MMDB: An ASN.1 Specification for Macromolecular Structure

Hitomi Ohkawa, James Ostell and Stephen Bryant

National Center for Biotechnology Information
National Library of Medicine, National Institute of Health, Bldg.38A, Rm.8N805
8600 Rockville Pike, Bethesda, MD 20894 USA
ohkawa@ncbi.nlm.nih.gov, ostell@ncbi.nlm.nih.gov, bryant@ncbi.nlm.nih.gov

Abstract

We present an exchange specification for data describing the three-dimensional structure of biological macromolecules. The specification was designed for MMDB, a Molecular Modeling Database supported by the National Center for Biotechnology Information (NCBI), based on information from the Protein Data Bank (PDB). In the MMDB specification, the chemical structures of molecules are described hierarchically as connectivity graphs, to directly support comparison by subgraph isomorphism or assignment algorithms. Three-dimensional coordinates are linked unambiguously to nodes in the chemical graph, so that homology-derived structures may be generated directly from alignment of chemically similar groups. In conversion to this form, data from PDB are extensively validated, so as to provide a description of chemical and spatial structure that is as accurate as possible. These changes in format and content of the known structure data are intended to support development of intelligent molecular modeling applications that make use of this invaluable information resource.

Description of MMDB

We present a data exchange specification for information describing the three-dimensional structure of biological macromolecules. The specification was designed for MMDB, a Molecular Modeling Database supported by the National Center for Biotechnology Information, NCBI. MMDB is based on information from the Protein Data Bank (Bernstein et al. 1977), modified in form and content to produce a macromolecular structure database readily usable by computational biologists and developers of molecular modeling software.

The MMDB specification is written in ASN.1, an ISO Open Systems Interconnection Standard used for formal, standardized data exchange above the level of specific software and hardware (Rose 1990, Ostell et al. 1994). Macromolecular structure data in this form may be read into computer memory using a suite of software tools also available from NCBI, in the form of C-language subroutine libraries (Ostell et al. 1994). This software automatically translates an ASN.1 stream into C data structures which are fully atomic, in the sense that all parsable data items from PDB are represented as individual numeric or character values. Software developers may therefore directly retrieve and manipulate data items relevant to molecular modeling by C subroutine call, instead of by parsing PDB text files. The C data structure declarations are produced automatically from the ASN.1 specification, and data item names and semantics are described fully by the MMDB specification presented here.

Molecular modeling involves comparison of the chemical structures of two molecules to produce an atom-by-atom mapping, from which the partial spatial structure of one molecule may be inferred from that of the other. The information required is an unambiguous description of chemical structure in the form of a chemical graph, and an unambiguous linking of spatial coordinate data to atoms forming the nodes of this graph. To facilitate molecular modeling MMDB therefore provides this information explicitly. Software may directly retrieve the data items needed for sequence alignment or subgraph isomorphism calculations, and need not encode the complex logic required to deduce covalent structure from atom and residue names and other conventions employed by PDB. Homology models derived in this way may also be represented explicitly.

Chemical graphs in MMDB are represented in a fashion similar to that proposed by the Chemical Abstracts Service in the CXF specification (Mockus & Steckert 1994, Mockus & Steckert 1995) and by the International Union of Crystallography in their mmCIF specification (Shindyalov et al. 1994, Shindyalov et al. 1995, Wodak et al. 1994). Biomolecular assemblies are organized as a chemical hierarchy of atoms, residues, molecules, with subgraphs for biopolymer residues given by reference to a standard dictionary. The standard subgraph dictionary distributed with MMDB includes the 20 amino acids naturally occurring in proteins and the 8 ribonucleotide and deoxyribonucleotide groups occurring in RNA and DNA. Construction of MMDB requires validation of PDB data against this dictionary, and therefore identifies a number

of inconsistencies and errors as occurrences of nonstandard residue groups. Subgraphs for these and true non-polymer components such as protein cofactors are constructed by reference to any explicit connectivity data provided by PDB, with validation by stereochemical calculations based on atomic coordinates.

Atomic coordinate data in MMDB retain all information provided by PDB, including crystallographic models with alternate conformations resulting from statistical disorder, and NMR-derived models represented as an ensemble of alternative structures. We have attempted to represent this information unambiguously, a process requiring considerable validation of any multiple-coordinate data provided by PDB. For many computational biology applications, however, it is useful to have a simplified model in which only a single "best" coordinate is provided for each atom in the chemical graph. To this end MMDB provides a single-coordinate-per-atom model as produced by the PKB analysis suite (Bryant 1989), a "view" of macromolecular structure which has been tested in many applications. MMDB also provides a further simplified single-coordinate-per-residue view, intended for graphical applications and rapid network transmission.

MMDB also allows for non-atomic representations of structure, such as density or surface models. These are not present in PDB, but the corresponding object types may nonetheless prove useful to computational biologists who encounter these common representations. Structural features are defined in MMDB as generic descriptors and sets of properties to be associated with atoms or residues, or a region in space. This definition is sufficient to represent secondary structure and site annotations as provided by PDB or proposed in mmCIF, but also general enough to accommodate new data. One might, for example, describe the electrostatic potential at points on a surface grid defined in the space of an atomic model from crystallography. One might similarly describe the local environment categories to be associated with a set of residues. These object types in MMDB are intended primarily to facilitate development of new applications.

Figure 1 shows a diagrammatic representation of the MMDB specification, giving an overview of relationships among data items and the design concepts behind MMDB. Appendix 1 lists the complete MMDB specification and constitutes the body of this paper. The specification itself includes detailed comments which explain data item semantics and the manner in which data items from PDB are mapped into MMDB. The specification, corresponding C structure definitions, and I/O routines are available via anonymous ftp from ncbi.nlm.nih.gov. Example C programs are also provided, including one that produces from MMDB a validated, PDB-formatted file. MMDB data files are available for ftp, but may also be accessed via client software addressing the Entrez server (NCBI 1994), which will provide in ASN.1 form data describing the three-dimensional structure of macromolecules, as well as their sequences, and citations to relevant scientific literature.

References

Benson, D.A.; Boguski, M.; Lipman, D.J.; and Ostell, J. 1994. Genbank. *Nucleic Acids Research* 22:3441-3444.

