Bioinformatics

Mark S Boguski

National Center for Biotechnology Information, Bethesda, USA

Computer databases, networks and software tools are essential materials and methods for biomedical research and are involved in almost every aspect of disease gene mapping and positional cloning. Public databases of DNA and protein sequences and genetic and physical map information are increasing rapidly in size and complexity and are also improving in quality, comprehensiveness, interoperability and access. A new generation of software tools for navigating through the biomedical literature has become available. Programs for sequence homology searching and genetic map construction have become more sophisticated, yet easier to use. Global computer networks are bringing state-of-the-art capabilities to all.

Current Opinion in Genetics and Development 1994, 4:383–388

Introduction

Information science and technology (informatics) became an essential concern to biologists following the development of rapid DNA sequencing methods in the mid-1970s. Fueled by the anticipated data deluge from the Human Genome Project, and assisted by the general advancement of computer hardware, software and networking technologies, bioinformatics is undergoing a period of explosive growth and development, a snapshot of which will be presented in this review. Bioinformatics is broad in scope and involves everything from laboratory automation and data acquisition to electronic publishing. The focus of this review will be the selected information analysis and retrieval tools which occupy a central role for all researchers that concern themselves with genetic and molecular sequence data and the biomedical literature.

The information landscape

The growth of information of potential interest to biomedical researchers is shown in Fig. 1. MEDLINE is an international bibliographic database provided by the US National Library of Medicine (NLM). MEDLINE currently contains over 7 million abstracts (Fig. 1a) derived from published articles in approximately 4000 scientific journals. NLM indexes nearly 400 000 articles per year. The literature devoted to genetics is growing at a rate somewhat faster than the biomedical literature as a whole and now contains over 60 000 articles (Fig. 1b). However, the number of articles directly relevant to molecular biology, biotechnology and genetics is at least three times greater than this, as evidenced by the number of MEDLINE records included in release 10.0 of the National Center for Biotechnology (NCBI) *Entrez* retrieval system (see below).

Gene mapping data is growing exponentially (Fig. 1c), as are data on nucleotide sequences, protein sequences and protein crystal structures (Fig. 1d–f). The Genome Data Base (GDB) is an on-line encyclopedia of human cytogenetic, linkage and physical mapping information [1]. For some of the latest results and data sets (e.g. [2•]), however, one also needs to be aware of a number of other repositories (Table 1). The amount of DNA sequence data in GenBank® [3] is doubling every 20 months, and this is even before the full impact of concerted genome sequencing is felt [4].

Over the past three years, an important new type of sequence data has become available [5]. In order to rapidly survey the expressed genome, several groups are using automated sequencing technology to generate 'single pass', partial cDNA sequences —'expressed sequence tags' (ESTs) — from a variety of animal and plant species [6,7]. Because of some of the unique characteristics of these data, a special database (dbEST) has been set up to facilitate usage by contributors and end-users alike [8•]. Release 2.10 of dbEST contains more than 34 000 sequences, with the majority coming from *Homo sapiens* (human), *Caenorhabditis elegans* (nematode), *Arabidopsis thaliana* (thale cress), *Oryza sativa* (rice), *Plasmodium falciparum* (malarial parasite) and *Zea mays* (maize). ESTs are being system-

Abbreviations

BLAST—Basic Local Alignment Search Tool; **CEPH**—Centre d'Étude du Polymorphisme Humain; **CERN**—Conseil Européen pour la Recherche Nucléaire; **CHLC**—Cooperative Human Linkage Center; **db**—database; **EST**—expressed sequence tags; **ftp**—file transfer protocol; **GBD**—Genome Data Base; **NCBI**—National Center for Biotechnology Information; **NCSA**—National Center for Supercomputer Applications; **NLM**—National Library of Medicine; **STS**—sequence-tagged site; **WWW**—World-Wide Web; **YAC**—yeast artificial chromosome.

