CRYSTAL GROWTH & DESIGN

Article

pubs.acs.org/crystal

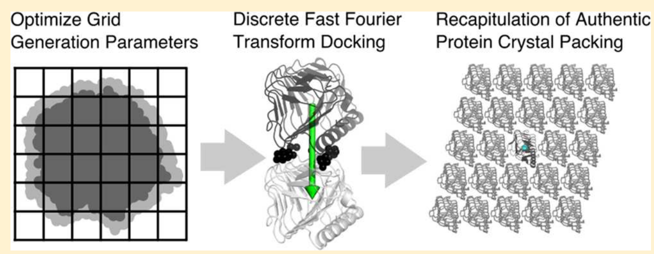
Optimizing Shape Complementarity Scoring Parameters for Recognition of Authentic Protein Crystal Packing Arrangements

Jeffrey A. Bennett and Christopher D. Snow*

1370 Campus Delivery, Colorado State University, Fort Collins, Colorado 80523-1370, United States

Supporting Information

ABSTRACT: The prediction of protein crystal packing arrangements and the *de novo* design of new crystal forms are notoriously challenging problems. For both problems, it is useful to enumerate high quality packing arrangements. To efficiently enumerate candidate crystal forms, we have adapted grid-based fast Fourier transform strategies first developed in the context of protein docking algorithms. To maximize performance of the grid-based, shape-complementarity scoring scheme, we have optimized parameters for the recognition of authentic protein crystal packing arrangements. To this end, extensive calculations were performed to assess a wide range of grid-representation parameter spaces for a panel of low-solvent protein crystals. The optimum parameters obtained from the computations successfully identify authentic protein crystal packing arrangements out of large pools of similar decoy packing arrangements.



1. INTRODUCTION

1.1. Motivation. Predicting crystal structures from molecular shape is a longstanding scientific challenge, and a solution would find applications in structural biology and pharmacology. In the former case, predicting packing arrangements could prove useful for structure determination via molecular replacement algorithms. In the latter case, the method might be adapted to predict crystal structures of small molecules such as pharmaceuticals. Pharmaceutical molecules frequently can crystallize into multiple arrangements (polymorphs) that can have very different pharmacokinetic properties.[1]

However, our current goal is not the prediction of packing arrangements for a specific sequence. Instead, our goal as protein engineers is to facilitate the design of novel protein crystals. For most applications that could use protein materials such as catalysis, separations, and light harvesting, fine control over the nanostructure of the material is useful and highly desirable. Pore size is an especially important feature to control since pore size is highly correlated to the mass transfer properties of the material.[2−4]

In the pursuit of designed protein crystals, "designability" is a key criterion.[5] To identify feasible packing arrangements, it would be very helpful to have an algorithm capable of efficiently generating realistic protein crystal packing arrangements. Candidate new crystal forms are more likely to be designable if they feature protein−protein interfaces with significant shape complementarity. Given high quality candidate crystal packing arrangements, crystal designers could proceed to use computational design methods[5−7] or alternative strategies such as surface entropy reduction.[8,9] Other factors being equal, crystals with low solvent fraction may be preferable design targets because protein crystals with less solvent tend to exhibit superior resolution via X-ray diffraction.[10] Dehydration postcrystallization can also improve diffraction resolution.[11−13]

Given the need to efficiently search through possible packing arrangements with high net shape complementarity, we adapted grid-representation methods that were originally developed for shape-based protein docking.[14] The free parameters for the algorithm (that control the grid representation of the protein building blocks) were trained to recapitulate protein crystals with dense packing arrangements.

1.2. Grid Representations of Protein Shape. This paper adapts a method originally outlined by Katchalski-Katzir[14] and refined by Gabb[15] to derive a three-dimensional (3D) grid representation for rigid protein building blocks. The algorithm should apply equally well to monomeric or oligomeric building blocks. However, for the cases presented here, the building blocks consisted of individual protein monomers.

