



# A simple technique to classify diffraction data from dynamic proteins according to individual polymorphs

Thu Nguyen,<sup>a</sup> Kim L. Phan,<sup>b</sup> Dima Kozakov,<sup>c</sup> Sandra B. Gabelli,<sup>b</sup> Dale F. Kreitler,<sup>d</sup> Lawrence C. Andrews,<sup>e</sup> Jean Jakoncic,<sup>d</sup> Robert M. Sweet,<sup>d</sup> Alexei S. Soares<sup>d\*</sup> and Herbert J. Bernstein<sup>e\*</sup>

Received 10 May 2021

Accepted 17 December 2021

Edited by K. Diederichs, University of Konstanz, Germany

**Keywords:** chymotrypsinogen; clustering; polymorphs; protein dynamics; unit-cell changes.

**PDB references:** chymotrypsinogen, 7kty; 7ktz; 7ku0; 7ku1; 7ku2; 7ku3

**Supporting information:** this article has supporting information at journals.iucr.org/d

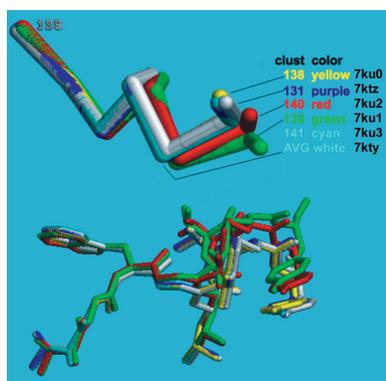
<sup>a</sup>Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-2424, USA, <sup>b</sup>Department of Medicine, Oncology, Biophysics and Biophysical Chemistry, Johns Hopkins University, 725 North Wolfe Street, Baltimore, MD 21205, USA, <sup>c</sup>Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA, <sup>d</sup>National Synchrotron Light Source II, Building 745, Brookhaven National Laboratory, PO Box 5000, Upton, NY 11973-5000, USA, and <sup>e</sup>Ronin Institute for Independent Scholarship, c/o NSLS-II, Building 745, Brookhaven National Laboratory, PO Box 5000, Upton, NY 11973-5000, USA. \*Correspondence e-mail: soares@bnl.gov, yayahjb@gmail.com

One often observes small but measurable differences in the diffraction data measured from different crystals of a single protein. These differences might reflect structural differences in the protein and may reveal the natural dynamism of the molecule in solution. Partitioning these mixed-state data into single-state clusters is a critical step that could extract information about the dynamic behavior of proteins from hundreds or thousands of single-crystal data sets. Mixed-state data can be obtained deliberately (through intentional perturbation) or inadvertently (while attempting to measure highly redundant single-crystal data). To the extent that different states adopt different molecular structures, one expects to observe differences in the crystals; each of the polystates will create a polymorph of the crystals. After mixed-state diffraction data have been measured, deliberately or inadvertently, the challenge is to sort the data into clusters that may represent relevant biological polystates. Here, this problem is addressed using a simple multi-factor clustering approach that classifies each data set using independent observables, thereby assigning each data set to the correct location in conformational space. This procedure is illustrated using two independent observables, unit-cell parameters and intensities, to cluster mixed-state data from chymotrypsinogen (ChTg) crystals. It is observed that the data populate an arc of the reaction trajectory as ChTg is converted into chymotrypsin.

## 1. Introduction

Proteins often undergo structural changes as part of their normal functioning. Crystal structures often reveal proteins in different conformations (called polymorphs). Crystallography explores an average structure of all of the molecules in the volume that the X-ray beam interrogates: an immense number of individual molecules with possibly significantly different polymorphs. These different structures might have revealed information on the dynamics of transitions among those polymorphs had they not all been averaged together. Because of this averaging, instead of seeing distinct states clearly, we may see only what looks like blurred thermal motion.

To reduce this problem, a typical structural study of a protein might involve constraining the molecule in one of its states by binding a ligand before crystal growth. For example, in the emerging field of biological data storage, proteins with



OPEN ACCESS

two distinct conformations (called polystates) are intentionally switched between them to represent binary code (0 and 1; Sethi, 2015). It seems likely that molecules in a single crystal form may show slightly different structures depending on the pH, the state of hydration *etc.* That is, they exhibit dynamic behavior that may or may not indicate changes related to their function. Here, we will adopt the term ‘polystates’ to refer to protein structural polymorphs that correspond to biologically relevant conformations of proteins, as well as to other significant state variations. To observe them, one might sample these variables carefully to find states where all molecules are the same in a crystal. We aim in this work to design a more general workflow.

Modern crystallographic practice provides opportunities to discover and analyze the sort of changes that we describe here: we use crystals small enough that each may contain only one polystate. In particular, data collection at a synchrotron source often includes the measurement of many partial sets of crystal diffraction data from many, often very small, crystals (Liu *et al.*, 2011; Giordano *et al.*, 2012; Rossmann, 2014; Assmann *et al.*, 2016; Bernstein *et al.*, 2017, 2020; Gao *et al.*, 2018). This is possible because fourth-generation synchrotron sources are very bright, the X-ray beams are very small, the 2D detectors employed are very fast and modern goniometers are very precise. With a detector operating at 200 Hz, a 360° rotational sweep with 0.2° per image will take approximately 9 s. These are standard experimental parameters at the FMX and AMX beamlines at National Synchrotron Light Source II (NSLS-II), where a beam size even smaller than the small crystals is employed. We could sample hundreds of tiny crystals (1–5 µm), each perhaps one polymorph, and then separate them into these polymorphs.

For this sort of treatment, one may mount crystals singly or, say, with half a dozen in each sample mount (loop or mesh micro-mount); in each case, data are obtained for each individual crystal. The crystals may come from different crystallization drops, or even different preparations, but they are all nominally isomorphous; our objective is to improve the quality of the data by merging multiple measurements. In practice one may partition the data from many crystals into different clusters, based on differences in intensities or unit-cell parameters. Each cluster may represent one step in the normal dynamic motion of the molecule. Starting from multiple independent samples increases the chances of having multiple states to observe.

As we mentioned, two popular criteria for clustering data sets are the similarity in unit-cell parameter values and in reflection intensities. Unit-cell databases have long been used for substance identification and are now used as a coarse screen for molecular-replacement candidates. Steno (1669), as cited in Authier (2013), noticed the constancy of interfacial angles of crystals. Reflection intensities represent the true structure, so similarity of reflections is a good metric to use in comparing data sets.

We demonstrate our approach using chymotrypsinogen (ChTg). ChTg is the precursor (zymogen) of the mammalian digestive enzyme chymotrypsin (CHT), one of several well

known serine proteases (Kunitz & Northrop, 1935; Siekevitz & Palade, 1960). This conversion is accomplished by several enzymatic cleavages. Firstly (in the digestive tract), trypsin cleaves the peptide bond between Arg15 and Ile16 to yield  $\pi$ -CHT, which is an active enzyme form. Secondly,  $\pi$ -CHT molecules autolyze one another to cleave the bonds between Leu13 and Ser14 to release Ser14-Arg15, and between Tyr146 and Thr147 and between Asn148 and Ala149 to release Thr147-Asn148. The resulting  $\alpha$ -chymotrypsin is formed by three chains tethered by disulfide bridges.

