SPECIAL PULL-OUT SECTION

GENOME DATABASES

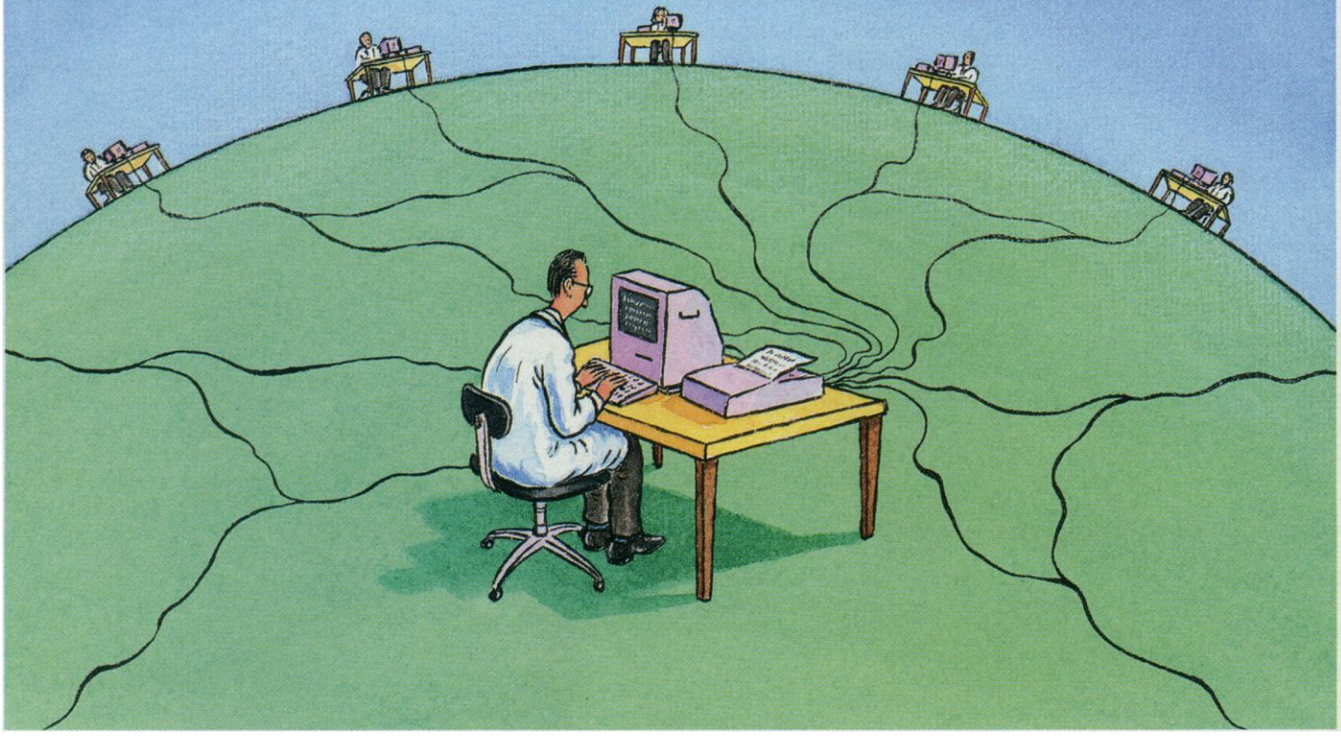


ILLUSTRATION: ROBERT KAUFMAN

The Human Genome Project will generate more data than any single project to date in biology. A major product of this 15-year, $3-billion effort will be either one huge database or perhaps a set of linked databases that will list the location of every one of the genome's 100,000 or so genes. The database(s) will also contain information on thousands of other landmarks along the chromosomes, and ultimately, the entire nucleotide sequence of the human genome and of several experimental animals.

Since the Genome Project began several years ago, a plethora of databases have been developed or are in the works. They range from the massive Genome Data Base at Johns Hopkins University, the central repository of all gene mapping information, to small databases focusing on single chromosomes or organisms. Some are publicly available, others are essentially private electronic lab notebooks. Still others limit access to a consortium of researchers working on, say, a single human chromosome. An increasing number incorporate sophisticated search and analytical software, while others operate as little more than data lists.

As a service to our readers, *Science*, in consultation with numerous experts in the field, has compiled this list of some key genome-related databases. We did not limit ourselves to map and sequence databases but also included the tools investigators use to interpret and elucidate genetic data, such as protein sequence and protein structure databases. Because a major goal of the Genome Project is to map and sequence the genomes of several experimental animals, including *E. coli*, yeast, fruit fly, nematode, and mouse, we have listed the available databases for those organisms as well. We also include several databases that are still under development, marked with a ❖—including some ambitious efforts that go beyond data compilation to create what are being called electronic research communities, enabling many users, rather than just one or a few curators, to add or edit the data and tag it as raw or confirmed.

Given the speed with which new databases are appearing, or existing ones are metamorphosing, our list cannot possibly be inclusive. For the most up-to-date information on the databases listed here, and on those we were unable to include, please consult either the Listing of Molecular Biology databases or the Directory of Biotechnology Information Resources (see below).

