International Tables for Crystallography (2006). Vol. F, Section 24.2.3, pp. 657–659.

24.2. The Nucleic Acid Database (NDB)

BY H. M. BERMAN, Z. FENG, B. SCHNEIDER, J. WESTBROOK AND C. ZARDECKI

24.2.1. Introduction

The Nucleic Acid Database (NDB) (Berman *et al.*, 1992) was established in 1991 as a resource for specialists in the field of nucleic acid structure. Its purpose was to gather all of the structural information about nucleic acids that had been obtained from X-ray crystallographic experiments and to organize them in such a way that it would be easy to retrieve the coordinates, the information about the experimental conditions used to derive these coordinates, and the structural information that could be derived from these coordinates. Since many NDB users are not crystallographers, the information provided by the database has been presented in such a way as to maximize its utility for various types of modelling and structure prediction.

Since the NDB was founded, many new technologies have presented new challenges and opportunities. The emergence of the World Wide Web has allowed for the creative and powerful dissemination and collection of data and information. The development of a standard interchange format for handling crystallographic data, the macromolecular Crystallographic Information File (mmCIF; Bourne *et al.*, 1997), has made it possible to ensure the integrity and consistency of the data in the archive. The NDB has used these resources to provide both a relational database and an archive of information to a global community.

Table 24.2.2.1. *The information content of the NDB*

(*a*) Primary experimental information stored in the NDB.

Structure summary – descriptor; NDB, PDB and CSD names; coordinate availability; modifications, mismatches and drugs (yes/no)
Structural description – sequence; structure type; descriptions about modifications, mismatches and drugs; description of asymmetric and biological units
Citation – authors, title, journal, volume, pages, year
Crystal data – cell dimensions; space group
Data-collection description – radiation source and wavelength; data-collection device; temperature; resolution range; total and unique number of reflections
Crystallization description – method; temperature; pH value; solution composition
Refinement information – method; program; number of reflections used for refinement; data cutoff; resolution range; *R* factor; refinement of temperature factors and occupancies
Coordinate information – atomic coordinates, occupancies and temperature factors for asymmetric unit; coordinates for symmetry-related strands; coordinates for unit cell; symmetry-related coordinates; orthogonal or fractional coordinates

(*b*) Derivative information stored in the NDB.

Distances – chemical bond lengths; virtual bonds (involving phosphorus atoms)
Torsions – backbone and side-chain torsion angles; pseudorotational parameters
Angles – valence bond angles, virtual angles (involving phosphorus atoms)
Base morphology – parameters calculated by different algorithms
Nonbonded contacts
Valence geometry r.m.s. deviations from small-molecule standards
Sequence pattern statistics

24.2.2. Information content of the NDB

Structures available in the NDB include RNA and DNA oligonucleotides with two or more bases either alone or complexed with ligands, natural nucleic acids such as tRNA, and protein–nucleic acid complexes. The archive stores both primary and derived information about the structures. The primary data include the crystallographic coordinate data, structure factors and information about the experiments used to determine the structures, such as crystallization information, data collection and refinement statistics. Derived information, such as valence geometry, torsion angles, base-morphology parameters and intermolecular contacts, is calculated and stored in the database. Database entries are further annotated to include information about the overall structural features, including conformational classes, special structural features, biological functions and crystal-packing classifications. Table 24.2.2.1 summarizes the information content of the NDB.

24.2.3. Data processing

Data processing includes data collection, integrity checking and validation of the entries. Once processing is completed, the data are entered into the database. This is accomplished using the integrated system that is illustrated in Fig. 24.2.3.1.

Structures are entered electronically into the NDB after they have been deposited directly by the experimentalist or by the NDB annotators, who scan the literature and the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al.*, 2000). The coordinate data may be deposited in any PDB format or in mmCIF format. The entries are transformed into mmCIF format and then annotated using a web-based tool (Westbrook, 1998). This tool operates on top

