Acta Cryst. (1998). D**54**, 1095–1104

The Nucleic Acid Database: A Resource for Nucleic Acid Science

Helen M. Berman,* Christine Zardecki and John Westbrook

NDB, Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087, USA.
E-mail: berman@adenine.rutgers.edu

(*Received 6 March 1998; accepted 4 June 1998*)

Abstract

The Nucleic Acid Database (NDB) distributes information about nucleic acid-containing structures. Here the information content of the database as well as the query capabilities are described. A summary of how the technology developed by this project has been used to develop other macromolecular databases is given.

1. Introduction

Nucleic acids have distinctive chemical and structural features. There are 12 conformation angles which describe the bonds interconnecting the base, sugars and phosphates, and 16 different parameters which describe the geometry of the base pairs. This potential conformational variability allows double-stranded DNA to be flexible enough to be coiled in the cell and for single-stranded RNA to fold into compact functional units. Single-crystal nucleic acid crystallography began in the 1970's, almost 20 years after the discovery of the double-helical form of DNA by Watson and Crick (Watson & Crick, 1953). By 1990, there were more than 150 structures of DNA and RNA oligonucleotides, t-RNA, and a handful of protein–nucleic acid complexes. The sample set was thus large enough to begin to ask questions about the effects of sequence and environment on the structures of these biological molecules. The vision behind the creation of the Nucleic Acid Database (NDB; Berman *et al.*, 1992) was to establish a resource which enables researchers to easily answer these questions and to facilitate, in general, research about nucleic acid structures. This resource would be realised through the creation of a database of information about nucleic acids which would have robust query capabilities. Thus, not only would the users be able to extract coordinates for a particular structure, but they would also be able to select groups of structures based on particular characteristics and extract the experimental information and derived features of these structures. In addition, it was desired to be able to make correlations among different characteristics of the structures so that eventually the database would become a predictive tool.

At the time that the NDB was established, nucleic acid structural information was found in both the Cambridge Structural Database (CSD; Allen *et al.*, 1979) and the Protein Data Bank (PDB; Bernstein *et al.*, 1977). Containing small nucleotides and dinucleoside phosphates, the CSD had a very robust query engine but it did not provide the specialized information required to understand nucleic acid structures. The PDB contained the larger oligonucleotides, but lacked a query interface; although it now provides a browser, it is not possible to extract derived data about groups of structures. The original scope of the NDB project was to create a 'value-added' database containing structural data for both RNA and DNA. The NDB is now the direct deposition site for these structures and has created a database with information about all nucleic acid-containing crystals. The ways in which the NDB is used to support research on nucleic acids are described here. Additionally, we describe how we have applied the technology developed by the NDB to other types of macromolecular databases.

2. Information content of the NDB

In addition to coordinate data, the NDB contains information about the experiments used to determine the structures, including crystallization information, data-collection information and refinement statistics. The database also stores derived information such as valence geometry, torsion angles, base morphology parameters and intermolecular contacts. Further annotation includes information about the overall structural features. These include the conformational classes, special structural features, biological functions, and crystal packing classifications. Table 1 summarizes these derived data and annotations. The organization of these features into a database makes it possible to gain new insights about structure using advanced techniques such as data mining.

3. Validation procedures for nucleic acids

Structures are added to the database *via* two routes: direct deposition (nucleic acids) and post processing of PDB files (protein–nucleic acid complexes). All data are transformed into mmCIF files which allows them to be

Table 1. *Summary of derived data and annotations*

(*a*) Annotations stored in the NDB
Structural features
NDB, PDB and CSD identifiers
Availability of coordinates
Availability of structure factors
Sequence
Conformation type
Description of modifiers of base, phosphate and sugar
Mismatched base-pairs
Name and binding type for drug
Description of base pairing
Description of asymmetric unit
Description of the biological unit
Crystal packing motif
(*b*) Derived data stored in the NDB
Covalent bond lengths and angles
Nonbonded contacts
Virtual bonds and angles involving P atoms
Backbone and side-chain torsion angles
Pseudorotation parameters
Base morphology parameters
Valence geometry r.m.s. deviations from small-molecule standards
Sequence pattern statistics

checked automatically against the mmCIF dictionary (Bourne *et al.*, 1997). This creates a uniform archive to ensure reliable query results. Structure checking is accomplished using a suite of programs that verify valence geometry, torsion angles, intermolecular contacts and chirality. The dictionaries used for checking the structures were developed from analyses of high-resolution small-molecule structures from the CSD (Clowney *et al.*, 1996; Gelbin *et al.*, 1996). The torsion-angle ranges were analyzed from high-resolution nucleic acid structures (Schneider *et al.*, 1997). One important outgrowth of this work 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).

