BNL--48477

DE93 006673

The Protein Data Bank - Present Status and Future Plans*

Thomas F. Koetzle, Enrique E. Abola, Frances C. Bernstein, Judith A. Callaway, Joseph C. Christian, Betty R. Deroski, Pamela A. Esposito, Arthur Forman, Patricia A. Langdon, John E. McCarthy, Regina K. Shea, John G. Skora, and Karen E. Smith

Chemistry Department, Brookhaven National Laboratory, Upton, NY 11973, USA

Abstract

The Protein Data Bank (PDB) archival database of three-dimensional structures of biological macromolecules, an international resource facility, contains information on protein, DNA, RNA, virus and carbohydrate structures. While the vast majority of PDB entries represent crystal structures, results from NMR and theoretical modeling studies also are included.

PDB, which in July 1992 contained 957 atomic coordinate entries, currently is experiencing a time of explosive growth. The present deposition rate is *ca.* 50 structures per month, doubling in less than two years. Responding to the challenge posed by this rising data flow, over the past 18 months PDB has attracted increased funding to implement important enhancements of the resource. A rapid pre-release of entries pending for input was inaugurated in April 1992, and a substantial fraction of the accumulated backlog of pending entries is now available *via* FTP and e-mail in pre-release form.

Extrapolation of current data rates suggests that by the year 2000 PDB may contain over 25 000 structures. PDB's plans, to manage this voluminous amount of data, include the development of PDB-AUTHORIN software to allow depositors to do most of the preparation and validation of their own entries, and a comprehensive upgrade of PDB contents to add new data items and convert the current interchange format to the Crystallographic Information File (CIF) standard established by the International Union of Crystallography (IUCr).

The Protein Data Bank - Present Status and Future Plans*

Thomas F. Koetzle, Enrique E. Abola, Frances C. Bernstein, Judith A. Callaway, Joseph C. Christian, Betty R. Deroski, Pamela A. Esposito, Arthur Forman, Patricia A. Langdon, John E. McCarthy, Regina K. Shea, John G. Skora, and Karen E. Smith

Chemistry Department, Brookhaven National Laboratory, Upton, NY 11973, USA

The Protein Data Bank (PDB) archival database for three-dimensional structures of biological macromolecules is an international resource facility located at Brookhaven National Laboratory (BNL), New York, USA. In operation since 1973, PDB distributes a database containing atomic coordinates, bibliographic citations, and primary sequence and secondary structure information, as well as crystallographic structure factors and 2D-NMR experimental data. Information is available on protein, DNA, RNA, virus and carbohydrate structures, and the primary mission of PDB is to collect, standardize, and provide access to this data. The PDB is international in scope, with affiliated centers and collaborations with groups in Canada, Europe, Japan, and the USA.

The July 1992 PDB release includes 957 atomic coordinate entries (855 proteins, enzymes, and viruses; 82 DNA's (*i.e.*, oligonucleotides); 10 RNA's; and 10 carbohydrates). Of these, 873 represent crystal structures, 41 are from NMR, and the remaining 43 entries describe results of theoretical modeling studies.

PDB currently is experiencing a time of explosive growth, in which the number of entries deposited is rising exponentially. The current deposition rate is *ca.* 50 new structures per month, doubling in less than two years (*cf.*, Figure 1). PDB has in the past been unable to keep up, and by the end of last year a backlog of *ca.* 300 atomic coordinate entries were pending for input (*cf.*, Figure 2). The mean time lag from deposition to release had reached 8 months, clearly an unacceptable delay.

Responding to the challenge posed by the rising data flow, over the past 18 months PDB has attracted increased funding to implement important enhancements of the resource. Additional staff have been hired, the PDB computing environment has been upgraded from VAX/VMS to high-performance SGI/UNIX workstations linked with powerful compute and file servers, resulting in markedly improved data processing throughput, and a rapid pre-release of atomic coordinate entries has been initiated *via* FTP and e-mail. To manage the increased amount of data, the SYBASE relational database management system (RDBMS) now is employed for key data entry functions and to track entries from deposition to final release. Substantial enhancements of data processing software, used to verify and validate data items, have been realized upon migration to the UNIX workstation environment. On-line access is now provided to bioinformatics resources such as the bibliographic and sequence databases maintained at the National Center for Biotechnology Information (NCBI), National Library of Medicine (*e.g.*, MEDLINE, PIR, SWISSPROT) so that, for example, primary sequence data submitted to PDB can be checked automatically against the relevant sequence databases.

