Computational crystallization

Irem Altan [a], Patrick Charbonneau [a, b], Edward H. Snell [c, d, *]

[a] Department of Chemistry, Duke University, Durham, NC 27708, USA
[b] Department of Physics, Duke University, Durham, NC 27708, USA
[c] Hauptman-Woodward Medical Research Institute, 700 Ellicott St., NY 14203, USA
[d] Department of Structural Biology, SUNY University of Buffalo, 700 Ellicott St., NY 14203, USA

ARTICLE INFO

ABSTRACT

Crystallization is a key step in macromolecular structure determination by crystallography. While a robust theoretical treatment of the process is available, due to the complexity of the system, the experimental process is still largely one of trial and error. In this article, efforts in the field are discussed together with a theoretical underpinning using a solubility phase diagram. Prior knowledge has been used to develop tools that computationally predict the crystallization outcome and define mutational approaches that enhance the likelihood of crystallization. For the most part these tools are based on binary outcomes (crystal or no crystal), and the full information contained in an assembly of crystallization screening experiments is lost. The potential of this additional information is illustrated by examples where new biological knowledge can be obtained and where a target can be sub-categorized to predict which class of reagents provides the crystallization driving force. Computational analysis of crystallization requires complete and correctly formatted data. While massive crystallization screening efforts are under way, the data available from many of these studies are sparse. The potential for this data and the steps needed to realize this potential are discussed.

1. Introduction

Macromolecular crystallography is a gateway to the detailed structure of biological molecules, and, to the biological processes in which they are involved. Its power as a scientific tool is recognized in the remarkable number of Nobel prizes that make direct reference to the technique. The main drawback of the approach is that — as its name indicates — it requires a crystal of the macromolecule of interest. There lies the crux of the problem.

The National Institutes of Health Protein Structure Initiative (PSI) targeting fold space (where all outcomes, crystallization and non-crystallization, were tracked) showed that out of ~45 K soluble, purified targets, ~8 K crystallized, of which only ~5 K resulted in a crystal structure. Later data from the PSI-Biology initiative, which focused on targets of compelling biological interest, showed that out of a further ~10 K targets, only ~2 K resulted in a crystal structure [1]. Results from these large datasets mirror those obtained at the large-scale crystallization screening center led by one

of us: the Hauptman-Woodward Medical Research Institute High-Throughput Screening (HTS) Laboratory [2]. An analysis of 96 biological macromolecular targets screened against a set of 1536 chemical cocktails gave 277 crystal leads (36 targets produced one or more crystals) from ~150 K experiments, and hence only ~0.2% of the experiments used for screening produced crystals [3,4]. These results thus provide a consistent illustration of the remarkably poor success rate of crystallization screening despite the power of the crystallographic technique itself.

A meta-analysis of large-scale crystallization screening centers reveals that out of ten soluble protein constructs, four are likely to crystallize, of which only one on average will produce a crystal structure [5]. Because the number of structures deposited in the Protein Data bank (PDB) that have been determined by X-ray crystallographic methods is now over 100 K, we infer that tens of millions of experiments have had outcomes other than diffracting crystals. Unfortunately, information on those experiments is not reported in the PDB, and in most cases nor are the conditions that produced a crystal but were not used in the final structure determination. We are thus left with datasets that incompletely capture the effort involved in the crystallization of macromolecules. If that were not bad enough, the data that is available often presents

multiple difficulties for automated attempts at data mining and prediction. In this article, we first describe the crystallization process, and then show that by utilizing the PDB-deposited structures and sequences along with the information whether or not that sequence produced a crystal, it is possible to provide predictive glimpses of the experimental outcome. We then review attempts at employing detailed screening information in order to provide further insight. We conclude with a discussion of the incentives for improved bookkeeping and with an outlook towards future research in computational crystallization.

2. Crystallization

Crystallization has been described as an "empirical art of rational trial and error guided by past results" [6]. For better and for worse, this still sums up the state of the art of the field. When presented with a biological macromolecular sequence it is currently impossible to know what conditions will result in crystallization or even if crystallization will occur at all. Basic predictive techniques can determine, for instance if the target may be associated with a membrane, providing an initial lead into the appropriate crystallization approaches, but little further guidance is available. The history of crystallization methods and the development of modern crystallization screens have been summarized elsewhere [7]. It is, however, important to note that the development of readily prepared commercial screens make initial crystallization screening experiments easier, but at the risk of standardizing the starting points and introducing a bias on outcome. We will not discuss the details of the available screens or approaches here, but rather sketch the features of these methods that are essential to describing the role and purpose of experimental crystallization data.

The main strategy in macromolecular crystallization is to gradually bring a target solubilized in an appropriate aqueous chemical cocktail (containing at least a precipitating agent and a buffer) to a region of supersaturation until a crystal nucleates and grows. Interestingly, experimental outcomes reveal that there are many landmarks on the screening landscape that are not crystals. Crystals form under solution conditions that fall between those producing clear drops (conditions under which macromolecular—solvent interactions are stronger than interactions between macromolecules) and those producing precipitate (conditions under which the opposite is true). Clear and precipitate are common outcomes but phase separation, skin-formation, and a combination of any of these outcomes can also occur. These outcomes are markers on the solubility landscape that can be used to direct experiments towards a crystal [8]. For the vast majority of reports, however, these outcomes are not captured, restricting significantly the insights that can be gained from failed crystallization experiments into profiling the phase response of a macromolecule to different chemistries.

To understand the potential insights it is useful to review the thermodynamical framework of crystallization screening experiments. First and foremost, macromolecular crystallization is a phase transition. To study the coexistence of the crystal and the solution forms of a macromolecule, one needs to determine the set of conditions at which the chemical potentials of the two phases are equal. The chemical potential determination from first-principles entails knowledge of effective interactions between macromolecules [9–12]. Unfortunately, although the individual forces involved in macromolecular interactions are well understood, there does not exist a simple and fast way to figure out how these forces collectively build up effective macromolecular interactions [13,14].

Both experimental and theoretical studies of crystallization inevitably involve the interpretation of phase diagrams. These recapitulate the conditions under which different phases are thermodynamically stable. For mixtures such as crystallization cocktails, i.e. the mixture of chemicals designed to drive crystallization, the phase diagram is multi-dimensional. While not true in general, these multi-dimensional macromolecular phase diagrams can often be projected on the temperature-macromolecular concentration plane. This projection changes the temperature scale, however, as the contents of the cocktail, such as co-solutes, alter the energy scale of macromolecular interactions (while the experimental temperature itself is barely changed). The composition of the final phase can also be complex. A macromolecular crystal, for instance, contains on average more than 50% of the mother liquor [15] and may embed co-solutes in fractions that differ from what is left in the crystallization cocktail.

