J Struct Funct Genomics (2014) 15:13–24
DOI 10.1007/s10969-014-9171-4

Carboxylic acids in crystallization of macromolecules: learning from successful crystallization experiments

Lesa R. Offermann · John Z. He · Nicholas J. Mank · William T. Booth II · Maksymilian Chruszcz

Received: 18 November 2013 / Accepted: 13 January 2014 / Published online: 23 January 2014

Abstract The production of macromolecular crystals suitable for structural analysis is one of the most important and limiting steps in the structure determination process. Often, preliminary crystallization trials are performed using hundreds of empirically selected conditions. Carboxylic acids and/or their salts are one of the most popular components of these empirically derived crystallization conditions. Our findings indicate that almost 40 % of entries deposited to the Protein Data Bank (PDB) reporting crystallization conditions contain at least one carboxylic acid. In order to analyze the role of carboxylic acids in macromolecular crystallization, a large-scale analysis of the successful crystallization experiments reported to the PDB was performed. The PDB is currently the largest source of crystallization data, however it is not easily searchable. These complications are due to a combination of a free text format, which is used to capture information on the crystallization experiments, and the inconsistent naming of chemicals used in crystallization experiments. Despite these difficulties, our approach allows for the extraction of over 47,000 crystallization conditions from the PDB. Initially, the selected conditions were investigated to determine which carboxylic acids or their salts are most often present in crystallization solutions. From this group, selected sets of crystallization conditions were analyzed in detail, assessing parameters such as concentration, pH, and precipitant used. Our findings will lead to the design of new crystallization screens focused around carboxylic acids.

Keywords Macromolecular crystallization · Carboxylic acids · Data mining · Protein Data Bank analysis

Abbreviations

PDB	Protein Data Bank
RD	Redundant dataset
nRD	Non-redundant dataset
PEG	Polyethylene glycol

Introduction

Structural biology and the structural determination process still remain primarily dependent on the formation of highly diffracting crystals [1]. The process leading to successful crystal formation, especially preliminary crystallization trials, is predominately empirical. Traditionally, crystallographers have employed methods of factorial screening [2], where a limited number of conditions are tested, or sparse-matrix screening [3] which offers a wide variety of conditions containing various pHs, buffers, precipitating agents, and additives. These methods of trial-and-error are made easier with many of commercially available crystallization screens, provided by several different companies, which come in a few different formats. These screens are based on successful experiments that provided macromolecular crystals suitable for structural analysis. Specialized screens are also available which specifically target nucleic acids, proteins, or protein complexes. In addition to

Electronic supplementary material The online version of this article (doi:10.1007/s10969-014-9171-4) contains supplementary material, which is available to authorized users.

L. R. Offermann · J. Z. He · N. J. Mank · W. T. Booth II · M. Chruszcz (✉)
Chemistry and Biochemistry, University of South Carolina, 631 Sumter Street, Columbia, SC 29208, USA
e-mail: chruszcz@mailbox.sc.edu

the increase in the number of available screens, the development of new methods for crystallization has also aided in alleviating some of the bottleneck congestion. Methods like counter diffusion [4], microfluidic crystallography [5], and nanovolume microcapillary protein crystallization [6] are providing scientists with alternative means of obtaining crystals in an effort to enhance efficiency. Furthermore, the technology associated with these techniques is also evolving.

Carboxylic acids and their salts constitute a large portion of successful crystallization experiments as reported to the Protein Data Bank (PDB) [7]. Furthermore, they were also shown to be successful cryoprotectants [8, 9]. Reported here is the extraction of such successful experiments and an in depth analysis of their components. This report, like the MORPHEUS protein crystallization screen report for example [10], originated from conditions derived from the PDB. Unlike the MORPHEUS screen however, our report is solely focused around carboxylic acids. Our efforts were to uncover trends that can be used to further alleviate the limiting step in structure determination. It is our hope to eventually develop a crystallization screen based on these findings where carboxylic acids are the main focus. Data mining has been used in the past in a similar approach to reveal trends in compositions of crystallization conditions [11, 12]; it is to our knowledge that carboxylic acids have not specifically been reviewed.

Experimental methods

The May 22, 2012 PDB release contained a total of 71,692 protein structures determined solely by X-ray. Those entries, in the mmcif format, were downloaded and statistically analyzed. For this study, the total number of deposited X-ray structures (71,692) is referred to as the redundant data set (RD). A sub-set of the RD, termed the non-redundant data set (nRD), was generated after the removal of similar sequences using 90 % sequence identity as the cut-off. The nRD is composed of 26,313 PDB entries. Crystallization conditions, as reported in the _exptl_crystal_grow.pdbx_details records, were extracted from the analyzed deposits. For this analysis, 'useful data' was considered to be those records containing information on chemical composition for the crystallization solution, which include chemical names, concentration, and/or pH. A similar approach was used for data acquisition to the Biological Macromolecular Crystallization Database [13]. All records that did not contain any information on the chemical composition were omitted. The final sets used for the analysis were comprised of 47,783 and 18,300 crystallization conditions for the RD and nRD, respectively.

Both the RD and the nRD were further analyzed and sub-categorized by carboxylic acid. The carboxylic acids of interest for this study were: acetate, citrate, formate, tartrate, malonate, malate, succinate, fumarate and oxalate. Very few entries contained fumarate or oxalate, therefore they were combined and categorized as 'other'. Additionally, entries containing Tacsimate^{TM}, a crystallization reagent developed by Hampton Research Corp. [14–16] were examined. Tacsimate^{TM}, available in pH 4, 5, 6, 7, 8, or 9, is composed of: 1.8 M malonic acid, 0.25 M ammonium citrate tribasic, 0.12 M succinic acid, 0.3 M DL-malic acid, 0.4 M sodium acetate trihydrate, 0.5 M sodium formate, and 0.16 M ammonium tartrate dibasic.

Concentration statistics were grouped into 'low', 'medium', and 'high' categories. The 'low' range represents concentrations of 100 mM or less. The 'medium' group contains concentrations greater than 100 mM up to 1 M, and the 'high' group contains concentrations greater than 1 M. Furthermore polyethylene glycols (PEGs) were also analyzed and divided into categories according to their average molecular weight. The 'low MW' category contained PEGs with average molecular weights <2,000. The 'medium MW' category contained PEGs between 2,000 and 6,000, and the 'high MW' category contained PEGs with weights >6,000. In all analyses the 'unknown' category contained results with insufficient or indeterminable data.