To accomplish the rapid pre-release, inaugurated in April 1992, PDB data validation software was adapted so that syntax and stereochemical checks could run essentially automatically, supplemented by a quick visual inspection. Any problems encountered at this stage must be addressed by the depositor(s) before an official PDB

* PDB is supported by a combination of US Government agency funds (work supported by the US National Science Foundation; the US Public Health Service, National Institutes of Health; and the US Department of Energy under contract DE-AC02-76CH00016) and user fees.

entry ID code is issued. Four weeks should usually be a sufficient period for depositors to respond to communications from PDB, after which time pre-release entries are expected to be available *via* FTP and e-mail. The status as of the July 1992 release is summarized in Figure 3. At that time 189 pre-release entries, representing about 40 percent of the backlog, were available and an additional 169 entries were awaiting depositor approval. The number of entries in preparation for pre-release was reduced to 62, or just a bit more than one month's data flow.

The 47 entries 'on hold' in Figure 3, *i.e.*, those where the depositor has requested that the data be withheld from public distribution for a specified period of time, represent slightly over 3 percent of the total. The International Union of Crystallography (IUCr) has recommended limitations in the period of time that data may be withheld of one year for atomic coordinates and four years for structure factors [*cf.*, *Acta Cryst. A45*, 658 (1989)]. These hold limitations have now been adopted in the US as official NIH grants and contracts policy. Currently, virtually all leading scientific journals worldwide publishing results from relevant crystallographic studies require deposition with the PDB as a prerequisite for publication.

Distribution of the PDB database has increased markedly in recent years. BNL distributed 630 copies on magnetic tape in 1991. The master release tapes are updated quarterly, to coincide with publication of the PDB Newsletter (worldwide circulation 5000). In-between quarterly releases, newly approved entries are added to the PDB e-mail server and FTP at BNL on a regular basis. Access to the PDB worldwide is provided by 11 affiliated centers, 5 in North America, 4 in Europe, and 2 in Japan. These centers, together with a number of commercial partners marketing 'value-added' products based on PDB, form a Protein Data Bank Service Organization (PDBSA).

In the fall of 1992, the PDB will formally inaugurate regular releases of the database on CD ROM.[1] A sample CD ROM recently has been distributed on a trial basis, and based on comments received it is expected that CD ROM availability will substantially broaden the PDB user base, particulary on PC's. BNL is considering the possibility of producing sufficient numbers of CD ROM's to allow these to be distributed by PDBSA centers as well as by BNL.

Also, in the fall of 1992, access to all atomic coordinate entries via e-mail and FTP will be provided by BNL. Initially, the number of entries permitted to be downloaded in a session will be limited by available bandwidth. However, in 1993 the PDB intends to acquire a direct dedicated T3 internet connection. Once this occurs, downloading of the entire atomic coordinate entry database (now 270 Mbytes) should be quite feasible. This will facilitate frequent updating *via* FTP of the database copies stored at PDBSA centers. These centers likewise will be encouraged to distribute PDB *via* FTP.

Projection of current data rates into the future suggests that PDB may include over 25 000 structures by the year 2000 (*cf.*, Figure 4). This trend makes it imperative for PDB to make plans to ensure that mechanisms are put in place as soon as possible to allow for the archiving and dissemination of information and knowledge emanating from these studies in the future. The consequences of failing to properly address these issues now could be disastrous, as any future remedial action is almost certain to be prohibitively expensive.

Briefly, PDB plans include the following initiatives beginning immediately to keep the database fully current, releasing new data entries directly upon receipt without compromising validation and data quality, and to upgrade the database so as to satisfy current and future needs of its growing international user community:

---

[1] The CD ROM will include structure factor entries as well as atomic coordinate data, plus some graphics and index/retrieval routines.

1. Development of procedures, including user-friendly PDB-AUTHORIN software, to allow depositors to do most of the preparation and validation of their own entries.

2. A comprehensive upgrade of PDB contents and of the data interchange format. Addition of new data items and conversion of the current fixed record length PDB interchange format to the Crystallographic Information File (CIF) standard established by the IUCr is about to commence. CIF will be used for both atomic coordinate *and* structure factor data.

3. Implementation of PDB in the SYBASE RDBMS, to support all database management and archive functions at BNL. PDB-SYBASE, once implemented, will form the basis of an on-line query system which BNL will make widely available.