4. The characteristics and uses of the database

4.1. *Basic features*

The core of the NDB project (Fig. 1) 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 containing five to 20 items. These tables contain both experimental and derived information. Example tables include: the citation table, which contains all the items that are contained 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. 2). In the first step, the user

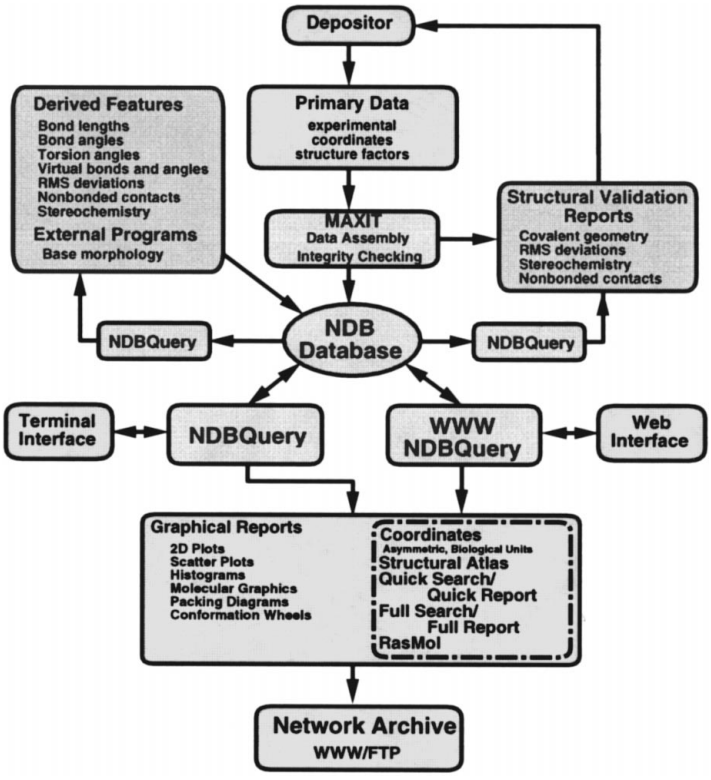


Fig. 1. Flow chart showing the organization of the Nucleic Acid Database project. The core of this project is the database.

Table 2. *Examples of the Boolean logic used to construct an NDB query*

Structure selection of B-DNAs with resolution $\leq$ 1.9 Å and R factor < 0.17 by authors A. Rich, R. E. Dickerson or O. Kennard.

Table	Attribute (Item)	Operator	Operand	Logical
structure_summary	Conformation_Type	=	B	AND
structure_summary	Classification	=	DNA	AND
r_factor	Upper_Resol_Limit	< =	1.9	AND
r_factor	R_Value	<	0.17	AND
citation	Authors	like	R.E.Dickerson	OR
structure_summary	Conformation_Type	=	B	AND
structure_summary	Classification	=	DNA	AND
r_factor	Upper_Resol_Limit	< =	1.9	AND
r_factor	R_Value	<	0.17	AND
citation	Authors	like	A.Rich	OR
structure_summary	Conformation_Type	=	B	AND
structure_summary	Classification	=	DNA	AND
r_factor	Upper_Resol_Limit	< =	1.9	AND
r_factor	R_Value	<	0.17	AND
citation	Authors	like	O.Kennard	

defines the selection criteria by combining different database items. As an example, one could select all B-DNA structures with resolution better than 2.0 Å, R factor better than 0.17 and that they were determined by Dickerson, Kennard or Rich. The logic for this constraint is shown in Table 2.

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 given above, a crystal data report or a backbone torsion-angle report can be generated easily. 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.

The NDB can also be used to create graphical reports. It is possible to produce pie charts, scattergrams and histograms describing any aspect of a selected group of structures. These capabilities were put to full use in deriving the ranges of torsion angles for different types of DNA helices (Schneider *et al.*, 1997). Fig. 3 shows some examples of the variety of reports one can make about torsion angles.

Another very popular and useful report is the NDB Atlas report page. Atlas pages are created directly from the NDB database for a particular structure and contain summary, crystallographic and experimental information about the structure, a molecular view of the biological unit and a crystal packing picture (Fig. 4). The Atlas section of the NDB WWW site contains Atlas entries for all of the structures in the database organized by structure type.

