Computer Programs

J. Appl. Cryst. (1992). **25**, 803–806

The database BIOSCAT: a tool for structure research by scattering and hydrodynamic methods.

By JÜRGEN J. MÜLLER,* HARALD PANKOW, BRUNHILDE POPPE and GREGOR DAMASCHUN, *Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Strasse* 10, *D-O*-1115 *Berlin, Germany*

(*Received* 20 *September* 1991; *accepted* 24 *June* 1992)

Abstract

The crystal structures of a large number of proteins and nucleic acids are known and the corresponding sets of coordinates are stored in the Brookhaven Protein Data Bank. For structure investigations of biological macromolecules in solution, scattering and hydrodynamical methods are powerful biophysical tools when starting the data interpretation on the basis of the crystal structure of the molecules. The database BIOSCAT covers the main structural parameters estimable by X-ray scattering, translation and rotation diffusion methods and the X-ray scattering intensities and low- and high-resolution real-space electron distance distribution functions of 70 biological macromolecules and of oligonucleotides in standard conformation. The parameters and the scattered intensities are calculated from the atomic coordinates using the improved cube method and the real-space functions are estimated *via* a termination-error-reduced Fourier sine transformation. The database access is organized by the program *PASSDB*, which can generally be used for 'readable' databases. A simple query language allows enquiries into the database without knowledge of a programming language. The program *CONVSQL* converts the database into normalized relations that can be handled by structured query languages (SQLs).

I. Introduction

Small- and intermediate-angle X-ray scattering, neutron or light scattering combined with hydrodynamical methods are well established methods in general use for investigations of the structure of biopolymers in solution. They can be included as components in a knowledge-based expert system (Altman & Jardetzky, 1986) when using the atomic coordinates of known structures as basic models for the molecules in solution under special ionic, pH or temperature conditions, interacting with ligands or modified by protein engineering (Trewhella, Carlson, Curtis & Heidorn, 1988; Hubbard, Hodgeson & Doniach, 1988; Fedorov & Denesyuk, 1978*a*). Similarities of structural features of molecules in crystal and in solution can be confirmed or changes detected by comparison of theoretical and observed solution scattering curves (Fedorov & Denesyuk, 1978*b*). Scattering may also be used for detection of structural similarities between functionally related but structurally noncharacterized macromolecules. The coupling of scattering and hydrodynamical data means that the determination of a solvation shell and the detection of intramolecular motion are possible (Müller, Glatter, Zirwer & Damaschun, 1983; Müller & Kayushina, 1987). Furthermore, scattering curves of a wide diversity of macromolecules are necessary examples for planning the experimental strategy for particular devices such as neutron sources of synchrotrons and a comprehensive collection of data is useful for setting up information retrieval systems for scattering and hydrodynamical data.

Structural molecular and hydrodynamic parameters of X-ray scattering and electron distribution functions have also been calculated for proteins and nucleic acids from their atomic coordinates, which are collected in the Brookhaven Protein Data Bank (PDB) (Bernstein *et al.*, 1977). These data are stored in the database BIOSCAT for quick reference.

II. Content of the database

The first release of the database BIOSCAT covers the data of 70 biological macromolecules and double-helical synthetic oligonucleotides (Table 1). Included are data of identical macromolecules investigated structurally with different refinements or resolution (*e.g.* lysozyme: 5LYZ, 6LYZ), with various ligands (*e.g.* deoxyribonucleic acid: 4BNA, 6BNA), crystallized in different crystal classes (*e.g.* lysozyme: 1LZT, 5LYZ), with and without crystallographically identified water molecules (*e.g.* transfer ribonucleic acid: 6TNA, 6TNA/W), for all α, β and (α, β) proteins (*e.g.* myoglobin: 1MBN, immunoglobulin FBA: 1FB4; subtilisin: 2SBT), for functionally and structurally related molecules (*e.g.* A-chain human hemoglobin: 2HHB/A; B-chain human hemoglobin: 2HHB/B; lamprey hemoglobin: 2LHB; leghemoglobin: 1LH4; B-chain horse hemoglobin: 2MHB/B; sperm whale met-myoglobin: 1MBN; sperm whale oxymyoglobin: 1MBO), for nucleic acids at different temperatures (*e.g.* 1BNA, 2BNA) and short and long synthetic double-helical oligonucleotides (*e.g.* ADF1, ADNA/20; ARN1, RNA/20).

