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

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

[a] Department of Computer Science, Stony Brook University, Stony Brook, NY, 11794-2424, USA
[b] Department of Medicine, Oncology, Biophysics and Biophysical Chemistry, Johns Hopkins University, 725 N Wolfe St., Baltimore, MD, 21205, USA
[c] Department of Applied Mathematics & Statistics, Stony Brook University, Stony Brook, NY, 11794-3600, USA
[d] National Synchrotron Light Source II, Bldg 745, Brookhaven National Laboratory, P.O. Box 5000, Upton, NY, 11973-5000, USA
[e] Ronin Institute for Independent Scholarship, Care of NSLS-II, Bldg 745, Brookhaven National Laboratory, P.O. Box 5000, Upton, NY, 11973-5000, USA

Correspondence email: soares@bnl.gov; yayahjb@gmail.com

Funding information: NIGMS (grant No. 1R21GM129570-01 to BNL; grant No. P41GM111244 to BNL); DOE BER (grant No. KP1605010 to BNL); DOE BES (grant No. DE-SC0012704 (KC0401040) to BNL); DoD CDMRP (grant No. BC151831 to SBG).

Synopsis

The dynamics of proteins can be explored from polymorphs observed by clustering of multiple data wedges.

Abstract

One often observes small but measurable differences in 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 are measured, deliberately or inadvertently, the challenge is to sort the data into clusters that may represent relevant biological polystates. Here we address this problem 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 conformation space. We illustrate this method using two independent observables – unit cell constants and intensities – to cluster mixed-state data from chymotrypsinogen (ChTg) crystals. We observe that the data populate an arc of the reaction trajectory as ChTg is converted into chymotrypsin.

Keywords:

chymotrypsinogen; clustering; polymorphs; protein dynamics; changes unit cell

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 the molecules in the volume 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 that 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 somehow 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 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 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 use 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 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 measurement of many partial sets of crystal diffraction data from many, often very small, crystals (Liu, Hendrickson, 2011) (Giordano *et al*., 2012) (Rossmann, 2014) (Assman *et al*., 2016) (Bernstein *et al*., 2017) (Gao *et al*., 2018) (Bernstein *et al*., 2020). 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 one's detector operating at 200 Hz, a 360-degree rotational sweep with 0.2 degree per image will take approximately 9 seconds. These are standard experimental parameters at the FMX and AMX beamlines, NSLS-II, where one employs a beam size even smaller than small crystals. We could sample hundreds of tiny crystals (1 to 5 μm), each perhaps one polymorph, and then separate them into those polymorphs.

For this sort of treatment, one may mount crystals singly or, say, half a dozen in each sample mount (loop or mesh micro-mount); in each case, data are taken 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 a merging of 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 datasets are similarity in lattice-parameter values, or 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 datasets.

We demonstrate our approach using chymotrypsinogen (ChTg). ChTg is the precursor (zymogen) for 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 π-chymotrypsin, which is an active enzyme form. Secondly π-CHT molecules autolyze one another to cleave the bonds between Leu13 - Ser14 to release Ser14 – Arg15, and between Tyr146 - Thr147 and Asn148 - Ala149 to release Thr147 - Asn148. The resulting α-chymotrypsin is formed by three chains held by disulfide bridges.

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

We first employed differences in cell parameters as a conventional method for clustering single crystal data. However, this was clearly inadequate, and we extended this to partitioning on similarities of 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 those data. We found that clustering by correlation of intensities also revealed the large unit cell differences 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 use of correlation coefficients to follow.

Methods

Crystallization

We determined the crystallization conditions for chymotrypsinogen (SIGMA) using the commercial Hampton Crystal Screen HT (Hampton Research, Inc), and set up with a TTP mosquito robot (TTP Labtech/ STP Inc). 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 with hanging-drop vapor diffusion with a fixed pH of 6.5, varying both dioxane (10% or 15%) and ammonium sulfate (1.0 M-2.0 M). The drop, containing 1 µL of the reservoir solution (1.0 M-2.0 M ammonium sulfate, 0.1 M MES pH 6.5, and 10% or 15% dioxane) and 1 µL of 10.0 mg/mL of enzyme, equilibrated over 0.5 mL of reservoir solution. The other two crystallization conditions had either 0.2 M ammonium acetate, 0.1 M sodium acetate trihydrate pH 4.6, and 30% w/v PEG 4000, or 0.5 M ammonium sulfate, 0.1 M sodium citrate, and 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 of these crystals appears in Fig. 1.

Data collection and structure-solving strategies

We used ChTg crystals to obtain several hundred datasets at energy of 13.48 keV (0.92 Å), collecting 120° per crystal by employing Brookhaven National Laboratory's National Synchrotron Light Source II at

beamline 17-ID-1(AMX) on a Dectris EIGER X 9M detector. The dataset for each crystal with sufficient data 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 that supplies these modules: *XDS* (Kabsch, 2010), *DIALS* (Winter *et al*., 2018), *PHENIX* (Afonine *et al*., 2012), *aimless*, and *pointless* (Evans, Murshudov, 2013). Tetragonal crystals of chymotrypsinogen diffracted between 2.0 - 2.4 Å.

We determined the structure by molecular replacement with PHASER (McCoy *et al*., 2007), using ChTg PDB ID 1EX3 as a model (Bernstein *et al*., 1977) (Berman *et al*., 2000) (Pjura *et al*., 2000). The data were refined to a final resolution using iterative rounds of refinement with *REFMAC* (Murshudov *et al*., 1997) (Murshudov *et al*., 2011) and manual rebuilding in *COOT* (Emsley, Cowtan, 2004). We then used this model to build our "average" structure from an average of all data sets at the "native" energy. While scaling the native-energy datasets, we had noted clusters in the data. Perhaps they originated from polymorphs in the ChTg crystals, which may represent dynamic behavior in the molecules.

Data-clustering program

We used a custom-modified version of the clustering pipeline *KAMO* (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 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.

In our new scheme, in a single workflow, to obtain initial "coarse" clusters one partitions data sets into groups according to the similarity of their 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 in github.com/nsls-ii-mx/blend and github.com/nsls-ii-mx/yamtbx.)

The approach involving Pearson Correlation Coefficient (CC) calculations to determine similarity scores requires that pairs of data sets have many measured amplitudes in common: one requires a reasonably complete set of structure factors. For this CC clustering, one requires 70% completeness. One can introduce a penalty for unmatched structure factors, and can get a solution with completeness as low as 20 - 40% (Bernstein *et al*., 2017); we are studying the effect of even lower completeness to apply that method to partial data sets. We will show that this clustering approach demonstrates how increasingly sensitive clustering methods can identify increasingly detailed structural differences (Figs. 4 -- 8).

In our study, when we ran *KAMO.multi_merge* with Blend, it informed us that the datasets belong to different space groups. Hence, we chose the space group with the largest population; we aimed to display the changes of the space group's structures based on structure factors. We used *KAMO* to divide the datasets 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 done prior to refinement is *point group clustering*; for example, our algorithms clustered the cymotrypsinogen data using the exemplar space group 89 of its point group, rather than space group 92)

The distance between datasets was calculated by d(i,j)=√(1-CC(i,j)). The method then used a hierarchical clustering analysis (Rokach, Maimon, 2005) with Ward's method (Ward, 1963) to find distinct groups of the chosen dataset. In Ward clustering, the datasets are considered first by building a small cluster out of the two closest datasets and then by adding one dataset at a time to whichever dataset or existing cluster results in a new cluster of smallest variance. 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 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 *et al.* (Strauss, von Maltitz, 2017) discusses some alternative linkage choices. The program outputs a dendrogram that illustrates the distances (differences) among clusters by the y axis (the height). To get a certain number of clusters which contain more similar datasets, we chose a height cutoff value k accordingly. The lower the k value, the more similar the datasets in each cluster are. Each cluster now relates to a structure built after merging datasets within it.