All records were obtained and analyzed using a combination of semi-automatic and manual approaches. The semi-automatic data was analyzed using the grep command in Linux. The manual data mining was quite involved due to the PDB's use of the free text format to collect information on the crystallization experiments. This format presented a challenge in accurate determination of a number of conditions due to the inconsistent naming of chemicals used in crystallization experiments.

Of all X-ray entries deposited to the PDB, the nRD was taken into account to avoid bias among entries. Statistics for the RD and nRD are quite comparable; therefore in this report more detailed investigations of the carboxylic acids are directed towards the nRD.

Results and discussion

Global statistics

Initially, the data collected from the PDB were separated into the RD and nRD. Out of all X-ray entries deposited to the PDB, as of May 22, 2012, 66.7 % of entries contained 'useful data' as reported in the _exptl_crystal_grow.pdbx_details records. In the nRD, where similar sequences were removed using 90 % sequence identity as the cut-off parameter, 69.5 % of entries contained 'useful data'. From this analysis 37.1 % of the entries in the RD and 37.2 % of entries in the

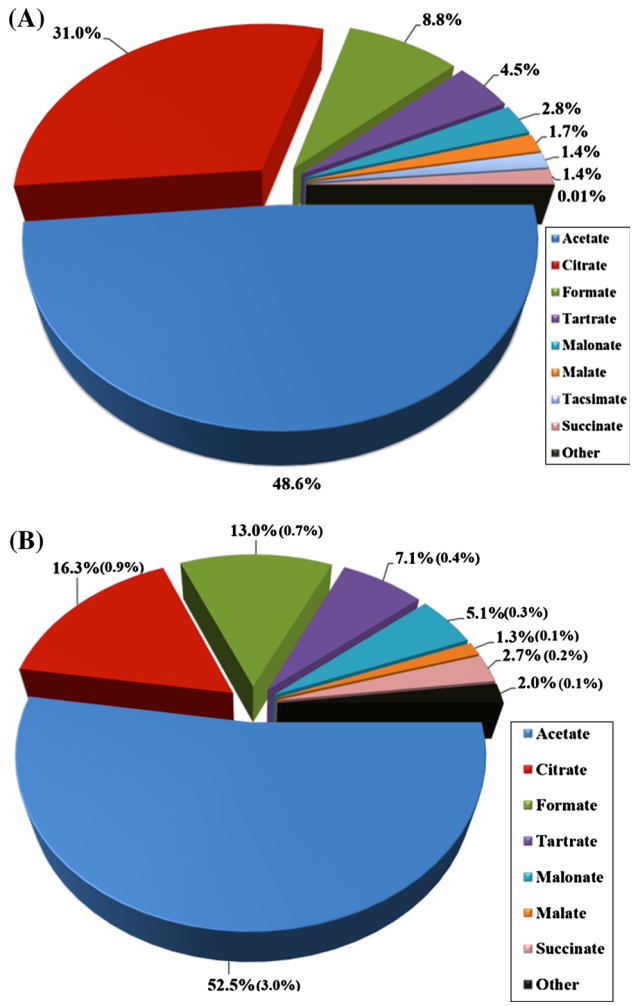

Fig. 1 Percentages of carboxylic acids found in the non-redundant data set, and the percentage of carboxylic acids found as ligands. **a** Statistics that correspond with the nRD. In all cases 'Other' refers to fumarate and oxalate. The non-redundant data set indicates that acetate is the most commonly reported carboxylic acid used in crystallization, followed by citrate, then formate. **b** Percentages of carboxylic acids found as ligands reported to the Protein Data Bank. Of the entries where a carboxylic acid was bound as a ligand, the statistical analysis of each individual carboxylic acid from that subset. Values in parentheses represent the percentage of the entire PDB that contained the specific carboxylic acid as a ligand

nRD contained a carboxylic acid and/or its salt. All entries containing 'useful data' and a carboxylic acid in both the RD and nRD were used for further analysis.

The RD and nRD were further investigated individually. In the RD, acetate was the most prevalent carboxylic acid at 50.8 %. Citrate was the second most frequently used carboxylic acid at 30.6 %, while formate was the third most commonly used carboxylic acid at 7.7 %. The prevalence of the remaining carboxylic acids investigated were all below 5 %. Additionally, Tacsimate™, while not a specific carboxylic acid, was analyzed and accounted for 1.7 % of entries containing carboxylic acids (Online Resource 1). The same statistical analysis was completed for the nRD with very similar results. Acetate accounted for 48.8 % of entries, 30.9 % contained citrate, and 8.7 % contained formate. Furthermore, in the nRD, Tacsimate™ accounted for 1.4 % of all entries containing a carboxylic acid (Fig. 1a). It is apparent that acetate, citrate, and formate are the most commonly used carboxylic acids in this analysis. This may be due to the fact that acetate and citrate are commonly used buffers in crystallization conditions. However, this simple explanation cannot be used in case of formate. Even though Tacsimate™ is composed of a mixture of carboxylic acids, it is not frequently listed in the crystallization conditions reported to the PDB (1.4 % of conditions containing a carboxylic acid in the nRD). A reason for this could be that it is used as an initial reagent for crystal growth, and then conditions are optimized further, including only one or two of the carboxylic acids from Tacsimate™.

Statistical analysis was also conducted on X-ray entries deposited to the PDB (March 20, 2013), where a carboxylic acid was found as a ligand. It was found that 4,495 of the 78,412 entries to the PDB had a carboxylic acid and/or its salt bound as a ligand. Overall 3.0 % of entries contained acetate as a ligand, while citrate and formate were found bound as a ligand 0.9 and 0.7 %, respectively. It shows that despite the fact that carboxylic acids are very popular components of crystallization conditions, they are relatively rarely observed in crystal structures. In this subset, acetate accounted for 52.5 % of the entries containing a carboxylic acid as a ligand, and this percentage is very similar to the observed frequency of acetate in crystallization conditions. However, citrate and formate are observed in 16.3 and 13.0 % of entries with a carboxylic acid as a ligand (Fig. 1b), which significantly differs from their distributions in reported crystallization conditions. Clearly, citrate is underrepresented and other compounds including formate, tartrate, malonate and succinate are significantly overrepresented. One possible explanation of this observation may be related to the fact that citrate is larger and is more charged than other analyzed carboxylic acids.