Writer: Jacqueline Courteau
Editor: Leslie Roberts
Graphic Design: Diana DeFrancesco
Research: Susanne Hiller

COMPREHENSIVE MAP AND SEQUENCE DATABASES

Human genetic map and sequence data are compiled in three large public databases. DNA and RNA nucleotide sequences are maintained by an international collaboration of GenBank®, the European Molecular Biology Laboratory (EMBL) Data Library, and the DNA Data Bank of Japan (DDBJ); they have worked together since 1987 to collect and distribute an international data pool. Human gene mapping data are compiled in the Genome Data Base (GDB); the complementary Online Mendelian Inheritance in Man (OMIM™) contains related phenotypic descriptions. GenInfo, under development by the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM), will form a "backbone" sequence database—a comprehensive yet flexible set of both nucleotide and protein sequence data to which other databases can add, refer, annotate, interpret, and extrapolate.

GENBANK®, EMBL, AND DDBJ

Contents

More than 65 million sequenced DNA and RNA base pairs from 3,000 species, including complete genomes of approximately 200 organelles, viruses, plasmids. Size: up to 177 MB of data and 80 MB of programs (depending on database and format).

Data Source

80% from direct electronic data submission; remainder from journal scans. Direct submissions reviewed, annotated, entered within 1 to 3 weeks; updates exchanged daily among the three databases. Data generally available on-line concurrently with or shortly after publication.

Format/Access/Cost

On-line access via network or direct modem connection (with various software and related down-loadable database files). Also distributed on magnetic tape, CD-ROM, and floppy diskette, with quarterly updates; distributed in various formats by commercial vendors.

GenBank®: Distributed at cost; prices range from $100 to $1100, depending on format. Free on-line access for limited search and connect time; annual subscriptions for $500-1200 offer more connect time, more sophisticated software.

EMBL: Tape distribution for non-profit organizations 75-150 deutsche marks (DM) per release; 300 DM for commercial users. CD-ROM for 150-300 DM (non-profit) or 600 DM (commercial). On-line access free through EMNET nodes in Europe. Free downloadable data files available from EMBL File Server.

DDBJ: Distributed on magnetic tapes for shipping charge only (provide tapes with request); updated twice a year. Free on-line access via network or direct modem connection; free access to sequence analysis software.

Contacts

GenBank®
c/o Intelligenetics, Inc.
700 E. El Camino Real
Mountain View, CA 94040, USA
Phone: 415-962-7364
Fax: 415-962-7302
E-mail: genbank@genbank.bio.net

EMBL Data Library
Postfach 10.2209, Meyerhofstrasse 1
6900 Heidelberg, GERMANY
Phone: 49-6221-387-258
Fax: 49-6221-387-519 or 387-306
E-mail: DataLib@EMBL-Heidelberg.DE

DDBJ
Takashi Gojobori
National Inst. of Genetics
Mishima, Shizuoka 411, JAPAN
Phone: 81-559-75-0771
Fax: 81-559-75-6040
E-mail: ddbj@ddbj.nig.ac.jp

GDB AND OMIM™

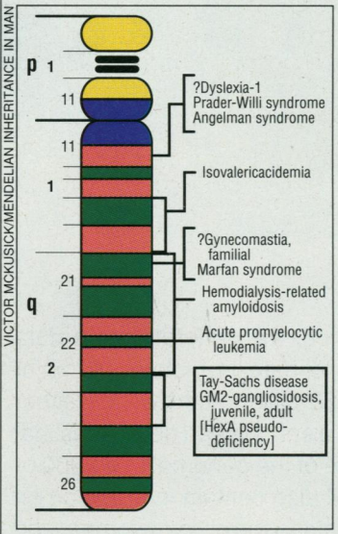
Contents

GDB: Human gene mapping information: genetic loci (genes, clinical phenotypes, and DNA segments) arranged by map and chromosome location; polymorphisms and alleles (including published genetic markers for the CEPH reference families); GenBank® cross-references for known sequences; probes; data sources; contacts. Under development: links to data for other species to allow homology searches. Size: 55 MB data plus 5 MB software.

OMIM™: Computer-readable form of data on human genetic traits and inheritance compiled and edited by Johns Hopkins geneticist Victor McKusick (*Mendelian Inheritance in Man*), organized by clinical disorders or traits: gene name; clinical observations; inheritance patterns; allelic variants; chromosome location and linkage; references. Size: 20 MB data plus 14 MB software.

Data Source

GDB: Literature; direct submission by human gene mapping committees for individual chromosomes. Lag-time from publication or submission to data entry varies from days to months; some direct submissions released prior to publication. Updates available on-line immediately in U.S., weekly in European nodes.



Chromosome 15.

OMIM™: Literature and other sources reviewed by McKusick; lag time varies. Updated daily for U.S. on-line access, weekly in European nodes.

Format/Access/Cost

Can log onto system using personal computers or terminals equipped with communications software; OMIM™ displays more sophisticated graphics on Sun workstations. On-line access free via direct modem dial-up or Internet; various international access nodes operating or planned.

Contact

GDB/OMIM User Support
Welch Medical Library
1830 E. Monument St., 3rd Floor
Baltimore, MD 21205, USA
Phone: 301-955-7058
Fax: 301-955-0054
E-mail: help@welch.jhu.edu

Special Features

GDB and OMIM™ are partially linked for cross-searching: One could query GDB about genes present on the long arm of chromosome 15, discover that Prader Willi syndrome is linked to a gene deletion there, then request information on clinical characteristics and inheritance patterns from OMIM™.

