Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Tapping the Protein Data Bank for crystallization information

Thomas S. Peat,[a]* Jon A. Christopher[a] and Janet Newman[b]

[a]OpenEye Scientific Software, 3600 Cerrillos Road, Suite 1107, Santa Fe, NM 87507, USA, and [b]In Stilla Consulting, 736 Arden Drive, Encinitas, CA 92024, USA

Correspondence e-mail: tom@eyesopen.com

Received 15 August 2005
Accepted 14 October 2005

A database application has been developed for the collection of crystallographic information. This database (the BDP) has been populated with the information found in the Protein Data Bank (PDB). The tool has been used to store crystallization data parsed out of the PDB and these data may be used to extend the crystallization information found in the Biological Macromolecule Crystallization Database (BMCD) and could be used to refine crystallization methodology. A standard is proposed for describing a crystallization experiment that will ease future crystallization data collations and analyses.

1. Introduction

Macromolecular crystallization is a poorly understood part of the process of structure determination and yet has been used in the solution of more than 80% of the 33 000+ structures available in the PDB (http://www.rcsb.org; Berman *et al.*, 2000). Much of the current knowledge is empiric rather than based on a fundamental understanding of the process of crystallization. Sparse-matrix crystallization screens based on knowledge of successful experiments were introduced to the general community in 1991 (Jancarik & Kim, 1991) and a number of screens (*e.g.* the Wizard Screens from Emerald BioSystems and the Crystal Screens from Hampton Research) have used this approach since. Prior to this, the incomplete factorial method of generating conditions had also been used for successful crystallization experiments (Abergel *et al.*, 1991; Carter & Carter, 1979). Programs such as *CRYSTOOL* (Segelke, 2001) rely on having a reasonable set of reagents in order to generate on-the-fly novel crystallization screens. For many years, the Biological Macromolecule Crystallization Database (BMCD; http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html; Gilliland *et al.*, 1994) provided some of this information and has been the basis of software tools to generate crystallization strategies (Hennessy *et al.*, 2000). However, the BMCD has not been updated since 1997, probably as the information in the database was hand-curated and it became too onerous a task to continue.

At about the same time that the BMCD stopped growing, the PDB introduced a new standard (Berman *et al.*, 2000) which included 'REMARK 280' specifically for capturing crystallization information. In theory, any structure deposited in the PDB from that time could have included crystallization details. In reality, less than half of the structures do include crystallization information and an even more modest percentage include useful information (Fig. 1).

doi:10.1107/S0907444905033202

Currently, the Protein Data Bank is a repository of information, as the name implies, rather than being a true relational database; however, there has been some effort to change this (Deshpande *et al.*, 2005). Much of the power of a relational database lies in the ability to determine relationships between data without prior knowledge of what those relationships might be. For this strength to be exploited, the integrity of the data inserted into the database is crucial. Mostly, this is achieved by formalizing data types, so that, for example, a date is consistently recorded in a valid date format. This narrows the ways in which data can be misinterpreted.

We designed a database (the BDP) for capturing macromolecular structure information to be used as an internal resource and have tested it by loading the publicly available data from the PDB into it. During this process, we learned a lot about the ways in which the integrity of these public data could be improved. Almost incidentally, we have collated a large (8000+) collection of crystallization conditions and have started to analyse these. Obtaining these records has required a significant scripting effort, mostly to overcome 'trivial' problems such as misspelling or alternate spellings of the same chemical. Even though extensive scripting tools were used, significant manual intervention was required to clarify specific problems.

2. Methods

The BDP is built using a generic script, allowing it to be database-agnostic. Our production version uses a PostgreSQL environment running under various flavours of Linux. Data are uploaded into the database using a Python script from standard PDB files. New information is downloaded from the PDB and uploaded into the BDP daily. The data captured include general information (including depositor, date *etc.*), details about the macromolecule (such as name, species, class, amino-acid sequence *etc.*), the atomic coordinates (x, y, z, q, B), the data-collection (wavelength, completeness, R, $R_{\rm free}$ *etc.*) and structure-solution details (beamline used, temperature) as well as the crystallization conditions. Heteroatoms are parsed out and validated by OEChem, a chemical library developed by OpenEye, which is the same tool currently used by PubChem to parse chemical information (http://pubchem.ncbi.nlm.nih.gov).

The crystallization data extraction first determines whether a REMARK 280 exists and, if it does, whether it can be appropriately processed; that is, if it contains a string with numbers and text which might look like a concentration, a unit and a chemical name. After that, string comparison is performed to assign a valid chemical name. This is repeated for each potential chemical identity. Deconvolution of the potential misspellings is performed using regular expressions. Data mining was performed using standard SQL queries using a desktop PC with SuSE Linux as the operating system and PostgreSQL as the database-management system.

