



ELSEVIER

Journal of Crystal Growth 183 (1998) 653–668

JOURNAL OF
**CRYSTAL
GROWTH**

Re-clustering the database for crystallization of macromolecules

Robert G. Farr, Jr., Alexander L. Perryman, Cleopas T. Samudzi*

Biochemistry Department, University of Missouri-Columbia, 117 Schweitzer Hall, Columbia, MO 65211, USA

Received 19 May 1997; accepted 8 August 1997

Abstract

The current version of the biological macromolecule crystallization database (BMCD version 3.0) was statistically analyzed using clustering techniques. This is an effort to look for trends that may be useful in the crystallization of new macromolecules. Our previous statistical analysis of the BMCD was performed on version 1.0 [C.T. Samudzi, M.J. Fivash, J.M. Rosenberg, *J. Crystal Growth* 123 (1992) 47]. That database contained information on a total of 1025 crystallization experiments for 820 biological macromolecules (about 35% of those entries were incomplete and, thus, inappropriate for analysis). Version 3.0 of the BMCD is more than 90% complete and contains information on a total of about 2300 crystallization experiments for approximately 1500 biological macromolecules [G.L. Gilliland, M. Tung, D.M. Bakerslee, J.E. Ladner, *Acta Cryst. D* 50 (1994) 408]. With significantly more data in the BMCD, the question is whether trends have changed. The SAS software [SAS Institute Inc., SAS/STAT, Version 6, 4th ed., vol. 1] was used throughout the analysis. The following crystallization parameters were used in defining an experiment: pH, temperature, molecular weight, macromolecular concentration, precipitant type and crystallization method. Using these parameters, a measure of the differences between experiments was developed. Groups or clusters of similar experiments were identified as those close together based upon this difference measure. The database was successfully resolved into 25 clusters. The pseudo-F statistic for 25 clusters was 306.30 and is statistically significant ($p < 0.0001$). Although eight of these clusters can be treated as outliers, the other 17 clusters provide useful information in recognizing new patterns and developing strategies for crystallization of macromolecules. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Statistical analysis; Crystallization conditions; Crystal growth

1. Introduction

1.1. General overview

Crystallization remains the rate-limiting step in the structure determination of macromolecules

using the single-crystal X-ray crystallography technique. Current literature contains a number of useful ideas that are designed to rationalize and systematize the process of crystallization of new macromolecules [4–13]. The use of statistics to evaluate available information on crystallization of macromolecules appears to be inherently objective. However, a complete statistical analysis requires examination of successful crystallization

*Corresponding author. Fax: +1 573 884 4812; e-mail: cleo@mercury.biochem.missouri.edu.

experiments as well as the unsuccessful attempts. But, since the unsuccessful crystallization experiments are not reported, the only option left is to look for trends from available data, i.e. using only successful crystallization experiments.

The current version of the biological macromolecule crystallization database (BMCD version 3.0) was statistically analyzed using clustering techniques. The previous statistical analysis of the BMCD was performed on version 1.0 [1]. That database contained information on a total of 1025 crystallization experiments for 820 biological macromolecules (about 35% of those entries were incomplete and/or inappropriate for analysis). Version 3.0 of the BMCD is about 90% complete and contains information on a total of about 2300 crystallization experiments for approximately 1500 biological macromolecules [2]. With significantly more data in the current BMCD (version 3.0), the question is whether there are differences in crystallization patterns.

The basic principle and theory of analysis of the BMCD using clustering is described in the previous analysis [1]. The only major difference between the previous report and the current analysis is the database. Thus, the reader is referred to our previous work for a detailed description of clustering. What should be emphasized here, however, is the choice of the following parameters to define a crystallization experiment; molecular weight, macromolecular concentration, pH, temperature, type of precipitant and crystallization method. Selection of these parameters was based both on their ability to differentiate between experiments and also on the following considerations. (1) There are many missing values (for the parameters) in the database. Therefore, only those entries with three or more values of the parameters present were extracted and analyzed. (2) Some of the most commonly mentioned conditions that affect crystal growth are molecular weight, macromolecular concentration, pH, ionic strength, temperature, precipitant type and crystallization method [4–6]. The BMCD, indeed, has no category called ‘precipitant’. Instead, there is a category called ‘chemical additions to the crystal growth medium’. Typically, this category lists three or four ‘chemical additions’ including a wide range of traditional

biological buffers (with concentrations about 0.10 M or less), chemicals historically described in the literature as ‘precipitants’ (e.g. ammonium sulfate, polyethylene glycols, sodium chloride, small molecular weight alcohols), prosthetic groups, cofactors, reducing agents, etc. Thus, one out of the three or four of these ‘chemical additions’ qualifies as the ‘precipitant’, and moreover the ‘precipitant’ happens to have a concentration that is 10–100 times greater than any other component. Thus, the authors use these arguments to decide which component qualifies as the ‘precipitant’ for a given experiment.

Although ionic strength and pI are recognized as strong determiners of macromolecular crystallization, the lack of consistent reporting of these parameters make it difficult for them to be included in a statistical analysis.

1.2. Reliability test

A general approach to assess and quantify the strength of regression models is to use the *F*-test defined as follows:

$$F = \frac{\text{mean square (model)}}{\text{mean square (error)}}$$

The value of *F* should be approximately 1.0 when the null hypothesis of no association holds and should become substantially larger if the model accounts for most of the data. The probability of getting an even larger *F* value if the null hypothesis is true is also computed. If this probability is small (approximately 0.05 or less), then it is safe to reject the null hypothesis. Since the *F*-test is computed on the assumption that the residuals are normally distributed, the distribution of the residuals should be checked [14]. Consider the case of fitting a cubic model $y = \alpha + \beta_1x + \beta_2x^2 + \beta_3x^3 + \varepsilon$ to data. The *F*-value is computed according to the equation below:

$$F = \frac{(SS_3 - SS_2)(dF_3 - dF_2)}{SS_2/dF_2},$$

where SS = sum squares and dF = degrees of freedom.

The probability, $P[(F(dF_3 - dF_2), dF_2, 1 - \alpha)]$ associated with this F -value is obtained from the standard tables of statistics. The clustering process used in this study has multiple dimensions and thus uses a modified F -test (hence pseudo- F).