❖ GENINFO ❖

Contents

Three separate databases—nucleotide sequences, protein sequences, and MEDLINE bibliographic abstracts—combined with windows-based software for sequence similarity searches and bibliographic information retrieval.

Data Source

Nucleotide and protein sequence data from GenBank, PIR, literature searches; bibliographic data from NLM.

Format/Access/Cost

Beta test version with software (entitled "Entrez: Sequences") distributed on CD-ROM; runs on Apple Macintoshes™ and IBM-compatible PCs equipped with Microsoft Windows™3.0 and CD-ROM player. Public release planned late 1991 by NTIS and NLM (anticipated cost: $200/year, with bimonthly updates). Also planned: free on-line access via Internet; data-only releases in GenBank® flat file format.

Contact

NLM Medlars Management Service
8600 Rockville Pike
Bethesda, MD 20894, USA
Phone: 800-638-8480
E-mail: info@ncbi.nlm.nih.gov

GENE EXPRESSION

Diverse specialized databases hold data on the structure and function of proteins, carbohydrates, and other biological molecules. Protein sequence data are gathered and maintained in the International Protein Sequence Database or Protein Information Resource (PIR) through international collaborators from the National Biomedical Research Foundation (NBRF), the Martinsried Institute for Protein Sequences (MIPS), and the Japanese International Protein Information Database (JIPID). Protein Data Bank (PDB) contains 3-D structures of biological macromolecules including proteins, DNA, RNA, viruses, and carbohydrates. The prototype BioMagResBank contains protein and peptide structure information from NMR solution studies, complementary to the crystallographic x-ray diffraction data gathered by PDB. SWISS-PROT and PROSITE integrate protein sequences with functional and structural data. The Cambridge Structural Database (CSD) contains published crystal structures from x-ray and neutron diffraction studies for small biological molecules (similar to PDB data on larger molecules). The Complex Carbohydrate Structure Database (CCSD) contains primary structures of polysaccharides, glycopeptides, and glycolipids.

PIR

Contents

30,000 amino acid sequence entries covering 8 million protein residues from many species (including some translations from nucleic acid sequences); cross-references to known nucleic acid sequences; data on x-ray crystallography and active site determination; annotations on functional features; references. Auxiliary data: preliminary, fragmented, predicted sequences. Size: up to 150 MB of data plus software (depending on database and format).

Data Source

Published literature; direct computer-readable submission by authors. Lag time 2 weeks to 6 months from publication to data entry, depending on database and whether data submitted directly. Quarterly updates collate data from all three sites; daily updates available on-line at each site.

Format/Access/Cost

Distributed on magnetic tape and on-line with companion databases and personal computer software.

NBRF: On-line access $350 per year; additional charges for set-up, hourly use. Magnetic tape with quarterly updates $250. Sequence searches performed on request at cost, generally $200-250.

MIPS: On-line access free. Magnetic tape 300 DM per release or update.

JIPID: On-line access 1000 yen/month plus 1 yen/10 seconds processor time and 1 yen/minute elapsed time. Magnetic tape, including software, 30,000 yen per release or update.

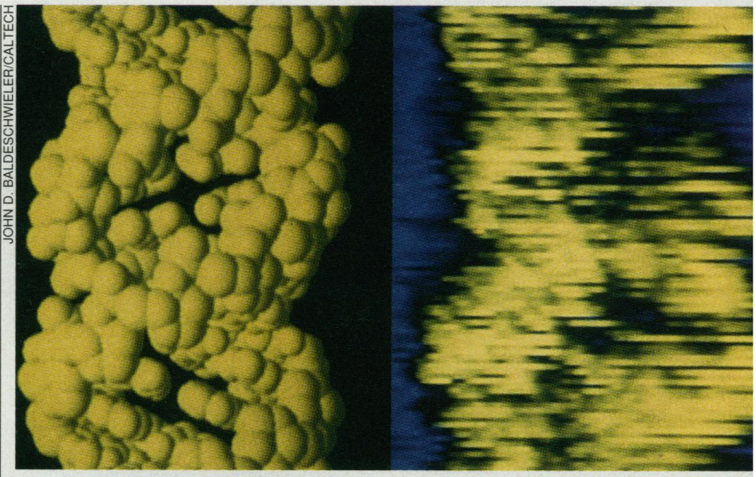
Contacts

Kathryn Sidman
NBRF
3900 Reservoir Rd. NW
Washington, DC 20007, USA
Phone: 202-687-2121
Fax: 202-687-1662
E-mail: pirmail@gunbrf.bitnet

MIPS
Hans-Werner Mewes, MPI/GEN
Max Planck Institut für Biochemie
8033 Martinsried, GERMANY
Phone: 49-89-8578-1
Fax: 49-89-8578-2655
E-mail: mewes@dm0mpb51.bitnet

JIPID
Akira Tsugita
Research Inst. for Biosciences
Science Univ. of Tokyo
Yamazaki, Noda 278, JAPAN
Phone: 81-471-23-9777
Fax: 81-471-22-1544
E-mail: tsugita@jpnsut31.bitnet

JOHN D. BALDESCHWIELER/CALTECH



Double helix. Atomic-scale image of the DNA double helix* (right) *compared to a computer graphic model.