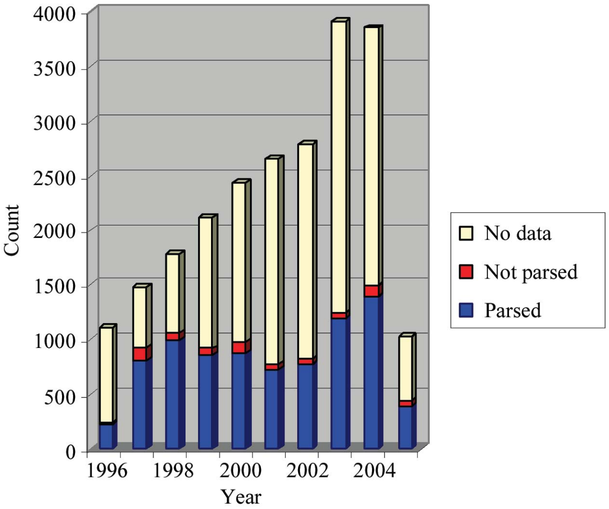


Figure 1
Graph showing the number of structures deposited each year since 1997. The computer script requires that there be a concentration (a number) and unit (text string) for each chemical species in the REMARK 280 field. The majority of the cases where data could not be automatically extracted had one or more chemical species with no corresponding concentration or unit value. The red portion of each bar shows those structures with a REMARK 280 which fulfilled the 'concentration, unit, species' requirement but were not parsed by the script. The blue portion shows the proportion of structures that were parsed by the script.

3. Results

As of 29 July 2005, the BDP contained 32 307 structures, of which 27 059 were solved using X-ray crystallography. Of these, 23 491 contain a REMARK 280 and 8289 contain interpretable crystallization information. Of the entries that did not yield any crystallization information, about 24% contain 'NULL' in the REMARK 280 field. Another 10% completely lack any numbers and contain only the names of the chemicals used in the experiment. In over 60% of cases the information is not complete (some chemicals have quantities and/or units and some do not) and the script simply fails in about 4.4% of the cases. In the process of checking the literature for missing values or dubious chemical names, we often found inconsistencies with the PDB crystallization records. In all cases we took the data published in the references given to be correct. In those cases where extraordinary claims were made in the PDB files and these could not be validated in the literature (*e.g.* crystallization at 100 K or at 373°C), we contacted the authors directly. In all cases so far, the anomalies were typing errors introduced by the authors during the process of entering the information into the PDB. These kinds of mistakes could be easily rectified in the future if the PDB enforced data typing and required reasonable values to be entered (*e.g.* there is no such date as '30 February').

Lack of data aside, the most egregious problem was the non-standard nomenclature of the chemicals used in the crystallization experiment. The extent of the spelling problem can be appreciated by looking at a collection of synonyms for ammonium sulfate (Fig. 2). Although this lists only some of the alternative representations of 'ammonium sulfate', it

Table 1
A list of the most commonly used chemicals found in the BDP from the crystallization data parsed out of the Protein Data Bank.

There are 38 chemical entities found over 100 times in the crystallization records when listed by frequency, with ammonium sulfate being the most commonly used single chemical for crystallization. As one might expect, buffers and additives such as dithiothreitol (DTT) are also found frequently in crystallization experiments. Summing the various sizes of polyethylene glycol (PEG) in this table gives a count of 3898 occurrences, which is almost double the frequency of ammonium sulfate; in fact, adding in some of the less commonly used PEGs brings PEG usage to over double that of ammonium sulfate.

Frequency of use	Chemical
2070	Ammonium sulfate
1623	Tris/Tris chloride
1271	PEG 4000
1118	HEPES/sodium HEPES
1116	Acetate/sodium acetate
934	PEG 8000
864	Citrate/sodium citrate
849	Sodium chloride
641	MES/sodium MES
626	Cacodylate/sodium cacodylate
600	Magnesium chloride
522	PEG 3350
501	DTT
478	Glycerol
474	MPD
466	PEG 6000
401	Calcium chloride
364	PEG 400
351	2-Propanol
340	Lithium sulfate
319	Ammonium acetate
295	Phosphate/sodium phosphate
283	Potassium phosphate
224	Sodium azide
187	Magnesium acetate
183	EDTA
177	Potassium chloride
171	PEG MME 2000
170	PEG MME 5000
168	Sodium potassium phosphate
148	Calcium acetate
146	Ethylene glycol
139	β-Mercaptoethanol
137	Formate/sodium formate
124	Bis-Tris
117	Ethanol
112	Imidazole
106	Ammonium phosphate

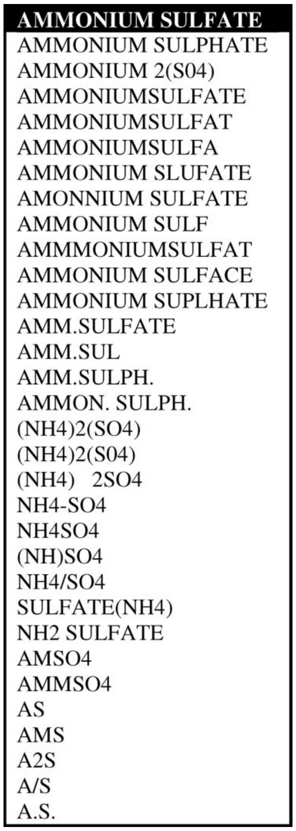
provides a good example of the variety of the ways a single chemical entity can be represented.