2. Methods

The general strategy used in analyzing the BMCD is outlined in Fig. 1. The first step involved developing programs that read specific fields of data from each entry in the BMCD. These fields consisted of the following parameters selected for clustering: molecular weight, macromolecular concentration, pH, temperature, type of precipitant and crystallization method. These parameters were divided into two groups for cluster analysis. The continuous parameters (molecular weight, macromolecular concentration, pH and temperature) were normalized to their maximum values. This resulted in each of these parameters being expressed in the range zero to one. Since each of the continuous parameters covers the same range, its influence on the distance measure between clusters is approximately the same as the others. The categorical parameters (precipitant type and crystallization method) are each represented by a tet-

rahedron with the distance between vertices set equal to one.

The PROC FASTCLUS (of the SAS software) [3] procedure for clustering was used to locate clusters in the BMCD. This procedure was also used in the previous analysis of the BMCD. The rationale of this procedure and its implementation are fully described elsewhere [1]. Using the crystallization parameters as described above, a measure of the differences between experiments was developed. Groups or clusters of similar experiments were identified as those close together based upon the difference measure. Descriptive statistics were performed on each cluster.

3. Results and discussion

3.1. Effectiveness of the clustering procedure

The number of clusters to be used was first determined by analyzing the database for appropriateness or suitability of up to 60 clusters. This is the same as asking the question 'What is the most meaningful way of sub-dividing the database into groups of similar crystallization experiments, and how many sub-groupings result?' Use of number clusters or sub-groupings less than five is generally not desirable since one or two of the clusters tend to have more than 1000 entries, and these would need to be further sub-clustered. Fig. 2 shows a plot of number of clusters versus the pseudo- F statistic. The pseudo- F statistic [3] is a measure of the success of the clustering process for a given number of clusters. The solid line with the '×' marks shows clustering from analysis of BMCD version 1.0, and the solid line with '○' marks shows the current clustering of BMCD version 3.0. Thus, the current clustering indicates that version 3.0 of the BMCD database (using the parameters chosen) is best described using 25 clusters. Further evaluation of the appropriateness of 25 clusters included the following: (1) Analysis of all the cluster groups for characteristics such as the average root-mean-square (rms) deviation within each cluster group, the average distance of the cluster group from the center of the cluster (centroid distance) and the number of experiments in each cluster group.

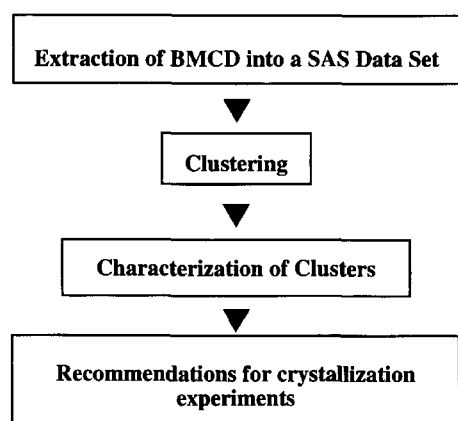


Fig. 1. Outline of steps involved in cluster analysis of the biological macromolecule crystallization database (BMCD) version 3.0.

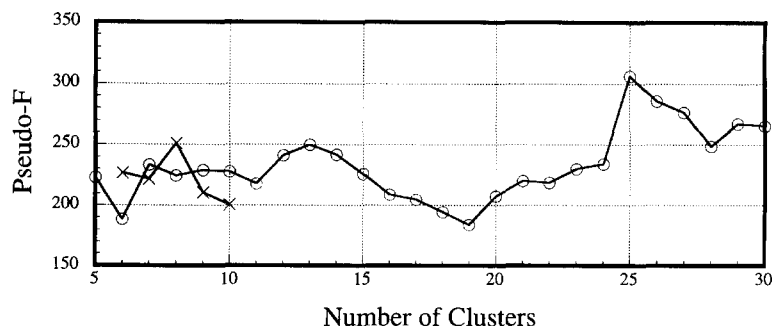


Fig. 2. A plot of the number of clusters versus the pseudo- F statistic. The peak at pseudo- F of 306.30 indicates that this database is best described using 25 clusters.

(2) Whether or not the cluster groups made any biological sense as far as similarity of macromolecules falling in the same cluster groups, crystallization experiments describing them, etc. Up to 60 numbers of clusters (or cluster groups) were evaluated for their ability to resolve the database into meaningful clusters based upon the above criteria. The value of pseudo- F is maximum for 25 clusters, and continues to decline much beyond 30 clusters. The pseudo- F statistic for 25 clusters is very statistically significant ($p < 0.0001$), but what cannot be said statistically is if 25 clusters are absolutely unique. Although Fig. 2 shows data for only up to 10 clusters from the previous analysis (with maximum pseudo- F at eight clusters), the pseudo- F value dropped after 10 clusters, and the resulting clusters could not be interpreted.

Table 1 shows the statistics of clustering for 25 clusters. The average rms deviation within each cluster is about seven times larger than in the previous cluster analysis, but the average cluster-to-cluster centroid distance is about 20 times larger than the previous study. This implies that although the individual experiments within each cluster were closer together (in our previous study), the clusters themselves were not as well resolved as in the current analysis. Clusters # 1, # 8, # 9, # 10, # 16, # 19, # 23 and # 25 contain very small numbers of experiments. A close examination of the characteristics of experiments in these clusters led the authors to treat the clusters as outliers. For example, cluster # 1 has only one crystallization experiment and it has the crystallization method

listed as 'pH-induced crystallization' (rare in the database) and the precipitant type is listed as 'ammonium bicarbonate' (also very rare), and cluster # 10 also has one experiment and it has a pH of 1.0 and a missing field (or value) for the crystallization method. Furthermore, all the macromolecules found in these outlier clusters are also found in the other 17 clusters. Although the authors concede that these rare or outlier clusters may contain useful information that may provide leads for some hard-to-crystallize macromolecules, they will not be considered for further discussion in this analysis.

Table 2 shows the distribution of macromolecules into clusters. Proteins (class I) are distributed relatively proportionally throughout most of the 17 clusters. Exceptions to this are clusters # 3 and # 15 which are dominated by viral assemblies and viral gene products (class VI). Nucleic acids (class III) and protein–nucleic acid complexes (class V) are primarily located in clusters # 17, # 18 and # 21. The few cases of nucleic acid–nucleic acid complexes (class IV) are also found in clusters # 17 and # 18. Thus, nucleic acid-containing macromolecules (class III, class IV and class V) are highly concentrated in three clusters (# 17, # 18 and # 21), and these clusters have characteristics (see below) that differ substantially from the three corresponding nucleic acid-containing clusters from the previous analysis.