We measured 146 diffraction data sets, each from a single crystal of ChTg crystallized in three different conditions: pH 4.6, pH 5.6 and pH 6.5. We discovered that the polymorphs that we observed resemble the partial conversion of ChTg to CHT. Almost all of the crystals that we successfully assembled into clusters, from which we could average data and solve the structure, were formed at pH 6.5. To detect different molecular states in individual crystals, we based the partitioning of the single-crystal data on the following two properties: unit-cell parameters and intensities.

We first employed differences in unit-cell parameters as a conventional method for clustering single-crystal data. However, this was clearly inadequate, and we extended this to partitioning on similarities in diffraction intensities. The criterion for similarity was the correlation coefficient calculated between pairs of measurements, thereby classifying according to the slightly different but stable conformational states that generated these data. We found that clustering by correlation of intensities also revealed the large unit-cell differences that were observed. In general, however, unit-cell differences are observable much earlier during structure determination than distinctive intensity differences, and they can provide preliminary clustering with the use of correlation coefficients to follow.

## 2. Methods

### 2.1. Crystallization

We determined crystallization conditions for chymotrypsinogen (Sigma) using the commercial Crystal Screen HT (Hampton Research) set up with a Mosquito robot (STP Labtech). Crystals were grown via hanging-drop vapor diffusion at 18°C from condition F11. To optimize the crystallization conditions further, we set up a 24-well tray using hanging-drop vapor diffusion with a fixed pH of 6.5, varying the concentration of both dioxane (10% or 15%) and ammonium sulfate (1.0–2.0 M). Each drop, consisting of 1 µl reservoir solution (1.0–2.0 M ammonium sulfate, 0.1 M MES pH 6.5 and 10% or 15% dioxane) and 1 µl 10.0 mg ml<sup>-1</sup> enzyme, was equilibrated over 0.5 ml reservoir solution. The other two crystallization conditions consisted of either 0.2 M ammonium acetate, 0.1 M sodium acetate trihydrate pH 4.6, 30% (w/v) PEG 4000 or 0.5 M ammonium sulfate, 0.1 M sodium citrate tribasic dihydrate pH 5.6, with 1.0 M lithium sulfate monohydrate in the reservoir. All crystals were

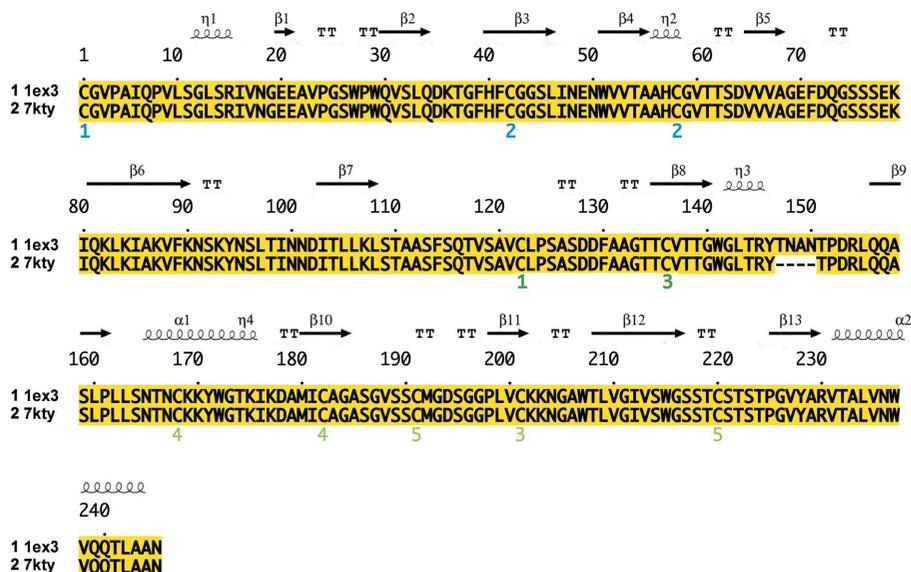
cryocooled in 3.5 M lithium sulfate. A typical drop containing these crystals is shown in Fig. 1.

### 2.2. Data collection and structure-solution strategies

We used ChTg crystals of ~50 μm in size and an ~5 μm beam to obtain several hundred data sets at an energy of 13.48 keV (0.92 Å), collecting 120° of data per crystal on beamline 17-ID-1 (AMX) at NSLS-II, Brookhaven National Laboratory using a Dectris EIGER X 9M detector. For each crystal with sufficient data, the data set was indexed, integrated and scaled using a version of the data-reduction pipeline *fast\_dp* (Winter & McAuley, 2011) modified to run in the local distributed computing environment, which supplies the following modules: *XDS* (Kabsch, 2010), *DIALS* (Winter *et al.*, 2018), *Phenix* (Afonine *et al.*, 2012), *AIMLESS* and *POINTLESS* (Evans & Murshudov, 2013). The tetragonal



**Figure 1**  
Representative ChTg crystals from crystallization conditions consisting of 1.0–2.0 M ammonium sulfate, 0.1 M MES pH 6.5, 10% or 15% dioxane.



**Figure 2**  
Sequence alignment of ChTg (PDB entry 1ex3) and the average structure (PDB entry 7kty). The loop residues 147–150 do not display electron density in PDB entry 7kty.

crystals of ChTg diffracted to between 2.0 and 2.4 Å resolution.

We determined the structure by molecular replacement with *Phaser* (McCoy *et al.*, 2007) using the structure of ChTg with PDB code 1ex3 as a model (Bernstein *et al.*, 1977; Berman *et al.*, 2000; Pjura *et al.*, 2000). The data were refined to their final resolution using iterative rounds of refinement with *REFMAC* (Murshudov *et al.*, 1997, 2011) and manual rebuilding in *Coot* (Emsley *et al.*, 2010). We then used this model to build our ‘average’ structure from an average of all data sets at an energy of 13.48 keV. Our average structure was broadly similar to previously published structures, with the exception of one loop that was not resolvable in our data (Thr147–Asn150; see Fig. 2) and one adjacent region that adopts a different conformation (Thr139–Tyr146; see Fig. 3). While scaling these data sets, we noted clusters in the data. It is possible that they originated from polymorphs in the ChTg crystals, which may represent dynamic behavior in the molecules.

### 2.3. Data-clustering program

We used a custom-modified version of the *KAMO* clustering pipeline (Yamashita *et al.*, 2018), which uses the clustering program *Blend* (Foadi *et al.*, 2013) to generate a dendrogram of the data sets. We expanded *KAMO* and *Blend* to allow two-factor clustering as follows. Unit-cell parameters and amplitudes contain independent information. One expects differences in unit-cell parameters to reflect changes in the outer shape of the structure, perhaps responding to the presence of internal or external ligands. On the other hand, differences in amplitudes will be sensitive to all conformational changes in the protein. Therefore, in a single workflow the new scheme obtains initial ‘coarse’ clusters according to the similarity of the crystallographic unit cells (space-group clusters) and then generates ‘fine’ clusters by a further partitioning of each cluster according to the similarity of the amplitude data. (The modified software is available at <https://github.com/nsls-ii-mx/blend> and <http://github.com/nsls-ii-mx/yamtbx>.)