These are major initiatives, involving a fundamental reorganization of PDB, and requiring the assembly of substantial software and database development teams at BNL. In moving toward its long-range goals, PDB will seek to incorporate the best developments from other groups into software packages assembled at BNL. To this end, direct collaborations have been established between PDB and a number of other laboratories, in the USA, Europe and Japan, involved in related validation and database development projects. A series of international workshops is planned, to coordinate these research and development efforts, to provide guidance with the CIF-based format upgrade which also must be coordinated with the IUCr, and generally to inform collaborators and users about the latest developments at PDB.

An International Advisory Board conducts periodic reviews of PDB resource operations. PDB is deeply grateful to present and past Advisory Board members for their advice and guidance. The current Board membership of five (C. Bugg, Chairman, I. Kuntz, R. Salemme, J. Thornton, and K. Watenpaugh) is in the process of being expanded to broaden its international representation and enlarge the range of disciplines covered.

Acknowledgement. It is a pleasure to ackowledge the valuable contributions of our many PDB consultants and collaborators, including Ethan Benatan, Helen Berman, Philip Bourne, Tamas Demeny, Anke Gelbin, Francois Major, Calton Pu, Jean Richelle, John Rose, John Ruble, Raymond Stevens, S. Swaminathan, and Shoshana Wodak.

Figure Captions

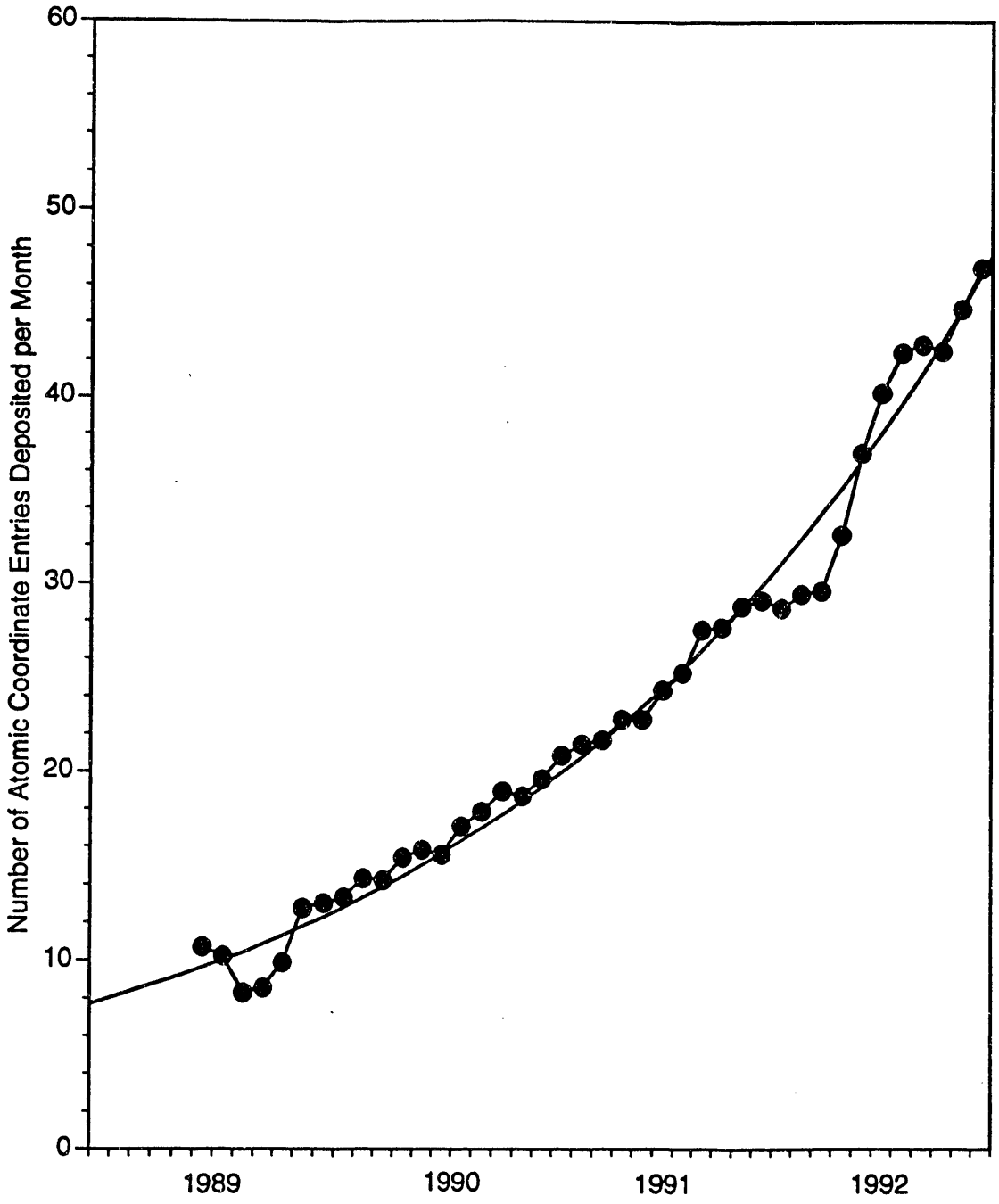
Figure 1. Running 12-month average number of atomic coordinate entries deposited in PDB per month since 1988. The curve shows an exponential fit to the experimental data points.