Tables 3–5 show the ability of the selected crystallization parameters (i.e. molecular weight, macromolecular concentration, pH, precipitant type and crystallization method) to discriminate

Table 1
Summary of statistics of the clustering process using 25 clusters

Cluster #	Number of experiments	Rms deviation ^a	Nearest cluster	Distance between cluster centroids ^b
# 1 ^c	1	–	# 18	41.8629
# 2	25	1.4229	# 6	12.5165
# 3	22	1.1417	# 15	14.2528
# 4	13	0.7925	# 18	24.4789
# 5	14	0.9003	# 25	16.8084
# 6	369	0.6997	# 18	6.9748
# 7	19	0.8162	# 10	20.5641
# 8 ^c	4	0.7021	# 14	8.4616
# 9 ^c	2	0.7033	# 17	38.9051
# 10 ^c	1	–	# 7	20.5641
# 11	25	0.8049	# 6	19.3535
# 12	223	0.7462	# 18	7.8763
# 13	56	0.8244	# 18	13.7440
# 14	34	0.6389	# 8	8.4616
# 15	23	1.1770	# 3	14.2528
# 16 ^c	4	0.4265	# 21	34.4254
# 17	388	0.9241	# 6	8.5661
# 18	618	0.7481	# 21	6.1762
# 19 ^c	2	0	# 3	14.6220
# 20	108	0.5360	# 18	10.8788
# 21	311	0.7471	# 18	6.1762
# 22	13	0.8407	# 18	28.6859
# 23 ^c	1	–	# 3	20.1339
# 24	15	1.0903	# 18	23.3055
# 25 ^c	2	1.0050	# 5	16.8084

^aA measure of the average distance between experiments within each cluster.

^bThe distance between the center of gravity of one cluster and that of its nearest neighbor.

^cThese clusters were removed from further consideration because of the small numbers of experiments.

between experiments. For most of the experiments, the continuous variables do not generally appear to individually discriminate between experiments. Table 3 shows that exceptions occur at the high and low ends of molecular weight (clusters # 3, # 14, # 15 and # 20), the high end of macromolecular concentration (cluster # 2), the high and low ends of pH (clusters # 17 and # 18), and the low end of temperature (cluster # 21).

Although there is some discrimination, the precipitants types in the present analysis (see Table 4) do not discriminate between experiments as strongly as in the previous study. On the other hand, Table 5 shows that crystallization method is still a powerful discriminator between experiments as previously reported. The

reader should be warned that there are discrepancies between the total number of crystallization experiments per cluster reported in Table 1 and those reported in Tables 2–5. There are two reasons for this: (1) If the value for the parameter is missing in the database and the entry (or crystallization experiment) has a minimum of three other parameters present, that entry is counted as present in Table 1. (2) There are many other precipitants such as sodium chloride, sodium/potassium phosphate, dioxane, dimethyl-sulfoxide, acetone, lithium chloride and sodium acetate (to name a few) that are used in the database but are not included in Table 4. Table 4 is an attempt to show patterns of only the most commonly used precipitant types.

Table 3
Descriptive statistics for the continuous variables per cluster

Continuous variable	Descriptive data per cluster																							
	Stats ^a	#2	#3	#4	#5	#6	#7	#11	#12	#13	#14	#15	#17	#18	#20	#21	#22	#24						
Molecular weight (kDa)	N	25	21	13	14	369	19	25	220	55	34	23	374	609	106	308	13	15						
	Mean	74.3	7890.5	90.4	32.4	71.9	63.9	73.6	119.8	246.8	7.8	4801.5	65.6	83.8	9.1	62.3	13.7	27.3						
	SD	118.9	948.6	144.2	33.1	115.0	132.8	82.4	174.9	1153.7	21.5	972.8	172.2	138.0	41.1	112.6	0	16.2						
	max	587.0	9400.0	400.0	140.4	1000.0	600.0	360.0	1700.0	8400.0	103.0	6200.0	1860.0	1500.0	400.0	1500.0	13.7	70.9						
	min	9.6	6300.0	0.6	9.0	1.9	8.2	14.5	2.9	9.4	0.3	3100.0	0.7	0.3	0.3	2.1	13.7	11.0						
Macromolecular concentration (mg/ml)	N	24	17	7	11	277	12	16	171	51	6	18	313	509	19	263	13	13						
	Mean	118.8	10.5	24.7	17.3	23.6	16.3	28.6	16.0	12.9	45.0	13.6	12.9	12.9	17.6	11.3	25.7	11.2						
	SD	45.6	10.2	20.1	19.4	19.2	19.6	25.6	12.5	9.4	43.2	9.3	12.0	11.4	21.8	8.9	2.8	6.6						
	max	250.0	40.0	60.0	70.0	70.0	70.0	100.0	70.0	50.0	100.0	40.0	30.0	12.5	35.0	50.0	35.0	30.0						
	min	75.0	2.5	1.6	2.0	0.8	1.5	1.6	0.001	2.3	10.0	4.0	0.5	1.0	0.5	1.0	23.5	2.7						
pH	N	24	20	8	14	341	19	25	212	48	4	22	382	580	21	274	9	15						
	Mean	6.6	7.2	6.9	6.2	6.8	6.4	6.3	7.1	6.9	7.5	6.8	5.0	7.4	6.5	6.9	6.7	6.7						
	SD	1.2	0.9	1.3	1.2	0.7	0.9	1.1	1.0	1.0	1.9	0.9	0.8	0.8	1.0	0.9	1.6	1.8						
	max	8.2	9.0	8.3	8.0	9.0	7.5	9.0	10.0	9.0	9.5	7.8	7.4	11.0	8.0	9.5	9.5	9.0						
	min	3.5	5.0	4.7	3.7	5.3	4.1	4.3	3.0	4.2	4.8	4.2	1.8	3.7	4.6	4.5	5.0	3.5						
Temperature °C	N	19	15	12	11	306	15	12	171	41	34	17	279	458	102	304	12	10						
	4 ± 4°C	10	9	1	2	99	6	5	79	19	3	4	54	33	10	298	1	2						
	20 ± 4°C	9	6	11	8	202	9	7	92	19	31	12	218	418	92	5	11	8						

^aStats are as follows: N = total number of entries or crystallization experiments with that variable present, mean = the mean value, SD = the standard deviation, max = the maximum value observed, and min = the minimum value observed.