Our procedure employs Pearson correlation coefficient (CC) calculations to determine similarity scores, requiring that pairs of data sets have many measured amplitudes in common: one must have a reasonably complete set of structure factors. For this CC clustering, 70% completeness is required. One can introduce a penalty for unmatched structure factors, and can obtain a solution with a completeness as low as 20–40% (Bernstein *et al.*, 2017); we are studying the effect of even lower completeness to apply this procedure to partial data sets. We will show that this

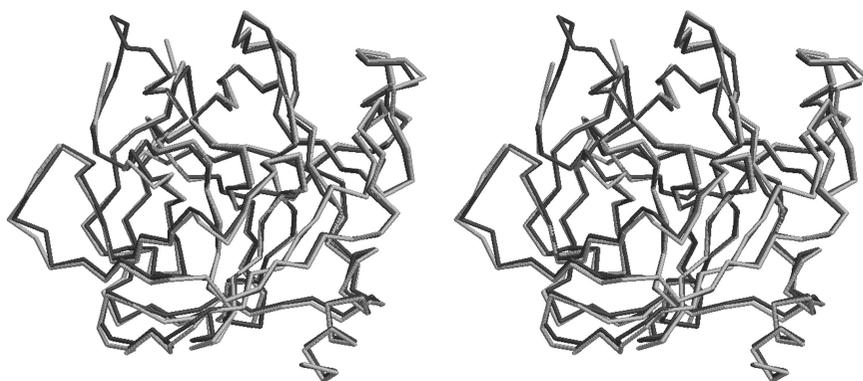
clustering approach demonstrates how increasingly sensitive clustering methods can identify increasingly detailed structural differences (Figs. 3–8).

In our study, when we ran *KAMO.multi\_merge* with *Blend*, it reduced the data in different point groups (the observed point groups were exemplified by space groups *P2*, *P222* and *P422*). We selected the data that reduced in *P422* because these data contained the largest number of data sets and correspond to the published space group for ChTg; we aimed to use the structure factors to display the differences in the structures corresponding to the data that reduced in *P422*. We used *KAMO* to divide the data sets of the chosen space group into different clusters based on intensity CC. Please note that the term *space-group clustering* is in common usage, but the technically correct term for clustering performed prior to refinement is *point-group clustering*; for example, our algorithms clustered the cymotrypsinogen data using the exemplar space group *P422* (No. 89) with its Laue point group (*4/m 2/m*

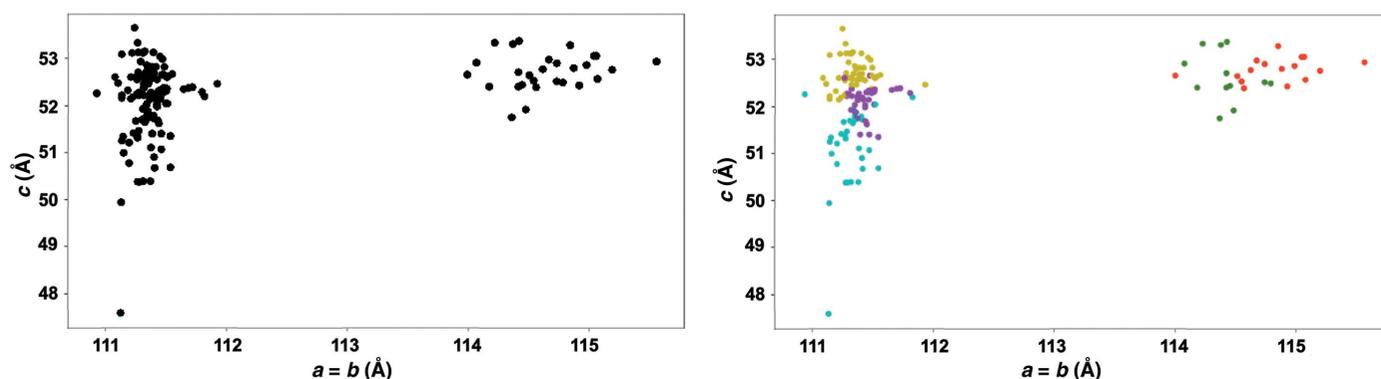
*2/m*), rather than the actual space group *P4<sub>1</sub>2<sub>1</sub>2* (No. 92) adopted by all crystals.

For the clustering step, we defined the ‘distance’ between pairs of data sets as  $d(i, j) = [1 - \text{CC}(i, j)]^{1/2}$ . Our procedure then used a hierarchical clustering analysis (Rokach & Maimon, 2005) with Ward clustering (Ward, 1963) to find distinct groups of the chosen data sets. In Ward clustering, the data sets are considered first by building a small cluster out of the two closest data sets and then adding one data set at a time to whichever data set or existing cluster results in a new cluster of smallest variance (mean of squares of distances). There are many other choices of what is called ‘linkage’ in forming a cluster dendrogram, such as using cluster centroids. Using the minimal variance allows the use of one simple distance matrix as input to the clustering algorithm, rather than requiring repeated calculation of distances among cells or, worse, among *hkl* vectors of structure factors, but it does tend to produce dendrograms for which the heights grow rapidly. Strauss & von Maltitz (2017) discuss some alternative linkage choices. The program outputs a dendrogram that illustrates the distances (differences) among clusters by the y axis (the joint variance). To obtain a certain number of clusters which contain more similar data sets, we chose a height-cutoff value *k* accordingly. The lower the *k* value, the more similar the data sets in each cluster are. Each cluster now relates to a structure built after merging data sets within it.

We want to understand how the unit-cell parameter values of data sets relate to the clusters determined by similarity in diffraction intensities. Since the space group is *P4<sub>1</sub>2<sub>1</sub>2* (No. 89/92), the unit-cell parameters *a* and *b* are equal and all



**Figure 3** Cross-eyed stereoview of the structural alignment of PDB entry 1ex3 (dark gray) and the average structure PDB entry 7kty (light gray). The *FATCAT* chain r.m.s.d. is 0.56 Å. The regions with significant differences are adjacent and appear at the upper left of this figure. Firstly, in the average structure the amino acids between Thr147 and Asn150 are missing. Secondly, in the average structure the amino acids between Thr139 and Tyr146 adopt a significantly different conformation.



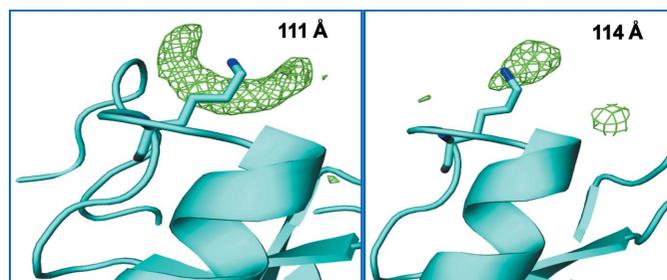
**Figure 4** Two main data clusters can be identified by inspection (the  $a = 111$  Å group and the  $a = 114$ – $115$  Å group). We observed that our data partitioned cleanly between 28 data sets with an  $a (= b)$  unit-cell parameter of approximately 114–115 Å and 118 data sets with an  $a (= b)$  unit-cell parameter of approximately 111 Å. The separation into the two unit-cell clusters is shown in the monochrome clustering on the left. The further division of these two clusters into amplitude-based clusters is shown by the colors on the right. The  $a = 114$ – $115$  Å unit-cell-based cluster contains the green and red clusters and the  $a = 111$  Å unit-cell-based cluster contains the cyan, purple and yellow clusters. Each of our data sets was sufficiently large that amplitude-based clustering could have been used from the start. However, many serial crystallography projects consist of narrow wedges of data, each of which might be too small to cluster effectively using amplitudes because amplitude-based clustering requires that data sets have a sufficient number of observations in common. This figure illustrates how an initial use of cell-based clustering might be used to bootstrap amplitude-based clustering.

unit-cell angles are  $90^\circ$ . Since there are only two free parameters, we could visually demonstrate how the intensity clusters relate to the unit-cell parameters  $a$  and  $c$  for each data set (Fig. 4).