There are 793 distinct chemicals found from these records, of which 415 are singletons; that is, they are found in only one crystallization condition. 38 chemicals are found in more than 100 crystallizations and Table 1 lists the 20 most popular chemicals. The most common chemical is ammonium sulfate (used 2070 times), mirroring what was seen in the BMCD, where it is the most abundant 'chemical additive' and was used in the crystallization of 497 molecules. However, if one adds together all of the PEG conditions seen in the BDP, PEG is found more than twice as often as ammonium sulfate (4652 occurrences). In the BMCD, PEGs are used to crystallize 599 macromolecules, which is approximately the same as the number for ammonium sulfate. This suggests that the relative popularity of PEG as crystallization agent has more than doubled since 1997. The emergence of sodium malonate as a crystallization reagent is reinforced by looking at the two collections of crystallization chemicals. The BMCD contains no chemical additive 'sodium malonate' (or malonate or malonic acid), whereas the BDP contains 38 entries which contain 'malonate'. This is scarcely surprising, as the paper which popularized this chemical was published in 2001 (McPherson, 2001).

Quoting from Gilliland *et al.* (1994),

> The primary motivation for creating the BMCD was to develop crystallization strategies . . .

and this was also one of the aims behind the parsing effort for the PDB data. The more extensive collection of crystallization conditions culled from the PDB should allow for a more accurate view into current successful crystallization experiments. As an initial foray into these data, we have looked at the concentration ranges of PEG found in the BDP: the results are shown in Fig. 3. Analysis of the commonly used PEGs shows that the small (liquid) PEGs show no clear preferred concentration: one may think of these PEGs as being used both as an additive and as a precipitant equally. With the medium-sized PEGs (molecular weights of 1000–8000),



AMMONIUM SULFATE
AMMONIUM SULPHATE
AMMONIUM 2(S04)
AMMONIUMSULFATE
AMMONIUMSULFAT
AMMONIUMSULFA
AMMONIUM SLUFATE
AMONNIUM SULFATE
AMMONIUM SULF
AMMMONIUMSULFAT
AMMONIUM SULFACE
AMMONIUM SUPLHATE
AMM.SULFATE
AMM.SUL
AMM.SULPH.
AMMON. SULPH.
(NH4)2(SO4)
(NH4)2(S04)
(NH4)   2SO4
NH4-SO4
NH4SO4
(NH)SO4
NH4/SO4
SULFATE(NH4)
NH2 SULFATE
AMSO4
AMMSO4
AS
AMS
A2S
A/S
A.S.

Figure 2
A representative sample of ammonium sulfate spelling alternatives from the Protein Data Bank. Some forms are clear misspellings, some are valid alternative spellings and some are more insidious (such as substituting a zero for an 'O'). Many are abbreviations that may or may not be commonly used by crystallographers and crystallizers around the world.

concentrations of 20–25% are frequently used for crystallization. In general it appears that the higher the molecular weight of the PEG, the tighter the distribution of its 'successful concentration'.

As the concentration of PEG found in the optimized condition will probably depend to some extent on the concentrations of that PEG found in the initial screens, it is interesting to compare the two (Fig. 4). This comparison is skewed, as not every protein is tested against all the commercial screens and usually the results of only a small number (often one) of positive screening experiments are used as the starting point for optimization. Furthermore, the set of available screens is redundant: the Jancarik and Kim screen is available from three different vendors (Crystal Screen from Hampton Research, Structure Screen 1 from Molecular Dimensions and The Classics Suite from Nextal). However, it appears that there is good agreement between the concentration of PEG 4000 found in the commercial screens and in optimized experiments, that the commercial screens could perhaps use PEG 8000 at slightly lower concentrations and that the tendency of the commercial screens to use PEG 400 at 30–35% is not mirrored at all in the data from the successful conditions that contain PEG 400. It could be argued that the role of a screen is to produce likely hits for further optimization, in which case the concentration of PEG 8K found in the screens is 'correct' in that it is a little higher than that found in the final refined conditions. If one continued with the same logic, the concentration of PEG 4K in the screens should be increased somewhat in order to push out more hits for optimization.

