Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The Nucleic Acid Database

Helen M. Berman,* John Westbrook, Zukang Feng, Lisa Iype, Bohdan Schneider and Christine Zardecki

Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087, USA

Correspondence e-mail:
berman@rcsb.rutgers.edu

Received 15 November 2001
Accepted 21 February 2002

The Nucleic Acid Database was established in 1991 as a resource to assemble and distribute structural information about nucleic acids. Over the years, the NDB has developed generalized software for processing, archiving, querying and distributing structural data for nucleic acid-containing structures. The architecture and capabilities of the Nucleic Acid Database, as well as some of the research enabled by this resource, are presented in this article.

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. Over the years, the NDB has developed generalized software for processing, archiving, querying and distributing structural data for nucleic acid-containing structures. The core of the NDB has been its relational database of nucleic acid-containing crystal structures. Recognizing the importance of a standard data representation in building a database, the NDB became an active participant in the mmCIF project and was the test-bed for this format. With a foundation of well curated data, the NDB created a searchable relational database of primary and derivative data with very rich query and reporting capabilities. This robust database was unique in that it allowed researchers to perform comparative analyses of nucleic acid-containing structures selected from the NDB according to the many attributes stored in the database.

In 1992, the NDB assumed responsibility for processing all nucleic acid crystal structures that were deposited into the PDB; it became a direct deposition site for those structures in 1996. In order to meet data-processing requirements, the NDB created the first validation software for nucleic acids (Feng, Westbrook *et al.*, 1998). Until 1998, protein/nucleic acid crystal structures deposited into the PDB were post-processed and then incorporated into the NDB. When the Research Collaboratory for Structural Bioinformatics assumed the management of the PDB in 1998, the tools developed by the NDB were used to process all macromolecular structures (Berman *et al.*, 2000). The NDB continues to provide a high level of information about nucleic acids and serves as a specialty database for its community of researchers.

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

Table 1
The information content of the NDB.

- Primary experimental information stored in the NDB
  - Structure summary – descriptor; NDB, PDB and CSD names; coordinate availability; modifications, mismatches and drug binding
  - 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
- Derivative information stored in the NDB
  - Distances – chemical bond lengths; virtual bonds (involving P atoms)
  - Torsions – backbone and side-chain torsion angles; pseudorotational parameters
  - Angles – valence-bond angles, virtual angles (involving P atoms)
  - Base morphology – parameters calculated by different algorithms
  - Non-bonded contacts
  - Valence-geometry RMS deviations from small-molecule standards
  - Sequence-pattern statistics

(Table 1). 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 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.

Some features are derived by different algorithms and it can be difficult to provide the most reliable values. Whenever possible, the NDB has tried to promote standards that allow structure comparison. An outstanding example of this was the problem associated with different values for base-morphology parameters produced by different programs (Babcock & Olson, 1994; Babcock *et al.*, 1993, 1994; Bansal *et al.*, 1995; Bhattacharyya & Bansal, 1989; Dickerson, 1998; El Hassan & Calladine, 1995; Gorin *et al.*, 1995; Kosikov *et al.*, 1999; Lavery & Sklenar, 1988, 1989; Lu *et al.*, 1997; Soumpasis & Tung, 1988; Tung *et al.*, 1994). This meant that it was not possible to compare any two structures using the numbers in the published literature and that it was necessary to recalculate these values for any analysis.

To help resolve this problem, the NDB co-sponsored the Tsukuba Workshop on Nucleic Acid Structure and Interactions (12–14 January 1999, AIST–NIBHT Structural Biology Centre, Tsukuba, Japan) to which all the key software developers in this field were invited. It was resolved that a single reference frame would be used to calculate these values and an agreement was reached about the definition of that reference frame (Olson *et al.*, 2001). This work fully quantifies the proposal for base morphology made previously at a meeting in Cambridge (Dickerson, 1989). All the programs are being amended so that they will produce very similar values for the parameters. The NDB has recalculated these values for all the structures in the repository and will make them available as output from NDB searches performed over the WWW (see §4 for more information).

3. Data validation and processing

The NDB has created a robust data-processing system that produces high-quality data that is readily loaded into a database. The full capability of this system was recently demonstrated by the successful processing of ribosomal subunits, which are very large and complex structures.

Early on, the NDB adopted the Macromolecular Crystallographic Information File (mmCIF: Bourne *et al.*, 1997) as its data standard. This format has several advantages from the point of view of building a database: (i) the definitions for the data items are based on a comprehensive dictionary of crystallographic terminology and molecular structure description, (ii) it is self-defining and (iii) the syntax contains explicit rules that further define the characteristics of the data items, particularly the relationships between data items (Westbrook & Bourne, 2000). The latter feature is important because it allows rigorous checking of the data.

Structures are deposited *via* the WWW using the *AutoDep* Input Tool (*ADIT*; Westbrook *et al.*, 1998) and then annotated using the same tool. *ADIT* operates on top of the mmCIF dictionary. In the next stage of data processing, a program

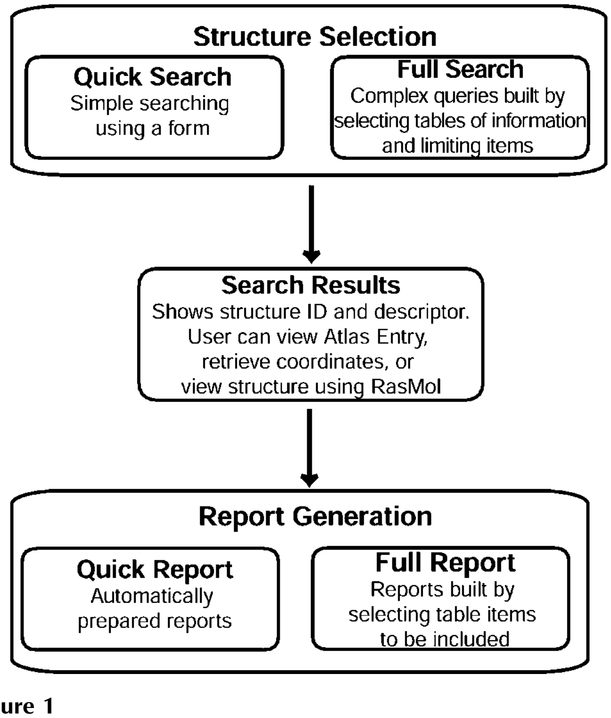

Structure Selection
Quick Search
Simple searching using a form
Full Search
Complex queries built by selecting tables of information and limiting items
Search Results
Shows structure ID and descriptor. User can view Atlas Entry, retrieve coordinates, or view structure using RasMol
Report Generation
Quick Report
Automatically prepared reports
Full Report
Reports built by selecting table items to be included