Each protein was embedded on an $N \times N \times N$ grid corresponding to a box around the protein, with nodes every η Å in each dimension. Values were assigned to each discrete grid point based on how close the grid point was to the nearest protein atom (Figure 1). A grid point was defined as inside the protein if the distance is less than r, to any atom in the protein. The value for these interior grid points was set to ρ (a large negative number). Grid points well outside the protein (more than $r + t$ Å from the nearest protein atom) were set to 0. Finally, the surface layer grid points that are more than r Å from

Received: May 22, 2016
Revised: July 26, 2016
Published: July 26, 2016

DOI: 10.1021/acs.cgd.6b00769
Cryst. Growth Des. 2016, 16, 5579−5583

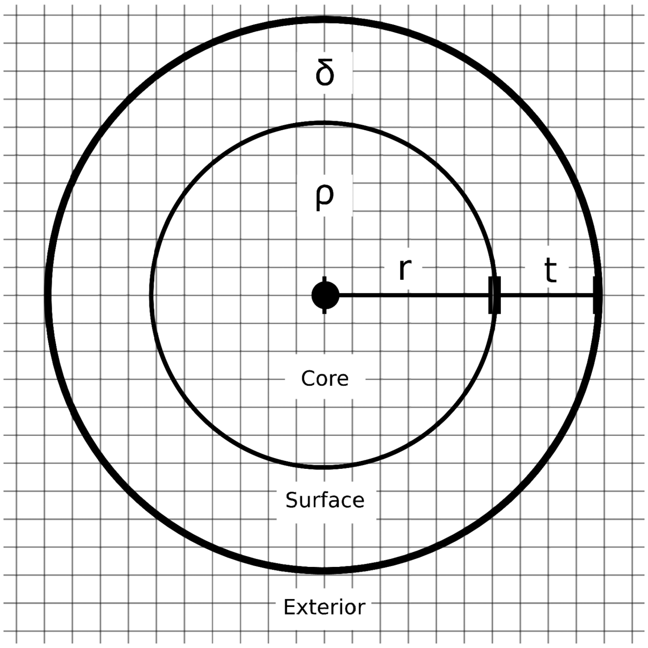


Figure 1. Katchalski-Katzir parameters for grid representation of a single atom center.

the closest protein atom, but less than $r + t$, were set to δ (a small positive number).

A second copy of the protein (the ligand in the original docking literature) receives the same treatment, but ρ and δ are set to different values, ρ is usually set to 1, and δ is set to 0. This grid construction scheme ensures a discrete correlation score of 0 when the proteins are not in contact, a favorable positive correlation when only the surface layer of the receptor overlaps with the ligand, and an unfavorable negative correlation if the proteins interpenetrate. Negative scores arise from the negative "core" receptor grid points ($\rho < 0$) multiplied with the positive ligand grid points ($\rho > 0$) leading to a negative value penalty for the core−core overlap. In this way, an $N \times N \times N$ overlap correlation map can be produced that contains the score for each relative displacement of the ligand protein with respect to the receptor protein in the defined grid space. The correlations for translations of the ligand protein can be quickly calculated using the discrete fast Fourier transform (DFFT).[14,15] Notably DFFT calculations can be readily accelerated using graphics card (GPU) calculations. Our Python/numpy-based code uses the pycuda module for GPU acceleration.[16] To sample different rotations, the process of grid generation and DFFT can be repeated.

Our ultimate application, identifying crystals that could be designed via surface mutations to the building block proteins warrants one notable change. Specifically, each protein has its side chains truncated to α and β carbons (proline residues also retain the γ carbon). The intent is to trim the most dynamic portions of the building block protein. Removing these groups will likely increase the difficulty of finding docking parameters that can successfully recapitulate the experimental crystal packing. However, parameters tuned to identify realistic packing arrangements without explicit surface side chains will be more useful for *de novo* protein crystal design.

