CRYSTAL GROWTH & DESIGN
Article
2010, Vol. 10
2785–2792

DOI: 10.1021/cg1004209

The C6 Web Tool: A Resource for the Rational Selection of Crystallization Conditions

Janet Newman,*,[†] Vincent J. Fazio,[†] Brian Lawson,[‡] and Thomas S. Peat[†]

[†]*CSIRO Molecular and Health Technologies, Parkville, Victoria 3052 Australia, and* [‡]*Department of Philosophy and Social Science, Santa Monica College, Santa Monica, California 90405*

Received March 29, 2010; Revised Manuscript Received April 21, 2010

ABSTRACT: We have created the C6 Web Tool that uses an underlying metric to compare the chemical similarity of two crystallization conditions and by extension the similarity of two crystallization screens. With over 220 crystallization screens currently available for purchase, it is difficult to know what each screen contains and when it is appropriate to use that screen. The C6 Web Tool can be found at http://c6.csiro.au and is available to the crystallization community at no charge. In addition to measuring the similarity of conditions and kits, the C6 Web Tool also provides the means to examine the conditions of a crystallization kit (or kits) in novel ways. Using the C6 Web Tool, researchers can efficiently select appropriate screens to use throughout the various stages of a crystallization project.

Introduction

Macromolecular structure determination via X-ray crystallography requires the growth of crystals which are large enough to be mounted in an X-ray beam and well ordered enough that they diffract those X-rays. The production of X-ray quality crystals from a protein sample is not straightforward, and is currently one of the bottlenecks in the process of determining the atomic coordinates of a protein molecule.[1] One of the problems is that crystallization screening requires searching for successful conditions in a vast parametric space.[2]

The Bio21 Collaborative Crystallisation Centre (C^3, www.csiro.au/c3) is a macromolecular crystallization service in Melbourne, Australia, that has over 100 national and international users: during the course of running the service we noticed that we fielded a lot of questions which were variants of the following: "What screens do you have that contain a lot of ammonium sulfate?", or "I've set up screen x, what screen would be good to try next?". These types of questions are quite hard to answer, as most screen descriptions are given condition by condition, and it is often difficult to get an overall impression of the chemical space covered by a screen.

To address these and other questions, we have developed a web-tool, C6 (Comparison of Crystallisation Conditions at the Collaborative Crystallisation Centre), which is based on the *quantification of the similarity* between two crystallization conditions. The C6 Web Tool can be found at the Web site http://c6.csiro.au, and has a number of analyses, including a report that shows the chemicals found in a screen or screens, and a report that returns which screens are most similar (or dissimilar) to a given screen.

Similarity. The idea of similarity has been used in the field of computational chemistry for decades as this concept allows for the simplification of the enormously complicated chemical space into a smaller space of chemical properties. This may be done in many different ways using, for example, fingerprinting methods, functional group similarity, or shape similarity.[3] The concept of similarity allows the clustering of chemicals in useful ways: one definition of "drug like" chemical space is the subset of chemical space where the molecules follow the famous Lipinski "Rule of 5".[4] More importantly, similarity allows for prediction: is an unknown chemical similar to members of the set of "drug-like" chemicals? The answer may be the basis for continuing to develop that particular chemical further.

We use a normalized similarity (distance) metric to assign a similarity value between two crystallization conditions. If two conditions contain exactly the same chemicals at the same concentration values and with the same units then the two conditions are identical and they are assigned a pairwise similarity distance value of "0". Two conditions that have no chemical species in common are maximally different and are assigned a pairwise similarity distance of "1". Conditions which have some chemical species in common result in a similarity distance between these two extremes.

We extend this idea of distance between any two conditions to define a quantitative measure of the difference between any two screens, or sets of conditions. The screen distance measure that has been developed is independent of the order of the two crystallization conditions, does not require the two screens being compared to contain the same number of conditions, and is commutative.

Crystallization Screens. Until 1991, most crystallization screening was done by creating grid screens by hand. These screens were often simple linear gradients of two or three chemicals, set up in 24-well culture dishes (Linbro plates).[5] A major change came about through the work of Jancarik and Kim,[6] who showed that an effective crystallization screen could be generated by creating a set of conditions based on chemicals that had been used in previously published (successful) crystallization conditions. Each crystallization condition contains one or more (often three) chemicals. They showed that this "sparse matrix" screen which contained only 50 conditions (and a mere 27 distinct chemical species) was an efficient starting point for a *de novo* crystallization experiment. Within five years of this initial paper on sparse matrix screening, the Jancarik and

*To whom correspondence should be addressed.

Published on Web 05/10/2010

Kim screen was available commercially (Crystal Screen, from Hampton Research).

Since then, there has been an explosion in the number of screens commercially available for initial crystallization screening. Along with the large number of sparse matrix type crystallization screens, there are also commercial grid screens as well as additive screens, which are intended for use during the optimization of initial crystallization hits. The sheer number of available crystallization reagents is daunting: as of January, 2010 there were over 220 commercial screens, or over 13 000 conditions which could be purchased. Crystallization space (which we define as the set of conditions that are used in the attempts to crystallize proteins) is certainly not limited to these commercially available conditions; however, most crystallization projects now start with a screening step using one or more commercial sparse matrix type screen(s). As there is always a limit to the number of crystallization experiments that can be set up − the limit may be determined by protein availability, access to screens, manpower to assemble the experiments or personnel to examine and analyze the experiments − it would be enormously useful to have a guide to help choose a limited set of appropriate crystallization conditions.

Many of the crystallization screens on the market are very similar in content, and knowing which screens are essentially duplicates may change which screens are selected to be used in a screening strategy. Determining which screens are closely related is not necessarily obvious, as although one vendor's screen may contain the same conditions as another vendor's product, the conditions are often found in a different order, confounding easy comparison. For example, a condition containing "0.4 M potassium sodium tartrate" is found in position A2 of Crystal Screen HT (Hampton Research), position D4 of Structure Screen 1 and 2 HT-96 (Molecular Dimensions), F6 of JBScreen Basic HTS (Jena BioScience) and C12 of The Classic Suite (Qiagen), and yet all these screens are essentially identical, except for the order of the conditions within the screen. Adding a further level of difficulty to crystallization screen comparison is the nonuniformity of chemical names within the crystallization community. "Polyethylene glycol monomethyl ether 5000" may be contracted to "MPEG 5K" by one vendor and PEG MME 5000" by another.

