Nucleic Acids Research, 2001, Vol. 29, No. 1 **11–16**

Database resources of the National Center for Biotechnology Information

David L. Wheeler*, Deanna M. Church, Alex E. Lash, Detlef D. Leipe, Thomas L. Madden, Joan U. Pontius, Gregory D. Schuler, Lynn M. Schriml, Tatiana A. Tatusova, Lukas Wagner and Barbara A. Rapp

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

Received October 3, 2000; Accepted October 4, 2000

ABSTRACT

In addition to maintaining the GenBank® nucleic acid sequence database, the National Center for Biotechnology Information (NCBI) provides data analysis and retrieval resources that operate on the data in GenBank and a variety of other biological data made available through NCBI's Web site. NCBI data retrieval resources include Entrez, PubMed, LocusLink and the Taxonomy Browser. Data analysis resources include BLAST, Electronic PCR, OrfFinder, RefSeq, UniGene, HomoloGene, Database of Single Nucleotide Polymorphisms (dbSNP), Human Genome Sequencing, Human MapViewer, GeneMap'99, Human–Mouse Homology Map, Cancer Chromosome Aberration Project (CCAP), Entrez Genomes, Clusters of Orthologous Groups (COGs) database, Retroviral Genotyping Tools, Cancer Genome Anatomy Project (CGAP), SAGEmap, Gene Expression Omnibus (GEO), Online Mendelian Inheritance in Man (OMIM), the Molecular Modeling Database (MMDB) and the Conserved Domain Database (CDD). 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® (1) nucleic acid sequence database, to which data is submitted directly by the scientific community, NCBI provides data retrieval systems and computational resources for the analysis of GenBank data and the variety of other biological data made available through NCBI.

The data accessible from NCBI's home page (http://www.ncbi.nlm.nih.gov) runs the gamut from short sequences representative of parts of genes to complete genomes, protein structures and clinical descriptions of genetic disorders. NCBI offers an array of computational resources to aid in the analysis of each type of data. For this overview, the NCBI suite of database resources is grouped into seven categories: database retrieval systems, sequence similarity search programs, resources for analysis of gene-level sequences, resources for chromosomal sequences, resources for genome-scale analysis, resources for the analysis of gene expression and phenotypes, and resources for protein structure and modeling. Table 1 provides an at-a-glance summary of these resources.

DATABASE RETRIEVAL TOOLS

Entrez

Entrez (2) is an integrated database retrieval system that accesses DNA and protein sequences, genome maps, population sets, protein structures from MMDB (3) and the biomedical literature via PubMed and Online Mendelian Inheritance in Man (OMIM), with embedded links to the NCBI taxonomy. The sequences in Entrez, especially protein sequences, are obtained from a variety of database sources [including GenBank protein translations, Protein Identification Resource (4), SWISS-PROT (5), Protein Research Foundation, Protein Data Bank (6) and RefSeq (7)], and therefore include more sequence data than GenBank alone. PubMed includes primarily the 10.7 million references and abstracts in MEDLINE®, with links to the full-text of more than 1100 journals available on the Web.

Entrez provides text searching of sequence or bibliographic records using simple Boolean queries, plus 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 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.

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

Table 1. A summary of selected web-based data resources, in addition to GenBank, provided by NCBI