III. Content of an entry

The information for one molecule is deposited in an entry. Each entry has a descriptive part and several data partitions. The descriptive part characterizes the molecule and the source of the atomic coordinates and also contains particular information about the BIOSCAT database. These records provide information that can be used to generate cross references of the entries.

The data section is partitioned into structure parameters, low-resolution shape models, hydrodynamical parameters,

* To whom correspondence should be addressed.

0021-8898/92/060803-04$06.00

Table 1. *BIOSCAT entries (release 1)*

The identification code is that of the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).

Proteins

156B	Cytochrome B562	1CHG	Chymotrypsinogen A
1CAC	Carbonic anhydrase	1CPV	Parvalbumin
1CYC	Ferrocytochrome C	1EST	Tosyl elastase
1FB4	Immunoglobulin Fab	1FB4	Immunoglobulin (L chain)
1GCR	Crystallin	1HHO	Hemoglobin A
1INS/AB	Insulin[(A + B) chain]	1INS/ABCD	Insulin[(A + B + C + D) chain]
1LH1	Leghemoglobin (met)	1LH4	Leghemoglobin (deoxy)
1LZT/W	Lysozyme with H_2O	1MB5	Myoglobin
1MBD	Myoglobin (deoxy)	1MBN	Myoglobin (met)
1MBO	Myoglobin (oxy)	1RN3	Ribonuclease A
1TGN	Trypsinogen	1TPO	β-Trypsin
2ALP	α-Lytic protease	2CAB	Carbonic anhydrase B
2CGA/B	Chymotrypsinogen A	2CHA	α-Chymotrypsin A
2CNA	Concanavalin A	2DHB	Hemoglobin (horse, deoxy)
2GCH	γ-Chymotrypsin A	2HHB	Hemoglobin (human, deoxy)
2HHB/A	Hemoglobin (human, A chain)	2HHB/B	Hemoglobin (human, B chain)
2LHB	Hemoglobin V	2MBN/W	Myoglobin with H_2O
2MHB/B	Hemoglobin (horse, B chain)	2PKA	Kallikrein A
2SBT	Subtilisin novo	2SGA	Proteinase A
2SOD	Superoxide dismutase	3C2C	Cytochrome C2
3CLN	Calmodulin	3PGK	Phosphoglycerate kinase
3RP2/A	Proteinase II	3SGB	Proteinase B
3TLN	Thermolysin	5CHA/A	α-Chymotrypsin (A chain)
5CHA/B	α-Chymotrypsin (B chain)	5CPA	Carboxypeptidase A
5LYZ	Lysozyme	5TNC	Troponin C
9PAP	Papain	6LYZ	Lysozyme

Nucleic acids

1BNA	DNA (B, CGCGAATTCGCG), 290K	2BNA	DNA (B, CGCGAATTCGCG), 16K
4BNA	DNA (B, CGCGAATTCGCG), 60MPD	6BNA	DNA (B, CGCGAATTCGCG) netropsin
6TNA	tRNA	6TNA/W	tRNA with H_2O
ADF1	DNA (A, CGCGCGCGCGCG)	ADNA/20	DNA (A, CGCGCGCGCGCGCGCGCGCG)
ARN1	RNA (A, CGCGCGCGCGCG)	ARNA/20	RNA (A, CGCGCGCGCGCGCGCGCGCG)
A-R1	RNA (A′, CGCGCGCGCGCG)	A-RN/20	RNA (A′, CGCGCGCGCGCGCGCGCGCG)
BDF1	DNA (B, CGCGCGCGCGCG), 36°	BDNA/20	DNA (B, CGCGCGCGCGCGCGCGCGCG), 36°
BDS1	DNA (B, CGCGCGCGCGCG), 34°	BDNS/20	DNA (B, CGCGCGCGCGCGCGCGCGCG), 34°
ZDNA	DNA (Z-I, CGCGCGCGCGCG)	ZIDN/20	DNA (Z-I, CGCGCGCGCGCGCGCGCGCG)

hydrodynamical models, scattered intensities and real-space functions. All stored data are briefly described in Table 2.