The arduous task of experimentally determining phase diagrams was carried out for a handful of proteins, including lysozyme, γ-crystallins, insulin, and myoglobin [16–22]. Despite the complexities introduced by the macromolecular structure and the additives in the crystallization cocktail, the phase diagrams of these proteins resemble that of simple liquids [23], in the sense that phases analogous to the crystal, liquid, vapor, and supercritical fluid are distinguishable. The phase diagram of simple liquids (shown in Fig. 1a), however, exhibits a vapor—liquid coexistence line that terminates at a stable critical point. Below this line (in the dark gray area) the system is unstable, and above the critical temperature, it is a supercritical fluid. One key difference is that for proteins the critical point typically lies below the crystal solubility line, meaning that the protein vapor and liquid phases are metastable. For proteins (Fig. 1b), below the critical point, the system has a propensity to aggregate in a disordered, percolating network, i.e., a gel; the gelation probability increasing with distance to the critical point [24,25]. Although such an aggregation is metastable, it can be long-lived, which arrests the crystallization process in the timescale of experiments. The area between the solubility line and the critical point is where crystallization is most likely to occur, and has consequently been dubbed the nucleation zone or crystallization gap [26,27].

Macromolecular phase diagrams are informative of the conditions in which crystallization is most likely to occur. However, their experimental determination is a task more challenging and time-consuming than even the most ambitious crystallization screen. Furthermore, theoretically determining phase diagrams to guide crystallization experiments suffers from the fundamental caveat that the macromolecular structure is required, because effective macromolecular interactions cannot be reliably determined
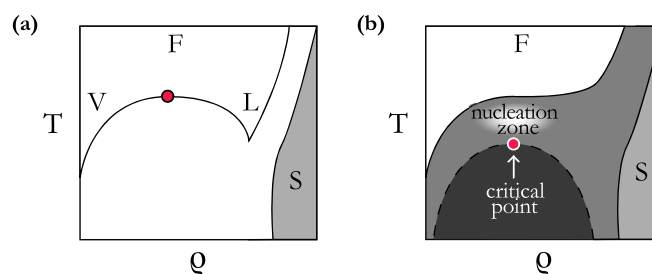

Fig. 1. Phase diagram of (a) a simple liquid and (b) a simple protein in the density—temperature plane (T-ρ). For crystallization cocktails, the projection of the multi-dimensional phase diagram onto the T-ρ plane typically involves rescaling the temperature axis by an aggregate function of the solution conditions. Vapor, liquid, solid, and fluid phases are denoted by V, L, S, and F, respectively. For macromolecules, the critical point [73] lies below the solubility line; for simple liquids, this critical point is located at the top of the (hidden) solubility curve. A supersaturated protein solution is typically found to be most likely to crystallize in the nucleation zone — the region between the solubility line and the critical point.

otherwise [64,73] Therefore, screening remains the most efficient way of probing the space of solution conditions [28], even though it can be done only very sparsely. Gaining useful insights about macromolecular crystal assembly requires extracting the key features of the phase diagrams from the limited information provided by experimental screening, and hope to use insights from simpler models to bridge the vast remaining gaps. Hence, although we have a reasonably good understanding of crystallization, the chemical space that needs to be sampled to make use of it is remarkably vast.

3. Experimental crystallization data

The protein data bank (PDB) is the international archive for all macromolecular structural models and, for crystallography, the X-ray diffraction data used to build these models. For every model in the PDB, a record exists of the crystallization conditions within an unstructured and non-mandatory remark field. Attempts have been made to enhance the usefulness of this record by standardizing chemical definitions and by enabling data queries [29]. The remarks themselves have also been analyzed revealing that almost 40% of the reported crystallization conditions can be formed by the optimization of a commercial screen [30]. Yet, although the PDB remains the largest source of successful crystallization data, over 30% of its structures contain no crystallization details. When those details are present, they are not systematically recorded, and of course, only the condition used to produce the structural model is captured. Despite these caveats, the information present is clearly useful. For example, Kirkwood et al. [31] used "cleaned" PDB data to confirm that acidic proteins (pI < 7) tend to crystallize at 1.0 pH unit above their pI and that basic proteins (pI > 7) tend to crystallize 1.5–3.0 pH units below their pI, in agreement with previous studies [32]. However, the PDB captures no information on other chemical conditions that may have yielded crystals, the range of chemical space sampled, or the other experimental outcomes that result. This is not surprising as the role of the PDB is to capture the structural model and not the complete experimental results to reproduce the result. Similarly, negative results are rarely reported in the literature. Much information that could be useful to understand the crystallization process itself is, however, lost.

Only a couple of dedicated crystallization condition databases exist. The Biomolecular Crystallization Database (BMCD) includes the subset of the information already captured in the PDB, i.e., about 43,000 entries [33], and the Marseille Protein Crystallization Database (MPCD) [34] combines an older version of the BMCD with the CYCLOP database (no longer available) having a total of about 15,000 entries. A few studies have used these databases to suggest improved crystallization strategies [35,36], but their overall influence has been limited. Over the last decade, however, a number of crystallization facilities have also been gathering datasets that contain the experimental details for each crystallization attempt, for both positive and negative results. This may open the road to types of analysis. Before detailing these efforts, we first review advances made with the datasets that existed prior to this advance.

4. Analysis of successful crystallization conditions

The most accessible and immediate source of information about a macromolecule is whether a model of its structure has been obtained from crystal or not up to this point in time. Because of the intensity of the decades-long efforts at obtaining such models, that information at least reflects in part the experimental difficulty or facility of crystallizing a target.

4.1. Prediction of crystallization propensity

Crystallization targets are typically chosen based on two criteria: (i) their biological interest, and (ii) their likeliness to crystallize. To evaluate the latter, a number of properties, such as the number of homologs that have been successfully crystallized, chain length, the frequency of charged residues, and the number of predicted transmembrane helices have long been taken into account [37]. More recently, a number of crystallization propensity prediction tools have also been developed using structural information from the PDB and information about the primary sequences that yield crystal structures or not from databases such as TargetDB [38] and PepcDB (now merged as TargetTrack) as well as UniRef50 [39]. These tools infer from the primary sequence alone the probability with which a macromolecule can be crystallized, which helps select targets more systematically.