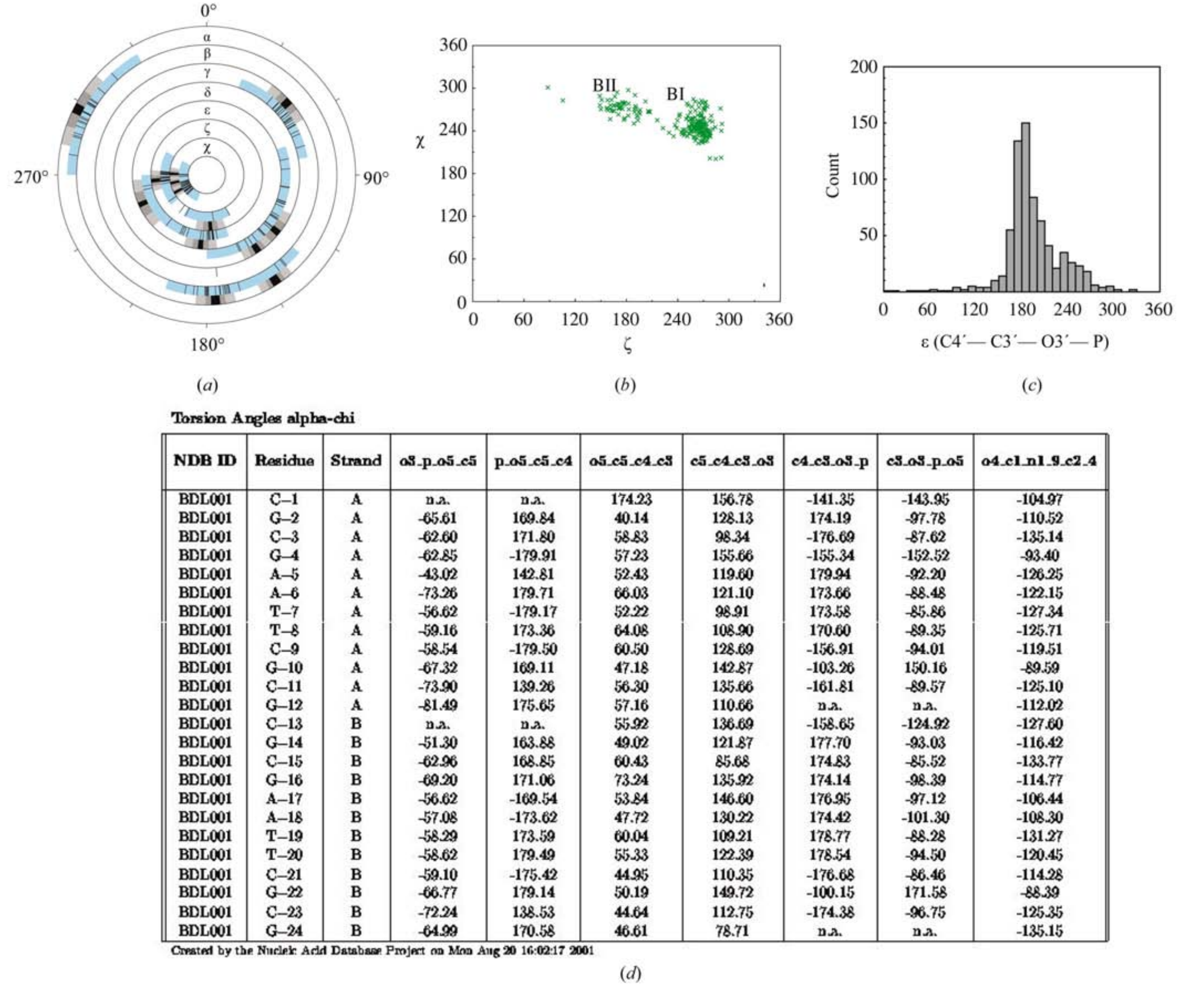
Figure 1
Flow chart demonstrating the two steps involved in searching the NDB: structure selection and report generation.

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 SEQRES PDB record and the residue names in the coordinate files. Once these integrity checks are completed, the structures are validated.

NUCheck (Feng, Westbrook *et al.*, 1998) verifies valence geometry, torsion angles, intermolecular contacts and the chiral centers 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; Allen, 2002). The torsion-angle ranges were derived from an analysis of well resolved nucleic acid structures (Schneider *et al.*, 1997). One important outgrowth of these 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, Vojtechovsky *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 released based upon its author-defined hold status.

4. The database and query capabilities

The core of the NDB project is a relational database in which all of the primary and derived data items are organized into



(*a*) (*b*) (*c*)

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
BDL001	C–1	A	n.a.	n.a.	174.23	156.78	-141.35	-143.95	-104.97
BDL001	G–2	A	-65.61	169.84	40.14	128.13	174.19	-97.78	-110.52
BDL001	C–3	A	-62.60	171.80	58.83	98.34	-176.69	-87.62	-135.14
BDL001	G–4	A	-62.85	-179.91	57.23	155.66	-155.34	-152.52	-93.40
BDL001	A–5	A	-43.02	142.81	52.43	119.60	179.94	-92.20	-126.25
BDL001	A–6	A	-73.26	179.71	66.03	121.10	173.66	-88.48	-122.15
BDL001	T–7	A	-56.62	-179.17	52.22	98.91	173.58	-85.86	-127.34
BDL001	T–8	A	-59.16	173.36	64.08	108.90	170.60	-89.35	-125.71
BDL001	C–9	A	-58.54	-179.50	60.50	128.69	-156.91	-94.01	-119.51
BDL001	G–10	A	-67.32	169.11	47.18	142.87	-103.26	150.16	-89.59
BDL001	C–11	A	-73.90	139.26	56.30	135.66	-161.81	-89.57	-125.10
BDL001	G–12	A	-81.49	175.65	57.16	110.66	n.a.	n.a.	-112.02
BDL001	C–13	B	n.a.	n.a.	55.92	136.69	-158.65	-124.92	-127.60
BDL001	G–14	B	-51.30	163.88	49.02	121.87	177.70	-93.03	-116.42
BDL001	C–15	B	-62.96	168.85	60.43	85.68	174.83	-85.52	-133.77
BDL001	G–16	B	-69.20	171.06	73.24	135.92	174.14	-98.39	-114.77
BDL001	A–17	B	-56.62	-169.54	53.84	146.60	176.95	-97.12	-106.44
BDL001	A–18	B	-57.08	-173.62	47.72	130.22	174.42	-101.30	-108.30
BDL001	T–19	B	-58.29	173.59	60.04	109.21	178.77	-88.28	-131.27
BDL001	T–20	B	-58.62	179.49	55.33	122.39	178.54	-94.50	-120.45
BDL001	C–21	B	-59.10	-175.42	44.95	110.35	-176.68	-86.46	-114.28
BDL001	G–22	B	-66.77	179.14	50.19	149.72	-100.15	171.58	-88.39
BDL001	C–23	B	-72.24	138.53	44.64	112.75	-174.38	-96.75	-125.35
BDL001	G–24	B	-64.99	170.58	46.61	78.71	n.a.	n.a.	-135.15