Berstein, F.C.; Koetzle, T.F.; Williams, G.J.B.; Meyer, E.F.; Brice, M.D.; Rodgers, J.R.; Kennerd, O.; Simanouchi, T.; and Tasumi, M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology* 112:535-542.

Bryant, S.H. 1989. PKB: A Program System and Data Base for Analysis of Protein Structure. *Proteins* 5:233-247.

Mockus, J., and Steckert, T.D. 1994. Chemical eXchange Format Version 1.0. Chemical Abstracts Service, A Division of the American Chemical Society.

Mockus, J., and Steckert, T.D. 1995. CXF - the Chemical eXchange Format. Gragg, C.E., and Mockus, J. eds. *Chemical Data Standards: Databases, Data Interchange, and Information Systems* - 2 vol., ASTM STP 1298. Philadelphia, PA: American Society for Testing and Materials.

NCBI. 1993. Entrez Release 6.0. Bethesda, MD: NCBI, NLM/NIH.

Ostell, J. 1994. NCBI Software Development Toolkit Version 1.9. NCBI, NLM/NIH, Bethesda, MD.

Rose, M.T. 1990. *The Open Book (A Practical Perspective on OSI).* Englewood Cliffs, NJ: Prentice Hall.

Shindyalov, I.N.; Chang, W; Pu, C; and Bourne, P.E. 1994. Macromolecular query language (MMQL): prototype data model and implementation. *Protein Engineering* 7(11): 1311-1322.

Shindyalov, L.N.; Chang, W; Cooper, J.A; and Bourne, P.E. 1995. Design and Use of a Software Framework to Obtain Information Derived from Macromolecular Structure Data. In Proceedings of the Twenty-Eighth Annual Hawaii International Conference on System Sciences, 207-216. Hawaii: The Institute of Electrical and Electronics Engineers, Inc.

Wodak, S. 1994. Proceedings of the European Macro-Molecular CIF Workshop. Brussels.

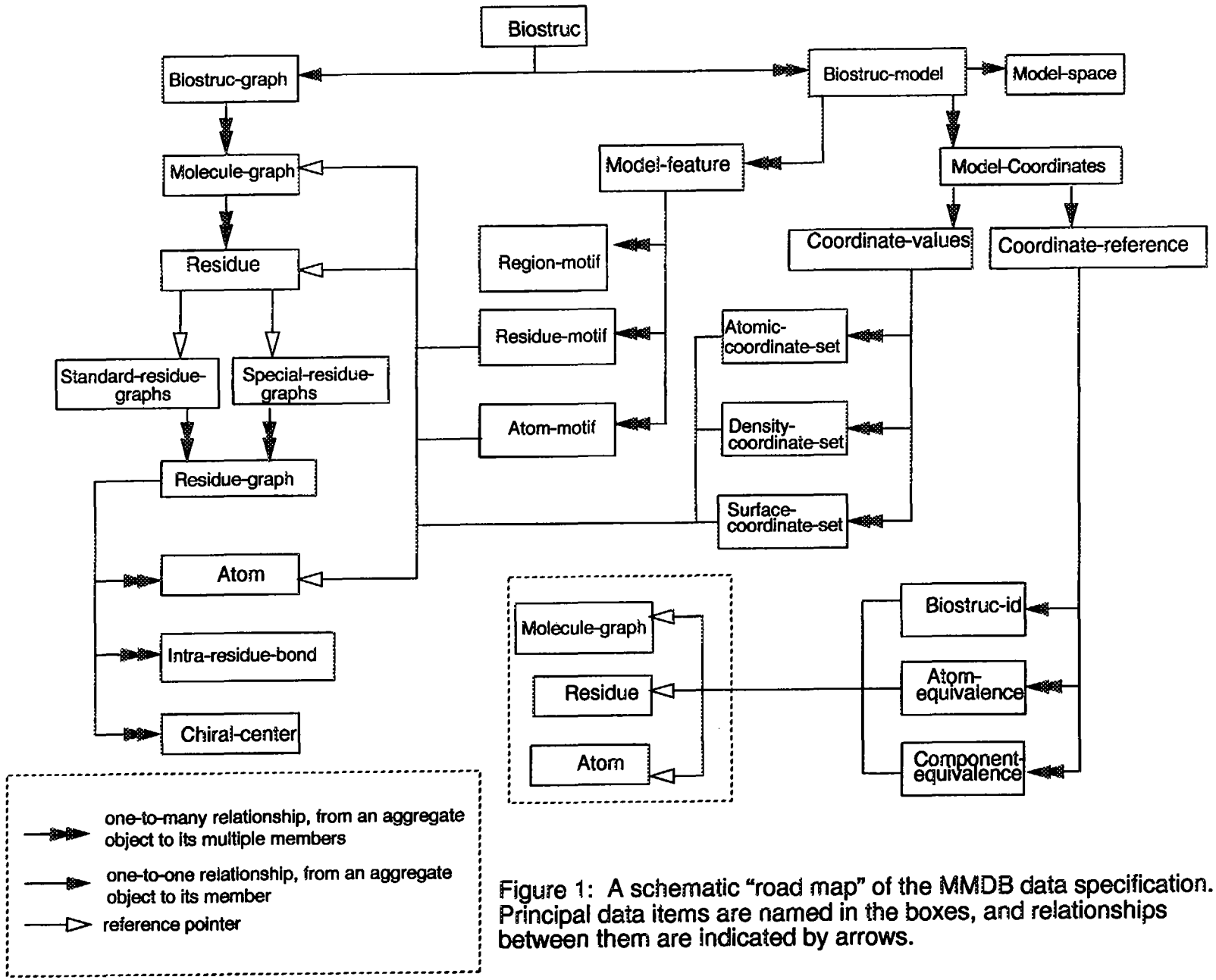

Figure 1: A schematic "road map" of the MMDB data specification. Principal data items are named in the boxes, and relationships between them are indicated by arrows.

