# Journal of MOLECULAR BIOLOGY

EDITOR-IN-CHIEF: J. C. KENDREW

Volume 112, Number 3, 25 May 1977

ACADEMIC PRESS
London New York San Francisco
A Subsidiary of Harcourt Brace Jovanovich, Publishers

JMOBAK 112 (3) 359–542 (1977)
ISSN No. 0022–2836

# JOURNAL OF MOLECULAR BIOLOGY




Published three times a month on the 5th, 15th and 25th at 24–28 Oval Road, London NW1 7DX, England by Academic Press Inc. (London) Limited.

1977: Volumes 109–117, 36 Issues. Inland, £199.35 including postage and packing; abroad, £239.40 including postage and packing.

Index and Cumulative Contents of Volumes 1 to 20, 21 to 40, 41 to 60: inland, £4.00 each; abroad, £4.75 each.

Subscription orders should be sent to Academic Press Inc. (London) Limited, 24–28 Oval Road, London NW1 7DX, with the exception of those originating in the U.S.A., Canada, Central America and South America; these should be sent to Academic Press Inc., 111 Fifth Avenue, New York, New York 10003. (Subscriptions from these countries: $470.85 inclusive of postage and air freight.) Second class postage paid at Jamaica, N.Y., U.S.A.

Air freight and mailing in the U.S.A. by Publications Expediting Inc., 200 Meacham Avenue, Elmont, N.Y. 11003, U.S.A.

*J. Mol. Biol.* (1977) **112**, 535–542

# The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures

The Protein Data Bank is a computer-based archival file for macromolecular structures. The Bank stores in a uniform format atomic co-ordinates and partial bond connectivities, as derived from crystallographic studies. Text included in each data entry gives pertinent information for the structure at hand (e.g. species from which the molecule has been obtained, resolution of diffraction data, literature citations and specifications of secondary structure). In addition to atomic co-ordinates and connectivities, the Protein Data Bank stores structure factors and phases, although these latter data are not placed in any uniform format. Input of data to the Bank and general maintenance functions are carried out at Brookhaven National Laboratory. All data stored in the Bank are available on magnetic tape for public distribution, from Brookhaven (to laboratories in the Americas), Tokyo (Japan), and Cambridge (Europe and worldwide). A master file is maintained at Brookhaven and duplicate copies are stored in Cambridge and Tokyo. In the future, it is hoped to expand the scope of the Protein Data Bank to make available co-ordinates for standard structural types (e.g. α-helix, RNA double-stranded helix) and representative computer programs of utility in the study and interpretation of macromolecular structures.

The Protein Data Bank† (1971,1973) was established in 1971 as a computer-based archival file for macromolecular structures. The purpose of the Bank is to collect, standardize, and distribute atomic co-ordinates and other data from crystallographic studies. As the number of solved protein and nucleic acid structures has grown to the point where some $10^7$ characters are necessary to represent the co-ordinate information currently held, the need for such a computer-readable file has become very clear, and demands for the Bank's services have increased accordingly. The Protein Data Bank is one of several data base activities in the field of crystallography, e.g. the Bibliographic (Kennard *et al.*, 1972) and Structural (Allen *et al.*, 1973) Data Files for organic and organometallic compounds, the Atlas of Macromolecular Structure on Microfiche (AMSOM) (Feldmann, 1977), the Bond Index to the Determination of Inorganic Crystal Structures (BIDICS)‡ and the Powder Diffraction File.§

## (a) *Scope*

The Protein Data Bank covers atomic co-ordinates, structure factors and phases from diffraction studies of macromolecules. Since most of this information is not generally published in the primary literature, the Bank depends for comprehensiveness on data supplied directly by the investigators. It is essentially a depository of data, held in computer-readable form, in contrast to other data banks that are based

† Protein Data Bank is a misnomer of historical origin, since the file now contains entries for a nucleic acid.

‡ I. D. Brown, Bond Index to the Determination of Inorganic Crystal Structures, McMaster University, Hamilton, Ontario, Canada, L8S 4M1.

