Chapter 24.3. The Biological Macromolecule Crystallization Database

D. T. Gallagher and M. Tung

24.3.1. Introduction

The crystallization of a biological macromolecule is a key step in determining its three-dimensional structure by X-ray diffraction. Even for proteins of known structure, it is very difficult to make predictions about crystal-growth conditions or crystal properties. Hence, in crystallizing macromolecules, empirical procedures are used that take advantage of the knowledge gained from past successes. Usually a large set of trials is carried out that varies parameters such as pH, temperature, ionic strength and macromolecule concentration. The number of experiments required for success is variable. In many cases the search ends quickly, either because the right choices were made early or because crystallization occurs over a broad range of conditions. Unfortunately, in other cases, a large number of experiments are required before the discovery of crystallization conditions, and in some cases no crystal conditions are found even after exhaustive searching.

After more than 50 years of experience in the production of diffraction-quality crystals, there is still no generally accepted strategy for searching for the crystal-growth conditions for a biological macromolecule. However, a number of systematic procedures and strategy suggestions have been put forth (*e.g.* McPherson, 1976; Blundell & Johnson, 1976; Carter & Carter, 1979; McPherson, 1982; Gilliland & Davies, 1984; Gilliland *et al.*, 1994, 1996; McPherson, 1999). Most current strategies employ a version of the 'fast screen' first popularized by Jancarik & Kim (1991). Fast screens are sets of preformulated solutions similar to those that have frequently produced crystals of other proteins in the past. Crystals are often found quickly in such experiments, but failure results in the need for a more general approach.

The purpose of the Biological Macromolecule Crystallization Database (BMCD) is to provide comprehensive information to facilitate the development of crystallization strategies for the production of crystals suitable for X-ray structural investigations (Gilliland & Davies, 1984). The BMCD includes entries for all classes of biological macromolecules for which diffraction-quality crystals have been obtained, including proteins, nucleic acids and their various complexes.

24.3.2. History of the BMCD

The BMCD has its roots in work that was initiated in Dr David Davies' laboratory at NIH in the late 1970s and early 1980s (Gilliland & Davies, 1984). While working on a variety of frustrating protein-crystallization problems, a large body of crystallization information was extracted from the literature. This eventually led to a systematic search of the literature and a compilation of data that included almost all of the crystallization reports of biological macromolecules available at the time. In 1983 the data, as an ASCII file, were submitted to the Protein Data Bank (PDB; see Chapter 24.1) for public distribution. The data included the crystallization conditions for 1025 crystal forms of more than 616 biological macromolecules.

In 1987, with assistance from the National Institute of Standards and Technology (NIST) Standard Reference Data Program, the data were incorporated into a searchable database and distributed with software that made it accessible using a personal computer. The database was released to the public in 1989 as the NIST/CARB (Center for Advanced Research in Biotechnology) Biological Macromolecule Crystallization Database, version 1.0 (Gilliland, 1988). In 1990, a second version was released (Gilliland & Bickham, 1990), and in 1994 the BMCD began including data from microgravity crystal-growth studies carried out in orbit (Gilliland *et al.*, 1994). Soon afterwards, the BMCD was migrated to a UNIX platform to facilitate internet access, and the software was rewritten to enable direct import of data from the PDB. During the period 1997–1998 this reprogramming superceded new data acquisition. Hence, there is a local minimum of crystal entries for that period. In 1999, data acquisition resumed and the size of the BMCD began to increase sharply. In 2008, the BMCD released version 4.01 (Tung & Gallagher, 2009), using the open-source database server PostgreSQL 8.1.3. This version also includes new search features capable of searching for arbitrary text and for ranges of five numeric parameters (pH *etc*.) as described below.

24.3.3. BMCD data

The BMCD stores information in 40 different data types generally divisible into three groups relating to the macromolecule, the crystallization and the crystal itself. Under macromolecule are the name, aliases, biological source genus, tissue *etc*., mutations, and numbers and sizes of subunit types. The macromolecule sequence is also included but is not yet searchable. The crystal-growth information contains the temperature, the pH, the concentrations of all chemical components including the macromolecule, growth time and the method. Crystal data include the space group, unit-cell parameters, molecules per cell and diffraction resolution. Also included are the V_m and solvent-fraction values. An additional section for each entry records references to published literature.

The BMCD4 (current release, version 4) 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 roughly covering the period 1997–2007. This represents the new model of data acquisition, utilizing the PDB's RSYNC download 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. This conversion involves extensive use of text-parsing scripts and human attention, as described for a similar data-acquisition project by Peat *et al.* (2005). For this processing, filtering and error checking are required in order to correctly interpret unformatted PDB information, especially the crystal conditions.