Table 4
Descriptive statistics for the major precipitant types per cluster

Major precipitant type	Descriptive data per cluster																							
	Stats ^a	#2	#3	#4	#5	#6	#7	#11	#12	#13	#14	#15	#17	#18	#20	#21	#22	#24						
Alcohols ^b (% conc.)	N	1	-	-	-	14	1	-	9	-	24	-	16	10	22	17	2	-						
	Mean	19.0	-	-	-	36.7	27.0	-	14.5	-	74.4	-	16.4	17.2	86.5	18.2	57.5	-						
	SD	0	-	-	-	25.2	0	-	7.6	-	10.1	-	13.7	15.6	28.3	19.4	3.5	-						
	max min	19.0 19.0	- -	- -	- -	70.0 8.0	27.0 27.0	- -	30.0 5.0	- -	100.0 70.0	- -	41.5 1.5	50.0 0.5	100.0 0.4	60 1.5	60.0 55.0	- -						
Ammonium sulfate (% saturation)	N	9	5	2	4	111	5	-	87	26	-	3	123	161	3	80	1	5						
	Mean	50.6	7.8	28.2	41.1	43.9	50.9	-	36.9	38.6	-	11.1	29.9	34.9	44.3	38.0	30.0	20.7						
	SD	20.9	6.9	8.2	34.6	19.9	18.9	-	21.4	17.9	-	1.9	18.0	20.1	14.0	20.5	0	11.4						
	max min	80.0 13.2	17.0 0.75	34.0 22.5	90.0 11.9	100.0 1.3	80.0 30.4	- -	95.0 2.2	75.0 0.7	- -	13.2 9.6	80.0 0.7	95.0 4.0	58.0 30.0	100.0 1.3	30.0 30.0	34.5 4.6						
MPD ^c (% conc.)	N	2	-	-	3	11	2	-	20	7	1	-	26	47	4	53	7	2						
	Mean	27.3	-	-	1.2	37.5	9.5	-	28.5	49.0	35.0	-	29.1	30.0	14.8	28.7	55.0	10.0						
	SD	21.6	-	-	0.9	12.5	12.0	-	20.1	4.1	-	-	23.5	19.5	10.2	18.9	0	0						
	max min	42.5 12.0	- -	- -	2.0 0.2	60.0 20.0	18.0 1.0	- -	60.0 2.0	53.0 40.0	35.0 35.0	- -	75.0 3.0	70.0 0.3	30.0 9.0	70.0 1.0	55.0 55.0	10.0 10.0						
PEG ^d (% conc.)	N	6	10	2	3	42	5	1	18	11	-	10	104	221	1	92	-	4						
	Mean	13.0	2.3	6.0	9.8	14.1	18.3	50.0	10.6	18.6	-	5.0	13.7	13.5	25.0	14.9	-	12.2						
	SD	5.4	2.2	2.8	3.7	10.1	6.1	0	7.5	9.8	-	3.9	8.9	8.6	0	10.3	-	7.1						
	max min	20.0 5.5	6.0 0.2	8.0 4.0	14.0 7.0	54.0 0.5	28.0 11.0	50.0 50.0	25.0 1.5	34.0 7.5	34.0 7.5	- -	10.0 0.4	40.0 0.03	60.0 0.01	25.0 25.0	57.0 0.1	- -	22.5 7.0					

^aStats are as follows: N = total number of entries or crystallization experiments with that variable present, mean = the mean value, SD = the standard deviation, max = the maximum value observed, and min = the minimum value observed.

^bIncludes low molecular weight alcohols such as butanol, ethanol, n-propanol, iso-propanol, glycerol and hexane-diol.

^cMPD stands for 2-methyl-2,4-pentanediol.

^dPEG stands for polyethylene glycol. Included here are all polyethylene glycols from PEG-400 to PEG-20,000.

Table 5
Distribution of crystallization methods by cluster

Crystallization method ^a	Number of crystallization experiments per cluster																							
	#2	#3	#4	#5	#6	#7 ^b	#11	#12	#13	#14	#15	#17	#18	#20	#21	#22	#24							
Bulk dialysis	1	-	-	-	-	-	-	64	-	-	-	16	-	-	-	-	-							
Microdialysis	1	4	-	-	-	-	108	-	-	-	2	19	7	-	-	-	-							
Dialysis	-	2	-	-	-	-	40	-	-	-	1	6	3	-	1	-	-							
Batch method	11	3	-	-	366	-	1	-	-	-	8	77	4	-	-	-	-							
Vapor diffusion	3	-	-	-	1	-	1	1	-	-	1	49	123	-	70	-	-							
Vapor diffusion on plates or slides	2	5	-	-	1	-	1	-	-	-	3	60	155	-	102	-	-							
Vapor diffusion in hanging drops	4	8	-	-	1	-	8	-	-	-	8	161	322	1	136	-	-							
Free interface diffusion	2	-	-	-	-	-	-	-	55	-	-	-	1	-	-	-	-							
Dialysis against distilled water	-	-	-	-	-	-	25	-	-	-	-	-	-	-	-	-	-							
Concentration by evaporation	-	-	-	-	-	-	-	-	-	-	-	-	1	107	-	-	-							
Temperature crystallization	-	-	-	-	-	-	-	-	-	34	-	-	-	-	-	-	-							
Direct addition of precipitant	-	-	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-							
Seeding techniques	-	-	-	13	-	-	-	-	-	-	-	-	-	-	-	-	15							
Freeze-thawing	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	13							

^aMcPherson [5] provides a complete description of these methods.

^bCrystallization experiments in cluster #7 have 'crystallization method' as a missing value.

3.2. Characteristics of the cluster groups

It is necessary to combine information from Tables 1–5 into some coherent description of individual clusters for the purpose of understanding general cluster patterns and trends. Thus, a presentation of the characteristics of each cluster is included below:

Cluster #2 contains mostly proteins (both enzymes and nonenzymatic) with molecular weights widely ranging from 10 to about 120 kDa. The macromolecular concentrations are the highest of any cluster, ranging from 75 to 250 mg/ml. Vapor diffusion and batch methods dominate this cluster, and the precipitants of choice are ammonium sulfate or PEG-4000, -6000 or -8000 (with approximately equal preferences). Close to neutral pH, and temperatures $4 \pm 4^\circ\text{C}$ and $20 \pm 4^\circ\text{C}$ are preferable.