Figure 2. Total number of atomic coordinate entries in the PDB, 1984-1991.

Figure 3. Status of atomic coordinate entries, July 1992.

Figure 4. Number of atomic coordinate entries in PDB projected through the year 2000. Qaudratic and exponential fits to the data from 1988 through the first quarter of 1992 are shown.

Figure 1



Running 12-month average number of atomic coordinate entries deposited per month since 1988. The curve shows an exponential fit to the experimental data points.

PDB

Figure 2

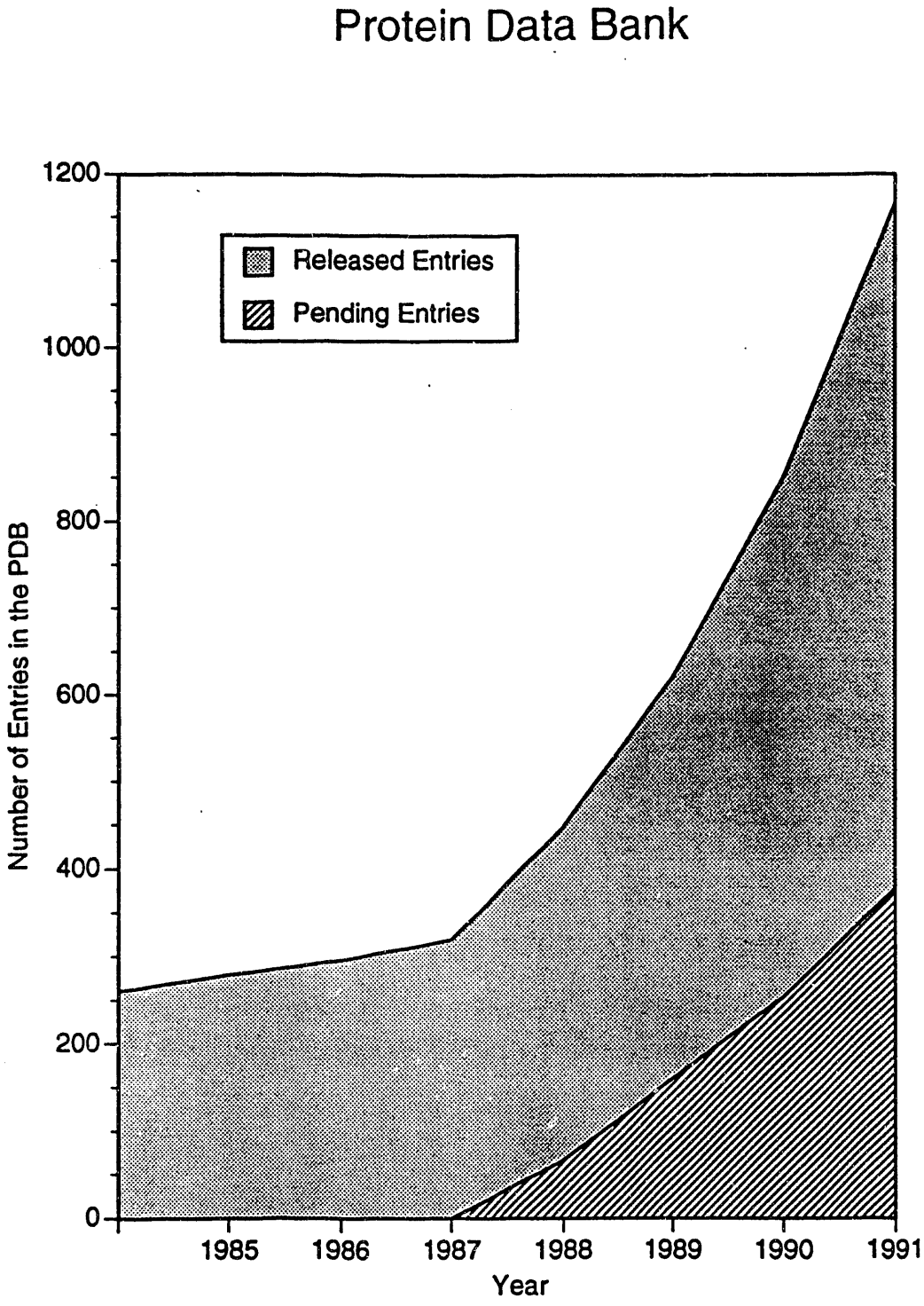

Protein Data Bank
Released Entries
Pending Entries
Number of Entries in the PDB
0
200
400
600
800
1000
1200
1985
1986
1987
1988
1989
1990
1991
Year