4.2. *Query capabilities*

4.2.1. *Character menu*. The most direct access to the database is through the use of SQL commands. However, given that this is a specialist language, a character menu has been constructed which allows access to all of the tables in the database. Queries and reports are constructed menu by menu. Both query and report constraints can be saved in a 'command' file which allows the user to access, revise and edit these constraint and report definitions at any time. These command files facilitate making multiple reports about a particular set of structures quickly and easily. This method is used to generate the summary

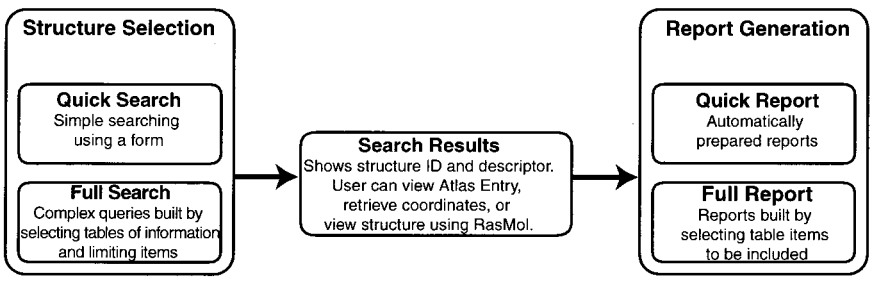
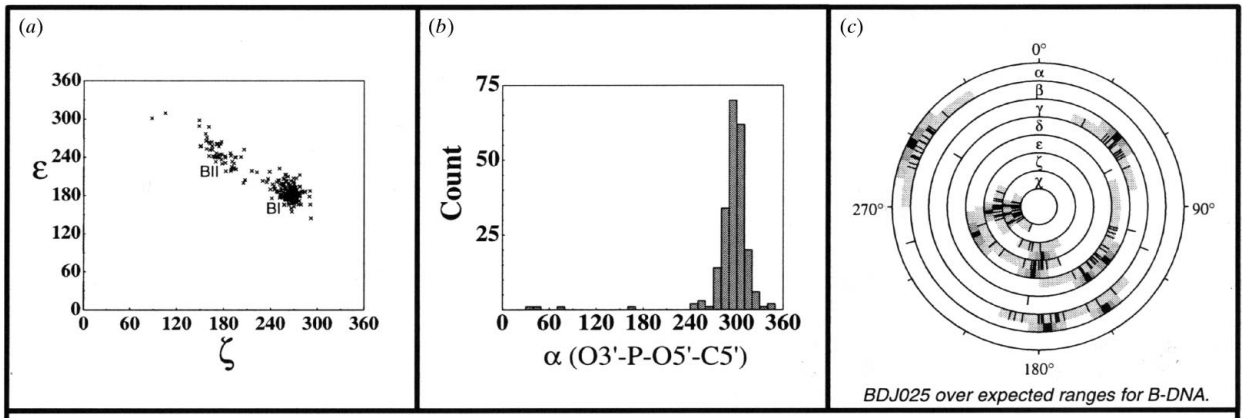


Fig. 2. Flow chart describing the steps involved in using the NDB: structure selection and report generation.

Table 3. *Structure Finder Quick Reports*

These prepared reports can be created for structures selected from the following databases using Structure Finder (http://ndbserver.rutgers.edu/)

Structure Finder database	Report name	Contains
NDB	NDB Status	Processing status information
	Cell Dimensions	Crystallographic cell constants
	Primary Citation	Primary bibliographic citations
	Structure Identifier	Identifiers, descriptor, coordinate availability
	Sequence	Strand ID, sequence, strand length
	NA Backbone Torsions	Sugar–phosphate backbone torsion angles using NDB residue numbering
	NA Backbone Torsions	Sugar–phosphate backbone torsion angles using PDB residue numbering
	Base Pair Parameters (global)	Global base-pair parameters calculated using *Curves* 5.1 (Lavery & Sklenar, 1988)
	Base Pair Step Parameters (local)	Local base-pair step parameters calculated using *Curves* 5.1
	Groove Dimensions	Groove dimensions using Stoffer and Lavery definitions from *Curves* 5.1
DNA-Binding Protein	Cell Dimensions	Crystallographic cell constants
	Primary Citation	Primary bibliographic citations
NMR Nucleic Acids	Primary Citation	Primary bibliographic citations
	Descriptor	Descriptor information
Proteins Plus	Cell Dimensions	Crystallographic cell constants
	Primary Citation	Primary bibliographic citations
	Structure Identifier	Identifiers, descriptor, coordinate availability
	Sequence	Strand ID, sequence, strand length
	Experimental Technique	Descriptor and experimental technique