Finally, we created a molecular structure from the average intensities from all of the crystals in each intensity cluster and also averaged structures for each of the different unit-cell parameter clusters (see Fig. 5; the data are available from the corresponding authors). To create structures that relate to each of these clusters, we employed the average structure defined above in the data-collection section as the starting model for structure determination and refinement of the structures of each of the clusters. All of the processes that we used to build the structures of the clusters and to subsequently refine them are automated with the help of *REFMAC* (Murshudov *et al.*, 1997, 2011). Following the automated refinement steps, we performed a manual check-and-refine step using *Coot* (Emsley *et al.*, 2010) to ensure that no serious errors remained from the automated process and corrected the refined model as needed. *FATCAT* (Ye & Godzik, 2004) allowed us to quantify the morphological differences among structure solutions.

#### 2.4. Illustrating the differences to identify physically meaningful clusters

Any software that uses observable parameters to generate clusters may generate a very large number of clusters. How is one to determine which clusters are physically meaningful? Dendrograms can illustrate the relationships among clusters, but one must illustrate physical relevance using structural tools, *i.e.* comparing the structures obtained from each of these clusters. We generated two software tools for this purpose (see <https://github.com/nsls-ii-mx/chymotrypsinogen>). Both tools use individual colors to differentiate among clusters, which we can then test for physical relevance, and both tools use two- or three-dimensional plots to illustrate an underlying physical characteristic of the structure.



**Figure 5**  
Differences in solvent between the  $a = 111 \text{ \AA}$  cluster and the  $a = 114\text{--}115 \text{ \AA}$  cluster.  $F_o - F_c$  electron difference density displaying two differences that we observed in solvent density between the  $a = 111 \text{ \AA}$  cluster and the  $a = 114\text{--}115 \text{ \AA}$  cluster (difference densities at  $2\sigma$  are shown in green for both data sets). Left: ribbon diagram of ChTg around Lys175 (cyan) for the  $a = 111 \text{ \AA}$  cluster. Right: ribbon diagram of ChTg around Lys175 (cyan) for the  $a = 114\text{--}115 \text{ \AA}$  cluster. This density was modeled as a water molecule.

We developed a tool to create color-coded coordinate ellipsoids. We plotted the  $xyz$  coordinates of the  $C^\alpha$  atom of a particular amino acid in the structure that we observed to be highly mobile among the clusters. We created color-coded ellipsoids that enclosed all  $C^\alpha$  atoms found from each of the individual clusters. The size of each ellipsoid indicates the variation of the coordinates within the corresponding cluster. Ideally, the size of each color-coded ellipsoid will not be very large compared with the separations among the centroids of the ellipsoids, indicating that each cluster represents a separable state. The code is available in the <https://github.com/nsls-ii-mx/chymotrypsinogen> git repository as the file <https://raw.githubusercontent.com/nsls-ii-mx/chymotrypsinogen/master/ellipsoid.py>. An example of the use of this graphic appears in Fig. 8.

We also plotted the  $a$  and  $c$  axis lengths for each data set that resides within an amplitude-based cluster (Fig. 4). Employing a dendrogram-plotting graphic tool from *KAMO*, we illustrated all data that originated from each postulated cluster in a different color (Fig. 6).

To detect subtle differences among the structures of the clusters, we used *FTMap* (Kozakov *et al.*, 2015), software that was designed to determine and characterize ligand-binding hotspots on the surfaces of proteins. The algorithm uses a library of 16 molecules as probes to discover potential patches on the surface of a structure where a molecule might bind. Differences in proposed surface binding could reveal otherwise unnoticeable physical differences among the structures.



**Figure 6**  
Amplitude-based clusters generated using *KAMO* (dendrogram). This dendrogram shows a representation of the similarity of pairs of data sets and of clusters of more data sets. They are arranged with the most similar clusters near each other and the connecting bar at a height corresponding to the distance between clusters. The difference was calculated using Ward's method for hierarchical clustering, which yields a composite metric that contains information from amplitude differences and from unit-cell differences. Our algorithm is described in Section 2.3. Structures were solved corresponding to each of these 145 clusters. We deposited the structure derived from refinement against structure factors, each of which was an average of that observation from all of these clusters, as PDB entry 7kyt. Through inspection of the derived structures we selected a height within the dendrogram at which to partition our data, giving five clusters. We then averaged all structure factors within each of the five distinct clusters and then refined against these to give cluster-average structures. We deposited the averaged structure from the green clusters as PDB entry 7ku1, from the red clusters as PDB entry 7ku2, from the cyan clusters as PDB entry 7ku3, from the purple clusters as PDB entry 7ktz and from the yellow clusters as PDB entry 7ku0.

**Table 1**  
Data collection and processing.

| PDB code                                 | 7kty                       | 7ku1                       | 7ku2                       | 7ku3                       | 7ktz                       | 7ku0                       |
|--|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| Description                              | Average                    | Green                      | Red                        | Cyan                       | Purple                     | Yellow                     |
| Cluster No.                              | 145                        | 139                        | 140                        | 141                        | 131                        | 138                        |
| No. of data sets                         | 146                        | 12                         | 16                         | 32                         | 37                         | 49                         |
| Wavelength (Å)                           | 0.9201                     | 0.9201                     | 0.9201                     | 0.9201                     | 0.9201                     | 0.9201                     |
| Temperature (K)                          | 100                        | 100                        | 100                        | 100                        | 100                        | 100                        |
| Detector                                 | EIGER X 9M                 |
| Distance (mm)                            | 100–200                    | 100–200                    | 100–200                    | 100–200                    | 100–200                    | 100–200                    |
| Rotation (°)                             | 0.2                        | 0.2                        | 0.2                        | 0.2                        | 0.2                        | 0.2                        |
| Total range (°)                          | 120                        | 120                        | 120                        | 120                        | 120                        | 120                        |
| Space group                              | <i>P</i> <sub>4</sub> ,2,2 |
| <i>a</i> , <i>b</i> (Å)                  | 114.49                     | 114.49                     | 115.57                     | 111.33                     | 111.49                     | 111.47                     |
| <i>c</i> (Å)                             | 51.90                      | 51.9                       | 52.92                      | 51.87                      | 52.02                      | 52.36                      |
| $\alpha$ , $\beta$ , $\gamma$ (°)        | 90                         | 90                         | 90                         | 90                         | 90                         | 90                         |
| Resolution (Å)                           | 2.00                       | 2.39                       | 2.19                       | 2.00                       | 2.00                       | 2.02                       |
| No. of reflections                       | 23850                      | 14139                      | 18927                      | 22533                      | 22696                      | 22261                      |
| Completeness (%)                         | 99.94                      | 99.80                      | 98.98                      | 99.81                      | 99.93                      | 99.59                      |
| $\langle I/\sigma(I) \rangle$            | 10.91                      | 9.48                       | 9.76                       | 10.66                      | 12.15                      | 10.85                      |
| Wilson <i>B</i> factor (Å <sup>2</sup> ) | 53.96                      | 61.87                      | 53.31                      | 33.72                      | 29.15                      | 30.16                      |

### 3. Results

#### 3.1. Data collection and protein structures

We collected 511 complete data sets and processed 325 of them using our data-reduction pipeline *fast\_dp\_nsls2*. Of these, 175 files had a resolution better than 4 Å. Finally, 146 data sets from the point group represented by space group No. 89 (*P*<sub>4</sub>22) were merged using *Blend* cell-based cluster analysis. The protein is a single chain of 245 residues, of which four residues (147–150) were not resolved.

