Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The Biomolecular Crystallization Database Version 4: expanded content and new features

Michael Tung and D. Travis Gallagher*

Center for Advanced Research in Biotechnology/
National Institute of Standards and Technology,
9600 Gudelsky Drive, Rockville,
Maryland 20850, USA

Correspondence e-mail:
gallaghe@umbi.umd.edu

Received 13 August 2008
Accepted 29 October 2008

The Biological Macromolecular Crystallization Database (BMCD) has been a publicly available resource since 1988, providing a curated archive of information on crystal growth for proteins and other biological macromolecules. The BMCD content has recently been expanded to include 14 372 crystal entries. The resource continues to be freely available at http://xpdb.nist.gov:8060/BMCD4. In addition, the software has been adapted to support the Java-based Lucene query language, enabling detailed searching over specific parameters, and explicit search of parameter ranges is offered for five numeric variables. Extensive tools have been developed for import and handling of data from the RCSB Protein Data Bank. The updated BMCD is called version 4.02 or BMCD4. BMCD4 entries have been expanded to include macromolecule sequence, enabling more elaborate analysis of relations among protein properties, crystal-growth conditions and the geometric and diffraction properties of the crystals. The BMCD version 4.02 contains greatly expanded content and enhanced search capabilities to facilitate scientific analysis and design of crystal-growth strategies.

1. Introduction

Because structural interactions among proteins and nucleic acids underlie biology, extensive efforts have been applied to the determination by X-ray crystallography of the structures of thousands of proteins from key organisms. These methods require high-quality crystals (diffracting to at least 3 Å resolution), so efficient methods of crystallizing proteins are actively being sought and have resulted in the accumulation of large amounts of crystal-growth data.

Proteins crystallize from pure (with respect to macromolecules) saturated solutions when additional chemical conditions meet two criteria. Firstly, the conditions must be appropriate to protect the protein from denaturation and other heterogeneity-inducing insults such as covalent modification of sensitive side chains. Secondly, out of all possible contact interactions among identical protein molecules, a single crystal-forming set must be selected. By subtle influences on protein hydration, ion binding and protonation, chemical conditions can alter the strengths of weak protein–protein interactions to make crystal-competent contact sets more thermodynamically and kinetically favored. In favorable cases a subset, usually small, of the thousands of possible sets of buffers, salts, alcohols, and other additives that induces growth of well diffracting crystals can be found. Current methods for crystallizing proteins rely primarily on sampling this infinite space of chemical compositions to identify a crystallogenic subspace (McPherson, 2001). Sampling usually begins with conditions similar to those that have worked previously for other proteins (Jancarik & Kim, 1991).

Even for a protein of known structure, it is very difficult to make predictions about its crystal conditions or crystal properties. For a novel protein of unknown structure, it is not possible to predict crystal geometry or crystallogenic conditions. However, some patterns have emerged from systematic analysis over large samples and with additional data on protein structures and crystallogenic conditions it is inevitable that predictions will improve. As data have accumulated, a literature of strategy suggestions has arisen. In theory, such knowledge-based strategies can enable focusing of searches so that, depending on protein properties, more probable regions of composition space can be searched. Several strategies and analyses have focused on particular classes of macromolecules, such as nucleic acids, whose distinct structure and charge give them characteristic crystal conditions [rich in 2-methyl-2,4-pentanediol (MPD) and divalent cations] and several specific commercial screens (Berger *et al.*, 1996). Subclasses of proteins are less distinctive, but several have been explored for trends in crystal conditions. For example, a commercial screen based on proprietary data analysis is available for kinases and an analysis of crystal conditions for protein–protein complexes has been published (Radaev *et al.*, 2006). This analytic approach depends on systematic correlation of protein properties (size, pI, solubility, function, structure, surface topography) with crystallogenic chemical composition. Scientific and commercial interest in this problem has led to the creation of several commercial databases and analytical tools focused on macromolecular crystallization.

Crystallization data typically include a list of chemicals, each at a specific concentration (including the concentration of the protein itself), a pH and a temperature. Since crystallogenesis is usually initiated by mixing a protein solution with a precipitant solution and then incubating this mixture in vapor contact with a reservoir solution within a sealed chamber, the chemical components, volumes and pH values of all three of these solutions (along with temperature) constitute the primary data. Beyond these primary data are additional variables, some dependent on the specific method used, such as the geometry and materials used to form and seal the chambers and the oil(s) used in batch methods. Secondary data also include crystal properties such as size, growth rate and habit, the important but difficult to define diffraction resolution of the crystals and the unit-cell parameters.

The Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB; Berman *et al.*, 2000) is the primary repository of protein structural data resulting from crystallography (along with the 15% of PDB structures obtained by noncrystallographic methods). The PDB also includes some information on how the crystals were obtained, but two problems limit the usefulness of the PDB as a scientific resource for crystal-growth data. Firstly, deposition of crystal-growth data into the PDB is optional, so that many depositors of protein structures decline to submit these data or submit incomplete information on crystal growth. In the current PDB with about 45 000 crystal structure entries, only about a third contain crystal-growth information with concentrations of the chemicals used, while another third contain lists of chemicals with no concentrations and the last third contain no crystal-growth chemical information at all. The second problem is that while the PDB stores the pH and temperature data numerically, all the chemical information, when present, is stored as a user-written unformatted text string. In this text, there are many abbreviations and synonyms for chemicals, including alternate ways to specify the same chemical and alternate ways to specify concentrations. This makes it difficult to process the data scientifically, *e.g.* to analyze correlations involving specific concentrations, chemicals or classes of chemicals. Despite these problems, the PDB contains a large amount of crystal-growth information and many crystallization research projects have attempted to assimilate and systematize its raw holdings (Page *et al.*, 2005; Peat *et al.*, 2005).

The BMCD (Biological Macromolecular Crystallization Database) was initiated as an archive of protein and nucleic acid crystallization data, including data obtained from extensive microgravity experiments in space (Gilliland, 1988). Previous versions of the BMCD drew data from published crystallization reports, by accessing the papers one by one; the publicly available resource has been used extensively in the development of crystal screening strategies and products (Hennessy *et al.*, 2000; Jancarik & Kim, 1991). Recently, methods of accessing and importing crystallization data from the PDB have been developed, enabling the BMCD to be updated significantly. This report describes new features in the BMCD version 4.02, principally its expanded data content and improved search capability, and gives examples of its use for scientific analysis. The operation of the BMCD v4.02 is also described briefly in terms of internal design and search features.

2. BMCD structure and data acquisition

The BMCD4 uses the open-source database server PostgreSQL 8.1.3 running on Unix/Linux operating systems to integrate its data, indices and search engine. Whereas earlier versions of the BMCD website were built with CGI (common gateway interface) software, the new BMCD4 makes extensive use of Java technology. The internal structure uses a set of 40 tables containing various data types (protein name, crystal pH, space group, chemical conditions *etc.*). The contents of the tables are indexed in a set of index files for efficient searching. The reporting feature (search result) gives a linked list of entries resulting from any given search. Individual entries are then presented with all the data for a particular crystal, including citations and notes, on a single web page.