Created by the Nucleic Acid Database Project on Mon Aug 20 16:02:17 2001

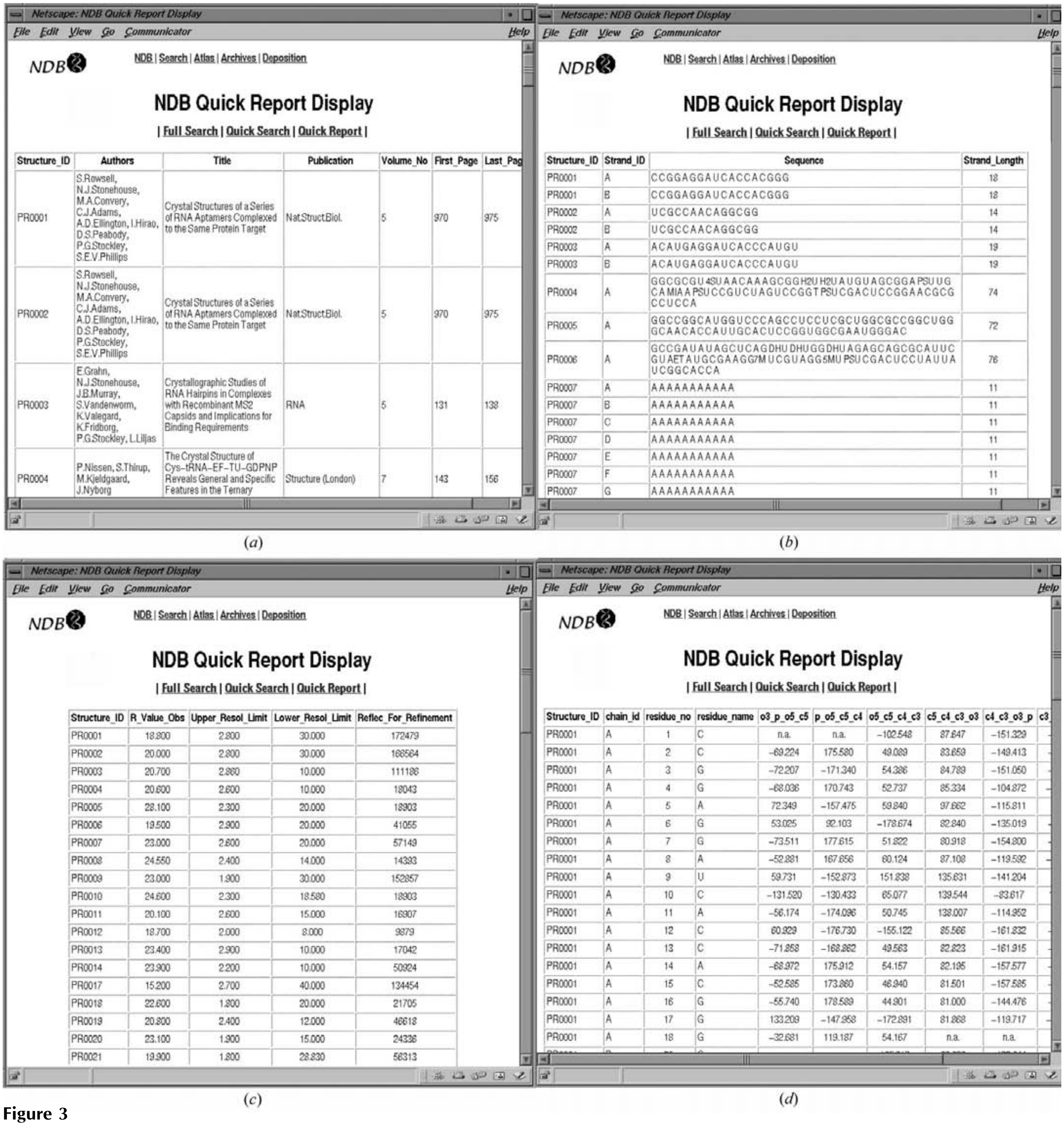
(*d*)

Figure 2

Examples of torsion-angle reports generated from the NDB. (*a*) Conformation wheel showing the torsion angles for BDL001 (Drew *et al.*, 1981). Black lines show actual values of torsion angles and cyan background their allowed range in the B-type DNA conformation (Schneider *et al.*, 1997). The gray shades in the outer rings show the average value(s) of the torsions in dark grey flanked by values of one and two estimated standard deviations in lighter grey. (*b*) Scattergram graph showing the relationship of χ *versus* ζ for all B-DNA. Two clusters, BI and BII, are labeled. (*c*) Histogram for ε (C4′—C3′—O3′—P) for all B-DNA. (*d*) A torsion-angle report for BDL001.

tables. At present, there are over 90 tables in the NDB, with each table containing 5–20 data 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. 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 or equal to 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 combina-


Netscape: NDB Quick Report Display
File Edit View Go Communicator Help
NDB
NDB | Search | Atlas | Archives | Deposition
NDB Quick Report Display
| Full Search | Quick Search | Quick Report |

Structure_ID	Authors	Title	Publication	Volume_No	First_Page	Last_Pag
PR0001	S.Rowsell, N.J.Stonehouse, M.A.Convery, C.J.Adams, A.D.Ellington, I.Hirao, D.S.Peabody, P.G.Stockley, S.E.V.Phillips	Crystal Structures of a Series of RNA Aptamers Complexed to the Same Protein Target	Nat.Struct.Biol.	5	970	975
PR0002	S.Rowsell, N.J.Stonehouse, M.A.Convery, C.J.Adams, A.D.Ellington, I.Hirao, D.S.Peabody, P.G.Stockley, S.E.V.Phillips	Crystal Structures of a Series of RNA Aptamers Complexed to the Same Protein Target	Nat.Struct.Biol.	5	970	975
PR0003	E.Grahn, N.J.Stonehouse, J.B.Murray, S.Vandenworm, K.Valegard, K.Fridborg, P.G.Stockley, L.Liljas	Crystallographic Studies of RNA Hairpins in Complexes with Recombinant MS2 Capsids and Implications for Binding Requirements	RNA	5	131	138
PR0004	P.Nissen, S.Thirup, M.Kjeldgaard, J.Nyborg	The Crystal Structure of Cys-tRNA-EF-TU-GDPNP Reveals General and Specific Features in the Ternary	Structure (London)	7	143	156

(*a*)

Netscape: NDB Quick Report Display
File Edit View Go Communicator Help
NDB
NDB | Search | Atlas | Archives | Deposition
NDB Quick Report Display
| Full Search | Quick Search | Quick Report |

Structure_ID	Strand_ID	Sequence	Strand_Length
PR0001	A	CCGGAGGAUCACCACGGG	18
PR0001	B	CCGGAGGAUCACCACGGG	18
PR0002	A	UCGCCAACAGGCGG	14
PR0002	B	UCGCCAACAGGCGG	14
PR0003	A	ACAUGAGGAUCACCCAUGU	19
PR0003	B	ACAUGAGGAUCACCCAUGU	19
PR0004	A	G G C G C G U 4SU A A C A A A G C G G H2U H2U A U G U A G C G G A PSU U G C A MIA A PSU C C G U C U A G U C C G G T PSU C G A C U C C G G A A C G C G C C U C C A	74
PR0005	A	GGCCGGCAUGGUCCCAGCCUCCUCGCUGGCGCCGGCUGG GCAACACCAUUGCACUCCGGUGGCGAAUGGGAC	72
PR0006	A	G C C G A U A U A G C U C A G DHU DHU G G DHU A G A G C A G C G C A U U C G U AET A U G C G A A G G 7M U C G U A G G 5MU PSU C G A C U C C U A U U A U C G G C A C C A	76
PR0007	A	AAAAAAAAAAA	11
PR0007	B	AAAAAAAAAAA	11
PR0007	C	AAAAAAAAAAA	11
PR0007	D	AAAAAAAAAAA	11
PR0007	E	AAAAAAAAAAA	11
PR0007	F	AAAAAAAAAAA	11
PR0007	G	AAAAAAAAAAA	11

(*b*)

Netscape: NDB Quick Report Display
File Edit View Go Communicator Help
NDB
NDB | Search | Atlas | Archives | Deposition
NDB Quick Report Display
| Full Search | Quick Search | Quick Report |

Structure_ID	R_Value_Obs	Upper_Resol_Limit	Lower_Resol_Limit	Reflec_For_Refinement
PR0001	18.800	2.800	30.000	172479
PR0002	20.000	2.800	30.000	168564
PR0003	20.700	2.860	10.000	111186
PR0004	20.600	2.600	10.000	18043
PR0005	28.100	2.300	20.000	18903
PR0006	19.500	2.900	20.000	41055
PR0007	23.000	2.600	20.000	57149
PR0008	24.550	2.400	14.000	14393
PR0009	23.000	1.900	30.000	152857
PR0010	24.600	2.300	18.580	18903
PR0011	20.100	2.600	15.000	16907
PR0012	18.700	2.000	8.000	9879
PR0013	23.400	2.900	10.000	17042
PR0014	23.900	2.200	10.000	50924
PR0017	15.200	2.700	40.000	134454
PR0018	22.600	1.800	20.000	21705
PR0019	20.800	2.400	12.000	46618
PR0020	23.100	1.900	15.000	24336
PR0021	19.900	1.800	28.830	56313

(*c*)

Netscape: NDB Quick Report Display
File Edit View Go Communicator Help
NDB
NDB | Search | Atlas | Archives | Deposition
NDB Quick Report Display
| Full Search | Quick Search | Quick Report |