Methods and Results

We used the "alias" feature of the CrystalTrak database application (Rigaku Automation, Carlsbad, California) as described earlier[7] to generate a list of commercial screens with consistent naming. However, simply generating a list of consistent names does not solve all the problems with chemical naming. Some valid chemical names are ambiguous: "ammonium citrate" can refer to triammonium citrate, diammonium hydrogen citrate, or ammonium dihydrogen citrate. In particular, the description of buffers in crystallization literature is often incomplete; a complete description of a buffer requires knowledge of the acid and conjugate base that was used to make the buffer. For example, a tris buffer is most often made by dissolving trizma base and adjusting to the desired pH with hydrochloric acid, yielding "tris chloride". We appreciate that this is more often called "tris hydrochloride", but we use "tris chloride" to be consistent with other tris-acid combination names. The pH of the trizma base could also be adjusted with sulfuric acid − which would give "tris sulfate". If a vendor provides only the information "tris" and a pH it is clear that the pH has been adjusted, but it is not clear how that adjustment was done. However, for the purposes of the condition-to-condition comparison, "tris" could map with equal validity to either "tris","tris chloride" or "tris sulfate". We have implemented a set of chemical classes to circumvent this problem. One of the chemical classes that we have created is the class "TRIS buffer" which contains the chemicals "trizma", "tris", "tris chloride", "tris sulfate", "tris acetate", and "tris phosphate". Any of the chemicals within a chemical class are considered identical in the comparisons used in the C6 Web Tool. Notice that this is only used for general comparisons: for user-defined searches the user may specify either a general search (TRIS buffer) or a specific chemical (e.g., tris sulfate). For a list of the chemical classes used in the C6 Web Tool and their contents, see Table 1.

Table 1. The Chemical Classes Set up in the C6 Web Tool[a]

chemical class	contains
TRIS class	tris, trizma, tris chloride, tris sulfate, tris acetate, tris phosphate
HEPES class	HEPES, sodium HEPES
MES class	MES, sodium MES, potassium MES
malonate class	malonic acid, sodium malonate
malate class	malic acid, sodium malate
citrate class	sodium citrate, citric acid, sodium citrate - citric acid
succinate class	sodium succinate, succinic acid

[a] Any member of a chemical class is considered equivalent to any other member of the same class in the C6 Web Tool underlying distance similarity metric.

We observed significant variation in the units used to describe the concentration of chemicals within the commercial crystallization conditions. Units which are simple derivatives are easy to interpret (M to mM, for example); however, units which are not equivalent (%v/v (percent by volume) and %w/v (percent by weight)) were sometimes found within the same screen for the same chemical (e.g., the MPD Suite from Qiagen). In this case, the density of MPD (2-methyl-2,4-pentanediol, density = 0.925 g/mL at 25 C was used to convert the %v/v value to %w/v for the purpose of comparison. If a chemical was described with more than one kind of unit in any of the conditions, then the unit for that chemical was converted to a common unit, (usually %w/v) for all conditions containing that chemical.

Initially, we determined the similarity distance between two conditions using a modification of the Canberra metric algorithm.[8] This algorithm considered the concentration difference between any chemical which was found in both conditions, and was normalized both for the solubility of the common chemical as well as for the number of chemicals in the two conditions, see eq 1. The output of this metric is a dissimilarity measure, where all values are between 0 and 1: a distance of 0 indicates that the two conditions are identical, and a value of 1 indicates that the two conditions have no chemical in common.

For two crystallizations conditions, i and j, each containing a number of chemical species s, the distance between them, D_{ij} may be given by eq 1.

$$D_{ij} = 1 \text{ (no chemicals in common)}$$

$$D_{ij} = \frac{1}{T}\sum_{t=1}^{t=T} \frac{|[s_{ti}] - [s_{tj}]|}{\max[s_t]} \tag{1}$$

A modified Canberra metric used to determine the distance between two conditions where T is the number of distinct chemical species in conditions i and j, $[s_{ti}]$ is the concentration of chemical t in condition i, $\max[s_t]$ is the maximum concentration found for chemical t within all the commercial crystallization conditions.

The difference in concentration for a chemical species found in both conditions is normalized to be a proportion of the maximum solubility of that chemical ($\max[s_t]$). As solubility information is not readily available for many of the chemicals used in crystallization, we select the maximum value of the chemical used within the entire collection of commercial crystallization conditions as a proxy for maximum solubility.

Table 2 shows a worked example using this approach.

Table 2. A Comparison of the Similarity of Three Conditions with the "Canberra" Metric (eq *1* in the Main Text)[a]

condition A	condition B	condition C
0.1 M HEPES pH 7.5	0.2 M sodium acetate	0.1 M HEPES pH 7.5
1.5 M ammonium sulfate	10 v/v MMT pH 5	40 v/v MPD
5 v/v MPD	30 v/v 2-propanol	0.2 M magnesium acetate 10 v/v 2-propanol

[a] The distance between conditions A and C is given by $D_{\mathrm{A,C}} = 1/5(0 + 1 + (40 - 5)/70 + 1 + 1) = 0.7$ where max[MPD] = 70 v/v. The distance between conditions B and C is given by $D_{\mathrm{B,C}} = 1/6(1 + 1 + (30 - 10)/60 + 1 + 1 + 1) = 0.88$ where max[2-propanol] = 60 v/v $D_{\mathrm{A,B}} = 1$, as these two conditions have no chemical species in common.