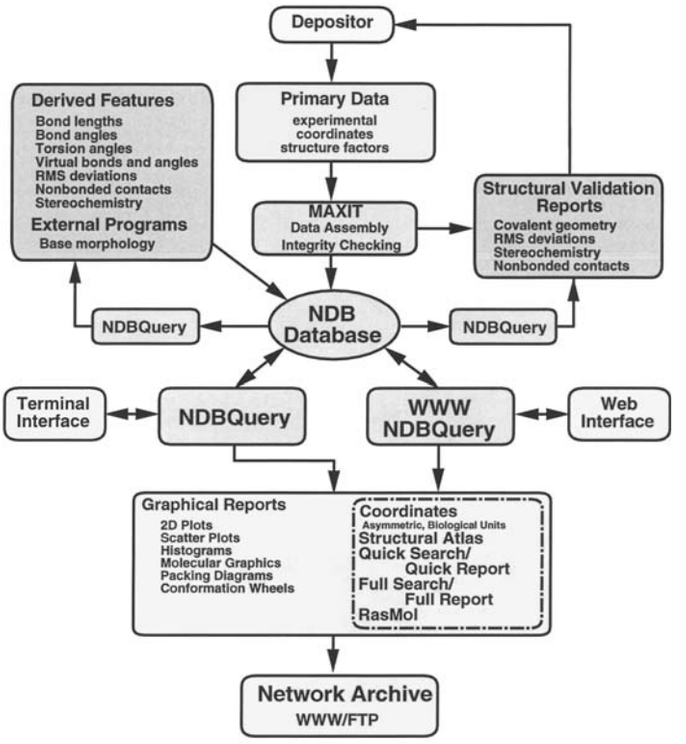


Fig. 24.2.3.1. Flow chart showing the organization of the Nucleic Acid Database Project. The core of this integrated system is the database.

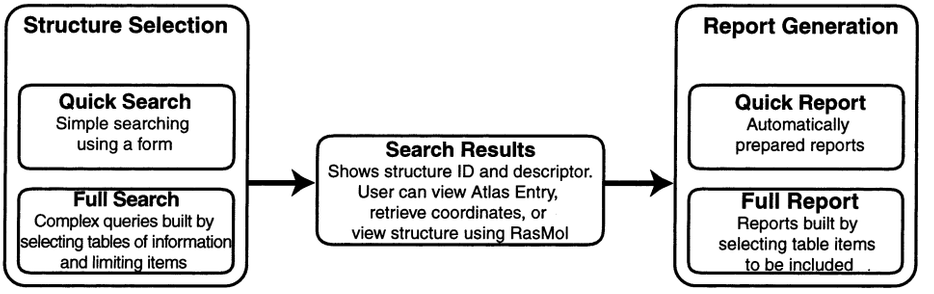
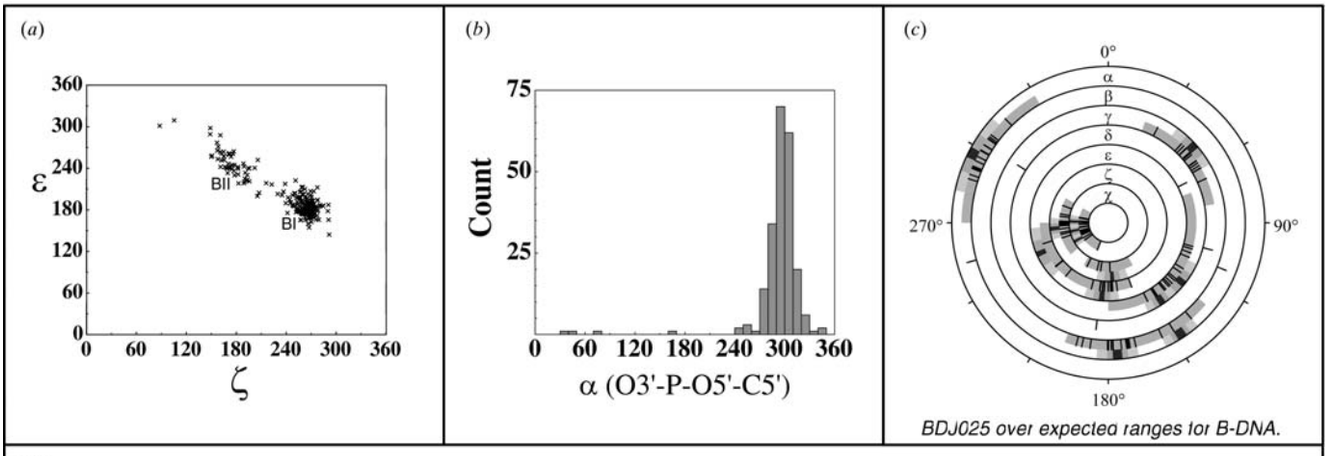


Fig. 24.2.4.1. Flow chart demonstrating the two steps involved in searching the NDB: structure selection and report generation.

of the mmCIF dictionary (Bourne *et al.*, 1997) and is used to incorporate experimental information to create a fully populated mmCIF format file. In the next stage of data processing, a program called *MAXIT* (*Macromolecular Exchange and Input Tool*; Feng, Hsieh *et al.*, 1998) checks and corrects atom numbering and ordering as well as the correspondence between the PDB SEQRES record and the residue names in the coordinate files. Once these integrity checks are completed, the structures are validated using a variety of programs.