Concentration and pH

Concentrations of the carboxylic acids were analyzed in the nRD without any filters, and without taking into account counter ions (Fig. 2a). The concentrations were categorized into 'low', 'medium', and 'high'. 'Low' represents concentration values up to, and including 100 mM, 'medium' represents concentration values >100 mM up to 1 M, and 'high' concentrations are greater than 1 M. For this analysis, Tacsimate™ was omitted. Acetate was most

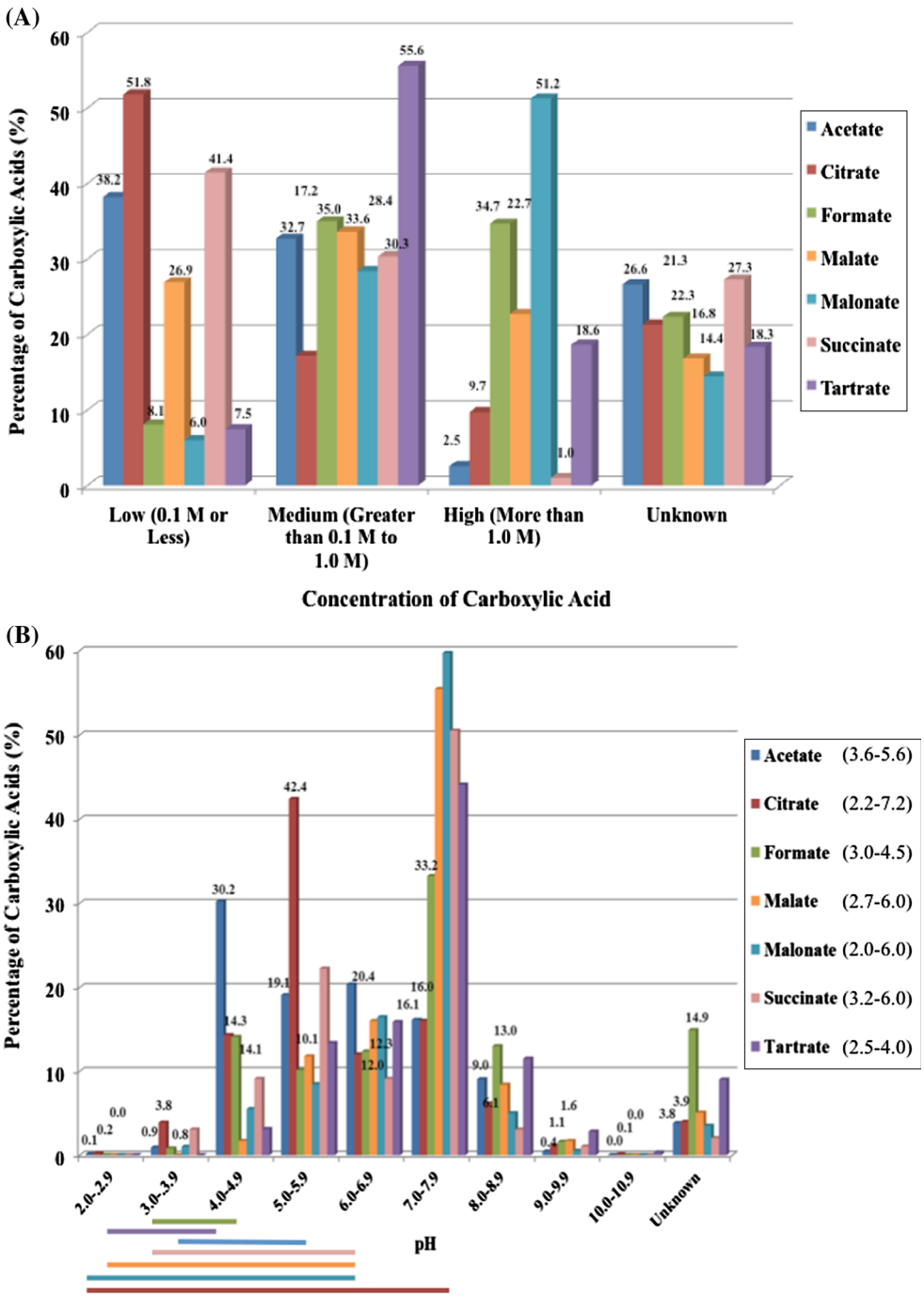

Fig. 2 Analysis of the concentration ranges and pH ranges of carboxylic acids in the non-redundant dataset. **a** Concentration ranges of carboxylic acids where 'low' represents values of 0.1 M or less, 'medium' contains values >0.1 M up to 1 M, and 'high' contains concentrations greater than 1 M. **b** pH ranges of the crystallization condition used that contains a carboxylic acid. For clarity, only values for acetate, citrate, and formate are displayed. *Bars* at the bottom of the figure and values in parenthesis following the acid indicate buffering ranges

commonly used at low and medium concentrations at 38.2 and 32.7 % respectively. Citrate was most prevalent at low concentration, which accounted for 51.8 % of citrate conditions, which again suggests that citrate is used mainly as a buffering agent. Formate was frequently found at medium and high concentrations at 35.0 and 34.7 % respectively. This observation, in the case of formate, is in agreement with results obtained by Radaev et al. [17] of conditions used for crystallization of protein complexes. It was found that not only is sodium formate one of the more common compounds used for crystallization of macromolecular complexes, but it is also used at high concentrations. This observation may be explained by the significantly higher solubility of formate salts in comparison with corresponding acetate, citrate and tartrate salts. Succinate was most commonly found at low concentrations, 41.4 %, which most likely is related to relatively low solubility of succinic acid and its salts. Malate and tartrate were most frequently found at medium concentration, 33.6 and 55.6 % respectively, and malonate was predominately used