The low-resolution structure parameters have been calculated directly from the atomic coordinates using classical mechanics or they are estimated from the X-ray scattering curve and electron distance distribution function (Damaschun, Müller, Pürschel & Sommer, 1969).

The stored low-resolution structure models are mostly triaxial ellipsoids. The inertia-equivalent ellipsoid and the models derived therefrom are calculated directly from the atomic coordinates (Taylor, Thornton & Turnell, 1983; Müller & Schrauber, 1992). They relate the atomic structure and low-resolution shape models of the macromolecules. The scattering-equivalent ellipsoid, determined by a nonlinear fitting procedure to the innermost part of the scattering curve of the macromolecule (Marquardt, 1963) has been included too, despite the fact that its axial dimensions are distorted by electron density fluctuations within the molecules (Müller & Schrauber, 1992).

The hydrodynamic parameters molar mass and partial specific volume (Cohn & Edsall, 1943) and the data describing the rotational frictional behaviour of the macromolecules on the basis of the enlarged inertia-equivalent ellipsoid (Müller, 1991; Müller & Schrauber, 1992) were calculated from the structure. The parameters that characterize the translational frictional properties of the molecule on the basis of a rotational ellipsoid are predicted from the X-ray scattering curve of the molecule in solution, as described recently by Kumosinski & Pessen (1982), Müller, Damaschun, Damaschun, Gast, Plietz & Zirwer (1984) and by Müller & Kayushina (1987).

The dimensions of the hydrodynamical models and of spheres and rotational and triaxial ellipsoids are stored in the next partition.

The scattered-intensity partition contains the X-ray scattering intensity of the molecule surrounded by solvents of various electron densities, the scattered intensity of the molecule solved in water and in a solvent with an electron density identical to the mean electron density of the molecule (contrast-matching point) and the intensity scattered by the homogeneous solvent-excluded body (infinite contrast). These intensities are calculated from atomic coordinates and

Table 2. *Content of an entry of the database BIOSCAT*

The parameters that have been directly calculated from the atomic structure are marked 'd.c.'.