Clusters #3 and #15 are dominated by viral assemblies and viral gene products. Molecular weights range from 3000 to 9000 kDa with macromolecular concentrations ranging from 3 to 14 mg/ml. Crystallization methods can be microdialysis, batch or the vapor diffusion methods. The precipitants can be PEG-6000, -8000 or ammonium sulfate (occasionally sodium chloride). Close to neutral pH and temperatures $4 \pm 4^\circ\text{C}$ and $20 \pm 4^\circ\text{C}$ are preferred.

Cluster #4 contains enzymes and nonenzymatic proteins of a wide range of molecular weights from 0.6 to 400 kDa. The macromolecular concentrations also vary widely from 1.6 to 60 mg/ml. The pH is 7.0 and temperature is almost exclusively $20 \pm 4^\circ\text{C}$. The crystallization method used here is almost exclusively direct addition of precipitants such as ammonium sulfate > any PEG > sodium chloride or ethanol.

Clusters #5 and #24 are very similar. They contain enzymes, nonenzymatic proteins with molecular weights ranging from 9 to 40 kDa. Macromolecular concentrations vary from 2 to 70 mg/ml and the pH from about 4.0 to 8.0. These clusters are dominated by seeding techniques, which means that initial crystallites are grown by other techniques and 'seeding' is then used to improve growth and quality using precipitants such as ammonium sulfate, lithium sulfate, MPD or any of

the PEGs (equal preferences). In cluster #5, the crystallization method used for the 'seeding' process can be one of a number of techniques (as listed in Table 5). However, in cluster #24 the technique used for the 'seeding' process is exclusively the vapor diffusion method.

Cluster #6 has predominantly proteins with molecular weight widely varying from 2 to 100 kDa. Macromolecular concentrations also vary widely from 1 to 70 mg/ml. The pH is mostly between 6.0 and 8.0. This cluster is dominated by the batch method. About two-thirds of the experiments were conducted at 20°C and one-third at 4°C . The order of preference for precipitants is ammonium sulfate > PEG-6000 > alcohols or Na/K phosphate.

Cluster #7 has very similar characteristics to *cluster #6*, with the exception that the crystallization method is missing in all the entries of *cluster #7*.

Cluster #11 contains mostly nonenzymatic proteins with molecular weights ranging from 15 to 360 kDa. Macromolecular concentrations vary widely from 1 to 100 mg/ml. The predominant crystallization method is dialysis against water or low ionic strength buffer with pH primarily between 5.0 and 7.0. The temperature is distributed evenly between the $4 \pm 4^\circ\text{C}$ and $20 \pm 4^\circ\text{C}$.

Cluster #12 is dominated by proteins, but also contains other macromolecular classes and, hence, molecular weight varies widely from 3 to 1700 kDa. Similarly, the macromolecular concentrations also vary widely from 0.001 to 40 mg/ml. The pH varies from 3.0 to 10.0, and the temperature is distributed evenly between the $4 \pm 4^\circ\text{C}$ and $20 \pm 4^\circ\text{C}$. The predominant crystallization methods involve dialysis (including microdialysis, bulk dialysis and dialysis). The order of preference of precipitants is ammonium sulfate > sodium chloride or phosphate > MPD > PEG-3350 or small molecular weight alcohols.

Cluster #13 has mostly proteins of molecular weight ranging from 9 to 150 kDa with macromolecular concentrations varying from 2 to 50 mg/ml. The pH is mostly around 7.0, and the temperature is evenly divided between the $4 \pm 4^\circ\text{C}$ and $20 \pm 4^\circ\text{C}$. This cluster uses almost exclusively the 'free interface diffusion' method, and the order

of preference for precipitants is ammonium sulfate > PEG-6000 > middle to high concentrations of MPD.

Cluster #14 contains exclusively nonenzymatic proteins with molecular weight ranging from 0.3 (small peptides and hormones) to 103 kDa. Macromolecular concentrations vary from 10 to 100 mg/ml and the pH between 5.0 and 9.0. This cluster uses temperature crystallization where the initial temperature is either much higher or lower than $20 \pm 4^\circ\text{C}$, but the final temperature is brought to $20 \pm 4^\circ\text{C}$. Ethanol is the most preferred precipitant (although MPD and methanol are also used).

Cluster #17 contains all the classes of macromolecules and, thus, the molecular weight widely ranges from 0.7 (small peptides and hormones) to 1900 kDa (viral assemblies). Likewise, the macromolecular concentrations also vary from 0.5 to 30 mg/ml. This cluster uses a number of crystallization methods including vapor diffusion, batch, bulk dialysis and microdialysis methods. Ammonium sulfate and PEG-6000 and -8000 are the most preferred precipitants (MPD, alcohols and sodium chloride are also used). The two most prominent features of this cluster are the low pH (mean of 5.0 ± 0.8) and high temperature preference ($20 \pm 4^\circ\text{C}$). No other cluster shows this low pH tendency as strongly as *cluster #17*.

Cluster #18 contains all the macromolecular classes, similar molecular weight range, similar macromolecular concentrations and similar temperature distribution to *cluster #17*. *Cluster #18*, however, strongly favors vapor diffusion methods and pH above 7.0. The order of preference of precipitants is PEG-6000, -4000 or -8000 > ammonium sulfate > MPD, alcohols and sodium chloride.

Cluster #20 contains mostly proteins and a few nucleic acids. The molecular weight range is from 0.3 to 400 kDa and the macromolecular concentration from 0.5 to 35 mg/ml. The pH ranges from 4.0–8.0 and temperature is predominantly high ($2 \pm 4^\circ\text{C}$). The crystallization method used here is 'concentration by evaporation', and therefore, the most preferred precipitants are mostly small molecular weight alcohols and organic solvents such

as ethanol, methanol, dioxane, acetonitrile, MPD, phenol and acetone.