(d)

res_no	res_name	α	β	γ	δ	ε	ζ	χ
1	C	n.a.	n.a.	-57.492	133.531	-129.486	-166.699	-98.607
2	G	-57.916	148.835	46.400	147.196	176.793	-108.109	-98.985
3	A	-67.234	178.651	56.203	129.589	179.965	-89.203	-114.326
4	T	-67.970	177.924	56.182	110.090	-175.340	-103.918	-120.950
5	C	-54.816	170.042	54.928	105.671	-165.549	-82.277	-126.345
6	G	-68.208	177.753	46.154	146.374	-141.504	-170.751	-89.892
7	A	-55.095	154.781	42.811	139.174	-176.812	-97.391	-111.854
8	T	-63.064	171.558	59.214	121.530	176.255	-101.698	-114.459
9	C	-44.728	171.339	49.127	124.453	-174.926	-79.477	-106.938
10	G	-72.496	-179.655	50.758	131.065	n.a.	n.a.	-86.621

truncated

Fig. 3. Examples of the different types of reports that can be generated from the NDB about torsion angles: (*a*) scattergram graph showing the relationship of ε(C4′—C3′—O3′—P) *versus* ζ(C3—O3′—P—O5′). The two clusters, BI and BII, are labeled; (*b*) histogram for α (O3′—P′—O5′—C5′) for all B-DNA; (*c*) 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.

reports produced regularly for the NDB WWW Archives.

4.2.2. *WWW interface*. The WWW 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, there are several items, including structure ID, author, classification and special features, which 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