SECRET [40] was one of the first such tools developed, and was shortly followed by others, including OB-Score [41], CRYSTALP [42], and ParCrys [43], but some of these tools (SECRET and CRYSTALP) have since been rendered obsolete by their low residue limit. In general, the success of these predictors depends strongly on the dataset with which they were trained. Newer training sets are thought to yield better predictors in part because advances in experimental techniques render crystallizable some sequences that used to be deemed uncrystallizable. Testing older prediction software with newer data thus often reveals their lower predictive power. Some of these tools also check features, such as the number of homologs of a given sequence in the PDB, which change with time. In addition, OB-Score, CRYSTALP2 [44], ParCrys, SVMCRYS [45], MetaPPCP [46], and CRYSpred [47] all use similar training datasets that, according to Mizianty and Kurgan [37], are outdated as they were prepared under the assumption that the sequences for which the experiments have been stopped are uncrystallizable.

Quantitatively, the reported accuracy of these tools can be measured by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where TP and TN are the number of true positives and negatives, and FP and FN are the number of false positives and negatives. Note that this quantity gives, for equal fractions of positive and negative results, a minimal accuracy of 50%. The Matthews correlation coefficient (MCC),

$$\text{MCC} = \frac{TP \quad TN \quad FP \quad FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \tag{2}$$

is also used, where a MCC = 0 implies that the prediction method does no better than randomly picking outcomes, while MCC = 1 implies a complete predictive power.

In what follows, we detail some of the tools available:

XtalPred calculates and predicts several protein properties from the primary sequence, including the average hydrophobicity, pI, the frequency of cysteines, methionines, tryptophans, tyrosines, and phenylalanines, disordered regions, and transmembrane helices. The chain is then assigned a score between one and five by comparing these properties to the distribution of properties in TargetDB. A score of five corresponds to very difficult to crystallize and one means "optimal" for crystallization [48]. This prediction system was later expanded by using random forest classifiers and by including surface ruggedness, surface entropy and the surface amino acid composition to the list of protein properties. It was also renamed XtalPred-RF [49]. Note that the

use of random forest classifiers was first prototyped in RFCRYS [50]. This method was selected because it was found to perform better compared to support vector machine (SVM) and artificial neural network methods. Both XtalPred and XtalPred-RF have accuracies around 70%, but the latter has a higher MCC of 0.47. PPCpred [37] and PredPPCrys [51] use a SVM classifier of exposed and buried amino acid compositions, chain disorder, the proximity of certain groups of amino acids and physico-chemical properties of individual and collocated amino acids. This large set of features (~800 for PPCpred and ~3000 for Pre-dPPCrys) is then reduced to obtain a best-performing set. In addition to predicting crystallization probabilities given a sequence, both PPCpred and PredPPCrys also predict the success of other experimental steps, such as purification and the pro-duction of well-diffracting crystals. PredPPCrys uses an enhanced feature selection process, a two-level SVM classifica-tion system, and an updated dataset. The PredPPCrys server can predict a protein's propensity of yielding well-diffracting crys-tals with 76% accuracy and an MCC of 0.43, while PPCpred, when analyzed with the same datasets, yields an accuracy of 64% and an MCC of 0.25.

SVMCRYS is an SVM-based method [45] that first used a set of 116 protein features, including amino acid frequencies (indi-vidually, in clusters and by amino acid type) and physico-chemical properties, such as solvent accessible surface area, hydrophobicity and side chain hydrophobicity. A reduced set of significant features was then identified. As stated above, this method uses an earlier training dataset, and hence its prediction accuracy drops when tested with newer datasets [51]. The re-ported accuracy was 87%, with MCC = 0.74. It should also be noted that when the same test set is used to assess XtalPred, the accuracy (78%) and MCC (0.58) were higher, while these values for ParCrys and OB-Score did not differ significantly from what was originally reported.

PPCinter [52] is another SVM-based predictor that predicts separately the outcome of the same experimental steps as PPCpred. This method uniquely takes into account inter-protein properties, such as the amino acid composition in protein re-gions predicted to be involved in protein−protein interactions. The reported accuracy (80%) and MCC (0.55) are about 3% and 0.03 higher, respectively, than PPCpred.

P_{XS} assigns a crystallization probability to proteins by incorpo-rating features that were found to be independent factors impacting crystallizability informed by logistic regression studies, namely the fraction of disordered residues, the average side chain entropies of exposed residues, the fraction of buried glycines, and the fraction of phenylalanine [53]. By contrast to the other prediction servers mentioned here, this system uses both the structure of crystallized proteins and their experi-mental propensity of crystallization success. However, this tool, as well as MCSG Z-score [54], is limited by its use of data from only one structural genomics center for training, and may thus perform poorly in predicting crystallizability of proteins that differ substantially from those in this relatively small training set. Logistic regression also cannot capture non-monotonic trends [55].

XANNpred is an artificial neural network trained on a dataset from both PDB and sequence data [56]. In total 428 protein properties were considered, including pI, average hydropho-bicity and the frequency of neighboring pairs of amino acids. The accuracy for this method is between 75% and 81% and MCC ranges from 0.50 to 0.63.

SCMCRYS is an ensemble method that creates "scoring cards" based on the identity of amino acid pairs separated by 0−9 residues as set of features [57]. The amino acid pairs are associated with a score derived from the frequencies of the pair in crystallizable and non-crystallizable proteins in the training set. These scores are then optimized to maximize the predic-tive power, and used to predict the crystallizability of the given protein sequence. This method uses the newer training data-sets of PPCpred. Its accuracy is found to be about 76% and has an MCC of 0.44. Its authors note, however, that the SVM-based PPCpred has a slightly better performance, because it includes more features and SVM can obtain more precise decision boundaries. One of the main purposes of this method is to create more easily interpretable results than those from ma-chine learning.

In order to compare performances of the available predictors on an equal footing, it is crucial that the same dataset (consisting of a wide variety of proteins) is used to calculate accuracy, MCC, and other such measures. Without such a comprehensive study, singling out a best method for prediction is difficult. Nevertheless, based on their relatively new training datasets, the wide range of protein features considered, and the appropriateness of the statis-tical methods chosen, tools such as PredPPCrys, PPCinter, and XtalPred-RF are expected to give better guidance in selecting crystallization targets.