Cluster #21 contains all the macromolecular classes, similar molecular weight range and similar macromolecular concentrations to *cluster #17*. The pH is about 7.0. The vapor diffusion methods dominate this cluster, and the preferential order for precipitants is PEG-4000, -6000, ammonium sulfate or MPD > small molecular weight alcohols. The most prominent feature of this cluster is low temperature ($4 \pm 4^\circ\text{C}$) preference.

Cluster #22 contains one enzyme (with a molecular weight of 13.7 kDa) which was crystallized using the 'freeze-thawing' method and varying the precipitants such as MPD, small alcohols and ammonium sulfate. The macromolecular concentrations used varied from 24 to 35 mg/ml, and the pH was also varied between 5.0 and 9.0. The final temperature after thawing is 25°C .

3.3. Recommended strategies for crystallization

Table 6 is a summary of suggested strategies based on the information from the cluster analysis. The macromolecules are organized into the same macromolecular classes as in Table 2. The proteins, however, are not further subdivided into their subclasses as in Table 2. They are, instead, divided into two categories by molecular weight: the very low molecular weight proteins and polypeptides (300–15 000 Da), and the remainder (16 000–600 000 Da). The reason is that, upon careful inspection of the data, the crystallization patterns of proteins appear to make more sense when divided into these two molecular weight groups than the functional subclasses. This gives a better picture when constructing recommendation for crystallizations.

The strategies (using Cryschem plates) that were developed in the previous report [1] for setting up crystallization experiments to explore two parameters at a time (for example, MPD versus pH) are still useful and recommended here for setting up initial experiments to explore variables recommended in Table 6. Figs. 3 and 4 are composite histograms for pH and macromolecular concentration for the summaries in Table 6. Some of the general patterns observed in our previous study were also

Table 6
Summary of overall strategy of crystallization of new macromolecules based on clustering

Macromolecular class	Outline of recommended strategies
I. Proteins ^a	
Molecular weight range: 300–12 000 Da	<p>Based on clusters # 2, # 14 and # 20</p> <ul style="list-style-type: none"> ● Crystallization method: the three methods showing great success are temperature crystallization, concentration by evaporation and vapor diffusion methods. ● Precipitants: use 20–60% (v/v) small alcohols, dioxane, DMSO with the first two methods, and 10–50% saturated ammonium sulfate with the vapor diffusion method. ● Macromolecular concentration range: 10–80 mg/ml. ● pH: explore full range from pH 4.0–9.0. ● Temperature: 20 ± 4°C is recommended.
Molecular weight range: 16 000–600 000 Da	<p>Based on clusters # 6, # 11, # 12, # 13, # 17, # 18 and # 21</p> <ul style="list-style-type: none"> ● Use batch method with high macromolecular concentrations (10–40 mg/ml), and precipitants such as 20–60% saturated ammonium sulfate, 5–30% (w/v) PEGs, 0.5–4 M NaCl or Na–K phosphate. ● Use vapor diffusion methods with 5–50% PEG-4000, -6000, -8000, or 10–60% saturated ammonium sulfate. Macromolecular concentration of 5–20 mg/ml is suitable. ● Use dialysis methods (including bulk and microdialysis) against water or low ionic strength buffer in the acidic pH range. Microdialysis against 20–80% saturated ammonium sulfate or 10–30% (w/v) sodium chloride is also recommended. ● Use the freeze-thawing (thawing to about 20°C) method with 30–60% (v/v) MPD as the precipitant. ● Use the free-interface diffusion method with 20–60% saturated ammonium sulfate or 10–30% (w/v) PEGs. ● Macromolecular concentration range: 5–30 mg/ml. ● pH: explore full range from pH 4.0–9.0. ● Temperature: explore full range from 4 ± 4–20 ± 4°C.
II. Protein–protein complexes Includes inhibitor complexes and other protein–protein association where the two or more associating polypeptides are not ordinarily part of the same molecule. Molecular weight range: 12 000–440 000 Da.	<p>Based on clusters # 6, # 17, # 18 and # 21</p> <ul style="list-style-type: none"> ● Crystallization method: the two methods equally successful are the vapor diffusion in hanging drops and the batch methods. ● Precipitants: use the following for either method; 10–60% saturated ammonium sulfate or 5–40% (w/v) PEG-4000 and -6000. ● Macromolecular concentration range: 4–30 mg/ml (use stoichiometric amounts). ● pH: explore full range from pH 4.0–9.0. ● Temperature: explore full range from 4 ± 4–20 ± 4°C.
III. Nucleic acids Includes DNA and RNA oligonucleotides. Molecular weight range: 800–34 000 Da.	<p>Based on clusters # 17, # 18, # 20 and # 21</p> <ul style="list-style-type: none"> ● Crystallization method: the two methods equally successful are the vapor diffusion on plates or slides and concentration by evaporation. ● Precipitants: use the following for either method; 10–50% (v/v) MPD, 30–90% (v/v) small molecular weight alcohols (see Table 4), dioxane, acetone and 10–50% saturated ammonium sulfate.

Table 6
Continued

Macromolecular class	Outline of recommended strategies
	<ul style="list-style-type: none"> ● Macromolecular concentration range: 3–10 mg/ml. ● pH: use acidic to neutral range, i.e. from pH 4.0–7.5. ● Temperature: focus on $20 \pm 4^\circ\text{C}$, also explore full range.
IV. Nucleic acid–nucleic acid complexes ^b Includes DNA–RNA hybrids. Molecular weight range: 6 000–10 000 Da.	<p>Based on clusters # 17 and # 18</p> <ul style="list-style-type: none"> ● Crystallization method: vapor diffusion methods on plates or slides. ● Precipitants: MPD, use 20–50% (v/v) concentrations. ● Macromolecular concentration range: 5–20 mg/ml. ● pH: explore full range but focus around pH 7.0. ● Temperature: first attempt $20 \pm 4^\circ\text{C}$, then explore full range.
V. Protein–nucleic acid complexes: Includes protein–DNA and protein–RNA complexes. Molecular weight range: 13 600–88 000 Da.	<p>Based on clusters # 17, # 18 and # 21</p> <ul style="list-style-type: none"> ● Crystallization method: vapor diffusion methods on plates or slides and hanging drops. ● Precipitants: use 5–20% (w/v) concentrations of PEG–4000 and –6000. (MPD, ammonium sulfate and sodium chloride worth trying). ● Macromolecular concentration range: 5–30 mg/ml (use excess nucleic acid stoichiometrically). ● pH: use acidic to neutral range, i.e. from pH 4.0–7.5. ● Temperature: explore full range from 4 ± 4–$20 \pm 4^\circ\text{C}$.
VI. Viral assemblies and viral gene products: Molecular weight range: 3 000–10 000 kDa.	<p>Based on clusters # 3 and # 15</p> <ul style="list-style-type: none"> ● Crystallization method: microdialysis, batch methods and vapor diffusion methods. ● Precipitants: ammonium sulfate and PEGs, use 1–15% (w/v) concentrations. ● Macromolecular concentration range: 2–20 mg/ml. ● pH: explore full range from pH 4.0–9.0. ● Temperature: explore full range, but focus on $4 \pm 4^\circ\text{C}$ and $20 \pm 4^\circ\text{C}$.