We want to understand how cell-parameter values of datasets relate to the clusters determined by similarity in diffraction intensities. Since the space group is P 4_12_12 [89/92], the cell parameters **a** and **b** are equal, and all cell angles are 90°. Since there are only two free parameters, so we could visually demonstrate how the intensity clusters relate to the cell parameters **a** and **c** for each dataset (Fig. 4).

Finally, we created a molecular structure from the average of intensities from all of the crystals in each intensity cluster, and also averaged structures for each of the different cell-parameter clusters (see Figure 5, 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 for each of the clusters' structures. All the processes we used to build clusters' structures and to refine them later are automated with the help of *REFMAC (Murshudov et al., 1997) (Murshudov et al., 2011)*. Following the automated refinement steps, we performed a manual check-and-refine step using *COOT* (Emsley, Cowtan, 2004) to ensure 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.

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.

We developed a tool to create color-coded coordinate ellipses. We plotted the XYZ coordinates for the C_α 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_α 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 to the separations

among the centroids of the ellipsoids, indicating that each cluster represents a separable state. The code is available in the github.com/nsls-ii-mx/chymotrypsinogen.git git repository in the file raw.githubusercontent.com/nsls-ii-mx/chymotrypsinogen/master/ellipsoid.py. An example of use of this graphic appears in Figure 8.

We also plotted the **a** and **c** axis lengths for each dataset 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 clusters' structures, we used *FTMap* (Kozakov *et al*., 2015), software designed to determine and characterize ligand-binding hot spots on proteins' surfaces. 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.

Results

Data collection and protein structures

We collected 511 complete data sets and processed 325 of them by our data-reduction pipeline fast_dp_nsls2. Of these175 files had a resolution better than 4 Å. Finally, 146 datasets from space group 89, P422, were merged using BLEND cell-based cluster analysis. The protein is a single chain of 245 residues, of which four residues (147 – 150) are not resolved.

We obtained our initial structure, PDB ID 7KTY, from a merge of all 146 datasets and called this the Average Structure (denoted thus in Table 1). 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_{merge} value (48%) but nevertheless our PDB ID 7KTY was a good fit to these data (R_{work} 19%, R_{free} 20%). This average structure is slightly different from the published PDB ID 1EX3 structure which we used as an initial phasing model. For example, PDB ID 7KTY has a missing loop from residue 147 to residue 150, which is a characteristic of mature α-chymotrypsin. Fig. 2 displays the sequence alignment between PDB ID 1EX3 ChTg and our structure, PDB ID 7KTY, with the elements of the secondary structure drawn on top.

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 on two samples will always give different average structures. Such differences often are not relevant in terms of dynamics or states when the differences are small compared to experimental error. The only way to determine if candidate clusters may correspond to biologically relevant polystates is to generate and examine corresponding structural models (typically atomic models) with appropriate real-space tools, such as FATCAT and *COOT*. In the case of the ChTg data, we could see from inspection that the data could divide into two large clusters corresponding to structures with **a** ≈ 111 Å axis and those with **a** ≈ 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 cell lengths separate clearly into two groups, while the **c** cell length varies less and is not clearly separable. There were significant solvent-region differences between the **a** = 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 1EX3), while the 111 Å **a** cluster data indicates that there is a missing loop from residue 147 to residue 150 (this region is also not observed in the average structure). Another thing we observed is the presence of strong density near Lys 175 in the **a** = 111 Å cluster data while the **a** = 114 -- 115 Å cluster data does not have this large artifact (Fig. 5).

Using *Blend* and *KAMO*, we obtained 145 clusters from the 146 ChTg data sets. We then generated structures after merging datasets 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 the reproducible differences could be accounted for by 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).

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 the datasets of the green and red clusters belong to the **a** = 114 – 115 Å cell-based cluster and datasets of the cyan, purple, and yellow clusters belong to the **a** = 111 Å cell-based cluster. If we increase the cut height to 1.5, we get the two intensities' sub-master clusters, one containing green and red clusters, the other containing cyan, purple, and yellow clusters. This means the intensity cluster result has a strong alignment with the cell parameter cluster result.

The five data clusters

We generated a dendrogram using Ward's method for hierarchical clustering with the height cutoff at 1.0 to get distinct groups of datasets. To observe the differences between the 145 structures generated using individual data clusters, we calculated the largest differences in physical coordinates at each residue's C_α. We observed that the most mobile area, particularly residue 139 to residue 145, is near the missing loop from residue 146 to residue 152. Note that 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 (Fig. 8). The largest differences were observed for residue 146 with more than 3 Å average positional differences (Fig. 8).

At each residue position, we plotted ellipsoids to illustrate the variation in the C_α coordinates observed in each of the structures corresponding to the green, red, cyan, purple, and yellow clusters. For example, the ellipsoid for position 146 illustrates that the C_α atoms in the green and red clusters have a much greater positional variation (the ellipsoids are bigger) compared to the cyan, purple, and yellow clusters. The ellipsoids show the variation of C_α coordinates of all structures belonging to each sub-master cluster. The sizes of the ellipsoids show that residue 146 of the green and the red cluster's structures varies a lot while the structures of the cyan, purple, and yellow cluster's structures do not change as much.

Table 1 shows that datasets which belong to the green and red cluster have **a** = **b** unit cell values around 114 – 115 Å, while datasets of the other clusters have these values around 111 Å (Fig. 4). Table 1 also reflects the fact that datasets belonging to the cyan, purple, and yellow clusters have higher resolution than those of 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 R_{work} and R_{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 the residues between 1 and 138 are well-aligned, but residues 139 to 146 increasingly diverge (Fig. 9).

Detecting dynamic behavior via ligand-binding hot spots

FTMap shows six binding hotpots for each of the five structures (Fig. 10) (Kozakov *et al.,* 2015). Among them, we observed the largest differences between the pockets of the red cluster's structure (7KU2, cluster 140) and the purple cluster's structure (7KTZ, cluster 131). Notably, the pockets with largest differences overlap with the binding site of the Bowman-Birk protease inhibitor. Since these two structures belong to each of the two different cell-based clusters, the differences provide strong evidence for the effectiveness of both cell-based and amplitude-based clustering in detecting polymorphs in the case of very small physical changes.

Note that the binding pocket for the Kazal-type inhibitor includes the Thr147→ Asn150 missing loop (ChTg Tyr146 makes two hydrogen bonds with 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 modelling increase our confidence in both methods. We observed additional similarities between the two methodologies, which we are currently investigating.

Discussion

Although both experimental work (Debrunner *et al*., 1982) and theoretical work (McCammon, 1984) established that dynamic behavior underlies most protein functions, crystallography was not regarded in the early years 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 *et al*., 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 *et al*., 1990)).

However, Stryer's 1968 assertion stands to this day in the sense that investigators rarely employ simple tools to identify dynamics from diffraction data, consequently most crystallographic contributions to dynamics continue to be fortuitous. We propose here a method 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 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 variations (Fig. 8)w.

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 (ability to operate on thin wedges of data that are often incomplete) and the strengths of diffraction-based clustering (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 method detected different polystates with coordinate differences less than 3Å 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 a simple method 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 (e.g. 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 of 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 to 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 the polystates that are present, and intuitive enough that investigators can identify useful clusters that represent biologically relevant polystates. Here, we presented an algorithm that accomplishes these goals.