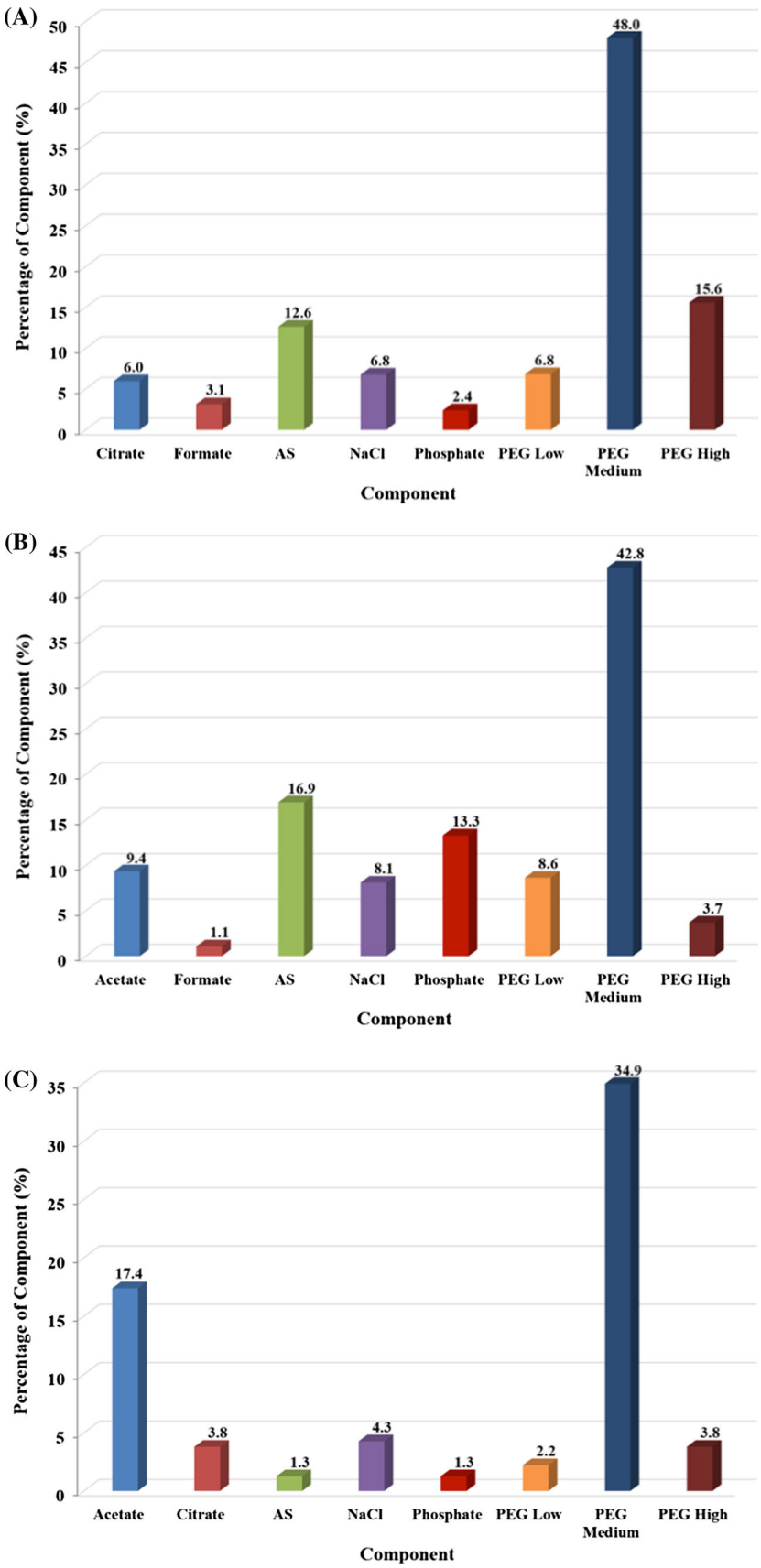

Fig. 3 Percentages of commonly used additives in conjunction with **a** Acetate. **b** Citrate. **c** Formate

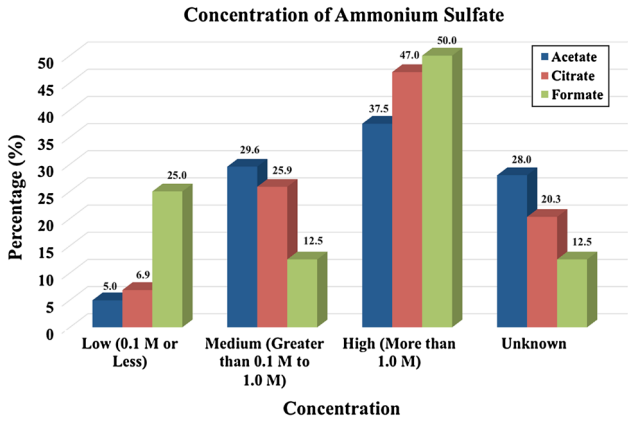

Fig. 4 Analysis of ammonium sulfate concentration ranges where 'low' represents values of 0.1 M or less, 'medium' contains values >0.1 M up to 1 M, and 'high' contains concentrations >1 M in combination with acetate (*blue*), citrate (*red*), or formate (*green*)

at high concentration, 51.2 %. Unfortunately, the unknown category represents 14.4–27.3 % of conditions where the data was insufficient or indeterminable.

Buffers utilized for crystallization conditions are typically used in the medium to low concentration range. If the carboxylic acids studied in this report were used only as buffers, this might explain why their concentrations are <1 M. The carboxylic acids, which were found in the high category, however, might be due to the carboxylic acid being used as a precipitant salt. The solubility of the carboxylic acid salts fall into the 'high' category (over 1 M), with the exception of succinate, therefore most of the carboxylic acids found in the medium to low concentration ranges are not related to solubility.

The nRD was also analyzed based on pH ranges, without the use of filters (Fig. 2b). The most commonly reported pH range for conditions containing acetate was 4.0–4.9. This range accounts for 30.2 % of all conditions containing acetate. 42.4 % of all conditions containing citrate were reported in the 5.0–5.9 pH range. These two carboxylic acids are typically reported within their buffering ranges and may indicate that both acetate and citrate, when used in a crystallization condition, are typically used as a buffer. Formate, malate, malonate, succinate, and tartrate are all

Table 1 Analysis of the concentration of ammonium sulfate in combination with the concentration of (**A**) Acetate (**B**) Citrate, and (**C**) Formate

A.

		Ammonium Sulfate			
		Low (0.1 M or less)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1.0 M)	Unknown
Acetate	Low (0.1 M or less)	4.3%	28.4%	33.9%	0.5%
	Medium (Greater than 0.1 M to 1.0 M)	0.5%	0.7%	2.5%	0.0%
	Higher (Greater than 1.0 M)	0.2%	0.0%	0.7%	0.0%
	Unknown	0.0%	0.5%	0.5%	27.5%

B.

		Ammonium Sulfate			
		Low (0.1 M or less)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1.0 M)	Unknown
Citrate	Low (0.1 M or less)	3.4%	22.8%	44.4%	0.3%
	Medium (Greater than 0.1 M to 1.0 M)	3.4%	0.8%	2.9%	0.0%
	Higher (Greater than 1.0 M)	0.0%	0.3%	0.3%	0.0%
	Unknown	0.0%	0.3%	0.8%	20.4%

C.