1.3. Searching Alternative P1 Crystal Structures. The scoring map produced above provides a rapid lookup of the score for two monomers given a relative displacement on the grid. As is, this "base" map is suitable to score the interactions between exactly two protein monomers. To rapidly assess prospective crystals, we must account for the relative interactions between *all neighboring monomers* defined by the crystal packing arrangement.

The triclinic *P*1 crystal space group is fully defined by three basis vectors (nine degrees of freedom) that specify the unit cell (Figure 2). Our convention was to represent possible *P*1 crystal

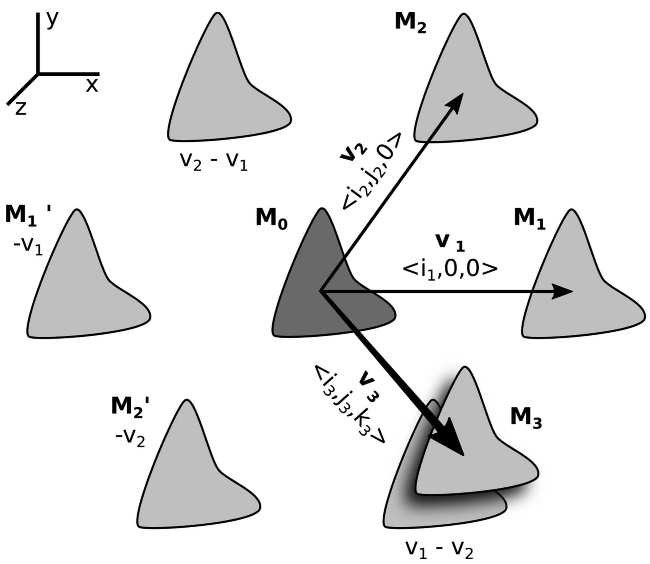


Figure 2. Notation for triclinic cell placements. A single copy of the protein building block (M_0) is placed at the origin with a specified orientation. Then, M_1 is placed along the *x*-axis. Next, M_2 is placed in the *x*−*y* plane. Finally, M_3 is placed in all three dimensions.

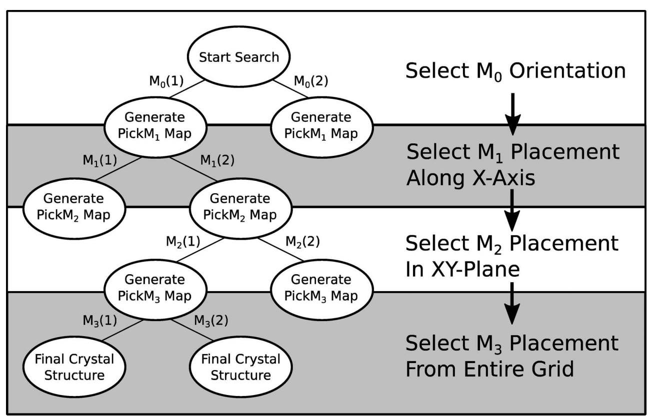


Figure 3. *P*1 crystal scoring decision tree.

forms as a decision tree (Figure 3). The first decision is the orientation of the initial monomer M_0 (three degrees of freedom). For M_0 rotations we use a precalculated uniform sampling of the 3D rotation group SO(3), calculated by Yershova.[17] As the translation of M_0 with respect to the Cartesian origin is arbitrary, M_0 is placed at the origin (Figure 4A). The second decision is where to place monomer M_1 along the *x*-axis (one degree of freedom) (Figure 4B,C). Third, M_2 is placed in the xy-plane (two degrees of freedom) (Figure 4D,E). Finally, M_3 is placed in xyz-space (three degrees of freedom) (Figure 4F,G). Together these choices yield a search tree for *P*1