Appendix 1

```
-- ****************************************************************************
-- Biological Macromolecule 3-D Structure Data Types for MMDB,
-- A Molecular Modeling Database
--
-- By Hitomi Ohkawa, Jim Ostell, and Steve Bryant
--
-- National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894 USA
--
-- February, 1995
-- ****************************************************************************

-- Contents of the MMDB database are currently based on files distributed by the Protein Data Bank, PDB. These data are changed in form, as
-- described in this specification. To some extent they are also changed in content, in as much as many data items implicit in PDB are made
-- explicit, and others are corrected or omitted as a consequence of validation checks. The semantics of MMDB data items are indicated by
-- comments within the specification below. These comments also explain in detail the manner in which data items from PDB have been
-- mapped into MMDB.

MMDB DEFINITIONS ::=

BEGIN

EXPORTS Biostruc, Biostruc-id;

IMPORTS Biostruc-graph FROM MMDB-Chemical-graph
        Biostruc-model FROM MMDB-Structural-model
        Biostruc-history FROM MMDB-Database-management
        Pub-set FROM NCBI-Pub;

-- A structure report or "biostruc" describes the components of a biomolecular assembly in terms of their name and descriptions, and a
-- chemical graph giving atomic formula, connectivity and chirality. It also gives one or more three-dimensional model structures, literally
-- a mapping of the atoms, residues and/or molecules of each component into a measured three-dimensional space. A model structure may
-- also describe certain structural features by name and type.

-- Note that a biostruc may contain multiple 3-dimensional models, meaning coordinate sets which describe alternative representations of
-- the spatial structure of the biomolecular assembly. Models derived from PDB files have two simplified "views" which are useful in
-- computational applications, a single-coordinate-per-atom model, and a single-coordinate-per-residue model. Both omit information
-- pertaining to ensemble models, alternate conformations and/or statistical disorder. Complete PDB structure descriptions which include
-- statistical disorder are provided as an additional model or models.

-- Note also that a biostruc may contain cross references to other databases, including citations to relevant scientific literature. These cross
-- references use object types from other NCBI data specifications, which are "imported" into MMDB.

Biostruc ::= SEQUENCE {
    id                  Biostruc-id,
    descr               Biostruc-descr,
    history             Biostruc-history,
    chemical-graph      Biostruc-graph,
    model-structure     SEQUENCE OF Biostruc-model }

-- A Biostruc-id is a collection identifiers for the molecular assembly. mmdb-id's are NCBI-assigned, and are intended to be unique and stable
-- identifiers. Other-id's are synonyms.

Biostruc-id ::= SEQUENCE {
    mmdb-id             Mmdb-id,
    other-id            SEQUENCE OF Other-id OPTIONAL }

Mmdb-id ::= INTEGER

Other-id ::= CHOICE {
    integer-id          INTEGER,
    character-id        VisibleString }

-- The description of a biostruc refers to both the reported chemical and spatial structure of a biomolecular assembly. PDB-derived
-- descriptors which refer specifically to the chemical components or spatial structure are not provided here, but instead as descriptors of the
-- biostruc-graph or biostruc-model. For PDB-derived structures the biostruc name is the PDB id-code. PDB-derived citations appear as
-- publications within the biostruc description. Biostruc citations also include a data-submission citation derived from PDB AUTHOR
-- records. Citations are described using the NCBI Pub-set specification, which is not repeated here.

Biostruc-descr ::= SEQUENCE {
    name                VisibleString OPTIONAL,
    comments            SEQUENCE OF VisibleString OPTIONAL,
    attribution         Pub-set }

END

MMDB-Chemical-graph DEFINITIONS ::=

BEGIN

EXPORTS Biostruc-graph, Molecule-id, Residue-id, Atom-id,

IMPORTS Pub-set FROM NCBI-Pub
        Org-ref FROM NCBI-Organism
        Seq-id FROM NCBI-Seqloc
        Biostruc-id FROM MMDB;

-- A biostruc graph contains the complete chemical graph of the biomolecular assembly. The assembly graph is defined hierarchically, in
-- terms of subgraphs of component molecules. For PDB-derived biostrucs, the molecules forming the assembly are the individual
-- biopolymer chains and any non-polymer or "heterogen" groups which are present.

-- The PDB-derived "compound name" field appears as the name within the biostruc-graph description. PDB "class" and "source" fields
-- appear as explicit attributes. PDB-derived structures are assigned an assembly type of "other" unless they have been further classified. If
-- they have, the source of the type classification appears as a citation within the assembly description.

-- Note that the biostruc-graph also includes as literals the subgraphs of any nonstandard residues present within it. For PDB-derived
-- biostrucs these subgraphs are constructed automatically, with validation as described below.

Biostruc-graph ::= SEQUENCE {
    descr               Biomol-descr,
    pdb-class           VisibleString OPTIONAL,
    pdb-source          VisibleString OPTIONAL,
    type                ENUMERATED {
                            physiological-form(1),
                            crystallographic-cell(2),
                            other(255) },
    molecule-graphs     SEQUENCE OF Molecule-graph,
    inter-molecule-bonds SEQUENCE OF Inter-residue-bond OPTIONAL,
    residue-graphs      SEQUENCE OF Residue-graph OPTIONAL }

-- A biomolecule description refers to the chemical structure of a molecule or component substructures. This descriptor type is used at the
-- level of assemblies, molecules and residues, and also for residue-graph dictionaries. The Org-ref object type is drawn from NCBI
-- taxonomy data specifications, and is not repeated here.

Biomol-descr ::= SEQUENCE {
    name                VisibleString,
    comments            SEQUENCE OF VisibleString OPTIONAL,
    organism            Org-ref OPTIONAL,
    attribution         Pub-set OPTIONAL }

-- A molecule chemical graph is defined by a sequence of residues. Nonpolymers are described in the same way, but may contain only a
-- single residue. Biopolymer molecules are identified within PDB entries according to their appearance on SEQRES recordership
-- formally define a biopolymer as such. Biopolymers are defined by the distinction between ATOM and HETATM coordinate records only
-- in cases where the chemical sequence from SEQRES is in conflict with coordinate data. The PDB-assigned chain code appears as the
-- name within the molecule descriptions of the biopolymers.

-- Nonpolymer molecules from PDB correspond to individual HETEROGEN groups, excluding any HETEROGEN groups which represent
-- modified biopolymer residues. These molecules are named according to the chain, residue type and residue number fields as assigned by
-- PDB. Any description appearing in the PDB HET record appear as a comment within the molecule description.
```

```