ENTRYNR	Number of the entry
IDPDB	Identification code from Protein Data Bank (PDB)
MOLNAME	Name of the molecule from Protein Data Bank
⋮	
SPRSVAC	Radius of gyration in vacuum (d.c.)
SPRSHOM	Radius of gyration of the solvent-excluded body (d.c.)
SPRSH20	Radius of gyration in water
SPRSCRO	Cross-sectional radius of gyration in water
SPRSTHI	Thickness radius of gyration in water
SPLARDIS	Maximum distance between two atoms in the molecule (d.c.)
SPLC	Correlation length of the shape
SPAC	Mean electron distance in the shape
SPSURHOM	Surface of the homogeneous solvent-excluded body
SPSURSHA	Surface of the shape
SPFC	Correlation area of the shape
SPVOLHOM	Volume of the solvent-excluded body (d.c.)
SPVOLSHA	Volume of the shape
SPVOLDRY	Dry volume of the molecule (d.c.)
SPMEDEL	Medium electron density in the molecule (d.c.)
SMIEEA,B,C	Inertia-equivalent ellipsoid (IEE), half axes *A*, *B*, *C* (d.c.)
SMIEE90A,B,C	IEE, covering 90% of the molecular volume, half axes *A*, *B*, *C* (d.c.)
SMIEE 100A,B,C	IEE, covering 100% of the molecular volume, half axes *A*, *B*, *C* (d.c.)
SMTHETA	Eulerian angle THETA between fixed-space and IEE systems (d.c.)
SMPSI	Eulerian angle PSI between fixed-space and IEE systems (d.c.)
SMPHI	Eulerian angle PHI between fixed-space and IEE systems (d.c.)
SMMASSX,Y,Z	Mass center of the molecule and IEE, *x*, *y*, *z* coordinates (d.c.)
SMSEEA,B,C	Scattering-equivalent ellipsoid, half axes *A*, *B*, *C*
HPPSVTH	Theoretical partial specific volume of the molecule (d.c.)
HPMCHEM	Molecular mass of the selected part of the molecule (d.c.)
HPMCHCRO	Mass per unit cross-sectional area
HPMCHTHI	Mass per unit area
HPFTRANS	Translational friction coefficient, ellipsoid of revolution
HPDTRANS	Translational diffusion coefficient, ellipsoid of revolution
HPSED	Sedimentation coefficient, ellipsoid of revolution
HPFTDRY	Translational friction coefficient, dry sphere
HPFTSHA	Translational friction coefficient, sphere with the shape volume
HPFTEIEE	Translational friction coefficient, sphere with the enlarged IEE (EIEE) volume
HPFRSPH	Rotational frictional coefficient, sphere with the volume of the EIEE (d.c.)
HPFREIEEA,B,C	Rotational friction coefficient, EIEE rotating around *A*, *B*, *C* (d.c.)
HPDREIEEA,B,C	Rotational diffusion coefficient, EIEE rotating around *A*, *B*, *C* (d.c.)
HPTREIEEA,B,C	Rotational correlation time TAU, axes *A*, *B*, *C* of the EIEE (d.c.)
HPTRHARM	Harmonic mean value of the correlation times (d.c.)
HMSTODRY	Stokes radius, predicted from the molecular dry volume (d.c.)
HMSTOSHA	Stokes radius, predicted from the molecular shape volume (d.c.)
HMSTOEIEE	Stokes radius, predicted from the molecular EIEE volume (d.c.)
HMTFREA,B,C	Translational friction-equivalent rotational ellipsoid, half axes *A*, *B*, *C*
HMREIEEA,B,C	Enlarged inertia equivalent ellipsoid (EIEE), half axes *A*, *B*, *C* (d.c.)
INI_HOM	Scattered intensity, solvent-excluded volume body
INI_VAC	Scattered intensity, molecule in vacuum (Debye)
INI_H2O	Scattered intensity, molecule in water
INI_MED	Scattered intensity, molecule in a solvent with the averaged electron density of the molecule
INI_ATOM	Scattered intensity, independent atoms in the molecule
RSC_LOW	Excess electron autocorrelation function, low resolution
RSC_HIGH4	Excess electron autocorrelation function, resolution 0.4 nm
RSP_LOW	Excess electron distance distribution, low resolution
RSP_HIGH4	Excess electron distance distribution, resolution 0.4 nm
RDP_DIR	Atom distance distribution, equally weighted atoms (d.c.)

a set of van der Waals radii (Müller, Pavlov & Fedorov, 1983) by using the improved cube method (Müller, 1983). The scattered intensity of the molecule in vacuum is calculated *via* Debye's (1959) equation.

The real-space functions cover the excess electron autocorrelation function of the molecule dissolved in water and the excess electron distance distribution function for two different structure resolutions, calculated from the theoretical scattering curve by a Fourier sine transformation (Müller, Damaschun & Schrauber, 1990). An atom distance distribution function for equally weighted atoms is determined directly from the atomic coordinates (Müller, Damaschun, Damaschun, Misselwitz, Zirwer & Nothnagel, 1984).

Additionally, 'remark' records exist to explain the special estimation conditions for the corresponding parameter. They provide error levels and describe the approximations used.

IV. Interrogating the database

The retrieval system *PASSDB* is used to extract information from the entries. *PASSDB* is a general system written in Fortran77 (VAX/VMS) for retrieving 'readable' databases and works on the basis of relational and Boolean operators (Gruber, 1990). Detailed information is summarized in the *PASSDB* manual.

The BIOSCAT tables may optionally be converted into normalized relations by the program *CONVSQL*. Then, these relations can be interrogated with a SQL-based databank management system (*e.g. ORACLE*).