		Ammonium Sulfate			
		Low (0.1 M or less)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1.0 M)	Unknown
Formate	Low (0.1 M or less)	0.0%	12.5%	25.0%	0.0%
	Medium (Greater than 0.1 M to 1.0 M)	12.5%	0.0%	12.5%	0.0%
	Higher (Greater than 1.0 M)	12.5%	0.0%	12.5%	0.0%
	Unknown	0.0%	0.0%	0.0%	12.5%

Cells highlighted in blue indicate combinations >10 %, while cells highlighted in green represent the unknown category of one component in combination with the unknown category of another component that are >10 %

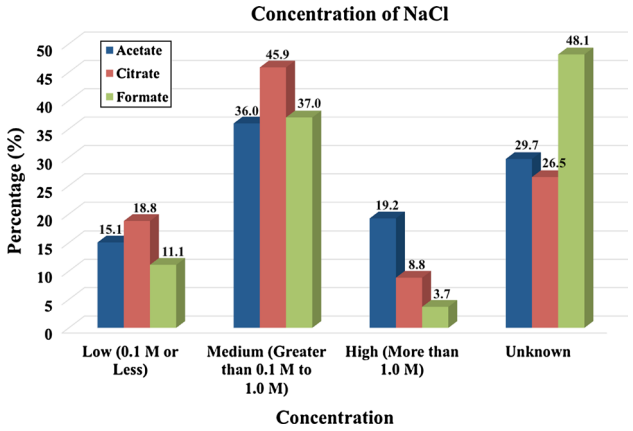

Fig. 5 Analysis of sodium chloride concentration ranges where 'low' represents values of 0.1 M or less, 'medium' contains values >0.1 M up to 1 M, and 'high' contains concentrations >1 M in combination with acetate (*blue*), citrate (*red*), or formate (*green*)

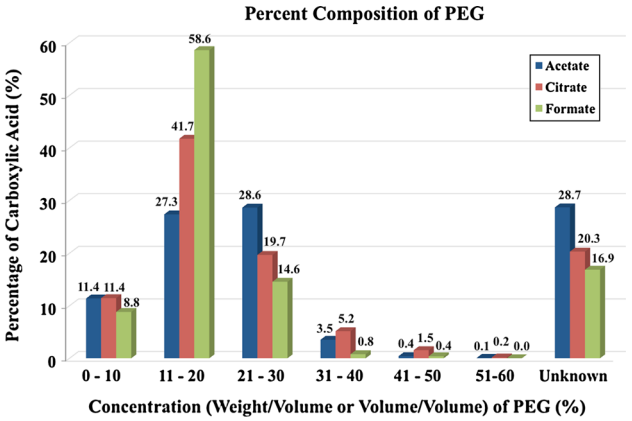

Fig. 6 Analysis of all polyethylene percentage ranges (weight/volume or volume/volume) in combination with acetate (*blue*), citrate (*red*), or formate (*green*)

Table 2 Analysis of the concentration of sodium chloride in combination with the concentration of (**A**) Acetate (**B**) Citrate, and (**C**) Formate

A.			**Sodium Chloride**		
		Low (0.1 M or less)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1.0 M)	Unknown
Acetate	Low (0.1 M or less)	12.7%	29.5%	17.3%	0.4%
	Medium (Greater than 0.1 M to 1.0 M)	1.7%	5.5%	0.4%	0.0%
	Higher (Greater than 1.0 M)	0.4%	0.0%	0.0%	0.0%
	Unknown	0.0%	0.8%	1.7%	29.5%

B.			**Sodium Chloride**		
		Low (0.1 M or less)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1.0 M)	Unknown
Citrate	Low (0.1 M or less)	15.5%	33.1%	8.3%	0.6%
	Medium (Greater than 0.1 M to 1.0 M)	2.2%	9.9%	0.6%	0.0%
	Higher (Greater than 1.0 M)	1.1%	2.2%	0.0%	0.0%
	Unknown	0.0%	0.6%	0.0%	24.3%

C.			**Sodium Chloride**		
		Low (0.1 M or less)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1.0 M)	Unknown
Formate	Low (0.1 M or less)	3.7%	11.1%	0.0%	0.0%
	Medium (Greater than 0.1 M to 1.0 M)	0.0%	11.1%	0.0%	0.0%
	Higher (Greater than 1.0 M)	7.4%	14.8%	3.7%	0.0%
	Unknown	0.0%	0.8%	0.0%	48.1%

Cells highlighted in blue indicate combinations >10 %, while cells highlighted in green represent the unknown category of one component in combination with the unknown category of another component that are >10 %

most commonly reported within the 7.0–7.9 pH range, 33.2, 55.5, 59.7, 50.5, and 44.1 % respectively. This occurrence may be due to conditions being prepared at physiological pH. Unknown conditions were much less common, 14.9 % or less, than what were determined for concentration. Moreover, pH values lower than 4, or higher

Table 3 Analysis of the molecular weight ranges of polyethylene glycol in combination with the concentration of (**A**) Acetate **(B)** Citrate, and (**C**) Formate

A.

		PEG			
		Low (Less than 2000)	Medium (2000-6000)	High (Greater than 6000)	Unknown
Acetate	Low (0.1 M or less)	4.4%	22.3%	5.2%	0.2%
	Medium (Greater than 0.1 M to 1.0 M)	2.2%	30.0%	9.3%	0.0%
	High (Greater than 1.0 M)	0.0%	0.2%	0.1%	0.0%
	Unknown	3.0%	15.4%	7.5%	0.1%

B.

		PEG			
		Low (Less than 2000)	Medium (2000-6000)	High (Greater than 6000)	Unknown
Citrate	Low (0.1 M or less)	10.0%	44.7%	4.7%	0.3%
	Medium (Greater than 0.1 M to 1.0 M)	2.5%	17.3%	0.3%	0.0%
	High (Greater than 1.0 M)	0.2%	0.1%	0.0%	0.0%
	Unknown	3.1%	15.0%	1.8%	0.1%

C.

		PEG			
		Low (Less than 2000)	Medium (2000-6000)	High (Greater than 6000)	Unknown
Formate	Low (0.1 M or less)	0.8%	13.0%	0.0%	0.4%
	Medium (Greater than 0.1 M to 1.0 M)	1.9%	56.3%	6.9%	0.4%
	High (Greater than 1.0 M)	0.8%	0.8%	0.8%	0.0%
	Unknown	1.9%	14.6%	1.5%	0.0%

Cells highlighted in blue indicate combinations >10 %

than 8 were far less frequently observed. Results of our analysis also demonstrate that the carboxylic acids are mainly used in the 4–9 pH range, and the overall distribution is similar to this observed range for all proteins reported in PDB [18].