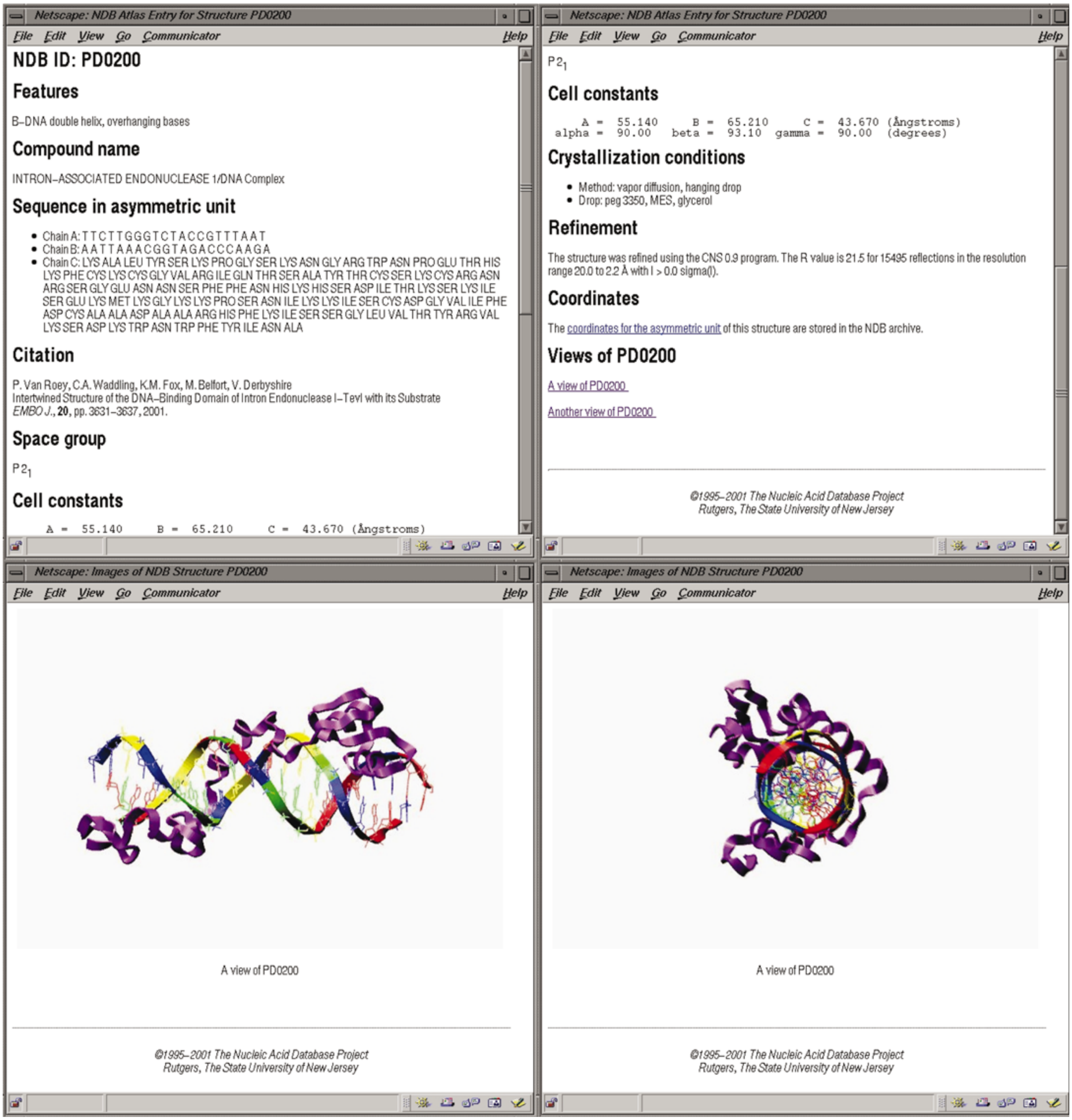
Structure_ID	chain_id	residue_no	residue_name	o3_p_o5_c5	p_o5_c5_c4	o5_c5_c4_c3	c5_c4_c3_o3	c4_c3_o3_p	c3_
PR0001	A	1	C	n.a.	n.a.	−102.548	87.647	−151.329	
PR0001	A	2	C	−69.224	175.580	49.089	83.659	−149.413	
PR0001	A	3	G	−72.207	−171.340	54.386	84.789	−151.050	
PR0001	A	4	G	−68.036	170.743	52.737	85.334	−104.872	
PR0001	A	5	A	72.349	−157.475	59.840	97.662	−115.811	
PR0001	A	6	G	53.025	92.103	−178.674	82.840	−135.019	
PR0001	A	7	G	−73.511	177.615	51.822	80.918	−154.800	
PR0001	A	8	A	−52.881	167.656	60.124	87.108	−119.592	
PR0001	A	9	U	59.731	−152.873	151.838	135.631	−141.204	
PR0001	A	10	C	−131.520	−130.433	65.077	139.544	−83.617	
PR0001	A	11	A	−56.174	−174.096	50.745	138.007	−114.952	
PR0001	A	12	C	60.929	−176.730	−155.122	85.566	−161.832	
PR0001	A	13	C	−71.858	−168.862	49.563	82.823	−161.915	
PR0001	A	14	A	−68.972	175.912	54.157	82.195	−157.577	
PR0001	A	15	C	−52.585	173.860	46.940	81.501	−157.585	
PR0001	A	16	G	−55.740	178.589	44.901	81.000	−144.476	
PR0001	A	17	G	133.209	−147.958	−172.891	81.868	−119.717	
PR0001	A	18	G	−32.681	119.187	54.167	n.a.	n.a.	

(*d*)

Figure 3
Examples of Quick Reports: (*a*) citation report for protein–RNA structures; (*b*) nucleic acid sequence report for protein–RNA structures; (*c*) refinement information for protein–RNA structures; (*d*) nucleic acid backbone torsions report for PR0001 (Rowsell *et al.*, 1998).

tion 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 the types of reports produced by the NDB are given in Fig. 2.

A WWW interface was designed to make the query capabilities of the NDB as widely accessible as possible. In the Quick Search/Quick Report mode, several items, including structure ID, author, classification and special features, can be


Netscape: NDB Atlas Entry for Structure PD0200
File Edit View Go Communicator Help

NDB ID: PD0200

Features

B–DNA double helix, overhanging bases

Compound name

INTRON–ASSOCIATED ENDONUCLEASE 1/DNA Complex

Sequence in asymmetric unit

- Chain A: TTCTTGGGTCTACCGTTTAAT
- Chain B: AATTAAACGGTAGACCCAAGA
- Chain C: LYS ALA LEU TYR SER LYS PRO GLY SER LYS ASN GLY ARG TRP ASN PRO GLU THR HIS LYS PHE CYS LYS CYS GLY VAL ARG ILE GLN THR SER ALA TYR THR CYS SER LYS CYS ARG ASN ARG SER GLY GLU ASN ASN SER PHE PHE ASN HIS LYS HIS SER ASP ILE THR LYS SER LYS ILE SER GLU LYS MET LYS GLY LYS LYS PRO SER ASN ILE LYS LYS ILE SER CYS ASP GLY VAL ILE PHE ASP CYS ALA ALA ASP ALA ALA ARG HIS PHE LYS ILE SER SER GLY LEU VAL THR TYR ARG VAL LYS SER ASP LYS TRP ASN TRP PHE TYR ILE ASN ALA

Citation

P. Van Roey, C.A. Waddling, K.M. Fox, M. Belfort, V. Derbyshire
Intertwined Structure of the DNA–Binding Domain of Intron Endonuclease I–TevI with its Substrate
EMBO J., **20**, pp. 3631–3637, 2001.