DOI: 10.1021/acs.cgd.6b00769
Cryst. Growth Des. 2016, 16, 5579−5583

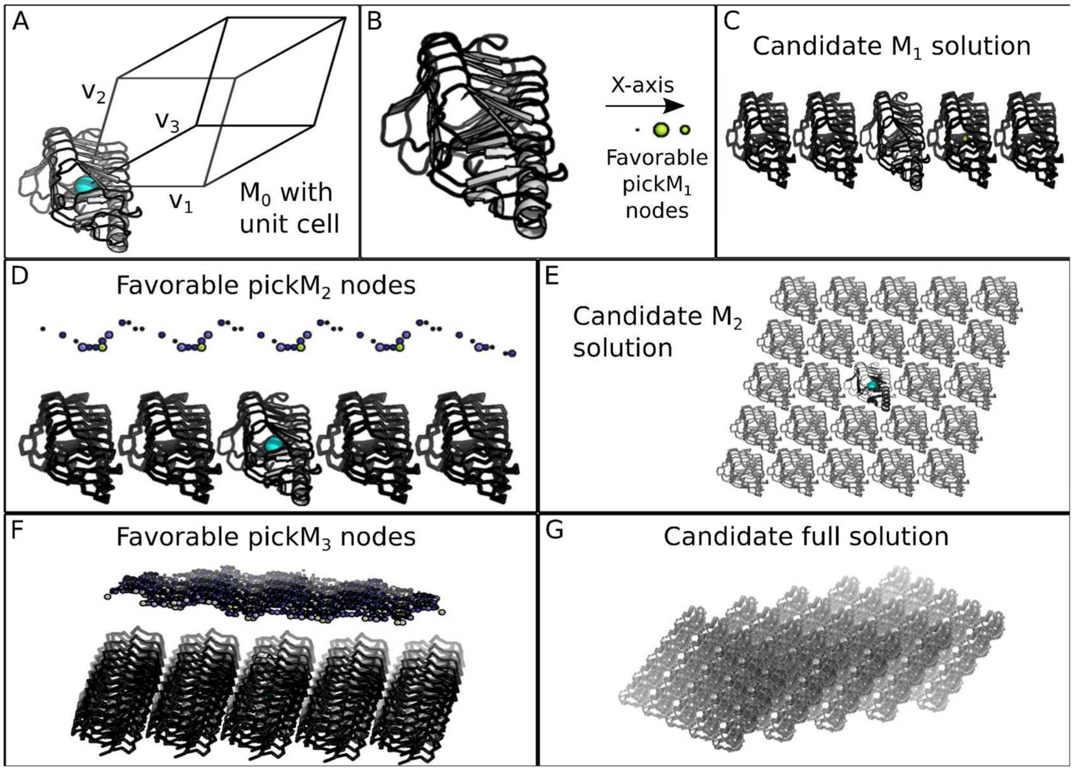


Figure 4. *P*1 crystal construction decision tree (A). Place M_0 with a specific orientation and calculate the base scoring array (B). Calculate top scoring M_1 grid displacement positions. (C). Place M_1 at a trial position along the *x*-axis and add induced symmetry copies. (D). Given the current M_0 and M_1 positions calculate the top scoring M_2 grid displacement positions. (E). Place M_2 at a trial position in the *xy*-plane and add induced symmetry copies. (F). Given the current collection of monomers induced by the previous M_1 ($\mathbf{v_1}$) and M_2 ($\mathbf{v_2}$) decisions, calculate the top scoring M_3 grid displacement positions. (G). Sample top scoring M_3 placements from the pickM3 array, each of which fully defines the *P*1 crystal. As necessary, additional branches of the decision tree can be sampled to consider alternative crystal packing arrangements.

crystals (Figure 3) with a maximal docking complementarity score. All nine degrees of freedom for a *P*1 crystal are retained.

