Table 24.5.2.2. *Demographics of the released data in the PDB as of 14 September 1999*

Experimental technique	Molecule type				
	Proteins, peptides, and viruses	Protein–nucleic acid complexes	Nucleic acids	Carbohydrates and other	Total
X-ray diffraction and other	7946	390	439	14	8789
NMR	1365	53	270	4	1692
Theoretical modelling	202	16	15	0	233
Total	9513	459	724	18	10714

contacts, and positions of water molecules. In consultation with the community, other structure checks will be implemented over the next few years.

The experimental data are also checked. Currently, X-ray crystallographic data are validated and plans for checking NMR data are in progress. For X-ray crystallographic structures, the structure factors are validated using *SFCHECK* (Vaguine *et al.*, 1999). This program extracts the deposited *R* factor, resolution and model information, and then compares them with values calculated from coordinate and structure-factor files. It also calculates an overall *B* factor, coordinate errors, an effective resolution and completeness. The summary of the density correlation shift and *B* factor are reported for each residue. As specific procedures are developed for checking NMR structures against experimental data, they will be incorporated into the PDB validation procedures.

24.5.2.3. *NMR data*

The PDB staff recognize that NMR data need a special development effort. Historically these data have been retro-fitted into a PDB format defined around crystallographic information. As a first step towards improving this situation, the PDB carried out an extensive assessment of the current NMR holdings and presented the findings to a task force consisting of a cross section of NMR researchers. The PDB is working with this group, the BioMagResBank (BMRB; Ulrich *et al.*, 1989) and other members of the NMR community to develop an NMR data dictionary along with deposition and validation tools specific for NMR structures.

24.5.2.4. *Data-processing statistics*

Production processing of PDB entries by the RCSB began on 27 January 1999. As of 1 July 1999, when the RCSB became fully responsible for the PDB, approximately 80% of all structures submitted to the PDB are deposited *via ADIT* and processed by the RCSB. Another 20% are submitted *via AutoDep* to the European Bioinformatics Institute (EBI), who process these submissions and forward them to the PDB for archiving and distribution. The average time from deposition to the completion of data processing including author interactions is two weeks. The number of structures with a HOLD release status remains at about 20% of all submissions; 57% are held until publication (HPUB); and 23% are released immediately after processing.

Table 24.5.2.2 shows the breakdown of the types of structures in the PDB. As of 14 September 1999, the PDB contained 10 714 publicly accessible structures with another 1169 entries on hold (not shown). Of these, 8789 (82%) were determined by X-ray methods, 1692 (16%) were determined by NMR and 233 (2%) were theoretical models. Overall, 35% of the entries have deposited experimental data.

24.5.3. The PDB database resource

24.5.3.1. *The database architecture*

In recognition of the fact that no single architecture can fully express the information content of the PDB, an integrated system of heterogeneous databases and indices that store and organize the structural data has been created. At present there are five major components (Fig. 24.5.3.1):

(1) The core relational database managed by Sybase (Sybase Inc., 1995) provides the central physical storage for the primary experimental and coordinate data described in Table 24.5.2.1. The core PDB relational database contains all deposited information in a tabular form that can be accessed across any number of structures.

(2) The final curated data files (in PDB format) and data dictionaries are the archival data and are present as ASCII files in the ftp archive.

(3) The POM-based databases (Shindyalov & Bourne, 1997) consist of indexed objects containing native (*e.g.* atomic coordinates) and derived properties (*e.g.* calculated secondary-structure assignments and property profiles). Some properties require no derivation, for example, *B* factors; others must be derived, for example, exposure of each amino-acid residue (Lee & Richards, 1971) or Cα contact maps. Properties requiring significant computation time, such as structure neighbours (Shindyalov & Bourne, 1998), are pre-calculated when the database is incremented to save considerable user-access time.

(4) The Biological Macromolecule Crystallization Database (BMCD; Gilliland, 1988) is organized as a relational database within Sybase and contains three general categories of literature-derived information: macromolecular, crystal and summary data.

(5) The Netscape LDAP server is used to index the textual content of the PDB in a structured format and provides support for keyword searches.

In the current implementation, communication among databases has been accomplished using the common gateway interface (CGI). An integrated web interface dispatches a query to the appropriate database(s), which then executes the query. Each database returns the PDB identifiers that satisfy the query, and the CGI program

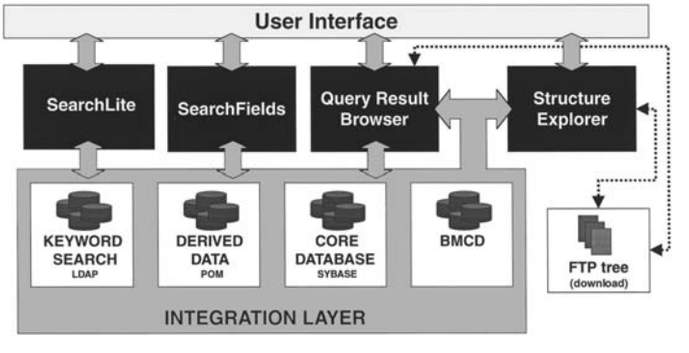

Fig. 24.5.3.1. The integrated query interface to the PDB.