Resource	Description
Database retrieval tools	
Entrez	Integrated database and retrieval system (2) for access to publicly available GenBank and other sequence data, mapping and complete genome data, 3-D structures, and the biomedical literature through PubMed and OMIM.
The Taxonomy Browser	Search tool for the NCBI taxonomy database which indexes over 79 000 organisms represented in the sequence databases with at least one nucleotide or protein sequence.
LocusLink	Database of official gene names and other gene identifiers offering a single query interface to curated sequences and descriptive information about genes, developed with international collaborators (7).
The BLAST family of sequence similarity search programs	
BLAST	The BLAST family of programs (including BLAST, PSI-BLAST, PHI-BLAST, BLAST2Sequences) perform rapid sequence-similarity searches of GenBank and specialized data sets (8–10).
Resources for gene-level sequences	
UniGene	The UniGene system (12) partitions GenBank sequences, including ESTs, into a non-redundant set of gene-oriented clusters; currently includes human, mouse, rat and zebrafish.
HomoloGene	A database of sets of homologous and orthologous UniGene clusters for human, mouse, rat, zebrafish and cow.
RefSeq	Database of reference sequence standards for mRNAs and proteins in the Entrez databases, curated by NCBI staff (7).
dbSNP	Database of Single Nucleotide Polymorphisms (dbSNP) that includes both single base nucleotide substiutions and short deletion and insertion polymorphisms deposited by the scientific community (15).
ORF Finder	Tool that performs a six-frame translation of a nucleotide query and returns a graphic that indicates the location of each ORF found.
Electronic PCR	Tool for locating STSs within a nucleotide sequence query by comparing the query with the dbSTS database of sequences and primer pairs.
Resources for chromosomal sequences	
Human Genome MapViewer	Genome browser showing an integrated view of the human genome maps, including both physical and genetic maps.
Human Genome Sequencing	Tracks progress and provides access to human genome sequencing data such as individual contigs and assemblies deposited by the Human Genome Project sequencing centers.
GeneMap'99	GeneMap'99 presents mapping information for 30 261 unique gene loci representing approximately half of the 60 000–80 000 genes contained in the human genome (13).
The Human-Mouse Homology Maps	Access to tables of genetic loci in homologous segments of DNA from human and the mouse.
The Cancer Chromosome Aberration Project (cCAP)	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 (18).
Resources for genome-scale analysis	
Entrez Genomes	The Entrez Genomes database (20) organizes and provides access to contributed genomic mapping and sequence data for over 900 species.
Clusters of Orthologous Groups (COGs)	Clusters of orthologous groups of proteins from completely sequenced bacteria, archaea, and eukaryote (20).
Retroviral Genotyping Tools	A web-based genotyping system for the analysis of retroviral genomes.
Resources for the analysis of patterns of gene expression and phenotypes	
The Cancer Genome Anatomy Project (CGAP)	Provides access to genetic data on normal, precancerous and malignant cells.
Gene Expression Omnibus (GEO)	A database for gene expression data obtained using a variety of experimental technologies such as gene-chips and SAGE (Serial Analysis of Gene Expression).
SAGEmap	SAGEmap offers many functions for the analysis of data generated by the SAGE technique.
Online Mendelian Inheritance in Man (OMIM)	Catalog of human genes and disorders, authored and edited by Dr Victor A. McKusick and colleagues at Johns Hopkins University (22).
Molecular structure	
The Conserved Domain Database (CDD)	The CDD combines data from the SMART and Pfam protein domain databases in the form of a library of PSI-BLAST PSSMs representative of each conserved domain. This library can be searched using NCBI's RPS-BLAST.
The Molecular Modeling Database (MMDB)	MMDB (3) is a structural database derived from the Protein Data Bank and accessible via the Entrez system.

The Taxonomy Browser

The NCBI taxonomy database indexes over 79 000 organisms that are represented in the sequence 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 of organisms. Searches of the NCBI taxonomy may be made on the basis of whole, partial or phonetically-spelled organism names, and direct links to organisms commonly used in biological research are also provided. The new Entrez Taxonomy system adds the ability to display custom taxonomic trees representing user-defined subsets of the full NCBI taxonomy.

LocusLink

The LocusLink database of official gene names and other gene identifiers, described elsewhere in this issue (7), was developed at NCBI in conjunction with several international collaborators, and offers a single query interface to curated sequences and descriptive information about genes.

THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The BLAST family of search programs (8,9) is provided for the most frequent type of analysis performed on GenBank, the sequence-similarity search. NCBI's Web interface to the standard BLAST 2.1 program accepts either a sequence or accession number and performs the search using either an identity matrix for blastn (nucleotide) searches or a PAM or BLOSUM scoring matrix for protein searches. BLAST produces a set of gapped alignments, with links to the full document records, accompanied by an alignment score and a measure of statistical significance, called the Expectation Value, for judging the quality of the alignment. Web BLAST provides a graphical overview of the alignments, color-coded by alignment score, which clearly shows the extent and quality of sequence similarities, as well as the disposition of gaps in the alignments. Web BLAST can also generate a taxonomically organized output that emphasizes taxonomic patterns of sequence-similarity.

The default databases searched by BLAST are the non-redundant (nr) nucleotide and protein databases constructed from the Entrez databases. Several specialized databases may also be searched, and searches may be restricted to sequences from a particular organism. Query sequences may be filtered for low complexity or human repeats. Customized BLAST pages allow queries against finished human genomic data, microbial genomes or the genomes of malaria-associated pathogens.

Specialized versions of BLAST are offered for the needs of protein similarity searching. Position Specific Iterated BLAST (PSI-BLAST) (9) initially performs a conventional BLAST search to produce alignments from which it constructs a position specific score matrix (PSSM). Subsequent BLAST iterations use this PSSM to find similarities in the database. Pattern Hit Initiated BLAST (PHI-BLAST) (10) requires both a query sequence and a pattern present within the query sequence. The pattern specifies an obligatory match between query and database sequences, about which optimal local alignments are constructed. Another variant, 'BLAST2Sequences' (11), compares two DNA or protein sequences and produces a dot-plot representation of the alignments it reports.

Basic BLAST 2.0 searches can also be performed by email through the address: blast@ncbi.nlm.nih.gov. Documentation can be obtained by sending the word 'help' to the server address.

RESOURCES FOR GENE-LEVEL SEQUENCES

UniGene

To manage the redundancy of the EST data, NCBI has developed UniGene (12), a system for automatically partitioning GenBank sequences, including ESTs, into a non-redundant set of gene-oriented clusters. There are currently five UniGene databases; for human, mouse, rat, zebrafish and cow. UniGene starts with entries in the appropriate organismic 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 1.8 million human ESTs in GenBank have been reduced 21-fold in number to approximately 84 000 sequence clusters. In a similar fashion, the mouse, rat, zebrafish, and cow ESTs have been organized as 73 000, 37 000, 10 000, and 5500 clusters, respectively. The human UniGene collection has been effectively used as a source of mapping candidates for the construction of a human gene map (13). In this case, the 3′ UTRs of genes and ESTs are converted to sequence-tagged sites (STSs) that are then placed on physical maps and integrated with pre-existing genetic maps of the genome. The UniGene collection has also been used as a source of unique sequences for the fabrication of 'chips' for the large-scale study of gene expression (14). 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.

HomoloGene

HomoloGene is a database of both curated and calculated orthologs and homologs for the human, mouse, rat, zebrafish and cow genes represented in UniGene and LocusLink. 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 clusters in which orthologous clusters in two organisms are both orthologous to the same cluster 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 in

UniGene cluster titles. The current datasets for the calculated orthologs and homologs and the Mutually Orthologous Pairs are also available via FTP.

RefSeq

The References Sequence (RefSeq) database, described elsewhere in this issue (7), provides curated reference sequences for mRNAs and proteins from human and other organisms.

A database of Single Nucleotide Polymorphisms (dbSNP)

The database of Single Nucleotide Polymorphisms (dbSNP), described elsewhere in this issue (15), serves as a repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms that are deposited by the research community.

ORF Finder

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

Electronic PCR

PCR-based assays for STSs can be used for gene identification and mapping. Electronic PCR (e-PCR) is a tool for locating STSs within a nucleotide sequence by comparing the query against the dbSTS database of STS sequences and primer pairs. The e-PCR application accepts either an accession number or sequence as input, and returns a table of links to matching dbSTS records as well as the primer pairs used to amplify each STS identified.