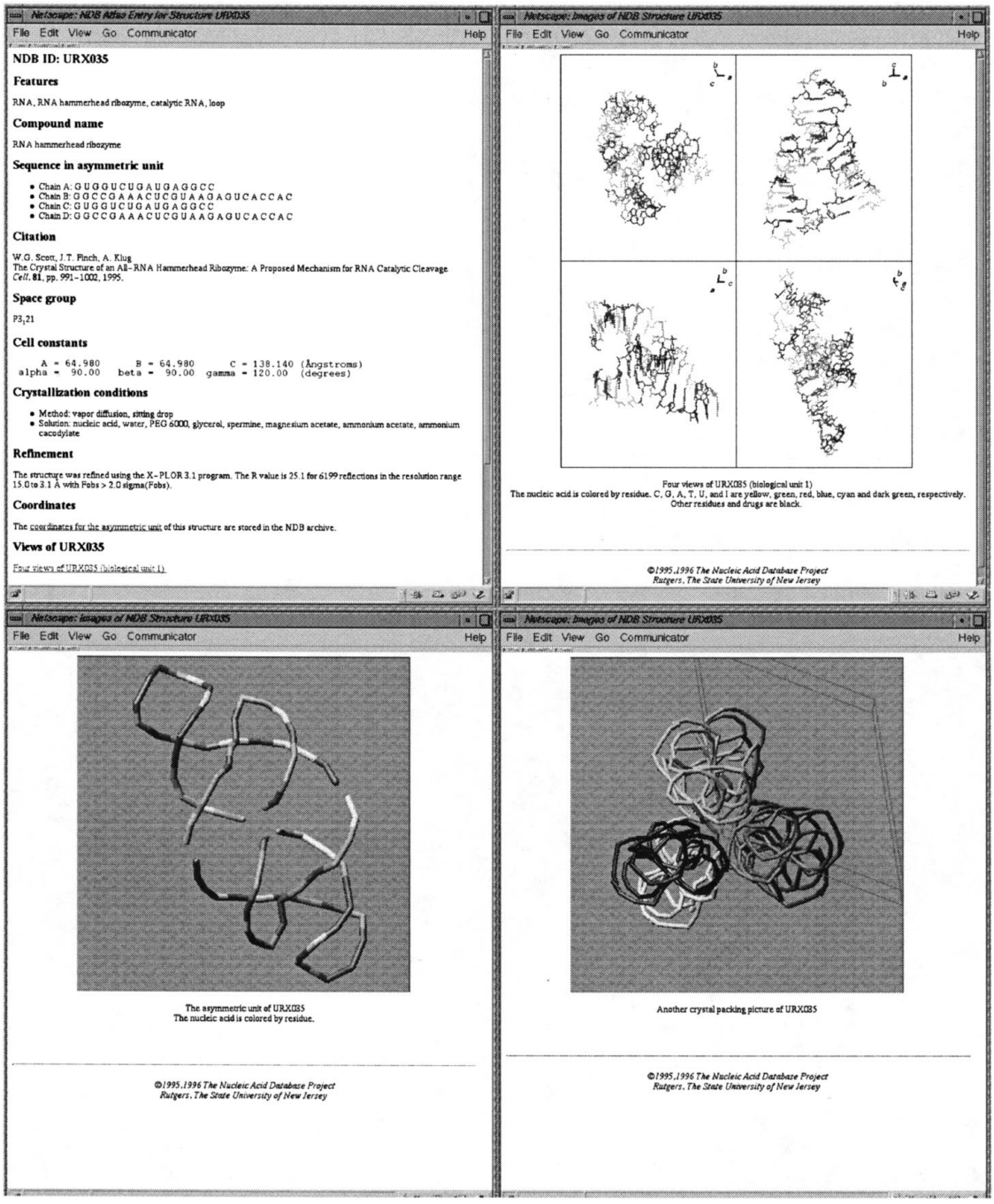


Fig. 4. NDB Atlas page for URX035 (Scott *et al.*, 1995) which highlights structural information that is contained in the database and provides images of the biological unit, asymmetric unit, and crystal packing of the structure.

will be presented with a list of IDs and descriptors of the structures 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 options, and the report is created automatically. For the structures in NDB, there are 12 different types of Quick Reports as shown in Table 3. These reports are particularly convenient for being able to quickly produce reports based on derived features such as torsion angles and base morphology.

In the Full Search/Full Report mode, it is possible to access most of the tables in the NDB and create more complex queries. Instead of selecting from a limited number of options, the user builds a search by selecting the tables, and then the items, that contain the desired features. These queries can use Boolean and logical operators to make complex queries. An example Full Search, finding transcription repressors that bind to DNA containing the step T G, is shown in Fig. 5.

After selecting structures using Full Search, a variety of reports can be written. The user selects the items that are to be displayed in a report by going through the tables that include the desired information. Multiple reports can be generated for the same group of selected structures; one report could list the DNA sequences, and then another report could present the base morphology of the bases in these sequences. Shortly, it will be possible to draw scattergrams and histograms interactively using JAVA applets to dynamically analyze the structural features.

Structure Selection using Full Search

To select structures using the NDB Full Search, the user finds the "table" that contains the information that is to be searched and then selects and limits that item. A search for transcription repressors that bend to DNA with TG steps is shown below.

Table Name	Attribute (Item)	Operator	Operand	Logical
sequence	Sequence	"like any"	T G	AND
binding_mode_regulatory	Type	"like any"	TRANSCRIPTION FACTOR/REPRESSOR	AND

After selecting the above items, select "Search Results" to view the list of structures that match this requirement. For each structure in the list, the user may retrieve the coordinates in PDB and mmCIF format, view the NDB Atlas entry, or view the structure using RasMol.

Report Generation using Full Report

Reports are created by simply selecting the desired items. The following tables could be used to create a sequence report which would show the strand ID and the protein or nucleic acid sequence.

Table Name	Attribute (Item)
sequence	Strand_ID
sequence	Sequence

To view the report select "Display Report". This report can be printed or saved. Select "Reset Report" to begin a new report.

The following tables could be used to create a base morphology report for the nucleic acid in the structures selected:

Table Name	Attribute (Item)
base_step_local_lavery	step_number
base_step_local_lavery	step_name
base_step_local_lavery	shift
base_step_local_lavery	slide
base_step_local_lavery	rise
base_step_local_lavery	twist
base_step_local_lavery	tilt
base_step_local_lavery	roll

Then select "Display Report" to view the report. This report can be printed or saved. Select "Reset Report" to begin a new report.

Fig. 5. Example of a complex query: selecting structures with a T G step from the NDB and generating reports.

Another variation on this query would be to select the protein–DNA complexes which contain a particular sequence pattern, *e.g.* ACA, and then write a report which gives the binding mode of these structures. A report showing the binding mode demonstrates that most of these structures are regulatory. To further refine the report, the user can also include the type of regulatory protein. The report that is produced is shown in Fig. 6.

Experimental features may be explored by constructing a query to search for structures with cell dimensions within a particular range. It is also possible to search for some aspects of crystallization conditions, although the information collected on crystallization by the NDB is more limited than that found at the Biological Macromolecule Crystallization Database (BMCD) (Gilliland *et al.*, 1994).