One should, however, keep in mind that most of these methods use structural properties predicted from their sequence alone, a process that is prone to error and thus fundamentally limits the accuracy of the predictors (and why crystallization is being attempted in the first place). This limitation is reflected in the relatively low MCC they produce. Although they provide guidance, these tools are far from being complete solutions to predicting outcomes of crystallization trials. Marked improve-ments require a better understanding of the microscopic prop-erties that promote crystallization and of the crystallization mechanisms themselves, as well as a continuous update of the training dataset.

4.2. Statistical analysis of databases and side chain entropy reduction

The structural models in the PDB have also been surveyed to identify macromolecular features that affect crystallization and thus more specifically guide mutagenesis strategies. The most developed of these approaches is the side chain entropy reduction (SER) method. The method is motivated by the observation that immobilizing side chains at crystal contacts between macromole-cules has an entropic cost than can disfavor their formation. This idea is also statistically supported by an analysis of the correlation between protein structures and crystallizability [55,58]. SER thus advocates mutating residues with high side chain entropies (SCE), such as lysine, to lower SCE residues, such as alanine. The approach was first successfully applied to Rho GDP-dissociation inhibitor, chosen for its high lysine and glutamate content and low crystal-lization propensity [59].

Although high SCE residues are predominantly found at the surface of macromolecules, choosing among possible mutagenesis sites from the input amino acid sequence alone is generally a non-trivial task. A more sophisticated algorithm for SER was thus developed and implemented on the SERp server. It suggests resi-dues for mutation based on: (i) predictions of the secondary structure and the surface SCE, and (ii) sequence homology to identify residues poorly conserved by evolution, and thus likely less functionally or structurally relevant [60]. Residues are individually scored based on these two criteria and, adjacent, high-scored res-idues are considered potential sites for mutagenesis. Because removing too many high SCE residues may reduce the solubility of

the macromolecule,[1] the algorithm also minimizes the number of proposed mutations for a given SER target. In the end SERp identifies mutations to lysine, glutamate, and glutamine residues. Although successful mutagenesis cases have also mutated asparagine, SERp does not allow it because that residue can also engage in favorable protein–protein interactions. For a number of test cases, the SER strategy has been shown to lead to the crystallization of otherwise recalcitrant macromolecules, and to improve the crystal quality of others [61]. Thus, protein engineering, of which SER mutagenesis is an example, is one of the means with which the experimental outcome can be improved [59,62,63].

It should be noted that, although SER help attack the crystallization problem, the physicochemistry of macromolecular crystallization is more intricate than the simple label suggests. For instance, mutating lysine to arginine is likely an alternative for the lysine to alanine mutation – despite them having nearly the same SCE – because favorable electrostatic interactions that enthalpically drive the formation of crystal contacts (with a nearby aspartate, for instance) could otherwise be lost [55,61,64]. Experimental studies indeed show that the interaction enthalpy of the aspartate-–arginine pair is $\Delta H = -6.3$ kJ/mol, but only $\Delta H = 1.3$ kJ/mol for the aspartate–lysine pair [65], which is partly due to the aspartate-–arginine pair's ability to form two hydrogen bonds, compared to the single hydrogen bond between aspartate and lysine [66]. In addition, hydrogen bonds, whose energy can range from ~2 kJ/mol to ~20 kJ/mol [67], can more than compensate the high SCE cost of binding with the backbone. These strong enthalpic contributions thus make it less surprising that some crystal contacts form despite their high SCE cost.

It should also be noted that while changing biochemical conditions for crystallization is straightforward, altering the macromolecule itself – as SER suggests – requires that it can be expressed, and therefore needs additional effort. Although it may have a dramatic impact on outcome, it is certainly not the easiest experimental route to follow.

5. Analysis of complete crystallization screening data – success and failure

Although sparse and imperfect, crystallization screening, either through commercial kits or individual laboratory efforts, remains the most efficient method of probing the phase diagram [28]. However, little to no information (other than overall success or failure) has been extracted from this screening process in general. In part, this is within the very nature of the scientific enterprise as most studies aim not to understand the fundamentals of crystallization, but rather to obtain a structural model of the macromolecule or complex studied. Compounded to this cross-purpose, only the final crystallization condition is usually reported – any initial screening results are masked by the process of optimization where an iterative process refines the initial promising screening result. Access to a data or a database with complete and detailed experimental information would allow a more complete study of the effect of chemical conditions on crystallization probability and outcome in relation with macromolecular properties. Given the limited but important information that has been gained from a

binary crystal or no crystal analysis of limited crystallization data, the results of such analyses could be instructive about how to optimize screen conditions and a lot more. We detail below two types of insights that have been teased out of this analysis as well as means to push this general idea forward.

5.1. New biological knowledge

We know that chemically related conditions are likely to result in similar outcomes, and conversely, chemically distinct conditions are likely to produce different outcomes. This observation is fundamental to the design of different crystallization screens for different classes of targets. This knowledge can be used to extract information from crystallization screening of a single macromolecular target if the chemical conditions can be mathematically described and compared.

Newman et al. [68] pioneered a similarity metric, termed the C6 metric, that assigns a quantitative value to the similarity between two or more crystallization cocktails, and allows those that are chemically similar (through obvious, or less-apparent relationships) to be distinguished from those that are chemically distinct. Using similar methods, related cocktails can be displayed with hierarchical clustering and experimental outcomes overlaid on the cluster [69]. Applying this approach to an exopolyphosphatase-related protein from *Bacteroides fragilis*, (BfR192) showed that biological information is contained within the collection of crystallization screening outcomes. A fact that may not be apparent from considering a single outcome alone – successful crystallization or not. As an example of this effect, BfR192, is a 343 residue protein with a molecular weight of 39.77 kDa that was screened in our laboratory using 1536 different cocktails. Fig. 2 overlays the location of specific cocktails that produced crystals on the clustering of 1536 different biochemical conditions. Out of the 28 clusters, 11 produced at least one crystal hit.

Analyzing the outcomes, 70 conditions out of the full 1536 produced initial crystallization hits. The highest frequency of hits was in cluster 13 followed by cluster 14, and then by cluster 12. In terms of the percentage of hits, cluster 12 ranks as the first with 19% followed by cluster 13 with 18% yielding hits. (The total number of cocktails in each cluster is variable due to the design of the screen incorporating commercial screens which operate on differing principles, *e.g.* grid screening, identifying particular chemical species, the use of multiple small molecules, cryogenic compatibility or incomplete factorial sampling of chemical space as used by the non-commercial condition sampling). Linking the clusters back to their chemistry, cluster 13 proved interesting in that sodium was present in 73% of the conditions as opposed to 47% for the 1536-condition-screen overall, potassium was present in 72% of the conditions versus 24% overall and finally phosphate was present in 100% of the conditions versus 16% overall.