-- Molecule types for PDB-derived molecule graphs are assigned by matching residue and atom names against the PDB-documented
-- standard types for protein, DNA and RNA, and against residue codes commonly-used to indicate solvent. Classification is by "majority
-- rule". If more than half of the residues in a biopolymer are standard groups of one type, then the molecule is of that type, and otherwise
-- classified as "other". Note that this classification does not preclude the presence of modified residues, but insists they constitute less than
-- half the biopolymer. Non-polymers are classified only as "solvent" or "other".

-- Note that a molecule graph may also contain a set of cross references to other databases, in particular to a biopolymer sequence database.
-- All biopolymer molecules in MMDB contain appropriate identifiers for the corresponding entry in the NCBI-Sequences database, in
-- particular the NCBI "gi" number, which may be used for sequence retrieval. The Seq-id object type is defined in the NCBI molecular
-- sequence specification, and not repeated here.

Molecule-graph ::= SEQUENCE {
    id                      Molecule-id,
    seq-id                  Seq-id OPTIONAL,
    descr                   Biomol-descr,
    type                    ENUMERATED {
                                dna(1),
                                rna(2),
                                protein(3),
                                other-biopolymer(4),
                                solvent(5),
                                other-nonpolymer(6),
                                unknown(255) },
    residue-sequence        SEQUENCE OF Residue,
    inter-residue-bonds     SEQUENCE OF Inter-residue-bond }

Molecule-id ::= INTEGER

-- Residues may be assigned a text-string name as well as an id number. PDB assigned residue numbers appear as the residue name.

Residue ::= SEQUENCE {
    id                      Residue-id,
    name                    VisibleString OPTIONAL,
    residue-graph           Residue-graph-ptr }

Residue-id ::= INTEGER

-- Residue graphs from different sources may be referenced within a molecule graph. The allowed choices are the nonstandard residue graphs
-- included in the present biostruc, residue graphs within other biostrucs, or residue graphs within tables of standard residue definitions.

Residue-graph-ptr ::= CHOICE {
    local                   Residue-graph-id,
    biostruc                Biostruc-ptr,
    dictionary              Dictionary-ptr }

Biostruc-ptr ::= SEQUENCE {
    biostruc-id             Biostruc-id,
    residue-graph-id        Residue-graph-id }

Dictionary-ptr ::= SEQUENCE {
    standard-graphs-id      Dictionary-id,
    residue-graph-id        Residue-graph-id }

-- A residue dictionary contains a collection of standard residue graphs. The standard residue graph dictionary supplied with the MMDB
-- database contains 20 standard L amino acids and 8 standard ribonucleotide groups. These graphs are complete, including explicit
-- hydrogen atoms and separate instances for the terminal polypeptide and polynucleotide residues.

Residue-graph-dictionary ::= SEQUENCE {
    id                      Dictionary-id,
    descr                   Biomol-descr,
    residue-graphs          SEQUENCE OF Residue-graph }

Dictionary-id ::= INTEGER

-- Residue graphs define atomic formulae, connectivity, chirality, and names. For standard residue graphs from the MMDB dictionary the
-- PDB-assigned residue-type code appears as the name within the residue graph description, and the full trivial name of the residue as a
-- comment within that description. For any nonstandard residue graphs provided with an MMDB biostruc the PDB-assigned residue-type
-- code similarly appears as the name within the description, and any information provided on PDB HET records as a comment within that
-- description.

-- Note that nonstandard residue graphs for a PDB-derived biostruc may be incomplete. Current PDB format cannot represent connectivity
-- for groups which are disordered, and for which no coordinates are given. In these cases the residue graph defined in MMDB represents
-- only the subgraph that could be identified from available ATOM, HETATM and CONECT records.

Residue-graph ::= SEQUENCE {
    id                      Residue-graph-id,
    descr                   Biomol-descr,
    type                    ENUMERATED {
                                deoxyribonucleotide(1),
                                ribonucleotide(2),
                                amino-acid(3),
                                other(255) },
    iupac-code              VisibleString OPTIONAL,
    atoms                   SEQUENCE OF Atom,
    bonds                   SEQUENCE OF Intra-residue-bond,
    chiral-centers          SEQUENCE OF Chiral-center OPTIONAL,
    formal-charge           Formal-charge OPTIONAL }

Residue-graph-id ::= INTEGER

-- Atoms in residue graphs are defined by elemental symbols and names. PDB-assigned atom names appear here in the name field, except
-- in cases of known PDB synonyms. In these cases atom names are mapped to the names used in the MMDB standard dictionary. This
-- occurs primarily for hydrogen atoms, where PDB practice allows synonyms for several atom types. For PDB atoms the elemental symbol
   is obtained by parsing the PDB atom name field, allowing for known special-semantics cases where the atom name does not follow the
-- documented encoding rule. Ionizable protons are identified within standard residue graphs in the MMDB dictionary, but not within
-- automatically-defined nonstandard graphs.

Atom ::= SEQUENCE {
    id                      Atom-id,
    name                    VisibleString OPTIONAL,
    iupac-code              VisibleString OPTIONAL,
    element                 ENUMERATED {
                                h(1), he(2), li(3), be(4), b(5),
                                c(6), n(7), o(8), f(9), ne(10),
                                na(11), mg(12), al(13), si(14), p(15),
                                s(16), cl(17), ar(18), k(19), ca(20),
                                sc(21), ti(22), v(23), cr(24), mn(25),
                                fe(26), co(27), ni(28), cu(29), zn(30),
                                ga(31), ge(32), as(33), se(34), br(35),
                                kr(36), rb(37), sr(38), y(39), zr(40),
                                nb(41), mo(42), tc(43), ru(44), rh(45),
                                pd(46), ag(47), cd(48), in(49), sn(50),
                                sb(51), te(52), i(53), xe(54), cs(55),
                                ba(56), la(57), ce(58), pr(59), nd(60),
                                pm(61), sm(62), eu(63), gd(64), tb(65),
                                dy(66), ho(67), er(68), tm(69), yb(70),
                                lu(71), hf(72), ta(73), w(74), re(75),
                                os(76), ir(77), pt(78), au(79), hg(80),
                                tl(81), pb(82), bi(83), po(84), at(85),
                                rn(86), fr(87), ra(88), ac(89), th(90),
                                pa(91), u(92), np(93), pu(94), am(95),
                                cm(96), bk(97), cf(98), es(99),
                                fm(100), md(101), no(102), lr(103),