The "Canberra" metric is a bare bones approach, and there are a number of factors which may increase the sensitivity of the distance metric, three of which we have implemented.

pH Values in Crystallization Conditions. Every crystallization condition must have a pH which can be measured, and clearly it is the combination of all the components considered together that determines the final pH of the condition. However, if a pH value is given at all for a commercial condition, it often refers only to the pH of the buffer component of the condition, rather than to the final pH of the condition, and over 23% of the commercial conditions do not record any pH value at all. Furthermore, the pH of the buffer component may not be a reliable estimation of the pH of the condition overall. If we assume that a chemical is used generally as a buffer close to its pK_a, then there is necessarily high correlation between the buffer component used in the condition and the pH of that buffer, given that the buffering range for most buffers is approximately the pK_a of that buffer ±1 pH unit. For example, TRIS buffers occurred 1875 times (out of 13 148 conditions) in the 222 commercial screens recorded in the C6 build of January 24, 2010. The average pH of the TRIS buffers was 8.217, the most common pH was 8.5, and the range of pH varied between pH 6.0 and pH 9.8.

Given that pH is an important factor in protein crystallization,[9] we recognized that including as much as information as possible about pH would be a great improvement to the metric. A pH comparison term was included in the metric which compares the pH value of the two conditions. The pH value used is the overall pH value of the condition (if available) OR the pH of the buffer component of the condition (if available) OR an interpolation between pH values, if more than one chemical in the condition has an associated pH value. If no pH value is available for a condition, then the pH term in eq 2 is not included.

Ionic Component of Chemicals. We assume that two chemicals that share either an anion or cation are "more similar to" each other than chemicals that have neither in common. Thus, sodium chloride would be considered more similar to ammonium chloride than it is to lithium sulfate, if only because both the sodium chloride and ammonium chloride contribute chloride ions to their respective crystallization conditions. We included a term in our "expanded" similarity distance algorithm which looks at the name of every chemical, and tests to see if this name can broken down into one or more cationic species and anionic species, using the CrystalTrak "ions" table as a list of known ions. Any matching cations or anions are then compared. The stoichiometry of the ions is not taken into account.

Polyethylene Glycols. The assumption can be made that PEG 3350 is more similar to PEG 4000 than it is to PEG 20000. We captured this in the "expanded" distance metric calculation as follows. Any chemical notation that contained the string "polyethylene glycol" is further analyzed for an average molecular weight − we defined as similar the PEGs if their molecular weights are within a factor of 2 of each other − for example, PEG 400 and PEG 600 were considered similar (400/600 = 0.6 ; 0.5 < 0.6 < 2), but PEG 400 and PEG 4000 were not (400/4000 = 0.1; 0.1 < 0.5).

We were aware that these extensions to the original "Canberra" metric are qualitative, and to compensate, we assigned a penalty to the extension terms. The size of the penalty is arbitrary, but was chosen as it should be approximately the magnitude as the correction itself. The current metric now employed in the C6 Web Tool is shown below:

$$D_{ij} = 1 \text{ (no species in common)}$$

$$D_{ij} = \frac{1}{(T+3)}\left(\left(\left(\sum_{t=1}^{T}\frac{|[s_{ti}]-[s_{tj}]|}{\max[s_t]}\right)+\left(\frac{|E(\mathrm{pH}_i)-E(\mathrm{pH}_j)|}{\mathrm{gul(pH)}-\mathrm{gll(pH)}}\right)\right)\right.$$
$$+\min\left(1,\left[\left(\frac{|[\mathrm{ion_i}]-[\mathrm{ion}_j]|}{\frac{(\max[\mathrm{ion}_i]+\max[\mathrm{ion}_j])}{2}}\right)+0.3\right]\right)$$
$$\left.+\min\left(1,\left[\left(\frac{|[\mathrm{PEG}_i]-[\mathrm{PEG}_j]|}{\frac{(\max[\mathrm{PEG}_i]+\max[\mathrm{PEG}_j])}{2}}\right)+0.2\right]\right)\right) \quad (2)$$

Distance similarity metric currently in use in the C6 Web Tool where T is the number of distinct chemical species in conditions i and j, $[s_{ti}]$ is the concentration of chemical t in condition i, $\max[s_t]$ is the maximum concentration found for chemical t within all the known crystallization conditions, $E(\mathrm{pH}_i)$ is an estimate of the pH of the pH of condition i, either the value of the pH for that condition, or the value of the buffer pH for that condition, or an interpolation between two pH values, if more than one are given for the same condition, gul(pH) is the overall maximum of pH seen in the commercially available crystallization conditions, gll(pH) is the overall minimum of pH seen in the commercially available crystallization conditions, $[\mathrm{salt}_i]$ is the concentration of salt i, used in the comparison of cations and anions, $[\mathrm{ion}_i]$ is the concentration of ion i, which is determined by $[\mathrm{salt}_i]$, and $[\mathrm{PEG}_i]$ is the concentration of PEG i.

Quantifying the Difference between Screens. Armed with a numeric estimation of the difference between any two conditions, we can extend this to look at overall differences between two sets of conditions. We use the expression shown as eq 3 below to compare two sets of conditions, or screens.

$$\mathrm{score}_{s,t} = \frac{1}{2}\left(\frac{1}{\mathrm{cond}_s}\sum_{i=1}^{\mathrm{cond}_s}\min_{j\in(1,\mathrm{cond}_t)}(D(c_{si},c_{tj}))\right.$$
$$\left.+\frac{1}{\mathrm{cond}_t}\sum_{i=1}^{\mathrm{cond}_t}\min_{j\in(1,\mathrm{cond}_s)}(D(c_{ti},c_{sj}))\right) \quad (3)$$

Comparison of two sets of conditions (screens) where cond_s is the number of conditions in screen s, $D(c_{si}, c_{tj})$ is the distance between condition i of screen s and condition j of screen t given by eq 2 above.

The expression averages the sum of the minimum distance for each condition in each screen over the two sets of conditions in order to ensure that comparing screen s to screen t returns the same value as comparing screen t to screen s. The screen comparison returns a value between 0 and 1, where 0 means that for each condition in screen s there is an identical condition in screen t.

Quantification of Internal Diversity within a Screen. A potentially useful application of the screen comparison application would be to compare a screen to itself, as this would give some idea of how similar each condition is within the screen; that is, it would give a measure of the diversity of the screen. For this analysis, we need to replace the "minimum" in the expression above with "average" as shown in eq 4

$$\mathrm{score}_{s,s} = \frac{1}{\mathrm{cond}_s}\sum_{i=1}^{\mathrm{cond}_s}\underset{j\in(1,\mathrm{cond}_s)}{\mathrm{average}}(D(c_{si},c_{sj})) \quad (4)$$

Comparison of a screen to itself (internal diversity measure) where cond_s is the number of conditions in screen s, $D(c_{si}, c_{sj})$ is the distance between condition i and j of screen s given by eq 2 above.