The initial structure was deposited in the PDB with undefined electron density. Based on the clustering analysis a sodium ion was placed in the structure, Fig. 3, which improved the match of the model to the electron density as did the inclusion of a potassium ion and of four phosphates. The R and R_{free} reduced from 22.3% to 25.9%–20.7% and 24.3%, respectively. These ions also proved to be biologically relevant. The location of the phosphate-binding pocket suggests that the phosphoryl moieties of polyP might anchor in this pocket. The putative active site has features that are consistent with active sites of other phosphatases involved in binding the phosphoryl moieties of nucleotide triphosphates [70]. A possible role of the active site phosphate is to ensure a proper substrate orientation and polarization of the phosphoryl P–O bond to increase the susceptibility of the P atom to nucleophilic attack. The space around the phosphate ions suggests that the cleft can bind longer polyP

---

[1] In addition to being potentially deleterious to protein solubility, such mutations could also reduce the very structural stability of the macromolecule. Nevertheless, it has been shown that stability does not correlate with its crystallizability, unless the protein is either unfolded (harder) or "hyperstable" (easier) [53] Price, W.N., 2nd, Y. Chen, S.K. Handelman, H. Neely, P. Manor, R. Karlin, R. Nair, J. Liu, M. Baran, J. Everett, S.N. Tong, F. Forouhar, S.S. Swaminathan, T. Acton, R. Xiao, J.R. Luft, A. Lauricella, G.T. DeTitta, B. Rost, G.T. Montelione, and J.F. Hunt, Nat Biotechnol, 27 (2009) 51-7.
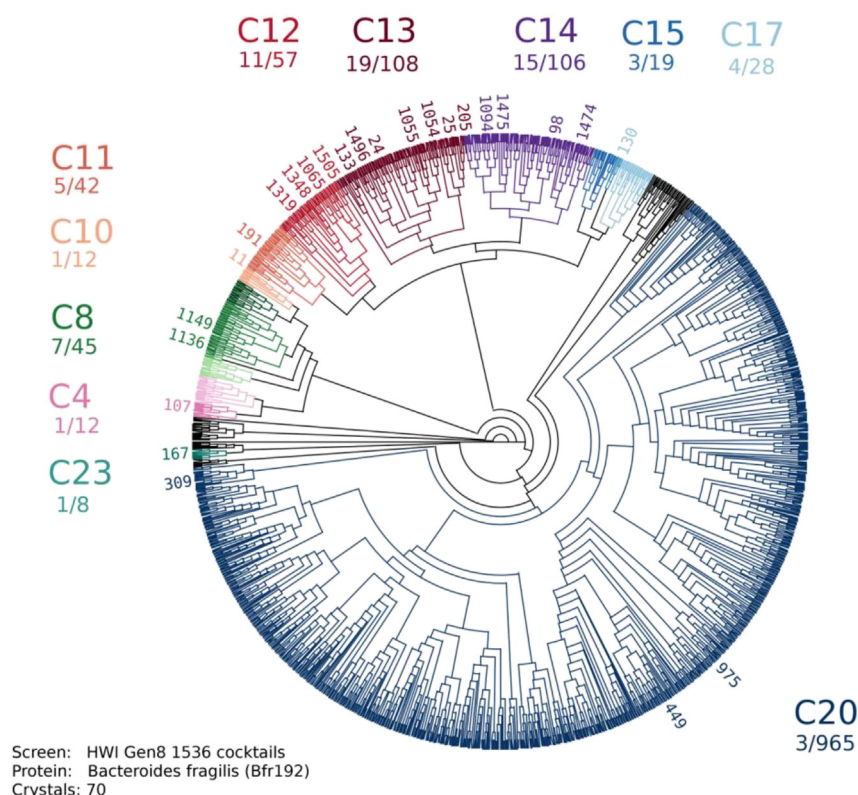
Fig. 2. Regions of crystallization space where hits for BfR192 were found. Out of the 28 clusters, 11 were identified containing at least one result determined visually to be crystalline. Adapted from Bruno et al. [69].
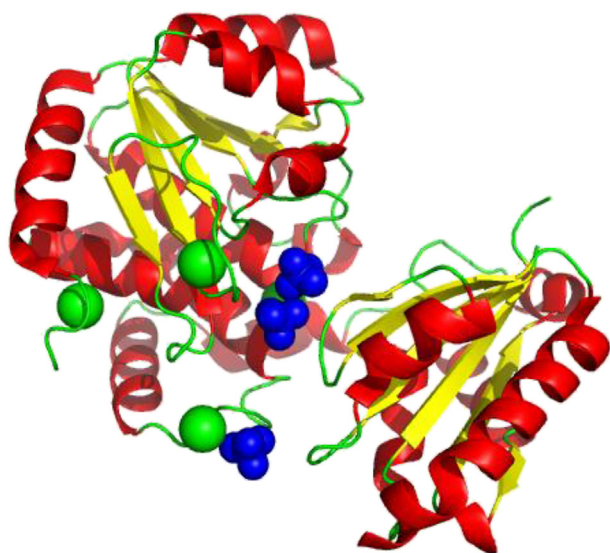

Fig. 3. Structure of the BfR192 exopolyphosphatase-related protein (PDB ID 4PY9) showing the two domains and highlighting the cleft containing the sodium, potassium and four phosphate ions [69].

substrates. An interesting observation was the lack of significant crystal formation in the PEG-based conditions (cluster C20), demonstrating one of the two classes of macromolecules revealed in the next section.

In short, the analysis of crystallization results from the initial screen aided by a mathematical description of chemical cocktail similarity identified the presence of ions and a putative biological function. Even for a single target and only considering crystal or no-crystal outcomes, the analysis of the crystallization results revealed new biological information.

5.2. Analysis of crystallization propensity

From the description of the prediction schemes detailed in Section 4, it is natural to wonder what cocktail properties lead to higher crystallizability. In order to productively conduct such a statistical analysis, it is crucial that the dataset assayed systematically contains both positive hits and negative results of crystallization experiments, as well as details of the experimental conditions.