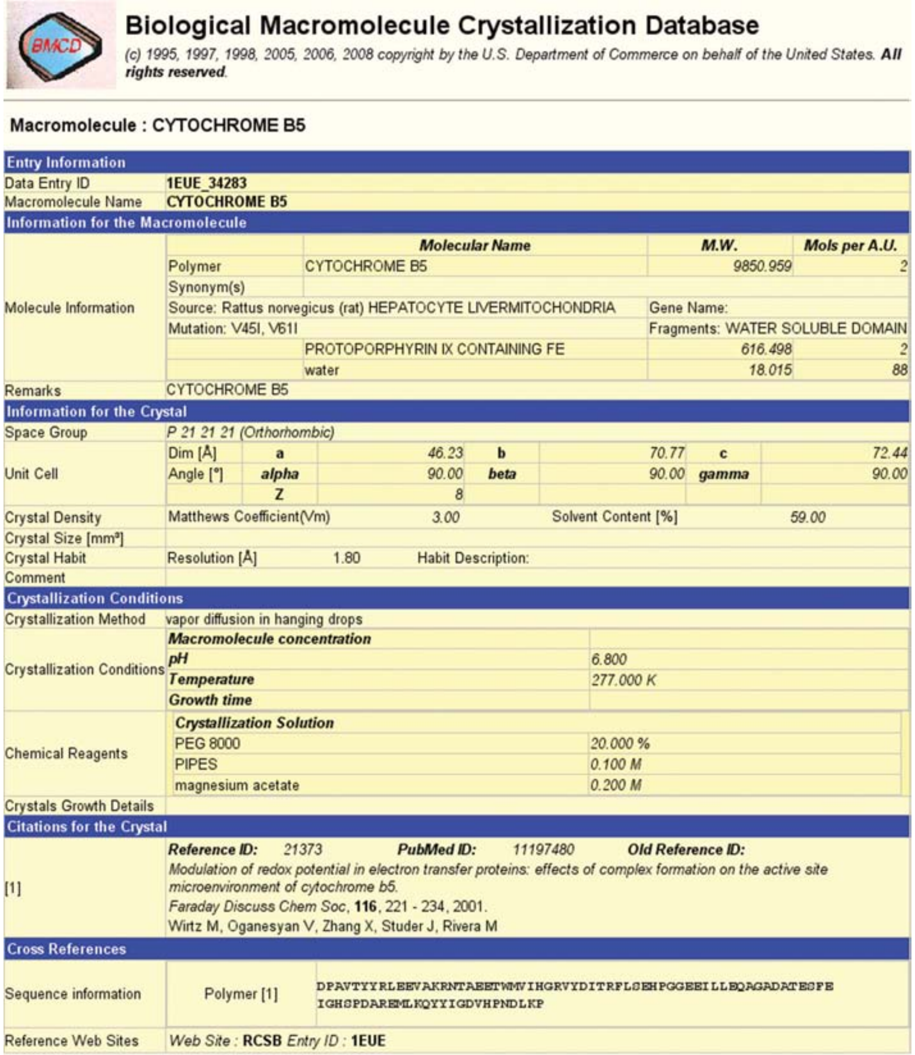
The BMCD4 contains two classes of entries. Those that belonged to BMCD version 3 (about 3500 entries, generally corresponding to information added before 1996) were obtained manually and tend to have complete information relating to crystal growth, including method-specific details. Approximately half of these entries derive from literature reports only and there is no structure in the PDB that directly corresponds to them. These entries have BMCD ID codes that begin with the letter M. The second class of entries are those that are new in version 4 and were obtained by retrieving and parsing data from the PDB. These entries correspond directly to PDB structures and the first four characters of the BMCD ID code are the same as the corresponding PDB code. For these entries, crystallization data are incomplete in some cases because the data are incomplete in the PDB.

In the current system of data acquisition, all new data are associated with a PDB entry. New data are imported from the PDB by use of its free download services, using the RSYNC utility to obtain XML files for each entry. The XML files are then processed by custom Java scripts to select data items of interest and convert them into database tables. When a PDB entry contains no crystallogen (chemical) information in its REMARK 280 fields, it is excluded from the BMCD. Entries that contain chemical names but with incomplete or absent concentration information are nevertheless imported because even without concentrations there is some information value in the chemical names. The conversion of REMARK 280 text into database tables involves the extensive use of custom text-processing scripts and extensive human attention, as described for a similar data-acquisition project by Peat *et al.* (2005). This processing, filtering and error checking is required in order to correctly interpret the unformatted raw information and preserve its scientific value.

3. Search features

The search engine was built using the search-engine library Lucene from Apache (Hatcher & Gospodnetic, 2004). The library is written entirely using Java technology to provide a rich query language. Information in the database can be searched by any text component of the entries, such as molecular name, EC number, biological source, space group, crystal system, article titles, author names, PubMed IDs and PDB IDs as well as specific combinations of these terms. The search engine also provides for tailored searches of specific data types and combinations of these by using AND, OR and NOT.

Two types of search are implemented. A simple text search is initiated by input on the home page. For example, a search for 'DTT OR mercaptoethanol' (single quotes are optional in actual input; the search string is case-insensitive) returns about a thousand entries that contain one or the other (or both) of these reducing agents, while a search for 'dtt mercaptoethanol' has the same result (OR is implicit). A search for "double mutant" returns the 13 entries that contain this phrase, while the search 'double AND mutant' returns a slightly larger superset that includes any entry with both the search terms. The asterisk character functions as a wildcard, enabling the search for 'tetra*' to yield a large set that includes tetragonal crystals as well as explicitly tetrameric proteins. Searching for 'mono* NOT monoclinic' retrieves non-monoclinic entries that contain the words monomeric, mononucleotide *etc*. The query syntax is described in a linked page provided by Lucene. The syntax for all BMCD searches is also explained on a linked information page.