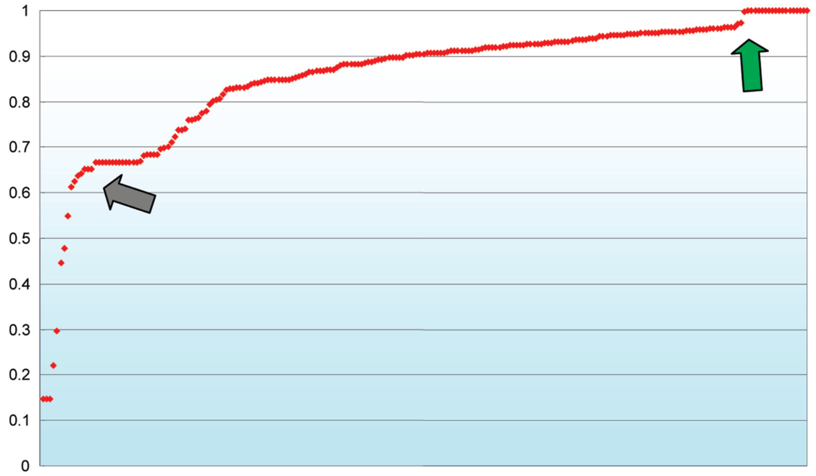


Figure 1. Scatter plot of the internal diversity of the 222 commercial screens in the C6 Web Tool build of 24 January, 2010. The "Canberra" metric (eq 1 in Methods and Results) and the mean distance comparison (eq 4) were used to generate an internal diversity value for each of the 222 commercial screens. The values of internal diversity range from 0.147 to 1. There is a break point at a low internal diversity value (gray arrow) − screens with internal diversity values below 0.6 tend to be grid screens. Another discontinuity is seen at high values of internal diversity (green arrow) − screens with internal diversity above a value of ≈0.95 tend to be additive screens.

A value of 0 returned by this self-comparison or internal diversity algorithm tells us that every condition within the screen is the same; a value of 1 means that no condition in the screen has any chemical species in common with any other condition in the screen. Grid screens will, in general, have low internal diversity values ("Grid Screen Sodium Malonate" from Hampton Research gives a value of 0.23), sparse matrix type screens return intermediate values of internal diversity ("Proplex" from Molecular Dimensions has a value of 0.89) and the only screens where we see values of 1 for internal diversity are additive screens. See Figure 1.

The C6 Web Tool. The extended metric described above in eq 2 has been captured in a web service http://c6.csiro.au, where the underlying code is written in C and python, and the data are updated automatically once a week. The C6 Web Tool uses the data about crystallization conditions and screens that are in the Bio21 C^3's CrystalTrak database. The CrystalTrak database contains descriptions of the commercial screens, as well as descriptions of any screen designed by a user of the Bio21 C^3 center. If a user logs onto the C6 Web Tool with their Bio21 C^3 username and password, this allows them to view information about commercial screens, in-house Bio21 C^3 screens as well as the screens of their own design which were created through the Bio21 C^3. Users logging on as a generic user have access to information about the commercial screens and the Bio21 C^3 screens.

As the screen to screen comparisons are computationally intensive, the distances between the screens are precalculated in order to avoid unacceptable delays for the Web site user. To generate all the pairwise distances currently takes about 8 h on a dual core 2.33 GHz-workstation.

The C6 Web Tool has a number of predefined reports, which include the following:

Single Screen Contents. This report returns a description of a screen, ordered by condition. The result of this report looks similar to the description of screens given by the vendors themselves, with the proviso that the chemical names used are the standard names used in CrystalTrak, rather than those provided by the vendor.

Single Screen Statistics. This report returns a description of a screen, sorted by chemical. The most abundant chemical is shown first, with a description of how many conditions contain that chemical, the average and standard deviation of the concentration over those conditions, the most commonly occurring concentration (mode), as well as minimum and maximum concentrations found for that chemical. Similar information is given for pH (average, standard deviation, mode minimum and maximum) if appropriate. Table 3 shows this report for Crystal Screen HT from Hampton Research.

Screen Internal Diversity. This report orders screens alphabetically. Screens with low internal diversity tend to be grid screens, and screens with high internal diversity are invariably additive screens. Intermediate values are generally sparse matrix screens. Currently, this report returns the internal diversity value calculated using the initial "Canberra" metric by default. See Figure 1.

Find Similar Screens. This report returns a list of screens ordered by similarity to a user-defined screen. By default, the screen list is ordered so that the screens which are most similar to the target screen are at the top of the list. See Figure 2.

Screens Group Statistics. This report returns the same information as the "Single screen statistics" report above, but for one or more screens selected by the user. This report returns not only the list of chemicals and the usage of the chemical, but also gives a count of the number of unique conditions within the selected screens. Using this report, it can be seen that the 222 commercial screens defined in the C6 Web Tool build of January 24, 2010 contained 13,148 conditions of which only 3918 were unique.

Find a Chemical in Screens. This report finds all the conditions which contain the target chemical, and returns a list of the ten screens with the largest number of conditions which contain the target chemical. This report also returns a list of screens that do not contain the target chemical. Using this report, one can, for example, determine that the NR-LBD and NR-LBD Extension HT-96 Screen from Molecular Dimensions contains the most ammonium acetate of any commercial screen, and further that the chemical ammonium acetate is found in 18 of the 96 conditions in this screen.

Find a Condition in Screens. This report returns a list of conditions (and the screen in which the condition is found) which are similar to a target crystallization condition. The target condition is created by the user, and does not have to correspond to any existing condition; however, it must be some combination of chemicals which are already in the C6 system. Each chemical selected by the user requires an associated concentration, unit and potentially pH. Once the condition has been defined, the report returns 10 conditions (well number, screen name) which show some similarity to the entered condition, as well as the distance from the query condition.