A molecular weight was calculated for each entry: this calculated molecular weight (cMW) is the sum of the atoms (excluding water atoms) for which there were coordinates in the PDB file. It must be emphasized that this molecular weight is for all of the atoms in the asymmetric unit (excluding water) and does not correspond to the molecular weight of a 'protein'. It is the molecular weight of the 'crystallographic unit', as the molecular weight is not given in the PDB record and there was no obvious way to automatically determine from the files what the true 'biological unit' might be or whether this was the relevant unit for crystallization queries. However, with these data we can extend our analysis of crystallization trends by looking at the relationship between the cMW and the concentration of crystallization chemical. Fig. 5 shows a comparison between the concentration of precipitant and the cMW for two common crystallization chemicals, ammonium sulfate and PEG 8K. From the graphs, it appears that larger cMW species require less precipitant for

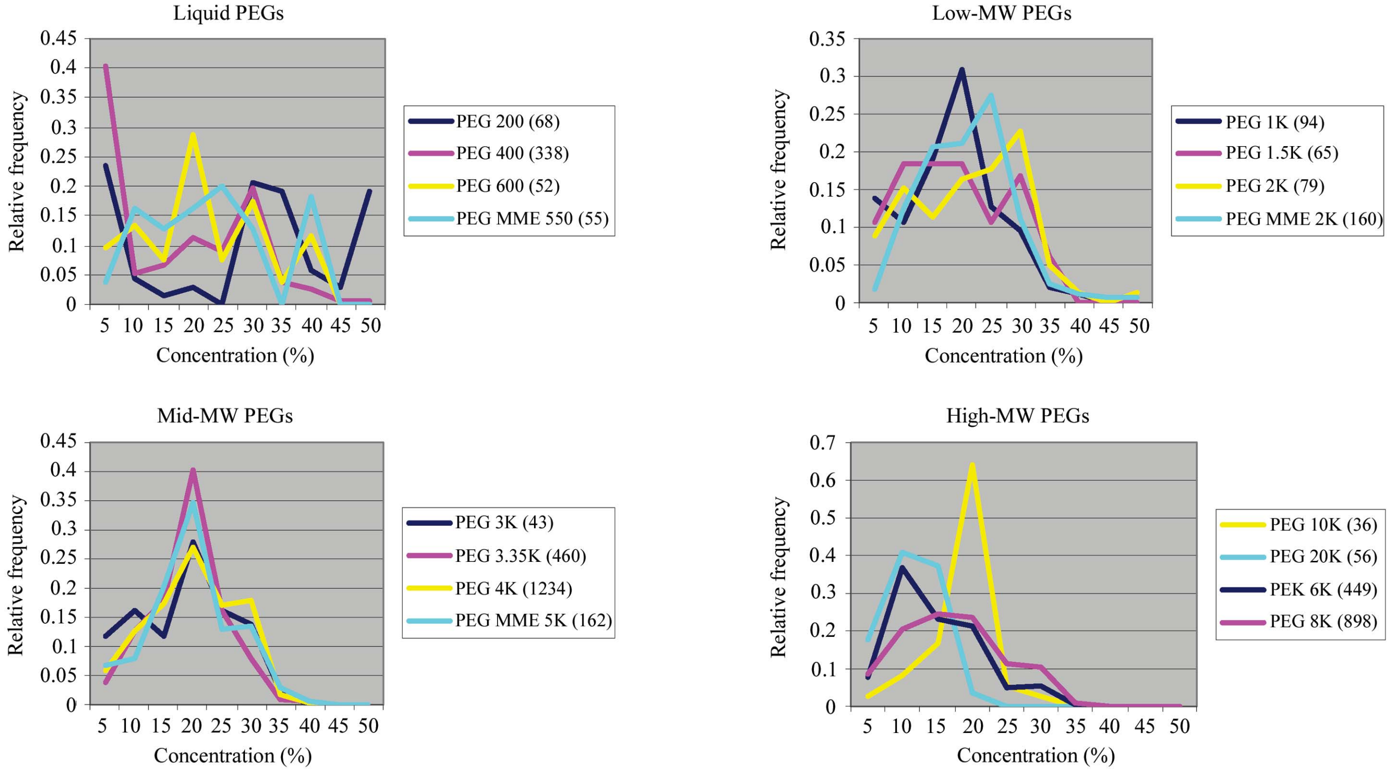


Figure 3

Charts of PEG concentrations *versus* relative frequency for the various sized PEGs. The PEG concentrations have been grouped into 5% ranges and each bin is described by the upper value of the concentration range, thus the bin '0–5%' is labeled '5%' on the graph. The 'relative frequency' is the count of the number of occurrences of a PEG at a given concentration divided by the total number of conditions containing that PEG. Only those PEGs that are found at least 30 times are shown; the number of usages is recorded in parentheses after the PEG name. The assignment of the different PEGs to four classes (Liquid, Low MW, Mid MW and High MW) was mostly for clarity of the graphical representation, but seems to capture groups of PEGs that behave somewhat similarly. There is a clear trend for the larger PEGs to be used at a specific concentration, whereas the liquid PEGs have no clear usage maximum and appear to be used both as additives and as precipitating agents (*i.e.* used both at low concentrations and at high concentrations).

crystallization. We compared two populations, small molecules (using 0–25 000 Da) and large molecules (all species above 100 000 Da), and performed a standard z-test on their average precipitant concentration. The average concentration of ammonium sulfate used for the small-molecule population is 1.99 *M* with a standard deviation of 0.80 *M* (n = 535); the average for the large molecules is 1.59 *M* with a standard deviation of 0.56 *M* (n = 208). The z value given these data is 7.69 (p > 0.0001); in other words, the difference in the average precipitant concentration is highly statistically significant. The same was performed for the PEG 8K data [the average for the low-cMW molecules is 19.8% with a standard deviation of 8.35% (n = 199); the average for the high-cMW molecules is 13.43% with a standard deviation of 6.69% (n = 149)] and this gave a z value of 7.89 (again, p > 0.0001).

