# Protein Data Bank depositions from synchrotron sources

**Jiansheng Jiang* and Robert M. Sweet**

*Biology Department, Brookhaven National Laboratory, USA. E-mail: jiang@bnl.gov*

A survey and analysis of Protein Data Bank (PDB) depositions from international synchrotron radiation facilities, based on the latest released PDB entries, are reported. The results (http://asdp.bnl.gov/asda/Libraries/) show that worldwide, every year since 1999, more than 50% of the deposited X-ray structures have used synchrotron facilities, reaching 75% by 2003. In this web-based database, all PDB entries among individual synchrotron beamlines are archived, synchronized with the weekly PDB release. Statistics regarding the quality of experimental data and the refined model for all structures are presented, and these are analysed to reflect the impact of synchrotron sources. The results confirm the common impression that synchrotron sources extend the size of structures that can be solved with equivalent or better quality than home sources.

**Keywords: PDB deposition; synchrotron radiation facilities; structural genomics.**

## 1. Introduction

The Protein Data Bank (http://www.rcsb.org/; Berman *et al.*, 2000; Sussman *et al.*, 1998) is the unique internationally recognized database for the depositing and archiving of biological macromolecular structures. By 31 December 2003, the PDB had released 23792 protein structures. Of these, 17102 entries were X-ray structures. Conventional ('home') X-ray sources (*e.g.* the rotating-anode-based sources) have been used for years to determine X-ray structures. Synchrotron radiation sources became available for structural biology in the 1980s, and have had a significant impact since the early 1990s. The availability of international synchrotron facilities was reviewed and summarized in Table 1 of Helliwell (1998). Because of the continual gradual construction of new beamlines for macromolecular crystallography, synchrotron sources are now widely available, and many depositions to the PDB depend on these sources. We have searched all beamlines that make depositions to the PDB, and developed an automatically updated database called PXLIB (http://asdp.bnl.gov/asda/Libraries/) to track what we believe is an objective measure of this productivity, which perhaps will be a useful tool for many in both the synchrotron and structural biological communities. We report on the work to produce the database, and also describe analyses on the data regarding the relative quality of structures determined at synchrotron X-ray sources.

The completion of the Human Genome Project at the end of the 20th century stimulated a new biological initiative known as 'structural genomics' (SG). The goal of SG is systematically to determine the structures of proteins and important macromolecules, with the selection of the target proteins being guided somehow by the knowledge of the genomes of key organisms. This will yield a large number of 'representative' protein structures in the near future. In North America the NIGMS/NIH-funded Protein Structure Initiative (http://www.nigms.nih.gov/psi/) was started in the year 2000 to develop nine SG centers (Table 1). International efforts were also established (http://www.isgo.org); a SG issue was published in *Nature Structural Biology* in 2000 to summarize the worldwide program. The SG centers are focusing not only on producing protein structures but also on the development of techniques, software and apparatus to 'pipeline' or automate the process of macromolecular structure determination (Chance *et al.*, 2002; Holton & Alber, 2004; Brunzelle *et al.*, 2003). Most SG centers are tied up with synchrotron facilities for the high-throughput structure determinations. Because it seems likely that the SG effort will become a particularly rich source of synchrotron-linked PDB entries, we pay particular attention to this advance. Table 1 lists the synchrotron sources involved with the US SG centers; http://www.rcsb.org/pdb/strucgen.html lists worldwide SG centers.

## 2. Methods in the development of PXLIB

The PDB began to collect information on X-ray experimental details in a comprehensive way in 1995. Only a few records can be found on synchrotron radiation information from the PDB database earlier than 1995, although the PDB was started in 1972 (Bernstein *et al.*, 1977), and a substantial amount of data were collected at synchrotron sources during the period 1985–1995. Initially, the information about synchrotron radiation source and beamline was placed in records labelled 'REMARK 18'. By 1995 the PDB was collecting and annotating X-ray experimental details more formally; this evolved into the 'REMARK 200' record (Protein Data Bank Contents Guide, 1996). 'REMARK 200 SYNCHROTRON' indicates that the X-ray data were produced by synchrotron radiation; 'REMARK 200 RADIATION SOURCE' indicates whether the radiation source is from a home source (*e.g.* a rotating anode) or a synchrotron facility

**Table 1**
Synchrotron facilities involved with the nine SG centers (funded by PSI NIGMS, USA, since 2000).

Number of structures reported to the PSI NIGMS Annual Meeting as of 2 December 2003 (personal communication). Information on the worldwide SG centers can be found at http://www.rcsb.org/pdb/strucgen.html.

| SG centres | Synchrotron sources | Structures solved | Structures in PDB |
|---|---|---|---|
| Berkeley Structural Genomics Center | ALS | 37 | Not provided |
| Center for Eukaryotic Structural Genomics | APS | 6 | 6 |
| Joint Center for Structural Genomics | SSRL, APS | 102 | 53 |
| Midwest Center for Structural Genomics | APS: SBC-CAT, DND-CAT | 107 | 100 |
| New York Structural Genomics Research Consortium | NSLS: X12C, X25, X9A, X9B<br>APS: 31ID | 102 | 75 |
| Northeast Structural Genomics Consortium | APS: NE-CAT | 85 | 75 |
| Southeast Collaboratory for Structural Genomics | APS: SER-CAT | 35 | Not provided |
| Structural Genomics of Pathogenic Protozoa | ALS, SSRL | 1 | 1 |
| TB Structural Genomics Consortium | NSLS: X8C<br>HASYLAB, ALS | 62 | 25 |
| **Summary** | | **537** | **335** |

 DOI: 10.1107/S0909049504013792  

**Table 2**
Synchrotron facility sources and reported beamline identifications.

The reported beamline IDs were taking from the PDB file.

| Synchrotron sites | Beamlines |
|---|---|
| US | |
| NSLS (National Synchrotron Light Source, BNL) | X12C, X12B, X25, X26C, X4A, X8C, X9A, X9B, X6A, X29† (may mix with '-'). (Several entries with X4C have been merged to X4A) |
| APS (Advanced Photon Source, ANL) | 5IDB, 14BMC, 14BMD, 14IDB, 17BM, 17ID, 19BM, 19ID, 22ID, 31ID, 32ID, 8BM (may mix with '-'). |
| SSRL (Stanford Synchrotron Radiation Laboratory) | BL7-1, BL9-1, BL9-2, BL1-5, BL11-1 |
| ALS (Advanced Light Source, LBNL) | 5.0.1, 5.0.2, 5.0.3, 8.2.1, 8.2.2, 8.3.1 |
| CHESS (Cornell High Energy Synchrotron Source) | A1, F1, F2 (may mix with '-') |
| CAMD (Center for Advanced Microstructure and Devices) | GCPCC |
| European | |
| ESRF (European Synchrotron Radiation Facility, Grenoble, France) | ID14-1, ID14-2, ID14-3, ID14-4 (obsolete EH1, EH2, EH3, EH4), ID2 (obsolete BL4), ID9 (obsolete BL9), ID13 (obsolete BL1), ID29, BM14 (obsolete BL19), BM30A, BM30B (obsolete BM1, BM2, D1, D2), BM16†, ID23† |
| HASYLAB (DESY/EMBL, Hamburg, Germany) | BW6, BW7A, BW7B, X11, X13, X31 |
| SRS (Daresbury, England) | 7.2, 9.5, 9.6, 9.8, 14.1, 14.2 |
| LURE (Orsay, France) | DW32 (closed in 2003), DW21B (closed in 2001), D41 (closed in 2000) |
| SOLEIL (operation in 2006) | PROXIMA-1†, PROXIMA-2† |
| MAXLAB (MAXII, Lund, Sweden) | I711 |
| ELETTRA (Trieste, Italy) | XRD1 (same as 5.2R) |
| SLS (Swiss Light Source, Villigen, Switzerland) | X06SA |
| BESSY (Berlin, Germany) | BESSY |
| Asian | |
| SPring-8 (Super Photon Ring, Japan) | 12B2, 24XU, 26B1, 38B1, 40B2, 41XU, 44B2, 44XU, 45PX, 45XU (may prefix with 'BL') |
| PF (Photon Factory, KEK, Japan) | 6A, 6B, 18B, NW12, 5†, 6C† (may prefix with 'BL') |
| PAL (Pohang Light Source, Korea) | 6B |
| SRRC (Hsinchu, Taiwan) | 17B2 |
| South America | |
| LNLS (Brazil) | PCR (protein crystallography) |

† New beamline.