^aThe proteins are divided into two molecular weight groups instead of subclasses as in Table 2.

^bThere is a paucity of data on crystallization of nucleic acid complexes.

observed in this analysis: e.g., the tendency of lower molecular weight proteins to have high macromolecular concentrations (20–80 mg/ml) and use MPD or small alcohols as precipitants; or the very high molecular weight clusters (clusters #3 and #15 dominated by viral assemblies) to have lower macromolecular concentrations. Also, the distribution of pH values for all experiments in the database still shows Gaussian character (with peak at pH 7.0, see Fig. 3) and the temperature distribution is still bimodal showing peaks around $4 \pm 4^\circ\text{C}$ and $20 \pm 4^\circ\text{C}$ (data not shown).

There are, however, some new features that were not as clear before: (1) There is a significant increase in the successful use of crystallization methods such as temperature crystallization, seeding techniques, freeze-thawing techniques and concentration by evaporation, and (2) Data also show an increase in the use of precipitants such as Na/K phosphate, acetone, dioxane, dimethyl-sulfoxide, lithium chloride, glycerol, potassium sulfate, sodium sulfate and ammonium chloride used in all crystallization techniques and for all macromolecules.

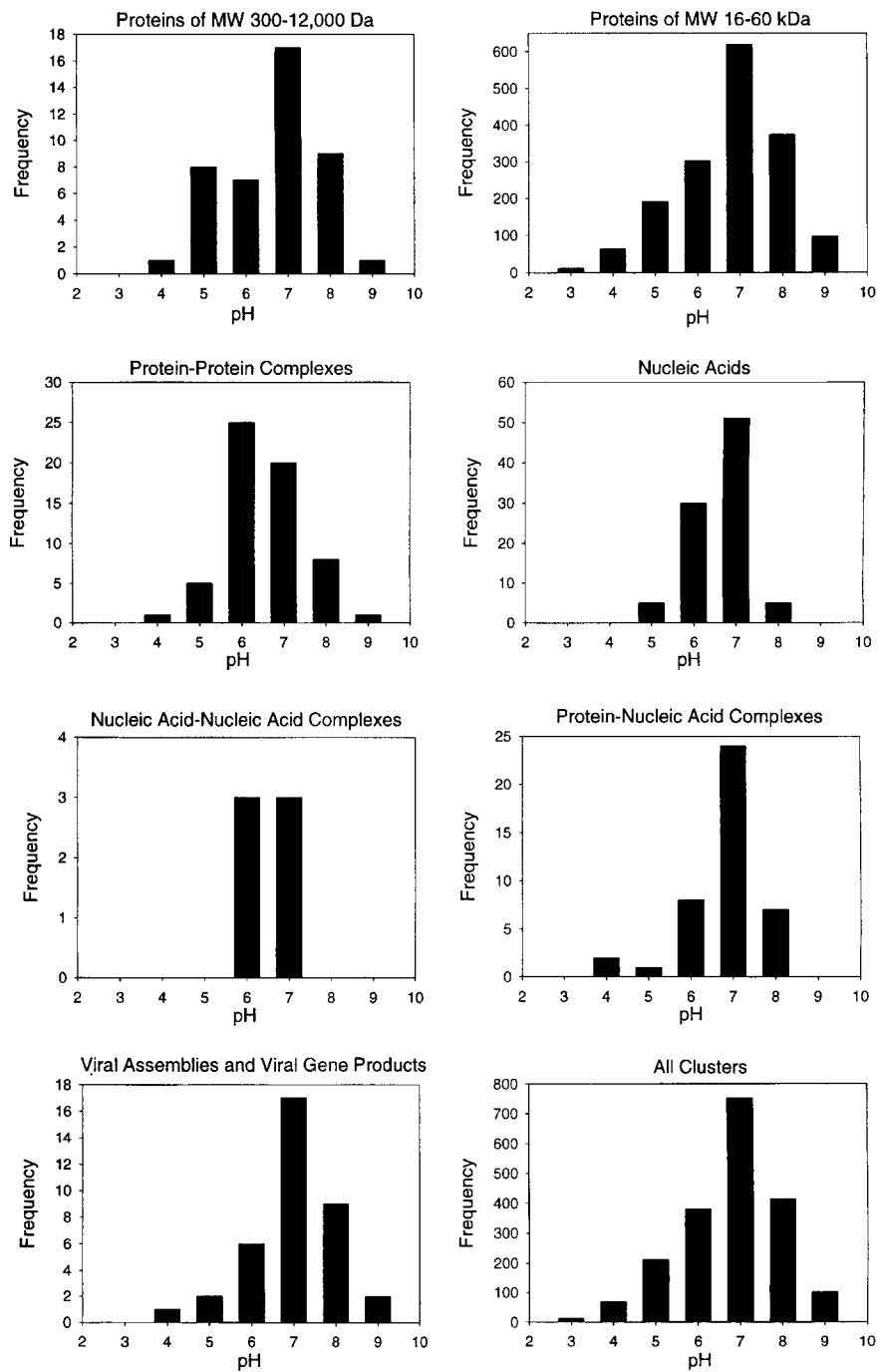


Fig. 3. Histograms showing the distribution of pH values within each macromolecular class. pH values are at midpoints of the intervals