Compounds most often seen in combination with acetate, citrate, and formate

As previously mentioned, acetate, citrate, and formate are the most prevalent carboxylic acids used as reported to the PDB. This report further investigates commonly used precipitating reagents used in combination with each of the carboxylic acids. A routinely used precipitating reagent is ammonium sulfate and was found in 12.6, 16.9 and 1.3 % of conditions that contained acetate, citrate, and formate respectively (Fig. 3a–c). Entries that contained ammonium sulfate were then filtered by acetate, citrate, or formate (Fig. 4). Typically, the concentration of ammonium sulfate used in combination with acetate, citrate, or formate is higher than 1 M. The second most frequent combination, between ammonium sulfate and acetate or citrate, was at medium or unknown concentrations of ammonium sulfate. Formate and low concentrations of ammonium sulfate however, are seen 25 % of the time, which is the second most frequent combination. Moreover, medium concentrations of ammonium sulfate and those entries with unknown values in conjunction with formate are equal at 12.5 %. Acetate, citrate, and formate concentrations were compared to the concentration of ammonium sulfate; results are shown in Table 1.

Sodium chloride, another salt used in crystallization solutions, was found in 6.8, 5.0, and 4.3 % of conditions that contained acetate, citrate, or formate respectively (Fig. 3a–c). Entries that contained sodium chloride were then filtered and reanalyzed to show the relative percentages of acetate, citrate, or formate. Disregarding the statistics for the unknown category, acetate, citrate, and formate follow the same trends. Typically, sodium chloride at medium concentration with each of the carboxylic acids is the most popular combination. The second most popular category is in combination with low concentrations of sodium chloride and the least favored combination are the carboxylic acids with high concentrations of sodium

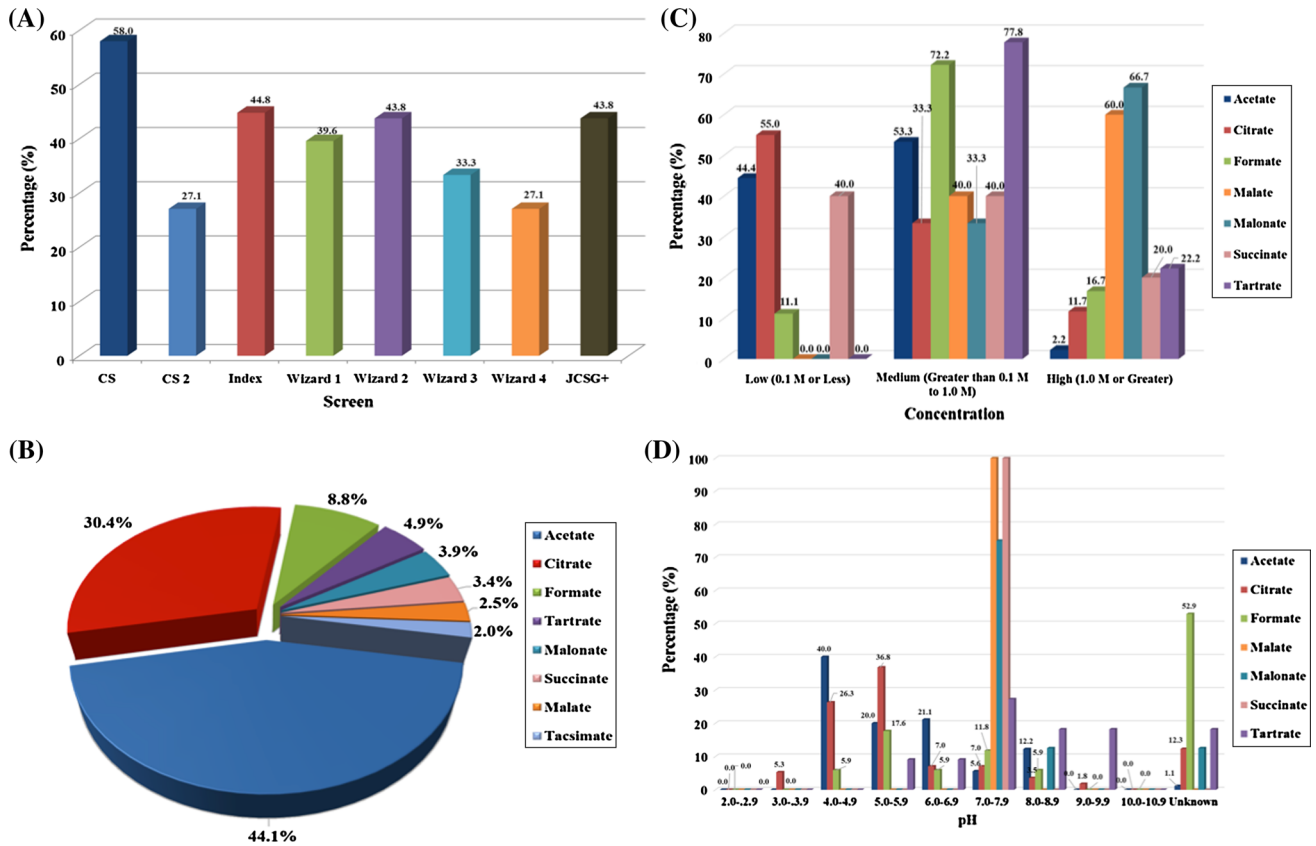

Fig. 7 Analysis of commercial screens Crystal Screen 1 (CS), Crystal Screen 2 (CS2), Index (Hampton Research), Wizard Screen 1-4 (Emerald Biosystems), and Joint Structure for Structural Genomics Plus (Molecular Dimensions). **a** Analysis of commercial screens containing a carboxylic acid. **b** Percentage of which carboxylic acid is present in which commercial screen. **c** pH ranges of the crystallization conditions in examined commercial screens used that contained a carboxylic acid. **d** Concentration ranges of crystallization conditions in examined commercial screens that contained a carboxylic acid

Table 4 Analysis of the concentration of ammonium sulfate in combination with the concentration of (**A**) Acetate (**B**) Citrate, for commercial screens

A.			**Ammonium Sulfate**		**B.**			**Ammonium Sulfate**	
		Low (0.1 M or lower)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1.0 M)			Low (0.1 M or lower)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1.0 M)
	Low (0.1 M or lower)	0.0%	33.3%	66.7%		Low (0.1 M or lower)	0.0%	44.4%	66.7%
Acetate	Medium (Greater than 0.1 M to 1.0 M)	0.0%	0.0%	0.0%	**Citrate**	Medium (Greater than 0.1 M to 1 M)	0.0%	0.0%	0.0%
	High (Greater than 1.0 M)	0.0%	0.0%	0.0%		High (Greater than 1.0 M)	0.0%	0.0%	0.0%

Cells highlighted in blue indicate combinations >10 %

chloride (Fig. 5). Additionally, the concentrations of acetate, citrate, and formate with the concentration of sodium chloride were analyzed (Table 2).