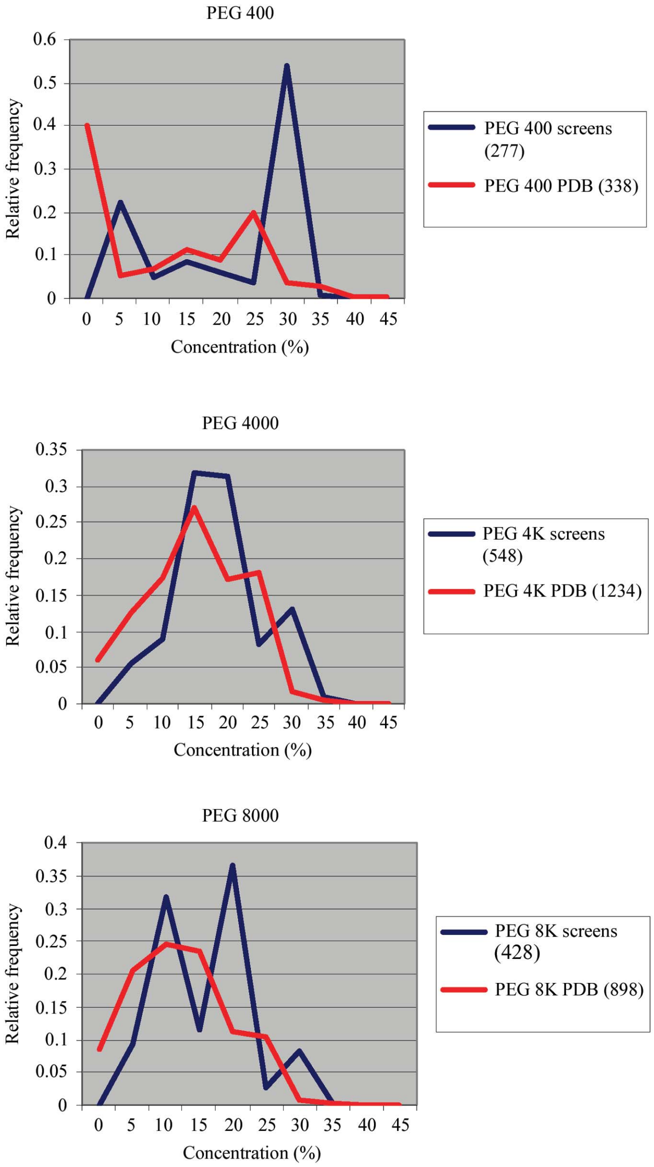


Figure 4
Comparison of the concentration ranges found in 88 commercial screens compared with the PDB for three common PEGs. The PEG concentrations have been grouped into 5% ranges and each bin is described by the upper value of the concentration range, thus the bin '0–5%' is labeled '5%' on the graph. The 'relative frequency' is the count of the number of occurrences of a PEG at a given concentration divided by the total number of conditions containing that PEG. The screens include those from Emerald BioSystems, Hampton Research, Jena BioSciences, Molecular Dimensions and Nextal. The number of times each PEG is found is given in parentheses.

4. Discussion

4.1. Is reporting increasing?

The crystallographic community seemed to be more interested in adding crystallization information when REMARK 280 was initially introduced: from Fig. 1, about 60% of the structures deposited in the years 1997–1998 contained crystallization information, whereas that number seems to have tapered off to 30% in later years. Indeed, while the number of deposited structures is growing rapidly, the number of depositions with crystallization information has stayed essentially the same. There seems to be no improvement in the quality of the data deposited: the number of PDB entries which did not meet the minimum requirement for data parsing (a concentration and unit for each chemical species) is also steady. This argues that the current method of capturing crystallization information needs some revision, as the rate of compliance should increase with greater familiarity with the