We obtained our initial structure, PDB entry 7kty, from a merge of all 146 data sets and called this the average structure (denoted thus in Tables 1 and 2). We used *REFMAC* and *Coot* to refine the structure and reduce the *R* value to about 18%.

Averaging all 146 data sets together resulted in a relatively high *R*<sub>merge</sub> value (48%), but nevertheless PDB entry 7kty was a good fit to these data (*R*<sub>work</sub> = 19%, *R*<sub>free</sub> = 20%). This average structure is slightly different from the published structure with PDB code 1ex3 which we used as an initial phasing model. For example, PDB entry 7kty has a missing loop from residue 147 to residue 150, which is a characteristic of mature  $\alpha$ -chymotrypsin. Fig. 2 displays the sequence alignment between PDB entry 1ex3 and our structure, PDB entry 7kty, with the elements of the secondary structure drawn on top.

#### 3.2. Clustering with unit cells and with amplitudes

Clustering software will generate data corresponding to candidate polystates, even in cases where truly distinct polystates are not actually present in the samples. Two independent data sets collected from two samples will always give different average structures. Such differences are often not relevant in terms of dynamics or states when the differences are small compared with the experimental error. The only way to determine whether candidate clusters may correspond to biologically relevant polystates is to generate and examine the corresponding structural models (typically atomic models) with appropriate real-space tools, such as *FATCAT* and *Coot*.

**Table 2**  
Structure solution and refinement.

| PDB code                                   | 7kty    | 7ku1  | 7ku2  | 7ku3  | 7ktz   | 7ku0   |
|--|---------|-------|-------|-------|--------|--------|
| Description                                | Average | Green | Red   | Cyan  | Purple | Yellow |
| Final <i>R</i> <sub>work</sub> (%)         | 19.13   | 22.08 | 20.97 | 18.15 | 16.18  | 16.74  |
| Final <i>R</i> <sub>free</sub> (%)         | 20.19   | 26.37 | 23.42 | 21.01 | 19.01  | 19.41  |
| No. of non-H atoms                         |         |       |       |       |        |        |
| Protein                                    | 1786    | 1771  | 1778  | 1794  | 1786   | 1786   |
| Ligand                                     | 15      | 5     | 5     | 15    | 15     | 15     |
| Water                                      | 99      | 10    | 44    | 195   | 249    | 243    |
| Total                                      | 1900    | 1786  | 1827  | 2004  | 2050   | 2044   |
| R.m.s. deviations                          |         |       |       |       |        |        |
| Bonds (Å)                                  | 0.01    | 0.01  | 0.01  | 0.01  | 0.01   | 0.01   |
| Angles (°)                                 | 0.86    | 1.00  | 0.86  | 0.77  | 0.80   | 0.74   |
| Average <i>B</i> factors (Å <sup>2</sup> ) |         |       |       |       |        |        |
| Protein                                    | 57.9    | 67.4  | 58.7  | 36.6  | 28.1   | 28.8   |
| Ligand                                     | 63.0    | 83.9  | 75.7  | 40.7  | 40.7   | 34.8   |
| Water                                      | 58.7    | 63.0  | 56.1  | 42.3  | 37.1   | 36.8   |
| Ramachandran plot (%)                      |         |       |       |       |        |        |
| Favored                                    | 97.47   | 96.20 | 97.06 | 97.47 | 98.31  | 98.73  |
| Allowed                                    | 1.69    | 2.95  | 2.52  | 2.11  | 1.27   | 0.84   |

In the case of the ChTg data, we could see from inspection that the data could be divided into two large clusters corresponding to structures with *a*  $\simeq$  111 Å and those with *a*  $\simeq$  114–115 Å (Fig. 4).

Employing only the observed diffraction intensities, we identified two main clusters that corresponded to the two main polymorphs that ChTg adopted in our crystals, based on the length of the *a* axis. In addition, there were five clusters that corresponded to biologically relevant polymorphs present in our data. The cell-based clustering shows that the unit-cell lengths separate clearly into two groups, while the *c* unit-cell length varies less and is not clearly separable. There were significant solvent-region differences between the *a* = 111 Å cluster and the *a* = 114–115 Å cluster (Fig. 4).

When comparing the structures corresponding to the *a* = 111 Å cluster and the *a* = 114–115 Å cluster, we observed that the *a* = 114–115 Å cluster data yields observable density for all 245 residues (similar to PDB entry 1ex3), while the *a* = 111 Å cluster data indicate that there is a missing loop from residue 147 to residue 150 (this region is also not

observed in the average structure). Another thing that we observed is the presence of strong density near Lys175 in the  $a = 111 \text{ \AA}$  cluster data, while the  $a = 114\text{--}115 \text{ \AA}$  cluster data does not have this large artifact (Fig. 5). We discuss these differences further below.

Using *Blend* and *KAMO*, we obtained 145 clusters from the 146 ChTg data sets. We then generated structures after merging data sets belonging to each of these 145 clusters, and we visually inspected each of them to find any recurring patterns. This visual inspection allowed us to determine that all of the reproducible differences could be accounted for using just five of the larger clusters (which we call the green, red, cyan, purple and yellow clusters). In other words, we chose the ‘height’ at which we cut the *KAMO* dendrogram so that five clusters contain the data corresponding to the relevant structures (Fig. 6). By comparing our results with identically processed data that were modified such that the unit-cell dimensions were either constant or randomized, we observed that the dominant contribution to the clustering correctness derives from differences in the data amplitudes, with differences in the unit-cell parameters playing a lesser role (see supporting information).

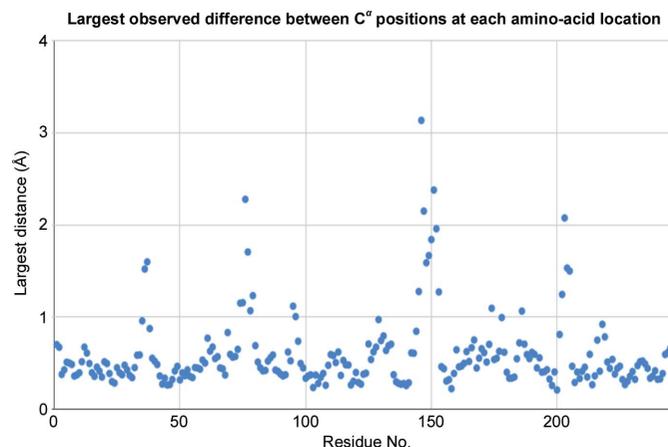
The 145 clusters could also be overlaid on the  $a$  and  $c$  axis diagram color-coded according to each of the five main clusters (Fig. 4). All of the data sets of the green and red clusters belong to the  $a = 114\text{--}115 \text{ \AA}$  cluster and the data sets of the cyan, purple and yellow clusters belong to the  $a = 111 \text{ \AA}$  cluster. If we increase the cut height to 1.5, we obtain sub-master clusters of the two intensities, one containing the green and red clusters and the other containing the cyan, purple and yellow clusters. This means the intensity cluster result has a strong alignment with the unit-cell parameter cluster result.