Figure 3

Status of Atomic Coordinate Entries

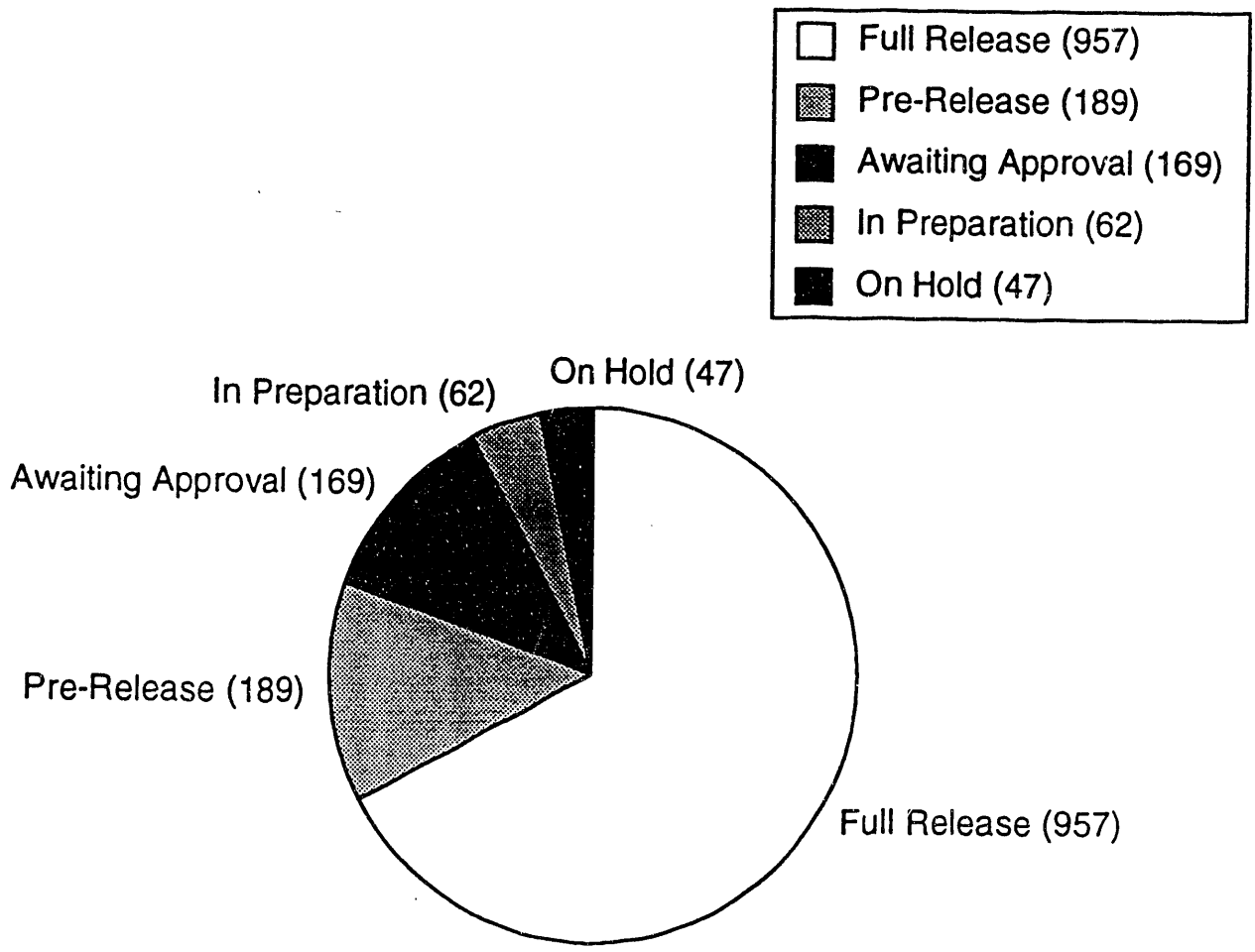
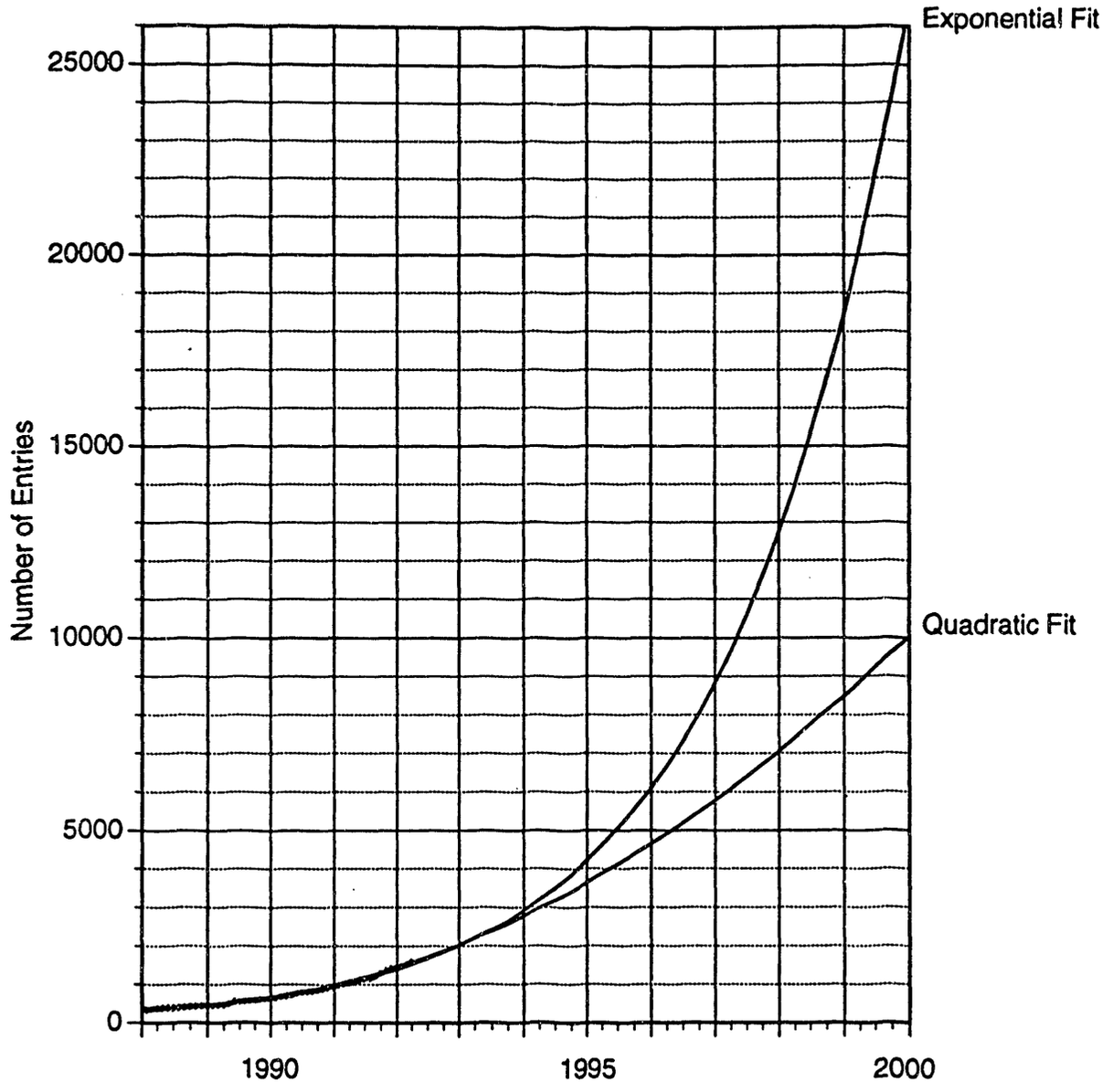


Figure 4


Exponential Fit
Quadratic Fit
Number of Entries
25000
20000
15000
10000
5000
0
1990
1995
2000

PDB

END

DATE
FILMED
5/10/93