§ American Society for Testing Materials, 1916 Race St., Philadelphia, PA. 19103, U.S A.

TABLE 1

*Protein data bank holdings*

| IDENT CODE | MOLECULE | DEPOSITOR | STATUS CODE |
|---|---|---|---|
| 1ADK | ADENYLATE KINASE | G. SCHULZ | A |
| 1ADH | ALCOHOL DEHYDROGENASE (ADP-RIB) | C.-I. BRANDEN | |
| 2ADH | ALCOHOL DEHYDROGENASE (ORTHOPHEN) | C.-I. BRANDEN | |
| 2CHA | ALPHA-CHYMOTRYPSIN (TOSYL) | D. BLOW | R |
| 3CHA | ALPHA-CHYMOTRYPSIN | A. TULINSKY | |
| 1FAB | ANTIGEN BINDING FRAGMENT (NEW) | R. POLJAK | |
| 1REI | BENCE-JONES IMMUNOGLOBULIN REI | O. EPP, R. HUBER | |
| 1CPV | CALCIUM-BINDING PARVALBUMIN SET 6A | R. KRETSINGER | |
| 2CPV | CALCIUM-BINDING PARVALBUMIN SET 6H | R. KRETSINGER | |
| 3CPV | CALCIUM-BINDING PARVALBUMIN SET 6I | R. KRETSINGER | |
| 1CAB | CARBONIC ANHYDRASE B | K. KANNAN | |
| 1CAC | CARBONIC ANHYDRASE C | K. KANNAN | |
| 1CPA | CARBOXYPEPTIDASE A | W. LIPSCOMB | |
| 1CHG | CHYMOTRYPSINOGEN | J. KRAUT | |
| 2CNA | CONCANAVALIN A | G. REEKE, G. EDELMAN | N |
| 3CNA | CONCANAVALIN A | K. HARDMAN | R |
| 1B5C | CYTOCHROME B5 | F. S. MATHEWS | |
| 1CYT | CYTOCHROME C (ALBACORE, OXIDIZED) | R. DICKERSON | |
| 2CYT | CYTOCHROME C (ALBACORE, REDUCED) | R. DICKERSON | |
| 1CYC | CYTOCHROME C (BONITO, HEART) | M. KAKUDO | |
| 1C2C | CYTOCHROME C2 | J. KRAUT | |
| 155C | CYTOCHROME C550 | R. TIMKOVICH | |
| 1EST | ELASTASE | H. WATSON | |
| 1FDX | FERREDOXIN | L. JENSEN | |
| 1FXN | FLAVODOXIN (CLOSTRIDIUM MP) | M. LUDWIG | |
| 1GCH | GAMMA-CHYMOTRYPSIN | COHEN,DAVIES,SILVERTON | P |
| 1GPD | GLYCERALDEHYDE-3-P-DEHYDROGENASE(LOBSTR) | M. ROSSMANN | N |
| 2MHB | HEMOGLOBIN (HORSE, AQUO MET) | LADNER, HEIDNER, PERUTZ | RP |
| 1DHB | HEMOGLOBIN (HORSE, DEOXY) | M. PERUTZ, G. FERMI | |
| 1HHB | HEMOGLOBIN (HUMAN, DEOXY) | M. PERUTZ, G. FERMI | |
| 1FDH | HEMOGLOBIN (HUMAN, FETAL, DEOXY) | J. FRIER | |
| 1LHB | HEMOGLOBIN (LAMPREY) | W. HENDRICKSON | |
| 1YHX | HEXOKINASE (YEAST) BIII | T. STEITZ | B |
| 1HIP | HIGH POTENTIAL IRON PROTEIN | J. KRAUT | |
| 2LDH | LACTATE DEHYDROGENASE | M. ROSSMANN | PD |
| 3LDH | LACTATE DEHYDROGENASE/NAD/PYRUVATE | M. ROSSMANN | PD |