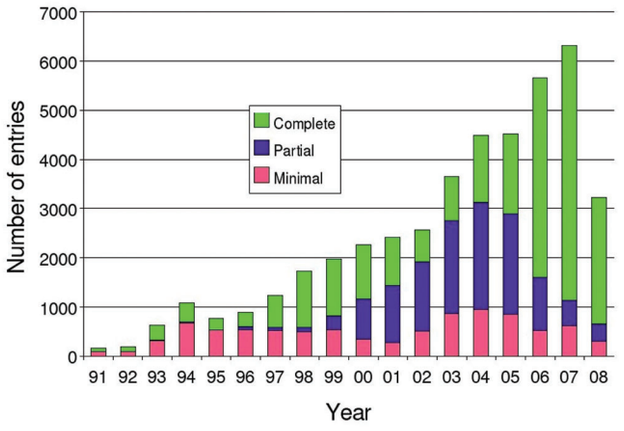


Figure 24.3.3.1
Trends in the completeness of BMCD crystallization conditions.

The resulting entries correspond directly to PDB structures, and the first four characters of the BMCD ID code are the same as the corresponding PDB code. Because the PDB deposition does not require crystal-growth information, many BMCD entries imported from the PDB contain incomplete information. BMCD entries can be divided into three groups according to the completeness of their crystal-growth information: those with 'complete' information, those listing chemicals but not concentrations ('partial'), and those that lack even chemical names but still have some information, such as pH ('minimal'). For a set of 43 698 BMCD entries including the currently available version 4.01 as well as 29 326 entries currently undergoing validation processing, the distribution of these three groups is shown in Fig. 24.3.3.1. As can be seen, most of the entries from 2004 and 2005 are incomplete, but more recently the trend appears to be in favour of recording complete conditions.

An additional complication arises from the wide assortment of synonyms and misspellings in the raw data. Since statistical analysis of chemical conditions requires standardization of chemical names, we are developing methods to interpret and convert both synonyms (*e.g.*, AmSO4, A.S. and AmS for ammonium sulfate) and common misspellings into standard chemical names. This process results in a large reduction in the number of unique chemical names, dominated by a few chemicals that have many synonyms and many misspellings. Efforts to standardize the input and archiving of crystallization information are also underway at major crystallography journals (Einspahr & Guss, 2005).

24.3.4. Web interface

The BMCD is an internet-accessible resource available through the website http://xpdb.nist.gov:8060/BMCD4 **(Please check URL)**. The web interface provides a convenient mechanism for browsing through the data contained in the BMCD. The user can examine individual entries, lists of entries selected by customizable search criteria or statistical trends based on the numbers of entries for various properties and parameter ranges. In addition, specific references are listed for most entries.

Two types of search are possible. 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 the 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 explained on linked information pages.

An advanced search is offered from its own query page, which 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 with parameters 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), 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; all must be 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 linked 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). By clicking on the ID code of one of the entries all the data pertaining to that entry are displayed.

24.3.5. Reproducing published crystallization procedures

The BMCD contains the information required to reproduce previously reported crystals for a biological macromolecule. Usually, a later batch of macromolecule will not behave identically to previous samples and some optimizing is required, beginning with the reported crystallization conditions. The

conditions may be simple to reproduce, but differences in the isolation and purification procedures, reagents, and crystallization methodology of different laboratories can dramatically influence the results. The crystallization conditions in the BMCD should be considered as a good starting point for the search or optimization that may require experiments that vary pH, macromolecule and reagent concentrations, and temperature.

An attempt to reproduce crystals of an isozyme of glutathione *S*-transferase from rat liver (Sesay *et al.*, 1987) is used to illustrate these points. The original crystallization conditions, for the enzyme purified from liver tissue, were archived as entry M0P3. Several years later the same enzyme was cloned and expressed in *Escherichia coli* for further structural studies. The crystals of the original enzyme grew in 3 to 5 days from vapour-diffusion experiments at 4 °C, with droplets containing a protein concentration of 11 mg ml^{-1}, 0.46% β-octylglucoside, 30–37% saturated ammonium sulfate and 0.1 *M* phosphate buffer pH 6.9 equilibrated against well solutions containing 60–74% ammonium sulfate.