In this context, the Northeast Structural Genomics consortium (NESG) database, which contains both the crystallization outcome and the full description of chemical conditions for 1536 cocktails, is an ideal model. From an analysis of a subset of its content, one could statistically reconstruct the experimental crystallization propensity, ξ, — the fraction of hits — based on structural features of the macromolecules alone [55]. Such a reconstruction revealed two mechanisms for crystallization: (i) driven by hydrophobic interaction and, (ii) reliant on specific electrostatic interactions. This nuance was also suggested by Price et al., in a study of a subset of the NESG database [53], although the statistical signature was not as clear. Interestingly, this further dataset enabled studying the response in ξ upon varying chemical conditions, which revealed that proteins identified to crystallize with the second mechanism have a propensity to crystalize with changes in salt concentration rather than molecular crowding caused by PEG, for example.

6. Technical considerations for computational crystallization

In order to expand the scope and capability of the above analysis, we discuss here four technical considerations.

6.1. Systematicity and completeness of datasets

A key prerequisite of the dataset for the study discussed in Section 5.2 was that ξ was obtained from experiments that span the same chemical space. Comparing results for different screens would be less reliable, and changing the number of screens would bias the "resolution" of ξ, especially when it is small. The details of the chemical composition are also essential. While some steps have been made to account for this in other studies with similarity metrics employed [55,68], more detailed knowledge of the chemical processes has to be captured. For instance, the counter-ion identity can be important to consider as different species can have different salting-out abilities. Proper bookkeeping of chemical conditions as well as individual and combined properties also allows efficient parsing of the data. Because increasing the number and diversity of proteins in the dataset should help reveal subtler trends, this effort could have a significant payoff.

6.2. Crystallization technique

For the studies described in Section 5, the initial method of crystallization was microbatch-under-oil. With the exception of concentration changes due to dehydration through the plate, the microbatch method is a static one. There is no dynamic change of component concentration as occurs in vapor diffusion, the predominant crystallization method. Although no systematic study of the matter has been conducted, one may refer to the physical basis of crystallization for insights. Under-oil experiments keep the solution volume constant, while vapor diffusion gradually supersaturates the protein solution as the droplet volume shrinks. Yet the typical increase in concentration of salt and other co-solutes change the pH and the chemical potential of the co-solutes relatively little, because both depend logarithmically on concentration. The protein concentration, however, may change sufficiently for the system to cross the phase boundary into the crystallization regime and to significantly lower the barrier to nucleation (Fig. 1). Hence, although the volume change and the consequent change in the chemical properties of the solution can likely be ignored in the analysis as a first approximation, the change in macromolecular concentration likely cannot.

6.3. Optimization of crystallization conditions based on screening

Commercial kits developed over many years of experience have made the initial crystallization screening process relatively easy for most laboratories [7]. The development of such screens involves taking an initial promising condition and changing the chemistry to optimize that condition. This, however, has two shortcomings: (i) optimization occurs around common fixed starting points, and (ii) focus may be on the single best initial result, causing separate but distinct clusters, as revealed in the example from Section 5.2, to be overlooked. Condition optimization is thus an important aspect that requires attention in analyzing crystallization outcome as it can impart an unintentional yet non-fictitious bias to the results.

6.4. Characterizing outcome

To best utilize the theoretical knowledge of crystallization, the experimental outcome has to be classified into distinct regions of the protein phase diagram. The phase diagrams obtained for model proteins (section 2) relied on very defined crystallization screens with few variables. In reality, screening a structurally uncharacterized macromolecule probes multiple variables. Given the number of experiments required to cover that chemical space, considerable effort is required to identify potential crystal conditions, let alone other outcomes. When these outcomes are noted for a single macromolecule, they can be used to construct a vector for crystallization within the surveyed chemical space [71]. However, human classification becomes an impediment when a multitude of macromolecules are studied.

In a detailed study led by one of us, around 150,000 images of the crystallization plates from 96 distinct macromolecules were visually classified by a team of eight viewers such that each image was surveyed by a minimum of three viewers. A control set unknown to the classifiers was included at the beginning, middle and end of the classification effort. Each image was visually classified into seven categories: clear, phase separation, precipitate, skin, crystals, junk, and unsure. Multiple classifications were allowed for all but "clear". This not only created a training set for automated image classification attempts, but also revealed the reliability of humans in the process. Using a deliberately strict definition of common classifications (classifications had to be exactly equal), the average agreement for classifications of the control set was 78%. When comparing the whole data set, approximately 48% of the images were unanimously classified by three separate viewers. Again a deliberately strict interpretation of agreement was used where multiple classifications, *e.g.* skin and precipitate, were regarded different from either skin, precipitate or the combination and a third component. This work establishes a baseline for automated classification success [4]. Using this training set and extensive feature extraction approach, automated classification can successfully recognize 80% of crystal containing images, 94% of clear images and 94% of precipitate. Although this approach is promising, the runtime is on the order of seconds per image using high-end graphical processing units [72], which limits its large-scale deployment.

Many other attempts have been made in automating crystallization screening classification. To date none are routinely available. A limiting factor in making use of the complete information produced by crystallization screening is the efficient characterization of outcome.

7. Outlook for the future

A fairly simple analysis of crystallization screening outcome from multiple experiments on a single macromolecule shows that useful biological information can be obtained. Not discussed here, but also a clear potential, is the ability to distinguish between different states of a macromolecule that may be present as a function of common crystallization conditions. A more global analysis on multiple macromolecules reveals two broad categories that correspond to salt-based crystallization approaches or other classes of precipitants. Membrane-bound macromolecules can be predicted and distinct crystallization approaches used in those cases. These observations, coupled with the success of commercial screens targeting specific macromolecular groups, *e.g.* nucleic acids, suggest that crystallographic screening results contain both real and useful information.

The theory of macromolecular crystallization is increasingly well developed. Phase diagrams are fairly well understood but the chemical space that needs to be sampled for a biological macromolecule remains humongous. With a more comprehensive description of chemical components and their interactions, experimental outcome classified on the basis of phase diagram regions, and accurate information on the macromolecular sample including

those that fail, successful predictive techniques for both screening and optimization should be in reach. Admittedly, these requirements are not trivial. Although they are not necessary for producing an individual structural model, they are *essential* to advance the field of crystallization, so that many more such models can be provided.

With only one in ten soluble macromolecules on average yielding a structural model today, advances in crystallization success will likely have to be more dramatic than deriving appropriate chemistry for optimizing conditions. Mutation of the sample to improve the crystallization properties appears to be the next step following biochemical or biophysical (*e.g.* temperature) approaches. A full computational analysis of initial crystallization screening would be more robust in determining the most recalcitrant samples and prioritizing the molecular biology approach.