**Table 3**
Information extracted from a PDB entry file in the statistics file.

| | | |
|---|---|---|
| IDCODE | PDB ID code | From the latest release |
| DEPOSIT | Date of PDB deposition (in format YYYYMMDD) | HEADER record |
| COLLECT | Date of data collection (in format YYYYMMDD) | REMARK 200 DATA COLLECTION |
| LAG | Time difference from data collection to PDB deposition (in months) | (DEPOSITION − COLLECT) |
| EXPDTA | Type of experiment (X-RAY, NMR, EM *etc.*) | EXPDTA |
| RESOL | Resolution (Å) | REMARK 2 RESOLUTION |
| R value | *R* value | REMARK 3 REFINEMENT |
| FREE-R | Free *R* value | REMARK 3 REFINEMNET |
| METHOD | X-ray crystallography method used for structure determination | REMARK 200 METHOD USED |
| PROGRAM | Program used to refine structure | REMARK 3 PROGRAM |
| SIZE | Numbers of amino acids in asymmetry unit | SEQRES |
| SITES | Synchrotron facility sites | REMARK 200 RADIATION SOURCE |
| BEAMLINES | Beamlines used | REMARK 200 BEAMLINE |
| COMPLETENESS | Data completeness | REMARK 200 COMPLETENESS |
| REDUNDANT | Data redundant | REMARK 200 DATA REDUNDANT |
| R-MERGE | Overall *R*-merge | REMARK 200 R MERGE |
| R-SYM | Overall *R*-sym | REMARK 200 R SYM |
| ISIGI | Overall $\langle I/\sigma(I)\rangle$ | REMARK 200 <I/SIG(I)> |
| UREFLEX | Number of unique reflections | REMARK 200 UNIQUE REFLECTIONS |
| BOND | R.m.s.d. bond length from ideal | REMARK 3 RMSD BOND LENGTH |
| ANGLE | R.m.s.d. bond angle from ideal | REMARK 3 RMSD BOND ANGLE |
| B-FACTOR | Overall average *B*-factor | REMARK 3 MEAN B VALUE |
| LUZZATI | Estimated coordinate error by the Luzzati plot | REMARK 3 ESD FROM C-V LUZZATI |
| SIGMAA | Estimated coordinate error by the plot | REMARK 3 ESD FROM C-V SIGMAA |
| ESU | Estimated coordinate error by ESU | REMARK 3 ESU BASED ON FREE R |
| NREFLEX | Number of reflections used in refinement | REMARK 3 NUMBER OF RELECTIONS |
| NATOM | Number of atoms in refinement | REMARK 3 NUMBER OF ATOMS |

site (*e.g.* NSLS); 'REMARK 200 BEAMLINE' indicates the particular beamline on which the data were collected. When two or more beamlines have been used, all beamline identifications should be presented in the deposition (separated by semicolons). In this study we parsed the latest released PDB entries and extracted all necessary information from the PDB flat files.

BIOSYNC (Structural Biology Synchrotron Users Organization, http://biosync.sdsc.edu) lists all beamline names in the USA but it does not provide beamline names for overseas. We have identified 82 synchrotron beamlines worldwide that have made depositions to the PDB, as listed in Table 2. Some beamline names have changed over time and non-standard beamline identifiers and multiple names have been deposited in the PDB entries. This leads to some difficulties in coding a program/script. We collect and include all possible alternative names and obsolete IDs in the brackets as 'synonyms' to a 'representative' beamline. Table 2 lists the 'representative' beamlines for each synchrotron facility site. The detailed 'synonyms' for each beamline are posted on the web site (http://asdp.bnl.gov/asda/Libraries/pdb_statis/). The 'representative' beamline may not be a single physical beamline. For example, at APS, sector 22 has two physical beamlines, 22ID and 22BM. Since 22BM has not yielded any PDB deposition, '22ID' in Table 2 represents both beamlines 22ID and 22BM. Some beamlines were closed, for example, LURE at ORSAY in France. However, the depositions that were made in the past, and there may be more depositions in the coming years, will still be present in the table. If a PDB deposition is received from a new beamline (after confirmation with the synchrotron source), the new beamline will appear in the table on the web site.

In order to catch the latest entries released from the PDB, we have installed a mirrored PDB database at BNL. The mirrored PDB database is updated every week on Wednesday. The PXLIB script parses every PDB file to extract the data that are needed for the statistics. Table 3 lists the information that is extracted and collected in statistics files. We count every occurrence of beamline IDs in all PDB depositions as the 'credits' for the synchrotron facilities. If multiple beamline IDs appear in the same record (separated by a semicolon), we count each occurrence as an equal fractional credit. For example, if both X25 (NSLS) and 19ID (APS) appeared on the same record of

'REMARK 200 BEAMLINE', both X25 and 19ID will have half a credit. In order to identify the unique beamline ID, we first abbreviate the entry by removing or adding prefixes and insertions, such as '-', 'BL', 'PX' or 'STATION', then we map it to the unique site ID and beamline ID as listed in Table 2. We make necessary corrections if some obvious errors are found in the PDB file. For example, 'SSRF' should be 'SSRL', 'X12C = PF' should be 'X12C = NSLS', 'BL711 = PF' should be 'I711 = MAXLAB'. There are some quoted IDs for beamlines but no indication of the synchrotron facility site. In this case, we found out the source ID from Table 2. Some PDB entries indicate the synchrotron facility site only but no information about the beamline; we denote them as in the 'UNC' (unclaimed) beamline for this facility source site.

PXLIB runs automatically on every Wednesday in synchrony with the PDB release. Therefore, the tables in the 'Latest Release' will automatically update every week. The data in the past weeks will be pushed down to 'Archives'. Meanwhile, a conversion script generates HTML pages to update the PXLIB web site. PXLIB also builds up a gallery of PDB images of structures for each beamline. The program PXLIB is a set of scripts written in the PERL and C-SHELL languages. The source codes are available and can be downloaded from http://asdp.bnl.gov/asda/Libraries/. PXLIB consists of two major steps: (i) extracting data from the mirrored PDB database, 'run_pdb_statis1.csh'; (ii) performing analysis and statistics on the extracted data, 'run_pdb_statis2.csh'.

## 3. Results and analysis

### 3.1. The numbers of PDB depositions from synchrotron sources

The numbers of PDB depositions from all synchrotron beamlines and various statistics derived from these have been posted at http://asdp.bnl.gov/asda/Libraries/pdb_statis/. The numbers in Table 4 are an example of the sort of data that can be derived; one can see that the tables on the PXLIB web site are referenced to the year when the PDB deposition was made. There is a policy at the PDB that the depositor is allowed to demand that the entry be held for up to one year, waiting for publication of the structure, before being released. Since there are some depositions that had been made in the deposition year but not yet released, the numbers in the current deposition year will be partial counts. Table 4(*a*) summarizes the PDB depositions from synchrotron sources as of 31 December 2003. The column headed XRAY is the total number of X-ray structures in the deposition year (since 1995), that headed SYNC is the total PDB depositions from synchrotron sources, and that headed HOMES is the PDB depositions from home sources (including unknowns). In Table 4(*b*) the total SYNC numbers are broken down to the sources from the US (North America), EURO (Europe), ASIAN (Asia) and S.AMER (South America). One may also browse the PXLIB web site to find data for individual beamlines.