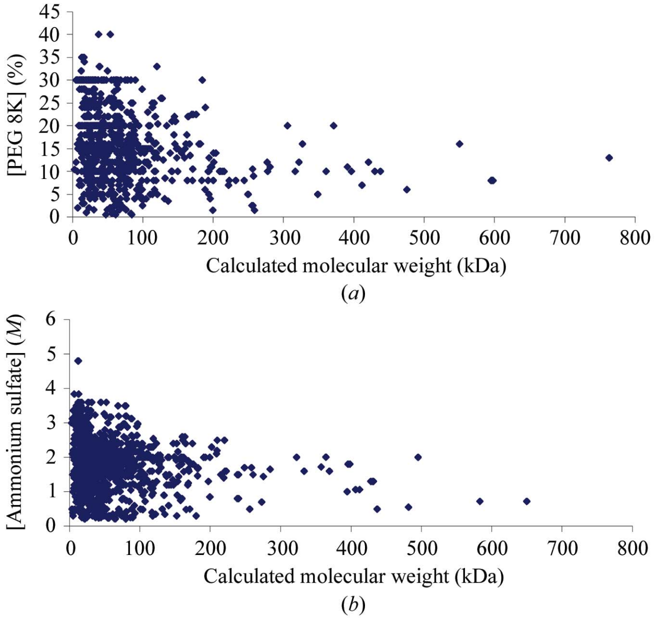


Figure 5
Calculated molecular weight (cMW) *versus* concentration for PEG 8K (*a*) and ammonium sulfate (*b*). The calculated molecular weight is the molecular weight of all of the atoms in the asymmetric unit after removing the waters. Ammonium sulfate is used both as an additive and as a precipitant and it is unclear where it stops being an additive and starts being a precipitant. The ammonium sulfate chart only shows data where ammonium sulfate concentrations of greater than 0.2 *M* were included. The cutoff of 0.2 *M* was chosen as the commercial screens tend to classify 0.2 *M* or less as a 'salt' or 'additive' and higher concentrations as 'precipitant'. When ammonium sulfate was found as a '%' unit, we calculated the equivalent molarity using 4.8 *M* as the saturation point at room temperature (*i.e.* 100%). There was no correspondingly obvious cutoff for PEG 8K used as a precipitant.

deposition tools. One possible way to improve full reporting of crystallization information is to give examples in the deposition tools that include all of the important details (amounts, units and chemical names).

Over 2000 lines of regular expressions are used to clean up the conditions: most of these deal with misspellings and alternative spellings of the chemical names (an example is shown in Fig. 2). The current list of distinct chemicals contains some ambiguous terms such as ‘citrate’. Citrate can be used in the acid form or can be purchased with a number of non-hydrogen counter ions (sodium, potassium and ammonium are common). We have grouped somewhat ambiguous terms with the sodium salt of the anion in Table 1 for chemicals that are most often considered as buffers (Tris, HEPES, MES *etc.*), as from experience these are the forms most often found in the experiment and the conjugate acids are often titrated to the correct pH by sodium hydroxide. In the BDP these are considered separate entities for data-mining purposes. Table 1 also shows us that six of the ten most popular chemicals in the database are chemicals used most often as buffers and that phosphate is not found as often as most of the other buffers.

The data that we have extracted comes from the PDB and cannot contain more ‘raw’ information than was in the original source. The difficulty in obtaining a molecular weight in order to perform the MW/concentration analysis is a case in point. The cMW that we used does not capture information about the oligomeric state of the protein nor the non-crystallographic symmetry of the crystals, so that the trends seen (Fig. 5) are a guide at best. Furthermore, for this analysis we had to make a guess at the conversion from ‘%’ to ‘M’ for the ammonium sulfate, as the units had to be consistent throughout. Our inability to compensate for missing information puts a limit on the information that can be extracted from the BDP. Although we feel confident that there are enough data points for the statistics we provide, having >23 000 conditions instead of ∼8300 conditions would have made the numbers much more convincing. The question ‘do very high resolution structures come from crystals grown in salt or in PEG?’ would require that we have a way of determining when a crystallization condition is a ‘salt’ condition and when it is a ‘PEG’ condition (many are both).

Although the data used in this study were not collated from the PDB completely automatically, the use of parsing scripts made the process significantly easier than it otherwise would have been and indeed made the whole process possible. The error checking and data analysis were also made easier by using scripts instead of performing such tasks manually. In addition, by having a single person run the scripts, checking for errors and contacting authors, the data were parsed in a more consistent manner than they would have been by a group of individuals. Although there are almost certainly still some errors in the BDP introduced by the script, even a hand-curated database like the BMCD has anomalies; for instance, there are two occurances of ‘see comments’ as a chemical additive.

The major barrier to this work was the lack of data standardization. Lack of a standard crystallization format also inhibits the exchange of crystallization information within the community. For example, it is often unclear whether the information given describes what is in the reservoir or in the drop. As ‘protein’ is found in over 800 records, this might indicate that in these cases the contents of the drop were being described.

Any standard format for recording crystallization conditions needs to be able to capture information about the protein solution, the crystallization solution (crystallant) and how the two were combined. An ideal implementation would require users to enter numbers and text in separate fields, with limited lists of choices: for example, the ‘units’ field would be chosen from a list such as ‘mg ml^{-1}’, ‘M’, ‘mM’, ‘μM’, ‘%(w/v)’ or ‘%(v/v)’. Other fields should have instantaneous checking for data type. Additionally, a set of standardized names for the chemicals used for crystallization would help immensely. The standard name of a chemical should not include waters of hydration (as crystallization deals with the chemicals in aqueous solution) and should include all counter ions: ‘sodium dihydrogen phosphate’ rather than just ‘sodium phosphate’ or worse ‘phosphate’. Although this does not capture what is in solution (for example, a buffered solution of sodium dihydrogen phosphate will contain a mix of $H_2PO_4^-$ and HPO_4^{2-}), it allows the solution to be reproduced. A commonly used quality check is the pH of the complete crystallant solution and so a field for the resultant pH is included. Similarly, the ratio of drop volume to reservoir volume is critical in many crystallization experiments and should be reported. As there are countless ways of tweaking a crystallization experiment, it is suggested that a free-format comment field be included in the format for capturing important but hard to categorize information.