                                other(254), unknown(255) },
    ionizable-proton        ENUMERATED {
                                true(1),
                                false(2),
                                unknown(255) } OPTIONAL }

Atom-id ::= INTEGER

-- Intra-residue-bond specifies connectivity between atoms in Residue-graph. Unlike Inter-residue-bond defined later, its participating
```

```
-- atoms are part of a residue subgraph dictionary, not part of a specific biostruc-graph.

-- For residue graphs in the standard MMDB dictionary bonds are defined from the known chemical structures of amino acids and nucleotides.
-- For nonstandard residue graphs bonds are defined from PDB CONECT records, with validation for consistency with coordinate data, and
-- from stereochemical calculation to identify unreported bonds. Validation and bond identification are based on comparison of inter-atomic
-- distances to the sum of covalent radii for the corresponding elements.

Intra-residue-bond ::= SEQUENCE {
    atom-id-1                   Atom-id,
    atom-id-2                   Atom-id,
    bond-order                  ENUMERATED {
                                    single(1),
                                    partial-double(2),
                                    aromatic(3),
                                    double(4),
                                    triple(5),
                                    other(6),
                                    unknown(255)} OPTIONAL }

-- Chiral centers are atoms with tetrahedral geometry. Chirality is defined by a chiral volume involving the chiral center and 3 other atoms
-- bonded to it. For any coordinates assigned to atoms c, n1, n2, and n3, the vector triple product (n1-c) dot ( (n2-c) cross (n3-c) ) must have
-- the indicated sign. The calculation assumes an orthogonal right-handed coordinate system as is used for MMDB model structures.

-- Chirality is defined for standard residues in the MMDB dictionary, but is not assigned automatically for PDB-derived nonstandard
   residues. If assigned for nonstandard residues, the source of chirality information is described by a citation within the residue description.

Chiral-center ::= SEQUENCE {
    c                           Atom-id,
    n1                          Atom-id,
    n2                          Atom-id,
    n3                          Atom-id,
    sign                        ENUMERATED { positive(1),
                                    negative(2) } }

-- Formal charge describes the ionization state and/or oxidation state of nonpolymer groups. Units are the difference in the number of electrons
-- present relative to the number required to formally neutralize nuclear charge. Formal charge is defined for residues in the MMDB standard
-- dictionary, but not for nonstandards.

Formal-charge ::= INTEGER

-- Inter-residue bonds are defined by a reference to two atoms. For PDB-derived structures bonds are identified from biopolymer connectivity
-- according to SEQRES and from other connectivity information from SSBOND and CONECT records. These data are validated and
-- unreported bonds identified by stereochemical calculation, using the same criteria as for intra-residue bonds, are also included.

Inter-residue-bond ::= SEQUENCE {
    atom-id-1                   Atom-ptr,
    atom-id-2                   Atom-ptr,
    bond-order                  ENUMERATED {
                                    single(1),
                                    partial-double(2),
                                    aromatic(3),
                                    double(4),
                                    triple(5),
                                    other(6),
                                    unknown(255)} OPTIONAL }

-- Atoms, residues and molecules within the current biostruc are referenced by hierarchical pointers.

Atom-ptr ::= SEQUENCE {
    molecule-id                 Molecule-id,
    residue-id                  Residue-id,
    atom-id                     Atom-id }

Atom-ptr-set ::= SEQUENCE OF Atom-ptr

END
```

```
MMDB-Structural-model DEFINITIONS ::=

BEGIN

EXPORTS Biostruc-model, Model-descr, Atom-ptrs, Component-ptr;

IMPORTS Molecule-id, Residue-id, Atom-id FROM MMDB-Chemical-graph
    Model-feature, Density-coordinate-set, Surface-coordinate-set,
    Residue-ptrs, RealValue FROM MMDB-structural-features
    Biostruc-id FROM MMDB
    Pub-set FROM NCBI-Pub;

-- A structural model maps chemical components into a measured three- dimensional space and defines named structural features.
-- PDB-derived biostrucs generally contain 3 models, corresponding to "views" of the structure of a biomolecular assemble with increasing
-- levels of complexity. Models are named, in their description, in way that indicates the complexity of the view. Contents of the PDB
-- "resolution" and "expdata" fields are provided as explicit attributes.

-- The model named "PKB Single Coordinate per Atom" represents a view suitable for most computational biology applications. It
-- provides complete atomic coordinate data for a "single best" model, omitting statistical disorder information and/or ensemble structure
-- descriptions provided in the source PDB file. Construction of the single best model is based on the assumption that the contents of the
-- "alternate conformation" field from pdb imply no correlation among the occupancies of multiple sites assigned to sets of atoms: the best
-- site is chosen only on the basis of highest occupancy. Note, however, that alternate conformation sets where correlation is implied are
-- generally constrained in crystallographic refinement to have uniform occupancy, and will thus be selected as a set. For ensemble models
-- the model which assigns coordinates to the most atoms is chosen. If numbers of coordinates are the same, the model occurring first in the
-- PDB file is selected. Any statistical disorder present in ensemble models is omitted by the highest-occupancy criterion. The single best
-- model includes complete coordinates for all nonpolymer components, but omits those classified as "solvent".

-- The model named "PKB Single Coordinate per Residue" represents a simple view intended for graphic displays and rapid transmission
-- over a network. It includes only alpha carbon or backbone phosphate coordinates for biopolymers. It is based on selection of
-- alpha-carbon and backbone phosphate atoms from the "PKB Single Coordinate per Atom" model.

-- The models named "PDB Model 1", "PDB Model 2", etc. represent the complete information provided by PDB, including full
-- descriptions of statistical disorder. The name of the model is based on the contents of the PDB MODEL record, with a default name of
-- "PDB Model 1" for PDB files which contain only a single model. Construction of these models is based on the assumption that contents
-- of the PDB "alternate conformation" field are intended to imply correlation among the occupancies of atom sets flagged by the same
-- identifier. The special flag " " (blank) is assumed to indicate sites occupied in all alternate conformations, and sites flagged otherwise,
-- together with " ", to indicate a distinct member of an ensemble of alternate conformations. Note that construction of ensemble members