Computational crystallization approaches are useful in many cases. Achieving their full potential requires significant increase in the amount and quality of available data. To make this happen in the high-throughput laboratory at the Hauptman-Woodward Medical Research Institute and partners in similar laboratories around the globe, which have privileged access and responsibility to collect and curate these datasets, aim to keep on pioneering the analysis and the best practices. Yet, this work can be a challenge to motivate in a conservative funding context. The potential for computational crystallization is thus here but remains to be fully realized.

Acknowledgments

References

[1] H.M. Berman, J.D. Westbrook, M.J. Gabanyi, W. Tao, R. Shah, A. Kouranov, T. Schwede, K. Arnold, F. Kiefer, L. Bordoli, J. Kopp, M. Podvinec, P.D. Adams, L.G. Carter, W. Minor, R. Nair, J.L. Baer, Nucleic Acids Res. 37 (2009) D365–D368.
[2] J.R. Luft, J. Wolfley, I. Jurisica, J. Glasgow, S. Fortier, G.T. DeTitta, J. Cryst. Growth 232 (2001) 591–595.
[3] E.H. Snell, A.M. Lauricella, S.A. Potter, J.R. Luft, S.M. Gulde, R.J. Collins, G. Franks, M.G. Malkowski, C. Cumbaa, I. Jurisica, G.T. DeTitta, Acta Crystallogr. Sect. D Biol. Crystallogr. 64 (2008) 1131–1137.
[4] E.H. Snell, J.R. Luft, S.A. Potter, A.M. Lauricella, S.M. Gulde, M.G. Malkowski, M. Koszelak-Rosenblum, M.I. Said, J.L. Smith, C.K. Veatch, R.J. Collins, G. Franks, M. Thayer, C. Cumbaa, I. Jurisica, G.T. Detitta, Acta Crystallogr. D Biol. Crystallogr. 64 (2008) 1123–1130.
[5] J. Newman, E.E. Bolton, J. Muller-Dieckmann, V.J. Fazio, D.T. Gallagher, D. Lovell, J.R. Luft, T.S. Peat, D. Ratcliffe, R.A. Sayle, E.H. Snell, K. Taylor, P. Vallotton, S. Velanker, F. von Delft, Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 68 (2012) 253–258.
[6] N.E. Chayen, T.J. Boggon, A. Cassetta, A. Deacon, T. Gleichmann, J. Habash, S.J. Harrop, J.R. Helliwell, Y.P. Nieh, M.R. Peterson, J. Raftery, E.H. Snell, A. Hadener, A.C. Niemann, D.P. Siddons, V. Stojanoff, A.W. Thompson, T. Ursby, M. Wulff, Q. Rev. Biophys. 29 (1996) 227–278.
[7] J.R. Luft, J. Newman, E.H. Snell, Acta Crystallogr. F Struct. Biol. Commun. 70 (2014) 835–853.
[8] J.R. Luft, J.R. Wolfley, E.H. Snell, Cryst. Growth Des. 11 (2011) 651–663.
[9] A. McPherson, Crystallization of Biological Macromolecules, Cold Spring Harbor Laboratory, 1999.
[10] Z.S. Derewenda, Acta Crystallogr. D Biol. Crystallogr. 66 (2010) 604–615.
[11] M.J. Anderson, C.L. Hansen, S.R. Quake, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 16746–16751.
[12] E. Saridakis, N.E. Chayen, Trends Biotechnol. 27 (2009) 99–106.
[13] P.G. Vekilov, A.A. Chernov, Solid State Phys. Adv. Res. Appl. 57 (2002) 1–147.
[14] Y.D. Devedjiev, Acta Crystallogr. Sect. F Struct. Biol. Commun. 71 (2015) 157–162.
[15] C.X. Weichenberger, P.V. Afonine, K. Kantardjieff, B. Rupp, Acta Crystallogr. D Biol. Crystallogr. 71 (2015) 1023–1038.
[16] F. Bonnete, M. Malfois, S. Finet, A. Tardieu, S. Lafont, S. Veesler, Acta Crystallogr. D Biol. Crystallogr. 53 (1997) 438–447.
[17] A. Lomakin, N. Asherie, G.B. Benedek, J. Chem. Phys. 104 (1996) 1646–1656.
[18] M. Muschol, F. Rosenberger, J. Chem. Phys. 107 (1997) 1953–1962.