Our data processing and clustering are all automatic to reduce the time of screening and analyzing the molecules. We also do manual checking to verify that the automated processes achieved reasonable fits to density. However, it is still a challenge for us if the data contain a lot of noise such as blurs or unindexable spots. This problem may be solved by future research on spot finding and auto-indexing. In addition, we would like to test if 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 if the tools could detect polystates well with datasets from other molecules so that we would have a comprehensive understanding about the efficiency of our clustering method.

Conclusions

Observing differences in protein structures, even small differences, could be meaningful and important. However, we usually miss the 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 cell-based and structure-factor-based clustering methods to detect polystates of molecules. We applied these methods on ChTg data and were able to detect polystates with very small differences among five clusters of datasets. From these clusters, we built molecular structures and verified the differences among them. The combined method should help scientists to discover minor changes in molecules that are hardly noticeable by the change of 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 cell-based clusters and their relations with structure factor-based clusters, and ellipsoids show the variations of physical coordinates of clusters' structures. Using the color-based plots, one could easily discriminate among groups of datasets. This visualization method is a fast way to screen many datasets, and to point out which ones are important for further investigation.

Figure 1. Representative ChTg crystals from the crystallization condition containing 1.0 – 2.0 M ammonium sulfate, 0.1 M MES pH 6.5, and 10% or 15% dioxane.

Figure 2. Sequence alignment of ChTg PDB ID 1EX3 and the average structure PDB ID 7KTY. The loop residues 147-150 do not display electron density in PDB ID 7KTY.

Figure 3. Cross-eyed stereo pair of the structural alignment of PDB ID 1EX3 (dark grey) and the average structure PDB ID 7KTY (light grey). The FATCAT chain RMSD is 0.56 Å. The regions with significant differences are adjacent and appear at the upper left of this figure. First, in the average structure the amino acids between Thr147 and Asn150 are missing. Second, 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 (**a** = 111 Å group and **a** = 114 – 115 Å group).

We observed that our data partitioned cleanly between 28 data sets with an **a** (=**b**) unit cell of approximately 114 -- 115 Å and 118 data sets with an a (=b) unit cell 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 those two clusters into amplitude-based clusters is shown by the colors on the right. The **a** = 114 -- 115 Å cell-based cluster contained the green and red clusters, and the **a** = 111 Å cell-based cluster contained 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 a first use of cell-based clustering might be used to boot-strap amplitude-based clustering.

Figure 5. Differences in solvent between the **a** = 111 Å cluster and the **a** = 114 – 115 Å cluster.

Fo-Fc electron difference density displaying two differences we observed in solvent density between the **a** = 111 Å cluster and the **a** = 114 -- 115 Å cluster (difference densities at 2-sigma shown in green for both data sets). Left: Ribbon diagram of ChTg around Lys175 (cyan) for the cluster **a** = 111 Å. Right: Ribbon diagram of ChTg around Lys175 (cyan) for the cluster **a** = 114 Å. This density was modeled as a water molecule.

Figure 6. Amplitude-based clusters generated using *KAMO* (dendrogram).

This dendrogram shows a representation of the similarity of pairs of data sets, and clusters of more data sets. They are arranged with the most similar ones 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 overall average structure as PDB ID 7KTY. We selected a height within the dendrogram at which to partition our data. We made our final choice to use five clusters through inspection of the derived structures. We then averaged all structure factors within each of the five distinct clusters to give a cluster-average structures. We deposited the averaged structure from the green clusters as PDB ID 7KU1, the red clusters as PDB ID 7KU2, the cyan clusters as PDB ID 7KU3, the purple clusters as PDB ID 7KTZ, and the yellow clusters as PDB ID 7KU0. Note that by "average structures", we mean structures derived from structure-factor averages.

Figure 7. Dot plot of differences between C_α 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 five intensity clusters, and noted the distances among the C_α carbons 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 α-chymotrypsin). There are also two shorter regions with smaller variation around Ser75 and Val200.

Figure 8. Using ellipsoids to illustrate variation in the C_α coordinate at position 146.

We calculated five ellipsoids for each residue position, corresponding to the observed variation in the C_α positions at a specific residue for the green, red, cyan, purple, and yellow clusters' data. 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_α 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_α 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.

Figure 9. Structural alignment of residues 138 -- 141 displaying the variation in position among the structures representing each cluster.

The overall average structure, 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 with the side chains. Remember that green and red come from structures with **a** = 114 – 115 Å and all the others have **a** = 111 Å.

Figure 10. Surface representation of the ChTg as calculated by *FTMap* comparing the hot spot areas of clusters 131 (7KTZ) and 140 (7KU2).

Left *FTMap* surface representation of the ChTg structure of model 131 (7KTZ, the purple cluster with **a** = 111.49 Å,) overlapped with a wire-frame rendering of critical parts of Bowman-Birk protease inhibitor complex with chymotrypsinogen, PDB ID 3RU4 (Barbosa *et al.*, 2007).

Right is an *FTMap* mapping result of model 140 (7KU2, the red cluster with **a** = 115.57 Â, as a Lee-Richards surface) overlapped with a wire-frame rendering of critical parts of Kazal Type inhibitor, PDB ID 1CGI (Hecht *et al*., 1991).

Table 1. Data collection and processing

PDB ID	7KTY	7KU1	7KU2	7KU3	7KTZ	7KU0
Description	Average	Green	Red	Cyan	Purple	Yellow
Cluster #	145	139	140	141	131	138
No of datasets.	146	12	16	32	37	49
Wavel. (Å)	0.9201	0.9201	0.9201	0.9201	0.9201	0.9201
Temp. (K)	100	100	100	100	100	100
Detector	E9M	E9M	E9M	E9M	E9M	E9M
Dist. (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	P 4_12_12	P 4_12_12	P 4_12_12	P 4_12_12	P 4_12_12	P 4_12_12
a, b (Å)	114.49	114.49	115.57	111.33	111.49	111.47
c (Å)	51.90	51.9	52.92	51.87	52.02	52.36
α, β, γ (°)	90	90	90	90	90	90
Resolution (Å)	2.00	2.39	2.19	2.00	2.00	2.02
Reflections #	23,850	14,139	18,927	22,533	22,696	22,261
Complet. (%)	99.94%	99.80%	98.98%	99.81%	99.93%	99.59%
I/σ(I)	10.91	9.48	9.76	10.66	12.15	10.85
Wilson B ($Å^2$)	53.96	61.87	53.31	33.72	29.15	30.16

Table 2. Structure solution and refinement

PDB ID 7KTY	7KU1	7KU2	7KU3	7KTZ	7KU0	
Description	Average	Green	Red	Cyan	Purple	Yellow
Final R_{work} (%)	19.13	22.08	20.97	18.15	16.18	16.74
Final R_{free} (%)	20.19	26.37	23.42	21.01	19.01	19.41
No. of non-H atoms						
Protein	1,786	1,771	1,778	1,794	1,786	1,786
Ligand	15	5	5	15	15	15
Water	99	10	44	195	249	243
Total	1,900	1,786	1,827	2,004	2,050	2,044
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 B factors (Å^2)						
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

Acknowledgments

Data for this study were measured at beamlines 17-ID-1 (AMX) and 17-ID-2 (FMX) at Brookhaven National Laboratory's National Synchrotron Light Source II (NSLS-II).

Our thanks to Gregg Crichlow for careful and thoughtful review of both the PDB depositions and of this paper.

Our thanks to Frances C. Bernstein for many hours of copy-editing.

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). Towards automated crystallographic structure refinement with phenix.refine. Acta Cryst. D68:4 352 - 367.