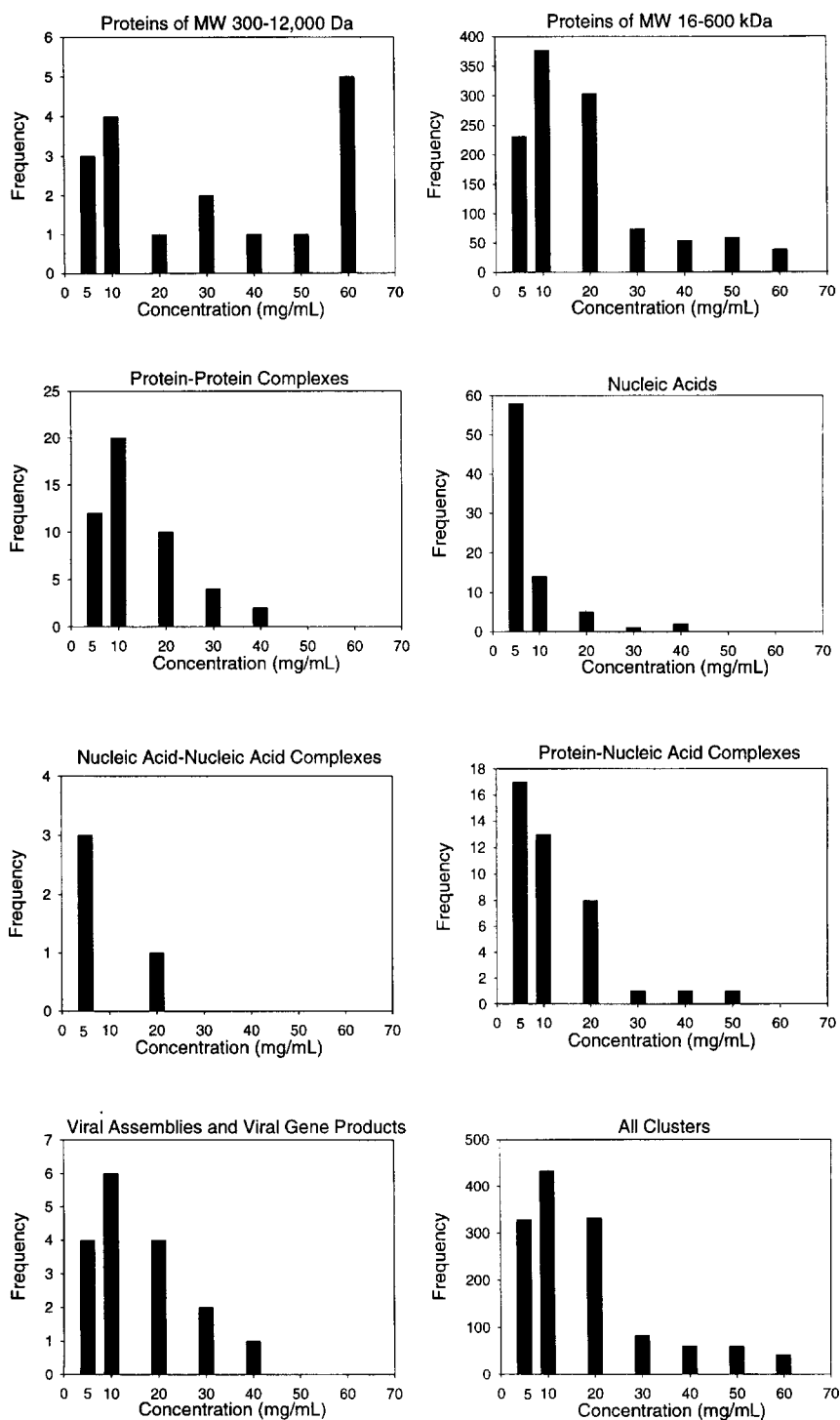


Fig. 4. Histograms showing the distribution of macromolecular concentration values within each macromolecular class. Macromolecular concentration values from 0.0 to 5.0 mg/ml are represented by the bar at 5 mg/ml; concentrations from 5.1 to 10.0 mg/ml are represented by the bar at 10 mg/ml, and so forth. Concentrations greater than 50.0 mg/ml are represented by the bar at 60 mg/ml.

4. Summary

The re-clustering of the updated BMCD version 3.0 was performed using the PROC FASTCLUS [3] procedure as before. The clustering statistics indicate that the database was well resolved into 25 statistically significant clusters. Analysis of these clusters shows patterns similar to the previous study as well as significant differences. Practical suggestions are made based on the analyses of the clusters.

Acknowledgements

The authors would like to thank Matthew J. Fivash (statistician) at the Frederick Cancer Research and Development Center for his valuable discussions and suggestions. This research was supported by the University of Missouri Research Board.

References

- [1] C.T. Samudzi, M.J. Fivash, J.M. Rosenberg, *J. Crystal Growth* 123 (1992) 47.
- [2] G.L. Gilliland, M. Tung, D.M. Bakerslee, J.E. Ladner, *Acta Cryst. D* 50 (1994) 408.
- [3] SAS Institute Inc., SAS/STAT, Version 6, 4th ed., vol. 1.
- [4] C.W. Carter Jr., C.W. Carter, *J. Biol. Chem.* 254 (1979) 12219.
- [5] A. McPherson Jr., *Meth. Biochem. Anal.* 23 (1976) 249.
- [6] A. McPherson, *Preparation and Analysis of Protein Crystals*, Wiley, New York, 1982.
- [7] G.L. Gilliland, *J. Crystal Growth* 90 (1988) 51.
- [8] A. McPherson, in: H. Michel (Ed.), *Crystallization of Membrane Proteins*, CRC Press, Florida, 1991, p. 1.
- [9] D. Hennessy, V. Gopalakrishnan, B.G. Buchanan, J.M. Rosenberg, D. Subramanian, *ISMB 94* (1994) 179.
- [10] A.J. Malkin, Y.G. Kuznetsov, T.A. Land, J.J. DeYoreo, A. McPherson, *Nature Struct. Biol.* 2 (1995) 956.
- [11] T. Cole, A. Kathman, S. Koszelak, A. McPherson, *Anal. Biochem.* 231 (1995) 92.
- [12] S.D. Durbin, G. Feher, *Ann. Rev. Phys. Chem.* 47 (1996) 171.
- [13] C.M. Roth, B.L. Neal, A.M. Lenhoff, *Biophys. J.* 70 (1996) 977.
- [14] D.G. Kleinbaum, L.L. Kupper, *Applied Regression Analysis and Other Multivariable Methods*, Duxbury Press, North Scituate, MA, 1978, p. 235.