinkOut.

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

LocusLink

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

THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The Basic Local Alignment Search Tool (BLAST) programs (9,10) perform sequence-similarity searches against a variety of sequence databases, returning a set of gapped alignments between the query and database sequences, and links to full database records, to UniGene, LocusLink, the MMDB or GEO. Sequences appearing in a BLAST alignment may be selected for bulk download. A BLAST variant, BLAST2Sequences (11), compares two DNA or protein sequences and produces a dot-plot representation of the alignments.

Each alignment returned by a BLAST search receives a score and a measure of statistical significance, called the Expectation Value (E-value), for judging its quality. Either an E-value threshold or a range can be specified to limit the alignments returned. BLAST takes into account the amino acid composition of the query sequence in its estimation of statistical significance. This composition-based statistical treatment, used in conventional protein BLAST searches as well as PSI-BLAST (10) searches, tends to reduce the number of false-positive database hits (12).

BLAST offers several output formats including the default 'pairwise' alignment, several 'query-anchored' multiple sequence alignment formats and a tabular 'Hit Table', which serves as an easily parsed summary of the BLAST results. In addition, BLAST can generate a taxonomically organized output that shows the distribution of BLAST hits by organism. The web BLAST interface allows both the initial search and the results displayed to be restricted to a database subset using standard Entrez search syntax. Web BLAST uses a standard URL-API that allows complete search specifications, including BLAST parameters, such as Entrez restrictions and the search query, to be contained in a URL posted to the web page.

A BLAST variant designed to search for nearly exact matches, called MegaBLAST (13), offers a web interface that handles batch nucleotide queries and operates up to 10 times more quickly than standard nucleotide BLAST. MegaBLAST is the default search program for NCBI's Genomic BLAST pages that search a set of genome-specific databases and generate, where possible, genomic views of the BLAST hits using the Map Viewer. MegaBLAST is also used to search the rapidly growing Trace Archive but is available for the standard BLAST databases as well. For rapid cross-species nucleotide queries of the Trace Archive as well as the standard BLAST databases, NCBI offers Discontiguous MegaBLAST, which uses a non-contiguous word match (14) 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.

BLink

BLAST Link (BLink) displays pre-computed protein BLAST alignments for each protein sequence in the Entrez databases. BLink can display subsets of these alignments by taxonomic

criteria, by database of origin, relation to a complete genome, membership in a COGs (15) or by relation to a 3D structure or conserved protein domain. BLink links are displayed for protein records in Entrez as well as within LocusLink reports.

RESOURCES FOR GENE-LEVEL SEQUENCES

UniGene

UniGene (16), is a system for automatically 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 now includes ESTs from 16 animals and 13 plants. Each UniGene cluster contains sequences that represent a unique gene, and is linked to related information, such as the tissue types in which the gene is expressed, model organism protein similarities, the LocusLink report for the gene and its map location. In the human UniGene June 2003 release (build 161), over 5.5 million human ESTs in GenBank have been reduced 50-fold in number to ~108 000 sequence clusters. The UniGene collection has been used as a source of unique sequences for the fabrication of microarrays for the large-scale study of gene expression (17). UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences.

ProtEST

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

The Trace Archive

A newly redesigned Trace Archive interface allows for more flexible searching and download of sequencing traces from a rapidly growing database of over 260 million whole-genome shotgun (WGS), shotgun, EST, clone end and finishing reads from more than 100 organisms.

HomoloGene

HomoloGene is a database of both curated and calculated gene orthologs and homologs and now covers 21 organisms. Curated orthologs include gene pairs from the Mouse Genome Database (MGD) at the Jackson Laboratory, the Zebrafish Information (ZFIN) database at the University of Oregon and from published reports. Computed orthologs and homologs, which are considered putative, are identified from BLAST nucleotide sequence comparisons between all UniGene clusters for each pair of organisms. The HomoloGene database can be queried using UniGene ClusterIDs, LocusLink LocusIDs, gene symbols, gene names and nucleotide accession numbers, as well as those terms found in UniGene cluster titles.

dbMHC

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

The Clinical section contains data generated by the 13th international HLA workshop and international HLA working group and includes sections presenting the results of the Anthropology project with global HLA allele frequencies and the human stem cell transplantation project.

Reference Sequence (RefSeq)

The References Sequence (RefSeq) database (6), provides curated references for transcripts, proteins and genomic regions, plus computationally derived nucleotide sequences and proteins. The complete RefSeq database is now being provided in the RefSeq directory on the NCBI FTP site. The first release contains over 1 million sequences, including more than 785 000 protein sequences, from about 2000 organisms. To register for the 'refseq-announce' mailing list and be informed of new releases or to read more about the RefSeq project, visit the RefSeq home page.