Biological Macromolecule Crystallization Database

(c) 1995, 1997, 1998, 2005, 2006, 2008 copyright by the U.S. Department of Commerce on behalf of the United States. **All rights reserved.**

Macromolecule : CYTOCHROME B5

Entry Information	
Data Entry ID	**1EUE_34283**
Macromolecule Name	**CYTOCHROME B5**

Information for the Macromolecule				
		Molecular Name	***M.W.***	***Mols per A.U.***
Molecule Information	Polymer	CYTOCHROME B5	9850.959	2
	Synonym(s)			
	Source: Rattus norvegicus (rat) HEPATOCYTE LIVERMITOCHONDRIA		Gene Name:	
	Mutation: V45I, V61I		Fragments: WATER SOLUBLE DOMAIN	
		PROTOPORPHYRIN IX CONTAINING FE	616.498	2
		water	18.015	88
Remarks	CYTOCHROME B5			

Information for the Crystal							
Space Group	*P 21 21 21 (Orthorhombic)*						
Unit Cell	Dim [Å]	**a**	46.23	**b**	70.77	**c**	72.44
	Angle [°]	***alpha***	90.00	***beta***	90.00	***gamma***	90.00
		Z	8				
Crystal Density	Matthews Coefficient(Vm)	3.00	Solvent Content [%]	59.00			
Crystal Size [mm^3]							
Crystal Habit	Resolution [Å]	1.80	Habit Description:				
Comment							

Crystallization Conditions		
Crystallization Method	vapor diffusion in hanging drops	
Crystallization Conditions	***Macromolecule concentration***	
	pH	*6.800*
	Temperature	*277.000 K*
	Growth time	
Chemical Reagents	***Crystallization Solution***	
	PEG 8000	*20.000 %*
	PIPES	*0.100 M*
	magnesium acetate	*0.200 M*
Crystals Growth Details		

Citations for the Crystal	
[1]	***Reference ID:*** *21373* ***PubMed ID:*** *11197480* ***Old Reference ID:*** *Modulation of redox potential in electron transfer proteins: effects of complex formation on the active site microenvironment of cytochrome b5.* *Faraday Discuss Chem Soc*, **116**, 221 - 234, 2001. Wirtz M, Oganesyan V, Zhang X, Studer J, Rivera M

Cross References		
Sequence information	Polymer [1]	DPAVTYYRLEEVAKRNTAEETWMVIHGRVYDITRFLSEHPGGEEILLEQAGADATESFE IGHSPDAREMLKQYYIGDVHPNDLKP
Reference Web Sites	*Web Site :* **RCSB** *Entry ID :* **1EUE**	

Figure 1
Information present in a typical BMCD entry. In this entry (1EUE_34283), only a few fields are blank, such as the crystal size. Note that the protein sequence is included near the end of the listing.

An advanced search is offered as a separate link from the home page. The advanced-search page accepts input in the form of numeric ranges for any of the five parameters macromolecule concentration, pH, temperature, resolution and year of publication. BMCD entries that lie within all the specified ranges are then listed as output. In addition, the advanced search accepts text input (like the simple search) to further specify the target set. For example, using the advanced search one could identify the 14 entries that contain the text string ‘adenosine’ and have pH between 3 and 6.6 and also have a diffraction resolution between 0.1 and 1.85 Å.

One additional search feature is the ability to request text matches within specific fields. Each entry has distinct text fields for protein name, organism name, space group, chemical names *etc.* (the full list of searchable fields, with examples, is linked to the search page under ‘More Information’) and these can be searched independently using a colon syntax. For example, the query `title:recognition` will find any entry whose publication title includes the word ‘recognition’. The query `title:"recognition helix"` will return the smaller set where the title includes this phrase. Multiple field searches may be combined using boolean operators. A term or phrase not preceded by a field name will be searched through the entire entry (general search). However, field searching and non-field searching cannot be combined in the same query. The way to combine field and general searching is to use the ‘Content:’ field, which is effectively a general search over all fields. Field names are case-specific. They are all completely lower case, except ‘Content’.

Here are a few examples of correct syntax for field searches: `title:"HIV-1 protease" AND spgrp:P61`, `title:antibody AND common_name:mouse`, `title:antibody AND Content:mouse AND NOT common_name:mouse`, `chem_name:aden* AND (author:mckay OR author:steitz)`.