Table 24.5.3.1. *Current query capabilities of the PDB*

(*a*) Query – single or iterative

Free text – any word in the PDB
Specific data items – compound name, author, description, deposition date, resolution, source, citation, cell dimensions, experimental method, data-collection method, refinement method, broad structure type, ligand (using the PDB HET records)
Property pattern – sequence, secondary structure
Structure similarity – 3D comparison

(*b*) Results analysis – single structure

Synopsis/Snapshot/Atlas – compound name, sequence, chemical components, citation, space group, cell constants, crystallization conditions, refinement details, structure views
Quick report – compound name, author, description, deposition date, resolution, source, citation, cell dimensions, experimental method, data-collection method, refinement method, geometry features
Full report – Quick report results plus secondary structure, chemical components, solvent
Property profiles – sequence, secondary structure
Links – see Table 24.5.3.2
Render – *RasMol*, *Chime*, *QuickPDB* (Java applet), VRML, *Protein Explorer*
Geometry – bond lengths, bond angles, dihedrals, close contacts, summary visual inspection

(*c*) Results analysis – multiple structure

Quick report – as above, but collated over multiple structures
Full report – as above, but collated over multiple structures
Structure neighbours – pairwise structure comparison

(*d*) Other query output options

mmCIF and PDB data files
Compressed files (gzip, tar, compressed)

integrates the results. Complex queries are performed by repeating the process and having the interface program perform the appropriate Boolean operation(s) on the collection of query results. A variety of output options are then available for use with the final list of selected structures.

The CGI approach (and in the future a CORBA-based approach) will permit other databases to be integrated into this system, for example, those containing extended data on different protein families. The same approach could also be applied to include NMR data found in the BMRB or data found in other community databases.

24.5.3.2. *Database queries*

Three distinct query interfaces are available for querying data within the PDB: *Status Query* (http://www.rcsb.org/pdb/status.html), *SearchLite* (http://www.rcsb.org/pdb/searchlite.html) and *SearchFields* (http://www.rcsb.org/pdb/cgi/queryForm.cgi). Table 24.5.3.1 summarizes the current query and analysis capabilities of the PDB. Fig. 24.5.3.2 illustrates how the various query options are organized.

SearchLite, which provides a single form field for keyword searches, was introduced in February 1999. All textual information within the PDB files as well as dates and some experimental data are accessible *via* simple or structured queries. *SearchFields*, accessible since May 1999, is a customizable query form that allows searching over many different data items, including compound, citation authors, sequence (*via* a *FASTA* search; Pearson & Lipman, 1988) and release or deposition dates.

Two user interfaces provide extensive information for results sets from *SearchLite* or *SearchFields* queries. The 'Query result browser' interface allows access to some general information, access to more detailed information in tabular format and the possibility of downloading whole sets of data files for result sets consisting of multiple PDB entries. The 'Structure explorer' interface provides information about individual structures as well as cross-links to many external resources for macromolecular structure data (Table 24.5.3.2). Both interfaces are accessible to other data resources through the simple CGI application programmer interface (API) described at http://www.rcsb.org/pdb/linking.html.

Table 24.5.3.3 indicates that usage has climbed dramatically since the system was first introduced in February 1999. Currently the PDB receives approximately 90000 web hits per day, or, on average, one query every second, seven days a week, 24 hours a day.

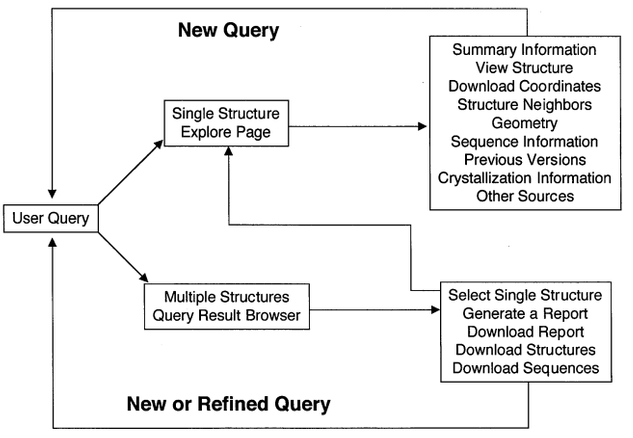

Fig. 24.5.3.2. The various query options that are available for the PDB.

references