PDB

Contents

3-D atomic coordinates from x-ray diffraction or NMR studies of each protein's several thousand atoms; secondary and crystallographic structure features; bond connectivity data; references. Size: 655 complete atomic coordinate entries (153 MB).

Data Source

Primarily direct computer-readable submission from researchers prior to, concurrent with, or after publication. Lag time: 10 months. Updates released and placed on-line quarterly.

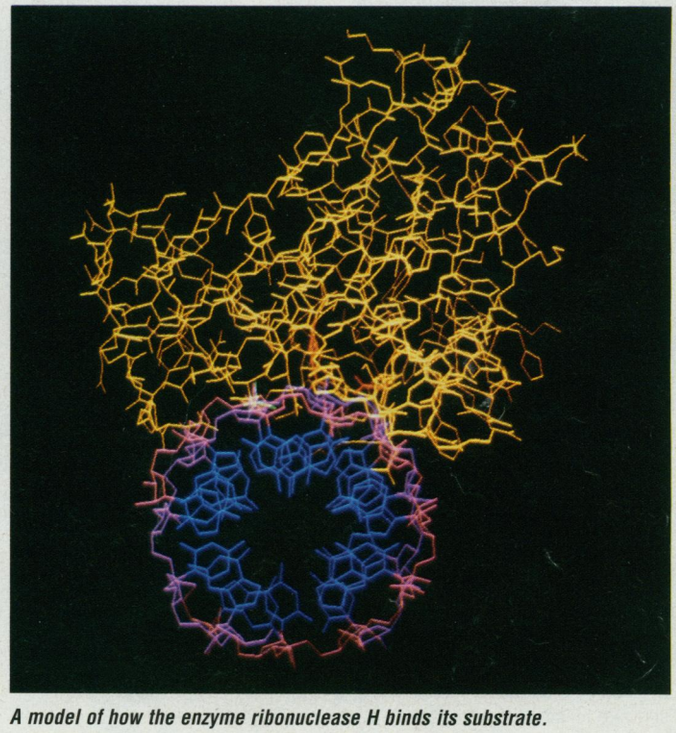
Format/Access/Cost

Free on-line access through several international distribution centers. Distributed on magnetic tape for $300-500 (depending on format) for initial release plus quarterly updates. Also available through commercial vendors.

Contact

Protein Data Bank
Chemistry Dept.—Bldg. 555
Brookhaven National Lab
Upton, NY 11973, USA
Phone: 516-282-3629
Fax: 516-282-5815
E-mail: pdb@bnlchm.bitnet

WAYNE HENDRICKSON/COLUMBIA



A model of how the enzyme ribonuclease H binds its substrate.

❖ BIOMAGRESBANK ❖

Contents
NMR spectroscopic data from solution NMR studies on proteins and protein fragments: chemical shifts; coupling constants; molecular species; sequence-related and atom-specific assignments; experimental conditions; citations; cross-references to PIR, PDB, CAS, CCSD. Covers 1,200 assignment groups, 20,000 shifts.

Data Source
Published literature; direct queries to authors for clarifications.

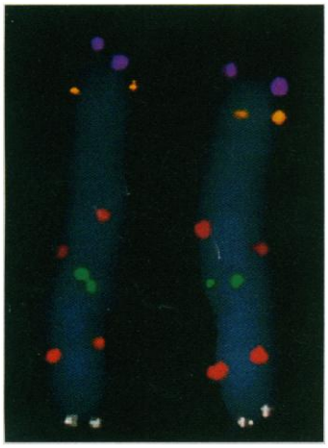
Format/Access/Cost
Prototype for personal computers, workstations now available; public distribution starting late 1991 will offer flat files or carrier files on magnetic tape or floppy diskette at minimal cost. On-line access planned.

Contact
B.R. Seavey or J.L. Markley
Biochemistry Dept.,
Univ. of Wisconsin
Madison, WI 53706-1569, USA
Phone: 608-263-9349
Fax: 608-262-3453

SWISS-PROT AND PROSITE

Contents
SWISS-PROT: 6 million base pairs of amino acid sequences from PIR, an automated translation of EMBL nucleotide sequences, and literature searches; taxonomic data; citations; annotations including protein function, modifications, functional domains and sites, similarities to other proteins, diseases associated with protein deficiency; cross-references to EMBL, PDB, PIR, OMIM™, PROSITE. Size: 12-15 MB data.
PROSITE: Biologically significant protein sequence sites and patterns—protein motifs; protein family; significance. Size: less than 1 MB data.

T. RIED AND D. WARD/YALE



Data Source
Published literature; research of database author (Amos Bairoch). Lag time: several weeks; updates released quarterly.

Format/Access/Cost
Distributed with service software on magnetic tape and CD-ROM with EMBL releases; data files downloadable from EMBL or GenBank® (see above).

Contact
EMBL or GenBank (see above)

Six DNA fragments mapped to human chromosome 5.