RESOURCES FOR CHROMOSOMAL SEQUENCES

Human Genome Sequencing

The Human Genome Sequencing (16) site shows chromosome-specific progress of the human sequencing project, provides access to individual contigs and assemblies, and offers chromosome-specific BLAST searches. Links to contributing genome sequencing centers are also provided. Sequence data may be downloaded by contig or chromosome.

Human Genome MapViewer

The Human Genome MapViewer can display the human genome data using up seven parallel chromosomal maps simultaneously. The maps displayed can be selected from a set of 19, and include cytogenetic maps, such as chromosomal ideograms, sequence-based maps, such as those showing contigs, genes, and SNPs, and radiation hybrid maps, such as the G3 and GB4 maps used to construct GeneMap '99. Queries against the entire human genome or particular chromosomes can be made using gene names or symbols, marker names, SNP identifiers, accession numbers and other identifiers. The Human Genome MapViewer is tightly integrated with other NCBI databases such as LocusLink and dbSNP. A MapViewer similar to the Human Genome MapViewer is also used to display the *Drosophila* genome data.

GeneMap'99

An international consortium was formed in 1994 to construct a human gene map by determining the locations of ESTs relative to a framework of well-characterized genetic markers (17). The current version of this map is the radiation hybrid map, GeneMap'99 (13), featuring 30 261 unique gene loci.

The Human–Mouse Homology Maps, mouse sequencing resources

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 Sequencing page, analogous to the Human Genome Sequencing page described above.

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 (18). Also provided are bacterial artificial chromosome (BAC) human chromosome mapping data provided through CCAP's fluorescent *in situ* hybridization (FISH) effort.

RESOURCES FOR GENOME-SCALE ANALYSIS

Entrez Genomes

The Entrez Genomes database (19) provides access to genomic data contributed by the scientific community for over 900 species whose sequencing and mapping is complete or in progress, and now includes more than 30 complete microbial genomes. Also included is a collection of 169 reference sequences for the complete genomes of eukaryotic organelles. Data can be accessed hierarchically starting from either an alphabetical listing or a phylogenetic tree for complete genomes in 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 for 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. 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 gene product. Any protein gene product that is a member of a COG (20) 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 (21), which allows the interactive display of 3-D structures and sequence alignments.

Clusters of Orthologous Groups (COGs)

The COGs database, described elsewhere in this issue (20), presents a compilation of orthologous groups of proteins from completely sequenced organisms representing phylogenetically distant clades.

Retroviral genotyping tools

Genotyping retrovirus sequences is important in the characterization of viral genetic diversity, tracking of epidemics and vaccine development. NCBI has developed a Web-based genotyping tool for the analysis of retroviral genomes. The genotyping method employs a blastn comparison between the retroviral sequence to be subtyped and a panel of reference sequences provided by the user. An HIV-1-specific subtyping tool uses a set of reference sequences taken from the principle HIV-1 variants.

RESOURCES FOR THE ANALYSIS OF PATTERNS OF GENE EXPRESSION AND PHENOTYPES

The Cancer Genome Anatomy Project (CGAP)

CGAP provides access to genetic data on normal, precancerous and malignant cells generated by the NCI's CGAP initiative. CGAP cDNA library information may be retrieved by text words, gene name, clone ID, tissue type, method of sample preparation, stage of tumor development or by UniGene Cluster ID. Expression profiles of cDNA libraries may be compared using either the Digital Differential Display (DDD) tool or the xProfiler. CGAP also includes a directory of tumor suppressor genes and oncogenes.

SAGEmap

Serial Analysis of Gene Expression (SAGE) refers to 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 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 specified by the user. SAGEmap is updated weekly, immediately following the update of UniGene.