Figures 3a–c clearly show that PEGs are the compounds that most often accompany acetate, citrate, and formate. For acetate and citrate, more than 50 % of the analyzed

Table 5 Analysis of the concentration of sodium chloride in combination with the concentration of (**A**) Acetate (**B**) Citrate, for commercial screens

A.			**Sodium Chloride**	
		Low (0.1 M or lower)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1 M)
Acetate	Low (0.1 M or lower)	0.0%	28.6%	71.4%
	Medium (Greater than 0.1 M to 1.0 M)	0.0%	0.0%	0.0%
	High (Greater than 1.0 M)	0.0%	0.0%	0.0%

B.			**Sodium Chloride**	
		Low (0.1 M or lower)	Medium (Greater than 0.1 M to 1.0 M)	High (Greater than 1.0 M)
Citrate	Low (0.1 M or lower)	0.0%	71.4%	14.3%
	Medium (Greater than 0.1 M to 1.0 M)	0.0%	14.3%	0.0%
	High (Greater than 1.0 M)	0.0%	0.0%	0.0%

Cells highlighted in blue indicate combinations >10 %

crystallization conditions contain PEGs, while for formate it is more than 40 % of all conditions. Entries containing PEG were also examined and divided based on average molecular weight, where 'low MW' refers to PEGs <2,000, 'medium MW' is comprised of PEGs 2,000–6,000, and 'high MW' contained PEGs >6,000. PEGs with medium molecular weight are most often observed in conditions containing acetate, citrate and formate. Conditions containing acetate differ somewhat from conditions with citrate and formate, as acetate more often is observed in combination with high molecular weight PEGs.

Conditions containing PEGs of various molecular weights in combination with acetate, citrate, and formate were also analyzed in detail. Entries that contained PEGs (of any size) were filtered by acetate, citrate, or formate, then sorted based on percentage of weight per volume (w/v) or volume per volume (v/v) depending on the average molecular weight of PEG used in the crystallization reagent. Acetate, citrate, and formate in combination with all PEGs followed the same trends. Disregarding the unknown category, the 11–20 % concentration range of PEG (w/v or v/v) was most commonly observed, followed by the 21–30 % range, then the 0–10 % range. About 17–29 % of all entries containing PEG had missing information, and therefore were categorized in the unknown bracket (Fig. 6).

All entries were categorized by carboxylic acid first, then whether or not it contained PEG. Of the entries that fulfilled carboxylic acid and PEG, the PEG was then divided into categories of low, medium, and high average molecular weight. In all cases, medium molecular weight PEG was most commonly used with acetate, citrate, and formate. Also, in all cases <1 % of entries fell into the unknown category indicating that when the use of PEG is reported to the PDB, the molecular weight is also disclosed. The concentrations of acetate, citrate, and formate with the average molecular weight of PEGs were also analyzed and are shown in Table 3.

Further analysis was completed in regards to acetate, citrate, and formate. Entries that contained the particular carboxylic acid were grouped and then sorted by how many conditions contained each of the other most common carboxylic acid, as well as if they contained phosphate. Acetate in combination with citrate or formate was seen at 6.0 or 3.1 % of all conditions containing acetate. Overall 2.4 % of conditions containing acetate also contained phosphate. Citrate in combination with acetate or formate was seen at 9.4 or 1.1 % of all conditions containing citrate. 13.3 % of conditions containing citrate also contained phosphate. Finally, formate in combination with acetate or citrate was seen at 17.4 or 3.8 % of all conditions containing formate and 1.3 % of conditions containing formate also contained phosphate.

Commercial screens

In order to investigate whether the crystallization conditions reported in the PDB differ significantly from those used in general use commercial screens, an additional portion of our analysis was the examination of carboxylic acids used in these screens. The commercial screens we investigated were: Crystal Screen 1 and 2 (Hampton Research), Index (Hampton research), Wizard Classic screens 1-4 (Emerald Bio), and JCSG+ (Molecular Dimensions) which were chosen as to not show bias towards a particular group of macromolecules. Of the commercial screens investigated, Crystal Screen 1 contained the highest percentage of conditions containing a carboxylic acid (58.0 %). Index, JCSG+, and Wizard Classic 1 and 2 had 46.9, 43.8, 39.6 and 43.8 % of conditions containing a carboxylic acid respectively. Crystal Screen 2, and Wizard Classic 3 and 4 contained the fewest number of conditions containing a carboxylic acid at 27.1, 33.3 and 27.1 %, respectively (Fig. 7a). All conditions containing a carboxylic acid from the commercial screens

Table 6 Analysis of the molecular weight ranges of polyethylene glycol in combination with the concentration of (**A**) Acetate (**B**) Citrate, and (**C**) Formate, for commercial screens

A.

		PEG		
		Low (Less than 2000)	Medium (2000-6000)	High (Greater than 6000)
Acetate	Low (0.1 M or less)	18.8%	20.8%	10.4%
	Medium (Greater than 0.1 M to 1.0 M)	6.3%	25.0%	18.8%
	High (Greater than 1.0 M)	0.0%	0.0%	0.0%

B.

		PEG		
		Low (Less than 2000)	Medium (2000-6000)	High (Greater than 6000)
Citrate	Low (0.1 M or less)	12.5%	41.7%	16.7%
	Medium (Greater than 0.1 M to 1.0 M)	4.2%	25.0%	0.0%
	High (Greater than 1.0 M)	0.0%	0.0%	0.0%

C.

		PEG		
		Low (Less than 2000)	Medium (2000-6000)	High (Greater than 6000)
Formate	Low (0.1 M or less)	0.0%	12.5%	0.0%
	Medium (Greater than 0.1 M to 1.0 M)	0.0%	87.5%	0.0%
	High (Greater than 1.0 M)	0.0%	0.0%	0.0%

Cells highlighted in blue indicate combinations >10 %

were pooled and analyzed. The commercial screens are in agreement with what we observed from our analysis of the PDB; where acetate, citrate, and formate are the three most commonly used carboxylic acids (Fig. 7b). Additionally, the pooled conditions were further analyzed to determine the carboxylic acid concentration and pH (Fig. 7c, d).