ISSN 0959-437X

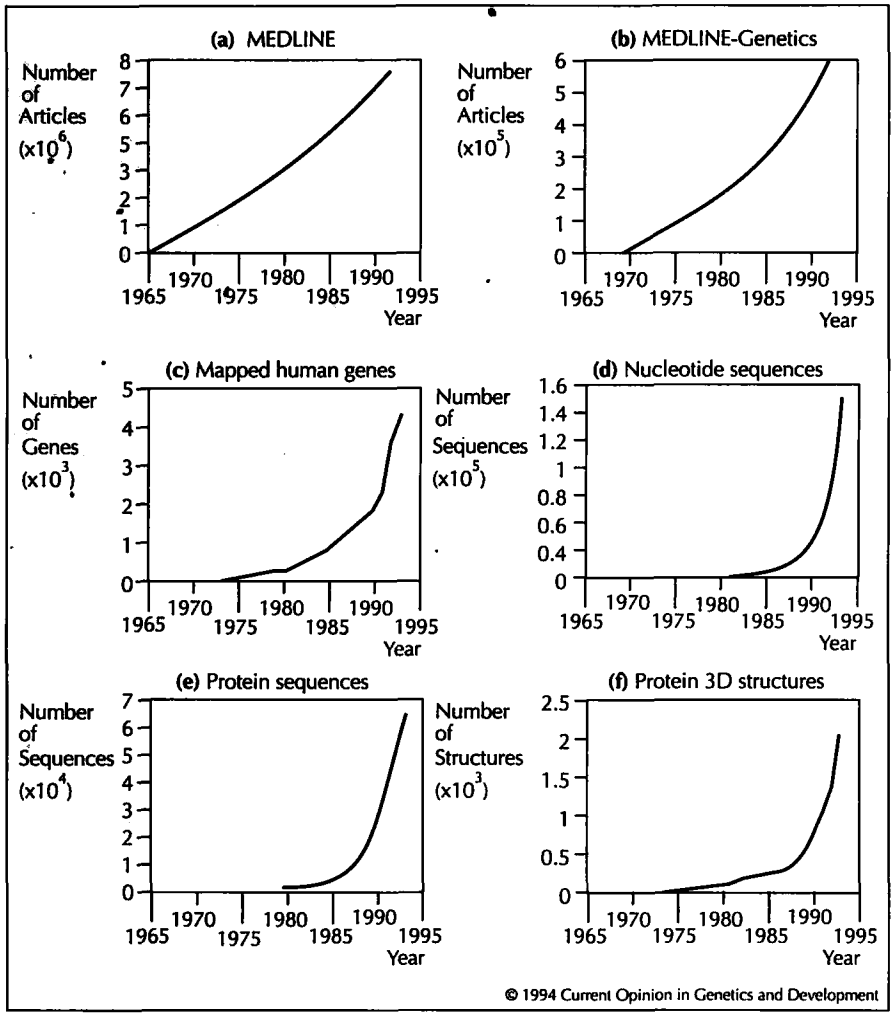

Fig. 1. Cumulative growth of biomedical research data. **(a)** Growth of the MEDLINE and **(b)** growth of its 'G5' (genetics) subset. **(c)** Growth in the number of mapped human genes compiled by Dr A Jamie Cuticchia from various Human Genome Mapping Workshop reports and the Genome Data Base (GDB) [1]. Recent data on the numbers of polymorphic loci and sequence-tagged sites (STSs) may be found in [28]. **(d–f)** Growth in the number of nucleotide sequences (d), protein sequences (e) and protein structures (f) lodged in GenBank® [3], PIR International [29] and the Brookhaven Protein Data Bank [30], respectively.

atically mapped [9], are becoming increasingly useful for identifying candidate genes (see below), and can greatly streamline the search for cognate genes among different organisms [10,11].

All of these data are available in electronic databases, but until recently researchers had been faced with the daunting task of mastering a number of complex search and retrieval systems and then integrating the information manually, post hoc. All of this has now changed as a result of a number of informatics efforts, some as part of the general development of computerized information resources and some under the direct auspices of biological research. An example of the latter case is GenBank® which has, under new management [3], undergone a major expansion in data coverage and annotation content and developed a host of new services for electronic data submission, sequence similarity searching and integrated information retrieval.

Exploring the information space

Certainly the most significant development in bioinformatics over the past year has been the tremendous proliferation of Internet resources devoted to the life sciences. Indeed, one 'virtual library' lists over 200 different databases or services for biology and medicine (see below). The Internet itself is a global assemblage of more than 30 000 interconnected computer networks, including US government, private and international networks, that has evolved over more than 20 years and is neither owned nor operated by any single organization [12••]. Internet resources consist of 'server' computers containing information (file servers), executable programs (compute servers), or both. These servers interact with 'client' computers which request information or services. The easiest (and most limited) way to access these services is through electronic mail (e-mail), but computer workstations and increasingly powerful personal computers running graphical user interfaces (GUIs) offer the most impressive capabilities, which include audio and video communications.