Proposed format

Crystallant

Amount	unit	chemical	pH
Amount	unit	chemical	pH
…			

Protein

Amount	unit	chemical	pH
Amount	unit	chemical	pH
…			

Drop details
drop component volumes (protein:crystallant:other)
temperature
method
crystallant pH
reservoir volume

Comments
Use this space to describe details

To illustrate, we take an example from a recent paper in *Cell* (Dürr *et al.*, 2005). This paper discusses the structure of the

ATP core protein Rad54 and a Rad54–DNA complex. The crystallization of SsoRad54cd is well described:

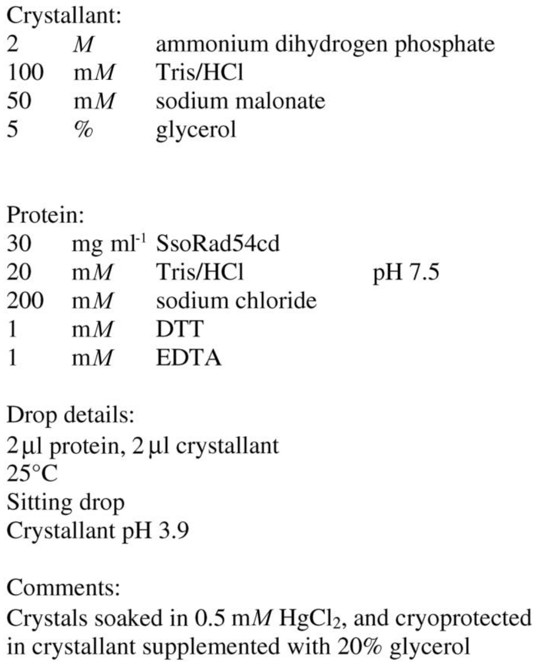
> We crystallized SsoRad54cd in space group $P4_12_12$ with one molecule per asymmetric unit by sitting drop vapor diffusion at 25°C after mixing 2 µl protein (30 mg/ml in 20 mM Tris/HCl [pH 7.5], 200 mM NaCl, 1 mM DTT, 1 mM EDTA) and 2 µl precipitant (2 M $(NH_4)H_2PO_4$, 100 mM Tris/HCl [pH 3.9], 50 mM sodium malonate, 5% glycerol). The crystals were soaked with 0.5 mM $HgCl_2$ for 2 hr, transferred into precipitant supplemented with 20% glycerol for 10 min, and flash frozen in liquid nitrogen.

The corresponding PDB entry (1z6a) is not so comprehensive:

```
REMARK 280 SOLVENT CONTENT, VS   (%): NULL
REMARK 280 MATTHEWS COEFFICIENT, VM (ANGSTROMS**3/DA): NULL
REMARK 280
REMARK 280 CRYSTALLIZATION CONDITIONS: AMMONIUM DIHYDROGEN PHOSPHATE,
REMARK 280 TRIS/HCL, NA-MALONATE, GLYCEROL, PH 3.9, VAPOR DIFFUSION,
REMARK 280 HANGING DROP, TEMPERATURE 323K
```

There are no amounts associated with the chemicals in this PDB entry, so this is an entry that could not have been parsed into the BDP. Furthermore, the temperature of the crystallization experiment is given as 323 K, which is 50°C, rather than the 25°C given in the original paper. Missing from both descriptions is the volume of the crystallant in the reservoir for this vapor-diffusion experiment.

Using the proposed format this entry would become



Crystallant:

2	*M*	ammonium dihydrogen phosphate
100	m*M*	Tris/HCl
50	m*M*	sodium malonate
5	%	glycerol

Protein:

30	mg ml^{-1}	SsoRad54cd	
20	m*M*	Tris/HCl	pH 7.5
200	m*M*	sodium chloride	
1	m*M*	DTT	
1	m*M*	EDTA	

Drop details:
2 µl protein, 2 µl crystallant
25°C
Sitting drop
Crystallant pH 3.9

Comments:
Crystals soaked in 0.5 m*M* $HgCl_2$, and cryoprotected
in crystallant supplemented with 20% glycerol

This captures many more of the details and is easy to parse, but there are still some problems. Firstly, '%' is ambiguous: it can mean weight (mass) percent, mole percent or volume percent. In the example given glycerol is a liquid and it is probably the latter. As most of the native liquids used in crystallization have densities in the range 0.785 g ml^{-1} (2-propanol) to 1.62 g ml^{-1} (1,1,1,3,3,3-hexafluoro-2-propanol), there is not too much uncertainty introduced by the looseness of the '%' definition. The pH is also badly defined. Often it is used to describe the pH of a component (the buffer) of a crystallization solution, but it is additionally used to describe the pH of the resulting solution itself. In our example, it appears that 'pH' is used to describe the pH of the Tris buffer (a component) in the 'protein' and is used to describe the resultant pH in the 'crystallant'. This of course is assuming that the investigator would not be using a stock solution of Tris–HCl (which has a pK_a of 8.1 at 298 K) set to pH 3.9. There is no volume given for the reservoir solution for this vapor-diffusion experiment, which may preclude other scientists reproducing these results.