The output of a search begins with the number of entries found, followed by a list of their BMCD codes, molecular names and biological sources. The molecular-name field usually includes the scientific name of the molecule along with common synonyms as previously described (Gilliland *et al.*, 1994). Then, by clicking on the ID code of one of the entries all the data pertaining to that entry is displayed, as shown in Fig. 1.

4. Examples of BMCD data analysis

Following are a few examples of analysis of data distributions in the BMCD4. These examples are provided primarily to show how the new search features can be used to probe relationships among crystal-growth parameters. Thus, the scientific interest in these findings is not the primary objective and more extensive analysis involving further and more detailed searches would be required before reaching scientific conclusions. These examples are intended to demonstrate the ease with which simple trends and correlations can be observed.

Using the newly implemented field-specific and numeric range search features, it is straightforward to obtain data such as, for example, the number of yeast protease entries with pH between 4 and 5. Such information enables histograms such as those in Figs. 2, 3, 4 and 5. Fig. 2 shows the distribution of BMCD4 content by major types of nonbuffer precipitants. A similar histogram could be made using the ‘Year of Publication’ field to show the distribution by year of cited publication. Searching in both these fields at once (chem_name:PEG and old-*versus*-recent year ranges) enables quick corroboration of the observation reported by Peat *et al.* (2005) that PEG conditions are becoming more prevalent (1715 of 5219 entries with publication year 1901–1996 include PEG, which is 33%, while 4439 of 9151 entries with publication year 1997–2007

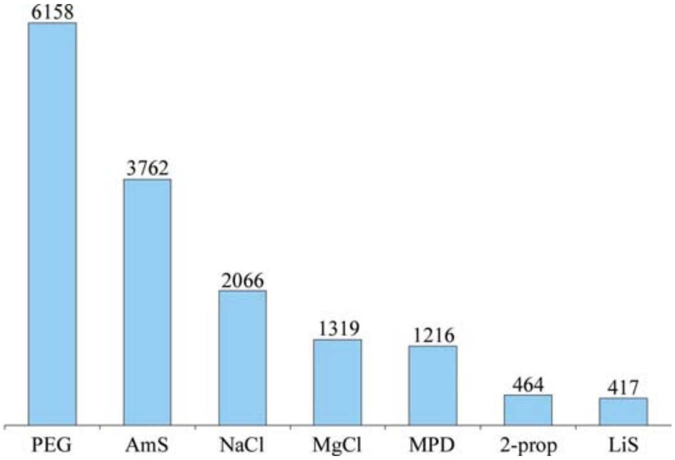


Figure 2
Histogram of BMCD4 entries by common precipitant types. Some chemical names have been abbreviated (AmS, ammonium sulfate; MgCl, $MgCl_2$, 2-prop, 2-propanol; LiS, lithium sulfate). These data were obtained by using searches such as `chem_name:"ammonium sulfate"`. Note that an entry that contains two of the chemicals (*e.g.* both PEG and AmS) would be counted twice in this graph.

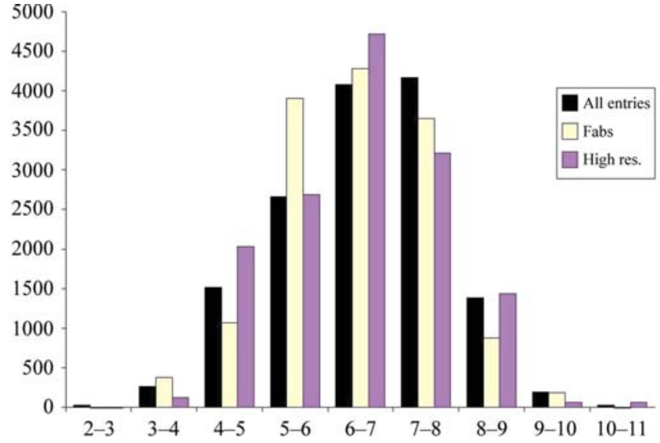


Figure 3
Histogram showing the number of BMCD4 entries as a function of the pH. Also shown are the pH distributions for two subsets: Fabs (antigen-binding fragments of antibodies) and high-resolution crystals (resolution between 0.9 and 1.2 Å). The two subsets have about the same size: 223 and 220 entries, respectively. The subset bars have been normalized against the full data set to facilitate comparison. It appears that Fab crystals are relatively rich in pH 5–6 conditions, while the high-diffracting crystals show slight preferences for pH 4–5 and pH 6–7.

