28–33 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg033

Database resources of the National Center for Biotechnology

David L. Wheeler*, Deanna M. Church, Scott Federhen, Alex E. Lash, Thomas L. Madden, Joan U. Pontius, Gregory D. Schuler, Lynn M. Schriml, Edwin Sequeira, 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, 2002; Accepted October 2, 2002

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 Web site. NCBI resources include Entrez, PubMed, PubMed Central (PMC), LocusLink, the NCBITaxonomy Browser, BLAST, BLAST Link (BLink), Electronic PCR (e-PCR), Open Reading Frame (ORF) Finder, References Sequence (RefSeq), UniGene, HomoloGene, ProtEST, Database of Single Nucleotide Polymorphisms (dbSNP), Human/Mouse Homology Map, Cancer Chromosome Aberration Project (CCAP), Entrez Genomes and related tools, the Map Viewer, Model Maker (MM), Evidence Viewer (EV), Clusters of Orthologous Groups (COGs) database, Retroviral Genotyping Tools, 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 is 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 overview, 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 is available for bulk download at 'ftp.ncbi.nih.gov'.

DATABASE RETRIEVAL TOOLS

Entrez

Entrez (2) is an integrated database retrieval system for DNA and protein sequences derived from several sources (1,3–6), the NCBI taxonomy, genome maps, population sets, gene expression data, protein structures from the Molecular Modeling Database (MMDB) (7), 3D and alignment-based protein domains, and the biomedical literature via PubMed, Online Mendelian Inheritance in Man (OMIM), and online Books. PubMed includes primarily 12 million references and abstracts in MEDLINE(R), with links to the full-text of more than 3000 journals available on the Web. The Books database now contains 12 online scientific textbooks.

Entrez enables text searching of databases ranging from those for sequences and the scientific literature, to those for structure and gene expression using simple Boolean queries, and provides extensive links to related information. Some links are simple cross-references, for example, from a sequence to the abstract of the paper in which it was reported, from a protein sequence to its corresponding DNA sequence or 3D-structure, or to alignments with other sequences. Other links are based on computed similarities among the sequences or MEDLINE abstracts. These pre-computed 'neighbors' allow rapid access for browsing groups of related records. A service called LinkOut expands the range of external links from individual database records to related outside services, including organism-specific genome databases.

The records retrieved by an Entrez search can be displayed in a wide variety of formats and downloaded singly or in large batches. Formatting options vary for records of different types.

*To whom correspondence should be addressed. Tel: +1 3014962475/+1 3014355950; Fax: +1 3014809241; Email: wheeler@ncbi.nlm.nih.gov

For example, display formats for GenBank records include the GenBank Flatfile, FASTA, XML, ASN.1, and others. Graphical display formats are offered for some types of records, including genomic records.

PubMed Central (PMC)

PMC is a digital archive of peer reviewed journals in the life sciences. Over 90 journals, including Nucleic Acids Research, now deposit the full text of their articles in PMC, which is available at http://www.pubmedcentral.gov/. 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.

Taxonomy

The NCBI taxonomy database indexes over 119 000 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 sequence and structural data for a particular organism or group. Searches of the NCBI taxonomy may be made on the basis of whole, partial or phonetically spelled organism names, and 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.

NCBI is developing the non-bibliographic applications of LinkOut, and expanding that project into the taxonomy and sequence domains of Entrez. Many outside resources currently maintain LinkOut links from Entrez entries, including model organism and taxonomic databases. NCBI has developed several tools to help LinkOut providers, including a simple flatfile format for specifying links and a Name/ID status page for tracking the current use of names and IDs in the taxonomy database.

LocusLink

LocusLink, developed at NCBI in conjunction with several international collaborators, offers a single query interface to curated sequences and descriptive information about genes and includes links to NCBI's Map Viewer, Evidence Viewer (EV), Model Maker (MM), BLAST Link (BLink), protein domains from NCBI's Conserved Domain Database, and many other gene-related resources. LocusLink is discussed in a separate article in this issue (6).

THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The Basic Local Alignment Search Tool (BLAST) programs (8,9) perform sequence-similarity searches against a variety of sequence databases, beginning with either a query sequence or a GenBank accession number. BLAST returns a set of gapped alignments between the query and similar database sequences, with links to the full database records and to other relevant databases such as UniGene or LocusLink. The sequences of any or all of the database hits appearing in a BLAST alignment may be selected for bulk download. A BLAST variant, BLAST2Sequences (10), compares two DNA or protein sequences using any of the standard BLAST programs and produces a dot-plot representation of the alignments it reports.

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 (9) searches, tends to reduce the number of false-positive database hits (11).

The default BLAST output format is the 'pairwise' alignment, however several 'query-anchored' multiple sequence alignment formats are available. An alignment option, called the 'Hit Table', provides a compact, tabular, easily parsable, summary of the BLAST results including, for each database hit, the positions of alignment starts and stops, scores, and Expectation Values. These outputs may be returned in HTML, XML, text, or as ASN.1. In addition, BLAST can generate a taxonomically organized output that shows the distribution of BLAST hits by organism in three formats.

A particularly powerful feature of the web BLAST interface allows searches to be restricted to a database subset using standard Entrez search strings; the same restrictions may be used to screen the output of an initially unrestricted search. These features provide the means to effectively construct a custom database for searching, or to process the output of a search to include only sequences of interest. 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 recent addition to the BLAST family, called MegaBLAST (12), facilitates batch nucleotide queries which can be pasted into a web page or uploaded from a file. MegaBLAST is designed to search for nearly exact matches and is up to 10 times faster than standard BLAST for such searches. MegaBLAST is provided to search entire eukaryotic genomes, but it is also used to search a rapidly growing database, called the Trace Archive, which contains over 125 million sequencing traces. The Trace Archive includes whole genome shotgun (WGS), shotgun, EST, clone end and finishing reads from over 30 organisms such as *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Zea mays* and *Caenorhabditis elegans*. To facilitate rapid cross-species nucleotide queries of the Trace Archive, NCBI offers a version of MegaBLAST called Discontinuous MegaBLAST that uses a non-contiguous word match (13) as the nucleus for its alignments. Searches using Discontinuous MegaBLAST are far more rapid than cross-species translated searches such as blastx, but maintain a competitive degree of sensitivity when comparing coding regions.

The NCBI-generated assembly of the human as well as other submitted genomic assemblies, such as those of the mouse and zebrafish, may be searched using specialized genome BLAST

pages. These pages search a set of genome-specific databases and generate, where possible, genomic views of the BLAST hits using the Map Viewer.

BLink

BLink displays pre-computed protein BLAST alignments for each protein sequence in the Entrez databases. BLink allows for the display of subsets of these alignments by taxonomic criteria, by database of origin, relation to a complete genome, membership in a Clusters of Orthologous Group (COG) (14) 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 (15), is a system for automatically partitioning GenBank sequences, including ESTs, into a non-redundant set of gene-oriented clusters. UniGene clusters ESTs from 10 animals and 7 plants, bringing the total number of organisms represented to 17. UniGene starts with entries in the appropriate organism division of GenBank, combines these with ESTs of that organism and creates clusters of sequences that share virtually identical 3′ untranslated regions (3′ UTRs). 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 database, over 3.6 million human ESTs in GenBank have been reduced 35-fold in number to over 104 000 sequence clusters. The UniGene collection has been used as a source of unique sequence for the fabrication of microarrays for the large-scale study of gene expression (16). UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences. UniGene clusters may be searched in several ways; by gene name, chromosomal location, cDNA library, accession number, and ordinary text words. Cluster sequences may also be downloaded by FTP.