CSD

Contents
2-D chemical information; 3-D atomic coordinates; connectivity (molecular topology); crystal data; references. Size: nearly 90,000 entries (250 MB).

Data Source
Published literature; lag time 6-9 months. Updates available continuously on-line or distributed semi-annually.

Format/Access/Cost
On-line or other access to data and software free to academic users through National Affiliated Centers in 31 countries. Distributed directly with software to industrial users for a fee.

Contact
Olga Kennard
Cambridge Crystallographic Data Centre
Univ. Chem. Lab
Cambridge CB2 1 EW, UK
Phone: 44-223-336408
Fax: 44-223-312288
E-mail: DGW1@UK.AC.CAM.PHX (Bitnet)

CCSD AND CARBBANK

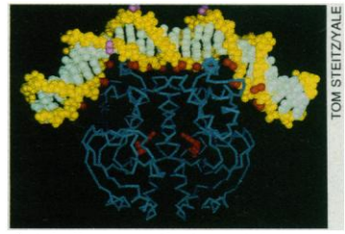
Contents
Polysaccharide, glycopeptide, and glycolipid primary structures larger than disaccharides; supplementary nonstructural information; references; cross-references to PDB. Size: 5,200 records (7.5 MB data) plus 1 MB software.

Data Source
Published literature collected by collaborating European and Japanese research teams and specialized curators in 30 countries; lag time approximately 6 months. Updates released semi-annually.

Format/Access/Cost
Distributed on IBM PC-compatible floppy diskettes: $500 for non-profit organizations; $1,500 for profit-making organizations. On-line access planned for 1992.

Special Features
CCSD graphically displays carbohydrate structures, indicating stereochemistry and configuration. The companion menu-driven program, CarbBank, allows users to manage and edit CCSD records, and search for text, complete structures, or substituents.

Contact
Dana Smith
Complex Carbohydrate Research Center,
Univ. of Georgia
Athens, GA 30602, USA
Phone: 404-542-4484
Fax: 404-542-4412
E-mail: CarbBank@UGA.bitnet or 76060.1127@Compuserve.com (Internet)



TOM STEITZ/YALE

Catabolic gene activator protein bound to duplex DNA.

BIBLIOGRAPHIC DATABASES

Four general bibliographic databases, available in many libraries, complement any genome data search. AGRICOLA covers agricultural literature and has recently increased coverage of plant genome data by scanning additional journals. BIOSIS Previews offers abstracts of biological literature in the entire life sciences area. CAS ONLINE compiles chemical abstracts and citations; its REGISTRY file includes more than 150,000 searchable protein and peptide sequences that have appeared in patent applications and published literature since 1957. MEDLARS offers many databases; its MEDLINE (medical literature abstracts) system covers most biomedical literature relevant for human genome research.

All are available on-line for direct modem communication through commercial on-line services and/or telecommunications networks; prices vary according to system, type of access, and length or type of search. BIOSIS is also available on CD-ROM, and AGRICOLA is available on both CD-ROM and magnetic tape.

Contacts

AGRICOLA
NAL Reference Branch, Rm. 111
10301 Baltimore Boulevard
Beltsville, MD 20705, USA
Phone: 301-344-4479
Fax: 301-344-5472

BIOSIS User Services,
2100 Arch St.
Philadelphia, PA 19103-1399, USA
Phone: 800-523-4806/215-587-4847
Fax: 215-587-2016

CAS Customer Service
P.O. Box 3012
Columbus, OH 43210, USA
Phone: 800-753-4227/614-447-3731
Fax: 614-447-3713

MEDLARS Management Section
NLM, Bldg. 38, Rm. 4N421
8600 Rockville Pike
Bethesda, MD 20894, USA
Phone: 800-638-8480
Fax: 301-496-0822

MODEL ORGANISMS

A major goal of the Genome Project is to map and sequence the genomes of several experimental organisms—including *E. coli*, mouse, nematode, and fruit fly. While nucleotide sequence data for many of these model organisms are contained in GenBank®, EMBL, and DDBJ, there is no central repository for their genetic map data. Numerous organism-specific databases have been initiated to collect such data; many attempt to link the genetic maps with physical maps or sequence data to provide powerful search and analysis capabilities.

GBASE, the Genomic Database of the Mouse, contains published genetic map data. The Encyclopedia of the Mouse Genome integrates various mouse genomic databases with software that generates graphical displays of genetic maps. *E. coli* Database (ECD) contains annotations to accompany nucleotide sequence data carried in GenBank®, EMBL, and DDBJ. The *Coli* Genetic Stock Center (CGSC) maintains genotypic descriptions and genetic map information for thousands of *E. coli* K12 strains. EcoSeq2, EcoMap, and EcoGene form a linked set of databases, still in prototype stage, integrating nucleotide sequences with physical and genetic map data. A *C. elegans* database (acedb) is the central repository of nematode data, combining genetic map data with physical map data and programs (formerly CEMAP). The Worm Community System (WCS), listed in "Beyond Databases," offers most acedb data with additional data and features. The *Drosophila* Information Database, still in the planning stages, will be a central repository for fruit fly data, integrating physical and genetic map data. The genetic map information alone is now available in machine-readable form as *Drosophila* Genetic Maps.