NUCheck (Feng, Westbrook & Berman, 1998) verifies valence geometry, torsion angles, intermolecular contacts and the chiral centres of the sugars and phosphates. The dictionaries used for checking the structures were developed by the NDB project from analyses (Clowney *et al.*, 1996, Gelbin *et al.*, 1996) of high-resolution small-molecule structures from the Cambridge Structural Database (CSD; Allen *et al.*, 1979). The torsion-angle ranges were derived from an analysis of high-resolution nucleic acid structures (Schneider *et al.*, 1997). One important outgrowth of these



(d)

Torsion Angles alpha-chi

NDB ID	Residue	Strand	o3_p_o5_c5	p_o5_c5_c4	o5_c5_c4_c3	c5_c4_c3_o3	c4_c3_o3_p	c3_o3_p_o5	o4_c1_n1_9_c2_4
BDJ025	C--1	A	n.a.	n.a.	-57.49	133.53	-129.49	-166.70	-98.61
BDJ025	G--2	A	-57.92	148.84	46.40	147.20	176.79	-108.11	-98.98
BDJ025	A--3	A	-67.23	178.65	56.20	129.59	179.97	-89.31	-114.33
BDJ025	T--4	A	-67.97	177.92	56.18	110.09	-175.34	-103.92	-120.95
BDJ025	C--5	A	-54.82	170.04	54.93	105.67	-165.55	-82.28	-126.34
BDJ025	G--6	A	-68.21	177.75	46.15	146.37	-141.50	-170.75	-89.89
BDJ025	A--7	A	-55.10	154.78	42.81	139.17	-176.81	-97.39	-111.85
BDJ025	T--8	A	-63.06	171.56	59.21	121.53	176.25	-101.70	-114.46
BDJ025	C--9	A	-44.73	171.34	49.13	124.45	-174.93	-79.48	-106.94
BDJ025	G--10	A	-72.50	-179.65	50.76	131.06	n.a.	n.a.	-86.62
BDJ025	C--11	B	n.a.	n.a.	-173.44	153.42	-129.94	178.02	-80.61
BDJ025	G--12	B	-60.51	139.09	49.69	140.22	159.98	-93.59	-99.46
BDJ025	A--13	B	-67.76	-171.69	57.76	135.24	-179.89	-91.37	-105.44
BDJ025	T--14	B	-54.09	173.97	44.44	107.40	177.82	-90.30	-110.90
BDJ025	C--15	B	-60.22	167.42	59.32	113.21	-172.95	-92.36	-109.72
BDJ025	G--16	B	-70.70	179.87	54.89	145.99	-107.84	-154.03	-93.23
BDJ025	A--17	B	-105.75	112.04	104.73	62.79	-172.66	-69.21	-157.78
BDJ025	T--18	B	-57.84	170.72	65.66	131.07	172.35	-88.07	-126.25
BDJ025	C--19	B	-64.90	-167.90	54.67	136.48	178.01	-89.49	-111.38
BDJ025	G--20	B	-53.12	-165.28	27.39	147.28	n.a.	n.a.	-79.35

Created by the Nucleic Acid Database Project on Tue Dec 22 16:48:58 1998

Fig. 24.2.4.2. Examples of reports generated from the NDB about torsion angles. (*a*) A scattergram showing the relationship of ε (C4′—C3′—O3′—P) *versus* ζ (C3′—O3′—P—O5′). The two clusters, BI and BII, are labelled. (*b*) A histogram for α (O3′—P—O5′—C5′) for all B-DNA. (*c*) A conformation wheel showing the torsion angles for structure BDJ025 (Grzeskowiak *et al.*, 1991) over the average values for all B-DNA. (*d*) A torsion-angle report for BDJ025.