Specialized tools: Open Reading Frame Finder, Spidey and Electronic PCR

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

Spidey is an alignment tool for eukaryotic genomic sequences that takes as input a set of mRNA accessions or FASTA sequences and aligns each to a single genomic sequence. Spidey takes into account predicted splice sites in constructing its alignments and can use one of four splice-site models (vertebrate, *Drosophila*, *Caenorhabditis elegans*, plant). Spidey returns exon alignments, protein translations and a summary showing the alignment quality and goodness of match to splice junction patterns for each putative exon.

Electronic PCR (e-PCR) locates Sequence Tagged Sites (STSs) within nucleotide sequences by searching against a non-redundant database of over 155 000 human and 92 000 non-human STSs called UniSTS. OrfFinder, Spidey and e-PCR are available via the 'Tools' link on the NCBI home page.

A database of single nucleotide polymorphisms

The database of single nucleotide polymorphisms (dbSNP) (18) is a repository for single base nucleotide substitutions and

short deletion and insertion polymorphisms that contains almost 6 million human SNPs as well as about 1.4 million from a variety of other organisms. Now an Entrez database, dbSNP can be queried from the NCBI home page. Searches for SNPs lying between two markers and batch downloads via Entrez are supported. SNP reports link to 3D visualizations of structures from the MMDB via NCBI's interactive macromolecular viewer Cn3D (19), which highlight amino acid changes implied by SNPs in coding regions.

RESOURCES FOR GENOME-SCALE ANALYSIS

Entrez Genomes

Entrez Genomes (20) provides access to genomic data contributed by the scientific community for species whose sequencing and mapping is complete or in progress. Entrez Genomes now includes over 140 complete microbial genomes, more than 1500 viruses, and over 425 reference sequences for eukaryotic organelles. Higher eukaryotic genomes are also included within Entrez Genomes such as the recent arrival, *Ciona intestinalis*. The Plant Genomes Central web page serves as a focal point for access to completed plant genomes, to information on plant genome sequencing projects or to plant-related resources at NCBI such as plant Genomic BLAST pages or Map Viewer.

Complete genomes can be accessed hierarchically starting from either an alphabetical listing or a phylogenetic tree for each of six principle taxonomic groups. One can follow the hierarchy to a graphical overview for the genome of a single organism, on to the level of a single chromosome and, finally, down to the level of a single gene. At the level of a genome or a chromosome, a Coding Regions view displays the location of each coding region, length of the product, GenBank identification number for the protein sequence and name of the protein product. An RNA Genes view lists the location and gene names for ribosomal and transfer RNA genes. At the level of a single gene, links are provided to pre-computed sequence neighbors for the implied protein with links to the COGs database if possible. A summary of COG functional groups is presented in both tabular and graphical formats at the genome level.

For complete microbial genomes, pre-computed BLAST neighbors for protein sequences, including their taxonomic distribution and links to 3D structures, are given in TaxTables and PDBTables, respectively. Pairwise sequence alignments are presented graphically and linked to the Cn3D macromolecular viewer (19), which provides interactive display of 3D structures and sequence alignments. The TaxPlot tool plots similarities in the proteomes of two organisms to that of a third, reference organism, and is available for both prokaryotic and eukaryotic genomes. Resources for the genomes of higher eukaryotes are discussed below.

Clusters of Orthologous Groups (COGs)

The COGs database (15), presents a compilation of orthologous groups of proteins from 66 completely sequenced organisms. A eukaryotic version, KOGs, is now available for seven eukaryotes including *Homo sapiens*, *C.elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana*. Alignments of sequence from COGS have been incorporated into the Conserved Domain Database described below.

Retroviral genotyping tools and the SARS Coronavirus resource

NCBI offers a web-based genotyping tool that employs a blastn comparison between a retroviral sequence to be subtyped and either a default panel of reference sequences or a panel provided by the user. An HIV-1-specific subtyping tool uses a set of reference sequences taken from the principle HIV-1 variants. The new SARS Coronavirus resource serves as a collection point for SARS-related information and provides the results of pre-computed analyses of the SARS genome, including continuously updated viral genomic alignments and alignments between predicted SARS proteins and other proteins of known 3D structure.

EUKARYOTIC GENOMIC RESOURCES

Map Viewer

The NCBI Map Viewer displays genome assemblies using sets of aligned chromosomal maps. A new Map Viewer home page organizes the available organisms by taxonomic group and provides links to both Map Viewer and Genomic BLAST pages. Map Viewer displays are available for the genomes of 19 organisms including *H.sapiens*, *Mus musculus* and *Rattus norvegicus*. The genomic maps displayed by the Map Viewer vary according to the data available for the subject organism and are selected from a set of cytogenetic maps, physical maps, maps showing predicted gene models, EST alignments with links to UniGene clusters and mRNA alignments used to construct gene models. Map Viewer displays link to related resources such as LocusLink, or tools such as the Evidence Viewer and Model Maker. Map Viewer can generate a tabular view of the current display that is convenient for export to other programs. Segments of a genomic assembly may be downloaded using the Map Viewer's 'Download/View Sequence' link in either GenBank or FASTA format.