STAN SHORT/JACKSON LABORATORIES

MOUSE: ENCYCLOPEDIA OF THE MOUSE GENOME

Contents
Genetic linkage data and maps; cytogenetic map information; *in situ* hybridization; breakpoints for translocation and rearrangement; homologies for mice, humans, and 23 other species; references. Size: 2-3 MB data plus software.

Data Source
Incorporates various independently maintained existing databases. Lag time varies by source. Updates released semi-annually.

Format/Access/Cost
Present version runs only on color Sun SPARCstation™ using SunOS 4.0.3 or 4.1; distributed free with documentation on high-density 3 1/2" floppy diskettes or 1/4" cartridge tape. X-Windows™ version for Macintoshes to be available late 1991.

Contact
Janice Ormsby
Jackson Lab, 600 Main St.
Bar Harbor, ME 04609, USA
Phone: 207-288-3371, ext. 1394
Fax: 207-288-5079
E-mail: davidnaman@jax.org

Special Features
Combines data with graphical tools to create displays of cytogenetic or linkage maps separately or in parallel. A click of the computer mouse will then pull up windows showing related or more detailed maps and references or other annotations.

MOUSE: GBASE

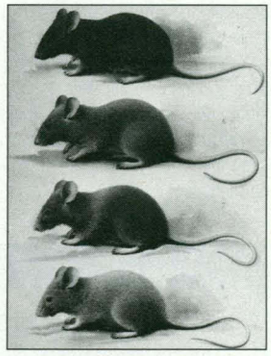
Contents
Mouse genetic map data: 2,400 mapped chromosomal loci and 800 unmapped loci; characterizations of alleles and polymorphic loci of 1,100 mouse strains; data on 600 mouse/human homologies; references; and data on effects of mutations in various loci (on-line version of *Mouse Locus Catalog*). Size: 18,000 lines of data (26 MB).

Data Source
Literature searches, direct submission of preprints or reprints by authors. Lag time 1-4 months; some direct submissions entered and accessible before publication. Data updates available immediately on-line; maps updated monthly.

Format/Access/Cost
Accessible on-line free via Internet from personal computers and terminals equipped with communications software. Free map printouts distributed monthly.

Contact
Thomas Roderick
Jackson Lab, 600 Main St.
Bar Harbor, ME 04609, USA
Phone: 207-288-3371
Fax: 207-288-5079
E-mail: thr@morgan.jax.org

Special Features
Accompanied by software that can generate laser-printable maps of the complete mouse genome or a subset of it, with overlays of known human homologies.



The inbred mouse.

E. COLI: ECD

Contents
Descriptive information to supplement *E. coli* nucleotide sequence data in EMBL, GenBank®: gene names; number of nonredundant base pairs; overlapping sequences; citations; EMBL and GenBank® cross-references. Ordered roughly according to linkage map position. Size: 1 MB.

Data Source
Existing databases; independent literature searches. Lag time varies.

Format/Access/Cost
Flat data files distributed free with EMBL releases; data and search program free on EMBL CD-ROM release or separately on floppy diskette.

Contact
EMBL (listed above) or
Manfred Kröger
Institut für Mikrobiologie und Molekularbiologie
Justus-Liebig-Universität Giessen,
Frankfurter Strasse 107
D-6300 Giessen, GERMANY
E-mail: kroeger@embl.bitnet

❖ E. COLI: CGSC ❖

Contents
Genotypic descriptions of 7,000 mutant derivatives of *E. coli* K-12: alleles; structural mutations; sex; and plasmids. Information on genes (sequenced and unsequenced), gene products, phenotypic properties, linkage map positions, references.

Data Source
Based on longstanding records of the *E. coli* Genetics Stock Center, curated by B. Bachman.

Format/Access/Cost
Database with prototype mapping utility runs on Sun workstations, can display genetic maps. Not yet distributed publicly, but responses to specific data requests supplied on disk, on tape, or via e-mail. Limited on-line access planned.

Contact
Mary Berlyn
Yale Univ., Osborn Memorial Lab
P.O. Box 6666
New Haven, CT 06511-7444, USA
Phone: 203-432-3536
Fax: 203-432-3854
E-mail: berlyn@yalemed.bitnet

E. COLI: ECOSEQ, ECOMAP, ECOGENE

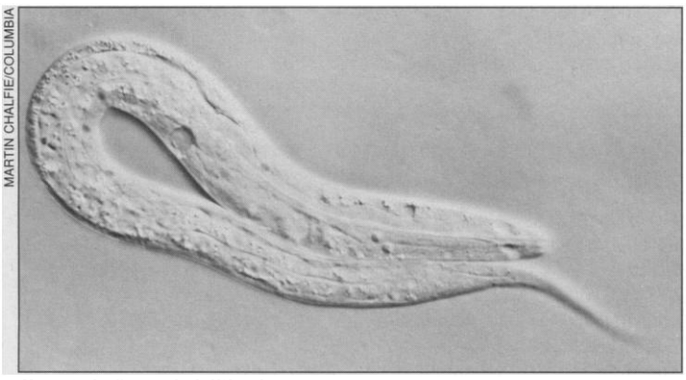
Contents
EcoSeq2 contains 1,644,238 base pairs of non-overlapping DNA sequences in FASTA format; EcoMap holds restriction map data, genetic map positions, and orientations of aligned DNA segments; EcoGene contains alignment and orientation information for over 1,100 *E. coli* genes. Linked by gene name and genomic position for cross-searching.