It is clear that by 1999 the synchrotron facilities contributed more than 50% of X-ray structures. The fraction leaps to 60% in 2000, and it continues a 5% growth annually. According to Table 4(*a*), PDB depositions from synchrotron sources will be more than 75% for year 2003. Since 1995, synchrotron sources from the US and Europe have contributed to 4505 and 3944 PDB structures, respectively (Table 4*b*). By the end of December 2003, the average depositions from the US (35 beamlines) and Europe (30 beamlines) were almost equal (130 depositions per beamline). Fig. 1(*a*) compares the number of depositions of each synchrotron facility site from US sources; Fig. 1(*b*) plots the depositions from European synchrotron sources, and Fig. 1(*c*) shows the depositions from Asian synchrotron sources. Since 1995, in the US the NSLS made the largest number of contributions to the PDB until 2002, at which point the third-generation APS took the lead. ESRF in Europe and SPring-8 in Asia are leading the PDB depositions.

**Table 4**
PDB depositions from international synchrotron facilities as of 31 December 2003: (*a*) depositions from all synchrotron sources (SYNC) and home sources (HOMES); (*b*) depositions from United States (US), European (EURO), Asian (ASIAN) and South American (S.AMER) synchrotron sources.

The column headed XRAY is the total number of X-ray structures for the deposition year. The total number of depositions from all experimental methods, including NMR *etc*, are not shown, but the column headed % is the percentage of contributions resulting from X-ray crystallography. The lysozyme-related structures are about 3% in the counts and most of them were deposited from home sources.

(*a*)

| YEAR | XRAY | SYNC | % | HOMES | % |
|---|---|---|---|---|---|
| 1995 | 897 | 170 | 19.0 | 726 | 80.9 |
| 1996 | 1113 | 279 | 25.1 | 832 | 74.8 |
| 1997 | 1472 | 499 | 33.9 | 975 | 66.2 |
| 1998 | 1763 | 670 | 38.0 | 1094 | 62.1 |
| 1999 | 2099 | 1080 | 51.5 | 1010 | 48.1 |
| 2000 | 2420 | 1482 | 61.2 | 935 | 38.6 |
| 2001 | 2612 | 1713 | 65.6 | 898 | 34.4 |
| 2002 | 2694 | 1897 | 70.4 | 800 | 29.7 |
| 2003 | 2032 | 1524 | 75.0 | 508 | 25.0 |
| SUM | 17102 | 9314 | 54.5 | 7778 | 45.5 |

(*b*)

| YEAR | US | % | EURO | % | ASIAN | % | S.AMER | % |
|---|---|---|---|---|---|---|---|---|
| 1995 | 50 | 5.6 | 91 | 10.1 | 29 | 3.2 | 0 | 0.0 |
| 1996 | 97 | 8.7 | 151 | 13.6 | 31 | 2.8 | 0 | 0.0 |
| 1997 | 199 | 13.5 | 270 | 18.3 | 30 | 2.0 | 0 | 0.0 |
| 1998 | 288 | 16.3 | 340 | 19.3 | 42 | 2.4 | 0 | 0.0 |
| 1999 | 542 | 25.8 | 458 | 21.8 | 76 | 3.6 | 4 | 0.2 |
| 2000 | 696 | 28.8 | 650 | 26.9 | 134 | 5.5 | 2 | 0.1 |
| 2001 | 862 | 33.0 | 691 | 26.5 | 155 | 5.9 | 5 | 0.2 |
| 2002 | 921 | 34.2 | 764 | 28.4 | 207 | 7.7 | 5 | 0.2 |
| 2003 | 850 | 41.8 | 529 | 26.0 | 141 | 6.9 | 4 | 0.2 |
| SUM | 4505 | 26.3 | 3944 | 23.1 | 845 | 4.9 | 20 | 0.1 |
| Beamlines† | 35 | – | 30 | – | 16 | – | 1 | – |

† Number of beamlines, see Table 2.

There is a rich source of information available on the PXLIB web site (http://asdp.bnl.gov/asda/Libraries/pdb_statis/). Selecting 'Latest Release' will present an overview of total PDB depositions from different areas; from there, selecting the location and the synchrotron facility IDs will lead to the tables for individual sites and beamlines. Then, within that, selecting a beamline will give a table of entries. Selecting 'Gallery' allows one to break out depositions from each facility and to display thumbnail images of all entries from each beamline; selecting 'Primary Citation' will list all publications that are related to this beamline; selecting the PDB ID code will lead to the PDB explore page (http://www.rcsb.org/pdb/cgi/explore.cgi?pdbid=1abc) which allows easy visualization of the contents of the PDB entry and the model using *Protein Explorer* (Martz, 2002). There are also alternate text forms rather than HTML, and raw data is provided in different formats. The source code is available from the web site, and potential users may contact the first author for assistance.

### 3.2. Lag time

The lag time is the difference between the date of data collection and the date of PDB deposition, in months. Table 5 shows, for each year, the average lag time and number of structures deposited for synchrotron and home sources. There is a clear difference between

the lag times for structures from synchrotron sources and home sources: 18 months and 25 months, respectively. Evidently the structures that came from synchrotron facilities somehow seemed to be more urgent, or perhaps were easier to complete, and therefore were deposited sooner. The average lag time over the PDB depositions for each beamline is also calculated and is available from the PXLIB web site.

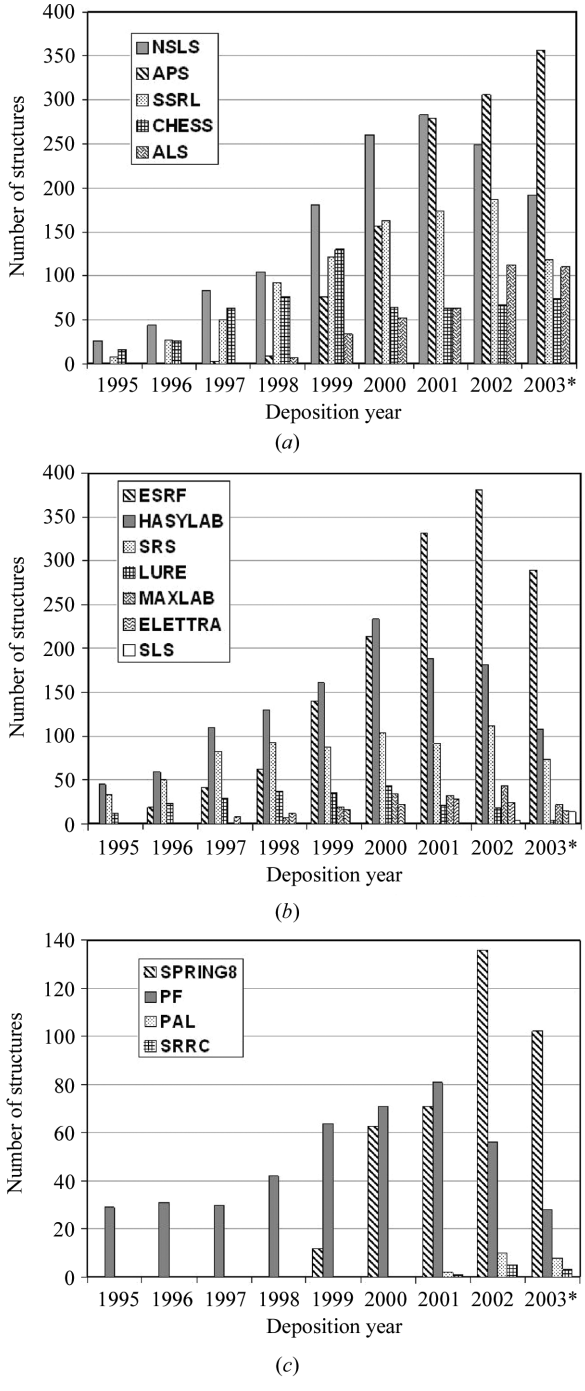


**Figure 1**
PDB depositions (*a*) from synchrotron sources in the US, (*b*) from European synchrotron sources and (*c*) from Asian synchrotron sources, as of 31 December 2003. The breakdown numbers by year are based on the deposition date. *The numbers in 2003 are partial counts since some deposited entries have not yet been released.