Space group

$P2_1$

Cell constants

```
     A =  55.140     B =  65.210     C =  43.670 (Ångstroms)
```

Netscape: NDB Atlas Entry for Structure PD0200
File Edit View Go Communicator Help

$P2_1$

Cell constants

```
     A =  55.140     B =  65.210     C =  43.670 (Ångstroms)
 alpha =  90.00   beta =  93.10  gamma =  90.00  (degrees)
```

Crystallization conditions

- Method: vapor diffusion, hanging drop
- Drop: peg 3350, MES, glycerol

Refinement

The structure was refined using the CNS 0.9 program. The R value is 21.5 for 15495 reflections in the resolution range 20.0 to 2.2 Å with I > 0.0 sigma(I).

Coordinates

The coordinates for the asymmetric unit of this structure are stored in the NDB archive.

Views of PD0200

A view of PD0200

Another view of PD0200

©1995–2001 The Nucleic Acid Database Project
Rutgers, The State University of New Jersey

Netscape: Images of NDB Structure PD0200
File Edit View Go Communicator Help

A view of PD0200

©1995–2001 The Nucleic Acid Database Project
Rutgers, The State University of New Jersey

Netscape: Images of NDB Structure PD0200
File Edit View Go Communicator Help

A view of PD0200

©1995–2001 The Nucleic Acid Database Project
Rutgers, The State University of New Jersey

Figure 4
NDB Atlas page for PD0200 (van Roey *et al.*, 2001) which highlights the structure's features, compound name, sequence, citation, space group, cell constants, crystallization conditions and refinement information. The Atlas page also links to the coordinates and to images (shown) of the entry.

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.

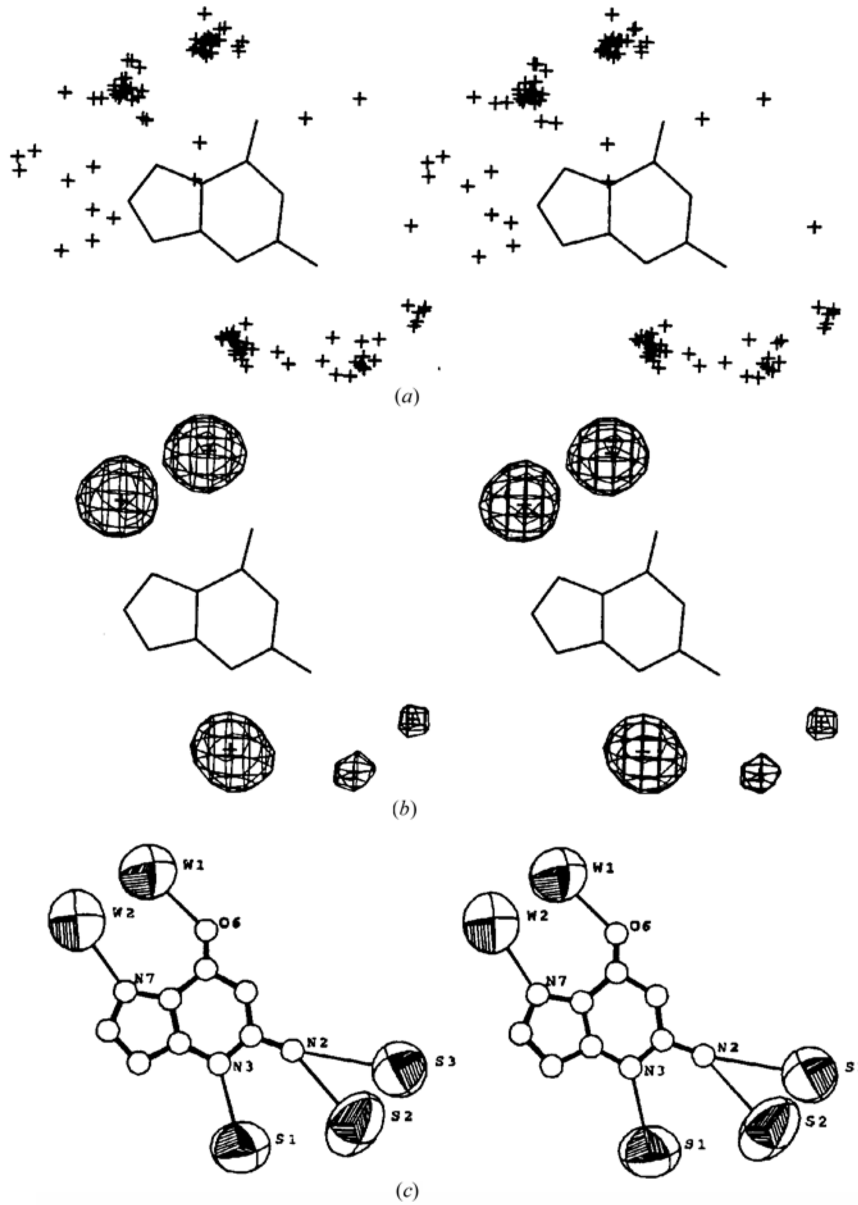


Figure 5
Water environment of guanine residues in structures in the NDB. (*a*) Scattergram of 101 water molecules within 3.4 Å of any atom of 42 guanines found in 14 B-DNA decamer structures. (*b*) Electron densities of the 101 water molecules plotted at the 4σ level. Each water is modeled as an O atom with an occupancy of 1/42; (*c*) An *ORTEP* (Johnson, 1976) plot of the current guanine B-DNA hydration sites after refinement. Plotted are 50% probability thermal ellipsoids. The key guanine atoms and hydration sites are labeled. All plots are in stereo. These figures are reprinted from Schneider & Berman (1995) with permission from the Biophysical Society.

Pre-formatted Quick Reports can then be generated for the structures in this result list. The user selects a report from a list of 13 report options (Table 2) and the report is created automatically. Multiple reports can be easily generated. These options are particularly convenient for quickly producing reports based on derived features, such as torsion angles and base morphology (Fig. 3).

In the Full Search/Full Report mode, it is possible to access most of the tables in the NDB to build more complex queries. Rather than being limited to items that are listed on a single page, 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.

After selecting structures using the Full Search, a variety of reports can be written. The report columns are selected from a variety of database tables and the full report is automatically generated. Multiple reports can be generated for the same group of selected structures; for example, reports on crystallization, base modification or a combination of these reports can be generated for a particular group of structures.

5. Data distribution

Coordinate files, database reports, software programs and other resources are available *via* the ftp server (ftp://ndbserver.rutgers.edu). In addition to links to information provided from the ftp server, the WWW server (http://ndbserver.rutgers.edu/) provides a variety of methods for querying the NDB (described above). These sites are updated continually.

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 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.

A very popular and useful report is the NDB Atlas report page (Fig. 4). An atlas

Table 2
Quick Reports available for the NDB.

Report Name	Contains
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
NA Backbone Torsions (NDB)	Sugar–phosphate backbone torsion angles using NDB residue numbers
NA Backbone Torsions (PDB)	Sugar–phosphate backbone torsion angles using PDB residue numbers
Base Pair Parameters	Global base-pair parameters calculated using *Standard Reference Frame* (Olson *et al.*, 2001)
Base Pair Step Parameters	Local base-pair step parameters calculated using *Standard Reference Frame*
Groove Dimensions	Groove dimensions using Stoffer and Lavery definitions from *CURVES*5.1

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. The entries for all structures in the database are organized by structure type in the NDB Atlas.

5.1. Mirrors

The NDB is based at Rutgers University (http://ndbserver.rutgers.edu/) and is currently mirrored at three other sites: the Institute of Cancer Research, UK (http://www.ndb.icr.ac.uk/), the San Diego Supercomputer Center, San Diego, California (http://ndb.sdsc.edu/NDB/) and the Structural Biology Centre, Tsukuba, Japan (http://ndbserver.nibh.go.jp/NDB/). These mirror sites are updated daily, are fully synchronous and contain the ftp directories, the web site and the full database.