GEO

The Gene Expression Omnibus (GEO) is an effort to build a data repository and retrieval system for gene expression data derived from any organism or artificial source. Gene expression data derived from spotted microarray (microarray), high-density oligonucleotide array (HDA), hybridization filter (filter) and serial analysis of gene expression (SAGE) data, are being accepted. Online tools for the interactive retrieval and analysis of this expression data are under development.

OMIM

NCBI provides Web access to the OMIM database, a catalog of human genes and genetic disorders authored and edited by Dr Victor A. McKusick at The Johns Hopkins University (22). The database contains information on disease phenotypes and genes, including extensive descriptions, gene names, inheritance patterns, map locations and gene polymorphisms. OMIM currently contains 11 925 entries, including data on 8594 established gene loci and 799 phenotypic descriptions, and is now searchable using the powerful Entrez interface.

THE MOLECULAR MODELING DATABASE

See Table 1 and (1).

THE CONSERVED DOMAIN DATABASE SEARCH

Conserved domains are structural modules that have been reused frequently during the process of evolution. The Conserved Domain Database (CDD) contains domains derived principally from two public protein domain collections, the Simple Modular Architecture Research Tool (Smart) (23), and Pfam (24). NCBI's Conserved Domain Search (CD-Search) service can be used to search a protein sequence for conserved domains in the CDD.

To produce the CDD a PSI-BLAST-type PSSM is calculated from each domain alignment in the SMART and Pfam databases. These PSSMs are then combined into a library that can be searched using Reverse Position-Specific BLAST (RPS-BLAST), a BLAST variant that searches a database of PSSMs with a protein sequence query. 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 (21).

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 What's New feature announces new and enhanced resources. Additional tools to guide users to NCBI's growing array of services are also being developed. 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. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
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. Wang,Y., Addess,K.J., Geer,L., Madej,T., Marchler-Bauer,A., Zimmerman,D. and Bryant,S.H. (2000) MMDB: 3D structure data in Entrez. *Nucleic Acids Res.*, **28**, 243–245.

4. 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. Updated article in this issue: *Nucleic Acids Res*. (2001), **29**, 29–32.
5. 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.
6. 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. Updated article in this issue: *Nucleic Acids Res*. (2001), **29**, 214–218.
7. Pruitt,K. and Maglott,D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
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. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3991.
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. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
13. Deloukas,P., Schuler,G.D., Gyapay,G., Beasley,E.M., Soderlund,C., Rodriguez-Tome,P., Hui,L., Matise,T.C., McKusick,K.B., Beckmann,S. *et al.* (1998) A physical map of 30,000 human genes. *Science*, **282**, 744–746.
14. Ermolaeva,O., Rastogi,M., Pruitt,K.D., Schuler,G.D., Bittner,M.L., Chen,Y., Simon,R., Meltzer,P., Trent,J.M. and Boguski,M.S. (1998) Data management and analysis for gene expression arrays. *Nature Genet.*, **20**, 19–23.
15. 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.
16. Jang,W., Chen,H.C., Sicotte,H. and Schuler,G.D. (1999) Making effective use of human genomic sequence data. *Trends Genet.*, **15**, 284–286.
17. Schuler,G.D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tome,P., Aggarwal,A., Bajorek,E. *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540–546.
18. Mitelman,F., Mertens,F. and Johansson,B. (1997) A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.*, **15**, 417–474.
19. Tatusova,T., Karsch-Mizrachi,I. and Ostell,J. (1999) Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics*, **15**, 536–543.
20. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
21. 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.
22. McKusick,V.A. (1998) Mendelian Inheritance in Man. *Catalogs of Human Genes and Genetic Disorders*, 12th Edn. The Johns Hopkins University Press, Baltimore, MD.
23. Schultz,J., Copley,R.R., Doerks,T., Ponting,C.P. and Bork,P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234.
24. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.