Discussion

Nomenclature Standardization. The basis of the metric which underpins the C6 Web Tool relies on having a set of standard descriptors for the crystallization conditions manufactured by the various vendors. Although there may be a good reason for a particular vendor choosing to name a chemical a particular way (for example, Hampton Research includes waters of hydration in the name of the chemicals in their product descriptions) if this information is not provided uniformly by all vendors, it becomes a barrier to cross-vendor

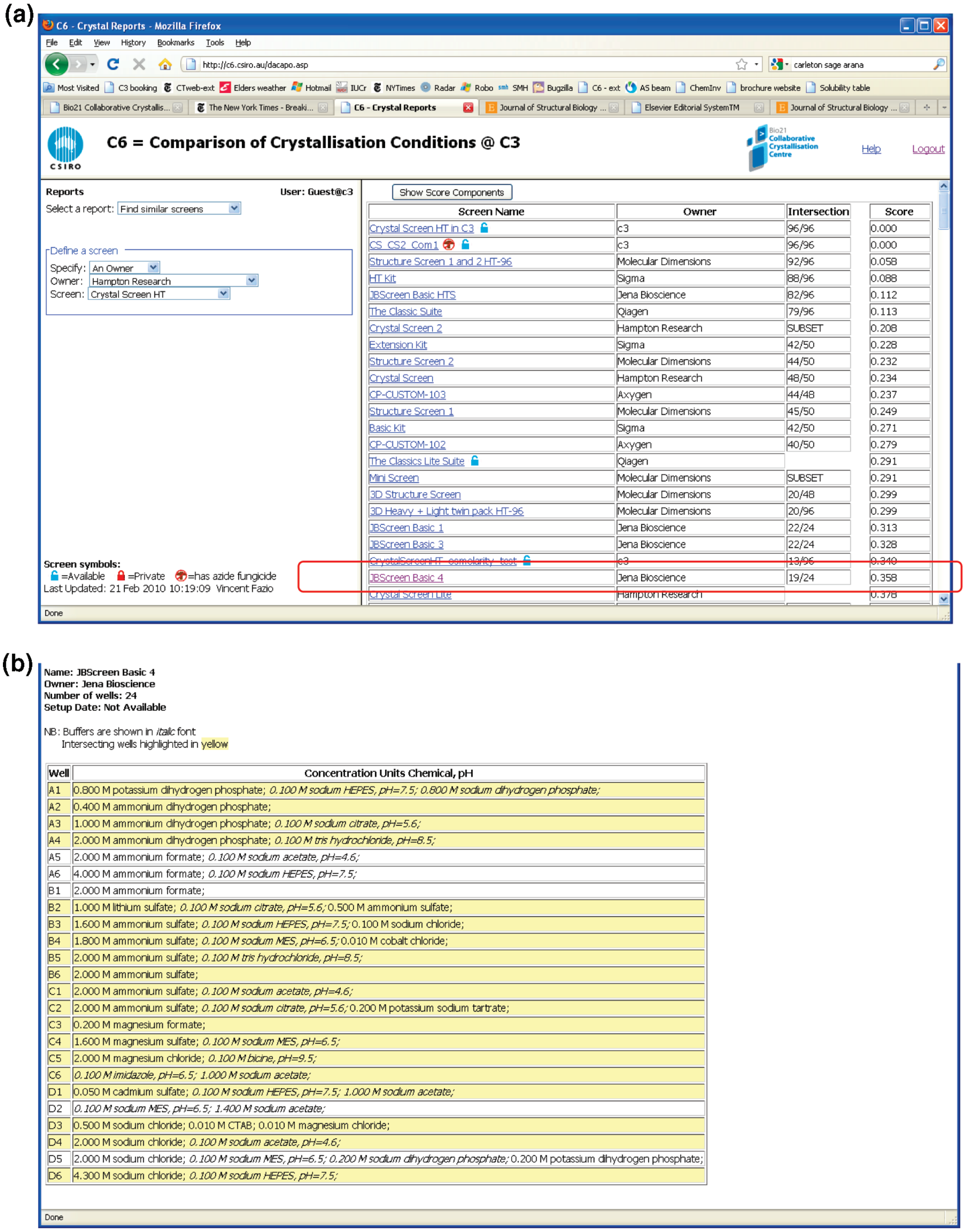
Table 3. Output of the "Single Screen Statistics" Report of the C6 Web Tool for the Screen *Crystal Screen HT* from Hampton Research[a]