Aller, P., Geng, T., Evans, G., Foadi, J. (2016) Applications of the BLEND Software to Crystallographic Data from Membrane Proteins. In: Moraes I. (Ed.) The Next Generation in Membrane Protein Structure Determination. Adv. Exp. Med. Biol. 922.

Assmann, G., Brehm, W., Diederichs, K. (2016). Identification of rogue datasets in serial crystallography. J. Appl. Cryst. 49:3 1021 - 1028.

Authier, A. (2013). Early days of X-ray crystallography. OUP, Oxford.

Barbosa, J. A. R., Silva, L. P., Teles, R. C., Esteves, G. F., Azevedo, R. B., Ventura, M. M., de Freitas, S. M. (2007). Crystal structure of the Bowman-Birk inhibitor from Vigna unguiculata seeds in complex with β-Trypsin at 1.55 Å resolution and its structural properties in association with proteinases. Biophys. J., 92:5 1638 - 1650.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000). The Protein Data Bank. Nucl. Acids Res. 28:1 235 - 242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr, E. F., Brice, M .D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:3 535 - 542.

Bernstein, H. J. (2000). Recent changes to RasMol, recombining the variants. TiBS 25:9 453 - 455.

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). Serial Crystallography with Multi-stage Merging of 1000's of Images. BioRxiv 141770.

Bernstein, H. J., Andrews, L. C., Diaz Jr, J. A., Jakoncic, J., Nguyen, T., Sauter, N. K., Soares, A. S., Wei, J. Y., Wlodek, M. R., Xerri, M. A. (2020). Best practices for high data-rate macromolecular crystallography (HDRMX). Struct. Dynamics, 7:1.

Caspar, D. L. D., Clarage, J., Salunke, D. M., Clarage, M. (1988). Liquid-like movements in crystalline insulin. Nature 332:6165 659 - 662.

Debrunner, P. G., Frauenfelder, H. (1982). Dynamics of proteins. Annu. Rev. Phys. Chem. 33:1 283 - 299.

DeLano, W. L. (2002). Pymol: An open-source molecular graphics tool. CCP4 Newsletter 40:1 82 - 92.

Diao, J. (2003). Crystallographic titration of cubic insulin crystals: pH affects GluB13 switching and sulfate binding. Acta Cryst., D59:4 670 - 676.

Dunbar, J., Yennawar, H. P., Banerjee, S., Luo, J., Farber, G. K. (1997). The effect of denaturants on protein structure. Protein Sci. 6:8 1727 - 1733.

Edwards, S. L., Poulos, T. L. (1990). Ligand binding and structural perturbations in cytochrome c peroxidase. A crystallographic study. J. Biol. Chem. 265:5 2588 - 2595.

Emsley, P., Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Cryst. D60:12 2126 - 2132.

Evans, P. R., Murshudov, G. N. (2013). How good are my data and what is the resolution? Acta Cryst. D69:7 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). Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography. Acta Cryst. D69:8 1617 - 1632.

Frauenfelder, H., Chen, G., Berendzen, J., Fenimore, P. W., Jansson, H., McMahon, B. H., Stroe, I. R., Swenson, J., Young, R. D. (2009). A unified model of protein dynamics. Proc. Nat. Acad. Sci. USA 106:13 5129 - 5134.

Gao, Y., Xu, W., Shi, W., Soares, A., Jakoncic, J., Myers, S., Martins, B., Skinner, J., Liu, Q., Bernstein, H., McSweeney, S. (2018). High-speed raster-scanning synchrotron serial microcrystallography with a high-precision piezo-scanner. J. Synch. Rad. 25:5 1362 - 1370.

Giordano, R., Leal, R. M., Bourenkov, G. P., McSweeney, S., Popov, A. N. (2012). The application of hierarchical cluster analysis to the selection of isomorphous crystals. Acta Cryst. D68:6 649 - 658.

Hecht, H. J., Szardenings, M., Collins, J., Schomburg, D. (1991). Three-dimensional structure of the complexes between bovine chymotrypsinogen A and two recombinant variants of human pancreatic secretory trypsin inhibitor (Kazal-type). J. Mol. Biol. 220:3 711 - 722.

Ho, B. K., Agard, D. A. (2009). Probing the flexibility of large conformational changes in protein structures through local perturbations. PLoS Comput Biol 5:4 e1000343.

Kabsch, W. (2010). XDS. Acta Cryst. D66, 125 - 132.

Kostov, K. S., Moffat, K. (2011). Cluster analysis of time-dependent crystallographic data: Direct identification of time-independent structural intermediates. Biophys. J. 100:2 440 - 449.

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., Xia, B., Beglov, D., Vajda, S. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. Nature protocols, 10(5), 733 - 755

Kunitz, M., Northrop, J. H. (1935). Crystalline chymo-trypsin and chymo-trypsinogen: I. Isolation, crystallization, and general properties of a new proteolytic enzyme and its precursor. J. Gen. Physiol. 18:4 433 - 458.

Liu, Q., Zhang, Z., Hendrickson, W. A. (2011). Multi-crystal anomalous diffraction for low-resolution macromolecular phasing. Acta Cryst. D67:1 45 - 59.

McCammon, J. A. (1984). Protein dynamics. Rep. Prog. Phys. 47:1 1 - 46.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., Read, R. J. (2007). Phaser crystallographic software. J. Appl. Cryst. 40:4 658 - 674.

Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. Acta Cryst. D67:4 355 - 367.

Murshudov, G. N., Vagin, A. A., Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst. D53:3 240 - 255.

Pjura, P. E., Lenhoff, A. M., Leonard, S. A., Gittis, A. G. (2000). Protein crystallization by design: chymotrypsinogen without precipitants. J. Mol. Biol. 300:2 235 - 239.

Ringe, D., Petsko, G. A. (1985). Mapping protein dynamics by X-ray diffraction. Progr. Biophy. Mol. Biol. 45:3 197 - 235.

Rokach, L., Maimon, O. (2005). Clustering methods. In Rokach, L., Maimon, O. Eds., Data mining and knowledge discovery handbook, (pp. 321 - 352). Springer, Boston, MA.

Rossmann, M. G. (2014). Serial crystallography using synchrotron radiation. IUCrJ 1:2 84 - 86.

Sanishvili, R. G., Margoliash, E., Westbrook, M. L., Westbrook, E. M., Volz, K. W. (1994). Crystallization of wild-type and mutant ferricytochromes c at low ionic strength: seeding technique and X-ray diffraction analysis. Acta Cryst. D50:5 687 - 694.

Santoni, G., Zander, U., Mueller-Dieckmann, C., Leonard, G., Popov, A. (2017). Hierarchical clustering for multiple-crystal macromolecular crystallography experiments: the ccCluster program. J. Appl. Cryst. 50:6 1844 - 1851.

Sayle, R. A. and Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. TiBS 20:9 374 - 376.

Schnell, J. R., Dyson, H. J., Wright, P. E. (2004). Structure, dynamics, and catalytic function of dihydrofolate reductase. Annu. Rev. Biophys. Biomol. Struct., 33 119 - 140.

Sethi, R., Toshiba America Electronic Components Inc. (2015). Semiconductor memory with integrated biologic element. U.S. Patent 9,208,864.

Siekevitz, P., Palade, G. E. (1960). A Cytochemical Study on the Pancreas of the Guinea Pig: V. In vivo Incorporation of Leucine-1-C14 into the Chymotrypsinogen of Various Cell Fractions. J. Cell Biol. 7:4 619 - 630.