**Table 5**
Average lag time (time between data collection and submission to the PDB) over the structures from all synchrotron and home sources based on the deposition date.

| | Synchrotron source | | Home source | |
|---|---|---|---|---|
| Deposition year | Number of structures | Lag (months) | Number of structures | Lag (months) |
| 1995 | 50 | 21 | 157 | 20 |
| 1996 | 271 | 21 | 670 | 23 |
| 1997 | 495 | 20 | 831 | 26 |
| 1998 | 597 | 19 | 895 | 25 |
| 1999 | 920 | 16 | 767 | 24 |
| 2000 | 1448 | 17 | 792 | 26 |
| 2001 | 1614 | 17 | 752 | 26 |
| 2002 | 1746 | 18 | 626 | 24 |
| 2003 | 1306 | 17 | 423 | 22 |
| Overall | 8447 | 18 | 5913 | 25 |

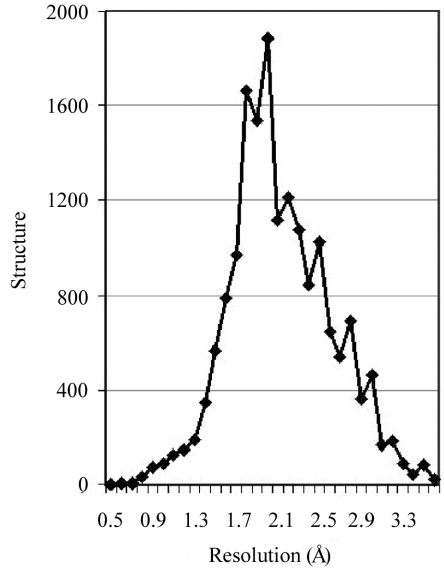


**Figure 2**
Distributions of the reported resolutions over 17144 X-ray structures. The distribution has a mean of 2.16 Å and a standard deviation of 0.53 Å.

### 3.3. Statistics on resolution and the size of the molecular structure

Various quality statistics for X-ray structures from the latest released PDB entries are summarized in Tables 6 and 7. It is intriguing to note that in Table 6 the mean values of resolution limit and both $R$ values are essentially identical for synchrotron and home sources. The difference is that the typical structure solved at a synchrotron source is more than half as large again as that solved on a home source, *i.e.* 738 residues and 453 residues, respectively. Several beamlines shown an average size of more than 1000 residues with a large standard deviation, which indicates that the beamline might have recorded data for larger macromolecular complex structures, such as ribosome. We analyse this difference later, but because of this similarity we combine the two for several analyses. The distribution of the reported resolutions over all deposited X-ray structures is plotted in Fig. 2. The average resolution is 2.2 Å with a standard deviation of 0.5 Å over 17144 structures. More than 90% of structures have the data falling in the resolution range between 1.5 Å and 3.0 Å. Only 6% of structures have a higher resolution than 1.5 Å, and 4% of structures have resolutions lower than 3.0 Å. The mean and standard deviation for each synchrotron site and beamline look similar except for a number of beamlines that have a few large structures (with 7.0 Å

to 8.0 Å resolution); a corollary of this is that several beamlines have a larger average size than the others.

### 3.4. Statistics on *R* values and the free *R* values

The crystallographic *R* value and free *R* value (Brünger, 1992, 1997) are reported in each PDB deposition to define how well coordinates of the model fit the X-ray data. Fig. 3 displays the distribution of the reported *R* values and free *R* values over 14746 structures. Each spot represents a structure. A number of free *R* values were reported less than *R* values in the plot (below the diagonal line $y = x$). These free *R* values might not be correct owing to the inefficient 'cross-validation'. Free *R* values and *R* values are highly correlated. The correlation coefficient over all reported X-ray structures is 0.814 (see Table 8). The *R* value has a mean of 0.20 and a standard deviation of 0.03 and the free *R* value has a mean of 0.25 and a standard deviation of 0.04 (see Table 7). The means of *R* values and free *R* values as a function of the reported resolution are plotted in Fig. 4. Error bars represent the associated standard deviations. The resolution bin is set to 0.2 Å. The number of the *R*/free-*R* values is significant in the resolution bins from 0.8 Å to 3.6 Å ($d_{\min}$). The standard deviations are smooth in the range between 0.027 and 0.037 for both the *R* values and free *R* values. The dependency of both *R* value and free *R* value on the resolution is clearly shown. We have further analysed the relation between the free *R* value and the resolution based on these statistics, which will be reported in a separate paper (Jiang, 2004).

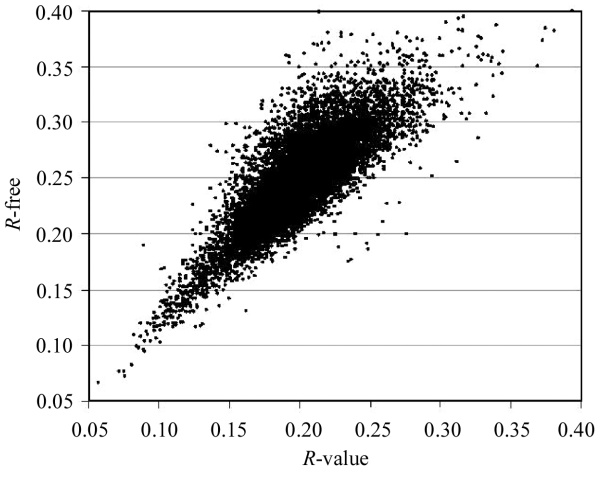


**Figure 3**
Correlation of *R* values and free *R* values. Each spot stands for one structure. The correlation coefficient over a total of 14746 structures is 0.814 (see Table 8). The *R* value distribution has a mean of 0.199 and a standard deviation of 0.032 and the free *R* value distribution has a mean of 0.248 and a standard deviation of 0.039.

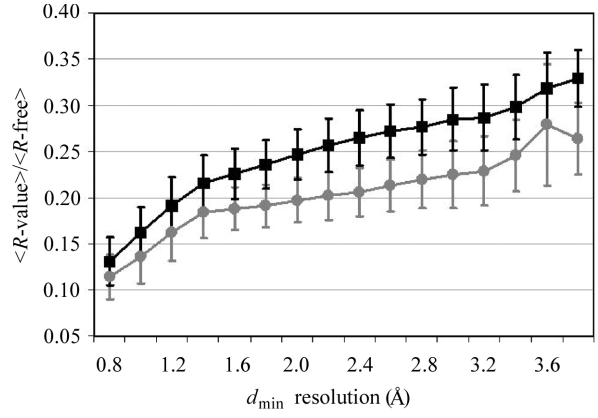


**Figure 4**
The means and the associated standard deviations of *R* values; the free *R* values are plotted as a function of reported resolution. The resolution bin is set to 0.2 Å. A total of 14087 X-ray structures are in these statistics. The grey line with circles represents *R* values; the dark line with squares represents free *R* values. The error bars represent the standard deviations.