ProtEST

ProtEST, a tool analogous to BLink, presents pre-computed BLAST alignments between protein sequences from model organisms and the 6-frame translations of UniGene nucleotide sequences. Protein sequences that are derived from conceptual translations or model transcripts are excluded. The eight model organisms included are *H. sapiens*, *M. musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *C. elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Escherichia coli*. ProtEST links are displayed in UniGene reports with model organism protein similarites. For each nucleotide sequence match, the ProtEST report shows the UniGene cluster ID, the GenBank accession number, and the percent identity between the protein and nucleotide translation in the aligned region. A link is also provided to the sequence trace in the NCBI Trace Archive, if available. ProtEST reports are updated in tandem with UniGene protein similarities.

HomoloGene

HomoloGene is a database of both curated and calculated gene orthologs and homologs for 14 organisms including *H. sapiens*, *M. musculus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *A. thaliana*, *Hordeum vulgare*, *Oryza sativa*, *Z. mays*. 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. HomoloGene also contains a set of triplet ortholog-based COG (14)-like clusters, which may include up to 14 members, in which the triplet orthologs in two organisms are both orthologous to the same gene in a third organism. For the three organisms human, mouse and rat, there are currently over 7000 of these self-consistent triplets. 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. The current datasets for the calculated orthologs and homologs and the Mutually Orthologous Pairs are also available via FTP.

References Sequence (RefSeq)

The RefSeq database (6), provides curated reference sequences for mRNAs, genomic sequences, computationally-derived sequences and proteins for human and over 1700 other organisms.

Open Reading Frame (ORF) Finder

ORF Finder performs a six-frame translation of a nucleotide sequence and returns a graphic that indicates the location of each ORF found. Restrictions on the size of the ORFs returned may be set. The protein translations of the ORFs detected can be submitted directly for BLAST similarity searching or searching against the COGs (see below) database.

Electronic PCR (e-PCR)

e-PCR is a tool for locating Sequence Tagged Sites (STSs) within a nucleotide sequence by searching against a non-redundant database of over 133 000 human and 84 000 non-human STSs called UniSTS.

A database of Single Nucleotide Polymorphisms (dbSNP)

The dbSNP (17) is a repository for single base nucleotide substitutions and short deletion and insertion polymorphisms. The dbSNP database contains almost 3 million human SNPs as well as about half a million from organisms including *M. musculus*, *Anopheles gambiae*, *D. rerio* and *A. thaliana*. The Web interface allows flexible searches by gene name and by cross-reference to other databases such as OMIM or the structure databases. Searches for SNPs lying between two markers and batch downloads are also supported. SNP reports link to structures from the MMDB, allowing the 3-D visualization, using NCBI's interactive macromolecular viewer

Cn3D (18), of amino acid changes implied by SNPs in coding regions.

RESOURCES FOR GENOME-SCALE ANALYSIS

Entrez Genomes

Entrez Genomes (19) provides access to genomic data contributed by the scientific community for over 1000 species whose sequencing and mapping is complete or in progress. Entrez Genomes now includes more than 86 complete microbial genomes and 302 RefSeq for eukaryotic organelles. Many higher eukaryotic genomes are also included within Entrez Genomes such as those of *H. sapiens*, *M. musculus*, *D. melanogaster*, *Anopheles gambiae*, *C. elegans* and *A. thaliana*.

In Entrez Genomes, 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 each level are one or more views, pre-computed summaries, and links to analyses appropriate at that level. For instance, 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. A 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 gene product. Any protein gene product that is a member of a COG (19) is linked to the COGs database. A summary of COG functional groups is also presented in 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 3-D structures, are given in TaxTables and PDBTables, respectively. Pairwise sequence alignments are presented graphically and linked to the Cn3D macromolecular viewer (18), which allows the interactive display of 3-D structures and sequence alignments. The TaxPlot tool, graphically compares 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.

COGs

The COGs database (14), presents a compilation of orthologous groups of proteins from completely sequenced organisms representing 44 species and 30 phylogenetically distant clades. The COGs are now also linked to the proteins of two higher eukaryotes; *C. elegans* and *D. melanogaster*.

Retroviral genotyping tools

The genotyping of retrovirus sequences is important in the characterization of viral genetic diversity, in the tracking of epidemics, and in vaccine development. NCBI offers a Web-based genotyping tool that employs a blastn comparison between a retroviral sequence to be subtyped and 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.