Steno, N (1669). De solido intra solidum naturaliter contento dissertationis prodromus, Stellae, Florence, Maar, Vol II, 181 - 227, No. XXVII.

Strauss, T., von Maltitz, M. J. (2017). Generalising Ward's method for use with Manhattan distances. PloS One 12:1 e0168288.

Stryer, L. (1968). Implications of X-ray crystallographic studies of protein structure. Annu. Rev. Biochem. 37:1 25 - 50.

Tilton Jr, R. F., Dewan, J. C., Petsko, G. A. (1992). Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320K. Biochem. 31:9 2469 - 2481.

Walsh, K. A., Neurath, H. (1964). Trypsinogen and chymotrypsinogen as homologous proteins. Proc. Nat. Acad. Sci. USA 52:4 884 - 889.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. J. Amer. Stat. Assoc. 58:301 236 - 244.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. Acta Cryst. D67: 4 235 – 242.

Winter, G., McAuley, K. E. (2011). Automated data collection for macromolecular crystallography. Methods, 55:1 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). DIALS: implementation and evaluation of a new integration package. Acta Cryst.. D74:2 85 - 97.

Yamashita, K., Hirata, K., Yamamoto, M. (2018). KAMO: towards automated data processing for microcrystals. Acta Cryst., D74:5 441 - 449.

Yang, L. Q., Sang, P., Tao, Y., Fu, Y. X., Zhang, K. Q., Xie, Y. H., Liu, S. Q. (2014). Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms. J. Bio. Struc. Dynam. 32:3 372 - 393.

Ye, Y., & Godzik, A. (2004). FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic acids research*, *32*(Web Server issue), W582–W585.

Yonetani, T., Laberge, M. (2008). Protein dynamics explain the allosteric behaviors of hemoglobin. Biochim. Biophys. Acta (BBA)-Proteins and Proteomics 1784:9 1146 - 1158.