[19] A. Stradner, G.M. Thurston, P. Schurtenberger, J. Phys. Condens. Matter 17 (2005) S2805–S2816.
[20] J.J. McManus, A. Lomakin, O. Ogun, A. Pande, M. Basan, J. Pande, G.B. Benedek, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 16856–16861.
[21] S. Talreja, S.L. Perry, S. Guha, V. Bhamidi, C.F. Zukoski, P.J.A. Kenis, J. Phys. Chem. B 114 (2010) 4432–4441.
[22] C. Gogelein, D. Wagner, F. Cardinaux, G. Nagele, S.U. Egelhaaf, J. Chem. Phys. 136 (2012).
[23] J.P. Hansen, I.R. McDonald, Theory of Simple Liquids, Academic Press, London ; New York, 1976.
[24] A.C. Dumetz, A.M. Chockla, E.W. Kaler, A.M. Lenhoff, Biophys. J. 94 (2008) 570–583.
[25] J.A. Thomson, P. Schurtenberger, G.M. Thurston, G.B. Benedek, Proc. Natl. Acad. Sci. U. S. A. 84 (1987) 7079–7083.
[26] N. Asherie, Methods 34 (2004) 266–272.
[27] M. Sleutel, J.F. Lutsko, D. Maes, A.E.S. Van Driessche, Phys. Rev. Lett. 114 (2015).
[28] A.C. Dumetz, A.M. Snellinger-O'Brien, E.W. Kaler, A.M. Lenhoff, Protein Sci. 16 (2007) 1867–1877.
[29] T.S. Peat, J.A. Christopher, J. Newman, Acta Crystallogr. D Biol. Crystallogr. 61 (2005) 1662–1669.
[30] V.J. Fazio, T.S. Peat, J. Newman, Acta Crystallogr. F Struct. Biol. Commun. 70 (2014) 1303–1311.
[31] J. Kirkwood, D. Hargreaves, S. O'Keefe, J. Wilson, Acta Crystallogr. F Struct. Biol. Commun. 71 (2015) 1228–1234.
[32] K.A. Kantardjieff, B. Rupp, Bioinformatics 20 (2004) 2162–2168.
[33] M. Tung, D.T. Gallagher, Acta Crystallogr. Sect. D Biol. Crystallogr. 65 (2009) 18–23.
[34] M. Charles, S. Veesler, F. Bonnete, Acta Crystallogr. D Biol. Crystallogr. 62 (2006) 1311–1318.
[35] Y. Liu, X.F. Zhang, C.Y. Zhang, Y.Z. Guo, S.X. Xie, R.B. Zhou, Q.D. Cheng, E.K. Yan, Y.L. Liu, X.L. Lu, Q.Q. Lu, H.M. Lu, Y.J. Ye, D.C. Yin, Crystengcomm 17 (2015) 5488–5495.
[36] H.M. Lu, D.C. Yin, Y.M. Liu, W.H. Guo, R.B. Zhou, Int. J. Mol. Sci. 13 (2012) 9514–9526.
[37] M.J. Mizianty, L. Kurgan, Bioinformatics 27 (2011) I24–I33.
[38] L. Chen, R. Oughtred, H.M. Berman, J. Westbrook, Bioinformatics 20 (2004) 2860–2862.
[39] B.E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C.H. Wu, Bioinformatics 23 (2007) 1282–1288.
[40] P. Smialowski, T. Schmidt, J. Cox, A. Kirschner, D. Frishman, Proteins Struct. Funct. Bioinforma. 62 (2006) 343–355.
[41] I.M. Overton, G.J. Barton, FEBS Lett. 580 (2006) 4005–4009.
[42] K. Chen, L. Kurgan, M. Rahbari, Biochem. Biophys. Res. Commun. 355 (2007) 764–769.
[43] I.M. Overton, G. Padovani, M.A. Girolami, G.J. Barton, Bioinformatics 24 (2008) 901–907.
[44] L. Kurgan, A.A. Razib, S. Aghakhani, S. Dick, M. Mizianty, S. Jahandideh, BMC Struct. Biol. 9 (2009) 50.
[45] K.K. Kandaswamy, G. Pugalenthi, P.N. Suganthan, R. Gangal, Protein Pept. Lett. 17 (2010) 423–430.
[46] M.J. Mizianty, L. Kurgan, Biochem. Biophys. Res. Commun. 390 (2009) 10–15.
[47] M.J. Mizianty, L.A. Kurgan, Protein Pept. Lett. 19 (2012) 40–49.
[48] L. Slabinski, L. Jaroszewski, L. Rychlewski, I.A. Wilson, S.A. Lesley, A. Godzik, Bioinformatics 23 (2007) 3403–3405.
[49] S. Jahandideh, L. Jaroszewski, A. Godzik, Acta Crystallogr. D Biol. Crystallogr. 70 (2014) 627–635.
[50] S. Jahandideh, A. Mahdavi, J. Theor. Biol. 306 (2012) 115–119.
[51] H. Wang, M. Wang, H. Tan, Y. Li, Z. Zhang, J. Song, PLoS One 9 (2014) e105902.
[52] J.Z. Gao, J. Hu, Z.H. Wu, J.S. Ruan, S.Y. Shen, M. Hanlon, K. Wang, Curr. Bioinforma. 9 (2014) 57–64.
[53] W.N. Price 2nd, Y. Chen, S.K. Handelman, H. Neely, P. Manor, R. Karlin, R. Nair, J. Liu, M. Baran, J. Everett, S.N. Tong, F. Forouhar, S.S. Swaminathan, T. Acton, R. Xiao, J.R. Luft, A. Lauricella, G.T. DeTitta, B. Rost, G.T. Montelione, J.F. Hunt, Nat. Biotechnol. 27 (2009) 51–57.
[54] G. Babnigg, A. Joachimiak, J. Struct. Funct. Genomics 11 (2010) 71–80.
[55] D. Fusco, T.J. Barnum, A.E. Bruno, J.R. Luft, E.H. Snell, S. Mukherjee, P. Charbonneau, PLoS One 9 (2014) e101123.
[56] I.M. Overton, C.A. van Niekerk, G.J. Barton, Proteins 79 (2011) 1027–1033.
[57] P. Charoenkwan, W. Shoombuatong, H.C. Lee, J. Chaijaruwanich, H.L. Huang, S.Y. Ho, PLoS One 8 (2013) e72368.
[58] M. Cieslik, Z.S. Derewenda, Acta Crystallogr. D Biol. Crystallogr. 65 (2009) 500–509.
[59] Z.S. Derewenda, Struct. (Camb.) 12 (2004) 529–535.
[60] L. Goldschmidt, D.R. Cooper, Z.S. Derewenda, D. Eisenberg, Protein Sci. 16 (2007) 1569–1576.
[61] Z.S. Derewenda, P.G. Vekilov, Acta Crystallogr. D Biol. Crystallogr. 62 (2006) 116–124.
[62] Z.S. Derewenda, Methods 34 (2004) 354–363.
[63] B. Rupp, Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology, Garland Science, 2009.
[64] D. Fusco, J.J. Headd, A. De Simone, J. Wang, P. Charbonneau, Soft Matter 10 (2014) 290–302.
[65] V. Petrauskas, E. Maximowitsch, D. Matulis, J. Phys. Chem. B 119 (2015) 12164–12171.

ARTICLE IN PRESS

[66] J.B. Mitchell, J.M. Thornton, J. Singh, S.L. Price, J. Mol. Biol. 226 (1992) 251–262.
[67] S.Y. Sheu, D.Y. Yang, H.L. Selzle, E.W. Schlag, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 12683–12687.
[68] J. Newman, V.J. Fazio, B. Lawson, T.S. Peat, Cryst. Growth Des. 10 (2010) 2785–2792.
[69] A.E. Bruno, A.M. Ruby, J.R. Luft, T.D. Grant, J. Seetharaman, G.T. Montelione, J.F. Hunt, E.H. Snell, PLoS One 9 (2014) e100782.
[70] L.W.T.a.L.T.D. Allan Matte, Structure 6 (1998) 413–419.
[71] E.H. Snell, R.M. Nagel, A. Wojtaszcyk, H. O'Neill, J.L. Wolfley, J.R. Luft, Acta Crystallogr. D Biol. Crystallogr. 64 (2008) 1240–1249.
[72] Y. Kotseruba, C.A. Cumbaa, I. Jurisica, in: J. Physics: Conference Series, High Performance Computing Symposium (HPCS 2011), Canada, 2011, 341, 2012, pp. 1–8.
[73] D. Fusco, P. Charbonneau, Colloids Surf. B 137 (2016) 22–31.