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

Model Maker

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

Previously observed exon splice patterns are indicated as guides to model building. Completed models may be saved locally or analyzed with OrfFinder.

Evidence Viewer

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

The Cancer Chromosome Aberration Project (CCAP)

The CCAP service is an initiative of the National Cancer Institute (NCI) and the NCBI. The data includes a compilation by F. Mitelman, F. Mertens and B. Johansson of recurrent neoplasia-associated chromosomal aberrations from the Cancer Chromosome Aberration Bank at the University of Lund, Sweden (21). The Spectral Karyotyping database, SKY, created jointly by the NCI and the NCBI, enables investigators to share their own SKY and Comparative Genomic Hybidization (CGH) data on chromosomal aberrations (http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi).

RESOURCES FOR THE ANALYSIS OF PATTERNS OF GENE EXPRESSION AND PHENOTYPES

SAGEmap

NCBI's SAGEmap (22) provides two-way mapping between SAGE tags and UniGene clusters. SAGEmap can also construct a user-configurable table of data comparing one group of SAGE libraries with another. SAGEmap is updated weekly, immediately following the update of UniGene and the data appear in the human genome Map Viewer as the SAGE track.

Gene Expression Omnibus (GEO)

GEO (23) is a data repository and retrieval system for any high-throughput gene expression, or molecular abundance data that may be derived from any organism. GEO currently contains expression data from spotted microarrays, oligonucleotide arrays, hybridization filters, protein MS and SAGE. The GEO repository accepts data deposits via the web or in batch. The repository can be browsed from the GEO home page, and may be queried using Entrez. At the time of writing, the repository contains high-throughput gene expression data from over 9000 hybridization experiments, has about 300 array definitions and over 80 million individual spot measurement data.

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 (24). The database contains information on disease phenotypes and genes, including extensive descriptions, gene names, inheritance patterns, map locations and gene polymorphisms. OMIM, now an Entrez database, contains about 14 000 entries, including data on over 10 000 established gene loci and phenotypic descriptions.

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

The NCBI Molecular Modeling Database (MMDB) (7) is built by processing entries from the Protein Data Bank (5). The structures in MMDB are linked to one another by VAST structure–structure neighboring, and to entries in the Conserved Domain Database (CDD) (25) by RPS-BLAST neighboring. The CDD contains over 11 000 position-specific score matrices representing domains taken from the Simple Modular Architecture Research Tool (SMART) (26), Pfam (27) and recently, from domain alignments derived from both COGs and KOGs. 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 that, coupled with a multiple sequence alignment of representatives of the domain hit, can be viewed with NCBI's 3D molecular structure viewer, Cn3D (19), now in version 4.1 and enhanced with advanced alignment-building tools that use the PSI-BLAST and threading algorithms. The Conserved Domain Architecture Retrieval Tool (CDART) (28) allows searches of protein databases on the basis of a conserved domain and returns the domain architectures of database proteins containing the query domain. Alignment-based protein domain information from the CDD and 3D domains from the MMDB are searchable via the Entrez interface.

FOR FURTHER INFORMATION

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

REFERENCES

1. Benson,D.A., Karsch-Mizrachi.,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
2. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.

3. Wu,C.H., Yeh,L.S.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
4. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
5. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
6. Pruitt,K. Tatusov,T. and Maglott,D. (2003) NCBI Reference Sequence Project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
7. Chen,J., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **31**, 474–477.
8. Sequeira,E. (2003) PubMed Central—three years old and growing stronger. *ARL*, **228**, 5–9.
9. Altschul,S.E., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. 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.
12. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
13. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
14. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
15. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
16. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
17. Ermolaeva,O., Rastogi,M., Pruitt,K.D., Schuler,G.D., Bittner,M.L., Chen,Y., Simon,R., Meltzer,P., Trent,J.M. and Boguski,M.S. (1998) Data management and analysis for gene expression arrays. *Nature Genet.*, **20**, 19–23.
18. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Pham,L., Smigielski,E. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
19. 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.
20. Tatusova,T., Karsch-Mizrachi,I. and Ostell,J. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
21. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.*, **15**, 417–474.
22. Lash,A.E., Tolstoshev C.M., Wagner L., Schuler G.D., Strausberg R.L., Riggins G.J. and Altschul S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **7**, 1051–1060.
23. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
24. McKusick,V.A. (1998) *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders*. 12th edn. Johns Hopkins University Press, Baltimore, MD.
25. Marchler-Bauer,A., Anderson,J., Fedorova,N., DeWeese-Scott,C., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A., Lanczycki,C. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
26. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
27. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L. and Sonnhammer,E.L.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
28. Geer,L.Y., Domrachev,M., Lipman,D.J. and Bryant,S.H. (2002) CDART: protein homology by domain architecture. *Genome Res.*, **12**, 1619–1623.