**Table 6**
Mean and standard deviation of crystallographic statistical properties over numbers of X-ray structures, for synchrotron (SYNC) and home (HOMES) sources.

The column headed Size ($\sigma$) shows the number of amino acid or nucleotide residues.

| | Structures | Resolution ($\sigma$) | *R* value ($\sigma$) | Free *R* ($\sigma$) | Size ($\sigma$) |
|---|---|---|---|---|---|
| SYNC | 9781 | 2.15 (0.60) | 0.205 (0.035) | 0.248 (0.041) | 737 (928) |
| HOMES | 7976 | 2.17 (0.44) | 0.191 (0.027) | 0.247 (0.035) | 452 (505) |
| All | 17757 | 2.16 (0.53) | 0.199 (0.033) | 0.248 (0.039) | 609 (781) |

**Table 7**
Statistics on the data and model qualities, combined for all X-ray sources (ALL), synchrotron sources (SYNC) and home sources (HOMES) (as of 31 December 2003).

The mean is over the numbers of structures reported in PDB deposition; $\sigma$ is the standard deviation associated with the mean. The estimated coordinates errors by Luzzati plot, $\sigma_A$ plot and ESU are based on the free *R* values.

| | ALL | | | SYNC | | | HOMES | | |
|---|---|---|---|---|---|---|---|---|---|
| Data and model qualities | Structures | Mean | $\sigma$ | Structures | Mean | $\sigma$ | Structures | Mean | $\sigma$ |
| Completeness (%) | 16076 | 93.2 | 8.05 | 9281 | 94.6 | 6.95 | 6795 | 91.4 | 9.04 |
| Redundancy | 13903 | 5.00 | 3.69 | 8043 | 5.42 | 3.98 | 5860 | 4.43 | 3.18 |
| *R*-merge | 12616 | 0.072 | 0.031 | 7012 | 0.071 | 0.031 | 5604 | 0.072 | 0.030 |
| *R*-sym | 5583 | 0.077 | 0.060 | 3293 | 0.074 | 0.059 | 2290 | 0.081 | 0.062 |
| $\langle I/\sigma(I)\rangle$ | 11956 | 15.8 | 10.3 | 7384 | 16.3 | 10.2 | 4572 | 15.0 | 10.6 |
| Number of reflections | 17290 | 44516 | – | 9601 | **57928** | – | 7689 | **27770** | – |
| Number of atoms | 16719 | 5005 | – | 9106 | **6071** | – | 7613 | **3731** | – |
| Ratio of atoms/reflections | 16387 | 0.150 | 0.094 | 8972 | 0.138 | 0.091 | 7415 | 0.166 | 0.095 |
| R.m.s.d. bond lengths (Å) | 16187 | 0.011 | 0.007 | 8927 | 0.011 | 0.007 | 7260 | 0.012 | 0.006 |
| R.m.s.d. bond angles (°) | 14535 | 1.69 | 0.58 | 7801 | 1.57 | 0.46 | 6734 | 1.83 | 0.68 |
| Average *B* values (Å$^2$) | 10096 | 30.8 | 14.7 | 6177 | **33.3** | 16.4 | 3919 | **26.7** | 10.7 |
| Estimated coordinates error Luzzati (Å) | 5560 | 0.329 | 0.112 | 3641 | 0.339 | 0.117 | 1919 | 0.311 | 0.099 |
| Estimated coordinates error by $\sigma_A$ (Å) | 5119 | 0.330 | 0.185 | 3533 | 0.335 | 0.194 | 1586 | 0.319 | 0.164 |
| Estimated coordinates error by ESU (Å) | 1140 | 0.188 | 0.120 | 917 | 0.185 | 0.123 | 223 | 0.202 | 0.105 |

**Table 8**
Correlation coefficients between pairs of statistical properties over the X-ray structures in the PDB.

SYNC is over the structures from synchrotron sources, HOMES is over the structures from home sources, and ALL is over all available structures (both SYNC and HOMES). cc = $\{(\langle xy\rangle - \langle x\rangle\langle y\rangle)/[(\langle x^2\rangle - \langle x\rangle^2)(\langle y^2\rangle - \langle y\rangle^2)]\}^{1/2}$. Occasions where the difference between SYNC and HOMES is especially large are emphasized with bold type.

| | Correlation coefficient ($x$, $y$) ALL SYNC HOMES | | | | |
|---|---|---|---|---|---|
| | Resolution | *R* value | *R*-free | *R*-merge | *B*-factor |
| *R* value | 0.513 | | | | |
| | **0.632 0.336** | | | | |
| *R*-free | 0.641 | 0.814 | | | |
| | 0.697 0.514 | 0.864 0.715 | | | |
| *R*-merge | 0.355 | 0.167 | 0.177 | | |
| | 0.339 0.386 | 0.179 0.172 | 0.185 0.168 | | |
| R.m.s.d. bond length from ideal | −0.046 | −0.136 | −0.050 | | |
| R.m.s.d. bond angle from ideal | −0.005 | −0.150 | 0.034 | | |
| *B*-factor | 0.538 | 0.538 | 0.498 | 0.085 | |
| | **0.638 0.358** | 0.571 0.354 | **0.573 0.308** | 0.103 0.063 | |
| Estimated coordinate error by Luzzati | 0.825 | 0.674 | 0.787 | 0.220 | 0.712 |
| | 0.854 0.767 | 0.697 0.613 | 0.801 0.754 | 0.187 0.323 | 0.758 0.578 |
| Estimated coordinate error by $\sigma_A$ | 0.802 | 0.548 | 0.657 | 0.299 | 0.684 |
| | 0.809 0.784 | 0.584 0.470 | 0.674 0.611 | 0.263 0.422 | 0.732 0.567 |
| Estimated coordinate error by ESU | 0.908 | 0.645 | 0.727 | 0.222 | 0.529 |
| | 0.905 0.933 | 0.654 0.615 | 0.729 0.717 | 0.233 0.203 | 0.551 0.420 |
| Ratio of atoms/reflections | 0.747 | 0.266 | 0.514 | 0.267 | 0.301 |
| | 0.753 0.765 | **0.421 0.149** | 0.574 0.437 | 0.232 0.314 | 0.461 0.157 |

### 3.5. Statistics on X-ray data qualities and coordinate model qualities

Some statistics representing the quality of the raw data and of the refined models, averaged over all structures, are summarized in Table 7. The completeness of the data measured has a mean of 93% with a standard deviation of 8%, and does not depend on resolution. The redundancy has a mean of 4.9 with a standard deviation of 3.6 (Fig. 5) and the distribution does not depend on resolution. $\langle I/\sigma(I)\rangle$ has a mean of 16 with a standard deviation of 10 (Fig. 6) and its distribution also does not depend on resolution. The statistical properties of the reported *R*-merge (12480 cases) and the reported *R*-sym (only 5487 cases) are very similar. *R*-merge has a mean of 0.074 with a standard deviation of 0.05 (Fig. 7*a*) while *R*-sym has a mean of 0.077 with a standard deviation of 0.06. The distribution of *R*-merge has a small dependence on resolution. The means of *R*-merge with associated standard deviations are plotted as a function of resolution in Fig. 7(*b*).