This selection process defines the unit cell for the crystal, and all remaining symmetry copies in the crystal can be obtained as linear combinations of the translation vectors ($\mathbf{v_1}$, $\mathbf{v_2}$, $\mathbf{v_3}$) used for M_1, M_2, and M_3. At this point, the three angles (α, β, and γ) for the triclinic parallelepiped are also defined. If the parallelepiped height is small compared to the size of the protein building block, it is possible that M_0 could interact with symmetry copies that originate in unit cells that are not adjacent to the central unit cell enclosing M_0. In other words, since the basis vectors are not orthogonal there could be symmetry copies that have a small Cartesian distance to M_0 but a large taxicab distance. For example, six translation steps are required to reach $3\,\mathbf{v_1} - 3\,\mathbf{v_3}$, but this position could be close to the origin. It is difficult to devise an efficient crystal search algorithm that includes all possible linear combinations. Instead, we adopt a simple strategy as necessary to ensure that all Cartesian neighbors of M_0 are housed within neighboring unit cells. Specifically, we apply a distance cutoff (e.g., 80 Å) when selecting possible $\mathbf{v_1}$, $\mathbf{v_2}$, and $\mathbf{v_3}$ to eliminate crystal forms in which M_1, M_2, or M_3 is overly distant from M_0.

2. METHODS

2.1. Crystal Packing Scoring. The crystal packing calculation starts with a discrete grid mapping of the input protein coordinates. Conversion to a grid representation relies on several parameters: grid (grid resolution in Angstroms, here 2 Å), *N* (number of nodes along one axis, here 128), *r* (radius in Å from each atom to be counted as protein interior), *t* (surface layer thickness in Å), ρ (core grid point value), and δ (surface grid point value). Parameters *r*, *t*, ρ, and δ were allowed to differ between the receptor and the ligand monomers when creating the map. For each varied parameter combination, a DFFT calculation was performed to produce a map that contains interaction correlation score versus relative ligand monomer placement coordinates in grid space. Using this base correlation map, we can evaluate candidate crystal packing configurations.

To score a crystal configuration, we must calculate the score for M_0 interacting with each contacting neighbor. This score will be 2-fold larger than the lattice energy of the crystal. For each successive monomer, a composite map for the placement energy can be created using a linear combination of transformed copies of the base correlation map. For the *P*1 space group, these transformations are limited to translational shifts and reflections (annotated code is presented in the Supporting Information). For example, the score for placing M_1 in the *P*1 crystal lattice is the sum of placing a monomer at $\mathbf{v_1}$ and at $-\mathbf{v_1}$. We therefore prepare a "PickM_1" array that includes the consequences of placing both monomers. The score for M_2 placements is similarly calculated with a PickM_2 array that places monomers at $\mathbf{v_2}$, $-\mathbf{v_2}$, $\mathbf{v_1} + \mathbf{v_2}$, $\mathbf{v_1} - \mathbf{v_2}$, $-\mathbf{v_1} + \mathbf{v_2}$, and $-\mathbf{v_1} - \mathbf{v_2}$. When building the PickM2 array, M_1 has already been placed earlier in the decision tree, resulting in $\mathbf{v_1}$ being constant for that layer of the decision tree. The last decision, M_3 placement, is modeled as the addition of 18 new monomers. Once these maps have been created the energy implications for placing a monomer at a certain grid position can simply be read off the grid element corresponding to that position. The total energy for the crystal is therefore the sum of placing monomers according to the three maps, PickM_1, PickM_2, and PickM_3, given the full crystal definition $\mathbf{v_1}$, $\mathbf{v_2}$, $\mathbf{v_3}$.