			concentration					pH				
number of conditions	name of chemical	units	mean	mode	min	max	std dev	mean	mode	min	max	std dev
18	ammonium sulfate	M	1.15	2.00	0.20	2.00	0.68					
17	sodium acetate	M	0.29	0.10	0.10	1.40	0.66	4.60	4.60	4.60	4.60	0.00
15	tris hydrochloride	M	0.10	0.10	0.10	0.10	0.00	8.50	8.50	8.50	8.50	0.00
11	sodium HEPES	M	0.10	0.10	0.10	0.10	0.00	7.50	7.50	7.50	7.50	0.00
11	sodium citrate - citric acid	M	0.24	0.10	0.10	1.60	0.97	5.68	5.60	5.60	6.50	0.58
10	polyethylene glycol 8000	w/v	19.40	30.00	8.00	30.00	8.16					
10	HEPES	M	0.10	0.10	0.10	0.10	0.000	7.50	7.50	7.50	7.50	0.00
10	polyethylene glycol 4000	w/v	25.30	30.00	8.00	30.00	9.35					
10	sodium chloride	M	1.32	2.00	0.10	4.30	1.46					
9	2-methyl-2,4-pentanediol	v/v	33.89	30.00	5.00	70.00	24.56					
8	sodium cacodylate	M	0.10	0.10	0.10	0.10	0.00	6.50	6.50	6.50	6.50	0.00
8	2-propanol	v/v	20.63	20.00	5.00	30.00	10.55					
8	MES	M	0.10	0.10	0.10	0.10	0.00	6.50	6.50	6.50	6.50	0.00
6	magnesium chloride	M	0.47	0.20	0.01	2.00	0.94					
5	trisodium citrate	M	0.44	0.20	0.20	1.40	0.70					
5	polyethylene glycol 400	v/v	24.00	30.00	2.00	30.00	13.37					
4	ammonium dihydrogen phosphate	M	0.90	1.00	0.20	2.00	0.70					
4	ammonium acetate	M	0.20	0.20	0.20	0.20	0.0000					
4	lithium sulfate	M	0.93	1.00	0.20	1.50	0.54					
3	dioxane	v/v	15.67	2.00	2.00	35.00	14.06					
3	calcium chloride	M	0.14	0.20	0.02	0.20	0.09					
3	potassium sodium tartrate	M	0.47	0.40	0.20	0.80	0.25					
3	Jeffamine M-600	v/v	20.00	10.00	10.00	30.00	8.16					
3	bicine	M	0.10	0.10	0.10	0.10	0.00	9.00	9.00	9.00	9.00	0.00
3	1,6-hexanediol	M	2.30	2.50	1.00	3.40	0.99					
3	potassium dihydrogen phosphate	M	0.32	0.05	0.05	0.80	0.34					
2	ethanol	v/v	15.00	10.00	10.00	20.00	5.00					
2	nickel(II) chloride	M	0.01	0.01	0.01	0.01	0.00					
2	cobalt chloride	M	0.01	0.01	0.01	0.01	0.00					
2	ethylene glycol	v/v	16.50	8.00	8.00	25.00	8.50					
2	polyethylene glycol 6000	w/v	10.00	10.00	10.00	10.00	0.00					
2	polyethylene glycol monomethyl ether 2000	w/v	25.00	20.00	20.00	30.00	5.00					
2	imidazole	M	0.55	1.00	0.10	1.00	0.45	6.75	6.50	6.50	7.00	0.25
2	sodium dihydrogen phosphate	M	0.45	0.10	0.10	0.80	0.35					
2	polyethylene glycol monomethyl ether 550	v/v	22.50	25.00	20.00	25.00	2.50					
2	polyethylene glycol 20000	w/v	11.00	10.00	10.00	12.00	1.00					
2	*tert*-butanol	v/v	30.00	25.00	25.00	35.00	5.00					
2	sodium formate	M	3.00	2.00	2.00	4.00	1.00					
2	magnesium acetate	M	0.20	0.20	0.20	0.20	0.00					
1	ethylene imine polymer	w/v	2.00	2.00	2.00	2.00	0.00					
1	CTAB	M	0.01	0.01	0.01	0.01	0.00					
1	zinc acetate	M	0.20	0.20	0.20	0.20	0.00					
1	cesium chloride	M	0.05	0.05	0.05	0.05	0.00					
1	polyethylene glycol 1000	w/v	10.00	10.00	10.00	10.00	0.00					
1	polyethylene glycol 10000	w/v	20.00	20.00	20.00	20.00	0.00					
1	polyethylene glycol monomethyl ether 5000	w/v	30.00	30.00	30.00	30.00	0.00					
1	zinc sulfate	M	0.01	0.01	0.01	0.01	0.00					
1	cadmium sulfate	M	0.05	0.05	0.05	0.05	0.00					
1	glycerol	v/v	12.00	12.00	12.00	12.00	0.00					
1	calcium acetate	M	0.20	0.20	0.20	0.20	0.00					
1	cadmium chloride	M	0.10	0.10	0.10	0.10	0.00					
1	polyethylene glycol 1500	w/v	30.00	30.00	30.00	30.00	0.00					
1	magnesium sulfate	M	1.60	1.60	1.60	1.60	0.00					
1	magnesium formate	M	0.20	0.20	0.20	0.20	0.00					
1	ammonium formate	M	2.00	2.00	2.00	2.00	0.00					
1	ferric chloride	M	0.01	0.01	0.01	0.01	0.00					

[a] This report also outputs the number of conditions in the screen (96); the number of distinct chemicals (56) and the internal diversity (0.9528).

comparison. We have imposed a consistent naming of chemicals across all the commercial screens, thus the descriptions of the screen conditions in the C6 Web Tool will not necessarily be exactly the same as those provided by a vendor.

Comparison of Conditions – "Input" vs. "Output" Methods. To determine the similarity between two crystallization conditions, one can consider two approaches. First, an estimation of two conditions' similarity can be obtained by "assaying" the two conditions in a crystallization experiment. In other words, by measuring how similar the results are of crystallization experiments set up using the two conditions. This may be considered an "output" approach. Second, the nature of the two conditions could be compared in some way which does not rely on the readout from an experiment. This may be denoted an "input" approach. The former "output" approach seems to be the preferred method, according to informal discussions by the authors with other structural biologists. Moreover, there are a number of crystallization resources that collect "output" information – for example, the Biomolecular Crystallization Database (BMCD),[10] the Marseille Protein Crystallization Database (MPCD),[11] and XtalBase.[12] However, the "output" approach is fraught with problems, the most serious being the


(a)
C6 - Crystal Reports - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://c6.csiro.au/dacapo.asp
C6 - Crystal Reports
CSIRO
Bio21 Collaborative Crystallisation Centre
Help
Logout
Show Score Components
Done
(b)
Done

C6 = Comparison of Crystallisation Conditions @ C3

Reports **User: Guest@c3**

Select a report: Find similar screens

Define a screen
Specify: An Owner
Owner: Hampton Research
Screen: Crystal Screen HT

Screen Name	Owner	Intersection	Score
Crystal Screen HT in C3 🔓	c3	96/96	0.000
CS_CS2_Com1 ☣ 🔓	c3	96/96	0.000
Structure Screen 1 and 2 HT-96	Molecular Dimensions	92/96	0.058
HT Kit	Sigma	88/96	0.088
JBScreen Basic HTS	Jena Bioscience	82/96	0.112
The Classic Suite	Qiagen	79/96	0.113
Crystal Screen 2	Hampton Research	SUBSET	0.208
Extension Kit	Sigma	42/50	0.228
Structure Screen 2	Molecular Dimensions	44/50	0.232
Crystal Screen	Hampton Research	48/50	0.234
CP-CUSTOM-103	Axygen	44/48	0.237
Structure Screen 1	Molecular Dimensions	45/50	0.249
Basic Kit	Sigma	42/50	0.271
CP-CUSTOM-102	Axygen	40/50	0.279
The Classics Lite Suite 🔓	Qiagen		0.291
Mini Screen	Molecular Dimensions	SUBSET	0.291
3D Structure Screen	Molecular Dimensions	20/48	0.299
3D Heavy + Light twin pack HT-96	Molecular Dimensions	20/96	0.299
JBScreen Basic 1	Jena Bioscience	22/24	0.313
JBScreen Basic 3	Jena Bioscience	22/24	0.328
CrystalScreenHT_osmolarity_test 🔓	c3	13/96	0.340
JBScreen Basic 4	Jena Bioscience	19/24	0.358
Crystal Screen Lite	Hampton Research		0.378