The r.m.s. bond-length deviation from the ideal has a mean of 0.012 Å with a standard deviation of 0.008 Å (Fig. 8*a*), and the r.m.s. bond angle deviation from the ideal has a mean of 1.7° with a standard deviation of 0.8° (Fig. 9*a*). In Fig. 8(*b*) and Fig. 9(*b*) are plotted the means and standard deviations against the resolution; neither depends on the resolution. The bond-length r.m.s.d. has a higher mean in the high-resolution structures ($d_{\rm min}$ < 1.5 Å) and a smooth lower mean (around 0.01) through the lower-resolution range. From this it can be interpreted that the lower-resolution ($d_{\rm min}$ > 1.5 Å) structures were heavily restrained to achieve a good geometry of the model (Engh & Huber, 1991). Fig. 10(*a*) shows the distributions of the overall *B*-factors. The average of the overall *B*-factors on 11736 structures is 31 Å$^2$ with a standard deviation of 16 Å$^2$. The means and standard deviations of the *B*-factors as a function of resolution are given in Fig. 10(*b*). It is no surprise that as $d_{\rm min}$ increases (lower resolution) the mean of the *B*-factors increases; the standard deviation also increases. The coordinate errors estimated by the Luzzati plot (Luzzati, 1952) derived from the cross-validated free *R* value have a mean of 0.33 Å with a standard deviation of 0.11 Å over 5447 structures; the coordinate errors estimated (cross-validated) by the $\sigma_A$ plot (Read, 1986, 1990) have a mean of 0.33 Å with a standard deviation of 0.18 Å over 5004 structures; the estimated coordinate errors by ESU [estimated standard uncertainty based on the free *R* value given by *REFMAC* (Murshudov *et al.*, 1997)] have a mean of 0.18 Å with a standard deviation of 0.12 Å over 1081 structures.

We have computed the correlation coefficients (cc) between pairs of some statistical properties over all released X-ray structures from the PDB (Table 8). Most data and model qualities are correlated to the resolution except for r.m.s. bond length and angle deviations from ideal. The estimated coordinate errors are highly correlated with the

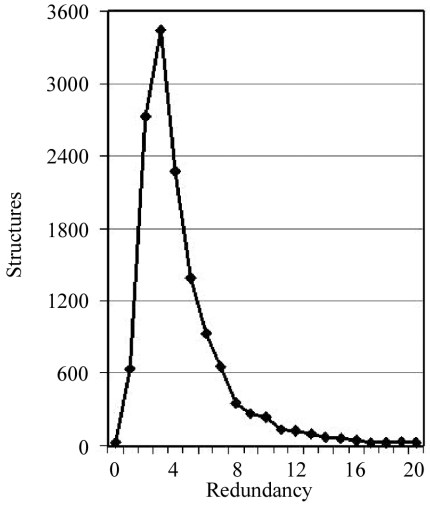


**Figure 5**
The distributions of the X-ray data redundancy. The number of structures as a function of redundancy. The redundancy has a mean of 4.9 with a standard deviation of 3.6.

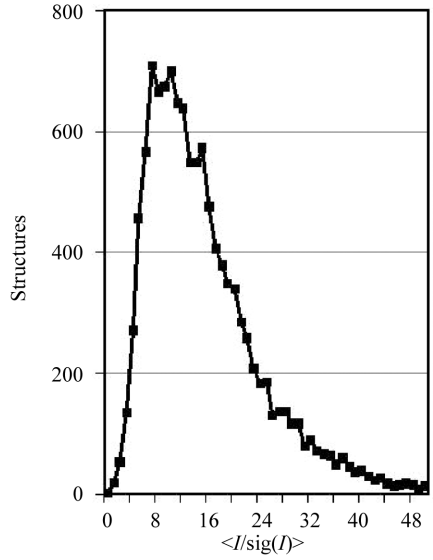


**Figure 6**
The distributions of data quality $\langle I/\sigma(I)\rangle$. The number of structures as a function of $\langle I/\sigma(I)\rangle$. $\langle I/\sigma(I)\rangle$ has a mean of 15.8 with a standard deviation of 10.4.

resolution, free $R$ value and $B$-factor (cc > 0.5), but only slightly correlated with $R$-merge (cc > 0.21).

### 3.6. Differences between synchrotron and home sources

Tables 6, 7 and 8 show any differences that might arise between data taken at synchrotrons and with home sources. As already mentioned, the data in Table 6 reveal the slightly surprising statistic that the most dramatic difference between the SYNC- and HOMES-based structures is the size of the structure, not the final quality. This is borne out in the thorough analysis of data and model quality in Table 7, where the only significant differences are shown in bold type. These suggest that the average number of reflections from synchrotron sources is much higher than those from home source, 57928 *versus* 27770. Also, since the average resolution limit between SYNC and HOMES is the same (Table 6), the average number of refined atoms follows: 6071 *versus* 3731. However, this relationship is not strictly true since, as the SYNC data tend to be slightly more complete, the ratio of atoms over reflections is slightly smaller in SYNC than in HOMES. Since it is this parameter-to-data ratio that defines the quality of a final refined structure, one might expect the SYNC structures to be slightly more accurate at a given resolution.

The other interesting discrepancy in Table 7 is that the average $B$-factors are significantly larger for SYNC data than for HOMES data. Not only are SYNC-derived structures larger on average, but they are also harder to determine because the $B$ factors are larger. This set of statistics [*i.e.* equivalent resolution and $R$ value, but bigger structures and weaker high-resolution data (larger numbers of reflections)] represents the crux of the power of synchrotron radiation for macromolecular crystallography.

It is worth noting that the correlations (Table 8) between $R$-value/$R$-free and the reported resolution are much higher for SYNC than for HOMES (0.63 *verses* 0.34 on $R$ value, 0.70 *verses* 0.51 on $R$-free). $B$-factors tend to be more highly correlated with SYNC-derived resolution and $R$ values than with HOMES-derived ones (Table 8).

### 3.7. X-ray crystallographic methods

X-ray crystallography methods are also reported in the PDB entry. The information can be extracted from the record of 'REMARK 200 METHODS USED'. However, the record is a text string that is difficult to parse, and about a quarter of entries do not have such information. We abbreviate the commonly used methods (http://asdp.bnl.gov/asda/Libraries/pdb_statis/latest/xme/METH.html) and then make necessary corrections and interpretations. We divide the methods into four major categories: molecular replacement, experimental phasing methods, direct methods and miscellaneous. The percentages of these four categories are 59%, 23%, 1% and 17%, respectively. The experimental phasing methods can be divided into

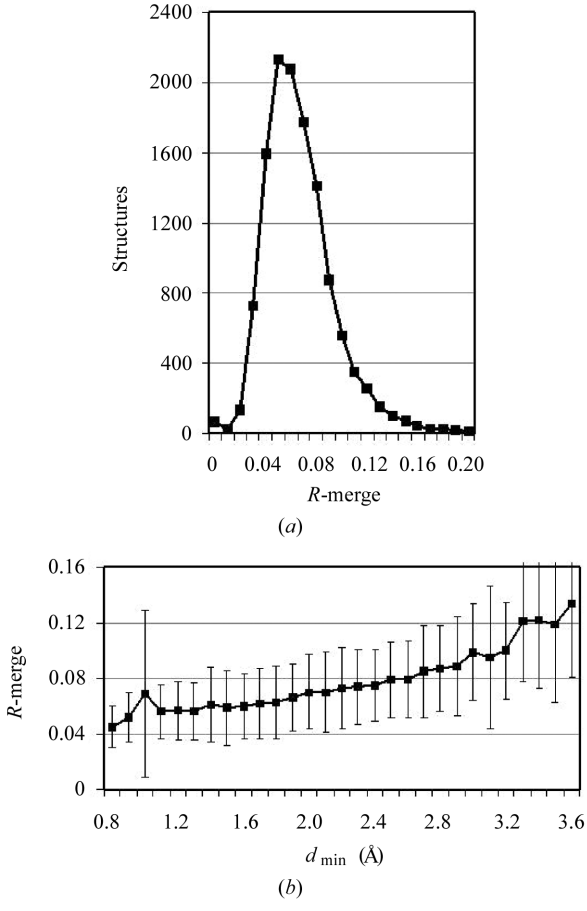


**Figure 7**
The distributions of $R$-merge. (*a*) The number of structures as a function of $R$-merge. $R$-merge has a mean of 0.074 with a standard deviation of 0.05. (*b*) The means of $R$-merge with associated standard deviations (error bars) are plotted as a function of resolution.