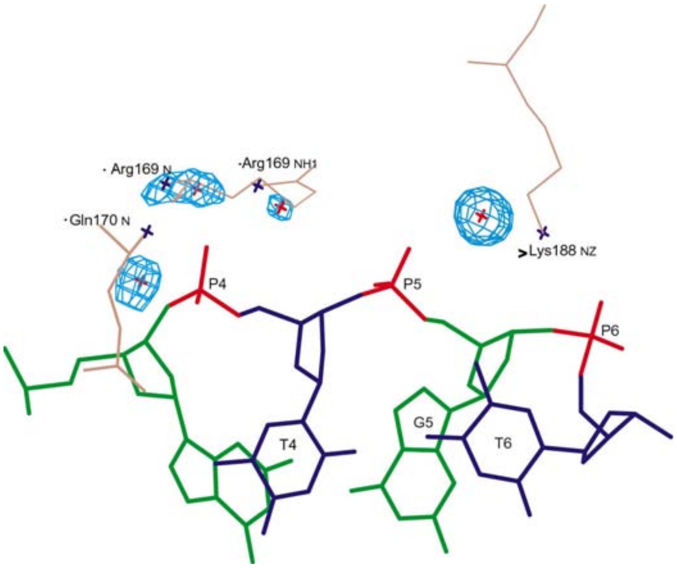


Figure 6
A view of the three residues in the consensus region for the high-resolution CAP–DNA$_{GCE}$ complex (Parkinson, Wilson *et al.*, 1996). The predicted phosphate hydration is drawn as pseudoelectron density in cyan, the interacting protein residues are shown in dark brown and the phosphate groups are red. The protein atoms that contact the DNA shown as blue crosses. The predicted sites are the red crosses. Reprinted from Woda *et al.* (1998) with permission from the Biophysical Society.

5.2. Community outreach

The NDB works closely with the research community to ensure that their needs are met. A newsletter is published electronically and provides information about the latest features of the system. To subscribe, send an e-mail to ndbnews@ndbserver.rutgers.edu. Very complex queries will be carried out by the staff in response to user requests *via* e-mail to ndbadmin@ndbserver.rutgers.edu.

6. Applications of the NDB

The NDB has been used to analyze characteristics of nucleic acids alone and complexed with proteins. The ability to select structures according to many criteria has made it possible to create appropriate data sets for study. A few examples are given here.

The conformational characteristics of A-, B- and Z-DNA were examined (Schneider *et al.*, 1997) using carefully selected examples of well resolved structures in these classes. Conformation wheels (Fig. 2*a*) for each conformation as well as scattergrams of selected torsion angles (Fig. 2*b*) were created. These diagrams can now be used to assess and classify new structures. Studies of B-DNA helices have shown that the base steps have characteristic values that depend on their sequence (Gorin *et al.*, 1995). Plots of twist *versus* roll are different for purine–purine, purine–pyrimidine and pyrimidine–purine

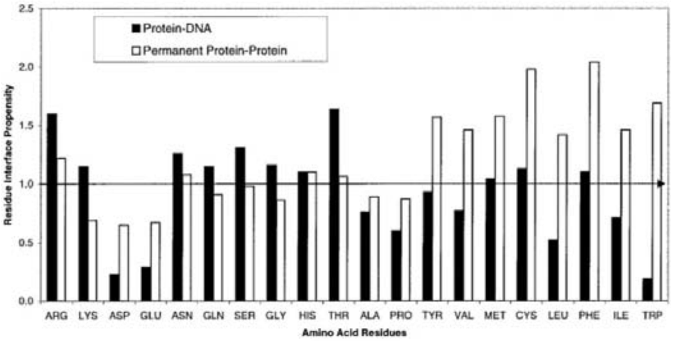


Figure 7
Histogram of the interface residue propensities calculated for 26 protein–DNA complexes and compared with those for permanent protein–protein complexes (Jones & Thornton, 1996). ‘Permanent’ complexes are those in which the components only exist as part of a complex; they do not exist in isolation. Generally, they have larger interfaces that are more hydrophobic and more complementary. A propensity of >1 indicates that a residue occurs more frequently in the interface than on the protein surface. The amino-acid residues have been ordered using the Faucher & Pliska (1983) hydrophobicity scale, with the most hydrophilic residues on the left-hand side and the most hydrophobic on the right-hand side of the graph. Reprinted with permission from Jones *et al.* (1999).

steps. This particular analysis has been extended to derive energy parameters for B-DNA sequences (Olson *et al.*, 1998).

In a series of systematic studies of the hydration patterns of DNA double helices, it was found that the hydration patterns around the bases are well defined and are local (Fig. 5; Schneider & Berman, 1995). That is, small changes in the conformation of the backbone do not affect the hydration around the bases. It was also found that there are more diffuse patterns around the phosphate backbone that are dependent on the conformational class of the DNA. These analyses were used to attempt to predict the binding sites of protein side chains on the DNA. In a series of protein–nucleic acid complexes, the hydration sites of the DNA were calculated and then compared with the location of the amino-acid side chain. The results were surprisingly good, in that in most cases the side-chain site and hydration sites were very close. This was true even in the case of a very bent DNA that is bound to catabolite activator protein (Fig. 6; Woda *et al.*, 1998).

Systematic studies of the interface in protein–nucleic acid complexes have been performed. In one analysis of protein–DNA complexes, 26 complexes were selected in which the proteins were non-homologous (Jones *et al.*, 1999). The results showed that there are amino-acid propensities at the interface that are markedly different to those in protein–protein complexes. It was also possible to place the complexes into three classes: double-headed, single-headed and enveloping (Figs. 7 and 8). A similar analysis has also been performed for protein–RNA complexes (Jones *et al.*, 2001). There have also been detailed analyses of the hydrogen-bonding patterns at the protein–DNA interface and it was found that CH···O bonds are surprisingly common (Mandel-Gutfreund *et al.*, 1998).

Some analyses have been performed on the relationship between crystal packing and conformation. Although there are more than 30 different crystals forms of B-DNA in the NDB, the actual number of packing motifs (Fig. 9) remains relatively small, with the most common motifs being minor groove–minor groove, stacking–lateral backbone and major groove–backbone (Timsit & Moras, 1992).

Minor groove–minor groove interactions in which the guanine of one duplex forms hydrogen bonds with the guanines of a neighboring duplex are seen not only in dodecamer structures, but also in an octamer sequence with three duplexes in the asymmetric unit (Urpi *et al.*, 1996). The second motif contains duplexes stacked above one another, with the adjoining phosphates forming lateral interactions. A large number of variations of this motif have been observed in decamer (Grzeskowiak *et al.*, 1991) and hexamer structures (Cruse *et al.*, 1986; Tari & Secco, 1995). The third type of packing involves the major groove of one helix interacting with the phosphate backbone of another (Timsit *et al.*, 1989). Sequence appears to be a large factor in determining these motifs, but it is not the only factor. For example, the first structures exhibiting the major groove–phosphate interactions contained a cytosine that formed a hydrogen bond to the phosphate. However, not all structures that show this motif have this hydrogen bond (Wood *et al.*, 1997). The particular sequence in this crystal is even more intriguing because it also crystallizes in another crystal form in which the terminal flips out to form a minor-groove interaction with another duplex (Spink *et al.*, 1995).