The concentrations of the carboxylic acids used in the commercial screens are similar to the concentrations of the carboxylic acids in the nRD. Acetate (53.3 %) and formate (72.2 %) were most commonly used at medium concentration, while low concentration was favored for citrate (55.0 %). Moreover, malate and malonate were typically used at high concentration, 60.0 and 66.7 % respectively, succinate was used equally at 40.0 % in both low and medium concentrations, and tartrate (77.8 %) was frequently used at medium concentration. The pH of the commercial screen conditions containing a carboxylic acid is in agreement with the pH of the conditions containing a carboxylic acid for the nRD. The only exception is malate, where it was found at high concentration in the commercial screens and found at medium concentrations in the nRD.

Acetate and citrate are most commonly used within their buffering ranges at 4.0–4.9 and 5.0–5.9 respectively. Formate, however, is unusual in that most of the conditions containing formate do not report a pH thus yielding almost 60 % of formate conditions in the 'unknown' category. As is seen in our analysis of the nRD, the majority of conditions are found between pH 4–8. Malonate, succinate, and tartrate are most commonly found in the 7.0–7.9 pH range, indicating their use at physiological pH, while 75 % of conditions containing malate are found in the 8.0–8.9 range. This range is higher than its buffering range and is greater than physiological pH.

The concentrations of acetate and citrate with the concentration of ammonium sulfate or the concentration of sodium chloride are shown in Tables 4 and 5. Statistics including formate in combination with ammonium sulfate or sodium chloride were not observed. The analysis average molecular weight of PEGs with acetate, citrate, and formate concentrations are shown in Table 6.

Conclusions

A significant portion of PDB entries, determined by X-ray crystallography, do not have information on crystallization

conditions, or the reported conditions do not have a detailed description [19, 20]. Nevertheless, even incomplete data was informative as to which of the carboxylic acids are most commonly used, what concentrations they are used at, and what pHs are most common. Not only did this analysis determine that acetate, citrate, and formate are the top three carboxylic acids used, it also investigated other components used in combination with the carboxylic acid. Furthermore, it prompted the analysis of commercial screens to determine whether or not the screens rely heavily on carboxylic acids. The commercial screens are in agreement that acetate, citrate, and formate are the most commonly used acids. This finding is consistent with the observation made by Gorrec [21], who noticed that the more often a reagent is used to formulate initial conditions, the more often it is later found in conditions that report quality diffracting crystals. The commercial screen conditions are also in agreement with this study in that the less utilized carboxylic acids studied here may not lead to successful crystallization experiments and therefore their occurrence in not only the PDB, but also commercial screens is much less frequent. Moreover, of the commercially available screens that were analyzed here, Crystal Screen 1 has the highest percentage of conditions containing carboxylic acids. Our observations show that some compounds, or combination of compounds, are significantly more successful in the production of diffracting crystals. Therefore, this analysis is leading the way for the development of a new crystallization screen focused around carboxylic acids. Our findings also are in agreement with results obtained by Kimber et al. [22] who analyzed crystallization conditions for several hundred proteins purified and crystallized using the same approach. It is also worth mentioning that Kimber et al. also noticed that citrate and formate salts were present in especially successful crystallization conditions. However, the same study showed that acetate and tartrate salts were less successful precipitants.

Carboxylic acids are very popular as components of crystallization conditions, yet most of these conditions correspond to the combination of relatively few compounds. Therefore, alternative crystallization strategies with the use of more diverse carboxylic acids as additives should be used to further increase crystallization probability [14, 16, 23].

Acknowledgments We would like to thank Lukasz Lebioda for helpful discussions and we would like to acknowledge the University of South Carolina for providing us with internal funding for this project.

References

1. McPherson A (1999) Crystallization of biological macromolecules. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
2. Carter CW, Carter CW (1979) J Biol Chem 254:12219–12223
3. Jancarik J, Kim SH (1991) J Appl Crystallogr 24:409–411
4. Ng JD, Gavira JA, García-Ruíz JM (2003) J Struct Biol 142:218–231
5. Leng J, Salmon J (2008) RSC 9:24–34
6. Gerdts CJ, Elliott M, Lovell S, Mixon MB, Napuli AJ, Staker BL, Mollert P, Stewart L (2008) Acta Crystallogr D64:1116–1122
7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
8. Bujacz G, Wrzesniewska B, Bujacz A (2010) Acta Crystallogr D66:789–796
9. Holyoak T, Fenn TD, Wilson MA, Moulin AG, Ringe D, Petsko GA (2003) Acta Crystallogr D59:2356–2358
10. Gorrec F (2009) J Appl Crystallogr 42:1035–1042
11. Hennessy D, Gopalakrishnan V, Buchanan BG, Rosenberg JM, Subramanian D (1994) Proc Int Conf Intell Syst Mol Biol 2:179–187
12. Samudzi CT, Fivash MJ, Rosenberg JM (1992) J Cryst Growth 123:47–58
13. Tung M, Gallagher DT (2009) Acta Crystallogr D65:18–23
14. Larson SB, Day JS, Cudney R, McPherson A (2007) Acta Crystallogr D63:310–318
15. McPherson A (2001) Protein Sci 10:418–422
16. McPherson A, Cudney B (2006) J Struct Biol 156:387–406
17. Radaev S, Li S, Sun PD (2006) Acta Crystallogr D62:605–612
18. Kantardjieff KA, Rupp B (2004) Biochemistry 20:2162–2168
19. Peat TS, Christopher JA, Newman J (2005) Acta Crystallogr D61:1662–1669
20. Newman J, Bolton EE, Muller-Dieckmann J, Fazio VJ, Gallagher DT, Lovell D, Luft JR, Peat TS, Ratcliffe D, Sale RA, Snell EH, Taylor K, Vallotton P, Velanker S, von Delft F (2012) Acta Crystallogr F68:253–258
21. Gorrec F (2013) J Appl Crystallogr 46:795–797
22. Kimber M, Vallee F, Houston S, Nečakov A, Skarina T, Evdokimova E, Beasley S, Christendat D, Savchenko A, Arrowsmith CH, Vedadi M, Gerstein M, Edwards AM (2003) Proteins: structure. Funct Genet 51:562–568
23. Larson SB, Day JS, Nguyen C, Cudney R, McPherson A (2008) Cryst Growth Des 8:3038–3052