### 3.3. The five data clusters

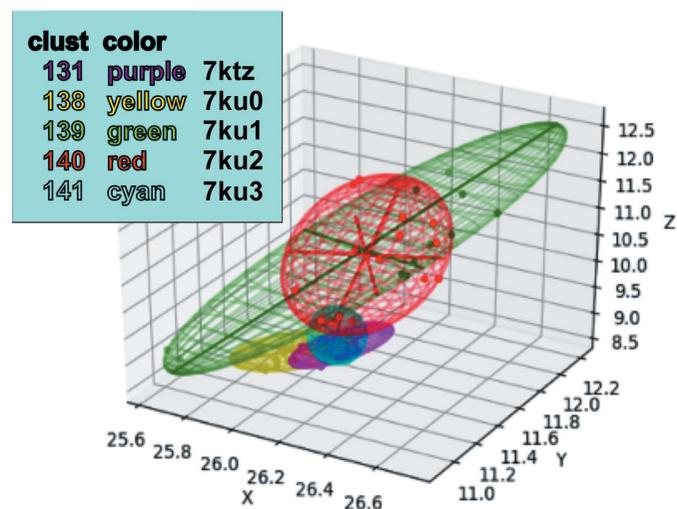
We generated a dendrogram using Ward’s method for hierarchical clustering with the height cutoff at 1.0 to obtain distinct groups of data sets. To observe the differences between the 145 structures generated using individual data clusters, we calculated the largest differences in physical coordinates at the  $C^\alpha$  atom of each residue (see Fig. 7). We observed that the most mobile area, particularly residues 139–145, is near the missing loop from residues 146–152. Note that the distinctive differences between ChTg (the zymogen) and CHT (the enzyme chymotrypsin) are the cleavages at the N-terminus and the gap between Tyr146 and Ala149. The largest differences were observed for residue 146, with average positional differences greater than  $3 \text{ \AA}$  (Fig. 8).

At each residue position, we plotted ellipsoids to illustrate the variation in the  $C^\alpha$  coordinates observed in each of the structures corresponding to the five clusters. For example, the ellipsoid for position 146 illustrates that the  $C^\alpha$  atoms in the green and red clusters have a much greater positional variation (the ellipsoids are bigger) compared with the cyan, purple and yellow clusters. The ellipsoids show the variation of  $C^\alpha$  coordinates of all structures belonging to each sub-master cluster. The sizes of the ellipsoids show that residue 146 of the

structures in the green and the red clusters varies a large amount, while the structures of the cyan, purple and yellow clusters do not change as much.



**Figure 7**  
Dot plot of the differences between the  $C^\alpha$  positions of each residue in the structures. To determine which regions of ChTg were most mobile in our data, we examined the five structures from the five intensity clusters and noted the distances among the  $C^\alpha$  atoms for each of the 146 amino acids. We plotted the largest value for each amino acid. The data illustrate one extended region with very large variation (between residues 146 and 151, in the vicinity of the missing loop that is a normal cleavage point for  $\alpha$ -chymotrypsin). There are also two shorter regions with smaller variation around Ser75 and Val200.



**Figure 8**  
Using ellipsoids to illustrate the variation in the  $C^\alpha$  coordinate at position 146. We calculated five ellipsoids for each residue position, corresponding to the observed variation in the  $C^\alpha$  positions at a specific residue for the data in the green, red, cyan, purple and yellow clusters. The lengths of the perpendicular axes were determined using the minimum-volume method (which minimizes the volume of the ellipsoid enclosing the data; see <https://github.com/nsls-ii-mx/chymotrypsinogen> and <https://raw.githubusercontent.com/nsls-ii-mx/chymotrypsinogen/master/ellipsoid.py>). This method optimizes the fit of each ellipse to the data, including the major axis in the direction of greatest variation. For example, at  $C^\alpha$  position 146 (shown here) the green cluster yielded 18 structures with large variation in the  $[0.2, -0.8, 0.0]$  direction. The volume of the ellipsoids indicates the overall variation in corresponding  $C^\alpha$  positions. For example, at position 146 the green and red clusters yielded structures with much larger positional variation than the cyan, purple and yellow clusters.

Table 1 shows that data sets which belong to the green and red clusters have  $a = b$  unit-cell parameters around 114–115 Å, while data sets in the other clusters have values around 111 Å (Fig. 4). Table 1 also reveals that data sets belonging to the

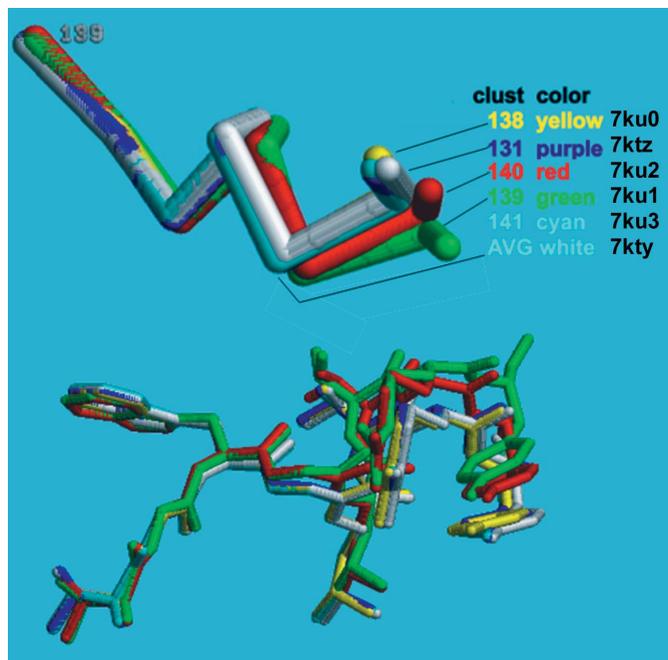
cyan, purple and yellow clusters have higher resolution than those belonging to the green and red clusters. The overall resolution of around 2 Å with good structure quality for each of the six structures is indicated by an  $R_{\text{work}}$  and  $R_{\text{free}}$  of about 20%.

When we align the model derived from the average cluster with the five major subclusters using *FATCAT* in rigid mode, all of the residues between 1 and 138 are well aligned, but residues 139–146 increasingly diverge (Fig. 9).