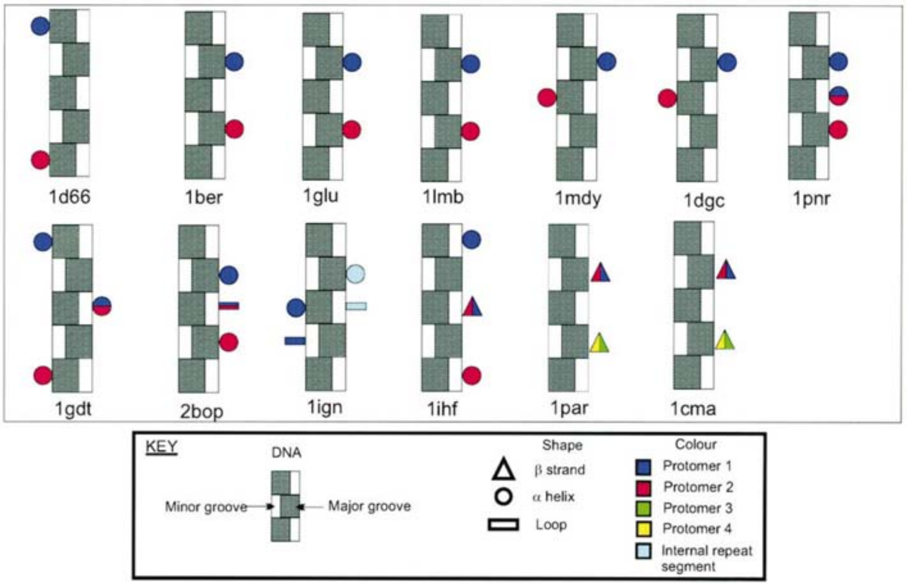


Figure 8
Simple model diagrams of protein–DNA complexes for double-headed binding proteins. The diagrams give an indication of the predominant secondary structure of the binding motif, protein symmetry and the type and relative position of the DNA groove bound. The secondary structure of the predominant binding motifs are indicated using different symbols analogous to those used in *TOPS* diagrams (Westhead *et al.*, 1998). Only one symbol of each type is indicated in any one groove, hence both a single sheet and two sheets are indicated by a single colored triangle. The symmetry of each protein is indicated by using a different color for each symmetry (or pseudosymmetry) related element. A single symbol shaded in two colors indicates that there are secondary structures of this type contributed by more than one symmetry-related element. Reprinted with permission from Jones *et al.* (1999).

The task of trying to determine the relative effects of base sequence and crystal packing on the values of the base-morphology parameters is hampered to some degree by the uneven distribution of the 16 different base steps among the different crystal types. Some steps like CG are very well represented in B structures, whereas others such as AC have very few representatives in the data set. Nonetheless, there are a few steps that occur in crystals with different packing motifs. An analysis of the CG steps across all crystal types show that its conformation is relatively insensitive to crystal packing and the distribution is similar to that found for all steps (see Berman *et al.*, 1996). On the other hand, the conformational variability of the CA step appears to depend not simply on crystal type but on the packing motif.

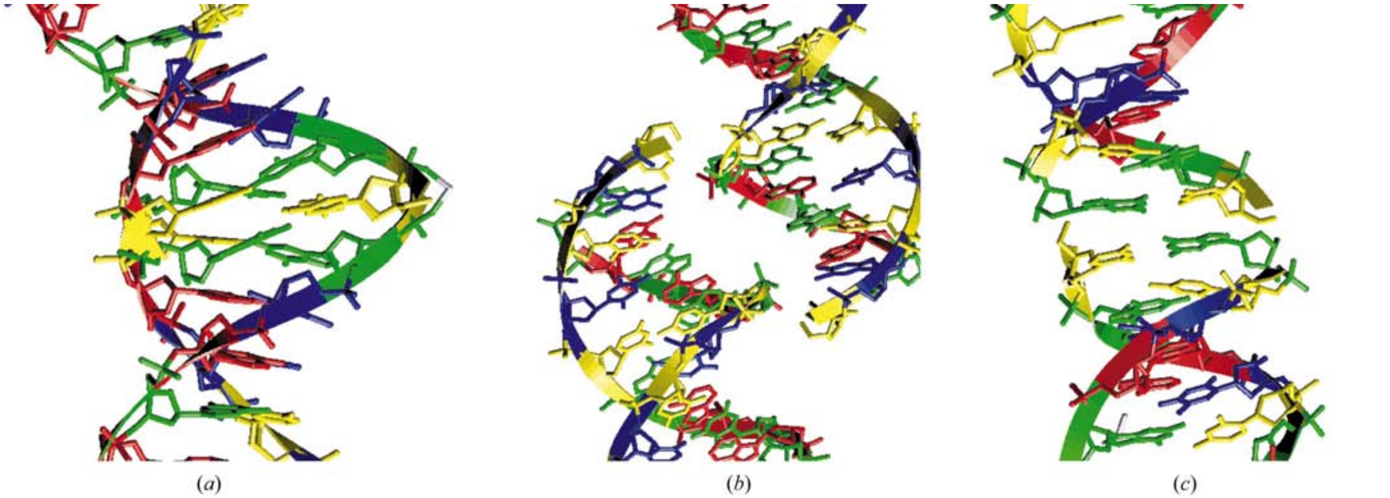


Figure 9
Examples of packing motifs in DNA duplexes in B- and A-DNA. From left to right: (*a*) minor groove–minor groove interactions in BDL042 (Leonard & Hunter, 1993); (*b*) major groove–backbone interactions in BDJ060 (Goodsell *et al.*, 1995); (*c*) stacking interactions in BDJ025 (Grzeskowiak *et al.*, 1991). The bases are colored green for guanine, yellow for cytosine, red for adenine and blue for thymine. Reprinted from Berman *et al.* (1996), copyright (1996), with permission from Elsevier Science.

The values of twist for CA steps in minor groove–minor groove motifs are smaller than those in the major groove–backbone motif. Very high values are displayed for CA steps in the stacking–lateral backbone motif. Plots of twist *versus* roll for CG steps show the distribution noted by others (Gorin *et al.*, 1995) and no clustering that depends on crystal type. On the other hand, the same plot for CA steps shows very distinctive differences that appear to depend on the packing motifs. It is important to note here that these motifs encompass several crystal types so that the structural variability observed is a function of a particular type of structural interaction rather than a particular crystal form. Before any definitive statements can be made about all the steps it will be necessary to have much more data.

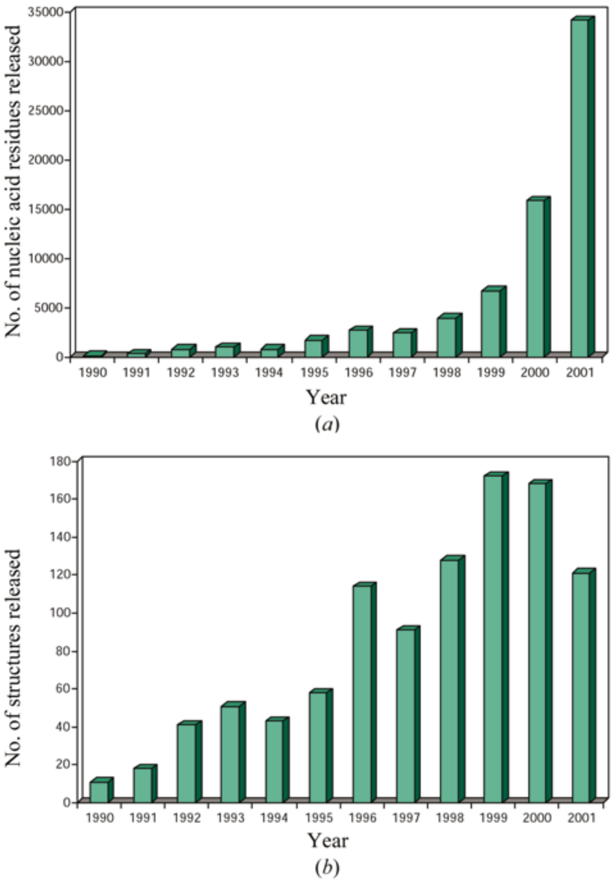


Figure 10
(*a*) The number of nucleic acid residues and (*b*) the number of structures released in the NDB as of 27 September 2001.