Data Source
Existing databases; literature sources; personal communications.

Format/Access/Cost
Flat data files with programs that run on Macintoshes, IBM PC-compatibles, and Unix™ machines free for downloading by anonymous.ftp via Internet or on floppy diskette (IBM or Macintosh format) or tape. Relational database version under development; CD-ROM distribution (with GenInfo, above) planned.

Contact
Kenneth E. Rudd
NCBI/NLM/NIH, Bldg. 38A, Rm. 8N805
8600 Rockville Pike
Bethesda, MD 20894, USA
Phone: 301-496-2475
Fax: 301-480-9241
E-mail: rudd@ncbi.nlm.nih.gov

NEMATODE: acedb


MARTIN CHALFIE/COLUMBIA

Nematode Caenorhabditis elegans.

Contents
Physical and genetic map data covering 900 localized genes and 21,000 cosmids; all known nucleotide sequences; references. Size: 25 MB.

Data Source
Contributed by collaborating laboratories; sequences often entered, accessible before publication; additional data verified, entered within weeks after publication. Updates released monthly; physical map updates available immediately.

Format/Access/Cost
Free on-line access via Internet (from various access nodes in North America and Europe) for use on Sun workstations; data, programs, and installation procedure can also be downloaded in ASCII format using anonymous.ftp. Macintosh™ version now being developed. Possible future CD-ROM distribution with GenInfo.

(See also WCS in Beyond Databases)

Contacts
Richard Durbin
MRC Molecular Bio. Lab
Cambridge CB2 2QH, UK
Phone: 44-223-402010
Fax: 44-223-402008
E-mail: rd@cele.mrc-lmb.cam.ac.uk

Jean Thierry-Mieg
Biochimie CNRS-INSERM
B.P. 5051
34033 Montpellier, FRANCE
Phone: 33-67-61-33-24
Fax: 33-67-52-15-59
E-mail: mieg@frmop11.bitnet

Special Features
Offers interconnected graphical displays of the genetic map, the physical map, and sequences, with click-of-the-mouse links to references, other annotations.

❖ DROSOPHILA INFORMATION DATABASE ❖

Contents (Planned)
Genetic map data (gene localizations, breakpoints, etc.); physical map data (cosmids, YACs, transcription units); stock availability; mutant descriptions (incorporating material from the *Drosophila* "Red Book"); bibliographic citations; cross-references to GenBank®, PIR.

Data Source
Existing databases and published sources, collected and/or contributed by collaborating researchers.

Format/Access/Cost
Windows-based system to include graphical display of chromosomal maps with click-of-the-mouse access to related information. To be available on-line through Internet and local access nodes, and downloadable through anonymous.ftp. Prototype release planned 1993.

Contact
T.C. Kaufman or K. Matthews
Howard Hughes Medical Inst.
Bio. Dept., Indiana Univ.
Bloomington, IN 47405, USA
Phone: 812-855-3033 or -5782
Fax: 812-855-2577
Email: KAUFMANTC@IUBACS or MATTHEWK@IUBACS

DROSOPHILA GENETIC MAPS

Contents
Genetic map data: loci; map position; gene products and function; cross-references to nucleotide and protein sequences; references. Size: 4,424 entries (1.3 MB). Aberration data: known aberrations, cytological breakpoints; references. Size: 11,488 entries (1.4 MB).

Data Source
Updated from *Drosophila Information Service*, volume 69. Lag time 1 month. Updates released semi-annually.

Format/Access/Cost
Free flat data files (ASCII format) can be downloaded from the Indiana Univ. or EMBL file servers or SEQNET (UK).

Contact
EMBL (see above), or
Michael Ashburner
Genetics Dept., Cambridge Univ.
Cambridge, CB2 3EH, U.K.
Phone: 011-44-223-333969
Fax: 011-44-223-333992
Email: ma11@phx.cam.ac.uk

❖ PLANTS ❖

The USDA plant genome research program is still in its infancy, but the National Agricultural Library (NAL) and the Agricultural Research Service (ARS) are gearing up to collect and maintain plant genome data through an experimental pilot database. In addition, NAL has expanded coverage of plant genome information in its bibliographic database, AGRICOLA (described above).

Contents (Planned)
Genome data (map information, sequences or references to GenBank®) for corn, soybeans, wheat, loblolly pine, possibly *Arabidopsis*. May add data on gene sets that control for complex phenotypic traits of characteristics of agricultural interest (yield, nutritional value, etc.), and pointers to germplasm sources and probes.

Data Source
Independent databases for each species, to be curated and contributed by collaborators.

Format/Access/Cost
Prototype for personal computer use may be available through NAL by 1992.

Contact
Susan McCarthy
Plant Genome Data & Info. Center
USDA/NAL/ISD, NAL Bldg., 14th floor
Beltsville, MD 20705, USA
Phone: 301-344-3875
Fax: 301-344-6098
E-mail: smccarthy@asrr.arsusda.gov (Internet)

Arabidopsis thaliana


ELIOT MEYEROWITZ/CALTECH

DATABASES OF DATABASES

Two databases of databases—the Listing of Molecular Biology databases (LiMB) and the Directory of Biotechnology Information Resources (DBIR)—offer additional and/or updated information on the selected databases listed here, as well as information on many data sources not covered.