-- according to these assumption requires two validation checks on PDB "alternate conformation" flags: they must be unique among sites
-- assigned to the same atom, and that the special " " flag must occur only for unique sites. Sites which violate the first check are flagged as
-- "u", for "unknown"; they are omitted from all ensemble definitions but are nonetheless retained in the coordinate list. Sites which violate
-- the second check are flagged "b" for "blank", and are included in an appropriately named ensemble.

Biostruc-model ::= SEQUENCE {
    descr                       Model-descr,
    model-space                 Model-space,
    model-coordinates           SEQUENCE OF Model-coordinates OPTIONAL,
    model-features              SEQUENCE OF Model-feature OPTIONAL }

Model-descr ::= SEQUENCE {
    name                        VisibleString,
    pdb-reso                    VisibleString OPTIONAL,
    pdb-method                  VisibleString OPTIONAL,
    comments                    SEQUENCE OF VisibleString OPTIONAL,
    attribution                 Pub-set OPTIONAL }

-- The model space defines measurement units and any external reference frame. Coordinates refer to a right-handed orthogonal system
-- defined on axes tagged x, y and z in the coordinate and feature definitions of a biostruc. Coordinates from PDB-derived structures are
-- reported without change, in angstrom units. The units of temperature and occupancy factors are not defined explicitly in PDB, but are
-- inferred from their value range.

Model-space ::= SEQUENCE {
    coordinate-units            ENUMERATED {
                                    angstroms(1),
                                    nanometers(2),
                                    other(3),
```

```
                                unknown(255)},
    thermal-factor-units        ENUMERATED {
                                    b(1),
                                    u(2),
                                    other(3),
                                    unknown(255)} OPTIONAL,
    occupancy-factor-units      ENUMERATED {
                                    fractional(1),
                                    electrons(2),
                                    other(3),
                                    unknown(255)} OPTIONAL,
    density-units               ENUMERATED {
                                    electrons-per-unit-volume(1),
                                    arbitrary-scale(2),
                                    other(3),
                                    unknown(255)} OPTIONAL,
    reference-frame             Reference-frame OPTIONAL }

-- An external reference frame is a pointer to another biostruc, with an optional operator to rotate and translate coordinates into its model
-- space. This item is intended for representation of homology-derived model structures, and is not present for structures from PDB.

Reference-frame ::= SEQUENCE {
    biostruc-id                 Biostruc-id,
    rotation-translation        Rot-trans-matrix OPTIONAL }

-- Atomic coordinates may be assigned literally or by reference to another biostruc. The reference coordinate type is used to represent
-- homology-derived model structures. PDB-derived structures have literal coordinates.

Model-coordinates ::= SEQUENCE {
    descr                       Model-descr,
    coordinates                 CHOICE {
                                    literal         Coordinates,
                                    reference       Coordinate-reference } }

-- Literal coordinates map chemical components into the model space. Three mapping types are allowed, atomic coordinate models,
-- density-grid models, and surface models. A model consists of a sequence of such coordinate sets, and may thus combine coordinate subsets
-- which have a different source. PDB-derived models contain a single atomic coordinate set, by definition, since they represent information
-- from a single source.

Coordinates ::= CHOICE {
    atomic                      SEQUENCE OF Atomic-coordinate-set,
    density                     SEQUENCE OF Density-coordinate-set,
    surface                     SEQUENCE OF Surface-coordinate-set }

-- Literal atomic coordinate values give location, occupancy and order parameters, and a pointer to a specific atom defined in the biostruc
-- graph. Temperature and occupancy factors have their conventional crystallographic definitions, with units defined in the model space
-- declaration. Conformation ensembles will be present only for certain "views" of PDB structures, as described above.

Atomic-coordinate-set ::= SEQUENCE {
    descr                       Model-descr,
    coordinate-values           Atomic-coordinates,
    conformation-ensembles      SEQUENCE OF Conformation-ensemble OPTIONAL }

-- Each data item of the atomic-coordinates is a vector with the same length, such that elements in the same offset in different components
-- represent properties of the same atomic coordinate. The "number-of-values" item of atoms, locations, temperature-factors, etc. are equal.
-- This representation is used to reduce the size of ASN.1 streams.

Atomic-coordinates ::= SEQUENCE {
    atoms                       Atom-pairs,
    locations                   Model-space-points,
    temperature-factors         Atomic-temperature-factors OPTIONAL,
    occupancies                 Occupancies OPTIONAL,
    alternate-conformation-ids  Alternate-conformation-ids OPTIONAL}

-- The atoms whose location is described by each coordinate are identified via a hierarchical pointer to the chemical graph of the
-- biomolecular assembly. Coordinates may be matched with atoms in the chemical structure by the values of the molecule, residue and
-- atom id's given here, which match exactly the items of the same type defined in Biostruc-graph.

Atom-pairs ::= SEQUENCE {
    number-of-values            INTEGER,
    molecule-ids                SEQUENCE OF Molecule-Id,
    residue-ids                 SEQUENCE OF Residue-Id,
    atom-ids                    SEQUENCE OF Atom-Id }

-- Coordinates are given as integer values, with a scale factor to convert to real values for each x, y or z, in the units indicated in model-
-- space. Integer values must be divided by the scale factor. This use of integer values reduces the ASN.1 stream size. The scale factors for
-- temperature factors and occupancies are given separately, but must be used in the same fashion to produce properly scaled real values.

Model-space-points ::= SEQUENCE {
    number-of-values            INTEGER,
    scale-factor                INTEGER,
    x                           SEQUENCE OF INTEGER,
    y                           SEQUENCE OF INTEGER,
    z                           SEQUENCE OF INTEGER }

Atomic-temperature-factors ::= CHOICE {
    isotropic                   Isotropic-temperature-factors,
    anisotropic                 Anisotropic-temperature-factors }

Isotropic-temperature-factors ::= SEQUENCE {
    number-of-values            INTEGER,
    scale-factor                INTEGER,
    b                           SEQUENCE OF INTEGER }

Anisotropic-temperature-factors ::= SEQUENCE {
    number-of-values            INTEGER,
    scale-factor                INTEGER,
    b-11                        SEQUENCE OF INTEGER,