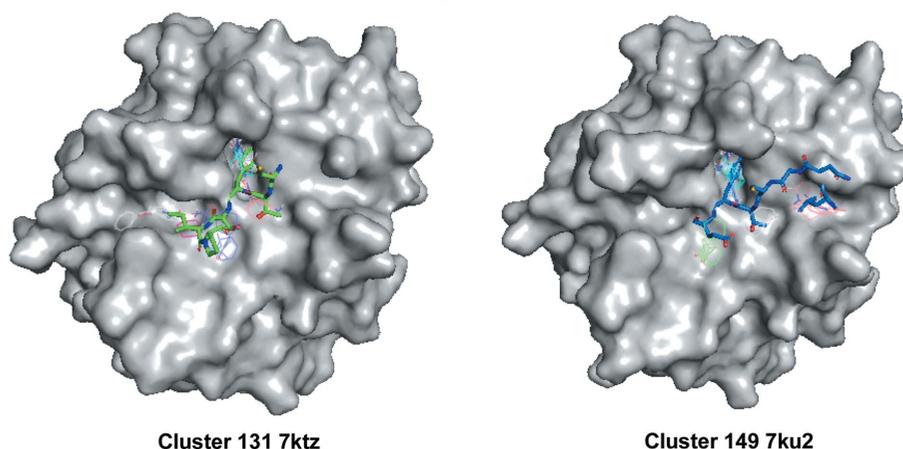
### 3.4. Detecting dynamic behavior via ligand-binding hotspots

*FTMap* shows six binding hotspots for each of the five structures (Fig. 10; Kozakov *et al.*, 2015). Among them, we observed the largest differences between the pockets in the structure of the red cluster (PDB entry 7ku2, cluster 140) and the structure of the purple cluster (PDB entry 7ktz, cluster 131). Notably, the pockets with the largest differences overlap with the binding site of the Bowman–Birk protease inhibitor. Since these two structures belong to the two different unit-cell-based clusters, the differences provide strong evidence for the effectiveness of both unit-cell-based and amplitude-based clustering in detecting polymorphs in the case of very small structural changes.

Note that the binding pocket for the Kazal-type inhibitor includes the missing Thr147→Asn150 loop (Tyr146 of ChTg makes two hydrogen bonds to the Kazal-type inhibitor: a direct hydrogen bond to Glu40 and a water-mediated hydrogen bond to Lys43). This would be a characteristic of the active enzyme chymotrypsin. The similarity between the results from data clustering and the results from computer modeling increase our confidence in both methods. We observed additional similarities between the two methodologies, which we are currently investigating.



**Figure 9**  
Structural alignment of residues 138–141 displaying the variation in position among the structures representing each cluster. The overall average structure, PDB entry 7kty, which is cluster 145 in the dendrogram, is colored white. PDB entries 7ktz, 7ku0, 7ku1, 7ku2 and 7ku3, which are clusters 131, 138, 139, 140 and 141, are colored purple, yellow, green, red and cyan, respectively. The top half shows the variation in the backbone alone. The bottom half shows the variation including the side chains. Remember that the green and red clusters are structures with  $a = 114$ – $115$  Å and all the others have  $a = 111$  Å.



**Figure 10**  
Surface representation of ChTg as calculated by *FTMap* comparing the hotspot areas of clusters 131 (PDB entry 7ktz) and 140 (PDB entry 7ku2). Left: *FTMap* surface representation of the ChTg structure of model 131 (PDB entry 7ktz, the purple cluster with  $a = 111.49$  Å) overlapped with a wire-frame rendering of critical parts of a Bowman–Birk plant-based protease-inhibitor complex of chymotrypsinogen (PDB entry 3ru4; Barbosa *et al.*, 2007). Right: an *FTMap* mapping result of model 140 (PDB entry 7ku2, the red cluster with  $a = 115.57$  Å, shown as a Lee–Richards surface) overlapped with a wire-frame rendering of critical parts of the pancreatic Kazal-type inhibitor (PDB entry 1cgi; Hecht *et al.*, 1991).

#### 4. Discussion

Although both experimental work (Debrunner & Frauenfelder, 1982) and theoretical work (McCammon, 1984) established that dynamic behavior underlies most protein functions, in the early years crystallography was not regarded as an appropriate tool for investigating protein dynamics. An early review of protein crystallography concluded by stating that ‘crystallographic methods are not suitable for the direct study of the dynamics of protein structure and interactions’ (Stryer, 1968).

However, the presence of diffuse scatter implied that there is dynamic behavior within protein crystals (Caspar *et al.*, 1988). Crystal structures soon illustrated examples of protein dynamics (Ringe & Petsko, 1985) that were induced by physical changes such as temperature (Tilton *et al.*, 1992), pH (Diao, 2003) and ionic strength (Sanishvili *et al.*, 1994) and induced by chemical changes by the addition of denaturants (Dunbar *et al.*, 1997) or ligands (Edwards & Poulos, 1990).

However, Stryer’s assertion stands to this day in the sense that investigators rarely employ simple tools to identify dynamics from diffraction data, and consequently most crystallographic contributions to dynamics continue to be fortuitous. Specifically, we propose here to provide experimenters with a nearly automatic procedure to suggest insights into the dynamics of proteins by a systematic surveying of diffraction data for the presence of clusters. Once crystallographers are equipped with appropriate tools to identify clusters within aggregates of diffraction data, results indicating dynamic behavior may emerge routinely in many protein crystallography projects.

A tool to identify dynamic contributions in diffraction data must be as automated as possible, must present results in a way that is easy to interpret and must be sensitive enough to identify small movements. The first of these requirements was simple to accommodate by deploying our software within the existing *KAMO* software package, which we easily integrated into our existing version of the *fast\_dp* automated data-analysis pipeline. The experimenter may include this test in the data-reduction pipeline with the flip of a switch, at a reasonably low cost in processing speed. We addressed the second requirement by incorporating visual tools such as systematic color annotation of clusters (Fig. 6), dot-plot visualization for structure variation (Fig. 7) and ellipsoid visualization for model variation (Fig. 8).

The most difficult benchmark was the ability to differentiate clusters where dynamic contributions are small and subtle. We tested our techniques using our data from ChTg, which was not known at the outset to exhibit dynamic behavior. Many of the changes that we identified involved just a few amino acids. By combining the strengths of unit-cell clustering (the ability to operate on thin wedges of data that are often incomplete) and of diffraction-based clustering (a sensitivity to very small structural changes), we believe that our technique will accurately identify relevant clusters of different structures hidden within highly similar data. Our procedure detected different polystates with coordinate differences of less than 3 Å in just

two amino acids. In addition, the visualization tools that we created (color-based ellipsoid and scatter plots) allow easy identification of the highly dynamic regions. This provides verification that our clusters are physically meaningful. These tools provide scientists with a simple procedure to screen their data for dynamic behaviors.

High-data-rate crystallography represents a large and growing fraction of all crystallographic data. At synchrotrons, serial crystallography and combinatorial crystallography (for example fragment screening) produce large streams of data from samples that are similar but not identical. One can cluster such data streams automatically, with visual results presented to scientists either to inform their main project or to yield serendipitous information that may expand their thinking about the system in question.

XFEL light sources generate even larger data streams, with individual diffraction images that are derived nearly instantaneously from very small protein crystals. The great reduction in the time and space averaging in XFEL data (compared with synchrotron data) further increases the likelihood of obtaining data from crystals that are in different resolvable polystates. We acknowledge that our software as it stands will not handle the partial data sets produced by the XFEL method. However, eventually the data-processing challenge will be the same: one needs a data-clustering algorithm that is robust enough to work with mixed-quality data, sensitive enough to partition all of the polystates that are present and intuitive enough that investigators can identify useful clusters that represent biologically relevant polystates. Here, we have presented an algorithm that accomplishes these goals.

Our data processing and clustering are automated to reduce the time spent screening and analyzing the molecules. We also perform manual checking to verify that the automated processes achieve reasonable fits to density. However, it is still a challenge if the data contain a lot of noise such as blurred or unindexable spots. This problem may be solved by future research on spot finding and auto-indexing. In addition, we would like to test whether different distance metrics could improve the accuracy of the clustering output and further improve the chances of detecting smaller potentially meaningful changes. We also will test whether the tools can detect polystates well in data sets from other molecules so that we can obtain a comprehensive understanding about the efficiency of our clustering procedure.