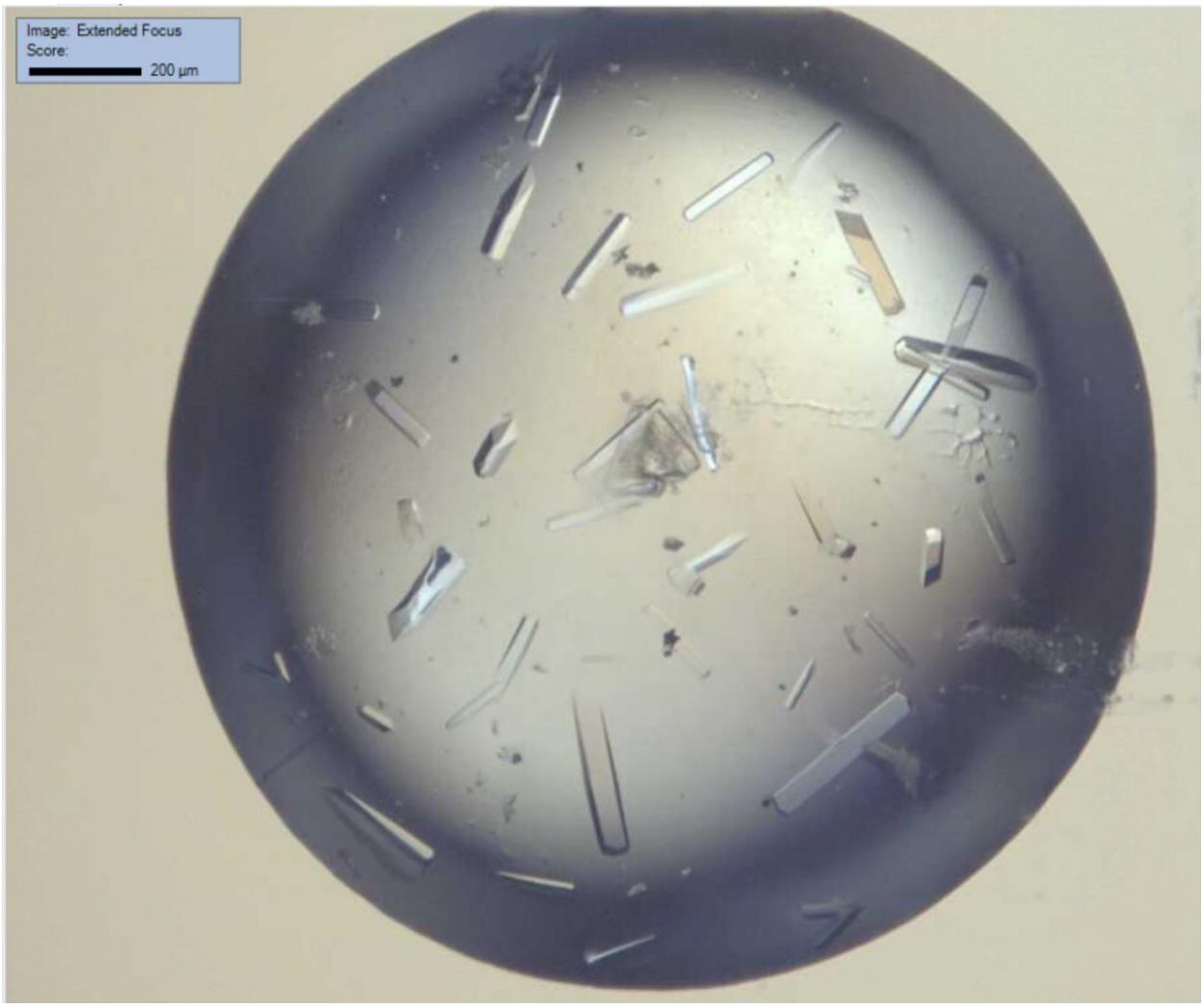


Figure 1

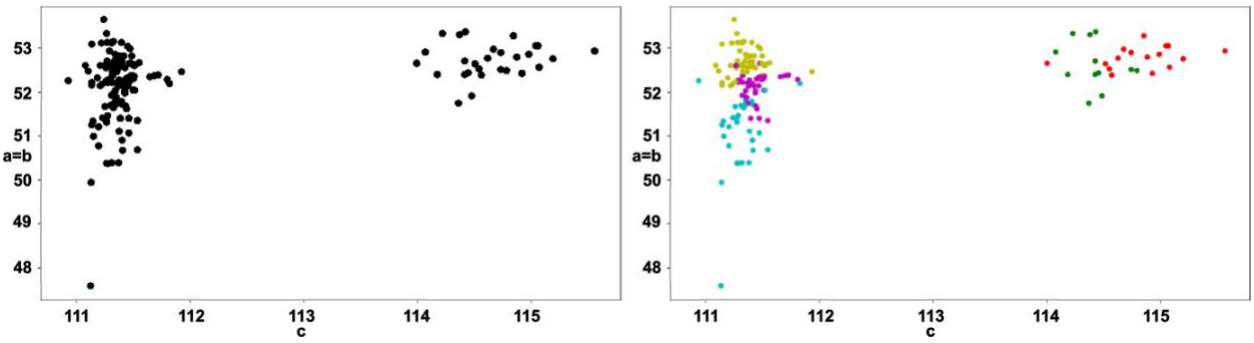


Figure 2

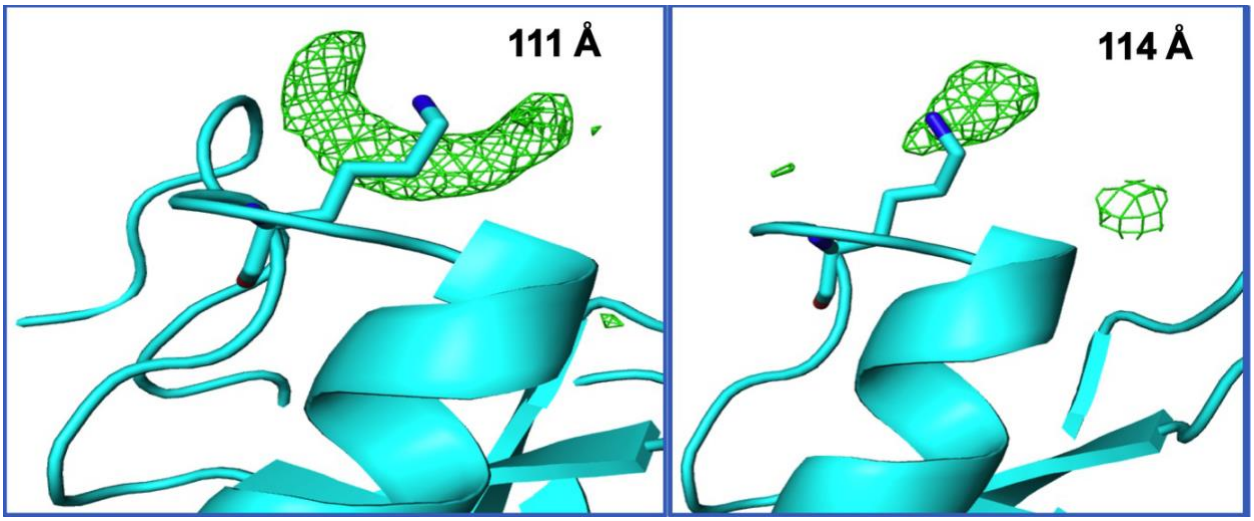


Figure 3

Figure 4

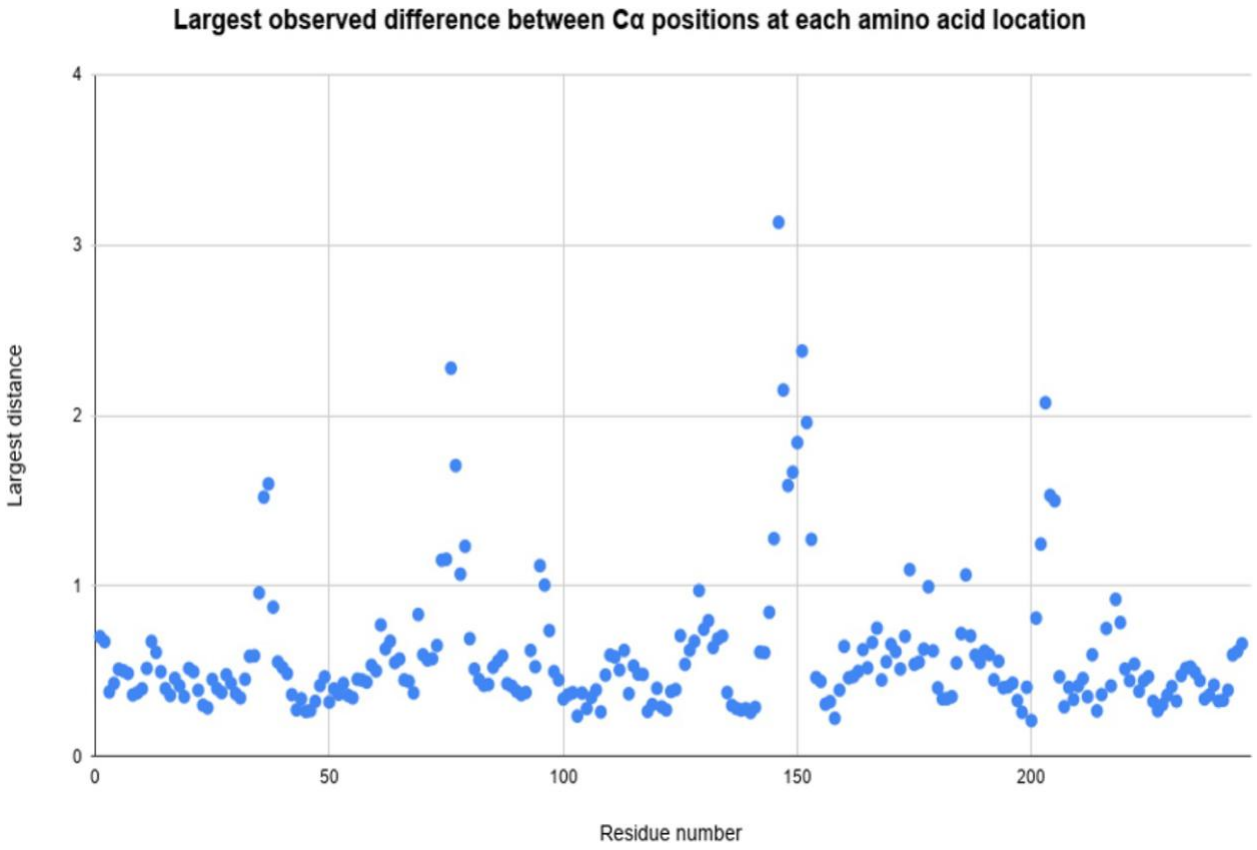

Largest observed difference between Cα positions at each amino acid location
Largest distance
Residue number
0
1
2
3
4
0
50
100
150
200