Screen symbols:
🔓=Available 🔒=Private ☣=has azide fungicide
Last Updated: 21 Feb 2010 10:19:09 Vincent Fazio

Name: JBScreen Basic 4
Owner: Jena Bioscience
Number of wells: 24
Setup Date: Not Available

NB: Buffers are shown in *italic* font
Intersecting wells highlighted in yellow

Well	Concentration Units Chemical, pH
A1	0.800 M potassium dihydrogen phosphate; *0.100 M sodium HEPES, pH=7.5; 0.800 M sodium dihydrogen phosphate;*
A2	0.400 M ammonium dihydrogen phosphate;
A3	1.000 M ammonium dihydrogen phosphate; *0.100 M sodium citrate, pH=5.6;*
A4	2.000 M ammonium dihydrogen phosphate; *0.100 M tris hydrochloride, pH=8.5;*
A5	2.000 M ammonium formate; *0.100 M sodium acetate, pH=4.6;*
A6	4.000 M ammonium formate; *0.100 M sodium HEPES, pH=7.5;*
B1	2.000 M ammonium formate;
B2	1.000 M lithium sulfate; *0.100 M sodium citrate, pH=5.6;* 0.500 M ammonium sulfate;
B3	1.600 M ammonium sulfate; *0.100 M sodium HEPES, pH=7.5;* 0.100 M sodium chloride;
B4	1.800 M ammonium sulfate; *0.100 M sodium MES, pH=6.5;* 0.010 M cobalt chloride;
B5	2.000 M ammonium sulfate; *0.100 M tris hydrochloride, pH=8.5;*
B6	2.000 M ammonium sulfate;
C1	2.000 M ammonium sulfate; *0.100 M sodium acetate, pH=4.6;*
C2	2.000 M ammonium sulfate; *0.100 M sodium citrate, pH=5.6;* 0.200 M potassium sodium tartrate;
C3	0.200 M magnesium formate;
C4	1.600 M magnesium sulfate; *0.100 M sodium MES, pH=6.5;*
C5	2.000 M magnesium chloride; *0.100 M bicine, pH=9.5;*
C6	*0.100 M imidazole, pH=6.5; 1.000 M sodium acetate;*
D1	0.050 M cadmium sulfate; *0.100 M sodium HEPES, pH=7.5; 1.000 M sodium acetate;*
D2	*0.100 M sodium MES, pH=6.5; 1.400 M sodium acetate;*
D3	0.500 M sodium chloride; 0.010 M CTAB; 0.010 M magnesium chloride;
D4	2.000 M sodium chloride; *0.100 M sodium acetate, pH=4.6;*
D5	2.000 M sodium chloride; *0.100 M sodium MES, pH=6.5; 0.200 M sodium dihydrogen phosphate;* 0.200 M potassium dihydrogen phosphate;
D6	4.300 M sodium chloride; *0.100 M sodium HEPES, pH=7.5;*

Figure 2. Figure 2a shows a screenshot of the "Find Similar Screens" report of the C6 Web Tool, when run with CrystalScreen HT from Hampton Research as the target. The left-hand frame is where the user selects the desired report and target screen, and the right-hand frame shows the results of the query. The "Scores" column shows the result of the minimum screen distance (eq 3) − the smaller the number, the more similar the screen to the target screen. The "Intersection" column in the right-hand frame shows how many conditions match a condition in the target screen exactly. Two screens may have a low distance score (i.e., be quite similar) even if there are no conditions in common between the two screens. In this example, The Classics Lite Suite from Qiagen has no intersection with Crystal Screen HT, but has quite a low score, and is thus these two screens are similar. If every condition in the screen matches a condition in the target screen then this is a "subset" of the target screen. Notice here that Crystal Screen II from Hampton Research is a subset of CrystalScreen HT, whereas Crystal Screen is not. That follows as Crystal Screen from Hampton Research is a 50 condition screen, of which only 48 are used to make up the CrystalScreen HT 96 condition screen. Clicking on the name of a screen in the "ScreenName" column in the right-hand frame of the page opens a new window, which shows a view of that screen, with the conditions that intersect with conditions of the target screen colored yellow. From (a) it can be seen that JBS Basic 4 (Jena BioSciences − outlined in red) has 19 conditions in common with CrystalScreen HT, and from (b), one can see that the JBS Basic 4 conditions A5, A6, B1, D2, and D5 are NOT found in the target screen.

intrinsic unreliability of protein crystallization; the stochastic nature of crystallization means that even the same crystallization condition set up at the same time by the same person with the same protein does not necessarily give the same results.[13]

An issue with the "output" approach is that it requires finding a representative set of proteins which are the basis of the comparison. The crystallization service provided by the Hauptmann Woodward Institute (HWI) has screened close to 12 000 proteins through their 1536 condition screen;[14] they have, without question, the best handle on the output approach to finding similar crystallization conditions. Given that most of us do not have a chance of ever putting 12 000 proteins through a set of screening conditions, how many proteins would be considered a good representative sample? Which proteins should be used? Moreover, it is not clear how the similarity between two crystallization results can be enumerated. The development of a crystallization condition metric requires that a numerical value be assigned to the difference between two conditions. It may be difficult to quantify the intrinsically qualitative comparisons of two crystallization experiments.