Eukaryotic Genomic Resources

Entrez Genomes links to Genome Resources webpages devoted to the sequencing of a number of eukaryotic organisms including *H. sapiens*, *M. musculus* and *D. melanogaster*. A page called Plant Genomes Central serves as a collection point for resources related to plant genome projects. Many genome projects have progressed to the point at which it is useful to have an interactive genome viewing tool with which to correlate the data present in various of genomic maps. NCBI has developed the Map Viewer for this purpose.

Map Viewer

The NCBI Map Viewer displays genome assemblies using sets of synchronized chromosomal maps. Map Viewer displays are available for the genomes of four vertebrates, including *H. sapiens* and *M. musculus* and *D. rerio*, three invertebrates, including *D. melanogaster* and *C. elegans*, seven plants, including *A. thaliana* and *O. satvia*, and two fungi, *S. cerevisiae* and *Schizosaccharomyces pombe*. The genomic maps displayed by the Map Viewer vary according to the data available for the subject organism. The maps can be selected from a set of cytogenetic maps, such as chromosomal ideograms, sequence-based maps, such as those showing contigs, genes, and SNPs, and physical maps, such as the G3 and GB4 human radiation-hybrid maps. Maps showing *ab initio* gene models, EST alignments with links to UniGene clusters, and mRNA alignments used to construct gene models are also available for some organisms. The rightmost map in a Map Viewer display, called the master map, generates an extended set of map-specific links to related resources. In the case of the Genes map, two of these links are to the EV and MM described below. In addition to its graphical display, the Map Viewer offers a tabular view of the data that is convenient for export to other programs for further analysis.

Queries against an entire genome or particular chromosomes can be made in the Map Viewer using gene names or symbols, marker names, SNP identifiers, accession numbers and other identifiers. The human version of the Map Viewer is tightly integrated with other NCBI databases such as LocusLink and dbSNP. Segments of a genomic assembly may be downloaded using the Map Viewer's 'Download/View Sequence' link for some genomes such as *H. sapiens*, *M. musculus* and *A. thaliana*. Supported download formats are GenBank and FASTA.

Model Maker (MM)

MM allows the construction of transcript models using novel 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. The MM interface consists of a graphical overview of transcript alignments to a genomic contig with each unique block of alignment collected and numbered as a putative exon. Transcript models are constructed by selecting

from this collection. As the 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 (EV)

The EV displays the alignments to a genomic contig of RefSeq transcripts, GenBank mRNAs, known or potential transcripts, and ESTs supporting a gene model. The EV produces a graphical summary of the alignments that indicates the coordinate range of the gene model on the genomic contig and the areas of alignment to the transcripts on separate tracks. EST alignment density along the contig is indicated on another track. A mismatch and an insertion/deletion track are also shown to highlight areas of disagreement between transcript sequences and the genomic sequence. Following the graphical summary are exon-by-exon alignments of all of the transcript sequences against the genomic contig, including flanking genomic sequence for each exon to show the presence or absence of splice sites. Any proteins annotated on the transcript sequences are also shown and mismatches between transcripts and the genomic contig or between proteins annotated on the aligned transcripts are highlighted.

The Human–Mouse Homology Maps

The Human–Mouse Homology Maps are tables of genetic loci in homologous segments of DNA from human and the mouse. The map is computed by integrating orthologs curated by the Mouse Genome Database with putative orthologs identified by homology. The maps are linked to GeneMap'99, OMIM, LocusLink, dbSTS, BLAST2Sequences and the Mouse Genome Database at The Jackson Laboratory. Other mouse genome resources can be found on the Mouse Genome Resources page.

The Cancer Chromosome Aberration Project (CCAP)

The CCAP service is an initiative of the National Cancer Institute (NCI) and 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 (20). The Spectral Karyotyping database, SKY, created jointly by NCI and NCBI, enables investigators to share their own SKY and Comparative Genomic Hybidization (CGH) data on chromsomal aberrations (http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi).