| 1LYZ | LYSOZYME (HEN EGG-WHITE, SET W2) | R. DIAMOND | P |
| 2LYZ | LYSOZYME (HEN EGG-WHITE, SET RS5D) | R. DIAMOND | P |
| 3LYZ | LYSOZYME (HEN EGG-WHITE, SET RS6A) | R. DIAMOND | P |
| 4LYZ | LYSOZYME (HEN EGG-WHITE, SET RS9A) | R. DIAMOND | P |
| 5LYZ | LYSOZYME (HEN EGG-WHITE, SET RS12A) | R. DIAMOND | P |
| 6LYZ | LYSOZYME (HEN EGG-WHITE, SET RS16) | R. DIAMOND | P |
| 1MDH | MALATE DEHYDROGENASE | L. BANASZAK | A |
| 1MBN | MYOGLOBIN (SPERM WHALE) | H. WATSON | |
| 2MBN | MYOGLOBIN (SPERM WHALE, MET) | T. TAKANO | |
| 3MBN | MYOGLOBIN (SPERM WHALE, DEOXY) | T. TAKANO | |
| 3PTI | PANCREATIC TRYPSIN INHIBITOR | R. HUBER | R |
| 8PAP | PAPAIN, NATIVE | J. DRENTH | R |
| 2PAP | PAPAIN (ACE-ALA-ALA-PHE-ALA, CYS-25) | J. DRENTH | |
| 3PAP | PAPAIN (CYS DERIV OF CYS-25) | J. DRENTH | |
| 4PAP | PAPAIN (OXIDIZED CYS-25) | J. DRENTH | |
| 5PAP | PAPAIN (TOS-LYS, CYS-25) | J. DRENTH | |
| 6PAP | PAPAIN (BZOXY-GLY-PHE-GLY, CYS-25) | J. DRENTH | |
| 7PAP | PAPAIN (BZOXY-PHE-ALA,CYS-25) | J. DRENTH | |
| 1PGK | PHOSPHOGLYCERATE KINASE (YEAST) | H. WATSON | A |
| 2PGK | PHOSPHOGLYCERATE KINASE (HORSE) | P. EVANS, D. PHILLIPS | B |
| 1PAB | PREALBUMIN (HUMAN, PLASMA) | S. OATLEY, D. PHILLIPS | |
| 1RNS | RIBONUCLEASE S | H. WYCKOFF | |
| 2RXN | RUBREDOXIN | L. JENSEN | ND |
| 1SNS | STAPHYLOCOCCAL NUCLEASE | F. A. COTTON, E. HAZEN | |
| 1SGB | STREPTOMYCES GRISEUS PROTEINASE B | M. JAMES | A |
| 1SBT | SUBTILISIN BPN' | J. KRAUT | |
| 2SBT | SUBTILISIN NOVO | J. DRENTH | |
| 1SOD | SUPEROXIDE DISMUTASE | J. AND D. RICHARDSON | A |
| 1TLN | THERMOLYSIN (UNREFINED) | B. MATTHEWS | |
| 2TLN | THERMOLYSIN (REFINED) | B. MATTHEWS | |
| 1SRX | THIOREDOXIN | B.-O. SODERBERG | A |
| 1TNA | TRANSFER RNA (YEAST, PHE) | J. SUSSMAN, S.-H. KIM | N |
| 2TNA | TRANSFER RNA (YEAST, PHE) | M. SUNDARALINGAM | P |
| 3TNA | TRANSFER RNA (YEAST, PHE) | JACK, LADNER, KLUG | P |
| 1TIM | TRIOSE PHOSPHATE ISOMERASE | I. WILSON, D. PHILLIPS | |
| 1PTN | TRYPSIN (NATIVE, PH8) | FEHLHAMMER,BODE,SCHWAGER | N |
| 2PTB | TRYPSIN(BENZAMIDINE INHIBITED, PH7) | FEHLHAMMER,BODE,SCHWAGER | RN |
| 1PTC | TRYPSIN/TRYPSIN INHIBITOR COMPLEX | BODE ET AL. | N |