It is hard to envisage a data format which is complete and unambiguous enough to enable a completely naïve reader to reliably reproduce the experiment. The pH example shown above shows a weakness that is probably quite common. Another source of confusion might be how one makes up the stocks used in a crystallization experiment. Does a 50% PEG 8000 stock consist of 50 g PEG 8000 dissolved in water to a final volume of 100 ml (50 mass/volume percent), 50 g PEG 8000 dissolved in 50 g water (50 mass percent) or 50 g PEG 8000 dissolved in 100 ml water?

If one tried to create a format that captured all salient information, the chances are that there would be almost 100% non-compliance in using it, as the detail required would ensure that adhering to the format necessitated prohibitive amounts of work. The format above tries to find a happy compromise between being complete and being easy to use.

5. Conclusions

We have built a relational database for capturing structural information that includes the crystallization conditions used in the structure solution. This database, the BDP, has been populated with data from the PDB. These data are now in a form that allows them to be used in further analyses. We have provided a few examples of some of the queries that might be performed and pointed out that more data would be helpful when looking for trends. Although there is a 'front end' to the database, most of the data mining for this study was performed directly using SQL. The drawback of this is that a user has to have some familiarity with SQL in order to access the data; the advantage is that there is no rigid GUI to dictate what information can be gleaned from the data.

Although the primary aim of this project was to collect all structural information into a database, there are some lessons we have learned about crystallization. We found that there have been changes in crystallization conditions since 1997. PEG conditions have become more prevalent, almost doubling in relative popularity. Malonate is now found in a significant number of successful crystallizations, although its recent entry into the field of crystallization chemicals precludes it from being in the top ten. It is, however, found in the top 100 chemicals. The higher molecular-weight PEGs tend to be found in relatively narrow distributions, whereas the liquid PEGs are found at all concentrations up to about

25%. The commercial screens perform a reasonable job of mirroring the optimized conditions from the PDB, but they could probably be tweaked to give somewhat better results. An example of this is seen in Fig. 4, which shows that the PEG 400 concentrations found in the commercial screens could be more usefully spread out across the concentration range 5–25%. The 'common lore' that states that large molecules require less precipitant in order to crystallize is supported by the data, at least for PEG 8K and ammonium sulfate.

The PDB is the primary resource for deposition of structural information in the biological community and is an obvious choice for the collection of crystallization information as well. The RCSB has done a tremendous job improving the web site(s) and making the data stored easier to access and has brought together many disparate sources of information for ease of use. Although the PDB has improved, the system in place for crystallization information is plagued with user non-compliance and data-typing flaws. We suggest a format for recording crystallization experiments, which although not ideal, is a step in the right direction to record this important information for the crystallographic community. The proposed mmCIF format for capturing crystallization data encompasses many of the features we propose here and will be a benefit to the community once it is fully implemented. However, without enforced data typing in the implementation, the value of this format will never be fully realised.

We would like to acknowledge Chris Boarman of Rigaku for his help with finding the concentration information from the commercial screens. We would also like to thank Anthony Nicholls for suggesting and supporting this project, the OEChem development team for OEChem, numerous authors for responding to e-mail requests for more information and to the PostgreSQL and Python communities for producing admirable free software for general use.

References

Abergel, C., Moulard, M., Moreau, H., Loret, E., Cambillau, C. & Fontecilla-Camps, J. C. (1991). *J. Biol. Chem.* **266**, 20131–20138.

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.

Carter, C. W. Jr & Carter, C. W. (1979). *J. Biol. Chem.* **254**, 12219–12223.

Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., Green, R. K., Flippen-Anderson, J. L., Westbrook, J., Berman, H. M. & Bourne, P. E. (2005). *Nucleic Acids Res.* **33**, D233–D237.

Dürr, H., Körner, C., Müller, M., Hickmann, V. & Hopfner, K. P. (2005). *Cell*, **121**, 363–373.

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

Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P. A. & Rosenberg, J. M. (2000). *Acta Cryst.* D**56**, 817–827.

Jancarik, J. & Kim, S.-H. (1991). *J. Appl. Cryst.* **24**, 409–411.

McPherson, A. (2001). *Protein Sci.* **10**, 418–422.

Segelke, B. W. (2001). *J. Cryst. Growth*, **232**, 553–562.