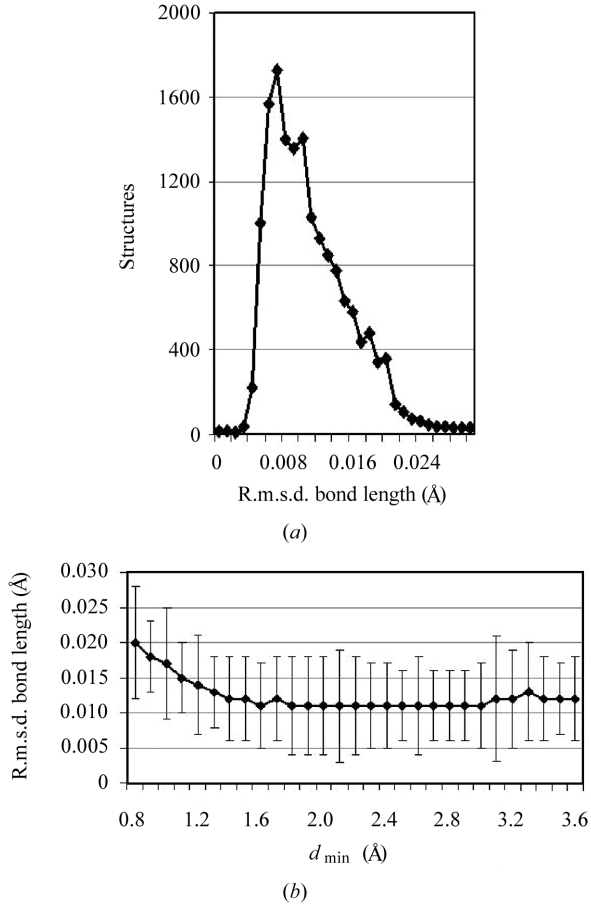


**Figure 8**
(*a*) The number of structures as a function of the r.m.s.d. bond length. The r.m.s.d. bond length from the ideal has a mean of 0.012 Å with a standard deviation of 0.008 Å over 15862 structures. (*b*) The means of the r.m.s.d. bond lengths with associated standard deviations (error bars) are plotted as a function of resolution.

two groups: the single methods (83%) and the combinations (17%). The single methods are composed of four individual phasing methods, multiple isomorphous replacement (MIR, 31%), single isomorphous replacement (SIR, 10%), multiwavelength anomalous diffraction (MAD, 52%) and single-wavelength anomalous diffraction (SAD, 7%). The combinations are the combination of anomalous scattering, isomorphous replacement and molecular replacement. Table 9 summarizes all reported X-ray crystallographic methods. Fig. 11 shows the numbers of reported experimental phasing methods (individual) referenced to the deposition year. One can see that the number of MAD structures has increased dramatically (Ogata, 1998), and it doubled in one year, from 2001 to 2002. One might also note that the ratio of SAD to MAD structures seems to be increasing precipitously. This single-wavelength anomalous method works best with tunable synchrotron radiation and the other qualities of synchrotron facilities: well collimated beams to provide good signal-to-noise, and state-of-the art detection systems.

## 4. Discussion

### 4.1. Number of PDB depositions from synchrotron sources

In this survey, all counts are based on the PDB depositions beginning in 1995. A substantial number of PDB depositions earlier than 1995 are excluded because the PDB did not systematically

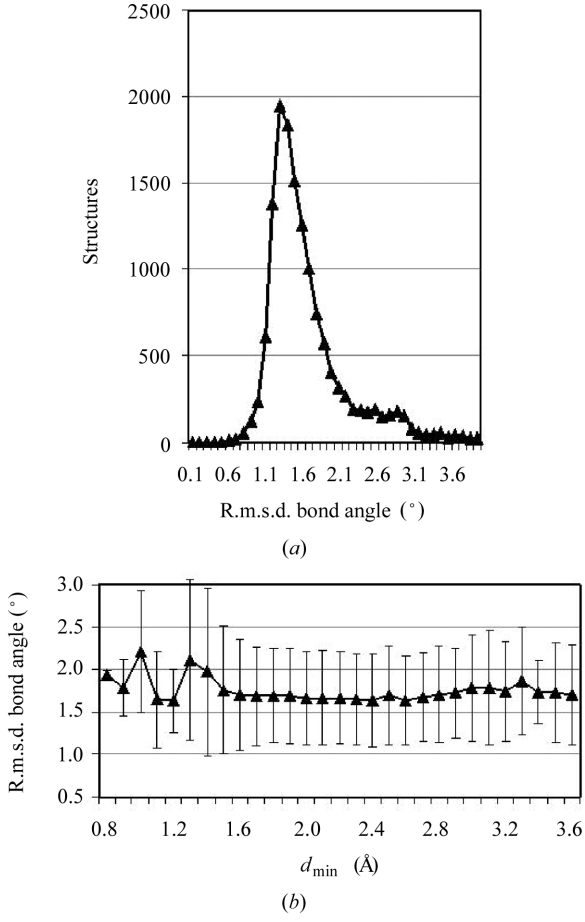


**Figure 9**
(*a*) The number of structures as a function of the r.m.s.d. bond angle. The r.m.s.d. bond angle from the ideal has a mean of 1.73° with a standard deviation of 0.8° over 14289 structures. (*b*) The means of the r.m.s.d. bond angle with associated standard deviations (error bars) are plotted as a function of resolution.

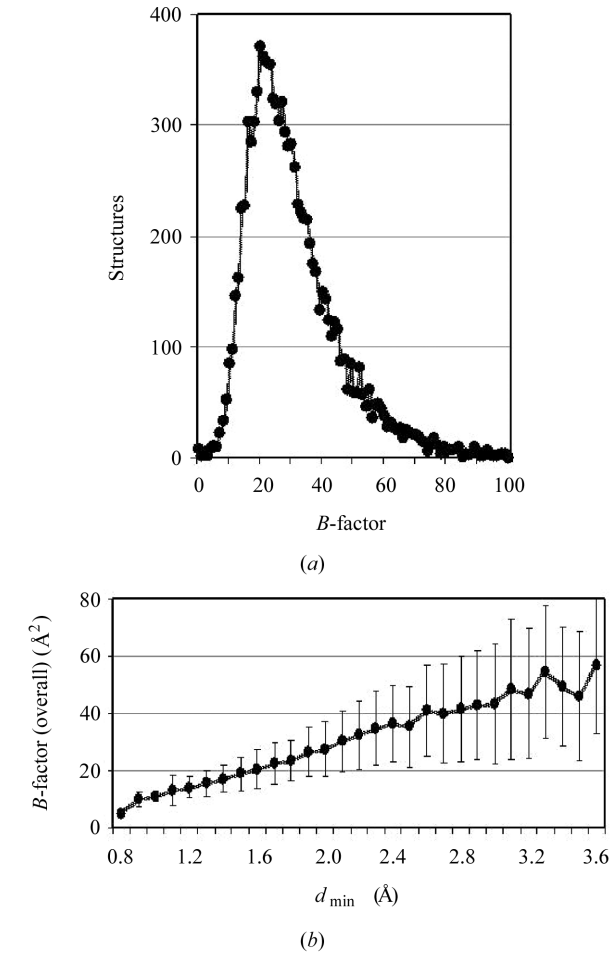


**Figure 10**
(*a*) The number of structures as a function of the overall *B*-factors. The average of the overall *B*-factors over 11736 structures is 31.1 with a standard deviation of 16.4 over 9811 structures. (*b*) The means of the overall *B*-factors with associated standard deviations (error bars) are plotted as a function of resolution.