STATUS CODES

| | |
|---|---|
| BLANK | STANDARD ENTRY AVAILABLE FOR DISTRIBUTION |
| A | ALPHA CARBON ATOMS ONLY |
| B | BACKBONE ONLY |
| D | NEW DATA HAS BEEN PROMISED |
| N | NEW ENTRY WITH DEPOSITOR FOR APPROVAL |
| P | IN PREPARATION |
| R | REPLACES ON HOLD OR OUT OF DATE PARAMETER SET |

on data abstracted from scientific publications. The Bank contains 77 atomic co-ordinate entries for 47 macromolecules (Table 1),† and 13 sets of structure factors and phases. The atomic co-ordinate entries, which include descriptive text and partial bond connectivities, conform to a uniform format (see below), but the structure factors and phases are stored in the format received from depositors. All co-ordinate entries are referred to depositors for verification, before being made available publicly through the Bank.

(b) *Record structure of atomic co-ordinate entries*

Atomic co-ordinate entries consist of records each of 80 characters.‡ Using the punched card analogy, columns 1 to 6 contain a record type identifier, and columns 7 to 70 contain data.§ Columns 71 to 80 are normally blank, but may contain sequence information which is added by the library-file management program UPDATE¶ used to maintain the file on the Brookhaven CDC CYBER 70/76 computing system. In order to facilitate retrieval of data from the file, the first four characters of each record define the unique record type, and the syntax of each record is independent of the order of records within any entry for a particular macromolecule. (In the master file, this order is always fixed.) Atomic co-ordinate data contributed by depositors are processed into the standard format with program MACMOL,‖ which also subjects the data to certain nomenclature and connectivity checking procedures.

A sample partial entry for the protein ribonuclease S is shown in Table 2.†† The unique code 1RNS identifying this entry is given in the HEADER record, along with the date these data were entered into the Bank, and a provisional classification based on function, intended for future use in indexing and subdividing the file. Text giving the name of molecule, species from which it has been obtained, authors, literature citations, and other general description are presented in records COMPND through REMARK. SEQRES gives the amino acid sequence, and FTNOTE records are footnotes keyed to particular residues or atoms. Records HELIX through TURN describe the secondary structure as stated or approved by the depositor. Record CRYST1 defines the unit cell, while ORIGX and SCALE respectively give transformations relating the orthogonal Ångström co-ordinates stored in the file to those originally supplied by the depositor (these frequently are referred to an oblique or non-isometric system) and to standard crystallographic fractional co-ordinates. ATOM records give the IUPAC-IUB (1969) standard atom names (IUPAC-IUB, 1970), and residue abbreviations (IUPAC-IUB, 1971), along with sequence identifiers (cf. SEQRES, above), co-ordinates in Ångström units, and occupancies and thermal

† In addition to current co-ordinate entries shown in Table 1, the Bank contains obsolete entries (for adenylate kinase tosyl, α-chymotrypsin, concanavalin A, lactate dehydrogenase, horse methemoglobin, papain, rubredoxin, benzamidine-inhibited trypsin and pancreatic trypsin inhibitor), which have been superseded by later, more accurate data. These obsolete data are available on special request.

‡ Originally, the Bank used a 140-character format, similar to that employed in the protein refinement programs of Diamond (1966,1971). The 140-character format has been superseded by the 80-character format.

§ A detailed description of the file formats is available from Brookhaven on request.

¶ Control Data Corporation, UPDATE Reference Manual, Publication No. 60342500, Control Data Corporation, Arden Hills, Minnesota, 1974.

‖ G. J. B. Williams, unpublished. For the 140-character data, program PROIN by E. F. Meyer was utilized.

†† The file is organized in a similar way for proteins and nucleic acids, although certain differences exist, e.g. with regard to details of atom and residue names.

DOCKET ALARM

# Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

## Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

## Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

## Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

## API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

**LAW FIRMS**
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

**FINANCIAL INSTITUTIONS**
Litigation and bankruptcy checks for companies and debtors.

**E-DISCOVERY AND LEGAL VENDORS**
Sync your system to PACER to automate legal marketing.