The recombinant enzyme required an optimization of these conditions to produce large single crystals. The recombinant protein crystallized best at the same temperature, with droplets containing a protein concentration of 12 mg ml^{-1}, 0.2% β-octylglucoside, 20–25% saturated ammonium sulfate, 1 m*M* EDTA and 0.025 *M* TrisHCl, pH 8.0 equilibrated against well solutions containing 40–50% ammonium sulfate. Both crystallization protocols required the presence of 1 m*M* (9*R*,10*R*)-9-*S*-glutathionyl-10-hydroxy-9,10-dihydrophenanthrene, a product inhibitor. Thus, in this example, most of the chemical concentrations shifted, and the pH shifted so much that a different buffer was used. Optimized crystals of the recombinant enzyme grew in 5 to 10 days (Ji *et al.*, 1994).

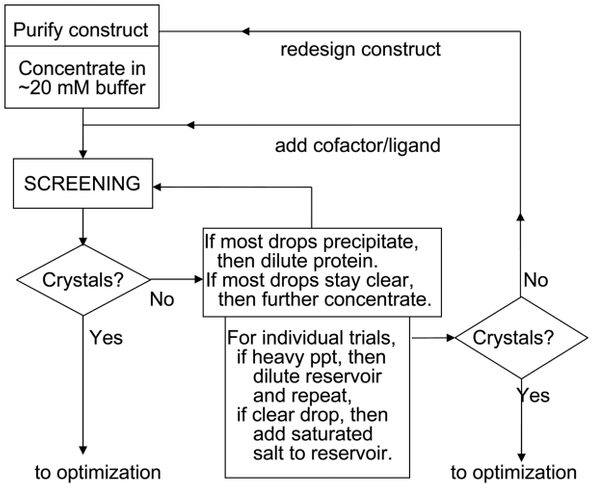
24.3.6. Crystallization screens

A crystallization screen is a set of formulations designed to be mixed with macromolecule solutions as an efficient search for crystallogenic conditions; several are now available commerically as sets of premixed solutions (*e.g.* Cudney *et al.*, 1994). Since the introduction of the fast screen by Jancarik & Kim (1991), almost all attempts to crystallize a new protein begin with a screen of one form or another. Early screens was based on the ideas put forth by Carter & Carter (1979) in their discussion of the use of incomplete factorial experiments to limit the search. The first screens were quite general and applicable to a wide range of biological macromolecules, but screens based on specific classes of molecules such as RNA (Scott *et al.*, 1995) or protein complexes (Radaev *et al.*, 2006) soon developed.

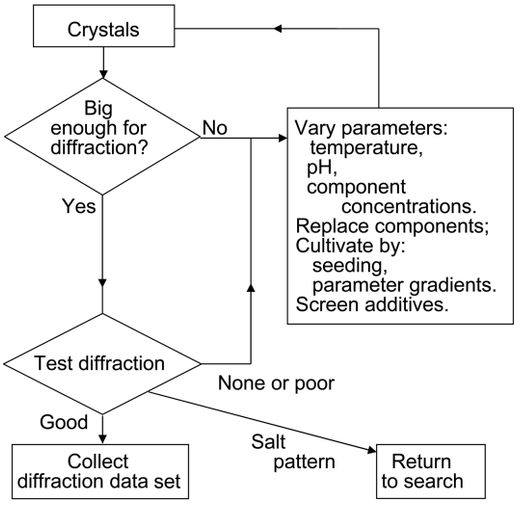
The design of crystal screens is based on previously successful crystallizations. The BMCD has been used extensively for the development of general screens, and can be used to facilitate the development of screens for specific classes of macromolecules. For example, if it were desired to produce a screen for endonucleases, a search of the BMCD would reveal the ranges of key parameters and the most prevalent reagents that have been used to crystallize these enzymes. From an examination of these parameters, a subset of potential crystallization conditions comprising an endonuclease screen could be developed. Another type of focused screen could be one designed to feature a particular crystallant. For example, if L-ornithine were theorized to promote the crystallization of proteins, a screen featuring this additive could be designed by surveying the BMCD for relevant precedents and parameter ranges. Similar methods could be used to focus the initial search for any specific crystallization project.

24.3.7. A general crystallization procedure

The information in the BMCD has been incorporated into general procedures for the crystallization of biological macromolecules not previously crystallized (Gilliland, 1988; Gilliland & Bickham, 1990; Gilliland *et al.*, 1994, 1996). One such general procedure is shown in Fig. 24.3.7.1. Briefly, in this procedure the purified biological macromolecule is concentrated (if possible) to 10 to 25 mg ml^{-1} and dialysed into 10 to 25 m*M* buffer at a neutral pH or at a pH favouring the solubility of the biopolymer. Other stabilizing agents such as EDTA and/or dithiothreitol may be included at low concentrations to stabilize the biological macromolecule during the crystallization trials. If the biomolecule requires some salt for stability or solubility, then it should be included, but ideally the additional components in the protein solution are kept to a minimum as they limit the search space.