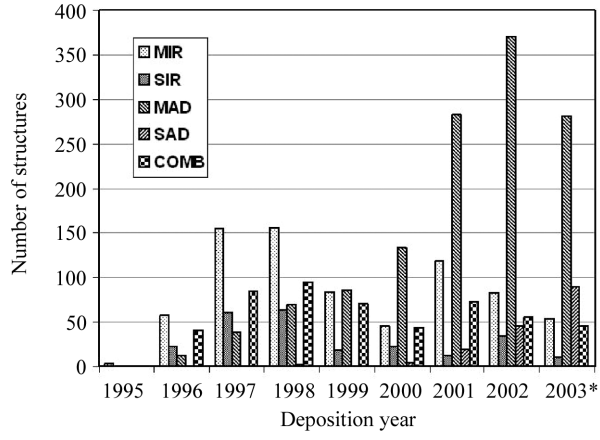


**Figure 11**
The numbers of reported experimental phasing methods, MIR, SIR, MAD, SAD and their combinations over the deposition years. The bars with stripes represent the structures from MAD and SAD and the bars with dots represent the structures from MIR and SIR. MAD doubles in 2001. *Year 2003 shows partial counts because some depositions have not yet been released.

**Table 9**
Reported X-ray crystallography methods in PDB depositions (as of 31 December 2003).

Abbreviations are listed in http://asdp.bnl.gov/asda/Libraries/xmethods/.

| Methods | Structures | % |
|---|---|---|
| Molecular replacement | 7548 | 59.2 |
| Experimental phasing | 2956 | 23.2 |
| Miscellaneous | 2128 | 16.7 |
| Direct methods | 119 | 0.9 |
| | | |
| Experimental phasing | | |
| Single | 2445 | |
| MIR | 757 | 31.0 |
| SIR | 248 | 10.1 |
| MAD | 1276 | 52.2 |
| SAD | 164 | 6.7 |
| Combinations | 511 | |
| MIRAS | 206 | 40.3 |
| SIRAS | 165 | 32.3 |
| MADIR | 46 | 9.0 |
| MR + AS | 26 | 5.1 |
| MR + IR | 59 | 11.6 |
| MR + AS + IR | 9 | 1.8 |
| | | |
| Miscellaneous | | |
| Fourier | 1174 | 55.2 |
| Refine | 158 | 7.4 |
| Model | 94 | 4.4 |
| Other | 673 | 31.6 |
| Miscellaneous | 29 | 1.4 |

collect information about the X-ray source. Some depositors might have misquoted the beamline IDs, and some errors exist in the primary PDB database. There may be a couple of beamlines that are in operation but do not appear in these statistics because they have not yet made a PDB deposition. We will add these or new beamlines once the PDB deposition is made.

Non-standard beamline IDs and the evolution of beamline names make it difficult to extract information from the primary PDB database. We have been communicating with the scientists at synchrotron facility sites to verify the deposited information on beamlines and we have made necessary corrections if there are errors. PXLIB will provide more accurate information on synchrotrons than the primary PDB database. We suggest that the synchrotron source community should have a standardization in defining beamline names. The users of synchrotron sources should be reminded to provide accurate beamline information to the PDB when they make a deposition.

[*However, the validation of synchrotron beamline information has involved checks and cross-checks with the synchrotron radiation facilities themselves made by ourselves and include the further suggestions of the referees to whom we are grateful. If any errors remain we welcome corrections that will, if necessary, lead to an updated analysis that will be submitted for publication to the* Journal of Synchrotron Radiation *at a later date. As part of this process, input from facilities on these matters should be addressed to the corresponding author, jiang@bnl.gov.*]

### 4.2. Impact on the PDB of SG and high-throughput methods

The SG approach differs from traditional structural biology by its organized 'pipeline'. The organized 'pipeline' includes all procedures from protein target selection, cloning and expression, crystallization, data collection, structure determination, and function research. The goal of SG is not only to determine the structures of proteins systematically, but also to develop high-throughput and automation methods that can be used in a production line. All of the SG centers are closely related to at least one synchrotron facility (Table 1). Several synchrotron beamlines have been supported by SG and reconstructed with a crystal-mounting mechanism ('robot') for high throughput. According to the report from PSI (Annual PSI Meeting 2003, NIGMS, Bethesda, MD), the nine PSI SG centers have determined 537 structures and deposited 335 structures to the PDB since funding was started in 2000 (Table 1, personal communications). Although we did not observe a big jump in the total numbers of PDB depositions since the start of SG in 2000, several beamlines participating in SG projects have seen a burst in PDB depositions. For example, 19ID at APS (MWSG) had more than 100 depositions in 2002 and 2003. Not only will the SG approach produce more structures, but also new technologies will speed up the whole process of structure determination. For example, the lag time on beamline 31ID/APS (SGX) is only three months and on X06SA at SLS is only six months. The average lag time (not shown) for selective beamlines that involve SG centers (as listed in Table 1) is less than 12 months compared with 18 months over all beamlines. More importantly, the SG approach produces more 'unique' structures (less than 30% sequence identity to the others) and more 'new' protein folds, because the new approach scans the whole genomes of all kinds in order to explore 'new' structures. Experimental phasing methods, particularly selenomethionine MAD/SAD phasing methods, are the major methods for SG because homologue models are not available for molecular replacement.

The author is grateful to Lonny Berman for contributions to this paper. The authors acknowledge support from NCRR/NIH grant P41-RR12408 (PI: R. M. Sweet) and NIGMS/NIH grant P50-GM62529 (PI: S. K. Burley). We would like to thank Lonny Berman, Mike Becker and Mark Chance for the consensus on the 'equal fractional credit' in counting multiple beamlines. Thanks also to Nancy Manning for the clarification on the 'REMARK' records, and Zheng Lin for help on the mirrored PDB database.

### References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, N., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucl. Acids Res.* **28**, 235–242.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.
Brunzelle, J. S., Shafaee, P., Yang, X., Weigand, S., Ren, Z. & Anderson, W. F. (2003). *Acta Cryst.* D**59**, 1138–1144.
Chance, M. R., Bresnick, A. R., Burley, S. K., Jiang, J. S., Lima, C. D., Sali, A., Almo, S. C., Bonanno, J. B., Buglino, J. A., Boulton, S., Chen, H., Eswar, N., He, G., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M. & Wang, L. K. (2002). *Protein Sci.* **11**, 723–738.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Helliwell, J. R. (1998). *Nature Struct. Biol. Suppl.* **5**, 614–617.
Holton, J. & Alber, T. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1537–1542.
Jiang, J.-S. (2004). *Acta Cryst.* D**60**. To be submitted.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
Martz, E. (2002). *Trends Biochem. Sci.* **27**, 107–109. (http://www.proteinexplorer.org/.)
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Ogata, C. M. (1998). *Nature Struct. Biol. Suppl.* **5**, 638–640.
Protein Data Bank Contents Guide (1996). *Protein Data Bank Contents Guide: PDB Format Description*, Version 2.1, 25 October 1996, http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Sussman, J. L., Lin, D., Jiang, J.-S., Manning, N. O., Prilusky, J., Ritter, O. & Abola, E. E. (1998). *Acta Cryst.* D**54**, 1078–1084.