Figure 5

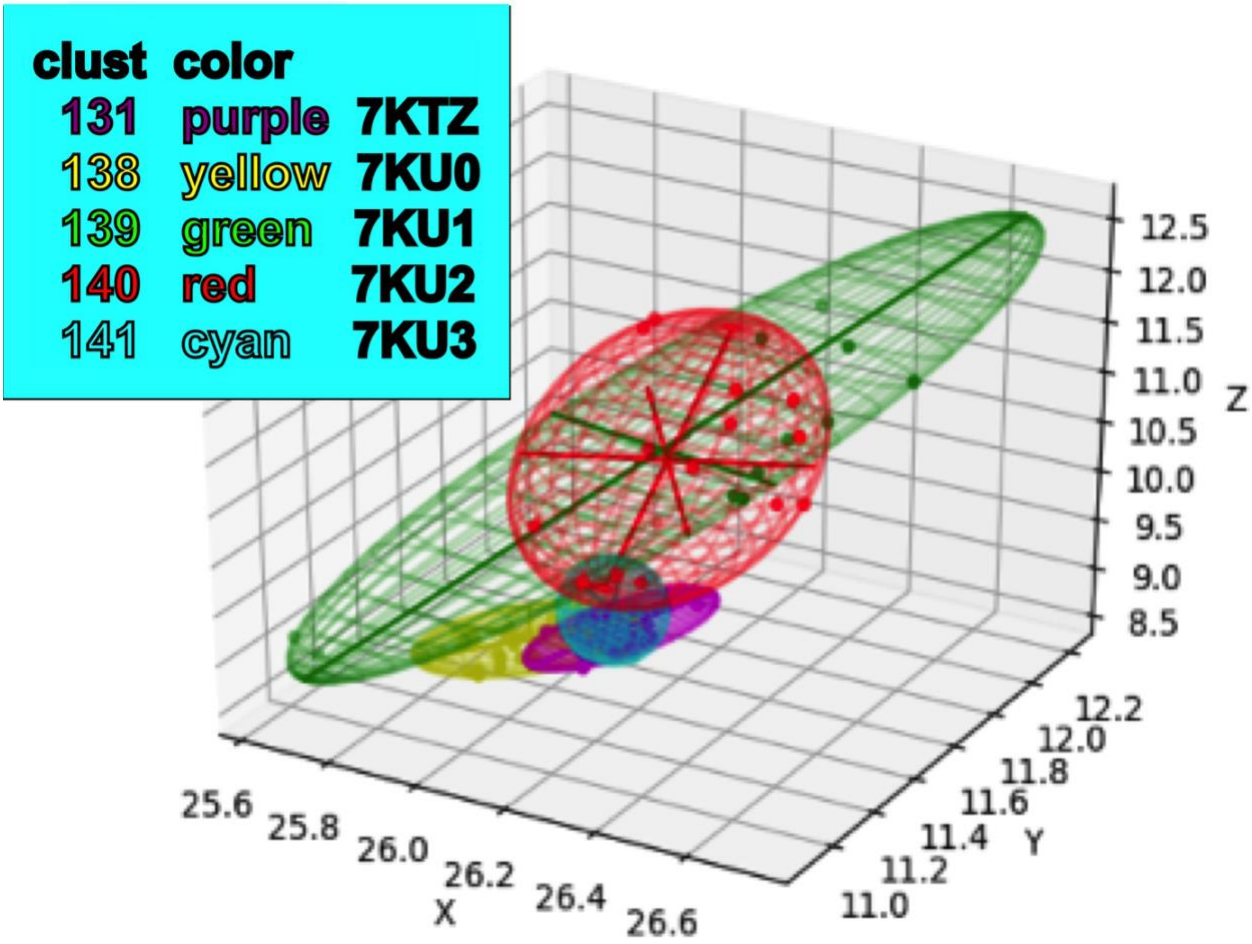
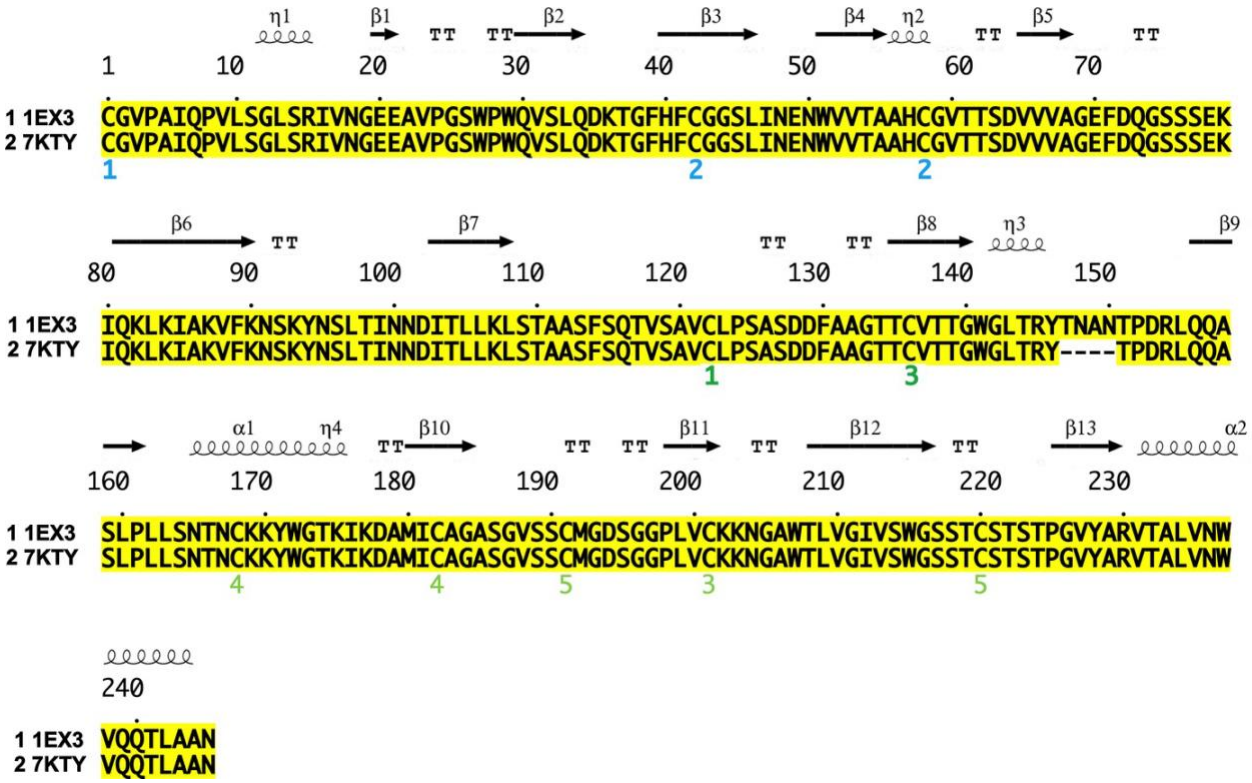

clust color
131 purple 7KTZ
138 yellow 7KU0
139 green 7KU1
140 red 7KU2
141 cyan 7KU3
12.5
12.0
11.5
11.0
Z
10.5
10.0
9.5
9.0
8.5
12.2
12.0
11.8
11.6
11.4
Y
11.2
11.0
25.6
25.8
26.0
26.2
X
26.4
26.6

Figure 6


η1 β1 TT TT β2 β3 β4 η2 TT β5 TT
1 10 20 30 40 50 60 70
1 1EX3 CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVTAAHCGVTTSDVVVAGEFDQGSSSEK
2 7KTY CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQDKTGFHFCGGSLINENWVVTAAHCGVTTSDVVVAGEFDQGSSSEK
1 2 2
β6 TT β7 TT TT β8 η3 β9
80 90 100 110 120 130 140 150
1 1EX3 IQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTCVTTGWGLTRYTNANTPDRLQQA
2 7KTY IQKLKIAKVFKNSKYNSLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTCVTTGWGLTRY----TPDRLQQA
1 3
α1 η4 β10 TT TT β11 TT β12 TT β13 α2
160 170 180 190 200 210 220 230
1 1EX3 SLPLLSNTNCKKYWGTKIKDAMICAGASGVSSCMGDSGGPLVCKKNGAWTLVGIVSWGSSTCSTSTPGVYARVTALVNW
2 7KTY SLPLLSNTNCKKYWGTKIKDAMICAGASGVSSCMGDSGGPLVCKKNGAWTLVGIVSWGSSTCSTSTPGVYARVTALVNW
4 4 5 3 5
240
1 1EX3 VQQTLAAN
2 7KTY VQQTLAAN
Figure 7