5. NDB Archives

The NDB Archives contain a large variety of information and tables useful for researchers. These include a variety of prepared reports that are sorted according to structure type (Fig. 7). The dictionaries of standard geometries of nucleic acids are here as well as parameter files for *X-PLOR* (Brünger, 1992). The ftp server provides coordinates for the asymmetric unit and

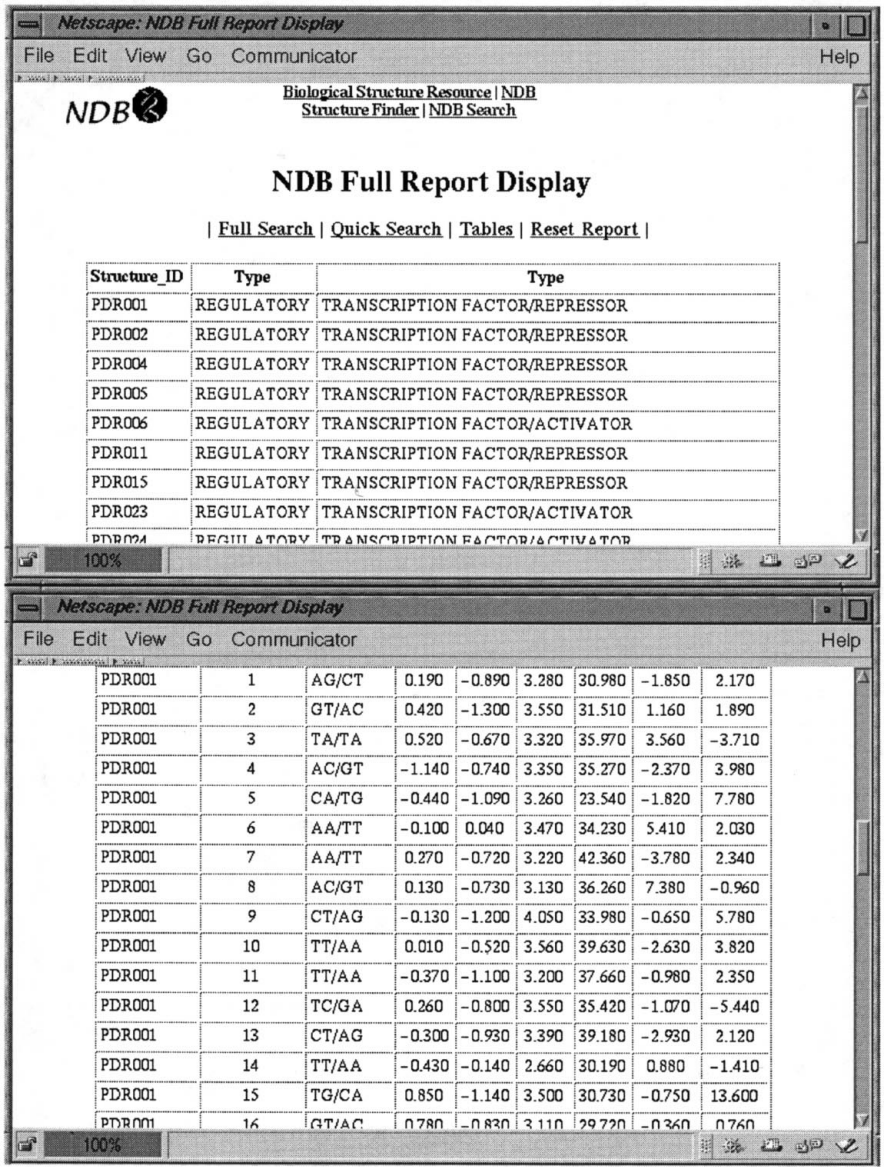


Fig. 6. Reports created for protein–DNA complexes containing the nucleic acid sequence ACA. (*a*) The top report displays the types of regulatory proteins that bind to the sequence ACA. (*b*) The bottom report displays the nucleic acid base morphology for the DNA in these complexes.

biological units in PDB and mmCIF formats, structure factor files, and coordinates for nucleic acid structures determined by NMR.

6. Other applications of NDB technology

The underlying technology of the NDB has also been used to create relational databases for other classes of macromolecules. The WWW interface, called Structure Finder, is designed so that specific tables of information can be turned on or off as appropriate for a particular structural class. The decision to turn a table on or off depends on two things: the quality of the underlying data and the appropriateness of that table to the structure class in the database. Four databases created by the NDB Project are described here and can be found under Structure Finder at the NDB Biological Structure Resource (http://ndbserver.rutgers.edu/). Tutorials for all of the Structure Finder databases are also available at this site.