contain PEG, which is 49%). This simple finding could of course be elaborated with more detailed sampling by years, PEG types *etc.* Figs. 3, 4 and 5 provide further examples of simple statistical analysis intended to demonstrate how to use the new search features to find correlations among BMCD data. These correlations can be useful from either a practical standpoint (*i.e.* focusing searches or designing crystal screens) or for their scientific interest. Thus in Fig. 3, two different BMCD subsets (Fab antibody fragments and high-resolution diffracting crystals) appear to have pH distributions that differ from the overall distribution of pH values in the BMCD. A complete investigation of these trends, which is beyond the scope of this report, would involve further statistical analysis and more detailed examination of buffers and other crystal-growth parameters.

Fig. 4 shows an example of how the biological source (BMCD field 'scientific_name' or 'common_name') can be probed for correlations with crystal-growth conditions. The five most common representatives from the eukaryotic and prokaryotic kingdoms are graphed overall and queried for the number of their crystals involving PEG and the number with sulfate. The data suggest that eukaryotic proteins crystallize from less ionic conditions than prokaryotic proteins. Deeper analysis would require consideration of additional factors such as historical trends in proteomic targets and in commercial screen compositions. Fig. 5 presents the distribution of diffraction resolutions for all BMCD entries and compares this overall distribution with those for two specific space groups with similar populations in the database. Space group $P1$ has 489 entries, while space group $P3_121$ has 463. It appears that crystals belonging to space group $P1$ generally diffract better than those belonging to the trigonal space group. A complete analysis of this correlation would also consider historical trends in space-group frequencies and solvent content. This finding appears to be of greater scientific than practical interest, but it is possible that such correlations could be applied to steer optimization efforts toward crystals with space groups or other properties that are more likely to lead to good diffraction.

5. Discussion

The BMCD has been recast on a PostgreSQL 8.1.3 platform using the Lucene query language and expanded to include over 14 000 entries in its continuing mission to make protein crystallization more predictable and scientific. The upgraded database and its website are intended to assist crystallographers worldwide to view crystallogenesis data, to identify correlations between protein properties and crystal conditions, to assist in selecting crystal conditions for a particular protein and for the design of crystal screening strategies.

A data standard intended to advance these general goals was proposed by Peat *et al.* (2005) and like-minded efforts to improve the archiving of crystallogenic data are in progress at IUCr journals (Einspahr & Guss, 2005). Since the deposition of crystallographic data into the RCSB and the submission of manuscripts to journals are correlated processes, it is important that depositors are not burdened with redundant or inefficient data entry. The mmCIF data standard, which accommodates crystallogenic data quantitatively and scientifically, is an especially promising instrument; ideally the mmCIF for each crystal structure, or some similar file, could serve as a master file with all important information, exchanged and amended as needed, suitably archived and available to the public.

The optimal format for archiving crystallogenic data is difficult to define owing to the vagaries of crystal-growth methods, which derive from the diversity of macromolecules and the laboratories that study them. Explicit recording of the primary data (as listed in §1) would cover most cases and would greatly abet crystal-growth science. However, this may

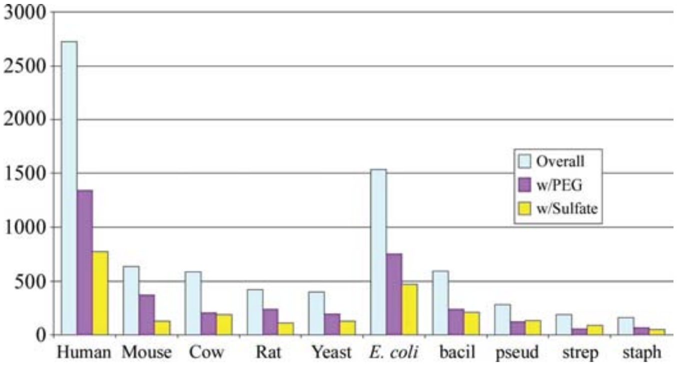


Figure 4
Histogram of BMCD4 crystals by source organism for the five most common eukaryotic and prokaryotic genera. Abbreviations used are bacil, *Bacillus*; pseud, *Pseudomonas*; strep, *Streptomyces*; staph, *Staphylococcus*. Also shown for each genus is the subcount of crystals with PEG and with sulfate (any sulfate). There appears to be a stronger preference for PEG conditions over sulfate conditions among the eukaryotes; if the numbers for each kingdom are summed, the overall ratio of PEG-grown to sulfate-grown crystals among eukaryotes is 1.78, while for prokaryotes this ratio is 1.29.