    b-12                        SEQUENCE OF INTEGER,
    b-13                        SEQUENCE OF INTEGER,
    b-22                        SEQUENCE OF INTEGER,
    b-23                        SEQUENCE OF INTEGER,
    b-33                        SEQUENCE OF INTEGER }

Occupancies ::= SEQUENCE {
    number-of-values            INTEGER,
    scale-factor                INTEGER,
    o                           SEQUENCE OF INTEGER }

-- An alternate conformation id is optionally associated with each coordinate. Aside from corrections due to the validation checks described
-- above, the contents of MMDB Alternate-conformation-ids are identical to the PDB "alternate conformation" field.

Alternate-conformation-ids ::= SEQUENCE {
    number-of-values            INTEGER,
    alternate-conf-id           SEQUENCE OF Alternate-conformation-id }

Alternate-conformation-id ::= VisibleString

-- Correlated disorder ensemble is defined by a set of alternate conformation id's which identify coordinates relevant to that ensemble.
-- These are defined from the validated and corrected contents of the PDB "alternate conformation" field as described above. A given
-- ensemble, for example, may consist of atom sites flagged by " " and "A" Alternate-conformation-ids. Names for ensembles are
-- constructed from these flags. This example would be named, in its description, "PDB Ensemble blank plus A".

-- Note that this interpretation is consistent with common PDB usage of the "alternate conformation" field, but that PDB specifications do
-- not formally distinguish between correlated and uncorrelated disorder in crystallographic models. Ensembles identified in MMDB thus
-- may not correspond to the meaning intended by PDB or the depositor. No information is lost, however, and if the intended meaning is
-- known alternative ensemble descriptions may be reconstructed directly from the Alternate-conformation-ids.

-- Note that correlated disorder as defined here is allowed within an atomic coordinate set but not between the multiple sets which may
-- define a model. Multiple sets within the same model are intended as a means to represent assemblies modeled theoretically from
-- experimentally determined structures for components, where correlated disorder between coordinate sets is not relevant.
```

```
Conformation-ensemble ::= SEQUENCE {
    id                          SEQUENCE OF Alternate-conformation-id,
    descr                       Model-descr }

-- Referenced coordinates identify another biostruc, any transformation to be applied to coordinates from that biostruc, and a mapping of
-- the chemical graph of the present biostruc onto that of the refereaced biostruc. They give an "alignment" of atoms in the current biostruc
-- with those in another, from which the coordinates of matched atoms may be retrieved. For non- atomic models "alignment" may also be
-- represented by molecule and residue equivalence lists. Referenced coordinates are a data item intended for representation of homology
-- models, with an explicit pointer to their source information. They do not occur in PDB-derived models.

Coordinate-reference ::= SEQUENCE {
    referenced-biostruc         Biostruc-id,
    transformation              Rot-trans-matrix OPTIONAL,
    coordinate-type             CHOICE {
                                    atomic          Atom-equivalence,
                                    density         Component-equivalence,
                                    surface         Component-equivalence } }

Atom-equivalence ::= SEQUENCE {
    atom                        Atom-pntrs,
    referenced-atom             Atom-pntrs }

-- Referenced density or surface coordinates give pointers to the chemical components of the present biostruc that are described by a grid or
-- surface, and pointers to the chemical components of another biostruc for which a density grid or surface has been defined.

Component-equivalence ::= SEQUENCE {
    contents                    Component-pntr,
    referenced-contents         Component-pntr }

Component-pntr ::= CHOICE {
    molecules                   Molecule-pntrs,
    residues                    Residue-pntrs,
    atoms                       Atom-pntrs }

Molecule-pntrs ::= SEQUENCE {
    number-of-ptrs              INTEGER,
    molecule-ids                SEQUENCE OF Molecule-id }

-- A rotation-translation matrix is defined by 12 numbers. The first 9 are a rotation matrix given by rows, that is with column indices varying
-- fastest. Coordinates, as a matrix with columns x, y, an z, are rotated to the target reference frame when post-multiplied by the rotation matrix.
-- The last 3 numbers are a translation. Addition to the rotated coordinates translates their origin to that of the reference frame.

Rot-trans-matrix ::= SEQUENCE {
    rot-11                      RealValue,
    rot-12                      RealValue,
    rot-13                      RealValue,
    rot-21                      RealValue,
    rot-22                      RealValue,
    rot-23                      RealValue,
    rot-31                      RealValue,
    rot-32                      RealValue,
    rot-33                      RealValue,
    tran-1                      RealValue,
    tran-2                      RealValue,
    tran-3                      RealValue }

END

MMDB-Structural-features DEFINITIONS ::=

BEGIN

EXPORTS Model-feature, Density-coordinate-set, Surface-coordinate-set,
    Residue-pntrs, RealValue;

IMPORTS Molecule-id, Residue-id FROM MMDB-Chemical-graph
    Model-descr, Atom-pntrs, Component-pntr FROM MMDB-Structural-model;

-- Named model features refer to sets of residues or atoms, or a region in the model space. A few specific feature types are allowed for
-- compatibility with PDB usage, but the purpose of a named model feature is simply to associate a text and/or citation-based description
-- with a set of atoms or residues, or a spatially-defined region of the model structure. They also support association of numeric or
-- character-defined properties with each residue or atom of a set.

-- PDB-derived secondary structure defines a single feature, represented as a sequence of residue motifs. The contents of PDB SITE and
-- FTNOTE records define features represented as sequences of atom motifs. NCBI-assigned core and secondary structure descriptions are
-- represented as a sequence of residue motifs, and cited as a data deposition in the feature description.

Model-feature ::= SEQUENCE{
    descr                       Model-descr,
    motif                       SEQUENCE OF Structural-motif }

Structural-motif ::=CHOICE {
    residue-motif               Residue-motif,
    atom-motif                  Atom-motif,
    region-motif                Region-motif }

-- Residue motifs describe secondary structure or other features defined on a set of residues. PDB secondary structure classifications from
-- HELIX, SHEET, and TURN records appear as residue motifs. The residue motif name is in this case derived from the PDB "identifier"
-- field.

Residue-motif ::= SEQUENCE {
    descr                       Model-descr,
    type                        ENUMERATED {
                                    helix(1),
                                    strand(2),