The description above has been simplified for clarity. Real crystals extend beyond a $3 \times 3 \times 3$ unit cell block; the placement of monomer M1 also induces symmetry copies at $2\,\mathbf{v_1}$, $-2\,\mathbf{v_1}$, $3\,\mathbf{v_1}$, $-3\,\mathbf{v_1}$, and so on. So long as the unit cells are not thin with respect to the building block

DOI: 10.1021/acs.cgd.6b00769
Cryst. Growth Des. 2016, 16, 5579−5583

Table 1. Test Protein Performance for Top Parameter Sets

ligand parameters (LP) and receptor parameters (RP)					# decoy crystals ranked above authentic crystal							net parameter set performance	
LP r [Å]	RP δ	RP ρ	RP t [Å]	RP r [Å]	1pwl	1t41	1vbw	2gzr	3q8j	4kdw	4qvr	median rank	mean rank
2.4	0.5	−200	1	2.4	0	0	0	0	0	4	0	0	0.57
2.4	1	−200	1	2.4	0	0	0	0	0	4	0	0	0.57
2.4	1.5	−200	1	2.4	0	0	0	0	0	4	0	0	0.57
2.4	2	−200	1	2.4	0	0	0	0	0	4	0	0	0.57
2.4	4	−200	1	2.4	0	0	0	0	0	6	0	0	0.86

protein, we have found that it is feasible to truncate this series, thereby limiting the modeled extent of the crystal to the unit cells that are (taxicab distance) neighbors of the central unit cell. For the illustration in Figure 4, the ±1 terms and ±2 terms were included, so the final model of the crystal includes 125 unit cells (5^3).

2.2. Parameter Set Scoring Calculations. For each of the 20 dense protein crystals in the training set (PDB codes: 1fw9, 1itx, 1lzn, 1qng, 1rfs, 1sn7, 1tt8, 1v7s, 1xcj, 1zvj, 2f4a, 2f4g, 2vb1, 2vt1, 3lpa, 3lpc, 3lzt, 4itk, 4lzt, and 4nuh), we identified the nearest discrete grid representation to the authentic crystal structure ($\mathbf{v}_1^{WT}$, $\mathbf{v}_2^{WT}$, and $\mathbf{v}_3^{WT}$). Then, for each candidate parameter set we stored the scores for the local solution space around the discretized wild-type crystal solution. Specifically, the *P*1 wild-type grid solution was varied ± one grid node, (2 Å), along the *x*, *y*, and *z*-axes for each of the three basis vectors. The perturbations result in numerous (3^9 = 19683) decoy solutions per parameter set per protein. Decoy crystalline packing arrangements represent relatively subtle variations of the authentic packing arrangement found in the experimental crystal structure. Notably, M_0 was left in its original orientation, so that recapitulating the wild-type solution is reduced to giving $\mathbf{v}_1^{WT}$, $\mathbf{v}_2^{WT}$, $\mathbf{v}_3^{WT}$ the top rank. Scores for all 19 682 perturbed packing arrangements were recorded and compared to the score for $\mathbf{v}_1^{WT}$, $\mathbf{v}_2^{WT}$, $\mathbf{v}_3^{WT}$.

The parameter space that was searched varied (independently) the receptor and ligand r = [1.2, 1.5, 1.8, 2.1, 2.4], the receptor t = [1.0, 1.5, 2.0, 2.5, 3], the receptor ρ = [−200, −30, −15, −10, −5], and the receptor δ = [0.5, 1, 1.5, 2, 4]. This parameter space resulted in 5^5 parameter combinations. The total number of local search calculations was 5^5 parameter sets times 20 dense proteins resulting in 62 500 different combinations of protein and parameter set to test. These calculations required approximately 9.65 CPU-years, running on ∼115 cores in a computer cluster for about a month. The speed of the algorithm calculation was increased in exchange for increased memory load by caching some of the partial maps in memory instead of reapplying array transformations. Caching reduced the time spent in the inner loop of the decision tree (varying $\mathbf{v}_3$ with fixed $\mathbf{v}_1$ and $\mathbf{v}_2$) from 9.1 to 6.7 s.