(*a*)



(*b*)

Figure 24.3.7.1
A general crystallization strategy based on the data contained in the BMCD. The overall strategy comprises a search phase (*a*) and an optimization phase (*b*).

Once the macromolecule solution has been prepared, commercial or customized fast screens are carried out using vapour-diffusion experiments. If no crystals appear, it is important to recognize that some trials may contain favourable chemical conditions but fail to produce crystals simply because the concentrations are too high or too low. Thus, each screen trial should be considered not as a single point in composition space but as a vector to be scanned, and any given trial cannot be considered to have failed until its region of transition (between clear drop and precipitation) is observed. If most of the trials produce heavy precipitate, then systematic dilution (either of the reservoirs or of the protein) is advised. If on the other hand most of the drops remain clear, then either more concentrated protein or more dehydrating reservoir conditions will help to bring the protein to its solubility limit in that milieu, and hence to its potential crystal-nucleation point. Individual trials can be manipulated to ensure that no potential crystallization conditions are missed simply because of too high or too low concentation.

Often the first crystals to appear are small, compound or otherwise poor in quality. In this case the optimization phase (Fig. 24.3.7.1*b*) begins by attempting to simplify the crystal conditions (can any components be omitted?), and then systematically varying the crystallization parameters (pH, temperature, chemical concentrations) in the hope of improving size and quality. These experiments generally incorporate controls corresponding to the best crystals obtained so far, so that reproducibility is continually assessed. When crystals are large enough (usually about 0.02 mm), diffraction is tested. Beyond optimizing existing parameters, new components can be substituted (*e.g.*, replacing PEG 4K with PEG 5K MME) or introduced *de novo*, using crystallizations of similar proteins or crystallizations that utilized similar conditions in the BMCD as a guide. Microseeding (often useful to obviate nucleation and to control populations) or macroseeding (to cultivate large, high-quality single crystals) may also be required to optimize crystal growth (McPherson, 1982, 1999).

If the fast screens produce no crystals, a more thorough approach can be undertaken. An analysis of the BMCD data reveals that out of the large number of reagents used as precipitating agents, a small set accounts for the majority of the crystals observed. The pH range for all crystals is quite large, but most proteins crystallize between pH 4.0 and 9.0. Even though temperature can be an important factor, crystallization experiments are usually set up at room (20 °C) or cold-room (5 °C) temperatures. Protein concentration varies quite markedly, but it appears that most experiments use from 5 to 15 mg ml^{-1}. After examining the data in the BMCD, the precipitating agents ammonium sulfate, polyethylene glycol 4000, 2-methyl-2,4-pentanediol and sodium–potassium phosphate might be selected for custom-screening efforts, with initial trials restricted to a pH range of 4.0 to 9.0 and temperatures of 5 and 20 °C. As a prescreening solubility assay, a small amount (2 µl) of the protein is mixed with several concentrations of each of the selected precipitants, and buffered at pH 4.5, 6.0 and 8.0, at both temperatures, monitoring by microscope for precipitation. This establishes the concentration ranges for the reagents for setting up hanging-drop (or any other commonly used technique) experiments. Next, separate sets of experiments that would sample the pH range in steps of 0.5 and reagent concentrations near, at and above those that induce precipitation of the protein would be set up at temperatures of 5 and 20 °C. The assessment of the results of experiments after periodic observations may show (for example by an abrupt precipitation at a particular reagent concentration, pH and/or temperature) a need for finer sampling of any or all of the parameters near the observed discontinuity. In parallel, or if the crystallization trials just described are unsuccessful, another set of experiments can be carried out that include the addition of small quantities of ligands, products, substrate, substrate analogues, monovalent or divalent cations, organic reagents *etc.* to the crystallization mixtures. If this does not prove fruitful, additional reagents may be selected with the aid of the BMCD and new experiments initiated.

24.3.8. The future of the BMCD

Further developments for the BMCD are in progress, with the general goal of increasing its capacity for data analysis, thereby facilitating both scientific understanding and practical applications. It is anticipated that text-parsing and error-correction scripts will soon enable efficient regular import of data from the PDB. This will enable the automatic parsing of all crystallization information into specific chemicals with numerically stored range-searchable concentrations to facilitate detailed statistical analysis. Additionally, future revisions of the database will incorporate taxonomic information on source organisms, classification of proteins, sequence analysis and more powerful searches with more user control of search outputs. The capabilities of the web resource will be expanded to include tools for the development of strategies for new crystallization problems.