                                    sheet(3),
                                    turn(4),
                                    other(255) },
    residue-ids                 Residue-pntrs,
    residue-props               SEQUENCE OF Properties OPTIONAL }


Residue-pntrs ::= SEQUENCE {
    number-of-ptrs              INTEGER,
    molecule-ids                SEQUENCE OF Molecule-id,
    residue-ids                 SEQUENCE OF Residue-id }

-- Properties are intended to represent anything describable by numbers or text, for example the surface accessibility of a residue, the
-- chemical shift of an atom, or a code for a local environment category. The intended meaning must be specified by the name and
-- description. The number of properties with a given description is assumed to be the same as the number of residues, atoms, or surface
-- points they describe, but this is not a requirement of the specification. No properties are defined for PDB-derived structural features.

Properties ::= SEQUENCE {
    name                        VisibleString,
    descr                       SEQUENCE OF VisibleString OPTIONAL,
    number-of-props             INTEGER,
    props                       CHOICE {
                                    numeric         Num-props,
                                    text            Text-props } }
Num-props ::= SEQUENCE {
    scale-factor                INTEGER,
    property                    SEQUENCE OF INTEGER }

Text-props ::= SEQUENCE OF VisibleString

-- Atom motifs describe binding sites, catalytic sites, or other features defined on a set of atoms. For PDB SITE and FTNOTE records atom
-- motifs are named according to "site identifier" or "footnote number" fields, and their description contains comments extracted from
-- FTNOTE records, and from REMARK records if these could be linked to a particular SITE record.

Atom-motif ::= SEQUENCE {
    descr                       Model-descr,
    atom-ids                    Atom-pntrs,
```

```
    atom-props              SEQUENCE OF Properties OPTIONAL }

-- Region motifs describe features of substructures defined by spatial location such as the portion of a density grid believed to contain a binding
-- site.

Region-motif ::= SEQUENCE {
    descr                   Model-descr,
    region-boundary         CHOICE {
                                sphere      Sphere,
                                cylinder    Cylinder,
                                brick       Brick,
                                surface     Surface-grid },
    region-props            SEQUENCE OF Properties OPTIONAL}

-- Geometrical primitives are used in the definition of region motifs, and also non-atomic coordinates. Spheres, cylinders and bricks are
-- defined by a few points in the model space.

Sphere ::= SEQUENCE {
    center                  Model-space-point,
    radius                  RealValue }

Cylinder ::= SEQUENCE {
    axis-top                Model-space-point,
    axis-bottom             Model-space-point,
    radius                  RealValue }

-- A brick is defined by the coordinates of eight corners. These are assumed to appear in the order 000, 001, 010, 011, 100, 101, 110, 111,
-- where the digits 0 and 1 refer to respectively to the x, y and z axes of a unit cube. Opposite edges are assumed to be parallel.

Brick ::= SEQUENCE {
    corner-000              Model-space-point,
    corner-001              Model-space-point,
    corner-010              Model-space-point,
    corner-011              Model-space-point,
    corner-100              Model-space-point,
    corner-101              Model-space-point,
    corner-110              Model-space-point,
    corner-111              Model-space-point}

-- A grid boundary is a set of points which are assumed to define a triangle mesh surface when each is connected to it's 3 nearest neighbors.

Surface-grid ::= SEQUENCE {
    number-of-points        INTEGER,
    scale-factor            INTEGER,
    x                       SEQUENCE OF INTEGER,
    y                       SEQUENCE OF INTEGER,
    z                       SEQUENCE OF INTEGER }

Model-space-point ::= SEQUENCE {
    x                       RealValue,
    y                       RealValue,
    z                       RealValue }

RealValue ::= SEQUENCE {
    scale-factor            INTEGER,
    scaled-integer-value    INTEGER }

-- Literal density coordinates define the chemical components whose structure is described by a density grid, parameters of this grid, and
-- density values.

Density-coordinate-set ::= SEQUENCE {
    descr                   Model-descr,
    contents                Component-ptr,
    grid-corners            Brick,
    grid-steps-x            INTEGER,
    grid-steps-y            INTEGER,
    grid-steps-z            INTEGER,
    fastest-varying         ENUMERATED {
                                x(1),
                                y(2),
                                z(3)},
    slowest-varying         ENUMERATED {
                                x(1),
                                y(2),
                                z(3)},
    scale-factor            INTEGER,
    density                 SEQUENCE OF INTEGER }

-- Literal surface coordinates define the chemical components whose structure is described by a surface, and the surface itself. The surface
-- may be either a regular geometric solid or a triangle-mesh of arbitrary shape.

Surface-coordinate-set ::= SEQUENCE {
    descr                   Model-descr,
    contents                Component-ptr,
    surface                 CHOICE {
                                sphere      Sphere,
                                cylinder    Cylinder,
                                brick       Brick,
                                surface     Surface-grid },
    surface-props           SEQUENCE OF Properties OPTIONAL}


END

MMDB-Database-management DEFINITIONS ::=

BEGIN

EXPORTS Biostruc-history, Biostruc-set;

IMPORTS Date FROM NCBI-General
    Biostruc, Biostruc-id FROM MMDB;

-- The history of a biostruc indicates it's origin and it's update history within MMDB, the NCBI-maintained molecular structure database.

Biostruc-history ::= SEQUENCE {
    replaces                Biostruc-replace OPTIONAL,
    replaced-by             Biostruc-replace OPTIONAL,
    data-source             Biostruc-source OPTIONAL }

Biostruc-replace ::= SEQUENCE {
    id                      Biostruc-id,
    date                    Date }

-- The origin of a biostruc is a reference to another database. PDB release date and PDB-assigned id codes are recorded here, as are the
-- PDB-assigned entry date and replacement history.

Biostruc-source ::= SEQUENCE {
    name-of-database        VisibleString,
    version-of-database     CHOICE {
                                release-date    Date,
                                lease-code      VisibleString } OPTIONAL,
    database-entry-id       VisibleString,
    database-entry-date     Date,
    database-entry-history  SEQUENCE OF VisibleString OPTIONAL}

-- A biostruc set is a means to collect ASN.1 data for many biostrucs in one file, for purposes of facilitating data management. The object
-- type does not imply similarity of the biostrucs grouped together.

Biostruc-set ::= SEQUENCE OF Biostruc

END
```