3. RESULTS

3.1. Scoring Parameter Sets. The primary goal of this study was to identify the most favorable parameter set for converting protein structure coordinates into a grid representation. For a parameter set figure of merit, we determined the rank for the $\mathbf{v}_1^{WT}$, $\mathbf{v}_2^{WT}$, $\mathbf{v}_3^{WT}$ solution in comparison to the 19 682 other decoy solutions. Ideally, the wild-type grid solution would correspond to a local energy minimum, as it is derived from the native crystal structure. The score (lower is better) given to each parameter set for each protein was the number of decoy crystal candidates that ranked above the wild-type grid solution. We output the ranking scores for all parameter sets and sorted the parameter sets by the median number of scores for each of the 20 training set proteins.

We observed several trends in the top scoring sets (Table 1). Favored parameter sets had a larger r value and a more negative ρ. These parameter values correspond to a smoother protein surface that strongly rejects core overlap. To test the favored parameter combinations, the calculation was repeated on a different set of dense protein crystals (PDB codes: 1pwl, 1t41, 1vbw, 2gzr, 3q8j, 4kdw, and 4qvr) to see if the previous parameter combinations could still discriminate between the authentic crystal form grid solution and the perturbed decoys. The optimum parameters also performed very well for the test set of proteins.

4. DISCUSSION

4.1. Overview. We were somewhat surprised that several parameter sets successfully ranked the wild-type grid solution as the best possible solution among similar decoy crystals. As shown in Table 1, six out of seven test proteins were ranked perfectly. This is more striking given the use of truncated monomer models in the creation of the initial correlation map. The high reliability of the parameters when scoring local solution space suggests that the protein−protein interfaces found in dense *P*1 crystals are indeed in a local free energy minima; surface complementarity scores could not be improved through 2 Å perturbations of the protein crystal packing arrangement. It is currently unknown if the solution space remains smooth as it moves further from the wild-type crystal solution.

One illuminating test protein, 4kdw, did not score as well as the other test proteins. The packing arrangements that ranked more favorably were more tightly packed than $\mathbf{v}_1^{WT}$, $\mathbf{v}_2^{WT}$, $\mathbf{v}_3^{WT}$. Upon inspection of the native 4kdw crystal structure we noted numerous interfacial calcium ions. These calcium ions (heteroatoms in the PDB model) were not included when deriving the Katchalski-Katzir/Gabb grid representations. Therefore, we attribute the less compact wild-type packing arrangement to the presence of the interfacial calcium ions.

4.2. Outlook. We suggest that the top parameter sets (Table 1) identified in our thorough search of grid parameter space will be useful for recognition of authentic protein crystal packing arrangements. The ability to enumerate realistic packing arrangements will be useful for efforts toward *de novo* prediction of protein crystal packing or in the identification of feasible packing arrangements for *de novo* protein crystal design. For some applications it may prove helpful in the future to reoptimize the grid generation parameters for finer grid resolution (e.g., 1 Å grid spacing). To enable a fully unbiased search of possible crystal packing arrangements we are currently working to generalize the crystal-building algorithm to other chiral space groups.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.cgd.6b00769.

A Python script that evaluates the shape complementarity score for a 3 × 3 × 3 unit cell block, for 19 682

DOI: 10.1021/acs.cgd.6b00769
Cryst. Growth Des. 2016, 16, 5579−5583

perturbed crystal packing arrangements similar to the experimental solution (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: Christopher.Snow@colostate.edu.

Funding

ACS PRF Grant #52404-DNI10. NSF DMR Grant #1506219.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Acknowledgment is made to the Donors of the American Chemical Society Petroleum Research Fund, for support (or partial support) of this research. Grant: #52404-DNI10.

ABBREVIATIONS

PDB, Protein Data Bank; 3D, three dimensional; DFFT, discrete fast Fourier transform; M_0, initial monomer; SO(3), 3D rotation group

REFERENCES

(1) Singhal, D.; Curatolo, W. Drug Polymorphism and Dosage Form Design: A Practical Perspective. *Adv. Drug Delivery Rev.* **2004**, *56*, 335−347.