With the recent increase in the number of RNA structures available there have been attempts to establish systems whereby it will be possible to systematically analyze these structures. The result of one of these studies has been the proposal of a classification scheme for the hydrogen bonds in the base pairs (Westhof & Fritsch, 2000). A new syntax (RNAML) has also been proposed for representing RNA structural features (http://www.smi.stanford.edu/projects/helix/rnaml/).

7. Changing face of the NDB

When the NDB began, the world of nucleic acid structures consisted of DNA and RNA oligonucleotides, a few protein–DNA complexes and some tRNA structures. Annotation of structural features was performed manually by visual inspection of molecular architectures. However, in the last ten years a whole new universe of nucleic acid structures has emerged (Fig. 10). There are many ribozyme structures and many different types of protein–nucleic acid complexes represented in over 500 structures. The newest additions to the archive – ribosomal structures – have increased the number of residues of RNA resident in the NDB several-fold (Moore, 2001).

One outcome of the systematic studies that have been carried out with data from the NDB has been improved classification schemes for understanding nucleic acids. These will be used to annotate structures contained within the NDB automatically, which will in turn improve the query capability of the NDB. This type of cycle shows the power of organizing information so that it is more accessible and can ultimately yield new knowledge.

The NDB Project is funded by the National Science Foundation and the Department of Energy. We would like to thank Wilma K. Olson and David Beveridge, co-founders of the NDB, for their collaborations on this project. This article is also appearing in the forthcoming book tentatively titled *Structural Bioinformatics* to be published by John Wiley & Sons, Inc. (P. E. Bourne & H. Weissig, Editors).

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

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.

Babcock, M. S. & Olson, W. K. (1994). *J. Mol. Biol.* **237**, 98–124.

Babcock, M. S., Pednault, E. P. D. & Olson, W. K. (1993). *J. Biomol. Struct. Dyn.* **11**, 597–628.

Babcock, M. S., Pednault, E. P. D. & Olson, W. K. (1994). *J. Mol. Biol.* **237**, 125–156.

Bansal, M., Bhattacharyya, D. & Ravi, B. (1995). *CABIOS*, **11**, 281–287.

Berman, H. M., Gelbin, A. & Westbrook, J. (1996). *Prog. Biophys. Mol. Biol.* **66**, 255–288.

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.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bhattacharyya, D. & Bansal, M. (1989). *J. Biomol. Struct. Dyn.* **6**, 635–653.

Bourne, P. E., 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, Connecticut, 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.

Cruse, W. B. T., Salisbury, S. A., Brown, T., Cosstick, R., Eckstein, F. & Kennard, O. (1986). *J. Mol. Biol.* **192**, 891–905.

Dickerson, R. E. (1989). *J. Mol. Biol.* **205** 787–791.

Dickerson, R. E. (1998). *Nucleic Acids Res.* **26**, 1906–1926.

Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K. & Dickerson, R. E. (1981). *Proc. Natl Acad. Sci. USA*, **78**, 2179–2183.

El Hassan, M. A. & Calladine, C. R. (1995). *J. Mol. Biol.* **251**, 648–664.

Faucher, J. & Pliska, V. (1983). *Eur. J. Med. Chem.* **18**, 369–375.

Feng, Z., Hsieh, S.-H., Gelbin, A. & Westbrook, J. (1998). NDB-120 *MAXIT: Macromolecular Exchange and Input Tool.* Rutgers University, New Brunswick, New Jersey, USA.

Feng, Z., Westbrook, J. & Berman, H. M. (1998). NDB-407 *NUCheck*. Rutgers University, New Brunswick, New Jersey, USA.

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

Goodsell, D. S., Grzeskowiak, K. & Dickerson, R. E. (1995). *Biochemistry*, **34**, 1022–1029.

Gorin, A. A., Zhurkin, V. B. & Olson, W. K. (1995). *J. Mol. Biol.* **247**, 34–48.

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

Johnson, C. K. (1976). *ORTEP*II. Report ORNL-5138. Oak Ridge National Laboratory, Tennessee, USA.

Jones, S., Daley, D. T. A., Luscombe, N. M., Berman, H. M. & Thornton, J. M. (2001). *Nucleic Acids Res.* **29**, 934–954.

Jones, S. & Thornton, J. M. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). *J. Mol. Biol.* **287**, 877–896.

Kosikov, K. M., Gorin, A. A., Zhurkin, V. B. & Olson, W. K. (1999). *J. Mol. Biol.* **289**, 1301–1326.

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

Lavery, R. & Sklenar, H. (1989). *J. Biomol. Struct. Dyn.* **6**, 655–667.

Leonard, G. A. & Hunter, W. N. (1993). *J. Mol. Biol.* **234**, 198–208.

Lu, X.-J., El Hassan, M. A. & Hunter, C. A. (1997). *J. Mol. Biol.* **273**, 668–680.

Mandel-Gutfreund, Y., Margalit, H., Jernigan, R. & Zhurkin, V. (1998). *J. Mol. Biol.* **277**, 1129–1140.

Moore, P. (2001). *Biochemistry*, **40**, 3243–3250.

Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, S. C., Heinemann, U., Lu, X.-J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C.-S., Westhof, E., Wolberger, C. & Berman, H. M. (2001). *J. Mol. Biol.* **313**, 229–237.

Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M. & Zhurkin, V. B. (1998). *Proc. Natl. Acad. Sci. USA*, **95**, 11163–11168.

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

Parkinson, G., Wilson, C., Gunasekera, A., Ebright, Y. W., Ebright, R. E. & Berman, H. M. (1996). *J. Mol. Biol.* **260**, 395–408.

Roey, P. van, Waddling, C. A., Fox, K. M., Belfort, M. & Derbyshire, V. (2001). *EMBO J.* **20**, 3631–3637.

Rowsell, S., Stonehouse, N. J., Convery, M. A., Adams, C. J., Ellington, A. D., Hirao, I., Peabody, D. S., Stockley, P. G. & Phillips, S. E. V. (1998). *Nature Struct. Biol.* **5**, 970–975.

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

Schneider, B. & Berman, H. M. (1995). *Biophys. J.* **69**, 2661–2669.

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

Soumpasis, D. M. & Tung, C. S. (1988). *J. Biomol. Struct. Dyn.* **6**, 397–420.

Spink, N., Nunn, C., Vojetchovsky, J., Berman, H. & Neidle, S. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 10767–10771.

Tari, L. W. & Secco, A. S. (1995). *Nucleic Acids Res.* **23**, 2065–2073.

Timsit, Y. & Moras, D. (1992). *Methods Enzymol.* **211**, 409–429.

Timsit, Y., Westhof, E., Fuchs, R. P. P. & Moras, D. (1989). *Nature (London)*, **341**, 459–462.

Tung, C.-S., Soumpasis, D. M. & Hummer, G. (1994). *J. Biomol. Struct. Dyn.* **11**, 1327–1344.

Urpi, L., Tereshko, V., Malinina, L., Huynh-Dinh, T. & Subirana, J. A. (1996). *Nature Struct. Biol.* **3**, 325–328.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.

Westbrook, J. & Bourne, P. E. (2000). *Bioinformatics*, **16**, 159–168.

Westbrook, J., Feng, Z. & Berman, H. M. (1998). RCSB-99 *ADIT – The AutoDep Input Tool.* Department of Chemistry, Rutgers, the State University of New Jersey, USA.

Westhead, D. R., Hatton, D. C. & Thornton, J. M. (1998). *Trends Biochem. Sci.* **23**, 35–36.

Westhof, E. & Fritsch, V. (2000). *Structure*, **8**, R55–R65.

Woda, J., Schneider, B., Patel, K., Mistry, K. & Berman, H. M. (1998). *Biophys. J.* **75**, 2170–2177.

Wood, A. A., Nunn, C. M., Trent, J. O. & Neidle, S. (1997). *J. Mol. Biol.* **269**, 827–841.