By contrast, the "input" approach to similarity searching has as its basis an assumption that it is some property of a crystallization condition that is responsible for the effects that are seen when that condition is used in a crystallization experiment. Arguably, it could be some combination of the physical properties of the condition; pH, viscosity, surface tension, water activity, conductance, etc. that should be compared. However, that information is not readily available for the commercial crystallization conditions, whereas the amount, units and type of chemical within each condition are. We made the potentially unjustified assumption that it is the concentration and type of chemicals that are the appropriate factors to compare between two conditions, and have based our distance metrics on these reported features of the crystallization conditions.

Neither of the metrics described above (eqs 1 and 2) are necessarily metrics in the topological definition of "metric", where the property of triangulation would be met.[15] Rather, we are using the term metric as it is used in the computer science definition of metric, which more broadly defines a metric as a measure of some sort.

Normalization of the difference in concentration of a chemical between two conditions is required as the solubility of the chemicals used in crystallization varies significantly. If two crystallization conditions contain the same chemical, but have a difference of 0.1 M in the concentration of that chemical, that difference may be insignificant (for example, if one condition contains 2.6 M $(NH_4)_2SO_4$ and the other 2.7 M $(NH_4)_2SO_4$) or significant (for example, if one contains 0.1 M NaF and the other 0.2 M NaF). One way of capturing the significance of the concentration difference is to normalize this difference in concentration by the solubility of the chemical. Thus, the 0.1 M difference above becomes a 2.5% difference (assuming a solubility of 4 M for $(NH_4)_2SO_4$), or becomes an 11% difference (assuming a solubility of 0.9 M for NaF). Max[s_t], (the highest observed concentration for the chemical s_t in commercial crystallization space) is used as a proxy for the solubility of a chemical, as information about the solubility of a chemical is not always available.

Conclusions

As crystallization of proteins can be challenging, practical tools which aid the crystallization process are of use to the structural biology community. The development of sparse matrix and other "pre-mixed" commercial screens has enormously aided the crystallization community, as has the creation of online databases of successful conditions. However, these resources in turn give rise to the need for new tools to enable one to obtain value from the information they provide. Given over 220 commercial screens, it becomes important to be able to ascertain the properties of these screens easily − which of the screens are similar, for example, or which screens can be combined together to give a superset of conditions with little redundancy. The collection of online tools which enumerate successful conditions becomes much more powerful if their information can be used to rationally choose one or more screens to start a new crystallization project.

The C6 Web Tool is a navigation tool through crystallization space and provides answers to these and other questions. Access to the tool can be requested by contacting the authors (or emailing janet.newman@csiro.au).

Acknowledgment. We are very grateful to Carleton Sage of Arena Pharmaceuticals, Anthony Nicholls of OpenEye Scientific Software. and Roger Sayle of NextMove Software for useful discussions over the course of almost a decade.

References

(1) Price, W. N.; Chen, Y.; Handelman, S. K.; Neely, H.; Manor, P.; Karlin, R.; Nair, R.; Liu, J.; Baran, M.; Everett, J.; Tong, S. N.; Forouhar, F.; Swaminathan, S. S.; Acton, T.; Xiao, R.; Luft, J. R.; Lauricella, A.; DeTitta, G. T.; Rost, B.; Montelione, G. T.; Hunt, J. F., Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nat. Biotechnol.* **2008**.

(2) Carter, C. W., Jr.; Carter, C. W. Protein crystallization using incomplete factorial experiments. *J. Biol. Chem.* **1979**, *254* (23), 12219–12223.

(3) (a) Nicholls, A.; MacCuish, N. E.; MacCuish, J. D. Variable selection and model validation of 2D and 3D molecular descriptors. *J. Comput.-Aided Mol. Des.* **2004**, *18* (7), 451–474. (b) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996.

(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23* (1−3), 3–25.

(5) Bergfors, T. M., *Protein Crystallization: Techniques, Strategies, and Tips*, 1st ed.; International University Line: San Diego, 1999; Vol. 1, p 306.

(6) Jancarik, J.; Kim, S. H. Sparse matrix sampling: a screening method for crystallization of proteins. *J. Appl. Crystallogr.* **1991**, *24* (4), 409–411.

(7) Peat, T. S.; Christopher, J. A.; Newman, J. Tapping the Protein Data Bank for crystallization information. *Acta Crystallogr.* **2005**, *D61* (12), 1662–1669.

(8) Lance, G. N.; Williams, W. T. A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput. J.* **1967**, *9* (4), 373–380.

(9) McPherson, A., *Crystallization of Biological Macromolecules*, 1st ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, 1999.

(10) Tung, M.; Gallagher, D. T. The Biomolecular Crystallization Database Version 4: expanded content and new features. *Acta Crystallogr. Sect. D* **2009**, *65* (1), 18–23.

(11) Charles, M.; Veesler, S.; Bonnete, F. MPCD: a new interactive online crystallization data bank for screening strategies. *Acta Crystallogr. Sect. D* **2006**, *62* (11), 1311–1318.

(12) Meining, W. XtalBase - a comprehensive data management system for macromolecular crystallography. *J. Appl. Crystallogr.* **2006**, *39* (5), 759–766.

(13) (a) Newman, J.; Xu, J.; Willis, M. C. Initial evaluations of the reproducibility of vapor-diffusion crystallization. *Acta Crystallogr.* **2007**, *D63* (7), 826–832. (b) Newman, J.; Fazio, V. J.; Caradoc-Davies,

T. T.; Branson, K.; Peat, T. S. Practical aspects of the SAMPL challenge: providing an extensive experimental data set for the modeling community. *J. Biomol. Screen* **2009**, *14* (10), 1245–1250.

(14) Snell, E. H.; Luft, J. R.; Potter, S. A.; Lauricella, A. M.; Gulde, S. M.; Malkowski, M. G.; Koszelak-Rosenblum, M.; Said, M. I.; Smith, J. L.; Veatch, C. K.; Collins, R. J.; Franks, G.; Thayer, M.; Cumbaa, C.; Jurisica, I.; DeTitta, G. T. Establishing a training set through the visual analysis of crystallization trials. Part I: approximately 150,000 images. *Acta Crystallogr.* **2008**, *D64* (11), 1123–1130.

(15) Koliha, J. J. *Metrics, Norms and Integrals: An Introduction to Contemporary Analysis*; World Scientific: Hackensack, NJ, 2008.