6.1. *Proteins Plus*

This database is created from all the files in the current PDB and is updated regularly. *MAXIT* (Feng *et al.*, 1997), a software tool developed by the NDB, extracts data from PDB files which are loaded into the database tables. A drawback of this database is that the PDB format has changed over the years so that abstracting information from these files is not entirely reliable. To remedy this, the NDB and the Macro-

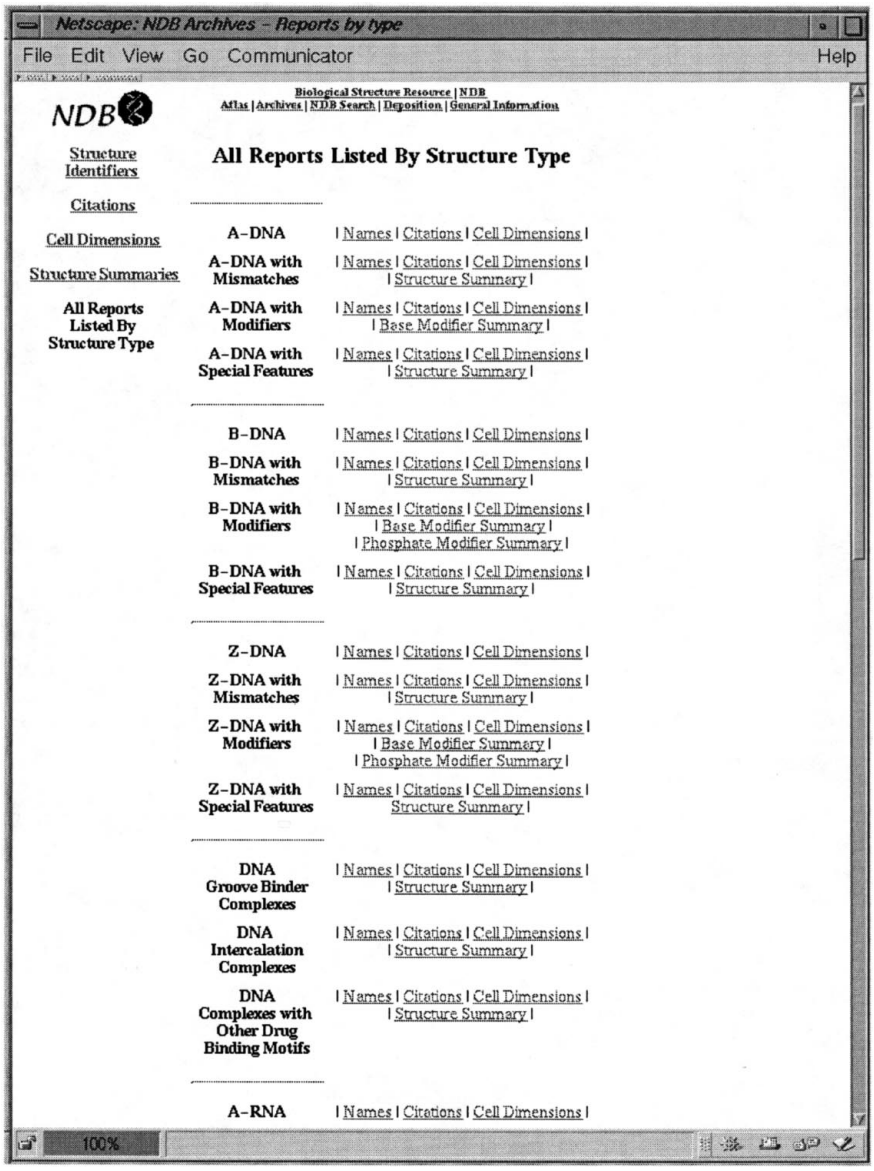


Fig. 7. Prepared Reports available from the Archives section. Reports are also available for A-RNA with mismatches, modifiers, and special features; RNA–drug complexes; t-RNA; ribozymes; single-stranded DNA and RNA; parallel-stranded DNA and RNA; DNA–RNA hybrids; DNA quadruplexes; protein–DNA enzymes, regulatory, structural, and other; protein–RNA enzymes, regulatory and structural.

molecular Structure Database (MSD) group at EBI are working to put the PDB files into a uniform format and add any missing data items. In the cases in which files have been recurated, the newly processed file is available from Proteins Plus. Once all the files have been remediated, many additional tables will be available for searching.

The Proteins Plus database can be searched using Quick Search/Quick Reports and Full Search/Full Reports, which are used in the same way as for the NDB. An example of a Proteins Plus query and report session might be to search for all myoglobin structures and then to create a report of all crystal data or another which lists the positions of the helices (Fig. 8).

6.2. *DNA binding proteins*

All proteins which bind to DNA have been fully curated and annotated and placed into a database. The database includes protein–DNA complexes as well as proteins that bind to DNA but do not have DNA in the crystal. The functions of these proteins are available as both search targets and report content.