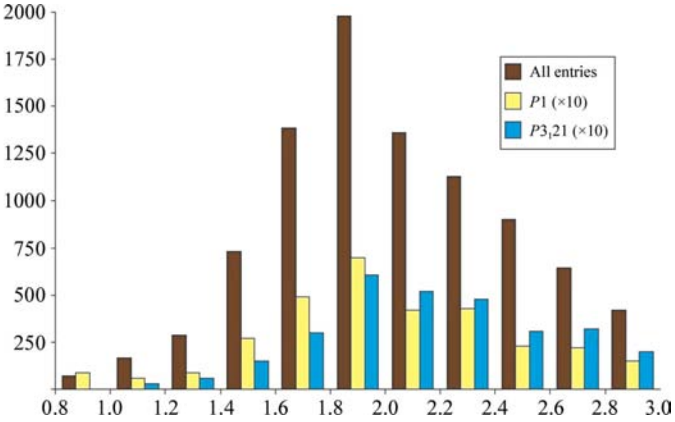


Figure 5
Histogram of BMCD4 entries by diffraction resolution in Å. The graph omits extremely low and high resolutions. Also shown are the distributions for two specific space groups. These two subsets are about the same size (489 $P1$ entries and 463 in space group $P3_121$) and their bar heights have been scaled up by a factor of 10 to facilitate comparison. It appears that the triclinic crystals tend to diffract better than the trigonal crystals.

be impractical for high-throughput laboratories using robots; in such situations it may be more efficient to record a tag corresponding to the screening condition; this could subsequently be translated by reference to a screen table and the detailed conditions substituted. Data entry inevitably involves some trade-offs between convenience to the depositor and quality of the archived data. Progress in this area is likely to depend on significant short-term investments in data-entry software. This software would ideally make the data-entry process user-friendly (*i.e.* clear and efficient, even automatic where possible) while at the same time enforcing on-the-fly data standardization and error checking as described by Peat *et al.* (2005).

Further growth and development of the BMCD are planned. Two goals for the next BMCD release are to parse all crystallogen information into specific chemicals with numerically stored range-searchable concentrations (to facilitate detailed statistical analysis) and to add tables of synonyms to enable integrated analysis of entries that use different names for the same chemical. Additionally, future revisions of the database will incorporate taxonomic information on source organisms, classification of proteins and more powerful searches with more user control of search outputs. Another goal for future BMCD development is to consolidate entries with their mutants, derivatives and complexes that crystallize isomorphously under similar conditions. Presently, many similar entries result from some searches; for some types of analysis and design it would be advantageous to eliminate this redundancy. Consolidating these similarly grown crystals will also help to identify potentially important cases where small changes in structure resulted in dramatic changes in crystal conditions.

The database and its publication are supported by the US National Institute of Standards and Technology. Assistance with database development and support from Gary Gilliland, Jane Ladner and Robert Goldberg is gratefully acknowledged. Certain commercial products are identified in this manuscript in order to specify the procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the products identified are necessarily the best available for the purpose.

References

Berger, I., Kang, C., Sinha, N., Wolters, M. & Rich, A. (1996). *Acta Cryst.* D**52**, 465–468.

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.

Einspahr, H. & Guss, M. (2005). *Acta Cryst.* F**61**, 1–2.

Gilliland, G. L. (1988). *J. Cryst. Growth*, **90**, 51–59.

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

Hatcher, E. & Gospodnetic, O. (2004). *Lucene in Action.* New York: Manning Publications.

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.

Page, R., Deacon, A. M., Lesley, S. A. & Stevens, R. C. (2005). *J. Struct. Funct. Genomics*, **6**, 209–217.

Peat, T. S., Christopher, J. A. & Newman, J. (2005). *Acta Cryst.* D**61**, 1662–1669.

Radaev, S., Li, S. & Sun, P. D. (2006). *Acta Cryst.* D**62**, 605–612.