RESOURCES FOR THE ANALYSIS OF PATTERNS OF GENE EXPRESSION AND PHENOTYPES

SAGEmap

Serial Analysis of Gene Expression (SAGE) is a technique for taking a snapshot of the messenger RNA population of a cell to obtain a quantitative measure of gene expression. NCBI's SAGEmap (21) service implements many functions useful in the analysis of SAGE data such as a two-way mapping between SAGE tag and UniGene. SAGEmap can also construct a user-configurable table of data comparing one group of SAGE libraries with another. Groups may be chosen for inclusion in the table on the basis of several expression criteria. SAGEmap is updated weekly, immediately following the update of UniGene and the data is reflected in the human genome Map Viewer as the SAGE track.

Gene Expression Omnibus (GEO)

The GEO (22) is a data repository and retrieval system for gene expression data derived from any organism or artificial source. Gene expression data derived from spotted microarray, high-density oligonucleotide array, hybridization filter and SAGE data, are available for download and accepted for deposit. At the time of writing, the repository contains high-throughput gene expression data on over 2300 samples.

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 and gene polymorphisms. OMIM currently contains 13 864 entries, including data on 10 290 established gene loci and 1019 phenotypic descriptions, and is now searchable using the powerful Entrez interface.

THE MOLECULAR MODELING DATABASE (MMDB), THE CONSERVED DOMAIN DATABASE SEARCH AND CDART

The NCBI MMDB, built by processing entries from the Protein Data Bank (5), is described in (7). The structures in the MMDB are linked to sequences in Entrez and to the Conserved Domain Database (CDD). The CDD contains PSI-BLAST-derived Position Specific Score Matrices representing domains taken principally from two public protein domain collections, the Simple Modular Architecture Research Tool (SMART) (24), and Pfam (25), but also draws from domains defined by NCBI researchers. 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 3-D molecular structure viewer, Cn3D (18). 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 conatining the query domain. Alignment-based protein domain information from the CDD and 3-D domains from the MMDB are searchable via the Entrez interface.

FOR FURTHER INFORMATION

Most of the resources described here include documentation, other explanatory material and references to collaborators and

data sources on the respective web sites. 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., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
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. Barker,W.C., Garavelli,J.S., Huong,H., McGarvey,P.B., Orcutt,B.C., Srinivarsarao,G.Y., Xiao,C., Yeh,L.S., Ledley,R.S., Janda,J., Pfeiffer,F., Mewes,H.W., Tsugita,A. and Wu,K. (2000) The Protein Information Resource (PIR). *Nucleic Acids Res.*, **28**, 41–44.
4. Kriventseva,E.V., Fleischmann,W., Zdobnov,E.M. and Apweiler,R. (2001) CluSTr: a database of Clusters of SWISS-PROT and TrEMBL proteins. *Nucleic Acids Res.*, **29**, 33–36.
5. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
6. Pruitt,K., Tatusov,T. and Maglott,D. (2003) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **31**, 34–37.
7. Marchler-Bauer,A., Anderson,J., Fedorova,N., DeWeese-Scott,C., Geer,L.Y., Hurwitz,D., Jackson,J.J., Jacobs,A., Lanczycki,C., Liebert,C., Madej,T., Marchler,G.H., Mazumder,R., Nikolskaya,A., Panchenko,A.R., Shoemaker,B.A., Song,J., Sridhar,R.B., Thiessen,P.A., Vasudevan,S., Wang,Y., Yamashita,R., Yin,J. and Bryant,S.H. (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **31**, 474–477.
8. 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.
9. 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.
10. 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.
11. 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.
12. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
13. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
14. 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.
15. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
16. 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.
17. 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.
18. 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.
19. Tatusova,T., Karsch-Mizrachi,I. and Ostell,J. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
20. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.*, **15**, 417–474.
21. 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.
22. 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.
23. McKusick,V.A. (1998) Mendelian Inheritance in Man. *Catalogs of Human Genes and Genetic Disorders*, 12th edn. The Johns Hopkins University Press, Baltimore, MD.
24. 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.
25. 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.