#### 5. Conclusions

Observing differences in protein structures, even small differences, could be meaningful and important. However, we usually miss changes that are very small since they are very hard to measure. In this paper, we show how one might use the combination of our unit-cell-based and structure-factor-based clustering procedure to detect polystates of molecules. We applied these methods to ChTg data and were able to detect polystates with very small differences among five clusters of data sets. From these clusters, we built molecular structures

and verified the differences among them. The combined procedure should help scientists to discover minor changes in molecules that are hardly noticeable from the change of unit-cell parameters only.

We have developed color-based visualization to assist investigators in screening their data for distinct groupings that may represent polystates: dendrograms to show correlations among intensities and scatter plots and ellipsoids to indicate differences in automatically refined structures. The dendrogram shows the members of clusters with custom height cutoffs and the differences among those clusters. The scatter plot quickly shows unit-cell-based clusters and their relations with structure factor-based clusters, and ellipsoids show the variations of the physical coordinates of structures of clusters. Using the color-based plots, one can easily discriminate among groups of data sets. This visualization method is a fast way to screen many data sets and to point out those that are important for further investigation.

### Acknowledgements

Data for this study were measured on beamlines 17-ID-1 (AMX) and 17-ID-2 (FMX) at the National Synchrotron Light Source II (NSLS-II), Brookhaven National Laboratory. Our thanks to Gregg Crichlow for careful and thoughtful review of both the PDB depositions and this paper. Our thanks to Frances C. Bernstein for many hours of copy-editing.

### Funding information

The following funding is acknowledged: National Institutes of Health/National Institute of General Medical Sciences (NIH/NIGMS; grant Nos. R21 GM129570-01 and P41 GM111244 to Brookhaven National Laboratory); US Department of Energy Biological and Environmental Research (grant No. KP 1605010 to Brookhaven National Laboratory); US Department of Energy Basic Energy Sciences [grant No. DE SC0012704 (KC 0401040) to Brookhaven National Laboratory]; US Department of Defense Congressionally Directed Medical Research Programs (grant No. BC151831 to SBG).

### References

- Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst.* **D68**, 352–367.
- Assmann, G., Brehm, W. & Diederichs, K. (2016). *J. Appl. Cryst.* **49**, 1021–1028.
- Authier, A. (2013). *Early Days of X-ray Crystallography*. Oxford University Press.
- Barbosa, J. A. R. G., Silva, L. P., Teles, R. C. L., Esteves, G. F., Azevedo, R. B., Ventura, M. M. & de Freitas, S. M. (2007). *Biophys. J.* **92**, 1638–1650.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
- Bernstein, H. J., Andrews, L. C., Diaz, J. A. Jr, Jakoncic, J., Nguyen, T., Sauter, N. K., Soares, A. S., Wei, J. Y., Wlodek, M. R. & Xerri, M. A. (2020). *Struct. Dyn.* **7**, 014302.
- Bernstein, H. J., Andrews, L. C., Foadi, J., Fuchs, M. R., Jakoncic, J., McSweeney, S., Schneider, D. K., Shi, W., Skinner, J., Soares, A. & Yamada, Y. (2017). *bioRxiv*, 141770.
- Caspar, D. L. D., Clarage, J., Salunke, D. M. & Clarage, M. (1988). *Nature*, **332**, 659–662.
- Debrunner, P. G. & Frauenfelder, H. (1982). *Annu. Rev. Phys. Chem.* **33**, 283–299.
- Diao, J. (2003). *Acta Cryst.* **D59**, 670–676.
- Dunbar, J., Yennawar, H. P., Banerjee, S., Luo, J. & Farber, G. K. (1997). *Protein Sci.* **6**, 1727–1733.
- Edwards, S. L. & Poulos, T. L. (1990). *J. Biol. Chem.* **265**, 2588–2595.
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* **D66**, 486–501.
- Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* **D69**, 1204–1214.
- Foadi, J., Aller, P., Alguel, Y., Cameron, A., Axford, D., Owen, R. L., Armour, W., Waterman, D. G., Iwata, S. & Evans, G. (2013). *Acta Cryst.* **D69**, 1617–1632.
- Gao, Y., Xu, W., Shi, W., Soares, A., Jakoncic, J., Myers, S., Martins, B., Skinner, J., Liu, Q., Bernstein, H., McSweeney, S., Nazaretski, E. & Fuchs, M. R. (2018). *J. Synchrotron Rad.* **25**, 1362–1370.
- Giordano, R., Leal, R. M. F., Bourenkov, G. P., McSweeney, S. & Popov, A. N. (2012). *Acta Cryst.* **D68**, 649–658.
- Hecht, H.-J., Szardenings, M., Collins, J. & Schomburg, D. (1991). *J. Mol. Biol.* **220**, 711–722.
- Kabsch, W. (2010). *Acta Cryst.* **D66**, 125–132.
- Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., Xia, B., Beglov, D. & Vajda, S. (2015). *Nat. Protoc.* **10**, 733–755.
- Kunitz, M. & Northrop, J. H. (1935). *J. Gen. Physiol.* **18**, 433–458.
- Liu, Q., Zhang, Z. & Hendrickson, W. A. (2011). *Acta Cryst.* **D67**, 45–59.
- McCammon, J. A. (1984). *Rep. Prog. Phys.* **47**, 1–46.
- McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* **D67**, 355–367.
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* **D53**, 240–255.
- Pjura, P. E., Lenhoff, A. M., Leonard, S. A. & Gittis, A. G. (2000). *J. Mol. Biol.* **300**, 235–239.
- Ringe, D. & Petsko, G. A. (1985). *Prog. Biophys. Mol. Biol.* **45**, 197–235.
- Rokach, L. & Maimon, O. (2005). *Data Mining and Knowledge Discovery Handbook*, edited by O. Maimon & L. Rokach, pp. 321–352. Boston: Springer.
- Rossmann, M. G. (2014). *IUCrJ*, **1**, 84–86.
- Sanishvili, R. G., Margoliash, E., Westbrook, M. L., Westbrook, E. M. & Volz, K. W. (1994). *Acta Cryst.* **D50**, 687–694.
- Sethi, R. (2015). US Patent 9208864.
- Siekevitz, P. & Palade, G. E. (1960). *J. Cell Biol.* **7**, 619–630.
- Steno, N. (1669). *De Solido Intra Solidum Naturaliter Contento Dissertationis Prodomus*. Florence: Accademia della Crusca.
- Strauss, T. & von Maltitz, M. J. (2017). *PLoS One*, **12**, e0168288.
- Stryer, L. (1968). *Annu. Rev. Biochem.* **37**, 25–50.
- Tilton, R. F. Jr, Dewan, J. C. & Petsko, G. A. (1992). *Biochemistry*, **31**, 2469–2481.
- Ward, J. H. Jr (1963). *J. Am. Stat. Assoc.* **58**, 236–244.
- Winter, G. & McAuley, K. E. (2011). *Methods*, **55**, 81–93.
- Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). *Acta Cryst.* **D74**, 85–97.
- Yamashita, K., Hirata, K. & Yamamoto, M. (2018). *Acta Cryst.* **D74**, 441–449.
- Ye, Y. & Godzik, A. (2004). *Nucleic Acids Res.* **32**, W582–W585.