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 equipment, instruments and materials are identified in this paper in order to specify the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials and equipment identified are necessarily the best available for the purpose.

References

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. New York: Academic Press.

Carter, C. W. Jr & Carter, C. W. (1979). *Protein crystallization using incomplete factorial experiments*. *J. Biol. Chem.* **254**, 12219–12223.

Cudney, R., Patel, S., Weisgraber, K., Newhouse, Y. & McPherson, A. (1994). *Screening and optimization strategies for macromolecular crystal growth*. *Acta Cryst.* D**50**, 414–423.

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

Gilliland, G. L. (1988). *A biological macromolecule crystallization database: a basis for a crystallization strategy*. *J. Cryst. Growth*, **90**, 51–59.

Gilliland, G. L. & Bickham, D. (1990). *The Biological Macromolecule Crystallization Database: a tool to assist the development of crystallization strategies*. *Methods Companion Methods Enzymol.* **1**, 6–11.

Gilliland, G. L. & Davies, D. R. (1984). *Protein crystallization: the growth of large-scale single crystals*. *Methods Enzymol.* **104**, 370–381.

Gilliland, G. L., Tung, M., Blakeslee, D. M. & Ladner, J. E. (1994). *Biological Macromolecule Crystallization Database, Version 3.0: new features, data and the NASA archive for protein crystal growth data*. *Acta Cryst.* D**50**, 408–413.

Gilliland, G. L., Tung, M. & Ladner, J. (1996). *The Biological Macromolecule Crystallization Database and NASA Protein Crystal Growth Archive*. *J. Res. Natl Inst. Stand. Technol.* **101**, 309–320.

Jancarik, J. & Kim, S.-H. (1991). *Sparse matrix sampling: a screening method for crystallization of proteins*. *J. Appl. Cryst.* **24**, 409–411.

Ji, X., Johnson, W. W., Sesay, M. A., Dickert, L., Prasad, S. M., Ammon, H. L., Armstrong, R. N. & Gilliland, G. L. (1994). *Structure of the xenobiotic substrate binding site of a glutathione S-transferase as*

revealed by X-ray crystallographic analysis of product complexes with the diastereomers of 9-(S-glutathionyl)-10-hydroxy-9,10-dihydrophenanthrene. *Biochemistry*, **33**, 1043–1052.

McPherson, A. (1976). *The growth and preliminary investigation of protein and nucleic acid crystals for X-ray diffraction analysis*. *Methods Biochem. Anal.* **23**, 249–345.

McPherson, A. (1982). *Preparation and Analysis of Protein Crystals*. New York: Wiley.

McPherson, A. (1999). *Crystallization of Biological Macromolecules*. New York: Cold Spring Harbor Laboratory Press.

Peat, T. S., Christopher, J. A. & Newman, J. (2005). *Tapping the Protein Data Bank for crystallization information*. *Acta Cryst.* D**61**, 1662–1669.

Radaev, S., Li, S. & Sun, P. D. (2006). *A survey of protein–protein complex crystallizations*. *Acta Cryst.* D**62**, 605–612.

Scott, W. G., Finch, J. T., Grenfell, R., Fogg, J., Smith, T., Gait, M. J. & Klug, A. (1995). *Rapid crystallization of chemically synthesized hammerhead RNAs using a double screening procedure*. *J. Mol. Biol.* **250**, 327–332.

Sesay, M. A., Ammon, H. L. & Armstrong, R. N. (1987). *Crystallization and a preliminary X-ray diffraction study of isozyme 3–3 of glutathione S-transferase from rat liver*. *J. Mol. Biol.* **197**, 377–378.

Tung, M. & Gallagher, D. T. (2009). *The Biomolecular Crystallization Database Version 4: expanded content and new features*. *Acta Cryst.* D**65**, 18–23.

Abstract

The Biological Macromolecule Crystallization Database (BMCD) is described. The database is available at http://xpdb.nist.gov:8060/BMCD4 **(please check URL)** and currently contains 14 372 entries with crystallization information for proteins, protein–protein complexes, nucleic acids, nucleic acid–nucleic acid complexes, protein–nucleic acid complexes and viruses. The information in the BMCD is applicable for the general study of protein crystallization, for reproducing crystals previously reported in the literature and for designing strategies to crystallize new macromolecules. Methods for utilizing the BMCD are presented, including examples of crystallization-strategy development.