It is impossible to capture in writing the richness and complexity of the Internet, so I will try to identify only a few ways and means by which to enter this 'information space' and try to point out some of the most useful resources for genetics and molecular biology (Table 1). An excellent tutorial on the Internet and its navigation may be found in [12••].

Table 1. Selected World-Wide Web (WWW) and Gopher servers.

Resource	WWW Uniform Resource Locator or Gopher address	Features and comments
World-Wide Web	http://info.cern.ch/	WWW project background; pointers to the world's on-line information; information on WWW software products; frequently asked questions (and answers).
NCSA Mosaic	http://www.ncsa.uiuc.edu/	Starting points for Internet exploration; Internet resources meta-index; "What's New" with NCSA and Internet.
Virtual Library: Biosciences	http://golgi.harvard.edu/	Comprehensive list and links to WWW resources for biology and medicine.
ExPASy	http://expasy.hcuge.ch/	Comprehensive library of documents describing e-mail servers, databases and software for molecular biology.
NCBI GenBank®	http://ncbi.nlm.nih.gov/	Information for submitting and updating sequences; GenBank® release notes; homology and text searching of sequence databases; *Entrez* MEDLINE browser.
Genome Data Base (GDB)	http://gdbwww.gdb.org/	Search GDB and On-line Mendelian Inheritance in Man (OMIM)
CEPH/Généthon	http://www.genethon.fr/	Human genome genetic and physical mapping data; search for information on YAC clones and STSs; QUICKMAP software for viewing CEPH map.
Cooperative Human Linkage Center (CHLC)	gopher://gopher.chlc.org/	Markers, genotype data and integrated maps.
UK Human Genome Mapping Project	gopher://menu.crc.ac.uk/	Primers, probes and chromosome abnormality database; cell lines from patients with genetic disorders or cytogenetic abnormalities.
MIT Genome Center	http://www-genome.wi.mit.edu/	Quarterly data releases from the human physical mapping and mouse genetic mapping projects.
Jackson Laboratory	http://www.jax.org/	Locus catalog and genetic maps of the mouse.
Dan's Favorite BioGopher	gopher://gopher.gdb.org/	Library catalogs around the world; world-wide campus phonebooks; search for people by name, location, research interest or funding agency.

Gopher, originally developed at the University of Minnesota, is an Internet search client that allows one to browse for resources on different servers, using menus, and to retrieve files of interest. Gopher client software is available free for all popular computers and can be obtained by anonymous file transfer protocol (ftp) — see [12••] or your local systems administrator for more details.

The World-Wide Web (WWW) originated at CERN, the European Laboratory for Particle Physics near Geneva, and is similar to Gopher in the sense that it is a client-server information system running over the Internet. WWW is based on a more versatile data model, however, which allows 'hypertext' links, or cross-references, among related information resources [12••]. The WWW server at CERN (Table 1) lists a variety of client programs that allow one to access WWW directly from a personal computer, including line-mode browsers for users of 'dumb' terminals, as well as full-screen browsers for use on VT100 terminals. One particularly notable client is Mosaic, produced by the US National Center for Supercomputer Applications (NCSA). Mosaic, available for X-windows/UNIX, Apple Macintosh, Microsoft Windows and other systems, utilizes a graphical user interface that permits retrieval of sound and images and demonstrates the potential for WWW as a distributed hypermedia system.

In my experience, Gopher and WWW servers are outstanding tools for browsing available resources and for finding specific items when one is uncertain about their accessibility or location. A sampling of Internet biology resources, available through Gopher or WWW, is pre-

sented in Table 1. For resources that one uses heavily and routinely, however, special, dedicated applications are often more efficient and superior in performance. Several of these are described below.

Network resources for analysis, search and retrieval

Every researcher searches the literature and almost every researcher needs to perform sequence 'homology' searching at one time or another. Somewhat smaller numbers need frequent access to protein crystal structures or consensus genetic map information, and only specialized research groups routinely compute on atomic coordinates or raw genetic and physical mapping data. The availability and ease of use of many informatics tools (and the data itself) are generally proportional to this distribution of end-users.