(2) Quiocho, F. A.; Richards, F. M. The Enzymic Behavior of Carboxypeptidase-A in the Solid State*. *Biochemistry* **1966**, *5*, 4062−4076.

(3) Cvetkovic, A.; Straathof, A. J. J.; Hanlon, D. N.; van der Zwaag, S.; Krishna, R.; van der Wielen, L. A. M. Quantifying Anisotropic Solute Transport in Protein Crystals Using 3-D Laser Scanning Confocal Microscopy Visualization. *Biotechnol. Bioeng.* **2004**, *86*, 389−398.

(4) Cvetkovic, A.; Picioreanu, C.; Straathof, A. J. J.; Krishna, R.; van der Wielen, L. A. M. Relation between Pore Sizes of Protein Crystals and Anisotropic Solute Diffusivities. *J. Am. Chem. Soc.* **2005**, *127*, 875−879.

(5) Lanci, C. J.; MacDermaid, C. M.; Kang, S.; Acharya, R.; North, B.; Yang, X.; Qiu, X. J.; DeGrado, W. F.; Saven, J. G. Computational Design of a Protein Crystal. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 7304−7309.

(6) King, N. P.; Sheffler, W.; Sawaya, M. R.; Vollmar, B. S.; Sumida, J. P.; André, I.; Gonen, T.; Yeates, T. O.; Baker, D. Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science* **2012**, *336*, 1171−1174.

(7) King, N. P.; Lai, Y.-T. Practical Approaches to Designing Novel Protein Assemblies. *Curr. Opin. Struct. Biol.* **2013**, *23*, 632−638.

(8) Derewenda, Z. S.; Vekilov, P. G. Entropy and Surface Engineering in Protein Crystallization. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 116−124.

(9) Goldschmidt, L.; Eisenberg, D.; Derewenda, Z. S. Salvage or Recovery of Failed Targets by Mutagenesis to Reduce Surface Entropy. *Methods Mol. Biol.* **2014**, *1140*, 201−209.

(10) Kantardjieff, K. A.; Rupp, B. Matthews Coefficient Probabilities: Improved Estimates for Unit Cell Contents of Proteins, DNA, and Protein-Nucleic Acid Complex Crystals. *Protein Sci.* **2003**, *12*, 1865−1871.

(11) Heras, B.; Edeling, M. A.; Byriel, K. A.; Jones, A.; Raina, S.; Martin, J. L. Dehydration Converts DsbG Crystal Diffraction from Low to High Resolution. *Structure* **2003**, *11*, 139−145.

(12) Kuo, A.; Bowler, M. W.; Zimmer, J.; Antcliff, J. F.; Doyle, D. A. Increasing the Diffraction Limit and Internal Order of a Membrane Protein Crystal by Dehydration. *J. Struct. Biol.* **2003**, *141*, 97−102.

(13) Heras, B.; Martin, J. L. Post-Crystallization Treatments for Improving Diffraction Quality of Protein Crystals. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2005**, *61*, 1173−1180.

(14) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Molecular Surface Recognition: Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 2195−2199.

(15) Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. E. Modelling Protein Docking Using Shape Complementarity, Electrostatics and Biochemical Information. *J. Mol. Biol.* **1997**, *272*, 106−120.

(16) Klöckner, A.; Pinto, N.; Lee, Y.; Catanzaro, B.; Ivanov, P.; Fasih, A. PyCUDA and PyOpenCL: A Scripting-Based Approach to GPU Run-Time Code Generation. *Parallel Computing* **2012**, *38*, 157−174.

(17) Yershova, A.; Jain, S.; LaValle, S. M.; Mitchell, J. C. Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration. *Int. J. Robotics Res.* **2009**, *29*, 801.

DOI: 10.1021/acs.cgd.6b00769
Cryst. Growth Des. 2016, 16, 5579−5583