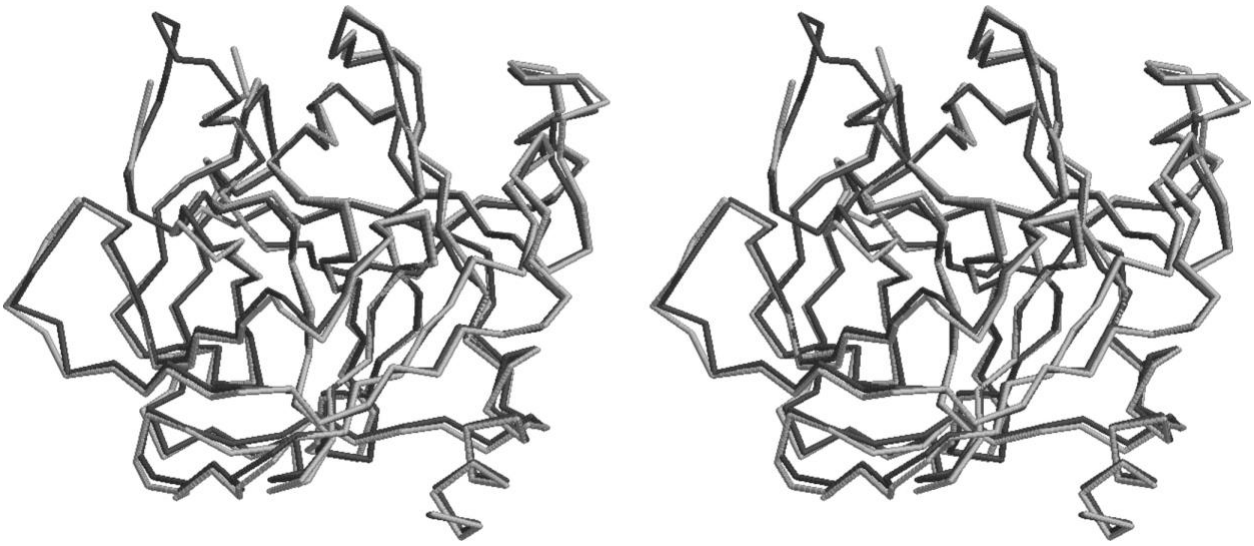


Figure 8

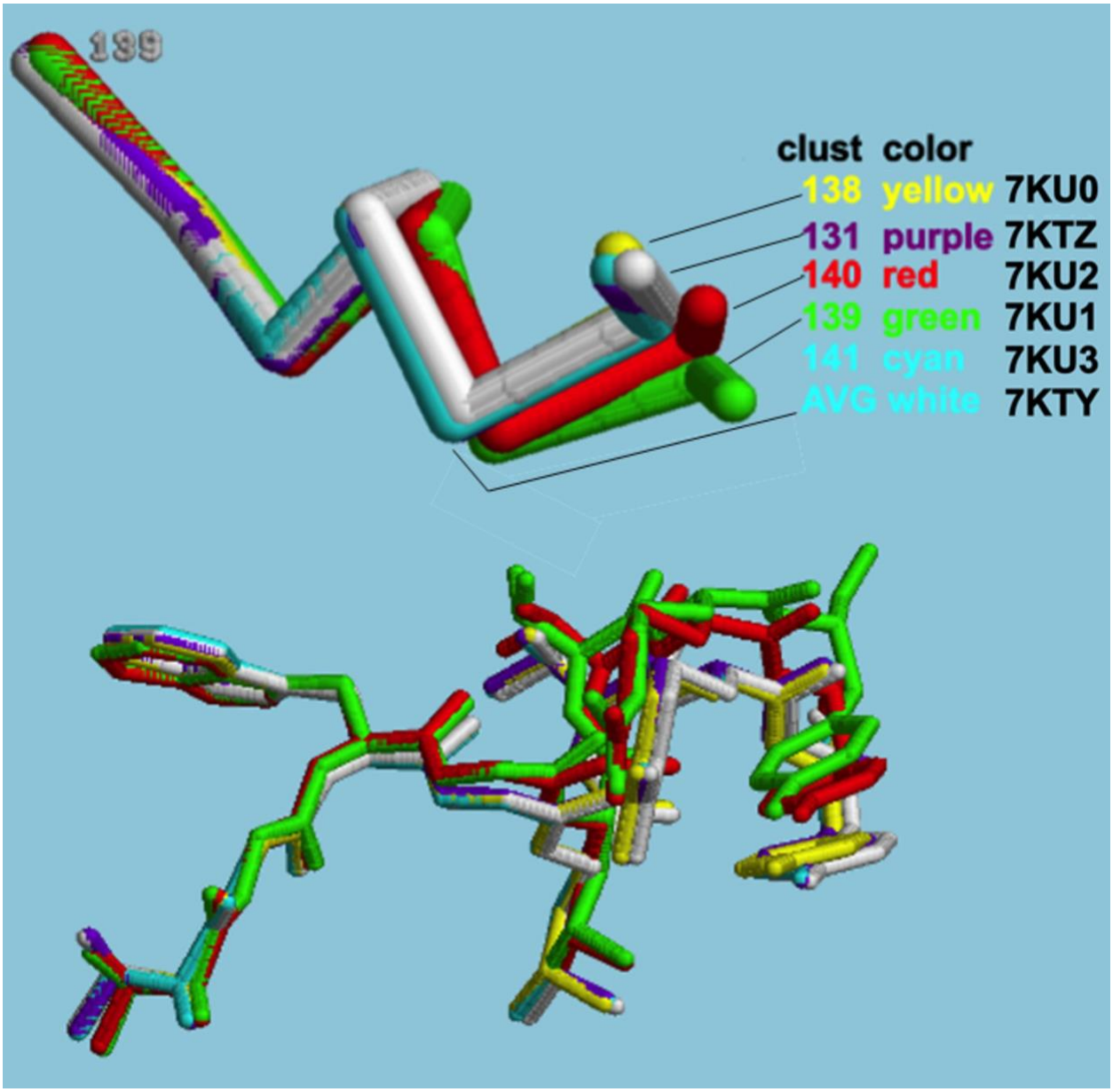

139
clust color
138 yellow 7KU0
131 purple 7KTZ
140 red 7KU2
139 green 7KU1
141 cyan 7KU3
AVG white 7KTY

Figure 9

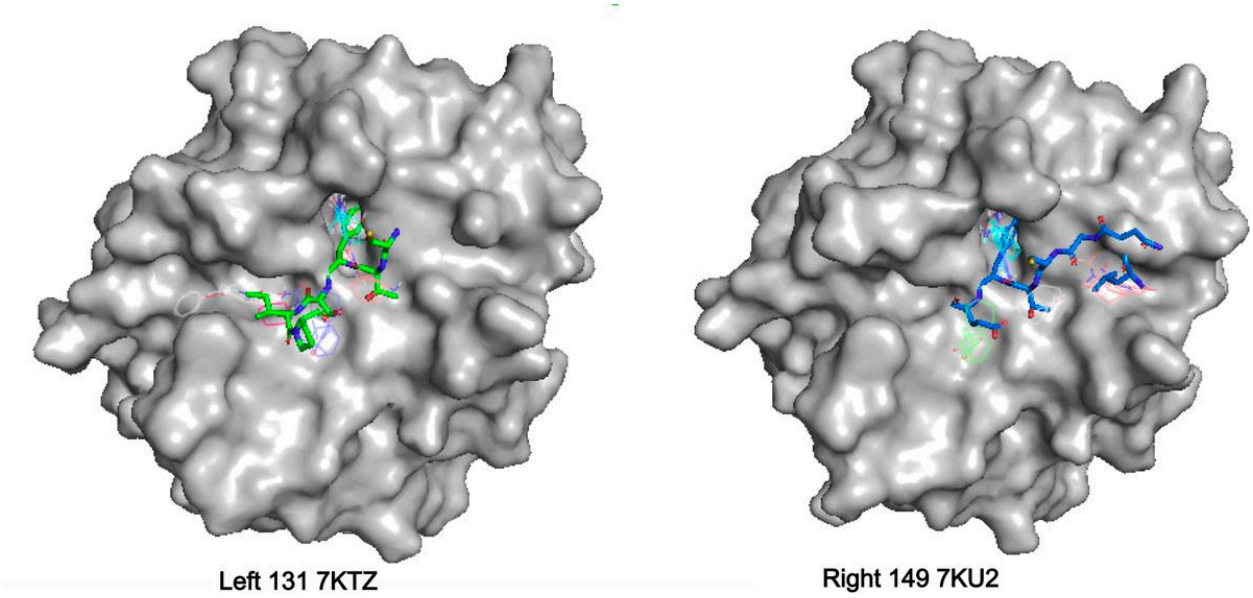


Figure 10