Many useful programs for database searching, sequence analysis and information retrieval are available via e-mail. Because of limited space, and my emphasis on the newer client-server technology, these will not be discussed here. The reader is directed to the short review by Henikoff [13], to the Indiana University BioGopher (ftp.bio.indiana.edu) and to the ExPASy WWW server (Table 1) for more information on e-mail services.

Integrated information retrieval

Entrez, produced by NCBI, is a versatile information retrieval system that integrates DNA and protein sequence data with MEDLINE abstracts in a graphical user interface, mouse-driven, point-and-click environment that runs on Apple Macintosh, Microsoft Windows and UNIX/X-windows computers [14•,15]. *Entrez* links the coding sequences of genes with their protein products and relates both types of sequence to the published literature. In addition, each release of *Entrez* includes computed 'homologies' for all DNA sequences and protein sequences derived from Basic Local Alignment Search Tool (BLAST) similarity searching (see below) of each database against itself. Relationships among MEDLINE records, and automatic retrieval of similar articles, are also available through a process known as 'neighboring', based on term frequency statistics. *Entrez* is available on CD-ROM (six releases per year) or as a free Internet service [14•]. The current release of *Entrez* contains 199956 MEDLINE records, 161160 DNA and 172595 protein sequences. For users of non-graphical systems, a text-based 'command line *Entrez* version' (CLEVER) has been developed by T Littlejohn, P Rioux and W Gilbert (contact ogmp@bch.umontreal.can by e-mail).

Homology searching

Sequence similarity (homology) searching is probably the most important analytic method that molecular biologists use and much effort has been directed at its development, usually focusing on the speed and sensitivity of the underlying algorithm. Recently, it has become apparent that a number of associated issues are equally important for maximizing the efficiency of search procedures [16•], including access to up-to-date, non-redundant sequence collections, improved scoring systems and methods for assessing the statistical significance of sequence similarities, and various types of sequence 'masking' to reduce or eliminate voluminous, spurious results that arise from locally biased amino acid composition, the presence of *Alu* repeats, *et cetera* in query sequences. These issues become critical for high-throughput, genome-scale sequencing efforts in which automated or semi-automated data analysis is required. Most of these features are embodied in the BLAST family of programs, which are available as a free network service from NCBI [16•].

Successful positional cloning depends upon the ability to locate exons or potential coding regions by either experimental or computational means (see review by Monaco [pp 360–365]). Sequence data derived from either 'genome survey sequencing' or exon 'trapping' or amplification procedures need to be analyzed by database homology searching at the earliest opportunity to identify candidate genes. Error-tolerant algorithms (BLASTX [17]) and scoring systems [18] are available for this purpose.

Exon identification

The best evidence for an exon (on the basis of sequence analysis alone) is the existence of a statistically significant similarity to a known gene product or EST. Some evidence suggests that most gene products that are shared across phyla are already present in the sequence databases [19]. This finding, together with the continuing rapid accumulation of partial cDNAs [8•], indicates that the chances of detecting exons by sequence similarity are getting better all the time. In the absence of similarity, however, methods for the identification of putative exons from intrinsic properties of sequences are essential. These properties include statistical differences between coding and non-coding nucleotide sequences, the presence of potential promoter sequences, splice junctions and polyadenylation signals, and other sequence attributes.

XGRAIL is a client-server implementation of a popular application for sequence analysis and the construction of gene models on the basis of intrinsic sequence characteristics [20]. XGRAIL client software is available for UNIX workstations from Sun Microsystems and Digital Equipment Corporation (e-mail to grailmail@ornl.gov). GeneID is another system for analyzing vertebrate DNA with predictions of exons and gene structure [21] and is available through a

Gopher interface (dna.cedb.uwf.edu). GeneID automatically tests predicted exons using a BLASTX [16•,17] protein database search; XGRAIL features several similarity search options.

Mapping data

GDB/Accessor is a new client-server program for searching the GDB (Table 1) and related databases, such as On-line Mendelian Inheritance in Man (OMIM), GenBank®/EMBL and SWISS-PROT. The software (Macintosh only) and a detailed user guide are available for free (e-mail to help@gdb.org).