Graphical representation of the stored data is possible *via HARVARD GRAPHICS* (product of Software Publishing Corporation SCP, USA) with or without using *PASSDB*.

V. Services and availability

Both the database BIOSCAT and the program *PASSDB* together with a user's manual are available on floppy disks, readable on an IBM-PC/AT compatible. A version for VAX/VMS computers is also available. The meaning and generation by Fortran77 programs of database entries are described in more detail in the manual.

The information deposited in the database will be updated by adding new or revised entries quarterly. Academic users interested in the database are requested to contact the authors. Distribution to academic users will be free of charge.

A PC version of the programs discussed above will be available in the future. Then the user can fill the BIOSCAT base with data from molecules of his own special interest and can choose between neutron or X-ray scattering.

This work was supported by the Deutsche Forschungsgemeinschaft, Bonn-Bad Godesberg.

References

ALTMAN, R. B. & JARDETZKY, O. (1986). *J. Biochem. (Tokyo)*, **100**, 1403–1423.

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, J. B., MEYER, E. F., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535–542.

COHN, E. J. & EDSALL, J. T. (1943). *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions*, edited by E. J. COHN & J. T. EDSALL, pp. 370–381, 428–431. New York: Reinhold.

DAMASCHUN, G., MÜLLER, J. J., PÜRSCHEL, H.-V. & SOMMER, G. (1969). *Monatsh. Chem.* **100**, 1701–1714.

DEBYE, P. (1959). *Z. Phys.* **156**, 256–264.

FEDOROV, B. A. & DENESYUK, A. I. (1978*a*). *FEBS Lett.* **88**, 114–117.

FEDOROV, B. A. & DENESYUK, A. I. (1978*b*). *J. Appl. Cryst.* **11**, 473–477.

GRUBER, M. (1990). *Das SQL Buch.* Düsseldorf, San Francisco, Paris, London, Soest: Sybex-Verlag.

HUBBARD, S. T., HODGESON, K. O. & DONIACH, S. (1988). *J. Biol. Chem.* **263**, 4151–4158.

KUMOSINSKI, T. F. & PESSEN, H. (1982). *Arch. Biochem. Biophys.* **219**, 89–100.

MARQUARDT, D. W. (1963). *J. Soc. Ind. Appl. Math.* **11**, 431–437.

MÜLLER, J. J. (1983). *J. Appl. Cryst.* **16**, 74–82.

MÜLLER, J. J. (1991). *Biopolymers*, **31**, 149–160.

MÜLLER, J. J., DAMASCHUN, G., DAMASCHUN, H., MISSELWITZ, R., ZIRWER, D. & NOTHNAGEL, A. (1984). *Biomed. Biochim. Acta*, **43**, 929–936.

MÜLLER, J. J., DAMASCHUN, G. & SCHRAUBER, H. (1990). *J. Appl. Cryst.* **23**, 26–34.

MÜLLER, J. J., DAMASCHUN, H., DAMASCHUN, G., GAST, K., PLIETZ, P. & ZIRWER, D. (1984). *Stud. Biophys.* **102**, 171–175.

MÜLLER, J. J., GLATTER, O., ZIRWER, D. & DAMASCHUN, G. (1983). *Stud. Biophys.* **93**, 39–46.

MÜLLER, J. J. & KAYUSHINA, R. (1987). *Stud. Biophys.* **120**, 15–22.

MÜLLER, J. J., PAVLOV, M. YU. & FEDOROV, B. A. (1983). *Stud. Biophys.* **97**, 121–128.

MÜLLER, J. J. & SCHRAUBER, H. (1992). *J. Appl. Cryst.* **25**, 181–191.

TAYLOR, W. R., THORNTON, J. M. & TURNELL, W. G. (1983). *J. Mol. Graph.* **1**, 30–38.

TREWHELLA, J., CARLSON, V. A. P., CURTIS, E. H. & HEIDORN, D. B. (1988). *Biochemistry*, **27**, 1121–1125.