DBIR

Contents
Descriptions of diverse biotechnology databases and information resources: contents; access; keywords; references; etc. 30% of listings cover international resources. Size: more than 1,600 entries (6 MB).

Data Source
Publications; other sources; personal communication. Existing entries reviewed, updated annually. Data entered within 1-2 months of verification. Revisions, additions released on-line monthly.

Format/Access/Cost
Available on-line for use on personal computers equipped with communications software through NLM's MEDLARS system (described below), by direct dial or through various telecommunication networks. Costs $6.00-26.00 per hour depending on time of day, type of connection, files searched.

Contact
NLM Specialized Info. Services Div.
8600 Rockville Pike
Bethesda, MD 20894, USA
Phone: 301-496-6531 or 496-1131

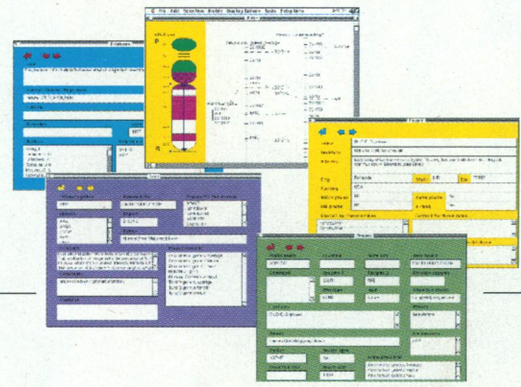
LIMB

Contents
Descriptions and access information covering more than 100 molecular biology databases: name; address; contents; collaborators; computer language, format and operating system; distribution format. Size: approximately 300K.

Data Source
Questionnaires filled out by database managers; publications, personal communication, or other secondary sources. Updates released sporadically.

Format/Access/Cost
Distributed free on floppy diskette for IBM PC-compatible computers (MS-DOS format with ASCII text files); by electronic mail; through various network servers; and in printed form.

Contact
Christian Burks
Los Alamos Nat'l Lab, T-10, MS K710
Los Alamos, NM 87545, USA
Phone: 505-667-6683
Fax: 505-665-3493
E-mail: limb@genome@lanl.gov



LAWRENCE BERKELEY LABORATORIES

BEYOND DATABASES

As the complexity and amount of genome data has increased, researchers have begun to move beyond conventional database structures and functions in a search for new ways to collect, link, share, analyze, and interpret data. Several prototype database systems offer not only the types of sophisticated analysis software and graphics displays incorporated by many of the databases listed above but also enable database users to add new data of their own, annotate or edit existing entries, indicate other users authorized to view or change that data, and share informal research information. The Chromosome Information System (CIS) and the Worm Community System (WCS) are two such prototypes.

❖ CIS ❖

CIS is designed for semi-private use by an individual laboratory or dispersed group of collaborators, rather than serving as a central public database. CIS offers a flexible database management system for genomic map information, allowing researchers to incorporate data from public sources and to add and define new data and data objects (such as pulsed-field gel images) of their own. Like computer-aided design systems for engineering, CIS allows changes in the graphic representation to be immediately reflected in the corresponding numeric data, and vice versa. (For example, a researcher viewing a map of a chromosome with markers X, Y, and Z could decide that the correct order was X, Z, Y, then move the image with a click of the mouse; corresponding numerical data describing the marker position would change simultaneously.) The prototype version was used to help display consensus maps at recent workshops for human chromosomes 3 and 21; its applicability for non-human organisms is being explored. CIS runs as a distributed software application in a TCP/IP network environment and can be accessed by collaborators over Internet. An X-Windows™ version for Unix™ workstations is being developed.

Contact
LBL Genome Computing Group
Lawrence Berkeley Lab, MS 50B-3238
Berkley, CA 94720, USA
E-mail: cis@lbl.gov (Internet) or cis@lbl (Bitnet)

❖ WCS ❖

WCS is an experiment in "building an electronic scientific community," a computer system that offers traditional database functions along with literature, informal information and research lore, mapping programs and graphics, and the ability for users to add their own notes and data into the system. WCS incorporates genomic data from the same sources as acedb (see above), then adds and interlinks other sources: a bibliographic database with citations and abstracts; electronic indexed versions of the informal but informative research newsletter, the *Worm Breeder's Gazette*; and other unpublished data (strain lists, a lab directory, etc.). In contrast to CIS, which aims to serve small groups of collaborators, WCS is intended to serve an entire community of more than 400 researchers, from experienced and specialized worm genome researchers to those new or peripheral to the field. WCS runs under X-Windows™ on Unix™ machines and can also be used remotely from Sun and Dec workstations and Macintosh™ personal computers. A free copy of the preliminary release for Unix™ workstations is available to interested researchers. Future releases will be widely accessible on-line across NSFNet.

Contact
Bruce Schatz
Life Sciences South Bldg., Univ. of Arizona
Tucson, AZ 85721, USA.
Phone: 602-621-9174; Fax: 602-621-3709
E-mail: schatz@cs.arizona.edu