6.3. *Nucleic acid NMR*

This database contains all NMR structures that contain nucleic acids. This database can be searched using Quick Search/Quick Report. Further curation of these structures is a future project.

7. Summary

The NDB Project has evolved over the years. The original NDB contained fewer than 100 crystal structures of nucleic acids; there are now over 700. In addition, the project has created a suite of speciality databases and technologies that will allow for the evolution of these databases. The strength of the search engines developed by this project is that they allow for rapid selection of structures by a wide variety of criteria. Once selected, the coordinates of these structures, along with the tabular and graphical report capabilities of the NDB, can be used to understand a large number of the characteristics of these structures. The curated and reliable files that are provided by the NDB can also be used by other independent programs.

8. Access

The home for the Nucleic Acid Database can be found on the NDB Biological Structure Resource Home page (http://ndbserver.rutgers.edu/). On this page are pointers to the NDB and Structure Finder, as well as to *à la mode* (Clowney & Westbrook, 1997), which is a database of ligands and monomer units, and to the mmCIF WWW site. In addition to backup mirror sites at Rutgers University, the NDB is mirrored at the European Bioinformatics Institute (Europe), NIBH-AIST (Japan), and San Diego Supercomputer Center (US), with additional public mirrors currently in development. These mirrors are kept synchronous by using software tools developed by the project.

Structure Selection: Myoglobin Structures

To search for myoglobin structures, use the Proteins Plus Quick Search to enter the word "myoglobin" in the Descriptor field provided. Select "Execute Selection" to view the Search Results list. For each structure in the list, the user may retrieve the coordinates in PDB and mmCIF format, view the NDB Atlas entry, or view the structure using RasMol.

Report Generation: Helical Report

Reports can be generated using either Quick Report or Full Report Search. For a list of possbile Quick Reports, see **Table 3**.

To create a report on helices, use the Proteins Plus Full Report. First select the table "helix". Then select the items Helix_id, Beg_res_num, Beg_res_name, End_res_num, and End_res_name. To view the report, select "Display Report". This report can be printed or saved. Select "Reset Report" to begin a new report.

Structure_ID	Helix_id	Beg_res_num	Beg_res_name	End_res_num	End_res_name
PDB1ABS	A	3	SER	18	GLU
PDB1ABS	B	20	ASP	35	SER
PDB1ABS	C	36	HIS	42	LYS
PDB1ABS	D	51	THR	57	ALA
PDB1ABS	E	58	SER	77	LYS
PDB1ABS	F	86	LEU	95	THR
truncated					
PDB5MBN	G	100	PRO	118	ARG
PDB5MBN	H	124	GLY	149	LEU

Fig. 8. An example of a Proteins Plus query and report.

This work has been funded by the National Science Foundation (BIR 95 10703).

References

Allen, F. H., Bellard, S., Brice, M. D., Cartright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *Acta Cryst.* B**35**, 2331–2339.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bourne, P., Berman, H. M., Watenpaugh, K., Westbrook, J. D. & Fitzgerald, P. M. D. (1997). *Methods Enzymol.* **277**, 571–590.

Brünger, A. T. (1992). *X-PLOR*, *Version* 3.1, *A System for X-ray Crystallography and NMR*, Yale University Press, New Haven, CT, USA.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 509–518.

Clowney, L. & Westbrook, J. D. (1997). *à la mode: A Ligand and Monomer Object Data Environment*, NDB-241, Rutgers University, New Brunswick, NJ, USA.

Feng, Z., Hsieh, S.-H., Gelbin, A. & Westbrook, J. (1997). *MAXIT: Macromolecular Exchange and Input Tool*, NDB-220, Rutgers University, New Brunswick, NJ, USA.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 519–528.

Gilliland, G. L., Tung, M., Blakeslee, D. M. & Ladner, J. E. (1994). *Acta Cryst.* D**50**, 408–413.

Grzeskowiak, K., Yanagi, K., Privé, G. G. & Dickerson, R. E. (1991). *J. Biol. Chem.* **266**, 8861–8883.

Lavery, R. & Sklenar, H. (1988). *J. Biomol. Struct. Dyn.* **6**, 63–91.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *Acta Cryst.* D**52**, 57–64.

Sayle, R. & Milner-White, J. E. (1995). *Trends Biochem. Sci.* **20**, 374.

Schneider, B., Neidle, S. & Berman, H. M. (1997). *Biopolymers*, **42**, 113–124.

Scott, W. G., Finch, J. T. & Klug, A. (1995). *Cell*, **81**, 991–1002.

Watson, J. D. & Crick, F. H. C. (1953). *Nature (London)*, **171**, 737–738.