Finally in this review, I describe several programs for working with chromosome mapping data. In contrast to the tools described above, these are not Internet client-server applications, but are included because they are important tools for gene mapping and positional cloning. QUICKMAP is a new graphical user interface for CEPH mapping data and allows one to visualize sequence-tagged sites (STSs) and yeast artificial chromosome (YAC) contigs anchored on the genetic map. (YAC contigs are clones whose inserts overlap.) QUICKMAP runs on Sun workstations and is so new that only minimal documentation was available at the time of writing (e-mail to rigault@genethon.fr for more information). MultiMap [22] is an 'expert system' that automates genetic linkage map construction. It requires a Lisp interpreter and the CRI-MAP program [23] on a Sun workstation (e-mail to multimap@genome1.hgen.pitt.edu). The Cooperative Human Linkage Center (CHLC; see Table 1) provides an e-mail server (contact info-server@chlc.org) to which one can send genotype data from CEPH reference families and obtain information on linked markers from the CHLC data sets [24]. SEGMAP is a program for constructing physical maps of chromosomes based on the STS content of YAC contigs [25]. SEGMAP is available for UNIX workstations (e-mail to pg@genome.wustl.edu).

Conclusions

Given the rapid accumulation of data and the proliferation of new software tools and network resources, this review will be dated before the ink is dry. The only way to keep abreast of important new developments is to master some of the instruments introduced herein and use them regularly. I would also recommend subscribing to the BIOSCI newsgroups [26]. In the future, however, 'intelligent agents' (software robots) will be programmed with our individual interests and will continuously scan the information space, automatically notifying us when any relevant new data or observations become available. This capability will fundamentally change the nature of scientific communication [27].

Acknowledgements

I thank Rose Marie Woodsmall, Steve Bryant and A Jamie Cuticchia for helping to obtain the data plotted in Fig. 1 and Francis Ouellette for suggestions on the manuscript.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Cuticchia AJ, Fasman KH, Kingsbury DT, Robbins RJ, Pearson PL: **The GDB Human Genome Data Base Anno 1993.** *Nucleic Acids Res* 1993, 21:3003–3006.

2. • Cohen D, Chumakov I, Weissenbach J: **A First-Generation Physical Map of the Human Genome.** *Nature* 1993, 366:698–701.

The full text of this article is electronically available on the CEPH/Généthon World-Wide Web (WWW) server (see Table 1). WWW hypertext links exist via D-segment numbers to the Genome Data Base (GDB) for a subset of markers.

3. Benson D, Lipman DJ, Ostell J: **GenBank.** *Nucleic Acids Res* 1993, 21:2963–2965.

4. Collins F, Galas D: **A New Five-Year Plan for the US Human Genome Project.** *Science* 1993, 262:43–46.

5. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al.*: **Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project.** *Science* 1991, 252:1651–1656.

6. Sikela JM, Auffray C: **Finding New Genes Faster Than Ever.** *Nature Genet* 1993, 3:189–191.

7. Davies K: **The EST Express Gathers Steam.** *Nature* 1993, 364:554.

8. • Boguski MS, Lowe TM, Tolstoshev CM: **dbEST — Database for "Expressed Sequence Tags".** *Nature Genet* 1993, 4:332–333.

Although all sequences in dbEST are also available in the EST Division of GenBank®, dbEST contains more value-added annotation (the latest homologies, mapping data and contact information for individual clones). Homology searching, Boolean queries and retrieval of ESTs by map location are available. One can access dbEST through the World-Wide Web (http://www.ncbi.nlm.nih.gov/) and by e-mail to retrieve@ncbi.nlm.nih.gov (type 'help' in the message body).

9. Polymeropolous MH, Xiao H, Sikela JM, Adams M, Venter JC, Merril CR: **Chromosomal Distribution of 320 Genes from a Brain cDNA Library.** *Nature Genet* 1993, 4:381–386.

10. Tugendreich S, Boguski MS, Seldin MS, Hieter P: **Linking Yeast Genetics to Mammalian Genomes: Identification and Mapping of the Human Homolog of CDC27 via the Expressed Sequence Tag (EST) Data Base.** *Proc Natl Acad Sci USA* 1993, 90:10031–10035.

11. Papdopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, Haseltine WA, Fleichmann RD, Fraser CM, Adams MD, *et al.*: **Mutation of a mutL Homolog in Hereditary Colon Cancer.** *Science* 1994, 263:1625–1629.

12. •• Krol E: ***The Whole Internet User's Guide & Catalog***, edn 2. Sebastopol, California: O'Reilly & Associates Inc; 1994.

If you read only one reference, it should be this one. The book, now in its second edition, is one of the best practical guides to the expanding electronic universe and may be ordered by e-mail from order@ora.com.

13. Henikoff S: **Sequence Analysis by Electronic Mail Server.** *Trends Biochem Sci* 1993, 18:267–268.

14. • **Network *Entrez*.** *NCBI News* 1993, 2:1.

NCBI News contains the latest information on *Entrez*, BLAST, and GenBank® access and direct submission and other databases and software tools. Free subscriptions may be obtained by contacting info@ncbi.nlm.nih.gov or 0101-301-496-2475.

15. Cockerill M: A Versatile Tool for Retrieving Molecular Sequences. *Trends Biochem Sci* 1994, 19:94–96.

16. • Altschul SF, Boguski MS, Gish W, Wootton JC: Issues in Searching Molecular Sequence Databases. *Nature Genet* 1994, 6:119–129.
BLAST is the de facto standard for sequence homology searching. This review covers the BLAST algorithm, scoring systems, alignment statistics and the new technique of query masking that dramatically improves the signal-to-noise ratio in database searches. The design, construction and availability of non-redundant sequence databases is also discussed. BLAST searches can be performed via e-mail or an Internet client-server application. Contact info@ncbi.nlm.nih.gov for more information.

17. Gish W, States DJ: Identification of Protein Coding Regions by Database Similarity Search. *Nature Genet* 1993, 3:266–272.

18. Claverie J-M: Detecting Frameshifts by Amino Acid Sequence Comparison. *J Mol Biol* 1993, 234:1140–1157.

19. Green P, Lipman D, Hillier L, Waterston R, States D, Claverie JM: Ancient Conserved Regions in New Gene Sequences and the Protein Databases. *Science* 1993, 259:1711–1716.

20. Mural RJ, Einstein JR, Guan X, Mann RC, Uberbacher EC: An Artificial Intelligence Approach to DNA Sequence Feature Recognition. *Trends Biotechnol* 1992, 10:66–69.

21. Guigo R, Knudsen S, Drake N, Smith T: Prediction of Gene Structures. *J Mol Biol* 1992, 226:141–157.

22. Matise TC, Perlin M, Chakravarti A: Automated Construction of Genetic Linkage Maps Using an Expert System (MultiMap): a Human Genome Linkage Map. *Nature Genet* 1994, 6:384–390.

23. Lander ES, Green P: Construction of Multi-Locus Genetic Linkage Maps in Humans. *Proc Natl Acad Sci USA* 1987, 84:2363–2367.

24. Buetow KH, Weber JL, Ludwigsen S, Scherpier-Heddema T, Duyk GM, Sheggield VC, Wang Z, Murray JC: Integrated Human-Wide Maps Constructed using the CEPH Reference Panel. *Nature Genet* 1994, 6:391–393.

25. Green ED, Green P: Sequence-Tagged Site (STS) Content Mapping of Human Chromosomes: Theoretical Considerations and Early Experiences. *PCR Methods Appl* 1991, 1:77–90.

26. Bleasby A, Griffiths P, Hines D, Marshall S, Staniford L, Hoover K, Kristofferson D: The BIOSCI Newsgroups — Computer Networks Changing Biology. *Trends Biochem Sci* 1993, 18:310–311.

27. Boguski M, McEntyre J: I Think Therefore I Publish. *Trends Biochem Sci* 1994, 19:71.

28. Cuticchia AJ, Chipperfield MA, Porter CJ, Kearns W, Pearson PL: Managing All Those Bytes: the Human Genome Project. *Science* 1993, 262:47–48.

29. Barker WC, George DG, Mewes H-W, Pfeiffer F, Tsugita A: The PIR-International Databases. *Nucleic Acids Res* 1993, 21:3089–3092.

30. *PDB Quarterly Newsletter* October 1993 (e-mail pdb@bnl.gov).

M Boguski, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, Maryland 20894 USA. E-mail: boguski@ncbi.nlm.nih.gov