Table 24.2.5.1. *Quick reports available from the NDB*

Report name	Contents
NDB status	Processing status information
Cell dimensions	Crystallographic cell constants
Primary citation	Primary bibliographic citations
Structure identifier	Identifiers, descriptor, coordinate availability
Sequence	Sequence
Nucleic acid sequence	Nucleic acid sequence only
Protein sequence	Protein sequence only
Refinement information	*R* factor, resolution and number of reflections used in refinement
Nucleic acid backbone torsions (NDB)	Sugar–phosphate backbone torsion angles using NDB residue numbers
Nucleic acid backbone torsions (PDB)	Sugar–phosphate backbone torsion angles using PDB residue numbers
Base-pair parameters (global)	Global base-pair parameters calculated using *Curves* 5.1 (Lavery & Sklenar, 1989)
Base-pair step parameters (local)	Local base-pair step parameters calculated using *Curves* 5.1
Groove dimensions	Groove dimensions using Stoffer & Lavery definitions from *Curves* 5.1

validation projects was the creation of the force constants and restraints that are now in common use for crystallographic refinement of nucleic acid structures (Parkinson *et al.*, 1996). The program *SFCHECK* (Vaguine *et al.*, 1999) is used to validate the model against the structure-factor data. The *R* factor and resolution are verified and the residue-based features are examined with this program. Once an entry has been processed satisfactorily, it is entered into the database.

24.2.4. The database

The core of the NDB project is a relational database in which all of the primary and derived data items are organized into tables. At present, there are over 90 tables in the NDB, with each table containing five to 20 data items. These tables contain both experimental and derived information. Example tables include: the citation table, which contains all the items that are present in literature references; the cell_dimension table, which contains all items related to crystal data; and the refine_parameters table, which contains the items that describe the refinement statistics.

Interaction with the database is a two-step process (Fig. 24.2.4.1). In the first step, the user defines the selection criteria by combining different database items. As an example, the user could select all B-DNA structures whose resolution is better than 2.0 Å, whose *R* factor is better than 0.17, and which were determined by the authors Dickerson, Kennard, or Rich. Once the structures that meet the constraint criteria have been selected, reports may be written using a combination of table items. For any set of chosen structures, a large variety of reports may be created. For the example set of structures given above, a crystal-data report or a backbone torsion-angle report can be easily generated, or the user could write a report that lists the twist values for all CG steps together with statistics, including mean, median and range of values. The constraints used for the reports do not have to be the same as those used to select the structures. Some examples of reports from the NDB are given in Fig. 24.2.4.2.

24.2.5. Data distribution

Data are made available *via* a variety of mechanisms, such as ftp and the World Wide Web. Coordinate files, reports, software programs and other resources are available *via* the ftp server (ndbserver.rutgers.edu). In addition to links to the ftp server, the web server provides a variety of methods for querying the NDB and accessing reports prepared from the database (http://ndbserver.rutgers.edu/).

24.2.5.1. *Archives*

The NDB archives, a section of the web site, contain a large variety of information and tables useful for researchers. Prepared reports about the structure identifiers, citations, cell dimensions and structure summaries are available and are sorted according to structure type. The dictionaries of standard geometries of nucleic acids as well as parameter files for *X-PLOR* (Brünger, 1992) are also available. The archives section links to the ftp server, providing coordinates for the asymmetric unit and biological units in PDB and mmCIF formats, structure-factor files, and coordinates for nucleic acid structures determined by NMR.

24.2.5.2. *Atlas*

A very popular and useful report is the NDB Atlas report page. An Atlas page contains summary, crystallographic and experimental information, a molecular view of the biological unit and a crystal-packing picture for a particular structure. Atlas pages are created directly from the NDB database (Fig. 24.2.5.1). The Atlas entries for all structures in the database are organized by structure type on the NDB web site.

24.2.5.3. *NDB searches*

A web interface was designed to make the query capabilities of the NDB as widely accessible as possible. To highlight the special features of NDB, the interface operates in two modes. In the quick search/quick report mode, several items, including structure ID, author, classification and special features, can be limited either by entering text in a box or by selecting an option from the pull-down menu. Any combination of these items may be used to constrain the structure selection. If none are used, the entire database will be selected. After selecting 'Execute Selection', the user will be presented with a list of structure IDs and descriptors that match the desired conditions. Several viewing options for each structure in this list are possible. These include retrieving the coordinate files in either mmCIF or PDB format, retrieving the coordinates for the biological unit, viewing the structure with *RasMol* (Sayle & Milner-White, 1995), or viewing an NDB Atlas page.

Preformatted quick reports can then be generated for the structures in this results list. The user selects a report from a list of 13 report options (Table 24.2.5.1), and the report is created

references